Estimating delivery dates in fashion e-tail

Inês Sofia de Sousa Carvalho

Master's Dissertation

Supervisor: Prof. Dr. Gonçalo Figueira

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia Industrial e Gestão

June 2016

"(...) e então perceberam que José Arcadio Buendía não estava tão louco como a família dizia, mas sim que era o único que dispusera de suficiente lucidez para vislumbrar a verdade de que também o tempo sofria precalços e acidentes e que, portanto, podia estilhaçar-se e deixar num quarto uma fracção eternizada."

Gabriel García Márquez in Cem Anos de Solidão

Abstract

E-commerce is a rising trading manner that is driving customers towards an increasingly demanding behaviour. Since e-customers still perceive the shopping activity as uncertain, information sharing is a vehicle to customer trust and hence customer retention. Sharing delivery status and estimated delivery dates with customers enhances their satisfaction and their repurchase intention.

The current project seeks to estimate delivery dates in a multi-brand luxury fashion e-seller. This delivery process spans several entities and stages, starting in worldwide scattered boutiques until final customers. As such, this project comprises six independent sub-problems that represent the order processing phases. For each one of these, a predicting model was created to determine the corresponding timespans, based on a range of independent factors that ought to characterize each specific order. This was supported by several data mining techniques such as data cleaning, classification and regression. Data cleaning and classification were performed in order to reshape data so that modelling results could be obtained or improved. Data cleaning mainly consisted in outlier removal and variable standardization, while classification's purpose was to decrease the number of levels of certain categorical factors. Modelling was an iterative process in which different techniques were explored, according to the previous modelling results. The main tool used in this stage were decision trees that created distinct order groups based on the combination of the chosen independent factors.

Due to the fact that each sub-problem comprises different processes and data, models' results diverged. With the exception of step 6 (*in transit*), results were satisfactory, which was evaluated based on several mean error measures. Concerning this step in particular, due to its importance in the scope of the delivery, factor *route* should be treated with more detail in order to improve model performance. Overall, expected timespan measurements were successfully obtained for each combination of order factors, which also constitutes a meaningful insight on what impacts on delivery performance. As this is an iterative process, once data is updated, other teams can be involved in the project in order to implement this tool. This is expected to positively impact on both business control and customer satisfaction.

Resumo

O *E-commerce* é um modelo de comércio em crescimento que está a instigar nos consumidores um padrão de exigência crescente. Uma vez que estes clientes ainda pressentem uma incerteza associada à compra *online*, a partilha de informação constitui um veículo para a sua confiança e, consequentemente, a sua retenção. Partilhar com o cliente o estado da encomenda, assim como a data prevista para a sua entrega, aumenta o seu nível de satisfação e intensifica a sua intenção de efetuar novas compras.

O presente projeto tem como objetivo estimar as datas de entrega das encomendas de um retalhista multimarca de moda de luxo. Este processo compreende diversas fases e entidades, começando nas *boutiques* espalhadas a nível global até ao consumidor final. Deste modo, este projeto inclui seis subproblemas independentes que representam as fases de processamento das encomendas. Um modelo de previsão foi criado para cada uma destas, a fim de determinar as respetivas durações, com base num conjunto de fatores independentes caracterizadores de cada encomenda. Isto foi levado a cabo através de diversas técnicas de *data mining*, nomeadamente: *data cleaning*, classificação e regressão. As duas primeiras permitiram uma reestruturação dos dados, de modo a que a sua modelação pudesse ser realizada. A fase de *Data cleaning* consistiu na remoção dos *outliers* e uniformização das variáveis, enquanto o propósito da classificação foi reduzir o número de níveis de determinados fatores categóricos. O processo de modelação foi iterativo, permitindo a exploração de diferentes técnicas, de acordo com os resultados dos modelos anteriores. A principal ferramenta usada nesta fase foram as árvores de decisão, que permitiram a criação de grupos de encomendas distintos, com base na combinação das variáveis independentes escolhidas.

Uma vez que cada subproblema é constituído por diferentes processos e dados, os resultados dos modelos divergiram entre si. À exceção do sexto (*em trânsito*), os resultados foram satisfatórios, tendo sido avaliados com base num conjunto de medidas de erro médias. Relativamente a este problema em particular, dada a sua importância no contexto dos processos de entrega, o fator *rota* deve ser tratado com maior minúcia a fim de melhorar a performance do modelo. No cômputo geral, a duração esperada das diferentes fases foi obtida com sucesso para cada combinação de fatores, o que constitui um conhecimento significativo sobre os preditores da qualidade de entrega. Uma vez que este se trata de um processo iterativo, após a atualização dos dados, outras equipas poderão ser envolvidas no projeto a fim de implementar esta ferramenta. É esperado que esta tenha um impacto positivo no controlo dos processos de negócio, assim como na satisfação do cliente.

