A Data-driven Statistical Approach to Customer Behaviour Analysis and Modelling in Online Freemium Games

Anusua Singh Roy

A thesis submitted in partial fulfilment of the requirements of Edinburgh Napier University, for the award of Doctor of Philosophy

December 2018

Declaration

I hereby declare that this work has not been submitted for any other degree or professional qualification, and the thesis is the result of my own independent work.

Anusua Singh Roy

December 2018

Acknowledgement

I am extremely grateful to my supervisors Professor Robert Raeside and Dr Moira Hughes and my colleague Dr Tao Chen for their help and support in the completion of this thesis. I would especially like to thank Robert for offering me this PhD opportunity, and for his continuous motivation throughout the course of my doctoral studies and assistance in coping with the stress of balancing a full time job with the write up of this thesis. I would also like to thank the examiners for their constructive feedback and suggestions. Lastly and most importantly, I am forever indebted to my family and friends for providing constant support, encouragement and entertainment during the challenging times I faced in my PhD journey.

Thank you all – without you this would not have been possible.

Abstract

The video games industry is one of the most attractive and lucrative segments in the entertainment and digital media, with big business of more than $150 billion worldwide. A popular approach in this industry is the online freemium model, wherein the game is downloadable free of cost, while advanced and bonus content have optional charges. Monetisation is through micro payments by customers and the focus is on maintaining average revenue per user and lifetime value of players. The overall aim of this research is to develop suitable data-driven methods to gain insight about customer behaviour in online freemium games, with a view to providing recommendations for successful business in this industry.

Three important aspects of user behaviour are modelled in this research - engagement, time until defection, and number of micro transactions made. A multiple logistic regression using penalised likelihood approach is found to be most suitable for modelling and demonstrates good fit and accuracy for assigning observations to engaged and non-engaged categories. Cox's proportional hazards model is adopted to analyse time to defection, and a negative binomial zero-inflated model results in the best fit to the data on micro payments. Cluster analysis techniques are used to classify the wide variety of customers based on their gameplay styles, and social network models are developed to identify prominent 'actors' based on social interactions. Some of the significant predictors of engagement and monetisation are amount of premium in-game currency, success in missions and competency in virtual fights, and quantity of virtual resources used in the game.

This research offers extensive insight into what drives the reputation, virality and commercial viability of freemium games. In particular it helps to fill a gap in understanding the behaviour of online game players by demonstrating the effectiveness of applying a data analytic approach. It gives more insight into the determinants of player behaviour than relying on observational studies or those based on survey research. Additionally, it refines statistical models and demonstrates their implementation in R to new and complex data types representing online customer behaviours.

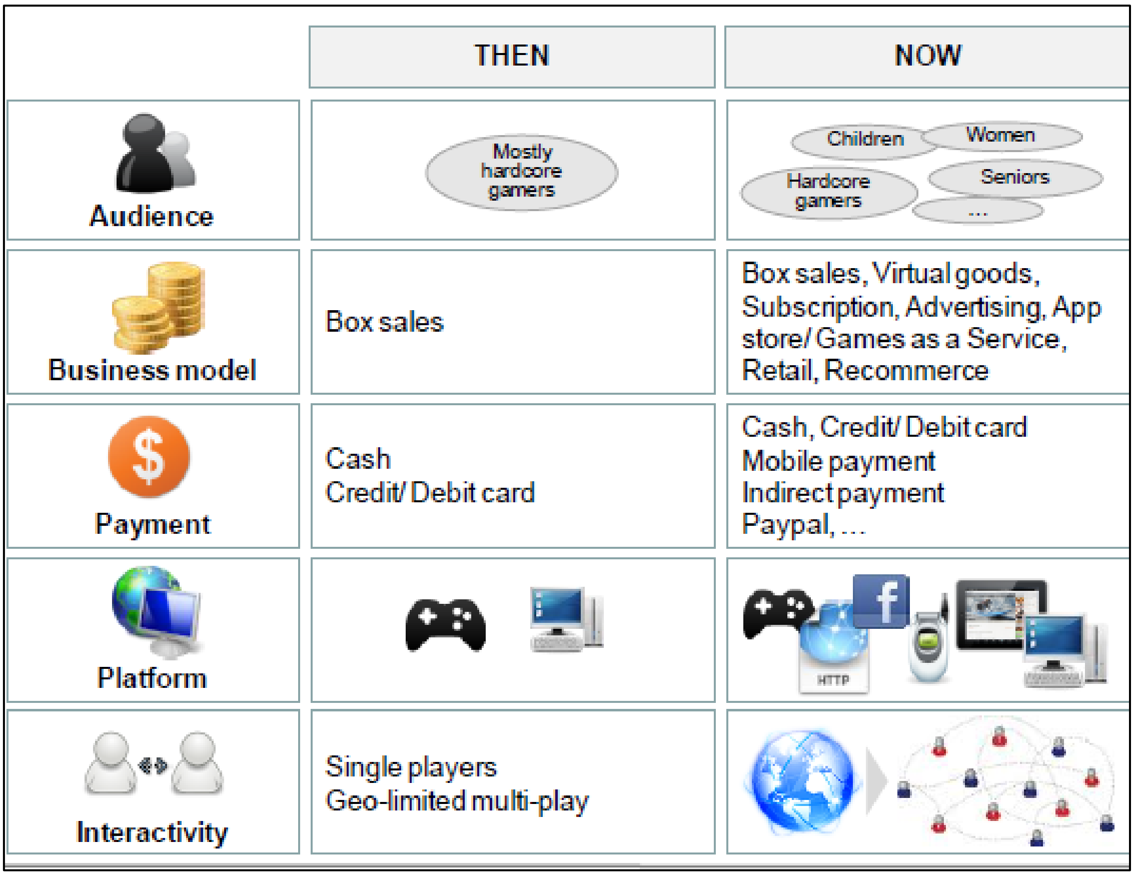
Contents

1. Introduction

1.1 The Video Games Industry

The video games industry is one of the most attractive segments in the entertainment and digital media. As described by Egenfeldt-Nielsen, Smith & Tosca (2016), "right at this moment, millions of people around the world are playing video games. One obvious way in which this matters is financial. The rising popularity of games translates into astounding amounts of cash" (p.7). A universal and integral characteristic across all video games is interactivity and intercommunication between the user and the game (Jansz & Martens, 2005). According to De Prato, Feijóo, Nepelski, Bogdanowicz, & Simon (2010), "this interactivity is pushed to the maximum in online gaming, where the gamer interacts not only with the game itself, but also, in many cases at the same time, with other gamers by means of the moves in the played game" (p.80). This interpersonal communication leads to social gaming in online games, particularly Massive Multiplayer Online (MMO) games that enable a user to play and interact with other users over an underlying network (De Prato et al., 2010). The sheer number of mobile devices in the market has fostered the emergence of these games that are either browser-based or available on online social networks such as Facebook, MySpace, Bebo and can be played on PCs as well as on smartphones and tablets (De Prato et al., 2010).

The video games industry has undergone an extraordinary metamorphosis over the last decade as demonstrated by Mueller-Veerse, Vocke, Vaidyanathan Rohini and Malatinska, (2011), suggesting that the emergence of online, social and mobile gaming has attracted a much wider audience to the industry over and beyond the traditional target group of established gamers. Even conventional MMO games that are PC or browser based are in a vital phase of restructure and development (Mueller-Veerse et al., 2011). Currently, several console games include an online element that enhances the gameplay experience and interactivity to an extent that may essentially override the significance of offline missions of the game (Chikhani, 2016). The commerce within the industry has also shifted its focus from box sales of PC or console games where a one-time purchase would provide access to full features and functionality of the game, to more complex current business models like subscription, pay-per-download and freemium (Mueller-Veerse et al., 2011). With the advent of smartphones and app stores in the market in 2007, there was a monumental shift towards mobile gaming, and the industry that was once dominated by a few companies, now has giants such as Apple and Google escalating as a result of sales earnings from their games app stores (Chikhani, 2016).



	THEN	NOW
Audience	Mostly hardcore gamers	Children Women Hardcore gamers Seniors ...
Business model	Box sales	Box sales, Virtual goods, Subscription, Advertising, App store/ Games as a Service, Retail, Recommerce
Payment	Cash Credit/ Debit card	Cash, Credit/ Debit card Mobile payment Indirect payment Paypal, ...
Platform		HTTP
Interactivity	Single players Geo-limited multi-play	

Figure 1.1: A radical transformation in the games industry (Mueller-Veerse et al., 2011)

1.2 The Free-to-Play Model

Osathanunkul, (2015) describes a particular type of business model in the video games industry called the Free-to-Play (F2P) model as an antithesis of the conventional Pay-to-Play (P2P) model of video games. In this model, clients are provided with access to a considerable part of fully functional game content before having to purchase any additional features (Solidoro, 2009; Marchand & Hennig-Thurau, 2013).

F2P models are mainly of three different kinds:

- Shareware: typically game demos offering some content free of cost as trial, with the assurance of full version upon payment after the end of the trial period (Osathanunkul, 2015; Sotamaa, 2005; Edwards, 2012).
- Freeware: small fully functional games that do not charge at all (or may charge an optional fee) and are available for unlimited time but with no access to the source code or limited rights to use (Kayne, 2017; Chen, 2005; Coleman & Dyer-Witheford, 2007).

- Freemium: games in which full versions are downloadable free of cost and a substantial part of the game accessed without having to make any purchase, while players are urged to pay for high-quality virtual products and advanced features and functionality during game play (Alha, Koskinen, Paavilainen, Hamari & Kinnunen, 2014; Mueller-Veerse et al., 2011; Hung, 2010).

F2P games are available across multiple platforms - gaming consoles, PC, smartphones and tablets and have penetrated their way into most game genres including Massively Multiplayer Online (MMO) games, social network games, multiplayer shooter games, mobile casual games, social casino games etc. (Alha et al., 2014). The boom of this business model on different platforms is evident. For example, on iOS, F2P is the leading revenue model among applications that rake in the highest amount of money (Appshopper, 2014). On PC, Team Fortress 2 which was initially released as a retail game in 2007, was re-launched as a F2P game in 2011 that led to a twelvefold rise in income (Miller, 2012); and most of the commercial social network games on Facebook have also adopted the F2P model (Paavilainen, Hamari, Stenros, & Kinnunen, 2013).

The thriving market of F2P games makes this a focus of this research, specifically the freemium model, which is the most prominent of the F2P models and are heading towards ubiquitous penetration with more and more platforms becoming available (Mueller-Veerse et al., 2011).

1.3 The Freemium Model

1.3.1 Gameplay

In this model, the principle commodity i.e. the game itself is offered free of cost, while superior or bonus content such as virtual game currency, additional game features or customisations have optional charges (Jacobs, 2015). Availability of freely downloadable full versions of the games (Evans, 2016) easily attracts a large number of users since no initial payments are required (Mueller-Veerse et al., 2011; Paavilainen et al., 2013; Psychguides, 2018).

To start with, freemium games offer straightforward access and use of its features through tutorials that help players become acquainted with basic actions while restricting access to advanced operations (Luban, 2011). Progress bars that determine player progression at each step and clear elucidation of game objectives ease players further into the gameplay through a coherent interface (Luban, 2011). Game developers design an addictive gameplay where initially the game is fun and easy (Psychguides, 2018), offering ample resources to explore its features as players level up (Luban, 2012). With progression through the game it becomes more challenging (Psychguides, 2018), virtual items and customisations become costly and new challenges and opportunities emerge (Luban, 2012). At this stage, carrying out principle tasks to advance in the game require virtual resources like food, energy, health, building materials, virtual (in-game) currency and similar things (Luban, 2011; Alha et al., 2014; Hum, 2014; Psychguides, 2018). The game mechanism is such that these resources are depleted quickly (while performing various actions), but require some amount of time to get replenished naturally (Luban, 2011), or in the case of in-game currency, can be gained only by fulfilling certain tasks or missions (Alha et al., 2014). As a result, the player is left with the frustrating choice of either waiting for the item to renew or giving in to the temptation of using real money to purchase them (including virtual currency) (Luban, 2011; Alha et al., 2014).

Designers employ several other mechanisms to stimulate the user's inclination towards freemium games. Some of these are,

- rewards for logging in to the game regularly each day (Dergousoff & Mandryk, 2015; Luban, 2011)
- daily indispensable tasks like harvesting crops or collecting rents from properties (Luban, 2011)
- new virtual resources, challenges, missions, features or relevant incentives from time to time (Luban, 2011; Psychguides, 2018; Hum, 2014)
- encouraging players to invite their friends to the game (Luban, 2011) to enable exchange of essential items as gifts (Luban, 2012)
- asking virtual help to complete certain tasks, thereby promoting social interaction and virality (Luban, 2011)

1.3.2 Revenue Generation

The monetisation model in freemium games is based on the micro-transaction or micropayment model (Whitson & Dormann, 2011; De Prato et al., 2010; Davidovici-Nora, 2014; Luban, 2012). In this model the elementary game is freely downloadable, but virtual goods, virtual currency, advanced content and game resources can be purchased by users in order to augment the overall game playing experience (Davidovici-Nora, 2014; Whitson & Dormann, 2011; De Prato et al., 2010).

A standard revenue structure used in most freemium games is the double currency model with one type of currency acquired during the course of game-play called virtual currency/grind currency and the other that needs to be paid for using real money called premium currency (Alha et al., 2014; Luban, 2012). Grind currency, earned by completing tasks in the game (i.e. by grinding away) requires time and effort and is usually used to purchase typical resources; while premium currency can only be acquired via real money transactions and can buy exclusive and premium content (Luban, 2012; Alha et al., 2014). In this dual currency setup, premium currency can be exchanged for virtual currency but not the other way round (Alha et al., 2014). However, in some games, small amounts of premium currency is often rewarded for first (or daily) logins to the game, special quest completions, sharing/liking the game on Facebook and referring the game to a friend. The purpose is to introduce the player to the value of the premium currency and hence motivate them to buy it at higher and complex levels of the game (Luban, 2012).

Alha et al. (2014) states that the monetisation structure of freemium games permits price points that are adaptable to players with varying levels of choices to pay for supplementary game features. Luban (2012) elucidates that some of the most widely sold items in freemium games are –

- Virtual resources that enable longer duration of gameplay by boosting players' skills or upgrading their resources
- Customisations of in-game characters or avatars
- Building units that assist players in completing missions/quests
- Health/energy recharges
- Collectibles not directly related to the gameplay but allow players to build an assortment of items and trade these with others,
- Time acceleration - some games have an "energy bar" that is exhausted as the player performs actions within the game environment. Virtual items worth premium currency like coffee, snacks or energy-bars can refill the "energy bar" without the player having to wait, thereby accelerating the speed of progression within the game.

Mueller-Veerse et al., 2011 explains that the thriving infiltration of micropayments has played a key role in the success of the freemium model and the expertise of games publishers in the suitable placement of virtual products is essential for its profitability.

In addition to the sale of in-game virtual products, several games publishers such as BigPoint, Gameforge, gPotato, 6waves and Perfect World employ indirect means of monetisation (Mueller-Veerse et al., 2011). For instance, the game studio dealunited allow the purchase of virtual products if players buy or try out a different independent product from the stores of one of their partners (Mueller-Veerse et al., 2011). Although the micro-transaction model largely thrives on the sale of virtual supplies and game extensions, advertising also has some share in the freemium games market (De Prato et al., 2010; Luban, 2012). Publishers like FreeCent gift virtual items when players watch in-game advertisements (Mueller-Veerse et al., 2011). Potential advertisers will only be interested once the game achieves a sufficient number of users, and hence this avenue for revenue generation cannot be used in the initial stages of game launch (Luban, 2012). Another monetisation technique adopted by game studios is Affiliate Marketing that involves "providing a player with hard money if he visits or registers on a partner site" (Luban, 2012, p.2).

1.3.3 Game Genre and Player Base

Freemium games are accessible online on diverse platforms such as smartphones, tablets, computers as well as gaming consoles (Xbox, Playstation etc.) (Alha et al., 2014). Their underlying structure is that of multiplayer games that thousands of individuals can play simultaneously in the same game environment, thereby promoting social interaction in the form of alliance, combat or contest (Nosrati, Karimi & Hariri, 2013). Thus, they are commonly known as Massive Multiplayer Online (MMO) games and have numerous categories depending on "their gameplay interaction rather than visual or narrative differences" (Nosrati et al., 2013, p.1).

The widely prevalent freemium MMO game genres, independent of game context or backdrop are –

- Role-Playing Games – Players adopt the character of a protagonist (cop, mafia, assassin, sailor etc.), usually in an imaginary world, exploring and solving puzzles in an interactive story-based gameplay (Nosrati et al., 2013). The fundamental aspects shared by these games are – quest completion to enable game progression, social communication, game character evolution through training and skill advancement and possession and handling of virtual reserves (ammunitions, armour, food, health recharge, companies and factories etc.) for game missions (Stahl, 2005, Nosrati et al., 2013).
- Strategy – These games aim to simulate a realistic experience for the player by giving them control over a situation like managing their base, collecting and managing resources, assembling a military force etc. (Stahl, 2005, Nosrati et al., 2013). For instance, the player could be a farmer trying to run his farm by managing his crops, trees, animals, other production buildings, energy etc. in an efficient manner to ensure continuous production in the farm. Other examples include, being the mayor of a city, running a hospital etc.

- Action – Players are required to destroy all enemies and obstacles they come across to be able to persist and advance game-play (Nosrati et al., 2013). This genre includes shooter, first person shooter and battle arena games (Stahl, 2005, Nosrati et al., 2013). The general setup involves two rival teams battling with each other in distinct games to dismantle the opponent's base for victory, with each player in the team regulating their individual game character. (Nosrati et al., 2013). These games also encourage cooperative team play (Nosrati et al., 2013).
- Others include social casino games, sports, racing and other casual/puzzle games (Stahl, 2005, Nosrati et al., 2013, Alha et al., 2014).

The user base of freemium games consists of different player types based on their playing styles, strategies, motivations, preferences and gratifications (Dixon, 2011). One of the early classifications of player types proposed by Bartle (1996) was the basis for several other more recent work in this area. These can be broadly categorised as follows –

- Killers, who have an aggressive style of gameplay and inflict themselves on others (Bartle, 1996). Similar to the play style of Conquerors who enjoy challenges and the feeling of victory after hardship (Bateman & Boon, 2005), and Griefers who bully their way to move forward (Bartle, 2004). Drachen, Canossa & Yannakakis (2009) have classified them as Veterans that are highly skilled performers. These players have committed mentalities with immersion as their motivation for play (Kallio, Mäyrä & Kaipainen, 2011, Yee, 2006).
- Achievers, who are predominantly interested in collecting experience points and progressing through game levels (Bartle, 1996). Closely related to the playing styles of Managers in the Bateman & Boon (2005) model, these players are committed to mastering the gameplay above everything else. Their motivation for play is achievement (advancement, mechanics, competition) (Yee, 2006). Achievers who are also organized in their approach are termed as Planners by Bartle (2004).
- Socialisers, who are more fixated in inter-player relationships, inviting friends to the game and playing together to complete missions and tasks (Bartle, 1996). Akin to Networkers and Friends in Bartle (2004) model. Their motivation stems from socialising, relationship and teamwork (Yee, 2006) and they exhibit social mentalities of gaming such as gaming with children, gaming with mates and gaming for company (Kallio et al., 2011). This type called Participants by Bateman & Boon (2005), are most content when playing with other people.
- Explorers, who enjoy traversing within the realm of the game world, curious about the finer details of game mechanism and features (Bartle, 1996). Bateman & Boon

(2005) have referred to these players as Wanderers who are keener to experience the exclusive and intriguing features of the game. These players exhibit some of the committed mentalities of gaming such as gaming for fun and entertainment and casual mentalities of killing time and relaxing (Kallio et al., 2011), with discovery and escapism as motivations for play (Yee, 2006). Subdivisions of Explorers defined by Bartle (2004) are Scientists (experimental gameplay that is meticulous and systematic) and Hackers (possessing an intuitive understanding of the virtual world).

- Some other varieties are
    - Politicians - candid manner of gameplay that may range from leadership to interfering (Bartle, 2004)
    - Solvers - play independently without asking for hints and answers thereby taking longer to finish the game (Drachen et al., 2009)
    - Opportunists - combination of Achievers and Explorers (Bartle, 2004)
    - Pacifists - make the slightest request for help with less than mediocre completion times (Drachen et al., 2009)
    - Runners - quick completion times but considerable number of requests for assistance (Drachen et al., 2009)

1.4 Importance of the Games Industry

In the following section, discussion is given of the significance of the video games industry, especially freemium games, in connection to the global economy and society. This justifies the usefulness in analysing consumer behaviours within the framework of online freemium games and the benefits that this research would make in that aspect.

1.4.1 Contribution to Economy

The video games industry is big business worldwide. The growth of the global games market was predicted to be a total of $143.5 billion in 2020 from $116.0 billion in 2017 (Wijman, 2017). Tech advisor Digi-Capital reported games software and hardware combined sales of more than $150 billion in 2017, reaching over $200 billion in 2021 with a flourishing compound annual growth rate of 7.9% over the following 5 years (Digi-Capital, 2017). Another report by ERA (2018) reveals an increase by 9.6% in video games sales (digital and physical combined) from £3.06 billion in 2016 to a £3.35 billion in 2017. UKIE (2017) states that the valuation of the UK games industry was almost £4.33 billion in customer spend in 2016, which was an increase of 1.2% from £4.28 billion in 2015.

This boom in the video games industry is primarily brought about by the predominance of mobile gaming (on smartphones and tablets), catering to the freemium model of online games. The mobile games market accounted for the majority of the 2017 global games market with revenues up to $50.4 billion i.e. 43%, while console and PC games contributed 29% and 28% respectively of the market share (Wijman, 2017). The report by Digi-Capital (2017) discusses the impact of mobile gaming on the games industry stating that this sector is expected to contribute to the most prominent part of the business growing from over $50.0 billion in 2017 to over $80.0 billion in 2021. The ERA (2018) report informs that the sales of digital games experienced a higher growth of 12.1% (£2.28 billion in 2016 to £2.56 billion in 2017), whereas that of physical games grew by only 2.1% (£776 million in 2016 to £792 million in 2017). Another report from games data and market research firm SuperData (2017) estimates earnings worth $22 billion from micro-transactions in freemium games in 2017, in contrast to $8 billion from full game investments the same year. According to a survey constituting of a sample of adults in the UK conducted by GoCompare (2017), it is found that around 37% play freemium games more than 5 days a week, and those that invest real money in these games spend £13.22 on average towards purchase of in-app products.

The big business of games has brought about a revival in global game investments as revealed by Digi-Capital (2017) with $2.8 billion invested worldwide over a year up to Q2 2017. Revenues generated by the global games market are comparable to the global sports business (approximately making $130 to $150 billion in total) as noted by Wijman (2017) who further points out that the ongoing growth rates of both markets will likely result in exceeding returns from the games business over sports.

TIGA (2018) informs that the largest video games sector in Europe is the UK video games industry, which in 2016, supplied £1.2 billion to the UK's GDP, raking in £514 million tax revenues for HM treasury. The industry provides direct or indirect employment to 33,637 individuals of which 11,893 are games developers (TIGA, 2018). UKIE (2017) reports, in accordance with sales data from ERA (2018), that 46.3% of the total worth of the UK entertainment sector is constituted by the UK video games industry, making it 1.25 times the size of the video market and 2.8 times the size of music industry.

Therefore, it is evident that the video games industry embodies big business and has much to offer to the global economy. The advent and expansion of high-class technology is spawning a growing market for freemium games that studios are tapping in to. With an outpour of indie game developers and small start-up companies contributing to the growth of the industry, jobs are being created increasingly. The freemium model of video games provides a viable investment opportunity, which creates a need for analysing the performance of these games and the behaviour of its consumers.

1.4.2 Impact on Society

The video games industry has several economic and social impacts– some being positive and some negative. While the contribution to economy has largely been a positive influence generating revenue and jobs, the possible effects of these games on society have received mixed response in the general media and scientific literature.

Gentile, Lynch, Linder & Walsh (2004) states that although video games are designed to be enjoyable, challenging and occasionally informative, many contain distressing content as well. Playing violent video games have been linked with an escalation of aggressive behaviour and physical fights and diminishing positive prosocial behaviours (Anderson & Bushman, 2001). In addition to unfavourable behavioural patterns, time spent playing electronic games have been found to be strongly correlated with childhood obesity (Vandewater, Shim & Caplovitz, 2004) and excessive game playing has been related to muscular and skeletal injuries (Brasington, 1990; Lemos, 2000) and video game related epileptic seizures (Trenite et al., 1999). Other negative health impacts of violent video game play include rise in physiological arousal such as blood pressure and heart rate (Anderson & Bushman, 2001) and stress hormones (Lynch, 1999). Several studies have also identified poorer educational performance as consequence of increasing play of video games (Harris & Williams, 1985; Creasey & Myers, 1986; Lieberman, Chaffee & Roberts, 1988; van Schie & Wiegman, 1997; Roberts, Foehr, Rideout & Brodie, 1999; Anderson & Dill, 2000; Walsh, 2000; Paschke, Green & Gentile, 2001).

Nevertheless, video games have been found to have favourable influences as well. In a study published by Huizinga (2014) that assessed the prominence of ‘play’ in cultures that have historically considered it to be a menial activity, it was observed that games build a “magic circle” that separate the player from the outside world and have no repercussions on anything present outside the circle. However, Egenfeldt-Nielsen et al. (2016) have been critical of this theory explaining some real-world outcomes that games have on peoples’ lives. These include consumption of time in daily life, impact on players’ moods that may in turn influence other activities performed by them, and affecting positive behaviour (for example: the game ‘America’s Army’ has been adopted by the American military as a recruitment tool that has been supposedly productive) (Egenfeldt-Nielsen et al., 2016). Simulation games that allow players to interact with virtual reality situations are likely to boost the learning of complex real-life skills like driving (Walter et al., 2001), flying airplanes and playing golf (Fery & Ponserre, 2001) and can be used as potential teaching tools. Rosser et al. (2007) demonstrates that video games may be a practical teaching mechanism to aid in the training of laparoscopic surgical skills. Strategy games may disseminate knowledge about the working of complicated structures such as cities or countries at war, in-game advertisements may promote branding of products in players' minds, and virtual items obtained during gameplay are often sold for real money on online trading platforms such as eBay (Egenfeldt-Nielsen et al., 2016). MMO games provide a rich social experience to players by promoting cooperative gameplay with different forms of intercommunication amongst players (Caplan, Williams & Yee, 2009; Ducheneaut & Moore, 2004). These include immediate and real-time communication between players to collaborate and participate with each other on various missions and adventures, accomplishing common goals as a team, and assisting each other via gifts and tips (Zhong, 2011). In a society with deteriorating community alliance and public trust (Livingstone & Markham, 2008; Putnam, 1995) and less opportunities to meet people (Williams, 2006), the virtual world of MMO freemium games can serve as a new scheme of association, interaction and collaboration in the society (Zhong, 2011). Other useful effects include better performance in neuro-psychological tests (Nielsen, Dahl, White & Grandjean, 1998), enhancement in visual-spatial skills (Dorval & Pepin, 1986), considerably superior eye-hand motor coordination on a pursuit rotor (Griffith, Voloschin, Gibb & Bailey, 1983), significantly quicker reaction times (Yuji, 1996), advancement in mental rotation performance (De Lisi & Wolford, 2002) and improvement in visual attention (Green & Bavelier, 2003).

Overall, video games are cultural, audio-visual commodities comparable to other entertainment media such as film, television and animation (TIGA, 2018) leading to its mass consumption and a viable market worldwide.

1.5 Research Aim and Objectives

This research study builds upon work conducted for a Scottish Government funded Knowledge Transfer Partnership (KTP) project that involved analysis of data from online freemium games. The purpose of this research is developed from the observations and limitations of that project.

The overall aim of this study is to develop suitable data-driven methods to gain insight about customer behaviour in online freemium games in order to enhance user engagement and maximise revenue generation. This will be done with a view to providing recommendations for successful business in the freemium games industry. The research aim is expected to be achieved by analytics of in-game player data, leading to the construction of statistical models of online behaviours that are implementable by the game studios in real-life scenarios. To do this, the following objectives will be pursued –

1. Maximisation of user engagement with the game – In any business, customer engagement and consequently retention is crucial in ensuring long-term profitability and this is no different in the case of the games industry. Engagement and retention of maximum number of players in a game is vital to its popularity, virality and therefore continuous and long-term revenue generation. This can be achieved by understanding the motivations for players to be engaged in the game and the reasons for their departure from it (also known as customer defection or customer churn or customer attrition).

2. Identification of the likelihood of customer defection at any given time in order to reduce drop-out and encourage remaining active – Another important metric considered in online businesses is the lifetime value (LTV) of customers, which determines the worth of users over their entire lifetime including the change in their value over time. In the context of online freemium games, this is closely associated with estimating at what time points players are most likely to abandon the game and thereby cease to be valuable, and the components of gameplay that induces this event.

3. Increase of the revenue derived from real currency purchases by users – An essential metric that online businesses are interested in sustaining is the average revenue per user (ARPU) which is an indicator of financial performance (Hindy, 2017). An insight into what triggers the first micropayment and what leads to subsequent purchases, the virtual items with maximum sales and characteristics that distinguish payers from non-payers will aid in understanding the monetary performance of games as well as factors driving micro-transactions. This in turn can advise aspects of game design that will promote monetisation thereby increasing revenue generation.

4. Identification of different types or clusters of users in online freemium games and how this is beneficial – A comprehensive idea about the customer base of any business is imperative in understanding its reach, which in turn guides marketing strategies for its expansion. The final objective of this research is to produce a method for identifying the wide variety of players that constitute the user base of online freemium games, such as the heavy users (Luban, 2011), casual gamers (Luban, 2011), whales

(Psychguides, 2018) etc. This will include further scrutiny of these groups with respect to their playing pattern, performance and value added in terms of proceeds and virality.

1.6 Contribution of the Research

This research is expected to contribute to industry and academia in the following ways –

- Providing specific recommendations for the monetisation of online freemium games:

Investigating the behavioural patterns of payers and non-payers using data driven methods to identify determinants of micro-transactions will lead to recommendations for revenue generation and increase in profitability of games. This can be invaluable learning for game studios to better understand how different game scenarios affect customers in their decision making for real money purchases. It will assist in the development of games that are customised for paying and non-paying users and their individual strategies and choices, resulting in an enhanced gameplay experience leading to improved business in this sector. This in turn will influence the global economy in which the online games business plays a vital role. As elucidated by Piggott (2015), prosperity of the freemium business model depends largely on motivating players to pay in micro transactions for improved content, and that "there has been relatively little study performed on this rapidly expanding business model and as such it can be considered as an area of research with plenty of 'low hanging fruit'" (p.15).

- Methodological contribution to the analysis of complex real-time data:

This study will implement statistical techniques in the analysis of large volumes of real-time user data. The procedures are required to be easily executable in real time and robust in effective handling of data that is likely to be skewed and characterised by long-tailed distributions and missing values. As explained by Shah, Horne & Capellá (2012), unless new skills can be developed, the naive analysis of big data will lead to poor understanding and missed opportunities at best and erroneous decision making in more serious cases. Therefore, the formulation of a standard analytic approach for such data will contribute to the armoury of tools for complex big data analysis and increase the reliability of inference from it.

- Expanding knowledge of consumer behaviour analysis for online businesses:

Although the specific area of focus for this research is consumer behaviour in online freemium games, customer interaction and usage in most online products follow a similar trend, with the majority of online businesses facing similar struggles with respect to customer attrition and monetisation. Therefore, the formulation of a statistical framework for analysis of customer behaviours in freemium games would not only add to the learnings in this field, but also extended to insights about user behaviour in other online platforms/apps such as Google, Amazon, Facebook, Netflix, etc. As highlighted by Harrison & Roberts (2011), "predictive models of player behaviour in video games is an open research topic that is receiving increasing attention in the literature" (p.1). Also stated by Bakkes, Spronck & van Lankveld (2012) "player behavioural modelling is a research area in game playing that is gaining attention from both game researchers and game developers" (p.1). Hence, this research would be an important addition to the literature of player behaviour modelling.

1.7 Research Challenges

In order to meet the aims, a thorough examination of players' behaviours within the game environment and during gameplay is required. Typically this involves statistical methods to classify and model customer/client behaviour. However, this is difficult and challenging to implement on data arising from the games industry in general because of –

- The sheer volume of data recorded from gameplay – thousands of players triggering millions of game events.
- Strategies and behaviours of players that are continually evolving as they progress further into the game, which means that there is pressure to engage in fast analysis and quick decision-making.
- Large number of different scenarios within games and between games that are compounded by differences between players and between regions - the inherent variation is enormous.

The wealth of data gathered from online games is invaluable for gaining actionable insights into player behaviour. However, this big data has its perils too. As reported by Lohr (2012), most of this data is unstructured and therefore not typically suited to traditional databases. The fundamental attributes of big data problems are exhibited namely "Volume, Variety and Velocity" – the so called three Vs (Zikopoulos, Eaton, DeRoos, Deutsch & Lapis, 2012; Fan, Han & Liu, 2014). The sheer volume of data being collected from users' gameplay is booming. The size of a sample of data comprising only about two weeks of gameplay is around 6 gigabytes. One of the complications posed by large amounts of data as explained by Zikopoulos et al. (2012) is that the proportion of data that can be evaluated, comprehended and examined diminishes with the increase in its size. Fan et al. (2014) further states that massive sample size induces concerns about excessive costs of computation and algorithmic uncertainty. Another characteristic of big data is its variety because of which it is convoluted. It is a mixture of relational data (that is conventional and methodically configured to fit perfectly into strict schemas) and crude semi-structured and unstructured data originating from web pages, social media and click stream data (Zikopoulos et al., 2012). The data recorded from gameplay is very similar to this thereby exhibiting the same complexity. Big data in the games industry is also high dimensional which is accompanied by “noise accumulation, spurious correlations and incidental homogeneity” – (Fan et al., 2014). The three Vs of big data generate issues, as mentioned by Fan et al. (2014), of statistical biases, heavy tail characteristics, errors in measurement, outliers and missing values which demands the advancement of more flexible and robust methods of analysis.

Any standard data set arising from gameplay events is expected to be filled with missing or null values, mixture of numbers and strings with special characters and an abundance of zeroes corresponding to rare events in a game. A considerable amount of time needs to be invested to clean and pre-process the data to make it suitable for even the simplest of analyses. Fine-grained analysis of large data is more likely to be exposed to the risk of false discoveries (Lohr, 2012). Due to the sheer nature of online games data, majority of the variables are expected to have skewed distributions with heavy-tails thereby making it difficult to apply standard statistical techniques.

1.8 Thesis Outline

The thesis starts with an introduction (current chapter) to the video games industry (especially the online freemium model), its socio-economic impacts, and its importance in the current context. This is followed by an outline of the overall goals of this study, objectives of this research and its contribution to academia and industry.

The next chapter is a critical review of existing literature on relevant analysis of data on consumer behaviour of online freemium games, including prevalent statistical methods for complex big data analysis and modelling.

This is followed by a chapter to describe and justify the research methodology and overall approach to be undertaken.

The five subsequent chapters form the crux of the thesis, starting with a description and exploration of the data used in this research, followed by its analysis and statistical modelling (including validation of the methods used) to fulfil the four research aims elucidated before.

The concluding chapter of the thesis summarises all the results of the research study into a final statistical framework for online consumer behaviour analysis with relevant recommendations for appropriate implementation of the information obtained. It also discusses limitations of the study and further work that could be attempted.

2. Literature Review of Analytics in Online Freemium Games

The overall aim of this research is to develop suitable data-driven methods to gain insight about consumer behaviour in online freemium games, with a view to providing recommendations for successful business in the freemium games industry. This is expected to be achieved by analytics of in-game player data leading to the construction of statistical models of online behaviours. In this chapter, detailed discussions of studies related to online gaming have been undertaken – what those studies have focussed on, their results and further work that needs to be achieved. As stated before, this work will rely on statistical analysis of data sets representing online customer behaviours, and hence the chapter also includes a review of existing statistical approaches to handling data sets of similar nature.

The chapter starts with a brief explanation of an approach governed by data and its advantages over a more subjective interview/questionnaire based procedure. It then attempts to examine how behaviours of players in online games have been studied and construed previously. This will lay the foundation for modelling player engagement and likelihood of micropayments, which is the subject of research objectives one and three. The literature review then progresses on to a study of customer defection or customer churn analysis to determine when players are most likely to quit the game, which is relevant to research objective two. Inspection of classical statistical methods employed for modelling data to address the above is undertaken. Additionally, methods to classify and determine clusters of the entire player base are scrutinised, hence giving an underpinning of research objective four, that of determining the wide variety of players constituting the customer base of online freemium games in terms of playing styles, performance and revenue generation. From these insights into understanding the various behaviour of customers in online freemium games, the research questions will be derived.

2.1 The Data Driven Approach

Conventional models of player behaviour have always been designed based on surveys, small-scale observation experiments, or knowledge engineering, wherein the resulting models were syntactically meaningful but restricted in their applicability (Harrison & Roberts, 2011). Dependence on surveys and questionnaires to develop models can be highly intuitive; however they only pertain to aspects of player behaviour that have been covered in the survey questions (Harrison & Roberts, 2011). Furthermore, the data retrieved from this procedure is not adequate or clean enough for building efficient models as response rates from surveys tend to be low and suffer from the social desirability bias implying that players are more likely to answer questions according to their personal judgement of what is expected of them by the questioner, rather than objectively (Harrison & Roberts, 2011).

Knowledge engineering is another means of investigating into player behaviours, which typically involves the task of collecting and inputting relevant information about players for use in knowledge based computer systems, for instance, a computer program (Studer, Benjamins & Fensel, 1998). This approach is also highly subjective as some pre-existing knowledge is required regarding the type of player and his preference towards certain actions and this is then correlated with the activities already present in the game for players to perform (Harrison & Roberts, 2011). However, more evolved behaviours cannot be easily defined and associated with complex in-game actions in this manner and is highly open to the individual perceptions of the modeller or analyst (Harrison & Roberts, 2011). Moreover, knowledge engineering involves hard coding groups of players with similar characteristics into the model even before observing their actual behaviour in a particular game and this is not very desirable (Harrison & Roberts, 2011). Observation experiments are a way to overcome the drawbacks of this method (Harrison & Roberts, 2011).

An alternative to the above procedures, as elucidated by Harrison & Roberts, (2011), is using a data-driven approach for creating models of player characteristics, the primary assumption being that an accurate prediction can be made of a player's actions in a given setting if sufficient data from other players in a similar situation has been analysed. A game’s player base is made up of a variety of players with different strategies, competencies, and likes and dislikes (Bartle, 1996). It is fair to assume that similar kinds of players are more likely to get attracted to similar content in a game, for example, fighting, exploring, managing resources etc. Hence, observing the behaviour of existing players could be used as a good basis for modelling future player behaviour (van Lankveld, Schreurs & Spronck, 2009). In a data-driven procedure, this would mean collecting data on game events triggered by active players and using that to build the models of current player behaviour. It could then be used to predict future behaviours as well.

The main concerns regarding this approach are model complexity and algorithmic efficiency. Most freemium games, especially massive multiplayer online games, have a considerably huge player base and several possible game events that can be triggered due to their complex and open-world nature. Therefore, the data collected is big data, typically millions of events on thousands of players, the analysis of which is naturally cumbersome.

2.2 Analysis of User Behaviours in Games

The focus of this section is to review literature pertinent to the research objectives one and three, that of maximising user engagement with the game and increasing revenue through in-game real currency purchases.

Contemporary computer games have sophisticated graphics and provide a large virtual setting often emulating the real world for players to explore (Jennett et al., 2008). The controls are highly advanced and enable players to move their characters and perform actions in a wide variety of ways, and even allow for multiplayer game-play (Jennett et al., 2008). Regardless of the genre of the game or how advanced or not its appearance, all successful games have one common aspect, the capacity to attract people (Jennett et al., 2008). In addition to the initial appeal of the game, it is also imperative for game designers and developers to keep in mind that players need to be committed to the game for it to do well. Prolonged engagement of the player with the game is crucial in adding value to the player experience (Schoenau-Fog, 2011).

Jennett et al. (2008) had described and measured immersion in computer games with the help of three experiments. On the assumption that gamers themselves are able to identify their state of immersion in a computer game, yet the entire concept of it is not coherently defined, one of the objectives was to develop an immersion questionnaire. Another aim was to study the correlation between the senses of immersion felt personally by candidates to more objective measures that were quantifiable. A questionnaire was developed for the first experiment in order to study the association between immersion in computer games and time taken to complete tasks in the real world. It was seen that higher the degree of engagement during game-play, longer it took players to subsequently complete a skill-based task unrelated to the game. The second experiment was designed to study the relationship between immersion in games and variations in the number of eye fixations over time. The same questionnaire that was developed for experiment 1 was used for this purpose as well. The data on number of eye fixations was analysed and it was found that when players were in a non-immersive state of being, there was a notable increase in their eye movements. The questionnaire in the third experiment was devised to probe into the relationship between immersion and the communication speed with the computer interface. It was learnt that, increasing the pace of computer interaction, though increased the level of immersion, also raised the candidate’s temporary condition of fear, nervousness and discomfort and caused negative influence. As a result, emotional absorption with the game was identified as a crucial factor leading to engagement and this was further supported by Brown & Cairns (2004). The questionnaire designed for experiments 1 and 2 successfully quantified immersion and revealed that the degree of immersion while playing a computer game was considerably higher than that observed during clicking tasks. Overall, the study by Jennett et al. (2008) addressed several matters concerning players’ engagement with video games. It attempted to evaluate immersion subjectively with the aid of questionnaires as well as objectively through physical attributes of players such as time to complete tasks and movements of the eye. It also demonstrated that immersion is not just a positive experience for players, but also brings with it negative feelings like unrest, apprehension and nervousness and these emotions only tend to increase with the level of engagement with the game.

Poels, De Kort & Ijsselsteijn (2007) studied the emotions and experiences people have when playing digital games. The approach taken by them was to assemble focus groups of different types of gamers to consolidate a provisional but exhaustive list of in-game experiences. An expert meeting was then set up to discuss the experimental findings with existing theoretical ones. At the end, in an expert meeting comprising of psychologists and both regular and rare gamers, knowledge gained from both theoretical studies as well as analysis of the focus groups was combined into a provisional model of game experience. This subjective and experimental technique provided the researchers with an extensively diverse set of digital gaming experiences and as opposed to fragmented literature, it “presented a more complete overview of how it feels to play digital games” (Poels et al., 2007, p.88). Additional studies are required to investigate and analyse the correlational and causal associations between various dimensions of game experience, as well as interaction between game experiences and game types, player categories and playing styles.

Schoenau-Fog (2011) used grounded theory to examine an aspect of player engagement by determining elements that could be related to the inclination towards playing video games. Here, player engagement was defined and distinguished from motivation to start playing a game or being drawn towards the game in the first instance. All information about the factors contributing to engagement was collected through surveys that covered questions related to overall player experiences during gaming in order to uncover what brings about engagement and disengagement with a game. The questions were kept flexible so as to enable participants to give unbiased responses without being influenced by predefined answers and classifications stemming from theoretical observations. Grounded theory was then applied to pick out the main causes of player engagement by classifying and coding the statements made by participants. These triggers for engagement were further verified for recurring cases or situations through focused coding and banded into initial categories. These categories were assessed for likeness and further classified into provisional categories with specific properties that held them together. In a further iteration, these groupings were then brought down to fewer numbers of conceptual categories. Four major factors came out as driving a player's engagement to games – objectives, activities, accomplishments and affects.

For Brown & Cairns (2004), the concept of engagement was clearly demarcated from immersion into a game – in fact, engagement is seen as the first stage of immersion and the "lowest level of involvement with a game and must occur before any other level" (p.1298). This study also used Grounded Theory to break down the concept of immersion into three stages – engagement, engrossment and total immersion. It was achieved through interviewing gamers of both genders and above the age of 17 who played video games routinely. Grounded theory was used to examine the interview through "open coding, identification of concepts and categories of concepts, and some axial coding, identification of relationships between categories" (p.1298). The outcome was recognition of three levels of involvement with a game – engagement, engrossment and total immersion. It was also found that captivation with a game was affected by time spent in the game and obstacles faced by players. These obstacles could be a mixture of human factors like concentration, endeavour and time dedicated to the game as well as game-related issues like graphics, exciting missions and storyline.

Medler, John & Lane (2011) collected player events data from actual gameplay of a multiplayer game called Dead Space 2 and used it to create a visual game analytic tool for studying player gameplay behaviour. The main graph in the tool was used to visualise data related to the total number of users being monitored, the ratio of kills to deaths, the number of game rounds played and won, experience points attained by players, statistics related to weapons and completion rates of game objectives. Abbasi, Ting & Hlavacs (2017) conducted a study to form a new instrument for analysing in-game engagement, which they defined as consumer video game engagement. The approach taken was to create an instrument based on the scale development method, to estimate the construct of consumer video game engagement.

Yee (2006) used factor analysis techniques to build an empirical model of player motivations, of which immersion was a component. The data used in this study was obtained from a 5-scale response of a questionnaire designed for this purpose. Cole & Griffiths (2007) used an online questionnaire and basic descriptive statistics to establish that social interactions in online gaming can be a significant aspect in contributing to the happiness and satisfaction in playing.

Monetisation of customers, as perceived in this research, refers to the tendency of users to spend real currency within the online freemium gaming environment, for the purchase of virtual items. Wohn (2014) analysed log data recorded by the server of an online social game called 'Puppy Red', consisting of demographic, behavioural and network variables, to determine the social factors that influence the prospect of spending money within the game. They fitted binomial logistic regression to a dichotomous outcome representing real money spent versus not spent, modelled the amount of money spent by active spenders using negative binomial regression, and compared the purchase patterns of low spenders with that of high spenders via t-tests. The social variable denoting donation of virtual items to other players was found to be the strongest positive factor impacting the tendency to spend real money as well as the amount of money spent. Another social variable, number of friends, also had positive association with the likelihood of real money spend, but had no influence on the amount spent. The amount of time spent on the game website and quantity of virtual currency earned did not impact the likelihood of spending real money. Furthermore, it was found that high spenders purchased decorative items with no utility value, whereas low spenders bought more useful and practical items.

Caetano (2017) conducted a survey comprising closed questions and Likert-type items to determine the history of mobile game usage of players in order to evaluate the hypothesized structural model for impulse buying in a micro-transaction mobile environment. They employed descriptive statistics, distribution tests and common factor analysis, Spearman's test to examine the relationship between drivers of micro-transactions and the propensity for impulse buying, Wilcoxon's test for the variation in purchase intention for different prices, and basic structured equation model path analysis using partial least squares regression to analyse the data. Their study found that an immersive and gratifying mobile game experience positively influenced impulse buying proneness. The intent to purchase was positively associated with social and emotional value but not to functional value, and declined with an increase in price. Features involving performance were found to be alluring to players as well as related to impulse buying.

King, Gainsbury, Delfabbro, Hing & Abarbanel (2015) studied the primary causes of overlap between different aspects of gameplay such as interactivity, monetisation, betting and wagering, and found that gambling can be described by monetisation characteristics comprising risk and pay-out to users. Hamari & Lehdonvirta (2010) studied the factors that caused customers to buy virtual products that were sold for real money, by aiming to understand the mechanics and rules constructed by game developers to promote virtual good sales. It was a subjective and theoretical approach that was based on marketing. (Heeks, 2009) focused on the study of virtual economies in MMO games that represent the production of premium in-game currencies, and suggested standard economic model fits to the data for their analysis.

Cheung, Shen, Lee & Chan (2015) implemented a survey design with participants in China answering an online questionnaire, to suggest a research model that investigated the factors contributing to active engagement in playing online games and the relationship between such engagement and the sales of online games. Their model was empirically tested using partial least squares regression, with the mediation effects examined in the structural model and internal consistency and convergent and discriminant validity of the instruments estimated. The study revealed that the amount of money spent in online games was motivated by both psychological and behavioural engagement, with the former being statistically significantly affected by satisfaction, customisation and social interaction within the game.

Review of the existing literature on user behaviour in games informs various notions surrounding player engagement and monetisation that will aid in the development of research questions associated with the relevant objectives (one and three) of this study. Overall, engagement, engrossment and total immersion were recognised as the three levels of involvement with a game, with immersion found to be a component of an empirical model of player motivations. Research undertaken to study user engagement in video games found that deep-rooted engagement of players with games was crucial in adding value to the player experience and consequently to successful games. Important determinants of engagement were emotional absorption with the game, game objectives, activities, accomplishments and affects, time spent in the game, and obstacles faced with respect to graphics, missions and storyline. A crucial factor contributing to the enjoyment and satisfaction in online gaming was social interactions between players. The methods used in these studies ranged from subjective analysis and descriptive statistics via questionnaires, and analysis of focus groups and expert meetings, to grounded theory applied to surveys, interviews of gamers, and development of an instrument to estimate the construct of engagement. One study investigated player events data from actual gameplay to design a visual analytic tool for studying player behaviour.

Research undertaken to study monetisation in games found that donation of virtual items, number of virtual friends, and the social and emotional value of games positively affected the likelihood of spending real money. The amount of money spent was positively associated with gifting virtual products, and motivated by behavioural engagement as well as psychological absorption that was influenced by satisfaction, customisation and social interaction within the game. It was also found that impulse buying tendencies were positively determined by an immersive and gratifying mobile game experience, and related to features involving performance. The studies predominantly used survey designs with online questionnaires, employing descriptive statistics, distribution tests and structural equation modelling. One study used log data collected from an online game server and fitted regression models to a dichotomous outcome denoting real money spent or not and the amount of money spent.

Thus, although player engagement and real currency transactions are established in the literature as crucial components of an enjoyable and satisfying gaming experience, the factors actually predicting these have not been analysed using real-time gameplay data and statistical techniques such as predictive modelling. Considering customer engagement as a binary outcome, and applying classification algorithms to contrast engaged group of players with the non-engaged have not been investigated. Additionally, studies have not examined the number of micro transactions as a response variable and the various customer behaviours that may have a relationship with the same. Therefore, there is a need for further analysis in this area, in order to understand user behaviour during actual game play that may predict engagement and the likelihood of making at least one micro transaction. This leads to the development of some questions that this research will attempt to answer, which are, what gameplay behaviours in online freemium games significantly predict increasing engagement amongst its users, and what facets of the player experience promote an increase in the quantity of real currency micro purchases by players.

2.3 Customer Churn Analysis

In this section, published studies of customer churn prediction in online games are reviewed, in order to fulfil the research objective two, that of identifying the probability of defection at any given time, and reducing player drop-out by encouraging to remain active.

Various industries such as banking, insurance, retailing, telecommunications, etc. investigate churn prediction (Kawale, Pal & Srivastava, 2009). Some of the widely used techniques for this purpose are classification and decision trees (Datta, Masand, Mani & Li, 2000; Hadden, Tiwari, Roy & Ruta, 2006; Ng & Liu, 2000), logistic regression (Buckinx & Van den Poel, 2005; Jones, Mothersbaugh & Beatty, 2000), latent semantic analysis (Morik & Köpcke 2004), survival analysis (Lu, 2002; Mavri & Ioannou, 2008), ordinal regression (Gopal & Meher, 2008), support vector machines (Zhao, Li, Li, Liu & Ren, 2005; Coussement & Van den Poel, 2008; Lee, Chiu, Chou & Lu, 2006) and random forests (Coussement & Van den Poel, 2008). In a comparison of random forests, support vector machines and logistic regression in churn prediction in newspaper services by Coussement & Van den Poel (2008), random forests were revealed to have better performance than support vector machines.

Several aspects of game playing strategies have been examined by studies, with a view to detect factors associated with defection times. Tarng, Chen & Huang (2008) focused on correlations between players' short term behaviour (average session time, average daily session count and average daily playtime) and long term behaviour (average length of consecutive days played, the average length of active period and the overall subscription time) and also explored whether players' gameplay behaviour in one time period will be continued in the following period (using correlation plots). Kuss, Louws & Wiers (2012) used a web-based questionnaire to test statistical associations between gaming behaviour, gaming-related problems and gaming motivations. They found that "gaming motivations escapism and mechanics significantly predicted excessive gaming and appeared as stronger predictors than time investment in game" (p.1). Chen, Huang & Lei (2009) conducted a study to understand the effect of network quality on the decision of a player to discontinue a game sooner than expected. They visually demonstrated the estimated probability that a player who has already played for a certain time will discontinue the game within the next 10 minutes through estimated hazard plots and survival plots for the observed game sessions. The relationship between the player departure process and the network conditions they face was explained via correlation analysis. The probability of premature departures was modelled with a logistic regression model. Kawale et al. (2009) suggested a churn prediction model that depends on social influence among players and their personal engagement with the game. They recommended a modified diffusion model that efficiently integrated social influence and player engagement to considerably enhance the prediction accuracy of churn prediction models. The approach taken by Hadiji et al. (2014) towards churn analysis was that of a binary classification test, wherein the knowledge of a player up to a particular time point is used to label them as churned or returning. The user's playtimes over sessions were modelled using temporal models in which individual observations recorded for each player were fitted with a power-law function. This was based on previous research by Bauckhage et al. (2012). The models by Hadiji et al. (2014) were comprehensive and not confined to the characteristics of any specific game, and their churn analysis technique evaluated a number of different classifiers such as neural networks, logistic regression, naïve Bayes and decision trees. The study by Bauckhage et al. (2012) on the impact of playtime for forecasting the decline of players' interest in a game, implemented lifetime analysis procedures to identify the Weibull distribution as an apt empirical distribution of total gameplay times, thereby suggesting that the progression of players' immersion into games followed a non-homogenous Poisson process with a power law intensity function. Demediuk et al. (2018) analysed historical data from 201 players of a multiplayer online freemium battle arena game called League of Legends to model and predicted the time until the player plays another match, conditional on a set of explanatory variables. The survival function at the population level was modelled using a Kaplan-Meier estimator, and standard and mixed effects Cox regression models were applied to investigate the effects of the explanatory variables. Hazard ratios for both the standard and mixed effects models were found to be relatively stable, and the rate of time until the next match is found to decrease with an increase in the length of play while increasing with an increase in the average time between sequential matches. Player competency had no significant contribution to the probability of match play. Periáñez, Saas, Guitart & Magne (2016) analysed data from high value players (also known as whales) of a mobile social game for modelling the time until churn. They visualised the churn problem using Kaplan-Meier survival curves stratified by whales, normal payers and non-payers, and predicted the departure time of whales from the game using survival ensembles with 1000 conditional inference trees. The latter method is compared with Cox regression and other binary classification techniques such as support vector machines, naïve Bayes and decision trees. The last purchase amount, days since last purchase and user level were found to be the most significant predictors of time to churn of whales from the game.

Review of the existing literature on customer churn in games informs knowledge about player defection, including its identification and causes, which will aid in the development of research questions associated with the relevant objective (two) of the study. Time investment in game was found to be not as strong a predictor of excessive gaming as escapism and mechanics, while social influence and engagement among players was observed to be related with their churn from the game. The decision to abandon a game prematurely was found to be affected by network quality. The rate of time until a subsequent match in a game was revealed to be negatively associated with the length of play, and positively associated with the average time between sequential matches. Finally, the time to churn for whales were seen to be significantly predicted by last purchase amount, days since last purchase and player level. A range of methods were used in these studies, including correlation plots of session and play times, analysis of web-based questionnaires, lifetime analysis with hazard and survival plots, modified diffusion models, classification algorithms, and gameplay times modelled as temporal models or Weibull distributions. Two studies actually examined the time to player churn using Kaplan-Meier survival curves, mixed effects Cox regression models and survival ensembles.

Predominantly, studies are found to have examined game play times and churn rate, rather than the time to churn. The two studies that have investigated the latter did not use a wide range of behavioural variables in their models, especially those indicating player performance and competence. One of these analysed a very small sample size of only 201 players and specifically examined the time until another match, while the other modelled survival times of high value players (whales) only. More than churn rate (proportion of players that defect), this study is interested in understanding when a player is at the highest risk of defection (time to churn), and what factors contribute towards that, so that remedial measures can be taken for player retention. This therefore is the gap which the research aims to fulfil, via analysing a considerable amount of real-time gameplay data with a view to identifying the risk of customer defection at any given time and the reasons for it. Therefore the research will endeavour to answer the question, that is, at what time points in the game progression are players most likely to defect and drop out and what causes this.

2.4 Cluster and Social Network Analysis of Players

The review conducted in this section serves to underpin the research objective four, pertaining to the identification of different types or clusters of users that constitute the customer base of online freemium games in terms of playing styles, performance and revenue generation.

Drachen, Sifa, Bauckhage & Thurau (2012) used cluster analysis techniques to high-dimensional telemetry data on player behaviour from two massively multiplayer online role playing games. They applied *k*-means and simplex volume maximization (SIVM) clustering algorithms to construct practical profiles of player behaviour after taking into account the game designs. Features that enabled determination of the most crucial mechanics of gameplay were chosen and they typically represented character performance, game features and play time. The study found that the *k*-means approach was effective in understanding the general distribution of player behaviours, whereas the SIVM technique was beneficial for describing players with extreme behaviours. A 6-7 cluster solution offered the best fit to the data with groups such as 'elites', 'stragglers', 'friendly pros', 'assassins' and 'veterans'. Online play time and levelling speed data were used by Drachen, Thurau, Sifa, & Bauckhage (2014) to create behavioural clusters employing a variety of unsupervised methods along with archetypal analysis with simplex volume maximization. Only two behavioural variables, number of days played and player level were incorporated in the different algorithms, and intuitively intelligible results were obtained from the *k*-means, *c*-means and archetypal analysis techniques. Principal component analysis and non-negative matrix factorisation resulted in not or only partly interpretable outcomes.

Hou (2012) adopted a cluster analysis approach for identifying the potential groups of behaviours of 100 gamers participating in an educational massively multiplayer online role playing game. Long-term detailed actions of users were collected, and behaviours such as teaming up, engaging in battles, learning, trading, and chatting were coded. A two-stage cluster analysis was performed on these coded behaviours, starting with a dendrogram based on Ward method in order to discover the optimal number of clusters, followed by the *k*-means cluster analysis. Three clusters were identified based on gamers' levels of participation (highest participation, high participation and ordinary participation), wherein the most committed group of gamers were found to be more attentive towards social interactions such as discussions with others and trading of items. The *k*-means technique was also employed by Tseng (2011) to segment a set of 228 online gamers in Taiwan, on the basis of their responses to a questionnaire for an online survey related to the motivations for playing online games including demographic data and other behaviours concerning online games. Exploratory factor analysis was used for reduction of dimensionality and to reveal the latent motivational factors, resulting in the emergence of two determinants - the need for exploration and the need for aggression. The *k*-means approach was used for player segmentation based on these two factors, which produced three groups that were significantly different in their consumer behaviours - the aggressive gamers, the social gamers, and the inactive gamers.

Bauckhage, Drachen & Sifa (2015) presented a detailed review focusing on the utilisation of three clustering methods – the *k*-means algorithm, matrix factorization methods, and spectral approaches, supported with examples of their applications to game related data.

They observed that the *k*-means technique achieved more extensively defined clusters in comparison to archetypal analysis. It directly assigned individuals to groups via cluster centroids, making it easier to interpret, whereas methods such as archetypal analysis, non-negative matrix factorisation and principal component analysis produce basis vectors spanning the space individuals belonged to. Inspite of the appeal of the *k*-means procedure, it suffers from a deficiency in terms of the method not being always feasible since it only works on averages. Some other studies that applied cluster analytic approaches to online game data were Halim, Atif, Rashid & Edwin (2017) who used data on the gameplay and the relationship between personality traits and players, to profile players into two clusters using four clustering techniques. They showed that gameplay can be used to forecast different aspects of personality using strategy game data. Saas, Guitart & Periánez (2016) applied time-series clustering algorithms based on measures such as dynamic time warping, discrete wavelet transform and agglomerative hierarchical method to temporal datasets of player activity in free to play games. They obtained clusters separately for the variables time played and purchases, and analysed the common player characteristics, demonstrating the extraction of sensible behavioural patterns of players.

Social network analysis may be an alternative approach to grouping the behaviours of users in online freemium games. Social networks, as explained by Borgatti, Everett & Johnson (2018), can be conceptualised as social systems that focus on the relationships among objects that comprise the system, which are called ‘actors’ or ‘nodes’. In the context of this research, actors are the players of online freemium games. When these actors (players) communicate with one another or form other links, the patterns of linkages can be used to derive clusters of actors who are connected and thereby considered similar by the principle of homophily (Newcomb, 1961). Borgatti et al. (2018) demonstrated the use of network variables as outcome variables, and a similar approach can be adopted in this study wherein network variables are used as the input variables in different clustering algorithms. Watts & Strogatz (1998) developed ‘small-world’ networks that can be highly clustered, and advanced metrics for assessing the cohesiveness of clusters. Longitudinal data recorded straight from a massively multiplayer online game was analysed by Ducheneaut, Yee, Nickell & Moore (2006) to understand gameplay and existing patterns of grouping of the players. The social environment present within the guild in the game was evaluated by building social networks to estimate the likelihood of sociability and quantification of joint activities within the guild. They found that some guilds were large enough to form a strong group of high performing players, and also that core members of a guild not only play with more number of guildmates, but also longer. Kirman & Lawson (2009) adopted a network analysis approach to study the social network within an online game and underline the most highly connected nodes as representative of the hardcore centre of the game. The data consisted of distinct interactions between 157 active players and a network graph was constructed based on nodes and distinct edges. The study established the existence of a distinct distribution of players and their playing styles, with the hardcore users constituting more than half of the interactions in the game. They also concluded that social games tend to be small world scale-free networks in which the growth of the user base followed a power law distribution.

Review of relevant literature identifying the different clusters of players reveals insights that will aid in the development of research questions associated with the relevant objective (four) of the study. It is evident from previously conducted studies that the behaviours of customers in an online gaming environment does indeed vary depending on their play style, competence and social interactions. The success of a game is likely to be dependent on identifying the type of users that constitute its player base, so that the game can be moulded accordingly to suit the varying tastes of its consumers. Past research have applied different clustering algorithms, of which the k-means approach was the predominant one, to identify various groups of players such as ‘assassins’, ‘stragglers’, highest participation group, ordinary participation group, ‘friendly pros’ and others. Network analysis was also performed in order to build and assess the social networks existing within games, and identify the patterns of inter communications between actors (or players).

The application of cluster and social network analyses of users is imperative to understand the widely varying customer base. Therefore, this research will attempt to use several different user gameplay features such as level of involvement with the game, performance in missions, rate of advancement within the game, and communications with fellow users to form unique groups of player behaviours. Different clustering approaches will be compared and evaluated to test their performance. The characteristics of these clusters will then be explored by means of different variables that will aid in defining the groups pragmatically, so that they can be targeted with a customised gameplay experience. Moreover, these clusters will be inspected with respect to their monetisation tendencies (micro transactions) with a view to classifying the high-paying profitable groups of players, an aspect not investigated by prior studies. Additionally, the social networks of players formed will shed light on valuable users that may potentially improve the popularity and virality of the games, also an area not focussed on by previous studies. This results in the formation of another research question, that is, what are the different categories of players that constitute the user base of freemium games in terms of playing styles, performance within the game and revenue generation. To this end, some classical cluster analysis approaches that may be useful in studying player behaviours as well as feasible to implement within an online gaming context are reviewed below.

2.4.1 K-means and K-medoids

One of the most popular and widely used clustering techniques is the k-means clustering algorithm (MacQueen, 1967; Anderberg, 2014) and an extension of that is the k-medoid algorithm (Kaufman & Rousseeuw, 1987). Both of these fall under the partitioning methods of clustering.

The process underlying k-means segregates data points into clusters, with a view to maximising intra-cluster likeness and minimising inter-cluster likeness, wherein the similarity (or likeness) within a cluster is measured by the average value of the members belonging to it (MacQueen, 1967). Although this technique is competent in handling large data sets, it suffers from a few drawbacks (Huang, 1998). It can only be used for data with defined mean, and fails if any of the variables used in the clustering procedure is categorical in nature (Nemala, 2009). Moreover, the number of clusters to be formed have to be pre-specified, and being a method that relies on mean values, it is easily affected by noisy data and outliers (Nemala, 2009). Considering these shortcomings, the k-medoids method was established as an improvement to k-means. Since one of the biggest shortfalls of k-means is that it fails to give correct results in the presence of outliers (as it works on averages), k-medoids was built to work on medoids rather than means. A medoid is the most centrally located data point in a data set or cluster that is representative of the data set or cluster and is always a member of it. The robustness of k-medoids over k-means stems from the fact that while k-means attempt to select the centre of the cluster, k-medoids choose the most centred data point in the cluster. As a result, extreme values will distort the true cluster centre in case of k-means whereas k-medoids will remain unaffected by them. Nevertheless, k-medoids is computationally expensive and require the user to supply in advance the number of clusters to separate the data into (Nemala, 2009).

2.4.2 K-modes and K-prototypes

Huang (1998) details two new developments over the k-means approach, and addresses the latter's deficiency with regards to application to categorical data. The first of these is the *k*-modes algorithm that is based on the statistical measure mode instead of mean and adopts a “simple matching dissimilarity measure for categorical objects” (p.285). It further attempts to reduce the cost function for clustering by updating the modes using a method formulated on frequency. The second is called the *k*-prototypes method, which merges the *k*-means technique with *k*-modes to define a dissimilarity measure that considers both quantitative and qualitative variables, thereby allowing for the grouping of mixed data sets. This approach is similar to the *k*-means in every aspect except that “it uses the *k*-modes approach to updating the categorical attribute values of cluster prototypes” (p.285)

A comparable procedure for clustering large data sets in high dimension is CLARA (Clustering for Large Applications) which was designed by Kaufman & Rousseeuw (1990). It incorporates a sampling methodology and the PAM (Partitioning Around Medoids) clustering algorithm, and can also handle categorical variables as it can apply any dissimilarity measure. However, Huang (1998) identified some advantages of their *k*-prototypes method over CLARA. While, CLARA uses a sampling procedure to cluster the data, *k*-prototypes can directly work on the entire data set. Consequently, optimisation of results from CLARA is also at the sample equivalent, which may not fundamentally translate to the actual data in case of a biased sample. Since CLARA works with samples, its competency depends on the same, and is likely to reduce with the size and complexity of the data. This shortcoming is clearly absent for the *k*-prototypes algorithm.

Huang (1998) further measured the effectiveness and scalability of *k*-modes and *k*-prototypes using real-world data sets. It was observed that the performance of both were adequate in distinguishing patterns in the data. The scalability tests revealed that the procedures were competent in grouping very big complicated data sets with respect to both the number of data points and the number of clusters. Overall, both algorithms had a linear increase in their run time with the increase in number of data points and number of clusters. Comparatively, *k*-modes tend to be much quicker than *k*-prototypes, given its discrete behaviour and requirement of much smaller number of iterations in order to converge.

2.4.3 Hierarchical Methods

In spite of their good efficiency and scalability on large data sets, the above clustering algorithms are efficient subject to prior awareness about the data, and suffer from the usual problem of determining the number of clusters naturally present in the data where such prior information is unavailable. Hierarchical methods of clustering are one of the simplest to perform computationally and do not experience the common drawback of having the need to pre-specify number of clusters (Nemala, 2009). Typically, the results are depicted in the form of a dendrogram (a tree-like structure).