Agradecimentos

Apesar do tempo se desenrolar de forma contínua, o ser humano tem necessidade de o fracionar e atribuir a determinados momentos uma importância especial. Sendo o fim do curso um desses momentos, penso ser pertinente expressar o meu apreço por aqueles que, por via do seu tempo e dedicação, enriqueceram o meu caminho até então.

Deste modo, e incontornavelmente, agradeço à minha mãe, cuja menção surge em primeiro lugar, não por tal corresponder à sucessão cronológica das minhas experiências interpessoais, mas pelo seu papel sem par entre as mesmas. O meu obrigada mais sincero e profundo. Contudo, uma vez que as suas inquestionáveis qualidades não são o bastante para permitir a minha existência, agradeço também ao meu pai, por sempre incutir em mim a curiosidade e o espírito crítico.

À minha Margarida, a personificação da doçura e dedicação, agradeço a partilha de tantos bons momentos e a compreensão demonstrada nestes últimos meses.

Aos meus amigos, pelas infinitas alegrias que partilhámos ao longo destes anos, gostaria de deixar uma mensagem carinhosa, não fosse o meu receio de ser alvo de chacota. Por favor, continuem assim.

Além destas pessoas que povoaram a minha vida até ao momento, tive também o prazer de encontrar seres humanos excecionais no contexto académico-laboral em que me inseri nos últimos meses. Neste sentido, começo por agradecer ao Professor Gonçalo Figueira, pela disponibilidade que sempre apresentou para a orientação deste projeto. No âmbito empresarial, agradeço primeiramente à Engenheira Rita Raposo, por me guiar neste projeto a um nível inesperado. Espero ter aprendido todas as lições. À Joana e à Luísa, por voluntariamente terem despendido do seu tempo a fim de melhorar este trabalho, um reconhecimento muito especial. Por fim, aos meus novos amigos, sem os quais não consigo imaginar esta experiência, agradeço o verdadeiro espírito de companheirismo que partilhamos.

Contents

List of Acronyms

ECT Expectation Confirmation Theory

GMT Greenwich Mean Time

MAE Mean Absolute Error

MAPE Mean Absolute Percentual Error

ME Mean Error

MPE Mean Percentual Error

MSE Mean Squared Error

List of Figures

List of Tables

1 Introduction

It is a truth universally acknowledged, that retail is being remodeled and rejuvenated. As culture evolves and working hours shrink throughout the decades (Basu et al. 2006), practical solutions to unburden citizens from their everyday tasks also proliferate, among them e-commerce. Additionally, e-commerce is a trading method that endorses a wide and heterogeneous spectrum of business models, ranging from undifferentiated items to luxury goods, domestic to transcontinental trade, culminating in total sales value of 1,671 billions of dollars in 2015, almost a billion of which in B2C retail (Statista 2016). Despite their success, e-commerce companies face critical challenges as their innovative character also lies in the paradoxical nature of their business processes.

One of the best examples of this phenomenon is the approach to customer-seller relationships that one might classify as tailored and considerate, despite of the physical distance between them. It is due to this remoteness that sellers need to further invest in customer support, to overcome risk aversion and conquer trust and satisfaction (Khan, Liang, and Shahzad 2015). By doing so, companies are investing in customer retention which, according to some studies, is the most cost-effective way of safeguarding their ultimate profit (Ng, Sumeet, and Kim 2007). Furthermore, sellers ought to recognize that e-customers opinions are highly scalable, as the internet also embodies an informative and social medium, an attribute that becomes more significant on the negative side of the satisfaction scale. Sellers should, therefore, be aware of the promotional value of their clients and manage them (and their expectations) accordingly.

Another antagonism that defines e-commerce is the increasing convenience from the customer's point of view, both from a time a space perspective, against the rising complexity from the supplier's side. The e-commerce supply chain is evolving into an even more sophisticated set of relationships, as e-sellers reach for their products, and therefore their multiple suppliers, globally. Several logistic providers are required to support these commercial relationships, resulting into an elaborate net of collaborators. In addition to this, customers are becoming increasingly demanding about shipping time (Y. Lin et al. 2016) and delivery points are also widespread, which further rises supply chain complexity .