Two types of hierarchical algorithms are commonly used – agglomerative and divisive. The agglomerative method takes a bottom-up route to clustering, whereby each data point is considered as a separate group initially, and based on a similarity function of groups nearest to each other, they are gradually combined "until the top most level of hierarchy is reached or until a termination condition holds" (Nemala, 2009, pp.12-13). Conversely, the divisive method adopts a top-down path, where the entire data set is considered as one parent group, and based on a degree of irrelevance, the most unrelated data point is disbanded from the main cluster and formed into a new cluster. This process is continued until the appropriate number of clusters is achieved or the inter-cluster distance between adjacent clusters is above a particular threshold distance (Nemala, 2009).

Few drawbacks of hierarchical techniques are discussed by Nemala (2009). The methods are deemed to be inflexible, and once a merge (agglomerative) or split (divisive) is performed, it is irreversible. This makes it computationally inexpensive but impossible to rectify erroneous decisions, leading to clusters of poor quality. It also suffers from the disadvantage of scaling issues because the procedure requires careful consideration and assessment of a large number of clusters before a split or a merge. The disadvantages of traditional hierarchical methods can be overcome by consolidating them with other suitable clustering techniques for mixed step clustering.

Zhang et al. (1996) described a clustering approach that was able to effectively handle very large datasets, called BIRCH - Balanced Iterative Reducing and Clustering using Hierarchies. The foundation of BIRCH lies in the idea of a clustering feature and CF tree. The algorithm as defined by Zhang et al. (1996) consists of four phases. Phase 1 mainly involves examining the entire big dataset and loading it into memory with the help of a CF tree that summarises the clustering information present in the data as accurately as possible by grouping together compact data points and eliminating scattered data points. This phase speeds up the subsequent phases by reducing the clustering problem ahead to a simpler problem and improves precision by removing outliers at the onset. Phase 2 is optional and focusses on investigating the leaf entries in the existing CF tree from in an attempt to build a smaller tree. Phase 3 applies a global or semi-global algorithm to cluster all leaf entries. The groups obtained after completion of phase 3 accumulates the distribution pattern dominant in the data largely, although small-scale errors may exist. Therefore, in the final phase 4, refining of the clusters takes place, which again is not mandatory and requires additional costs. The study by Zhang et al. (1996) also tested the performance of BIRCH regarding certain parameters, observing that the method worked in a steady manner unless the initial threshold value is disproportionately high relative to the data set.

2.5 Approaches for Handling Online Data

As explained by Adler, Feldman & Taqqu (1998), data collection has resulted in the emergence of two distinct categories, good data and bad data, wherein, good data is easy to explore and analyse and can be modelled using standard probability distributions, while bad data is usually highly skewed, contains important outliers and is often incomplete with missing values. Thus, it is not straightforward to assess this type of data and there is an obvious lack of well-developed statistical techniques to analyse it (Adler et al., 1998).

Based on prior experience of working with gameplay data from online freemium games, it is assumed that the data to be used in this research will be massive, typically millions of events on thousands of players, the analysis of which is naturally cumbersome with respect to frequentist methods and modelling. Several game events are rare or infrequent which many players do not attempt or do so only a few times. Data collection on online events in real time can suffer from lagged, duplicated or missing events. Therefore, data on gameplay events, at times emerge as “bad data” as they are likely to contain null values or be heavily influenced by zeroes or contain variables with heavy tailed or unknown distributions. This would typically require a lot of cleansing before any analysis could be performed. Some distributions that have been used to model such data have been discussed below.

2.5.1 Modelling Rare or Unpopular Events

In online gaming, initiating a real currency micro transaction, which is the subject of the third research objective, is a rare event, and in this section the statistical modelling of such rare events is reviewed.

Probability distributions such as Zipf, Power Law and Pareto are applicable in representing incidents where the occurrence of large events is rare while that of smaller events are fairly frequent (Adamic & Huberman, 2002). Online freemium games are often seen to follow a similar trend in that, certain in-game actions are seldom performed by players. For instance, players may rarely indulge in social interactions such as sending or accepting friend invites or messages; they may not explore some areas of the gameplay as much as others; and may also not indulge in in-game transactions often. In such scenarios, small values for variables associated with the above aspects tend to be quite common whereas high values for the same will be rather infrequent. The above mentioned probability distributions can then be assumed for this type of data.

Zipf's law commonly specifies the frequency of an event happening in relation to its rank; while Pareto's law is more concerned with income distributions, and addresses the frequency of cases with an income greater than a particular value rather than the rank of income; and the power law distribution explains the total count of cases with an income exactly equal to a particular value instead of that greater than the said value (Adamic & Huberman, 2002).

Adamic & Huberman (2002) investigated online data such as frequency of visits to a website, count of pages within a website and number of links to a page within a site, that were widely found to exhibit power law distributions. Zipf's law was found to be the standard when considering internet data, such as selection of websites visited by individuals and formation of virtual societies between them (Adamic & Huberman, 2002). Cha, Kwak, Rodriguez, Ahn & Moon (2007) investigated the distribution of non-popular video content in User Generated Content (UGC) services, in order to address whether UGC popularity followed a Power Law distribution. It was found that the distributions of popularity of videos across four typical classes on Youtube and Daum (a web portal) all follow the Power Law model, while non-popular videos on Netflix followed a distribution that was not Power Law (Cha et al., 2007).

2.5.2 Modelling Distributions that are Skewed and Contain Excess Zeroes

Typically, probability distributions of dependent variables in online gameplay data, such as total time spent playing the game or magnitude and number of micro payments are heavily left skewed, and special data treatment is required in such cases. Incorporation of methods to deal with heavy skew and an over preponderance of zeros are reviewed in the next section.

Zero-inflated models are a class of adjusted count models that aim to explain for the excess zeroes in a probability distribution, and consists of two processes estimating two equations, one producing the zeroes and one producing the positive values (Winkelmann, 2008). Hurdle models are another class of models that can explain the excess zeroes in a distribution, and are based on the concept that the binary outcome of a zero or positive value is driven by a binomial probability model, while the positive values have a conditional distribution driven by a zero-truncated count model (McDowell, 2003).

Skewed distributions frequently occur when dealing with real world data. Bi, Faloutsos & Korn (2001) suggest a novel probability distribution called the Discrete Gaussian

Exponential (DGX) to obtain exceptional model fits of a data in extensively variable frameworks. Bi et al., (2001) conducted a study to develop a statistically robust process for parameter estimation of the DGX model using a maximum likelihood estimator, wherein the model was tested on vastly different real world data sets including internet click stream data, and it was found that for all instances the DGX model efficiently fit the data distributions.

2.5.3 Statistical Modelling of Gameplay Data

Understanding the way people play online freemium games is crucial in determining whether they enjoy the game or find it monotonous, the degree of their engagement with the game and consequently the risk of defection from it, and the likelihood of making real currency micro transactions. This section deals with methods that provide insights regarding the above and is therefore relevant to research objectives one, two and three.

There are several variables comprising a dataset of in-game events that reflect on the above measures. Some of these include but are not limited to, the total time played, number of game events triggered, number of missions successfully accomplished, other measures of competency against opponents, use of virtual resources available in the game etc. The development of statistical models of these variables is necessary to make estimation and predictions about player engagement, time until defection from the game and inclination to make real monetary purchases. Some classical modelling approaches that are able to address these and are also implementable in real-life are reviewed below.

2.5.3.1 Multiple Logistic Regression

Regression models are one of the most common and widely used modelling techniques that allow prediction of a response variable from one or multiple explanatory (or dependent) variables, and are simple to perform and inexpensive in terms of processing time or time/money spent for data collection (Stockburger, 2001). A specific class of regression models, called the multivariate logistic regression, is used to predict a response variable with binary outcome (success or failure) with the help of multiple explanatory variables Pampel (2000). The multivariate logistic regression approach is deemed to be apt for modelling user engagement with possible outcomes, engaged (representing success) or not engaged (representing failure). A study by Aguilera, Escabias & Valderrama, (2006) adopted a “reduced set of optimum principal components of the original predictors” (p.1) as covariates of the logistic model, and found an enhancement in the estimation power of the parameters of the logistic model even in the presence of multicollinearity, and also addressed the issue of large number of predictor variables by facilitating dimensionality reduction. Instead of the usual maximum likelihood estimation, Aguilera et al., (2006) used principal components regression and partial least squares linear regression to scale down the model dimensions and make parameter estimation more accurate.

The regression approaches discussed are important in helping address the first and third research objectives, that of understanding player engagement and number of micro transactions made.

2.5.3.2 Survival Modelling

Survival time refers to the time taken for a particular event of interest to occur, and the main aim of investigating survival data is to be able to predict the probability of the response variable i.e. the survival time (Lee & Wang, 2003). In the context of this research, the event of interest is time until defection from the game for customers. The following discussion is focused on the second research objective, that of determining the time point in game progression that players are most likely to defect and drop out and what factors drive this

The Kaplan-Meier method of estimating survival times is a non-parametric maximum likelihood approach that does not require the assumption of an underlying probability distribution, however allows no scope for considering any other explanatory variables that might have an influence on the survival curve (Bewick, Cheek & Ball, 2004). This can be overcome by an alternative model, Cox's proportional hazards model, which does not assume anything regarding the underlying probability distribution of the hazard function and the hazard ratio between two groups of observations is independent of time (Bewick et. al., 2004). In case of a large number of explanatory variables and relatively smaller number of samples, a study by Datta, Le-Rademacher & Datta (2007) established partial least squares and least absolute shrinkage and selection operator as useful methods for modelling survival data, with the latter being superior with regards to error in prediction, when the set of explanatory variables contain a modest to substantial proportion of irrelevant or noisy variables.

2.6 Summary of Findings from the Literature

The literature review undertaken in this chapter helps to understand the nature and amount of work that has been done in this particular field of research. A brief summary of the learnings from the literature review, relevant to the aim and scope of this research presented here.

Adopting a data-driven approach to player behavioural modelling is a much better alternative than surveys, small-scale observation experiments or knowledge engineering (Harrison et al, 2011). Player engagement is one of the most important criteria for successful games and is crucial in adding value to player experience (Jennett et al., 2008) and (Schoenau-Fog, 2011). Majority of the studies conducted to understand player engagement or investigate player immersion in games have been accomplished through surveys (Schoenau-Fog 2011), questionnaires (Jennett et al., 2008, Yee 2006, Cole & Griffiths, 2007), focus groups (Poels et. al., 2007) and interviews (Brown & Cairns, 2004). Although Medler et al. (2011) collected player events data from actual gameplay, it was used it to create a visual game analytic tool for studying player gameplay behaviour. Predictive modelling of user engagement using real-time gameplay data was not extensively investigated in the literature. This gave rise to a question that this research will attempt to answer, that is, what gameplay behaviours in online freemium games significantly predict increasing engagement amongst its users.

Predominant methods for the analysis of customer defection in online games included correlations between players' behaviours (Tarng et al., 2008), statistical associations (Kuss et al., 2012), visualisations of hazard and survival plots (Chen et al., 2009), logistic regression models (Chen et al., 2009), diffusion models (Kawale et.al., 2009), binary classification tests and comparisons with neural networks, logistic regression, naïve Bayes and decision trees (Hadiji et al., 2014), lifetime analysis identifying a Weibull distribution of total gameplay times (Bauckhage et al., 2012), mixed effects Cox regression models (Demediuk et al., 2018), and survival ensembles with conditional inference trees (Periáñez et al., 2016). Generally studies are found to have examined game play times and churn rate, rather than the time to churn. Those that studied the latter did not use a wide range of behavioural variables, or specifically modelled the survival times of high value players (whales) only. More than churn rate, this research is interested in understanding when a player is at the highest risk of defection, also known as the time to churn, and what factors contribute towards that, so that remedial measures can be taken for retention. Therefore it will endeavour to answer the question, that is, at what time points in the game progression are players most likely to defect and drop out and what causes this.

Some of the analyses of behaviours of players regarding real currency micro transactions included investigations into gambling tendencies of players (King et al, 2015), understanding the causes of virtual products purchase through a subjective marketing approach (Hamari & Lehdonvirta, 2010) and study of virtual economies through fitting of economic models to the data (Heeks, 2009). Statistical modelling of the number of real currency micro transactions made by players was not found to be examined in the literature. Therefore in this research, transaction behaviours will be analysed using regression modelling of variables reflecting the tendencies of users to invest in micro purchases within the game, attempting to answer the question, that is, what facets of the player experience promote an increase in the quantity of real currency micro purchases by players.

Previously conducted studies on cluster analysis made it evident that the behaviours of customers in an online gaming environment does indeed vary depending on their play style, competence and social interactions. Identifying different groups of users that constitute the player base will contribute towards the success of the game by allowing for it to be moulded accordingly to suit the tastes of its various audiences. This makes the application of cluster and social network analyses of customers crucial, which the research intends to implement in order to address another research question, that is, what are the different categories of players that constitute the user base of freemium games in terms of playing styles, performance within the game and revenue generation.

An examination of relevant statistical approaches identified some useful models for complex and skewed data sets. The zipf, power law and pareto were identified for rare or unpopular events (Adamic & Huberman, 2002), hurdle and zero-inflated models for distributions with excess zeroes, and discrete gaussian exponential models for skewed distributions (Zhiqiang et. al., 2001). Performing survival analysis at different periods in time, for example after a major patch or character rework, may provide deeper insight into their impact on player churn.

Overall, it was found that the existing literature related specifically to statistical analysis and modelling and clustering of player behaviours of online freemium games based on real time game events is sparse and inadequate. The research intends to overcome this by attempting to fit probabilistic statistical models to data arising from actual gameplay of users within an online freemium game, and subsequently address the research aims and questions.

2.6.1 Research Questions

As already established, the enormous business driving the online freemium games market and its contribution to the worldwide economy, employment and society makes it a lucrative industry to focus on. Downloadable free of cost but regulated by micro-payments, these games add to the intricate relationship between game design and business planning, leading to the formation of systems that manage the sequential flow of player experience to monetise player immersion (Evans, 2016). Knowledge gained from the KTP project revealed that the prime interest for freemium game studios is to understand what makes their customers engage with their games and drives them to make real currency transactions. This insight is believed to be beneficial in the development of games that maximise user engagement and monetisation, thereby increasing profitability and success. Majority of game studios were found to be data-driven, in addition to being creative-driven, with processes implemented within the game that enabled them to capture real-time gameplay data from its users. As highlighted by Luban (2011), freemium games are seldom released into the market with complete content, and hence gameplay data from users is necessary in the improvement of the game and revival of its popular features.

Motivated by an extensive review of the literature and identifying potential gaps in it, the following research questions are what this study intends to address, using real-time data on gameplay events and quantitative analysis approaches. As determined in the literature review, this is a more robust approach as it will eliminate data gaps and biases that typically arise from qualitative methods of data collection (Harrison et al, 2011).

1. What gameplay behaviours in online freemium games significantly predict increasing engagement amongst its users?

Development of statistical models to explore and identify the specific aspects of users' gameplay that cause engagement or not, which can then guide the creative design process of games in making the experience more appealing and enjoyable to its consumers thereby minimising attrition.

2. At what time points in the game progression are players most likely to defect and drop out and what causes this?

Development of survival models to investigate this, as anticipating when certain players are about to drop out will enable developers to customise their game, targeting these players with assistance to overcome any obstacles in their progression, which may cause defection.

3. What facets of the player experience promote an increase in the quantity of real currency micro purchases by players?

Examination of existing purchasing trends of customers in online freemium games (operating on micropayment revenue model), thereby establishing a regression model to determine the incentives for the number of real currency transactions that will benefit ARPU.

4. What are the different categories of players that constitute the user base of freemium games in terms of playing styles, performance within the game and revenue generation?

Cluster analytic and social network techniques for identifying the wide variety of players that constitute the user base of online freemium games, including additional scrutiny of the player groups with respect to their playing pattern, performance and value added in terms of proceeds and virality.

In order to achieve the research aims and answer the above questions, and following on from a review of the literature on online gameplay data and relevant statistical approaches, the study will focus on the following steps –

- Collection of data representing players' game events over time from a typical online freemium game called eRepublik
- Development of a statistical framework for analysis of user behaviours and tendencies by adapting suitable statistical modelling techniques to address the research questions
- Utilising the analysis results to provide implementable recommendations to games publishers and developers for utmost engagement of their users and customisable games that will optimise their revenue

These will be fulfilled through the research methodology and methods elucidated later in the thesis.

3. Research Methodology

Provided in this chapter is a detailed account of the methodology adopted to conduct the research, in order to meet the objectives and answer the research questions. Brown (2006) defines methodology as "the philosophical framework within which the research is conducted or the foundation upon which the research is based" (p.12). Glatthorn and Joyner (2005) emphasize three concepts relevant to research methodology as research perspective, research type and research method. These are elucidated later in the chapter. First, the general purpose of this study is reiterated.

3.1 Revisiting Research Aims

The overall aim of this study is to develop suitable data-driven methods to gain insight about consumer behaviour in online freemium games, with a view to provide recommendations for successful business in the freemium games industry. This is expected to be achieved by analytics of in-game player data leading to the construction of statistical models of online behaviours. The four aspects of customer behaviour that the research addresses are – engagement, monetisation, attrition and classification.

The objectives of the research as described before are to review the theoretical concepts of relevant statistical methods in customer behaviour analysis, use players' gameplay data to construct a statistical framework for analysis of behaviours by adopting suitable modelling techniques, and thereby provide implementable recommendations for improved business in the online freemium games industry.

3.2 Research Philosophy or Research Perspective

Review of existing literature on player behaviour in video games revealed that investigation of player immersion has usually been accomplished through surveys, questionnaires and focus groups. Harrison and Roberts (2011) demonstrated that adopting a data-driven approach to player behavioural modelling could more conveniently produce models that are more robust than using surveys, small-scale observation experiments or knowledge engineering. Moreover, gameplay data from existing players can be used to build predictive models for behaviours of future players (Harrison & Roberts, 2011).

Thus, the research perspective adopted for this study is the quantitative perspective, which as stated by Glatthorn and Joyner (2005), is deduced from a positivist epistemology. The positivism research philosophy (Bryman & Bell, 2015) is characterised by rational and empirical research (O'Leary, 2004; Glatthorn & Joyner, 2005) and post positivism follows a deterministic principle in which causes are likely to drive effects or outcomes (Creswell, 2009). The study is experimental in nature, focusing on data collection, quantifiable observations or measurements leading to statistical analyses and correlational research (Glatthorn & Joyner, 2005; Dudovskiy, 2016a). It follows deductive reasoning, which starts with the construction of a theory based on prior knowledge and review of previous work (Babbie, 2015) on online gameplay behaviours and their influence on the business in freemium games industry. By adopting a positivism paradigm based on deductive approach, the research study is able to stipulate primary concepts and variables (Babbie, 2015), recognise and evaluate the causes that drive outcomes (Creswell, 2009), quantitatively measure theories and concepts (Bryman & Bell, 2015) and moderately generalise the findings (Dudovskiy, 2016b). The philosophy followed in this research will aid in fulfilling the purpose of explaining and predicting (Dudovskiy, 2016a), being empirically discernible and developing and testing hypotheses during the research process (Dudovskiy, 2016a), as well as being generalised through statistical probability utilising large random samples (Ramanathan, 2009).

3.3 Research Type

Research type as explained by Glatthorn and Joyner (2005) is used to determine the predominant research approach employed in a study. Since this study follows a quantitative perspective, the appropriate research type adopted is a combination of descriptive, quasi-experimental and causal-comparative research.

The descriptive approach is undertaken right after data collection in order to depict the identified variables and phenomena (Glatthorn & Joyner, 2005; BCPS, 2017) and is particularly beneficial in the preliminary stages of the research (Glatthorn & Joyner, 2005). In this stage, exploratory data analysis is performed and a description of the prevailing customer behaviour in the game is outlined in terms of important variables that represent these behaviours. Relationships between variables are not examined at this stage of the research (Glatthorn & Joyner, 2005).

A quasi-experimental and causal-comparative approach constitutes the major part of this study. As elucidated by Glatthorn and Joyner (2005) and BCPS (2017), causal-comparative research attempts to analyse the reasons behind the occurrence of an event through demonstrating a cause-effect relationship between variables. These studies are also known as “ex post facto research” since the causes leading to an event are examined after they have produced the event (Glatthorn & Joyner, 2005). From the perspective of this study, causes of events such as customer engagement, customer defection, and in-game transactions are investigated after the occurrence of these events. This causal-comparative approach is undertaken in conjunction with a quasi-experimental approach involving the identification of explanatory (or independent) variables that already exist in the domain of customer behaviour in online freemium games, and are not contrived by the researcher (BCPS, 2017). Statistical modelling of relevant dependent variables is then performed using independent variables that are expected to have an effect on these outcomes. As construed by Glatthorn and Joyner (2005) and BCPS (2017), the quasi-experimental approach (as adopted in this study) does not incorporate random assignment of groups and aims to explore pre-existing ones that are characteristic to consumer behaviour in online games.

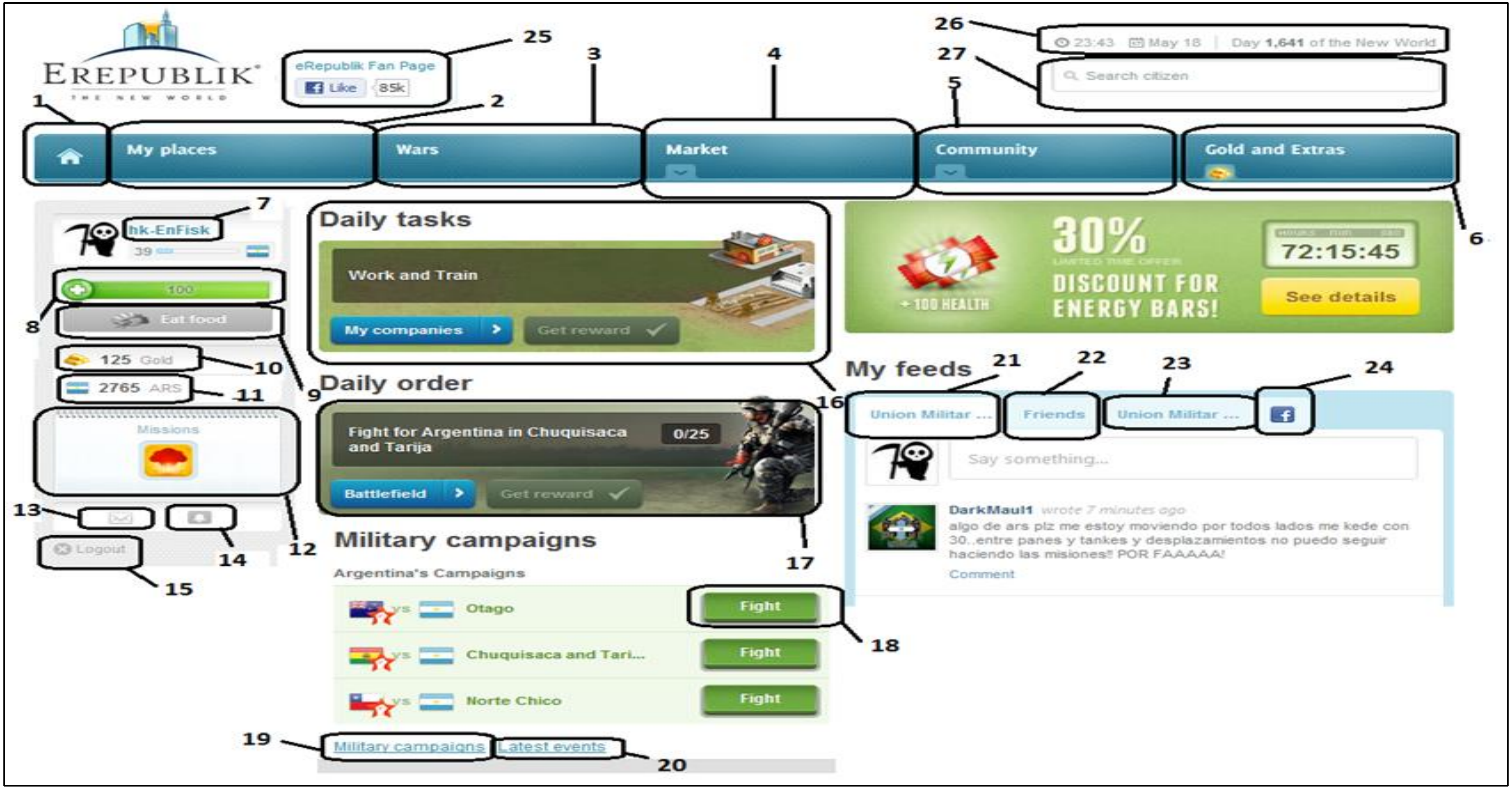
3.4 Research Method

Data collection, description and analysis form the crux of research methods (Dudovskiy, 2016a), which are described by Walliman, (2011) as “the practical techniques used to carry out research” (p.29). The following sections present an outline of the methods that will be undertaken in this research and detailed in the upcoming chapters

3.4.1 Data Collection and Description

The data for this study is obtained from a conventional online freemium game called eRepublik, which incorporates most aspects of online freemium games concerning its overall structure, gameplay and revenue model.

eRepublik is a freemium massively multiplayer online (MMO) game that is browser-based and can be played on the web. According to eRepublik (2015), it is a strategy game, taking place in a virtual world called ‘The New World’ in which players are considered as ‘citizens’. There are different game-playing approaches that players can choose. They can survive as private citizens, working fighting and voting for their state. They can participate in the local and national politics through voting for suitable leaders or becoming one themselves. They can also formulate economic and social policies for their nation, as well as attack neighbouring countries by launching wars against them.



DarkMaul1 wrote 7 minutes ago

algo de ars plz me estoy moviendo por todos lados me kede con 30..entre panes y tankes y desplazamientos no puedo seguir haciendo las misiones!! POR FAAAAA!

Figure 3.1: Main gameplay screen of eRepublik (eRepublik, 2015)

As is evident from a screenshot of the game in Figure 3.1, it has a variety of areas for players to focus on and develop their gameplay strategies. The ‘Wars’ tab can be used to get involved in fights with other players, ‘Market’ will allow purchase of food, weapons etc. for survival, ‘Community’ is where players can perform political and media related activities, and ‘My places’ will provide them with opportunity to train and upgrade. The game features two virtual currencies - the premium currency ‘gold’ that has to be purchased with real money, and the grind currency ‘national currency’ that can be acquired during game progression. Thus, with its wide variety of game content, eRepublik is a classic example of an online freemium game that can facilitate analysis of different user behaviours for studying customer engagement, monetisation and defection.

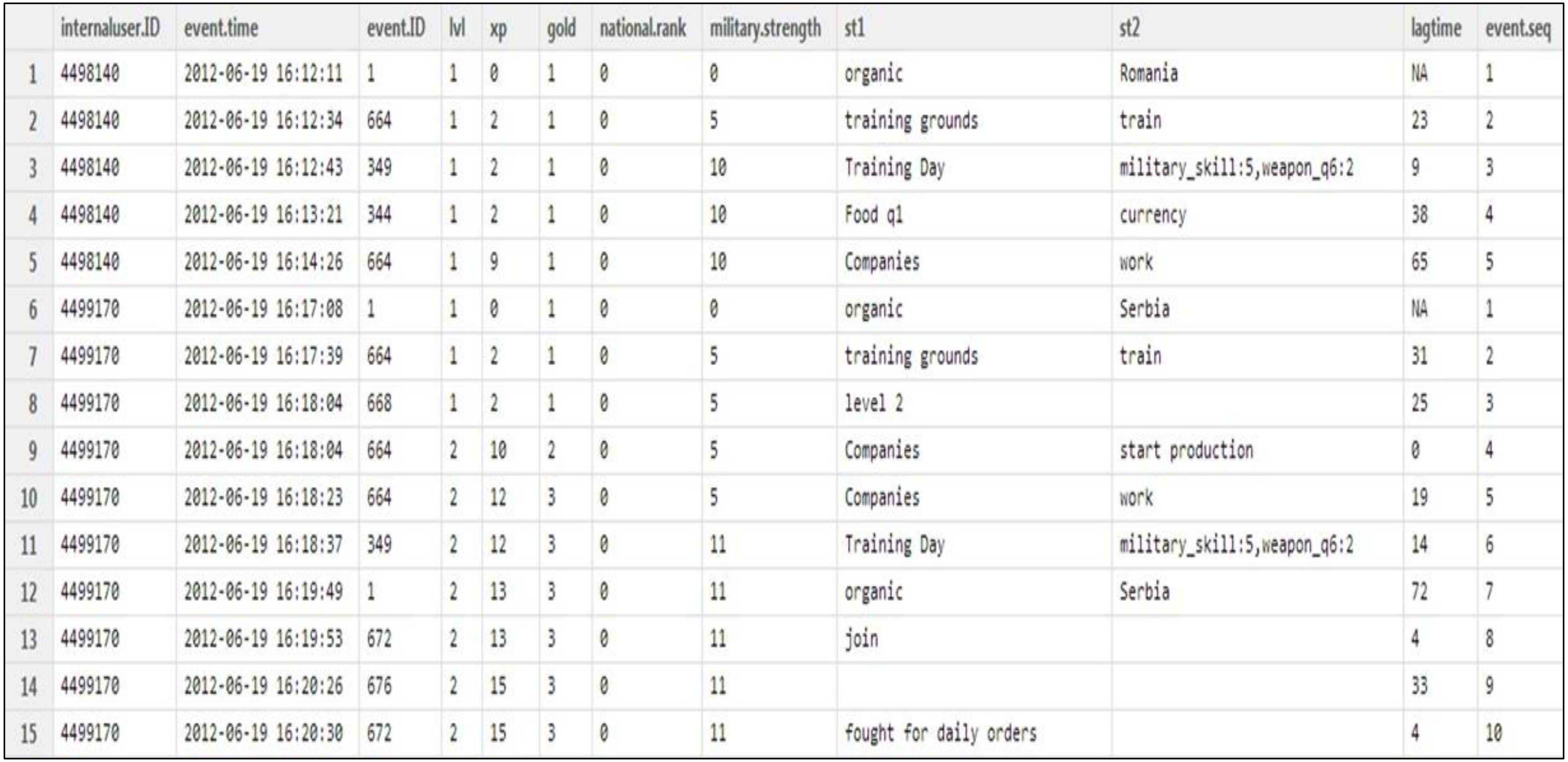
eRepublik data used in this research was provided by a UK based games analytics and marketing company. The company in turn acquired this data from the games publisher eRepublik Labs who was their client. The game developers at eRepublik used suitable techniques to track and log in-game events of players, which were then compiled and stored on the analytics company’s cloud servers, and transferred to the researcher in the form of a CSV file. The file is then read in to the statistical programming language R 1.1.456 (R Core Team, 2017) for further examination and analysis.

The data represents in-game events triggered by players as they install, login and advance within the game. It is continuous, time-stamped and recorded in real-time.

Illustrated in table 3.1 is an extract of the raw events data from eRepublik. The original, is big data that is typically a flat file of event history consisting of records for 40716 players on 101 variables. This amounts to more than 3 million rows of observations (3204027). Each row corresponds to an event triggered in real time by an individual player. The columns describe the variables associated with that particular event.

Intricate details regarding the collection, storage and description of data used in this study are extensively discussed in the next chapter.

Table 3.1: A subset of the raw events data from eRepublik



	internaluser.ID	event.time	event.ID	lvl	xp	gold	national.rank	military.strength	st1	st2	lagtime	event.seq
1	4498140	2012-06-19 16:12:11	1	1	0	1	0	0	organic	Romania	NA	1
2	4498140	2012-06-19 16:12:34	664	1	2	1	0	5	training grounds	train	23	2
3	4498140	2012-06-19 16:12:43	349	1	2	1	0	10	Training Day	military_skill:5,weapon_q6:2	9	3
4	4498140	2012-06-19 16:13:21	344	1	2	1	0	10	Food q1	currency	38	4
5	4498140	2012-06-19 16:14:26	664	1	9	1	0	10	Companies	work	65	5
6	4499170	2012-06-19 16:17:08	1	1	0	1	0	0	organic	Serbia	NA	1
7	4499170	2012-06-19 16:17:39	664	1	2	1	0	5	training grounds	train	31	2
8	4499170	2012-06-19 16:18:04	668	1	2	1	0	5	level 2		25	3
9	4499170	2012-06-19 16:18:04	664	2	10	2	0	5	Companies	start production	0	4
10	4499170	2012-06-19 16:18:23	664	2	12	3	0	5	Companies	work	19	5
11	4499170	2012-06-19 16:18:37	349	2	12	3	0	11	Training Day	military_skill:5,weapon_q6:2	14	6
12	4499170	2012-06-19 16:19:49	1	2	13	3	0	11	organic	Serbia	72	7
13	4499170	2012-06-19 16:19:53	672	2	13	3	0	11	join		4	8
14	4499170	2012-06-19 16:20:26	676	2	15	3	0	11			33	9
15	4499170	2012-06-19 16:20:30	672	2	15	3	0	11	fought for daily orders		4	10

3.4.2 Data Analysis

Investigation and scrutiny of data form the most crucial aspect of this research. As a quantitative study, all inferences surrounding the research questions are drawn from appropriate analysis of data. All analysis in this study are conducted using the statistical programming language R 1.1.456 (R Core Team, 2017). An overview of the data analysis stages is presented below which are covered in detail in the subsequent chapters.

3.4.2.1 Data Cleaning and Exploratory Analysis

Once the data is acquired and read in to R, it is first processed and refined for analysis. This is because online data recorded in real-time is likely to be exposed to technical glitches during the data collection procedure leading to missing, repetitive or incorrect information. Some instances of this could be – duplicate events (i.e. the same event triggered at the same time by the same user), time lag between events, variable values not being updated with the generation of a new event, etc. Moreover, the data inherently may record rare or infrequent events resulting in highly skewed distributions of variables or those heavily influenced by zeroes and extreme values. Hence, prior to performing core analyses, it is imperative to process the data to make it suitable for advanced statistical investigations.

The data cleansing is followed by exploratory data analysis, which includes rudimentary procedures to uncover essential and integral characteristics of the data in order to develop a robust understanding of subsequent examinations (Cox, 2017; Hartwig & Dearing, 1979). Chapter 4 deals in depth with this. In essence, it involves elimination of duplicate data, estimation or elimination of null and missing values and probability distributions of potential variables of interest.

3.4.2.2 Statistical Modelling of Outcomes of Interest

This section addresses the research aims and objectives. It principally involves building statistical models of player behavioural data to study and predict the three important facets associated with successful games – player retention, time to defect and micro-payments by players.

Response variables indicating players' engagement with the game, the total time played before defecting and the frequency of real currency transactions are defined. These outcomes are then modelled accounting for certain explanatory variables that depict players' interactions with the game. Thorough examination of the influence of these variables on the outcome of interest is undertaken along with an attempt to predict the response from them. Some of the models considered are – multiple logistic regression to examine player engagement, survival models (for example Cox's proportional hazard model) for time to defect from game, and count models for player transactions. Model fitting to relevant data is followed by model diagnostics, and observation and evaluation of the results.

3.4.2.3 Classification of Player Behaviours

A widely varied player base is recognised in the online freemium games genre that trigger diverse customer behaviours regarding engagement, monetisation and defection. A comprehensive insight into these player characteristics will aid in establishing the motivations for making a purchase in the game or staying associated with it. It will also enable profiling of players thereby producing recommendations for customisable games.

A multivariate method of classification called cluster analysis is explored to address this. Variables representing different playing styles and competencies are selected based on knowledge derived from playing the game as well as exploratory analysis and modelling. Frequentist clustering techniques using hierarchical algorithms and partitioning algorithms (*k*-means and *k*-medoids) are applied in the first instance. In case of a combination of quantitative and qualitative variables, *k*-modes and *k*-prototypes are used. Finally, social network analysis is performed for classification of users and the results compared with the above techniques.

3.4.2.4 Validation of Statistical Models

On completion of the model building procedure, it is essential to establish the validity of these models as the best representation of given data. The efficiency of model performance is assessed in terms of computational costs and processing time, ease of implementation in real time and accuracy of prediction. Model strength is examined using statistical significance, robustness of results, explanatory power, forecasting power from a sample and model simplicity.

Cross-validation is a widely recognised practice in model evaluation methods. It is used in logistic regression (Hahs-Vaughn & Lomax, 2012), estimation of prediction errors (Efron & Tibshirani, 1994) and in general evaluation of predictive models (Konishi & Kitagawa, 2008). Hahs-Vaughn and Lomax (2012) state that when the sample size is adequate, 75%-80% of the data can be used as a primary sample to determine the model while the remaining data can serve as a holdout sample to ascertain model accuracy. "If classification accuracy of the holdout sample is within 10% of the primary sample, this provides evidence of the utility of the logistic regression model" (Hahs-Vaughn &

Lomax, 2013, p.1114). Efron and Tibshirani (1993) assert that, prediction error is more realistically estimated by using a test sample that is independent from the training sample.

This study adopts the cross-validation technique to test model validity. A part of the available data (training sample) is used to construct the statistical models while the remaining data (test sample) is used to validate those. Cluster analysis of playing styles will also be assessed using relevant validation techniques since classification problems give rise to prediction errors (Efron & Tibshirani, 1993).

This is a crucial step to the research, as it will facilitate the development of an overall analytical framework for customer behavioural data in the field of online freemium games.

3.5 Research Ethics

The ethical considerations of this research are elucidated as follows.

This is an independent research conducted with the sole purpose of adding knowledge to the area of application of statistical methods to investigate big data representing consumer behaviours online. The data is obtained with permission from an analytics company that in turn receive it from the game studio eRepublik Labs who is their client. Collection of customer data is extensively explained in the privacy policy of eRepublik Labs that users sign up to once they decide to install and log in to the game. This study ensures the protection of the privacy of users. The data used does not contain any information that compromises the confidentiality of the customers. No demographic or personal data is collected and users are allocated unique IDs that are numeric in nature, which ensures that they are non-identifiable. Efforts are undertaken to guarantee that all analyses and discussions in this research study are objective and unbiased.

3.6 Summary

Table 3.2 summarises the overall research methodology, linking the research questions specified in the previous chapter with the method and analysis

The following chapters detail the collection, storage and exploratory analysis of the data; development of statistical models to examine user engagement, time to defect and monetisation; classification of players and their playing styles via clustering techniques; evaluation and validation of all methods adopted in the study; and the final conclusion from this research.

Table 3.2: Research questions linked to methodology

Research Question	Methodology and Method	Analysis
What gameplay behaviours in online freemium games significantly predict increasing engagement amongst its users?	Real-time data indicating if a player is engaged or not and comprising other gameplay related variables. Predictive modelling and validation	Multiple logistic regression
At what time points in the game progression are players most likely to defect and drop out and what causes this?	Real-time data on the churn status of players (whether they have defected or not) and comprising other gameplay related variables. Lifetime analysis and survival modelling and validation	Kaplan-Meier estimation and Cox proportional hazards model
What facets of the player experience promote an increase in the quantity of real currency micro purchases by players?	Real-time data on the number of real currency micro transactions and comprising other gameplay related variables. Count regression models and validation	Discrete models including zero-inflated and hurdle models
What are the different categories of players that constitute the user base of freemium games in terms of playing styles, performance within the game and revenue generation?	Real-time gameplay data covering various aspects of player behaviour Cluster analysis and social network analysis	Hierarchical and partitioning algorithms for clustering and network analysis using sociograms

4. The Data and Exploratory Analysis

This chapter is an exhaustive illustration of the data and its characteristics. An understanding of the nature of data to be used in the analysis and modelling is a necessary pre-requisite for proceeding on to more advanced and complex methods that will address the research questions. Complete knowledge about the structure and composition of the data is expected to be useful in guiding decisions related to the statistical models and approaches adopted within the research.

4.1 Collection and Storage

Data used for this research study arises from a freemium MMO game called eRepublik, published by the studio eRepublik Labs. It is acquired by the researcher from a games analytics and marketing company, which in turn obtains it from the game studio itself.

Typically, data collection from games takes place through "play testing a limited population" (Medler, John & Lane, 2011, p.3). However, analysis of this data can be tedious and complicated with regards to statistical significance, in which case a useful alternative would be telemetric recording of gameplay data which enables large amounts of it to be accumulated and analysed (Medler et al., 2011). Gameplay data originates during the course of users actually playing a game (Medler, 2011), via telemetric software built into the game's programming that tracks and archives the actions undertaken and in-game events triggered by players (Medler et al., 2011; Thompson, 2007). Medler et al., (2011) describe a typical metric logged by games as "when a player begins a level or performs an action like jumping while in the midst of gameplay" (p.1). They also explain the data collection process for a game called Dead Space 2 as documenting events generated by players "by adding telemetric "hooks" or functions into a game's code which send event data to a separate server location whenever an event in the game is triggered" (p.3). Similar techniques are used by the game developers at eRepublik Labs to track and log in-game events, which are then passed on to the analytics company's servers.

Continuous raw data from the game in real time is initially stored on the company's Amazon cloud servers. A fragment of this, representing only new users and their gameplay for an adequate amount of time, is then downloaded using Microsoft SQL server and transferred to the researcher in the form of a CSV file. This data is primarily stored in a University laptop, while backups are created and saved on an external hard drive and the University's cloud storage for students (OneDrive).

Medler et al., (2011) affirms that behaviour data is valuable for game designers in order to investigate playing styles and is by far more effective than data from "self-reporting surveys or controlled play tests" (p.1). This supports the fact that the collection and nature of the data used in this research will facilitate robust analysis of player behaviours through statistical modelling and other sophisticated analyses.

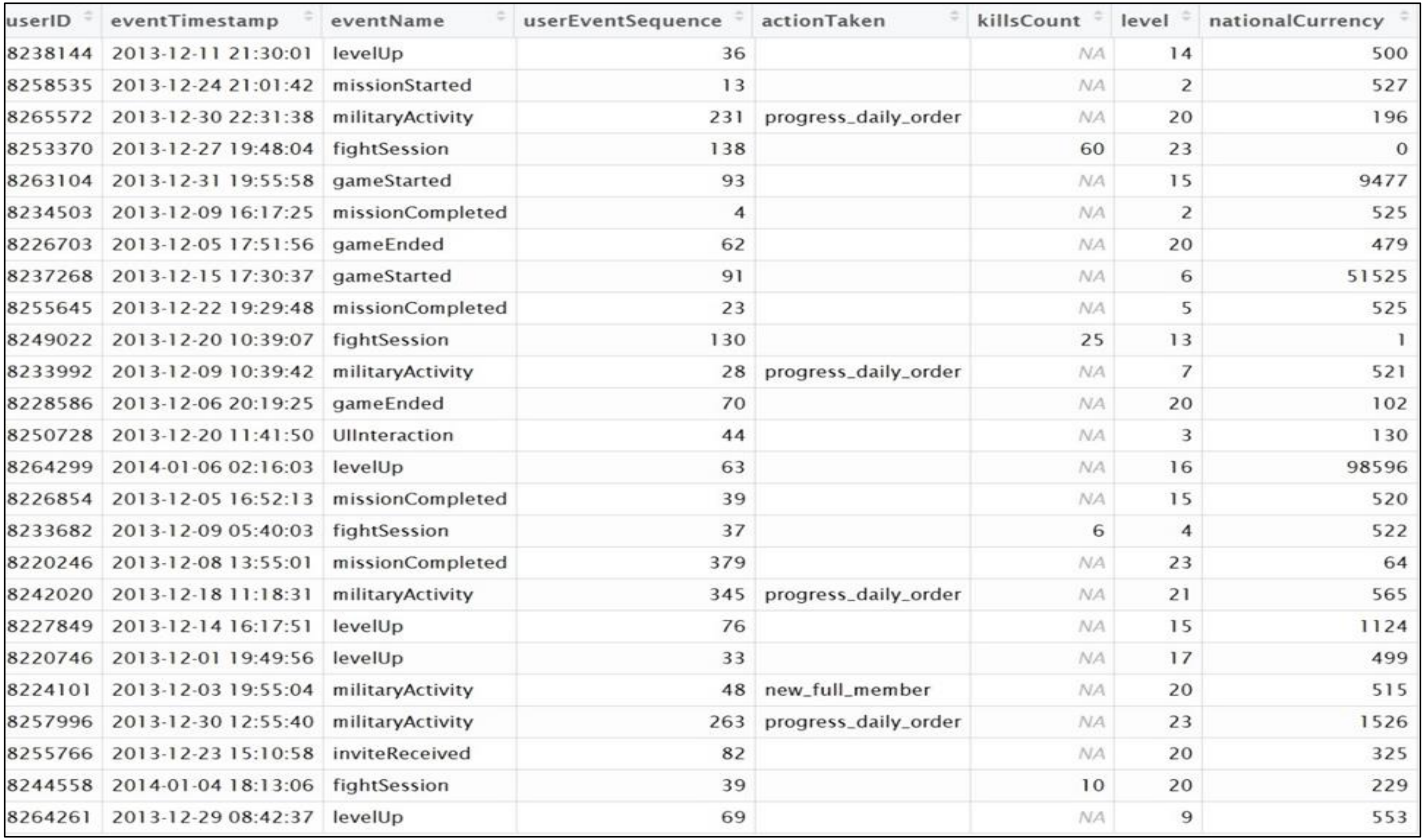
4.2 Reading and Nature

Based on the number of active new players within the game and subsequent events triggered by them, data corresponding to eight weeks of gameplay is gathered for robust analysis within this research study. This data is extracted from the cloud server of the games analytics company and exported as a CSV file, which is then imported into R 1.1.456 (R Core Team, 2017). Standard R programming functions and codes are used to then read the data and carry out preliminary investigations about the variables.

The first step is to understand the basic structure of the data, which amounts to more than 3 million rows of observations (3204027) and 120 variables. It is a large data set comprising of 40716 active new players between 11 November 2013 and 29 December 2013, triggering a total of 2375700 events in the period 11 November 2013 and 6 January 2014. The time periods for analysis are selected in a way that ensures all new players have at least 9 days of having their gameplay monitored. That is, even users that join the game on 29 December 2013 will be followed for 9 days (until 6 January 2014) to explore how they interact with the game.

A small fraction of the dataset that is read in to R is displayed in table 4.1. It is a data frame created by selecting a random sample of 25 rows of observations and 8 variables of interest from the complete dataset. Overall, it is a continuous and time-stamped data representing in-game events triggered by users in real time. Each row symbolises an event triggered and each column denotes a variable associated with that event. For example, the first row indicates an event generated on 11th December 2013 at 21:30:01 hours by a player with the ID 8238144. It is a level-up event and is the 36th event of the user, who at that point in the game was at level 14 with 500 units of national currency (grind currency) in possession.

Table 4.1: A subset of the eRepublik events data



userID	eventTimestamp	eventName	userEventSequence	actionTaken	killsCount	level	nationalCurrency
8238144	2013-12-11 21:30:01	levelUp	36		NA	14	500
8258535	2013-12-24 21:01:42	missionStarted	13		NA	2	527
8265572	2013-12-30 22:31:38	militaryActivity	231	progress_daily_order	NA	20	196
8253370	2013-12-27 19:48:04	fightSession	138		60	23	0
8263104	2013-12-31 19:55:58	gameStarted	93		NA	15	9477
8234503	2013-12-09 16:17:25	missionCompleted	4		NA	2	525
8226703	2013-12-05 17:51:56	gameEnded	62		NA	20	479
8237268	2013-12-15 17:30:37	gameStarted	91		NA	6	51525
8255645	2013-12-22 19:29:48	missionCompleted	23		NA	5	525
8249022	2013-12-20 10:39:07	fightSession	130		25	13	1
8233992	2013-12-09 10:39:42	militaryActivity	28	progress_daily_order	NA	7	521
8228586	2013-12-06 20:19:25	gameEnded	70		NA	20	102
8250728	2013-12-20 11:41:50	UIInteraction	44		NA	3	130
8264299	2014-01-06 02:16:03	levelUp	63		NA	16	98596
8226854	2013-12-05 16:52:13	missionCompleted	39		NA	15	520
8233682	2013-12-09 05:40:03	fightSession	37		6	4	522
8220246	2013-12-08 13:55:01	missionCompleted	379		NA	23	64
8242020	2013-12-18 11:18:31	militaryActivity	345	progress_daily_order	NA	21	565
8227849	2013-12-14 16:17:51	levelUp	76		NA	15	1124
8220746	2013-12-01 19:49:56	levelUp	33		NA	17	499
8224101	2013-12-03 19:55:04	militaryActivity	48	new_full_member	NA	20	515
8257996	2013-12-30 12:55:40	militaryActivity	263	progress_daily_order	NA	23	1526
8255766	2013-12-23 15:10:58	inviteReceived	82		NA	20	325
8244558	2014-01-04 18:13:06	fightSession	39		10	20	229
8264261	2013-12-29 08:42:37	levelUp	69		NA	9	553

The data is comprised of unique player identifiers preserving the anonymity of users, and a set of attributes associated with each event generated by these users. Similar to that explained by Medler et al. (2011), these attributes generally illustrate the name and type of event, timestamp of when it was triggered, and other distinct features of the event depicting player actions. These activities are a combination of diegetic actions relevant to "the game's total world of narrative action" (Galloway, 2006, p.7), and nondiegetic actions consistent with "gamic elements that are inside the total gamic apparatus yet outside the portion of the apparatus that constitutes a pretend world of character and story" (Galloway, 2006, pp.7-8). Some of the diegetic actions constituting this data are fight summary, military activity, political activity, upgrade, transaction etc., while the nondiegetic ones are game started, game ended, UI interaction and help. Event information is in the form of key-value pairs (Medler et al., 2011), where each pair constitutes the name of the variable associated with an event and its value. In this dataset, some of the relevant key-value pairs contain information regarding the type of event triggered, level and experience points attained, type of military action undertaken, amount of gold (premium currency) and national currency (grind currency) owned, number of kills per session of fight, amount of weapon damage inflicted on opponents, military rank achieved and so forth.

4.3 Cleaning

Data cleaning concerns the discovery and elimination of anomalies and fallacies in the data for the purpose of improving its quality (Rahm & Do, 2000). Hernández and Stolfo (1998) and Lee, Lu, Ling, and Ko (1999) discuss some prevalent complications with large archives of data as that containing duplicate records for the same entities, errors during the data entry stage leading to inaccurate or missing data, and other kinds of inconsistencies and inadequacies. Medler et al. (2011) reveals that "the creation of telemetric hooks to track player behaviour is a software process requiring the use of an API and as such is as prone to error as any other part of the software" (p.7). Therefore, it is imperative to keenly scrutinize the data so as to assure that event hooks in the game are meticulously put in place to exhibit player behaviour. Identification of missing values, multiple copies of the same event and other anomalies greatly reduce the risk of aggregated calculations appearing skewed, that would otherwise affect data interpretation.

Observations from the data cleaning procedure are elucidated below.

- The data typically contains some variables that are redundant to this particular research, as these do not contribute towards insights on customer behaviours. This makes the selection of suitable variables of interest a crucial first step. It is achieved through an extensive study of the variables in conjunction with an understanding of the game and its various elements. Additionally, certain variables containing only null or missing values are also eliminated from the analysis. This step results in 19 superfluous variables being discarded from analysis.
- Some of the candidate variables have an atypical formatting, such as the date-time variables 'eventTimestamp' and 'firstRegistered', which are recorded as character variables. These are converted to date-time classes representing calendar dates and times and are also used to derive new variables storing only the date part. The converted and new variables are more appropriate for performing complex operations and functions.
- There is prevalence of duplicate records in this data. These are identified based on observations having the same user ID along with the exact same time, name and ID of the event triggered. These are found to usually correspond to level up, mission complete or message sent/received events Duplicate events are not accounted for in the analysis.

4.4 Exploratory Analysis

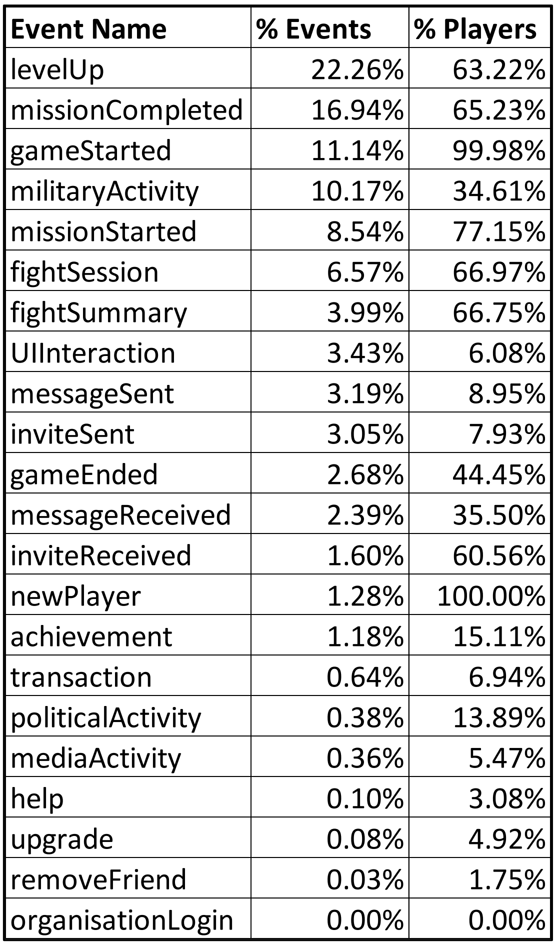
Following the data cleaning process is exploratory data analysis, a concept coined by Tukey (1977), which lets the data direct the choice of pertinent statistical models by lessening prior hypotheses or assumptions (Velleman & Hoaglin, 1981). Cox and Jones (1981) assert that exploratory data analysis aids in the study and recognition of the most predominant characteristics of a data set, which then promotes additional inspections of the data. It is a predecessor of confirmatory data analysis where "attention is focused on model specification, parameter estimation, hypothesis testing and firm decisions about data" – (Cox & Jones, 1981, p.135).

The exploratory analysis techniques used here may be useful in advising the construction of more advanced and complicated models. It may also assist in discarding or emphasizing likely hypotheses about player behaviour that can be tested using the data.

Initially, distributions of potential variables of interest are examined, which results in two broad classifications - generic ones that are likely to exist across most online freemium games, and distinct ones that are exclusive to eRepublik's gameplay. Examples of generic variables include user ID, event timestamp, event name, mission name, level, experience points, session ID and transaction ID. Examples of game-specific variables include action taken (denoting a particular activity performed by the player such as becoming a full member of their team or completing their daily order), kills count (recording the number of opponents killed in a fight), gold (amount of premium currency gold that the user has at a point in time) and military strength (a measure of their strength within the military, directly correlated with their military rank).

The foremost and elementary aspect studied is the distribution of the types of events comprising the data. Table 4.2 displays the types of events, arranged according to their frequency of occurrence, and the proportion of users triggering each event.

Table 4.2: Distribution of Event Types



Event Name	% Events	% Players
levelUp	22.26%	63.22%
missionCompleted	16.94%	65.23%
gameStarted	11.14%	99.98%
militaryActivity	10.17%	34.61%
missionStarted	8.54%	77.15%
fightSession	6.57%	66.97%
fightSummary	3.99%	66.75%
UIInteraction	3.43%	6.08%
messageSent	3.19%	8.95%
inviteSent	3.05%	7.93%
gameEnded	2.68%	44.45%
messageReceived	2.39%	35.50%
inviteReceived	1.60%	60.56%
newPlayer	1.28%	100.00%
achievement	1.18%	15.11%
transaction	0.64%	6.94%
politicalActivity	0.38%	13.89%
mediaActivity	0.36%	5.47%
help	0.10%	3.08%
upgrade	0.08%	4.92%
removeFriend	0.03%	1.75%
organisationLogin	0.00%	0.00%

Level up is the most frequently occurring event, followed by mission completion, game start and military activity events, constituting about 61% of the total events triggered. The table also reveals that the military and fight component of the gameplay dominates over the political and media aspects. Although 77% of players start a mission, 65% complete them, thereby indicating a gap of 12% who are unsuccessful. Only about 7% of users generate a transaction event, which make up even less than 1% of the total events triggered. The upgrade event, which can be deemed as an indicator of game progression is also one of the unpopular events, being generated by less than 5% of the player population. Although not particularly affecting behavioural study, an anomaly found in the data shows that not every game start event has an analogous game end event. Around 11% of events are game starts whereas only about 3% are game ends; and almost all players have a game start event, while only about 44% have a game end. This is reflective of the manner in which events are recorded by the game studio, in which, if the user closes a game session without a proper log out, a game end event is not generated. Thus, scrutiny of the events distribution provides a basic insight into how users are interacting with the game and its various components.

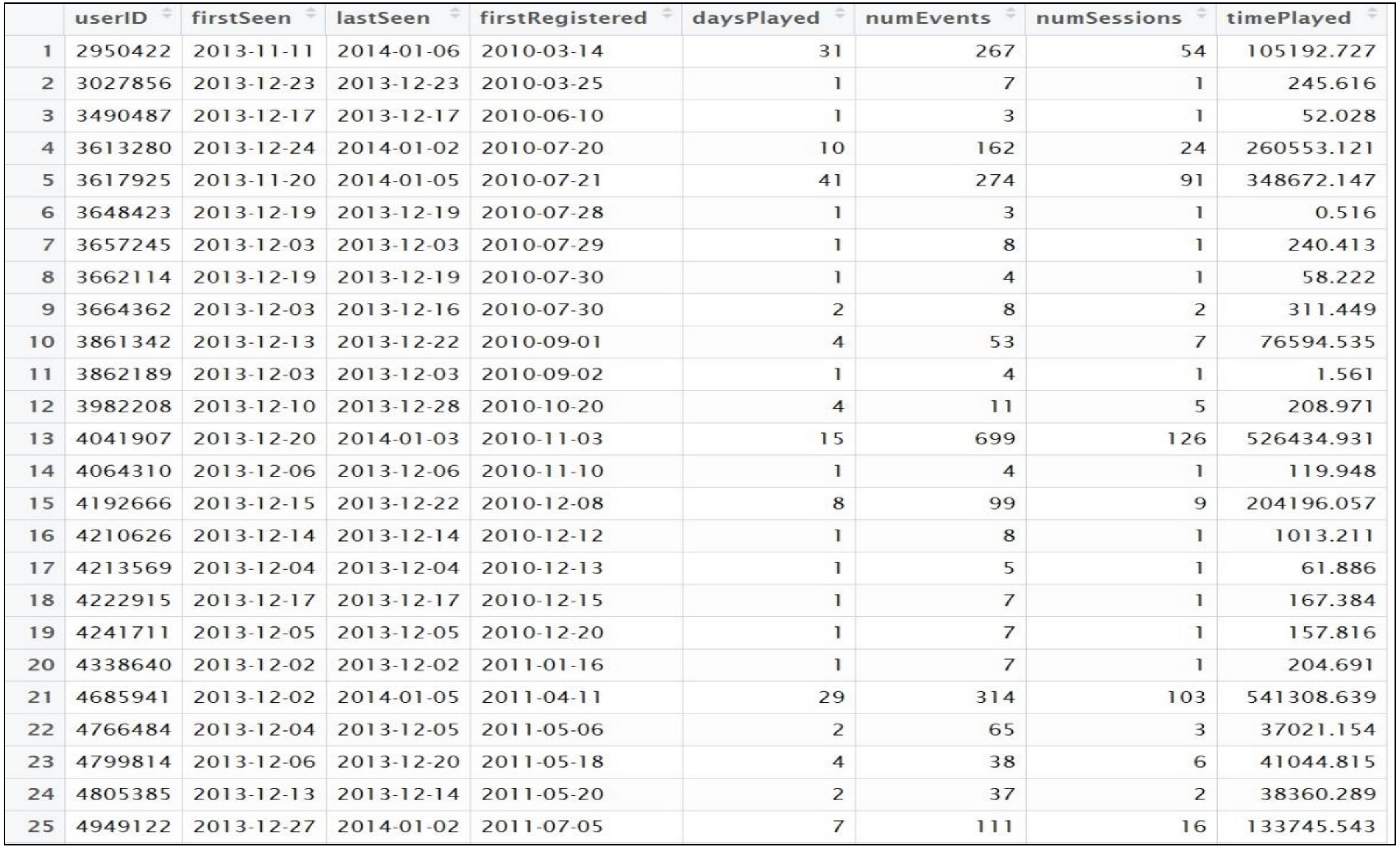
Medler et al. (2011) state that the time to query data is often a complication when handling large datasets where investigating each element can be tedious, and suggest "aggregate data for quick analysis" (p.7) as a solution to this problem. Given this data is considerably big, user-level data frames are constructed by aggregating relevant variables per user from the raw events history, thereby prompting efficient computation and analysis henceforth. Essentially, it produces two kinds of user-level data – general gameplay metrics based on generic variables and player behaviour metrics based on game-specific variables.

The data frame outlined in table 4.3 represents user-level metrics of general gameplay calculated from the raw events file. This is a summary data consisting of important statistics for each user such as the number of days they played the game for, total number of events triggered and total time played.

The data frame outlined in table 4.4 illustrates user-level metrics of player behaviour calculated from the raw events file. This data comprises of variables that highlight the behavioural aspects of gameplay such as competency in game missions (average kill to hit ratio), social interaction with other users (friends), premium and grind currencies owned (average gold and average national currency) within the game, and overall performance (level).

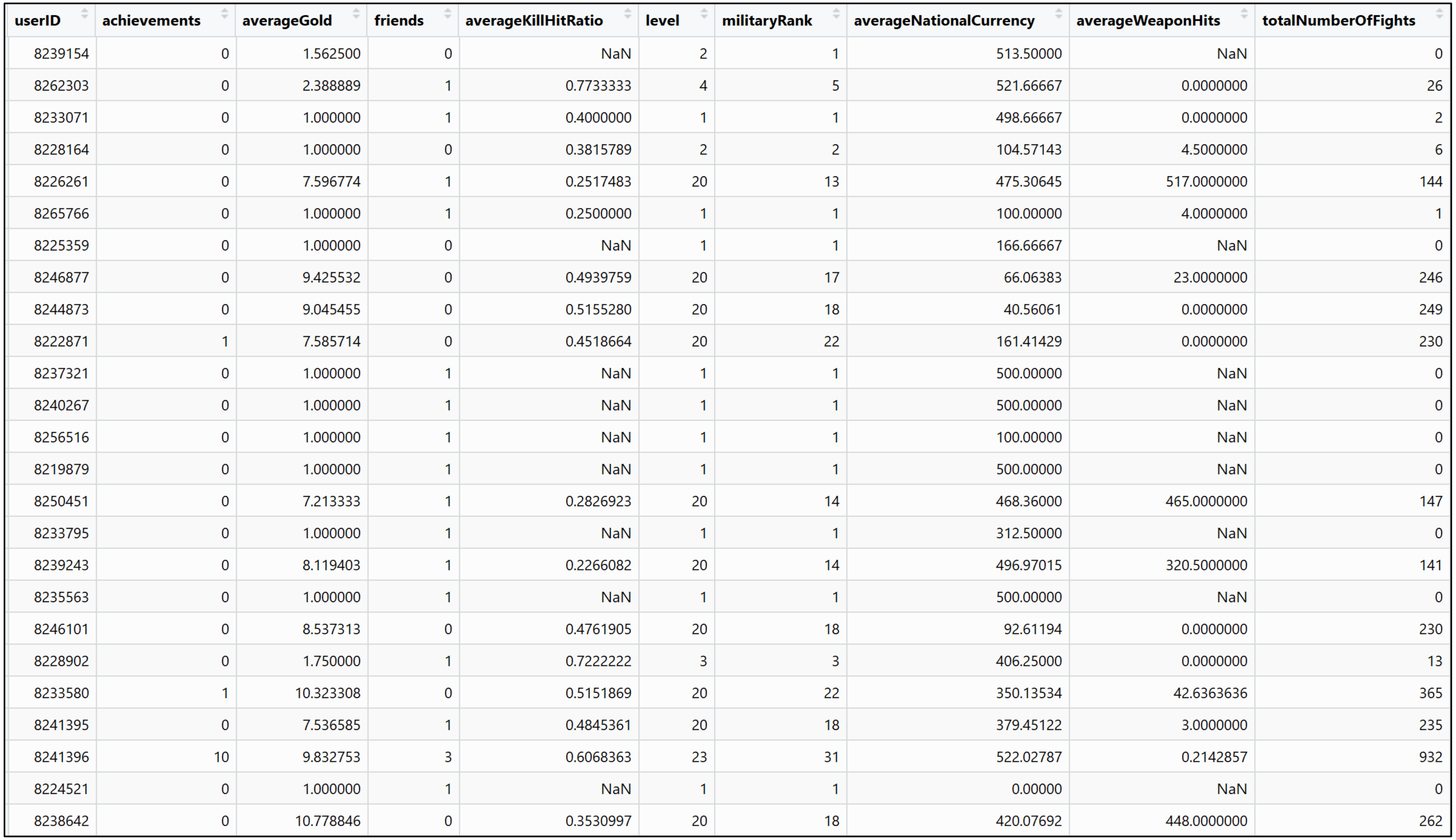
These player-level summary tables are more suitable for quick querying and analysis. Hence, it is more feasible to base modelling and other complex methods on these types of data sets, while still leaving the raw events data intact to access and use as and when needed.

Table 4.3: A subset of the general gameplay metrics



	userID	firstSeen	lastSeen	firstRegistered	daysPlayed	numEvents	numSessions	timePlayed
1	2950422	2013-11-11	2014-01-06	2010-03-14	31	267	54	105192.727
2	3027856	2013-12-23	2013-12-23	2010-03-25	1	7	1	245.616
3	3490487	2013-12-17	2013-12-17	2010-06-10	1	3	1	52.028
4	3613280	2013-12-24	2014-01-02	2010-07-20	10	162	24	260553.121
5	3617925	2013-11-20	2014-01-05	2010-07-21	41	274	91	348672.147
6	3648423	2013-12-19	2013-12-19	2010-07-28	1	3	1	0.516
7	3657245	2013-12-03	2013-12-03	2010-07-29	1	8	1	240.413
8	3662114	2013-12-19	2013-12-19	2010-07-30	1	4	1	58.222
9	3664362	2013-12-03	2013-12-16	2010-07-30	2	8	2	311.449
10	3861342	2013-12-13	2013-12-22	2010-09-01	4	53	7	76594.535
11	3862189	2013-12-03	2013-12-03	2010-09-02	1	4	1	1.561
12	3982208	2013-12-10	2013-12-28	2010-10-20	4	11	5	208.971
13	4041907	2013-12-20	2014-01-03	2010-11-03	15	699	126	526434.931
14	4064310	2013-12-06	2013-12-06	2010-11-10	1	4	1	119.948
15	4192666	2013-12-15	2013-12-22	2010-12-08	8	99	9	204196.057
16	4210626	2013-12-14	2013-12-14	2010-12-12	1	8	1	1013.211
17	4213569	2013-12-04	2013-12-04	2010-12-13	1	5	1	61.886
18	4222915	2013-12-17	2013-12-17	2010-12-15	1	7	1	167.384
19	4241711	2013-12-05	2013-12-05	2010-12-20	1	7	1	157.816
20	4338640	2013-12-02	2013-12-02	2011-01-16	1	7	1	204.691
21	4685941	2013-12-02	2014-01-05	2011-04-11	29	314	103	541308.639
22	4766484	2013-12-04	2013-12-05	2011-05-06	2	65	3	37021.154
23	4799814	2013-12-06	2013-12-20	2011-05-18	4	38	6	41044.815
24	4805385	2013-12-13	2013-12-14	2011-05-20	2	37	2	38360.289
25	4949122	2013-12-27	2014-01-02	2011-07-05	7	111	16	133745.543

Table 4.4: A subset of the player behaviour metrics



userID	achievements	averageGold	friends	averageKillHitRatio	level	militaryRank	averageNationalCurrency	averageWeaponHits	totalNumberOfFights
8239154	0	1.562500	0	NaN	2	1	513.50000	NaN	0
8262303	0	2.388889	1	0.7733333	4	5	521.66667	0.0000000	26
8233071	0	1.000000	1	0.4000000	1	1	498.66667	0.0000000	2
8228164	0	1.000000	0	0.3815789	2	2	104.57143	4.5000000	6
8226261	0	7.596774	1	0.2517483	20	13	475.30645	517.0000000	144
8265766	0	1.000000	1	0.2500000	1	1	100.00000	4.0000000	1
8225359	0	1.000000	0	NaN	1	1	166.66667	NaN	0
8246877	0	9.425532	0	0.4939759	20	17	66.06383	23.0000000	246
8244873	0	9.045455	0	0.5155280	20	18	40.56061	0.0000000	249
8222871	1	7.585714	0	0.4518664	20	22	161.41429	0.0000000	230
8237321	0	1.000000	1	NaN	1	1	500.00000	NaN	0
8240267	0	1.000000	1	NaN	1	1	500.00000	NaN	0
8256516	0	1.000000	1	NaN	1	1	100.00000	NaN	0
8219879	0	1.000000	1	NaN	1	1	500.00000	NaN	0
8250451	0	7.213333	1	0.2826923	20	14	468.36000	465.0000000	147
8233795	0	1.000000	1	NaN	1	1	312.50000	NaN	0
8239243	0	8.119403	1	0.2266082	20	14	496.97015	320.5000000	141
8235563	0	1.000000	1	NaN	1	1	500.00000	NaN	0
8246101	0	8.537313	0	0.4761905	20	18	92.61194	0.0000000	230
8228902	0	1.750000	1	0.7222222	3	3	406.25000	0.0000000	13
8233580	1	10.323308	0	0.5151869	20	22	350.13534	42.6363636	365
8241395	0	7.536585	1	0.4845361	20	18	379.45122	3.0000000	235
8241396	10	9.832753	3	0.6068363	23	31	522.02787	0.2142857	932
8224521	0	1.000000	1	NaN	1	1	0.00000	NaN	0
8238642	0	10.778846	0	0.3530997	20	18	420.07692	448.0000000	262

After having developed user-level data tables, probability distributions of some of the key variables are investigated. These include total time played representing user engagement, total no. of transactions representing monetisation of users, amount of gold (premium currency) possessed which might be an influencing factor for micro-purchases and mission completion rate denoting customer performance and hence immersion with the game. Kernel density plots depicting probability distributions are presented below.

The distribution of total gameplay time (in minutes) of users is illustrated in figure 4.1. It is seen to be highly skewed with a very long right tail that starts tapering off to almost zero from 3000 minutes (50 hours) worth of gameplay onwards. Summary statistics reveal about 15% of users having a total gameplay time of less than a minute, with a median of 99 minutes (less than 2 hours) and first and third quartiles 3 minutes and 682 minutes (~11 hours) respectively. The presence of outliers in the distribution results in a mean of 952 minutes (~16 hours). Examination of the distribution of total gameplay time of users makes it evident that a considerable proportion of the population dropout very early from the game with no substantial gameplay to study. These players are excluded from the analyses henceforth. Moreover, with the mean gameplay time being remarkably higher than its median, the extreme values are indicative of a group of users that are actively engaged with the game, the study of which will aid in understanding the catalysts of monetisation and survival within the game.

The distribution of number of real currency purchases by customers is depicted in figure 4.2. It is a discrete distribution, observed to be heavily inflated with zeroes. Summary statistics indicate that approximately 99% of the sample under study have zero purchases in the time period being analysed. Although the maximum no. of transactions made is 20, the median, mean, first and third quartiles are all approximately zero. This suggests that micro-transactions are extremely rare events, thereby making user monetisation one of the biggest challenges facing game studios. Moreover, of the 1% that make payments within the game, more than half (~54%) are one-off payers. Monetisation analysis will involve study of the behaviours of this 1%, with further investigations distinguishing the 54% one-time payers to the remaining 46% repeat payers.

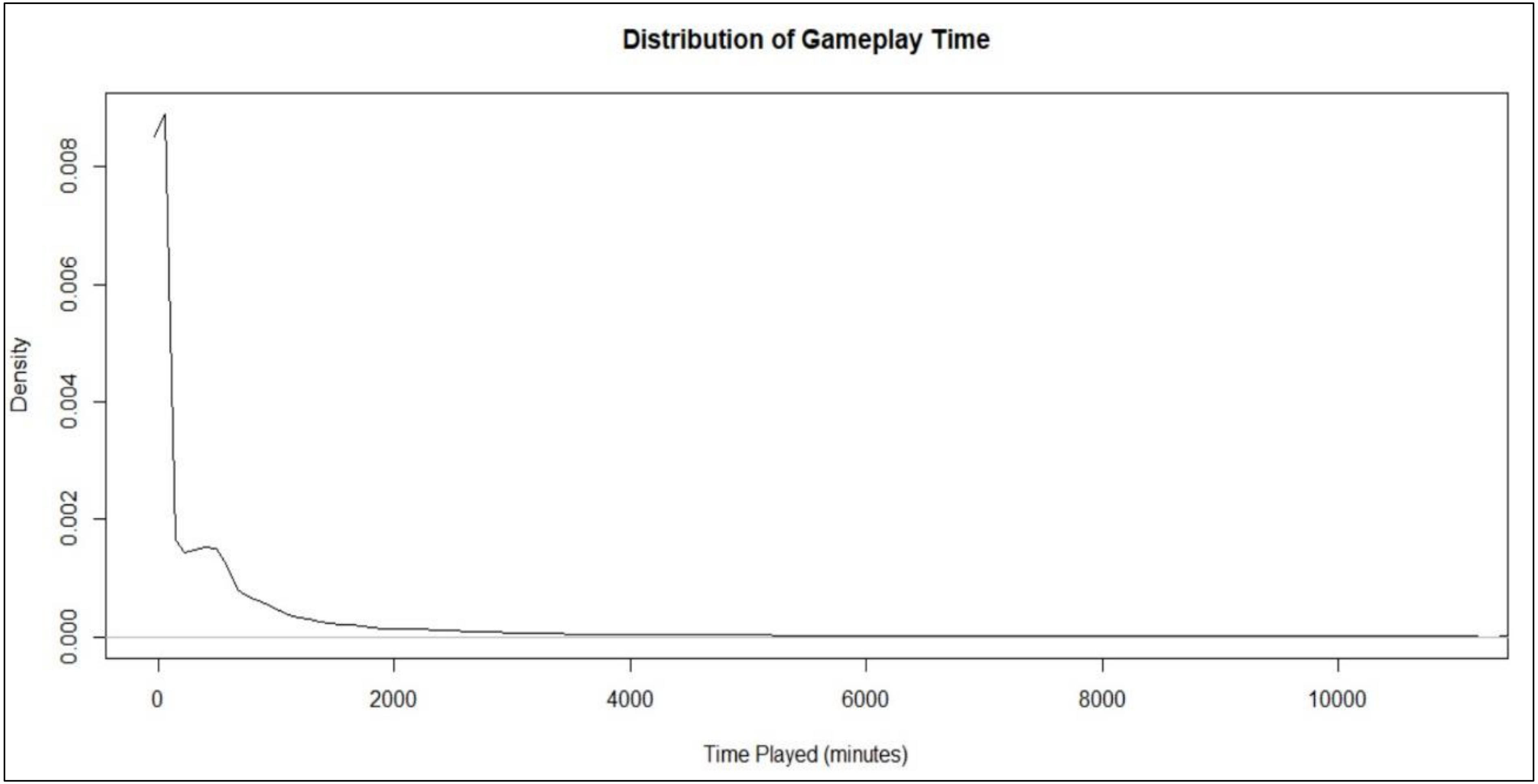


Figure 4.1: Distribution of users' gameplay times

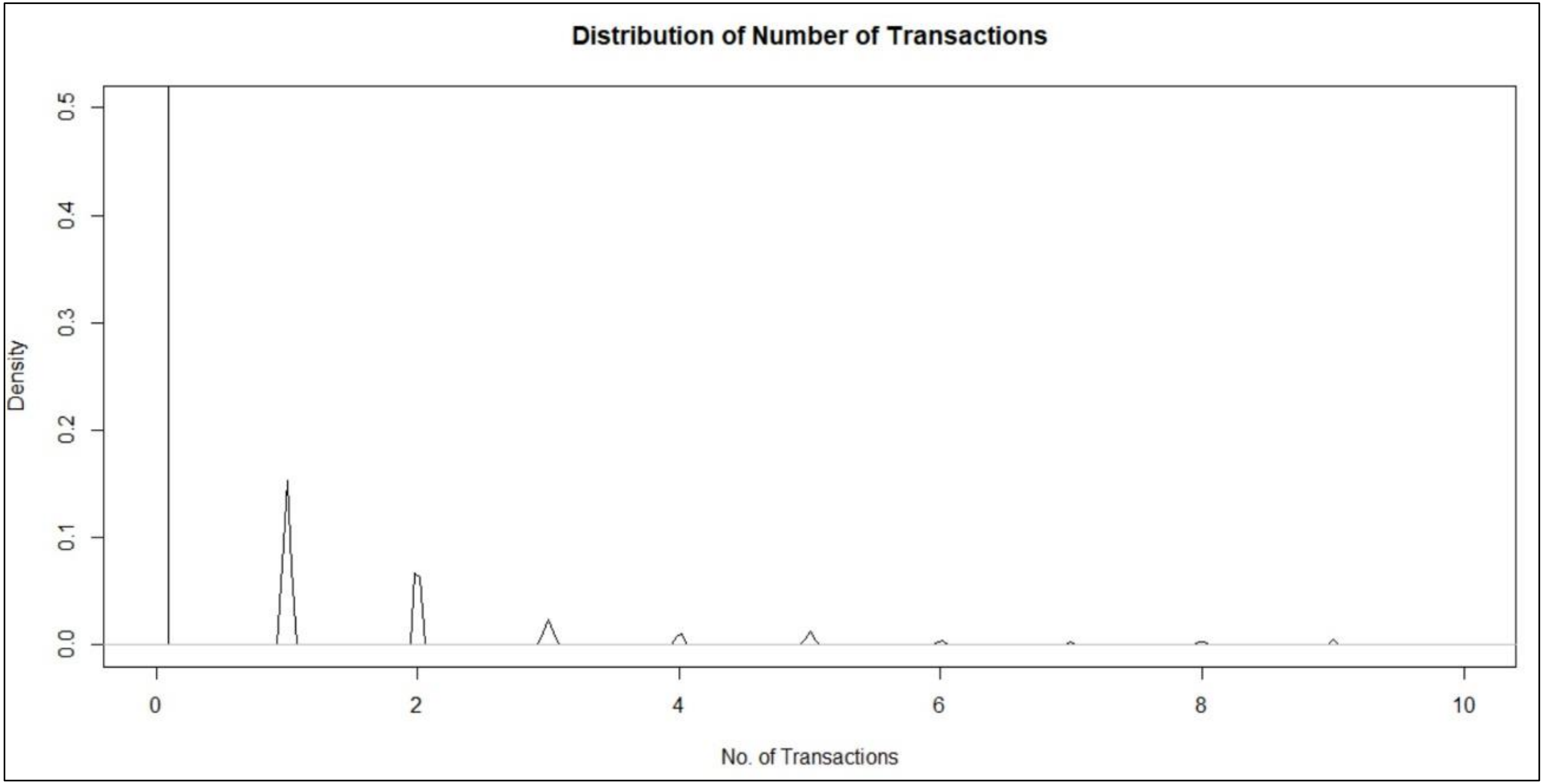


Figure 4.2: Distribution of number of player transactions

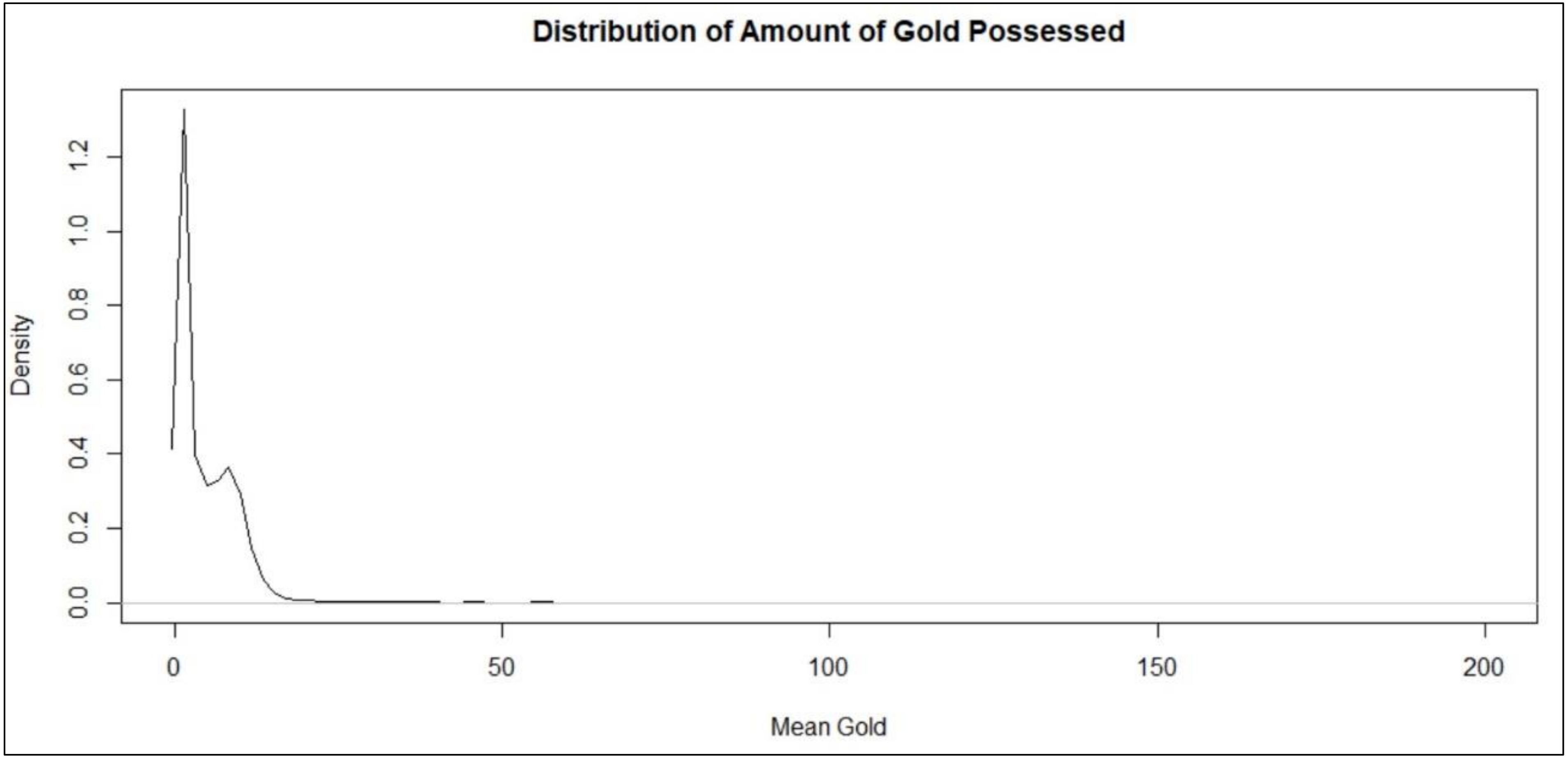


Figure 4.3: Distribution of amount of gold (premium currency) possessed by players

The average amount of premium currency, called gold in eRepublik, owned by players is distributed as shown in figure 4.3. Similar to the previous variables, this also represents a skewed right tailed distribution. The central tendency of the distribution represented by median and mean is approximately 3 and 5 respectively, the first and third quartiles being 1 and 8 respectively. Although only about 2% of users do not own any premium currency, majority of them (~60%) have only 1-5 units of gold on an average. The attainment of premium currency may indicate how well players are performing in tasks and missions since gold is usually awarded to them on completion of certain challenges. It might also drive micro-purchases, as gold being a premium currency can be used to buy premium virtual goods that would enhance the user's gaming experience. Since there are a considerable number of early churners in this game, the mean gold count being low is not entirely unexpected. However, a bulk of the users possessing 1-5 units of gold might signify poor proficiency for engaged players as well. This can be further verified by inspecting the completion rates of users in missions.

The distribution of mission completion rates of users shown in figure 4.4.is bimodal in nature, with distinct peaks at the 0 and 1 marks. Around 23% of the player sample in this research do not start a mission at all and therefore have no valid completion rates. This can be attributed to those that drop out within a short span of time. Of the players that attempt missions, 12% have a completion rate of 0%, while 10% have a perfect completion rate of 100%. The average completion rate is about 63% and the distribution demonstrates that usually about half the missions started are also completed. This justifies the earlier observation that eRepublik users are generally not highly competent in their gameplay.

Insights gained from preliminary exploratory analyses are utilised during statistical modelling of relevant outcomes in the upcoming chapters.

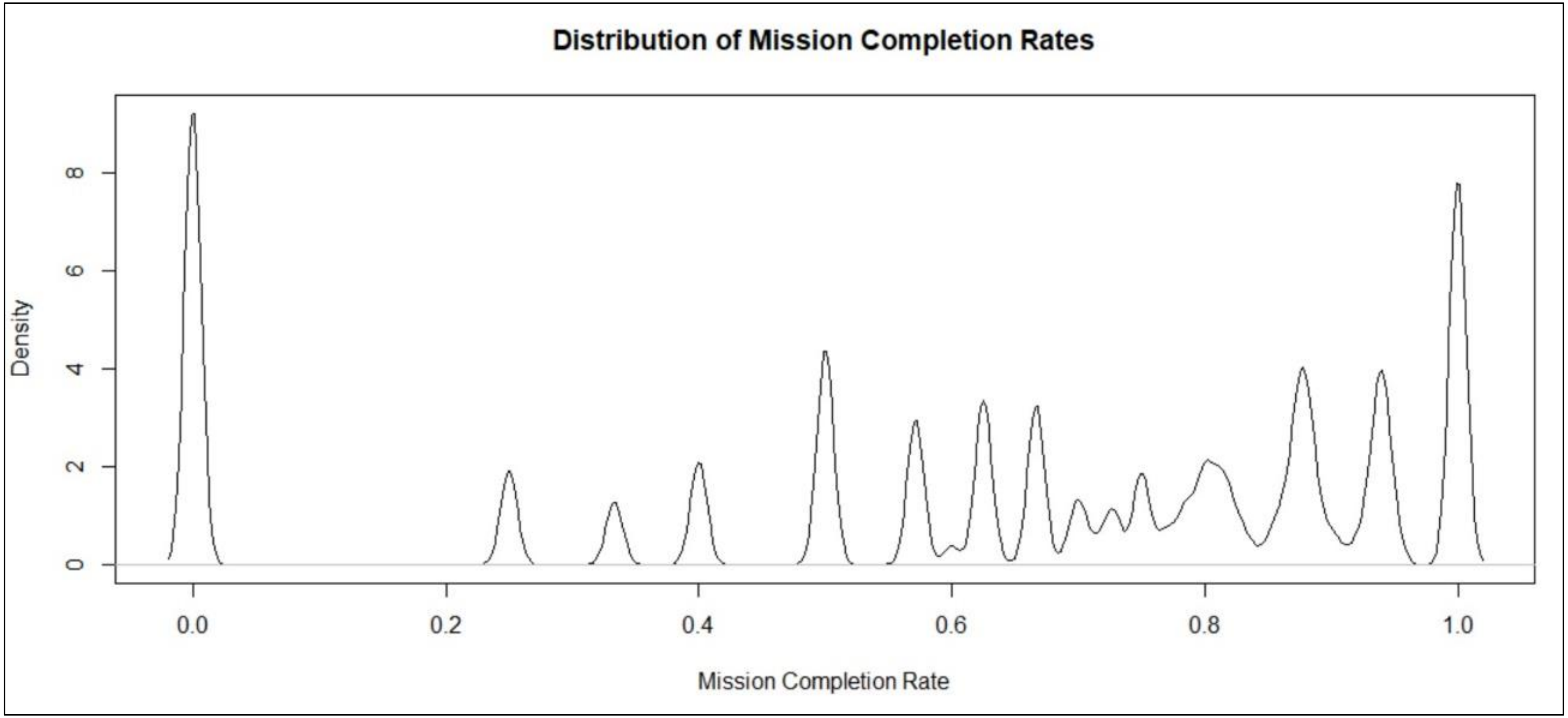


Figure 4.4: Distribution of completion rates of players in missions

5. Analysis of Customer Engagement

It is not imperative to engage customers of premium games since the studio has already managed to sell these to the individuals, however, user immersion or engagement is undoubtedly an integral facet of freemium games for initiating virality and monetisation (Iterable, 2017). It is one of the most important criteria for successful games (Jennett et al., 2008) and crucial in adding value to player experience (Schoenau-Fog, 2011). Schoenau-Fog (2011) demarcates player engagement from the motivation to start playing a game or being drawn towards the game, and assumes it to be "the level of continuation desire experienced in-game, during play or over a longer period of time, when players dedicate themselves to coming back and playing a game again and again" (p.4).

In this chapter the various components of user engagement learnt during the literature review, especially the above-mentioned repeated fascination of returning to the game over time are examined and the aim is to build a predictive model of customer engagement in online freemium games.

5.1 Indicators of Engagement

Preliminary investigation of the data specifies some generic variables that may be considered as barometers of user engagement. These are number of days played, number of events triggered, number of sessions played and total time played, which in turn are likely to impact game specific variables such as number of missions started, highest level achieved and highest military rank achieved. Since this research aims to provide a general framework of statistical analysis, engagement is primarily represented by the generic variables, such as total gameplay time (in minutes) of users, the distribution of which is displayed in figure 4.1. The measures of engagement are potentially heavily skewed with long right-tailed distributions.

Although each of the 40716 active new players have a period of at least 8 days to engage with the game based on the time period (i.e. 11 November 2013 – 06 January 2014) analysed, a vast majority (~80%) play the game for 1 – 3 days, with almost half the sample (~48%) playing for only one day. Negligible number of players (~1%) interact with the game for more than 30 days, thereby demonstrating that it is extremely rare for users to indulge in as much a month's worth of gameplay.

Furthermore, the average number of events triggered were 58, but the median only 22, with about 35% users generating less than 10 events. Based on the exploratory inspections, it is known that some events produced are compulsory without the player actively choosing to perform actions within the game environment. For instance, game start and end events, new player event, organisation login, fight summary, message received and invite received are events that are inevitably generated and do not imply a significant in-game activity. Therefore, number of events triggered may not be the best metric to represent user engagement with the game.

The number of game sessions played is another measure of user engagement, wherein a game session is denoted by the period between a game login and logout. Players can have multiple sessions in a day. Although 48% of users play a single day, 30% play a single session, implying that there are about 18% that play more than one session in a day but do not return for a second day of gameplay. Additionally, a bulk of the users (around 77%) play less than 5 sessions. The 30% that do not use the product for more than one session have an average playtime of about 3 minutes, and are not likely to contribute any knowledge regarding engagement or immersion. These players presumably are not attracted to the game, and are best excluded from the analysis.

The total time invested in playing the game is inherently the most fundamental measure of engagement. As examined in the exploratory analysis, the distribution of this variable is highly skewed with a very long right tail that starts tapering off to almost zero from 3000 minutes (50 hours) worth of gameplay onwards. About 15% of users have a total gameplay time of less than a minute. Depicted in figure 5.1 is the relationship between total gameplay time (in minutes) and total number of days played. It demonstrates how the total time played changes with the change in number of days played.

The plot shows an approximate positive linear relationship between the two variables. Correlation coefficient tests are also statistically significant, exhibiting strong positive correlations ($r=0.74$, $\rho=0.74$, $\tau=0.61$; $p<0.001$).

Study of the various metrics of game immersion leads to the judgement that total number of days played and total time played are expectedly the best indicators of player engagement with the game.

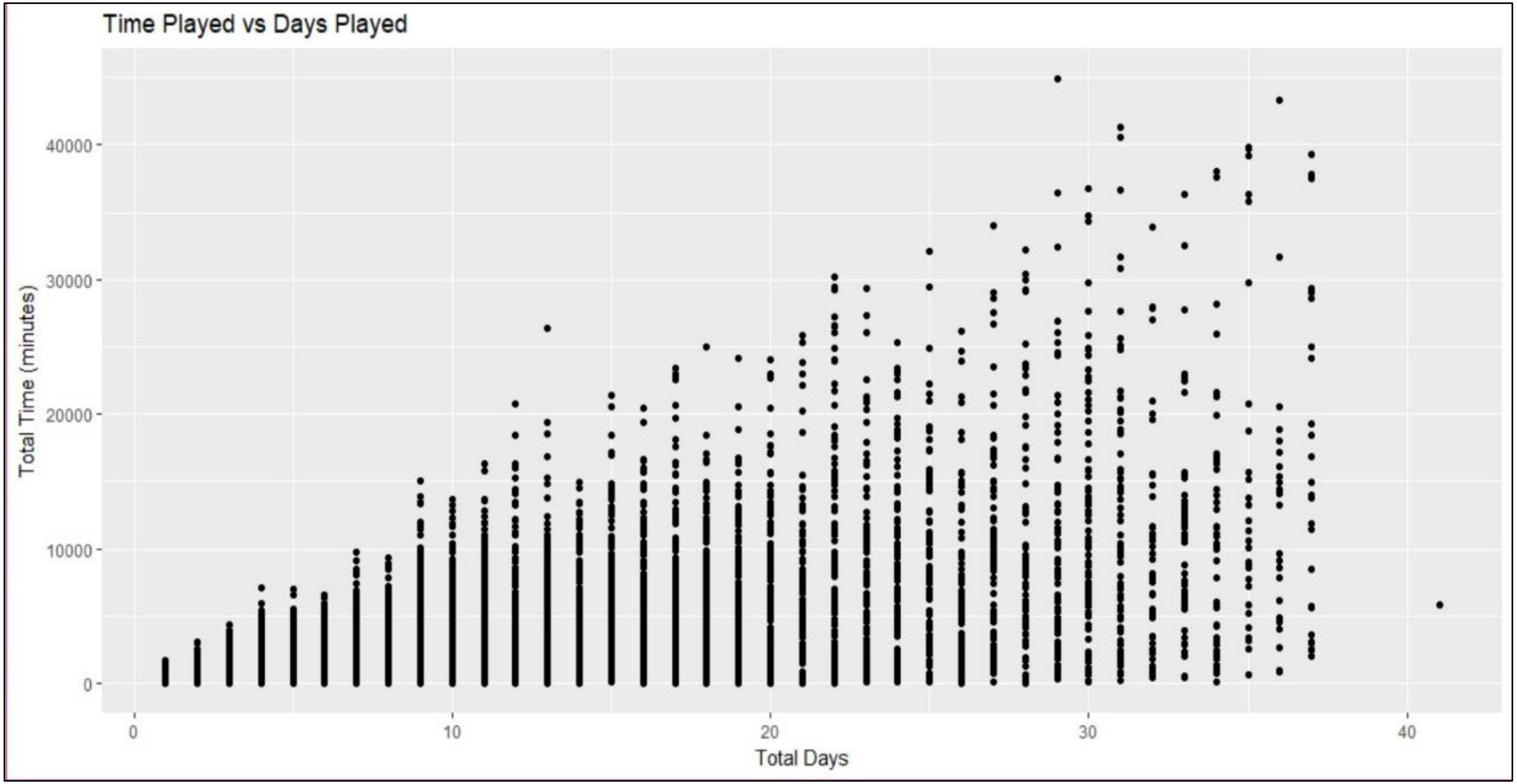


Figure 5.1: Distribution of total gameplay time against total number of days played

5.2 The Non-Engaged

As stated before, around 30% of the sample interact with the game for a single session only, i.e. they do not return after their very first logout. They play an average of approximately 3 minutes, with 75% of these users playing less than or equal to this. The mean number of events triggered by them is 5, with 75% triggering not more than 6 events. A large majority (~93%) of these events are the mandatory ones such as game start and end, new player and invite received events, albeit mission start and end events also constitute a part of this. Less than half start a mission (~42%), which are usually part of the tutorial, but the majority of them (~62%) do not complete any mission. Thus, it can be concluded that these players try out the game by logging in and starting a mission, however it is not something that suit their preference as a result of which they quit immediately thereafter never to return. This group of users are not examined, as they do not really connect with the gaming platform and hence not exhibit adequate consumer behaviours to study.

There is another category of customers who use the platform for a substantial amount of time, but is not fully immersed into it. They display behaviours that may indicate how they are communicating with the product and possible reasons for their non-immersion into it. Moreover, contrasting this group of players with the engaged ones is vital in gaining insight about what motivates user engagement and what actions discourage it.

Conceptualised in this research is that users who have failed to immerse themselves in the gaming experience, although they have had enough exposure to be absorbed into it, can be defined as non-engaged. In accordance with this and the distributions of engagement indicators - total time played and number of days played, non-engaged players are defined in the scope of this study as those that have played the game for 2 - 6 days with a total playtime of at least 5 hours, and do not return to it within at least a week. This will ensure that the users identified as non-engaged have initially connected with the game by returning to play for more than one day and also had sufficient experience of its features having played for at least 5 hours in total. However, they have not been completely engrossed in it, having less than a week's worth of gameplay and not using the platform consistently i.e. within at least a week.

The non-engaged users defined in this study constitute about 19.6% of the total user sample. Illustrated in figure 5.2 are the distributions of their time spent playing the game and performance in missions, the dashed lines represent the means of the distributions.

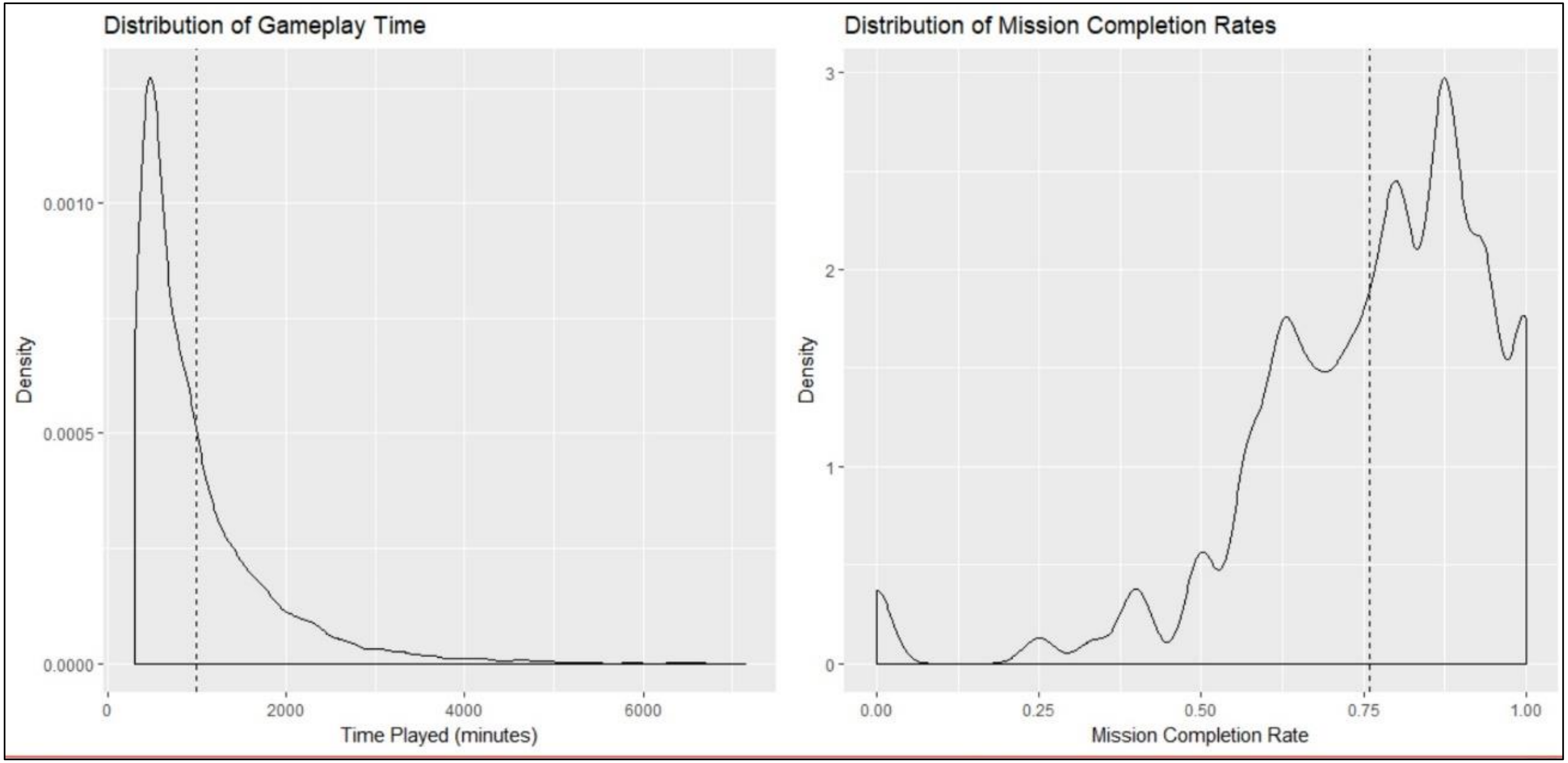


Figure 5.2: Distributions of total gameplay time and mission completion rates for non-engaged users

They invest 3 days on an average playing the game with a median gameplay time of 12 hours. They attempt 11 missions on an average, which is about 22% of the total number of missions in the analysis period. Their average completion rate in these missions is approximately 76%, albeit a tiny peak is observed at the 0% completion rate. This initial examination shows that the non-engaged are fairly involved and have a moderately high success rate, but with a few absolute failures in missions, although given the amount of time spent on the platform, they do not seem to participate in many missions. This leads to some further scrutiny of the type of actions these users tend to perform within the game environment. A big majority of their actions (~71%) include pursuing the daily order of the game, which is essentially a fighting order issued by the military unit to which the user belongs. An implication of this could be that non-engaged users are more inclined towards the fighting and action aspect of the gameplay than being motivated to completing missions.

5.3 The Engaged

The literature review revealed several nuances of customer engagement in online freemium games, of which the most crucial notion reiterated was the repeated return of users back to the platform over a long period of time. This was indicative of their absolute immersion into the game and aspiration to continue playing over and beyond the stages of initial attraction. Taking this into account, engaged customers of the game are defined in this research as those having played for a week or more and returned to it at least once in less than 4 days. This establishes players' continued use of the platform for long durations of time.

The engaged users defined in this study constitute about 8% of the total user sample. The distributions of their time spent playing the game and performance in missions are shown in figure 5.3, where the dashed lines represent the means of the distributions.

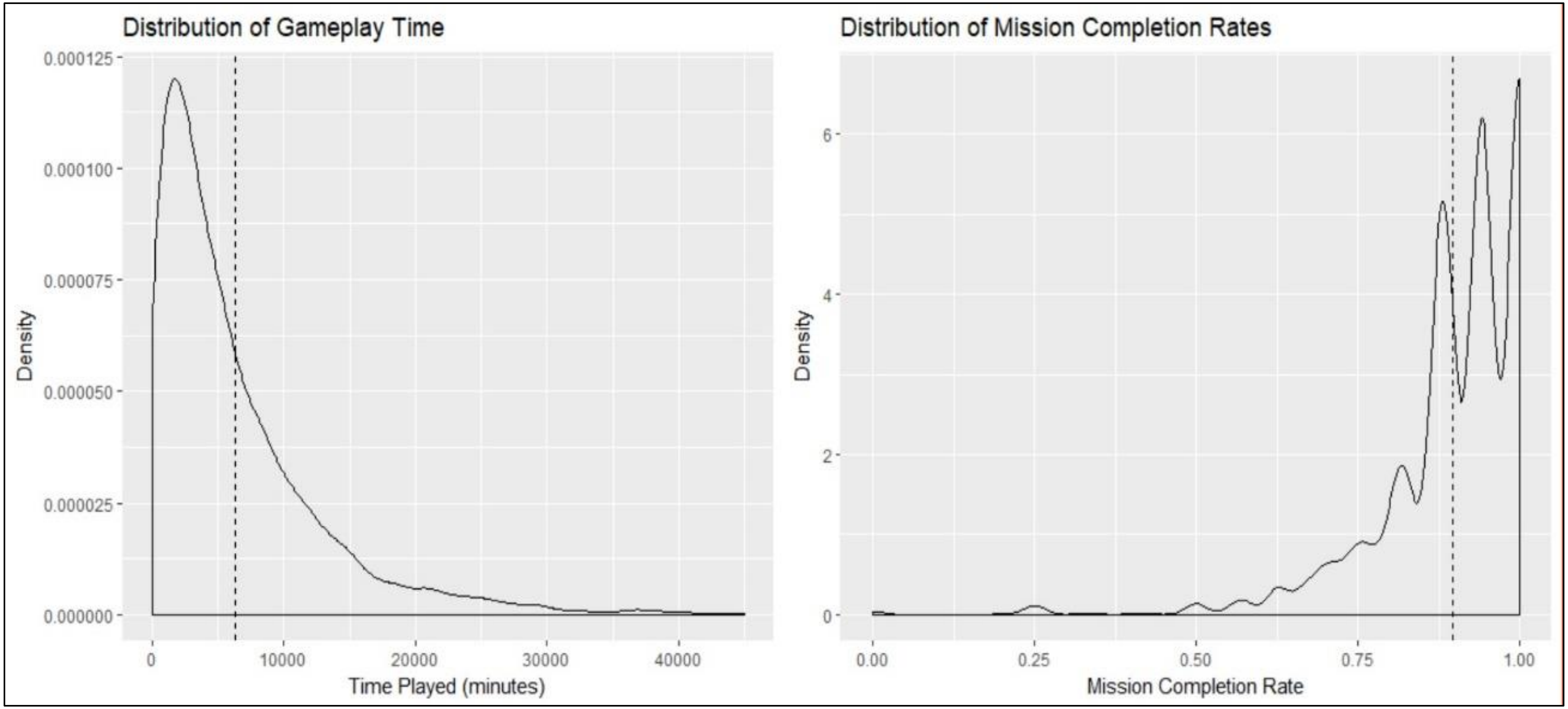


Figure 5.3: Distributions of total gameplay time and mission completion rates for engaged users

The engaged invest 18 days on an average playing the game with a median gameplay time of about 72 hours. They attempt 16 missions on an average, which is about 33% of the total number of missions in the analysis period. Their average completion rate in these missions is approximately 90%. In comparison with the non-engaged, engaged players are expectedly much more involved and have a very high success rate with almost negligible completion rates of 0%, although just like the non-engaged, given the amount of time spent on the platform, they do not seem to participate in many missions either. Similar to the non-engaged group, a large majority of their actions (~76%) include pursuing the daily order of the game. Hence it is evident that overall, players generally prefer to fight the virtual wars in the game rather than undertake set tasks provided to them by the platform.

5.4 Modelling User Engagement

The statistical model to be developed in this chapter is intended to identify factors evoking engagement and immersion in users of online freemium games, with a view to contrasting engaged players from the non-engaged ones. The outcome of interest here is player engagement, a binary variable assuming values 1(engaged) and 0 (non-engaged). The model is intended to estimate the significance and amount of association between the outcome variable and a set of predictors. Significant predictors will be inferred as the influencing determinants of player engagement. The model is also able to control for confounding variables and predict odds of engagement for different scenarios.

Logistic regression (or multiple logistic regression) model, introduced by Berkson (1944), is a class of generalized linear model involving a binary response variable and one or more independent explanatory variables or predictors (Hilbe, 2009). These models use the iterative re-weighted least squares estimating algorithm, which is simplistic and encounter only minor convergence complexities, thereby being a robust method of estimation (Hilbe, 2009). As demonstrated by Cabrera (1994), these models are optimal for dealing with dichotomous outcomes since they do not function under the rigid assumptions of normality, linearity and continuity as ordinary least square regression models do. Therefore, a logistic regression modelling approach is considered in the context of player engagement as one that may be most suitable within the scope of this research.

5.4.1 The Response and Predictor Variables

The dependent variable to be modelled is the outcome denoting engagement of users. It is dichotomous and assumes the value 1 for engaged users and 0 for non-engaged users.

The total number of engaged and non-engaged customers in the analysis period (11 November 2013 - 6 January 2014) are 3216 and 7974 respectively.

Primary analyses suggested that players are more involved in military related activities and prone to partake in the virtual fights rather than the rigid set of tasks already laid out by the game. However, it was also learned that engaged players achieve marginally better completion rates in missions than the non-engaged. Considering all this as well as the knowledge gained from actually playing the game, it is hypothesized that the variables that could potentially explain and influence player engagement are as follows –

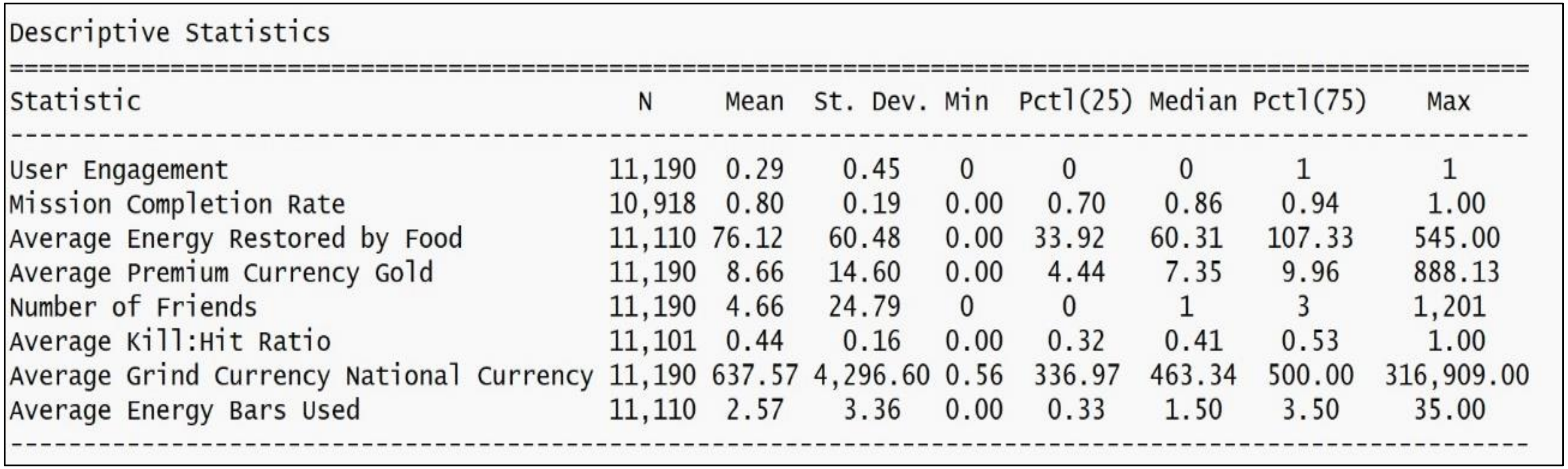
- Characterising player competency such as mission completion rate and average kill-hit ratio in virtual wars
- Denoting possession of virtual currencies such as average gold (premium currency) and average national currency (grind currency)
- Representing the use of available in-game resources to improve gameplay such as average energy bars used and average energy restored by food
- Social variables like number of friends

Prior to proceeding with statistical modelling, an insight on the underlying structure of the data used is acquired through basic descriptive statistics.

Displayed in table 5.1 are summary statistics of the dependent variable (user engagement) and the set of independent variables described before. The measures reported are number of observations, mean, standard deviation, minimum, first quartile (25th percentile), median, third quartile (75th percentile) and maximum of the distributions of each variable. It is evident from the table that some variables contain missing values, since the total number of cases should be 11190 (3216 engaged and 7974 non-engaged) for all.

Further scrutiny of the events triggered by users that have variables with missing values is conducted in order to estimate these. For variable mission completion rate, missing values were found to occur for players that did not start a mission at all, which made the calculation of completion rates invalid. In this case, first, the average number of days played by engaged and non-engaged players who have missing completion rates are calculated separately. Then, missing completion rates are approximated by the median completion rates of engaged and non-engaged users (separately) that play the same number of days as the average calculated before and have valid completion rates.

Table 5.1: Summary statistics of the data to be modelled



```
Descriptive Statistics
=================================================================================================================
Statistic                                   N      Mean   St. Dev.  Min   Pctl(25) Median Pctl(75)     Max
-----------------------------------------------------------------------------------------------------------------
User Engagement                          11,190  0.29     0.45      0       0        0       1          1
Mission Completion Rate                  10,918  0.80     0.19     0.00    0.70     0.86    0.94       1.00
Average Energy Restored by Food          11,110 76.12    60.48     0.00   33.92    60.31  107.33     545.00
Average Premium Currency Gold            11,190  8.66    14.60     0.00    4.44     7.35    9.96     888.13
Number of Friends                        11,190  4.66    24.79      0       0        1       3       1,201
Average Kill:Hit Ratio                   11,101  0.44     0.16     0.00    0.32     0.41    0.53       1.00
Average Grind Currency National Currency 11,190 637.57 4,296.60 0.56  336.97   463.34  500.00  316,909.00
Average Energy Bars Used                 11,110  2.57     3.36     0.00    0.33     1.50    3.50      35.00
-----------------------------------------------------------------------------------------------------------------
```

The variables average energy restored by food and average energy bars used contain missing values in instances where the player does not participate in any of the virtual fights and thus is never required to consume food or energy bars replenish energy, thereby making the averages invalid. In such cases, the missing values are replaced by zeroes indicating that these users do not avail of virtual resources in the game at all. The missing values for average kill:hit ratio arises when players do not make any hits against their opponents in the virtual wars, leading to no kills and therefore an average kill:hit ratio being null. Estimation of these nulls are handled in the same way as the approximation of missing mission completion rates.

5.4.2 Building the Model

The response variable engagement is denoted by Y and the set of independent variables are (X_1, X_2, …, X_p). The multiple logistic regression model to be fitted, as described in Hosmer Jr, Lemeshow and Sturdivant (2013), is as follows –

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (1)$$

Where, $\pi(x)$ =P(Y=1|X=x) is the conditional probability that the user is engaged given particular values $x_1, x_2, \ldots, x_\mathrm{p}$ for the predictors, and $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters to be estimated by the model.

At the outset, the data is partitioned into training and test samples based on a 75:25 ratio. The training set is used to build the multiple logistic regression model and learn about the relationship between the dependent and explanatory variables. The final model developed is validated and its performance estimated using the test set.

A multiple logistic regression model is fitted to the training data with user engagement as the response and the previously discussed variables hypothesized to affect the response as potential predictors. Parameter estimation by the model is via maximum likelihood estimation using an iterative reweighted least squares algorithm (Hosmer Jr et al., 2013, Pampel, 2000). An attempt to fit this model is found to result in a perfect separation of the data, which as explained by Albert and Anderson (1984), occurs if there exists a vector "that correctly allocates all observations to their group" (p.3), resulting in at least one independent variable accurately predicting the dichotomous response (Zorn, 2005, Rainey, 2016), thereby perfectly splitting the 0's and 1's.

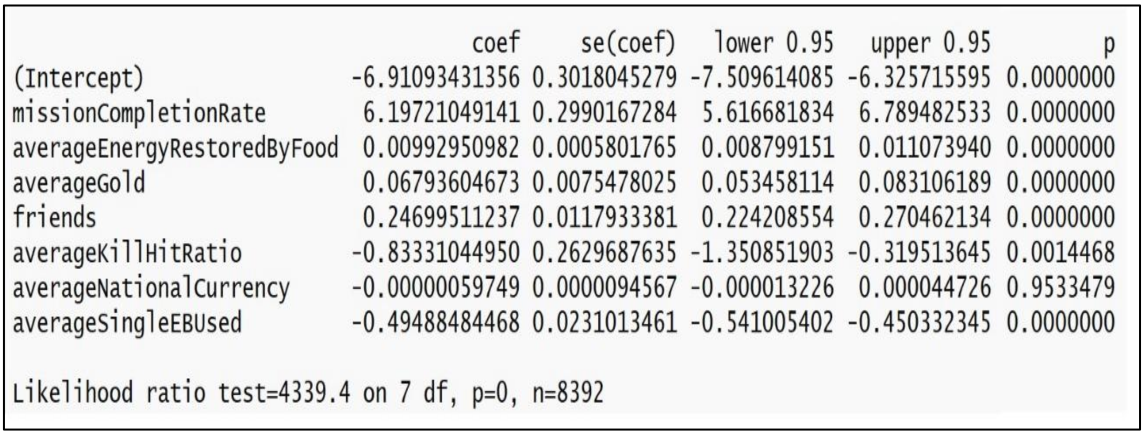
Perfect separation is the source of issues wherein maximum likelihood estimates are not finite resulting in parameter estimates that diverge to infinity, and the standard procedure for the calculation of standard errors falls short (Albert & Anderson, 1984, Zorn, 2005). This problem is overcome by adopting a penalized likelihood function, proposed by Firth (1993), as a substitute of the usual likelihood function. Firth (1993) demonstrates that the asymptotic bias of the maximum likelihood estimator may be eliminated by just penalizing the likelihood by the Jeffrey's invariant prior (Jeffreys, 1946). Heinze and Schemper (2002) and Zorn (2005) discuss and establish through comprehensive empirical study that Firth's penalized likelihood approach offer an impeccable and credible fix to the issue of separation and can be effortlessly implemented without modifying the understanding of standard models. For these reasons and also because "Firth's approach is asymptotically equivalent to (optimal) maximum–likelihood methods in large samples" (Zorn, 2005, p.168), this method was adopted in the multiple logistic regression model in this research.

The regression model is fitted to the training data as described before, this time using the penalized likelihood process. The results from fitting the multiple logistic regression model are presented in table 5.2. The parameters of the model, estimated using penalized maximum likelihood, their standard errors, the more preferable profile penalized likelihood 95% confidence interval (Heinze & Schemper, 2002) and the p-values indicating statistical significance of the model's explanatory variables are reported. All covariates, barring average national currency are found to be statistically significant at the 1% level of significance (p<0.01).

Since this research aims to achieve the model with the best fit and at the same time curtail the number of parameters, a reduced model is fitted after excluding the non-significant variable (Hosmer Jr et al., 2013). The results from this model fit are displayed in table 5.3.

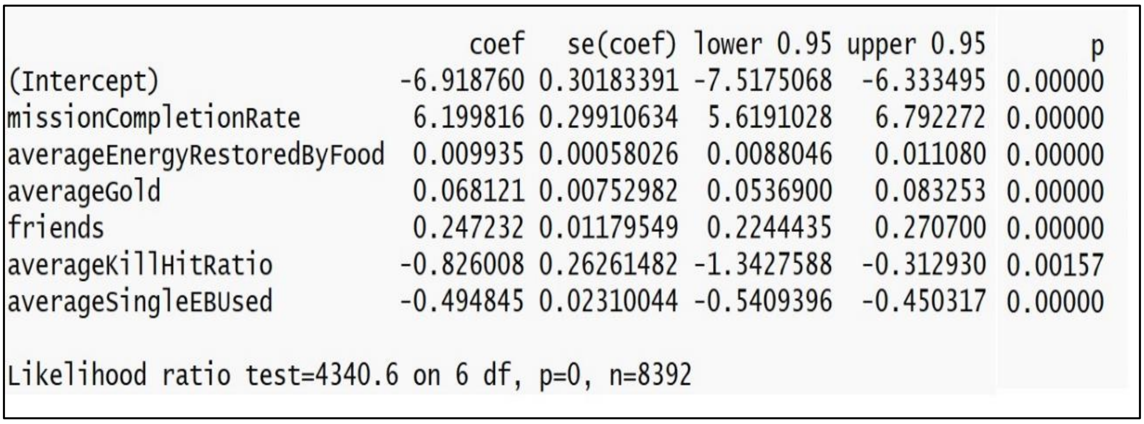
All explanatory variables in the reduced model are statistically significant at the 1% level of significance (p<0.01). The likelihood ratio test, widely considered to be the most optimal and effective (Engle, 1984, Bewick, Cheek & Ball, 2005), evaluates the overall significance of the coefficients for the independent variables in the model (Hosmer Jr et al., 2013) by contrasting the likelihood of acquiring the data when the parameters are zero against the likelihood of acquiring the data assessed at the maximum likelihood estimates of the parameters (Bewick et al., 2005). The penalized likelihood ratio test statistic for the reduced model is 4340.6 (df=6) with a significant p-value (p<0.01). This implies rejection of the null hypothesis at the 1% level of significance, concluding that the penalized maximum likelihood estimates of the current model are more likely to result in the data, and therefore the reduced model holds. Moreover, a penalized likelihood ratio test is conducted to compare the full and the reduced (or nested) model, which follows a chi-square distribution and assumes a null hypothesis that the coefficient for the omitted variable is zero (Hosmer Jr et al., 2013). The test results in a non-significant p-value of P[χ^2(1)>0.0034]=0.95, indicating a lack of evidence in rejecting the null hypothesis, thereby inferring that the nested model is as good as the full model and inclusion of the variable average national currency renders no improvement. Model comparison results are further corroborated by examining the AIC (Akaike's Information Criterion) scores which decide the preference for the optimal model as that which minimizes the AIC score (Akaike, 1973). The AIC scores for the full model and reduced model are -4325.4 and -4328.6 respectively, signifying that the reduced model is the ideal.

Table 5.2: Summary table of results from fitting the full model



```
                                     coef     se(coef)   lower 0.95   upper 0.95         p
(Intercept)                -6.91093431356 0.3018045279 -7.509614085 -6.325715595 0.0000000
missionCompletionRate       6.19721049141 0.2990167284  5.616681834  6.789482533 0.0000000
averageEnergyRestoredByFood 0.00992950982 0.0005801765  0.008799151  0.011073940 0.0000000
averageGold                 0.06793604673 0.0075478025  0.053458114  0.083106189 0.0000000
friends                     0.24699511237 0.0117933381  0.224208554  0.270462134 0.0000000
averageKillHitRatio        -0.83331044950 0.2629687635 -1.350851903 -0.319513645 0.0014468
averageNationalCurrency    -0.00000059749 0.0000094567 -0.000013226  0.000044726 0.9533479
averageSingleEBUsed        -0.49488484468 0.0231013461 -0.541005402 -0.450332345 0.0000000

Likelihood ratio test=4339.4 on 7 df, p=0, n=8392
```

Table 5.3: Summary table of results from fitting the reduced model



```
                                 coef   se(coef) lower 0.95 upper 0.95       p
(Intercept)                 -6.918760 0.30183391 -7.5175068  -6.333495 0.00000
missionCompletionRate        6.199816 0.29910634  5.6191028   6.792272 0.00000
averageEnergyRestoredByFood  0.009935 0.00058026  0.0088046   0.011080 0.00000
averageGold                  0.068121 0.00752982  0.0536900   0.083253 0.00000
friends                      0.247232 0.01179549  0.2244435   0.270700 0.00000
averageKillHitRatio         -0.826008 0.26261482 -1.3427588  -0.312930 0.00157
averageSingleEBUsed         -0.494845 0.02310044 -0.5409396  -0.450317 0.00000

Likelihood ratio test=4340.6 on 6 df, p=0, n=8392
```

5.4.3 Model Diagnostics and Validation

5.4.3.1 Assessment of Model Assumptions

Hosmer, Taber and Lemeshow (1991) assert that that legitimacy of the interpretations from a statistical model hinges on how aptly the model describes or fits the observed data, the failure of which may cause inaccurate or false conclusions to be derived from the model. Diagnostic measures are computed for the model derived above which advise if the model assumptions are met and it fits the data well. Transgressions from the assumptions may contribute towards weak estimates, biased coefficients and unreasonable inferences (Menard & Menard, 2010).

The logistic regression model assumes linearity in the logit function, denoting that the relationship between the log odds of the outcome, $\log(\frac{P(Y=1)}{1-P(Y=1)})$, and the model predictors is linear (Menard & Menard, 2010). This is visually inspected by means of smoothed scatter plots (using LOESS (Cleveland, 1979, Cleveland & Devlin, 1988)) of the logit against the independent variables.

The plots in figure 5.4 display the relationship between the log odds of the outcome that a user is engaged and the variables that are likely to predict it, which is seen to be fairly linear.

Another assumption of this model, elucidated by Stoltzfus (2011), is the absence of outliers that are heavily influential and cause the predicted outcomes for the sample to be exceedingly inconsistent with the actual outcomes. Figure 5.5 is a plot of deviance residuals, used to determine possible outliers in the model (Sarkar, Midi & Rana, 2011, Stoltzfus, 2011).

The influence index plot exhibited by figure 5.5 showing the deviance residuals from the fitted model, reveals a few outliers on both the positive and negative scales. However, not all outliers are influential cases, and as explained by Stoltzfus (2011), the identification of influential outliers involves comparing the model fit and parameter estimates from the full data with that obtained after exclusion of the identified outliers. Conditional on the degree of change, the outliers deemed to have a strong influence on the model are discarded, while the ones without a massive effect are retained (Stoltzfus, 2011). Hence, the model is fitted again after discarding the outliers recognised in figure 5.5 and the summary of the fit, presented in table 5.4 is analysed to see if it is considerably different from the output in table 5.3.

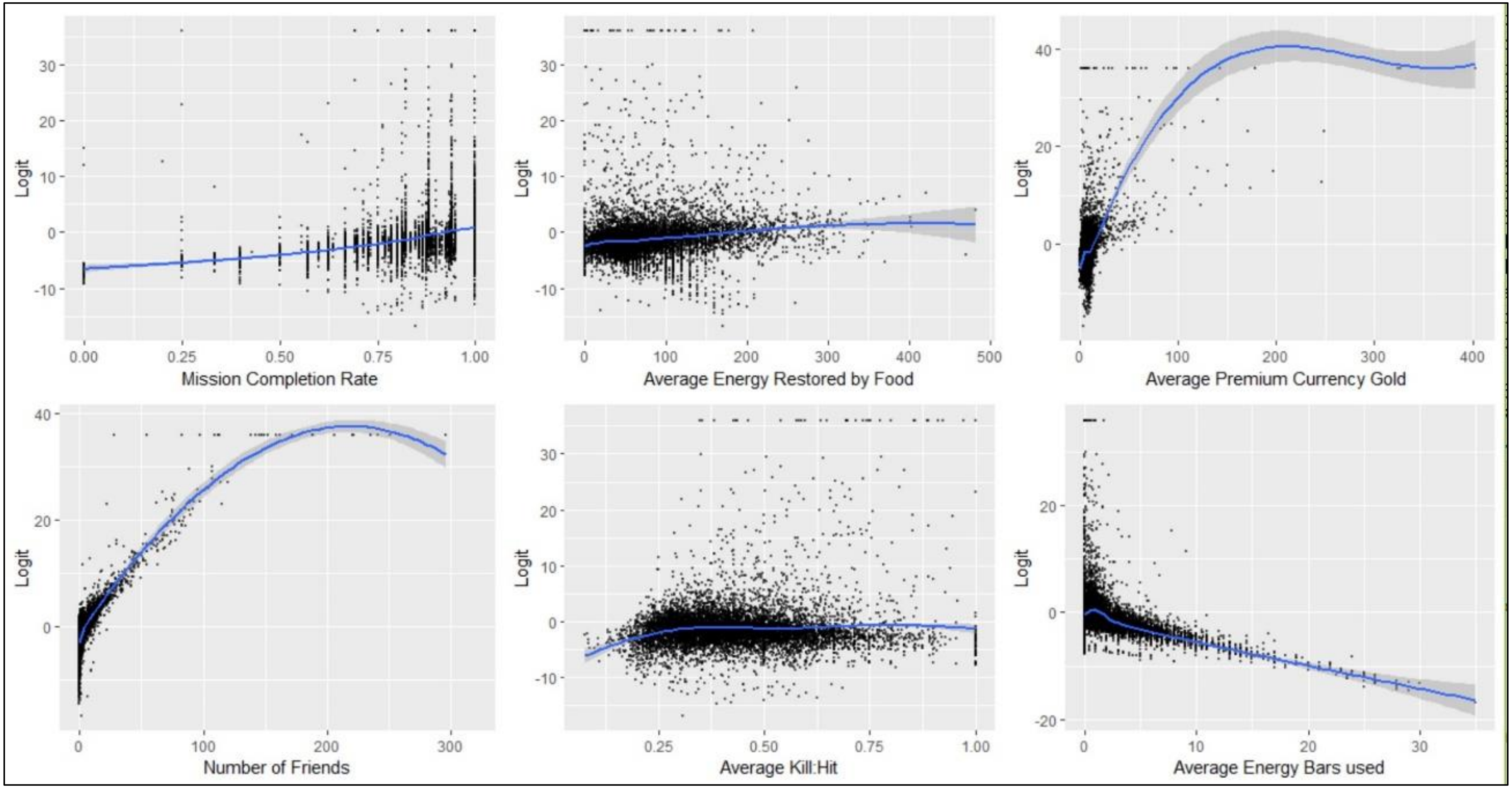


Figure 5.4: Smoothed scatter plots of the logit function against predictor variables in the model

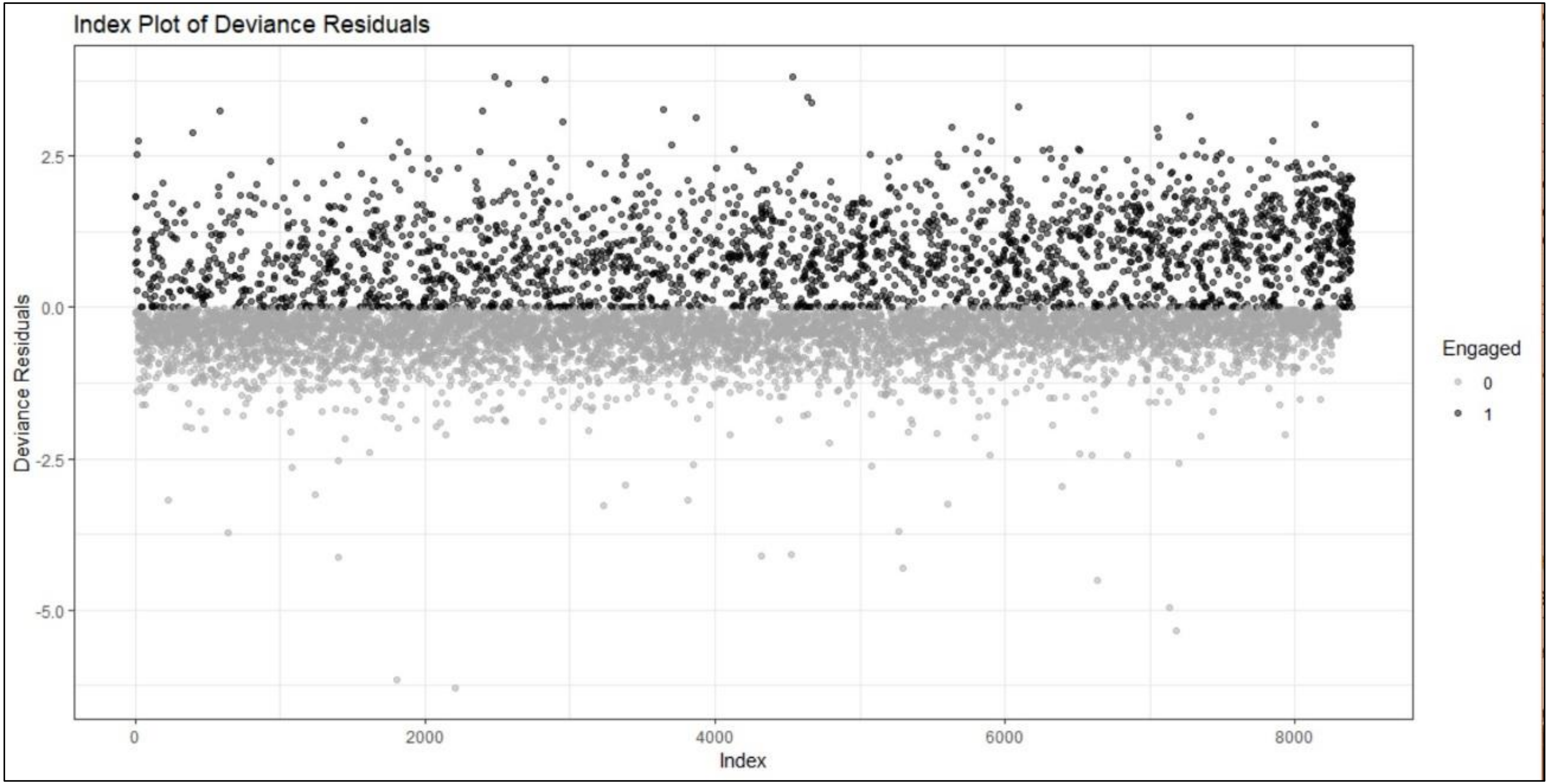
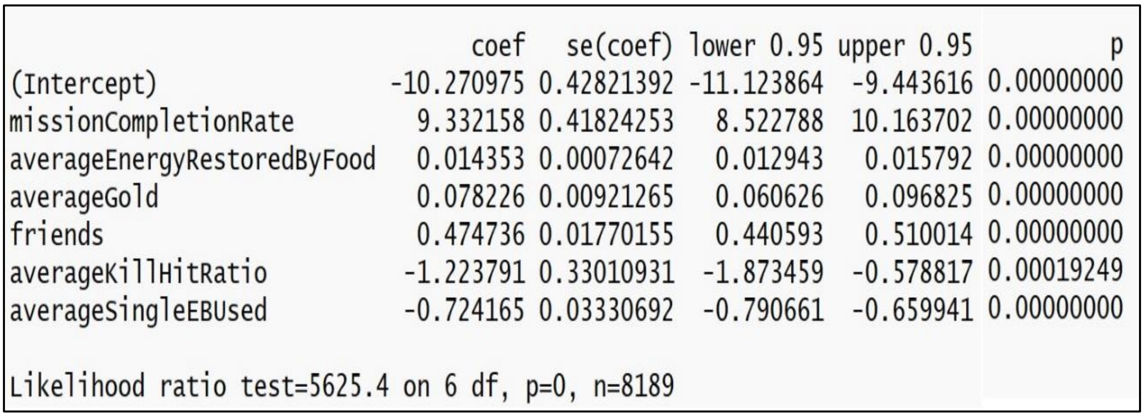


Figure 5.5: Index plot of the influence diagnostic - deviance residual

Table 5.4: Summary table of results from fitting the model after deletion of outliers



```
                                 coef   se(coef) lower 0.95 upper 0.95          p
(Intercept)                -10.270975 0.42821392 -11.123864  -9.443616 0.00000000
missionCompletionRate        9.332158 0.41824253   8.522788  10.163702 0.00000000
averageEnergyRestoredByFood  0.014353 0.00072642   0.012943   0.015792 0.00000000
averageGold                  0.078226 0.00921265   0.060626   0.096825 0.00000000
friends                      0.474736 0.01770155   0.440593   0.510014 0.00000000
averageKillHitRatio         -1.223791 0.33010931  -1.873459  -0.578817 0.00019249
averageSingleEBUsed         -0.724165 0.03330692  -0.790661  -0.659941 0.00000000

Likelihood ratio test=5625.4 on 6 df, p=0, n=8189
```

Contrasting the results from tables 5.3 and 5.4, it is evident that the overall fit of the models are not drastically different from each other in terms of variable significance, neither are the penalized maximum likelihood estimates and their standard errors. Therefore, it can be concluded that although there is presence of some outlier observations used in the model, these are not influential and do not affect the significance of the model and inferences from it.