Moreover, there are other dimensions of delivery performance that shape customers' reaction to the provided shopping experience. A major example of this is the fact that sharing delivery-related information with the client leads to satisfied and loyal customers. It is convenient to recall that one of the major barriers to e-commerce is the risk perceived by buyers concerning this activity. This uncertainty, though it ceases at the delivery moment, can be meanwhile mitigated by means of information sharing. Subsequently, communicating delivery status increases service quality and e-trust. The sources of this information may, however, be spread among the entities previously mentioned, which hampers its transmission, unless it is integrated beforehand. This integration may be impractical, due to technological or organizational reasons, consequently, the remaining solution would be for the seller to forecast delivery time, based on former data. As buyers, sellers and carriers are distributed globally, and due to their heterogeneity and high number, the present case falls into this category. Hence, Bid Data

methods were applied throughout this project in order to create a predicting tool based on order-related data.

Prior to the unfolding of the specific attributes of this challenge and the company itself, it is relevant to highlight that satisfaction is one of the primary factors impacting over customer retention, along with e-trust. E-satisfaction, in turn, is significantly dependent on information and delivery quality. Delivery dimensions are, for this reason, not merely niche operational components of e-commerce as they have a real impact on customer retention and, henceforth, corporate vitality.

1.1 Farfetch

The present case study is based upon Farfetch e-commerce company. It consists of a multi-brand luxury fashion platform that bridges the physical gap between worldwide high-class boutiques and its current and potential customers. At the moment, more than 600 boutiques (200 of those restricted to the Brazilian market) have so far supplied over 2 million orders to nearly 200 different countries and independent regions, despite Farfetch's recent creation in 2008. The company's success can be further expressed by its recent evaluation of 1.5 billion dollars (2016). Additionally, Farfetch finds itself in a growth period and has been the target of massive investment in the past few years. As a result, its workforce has also been increasing at a steep rate, within different countries. Aligned with this all-encompassing nature, Farfetch offices are geographically positioned to guarantee operational support in the main time zones where stakeholders can be found. Accordingly, the majority of its workforce is placed in Europe, where the best-selling stores are established, specifically in London, Porto, Guimarães and Moscow. The remaining offices are located in Los Angeles, New York, Shanghai, Hong Kong, Tokyo and São Paulo. This distribution is attributed as the United States are the most significant market, followed by the United Kingdom, Australia, Brazil, Hong Kong, Germany, Russia, South Korea, France, Japan and China. In order to maintain a global and unified feel, Farfetch has developed multiple platforms, which gain relevance in the scope of the present matter, fed by several internal databases.

Concerning its business model, and unlike its direct competitors, Farfetch does not possess the items that are displayed online for sale, since they are owned by the boutiques. Whereas Farfetch profits both from non-existing inventory expenses and vast product availability, this business model also endorses some drawbacks, such as substantial *stockout* exposure and increased delivery complexity. On top of this, shipping is not the only activity whose time needs to be forecasted, as several processes are conducted and numerous entities are involved to guarantee that order requirements, such as flawless packaging, are fulfilled and fraudulent shopping is uncovered. As a result, estimating the total delivery time represents a challenge. However, given the fact that sharing this information with customers embodies an influential asset over their trust, the present project seeks to overcome it.

Due to the luxurious nature of the business, Farfetch customers exhibit a particularly demanding attitude, in comparison with the average consumer behavior, which is also applicable to the delivery service. As a matter of fact, in Farfetch, 11% of customer service contacts are related to delivery time, an element reflected also in customers' evaluation of the company. This is aligned with the data conveyed by Eurostat (2015), which identifies late delivery as the main problem faced by online customers, the source of 16% of dissatisfaction issues. In addition, high-class fashion purchasing has a significant hedonic component, a dimension that customers often find more powerful in traditional retail. In light of these facts, customer satisfaction in the present company depends on meeting foremost expectations, as to guarantee an unblemished delivery performance, in the midst of a net of competitors.

1.2 Current status vs. projected solution

It is now suitable to scrutinize the currently offered service for the sake of grasping what is lacking and establishing a strategy to attain the intended results. At the moment, during the checkout process, the customer is given an estimated delivery date that covers both shipping and additional activities introduced beforehand. However, this estimation is rather unpolished, grouping orders in 3 groups, disregarding the effect of promotional campaigns, uneven boutique performances, and other significant factors. Complementary to the information provided upon ckeckout, when the courier picks the order in the boutique, the customer receives an e-mail in which a link to the courier's website is included. In this page, order progress and estimated delivery time are displayed.

Since the current forecasting model is very generic, it typically delivers too conservative forecasts so that it is able to embrace the wide spectrum of cases that are included in each considered category. Consequentially, although orders hardly ever arrive after estimated delivery date, customers initially perceive shipping service as mediocre. Hence, as an alternative to this solution, the proposed model has its roots in identifying the main factors that impact on delivery date. To do so, the processes comprised between the purchasing moment and delivery date must be mapped beforehand. This period can be divided in 6 steps that will be described further ahead. Only two within these steps will be of Farfetch's operational domain, whereas the remaining are related to the boutique or the carrier. The project is respectively divided in six independent forecasting problems, each one of them determined by distinct factors. Ultimately, the estimated delivery time will approximately be the sum of these forecasts (Figure 1).