The final assumption tested with respect to a logistic regression model is the presence of multicollinearity, “which occurs when there are strong linear dependencies among the explanatory variables” (Allison, 2012, p.60). The variance inflation factor (VIF), defined by O’Brien (2007) as the amount by which the estimated variance of the model coefficients is magnified when multicollinearity exists, is a feasible and insightful measure of the existence of serious multicollinearity (O’Brien, 2007). The VIF calculated for the fitted model is reported in table 5.5.

By definition, the VIF quantifies the variance of the coefficients that is inflated due to the presence of correlation among the model’s predictors, and thus a value of 1 for a given independent variable signifies that there is no correlation between that and the remaining independent variables resulting in the variance not being inflated at all. Considering this, as well as a cut-off value of 2.5 for the VIF as suggested by Allison (2012), multicollinearity is not an issue for the model fitted here since VIF<2.5 (in fact slightly greater than 1) for all predictors incorporated in the model.

Table 5.5: Variance inflation factor corresponding to predictor variables in the model

Predictor Variable	**VIF**
Mission Completion Rate	1.035
Average Energy Restored by Food	1.122
Average Premium Currency Gold	1.043
Number of Friends	1.028
Average Kill:Hit Ratio	1.092
Average Energy Bars Used	1.195

5.4.3.2 Assessment of Predictive Power

As stated at the beginning of this chapter, one of the fundamental aims is to build a predictive model of customer engagement in online freemium games, with a view to creating models with good generalisation accuracy. The purpose of the predictive modelling algorithm is to forecast the target variable based on new or future values of explanatory variables (Shmueli, 2010). Breiman (2001) suggest the use of a holdout sample (by laying aside a test data set) to evaluate the predictive accuracy of the model when the sample size is large. This approach is undertaken in order to validate the model developed. The probabilities $\pi(x)$ =P(Y=1|X=x) from equation (1) are predicted using the test sample. A stringent cut-off of 0.8 is used to determine the 1's (engaged) against the 0's (non-engaged), i.e. if the predicted probability is greater than 0.8, the response is classified as 1, else 0.

A confusion matrix is used to describe the efficiency of the solution to a classification problem, via measurement of the performance of a classification model (Patil & Sherekar, 2013). The proficiency of the logistic regression algorithm (a classification model for dichotomous data) implemented here, is measured by the confusion matrix which comprises of information regarding true and predicted classifications by the model (Patil & Sherekar, 2013). Outcome statistics for the confusion matrix calculated from the test sample are illustrated in table 5.6.

Table 5.6: Outcome statistics for the confusion matrix based on test data

Statistic	**Value**
Precision	0.928
Recall	0.386
Accuracy	0.808
Balanced Accuracy	0.686
Kappa	0.449

Sokolova and Lapalme (2009) assert that precision and recall are prevalent performance metrics since they emphasize the retrieval of cases denoting response ‘success’ (user engagement here) and do not consider the correct classification of response ‘failures’ (non-engagement). Precision, defined as the “fraction of retrieved instances that are relevant” (Patil & Sherekar, 2013, p.258) is “a measure of exactness or quality” (Patil & Sherekar, 2013, p.258) and represents that out of all the predictions for engaged, 93% of cases are actually engaged. Recall, defined as the “fraction of relevant instances that are retrieved” (Patil & Sherekar, 2013, p.258) is “measure of completeness or quantity” (Patil & Sherekar, 2013, p.258) and represents that out of all the true engaged cases, 39% are predicted as engaged. Although this reflects that a considerable amount of engaged cases are predicted as non-engaged, its implications are not that critical. In practice, predicted non-engaged customers will be offered with improved gaming experience (to convert them to engage before they drop out) and therefore even if this is mistakenly extended to some already engaged users, the impact will be positive. On the contrary, it is vital to minimise prediction of non-engaged cases as engaged, since that could deprive users at a risk of quitting the game, of an enhanced gameplay experience, by erroneously considering them as already engaged. The confusion matrix reveals only 25 (~1.3%) instances where true non-engaged cases are predicted as engaged. The overall accuracy of the model, denoting how frequently it produces correct classifications is fairly high at 0.81. The balanced accuracy, described by Brodersen, Ong, Stephan and Buhmann (2010) as a generalisability measure representing “the average accuracy obtained on either class” (p.3122), used to overcome the bias introduced by unbalanced data with different numbers of representations from either category (1 and 0), is moderate at 0.69. Finally, the Kappa statistic (Cohen, 1960), a measure of performance of the classification model assessed here relative to its performance by random chance, is 0.45, which according to Landis and Koch (1977) signifies a moderate strength.

In addition to the confusion matrix, ROC (Receiver Operating Characteristic) curve analysis is also implemented in order to quantify how efficiently the fitted model can discriminate between the two groups of customers – engaged and non-engaged. The ROC curve, elucidated by Swets (1979) and Metz (1986), assesses the trade-off between the true positive rate and the false positive rate across varying thresholds for assigning observations to the class represented by 1 (the threshold originally used here is 0.8). The true positive rate relates to the cases correctly predicted as engaged, while the false positive rate is relates to the cases incorrectly predicted as engaged. Figure 5.6 depicts the ROC curve derived for the fitted model predicting user engagement by using the test sample.

The ROC curve in figure 5.6 is summarised with the help of an index called AUC denoting the area under the ROC curve, which represents the probability that a randomly drawn engaged user is more likely to be classified as engaged than a randomly drawn non-engaged user (Hajian-Tilaki, 2013). The AUC for the fitted model is 0.913, indicating that the fitted model is quite adept in discriminating between the engaged and non-engaged cases comprising its response variable.

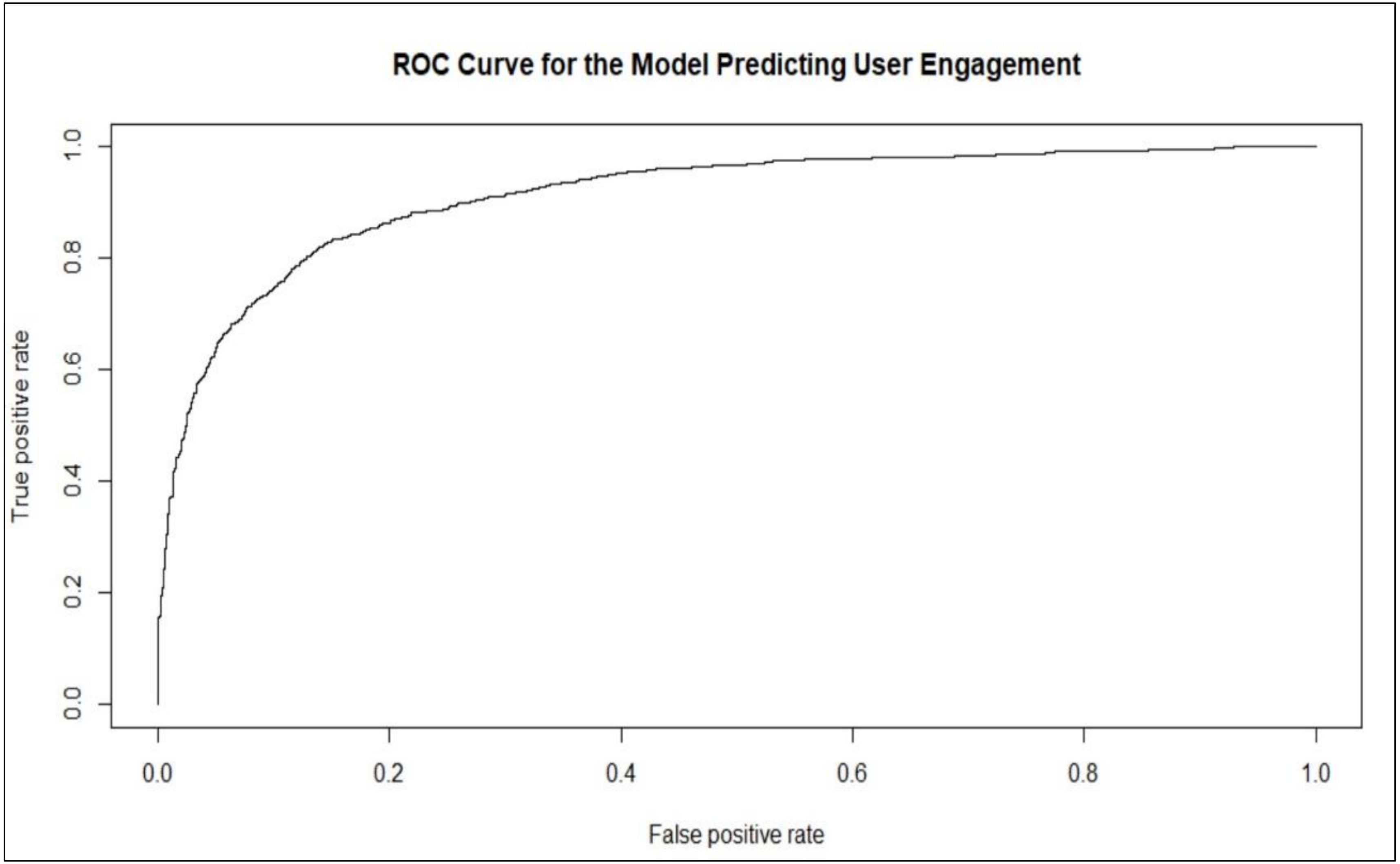


Figure 5.6: ROC curve for the fitted multiple logistic regression model predicting customer engagement

5.5 Results and Interpretations

Evaluation of model assumptions and predictive validity demonstrate fairly good fit and accuracy despite assuming a rigid cut-off probability of 0.8 for assigning observations to the engaged and non-engaged categories. Therefore, the model fitted earlier is accepted as reasonably explaining and predicting customer engagement in online freemium games, and inferences drawn from it are subsequently elucidated below.

The final fitted model following from equation (1) is of the form –

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = -6.92 + 6.20(mission\ completion\ rate) + \\ 0.01(average\ energy\ restored\ by\ food) + \\ 0.07(average\ premium\ currency\ gold) + 0.25(number\ of\ friends) - \\ 0.83(average\ kill{:}\ hit\ ratio) - 0.49(average\ energy\ bars\ used) \qquad (2)$$

The model describes that the variables significant in predicting users being engaged are – the rate at which they complete missions, amount of premium currency possessed, the number of virtual friends they have, the ratio of kills to hits in a virtual fight and the quantity of virtual resources (food and energy bars) utilised to enhance their in-game performance. The model coefficients from table 5.3 are examined to gain insight on the effect of a unit change in the explanatory variables on the log odds of engagement. Increase in all the independent variables, with the exception of kill:hit ratio and amount of energy bars used, indicate greater likelihood of users being engaged than not. Hence, these variables are deemed to have a positive impact on customer engagement and players on the verge of quitting should be guided towards improving these metrics. On the contrary, increase in kills:hits or consumption of energy bars have an adverse effect on engagement, with users being less likely to be engaged, and as such players should be directed towards controlling these statistics. This is because, with greater participation in virtual wars (i.e.kills:hits), more is the energy required by players to fight their opponents. This may lead to an increased consumption in energy bars (used to replenish energy), which in turn may result in quicker depletion of that resource. A lack of energy bars may prevent players from acquiring the energy needed to be involved in virtual fights, thereby causing them to not fully connect with or be unduly interested in the game

The predictive model developed can be executed in practical situations by game developers to forecast engaged and non-engaged customers at set intervals. Explanatory variables can be computed for the player base approximately every fortnight to predict players that are engaged, and considering the remaining as not immersed with the game and thus at a risk of dropping. The engaged players may be rewarded with premium content to keep them further involved in the gameplay, while the latter can be assisted towards a more enhanced gameplay experience to prevent them from quitting. The game studios are undoubtedly better equipped to come up with preventive strategies in this respect, however based on the understanding from the model fitted, some approaches that could be undertaken are suggested. Users identified as non-engaged could be -

- provided with hints and tips to improve their competency in missions
- guided towards exhausting the available virtual resources to improve their in-game characters strength and skills
- simultaneously also advised to train and be involved in other activities that may help replenish these important resources
- rewarded with premium currency or virtual resources that are on the verge of depletion
- offered suggestions about striking a balance between participating in virtual fights as well as exploring other aspects of gameplay so that they have a varied experience and judicious use of the required resources to indulge in fights
- directed towards socialising with other players

Finally, although the model is developed and tested using data from a specific game eRepublik, the predictor variables' interpretations, implementation of the model in the real world and suggestions are generalisable to a vast majority of online freemium with a similar structure.

6. Prediction of Time to Defection of Customers

The business model of freemium games is largely built upon a digital marketing phenomenon called the network effect (Pahwa, 2017), which states that, with increase in the use of a service or product, its worth also increases (Dutta, 2018). The freemium model has also led to the emergence of a new concept known as the Newtonian Engagement thereby identifying engagement as the fundamental catalyst for freemium games (Pahwa, 2017). One of the challenges of this business model is to maintain engagement and addiction whereby users are encouraged to return and purchase premium services in the long-run (Pahwa, 2017). A crucial factor that dictate the revenue model of freemium games is retention of players to the game and consequently prediction of players' time to defect or drop out of it. From the point of view of game designers and publishers, predicting how long players will continue in the game after they join is critical to the popularity of the game and consequent earnings from it.

6.1 Need for Analysing Customer Defection Times

Players generally resign from a game if they are not content with its structure, composition and features, which makes this a good indicator of low user satisfaction (Tarng et al., 2008). The ability to predict when customers drop out of the game may provide an opportunity to developers to enhance their gameplay experience, thereby preventing them from leaving (Tarng et al., 2008). As explained by Hadiji et al., (2014), the revenue produced by freemium games is heavily reliant on in-game purchases and in-game advertising, and hence predictive models of future player behaviour are critical data-driven methods to acquire knowledge about design, advancement and business of these games.

In this chapter, survival analysis of players' lifetimes in the game is conducted and the resulting model is used to predict when a player is likely to drop out and what factors influence that behaviour. The model is based on variables that explain the different aspects of users' interactions with and performance in the virtual world of the game. This is also called churn analysis or churn prediction, an essential aspect of online businesses (Kawale et al., 2009), which is described by Hadiji, et al., (2014) as the process of determining subscribers of an online or mobile platform that are expected to quit in the future i.e. inclined to cancel their subscription to the service (Kawale et al., 2009). These players are typically called churners and the churn rate is defined as the ratio of churners to non-churners over time (Mutanen, Ahola & Nousiainen, 2006). Churn rates are usually observed to be high with associated frequency distributions being heavily skewed, thereby implying that most players leave in the initial stages of gameplay (El-Nasr, Drachen & Canossa, 2013, Fields & Cotton, 2011, Luton, 2013, Sifa, Bauckhage & Drachen, 2014). The significance of predicting player churn is highlighted by Hadiji, et al., (2014) as an opportunity to distinguish players that are detached with the game and at a risk of defecting, and therefore implement mechanisms that motivate them to engage with the game and make in-app purchases in order to boost revenue. This is particularly relevant to game publishers as it provides fundamental insight into the behaviour of players and various causes for them abandoning the game such as "personal commitment, competing products, shifting interest to social influence" (Kawale et al., 2009, p.1). The successful performance of a game is generally measured by retention and monetisation - concepts that are closely intertwined with one another (Fields & Cotton, 2011, Luton, 2013, Drachen, Thurau, Togelius, Yannakakis & Bauckhage, 2013).

6.2 Survival Analysis

Survival analysis is an assortment of statistical techniques to analyse the expected duration of time until an event occurs (Kleinbaum & Klein, 2011). This involves the study of positive valued random variables that usually denote the "time to the failure of a physical component" or the "time to the death of a biological unit" or the "time to the learning of a skill" (Miller Jr, 2011). Survival data analysis is accomplished through survival models, in which the response variable denotes the waiting time for a distinct event of interest to happen; at the time of data analysis, the event of interest is yet to occur for some individuals/units which results in censoring of certain observations; and one of the objectives of the analysis is to investigate or control the effect of explanatory variables on the time to the event of interest (Rodríguez (2007).

6.2.1 Survival and Hazard Functions

Two essential concepts associated with describing survival distributions are hazard rates and survival times (Hosmer, Lemeshow & May, 2011, Moore, 2016). Hazard rate represents the conditional probability of an event of interest transpiring at a specific time interval t, while survival time signifies the duration or time period prior to that event occurring (Mills, 2011). Mathematically, these concepts can be illustrated as follows.

Given a non-negative continuous random variable T with probability density function $f(t)$ and cumulative distribution function $F(t)$, the probability of the event of interest to have happened by time period t is denoted by $F(t) = P(T < t)$ (Rodríguez, 2007). T

symbolises survival times and is non-continuous only in the analysis of discrete-time models (Mills, 2011).

The complement of $F(t)$ is defined as the survival function, $S(t) = 1 - F(t) = P(T \geq t) = \int_t^\infty f(x)dx$ which explains the probability of survival up to time point t (Rodríguez, 2007).

The distribution of T is alternatively summarized by the hazard rate or hazard function $h(t) = \frac{f(t)}{S(t)}$ which describes the instantaneous risk of the event taking place in the time interval $[t, t + dt)$ given survival until t without experiencing the event (Mills, 2011).

The survival and hazard functions are described by Mills (2011) as counterparts of each other in that the survival function is concerned with not encountering the event of interest (i.e. surviving), whereas the hazard function deals with experiencing the event of interest (i.e. failing).

6.2.2 Censoring of Observations

Survival analysis models allow for the provision of handling missing data, which is also known as censoring (Springate, 2014, Mills, 2011, Rodríguez, 2007). Some individuals/units have already experienced the event of interest and information is available on their exact survival times, whereas for some others the event is yet to occur and the survival times are assumed to be after the termination of the study i.e. they surpass the observation time (Rodríguez, 2007, Springate, 2014). For the latter, the survival times are only partially observed and they are said to be right censored, which is the most prevalent type of censoring (Moore, 2016). A less frequent phenomenon is that of left censoring where the event of interest has already taken place for some individuals/units before the beginning of the study (Gomez, Julià, Utzet & Moeschberger, 1992).

6.2.3 Types of Survival Models

Rodríguez (2007) explains that the population of individuals/units studied is usually heterogeneous and their lifetimes are not dictated by the same survival function $S(t)$, instead, influenced by a set of explanatory variables or covariates whose effects are to be modelled. This is achieved through survival models, which can be broadly classified as non-parametric, semi-parametric and parametric models (Mills, 2011, Rodríguez, 2005). An outline of these models is as follows –

- Non-parametric models – lifetable estimates (Mills, 2011), Kaplan-Meier estimators (Mills, 2011, Rodríguez, 2005), Mantel-Haenszel approach for comparing multiple survival functions (Rodríguez, 2005) and Cox's partial likelihood estimation

(Rodríguez, 2005). These are useful tools for descriptive analyses but do not incorporate multiple covariates or multivariate controls into the model (Mills, 2011).

- Semi-parametric models – Cox's proportional hazards model (Guo & Zeng, 2014, Mills, 2011), proportional odds model (Guo & Zeng, 2014, Bennett, 1983), piecewise exponential model (Rodríguez, 2007, Mills, 2011), additive hazards model (Guo & Zeng, 2014) and accelerated failure time model (Guo & Zeng, 2014). These models are flexible and make less restrictive assumptions about the baseline hazard function, which renders a broader applicability (Rodríguez, 2007, Mills, 2011). Although there is provision for the incorporation of multiple covariates and multivariate controls, the models may not be as suitable in testing the change of hazard over time (Mills, 2011).
- Parametric models – exponential, gamma, Weibull, generalized F distributions, logistic, Gaussian, complementary log-log, log-logistic, log-normal, Gompertz, Makeham, extreme value, Rayleigh etc. (Rodríguez, 2007, Mills, 2011). These models can estimate parameters more accurately, allow multivariate analysis, handle both discrete and continuous covariates and also achieve predictive modelling (Mills, 2011). However, these are very susceptible to the explanatory variables being inserted or dropped from the model, and can result in strongly biased estimates if the model assumptions are not met or the distribution of hazard function wrongly described (Mills, 2011).

6.3 Method Adopted for Customer Defection Time Analysis

Taking into account existing approaches towards churn analysis and a general overview of survival analysis, the procedure adopted for investigating defection times of players' of online freemium games is survival analysis and Cox's proportional hazards modelling. The rationale behind this choice is by virtue of the exclusive qualities possessed by survival analysis techniques and Cox's model that makes these a more favourable approach than other methods in this case.

Survival analysis incorporates knowledge about the time period of a study thereby accounting for censoring of observations (Mills, 2011). This makes it a superior method to logistic regression which does not employ knowledge about the time point during the follow-up period at which the event of interest takes place (Green & Symons, 1983), hence disregarding the duration of the study period (Myers, Hankey & Mantel, 1973), making it a complicated procedure that may be hard to justify (Green & Symons, 1983). Survival analysis provides insight about the outcome of interest and evaluates the time to an event, hence allowing the comparison of survival times between multiple groups and study the association between the model covariates and survival time (Mills, 2011). Survival models are usually called dynamic or process models (Aalen, Borgan & Gjessing, 2008, Willekens, 1991) because of their distinctive characteristic of being able to assess explanatory variables that vary over time, a trait not found in regression models such as ordinary least squares or logistic (Mills, 2011). Cleves (2008) mentions another advantage of survival models over ordinary least squares regression with respect to the assumption of normality. OLS regression assumes that the residuals are distributed normally which may often be an illogical assumption for survival data on time to a particular event. Although linear regression models are exceptionally robust to departure from normality, survival data distributions, in addition to being non-normal, are mostly non-symmetric and often bimodal, characteristics to which the linear regression is not robust. In such cases, a more plausible distributional assumption for the residuals is provided by survival models, making it the more favourable choice.

The comparative study between logistic regression and survival models (specifically Cox's proportional hazards model) conducted by Green & Symons (1983) concludes that for an adequately short follow-up period, results from both models tend to be similar. However, with the increase in follow-up time, standard error of the estimates from the proportional hazard model decreases, making it more accurate, whereas that from the logistic regression model increases with longer follow-up. Green & Symons (1983) also emphasize the appropriateness of Cox's model as providing a better fit to the data, since it includes more information by making use of the time of the response as well. This model has an undefined baseline hazard function with no chance assumptions about the distribution or shape of it (Springate, 2014), thereby enhancing its suitability particularly in cases where the basic model assumptions are not met (Bender, Augustin & Blettner, 2005). Therefore, when investigating survival time data with no knowledge or bias regarding its underlying distribution, Cox's proportional hazards model is an appealing approach. Additionally, the model can account for both quantitative and qualitative predictor variables and determine the concurrent effects of multiple risk factors on the time to an event (Easy Guides, 2016). Moreover, being a hazards model, it employs the hazard function to model survival times, which is convenient in the presence of censored data (Bender et al., 2005).

6.4 Defection Time Data of Customers

As stated before, gameplay data acquired from 40716 new users of eRepublik in the period 11th November 2013 and 6th January 2014 forms the basis of this analysis.

However, as discussed in the last chapter, about 30% of this player base interact with the game for a single session only, i.e. they do not return after their very first logout. A comprehensive analysis of these players revealed that they are not attracted to the game right from the onset and do not exhibit adequate consumer behaviours to study, and therefore are excluded from the survival analysis.

For the purpose of this analysis, the users are then defined as defected or not, wherein, the event of interest, customer defection from the game, is construed as having used the gaming platform for at least 5 hours, but not having logged in for more than two weeks from the end of the analysis period. This will ensure that the players identified as defected have connected with and had sufficient experience of the features of the game, having played for 5 hours at least; but can be assumed to have dropped out since they have not been seen playing the game for more than two weeks from the end of the analysis period. The rationale for this stems from the fact that online freemium games like eRepublik usually employ daily reward mechanisms and daily tasks to keep players engrossed and returning, and thus absence for more than two weeks could be quite certainly considered as quitting the game. The survival times for customers considered defected are indicated by their total gameplay time (in minutes) up to the point they are last seen on the platform i.e. until they defect. The remaining users are assumed to have 'survived' i.e. not defected at least until the end of the analysis period, and their survival times are right censored and is indicated by their total gameplay time (in minutes) up to the end of the analysis period.

The final data used in survival analysis is a player-level summary consisting of total gameplay time (in minutes) representing survival times, defection status assuming values 1 (defected) and 0 (censored or not defected within the analysis period) and several other user specific variables (covariates) that depict gameplay behaviour and performance. The total number of defected and censored customers within the analysis period is 8506 and 8185 respectively.

6.5 Survival Analysis of Time to Defection

The overall purpose is to study and model survival time or time to churn or defect, and determine the factors that influence them. The survival data detailed above is applied in the statistical techniques comprising survival analysis and appropriate models are formed.

6.5.1 Kaplan-Meier Estimates

The Kaplan-Meier estimate (Kaplan & Meier, 1958), a non-parametric maximum likelihood estimate of the survival function $S(t)$, is computed. It is a descriptive tool for exploratory analysis, and is noted by Kaplan and Meier (1958) as useful for dealing with partial information on survival times i.e. censored observations.

Figures 6.1 and 6.2 illustrate the survival curves (using Kaplan-Meier estimates) for all users and for users stratified by their paying status (0: non-payers and 1: payers). The Kaplan Meier curve is a plot of the estimated survival function. It is a non-continuous stepwise estimate of survival probabilities (y-axis) for different intervals of time (x-axis) (Rich et al., 2010).

The Kaplan-Meier curve in figure 6.1 shows the decreasing probabilities for survival of customers with the passage of time. The dashed lines are the upper and lower confidence intervals. It is a self-explanatory curve from which, the probabilities of survival at different time points and the corresponding margin of expected errors represented by the confidence intervals can be visually judged. For instance, the probability of surviving in the game i.e. continue playing it after about 10000 minutes (~7 days) of gameplay is approximately 0.35 or 35%. The median survival time of these players is 1630 minutes (~27 hours) with 95% confidence intervals [1560, 1710] minutes.

Similarly, the curve is figure 6.2 shows the decreasing probabilities for survival, separately of customers with a real currency transaction and those without one, with the passage of time. The black line represents the survival curve for paid users, while the grey line represents the same for non-paid users. Moreover, the plot also shows the p-value of a log rank test, where p<0.0001 indicates a significant result at the 5% level of significance, thereby implying that the survival probabilities for individuals that make monetary purchases in the game is significantly different (better in this case) than those that do not. This is very pronounced in the graph as well ,where the probability of surviving in the game after about 10000 minutes (~7 days) of gameplay is approximately 0.30 for non-payers, whereas the same for payers is almost 0.80.

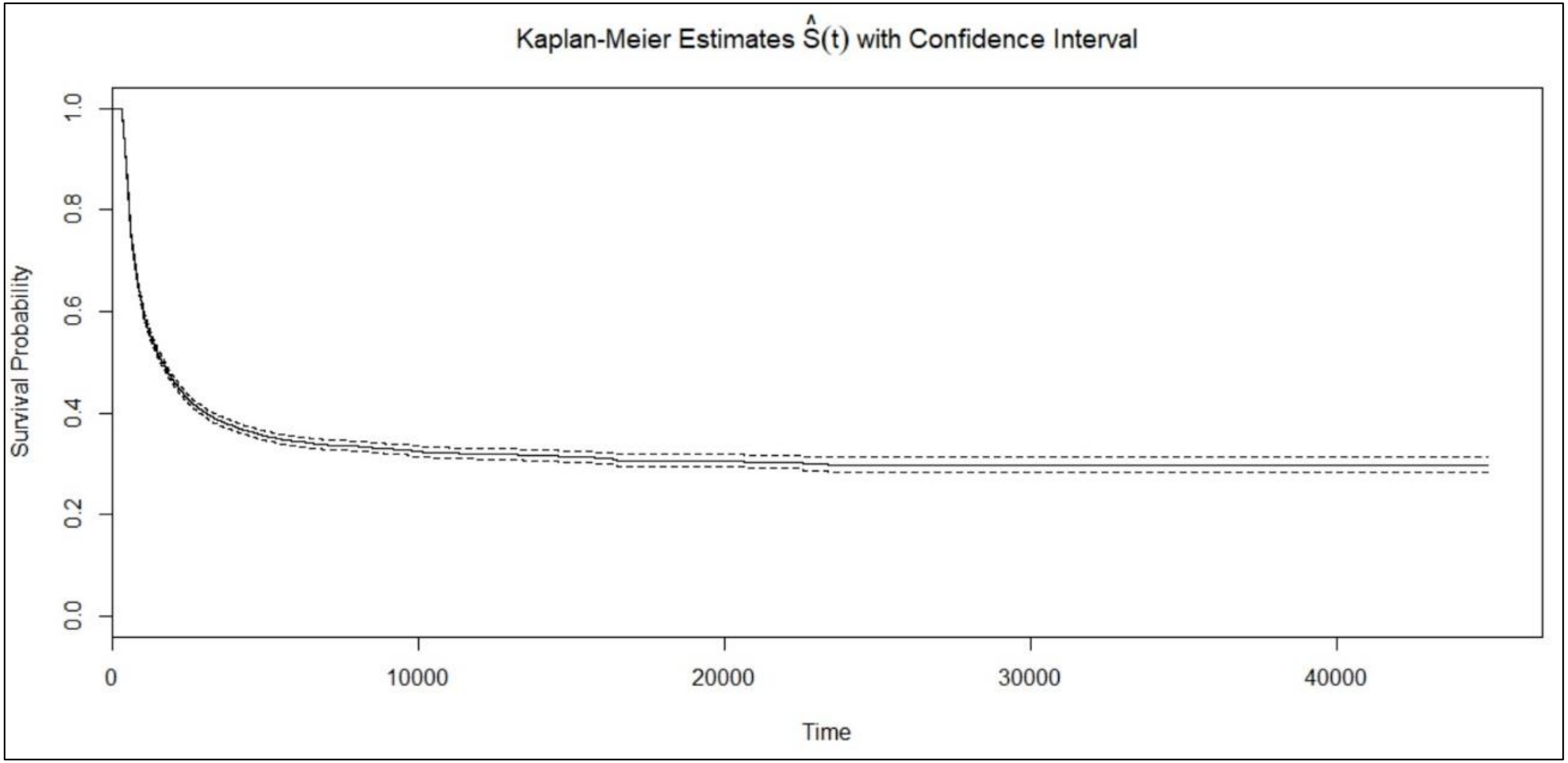


Figure 6.1: Kaplan-Meier survival curve for all players

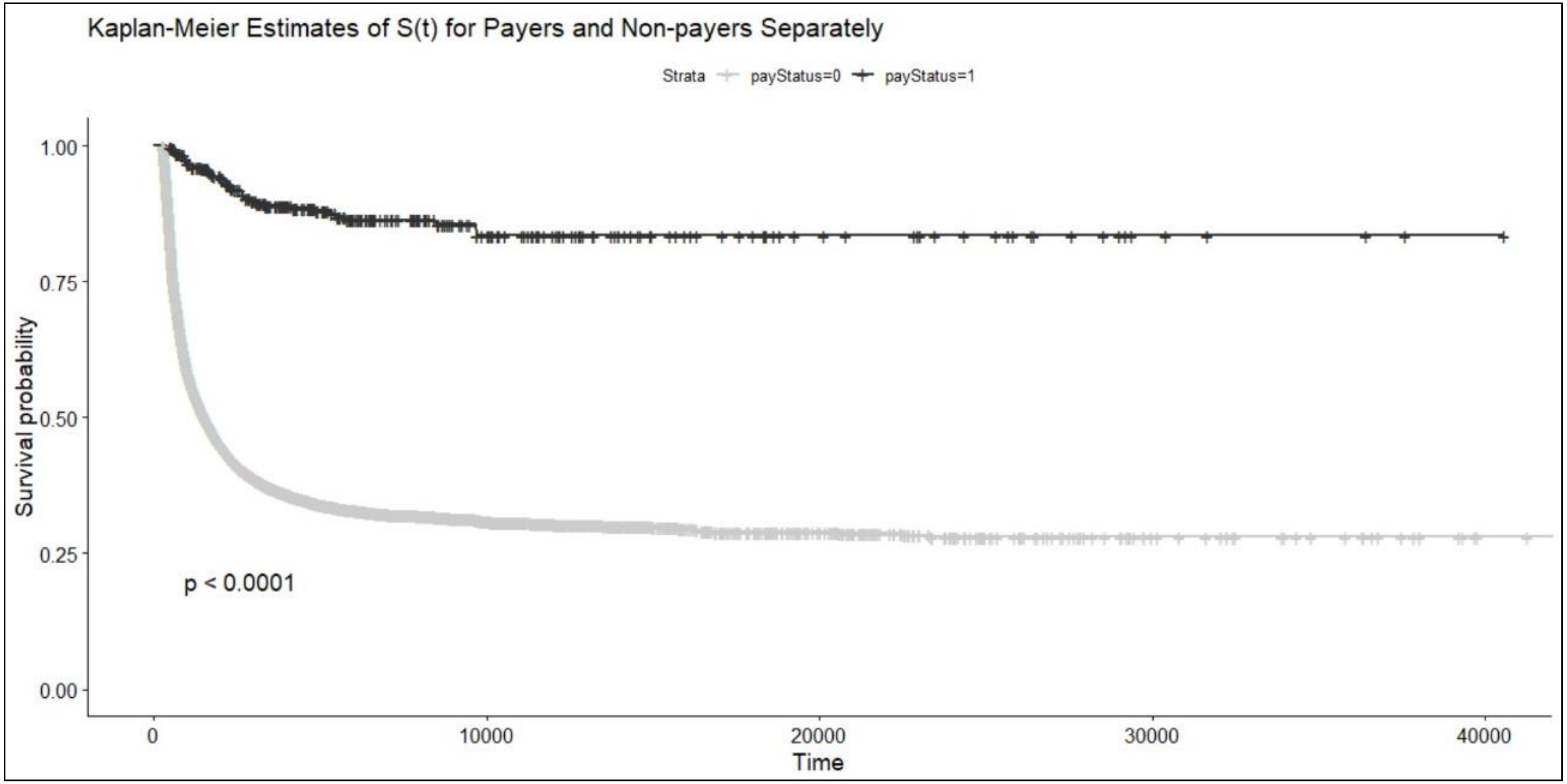


Figure 6.2: Kaplan-Meier survival curve for players stratified by their paying status

6.5.2 Cox's Proportional Hazards Model

Kaplan Meier estimator being a non-parametric method for survival analysis is univariate in nature and fails to incorporate multiple covariates or multivariate controls into the analysis, thereby explaining survival based on any one factor being considered (for example, paying status as demonstrated above) and disregarding the effect of others. Moreover, it is effective only in the case of a categorical explanatory variable and not suitable for handling quantitative variables (Easy Guides, 2016).

Therefore, a more appropriate approach by means of a semi-parametric survival model called Cox's proportional hazards model (Cox, 1972) is adopted. It is typically a regression model (Easy Guides, 2016) used to analyse the effects of covariates on the survival time of individuals/units (Springate, 2014) and most widely used in medical research (Cox, 1972). As already stated, the model can account for both quantitative and qualitative predictor variables and determine the concurrent effects of multiple risk factors on the time to an event (Easy Guides, 2016).

The Cox's proportional hazards model is given by $h(t|x) = h_0(t)e^{\widehat{\beta}x}$ (1)

Where t denotes time, $h_0(t)$ is the baseline hazard function which the model does not estimate and is defined as the hazard function corresponding to all covariates taking the value zero (Royston, 2011), β is the vector of regression coefficients and $\hat{\beta}$ its corresponding estimates, and x is the set of explanatory variables (Bender et al., 2005).

6.5.2.1 Model Covariates

Cox's proportional hazards model is fitted to the data with user-level covariates that reflect their playing styles and performance within the game. The primary objective is to estimate the parameters β i.e. $\hat{\beta}$ using partial likelihood estimates (Cox, 1975) in order to investigate the effect of the model covariates on the rate of defecting from the game (i.e. the hazard rate) at a specific time point (Easy Guides, 2016). The explanatory variables considered in this model are the same ones examined when modelling customer engagement. This is because it is conceptualised that the factors that may affect user engagement is also likely to influence survival times of users in the game and hence their defection from the game as well. Attempt is made to verify this using Cox's model, while also considering an additional covariate paying status, by virtue of the fact that the Kaplan-Meier estimation process has already indicated that a significant difference may exist between individuals depending on whether they make a real money in-game transaction or not.

Descriptive statistics of the predictor variables have been extensively studied and approaches undertaken to handle missing observations detailed in the previous chapter (section 5.4.1). The only additional variable included here is paying status, which is distributed as 0 (non-payers): 16344 and 1 (payers): 347.

6.5.2.2 Model Fitting

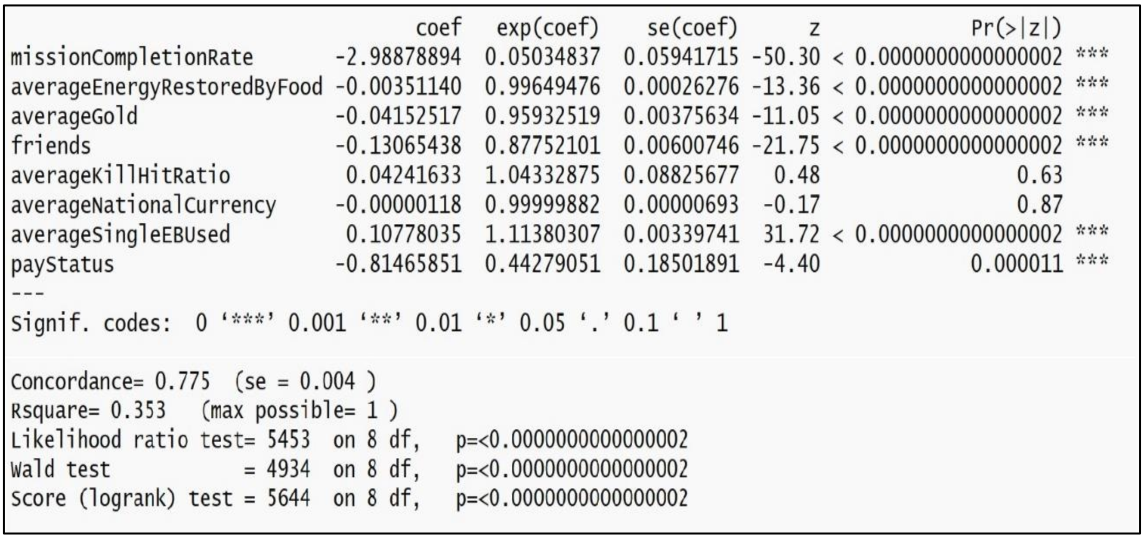
As before, the survival data is first partitioned into training and test samples based on a 75:25 ratio. The training set is used to build the Cox's proportional hazards model and learn about the relationship between the hazard function and explanatory variables. The final model developed is validated and its performance estimated using the test set.

A multivariate Cox's proportional hazards model, following from equation (1) is fitted to the training data using survival time and covariates that have been described above. Estimation of the parameters in the model is via maximisation of the partial likelihood (Nikulin & Wu, 2016) and handling of ties is by the Efron approximation (Efron, 1977). The results from the model fitting are displayed in table 6.1.

The results from fitting the multivariate Cox's proportional hazards model are presented in table 6.1, which reports the following. The estimated parameters $\hat{\beta}$ of the model, exponentiated coefficients that are hazard ratios representing multiplicative effect of a covariate on the hazard function (Springate, 2014, Fox & Weisberg, 2018), standard errors of the parameter estimates, the Wald (z) statistics testing the null hypothesis that the corresponding $\hat{\beta}$ coefficients are 0, (Fox & Weisberg, 2018) and their associated p-values indicating statistical significance of the covariates are reported. All covariates, barring average kill:hit and average grind currency (national currency) are found to be statistically significant at the 1% level of significance ($p<0.01$).

Survival curves, using estimates from the fitted Cox's model, for all users and for users stratified by their paying status (0: non-payers and 1: payers) are visualised in figures 6.3 and 6.4.

Table 6.1: Summary table of results from fitting Cox’s proportional hazards model



```
                                coef   exp(coef)    se(coef)      z              Pr(>|z|)
missionCompletionRate       -2.98878894  0.05034837  0.05941715 -50.30 < 0.0000000000000002 ***
averageEnergyRestoredByFood -0.00351140  0.99649476  0.00026276 -13.36 < 0.0000000000000002 ***
averageGold                 -0.04152517  0.95932519  0.00375634 -11.05 < 0.0000000000000002 ***
friends                     -0.13065438  0.87752101  0.00600746 -21.75 < 0.0000000000000002 ***
averageKillHitRatio          0.04241633  1.04332875  0.08825677   0.48                  0.63
averageNationalCurrency     -0.00000118  0.99999882  0.00000693  -0.17                  0.87
averageSingleEBUsed          0.10778035  1.11380307  0.00339741  31.72 < 0.0000000000000002 ***
payStatus                   -0.81465851  0.44279051  0.18501891  -4.40              0.000011 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Concordance= 0.775  (se = 0.004 )
Rsquare= 0.353   (max possible= 1 )
Likelihood ratio test= 5453  on 8 df,   p=<0.0000000000000002
Wald test            = 4934  on 8 df,   p=<0.0000000000000002
Score (logrank) test = 5644  on 8 df,   p=<0.0000000000000002
```

Plots of the survival functions estimated by Cox’s proportional hazards model are displayed in figures 6.3 and 6.4. As usual, these show the decreasing probabilities of survival with the passage of time, and the corresponding margin of expected errors represented by the confidence intervals. In figure 6.3, it is observed that the probability of surviving in the game i.e. continue playing it after about 10000 minutes (~7 days) of gameplay approximately 0.23 or 23%. In figure 6.4, the black line represents the survival curve for paid users, while the grey line represents the same for non-paid users. Pay status is already seen to be a significant variable in the fitted Cox’s model, thereby implying that the survival probabilities for individuals that make monetary purchases in the game is significantly different (better in this case) than those that do not. This is also evident from the graph, where the probability of surviving in the game after about 10000 minutes (~7 days) of gameplay is slightly greater than 0.20 for non-payers, whereas the same for payers is about 0.55.

A different perspective of the model fitting results can be demonstrated via a forest plot in figure 6.5.

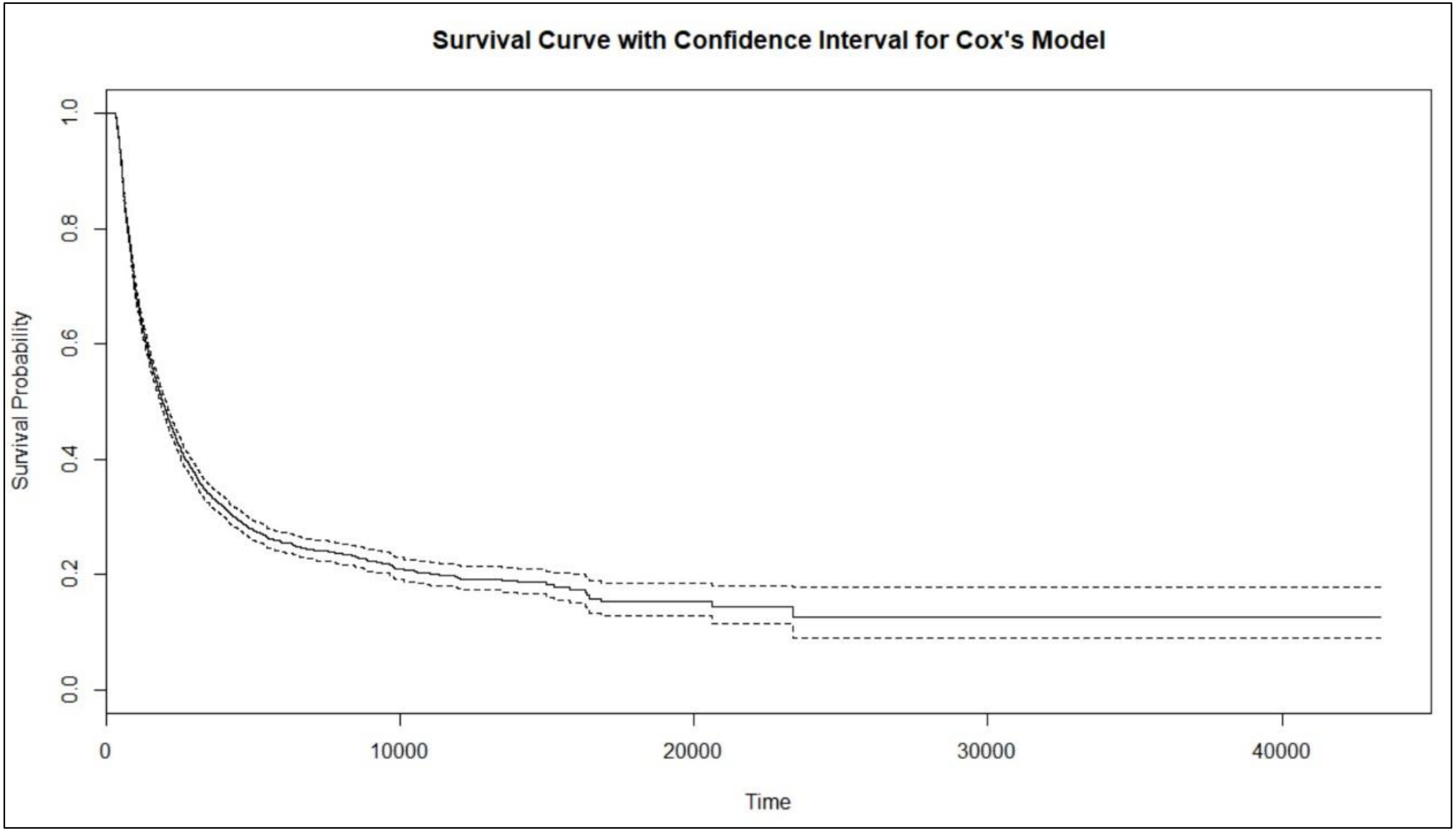


Figure 6.3: Survival curve for Cox’s proportional hazards model for all players

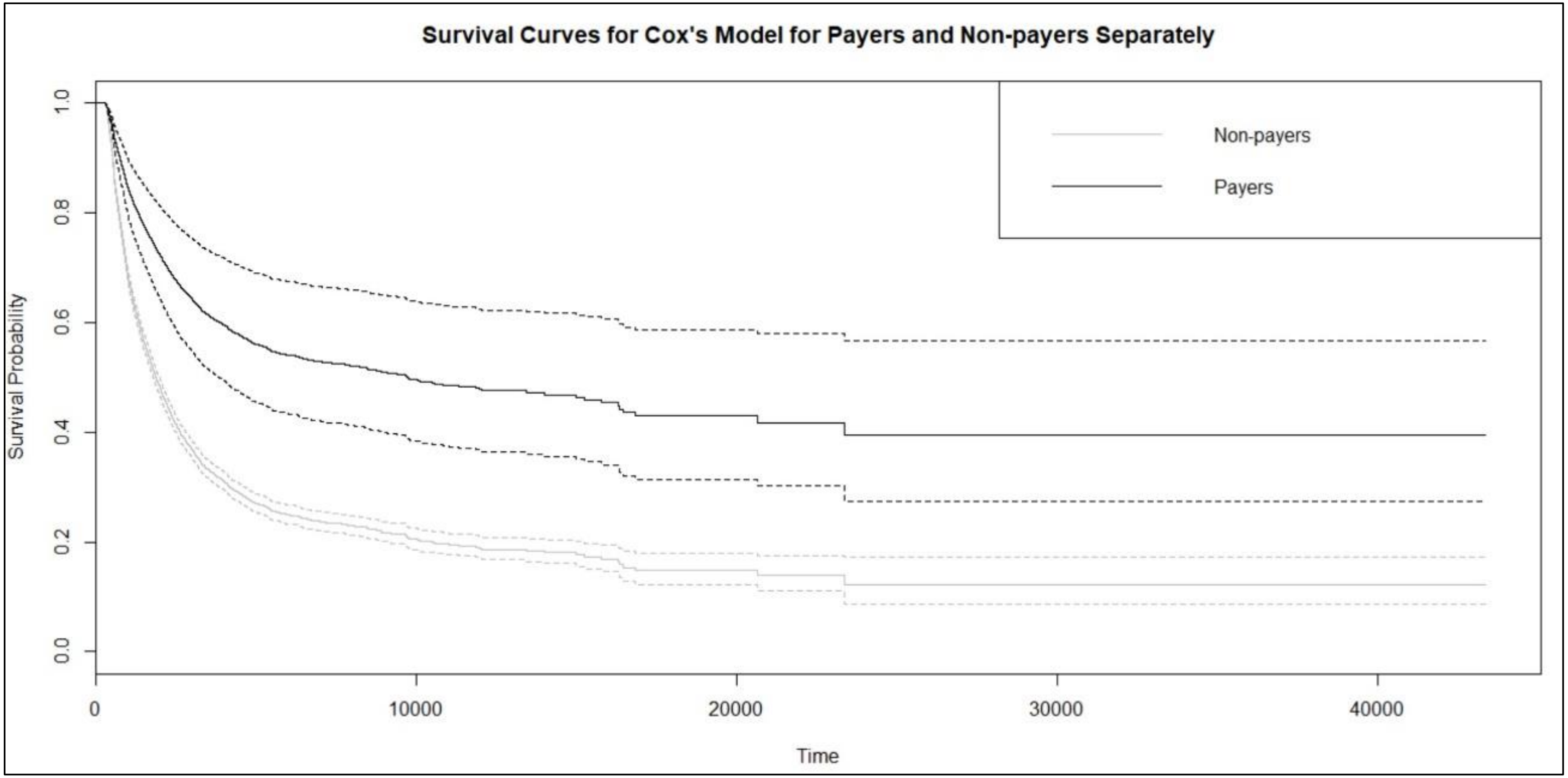


Figure 6.4: Survival curves for Cox’s proportional hazards model for players stratified by their paying status

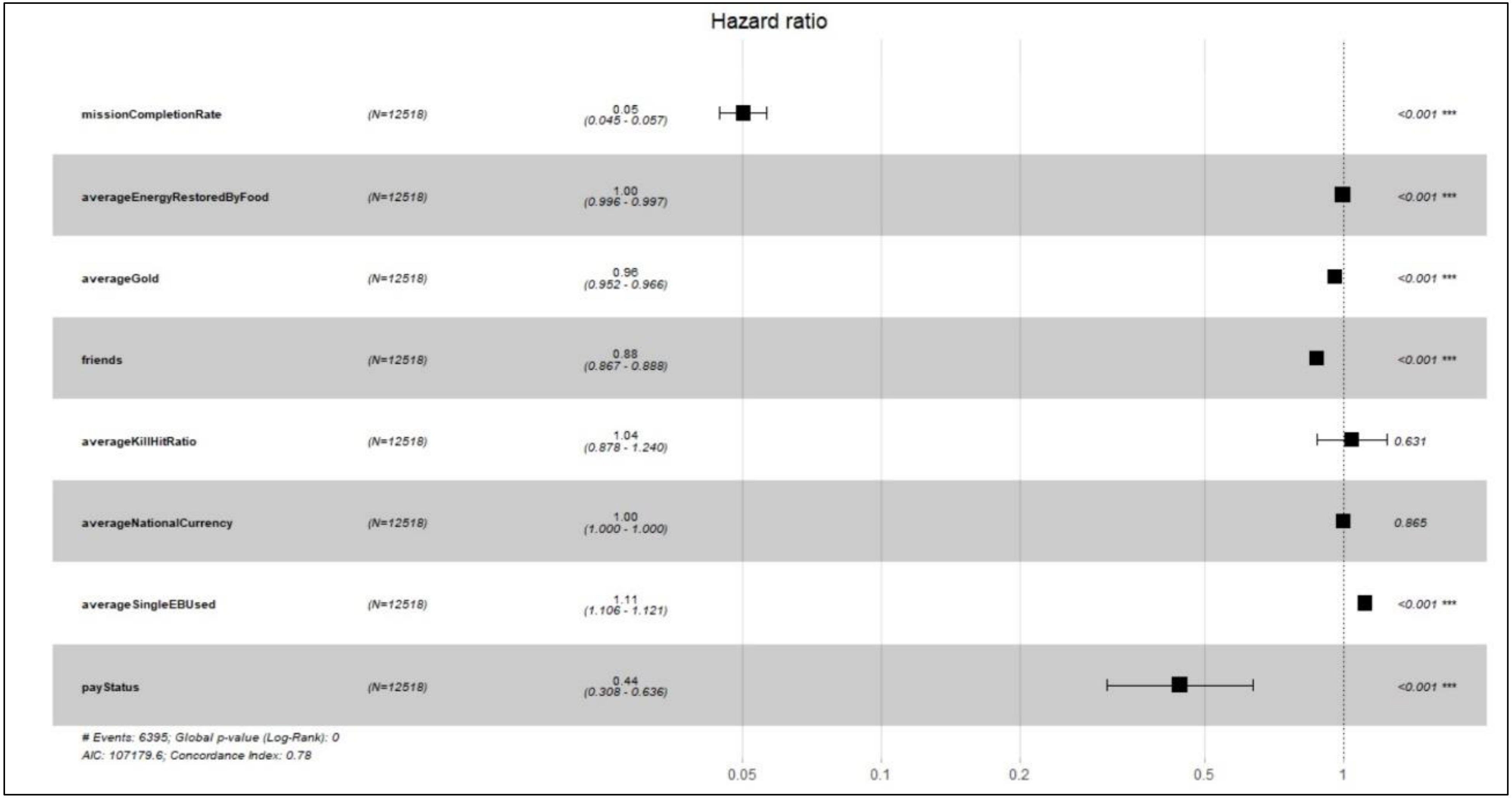


Figure 6.5: Forest plot visualisation of the Cox's proportional hazards model

The forest plot shows the hazard ratios derived from fitting the Cox's model for all explanatory variables that were included. A hazard ratio equal to 1 indicates no effect, whereas a hazard ratio <1 implies reduced risk, and a hazard ratio >1 means increased risk (Springate, 2014, Easy Guides, 2016). So for example, holding all other covariates constant, an additional unit of premium currency gold reduces the risk of defecting by a factor of 0.96, i.e. by 4%. Similarly, each increase in the use of energy bars increases the risk by a factor of 1.11., i.e. by 11%. The results and its implications are discussed in detail in a later section.

6.5.2.3 Model Evaluation

This section verifies the assumptions, goodness of fit and predictive ability of the Cox's proportional hazards model fitted earlier.

The fundamental postulation of Cox's proportional hazards model is that it makes no assumption regarding the shape of the underlying hazard function over time (Harrell Jr., 2015). This is confirmed from the plots in figure 6.6 that checks for systematic trends to verify if the hazards associated with different variables is constant over time.

The plots do not demonstrate any specific trends over time and remain fairly constant, albeit with a few outliers. Thus, the proportional hazards assumption of the Cox's model appears to be satisfied.

Statistics showing the overall goodness of fit of the model is displayed at the bottom of table 6.1. The omnibus null hypothesis that all the $\hat{\beta}$s are 0 are tested using three asymptotically equivalent tests, likelihood ratio, Wald and score (logrank) tests (Fox & Weisberg, 2018). The p-values for all these tests are statistically significant (p<0.01), thereby providing evidence to reject the null hypothesis and indicate global statistical significance of the fitted model. The coefficient of determination R^2 is a measure of the variance explained by the fitted model, and aids in understanding the ability of the explanatory variables to predict the time until defection of customers (Gillespie & Mccullough, 2006). The R^2 for this model is 0.35, indicating low moderate robustness of the fitted model. However, this measure is dependent on the censorship distribution of the data and can be quite low in some cases (Müller, 2004). A better measure of usefulness of the model is the more popular Harrell's c-index of concordance (Harrell Jr, Califf, Pryor, Lee & Rosati, 1982), which is a global index for validating the predictive ability of a survival model, and defined as the fraction of pairs in the data for which observations with shorter survival times have larger risk scores predicted by the model. The c-index of concordance for this model is 0.77, indicating good predictive and discriminatory power of the fitted model.

The predictive accuracy of the fitted model is validated using the holdout sample technique, applying the same procedure as that elucidated in the previous chapter (section 5.4.3.2). Survival probabilities are predicted using the test sample from the model that was fitted earlier using the training sample. A confusion matrix is then obtained, that contains information regarding the true and predicted classifications of customer defection by the model are then obtained. Outcome statistics for the confusion matrix calculated from the test sample are illustrated in table 6.2.

Table 6.2: Outcome statistics for the confusion matrix based on test data

Statistic	**Value**
Precision	0.520
Recall	0.762
Accuracy	0.523
Balanced Accuracy	0.520

The definitions for all these measures have been detailed in section 5.4.3.2 and are hence not reiterated here. Precision denotes that out of all the predictions for user defection, 52% cases have actually defected. Recall signifies that out of all the true defected cases, 76% are predicted as defected. This implies that there is a chance that 24% of cases at a risk of defection may not be correctly predicted as that, in which case preventive measures may fail to be implemented for these. The overall accuracy of the model, denoting how frequently it produces correct classifications is moderate at 0.52. The balanced accuracy is also moderate at 0.52.

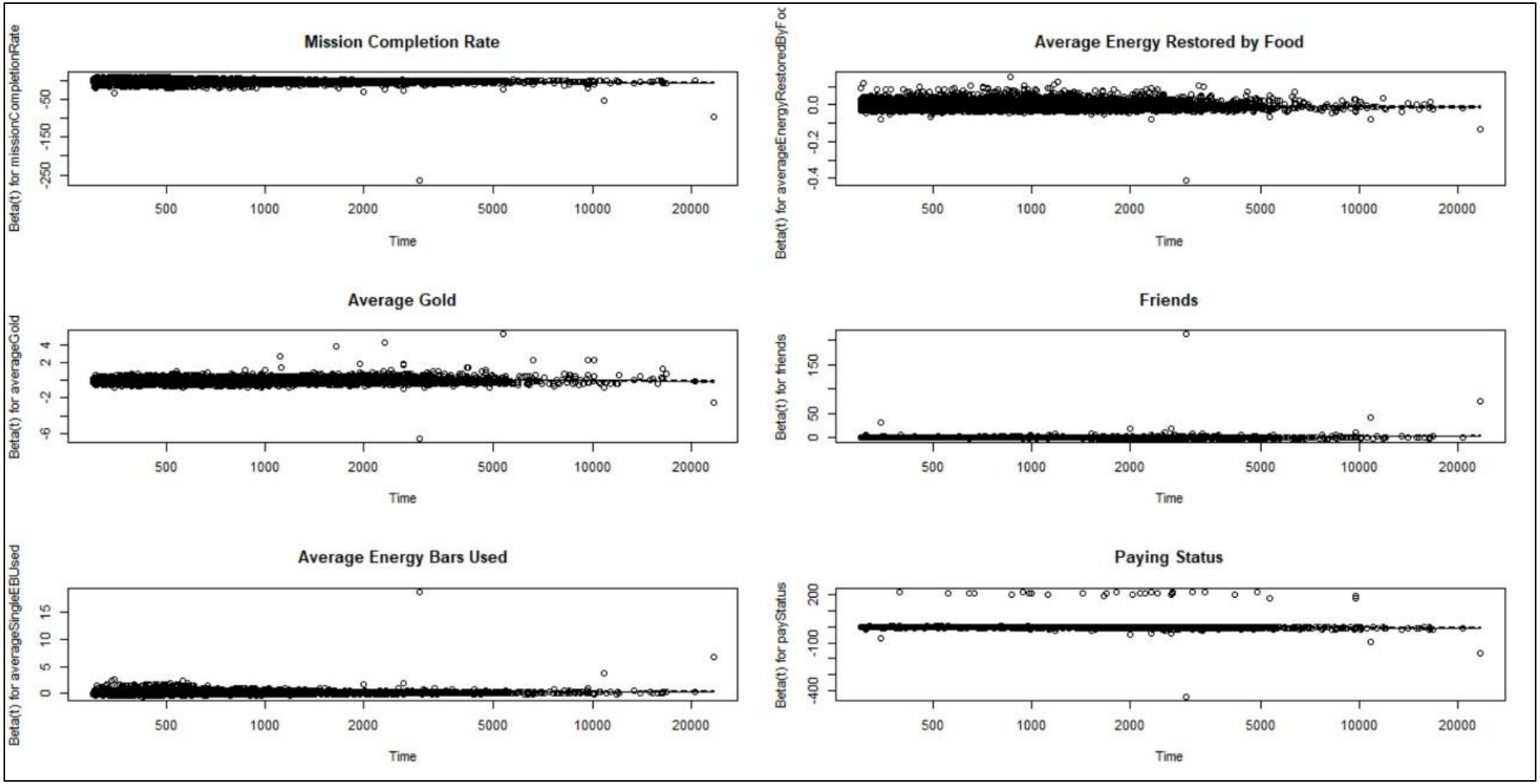


Figure 6.6: Plot verifying the proportional hazards assumption of Cox's model

6.6 Results and Interpretations

The survival curves using Kaplan-Meier estimates and estimates from the fitted Cox's proportional hazards model are comparable to some extent, showing the decreasing probabilities for survival of eRepublik users with the passage of time. However, the distribution corresponding to the Cox's model is more conservative as it takes into account the influence of multiple covariates on the survival probability. Both methods reveal that survival probabilities for individuals that make monetary purchases in the game is significantly different (better) than those that do not.

The final fitted Cox's proportional hazards model results in the variables mission completion rate of users, their energy that was restored by virtual products such as food, premium currency gold owned by players, virtual friends they have, virtual items like energy bars used and their paying status as being statistically significant in affecting the hazard rate or conversely survival time of customers. Increase in the variables mission completion rate, premium currency gold and friends are associated with a reduced risk of defection, while an increase in energy bars used is associated with an increased risk of defection. Also, a paying status of 1 (i.e. payer) reduces the risk of defecting by a factor of 0.44, i.e. by 56%. Finally, although a significant variable, the hazard ratio for energy restored by food is approximately 1, thereby indicating no effect. Therefore it reveals that, higher a player's competency in tasks, more his possession of premium currency, greater his tendency to make friends and more his propensity to make real money purchases in the game, less are the risk of him dropping out of the game at any point in time. However, more a player utilises a virtual item such as energy bar, higher is the risk of him defecting from the game at any point in time. The replenishment of energy by consumption of food is not seen to have an effect on the risk of churning from the game.

The survival analysis approach undertaken here can be adopted by game studios to provide insight on the user gameplay variables that are likely to have an impact on the risk of defection of customers. Moreover, the model developed can be used to predict the probability of survival, and consequently whether a customer is likely to drop out of the game or not, given that they have survived up to a specific time point of interest. Thus, the approach can be used to determine how likely a particular set of customers are to defect from the game, say after 10 days, or a month and so on. This kind of information may be exceptionally valuable for game developers to understand when their customer is at the highest risk of dropping out. This can in turn allow them to target these users with appropriate measures to enhance their gameplay, providing a more enjoyable experience that will inspire them to continue in the game, which in turn will be useful for the game's productivity, popularity and revenue generation.

7. Modelling Monetisation Behaviours of Customers

Online games that adopt the free-to-play (F2P) model can be played without having to purchase the game (Mueller-Veerse et al., 2011, Evans, 2016,). A substantial part of the game is open to players to access freely, while players are urged to pay for high-quality virtual products during game play (Solidoro, 2009, Luban, 2011). The freemium model is the most prominent of the F2P models (Mueller-Veerse et al., 2011). Freemium games are available free of cost, while players are encouraged to make micro payments to access advanced features, advantages and premium virtual goods (Alha et al., 2014).

7.1 Types of Virtual In-game Currencies

In a freemium model, players can set up an account free of cost and play whenever they want to. However, in order to acquire additional benefits that will help easy progress in the game, they are required to spend in-game currencies to purchase virtual products. There are two types of virtual in-game currencies available – grind currency and premium currency (Saldana, 2014). Grind currency is difficult to procure, and can be acquired during the course of gameplay upon performing certain repetitive tasks i.e. 'grinding' throughout the game, such as acquiring collectibles, training the player's virtual avatar, defeating opponents etc. (Saldana, 2014) . Premium currency on the other hand is instantly obtainable through micropayments using real money, and sometimes awarded in very small amounts to players when they complete special tasks (GameSparks, "n.d."). Grind currency is free to obtain but highly time consuming whereas premium currency is easily attainable but costs real money.

7.2 The Monetisation Design

The process of grinding through the game to gather grind currency is tedious and takes up a great deal of time. The required actions have to be performed repeatedly and several times before a reasonable amount of currency is accumulated to be used in the game. Therefore, players who are impatient or eager to get to a certain stage can avoid grinding and simply purchase the premium currency to rapidly advance forward in the game. This is the fundamental business approach of freemium games.

In the freemium model, the focus is on maintaining the average revenue per user (ARPU) and lifetime value (LTV) of the player instead of the one-off retail price from box sales (Hindy, 2017). Miscellaneous channels are available for making micropayments within the game and these have been the driving force behind this business model. As a result, it is imperative for game designers and publishers to understand the incentives for micro transactions by players in order to develop successful games that will monetise their users.

The aim of this chapter is to understand the factors that drive monetisation behaviour in customers of online freemium games. This is done by fitting an appropriate model to a response variable that stores payment information about players. The statistically significant explanatory variables in the model provide information about the behavioural characteristics that motivate players to make real money transactions in the game. This knowledge is then used to develop strategies that may inspire users to invest in micropayments, thereby making successful games that are appealing to customers and increasing their prospect of revenue generation.

7.3 Monetisation Structure in eRepublik

eRepublik is an online multiplayer game that follows a freemium business model and its monetisation structure. The following description of the monetisation system in eRepublik is a result of knowledge gained from actually playing the game, and from eRepublik Official Wiki, (2018).

The grind currency in the game is called ‘National Currency’, which can be acquired during the course of gameplay. The most common way of obtaining national currency is as a form of salary by working in any country. It can also be attained as gifts from other players who possess that currency, and as rewards on completion of certain missions. National currency is used to buy non-premium items in the game such as food, weapons and raw materials. It can be offered as salary by players to others that work for their company. The premium currency in eRepublik is called ‘Gold’, which can be acquired through purchase using real money (€). Some other ways to procure it are as bonus upon accomplishing in-game achievements, levelling up, inviting other users to the game, and if an invited user purchases gold. Gold is used to access more advanced and superior features such as creating and buying companies, upgrading companies, creating political parties, creating a military unit, buying training boosters, energy bars etc. Moreover, both grind and premium currencies can be exchanged for each other i.e. gold can be bought with national currency and vice versa.