Concerning each one of these problems, though their mathematical features are analogous, according to the specific contexts of the micro-models, gathering and treating data are rather dissimilar activities. While some of them can be explained based on pre-determined existing variables, others require creating and calculating new metrics or clustering data. Also the number of independent variables is heterogeneous. On top of this, the values involved in each one of the steps differ in magnitude. Hence, the independent variables that assume higher expected values and variability are to be given more attention in this project. These are typical outlines of Big Data problems. Akter and Wamba (2016) define it using the "five v's" regarding data characteristics: volume, velocity, variety, veracity and value. Big Data Analytics (or Data Mining) tools are used to create value from this kind of source, reshaping data into meaningful information. As such, companies are turning to tools like the ones used in this project, most importantly, classification, data cleaning and modelling. Figure 1 represents the main features of the current project.

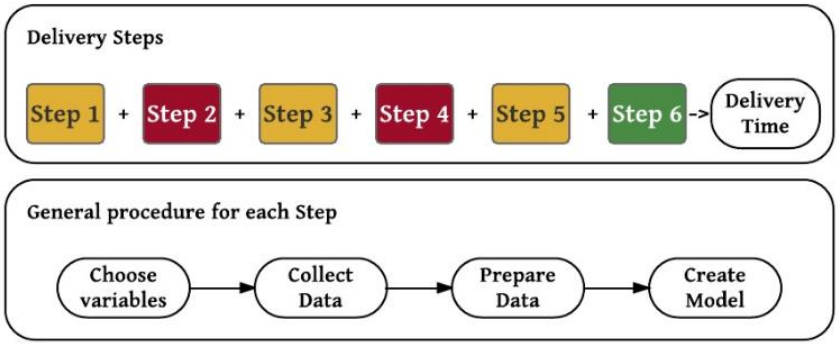


Figure 1 - Project overview

The current dissertation reports the project *Estimated Delivery Date*, a cross-functional development in Farfetch. Following this first exposition of the project, its theoretical purpose and methodology are further solidified in the next chapter. Subsequently, *The Challenge* is described on an extensive level, covering both process description, first numerical features of

data and preliminary decisions (Chapter 3). The main stages of *The Project* are then recounted, from *Data collection*, towards *Modelling* and results (Chapter 4). The last chapter displays a critical overview of the project and its main outcomes and it also suggests some enhancements, to be adopted further ahead.

2 State of the Art

2.1 E-commerce

Internet, a crucial technological offspring of the previous century, has paved the way for the birth of e-commerce which, in turn, revolutionized the commerce paradigm and led to an upgrade in business processes by demanding more efficient, flexible and responsive companies (Labajos and Jimenez-Zarco 2016) . Accordingly, some studies stress the positive relationship between the adoption of e-commerce and labor productivity (Falk and Hagsten 2015).

Despite being a relatively recent way of doing commerce, e-commerce is becoming increasingly popular among Internet users, ranking third place among the most common internet activities, only surpassed by e-mail and web browsing (Bhaskar and Kumar 2015). In fact, e-commerce B2C sales have been growing steadily and are estimated to reach 1.16 trillion U.S. dollars in 2016 (Statista 2016). World economy is facing major challenges worldwide, the online luxury market has also been growing in the past few years and this trend is expected to endure during the following (Statista 2016). However, even taking into account the diverse population levels among countries, customers' interest in luxury e-tail is massively concentrated in a few countries; the United States have 66% of the visitors of multibrand luxury retail websites, while South Korean and Japanese customers, the second and third most frequent customers respectively, only comprise 4% each, followed by China and the UK, with 3% of the global visitors (Guercini and Runfola 2015).

From the company's side, e-commerce allows the seller not only to reach more customers, due to the lower costs of distributing information in comparison with traditional retail, but also through a personalized approach, enabling closer buyer-seller relationships and a more effective market segmentation (Labajos and Jimenez-Zarco 2016). However, due to the risk perceived by the customer, e-commerce faces several challenges (such as the loss of potential customers) that should be dealt with by improving service quality (Y. Lin et al. 2016).

2.1.1 *The customer side*

Internet has restructured the interaction between customers and companies into a continuous and dynamic relationship, empowering the first to a more informed and active role in the selling process, enabling the selling process, and the value chain itself, to become clearer (Nuseir 2016). In spite of this, research has so far failed to expose the precise nature of the interactions between the shopping experience offered by the seller and the resulting recognition of the customer (Y. Lin et al. 2016).