The predominant game elements in eRepublik are working, training and fighting, the performance of all of which consume energy. Thus, a player's energy can be assumed to be the main characteristic that allows them to move forward in the game. There are multiple ways to gain energy such as – limited units of energy can be recovered every 6 minutes by eating food (which costs national currency); players can instantly recover energy by purchasing and using energy bars and first aid kits (both of which costs gold); and energy can also be accumulated through progression to a new level each time. Hence, grind currency (national currency) is useful only when a player is willing to wait to refill a limited and small amount of his in-game resource (energy). In order to replenish the resource promptly without having to wait, the player has to spend premium currency (gold). Similarly, only some basic companies require raw materials purchased with national currency to be set up. Building a new company requires massive amounts of national currency, thereby requiring a very long waiting time if users wish to grind through, while some such as factories or large storages are only available for gold. Moreover, training grounds and advanced buildings can only be procured in exchange for gold.

Overall, the monetisation mechanism used in eRepublik is that all basic items or limited quantities of items are available for national currency. Any resource that is of high quality or prime importance in the game, or can be instantaneously replenished, is only available for gold (or in some cases very high quantities of national currency that will obviously take time to collect). Therefore it seems that, for quick progression, immediate success, and an edge over opponents, it is imperative for users to invest in premium currency gold, which in turn is purchased through real money (€) micro transactions

7.4 Preliminary Analysis of Micro Transactions

Following a description of the revenue structure in eRepublik, some exploratory analysis to quantitatively determine the monetisation of customers in the game is implemented. As usual, new and active users from 11th November 2013 to 29th December 2013 and their gameplay up to 6th January 2014 is considered, resulting in 40716 individuals. Again, the 30% of the sample that interact with the game for a single session only and do not connect with it right from the onset, and hence fail to exhibit adequate consumer behaviours to study, are excluded from the analysis.

Transaction events triggered by users are first extracted from the history of in-game events and scrutinised. Three types of transactions are prevalent within the game – that involving premium currency gold, constituting about 86% of all transaction events; that involving real currency €, constituting about 12% of all transaction events; and that involving loyalty points, constituting about 2% of all transaction events. Loyalty points are only available to special customers that are part of the loyalty program of eRepublik, which requires a purchase of at least €50 worth of gold, and includes several benefits such as not requiring to train or work, exclusive gold bonuses, committed customer support and availability of more special offers (eRepublik Official Wiki, 2015). Virtual products such as energy bars, energy centres and storages are bought with loyalty points. All other virtual items are bought with gold, the most popular being energy bars, followed by bazooka parts (scope, ammo, barrel etc.), and permits to start newspapers and political parties. Since the focus of this analysis is to study user behaviours related to real currency micro transactions within the game, the 12% transaction events that involve € are of prime importance here.

The individuals that invest in at least one transaction (in €) are examined and it is found that only approximately 1.3% of the sample make micro purchases within the game. These users are called payers and the objective is to understand what aspects of gameplay induces them to make payments. For this purpose, a group of individuals that have experienced the game to roughly the same extent are studied, and the factors that significantly contribute to only a subset of them indulging in micro transactions are determined. These are conceptualised to be the greatest motivators of payments as otherwise the group of users have had a similar exposure to and experience with the game. In order to define this set of users that are of interest to this analysis, the total gameplay time of payers are investigated, and it is found that all payers have been engaged with the game for at least 75 minutes before making a purchase. Therefore, the final group of players that are considered for this analysis are all those who have a total gameplay time of at least 75 minutes, thereby allowing for them to have a similar and comparable involvement with the gaming platform. This results in 20854 individuals, including 359 (~1.7%) payers with a total of 788 micro transactions amounting to €6802.30 in all, during the analysis period.

The distributions of the number of payments and amount spent (€) are illustrated in figures 7.1 and 7.2 respectively.

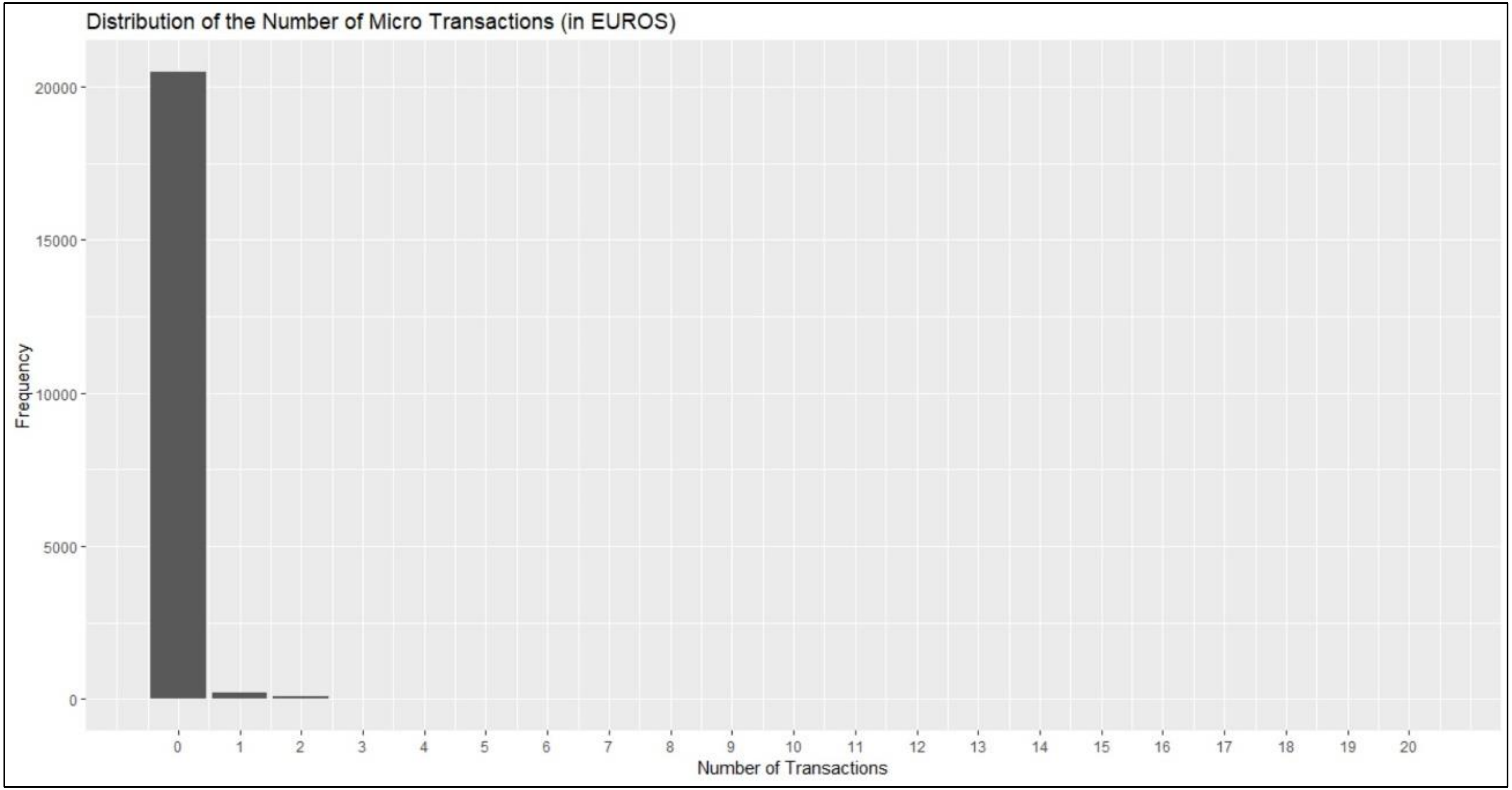


Figure 7.1: Distribution of the number of micro transactions (in €) made by players

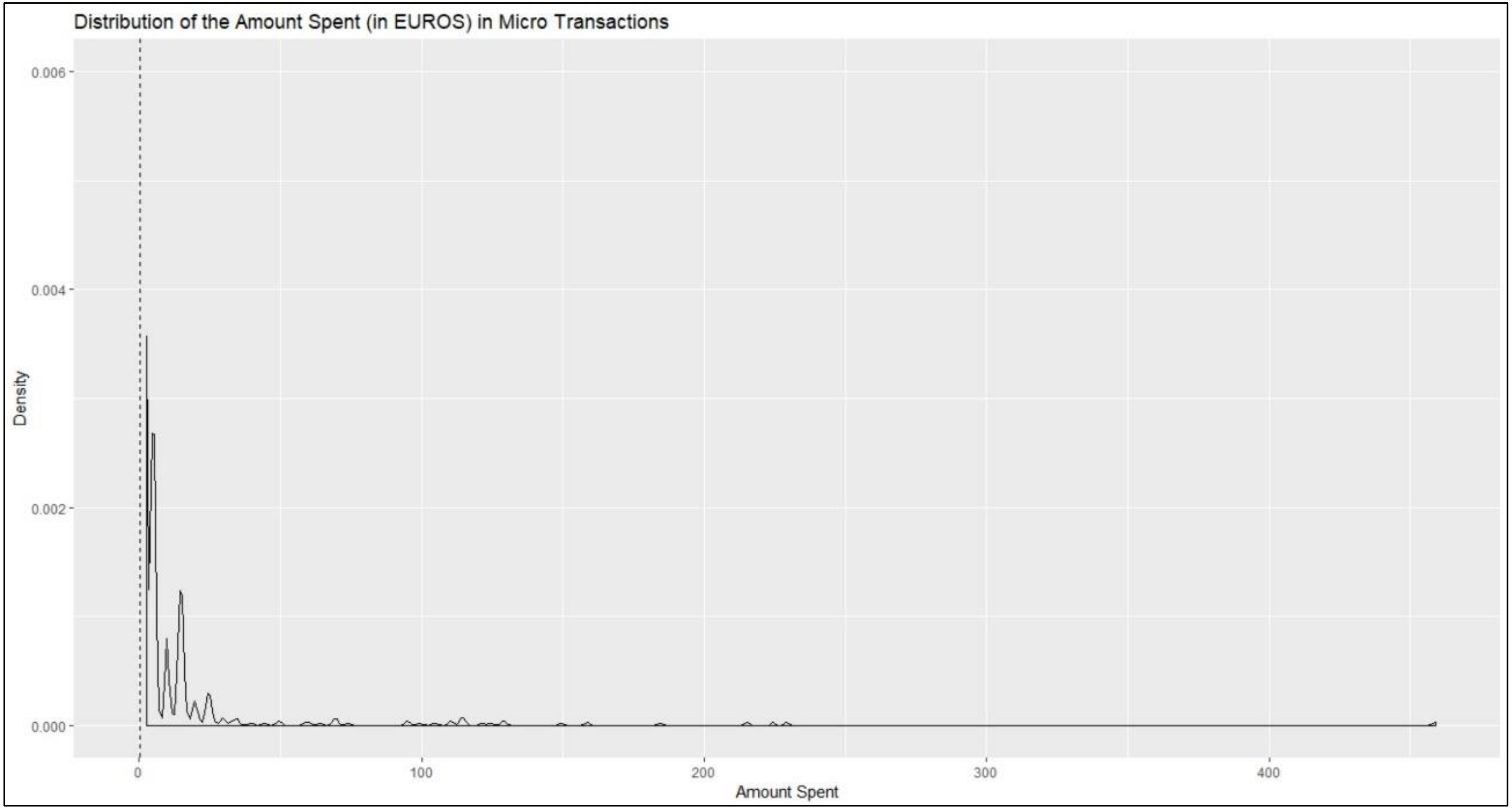


Figure 7.2: Distribution of the amount spent (in €) in micro transactions by players

Figure 7.1 is a bar plot of the number of micro payments made, and figure 7.2 shows a kernel density curve of the amount spent (in €), which is an effective non-parametric approach of visualising a distribution fundamental to a continuous variable (Cai, 2013). Both distributions appear to be extremely skewed with long right tails. The dashed line in figure 7.2 represents the average amount spent, which is found to be €0.33. The number of payments is a discrete variable ranging from 0 to 20, whereas the amount spent (€) is a continuous variable ranging from €0 to €459. An overwhelming majority (~98.3%) do not invest in a micro transaction at all. The average number of payments and average amount spent are 0 and €0.33 respectively, while all three quartiles of both the variables are 0. This indicates that the distributions are heavily inflated with zeroes.

Further examination focuses on only the payers and the distributions of their number of payments and amount spent, as displayed in figures 7.3 and 7.4.

Figure 7.3 is a bar plot of the number of micro payments made by payers, and figure 7.4 shows a kernel density curve of the amount spent (in €) by payers. Although the massive chunk of non-paying users are removed in this case, the distributions still appear to be skewed with long right tails. More than half the paying population (~54%), makes only one monetary transaction throughout their course of gameplay, approximately 40% invest in 2-5 payments, and a minority (~6%) make 6-20 payments. About 92% of payers invest a total of less than €50 during their time in the game. The average number of payments and average amount spent are 2 and €18.95 (represented by the dashed line in figure 7.4) respectively, but the median number of payments and median amount spent are 1 and €4.90.

Overall, it can be concluded that that the number of customers that make micro payments in real currency (€) is minimal. The distributions for both number of payments and amount spent are heavily right skewed and inflated with zeroes, suggesting that more number of payments and high amount spent are much less likely than small number of payments and less amount spent. This clearly implies that users are far less inclined to make spend money in the game, and even those that do, indulge in small number of micro transactions that are worth less amounts of money.

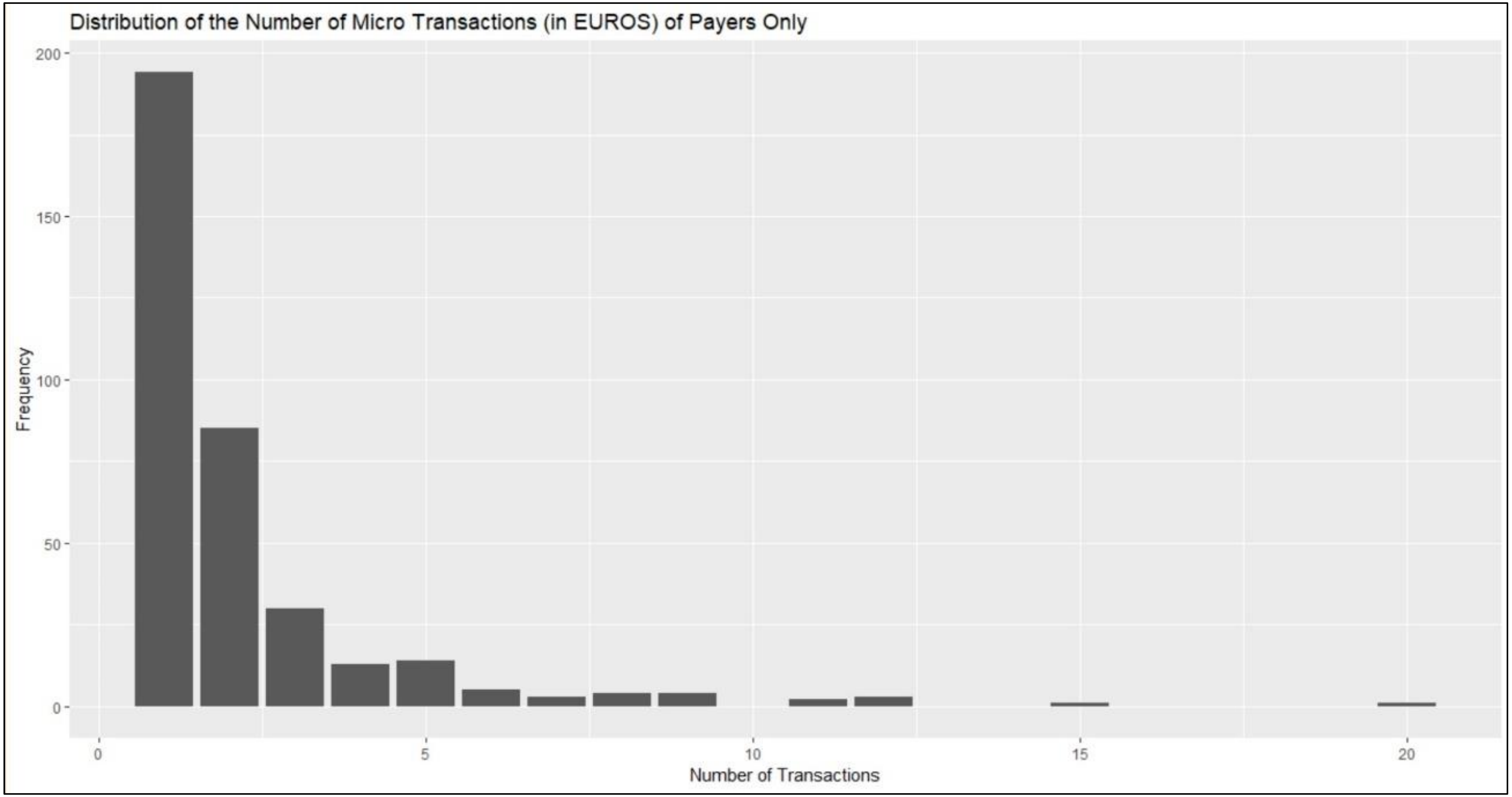


Figure 7.3: Distribution of the number of micro transactions (in €) made by payers

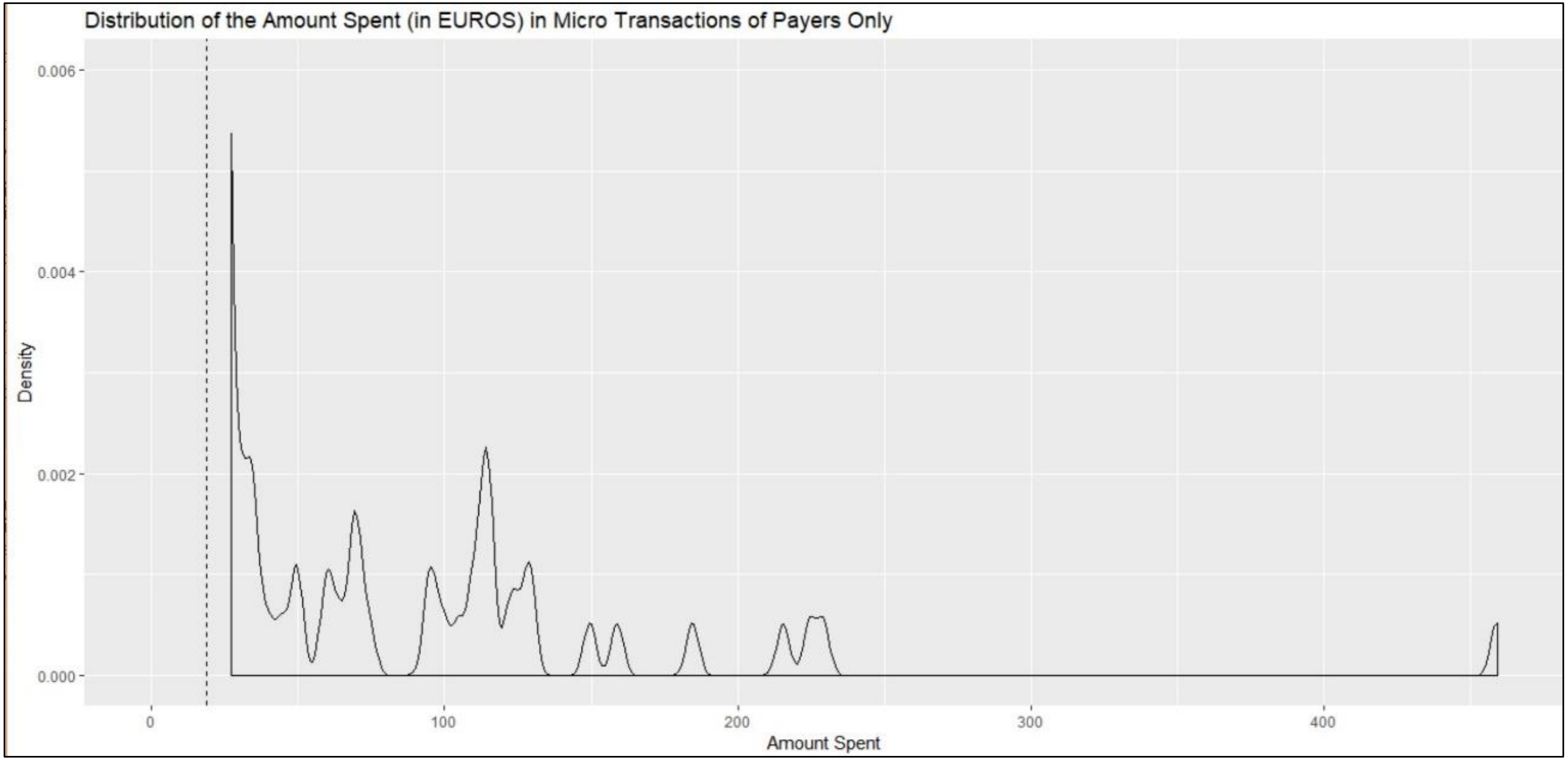


Figure 7.4: Distribution of the amount spent (in €) in micro transactions by payers

7.5 Data Representing Micro Transactions to be Modelled

Investigation of the distributions of the number and worth (in €) of micro payments by customers makes it evident that not only is it imperative to understand what drives users to make their first transaction, but also what motivates them to repeatedly invest money in the game. This is because of the two striking aspects of monetisation revealed earlier – players mostly tend to remain in the game, grinding their way through, with only about 1.7% making a payment; and of those that do, more than half are only one-off payers with about 6% that make more than 5 payments. Thus, this study attempts to understand the gameplay elements that promotes an increase in the number of payments made by users, starting from their very first transaction to multiple ones after that. This will provide useful insight into the determinants of market performance of the game and its likelihood of generating big revenue. Accordingly, the outcome variable of interest here is number of micro payments by users, to which an appropriate model is to be fitted and the significant variables examined.

A statistical model of the outcome variable will be able to elucidate the association between number of payments and a set of other variables called explanatory variables. It is stochastic, i.e. based on probability functions, as stated by Hilbe, (2014) that all parametric statistical models are determined by a latent probability distribution. The fitting of the model will entail estimation of the unknown parameters of the underlying probability distribution that are regarded as the best representation of the data that is being modelled (Hilbe, 2014). The covariates whose effects on the number of micro payments are to be determined, depict the energy restored by food, amount of premium currency gold possessed, kill:hit in virtual fights, amount of grind currency national currency possessed, energy bars used, and completion rate in missions.. The selection of these variables are based on judgement from playing the game and observing game dynamics.

In a similar manner as before, the data is first partitioned into training and test samples based on a 75:25 ratio. The training set is used to build the payments model and learn about the relationship between the number of payments and explanatory variables. The final model developed is validated and its performance estimated using the test set. However, due to the extremely small number of payers in the initial data set, it is found that the test sample, which is based on 25% of the data, consist of only 93 payers. It is unlikely to be able to draw any robust conclusions from the transactional behaviour of only 93 individuals regarding what influences players to make micro payments.

Therefore, this approach is not adopted for this part of the analysis, and the entire data set is used in the model building process.

7.6 Model Building and Evaluation

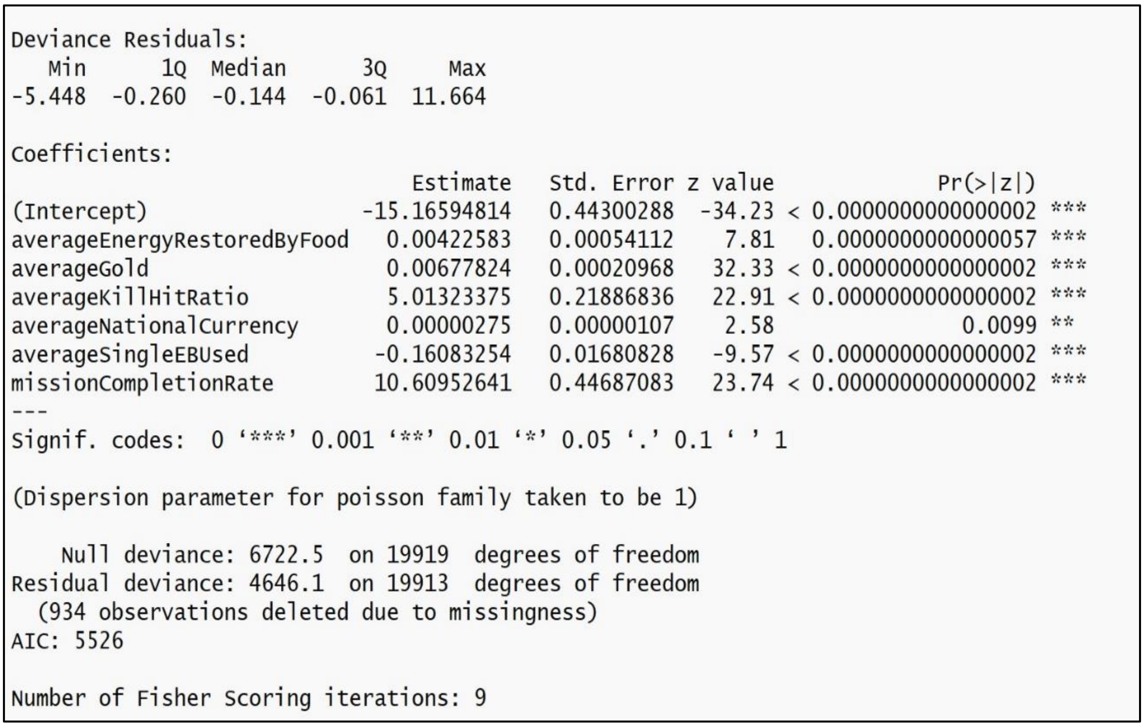
The modelling procedure is started with fitting a standard Poisson model to the response variable number of micro payments, representing a discrete count distribution, since Poisson distribution is the benchmark parametric model for count data (Cameron & Trivedi, 2013). The Poisson model may not be the best fit for this data since the variance of the distribution of payments (0.17) is not equal to its mean (0.04), which violates the equidispersion property of the Poisson (Hilbe, 2014). However, Cameron & Trivedi (2013) and Hilbe (2014) state that when modelling real-life data, the equidispersion criterion of Poisson is often violated, leading to underdispersed or overdispersed models. The expected number of counts being equal to 0 in a Poisson distribution with mean 0.04 is given by, $e^{-0.04}\ \frac{0.04^{\,0}}{0!}$ = 0.96. Thus, it is expected that about 96% of the counts in the model will be 0, which does not vary much from the observed percentage of payments equal to 0 in the data (corresponding to non-payers) being approximately 98%. Therefore, a Poisson model is fitted to the outcome variable to verify how well or not it explains the data.

7.6.1 Poisson Regression Model

A Poisson regression model is fitted to the outcome variable incorporating covariates that have been discussed before, and the results reported in table 7.1.

Table 7.1 displays results from fitting a Poisson model to the response variable number of micro payments, using a set of explanatory variables representing energy restored by food, amount of premium currency gold possessed, kill:hit in virtual fights, amount of grind currency national currency possessed, energy bars used and completion rate in missions. Parameter estimates and their standard errors, Wald (z) statistics testing the null hypothesis that the corresponding coefficients are 0, (Fox & Weisberg, 2018) and their associated p-values indicating statistical significance of the covariates are reported. All independent variables are found to be statistically significant at the 1% level of significance with p<<0.01. The standard errors of the estimates vary from moderately small to quite small, indicating reasonably good precision. However, this may be because of the large data set being modelled, or due to an underestimation of the standard errors resulting from poor model fit.

Table 7.1: Summary table of results from fitting a Poisson model



```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-5.448  -0.260  -0.144  -0.061  11.664

Coefficients:
                                Estimate   Std. Error z value             Pr(>|z|)
(Intercept)                 -15.16594814   0.44300288  -34.23 < 0.0000000000000002 ***
averageEnergyRestoredByFood   0.00422583   0.00054112    7.81   0.0000000000000057 ***
averageGold                   0.00677824   0.00020968   32.33 < 0.0000000000000002 ***
averageKillHitRatio           5.01323375   0.21886836   22.91 < 0.0000000000000002 ***
averageNationalCurrency       0.00000275   0.00000107    2.58               0.0099 **
averageSingleEBUsed          -0.16083254   0.01680828   -9.57 < 0.0000000000000002 ***
missionCompletionRate        10.60952641   0.44687083   23.74 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6722.5  on 19919  degrees of freedom
Residual deviance: 4646.1  on 19913  degrees of freedom
  (934 observations deleted due to missingness)
AIC: 5526

Number of Fisher Scoring iterations: 9
```

The Pearson based dispersion statistic is calculated as 2.79. The ideal value of the Pearson based dispersion statistic for a Poisson model is 1, and any variation from this indicates that the model is extradispersed (Hilbe, 2014). Therefore, the value of 2.79 suggests presence of extradispersion, more specifically overdispersion since it is greater than 1 (Hilbe, 2014). The observed variance of the distribution 0.17 is greater than the expected variance 0.04, which further exhibits a case for overdispersion (Hilbe, 2014).

A test of goodness of fit for a Poisson model is the chi-squared test, in which the residual deviance is compared to a χ^2 distribution with appropriate degrees of freedom; although as discussed by Boyle, Flowerdew & Williams (1997), this test is likely to be unreliable when the data is sparse with several cases having 0 counts. The residual deviance for the model is 4646.1, which is much smaller than both the residual degrees of freedom 19913 as well as the lower critical χ^2 value 19302. In a typical scenario, underdispersion would be considered a valid inference from this, but it contradicts the reasoning from the dispersion statistic, and observed and expected variances. Moreover, there is indeed a sparse distribution of data on number of payments as it is heavily right skewed with about 98.3% of cases recording a value of 0 payments. Hence, the plenitude of zeroes overrules the appropriateness of the $\chi 2$ goodness of fit statistic, including the possibility of underdispersion (Boyle et al., 1997).

Therefore, the fitted model appears to be overdispersed, suggesting that Poisson distribution is a poor fit for this data and as such, the parameter estimates and their statistical significance hold no value.

7.6.2 Negative Binomial Model

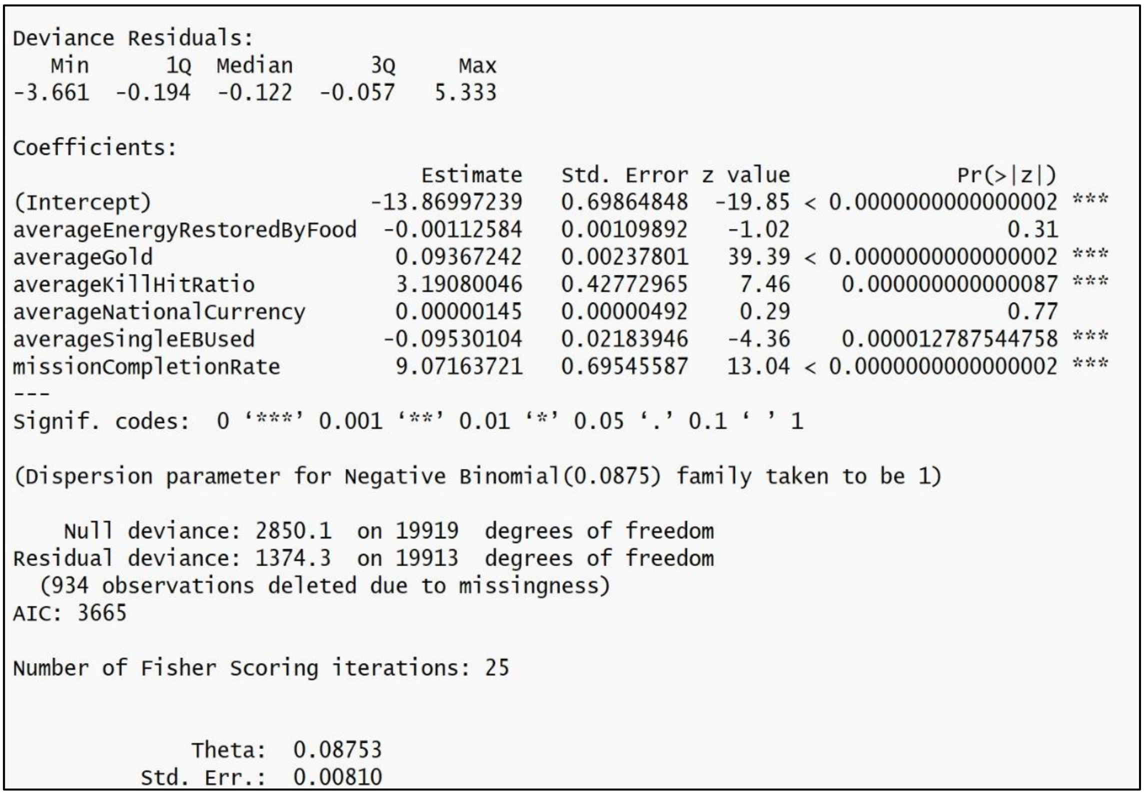
The most prominent method of handling overdispersed count data is to model it using a negative binomial model, which is a two-parameter model, comprising parameters mean and dispersion (Cameron & Trivedi, 2013). The dispersion parameter allows the model to adjust for the excess variability or heterogeneity in the data that the Poisson model fails to, and the mean of a negative binomial has the same interpretation as a Poisson mean, although the variance has a much wider extent than that accepted by the Poisson variance (Hilbe, 2014).

The data is fitted with a negative binomial model, consisting of the same set of covariates used before, and the results are stated in table 7.2.

Table 7.2 displays results from fitting a negative binomial model to the response variable number of micro payments, using a set of explanatory variables representing energy restored by food, amount of premium currency gold possessed, kill:hit in virtual fights, amount of grind currency national currency possessed, energy bars used and completion rate in missions. Parameter estimates and their standard errors, Wald (z) statistics and their associated p-values indicating statistical significance of the covariates are reported. All independent variables, except energy restored by food and quantity of national currency owned by players, are found to be statistically significant at the 1% level of significance with $p<<0.01$. The standard errors of the estimates vary from moderately small to quite small, indicating reasonably good precision. As discussed before, this could be due to the large amount of data being modelled, or an underestimation of the standard errors resulting from poor model fit.

The Pearson based dispersion statistic is calculated as 1.54, better than that obtained from the fitted Poisson model (i.e. 2.79). Although the dispersion shows decrease, it continues to be more than 1, thereby indicating that the data still has more variability than what the fitted model accounts for i.e. presence of overdispersion. The observed variance of the distribution 0.17 also remains greater than the expected variance 0.06, further confirming overdispersion.

Table 7.2: Summary table of results from fitting a negative binomial model



```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.661  -0.194  -0.122  -0.057   5.333

Coefficients:
                              Estimate   Std. Error z value            Pr(>|z|)
(Intercept)               -13.86997239   0.69864848  -19.85 < 0.0000000000000002 ***
averageEnergyRestoredByFood  -0.00112584   0.00109892   -1.02                 0.31
averageGold                 0.09367242   0.00237801   39.39 < 0.0000000000000002 ***
averageKillHitRatio         3.19080046   0.42772965    7.46   0.000000000000087 ***
averageNationalCurrency     0.00000145   0.00000492    0.29                 0.77
averageSingleEBUsed        -0.09530104   0.02183946   -4.36   0.000012787544758 ***
missionCompletionRate       9.07163721   0.69545587   13.04 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.0875) family taken to be 1)

    Null deviance: 2850.1  on 19919  degrees of freedom
Residual deviance: 1374.3  on 19913  degrees of freedom
  (934 observations deleted due to missingness)
AIC: 3665

Number of Fisher Scoring iterations: 25


              Theta:  0.08753
          Std. Err.:  0.00810
```

A common approach for model selection between non-nested models is to compare information criteria, Akaike information criterion (AIC) values and Bayesian information criterion (BIC) values for the different models (Kharrat & Boshnakov, 2017), wherein smaller values are preferred (Hilbe, 2014). The AIC and BIC values corresponding to the fitted negative binomial are 3664.7 and 3720 respectively, while those from the fitted Poisson are 5525.5 and 5580.8, implying that the negatively binomial fit is superior than the Poisson fit for this data.

However, overdispersion continues to be a persistent issue in the modelling process, and the negative binomial model does not appear to be the most suitable in this scenario. Thus, the parameter estimates and their statistical significance obtained from the model fit are likely to be meaningless. Moreover, there are excessive zero counts in the data that have not been adjusted for, which may also be a potential cause for overdispersion. This prompts the need for alternative count models that can deal with excess zeroes.

7.6.3 Model for Excess Zeroes

As established in the previous sections, the response variable number of micro payments contains zeroes that are excess in number to that expected in standard count models like the Poisson and negative binomial. Both standard models, when fitted to the data, display overdispersion, which may be a likely result of the large number of zeroes. Hence, they are not robust leading to inaccurate inferences.

This instigates the need to implement certain approaches to modelling count distributions that include more zeroes than are expected in the standard Poisson or negative binomial distributions. Two kinds of models are generally employed to deal with overdispersion and excess zeroes, hurdle models and zero-inflated models (Hilbe, 2014). These are suitable in cases where the attributes that explain a zero value are conceived to be different from those that explain variation further along the count distribution. The transactional behaviour of players could be considered a similar scenario, in which the aspects of gameplay that motivate the transition of non-payers to payers may be different to the factors that influence repeat payments by those that are already payers.

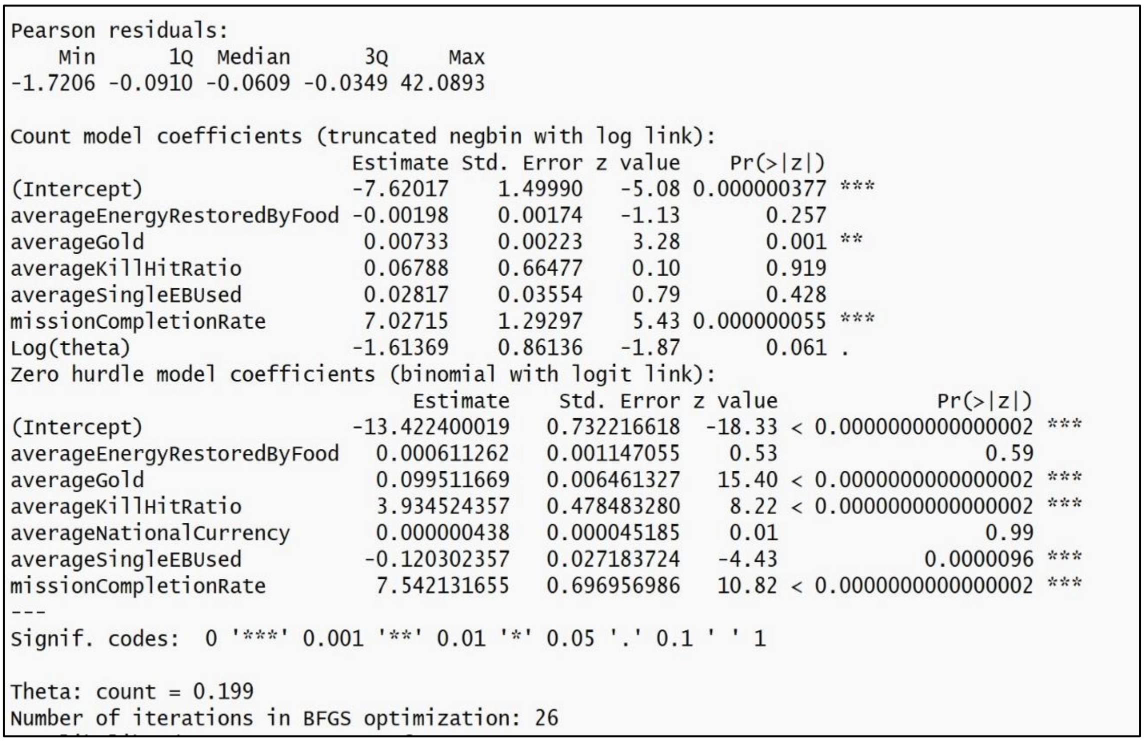
7.6.3.1 Hurdle Models

Hilbe (2014) explains that the fundamental notion of a hurdle model is to partition the model into two parts - a binary process to produce positive counts versus zero counts, which is typically modelled using a binary model like logit or probit; and a process producing only positive counts., which is modelled with a zero-truncated model. Therefore, hurdle models unequivocally incorporate two different data generation processes, as they treat the process of moving from zero to one in a manner dissimilar to that from moving from one to greater than one.

Construction of an appropriate negative binomial-logit hurdle model for the outcome number of micro payments is demonstrated below. The binary model used is a standard logit model, whereas the zero-truncated model used is the zero-truncated negative binomial model. This is because it has already been established that the data exhibits overdispersion and there is an improvement in model fit for the negative binomial over the Poisson.

The initial model contains the full set of explanatory variables that were used in the previous two instances, and the results are reported in table 7.3. The fitted negative binomial-logit hurdle model consists of two parts. The binary process, called the zero hurdle model, results in coefficients (the second set of coefficients in table 7.3) that are associated with the probability of having a positive count instead of a zero. All the independent variables in this part, except energy restored by food and quantity of national currency possessed, are statistically significant with $p<<0.01$. The count part of the model, results in coefficients (the first set of coefficients in table 7.3) in which all cases for which the outcome is zero are disregarded, and only values equal to or greater than 1 are modelled as a truncated negative binomial model. In this part, the intercept, quantity of gold owned and completion rate in missions are statistically significant with $p<0.01$.

Table 7.3: Summary table of results from fitting an initial negative binomial-logit hurdle model

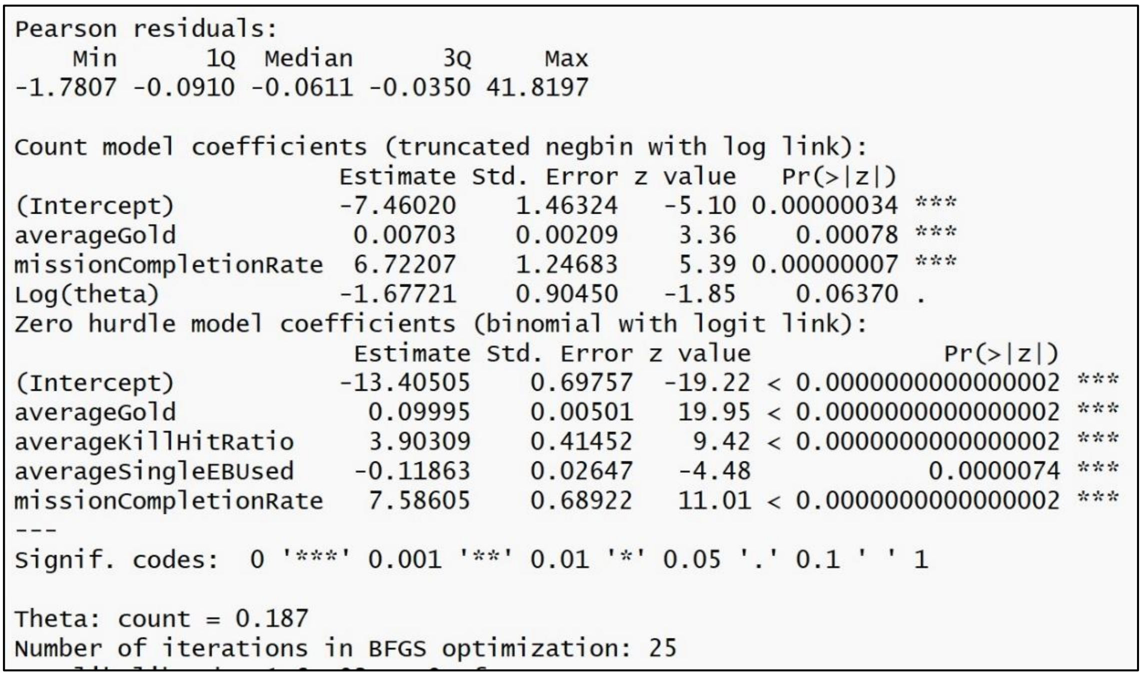


```
Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.7206 -0.0910 -0.0609 -0.0349 42.0893

Count model coefficients (truncated negbin with log link):
                           Estimate Std. Error z value    Pr(>|z|)
(Intercept)                -7.62017    1.49990   -5.08 0.000000377 ***
averageEnergyRestoredByFood -0.00198    0.00174   -1.13       0.257
averageGold                 0.00733    0.00223    3.28       0.001 **
averageKillHitRatio         0.06788    0.66477    0.10       0.919
averageSingleEBUsed         0.02817    0.03554    0.79       0.428
missionCompletionRate       7.02715    1.29297    5.43 0.000000055 ***
Log(theta)                 -1.61369    0.86136   -1.87       0.061 .
Zero hurdle model coefficients (binomial with logit link):
                                Estimate    Std. Error z value             Pr(>|z|)
(Intercept)                -13.422400019   0.732216618  -18.33 < 0.0000000000000002 ***
averageEnergyRestoredByFood   0.000611262   0.001147055    0.53                 0.59
averageGold                   0.099511669   0.006461327   15.40 < 0.0000000000000002 ***
averageKillHitRatio           3.934524357   0.478483280    8.22 < 0.0000000000000002 ***
averageNationalCurrency       0.000000438   0.000045185    0.01                 0.99
averageSingleEBUsed          -0.120302357   0.027183724   -4.43            0.0000096 ***
missionCompletionRate         7.542131655   0.696956986   10.82 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta: count = 0.199
Number of iterations in BFGS optimization: 26
```

A goodness of fit test for the hurdle model is similar to that stated before, whereby the AIC statistic is compared to that from alternative count models, and smaller values preferred. The AIC value corresponding to the fitted hurdle model is 3227.1, which is lower than the AIC of 3664.7 from the negative binomial model. Thus, the hurdle model demonstrates a better fit to the data than the negative binomial, which in turn is an improvement over the Poisson as already established.

The final hurdle model is now fitted after removal of the non-significant variables, and the results stated in table 7.4. All covariates in both the binary and the count parts of the fitted model show statistical significance. Since the final model is nested within the initial model, the prevalent likelihood ratio test for comparison of nested models (Zeileis, Kleiber & Jackman, 2008) is implemented, which results in p=0.83, indicating a lack of evidence in rejecting the null hypothesis that the larger model is not a significant improvement over the nested one. Therefore, the less complex negative binomial-logit hurdle model is the preferable one and exhibits better fit to the data.

Table 7.4: Summary table of results from fitting the final negative binomial-logit hurdle model



```
Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.7807 -0.0910 -0.0611 -0.0350 41.8197 

Count model coefficients (truncated negbin with log link):
                      Estimate Std. Error z value   Pr(>|z|)    
(Intercept)           -7.46020    1.46324   -5.10 0.00000034 ***
averageGold            0.00703    0.00209    3.36    0.00078 ***
missionCompletionRate  6.72207    1.24683    5.39 0.00000007 ***
Log(theta)            -1.67721    0.90450   -1.85    0.06370 .  
Zero hurdle model coefficients (binomial with logit link):
                       Estimate Std. Error z value            Pr(>|z|)    
(Intercept)           -13.40505    0.69757  -19.22 < 0.0000000000000002 ***
averageGold             0.09995    0.00501   19.95 < 0.0000000000000002 ***
averageKillHitRatio     3.90309    0.41452    9.42 < 0.0000000000000002 ***
averageSingleEBUsed    -0.11863    0.02647   -4.48            0.0000074 ***
missionCompletionRate   7.58605    0.68922   11.01 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Theta: count = 0.187
Number of iterations in BFGS optimization: 25 
```

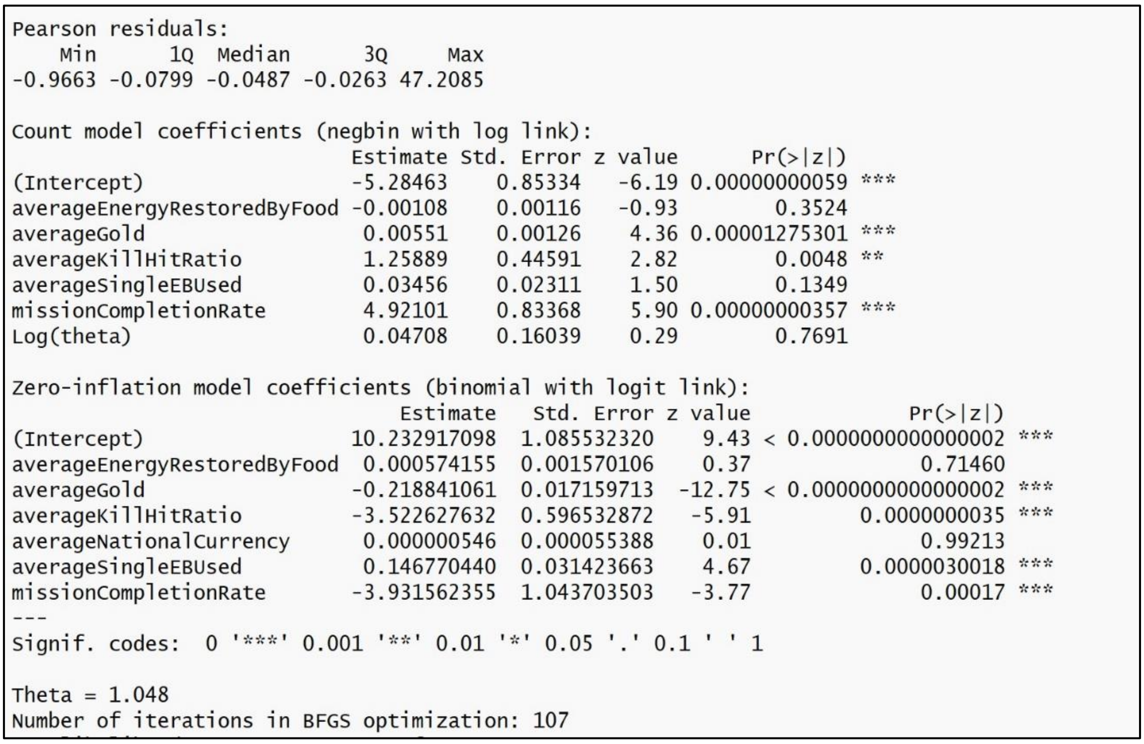
7.6.3.2 Zero-Inflated Models

An alternative to the hurdle model, when dealing with data having excess zeroes is the zero-inflated model. According to Hilbe (2014), zero-inflated models can be considered as “finite mixture models (i.e., where there are supposedly two data-generating mechanisms, one generating 0’s and one generating the full range of counts)” (p.196). Contrary to the hurdle model, which is a two-part model whose each constituent can be modelled separately, zero-inflated models do not explicitly divide the data into two parts. Some zeroes are characterised by structure, wherein they never meet the threshold to become a positive count, while others exist through choice, but could have been positive counts, thereby leading to the presence of zeroes that overlap and are estimated by both the binary and count mechanisms (Hilbe, 2014).

Construction of suitable zero-inflated negative binomial model for the response number of micro payments is demonstrated below. The binary component is modelled as logit whereas the count component is modelled as negative binomial. This is because the data being modelled is typically overdispersed, therefore making it more reasonable to use negative binomial rather than Poisson as the distributional basis for the zero-inflated model.

The initial model fitted contains the full set of covariates that were originally considered, and the results are reported in table 7.5.

Table 7.5: Summary table of results from fitting an initial zero-inflated negative binomial model

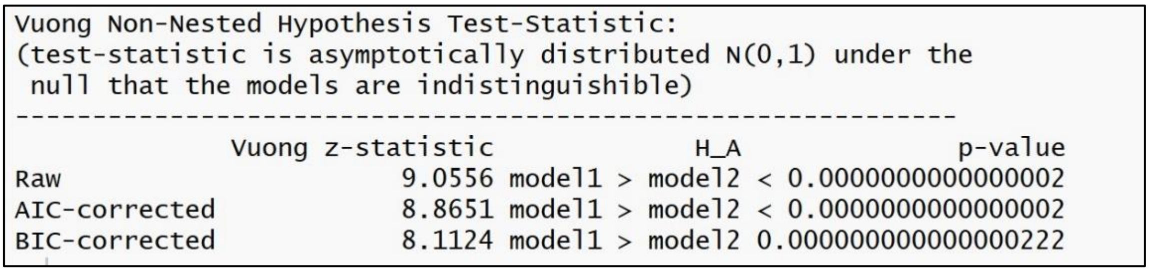


```
Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.9663 -0.0799 -0.0487 -0.0263 47.2085

Count model coefficients (negbin with log link):
                             Estimate Std. Error z value      Pr(>|z|)
(Intercept)                  -5.28463    0.85334   -6.19 0.00000000059 ***
averageEnergyRestoredByFood -0.00108    0.00116   -0.93        0.3524
averageGold                   0.00551    0.00126    4.36 0.00001275301 ***
averageKillHitRatio           1.25889    0.44591    2.82        0.0048 **
averageSingleEBUsed           0.03456    0.02311    1.50        0.1349
missionCompletionRate         4.92101    0.83368    5.90 0.00000000357 ***
Log(theta)                    0.04708    0.16039    0.29        0.7691

Zero-inflation model coefficients (binomial with logit link):
                                 Estimate   Std. Error z value             Pr(>|z|)
(Intercept)                  10.232917098  1.085532320    9.43 < 0.0000000000000002 ***
averageEnergyRestoredByFood  0.000574155  0.001570106    0.37              0.71460
averageGold                  -0.218841061  0.017159713  -12.75 < 0.0000000000000002 ***
averageKillHitRatio          -3.522627632  0.596532872   -5.91         0.0000000035 ***
averageNationalCurrency       0.000000546  0.000055388    0.01              0.99213
averageSingleEBUsed           0.146770440  0.031423663    4.67         0.0000030018 ***
missionCompletionRate        -3.931562355  1.043703503   -3.77              0.00017 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.048
Number of iterations in BFGS optimization: 107
```

The fitted zero-inflated negative binomial model results are presented as two parts, the count component (first set of coefficients), and the binary component (second set of coefficients) that represent the probability that a case has a zero outcome recorded. For the count component, the intercept and the explanatory variables denoting amount of gold possessed, kill:hit in virtual fights and completion rate in missions achieve statistical significance with p<<0.01. For the binary component, all independent variables, except energy restored by food and quantity of national currency possessed, are statistically significant with p<<0.01.

Hilbe (2014) discuss a non-nested test of a zero-inflated model against its non-inflated counterpart called the Vuong test (Vuong, 1989), the null hypothesis of which states that both models are equally well fitted to the data. A statistically significant result would indicate that the zero-inflated model is preferred over the non-inflated one. However, the test is biased towards favouring the zero-inflated model, due to which AIC and BIC based correction factors are developed, which adjust for the extra parameters in the zero-inflated component (Desmarais & Harden, 2013). Thus, a Vuong test is performed in order to compare the zero-inflated negative binomial with the standard negative binomial model fitted earlier, and the results are illustrated in table 7.6.

Table 7.6: Vuong test comparing the zero-inflated negative binomial and the standard negative binomial models



```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
------------------------------------------------------------
              Vuong z-statistic             H_A            p-value
Raw                      9.0556 model1 > model2 < 0.0000000000000002
AIC-corrected            8.8651 model1 > model2 < 0.0000000000000002
BIC-corrected            8.1124 model1 > model2 0.000000000000000222
```

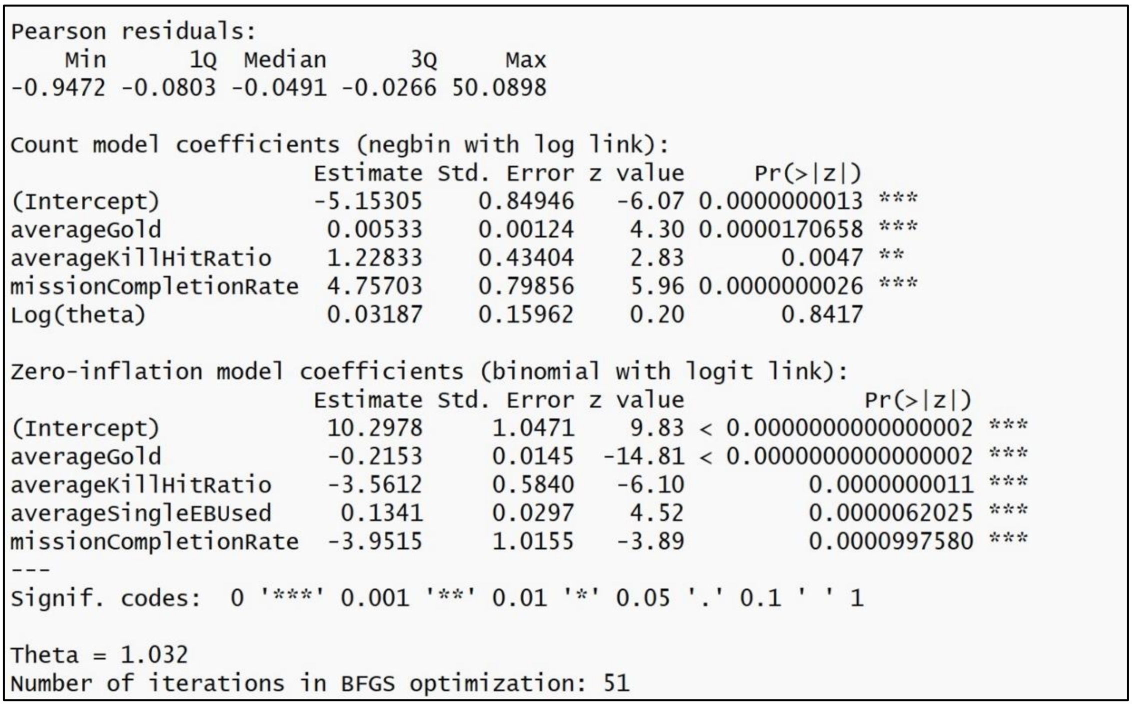
Model1 in table 7.6 denotes the zero-inflated negative binomial model, while model2 denotes the standard negative binomial model. The Vuong statistic as well as the AIC and BIC based correction factors are statistically significant, implying that there is evidence to reject the null hypothesis and the zero-inflated negative binomial is a better fit to the data than its non-inflated counterpart.

The final zero-inflated model is now fitted after removal of the non-significant variables, and the results stated in table 7.7.

All covariates in both the binary and the count parts of the fitted model show statistical significance. Since the final model is nested within the initial model, the prevalent likelihood ratio test for comparison of nested models (Zeileis et al., 2008) is implemented, which results in p=0.42, indicating a lack of evidence in rejecting the null hypothesis that the larger model is not a significant improvement over the nested one. Therefore, the less complex zero-inflated negative binomial model is the preferable one and exhibits better fit to the data.

Comparing the final fitted negative binomial-logit hurdle model displayed in table 7.4 and the final fitted zero-inflated negative binomial model displayed in table 7.7, it can be seen that the hurdle model is a nested version of the zero-inflated model, wherein both models are alike, except that the hurdle does not include the kill:hit variable in its count part. Therefore, a further comparison between the hurdle and the zero-inflated models is performed using the likelihood ratio test for nested models, which results in p<<0.00, implying that there is evidence to reject the null hypothesis and that the larger zero-inflated model is a significant improvement over the nested hurdle model.

Table 7.7: Summary table of results from fitting the final zero-inflated negative binomial model



```
Pearson residuals:
    Min      1Q  Median      3Q     Max 
-0.9472 -0.0803 -0.0491 -0.0266 50.0898 

Count model coefficients (negbin with log link):
                      Estimate Std. Error z value     Pr(>|z|)    
(Intercept)           -5.15305    0.84946   -6.07 0.0000000013 ***
averageGold            0.00533    0.00124    4.30 0.0000170658 ***
averageKillHitRatio    1.22833    0.43404    2.83       0.0047 ** 
missionCompletionRate  4.75703    0.79856    5.96 0.0000000026 ***
Log(theta)             0.03187    0.15962    0.20       0.8417    

Zero-inflation model coefficients (binomial with logit link):
                      Estimate Std. Error z value             Pr(>|z|)    
(Intercept)            10.2978     1.0471    9.83 < 0.0000000000000002 ***
averageGold            -0.2153     0.0145  -14.81 < 0.0000000000000002 ***
averageKillHitRatio    -3.5612     0.5840   -6.10         0.0000000011 ***
averageSingleEBUsed     0.1341     0.0297    4.52         0.0000062025 ***
missionCompletionRate  -3.9515     1.0155   -3.89         0.0000997580 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Theta = 1.032 
Number of iterations in BFGS optimization: 51 
```

7.7 Results and Interpretations

The model building procedure results in the zero-inflated negative binomial model exhibiting the best fit to the data on number of micro payments by customers. At each stage of the model building process, the fitted models were evaluated and compared with alternative ones using different goodness of fit statistics. Since it was not possible to use a holdout sample for this analysis, the final fitted model could not be validated regarding its ability to predict future data points.

The zero-inflated negative binomial model, with covariates amount of gold possessed, kill:hit in virtual fights and completion rate in missions (for the count component), and amount of gold possessed, kill:hit in virtual fights, quantity of energy bars used and completion rate in missions (for the binary component) is considered to be the most appropriate in modelling number of micro transactions made by users of eRepublik. Increase in the quantity of premium currency gold, higher ratio of kills to hits i.e. more competency in virtual fights, and better completion rates in missions contribute towards users investing in more number of real currency transactions and consequently multiple payments. This can be rationalised by the fact that as players get better at the game and develop more dependency on quick progression through utilising gold rather than grinding through, the more their need and motivation for premium currency increases. They start to realise and reap the benefits of using it, and hence are willing to invest real money in purchasing more. On the contrary, decrease in the quantity of premium currency gold, lower ratio of kills to hits i.e. less competency in virtual fights, higher use of energy bars, and poor completion rates in missions improves the likelihood of a non-payer converting to a payer. This suggests that generally the more a player struggles in his gameplay, higher will be his inclination to invest in micro transactions, which is understandable intuitively. Additionally, the use of energy bars and the amount of premium currency owned by users also influence their tendency to make real money payments. As players use up their energy bars and gold resulting in depletion of both resources, they are unable to advance through the game since energy bars are the most important virtual items required for participation in tasks and fights, as well as explore other aspects of gameplay. Moreover, having a low amount of gold coupled with not being able to earn any through good performance in missions and fights against opponents would compel users to purchase it through micro payments.

8. Classification of Customers' Gameplay and Social Behaviours Online

The eminence of online freemium games leading to the plethora of players in that industry produce customer behavioural data, which is vital to deduce insights from, in this current competitive world (Bauckhage et al., 2015). Ling and Yen (2001) recognise customer relationship management as a support towards business schemes to develop stable, long lasting and lucrative relationships with customers. A critical aspect of customer relationship management, established by Kracklauer, Mills and Seifert (2004) is customer identification, one of whose elements encompass customer segmentation (Ngai, Xiu & Chau, 2009), which "separates the market (i.e. the consumers) into several groups that are internally homogeneous and heterogeneous vis-à-vis the external members" (Brito, Soares, Almeida, Monte & Byvoet, 2015, p.1). Segmentation of customers allow organisations to focus on promotional, marketing and product development endeavours that are better tailored to the tastes of the target customer segments (Ngai et al., 2009). Hence, customer behavioural data emerging from online freemium games are studied with a view to identification and classification of the existing customer base by means of segmentation.

8.1 Cluster Analysis

Behavioural data sets from the online freemium games industry tend to be of large scale, high dimensional, varied and evolve with time (Bauckhage et al., 2015, Drachen et al., 2012). The ability to assess the gameplay strategies adopted by users is an intrinsic part of the game development process, enabling studios to tweak and enhance the product to better cater to the tastes of their consumers (Drachen et al., 2009). Cluster analysis offers a useful method to overcome the complex data, revealing behavioural patterns that can be used to create profiles of playing styles, thereby allowing game studios to develop customised gaming experiences for their users and thus eventually improving revenue generation (Bauckhage et al., 2015, Drachen et al., 2012).

Cluster analysis is defined as an exploratory multivariate procedure to discover and reveal natural groups existing within the data (Klimberg & McCullough, 2017, Everitt, Landau, Leese & Stahl, 2011, Anderberg, 2014). It is an unsupervised learning technique for pattern recognition (Everitt et al., 2011), and the purpose is to "form groups in such a way that objects in the same group are similar to each other, whereas objects in different groups are as dissimilar as possible" (Kaufman & Rousseeuw, 2009, p.1). Therefore, cluster analytic approaches are adopted to segregate and describe the variety of user behaviours existing in online freemium games.

8.1.1 Gameplay Behaviour Variables

In order to investigate the behaviour of users within the game environment, it is first imperative to establish the variables representing gameplay behaviours. As stated before, there are a total of 40716 active new players between 11 November 2013 and 29 December 2013, whose game events in the period 11 November 2013 and 6 January 2014 are analysed. However, it was found that about 30% of this player base interact with the game for a single session only, i.e. they do not return after their very first logout. A comprehensive analysis of these players (discussed in chapter 5) revealed that they do not really connect with the gaming platform and hence not exhibit adequate consumer behaviours to study. Hence, these users are excluded from the cluster analysis. Furthermore, the intent is to categorise an active user base that have connected with the game, and thus, the already churned players are also not considered in the clustering procedure.

The final data set of customer gameplay behaviour consists of players that are committed to the game, having played for at least a week. This constitutes 4868 (~12%) individuals of the total user base, and 19 variables depicting their playing styles and strategies adopted within the gaming world. These attributes are reflected through –

- achievements accomplished that result in virtual rewards
- participation in various in-game activities related to the military, politics and media
- use of virtual weapons during in-game fights
- rate at which energy (a crucial element required to compete in fights and tasks) is recovered
- damage inflicted on opponents in virtual battles
- damage inflicted on opponents in military campaigns
- use of virtual items such as food to replenish energy
- use of virtual items such as energy bars to replenish energy
- possession of premium currency gold
- performance in virtual fights represented by the ratio of kills to hits
- player's level in the game
- player's military rank in the game
- possession of grind currency national currency

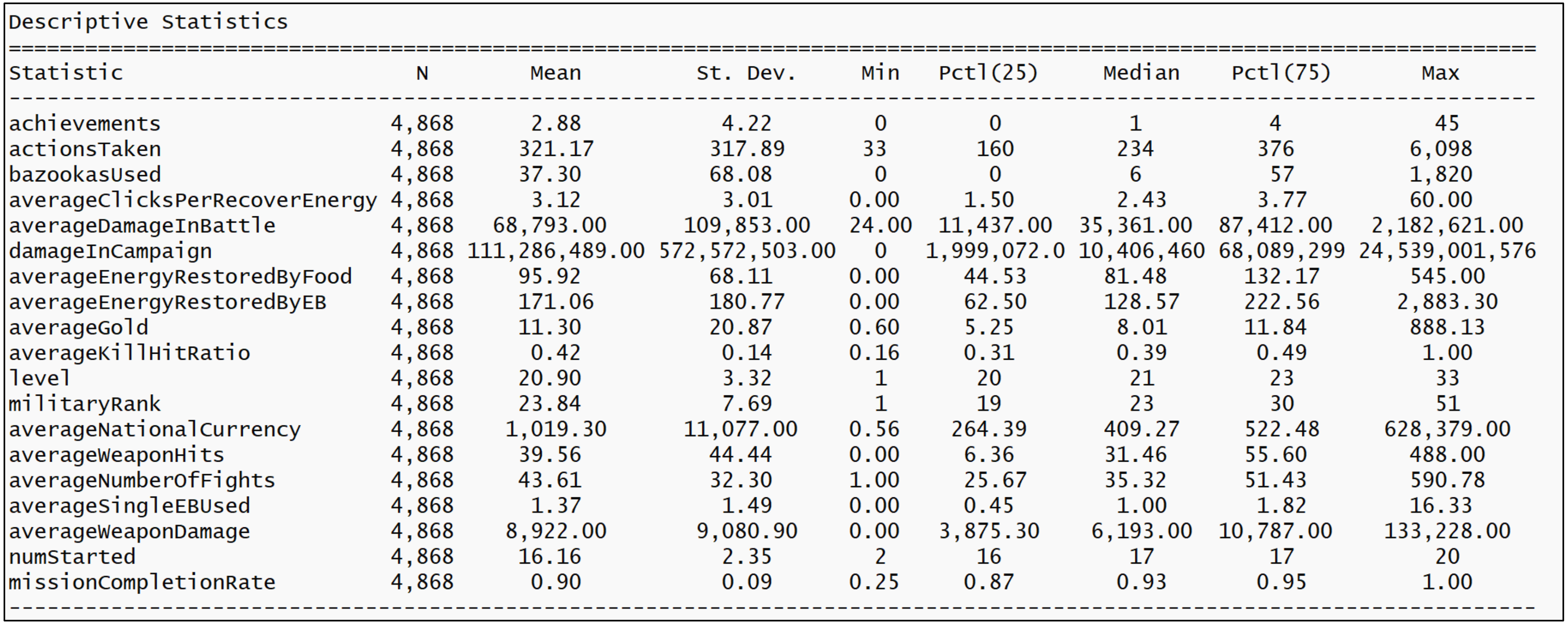
- hits or damage imposed with weapons (a virtual item used in in-game wars against opponents)
- participation in virtual fights
- use of energy bars (an important virtual item necessary for majority of in-game activities such as military campaigns, fights, train the player, build establishments etc.)
- weapon damage dealt with by players
- pre-determined in-game missions initiated
- completion rate in missions

Prior to performing a clustering algorithm, the behavioural variables are inspected and some fundamental statistics reported in table 8.1.

Table 8.1 illustrates a set of descriptive statistics for the user behavioural attributes that will be considered in the clustering procedure. As can be seen, the variables have widely varying means, medians and standard deviations (and equivalently variances). Detection of outliers in the data set is carried out using the Mahalanobis distance, which is a popular measure based on estimated parameters of the multivariate distribution (Ben-Gal, 2005), and incorporates inter-variable dependencies that enable comparison of variable combinations (Hodge & Austin, 2004). Approximately 17% of the cases consist of extreme observations or outliers. Inspection of the quartiles and range for variables reveals fairly heavy tailed skewed distributions. Considering all the above information, the behavioural data is standardized in order to put them on the same scale so that the clustering algorithm does not depend on an arbitrary variable unit and all variables are comparable and given equal weightage.

In the following sections, two prevalent types of clustering methods are implemented, hierarchical and non-hierarchical approaches, and the results compared.

Table 8.1: Summary statistics of the behavioural data to be used in the clustering algorithm



```
Descriptive Statistics
========================================================================================================================
Statistic                     N         Mean          St. Dev.     Min    Pctl(25)     Median     Pctl(75)       Max
------------------------------------------------------------------------------------------------------------------------
achievements                4,868       2.88            4.22        0        0           1           4            45
actionsTaken                4,868      321.17          317.89      33       160         234         376          6,098
bazookasUsed                4,868      37.30           68.08        0        0           6           57          1,820
averageClicksPerRecoverEnergy 4,868     3.12            3.01      0.00     1.50        2.43        3.77         60.00
averageDamageInBattle       4,868    68,793.00      109,853.00   24.00  11,437.00   35,361.00   87,412.00   2,182,621.00
damageInCampaign            4,868 111,286,489.00 572,572,503.00  0    1,999,072.0 10,406,460  68,089,299 24,539,001,576
averageEnergyRestoredByFood 4,868     95.92           68.11      0.00     44.53       81.48      132.17        545.00
averageEnergyRestoredByEB   4,868     171.06         180.77      0.00     62.50      128.57      222.56       2,883.30
averageGold                 4,868     11.30           20.87      0.60     5.25        8.01       11.84         888.13
averageKillHitRatio         4,868      0.42            0.14      0.16     0.31        0.39        0.49          1.00
level                       4,868     20.90            3.32        1       20          21          23            33
militaryRank                4,868     23.84            7.69        1       19          23          30            51
averageNationalCurrency     4,868    1,019.30       11,077.00    0.56    264.39      409.27      522.48      628,379.00
averageWeaponHits           4,868     39.56           44.44      0.00     6.36       31.46       55.60        488.00
averageNumberOfFights       4,868     43.61           32.30      1.00     25.67       35.32       51.43        590.78
averageSingleEBUsed         4,868      1.37            1.49      0.00     0.45        1.00        1.82         16.33
averageWeaponDamage         4,868    8,922.00        9,080.90    0.00   3,875.30    6,193.00   10,787.00    133,228.00
numStarted                  4,868     16.16            2.35        2       16          17          17            20
missionCompletionRate       4,868      0.90            0.09      0.25     0.87        0.93        0.95          1.00
------------------------------------------------------------------------------------------------------------------------
```

8.1.2 Hierarchical Method

The hierarchical approach of forming clusters, introduced by Ward (1963) is based on the notion of producing a hierarchically classified sequence of partitions for the data, based on some adjacency measure for each pair of observations, starting from single less inclusive groups and proceeding sequentially on to larger more inclusive groups (Bridges Jr, 1966, Köhn & Hubert, 2015). This is analogous to the agglomerative algorithm executed here, in which each data point is considered as its own cluster and the distance or similarity measure is calculated between each cluster and all other clusters, resulting in the proximate clusters being successively combined until there remains a single huge cluster (Klimberg & McCullough, 2017, Kassambara, 2017). The commonly used Euclidean distance is calculated as a similarity measure between pairs of observations, while the centroid linkage method which is more robust to outliers (SAS Institute, Inc., 1985, Klimberg & McCullough, 2017), is used as a measure of the proximity between two clusters of observations.

Figure 8.1 displays a visual representation of the hierarchical clustering method by a tree-like structure called the dendrogram (Klimberg & McCullough, 2017, Kassambara, 2017).

The dendrogram resulting from hierarchical clustering using the centroid linkage method is found to be unintelligible. The centroid method is non-monotonic and the resulting dendrogram may be affected by inversions or reversals that are hard to interpret (Manning, Raghavan & Schütze, 2008), which seems to be the case here. The hierarchical clustering algorithm is performed again, using the next best alternatives, Ward's minimum variance method (Ward, 1963) and average linkage. Punj & Stewart (1983) assert that Ward's minimum variance method and average linkage seem to surpass the other approaches, with Ward's method performing better than average linkage barring when outliers are present. Figures 8.2 and 8.3 illustrate dendrograms resulting from Ward's method and average linkage method respectively.

Both dendrograms in figures 8.2 and 8.3 are not entirely comprehensible due to the large number of observations overlapping each other in plot area. However, figure 8.2 showing Ward's method is comparatively interpretable, and based on the height on the y-axis that indicates proximity between observations (higher the height, less similar are the observations), roughly 4 different groups seem to exist within the data, presented in figures 8.4 and 8.5.

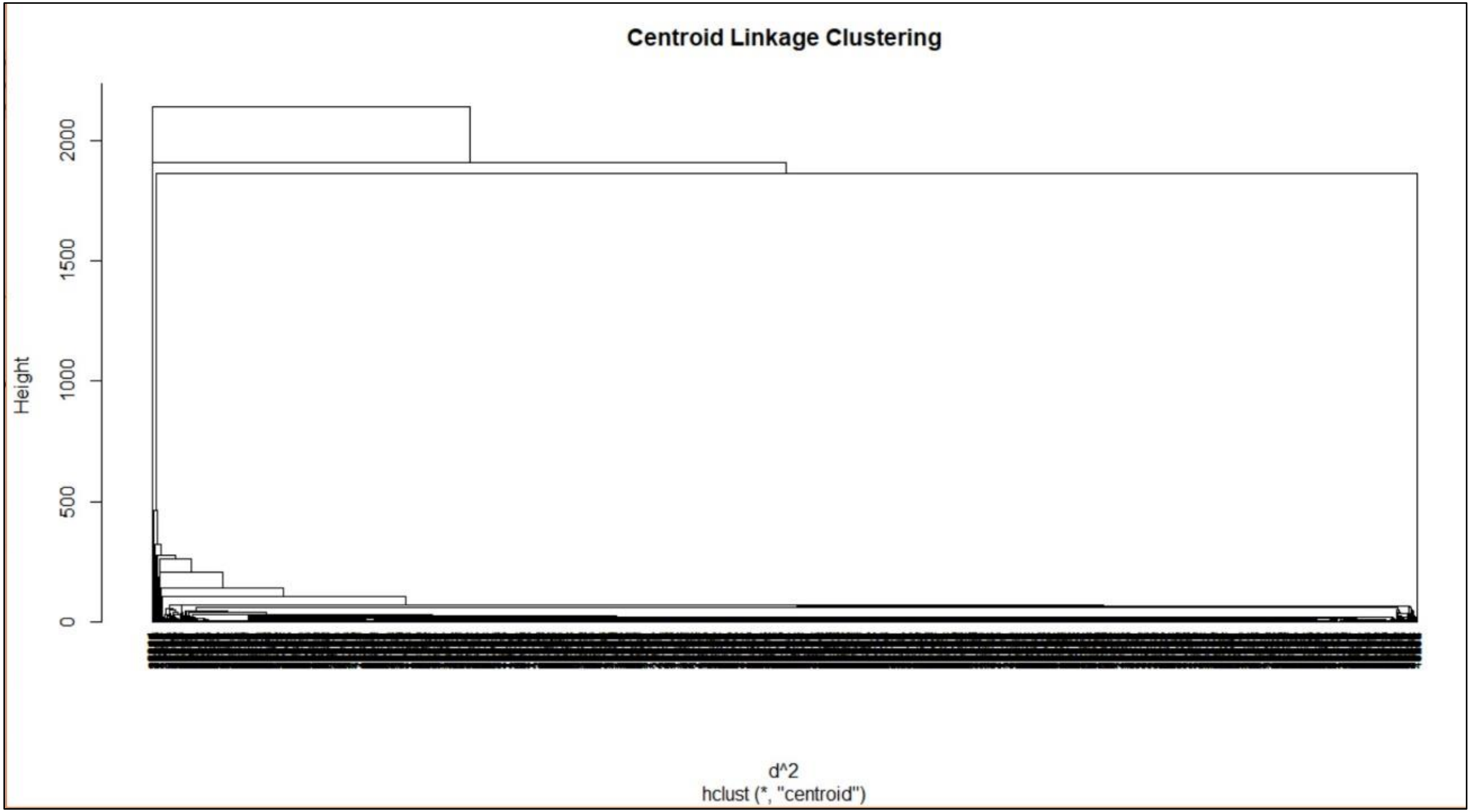


Figure 8.1: Dendrogram showing hierarchical clustering of player behaviour using the centroid linkage method

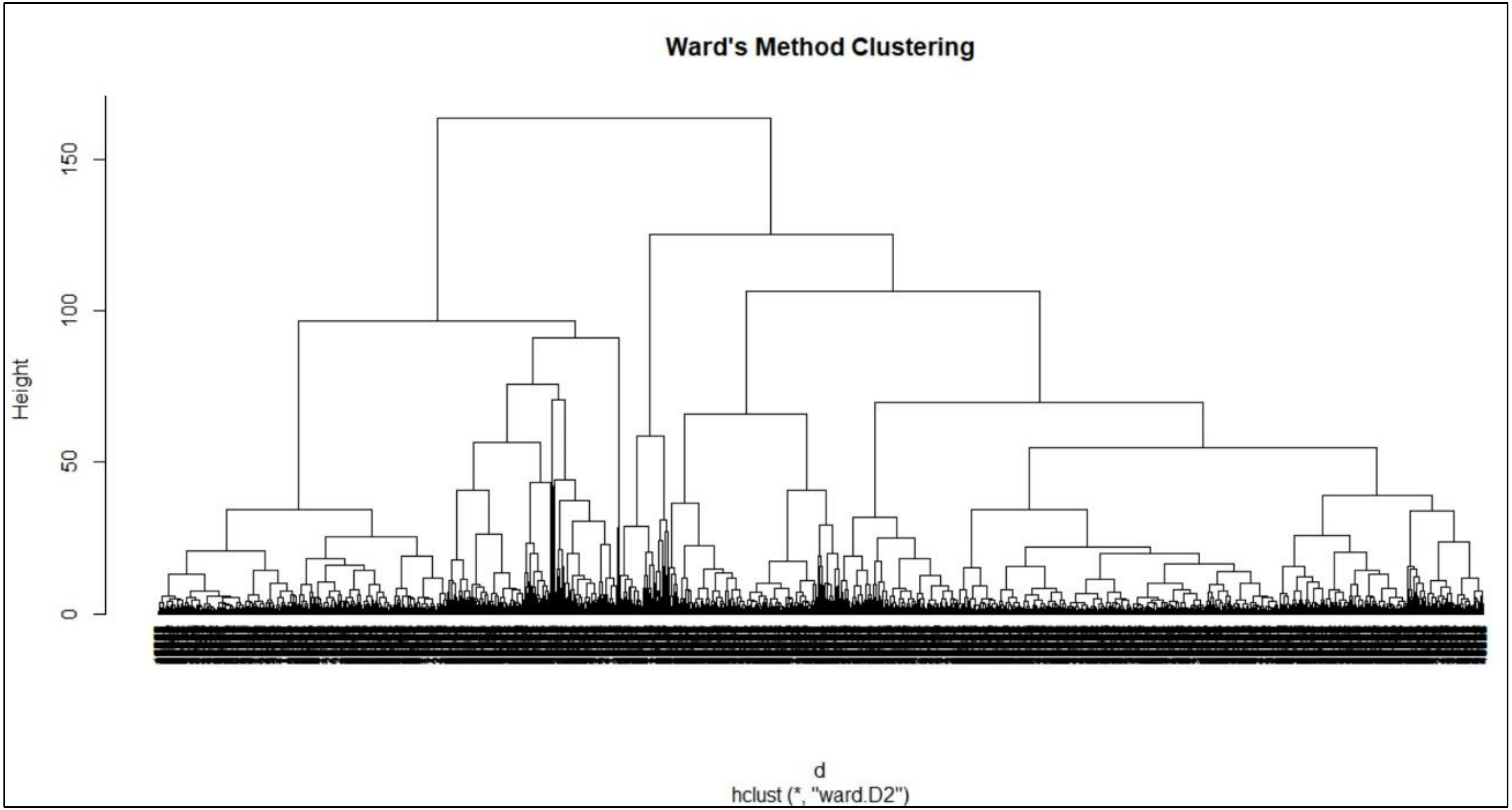


Figure 8.2: Dendrogram showing hierarchical clustering of player behaviour using Ward’s minimum variance method

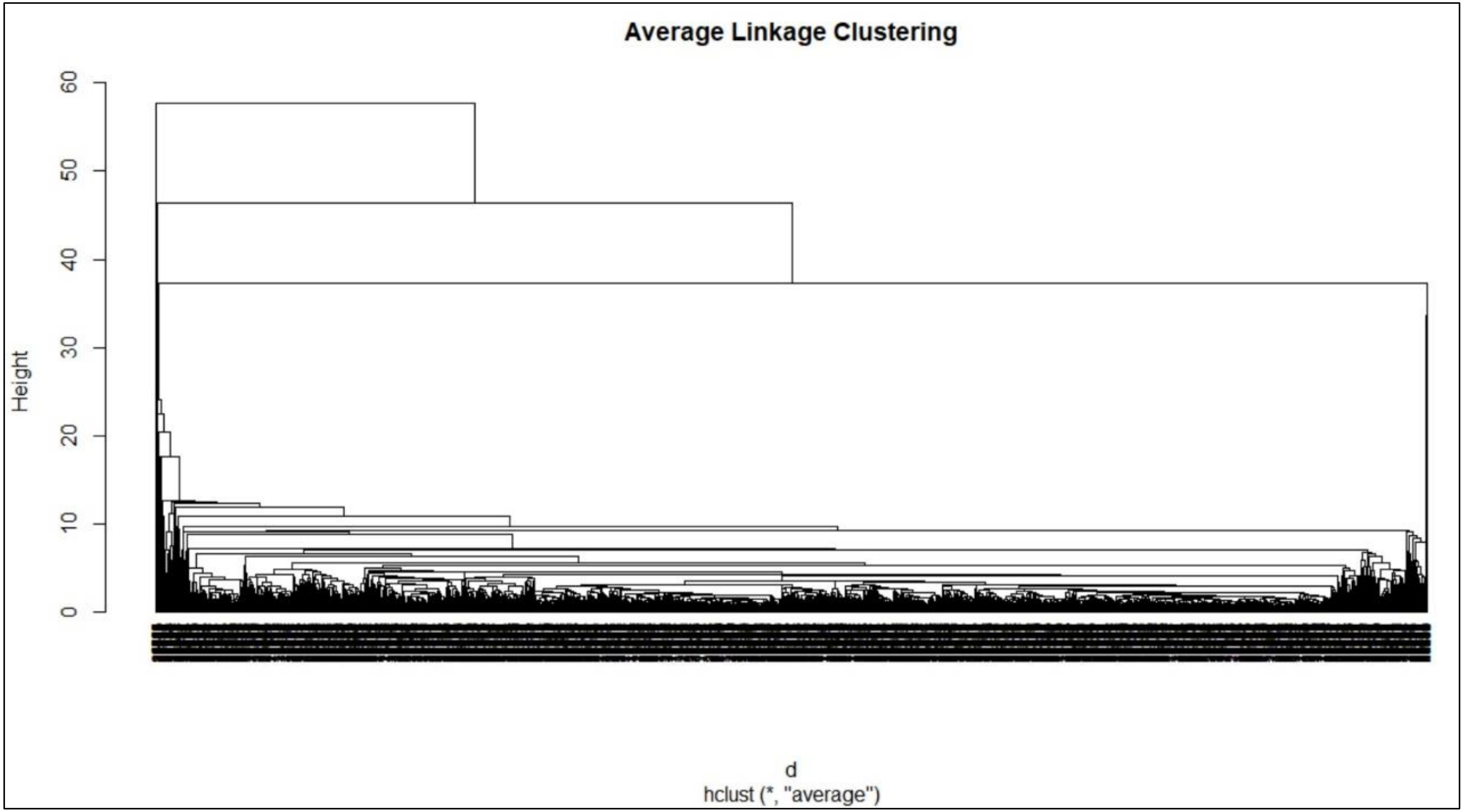


Figure 8.3: Dendrogram showing hierarchical clustering of player behaviour using average linkage method

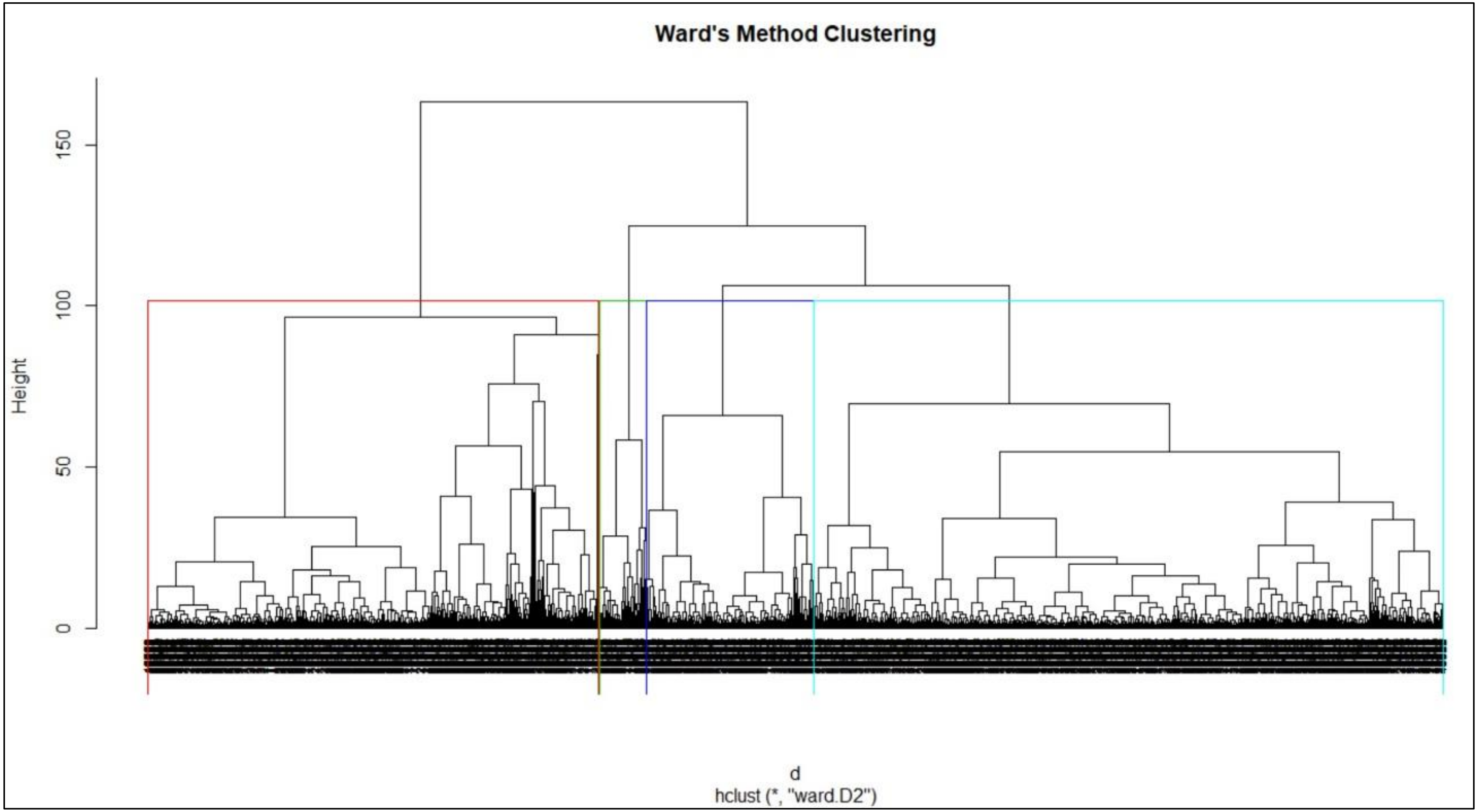


Figure 8.4: Dendrogram showing hierarchical clustering using Ward's minimum variance method with four clusters of players

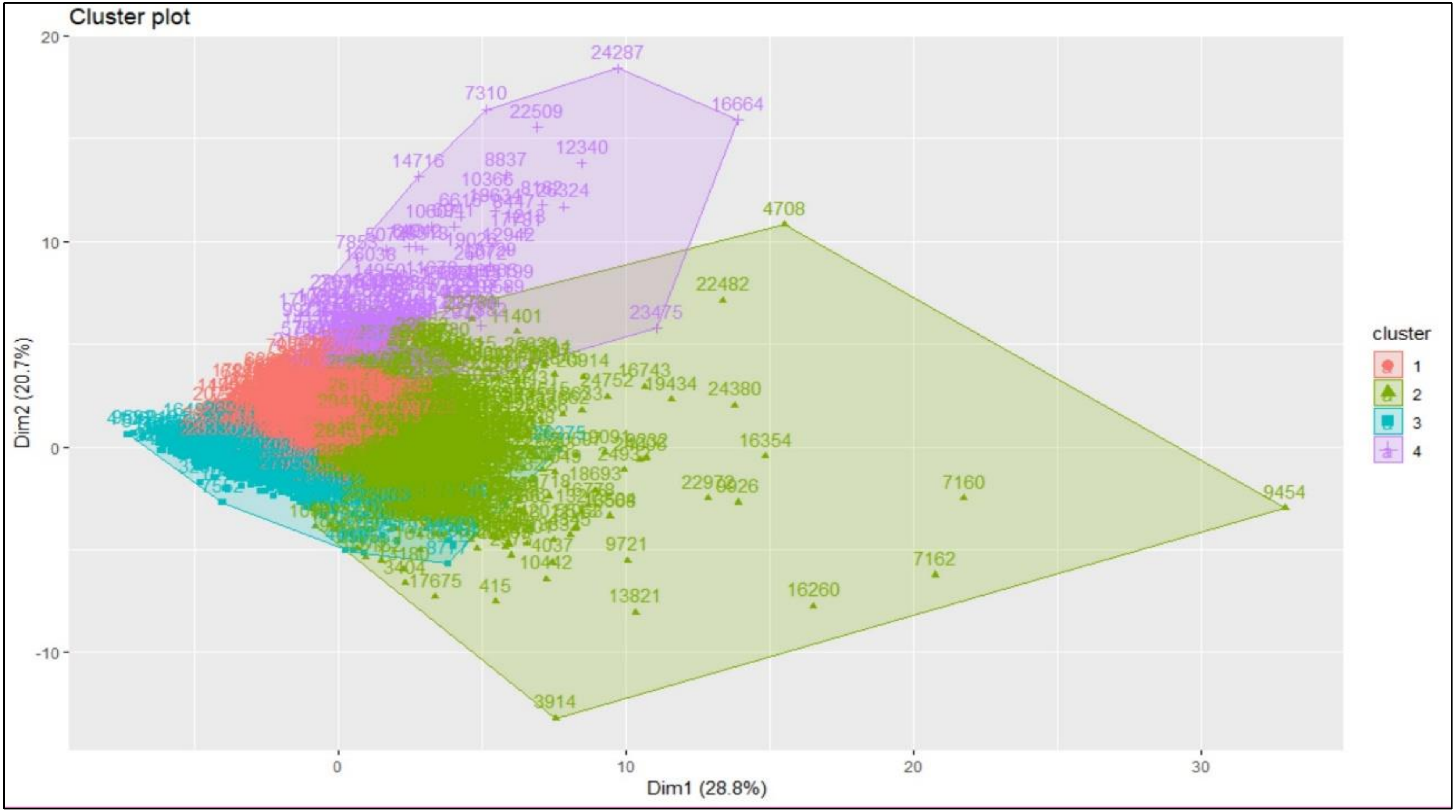


Figure 8.5: Scatter plot showing hierarchical clustering using Ward's minimum variance method of four clusters of players

The 4868 users, grouped into 4 clusters displayed in figures 8.4 and 8.5 appear to be distributed as follows –

- Cluster 1 consists of 2365 individuals and is compactly grouped together without the presence of outliers
- Cluster 2 consists of 1694 individuals, but is a widely spread out group with some extreme observations
- Cluster 3 consists of 631 individuals and also form a tight knit group
- Cluster 4 consists of 178 individuals, and is a moderately spread out with some extreme observations

Klimberg and McCullough (2017) state a bias of the Ward's method towards generating similar number of observations. Although this is not evident in the results attained here, the 4-cluster solution obtained from a hierarchical algorithm using Ward's method of minimum variance needs to be further verified by comparison with a different approach, i.e. the partitioning method of clustering. This is also required since the dendrograms produced are not very coherent, leading to incorrect inferences regarding the clustering structure.

8.1.3 Partitioning Method

This technique for clustering employs a partitioning algorithm to classify observations into *k* partitions iteratively until an objective partitioning criterion is optimised, which results in observations within a partition being similar whereas observations of different partitions are dissimilar (Pal, Ray & Ganivada, 2017). The two most widely used partitioning algorithms *k*-means and *k*-medoids are considered in this analysis. The *k*-means approach (Hartigan, 1975, Hartigan & Wong, 1979) uses the sum of disagreement between a data point and its centroid as the objective criterion and expresses each of the *k* clusters by the mean or weighted average (centroid) of its observations (Berkhin, 2006). The *k*-medoids procedure (Kaufman & Rousseeuw, 1990) uses the average distance between a data point and the corresponding medoid as the objective criterion and expresses each of the *k* clusters by the medoid of its observations (Berkhin, 2006). The *k*-means approach, as highlighted by Kassambara (2017) and Berkhin (2006) suffers from certain drawbacks such as being sensitive to anomalous data points (noise) and outliers, greatly dependent on the initial partitions based on centroids, not scalable and ineffective when used with a global cluster (Arora & Varshney, 2016). The behavioural data used in this study is heavily skewed and contains considerable number of extreme points, therefore, the more robust (Arora & Varshney, 2016) *k*-medoids approach is adopted since "the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is insensitive to the presence of outliers" (Berkhin, 2006, p.37).

The PAM (Partitioning Around Medoids) algorithm (Kaufman & Rousseeuw, 1990) of *k*-medoids clustering is implemented to the customer behavioural data, with *k*=4 (following from the hierarchical approach). The *k*-medoids partitioning results are visualised in figure 8.6 and the clustering information is presented in table 8.2.

A visual representation of the grouping structure is illustrated in figure 8.6 and statistics for the groups formed are reported in table 8.2. The statistics describe the number of individuals in each group, the maximum dissimilarity between the observations and the medoid in each cluster, the average dissimilarity between the observations and the medoid in each cluster, the maximum dissimilarity between pairs of observations within each group (diameter) and the minimum dissimilarity between pairs of observations belonging to two different groups (separation).

Cluster 1 is the largest and internally most homogenous, with dissimilarity measures within the cluster being the least (max_diss, av_diss and diameter). Cluster 3, the second largest group is the most spread out, with a huge difference between the maximum and average dissimilarity measures, and considerably heterogeneous both internally and externally with large values for diameter and separation. A similar structure is observed in cluster 2, albeit slightly less spread out and moderately heterogeneous internally and externally. Internal (within-cluster) homogeneity and external (between-clusters) heterogeneity is the goal of any clustering method, and clusters 1 and 4, although have good to moderate within-cluster homogeneity, are less heterogeneous between-clusters.

Comparison of figures 8.5 and 8.6 suggests that the overall grouping structure of the customer behavioural data in this study produced by Ward's hierarchical approach and *k*-medoids partitioning approach seems to be analogous with each other. Comparing the allocation of individuals to clusters by the two approaches reveals that about 75% of cases are assigned to the same cluster by the two techniques. Moreover, the *k*-medoids algorithm seems to generate a relatively more even distribution of individuals to the 4 clusters as compared to the Ward's approach.

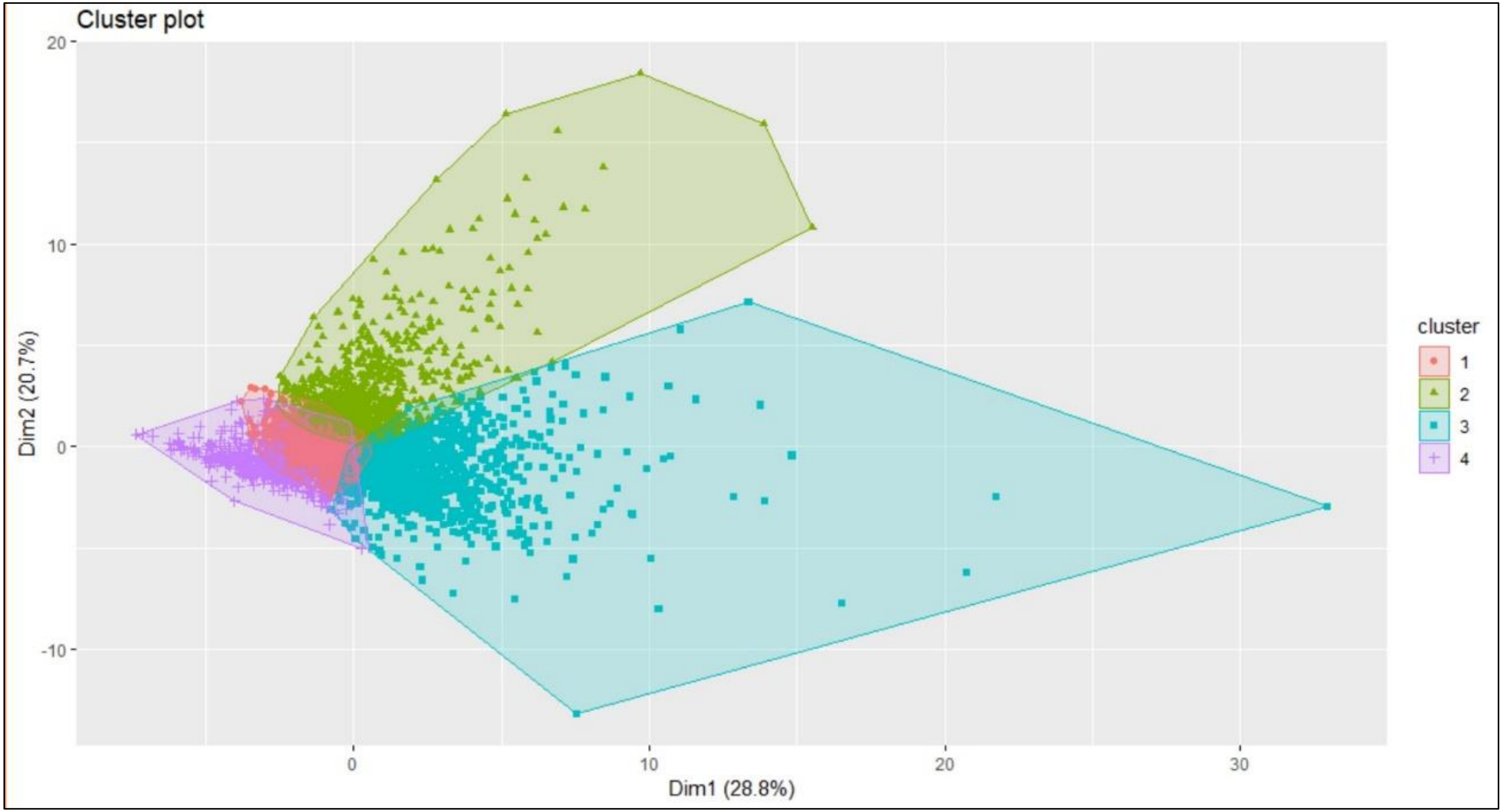
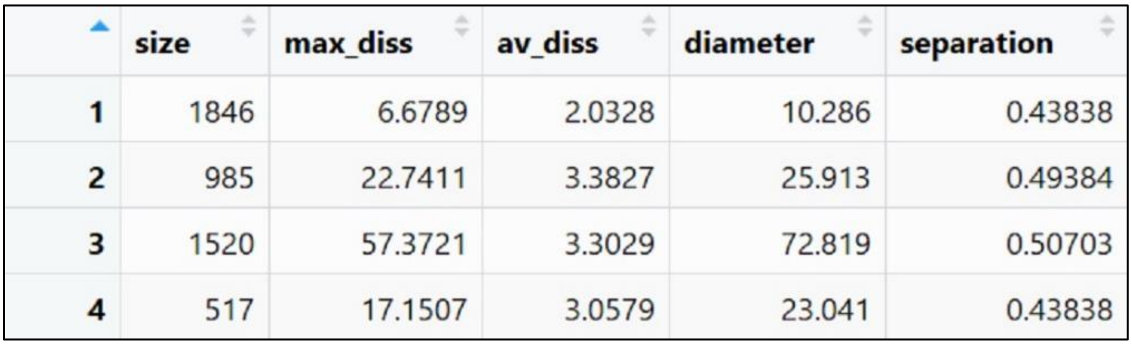


Figure 8.6: Scatter plot showing *k*-medoids clustering using the PAM algorithm of four clusters of players

Table 8.2: Clustering information for the *k*-medoids PAM approach



	size	max_diss	av_diss	diameter	separation
1	1846	6.6789	2.0328	10.286	0.43838
2	985	22.7411	3.3827	25.913	0.49384
3	1520	57.3721	3.3029	72.819	0.50703
4	517	17.1507	3.0579	23.041	0.43838

8.1.4 Evaluation

Prior to the interpretation of groups with respect to user gameplay behaviours, the goodness of clustering results are assessed and compared via clustering validation (Maulik & Bandyopadhyay, 2002). Primarily, there are two types of validation – internal and external, of which the latter requires external information not existing in the data, such as knowledge of the real cluster number in advance (Liu, Li, Xiong, Gao & Wu, 2010). Since there is no external knowledge available to the researcher barring the gameplay data, only internal validation procedures that estimate the merits of a clustering structure without relying on external information (Tan, Steinbach & Kumar, 2005) will be focused on.

A good clustering method should be able to split the data into groups such that entities in the same cluster are as similar as possible, whereas those in different clusters are as disassociated as possible (Tan et al., 2005, Zhao & Karypis, 2002). Thus, internal validation measures that reflect the compactness and separation of the cluster partitions are selected (Brock, Pihur, Datta & Datta, 2011) as assessors of cluster quality.

Compactness determines how close objects are within the same cluster, with lower intra-cluster variation indicating good compactness (Liu et al., 2010, Brock et al., 2011). Separation judges how distant and disconnected the clusters are from each other, generally through measurement of the distance between cluster centroids or pairwise minimum distance between data points belonging to different clusters (Liu et al., 2010, Brock et al., 2011). Compactness and separation measures are combined together and captured by two indices called Silhouette Width and Dunn Index (Brock et al., 2011).

Silhouette Width (Rousseeuw, 1987) denotes the mean of the Silhouette value for each observation, where Silhouette value quantifies the amount of confidence in the assignment of an observation to a particular cluster, and ranges from -1 (poorly grouped observations) to 1 (well grouped observations). Figures 8.7 and 8.8 show Silhouette plots for the two clustering approaches demonstrating the average Silhouette Width across all clusters and that for each cluster.

The average Silhouette Width overall corresponding to both approaches is approximately 0.2, indicating a low moderate quality of grouping structure for observations. Cluster 1 represents the set of most well grouped observations with a Silhouette Width of approximately 0.4 for both methods. The remaining clusters have some observations with a negative silhouette coefficient (seen in figure 8.8) which may imply that they are not in the right cluster.

Dunn Index (Dunn, 1974) is the ratio of the minimum average dissimilarity between two clusters (inter-cluster distance) to the maximum average within cluster dissimilarity (intra-cluster distance). The objective of grouping is to inflate the inter-cluster distance while diminishing the intra-cluster distance between groups, and hence the larger the Dunn Index, the more compact and well separated the clusters are (Saitta, Raphael & Smith, 2007). The Dunn Indices for hierarchical and *k*-medoids are 0.75 and 1.0 respectively, thereby indicating that *k*-medoids produce slightly more compact and well-separated clusters. However, it is to be noted that the Dunn Index is quite conservative and sensitive to outliers, and “useful for identifying clean clusters in data sets containing no more than hundreds of points” (Saitta et al., 2007, p.4).

Overall, some general cluster validation statistics for both algorithms are displayed in table 8.3 to further examine them in contrast to each other. These are distance-based statistics that can be used for cluster validation and comparison.

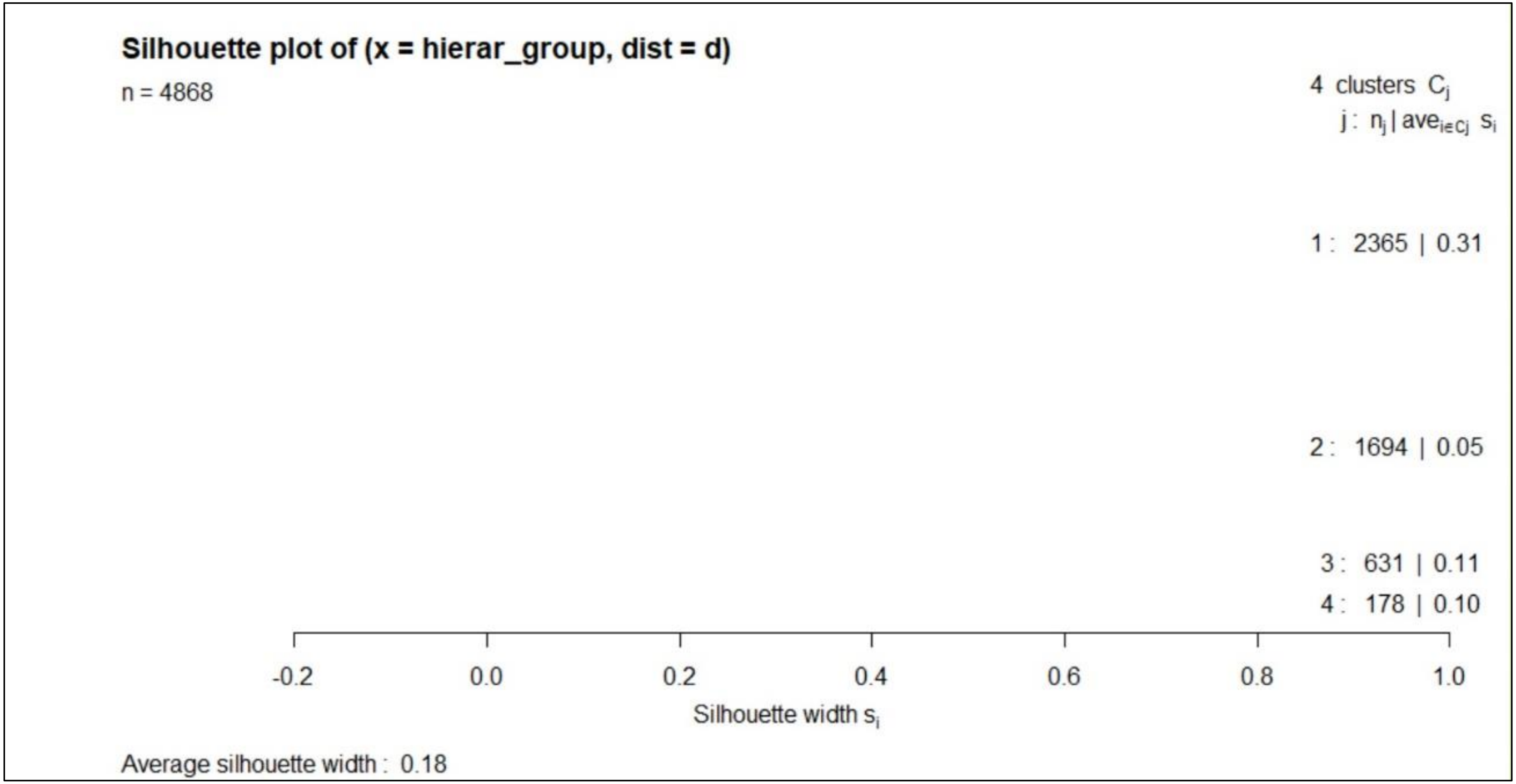


Figure 8.7: Silhouette plot for hierarchical clustering using Ward's minimum variance method

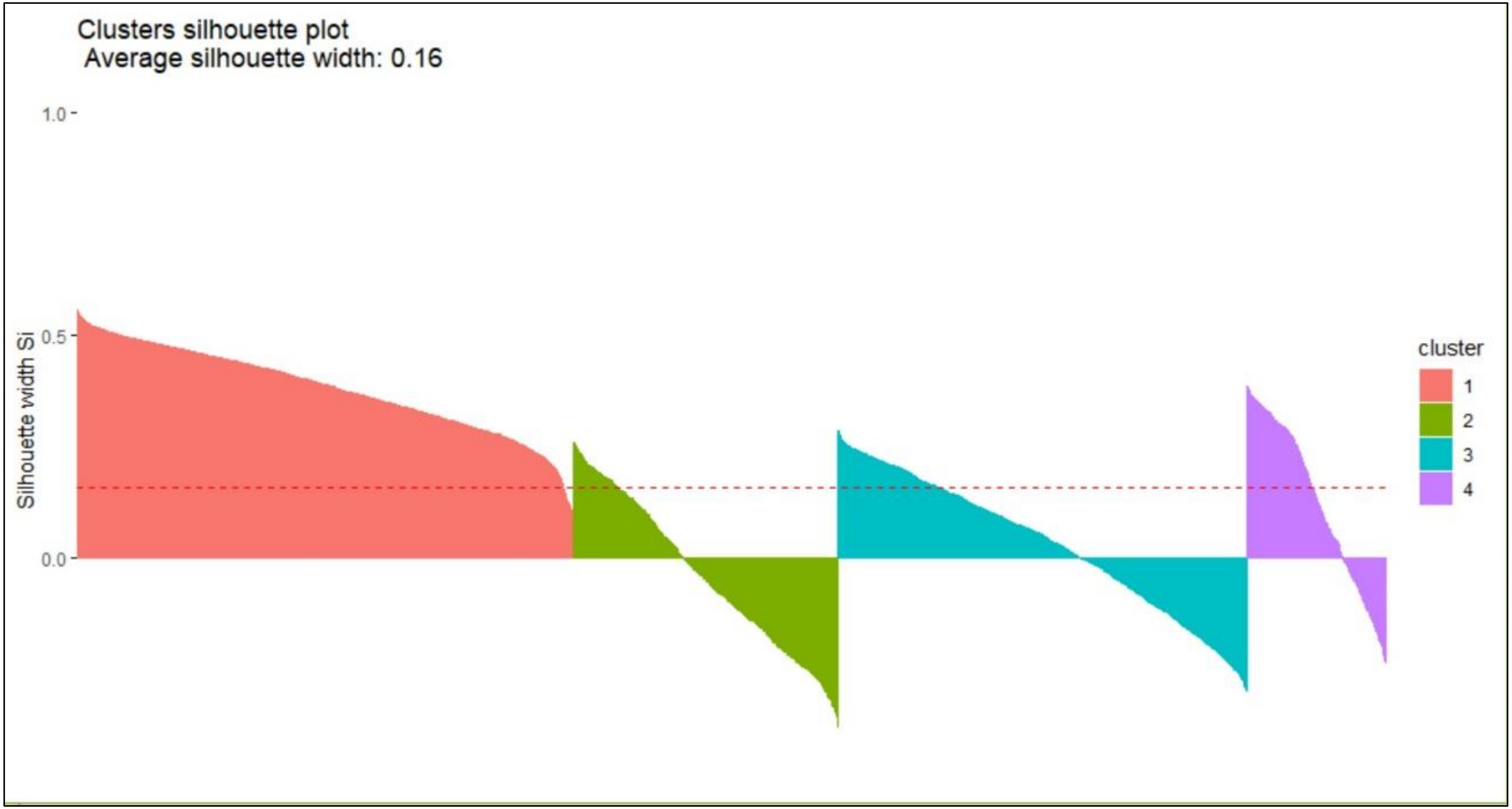


Figure 8.8: Silhouette plot for *k*-medoids clustering using the PAM algorithm

Table 8.3: Distance based cluster statistics for hierarchical and partitioning cluster analyses

Statistics	**Ward's Hierarchical**	***k*-Medoids PAM**
Total size	4868	4868
Cluster size	2365, 178, 1694, 631	1846, 985, 1520, 517
Maximum intra-cluster distances (diameter)	11.91, 25.91, 72.82, 20.8	10.29, 25.91, 72.82, 23.04
Minimum distances between a point in a cluster and a point in another cluster (separation)	0.61, 1.17, 0.61, 0.8	0.44, 0.49, 0.51, 0.44
Average inter-cluster distance	5.84	5.65
Average intra-cluster distance	3.9	3.83
Widest intra-cluster gap	31.31	31.31

The table reveals a slightly more even distribution of observations to the clusters for *k*-medoids. The average inter-cluster distance is greater (i.e. better) for Ward's, but the average intra-cluster distance is smaller (i.e. better) for *k*-medoids.

Overall, it can be seen that the two approaches towards clustering of customer behavioural data, the hierarchical Ward's minimum variance and the partitioning *k*-medoids do not considerably vary from each other. The groupings formed by both are of a low moderate quality; however, this does not seem to improve appreciably when other algorithms such as the centroid or average linkage methods or *k*-means are applied. Therefore, the final grouping of observations follow the procedure specified below –

- About 75% of observations are classified into the same group by both methods, so these are left intact
- The remaining observations for which the groupings are mismatched, the Silhouette values (degree of confidence in the assignment of an observation to a particular cluster) are computed and the class membership corresponding to the method with the higher Silhouette value is accepted. In case of an observation with a negative Silhouette coefficient (implying that it may not be in the correct cluster), the neighbouring cluster to which it is closest to is determined, and the observation is assigned to that.

This approach leads to an improvement in agreement of cluster memberships by the two techniques, wherein 87% of the cases are classified into the same group by both methods.

8.1.5 Results and Interpretation

Following the assignment of users into respective clusters employing the approach elucidated above, their characteristics concerning general gameplay, performance and competence, engagement and advancement, and monetisation are studied. Table 8.4 exhibits summary statistics of certain variables that can intuitively describe these characteristics for players.

The playing styles of users belonging to each group, formed as a result of conducting cluster analysis of their behavioural data, is demonstrated in table 8.4. This is done by means of appropriate variables that highlight their general in-game strategies and performance. The last column represents the total amount of real currency transactions (€) made by each group. Four distinct gameplay styles of eRepublik customers, based on table 8.4 are outlined below –

- Engaged, competent in missions, achievers & high paying (Cluster 3): This is the most profitable group for game studios. It makes the maximum amount of real currency purchases, thereby contributing highly to the revenue generation. The players are very engaged with the game and make the highest progress, as indicated by their in-game achievements and military rank. They are more competent in missions (mission completion rate) than virtual fights (average kill:hit). They prefer to spend money on premium currency (average gold) to help them advance through the game than grinding their way through it through grind currency (average national currency). They do not require to use many energy bars (virtual item) due to their judicious participation in fights and other activities as well as general competency.

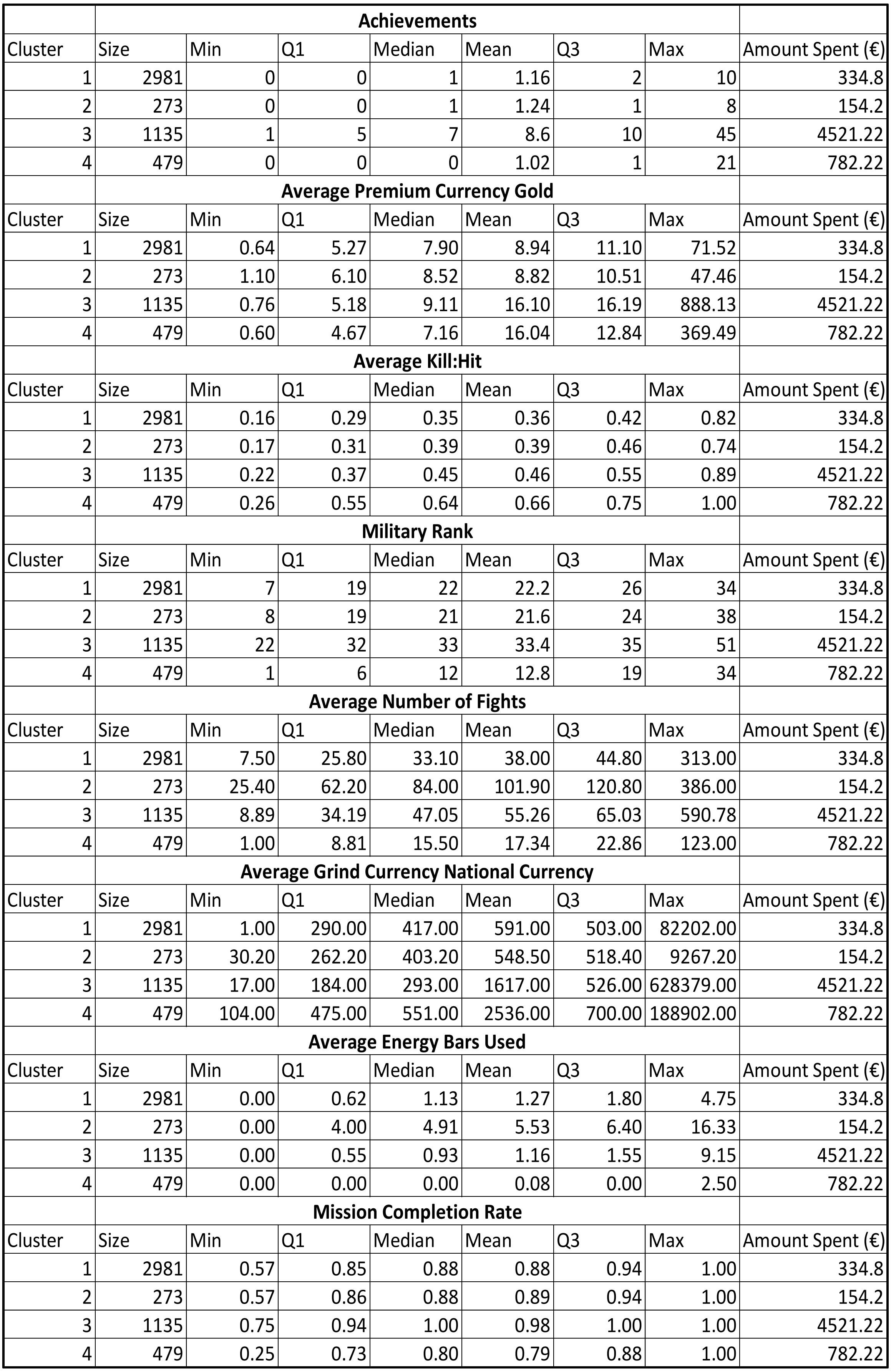
- Slow progressors, competent in fights, low achievers & moderately high paying (Cluster 4): This constitutes the second highest paying group of users. These players, although enjoy the game (as seen from their tendency to invest money in it), make the slowest progress, as indicated by their in-game achievements and military rank. They are most competent in virtual fights and least competent in missions among all other groups. Although they do spend money on premium currency to help them advance through the game, they tend to wait and grind their way through the most amongst all other groups, which may be a reason for their slow progress. They use the least quantity of energy bars due to their least amount of participation in fights coupled with their maximum success in it.

- Moderately engaged, moderate competence in missions, moderate achievers & moderately low paying (Cluster 1): This group represents a conservative set of players with average qualities and is the third highest paying group. The players are engaged with the game to some extent, having moderate in-game achievements and military rank. They have an average performance in missions and the worst performance in virtual fights compared to all other groups. They are balanced in their purchase of premium currency as well as their tendency to drudge through the game for grind currency. The use a modest quantity of energy bars and do not partake in many virtual fights.

- Moderately engaged, moderate competence overall, moderately low achievers & low paying (Cluster 2): This group of players make the least amount of real currency purchases and do not contribute much to the monetisation of the game. They are somewhat immersed into the game, with a moderately low in-game achievements and military rank. They have an average performance in missions and slightly better performance in virtual fights than cluster 1. They invest the least in premium currency, however do not seem to grind their way through a lot either. They use the highest quantity of energy bars due to their most amount of participation in fights compared to all other groups.

From the descriptions of the different user segments that exist, it is evident that players have varied focuses and adopt different tactics to advance through the game. While some are more involved in virtual wars and military campaigns that are flexible and created by other players themselves, others follow a fixed path set by the game and strive to complete the in-game missions. The performance of players also vary from participating in missions or in virtual fights. Several users prefer to take their time and work towards obtaining grind currency for free in order to buy virtual products that will help them in their gameplay. However, others, especially high paying players, are impatient and willing to spend real money to purchase premium currency that will fast track them through the game and provide access to more sophisticated virtual products. The use of virtual products (such as energy bars) also differ between players, depending on the approach they take towards the gameplay and the type of in-game features they explore.

Table 8.4: Descriptive statistics of variables representing gameplay behaviour of players in each cluster



				Achievements				
Cluster	Size	Min	Q1	Median	Mean	Q3	Max	Amount Spent (€)
1	2981	0	0	1	1.16	2	10	334.8
2	273	0	0	1	1.24	1	8	154.2
3	1135	1	5	7	8.6	10	45	4521.22
4	479	0	0	0	1.02	1	21	782.22
				Average Premium Currency Gold				
Cluster	Size	Min	Q1	Median	Mean	Q3	Max	Amount Spent (€)
1	2981	0.64	5.27	7.90	8.94	11.10	71.52	334.8
2	273	1.10	6.10	8.52	8.82	10.51	47.46	154.2
3	1135	0.76	5.18	9.11	16.10	16.19	888.13	4521.22
4	479	0.60	4.67	7.16	16.04	12.84	369.49	782.22
				Average Kill:Hit				
Cluster	Size	Min	Q1	Median	Mean	Q3	Max	Amount Spent (€)
1	2981	0.16	0.29	0.35	0.36	0.42	0.82	334.8
2	273	0.17	0.31	0.39	0.39	0.46	0.74	154.2
3	1135	0.22	0.37	0.45	0.46	0.55	0.89	4521.22
4	479	0.26	0.55	0.64	0.66	0.75	1.00	782.22
				Military Rank				
Cluster	Size	Min	Q1	Median	Mean	Q3	Max	Amount Spent (€)
1	2981	7	19	22	22.2	26	34	334.8
2	273	8	19	21	21.6	24	38	154.2
3	1135	22	32	33	33.4	35	51	4521.22
4	479	1	6	12	12.8	19	34	782.22
				Average Number of Fights				
Cluster	Size	Min	Q1	Median	Mean	Q3	Max	Amount Spent (€)
1	2981	7.50	25.80	33.10	38.00	44.80	313.00	334.8
2	273	25.40	62.20	84.00	101.90	120.80	386.00	154.2
3	1135	8.89	34.19	47.05	55.26	65.03	590.78	4521.22
4	479	1.00	8.81	15.50	17.34	22.86	123.00	782.22
				Average Grind Currency National Currency				
Cluster	Size	Min	Q1	Median	Mean	Q3	Max	Amount Spent (€)
1	2981	1.00	290.00	417.00	591.00	503.00	82202.00	334.8
2	273	30.20	262.20	403.20	548.50	518.40	9267.20	154.2
3	1135	17.00	184.00	293.00	1617.00	526.00	628379.00	4521.22
4	479	104.00	475.00	551.00	2536.00	700.00	188902.00	782.22
				Average Energy Bars Used				
Cluster	Size	Min	Q1	Median	Mean	Q3	Max	Amount Spent (€)
1	2981	0.00	0.62	1.13	1.27	1.80	4.75	334.8
2	273	0.00	4.00	4.91	5.53	6.40	16.33	154.2
3	1135	0.00	0.55	0.93	1.16	1.55	9.15	4521.22
4	479	0.00	0.00	0.00	0.08	0.00	2.50	782.22
				Mission Completion Rate				
Cluster	Size	Min	Q1	Median	Mean	Q3	Max	Amount Spent (€)
1	2981	0.57	0.85	0.88	0.88	0.94	1.00	334.8
2	273	0.57	0.86	0.88	0.89	0.94	1.00	154.2
3	1135	0.75	0.94	1.00	0.98	1.00	1.00	4521.22
4	479	0.25	0.73	0.80	0.79	0.88	1.00	782.22

Clearly, an insight into the wide variety of the player base and the different facets of gameplay is valuable learning for game developers and publishers. This will help them understand the characteristics of profitable players that generate revenue for the game, the approach taken by highly competent users, as well as the tendencies of struggling players that do not help much in the monetisation process. Accordingly, steps can be taken to offer rewards as encouragement to players doing well and provide hints, tips and free incentives to players struggling. Moreover, overall game development can be further enhanced by focusing on popular gameplay features that are being frequently explored and enjoyed by the engaged and lucrative customers.

8.2 Social Network Analysis

The constant growth in the online freemium games business leading to an increasingly abundant number of its users, serves as an enormous playing field for virtual societies and communities, which makes possible the extensive study of virtual social communications and interactions (Schatten, Tomičić & Đurić, 2015). Alsén, Runge, Drachen & Klapper (2016) emphasize that social gameplay is crucial for engagement and monetisation in casual games, especially on mobile platforms, and it is beneficial for games developers to direct their attention to the social aspects and interactions in their games in order to build profitable products. Runge, (2017) discuss that social components of games such as exchanging messages and virtual items in the form of gifts, not only promote its viral growth, but also influence players' engagement with the game and the revenue generated by it. Hence, the ability to comprehend the social structure that exists within a game's user base in terms of how players interact with each other is vital for its growth and commercial success.

Social network analysis, defined by Scott (2017), constitutes a set of methodological procedures that intends to probe into and explain the possible patterns that exist in the social relationships developed by individuals and groups with each other. It is "motivated by a structural intuition based on ties linking social actors" (McCulloh, Armstrong & Johnson, 2013, p.1) and in addition to providing a visual imagery of the social connection of these actors, also aids in the accurate measurement and representation of the structural relations (Knoke & Yang, 2008) using empirical data (McCulloh et al., 2013). The section below attempts to apply some social network concepts to discern the social behaviour of users of eRepublik.

8.2.1 Network Composition

To examine virtual associations between players with respect to a social setting within the game environment, the events message sent and message received are considered. These events capture private communications between users, that may include messages promoting virtual military campaigns initiated by leaders of a military unit, invitations to join different military units or political parties, exchange of virtual items as gifts between 'friends' etc. Due to the ethical considerations of this research, actual content of the messages interchanged are not available or accessible to the researcher. The above information is known solely through extensive use of the gaming platform as an eRepublik player. Thus, the social network analysis implemented here will not be able to explain the definite nature of the social relationship between individuals. Nonetheless, it is useful to gain insight on the communication between players in terms of identifying the important members within the social structure of the game.

Of the total player base of 40716 users, those that have been involved in messaging interactions are first extracted. The structure of the game is such, that new players that login to the platform for the first time are always greeted with a generic welcome message and invitation to join a military unit. These messages are not indicative of a real interaction between users and hence not considered in the analysis. The final data set of customer social behaviour consists of 514 players that have communicated with each other through messages that are suggestive of substantial in-game social interactions throughout the course of their gameplay. Three variables are taken into account – user ID of the sender of a message, user ID of the recipient of a message and the frequency of intercommunication between them.

8.2.2 Network Plot and Properties

In this analysis, each individual player from the 514 considered, is an 'actor', and the relation between a pair of actors, represented through the exchange of messages is a 'tie'. The ties are directed, since one actor initiates and the other receives. Additionally, the ties have a strength associated with them in the form of the intensity of the interaction, which is measured by the frequency of messages exchanged. The tie strength is recorded as 'weights' (Opsahl, Agneessens & Skvoretz, 2010), which leads to the construction of a weighted directed social network structure. Figure 8.9 depicts a sociogram (Knoke & Yang, 2008), illustrating the relations among actors in a confined social structure.

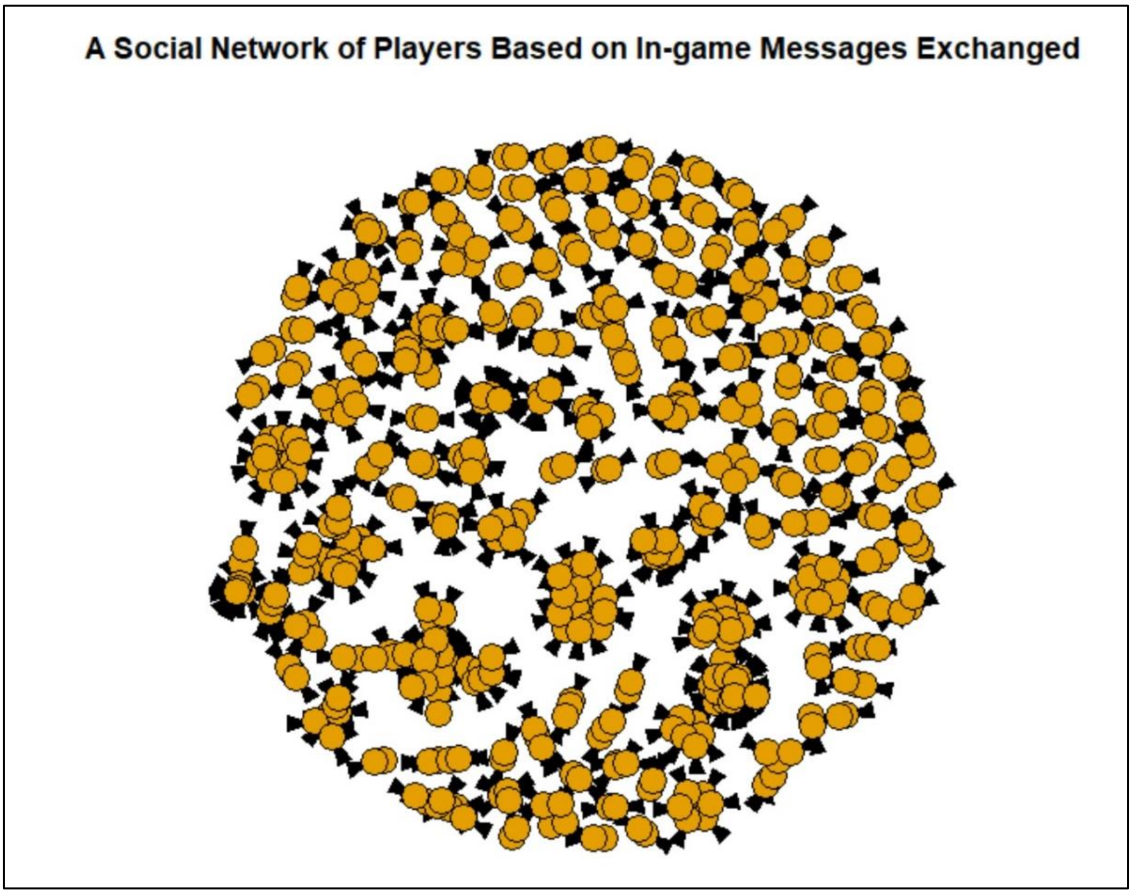

A Social Network of Players Based on In-game Messages Exchanged

Figure 8.9: Sociogram representing the social network of players based on in-game messages exchanged

The social network plot in figure 8.9, which shows the set of actors represented by nodes (points), and the ties between any pair of actors illustrated by edges (lines between a pair of points), enables visualisation of the overall structure of the network. Since the relations are directed from one actor to another, the lines have arrowheads indicating that direction (Knoke & Yang, 2008). Overall, the graph is disconnected (Wasserman & Faust, 1994, Knoke & Yang, 2008), since not every pair of nodes is connected by a path. The sociogram seems to reveal several small clusters of nodes (i.e. individuals) that are clumped together, indicating an interchange of messages between them. There appears to be a few pairs of nodes that are isolated, implying that these users only interact with each other and not to the rest of the player base, however this is not clear from the plot itself as the points are superimposed onto one another due to the amount of data being plotted. The large volume of data also results in the direction and strength of ties being indiscernible, as well as displaying the user IDs (for identification of players) for each node impractical. Therefore, although network visualisation is valuable in order to study the social behavioural structure at a high level, it is imperative to quantitatively detail the characteristics of the network and its actors (nodes).

First, the size of network as a whole is described numerically. The diameter of the network, defined as the length (number of edges) of the longest geodesic distance between a pair of nodes (Wasserman & Faust, 1994, Newman, 2003), is 382, which means that the distance between the nodes that are farthest from each other is 382 units. The mean distance, which is the average geodesic path length (number of edges) between any pair of nodes (Newman, 2003) is 1.99. The large difference between the diameter and the mean distance is indicative of the fact that the overall spread of the network structure is very wide, in which numerous small clusters of nodes and actors are densely connected to each other than to the rest of the network.

This is further investigated by means of the network communities that exist within the structure. The network is partitioned into connected components (Wasserman & Faust, 1994), in which each component is a collection of nodes that are connected to each other (i.e. a path exists between all pairs of nodes in a component), but a node in a particular component is not connected (i.e. there is no path) to any node not belonging to that component. The number of communities or connected components found in this network is 131, which is in line with the previous observation that numerous small clusters exist in this network. Figure 8.10 is a graphical display of the distribution of the size of these clusters.