Customer retention has been widely stated as the best way of keeping a healthy customer set since, according to Reichheld and Schefter (2000), it is much more expensive to acquire a new customer than keeping an old one, which is particularly accurate in e-commerce (Kong, Kee, and Ireland 2003). To increase retention rate, it is compulsory to understand what drives customers in their shopping experience. Hedonic characteristics of the shopping experience,

such as a pleasant and interactive navigation experience, have been proved to be determinant factors in both e-commerce intention and adoption (Machado 2005).

Further understanding of customer behavior and motivations is critical, as a significant share of online searches still does not lead to purchases (Al-maghrabi and Dennis 2009). Yet online shopping experience is so different from traditional retail that previous Marketing principles are not suitable for this purpose (Machado 2005). Interactivity has been widely stated as one of the most important advantages of e-business forms, as customers can now be delivered various sorts of precise and helpful information supported by data mining technology, increasing customer loyalty (Nuseir 2016).

The Expectation Confirmation Theory (ECT) explains repurchase intentions as the result of the comparison between initial expectations and reality. This, however, fails to include possible deviations in the customer expectations due to shopping experience (Al-maghrabi and Dennis 2009). Concerning the factors that influence customer retention, satisfaction, trust and loyalty are the most significant concepts in the literature.

Satisfaction may be explained by ECT as an emotional state determined by the comparison between expectations and reality (Khan, Liang, and Shahzad 2015; Kong, Kee, and Ireland 2003). According to Lin, Wu, and Chang (2011), customer satisfaction is mainly determined by system, information and service quality. Hence, it is essential for companies to understand the customers' point of view on their services, especially in e-commerce context, where it is much easier for the customer to compare prices and start buying elsewhere (Khan, Liang, and Shahzad 2015). Moreover, satisfaction is widely pointed out as one of the crucial factors impacting on repurchase intention (C. Lin, Wu, and Chang 2011). The lack of trust, which may be the result of e-commerce perceived risky nature (Kong, Kee, and Ireland 2003) is also one of the most important factors hindering on-line sales and customer retention (Bulut 2015).

At last, customer loyalty may be reached by maintaining a cooperative relationship and providing an easy and personalized shopping experience. (Bhaskar and Kumar 2015). Due to the nature of the e-commerce medium, building customer loyalty has the further advantage of them spreading their satisfaction faster and further, a significant opportunity to grasp new customers (Zhou, Dai, and Zhang 2007). However, these customers are also more challenging, demanding an experience that fits or surpasses their needs, usually for free (Al-maghrabi and Dennis 2009).

Although there is an extensive literature collection seeking to find the determinants of repurchase intention, it has up to now been unsuccessful in uncovering the relationships among customer psychological traits such as satisfaction, trust and loyalty, as there is little focus on the matter and results are conflicting (Bulut 2015). For this reason, the literature can be divided in two groups or views: the transactional view, according to which customer satisfaction is a determinant factor of customer loyalty, and the relational view, that assumes trust to have this role (Li, Browne, and Wetherbe 2006). However, some studies support the existence of a direct relationship between trust and satisfaction (Polites et al. 2012), although opposite causal links can be found in literature (Khalifa and Liu 2007; Rose et al. 2012). Others define satisfaction as a necessary, yet non-sufficient, determinant of trust (Bhaskar and Kumar 2015).

Complementary, some researchers state that trust is a mediating effect that increases the influence of satisfaction on repurchase intention (Ha, Janda, and Muthaly 2010), while others simply argue that these are the two main predictors of loyalty, regardless their mutual influence (Valvi and West 2013). At last, another relevant dimension to understand customer loyalty and retention is relationship quality, which is widely believed to be the result of satisfaction, trust, commitment, among other factors (Wulf, Odekerken-Schröder, and Iacobucci 2001).

2.1.1.1 Luxurious fashion customers

Following the previous introduction to general consumer behavior in online shopping, outlining the main characteristics of luxury e-buyers with focus on fashion products is required. According to Hansen and Jensen (2009), fashion-related buying activities generally possess a strong hedonic dimension, which can be a hindering factor for online fashion sales, as customers usually perceive traditional shopping experience as more entertaining than online shopping experience, but may also benefit from hedonic-driven shopping impulses (Santos, Hamza, and Nogami 2016). Research also points out that customers' satisfaction with delivery service depends on the kind of product purchased, as specialty goods (those that the customer would insist on buying and would be willing to pay a higher price for) induce a more demanding customer (Cao and Mokhtarian 2009).

A peculiar characteristic of this segment is that, unlike the majority of B2C business cases, luxury companies do not seek to grow as much as possible, as this would imply losing their exclusivity status, which explains the conflicting nature between e-commerce and luxury brands (Guercini and Runfola 2015). Due to this seeming incompatibility, many luxury brands have first offered resistance towards e-commerce. This phenomenon is currently being reversed, as research points out that e-presence is a favorable asset for luxury brands, since current technology can already deliver a pleasant online shopping experience. Moreover, luxury consumers are increasingly keen on e-channels, both for information-seeking and buying purposes, especially in emerging markets, an opportunity for luxury brands to grow without relinquishing their exclusivity appeal (Chen and Zhang 2011) .

2.1.2 *Globalization, logistics and risk*

Since many items are inaccessible to customers via traditional channels due to political barriers, international trade has triggered the current e-sales flourishing phenomenon and overcome several geographic drawbacks (Alden, Steenkamp, and Batra 2006). Moreover, Terzi (2011) adds that e-commerce will reshape the nature of trade barriers and, furthermore, Martens (2013) supports the idea that distance-related costs will also be mitigated. In spite of this, distance still encompasses several challenges for e-commerce.

Firstly, as the delivery moment constitutes the first contact between the customer and the purchased item, logistical support should be carefully provided in order to enhance prior customer satisfaction. As a result, fast communication and quick, timely and effective delivery are critical (Azar et al. 2015). Kumawat and Tandon (2014) found that delivery performance positively influences customer e-satisfaction and e-loyalty, an idea reinforced by C. Lin, Wu, and Chang (2011) that define information, service and delivery quality as three of the main factors leading to customer satisfaction.

Secondly, as customers still regard traditional retail as safer than on-line shopping (Bulut 2015), another vital dimension shaping both e-commerce adoption and e-satisfaction is perceived risk, which can be explained as a consequence of the impersonal environment and the unpredictable outcomes of e-shopping, culminating in fears in the customer's mind (Khan, Liang, and Shahzad 2015). This is a multi-dimensional factor that comprises different kinds of risk, including risk of failed delivery or related time-consumption, both negative contributors to customer satisfaction and repurchase intention that can be reduced by increasing the quality of logistics (Y. Lin et al. 2016). According to Mentzer, Gomes, and Krapfel (1989) and Emerson and Grimm (1996) this dimension is shaped by product availability, delivery quality, accurateness and on-going status communication.

On top of this, customers' perception of delivery performance and its effects on satisfaction are further disguised when the courier and the seller are separate entities. In fact, as from the customer point of view there is not a clear line dissociating these entities, the performance of

the first can damage or strengthen customers' relationship with the company, and the other way around (Y. Lin et al. 2016). Therefore, stimulating cooperation and monitoring outsourced logistic services are critical points to preserve a satisfied and loyal set of customers. However, these services can only be controlled if the seller has some analytical knowledge about the activity of the courier. This can be obtained using Big Data methods, which will be covered in section 2.3.

2.2 Big Data methods

With a total sales volume surpassing one million orders in the one-year period under study and a thousand combinations of origin and destination countries, the current predictive problem certainly falls under the category of Big Data Analytics (BDA).

Similarly to other recent fields of study, Big Data can be defined from different perspectives and consensus has not been reached so far. Schroeck et al. (2012), for example, define it by describing its main characteristics: high volume, heterogeneous and real-time data and information and diverse analytic methods and purposes. Likewise, White (2012) defines Big Data using "the five Vs": volume, velocity, variety, veracity and value (Akter and Wamba 2016). Value is the intended outcome of Data Mining activities, such as accurate forecasting information and, ultimately, information-based business decisions (Davenport, Harris, and Shapiro 2010), empowered customers and loyalty-based relationships (Gefen 2002). Other authors stress the existence of different data sources, usually physically sparse, combined into a unique database and its consequences for database-related tasks. E-commerce firms were among the earliest adopters of BDA, since this is critical to their survival and strongly connected to their business field. Data concerning e-commerce and BDA can be found in Table 1. According to McAfee and Brynjolfsson (2012) e-commerce companies that have done so have consequentially experienced a 5 to 6% productivity surplus (Akter and Wamba 2016). In this case, data is usually heterogeneous, ranging from structured to unstructured, since it can be generated from customer experience (inserted data or traces of his/her web electronic path), from interaction with suppliers and partners or even internal processes. However, these companies should be aware of the privacy issues that are implied by the use of customer-related data, since, although internet users generically want free and tailored services, they generically want to safeguard their privacy (Hull 2015).