As is evident from the plot, the maximum number of communities are formed by only pairs of actors or individuals – there are 72 communities of size 2. The largest community is comprised of 36 users, followed by another comprised of 34 users and another with 21 users. These are the groups deemed important and useful for promoting social virality as they indulge in many virtual social interactions across the wider player base. The members of these groups are identifiable through their user IDs, and it is found that these are all paying clusters of individuals with total real currency transactions worth €245.18, €9.8 and €19.8 respectively. This makes them not only socially but also commercially viable.

Having determined the overall structure and attributes of the social network for this data, the characteristics of its actors (eRepublik players) are now studied. This is accomplished with the help of some fundamental relationship measures for social networks.

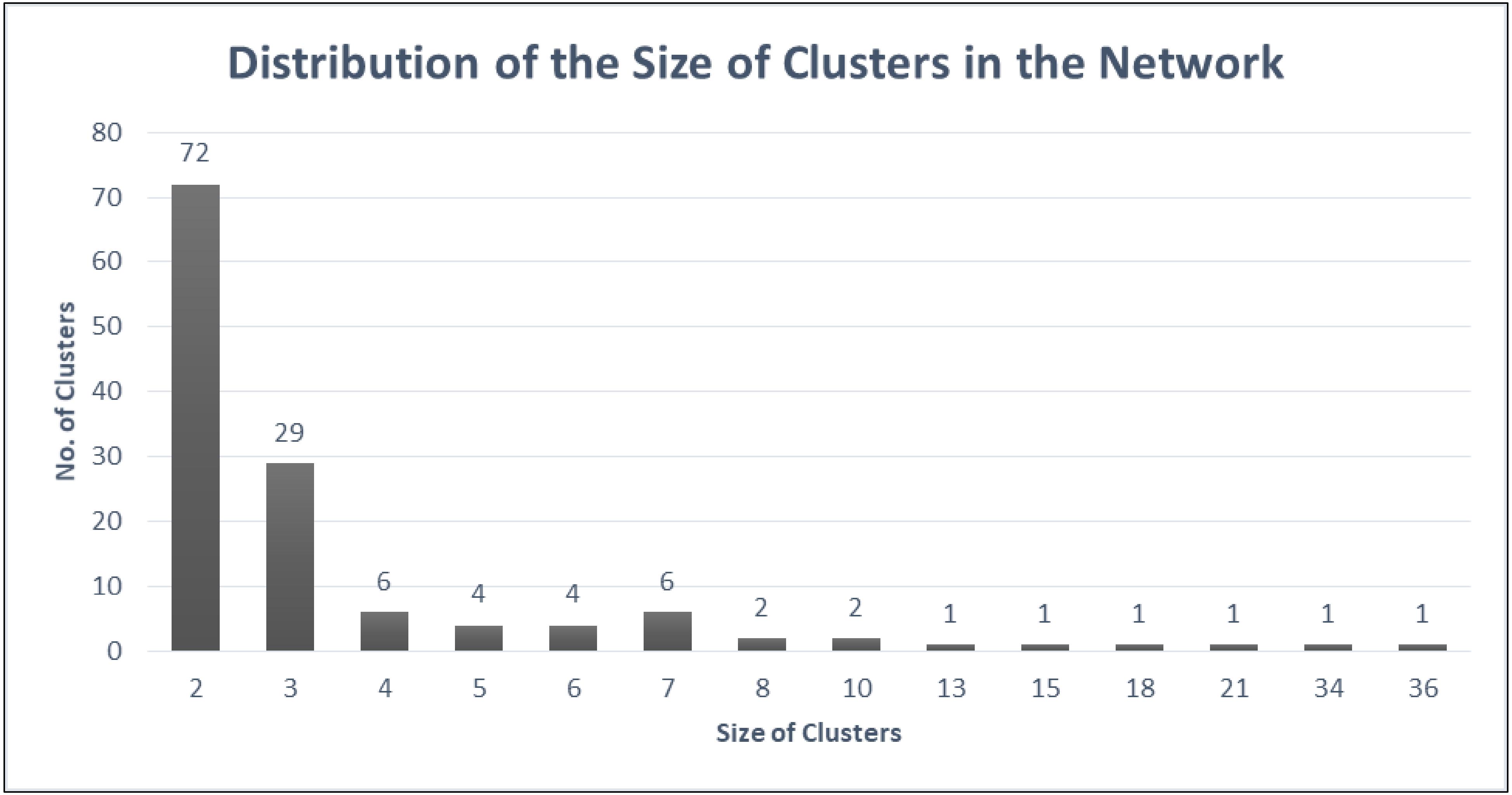


Figure 8.10: Distribution of the size of social network communities based on in-game messaging interactions

8.2.3 Clustering Social Network Measures

An essential goal of social network analysis is to establish the significant and more valuable actors (Knoke & Yang, 2008), achieved by the concept of centrality, which quantifies "graph theoretic ideas about an actor's prominence within a complete network by summarising the structural relations among all nodes" (Knoke & Yang, 2008, p.62.) Traditional centrality measures – degree, closeness and betweenness (Freeman, 1977, Freeman, 1978) are computed and applied in clustering algorithms to determine groups of prominent actors within the data set. Specifically, the variables used in the analysis are –

- Weighted degree centrality: a weighted measure of degree centrality, which evaluates the extent to which a particular node connects to the remaining nodes in a social network, by taking into account the tie strength (weights) between nodes (Wasserman & Faust, 1994). Two types of degree centrality are considered, in-degree centrality that involves incoming links, and out-degree centrality that involves outgoing links. The nodes with higher in-degree denote messages received and are considered more prestigious, whereas those with higher out-degree denote messages sent and are considered more central (Knoke & Yang, 2008).

- Closeness centrality: a measure of the proximity of a node to all other nodes in a social network (Sabidussi, 1966). Two types are considered, in-closeness centrality that evaluates the degree to which a node can be easily reached (in terms of shortest distance) from other nodes, and out-closeness centrality that evaluates the degree to which a node can easily reach (in terms of shortest distance) out to other nodes. Higher the values for closeness, less central the nodes are considered to be.

- Betweeenness centrality: a measure of the extent to which other actors dominate or arbitrate the relations between pairs of actors that are not directly connected, wherein, higher the betweenness centrality, more is the control or mediating power the actor has on the relations in the network (Knoke & Yang, 2008).

The variables are normalised, scaled, and applied to clustering algorithms. A three-cluster solution of the *k*-medoids PAM approach is found to be most suitable and the summary statistics of the (normalised) centrality measures for different groups are presented in table 8.5.

Table 8.5: Descriptive statistics of normalised centrality measures representing social behaviour of players in each cluster

	In-degree Centrality						
Cluster	Size	Min	Q1	Median	Mean	Q3	Max
1	413	0.00	0.01	0.01	0.02	0.02	0.74
2	70	0.00	0.02	0.04	0.13	0.09	0.80
3	31	0.01	0.01	0.01	0.07	0.04	0.58
	Out-degree Centrality						
Cluster	Size	Min	Q1	Median	Mean	Q3	Max
1	413	0.00	0.00	0.01	0.02	0.02	0.70
2	70	0.01	0.02	0.05	0.17	0.14	2.30
3	31	0.00	0.00	0.00	0.00	0.00	0.02
	In-closeness Centrality						
Cluster	Size	Min	Q1	Median	Mean	Q3	Max
1	413	0.00195	0.00195	0.00195	0.00195	0.00195	0.00197
2	70	0.00195	0.00196	0.00197	0.00197	0.00199	0.00200
3	31	0.00198	0.00198	0.00200	0.00199	0.00200	0.00201
	Out-closeness Centrality						
Cluster	Size	Min	Q1	Median	Mean	Q3	Max
1	413	0.00195	0.00195	0.00195	0.00195	0.00195	0.00198
2	70	0.00195	0.00199	0.00200	0.00200	0.00202	0.00204
3	31	0.00195	0.00195	0.00195	0.00195	0.00195	0.00196
	Betweenness Centrality						
Cluster	Size	Min	Q1	Median	Mean	Q3	Max
1	413	0.0000000	0.0000000	0.0000000	0.0000011	0.0000000	0.0000381
2	70	0.0000000	0.0000000	0.0000057	0.0000856	0.0001390	0.0005977
3	31	0.0000000	0.0000000	0.0000000	0.0000057	0.0000000	0.0001142

The clustering of players using social network variables representing various measures of centrality results in three groups of sizes – 413, 70 and 31. The partitions formed are not balanced, with group 1 containing the vast majority of users, and groups 2 and 3 relatively smaller.

Group 2 has the highest in-degree and out-degree centrality, indicating that it contains actors that are both more prestigious (messages received) as well as central (messages sent). Overall, group 2 comprises the most prominent individuals of the player base in terms of social behaviour. Comparison of groups 1 and 3 shows that group 3 has slightly higher in-degree centrality, implying the presence of more prestigious users, whereas group 1 has relatively more out-degree centrality, indicating the presence of more central users. The groups do not show significant variations with respect to the measures of closeness and betweenness. The betweenness is found to be extremely low overall, with group 2 showing comparatively higher values, signifying that the actors or individuals in the network have almost negligible control or mediating power on the relations in the network.

8.2.4 Evaluation and Interpretation

Although validation statistics for cluster analysis of the social network variables indicate moderately high quality of grouping structure for observations (average Silhouette Width = 0.73, i.e. quite well grouped observations), the segmentation of users is mainly appreciable in terms of the in-degree and out-degree centrality measures, while the others show no noticeable variation between groups. This may be because only one type of social interactions between individuals is considered in this analysis, i.e. messages sent and received, or may be a result of the natural pattern of communication existing between eRepublik players.

Nevertheless, social network analysis and clustering techniques of network variables appear to be able to identify groups of prominent users in terms of their social behaviour. These players are valuable to the game studios for improving their viral growth. As stated by Knoke and Yang (2008), "direct contacts and more intensive interactions dispose entities to better information, greater awareness, and higher susceptibility to influencing or being influenced by others". Therefore, identifying and targeting socially active users in a way that enhances their gameplay will be beneficial to the overall advancement and commercial success of the game. Since these users are found to engage in communications with other players in addition to usual gameplay, they may be more likely to invite their friends and other people to the game, thereby increasing its customer base and eventually revenue.

9. Conclusion

This is the closing chapter of the thesis and its purpose is to compile all the information and knowledge acquired in the process of conducting this study and demonstrate how it achieves the goal of this research. It is followed by specifying the contributions of this research and ends with a discussion on some restrictions and advice on future work.

9.1 Addressing Research Aims

First and foremost, the questions and aims of the research are recapitulated. The approach that was employed to accomplish each aim and the associated findings from it are also summarised. The overall goal was to develop suitable data-driven methods to gain insight about consumer behaviour in online freemium games, with a view to providing recommendations for successful business in the freemium games industry. This approach was adopted in order to overcome weaknesses (Harrison et al, 2011) in research undertaken that included small scale experiments involving a few variables or survey and questionnaire based methods (Jennett et al., 2008, Schoenau-Fog, 2011, Yee 2006, Cole & Griffiths, 2007, Poels et. al., 2007, Brown & Cairns, 2004). Gameplay data from a particular game called eRepublik, which is representative of a typical online freemium multiplayer game, was used in all statistical analyses that were carried out in R.

The research questions and aims, and corresponding methods and findings are as follows –

1. What gameplay behaviours in online freemium games significantly predict increasing engagement amongst its users?

The aim was to develop techniques to explore and identify specific aspects of users' gameplay that cause engagement or not, which can then guide the creative design process of games in making the experience more appealing and enjoyable to its consumers thereby minimising attrition and enhancing revenue.

A multiple logistic regression model using the penalised likelihood approach was developed to model player engagement. Evaluation of model assumptions and predictive validity demonstrated fairly good fit and accuracy for assigning observations to the engaged and non-engaged categories. The variables significant in predicting users being engaged were – the rate at which they complete missions, amount of premium currency possessed, the number of virtual friends they have, the ratio of kills to hits in a virtual fight and the quantity of virtual resources (food and energy bars) utilised to enhance their in-game performance. Therefore, this procedure was able to investigate the gameplay of users and provide insight on the aspects that promoted engagement amongst players.

2. At what time points in the game progression are players most likely to defect and drop out and what causes this?

The aim was to develop a process to investigate at what stage players are most likely to abandon the game and thereby cease to be valuable, and the components of gameplay that induces this event. This is because, anticipating when certain players are about to drop out will enable developers to customise their game, targeting these players with assistance to overcome any obstacles in their progression, which may cause defection.

Survival analysis methods and Cox's proportional hazards modelling was adopted to analyse time to customer defection. Survival curves using Kaplan-Meier estimates and estimates from the fitted Cox's proportional hazards model demonstrated decreasing probabilities for survival of users with the passage of time, and also showed the survival probabilities at specific time points in their gameplay. The overall robustness of the fitted Cox's model was moderate and it revealed the variables mission completion rate of users, their energy that was restored by virtual products such as food, premium currency gold owned by players, virtual friends they have, virtual items like energy bars used and their paying status as being statistically significant in affecting the hazard rate or conversely survival time of customers. Therefore, this technique was able to examine survival data of users and compute probabilities of their survival (and conversely risk of defection) at different time points, while also determining what components of gameplay affect the risk of dropping out at specific time points i.e. time to customer churn.

3. What facets of the player experience promote an increase in the quantity of real currency micro purchases by players?

The goal was to examine existing purchasing trends of customers in online freemium games (operating on micropayment revenue model), thereby establishing an approach to determine the incentives for real currency transactions that will benefit ARPU

Preliminary analyses found that 12% of player transactions involve real currency in €, 86% involve premium in-game currency gold and 2% involve loyalty points. An overwhelming majority of the total number of users did not make a purchase at all and of those that did, approximately 200 made only 1 payment. More than 90% of payers spend less than €50 during the entire course of their gameplay. Negative binomial-logit hurdle and zero inflated negative binomial models were developed to model the number of real currency purchases made by customers, of which the zero-inflated model demonstrated better fit to the data. The variables that contributed significantly to the number of (€)

purchases were the amount of premium currency possessed, the ratio of kills to hits in a virtual fight, completion rates in missions and quantity of energy bars used. Thus, this approach scrutinised the real currency transaction tendencies of customers and was able to develop a model that analysed factors which motivated players to indulge in purchases within the game environment thereby positively affecting the revenue generation of the product.

4. What are the different categories of players that constitute the user base of freemium games in terms of playing styles, performance within the game and revenue generation?

The aim was to produce a method for identifying the wide variety of players that constitute the user base of online freemium games, including further scrutiny of these classes with respect to their playing pattern, performance, and value added in terms of proceeds and virality.

Cluster analysis techniques and social network analysis were adopted to classify the existing player base into different groups based on their gameplay behaviour and social behaviour. The hierarchical Ward's minimum variance and the partitioning *k*-medoids methods of clustering were used and it was found that they performed similarly. Based on these algorithms, four different groups of users were established and their characteristics concerning general gameplay, performance and competence, engagement and advancement, and monetisation were studied. The groups identified in essence were – a) the engaged, competent in missions, achievers & high paying, b) the slow progressors, competent in fights, low achievers & moderately high paying, c) the moderately engaged, moderate competence in missions, moderate achievers & moderately low paying, and d) the moderately engaged, moderate competence overall, moderately low achievers & low paying. Social network analysis was used to investigate the network of users based on their in-game messaging patterns as well as cluster analysis of social network variables representing actor centrality. Three groups of players were obtained, and one group was clearly identified as that which included the most prominent actors of the social network. Thus, this method was able to develop a means to be able to identify the various types of individuals that constitute the user base of online freemium games and explain their playing styles, monetisation habits and socialisation. This could aid in the development of a better game product that is more customised to the preferences and styles of the different types of players constituting its customer base. It would contribute to an enhanced gaming experience for users and therefore increase their likelihood of both engagement and monetisation within the game, making it a more lucrative product.

9.2 Significance of the Research

This research has demonstrated that statistical methods can be applied to online data to develop an understanding of player behaviour. This allows generation of information that is more understandable and actionable compared to “black box” approaches such as neural networks and machine learning. Also, it is more reliable information as the data is generated from actual gameplay rather than out of game surveys or artificial laboratory experiments. One of the objectives of developing suitable data-driven methods to gain insight about consumer behaviour was to be able to provide recommendations for successful business in the freemium games industry. Thus, the usefulness of this research study lies in its ability to build a statistical framework for analysis of customer behaviours in online freemium games with a view to improving profitability and popularity of the product.

9.2.1 Recommendations for Analysis of Online Gameplay Data

The statistical methods outlined here can be considered as a guide for the analytics divisions of game studios to employ a more sophisticated procedure for evaluating the determinants of the performance of their games in terms of user engagement and monetisation, which goes beyond basic descriptive statistics. The analysis framework includes the following steps that game publishers can implement at each step, which may then allow them to enhance their product and customise it to suit the tastes of their users, thereby providing them with a more enjoyable experience.

- First, a set of variables depicting the gameplay patterns of players are computed. This should be an iterative process, wherein the variable values are updated at regular intervals of time as players progress through the game.

- Preliminary analyses to be conducted to confirm the existing gameplay patterns in terms of the events being triggered that would indicate the aspects of the game that players are enjoying more, performance and competency of users in missions, tasks and other competitive areas of the game, the amount of time (in days, hours, minutes etc.) that players are investing in the game and so on. This would also aid in identifying customers that do not contribute to the analysis at all by virtue of them not connecting with the game right away.

- This is followed by an attempt to understand user engagement, adopting the methods used in this study. Engagement is defined in an appropriate way and modelled using a multiple logistic regression model. After developing the most suitable model using the existing player base, it can help understand the factors influencing engagement, as well as make future predictions for players to be engaged or not. The ones that are observed to be non-engaged can then be targeted in terms of the variables significant in affecting engagement, and methods implemented to motivate engagement in them.

- Similarly, time until customers defect from the game is studied employing the procedures detailed in this research. Players that have defected from the game are identified and survival times defined. This is followed with an evaluation of survival probabilities at different time points of interest using non-parametric (Kaplan-Meier) and semi-parametric (Cox's proportional hazards model) methods. The Cox's model will be able to highlight the risk factors associated with customers quitting the game (i.e. stopping use of the product), and can also be used to predict the survival probabilities of users at any given time point of interest, thereby identifying those at the highest risk of defection at those time points. Additionally, performing survival analysis at different periods in time, for example after a major patch or character rework, may provide deeper insight into player defection. Once this is done, remedial measures can be taken in terms of the model significant variables, to prevent these players from dropping out.

- Approaches to customer monetisation are then undertaken as elucidated in this study. Real currency transactions are identified, depending on the pricing structure of the game, and primary investigations of the distributions of the number and amount of real currency transactions as well as popular items of purchase carried out. Hurdle and zero inflated models are employed to assess the features of gameplay that motivate the number of payments made by users. These features are then improved for all users in general, so that there can be a positive impact on customers inspire them to invest monetarily in the product.

- Finally, cluster analytic and social network methods are performed on gameplay related and social behaviour related variables, to scrutinise the existing customer base with a view to classifying them into distinct groups. A combination of hierarchical and partitioning algorithms as demonstrated in this research can be similarly adopted to form clusters of users, whose playing styles can then be described using summary statistics. Social behavioural variables recorded by the game can also be used to build networks of players that communicate with each other in various aspects (in-game messaging being an example used in this study). Measures of centrality extracted from the social network analysis can then be used in a clustering procedure to further categorize players.

Overall, this research has found that classical statistical techniques and modelling approaches work in investigating and explaining the vital measures associated with popularity (in terms of user engagement) and profitability (in terms of revenue generation) of online freemium games. The methods are practical and feasible to be applied with real world data arising from the games industry. The solutions, in terms of variable significance and future predictions are also pragmatic and useful in generating remedies to improve engagement and monetisation and reduce attrition.

9.2.2 Recommendations for Strategies for Game Design

The insights gained from implementing the analysis framework described above will equip the developers and publishers to design strategies for an improved game and gaming experience that will favour its customers. Some recommendations that the research suggests are –

- provisions of hints and tips to weak and struggling users to overcome missions and tasks,
- rewarding players not engaged with the game or at a risk of defection with premium virtual products that may increase or hold their degree of immersion,
- allowing paid customers to access advanced features of the game to give them to give them a snippet of more exciting missions and storylines, and
- motivating non-paying customers to make their first purchase through offering heavy discounts on extremely premium items that will greatly benefit the quality of their gameplay, while at the same time not going overboard with the promotions.

9.3 Research Contributions

The main contribution of this research is towards the online freemium games industry and its business. As explained in detail, it offers extensive insight into what drives the reputation, virality and commercial viability of gaming platforms, what methods can be adopted to analyse vast amounts of gameplay data and finally some recommendations on steps that can be taken to improve business in this sector. This can help game studios generate more revenue, which in turn will contribute to an improvement in the global economy as the games industry is seen as a growth sector. As highlighted by Harrison & Roberts (2011), "predictive models of player behaviour in video games is an open research topic that is receiving increasing attention in the literature" (p.1).

Additionally, there is an academic contribution of the research to developing methods for statistical analysis and modelling using similar data sets that represent customer behaviours in other e-commerce apps and online platforms such as Google, Amazon, Facebook, Netflix, etc. It has a methodological contribution to processing and analysing data sets that are typically highly skewed with long right tails, contain duplicate records and null/missing values. Therefore, it adds to the armoury of tools for analysis and modelling of online customer behavioural data and increase reliability of inference from it. The research has refined statistical models and demonstrated how they can be applied in R to new data types and complex areas to allow insight to be made.

Finally, the research is able to demonstrate that all analyses, modelling and predictions can be developed and implemented in the R programming language, thereby establishing how open access software can be successfully applied in a commercial setting. This is an important contribution in demonstrating the versatility of R and presents methods which can be implemented by game designers. It points to the possibility of incorporating R in game code to allow instantaneous real time analysis and in game autonomous decision making to enhance player experiences.

9.4 Limitations and Future Work

Although the research here focused on data obtained from one game, eRepublik, other online freemium games have been investigated, but because of data access issues and commercial confidentiality these have not been reported. Overall, the methodology was found to be generalisable, however more work is needed on establishing this.

The research only reports and explains classical frequentist statistical techniques that are implementable to real world data and have been able to produce sensible results, having tested certain alternatives. Future analysis can be carried out using alternative approaches such as Bayesian analysis and modelling. These can then be compared with the frequentist methods developed here to verify which ones are most appropriate and can be easily adopted by game studios in practice. There is also scope to develop approaches to allow those with little statistical background to understand the models and the methods used, and this points to the need to develop visualisations of the data analysis.

The social network analysis incorporated social behaviours concerning sending and receiving in-game messages only. This can be further improved by collecting and including other variables that are indicators of social interactions between players. Additionally, the model of number of micro payments made by users could not be validated regarding its ability to predict future data points using a holdout sample technique. This was because of the sparse distribution of payers in the sample considered for analysis. It can be overcome by using a longer period of data collection and analysis that may result in more payers in the data set, allowing for partitioning into test and training samples, and therefore predictive validation.

In the process of preserving ethical considerations of the research, no data on customer demographics is collected and used. This can be a limitation to the generalisability of the research in terms of player behaviour of different genders, age groups, cultures etc. Future work is needed to investigate into these dimensions of online customer behaviours, albeit only if it is able to secure informed consent of the participants.

References

Aalen, O., Borgan, O., & Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View.* Springer Science & Business Media.

Abbasi, A. Z., Ting, D. H., & Hlavacs, H. (2017). Engagement in Games: Developing an Instrument to Measure Consumer Videogame Engagement and its Validation. *International Journal of Computer Games Technology*, 2017.

Adamic, L. A., & Huberman, B. A. (2002). Zipf's Law and the Internet. *Glottometrics, 3*(1), 143-150.

Adler, R., Feldman, R., & Taqqu, M. (Eds.). (1998). *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Boston: Birkhauser.

Aguilera, A. M., Escabias, M., & Valderrama, M. J. (2006). Using Principal Components for Estimating Logistic Regression with High-dimensional Multicollinear data. *Computational Statistics & Data Analysis, 50*(8), 1905-1924.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Proceeding of the Second International Symposium on Information Theory, B.N. Petrov and F. Caski, eds., Akademiai Kiado, Budapest*, 267-281.

Albert, A., & Anderson, J. A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika, 71*(1), 1-10.

Alha, K., Koskinen, E., Paavilainen, J., Hamari, J., & Kinnunen, J. (2014). Free-to-play games: Professionals' perspectives. *Proceedings of nordic digra*.

Allison, P. D. (2012). *Logistic Regression using SAS: Theory and Application* (2nd ed.). North Carolina: SAS Institute Inc.

Allison, P. D. (2014). *Event History and Survival Analysis: Regression for Longitudinal Event Data* (Vol. 46). Sage.

Alsén, A., Runge, J., Drachen, A., & Klapper, D. (2016). Play With Me? Understanding and Measuring the Social Aspect of Casual Gaming. *Player Analytics: Papers from the AIIDE Workshop AAAI Technical Report WS-16-23*, 115-121.

Anderberg, M. R. (2014). *Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*. New York, London: Academic press.

Anderson, C. A., & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological science, 12*(5), 353-359.

Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of personality and social psychology, 78*(4), 772.

Appshopper.com. (2014). *AppShopper*. Retrieved from http://appshopper.com/

Arora, P., & Varshney, S. (2016). *Procedia Computer Science, 78*, 507-512.

Babbie, E. R. (2015). *The Practice of Social Research* (14th ed.). Cengage Learning.

Bakkes, S. C., Spronck, P. H., & van Lankveld, G. (2012). Player behavioural modelling for video games. *Entertainment Computing, 3*(3), 71-79.

Bartle, R. (1996). Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD research, 1*(1), 19.

Bartle, R. A. (2004). *Designing virtual worlds*. New Riders.

Bateman, C., & Boon, R. (2005). *21st Century Game Design (Game Development Series)*. Charles River Media, Inc.

Bauckhage, C., Drachen, A., & Sifa, R. (2015). Clustering game behavior data. *IEEE Transactions on Computational Intelligence and AI in Games, 7*(3), 266-278.

Bauckhage, C., Kersting, K., Sifa, R., Thurau, C., Drachen, A., & Canossa, A. (2012). How players lose interest in playing a game: An empirical study based on distributions of total playing times. *Computational Intelligence and Games (CIG), 2012 IEEE conference*, 139-146.

BCPS. (2017, June). *Develop a Research Proposal - Planning the Methodology - The Quantitative Pathway*. Retrieved from https://www.bcps.org/offices/lis/researchcourse/develop_writing_method_quantitative.html

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine, 24*(11), 1713-1723.

Ben-Gal, I. (2005). Outlier Detection. In: Maimon O., Rokach L. (eds) *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in medicine*, 2(2), 273-277.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In: Kogan J., Nicholas C., Teboulle M. (eds) *Grouping Multidimensional Data*. Berlin, Heidelberg: Springer.

Berkson, J. (1944). Application of the Logistic Function to Bio-assay. *Journal of the American Statistical Association, 39*(227), 357-365.

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics Review 12: Survival Analysis. *Critical care, 8*(5), 389.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics Review 14: Logistic Regression. *Critical Care, 9*(1), 112.

Bi, Z., Faloutsos, C., & Korn, F. (2001). The DGX Distribution for Mining Massive, Skewed Data. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 17-26.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). *Analyzing Social Networks* (2nd ed.). London: Sage.

Boyle, P., Flowerdew, R., & Williams, A. (1997). Evaluating the Goodness of Fit in Models of Sparse Medical Data: A Simulation Approach. *International Journal of Epidemiology, 26*(3), 651–656.

Brasington, R. (1990). Nintendinitis. *The New England journal of medicine, 322*(20), 1473.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*(3), 199-231.

Bridges Jr, C. C. (1966). Hierarchical Cluster Analysis. *Psychological Reports, 18*(3), 851-854.

Brito, P. Q., Soares, C., Almeida, S., Monte, A., & Byvoet, M. (2015). Customer Segmentation in a Large Database of an Online Customized Fashion Business. *Robotics and Computer-Integrated Manufacturing, 36*, 93-100.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2011). clValid, an R Package for Cluster Validation. *Journal of Statistical Software (Brock et al., March 2008)*.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and its Posterior Distribution. *20th International Conference on*

Pattern Recognition, Istanbul, 2010, 3121-3124. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.717.8158&rep=rep1&type=pdf

Brown, E., & Cairns, P. (2004). A Grounded Investigation of Game Immersion. *Human Factors in Computing Systems*, 1297-1300.

Brown, R. B. (2006). *Doing your dissertation in business and management: the reality of researching and writing*. Sage.

Bryman, A., & Bell, E. (2015). *Business Research Methods* (4th ed.). Oxford: Oxford University Press.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*(1), 252-268.

Cabrera, A. F. (1994). Logistic Regression Analysis in Higher Education: An Applied Perspective. *Higher education: Handbook of theory and research, 10*, 225-256.

Caetano, R. G. F. (2017). *Main drivers for microtransactions as impulse purchases in e-commerce* (Doctoral dissertation).

Cai, E. (2013). *Exploratory Data Analysis: Kernel Density Estimation – Conceptual Foundations*. Retrieved from https://chemicalstatistician.wordpress.com/2013/06/09/exploratory-data-analysis-kernel-density-estimation-in-r-on-ozone-pollution-data-in-new-york-and-ozonopolis/

Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data* (2nd ed.). New York: Cambridge University Press.

Caplan, S., Williams, D., & Yee, N. (2009). Problematic Internet use and psychosocial well-being among MMO players. *Computers in human behavior, 25*(6), 1312-1319.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. (2007). I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 1-14.

Chen, K. T., Huang, P., & Lei, C. L. (2009). Effect of network quality on player departure behavior in online games. *IEEE Transactions on Parallel and Distributed Systems, 20*(5), 593-606.

Chen, S. (2005). *Freeware vs Shareware vs Open Source*. Retrieved from http://opensourcestrategies.blogspot.co.uk/2005/09/freeware-vs-shareware-vs-open-source.html

Cheung, C. M., Shen, X. L., Lee, Z. W., & Chan, T. K. (2015). Promoting sales of online games through customer engagement. *Electronic Commerce Research and Applications, 14*(4), 241-250.

Chikhani, R. (2016). *The History Of Gaming: An Evolving Community*. Retrieved from https://techcrunch.com/2015/10/31/the-history-of-gaming-an-evolving-community/

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association, 74*(368), 829-836.

Cleveland, W. S., & Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association, 83*(403), 596-610.

Cleves, M. (2008). *An introduction to survival analysis using Stata* (2nd ed.). Stata Press.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cole, H., & Griffiths, M. D. (2007). Social interactions in massively multiplayer online role-playing gamers. *Cyberpsychology & behavior, 10*(4), 575-583.

Coleman, S., & Dyer-Witheford, N. (2007). Playing on the digital commons: collectivities, capital and contestation in videogame culture. *Media, culture & society, 29*(6), 934-953.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications, 34*(1), 313-327.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society – B, 34*, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika, 62*, 269-276.

Cox, N. J., & Jones, K. (1981). *Exploratory data analysis*. Retrieved from https://www.researchgate.net/profile/Kelvyn_Jones/publication/256802405_Exploratory_data_analysis/links/00463523c715d113ce000000.pdf

Cox, V. (2017). *Translating Statistics to Make Decisions: A Guide for the Non-Statistician*. Retrieved from https://link.springer.com/content/pdf/10.1007%2F978-1-4842-2256-0.pdf

Creasey, G. L., & Myers, B. J. (1986). Video games and children: Effects on leisure activities, schoolwork, and peer involvement. *Merrill-Palmer Quarterly, 32*(3), 251-262.

Creswell, J. W. (2009). *Research design: qualitative, quantitative, and mixed methods approaches* (3rd ed.). Sage.

Datta, P., Masand, B., Mani, D. R., & Li, B. (2000). Automated Cellular Modeling and Prediction on a Large Scale. *Artificial Intelligence Review (2000), 14*(6), 485–502.

Datta, S., Le-Rademacher, J., & Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics, 63*(1), 259-271.

Davidovici-Nora, M. (2014). Paid and free digital business models innovations in the video game industry. *Digiworld Economic Journal, 94,* 83-102.

De Lisi, R., & Wolford, J. L. (2002). Improving children's mental rotation accuracy with computer game playing. *The Journal of genetic psychology, 163*(3), 272-282.

De Prato, G., Feijóo, C., Nepelski, D., Bogdanowicz, M., & Simon, J.P. (2010). Born Digital / Grown Digital: Assessing the Future Competitiveness of the EU Video Games Software Industry. *Joint Research Centre, Institute for Prospective Technological Studies*. Retrieved from http://publications.jrc.ec.europa.eu/repository/handle/JRC60711

Demediuk, S., Murrin, A., Bulger, D., Hitchens, M., Drachen, A., Raffe, W. L., & Tamassia, M. (2018). Player retention in league of legends: a study using survival analysis. *Proceedings of the Australasian Computer Science Week Multiconference,* 43.

Dergousoff, K., & Mandryk, R. L. (2015, April). Mobile gamification for crowdsourcing data collection: Leveraging the freemium model. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems,* 1065-1074.

Desmarais, B. A., & Harden, J. J. (2013). Testing for Zero Inflation in Count Models: Bias Correction for the Vuong Test. *The Stata Journal, 13*(4), 810-835.

Digi-Capital. (2017, July). *Games software/hardware over $150B in 2017, $200B by 2021, record $2.8B invested*. Retrieved from http://www.digi-capital.com/news/2017/07/games-softwarehardware-over-150b-in-2017-200b-by-2021-record-2-8b-invested/#.Wl3k9Jq7JoB

Dixon, D. (2011). Player types and gamification. *Proceedings of the CHI 2011 Workshop on Gamification*.

Dorval, M., & Pepin, M. (1986). Effect of playing a video game on a measure of spatial visualization. *Perceptual and motor skills, 62*(1), 159-162.

Drachen, A., Canossa, A., & Yannakakis, G. N. (2009). Player modeling using self-organization in tomb raider: underworld. *2009 IEEE Symposium on Computational Intelligence and Games*, 1-8.

Drachen, A., Sifa, R., Bauckhage, C., & Thurau, C. (2012). Guns, swords and data: clustering of player behavior in computer games in the wild. *2012 IEEE Conference on Computational Intelligence and Games (CIG)*, 163-170.

Drachen, A., Thurau, C., Sifa, R., & Bauckhage, C. (2014). A comparison of methods for player clustering via behavioral telemetry. *arXiv:1407.3950*.

Drachen, A., Thurau, C., Togelius, J., Yannakakis, G., & Bauckhage, C. (2013). Game Data Mining. In El-Nasr, M.S., Drachen, A., & Canossa, A. (Eds.), *Game Analytics: Maximizing the Value of Player Data*, 205-253. London: Springer-VS.

Ducheneaut, N., & Moore, R. J. (2004). The social side of gaming: a study of interaction patterns in a massively multiplayer online game. *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work,* 360-369.

Ducheneaut, N., Yee, N., Nickell, E., & Moore, R. J. (2006). Alone together?: exploring the social dynamics of massively multiplayer online games. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 407-416.

Dudovskiy, J. (2016a). *Research Methodology: Positivism Research Philosophy*. Retrieved from https://research-methodology.net/research-philosophy/positivism/

Dudovskiy, J. (2016b). *Research Methodology: Deductive Approach (Deductive Reasoning)*. Retrieved from https://research-methodology.net/research-methodology/research-approach/deductive-approach-2/

Dudovskiy, J. (2016c). *Research Methodology: Deductive Approach (Deductive Reasoning)*. Retrieved from https://research-methodology.net/research-methods/

Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, *4*(1), 95-104.

Dutta, A. (2018). *What Is The Network Effect? Why Is It Valuable?*. Retrieved from https://www.feedough.com/network-effect/

Easy Guides. (2016). *Cox Proportional-Hazards Model*. Retrieved from https://www.r-bloggers.com/cox-proportional-hazards-model/

Edwards, B. (2012, January 22). *The 12 Greatest PC Shareware Games of All Time*. Retrieved from https://www.pcworld.com/article/248494/the_12_greatest_pc_shareware_games_of_all_time.html

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American statistical Association*, *72*(359), 557-565.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Egenfeldt-Nielsen, S., Smith, J. H., & Tosca, S. P. (2016). *Understanding video games: The essential introduction*. Devon, United Kingdom: Routledge.

El-Nasr, M.S., Drachen, A., & Canossa, A. (2013*). Game Analytics: Maximizing the Value of Player Data*. London: Springer-Verlag.

Engle, R. F. (1984). Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. *Handbook of Econometrics, 2*, 775-826.

ERA. (2018, January 03). *Streaming Boom Powers Entertainment Market To New All-Time-High Of £7.24bn In 2017*. Retrieved from https://eraltd.org/news-events/press-releases/2018/streaming-boom-powers-entertainment-market-to-new-all-time-high-of-724bn-in-2017/

eRepublik Official Wiki. (2015). *Loyalty Program*. Retrieved from https://wiki.erepublik.com/index.php/Loyalty_program eRepublik Official Wiki. (2018). *Welcome to eRepublik Encyclopedia, the Official On-line Encyclopedia of the Browser Game eRepublik*. Retrieved from https://wiki.erepublik.com/index.php/Main_Page

eRepublik. (2015). *EREPUBLIK*. Retrieved from https://www.erepublik.com/en

Evans, E. (2016). The economics of free: Freemium games, branding and the impatience economy. *Convergence*, *22*(6), 563-580.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th ed.). United Kingdom: John Wiley & Sons.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.

Fery, Y. A., & Ponserre, S. (2001). Enhancing the control of force in putting by video game training. *Ergonomics, 44*(12), 1025-1037.

Fields, T., & Cotton, B. (2011). *Social Game Design: Monetization Methods and Mechanics*. Morgan Kaufmann.

Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika, 80*(1), 27-38.

Fox, J., & Weisberg, S. (2018). *An R Companion to Applied Regressio*n (3rd ed.). Los Angeles: Sage Publications.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry, 40*(1), 35-41.

Freeman, L. C. (1978). Centrality in Social Networks: Conceptual Clarification. *Social Networks, 1*(3), 215-239.

Galloway, A. R. (2006). *Gaming: Essays on Algorithmic Culture (Electronic Mediations, Volume 18)*. Minneapolis: University of Minnesota Press.

GameSparks. ("n.d."). *Looking at In-Game Currencies*. Retrieved from https://www.gamesparks.com/blog/looking-at-in-game-currencies/

Gentile, D. A., Lynch, P. J., Linder, J. R., & Walsh, D. A. (2004). The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. *Journal of adolescence, 27*(1), 5-22.

Gillespie, B. W., & Mccullough, K. (2006). *Use of Generalized R-squared in Cox Regression. APHA Scientific Session and Event Listing*. Retrieved from https://aphanew.confex.com/apha/134am/techprogram/paper_135906.htm

Glatthorn, A. A., & Joyner, R. L. (2005). *Writing the Winning Thesis Or Dissertation: A Step-by-Step Guide* (2nd ed.). Corwin Press.

GoCompare. (2017, July 13*). UK adults spend almost £90 million* on 'freemium games'*. Retrieved from http://www.gocompare.com/press-office/2017/07/freemium-games/

Gomez, G., Julià, O., Utzet, F., & Moeschberger, M.L. (1992). Survival Analysis For Left Censored Data. In Klein, J. P., & Goel, P. K. (Eds), *Survival Analysis: State of the Art, Nato Science (Series E: Applied Sciences), 211*, 269-288. Dordrecht: Springer.

Gopal, R. K., & Meher, S. K. (2008, May). Customer churn time prediction in mobile telecommunication industry using ordinal regression. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 884-889.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*(6939), 534.

Green, M. S., & Symons, M. J. (1983). A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of chronic diseases, 36*(10), 715-723.

Griffith, J. L., Voloschin, P., Gibb, G. D., & Bailey, J. R. (1983). Differences in eye-hand motor coordination of video-game users and non-users. *Perceptual and motor skills, 57*(1), 155-158.

Guo, S., & Zeng, D. (2014). An overview of semiparametric models in survival analysis. *Journal of Statistical Planning and Inference*, 151-152, 1-16.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2006). Churn Prediction using Complaints Data. *World Academy of Science, Engineering and Technology, 13,* 158-163.

Hadiji, F., Sifa, R., Drachen, A., Thurau, C., Kersting, K., & Bauckhage, C. (2014). Predicting player churn in the wild. *Computational intelligence and games (CIG), 2014 IEEE conference*, 1-8.

Hahs-Vaughn, D. L., & Lomax, R. G. (2012). *An Introduction to Statistical Concepts: Third Edition* (3rd ed.). New York, NY: Routledge.

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine, 4*(2), 627.

Halim, Z., Atif, M., Rashid, A., & Edwin, C. A. (2017). Profiling players using real-world datasets: Clustering the data and correlating the results with the big-five personality traits. *IEEE Transactions on Affective Computing*.

Hamari, J., & Lehdonvirta, V. (2010). *Game design as marketing: How game mechanics create demand for virtual goods*.

Harrell Jr, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the Yield of Medical Tests. *Jama*, *247*(18), 2543-2546.

Harrell Jr., F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (2nd ed.). Switzerland: Springer.

Harris, M. B., & Williams, R. (1985). Video games and school performance. *Education, 105*(3).

Harrison, B., & Roberts, D. L. (2011). Using Sequential Observations to Model and Predict Player Behavior. *Proceedings of the 6th International Conference on Foundations of Digital Games,* 91-98.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York, NY: John Wiley & Sons

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100-108.

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis (Sage University Paper Series on Qualitative Research Methods, Vol. 16)*. Newbury Park, CA: Sage.

Heeks, R. (2009). Understanding" gold farming" and real-money trading as the intersection of real and virtual economies. *Journal for Virtual Worlds Research, 2*(4).

Heinze, G., & Schemper, M. (2002). A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine, 21*(16), 2409-2419.

Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery, 2*(1), 9-37.

Hilbe, J. M. (2009). *Logistic Regression Models*. Florida: CRC Press.

Hilbe, J. M. (2014). *Modeling Count Data*. New York: Cambridge University Press.

Hindy, J. (2017). *2016 Recap: 90% of Google Play's Revenue came from Games (and More Fun Stats!)*. Retrieved from https://www.androidauthority.com/2016-recap-90-percent-google-play-revenue-gaming-fun-stats-743626/

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85-126.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*, (Vol. 398). New Jersey: John Wiley & Sons.

Hosmer, D. W., Lemeshow, S., & May, S. (2011). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data* (2nd ed.). New Jersey: John Wiley & Sons.

Hosmer, D. W., Taber, S., & Lemeshow, S. (1991). The importance of assessing the fit of logistic regression models: a case study. *American Journal of Public Health, 81*(12), 1630-1635.

Hou, H. T. (2012). Exploring the behavioral patterns of learners in an educational massively multiple online role-playing game (MMORPG). *Computers & Education, 58*(4), 1225-1233.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery, 2*(3), 283-304.

Huizinga, J. (2014). *Homo Ludens Ils 86* (Vol. 3). Routledge.

Hum, S. (2014, January 20). *The psychology of freemium games – lessons and insights that can improve your business*. Retrieved from https://www.referralcandy.com/blog/how-psychology-behind-freemium-games-can-improve-business/

Hung, J. (2010). Economic essentials of online publishing with associated trends and patterns. *Publishing Research Quarterly*, *26*(2), 79-95.

Iterable. (2017). *The 5 stages of user engagement in mobile gaming*. Retrieved from https://iterable.com/blog/the-5-stages-of-user-engagement-in-mobile-gaming/

Jacobs, H. (2015, March 19). *Gaming guru explains why 'freemium' is actually the best business model for multiplayer video games*. Retrieved from http://uk.businessinsider.com/sean-plott-explains-why-he-thinks-freemium-games-are-the-best-business-model-for-both-players-and-developers-2015-3

Jansz, J., & Martens, L. (2005). Gaming at a LAN event: the social context of playing video games. *New media & society, 7*(3), 333-355.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London A, 186*(1007), 453-461.

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and Defining the Experience of Immersion in Games. *International Journal of Human-computer Studies, 66*(9), 641-661.

Jones, M. A., Mothersbaugh, D. L., & Beatty, S. E. (2000). Switching barriers and repurchase intentions in services. *Journal of retailing, 76*(2), 259-274.

Kallio, K. P., Mäyrä, F., & Kaipainen, K. (2011). At least nine ways to play: Approaching gamer mentalities. *Games and Culture, 6*(4), 327-353.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association, 53*(282), 457-481.

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. STHDA.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. *Statistical Data Analysis based on the L1 Norm. Y. Dodge, Ed*, 405-416.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis* (99th ed.). New Jersey: John Wiley & Sons.

Kawale, J., Pal, A., & Srivastava, J. (2009). Churn Prediction in MMORPGs: A Social Influence Based Approach. *2009 International Conference on Computational Science and Engineering*, 423-428.

Kayne, R. (2017, December 29). *What is Freeware?*. Retrieved from http://www.wisegeek.org/what-is-freeware.htm

Kharrat, T., & Boshnakov, G. N. (2017). *Model Selection and Comparison*. Retrieved from https://cran.r-project.org/web/packages/Countr/vignettes/ModelSelectionAndComparison.pdf

King, D. L., Gainsbury, S. M., Delfabbro, P. H., Hing, N., & Abarbanel, B. (2015). Distinguishing between gaming and gambling activities in addiction research. *Journal of Behavioral Addictions, 4*(4), 215-220.

Kirman, B., & Lawson, S. (2009). Hardcore classification: Identifying play styles in social games using network analysis. *International Conference on Entertainment Computing*, 246-251.

Kleinbaum, D. G., & Klein, M. (2011). *Survival Analysis: A Self-Learning Text* (3rd ed.). New York: Springer

Klimberg, R., & McCullough, B. D. (2017). *Fundamentals of Predictive Analytics with JMP* (2nd ed.). Cary: SAS Institute.

Knoke, D., & Yang, S. (2008). *Social Network Analysis* (2nd ed.). California: Sage.

Köhn, H. F., & Hubert, L. J. (2015). Hierarchical Cluster Analysis. *Wiley StatsRef: Statistics Reference Online*, 1-13.

Kracklauer, A. H., Mills, D. Q., & Seifert, D. (2004). Customer Management as the Origin of Collaborative Customer Relationship Management. In: Kracklauer A.H., Mills D.Q., Seifert D. (eds) *Collaborative Customer Relationship Management*. Berlin, Heidelberg: Springer.

Kuss, D. J., Louws, J., & Wiers, R. W. (2012). Online gaming addiction? Motives predict addictive play behavior in massively multiplayer online role-playing games. *Cyberpsychology, Behavior, and Social Networking, 15*(9), 480-485.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174.

Lee, E. T., & Wang, J. (2003). *Statistical Methods for Survival Data Analysis* (3rd ed.). New Jersey: John Wiley & Sons.

Lee, M. L., Lu, H., Ling, T. W., & Ko, Y. T. (1999). Cleansing Data for Mining and Warehousing. In Bench-Capon. T.J., Soda. G., & Tjoa. A.M. (Eds.), *Database and Expert Systems Applications (Lecture notes in computer science, 1677*, 751-760.

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis, 50*(4), 1113-1130.

Lemos, R. (2000, April 26). *Nintendo Issues Game Gloves*. Retrieved from https://www.gamespot.com/articles/nintendo-issues-game-gloves/1100-2541755/

Lieberman, D. A., Chaffee, S. H., & Roberts, D. F. (1988). Computers, mass media, and schooling: Functional equivalence in uses of new media. *Social Science Computer Review, 6*(2), 224-241.

Ling, R., & Yen, D. C. (2001). Customer Relationship Management: An Analysis Framework and Implementation Strategies. *Journal of Computer Information Systems, 41*(3), 82-97.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. *2010 IEEE International Conference on Data Mining*, 911-916.

Livingstone, S., & Markham, T. (2008). The contribution of media consumption to civic participation. *The British journal of sociology, 59*(2), 351-371.

Lohr, S. (2012). The Age of Big Data. *New York Times*, *11*(2012).

Lu, J. (2002). Predicting customer churn in the telecommunications industry—An application of survival analysis modeling using SAS. *SAS User Group International (SUGI27) Online Proceedings*, 114-27.

Luban, P. (2011). *The Design of Free-To-Play Games: Part 1*. Retrieved from https://www.gamasutra.com/view/feature/134920/the_design_of_freetoplay_games_.php?page=2

Luban, P. (2012). *The Design of Free-to-Play Games, Part 2*. Retrieved from https://www.gamasutra.com/view/feature/134959/the_design_of_freetoplay_games_.php

Luton, W. (2013). *Free-to-Play: Making Money From Games You Give Away*. New Riders.

Lynch, P. J. (1999). Hostility, Type A behavior, and stress hormones at rest and after playing violent video games in teenagers. *Psychosomatic Medicine, 61*(1), 113.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1*(14), 281-297.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Retrieved from https://nlp.stanford.edu/IR-book/html/htmledition/centroid-clustering-1.html

Marchand, A., & Hennig-Thurau, T. (2013). Value creation in the video game industry: Industry economics, consumer benefits, and research opportunities. *Journal of Interactive Marketing, 27*(3), 141-157.

Maulik, U., & Bandyopadhyay, S. (2002). Performance Evaluation of Some Clustering Algorithms and Validity Indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(12), 1650-1654.

Mavri, M., & Ioannou, G. (2008). Customer switching behaviour in Greek banking services using survival analysis. *Managerial Finance, 34*(3), 186-197.

McCulloh, I., Armstrong, H., & Johnson, A. (2013). *Social Network Analysis with Applications*. New Jersey: John Wiley & Sons.

McDowell, A. (2003). From the Help Desk: Hurdle Models. *The Stata Journal, 3*(2), 178-184.

Medler, B. (2011). Player dossiers: Analyzing gameplay data as a reward. *Game Studies, 11*(1).

Medler, B., John, M., & Lane, J. (2011). Data Cracker: Developing a Visual Game Analytic Tool for Analyzing Online Gameplay. *SIGCHI Conference on Human Factors in Computing Systems*, 2365-2374.

Menard, S., & Menard, S. W. (2010). Logistic Regression: From Introductory to Advanced Concepts and Applications. California: Sage.

Metz, C. E. (1986). ROC Methodology in Radiologic Imaging. *Investigative Radiology, 21*(9), 720-733.

Miller Jr, R. G. (2011). *Survival analysis* (Vol. 66). John Wiley & Sons.

Miller, P. (2012, March 7). *GDC 2012: How Valve made Team Fortress 2 free-to-play.* Retrieved from http://www.gamasutra.com/view/news/164922/GDC_2012_How_Valve_made_Team_Fortress_2_freetoplay.php

Mills, M. (2011). *Introducing Survival and Event History Analysis*. London: SAGE Publications.

Moore, D. F. (2016). *Applied Survival Analysis Using R*. Switzerland: Springer.

Morik, K., & Köpcke, H. (2004). Analysing customer churn in insurance data–a case study. *Knowledge Discovery in Databases: PKDD 2004*, 325-336.

Mueller-Veerse, F., Vocke, J., Vaidyanathan Rohini, D., & Malatinska, I. (2011). *Online, social, and mobile: The future of the video games industry. Cartagena-capital.* Retrieved from http://www.cartagena-capital.com/news-and-events/news/256-online-social-and-mobilethe-future-of-the-video-games-industry

Müller, M. (2004). Goodness-of-fit Criteria for Survival Data. *Sonderforschungsbereich 386, Paper 382*. Retrieved from https://epub.ub.uni-muenchen.de/1752/1/paper_382.pdf

Mutanen, T., Ahola, J., & Nousiainen, S. (2006). Customer churn prediction - a case study in retail banking. *ECML/PKDD Workshop on Practical Data Mining*, 13–19.

Myers, M. H., Hankey, B. F., & Mantel, N. (1973). A logistic-exponential model for use with response-time data involving regressor variables. *Biometrics*, 257-269.

Nemala, V. (2009). Efficient clustering techniques for managing large datasets. *UNLV Theses, Dissertations, Professional Papers, and Capstones*, 72.

Newcomb, T.M., (1961). *The Acquaintance Process*. New York: Holt, Rinehart and Winston.

Newman, M. E. (2003). The Structure and Function of Complex Networks. *SIAM Review*, *45*(2), 167-256.

Ng, K., & Liu, H. (2000). Customer Retention via Data Mining. *Artificial Intelligence Review (2000)*, *14*(6), 569–590.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification. *Expert Systems with Applications, 36*(2), 2592-2602.

Nielsen, U., Dahl, R., White, R. F., & Grandjean, P. (1998). Computer assisted neuropsychological testing of children. *Ugeskrift for laeger, 160*(24), 3557-3561.

Nikulin, M., & Wu, H. D. I. (2016). *The Cox Model and Its Applications*. Berlin Heidelberg: Springer.

Nosrati, M., Karimi, R., & Hariri, M. E. H. D. I. (2013). General trends in multiplayer online games. *World Applied Programming, 3*(1), 1-4.

O’brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity, 41*(5), 673-690.

O'Leary, Z. (2004). *The essential guide to doing research*. Sage.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths. *Social Networks 32*(3), 245-251.

Osathanunkul, C. (2015). A classification of business models in video game industry. *International Journal of Management Cases, 17*(1), 35-44.

Paavilainen, J., Hamari, J., Stenros, J., & Kinnunen, J. (2013). Social network games: Players’ perspectives. *Simulation & Gaming, 44*(6), 794-820.

Pahwa, A. (2017). *Freemium Business Model | The Psychology of Freemium*. Retrieved from https://www.feedough.com/freemium-business-model/

Pal, S. K., Ray, S. S., & Ganivada, A. (2017). *Granular Neural Networks, Pattern Recognition and Bioinformatics*. Springer.

Pampel, F. C. (2000). *Logistic Regression: A Primer* (Vol. 132). California: Sage.

Paschke, M. B., Green, E., & Gentile, D. A. (2001). The physiological and psychological effects of video games. In *Poster presented at the 36th Annual Minnesota Undergraduate Psychology Conference, St. Paul, MN.*

Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal of Computer Science and Applications, 6*(2), 256-261.

Periáñez, Á., Saas, A., Guitart, A., & Magne, C. (2016). Churn prediction in mobile social games: Towards a complete assessment using survival ensembles. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 564-573.

Piggott, J. (2015). Systematic Review of Gameplay Requirements for Massively Multiplayer Online Games. *10.13140/RG.2.1.3647.3761*.

Poels, K., De Kort, Y., & Ijsselsteijn, W. (2007). It is Always a Lot of Fun!: Exploring Dimensions of Digital Game Experience using Focus Group Methodology. *Proceedings of the 2007 Conference on Future Play*, 83-89.

Psychguides.com. (2018). *The Psychology of Freemium*. Retrieved from https://www.psychguides.com/interact/the-psychology-of-freemium/

Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research, 20*, 134-148.

Putnam, R. (1995). Bowling alone: America's declining social capital. *Journal of Democracy, 6*(1), 65-78.

R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. Retrieved from https://www.R-project.org/.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *Bulletin of the Technical Committee on Data Engineering, IEEE Computer Society, 23*(4), 3-13.

Rainey, C. (2016). Dealing with Separation in Logistic Regression Models. *Political Analysis, 24*(3), 339-355.

Ramanathan, T. R. (2009). *The role of organisational change management in offshore outsourcing of information technology services: Qualitative case studies from a multinational pharmaceutical company*. Universal-Publishers.

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. *Otolaryngology-Head and Neck Surgery, 143*(3), 331-336.

Roberts, D. F., Foehr, U. G., Rideout, V. J., & Brodie, M. (1999, November). *Kids and media @ the new millenium*. Menlo Park, CA: Kaiser Family Foundation. Retrieved from https://kaiserfamilyfoundation.files.wordpress.com/2013/01/kids-media-the-new-millennium-report.pdf

Rodríguez, G. (2005). *Non-Parametric Estimation in Survival Models*. Retrieved from http://data.princeton.edu/pop509/NonParametricSurvival.pdf

Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*. Retrieved from http://data.princeton.edu/wws509/notes/

Rosser, J. C., Lynch, P. J., Cuddihy, L., Gentile, D. A., Klonsky, J., & Merrell, R. (2007). The impact of video games on training surgeons in the 21st century. *Archives of surgery, 142*(2), 181-186.

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics, 20,* 53-65.

Royston, P. (2011). *Estimating a smooth baseline hazard function for the Cox model*. London: Department of Statistical Science, University College London.

Runge, J. (2017). *Social Gaming Was Never Just a Tactic for Viral Growth…* Retrieved from https://www.gamasutra.com/blogs/JulianRunge/20170124/289798/Social_Gaming_Was_Never_Just_a_Tactic_for_Viral_Growth.php

Saas, A., Guitart, A., & Periánez, A. (2016). Discovering playing patterns: Time series clustering of free-to-play game data. *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, 1-8.

Sabidussi, G. (1966). The Centrality Index of a Graph. *Psychometrika, 31*(4), 581-603.

Saitta, S., Raphael, B., & Smith, I. F. (2007). A Bounded Index for Cluster Validity. *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, 174-187.

Saldana, G. (2014). *Here are Some In-app Purchases that'll Break your Bank*. Retrieved from https://www.gamesradar.com/uk/most-expensive-in-app-purchases/

Sarkar, S. K., Midi, H., & Rana, S. (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences, 11*(1), 26-35.

SAS Institute, Inc. (1985). *SAS Users Guide: Statistics* (5th ed.). Cary, NC: SAS Institute.

Schatten, M., Tomičić, I., & Đurić, B. O. (2015). Multi-agent Modeling Methods for Massivley Multi-Player On-Line Role-Playing Games. *38th International Convention Mipro 2015*.

Schoenau-Fog, H. (2011). The Player Engagement Process-An Exploration of Continuation Desire in Digital Games. *Proceedings of DiGRA 2011 Conference: Think Design Play*. Retrieved from http://www.digra.org/wp-content/uploads/digital-library/11307.06025.pdf

Scott, J. (2017). *Social Network Analysis* (4th ed.). Croydon: Sage.

Shah, S., Horne, A., & Capellá, J. (2012). Good data won't guarantee good decisions. *Harvard Business Review, 90*(4).

Shmueli, G. (2010). To Explain or to Predict?. *Statistical Science, 25*(3), 289-310.

Sifa, R., Bauckhage, C., & Drachen, A. (2014). The playtime principle: large-scale cross-games interest modeling. *Computational intelligence and games (CIG), 2014 IEEE conference*, 1-8.

Sokolova, M., & Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management, 45*(4), 427-437.

Solidoro, A. (2009, April). The evolution of the creative industries as a model of innovation. *Proceedings of the 10th Workshop di Organizzazione Aziendale, 29th–30th April, Cagliari, Italy*.

Sotamaa, O. (2005). "Have Fun Working with Our Product!": Critical Perspectives On Computer Game Mod Competitions. In *DiGRA Conference*.

Springate, D. (2014). *Introduction to Survival Analysis in R*. Retrieved from https://rpubs.com/daspringate/survival

Stahl, T. (2005). *Video Game Genres*. Retrieved from http://www.thocp.net/software/games/reference/genres.htm

Stockburger, D. (2001). *Introductory Statistics: Concepts, Models, and Application*s (2nd ed.). Atomic Dog Pub.

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine, 18*(10), 1099-1104.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *Data and Knowledge Engineering, 25*(1), 161-198.

SuperData. (2017, November 21). *Battlefront II goofed, but gamers are still spending more on additional content.* Retrieved from https://www.superdataresearch.com/battlefront-ii-goofed-but-its-the-future/

Swets, J. A. (1979). ROC Analysis Applied to the Evaluation of Medical Imaging Techniques. *Investigative Radiology, 14*(2), 109.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*. USA: Addison-Wesley Longman.

Tarng, P. Y., Chen, K. T., & Huang, P. (2008). An analysis of WoW players' game hours. *Proceedings of the 7th ACM SIGCOMM Workshop on Network and System Support for Games*, 47-52.

Thompson, C. (2007). Halo 3: How Microsoft labs invented a new science of play. *Wired Magazine, 15*(9). Retrieved from http://josquin.cti.depaul.edu/~rburke/courses/f07/gam224/doc/science_of_play.pdf

TIGA. (2018). *About the UK Video Games Industry*. Retrieved from http://tiga.org/about-tiga-and-our-industry/about-uk-video-games-industry

Trenite, D. G. A., Silva, A. M., Ricci, S., Binnie, C. D., Rubboli, G., Tassinari, C. A., & Segers, J. P. (1999). Video-game epilepsy: A European study. *Epilepsia, 40*(s4), 70-74.

Tseng, F. C. (2011). Segmenting online gamers by motivation. *Expert Systems with Applications, 38*(6), 7693-7697.

Tukey, J. W. (1977). *Exploratory data analysis* (18th ed.). Reading, PA: Addison-Wesley.

UKIE. (2017). *The games industry in numbers*. Retrieved from https://ukie.org.uk/research

Van Lankveld, G., Schreurs, S., & Spronck, P. (2009). Psychologically Verified Player Modelling. *GAMEON*, 12-19.

Van Schie, E. G. M., & Wiegman, O. (1997). Children and videogames: Leisure activities, aggression, social integration, and school performance. *Journal of applied social psychology, 27*(13), 1175-1194.

Vandewater, E. A., Shim, M. S., & Caplovitz, A. G. (2004). Linking obesity and activity level with children's television and video game use. *Journal of adolescence, 27*(1), 71-85.

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston: Duxbury Press.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses. *Econometrica: Journal of the Econometric Society*, 307-333.

Walliman, N. (2011). *Research methods: The basics*. Oxon: Routledge.

Walsh, D. (2000). Testimony submitted to the United States Senate Committee on Commerce, Science, and Transportation. *Hearing on the impact of interactive violence on children*.

Walter, H., Vetter, S. C., Grothe, J. O., Wunderlich, A. P., Hahn, S., & Spitzer, M. (2001). The neural correlates of driving. *Neuroreport, 12*(8), 1763-1767.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*(301), 236-244.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440.

Whitson, J. R., & Dormann, C. (2011). Social gaming for change: Facebook unleashed. *First Monday, 16*(10).

Wijman, T. (2017, November 28). *New Gaming Boom: Newzoo Ups Its 2017 Global Games Market Estimate to $116.0Bn Growing to $143.5Bn in* 2020. Retrieved from https://newzoo.com/insights/articles/new-gaming-boom-newzoo-ups-its-2017-global-games-market-estimate-to-116-0bn-growing-to-143-5bn-in-2020/

Willekens, F. J. (1991). Life table analysis of staging processes. In Becker, H. A. (Ed), *Life Histories and Generations, ISOR, University of Utrecht, 2*, 477-518.

Williams, D. (2006, January). Why game studies now? *Games and Culture, 1*(1), 1-4.

Winkelmann, R. (2008). *Econometric Analysis of Count Data* (5th ed.). Berlin Heidelberg: Springer.

Wohn, D. Y. (2014). Spending real money: purchasing patterns of virtual goods in an online social game. *SIGCHI Conference on Human Factors in Computing Systems*, 3359-3368.

Yee, N. (2006). Motivations for play in online games. *CyberPsychology & behavior, 9*(6), 772-775.

Yuji, H. (1996). Computer games and information-processing skills. *Perceptual and motor skills, 83*(2), 643-647.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software, 27*(8), 1-25.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record, 25*(2), 103-114.

Zhao, Y., & Karypis, G. (2002). Evaluation of Hierarchical Clustering Algorithms for Document Datasets. In *Proceedings of the eleventh international conference on Information and knowledge management*, 515-524.

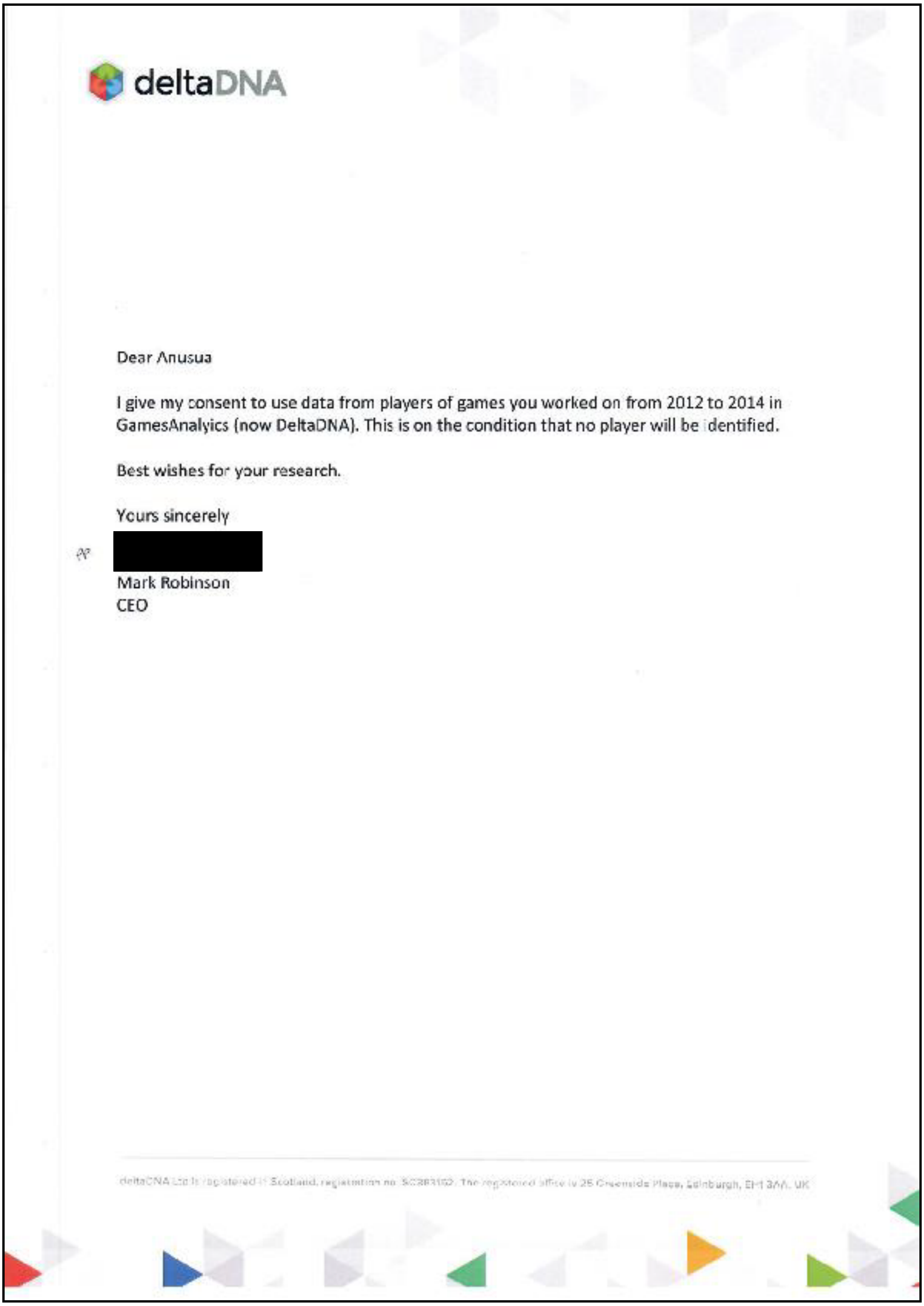
Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005). Customer churn prediction using improved one-class support vector machine. *Advanced data mining and applications*, 731-731.

Zhong, Z. J. (2011). The effects of collective MMORPG (Massively Multiplayer Online Role-Playing Games) play on gamers' online and offline social capital. *Computers in human behavior, 27*(6), 2352-2363.

Zikopoulos, P. C., Eaton, C., DeRoos, D., Deutsch, T., & Lapis, G. (2012). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. New York: Mcgraw-hill.

Zorn, C. (2005). A Solution to Separation in Binary Response Models. *Political Analysis, 13*(2), 157-170.

Appendix A: Permission Letter for Use of Data



deltaDNA

Dear Anusua

I give my consent to use data from players of games you worked on from 2012 to 2014 in GamesAnalyics (now DeltaDNA). This is on the condition that no player will be identified.

Best wishes for your research.

Yours sincerely

[redacted]

Mark Robinson
CEO

deltaDNA Ltd is registered in Scotland, registration no. SC383152. The registered office is 25 Greenside Place, Edinburgh, EH1 3AA, UK

Appendix B: R Code for Analysis and Modelling

```
#Load packages

library(data.table)

library(ggplot2)

library(pscl)

library(stargazer)

library(plyr)

library(gridExtra)

library(logistf)

library(brglm)

library(ROCR)

library(splines)

library(car)

library(Biobas)

library(caret)

library(Rcpp)

library(factoextra)

library(fpc)

library(dbscan)

library(cluster)

library(NbClust)

library(clValid)

library(igraph)

library(survival)

library(survminer)

library(caTools)

library(pec)

library(MASS)
library(COUNT)

library(pscl)

library(lmtest)


options(scipen=99,digits=5)


-----------------------------------------------------------------------------------------------------

# Read and view data

events<-read.csv("D:/Temp Data/events_erep_20140107.csv",header=F)

View (events[1:100,])


-----------------------------------------------------------------------------------------------------

# Drop unwanted variables

events<-
subset(events,select=c(V1,V2,V3,V5,V6,V7,V8,V9,V11,V13,V15,V16,V17,V18,V19,
V20,V21,V22,V24,V25,V26,V27,V31,V33,V35,V36,V37,V38,V39,

V40,V41,V42,V43,V45,V49,V51,V52,V53,V54,V55,V56,V57,V58,V59,V61,V63,V64
,V65,V69,V70,V71,V72,V73,V74,V75,V76,V77,V78,V79,V80,V81,V82,V83,V84,

V85,V87,V88,V89,V90,V91,V92,V93,V94,V95,V96,V97,V98,V99,V101,V102,V103,
V104,V105,V106,V107,V108,V109,V110,V111,V112,V113,V114,V115,V116,V117,

V118,V119,V120))


-----------------------------------------------------------------------------------------------------

# Assign names to variables

names(events)<-
c("esEventID","userID","segmentName","eventTimestamp","eventName","eventLevel"
,"msSinceLastEvent","userEventSequence",

"userSessionSequence","userRevenueEventSequence","UIAction","UIName","UIType"
,"achievementID","achievementName","achievements",
"acquisitionChannel","actionTaken","bazookaDamage","bazookasUsed","bigBombsUsed","bombsDamage","clicksPerRecoverEnergy","combatOrderRevenue",
"convertedProductAmount","damageForPatriotMedal","damageInBattle","damageInCampaign","division","doubleEnergyBarsUsed","V41",
"energyRestoredByEB","energyRestoredByFood","eventID","V49","firstRegistered","foodQ1Used","foodQ2Used","foodQ3Used","foodQ4Used",
"foodQ5Used","foodQ6Used","foodQ7Used","friendsCount","gold","V63","hitsCount","inviteType","isInviteAccepted","isResistance","isTutorial",
"killsCount","level","levelUpName","loyaltyLevel","mainEventID","militaryRank","militaryStrength","missionID","missionName",
"nationalCurrency","noWeaponHits","numberOfFights","parentEventID","placeVisited","productAmount","productCategory","productID","productName",
"productType","rankPoints","recipientID","recipientUserID","referrer","V96","rewardName","rocketsDamage","rocketsUsed","senderID","sessionID",
"V103","singleEnergyBarsUsed","smallBombsUsed","transactionID","transactionName","transactionType","transactionVector","transactorID",
"uniqueTracking","weaponDamage","weaponsQ1Used","weaponsQ2Used","weaponsQ3Used","weaponsQ4Used","weaponsQ5Used","weaponsQ6Used",
"weaponsQ7Used","xp")


-----------------------------------------------------------------------------------------------------

#  Convert timestamps

events$eventTimestamp<-strptime(events$eventTimestamp,"%d%b%Y %H:%M:%S")

events$firstRegistered<-strptime(events$firstRegistered,"%Y-%m-%d %H:%M:%S")


-----------------------------------------------------------------------------------------------------

# Search for duplicate events and remove (optional) if required

duplicate_events<-events[duplicated(events[c("userID","eventTimestamp","eventName","eventLevel","mainEventID")]),]
```

```
View(table(duplicate_events$msSinceLastEvent,useNA="ifany"))
View(duplicate_events[is.na(duplicate_events$msSinceLastEvent),])
duplicate_events<-duplicate_events[which(duplicate_events$msSinceLastEvent==0),]
View(table(duplicate_events$eventName))

-----------------------------------------------------------------------------------------------------
# Random selection of few events
temp<-
events[,c("userID","eventTimestamp","eventName","userEventSequence","actionTaken
","killsCount","level","nationalCurrency")]
temp<-temp[sample(nrow(temp),25),]

-----------------------------------------------------------------------------------------------------
# Create new variables
events$eventDate<-rep(NA,length(events$userID))
events$eventDate<-as.Date(events$eventTimestamp)

events$firstRegisteredDate<-rep(NA,length(events$userID))
events$firstRegisteredDate<-as.Date(events$firstRegistered)

-----------------------------------------------------------------------------------------------------
# Check for new players
new_players<-unique(events$userID[which(events$eventName=="newPlayer")])
setdiff(new_players, general_metrics$userID)
setdiff(general_metrics$userID,new_players)

-----------------------------------------------------------------------------------------------------
# User level metrics - General stats
```

```
general_metrics<-
data.table(events)[,list(firstSeen=min(eventDate),lastSeen=max(eventDate),firstRegistered=min(firstRegisteredDate,na.rm=T),
daysPlayed=length(unique(eventDate)),numEvents=max(userEventSequence),numSessions=max(userSessionSequence),timePlayed=sum(as.numeric(
msSinceLastEvent),na.rm=T)),by=userID]

general_metrics$timePlayed<-general_metrics$timePlayed/1000
general_metrics$timePlayed_mins<-general_metrics$timePlayed/60

general_metrics$sinceLastPlayed<-rep(NA,length(general_metrics$userID))
general_metrics$sinceLastPlayed<-as.Date("2014-01-06")- general_metrics$lastSeen

View(table(general_metrics$firstSeen))
View(table(general_metrics$lastSeen))

min(general_metrics$firstSeen)
max(general_metrics$lastSeen)
sum(general_metrics$numEvents)
length(unique(events$userID))
max(general_metrics$firstSeen)
min(events$eventDate)
max(events$eventDate)

dates<-data.frame(general_metrics$sinceLastPlayed,general_metrics$daysPlayed)
dates_frequency<-table(dates)
write.csv(dates_frequency,"Retention Matrix.csv")
```

```
general_metrics<-as.data.frame(general_metrics)


-----------------------------------------------------------------------------------------------------

# User level metrics - Gameplay stats

events$killHitRatio<-rep(NA,length(events$userID))

events$killHitRatio<-events$killsCount/events$hitsCount


events$energyRestoredByEB[which(events$energyRestoredByEB<0)]<-
abs(events$energyRestoredByEB[which(events$energyRestoredByEB<0)])

events$energyRestoredByFood[which(events$energyRestoredByFood<0)]<-
abs(events$energyRestoredByFood[which(events$energyRestoredByFood<0)])



gameplay_metrics<-
data.table(events)[,list(achievements=length(unique(achievementID[!is.na(achievement
ID)])),actionsTaken=length(actionTaken[

!is.na(actionTaken)]),bazookasUsed=sum(bazookasUsed,na.rm=TRUE),averageClicksP
erRecoverEnergy=mean(clicksPerRecoverEnergy,na.rm=TRUE),

averageDamageInBattle=mean(damageInBattle,na.rm=TRUE),damageInCampaign=su
m(as.numeric(damageInCampaign),na.rm=TRUE),

averageEnergyRestoredByFood=mean(energyRestoredByFood,na.rm=TRUE),averageE
nergyRestoredByEB=mean(energyRestoredByEB,na.rm=TRUE),averageGold=

mean(gold,na.rm=TRUE),averageHits=mean(hitsCount,na.rm=TRUE),friends=max(frie
ndsCount,na.rm=TRUE),averageKills=mean(killsCount,na.rm=TRUE),

averageKillHitRatio=mean(killHitRatio,na.rm=TRUE),level=max(level,na.rm=T),milita
ryRank=max(militaryRank,na.rm=TRUE),averageNationalCurrency=

mean(nationalCurrency,na.rm=TRUE),averageWeaponHits=mean(noWeaponHits,na.r
m=TRUE),totalNumberOfFights=sum(numberOfFights,na.rm=TRUE),

averageNumberOfFights=mean(numberOfFights,na.rm=TRUE),averageSingleEBUsed
=mean(singleEnergyBarsUsed,na.rm=TRUE),averageWeaponDamage=mean(
weaponDamage,na.rm=TRUE)),by=userID]

gameplay_metrics<-merge(gameplay_metrics,general_metrics[,c(1,11)],by="userID")

gameplay_metrics$relative_friends<-
gameplay_metrics$friends/gameplay_metrics$timePlayed_mins

gameplay_metrics<-as.data.frame(gameplay_metrics)

temp<-gameplay_metrics[,c(1,2,10,12,14,15,16,17,18,19)]
temp<-temp[sample(nrow(temp),25),]

-----------------------------------------------------------------------------------------------------
# User level metrics - Mission stats
View(table(events$missionName,exclude=NULL))
View(table(events$missionID,exclude=NULL))

mission_completion<-
data.table(events)[,list(numStarted=length(unique(userID[which(eventName=="mission
Started")])),numCompleted=length(unique(
userID[which(eventName=="missionCompleted")]))),by=missionName]

mission_metrics<-
data.table(events)[,list(numStarted=length(unique(missionName[which(eventName=="
missionStarted")])),numCompleted=length(
unique(missionName[which(eventName=="missionCompleted")]))),by=userID]

mission_metrics$missionCompletionRate<-
mission_metrics$numCompleted/mission_metrics$numStarted
```

```
mission_metrics$missionCompletionRate[which(mission_metrics$missionCompletionRate>1)]<-1

View(table(mission_metrics$numStarted,useNA="ifany"))
View(table(mission_metrics$numCompleted,useNA="ifany"))

mission_metrics<-as.data.frame(mission_metrics)

-----------------------------------------------------------------------------------------------------
# User level metrics - Transaction stats
transaction_events<-events[which(events$eventName=="transaction"),]
View(unique(transaction_events$userID))

transaction_metrics<-data.table(transaction_events)[,list(numPayments=length(unique(transactionID[which(productCategory=="REAL_CURRENCY" &
transactionVector=="SPENT")])),amountSpent=sum(productAmount[which(productCategory=="REAL_CURRENCY" & transactionVector=="SPENT")])),by=userID]


transaction_metrics<-merge(transaction_metrics,general_metrics[,c(1,8,10)],by="userID",all.y=T)
transaction_metrics<-merge(transaction_metrics,gameplay_metrics[,c(1,2,5,10,14,16,17,20)],by="userID",all.y=T)
transaction_metrics<-merge(transaction_metrics,mission_metrics[,c(1,4),with=FALSE],by="userID",all.y=T)
transaction_metrics[is.na(transaction_metrics)]<-0

transaction_metrics$amountSpent<-transaction_metrics$amountSpent/100
```

```
View(table(transaction_metrics$numPayments))

View(table(transaction_metrics$amountSpent))


transaction_metrics<-as.data.frame(transaction_metrics)


------------------------------------------------------------------------------------------------------

# Events Triggered

View(table(events$eventName,useNA="ifany"))


events_distribution<-
data.table(events)[,list(numPlayers=length(unique(userID))),by=eventName]


------------------------------------------------------------------------------------------------------

# Distribution of variables

plot(density(general_metrics$timePlayed_mins,adjust=0.2),xlim=c(0,11000),xlab="Tim
e Played (minutes)",main="Distribution of Gameplay Time")

summary(general_metrics$timePlayed_mins)

length(unique(general_metrics$userID[which(general_metrics$timePlayed_mins<1)]))

length(unique(general_metrics$userID[which(general_metrics$timePlayed_mins>240)]
))


plot(density(transaction_metrics$numPayments,adjust=0.2),xlim=c(0,10),ylim=c(0,0.5),
xlab="No. of Transactions",main="Distribution of Number of Transactions")

View(table(transaction_metrics$numPayments,useNA="ifany"))

summary(transaction_metrics$numPayments)


plot(density(gameplay_metrics$averageGold,adjust=0.2),xlim=c(0,200),xlab="Mean
Gold",main="Distribution of Amount of Gold Possessed")
```

```
View(table(gameplay_metrics$averageGold,useNA="ifany"))
summary(gameplay_metrics$averageGold)
length(unique(gameplay_metrics$userID[which(gameplay_metrics$averageGold<1)]))
length(unique(gameplay_metrics$userID[which(gameplay_metrics$averageGold==1)]))
length(unique(gameplay_metrics$userID[which(gameplay_metrics$averageGold>=1 &
gameplay_metrics$averageGold<=5)]))
length(unique(gameplay_metrics$userID[which(gameplay_metrics$averageGold>50)]))

plot(density(mission_metrics$missionCompletionRate[!is.na(mission_metrics$missionCompletionRate)],adjust=0.2),xlab="Mission Completion Rate",main=
"Distribution of Mission Completion Rates")
View(table(mission_metrics$missionCompletionRate,useNA="ifany"))
summary(mission_metrics$missionCompletionRate[which(mission_metrics$numStarted>0)])

-----------------------------------------------------------------------------------------------------
# Modelling Engagement
View(table(general_metrics$daysPlayed,useNA="ifany"))
summary(general_metrics$daysPlayed)

View(table(general_metrics$numEvents,useNA="ifany"))
summary(general_metrics$numEvents)
length(unique(general_metrics$userID[which(general_metrics$numEvents<10)]))
View(table(general_metrics$numEvents[which(general_metrics$daysPlayed==1)]))
summary(general_metrics$numEvents[which(general_metrics$daysPlayed==1)])


View(table(general_metrics$numSessions,useNA="ifany"))
```

```
summary(general_metrics$numSessions)
length(unique(general_metrics$userID[which(general_metrics$numSessions<5)]))
summary(general_metrics$timePlayed_mins[which(general_metrics$numSessions==1)])

ggplot(general_metrics,aes(x=daysPlayed,y=timePlayed_mins))+geom_point()+labs(y="Total Time (minutes)",x="Total Days")+ggtitle("Time Played vs Days Played")

cor.test(general_metrics$daysPlayed,general_metrics$timePlayed_mins,method="pearson")
cor.test(general_metrics$daysPlayed,general_metrics$timePlayed_mins,method="spearman")
cor.test(general_metrics$daysPlayed,general_metrics$timePlayed_mins,method="kendall")

VEarly_Lapsers<-general_metrics[which(general_metrics$numSessions==1),]
VEarly_Lapsers_events<-events[which(events$userID %in% VEarly_Lapsers$userID),]

summary(VEarly_Lapsers$daysPlayed)
summary(VEarly_Lapsers$numEvents)
summary(VEarly_Lapsers$timePlayed_mins)

View(table(VEarly_Lapsers_events$eventName,useNA="ifany"))
View(table(VEarly_Lapsers_events$missionName,useNA="ifany"))

summary(general_metrics$timePlayed_mins[general_metrics$userID %in% VEarly_Lapsers$userID])
View(table(mission_metrics$numStarted[which(mission_metrics$userID %in% VEarly_Lapsers$userID)],useNA="ifany"))
View(table(mission_metrics$missionCompletionRate[which(mission_metrics$userID
%in% VEarly_Lapsers$userID)],useNA="ifany"))

general_metrics$engagementStatus<-rep(NA,length(general_metrics$userID))
general_metrics$engagementStatus[which(general_metrics$daysPlayed>=7 &
general_metrics$sinceLastPlayed<=3)]<-1
general_metrics$engagementStatus[which(general_metrics$daysPlayed>=2 &
general_metrics$daysPlayed<=6 & general_metrics$timePlayed_mins>=300 &
general_metrics$sinceLastPlayed>7)]<-0
View(table(general_metrics$engagementStatus,useNA="ifany"))

lapsed<-general_metrics$userID[which(general_metrics$engagementStatus==0)]
engaged<-general_metrics$userID[which(general_metrics$engagementStatus==1)]

gameplay_metrics<-merge(gameplay_metrics,general_metrics[,c(1,10)],by="userID")
mission_metrics<-merge(mission_metrics,general_metrics[,c(1,10)],by="userID")

summary(general_metrics$daysPlayed[which(general_metrics$engagementStatus==0)])
summary(general_metrics$timePlayed_mins[which(general_metrics$engagementStatus
==0)])
summary(general_metrics$numEvents[which(general_metrics$engagementStatus==0)])
summary(mission_metrics$numStarted[which(mission_metrics$engagementStatus==0)
])
summary(mission_metrics$missionCompletionRate[which(mission_metrics$engageme
ntStatus==0 & mission_metrics$numStarted>0)])

plot1<-
ggplot(general_metrics[which(general_metrics$engagementStatus==0),],aes(x=timePla
yed_mins))+geom_density()+geom_vline(aes(xintercept=mean(
timePlayed_mins)),linetype="dashed",size=0.6)+xlab("Time Played
(minutes)")+ylab("Density")+ggtitle("Distribution of Gameplay Time")
plot2<-ggplot(mission_metrics[which(mission_metrics$engagementStatus==0 &
mission_metrics$numStarted>0),],aes(x=missionCompletionRate))+geom_density()+
geom_vline(aes(xintercept=mean(missionCompletionRate)),linetype="dashed",size=0.6
)+xlab("Mission Completion Rate")+ylab("Density")+ggtitle(
"Distribution of Mission Completion Rates")
grid.arrange(plot1,plot2,nrow=1,ncol=2)

View(table(events$actionTaken[which(events$userID %in% lapsed)],useNA="ifany"))


summary(general_metrics$daysPlayed[which(general_metrics$engagementStatus==1)])
summary(general_metrics$timePlayed_mins[which(general_metrics$engagementStatus
==1)])
summary(general_metrics$numEvents[which(general_metrics$engagementStatus==1)])
summary(mission_metrics$numStarted[which(mission_metrics$engagementStatus==1)
])
summary(mission_metrics$missionCompletionRate[which(mission_metrics$engageme
ntStatus==1 & mission_metrics$numStarted>0)])

plot1<-
ggplot(general_metrics[which(general_metrics$engagementStatus==1),],aes(x=timePla
yed_mins))+geom_density()+geom_vline(aes(xintercept=mean(
timePlayed_mins)),linetype="dashed",size=0.6)+xlab("Time Played
(minutes)")+ylab("Density")+ggtitle("Distribution of Gameplay Time")
plot2<-ggplot(mission_metrics[which(mission_metrics$engagementStatus==1 &
mission_metrics$numStarted>0),],aes(x=missionCompletionRate))+geom_density()+
geom_vline(aes(xintercept=mean(missionCompletionRate)),linetype="dashed",size=0.6
)+xlab("Mission Completion Rate")+ylab("Density")+ggtitle(
"Distribution of Mission Completion Rates")
grid.arrange(plot1,plot2,nrow=1,ncol=2)

View(table(events$actionTaken[which(events$userID %in%
engaged)],useNA="ifany"))



model_data<-general_metrics[,c(1,10)]
model_data<-merge(model_data,mission_metrics[,c(1,4)],by="userID")
model_data<-
merge(model_data,gameplay_metrics[,c(1,8,10,12,14,17,21,24)],by="userID")
model_data<-model_data[!is.na(model_data$engagementStatus),]

summary(model_data[,c(-1,-10)])

stargazer(model_data[,c(-1,-10)],type="text",title="Descriptive
Statistics",digits=2,covariate.labels=c("User Engagement","Mission Completion Rate",
"Average Energy Restored by Food","Average Premium Currency Gold","Number of
Friends","Average Kill:Hit Ratio","Average Grind Currency National Currency",
"Average Energy Bars
Used"),mean.sd=TRUE,min.max=TRUE,median=TRUE,iqr=TRUE)



model_data_final<-model_data

chk<-model_data$userID[is.na(model_data$missionCompletionRate)]
chk1<-general_metrics[which(general_metrics$userID %in% chk &
general_metrics$engagementStatus==0),]
mean(chk1$daysPlayed)
```

```
chk2<-general_metrics$userID[which(general_metrics$daysPlayed==2 &
general_metrics$engagementStatus==0)]

median(mission_metrics$missionCompletionRate[which(mission_metrics$userID
%in% chk2)],na.rm=TRUE)

model_data_final$missionCompletionRate[is.na(model_data_final$missionCompletion
Rate) & model_data_final$engagementStatus==0]<-0.7


chk<-model_data$userID[is.na(model_data$missionCompletionRate)]

chk1<-general_metrics[which(general_metrics$userID %in% chk &
general_metrics$engagementStatus==1),]

mean(chk1$daysPlayed)

chk2<-general_metrics$userID[which(general_metrics$daysPlayed==15 &
general_metrics$engagementStatus==1)]

median(mission_metrics$missionCompletionRate[which(mission_metrics$userID
%in% chk2)],na.rm=TRUE)

model_data_final$missionCompletionRate[is.na(model_data_final$missionCompletion
Rate) & model_data_final$engagementStatus==1]<-0.94


model_data_final$averageEnergyRestoredByFood[is.na(model_data_final$averageEner
gyRestoredByFood)]<-0


model_data_final$averageSingleEBUsed[is.na(model_data_final$averageSingleEBUse
d)]<-0


chk<-model_data$userID[is.na(model_data$averageKillHitRatio)]

chk1<-general_metrics[which(general_metrics$userID %in% chk &
general_metrics$engagementStatus==0),]

mean(chk1$daysPlayed)

chk2<-general_metrics$userID[which(general_metrics$daysPlayed==3 &
general_metrics$engagementStatus==0)]
```

```
median(gameplay_metrics$averageKillHitRatio[which(gameplay_metrics$userID %in% chk2)],na.rm=TRUE)

model_data_final$averageKillHitRatio[is.na(model_data_final$averageKillHitRatio) & model_data_final$engagementStatus==0]<-0.42


chk<-model_data$userID[is.na(model_data$averageKillHitRatio)]

chk1<-general_metrics[which(general_metrics$userID %in% chk & general_metrics$engagementStatus==1),]

mean(chk1$daysPlayed)

chk2<-general_metrics$userID[which(general_metrics$daysPlayed==19 & general_metrics$engagementStatus==1)]

median(gameplay_metrics$averageKillHitRatio[which(gameplay_metrics$userID %in% chk2)],na.rm=TRUE)

model_data_final$averageKillHitRatio[is.na(model_data_final$averageKillHitRatio) & model_data_final$engagementStatus==1]<-0.43



set.seed(100)

sample<-sample.split(model_data_final$userID,SplitRatio=0.75)

train_data<-subset(model_data_final,sample==TRUE)

test_data<-subset(model_data_final,sample==FALSE)


View(table(train_data$engagementStatus))

View(table(test_data$engagementStatus))


model<-glm(engagementStatus~missionCompletionRate+averageEnergyRestoredByFood+averageGold+friends+averageKillHitRatio+averageNationalCurrency+

averageSingleEBUsed,family=binomial(link='logit'),data=train_data)

summary(model)
```

```
model<-
logistf(engagementStatus~missionCompletionRate+averageEnergyRestoredByFood+av
erageGold+friends+averageKillHitRatio+averageNationalCurrency+
averageSingleEBUsed,data=train_data)
summary(model)
model$method
model$method.ci
add1(model)
drop1(model)
forward(model)
backward(model)
extractAIC(model)

model1<-
update(model,engagementStatus~missionCompletionRate+averageEnergyRestoredByF
ood+averageGold+friends+averageKillHitRatio+averageSingleEBUsed)
summary(model1)
anova(model,model1)
extractAIC(model1)

model1_alt<-
brglm(engagementStatus~missionCompletionRate+averageEnergyRestoredByFood+av
erageGold+friends+averageKillHitRatio+averageSingleEBUsed,family=
binomial(link='logit'),data=train_data)
summary(model1_alt)

logit<-predict(model1_alt,type="link")
```

```
plot1<-
ggplot(train_data,aes(missionCompletionRate,logit))+geom_point(size=0.5,alpha=0.5)+
geom_smooth(method="loess")+xlab("Mission Completion Rate")+ylab(
"Logit")
plot2<-
ggplot(train_data,aes(averageEnergyRestoredByFood,logit))+geom_point(size=0.5,alph
a=0.5)+geom_smooth(method="loess")+xlab(
"Average Energy Restored by Food")+ylab("Logit")
plot3<-
ggplot(train_data,aes(averageGold,logit))+geom_point(size=0.5,alpha=0.5)+geom_smo
oth(method="loess")+xlab("Average Premium Currency Gold")+ylab(
"Logit")
plot4<-
ggplot(train_data,aes(friends,logit))+geom_point(size=0.5,alpha=0.5)+geom_smooth(m
ethod="loess")+xlim(0,320)+xlab("Number of Friends")+ylab("Logit")
plot5<-
ggplot(train_data,aes(averageKillHitRatio,logit))+geom_point(size=0.5,alpha=0.5)+geo
m_smooth(method="loess")+xlab("Average Kill:Hit")+ylab("Logit")
plot6<-
ggplot(train_data,aes(averageSingleEBUsed,logit))+geom_point(size=0.5,alpha=0.5)+g
eom_smooth(method="loess")+xlab("Average Energy Bars used")+ylab(
"Logit")
grid.arrange(plot1,plot2,plot3,plot4,plot5,plot6,nrow=2,ncol=3)


n=8392
model1_alt_residuals<-rstandard(model1_alt,type="deviance")
length(unique(model1_alt_residuals[which(abs(model1_alt_residuals)>2)]))
model1_alt_residuals<-data.frame(model1_alt_residuals)
model1_alt_residuals$index=1:n
model1_alt_residuals<-cbind(model1_alt_residuals,train_data$engagementStatus)
```

```
names(model1_alt_residuals)<-c("Residual","Index","Engaged")
model1_alt_residuals$Engaged<-as.factor(model1_alt_residuals$Engaged)
ggplot(model1_alt_residuals,aes(Index,Residual))+geom_point(aes(color=Engaged),alpha=0.5)+theme_bw()+xlab("Index")+ylab("Deviance Residuals")+
ggtitle("Index Plot of Deviance Residuals")+scale_color_manual(values=c("0"="dark grey","1"="black"))

train_data1<-cbind(train_data,model1_alt_residuals)
train_data1<-train_data1[which(abs(train_data1$model1_alt_residuals)<=2),]
model_chk<-logistf(engagementStatus~missionCompletionRate+averageEnergyRestoredByFood+averageGold+friends+averageKillHitRatio+averageSingleEBUsed,data=
train_data1)
summary(model_chk)

vif(model1_alt)

model1_alt_pred<-predict(model1_alt,newdata=test_data,type="response")
model1_alt_pred<-ifelse(model1_alt_pred>0.8,1,0)

confmat<-confusionMatrix(data=as.factor(model1_alt_pred),reference=as.factor(test_data$engagementStatus),positive="1")
View(confmat$table)
confmat$positive
confmat$overall
confmat$byClass

model1_alt_pred2<-predict(model1_alt,newdata=test_data,type="response")
```

```
pr<-prediction(model1_alt_pred2,test_data$engagementStatus)

prf<-performance(pr,measure="tpr",x.measure="fpr")

plot(prf,main="ROC Curve for the Model Predicting User Engagement")


auc<-performance(pr,measure="auc")

auc<-auc@y.values[[1]]

auc


########################################################################

model2_alt<-
brglm(engagementStatus~missionCompletionRate+averageEnergyRestoredByFood+bs(
averageGold,knots=c(7,250))+friends+averageKillHitRatio+

averageSingleEBUsed,family=binomial(link='logit'),data=train_data)

summary(model2_alt)


hoslem.test(train_data$engagementStatus,fitted.values(model1_alt))


pR2(model1_alt)


bestlogisticmodel_data<-within(model_data_final,{

  userID<-NULL

  averageNationalCurrency_scaled<-NULL

  relative_friends<-NULL

  y<-engagementStatus

  engagementStatus<-NULL

  relative_friends<-NULL

}

)
```

```
model_bestglm<-bestglm(Xy=bestlogisticmodel_data,family=binomial,IC="AIC",method="exhaustive")

model_bestglm$BestModels

summary(model_bestglm$BestModel)

#############################################################################


--------------------------------------------------------------------------------------------------------------------

# Modelling Transactions

# Check user 8241590 for real money transaction

transaction_events2<-transaction_events[!(transaction_events$userID %in% VEarly_Lapsers$userID),]

#the above results in the same as transaction events and hence not needed


View(table(transaction_events$productName[which(transaction_events$transactionVector=="SPENT")]))


View(table(transaction_events$productName[which(transaction_events$transactionVector=="RECEIVED")]))


id<-transaction_events$transactionID[which(transaction_events$transactionVector=="SPENT" & transaction_events$productName=="loyalty points")]

View(table(transaction_events$productName[which(transaction_events$transactionVector=="RECEIVED" & transaction_events$transactionID %in% id)]))


transaction_metrics2<-transaction_metrics[!(transaction_metrics$userID %in% VEarly_Lapsers$userID),]

transaction_metrics2<-transaction_metrics2[,c(1:3)]
```

```
transaction_metrics2<-merge(transaction_metrics2,general_metrics[,c(1,10,11)],by="userID")
transaction_metrics2<-merge(transaction_metrics2,gameplay_metrics[,-c(23,24,25)],by="userID")
transaction_metrics2<-merge(transaction_metrics2,mission_metrics[,c(1,2,4)],by="userID")

View(table(transaction_metrics$numPayments))

summary(transaction_metrics2$timePlayed_mins[which(transaction_metrics2$numPayments>0)])

transaction_metrics2<-transaction_metrics2[which(transaction_metrics2$timePlayed_mins>74),]

transaction_metrics2$amountSpent<-transaction_metrics2$amountSpent/100

View(table(transaction_metrics2$numPayments))
View(table(transaction_metrics2$amountSpent))

sum(transaction_metrics2$numPayments)
sum(transaction_metrics2$amountSpent)


ggplot(transaction_metrics2,aes(numPayments))+geom_bar()+scale_x_continuous(breaks=0:20)+xlab("Number of Transactions")+ylab("Frequency")+ggtitle(
"Distribution of the Number of Micro Transactions (in EUROS)")
```

```
ggplot(transaction_metrics2,aes(amountSpent))+geom_histogram(binwidth=10)+xlab("Amount Spent")+ylab("Frequency")+ggtitle("Distribution of the Amount Spent (in EUROS) in Micro Transactions")
```

```
ggplot(transaction_metrics2,aes(x=amountSpent))+geom_density()+geom_vline(aes(xintercept=mean(amountSpent)),linetype="dashed",size=0.6)+ylim(0,0.006)+xlab("Amount Spent")+ylab("Density")+ggtitle("Distribution of the Amount Spent (in EUROS) in Micro Transactions")
```

```
summary(transaction_metrics2$numPayments)
summary(transaction_metrics2$amountSpent)
```

```
ggplot(transaction_metrics2[which(transaction_metrics2$numPayments>0),],aes(numPayments))+geom_bar()+xlab("Number of Transactions")+ylab("Frequency")+ggtitle("Distribution of the Number of Micro Transactions (in EUROS) of Payers Only")
```

```
ggplot(transaction_metrics2[which(transaction_metrics2$numPayments>0),],aes(amountSpent))+geom_histogram(binwidth=10)+xlab("Amount Spent")+ylab("Frequency")+ggtitle("Distribution of the Amount Spent (in EUROS) in Micro Transactions of Payers Only")
```

```
ggplot(transaction_metrics2[which(transaction_metrics2$numPayments>0),],aes(x=amountSpent))+geom_density()+geom_vline(aes(xintercept=mean(amountSpent)),linetype="dashed",size=0.6)+ylim(0,0.006)+xlab("Amount Spent")+ylab("Density")+ggtitle("Distribution of the Amount Spent (in EUROS) in Micro Transactions of Payers Only")
```

```
summary(transaction_metrics2$numPayments[which(transaction_metrics2$numPayments>0)])
summary(transaction_metrics2$amountSpent[which(transaction_metrics2$numPayments>0)])

set.seed(100)
sample<-sample.split(transaction_metrics2$userID,SplitRatio=0.75)
train_data<-subset(transaction_metrics2,sample==TRUE)
test_data<-subset(transaction_metrics2,sample==FALSE)

View(table(train_data$numPayments))
View(table(test_data$numPayments))



payments_null<-glm(numPayments~1,family="poisson",data=transaction_metrics2)
summary(payments_null)
# pchisq(6931.6,20853,lower.tail=FALSE)
qchisq(0.001,df=21309)
pearson_chisq<-sum(residuals(payments_null,type="pearson")^2)
dispersion<-pearson_chisq/payments_null$df.residual
modelfit(payments_null)
# print(pchisq(pearson_chisq,21309,lower.tail=FALSE),digits=15)
expected_values<-predict(payments_null,type="response")
expected_mean<-mean(predicted_values)

payments_poi<-glm(numPayments~averageEnergyRestoredByFood+averageGold+averageKillHitRatio
+averageNationalCurrency+averageSingleEBUsed+missionCompletionRate,
family="poisson",data=transaction_metrics2)
summary(payments_poi)
```

```
# pchisq(4325.4,21305,lower.tail=FALSE)
qchisq(0.001,df=19913)
pearson_chisq<-sum(residuals(payments_poi,type="pearson")^2)
dispersion<-pearson_chisq/payments_poi$df.residual
modelfit(payments_poi)
# print(pchisq(pearson_chisq,21309,lower.tail=FALSE),digits=15)
expected_values<-predict(payments_poi,type="response")
expected_mean<-mean(expected_values)

payments_nb<-glm.nb(numPayments~averageEnergyRestoredByFood+averageGold+averageKillHitRatio+averageNationalCurrency+averageSingleEBUsed+missionCompletionRate,
data=transaction_metrics2)
# payments_nb<-nbinomial(numPayments~averageEnergyRestoredByFood+averageGold+averageKillHitRatio+averageNationalCurrency+averageSingleEBUsed+
# missionCompletionRate,data=transaction_metrics2)
summary(payments_nb)
alpha<-1/payments_nb$theta
pearson_chisq<-sum(residuals(payments_nb,type="pearson")^2)
dispersion<-pearson_chisq/payments_nb$df.residual
modelfit(payments_nb)
# print(pchisq(pearson_chisq,21309,lower.tail=FALSE),digits=15)
expected_values_poi<-predict(payments_poi,type="response")
expected_variance<-mean(expected_values_poi)+(alpha*(mean(expected_values_poi)^2))
```

```
payments_hurdle<-
hurdle(numPayments~averageEnergyRestoredByFood+averageGold+averageKillHitRatio+averageSingleEBUsed+missionCompletionRate|
averageEnergyRestoredByFood+averageGold+averageKillHitRatio+averageNationalCurrency+averageSingleEBUsed+missionCompletionRate,
data=transaction_metrics2,dist="negbin",zero.dist="binomial",link="logit")
summary(payments_hurdle)
AIC(payments_hurdle)
BIC(payments_hurdle)

payments_hurdle2<-
hurdle(numPayments~averageGold+missionCompletionRate|averageGold+averageKillHitRatio+averageSingleEBUsed+missionCompletionRate,data=
transaction_metrics2,dist="negbin",zero.dist="binomial",link="logit")
summary(payments_hurdle2)

lrtest(payments_hurdle,payments_hurdle2)

payments_zeroinf<-
zeroinfl(numPayments~averageEnergyRestoredByFood+averageGold+averageKillHitRatio+averageSingleEBUsed+missionCompletionRate|
averageEnergyRestoredByFood+averageGold+averageKillHitRatio+averageNationalCurrency+averageSingleEBUsed+missionCompletionRate,data=transaction_metrics2,dist=
"negbin")
summary(payments_zeroinf)
AIC(payments_zeroinf)
BIC(payments_zeroinf)

vuong(payments_zeroinf,payments_nb)
```

```
payments_zeroinf2<-
zeroinfl(numPayments~averageGold+averageKillHitRatio+missionCompletionRate|ave
rageGold+averageKillHitRatio+averageSingleEBUsed+
missionCompletionRate,data=transaction_metrics2,dist="negbin")
summary(payments_zeroinf2)

lrtest(payments_zeroinf,payments_zeroinf2)
lrtest(payments_zeroinf2,payments_hurdle2)

#################################################################################

var(transaction_metrics_subset$numPayments)
summary(transaction_metrics_subset$numPayments)
summary<-stat.desc(transaction_metrics_subset$numPayments,basic=F)
stargazer(summary,type="html",title="Summary Statistics for Number of
Payments",digits=4,out="payments_summarytable.htm")



payments_hurdle2<-
hurdle(numPayments~achievements+averageGold+averageKillHitRatio+averageClicks
PerRecoverEnergy|achievements+averageGold+
averageKillHitRatio+missionCompletionRate+averageClicksPerRecoverEnergy,data=tr
ansaction_metrics_subset,dist="negbin",zero.dist="binomial",link=
"logit")
summary(payments_hurdle2)
AIC(payments_hurdle2)
BIC(payments_hurdle2)
```

```
payments_hurdle3<-hurdle(numPayments~achievements+averageGold+averageKillHitRatio+militaryRank|achievements+averageGold+averageKillHitRatio+

missionCompletionRate+averageClicksPerRecoverEnergy+militaryRank,data=transaction_metrics_subset,dist="negbin",zero.dist="binomial",link="logit")

summary(payments_hurdle3)

AIC(payments_hurdle3)

BIC(payments_hurdle3)


payments_hurdle4<-hurdle(numPayments~achievements+averageGold+averageKillHitRatio+averageNationalCurrency|averageGold+averageKillHitRatio+

averageClicksPerRecoverEnergy+militaryRank+averageNationalCurrency,data=transaction_metrics_subset,dist="negbin",zero.dist="binomial",link="logit")

summary(payments_hurdle4)

AIC(payments_hurdle4)

BIC(payments_hurdle4)


payments_hurdle5<-hurdle(numPayments~achievements+averageGold+averageKillHitRatio+averageNumberOfFights|averageGold+averageKillHitRatio+

averageClicksPerRecoverEnergy+militaryRank+averageNumberOfFights,data=transaction_metrics_subset,dist="negbin",zero.dist="binomial",link="logit")

summary(payments_hurdle5)

AIC(payments_hurdle5)

BIC(payments_hurdle5)


payments_hurdle6<-hurdle(numPayments~achievements+averageGold+averageKillHitRatio|averageGold+averageKillHitRatio+militaryRank+averageNumberOfFights
,data=transaction_metrics_subset,dist="negbin",zero.dist="binomial",link="logit")
summary(payments_hurdle6)
AIC(payments_hurdle6)
BIC(payments_hurdle6)

payments_zeroinf2<-
zeroinfl(numPayments~achievements+averageGold|achievements+averageGold+averag
eKillHitRatio,data=transaction_metrics_subset,
dist="negbin")
summary(payments_zeroinf2)

payments_zeroinf2<-
zeroinfl(numPayments~achievements+averageGold+scale(timePlayed)|achievements+a
verageGold+averageKillHitRatio+scale(
timePlayed),data=transaction_metrics_subset,dist="negbin")
summary(payments_zeroinf2)
AIC(payments_zeroinf2)
BIC(payments_zeroinf2)

payments_zeroinf3<-
zeroinfl(numPayments~achievements+averageGold+averageClicksPerRecoverEnergy|a
chievements+averageGold+averageKillHitRatio+
averageClicksPerRecoverEnergy,data=transaction_metrics_subset,dist="negbin")
summary(payments_zeroinf3)
AIC(payments_zeroinf3)
BIC(payments_zeroinf3)

pchisq(2*(logLik(payments_zeroinf3)-
logLik(payments_zeroinf2)),df=2,lower.tail=FALSE)
```

```
# payments_zeroinf3 is preferred

payments_zeroinf4<-zeroinfl(numPayments~achievements+averageGold+militaryRank|achievements+averageGold+averageKillHitRatio+averageClicksPerRecoverEnergy+militaryRank,data=transaction_metrics_subset,dist="negbin")
summary(payments_zeroinf4)
AIC(payments_zeroinf4)
BIC(payments_zeroinf4)

pchisq(2*(logLik(payments_zeroinf4)-logLik(payments_zeroinf3)),df=1,lower.tail=FALSE)
# payments_zeroinf4 is preferred

payments_zeroinf5<-zeroinfl(numPayments~achievements+averageGold+militaryRank+averageNationalCurrency|averageGold+averageKillHitRatio+averageClicksPerRecoverEnergy+militaryRank+averageNationalCurrency,data=transaction_metrics_subset,dist="negbin")
summary(payments_zeroinf5)
AIC(payments_zeroinf5)
BIC(payments_zeroinf5)

payments_zeroinf6<-zeroinfl(numPayments~achievements+averageGold+militaryRank+averageNumberOfFights|averageGold+averageKillHitRatio+averageClicksPerRecoverEnergy+militaryRank+averageNumberOfFights,data=transaction_metrics_subset,dist="negbin")
summary(payments_zeroinf6)
```

```
AIC(payments_zeroinf6)
BIC(payments_zeroinf6)

payments_zeroinf7<-
zeroinfl(numPayments~achievements+averageGold|averageGold+averageKillHitRatio+
militaryRank+averageNumberOfFights,data=
transaction_metrics_subset,dist="negbin")
summary(payments_zeroinf7)
AIC(payments_zeroinf7)
BIC(payments_zeroinf7)

pchisq(2*(logLik(payments_zeroinf6)-
logLik(payments_zeroinf7)),df=3,lower.tail=FALSE)
# payments_zeroinf7 is preferred

############################################################################
-------------------------------------------------------------------------------------------------------------------
# Classification of customer behaviours
cluster_data<-general_metrics[!(general_metrics$userID %in%
VEarly_Lapsers$userID),]
cluster_data<-cluster_data[,c(1,11)]
cluster_data<-merge(cluster_data,gameplay_metrics[,-c(23,24,25)],by="userID")
cluster_data<-merge(cluster_data,mission_metrics[,c(1,2,4)],by="userID")
cluster_data<-cluster_data[,-c(2,13)]

id<-general_metrics$userID[which(general_metrics$daysPlayed>=7)]

cluster_data<-cluster_data[which(cluster_data$userID %in% id),]
```

```
cluster_data2<-cluster_data[,-c(1,11,12,18)]
cluster_data2<-na.omit(cluster_data2)

stargazer(cluster_data2,type="text",title="Descriptive Statistics",digits=2,mean.sd=TRUE,min.max=TRUE,median=TRUE,iqr=TRUE)

maha_dist<-mahalanobis(cluster_data2,colMeans(cluster_data2),cov(cluster_data2),tol=8.74725e-21)

cluster_data_outliers<-cluster_data2
cluster_data_outliers$maha_dist<-round(maha_dist,1)
cluster_data_outliers$outlier<-"No"
cluster_data_outliers$outlier[cluster_data_outliers$maha_dist>20]<-"Yes"
View(table(cluster_data_outliers$outlier,useNA="ifany"))

cluster_data2_scaled<-scale(cluster_data2)

# cluster_data2<-cluster_data[,-c(1,2,3,4,6,7,9,11,12,14,15,16,18,21,22)]



d<-dist(cluster_data2_scaled,method="euclidean")


hierar_cluster<-hclust(d^2,method="centroid")
plot(hierar_cluster,cex=0.6,hang=-1,main="Centroid Linkage Clustering")

hierar_cluster<-hclust(d,method="ward.D2")
plot(hierar_cluster,cex=0.6,hang=-1,main="Ward's Method Clustering")
```

```
rect.hclust(hierar_cluster,k=4,border=2:5)
hierar_group<-cutree(hierar_cluster,k=4)
View(table(hierar_group))
fviz_cluster(list(data=cluster_data2_scaled,cluster=hierar_group))

hierar_cluster<-hclust(d,method="average")
plot(hierar_cluster,cex=0.6,hang=-1,main="Average Linkage Clustering")



set.seed(9)
kmedoids_cluster<-pam(cluster_data2_scaled,k=4,metric="euclidean")
View(kmedoids_cluster$medoids)
View(kmedoids_cluster$clusinfo)
View(kmedoids_cluster$objective)
fviz_cluster(kmedoids_cluster,cluster_data2_scaled,ellipse=TRUE,geom="point")

cluster_data2<-
cbind(cluster_data2,cluster=kmedoids_cluster$clustering,cluster2=hierar_group)
cluster_data2$equal<-rep(0,length(cluster_data2$cluster))
cluster_data2$equal[which(cluster_data2$cluster==1 & cluster_data2$cluster2==1)]<-1
cluster_data2$equal[which(cluster_data2$cluster==2 & cluster_data2$cluster2==4)]<-1
cluster_data2$equal[which(cluster_data2$cluster==3 & cluster_data2$cluster2==2)]<-1
cluster_data2$equal[which(cluster_data2$cluster==4 & cluster_data2$cluster2==3)]<-1
View(table(cluster_data2$equal))

fviz_silhouette(kmedoids_cluster)
plot(silhouette(hierar_group,d))
summary(silhouette(hierar_group,d))
```

```
kmedoids_stats<-cluster.stats(d,kmedoids_cluster$clustering)
kmedoids_stats$dunn2
hierar_stats<-cluster.stats(d,hierar_group)
hierar_stats$dunn2

kmedoids_stats$n
kmedoids_stats$cluster.size
kmedoids_stats$noisen
kmedoids_stats$diameter
kmedoids_stats$separation
kmedoids_stats$average.between
kmedoids_stats$average.within
kmedoids_stats$widestgap

hierar_stats$n
hierar_stats$cluster.size
hierar_stats$noisen
hierar_stats$diameter
hierar_stats$separation
hierar_stats$average.between
hierar_stats$average.within
hierar_stats$widestgap

hierar_sil<-silhouette(hierar_group,d)
hierar_sil<-matrix(hierar_sil,ncol=3)
hierar_sil<-data.frame(hierar_sil)
hierar_sil$hierar_clust<-rep(NA,length(hierar_sil$X1))
```

```
hierar_sil$hierar_clust[which(hierar_sil$X3>0)]<-hierar_sil$X1[which(hierar_sil$X3>0)]
hierar_sil$hierar_clust[which(hierar_sil$X3<0)]<-hierar_sil$X2[which(hierar_sil$X3<0)]

kmedoids_sil<-silhouette(kmedoids_cluster$clustering,d)
kmedoids_sil<-matrix(kmedoids_sil,ncol=3)
kmedoids_sil<-data.frame(kmedoids_sil)
kmedoids_sil$kmedoids_clust<-rep(NA,length(kmedoids_sil$X1))
kmedoids_sil$kmedoids_clust[which(kmedoids_sil$X3>0)]<-kmedoids_sil$X1[which(kmedoids_sil$X3>0)]
kmedoids_sil$kmedoids_clust[which(kmedoids_sil$X3<0)]<-kmedoids_sil$X2[which(kmedoids_sil$X3<0)]

cluster_data2<-cbind(cluster_data2,sil=kmedoids_sil$X3,sil2=hierar_sil$X3,kclust=kmedoids_sil$kmedoids_clust,hiclust=hierar_sil$hierar_clust)

cluster_data2$hiclust2<-rep(NA,length(cluster_data2$hiclust))
cluster_data2$hiclust2[which(cluster_data2$hiclust==1)]<-1
cluster_data2$hiclust2[which(cluster_data2$hiclust==2)]<-3
cluster_data2$hiclust2[which(cluster_data2$hiclust==3)]<-4
cluster_data2$hiclust2[which(cluster_data2$hiclust==4)]<-2

cluster_data2$equal2<-rep(0,length(cluster_data2$cluster))
cluster_data2$equal2[which(cluster_data2$kclust==cluster_data2$hiclust2)]<-1

cluster_data2$finalclust<-rep(NA,length(cluster_data2$cluster))
```

```
cluster_data2$finalclust[which(cluster_data2$equal2==1)]<-
cluster_data2$kclust[which(cluster_data2$equal2==1)]

cluster_data2$finalclust[which(cluster_data2$equal2==0 &
cluster_data2$sil>cluster_data2$sil2)]<-
cluster_data2$kclust[which(cluster_data2$equal2==0 &
cluster_data2$sil>cluster_data2$sil2)]

cluster_data2$finalclust[which(cluster_data2$equal2==0 &
cluster_data2$sil2>cluster_data2$sil)]<-
cluster_data2$hiclust2[which(cluster_data2$equal2==0 &
cluster_data2$sil2>cluster_data2$sil)]


cluster_data2<-cluster_data2[,-c(20,21,22,23,24,25,26,27,27)]


summary(cluster_data2$achievements[which(cluster_data2$finalclust==4)])
summary(cluster_data2$averageGold[which(cluster_data2$finalclust==4)])
summary(cluster_data2$averageKillHitRatio[which(cluster_data2$finalclust==4)])
summary(cluster_data2$militaryRank[which(cluster_data2$finalclust==4)])
summary(cluster_data2$averageNumberOfFights[which(cluster_data2$finalclust==4)])
summary(cluster_data2$averageNationalCurrency[which(cluster_data2$finalclust==4)]
)
summary(cluster_data2$averageSingleEBUsed[which(cluster_data2$finalclust==4)])
summary(cluster_data2$missionCompletionRate[which(cluster_data2$finalclust==4)])


cluster_data_tmp<-na.omit(cluster_data)
cluster_data_tmp<-
merge(cluster_data_tmp,transaction_metrics[,c(1,2,3)],by="userID",all.x=T)
cluster_data_tmp<-cluster_data_tmp[,c(1,24,25)]
```

```
cluster_data2<-cbind(cluster_data2,numPayments=cluster_data_tmp$numPayments,amountSpent=cluster_data_tmp$amountSpent)

sum(cluster_data2$numPayments[which(cluster_data2$finalclust==3)])
sum(cluster_data2$amountSpent[which(cluster_data2$finalclust==4)])

########################################################################
fviz_nbclust(cluster_data2_scaled,pam,method="silhouette")
gap_stat<-clusGap(cluster_data2_scaled,FUN=pam,K.max=10,B=50)
fviz_gap_stat(gap_stat)

res.agnes<-agnes(cluster_data2_scaled,diss=FALSE,method="average")
res.agnes$ac
pltree(res.agnes,cex=0.6,hang=-1,main="Dendrogram of agnes")

set.seed(9)
km_res<-kmeans(cluster_data2,4,nstart=25)
View(km_res$centers)
View(km_res$size)
fviz_cluster(km_res,cluster_data2,ellipse=TRUE,geom="point")

kNNdistplot(cluster_data2_scaled,k=4)
abline(h=2.8,lty=2)

set.seed(27)
db<-dbscan(cluster_data2_scaled,2.8,4)
db
```

```
set.seed(27)

hdb_cluster<-hdbscan(cluster_data2,minPts=5)

hdb_cluster

########################################################################


#Social network analysis

messages_sent<-events[which(events$eventName=="messageSent"),]

messages_sent<-messages_sent[,c("userID","recipientID")]

messages_sent<-messages_sent[!is.na(messages_sent$recipientID),]

length(unique(messages_sent$userID))


messages_received<-events[which(events$eventName=="messageReceived"),]

messages_received<-messages_received[,c("userID","senderID")]

messages_received<-messages_received[ ,c("senderID","userID")]

colnames(messages_received)<- c("userID","recipientID")

length(unique(messages_received$userID))


messages<-rbind(messages_sent,messages_received)

messages<-ddply(messages,.(userID,recipientID),nrow)

colnames(messages)<- c("userID","recipientID","weights")

messages<-messages[which(messages$userID %in% general_metrics$userID &
messages$recipientID %in% general_metrics$userID),]

messages<-messages[with(messages,order(userID,recipientID)), ]

messages<-messages[which(messages$userID!=messages$recipientID),]

messages<-messages[which(messages$weights>2),]


length(unique(messages$userID))
length(unique(messages$recipientID))
length(union(unique(messages$userID),unique(messages$recipientID)))

messages_graph<-graph.data.frame(messages,directed=TRUE)
messages_adj<-as_adjacency_matrix(messages_graph,sparse=FALSE,attr="weights")
messages_adj_graph<-
graph.adjacency(messages_adj,mode="directed",weighted=TRUE,diag=FALSE)
View(E(messages_adj_graph)$weight)
messages_adj_graph<-
simplify(messages_adj_graph,remove.multiple=F,remove.loops=T)
plot.igraph(messages_adj_graph,layout=layout.fruchterman.reingold,edge.width=0.6,ed
ge.arrow.size=0.6,edge.arrow.width=0.6,edge.color="black",vertex.label=NA,
vertex.size=8,main="A Social Network of Players Based on In-game Messages
Exchanged")

diameter(messages_adj_graph,directed=TRUE)
farthest_vertices(messages_adj_graph,directed=TRUE)

mean_distance(messages_adj_graph,directed=TRUE)

is_connected(messages_adj_graph)
communities<-components(messages_adj_graph,mode="weak")
communities$no
count_components(messages_adj_graph,mode="weak")
View(table(communities$membership))
View(communities$csize)
grps<-groups(communities)
View(grps)
```

```
sum(transaction_metrics$amountSpent[which(transaction_metrics$userID %in%
grps$`14`)])
sum(transaction_metrics$amountSpent[which(transaction_metrics$userID %in%
grps$`112`)])
sum(transaction_metrics$amountSpent[which(transaction_metrics$userID %in%
grps$`51`)])
sum(transaction_metrics$amountSpent)



messages_metrics<-data.frame(V(messages_adj_graph)$name,degree<-
strength(messages_adj_graph,mode="all"),indegree=strength(messages_adj_graph,mod
e="in"),
outdegree=strength(messages_adj_graph,mode="out"),closeness=closeness(messages_a
dj_graph,mode="all",normalized=TRUE),incloseness=closeness(messages_adj_graph,
mode="in",normalized=TRUE),outcloseness=closeness(messages_adj_graph,mode="ou
t",normalized=TRUE),btweenness=betweenness(messages_adj_graph,directed=TRUE,
nobigint=TRUE,normalized=TRUE))
colnames(messages_metrics)<-
c("userID","degree","indegree","outdegree","closeness","incloseness","outcloseness","b
etweenness")

messages_metrics$degree<-messages_metrics$degree/513
messages_metrics$indegree<-messages_metrics$indegree/513
messages_metrics$outdegree<-messages_metrics$outdegree/513

messages_metrics2<-messages_metrics[,c(3,4,6,7,8)]

set.seed(9)
kmedoids_messages<-pam(scale(messages_metrics2),3)
View(kmedoids_messages$medoids)
View(kmedoids_messages$clusinfo)

messages_metrics2<-cbind(messages_metrics2,cluster=kmedoids_messages$clustering)

summary(messages_metrics2$indegree[which(messages_metrics2$cluster==3)])
summary(messages_metrics2$outdegree[which(messages_metrics2$cluster==3)])
summary(messages_metrics2$incloseness[which(messages_metrics2$cluster==3)])
summary(messages_metrics2$outcloseness[which(messages_metrics2$cluster==3)])
summary(messages_metrics2$betweenness[which(messages_metrics2$cluster==3)])

d<-dist(scale(messages_metrics2),method="euclidean")
kmedoids_messages_stats<-cluster.stats(d,kmedoids_messages$clustering)
kmedoids_messages_stats$avg.silwidth
kmedoids_messages_stats$dunn2

#######################################################################

clust_infmap<-cluster_infomap(messages_adj_graph,e.weights=messages$weights)
modularity(clust_infmap)
par(mar=c(0,0,0,0));plot(clust_infmap,messages_adj_graph)
membership(clust_infmap)
length(communities(clust_infmap))

sna_invites<-events[which(events$eventName=="inviteReceived" & events$isInviteAccepted==1),]
sna_invites<-invites[,c("userID","senderID")]
sna_invites<-sna_invites[ , c("senderID","userID")]
colnames(sna_invites)<- c("userID","recipientID")
```

```
sna_invites$userID<-as.character(sna_invites$userID)
sna_invites$recipientID<-as.character(sna_invites$recipientID)
sna_invites<-ddply(sna_invites,.(userID,recipientID),nrow)
sna_invites<-sna_invites[which(sna_invites$V1==1),]
sna_invites<-sna_invites[which(sna_invites$userID %in% general_metrics$userID &
sna_invites$recipientID %in% general_metrics$userID),]
sna_invites$V1<-NULL

sna_invites_graph<-graph.data.frame(sna_invites,directed=TRUE)
sna_invites_adj<-as_adjacency_matrix(sna_invites_graph,sparse=FALSE,attr=NULL)
sna_invites_adj_graph<-
graph.adjacency(sna_invites_adj,mode="directed",weighted=NULL,diag=FALSE)
sna_invites_adj_graph<-
simplify(sna_invites_adj_graph,remove.multiple=F,remove.loops=T)
plot.igraph(sna_invites_adj_graph,layout=layout.fruchterman.reingold,edge.width=0.6,e
dge.arrow.size=0.6,edge.arrow.width=0.6,edge.color="black",
vertex.label=NA,vertex.size=8)

sna_invites_metrics<-
data.frame(V(sna_invites_adj_graph)$name,indegree=degree(sna_invites_adj_graph,mo
de="in"),outdegree=degree(
sna_invites_adj_graph,mode="out"),incloseness=closeness(sna_invites_adj_graph,mode
="in"),outcloseness=closeness(sna_invites_adj_graph,mode="out"),
btweenness=betweenness(sna_invites_adj_graph,directed=TRUE,nobigint=TRUE,norm
alized=TRUE))
colnames(sna_invites_metrics)<-
c("userID","indegree","outdegree","incloseness","outcloseness","btweenness")

set.seed(14)
kmedoids_sna_invites<-pam(sna_invites_metrics[,c(2,3,4,5,6)],5)
```

```
View(kmedoids_sna_invites$medoids)
View(kmedoids_sna_invites$clusinfo)


set.seed(3)
kmeans_messages<-kmeans(scale(messages_metrics2),3,nstart=25)
View(kmeans_messages$size)
View(kmeans_messages$centers)


kmeans_sna_messages$cluster<-as.factor(kmeans_sna_messages$cluster)
ggplot(sna_messages_metrics[,c(2,3,4,5,6)],aes(incloseness,outcloseness,color=kmeans_sna_messages$cluster))+geom_point()
ggplot(sna_messages_metrics[,c(2,3,4,5,6)],aes(indegree,outdegree,color=kmeans_sna_messages$cluster))+geom_point()


#######################################################################


------------------------------------------------------------------------------------------------------------
# Predicting time to defection
survival_data<-general_metrics[!(general_metrics$userID %in% VEarly_Lapsers$userID),]
survival_data<-general_metrics[which(general_metrics$timePlayed_mins>=300),]
survival_data$defectionStatus<-rep(0,length(survival_data$userID))
survival_data$defectionStatus[which(survival_data$sinceLastPlayed>14)]<-1


View(table(survival_data$defectionStatus,useNA="ifany"))


survival_data<-survival_data[,c(1,11,12)]
survival_data<-merge(survival_data,mission_metrics[,c(1,2,4)],by="userID",all.x=T)
```

```
survival_data<-merge(survival_data,gameplay_metrics[,c(1,8,9,10,12,14,17,20,21)],by="userID",all.x=T)
survival_data<-merge(survival_data,transaction_metrics[,c(1,2,3)],by="userID",all.x=T)

survival_data$payStatus<-rep(0,length(survival_data$userID))
survival_data$payStatus[which(survival_data$numPayments>0)]<-1

km<-survfit(formula=Surv(survival_data$timePlayed_mins,survival_data$defectionStatus==1)~1,data=survival_data,type="kaplan-meier",conf.type="log")
km
plot(km,main=expression(paste("Kaplan-Meier Estimates ", hat(S)(t), " with Confidence Interval")),xlab="Time",ylab="Survival Probability")

km_payStatus<-survfit(formula=Surv(survival_data$timePlayed_mins,survival_data$defectionStatus==1)~payStatus,data=survival_data,conf.type="log")
km_payStatus
ggsurvplot(km_payStatus,data=survival_data,pval=TRUE,palette="grey")+ggtitle("Kaplan-Meier Estimates of S(t) for Payers and Non-payers Separately")

View(table(survival_data[,16]))

survival_data$missionCompletionRate[is.na(survival_data$missionCompletionRate) & survival_data$defectionStatus==1]<-0.7
survival_data$missionCompletionRate[is.na(survival_data$missionCompletionRate) & survival_data$defectionStatus==0]<-0.94
```

```
survival_data$averageEnergyRestoredByFood[is.na(survival_data$averageEnergyRestoredByFood)]<-0


survival_data$averageSingleEBUsed[is.na(survival_data$averageSingleEBUsed)]<-0


survival_data$averageKillHitRatio[is.na(survival_data$averageKillHitRatio) & survival_data$defectionStatus==1]<-0.42
survival_data$averageKillHitRatio[is.na(survival_data$averageKillHitRatio) & survival_data$defectionStatus==0]<-0.43


survival_data_final<-survival_data[,c(1,2,3,5,6,8,9,10,11,13,16)]



set.seed(27)
sample<-sample.split(survival_data_final$userID,SplitRatio=0.75)
train_data<-subset(survival_data_final,sample==TRUE)
test_data<-subset(survival_data_final,sample==FALSE)

View(table(train_data$defectionStatus))
View(table(test_data$defectionStatus))



cph<-coxph(Surv(train_data$timePlayed_mins,train_data$defectionStatus==1)~missionCompletionRate+averageEnergyRestoredByFood+averageGold+friends+
averageKillHitRatio+averageNationalCurrency+averageSingleEBUsed+payStatus,data=train_data)
summary(cph)
```

```
plot(survfit(cph),xlab="Time",ylab="Survival Probability",main="Survival Curve with
Confidence Interval for Cox's Model")

payStatus_data<-
with(train_data,data.frame(payStatus=c(0,1),missionCompletionRate=rep(mean(missio
nCompletionRate),2),averageEnergyRestoredByFood=rep(
mean(averageEnergyRestoredByFood),2),averageGold=rep(mean(averageGold),2),frien
ds=rep(mean(friends),2),averageKillHitRatio=rep(mean(averageKillHitRatio),2),
averageNationalCurrency=rep(mean(averageNationalCurrency),2),averageSingleEBUse
d=rep(mean(averageSingleEBUsed),2)))

plot(survfit(cph,newdata=payStatus_data),conf.int=TRUE,main="Survival Curves for
Cox's Model for Payers and Non-payers Separately",xlab="Time",ylab=
"Survival Probability",col=c("grey","black"))
legend("topright",legend=c("Non-payers","Payers"),lty=1,col=c("grey","black"))

ggforest(cph,data=train_data)

test_cph<-cox.zph(cph,transform="log")
par(mfrow=c(3,2))
plot(test_cph[1],main="Mission Completion Rate")
plot(test_cph[2],main="Average Energy Restored by Food")
plot(test_cph[3],main="Average Gold")
plot(test_cph[4],main="Friends")
plot(test_cph[7],main="Average Energy Bars Used")
plot(test_cph[8],main="Paying Status")

pred_error<-
pec(list("Cox"=cph,x=TRUE),Hist(timePlayed_mins,defectionStatus)~missionCompleti
onRate+averageEnergyRestoredByFood+averageGold+friends+
averageSingleEBUsed+payStatus,data=test_data)

cph1<-
coxph(Surv(train_data$timePlayed_mins,train_data$defectionStatus==1)~missionCompletionRate+averageEnergyRestoredByFood+averageGold+friends+
averageKillHitRatio+averageNationalCurrency+averageSingleEBUsed+payStatus,data=train_data,x=TRUE,y=TRUE)
summary(cph1)

pred_cph<-
predictSurvProb(cph1,newdata=test_data,times=test_data$timePlayed_mins)
prob<-diag(pred_cph)

test_data<-cbind(test_data,probability=prob)

test_data$pred_defectionStatus<-ifelse(test_data$probability>0.5,1,0)

confmat<-
confusionMatrix(data=as.factor(test_data$pred_defectionStatus),reference=as.factor(test_data$defectionStatus),positive="1")
View(confmat$table)
confmat$positive
confmat$overall
confmat$byClass


########################################################################
```

```
pred_error<-
pec(list("Cox"=cph1),Hist(timePlayed_mins,defectionStatus)~missionCompletionRate+
averageEnergyRestoredByFood+averageGold+friends+

averageSingleEBUsed+payStatus,data=test_data)

plot(pred_error)


pred_error<-
pec(list("Cox"=cph1),formula=Surv(timePlayed_mins,defectionStatus==1)~1,data=test
_data)
```

Appendix C: List of Variables in the Raw Data

The variables in the raw data set and a brief description of the type of information they represent are detailed below.

Information on player ID and the type of event triggered –

1. esEventID,
2. userID
3. eventTimestamp
4. eventName
5. eventLevel
6. msSinceLastEvent
7. userEventSequence
8. userSessionSequence
9. userRevenueEventSequence
10. eventID
11. firstRegistered
12. mainEventID
13. parentEventID
14. uniqueTracking
15. sessionID

Information on general gameplay –

16. achievementID
17. achievementName
18. achievements
19. level
20. levelUpName
21. loyaltyLevel
22. xp
23. UIAction
24. UIName
25. UIType
26. acquisitionChannel
27. actionTaken
28. placeVisited

29. rewardName

Mission related information –

30. isTutorial
31. missionID
32. missionName

Military related information –

33. militaryRank
34. militaryStrength
35. damageForPatriotMedal
36. damageInBattle
37. damageInCampaign
38. division
39. rankPoints
40. combatOrderRevenue
41. isResistance

Player stats in virtual fights –

42. bazookaDamage
43. bazookasUsed
44. bigBombsUsed
45. bombsDamage
46. hitsCount
47. killsCount
48. noWeaponHits
49. numberOfFights
50. weaponDamage
51. rocketsDamage

Use of virtual resources –

52. foodQ1Used

53. foodQ2Used
54. foodQ3Used
55. foodQ4Used
56. foodQ5Used
57. foodQ6Used
58. foodQ7Used
59. weaponsQ1Used
60. weaponsQ2Used
61. weaponsQ3Used
62. weaponsQ4Used
63. weaponsQ5Used
64. weaponsQ6Used
65. weaponsQ7Used
66. rocketsUsed
67. smallBombsUsed

Energy (main virtual resource) related information –

68. clicksPerRecoverEnergy
69. doubleEnergyBarsUsed
70. energyRestoredByEB
71. energyRestoredByFood
72. singleEnergyBarsUsed

Virtual currency owned –

73. gold
74. nationalCurrency

Social variables –

75. friendsCount
76. inviteType
77. isInviteAccepted
78. recipientID
79. recipientUserID

80. referrer

81. senderID

Transaction information –

82. productAmount

83. productCategory

84. productID

85. productName

86. productType

87. transactionID

88. transactionName

89. transactionType

90. transactionVector

91. transactorID

92. convertedProductAmount