Table 1 - Global growth in e-commerce and big data analytics (BDA)

Year	Growth in the number of e-commerce customers worldwide (in millions)	Growth in e-commerce sales per customer worldwide (in US$)	Growth in big data analytics (BDA) market worldwide (in billions)
2011	792.6	1162	7.3
2012	903.6	1243	11.8
2013	1015.8	1318	18.6
2014	1124.3	1399	28.5
2015	1228.5	1459	38.4
2016	1321.4	1513	45.3

Source: Adapted from emarketer (2013) and (Piatetsky, 2014)

Focusing on Big Data Analytics contribution to the present work, customer expectations towards order tracking and delivery date estimation are often unfulfilled due to high supply chain complexity. BDA can solve this by gathering data from several entities (logistics partners, for example) and creating methods to share this information and determine expected delivery time.

Big Data methods require both technical and analytical skills that should be combined with business knowledge and communication competences in order to generate value (Davenport,

T.H. 2012). Structuring BDA processes in sequential stages, they can be divided in 6 main steps (Figure 2), beginning with *Business Understanding*, when the problem is defined from a business point of view. This is followed by *Data Understanding*, the initial approach to data that should lead to the first insights concerning its nature and consequentially defines how *Data Preparation* should be conducted. By the end of the third stage, data is ready for *Modeling*, a stage that comprises several iterations until the right model is validated in the *Evaluation* stage, a process conducted by comparing the business requirements defined in the first stage with the model attributes (Akter and Wamba 2016). At last, in the *Deployment* stage, model outputs are reshaped in order to provide meaningful information for the target audience.

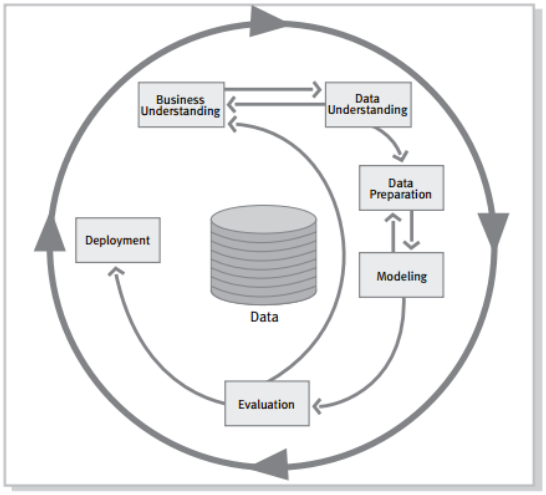


Figure 2 - Big Data Main Steps

2.2.1 *Data preparation: Classification and Outlier detection*

Predictive methods cannot often be implemented in raw data, since redundant information, outliers and non-normalized data can hinder their performance. Additionally, independent continuous or categorical variables may have to be classified in clusters prior to modelling.

Concerning classification, the act of grouping variable values into clusters, this may be performed according to one or more variables. Within multivariate data contexts, multivariate data analysis is an important tool to understand and interpret data frames. Each data register represents an individual that is characterized by diverse information stored in variables.

One widely used classification methodology is cluster analysis, whose goal is to find separate clusters of homogenous elements. This means that values within each cluster should be similar to each other, while as distinct as possible to those of other clusters. Hierarchical Classification methods produce embedded partitions which can be divided in agglomerative or divisible. Results are most often displayed in a *dendrogram*. When applied to a data frame, Non-Hierarchical Classification creates as many partition clusters as requested. (Sousa and Nicolau 2001).

Generically, outliers can be defined as values that are inconsistent or significantly different from the remaining observations (Wang 2014) and, therefore, lead to suspicion regarding their reliability (Enderlein 1987). Although, in many cases, this may be due to mistyping or measuring errors, outliers can also be indicators of atypical, yet factual, occurrences, which means that these values are highly informative (O. Maimon and Rokach 2010). On the other hand, since erroneous values lead to model miscalculations and incorrect results, those should be spotted and removed before modeling (Hawkins et al. 2002).

Typically in prediction problems, observations are multivariable, i.e. there is more than one independent variable. Therefore, the combination of these values should be considered in order

to achieve an unbiased classification of the dependent values. Hence, a univariate outlier detection can be performed within each group. Outlier detection methods can be divided in two groups: parametric and non-parametric, that are usually more appropriate in the Big Data context, since it is not compulsory for data to fit a certain statistical distribution to use them (Papadimitriou et al. 2003). Within the last group, the boxplot rule can be found. Boxplots, formally introduced by Tukey in 1977, are graphics that display the median, first and third quartiles values of continuous univariate data plus two whiskers that separate potential outlier regions, where values are depicted as circles, from the remaining values. These values can be calculated by adding 1.5 interquartile ranges to the third quarter value, for the upper whisker, or subtracting 1.5 interquartile ranges to the first quarter, for the lower whisker (Figure 3). Values found outside these limits are classified as mild outliers. In case whiskers are calculated by multiplying the interquartile range by 3, instead of 1.5, extreme outliers can be determined. This method delivers good results when applied to skewed data, due to the fact that highly outlier-sensitive parameters, such as the mean or standard deviation are not used as outlier criteria inputs. (Seo 2006).

Although ideally all outliers would be spotted and treated accordingly, this is usually not the case, especially in BDA context. Consequentially, there are erroneous values that are kept in the dataset, while some irregular, yet realistic ones, are perceived as outliers, being, therefore, removed. More concretely, the presence of outliers could be analyzed by performing a hypothesis test where the null hypothesis would be "the value is not an outlier". In this case, false negatives would be the unspotted outliers, while false positives would be the discarded regular values. Although in an optimal solution both would be minimized, by choosing the boundary between rejecting and non-rejecting area, there is a trade-off between both, which should be considered according to the business context. On one hand, false positives are more costly, since they lead to the loss of healthy data and important occurrences may be masked, while, on the other, false negatives may induce model inaccuracies and delay the BDA process (Wang 2014).

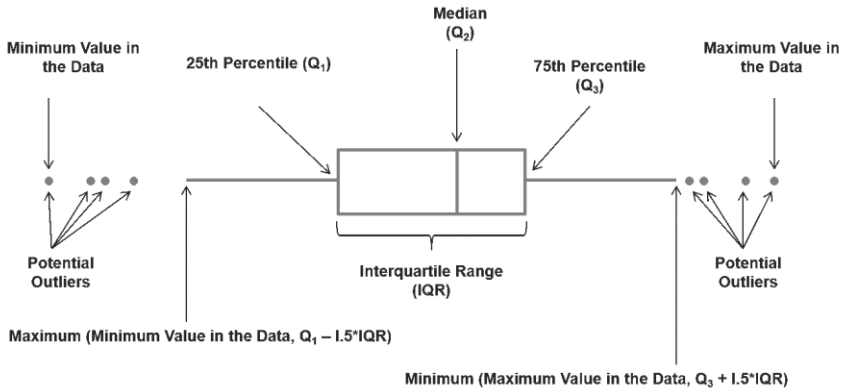


Figure 3 - Boxplot

2.2.2 *Modelling*

The purpose of this BDA activity is to estimate the outcome of a certain event, an exercise known as prediction, since the estimated values follow a continuous distribution. These values are assigned to the dependent variables of the model and represent the unknown attributes of tuples, or records. Each tuple is a vector with *n* dimensions, where the remaining *n-1* dimensions are independent variables and, by assembling all tuples in matrix, the dataset is obtained (Han and Kamber 2006).

According to Mayor (2015), regression is a widely used method in prediction problems and its goal is to define a relationship among independent and dependent variables. To do so, a part of the dataset, the *training set*, is selected in order to build the model, while the remaining tuples,

the *testing set*, are used afterwards to evaluate its accuracy (similarity between predicted and actual values). Furthermore, the quality of the model is defined by its speed, robustness, scalability and interpretability. Linear regression can be relied upon when data attributes (both dependent and independent) are normally distributed, which is not the case of the current project. Alternatively, classification trees can be found among the methods used to predict a certain variable outcome even if this variable does not follow any theoretical distribution. (Mayor 2015). Given the *training dataset*, decision trees take successive decisions based on attributed values, so that the best classes are obtained. Conditional inference trees can do this by recursively subdividing branches in two paths in each split point, which is designated as node. Nodes can be *internal* or *terminal*. *Internal* nodes constitute decision points, while *terminal* nodes represent the final groups produced by the model. Figure 4 represents a generic decision tree.

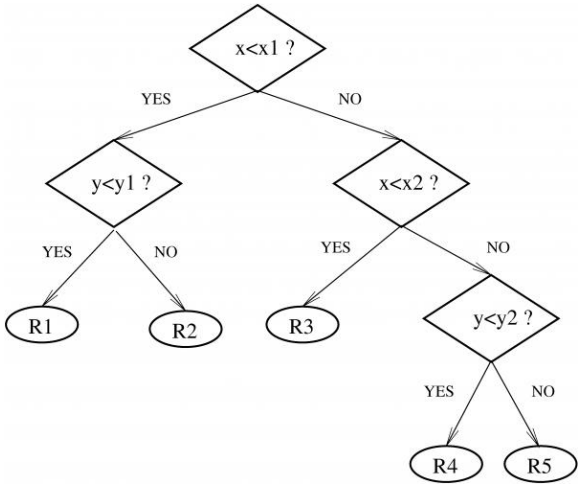


Figure 4 - Decision tree example

Splitting procedures commonly lead to two major problems: overfitting (too many final nodes) and biased splitting criteria. The last one happens when attributed values offer too many division possibilities or when missing values can be found in data. (Hothorn, Hornik, and Zeileis 2006).

R language constitutes a wide-ranging tool for Data Mining activities, such as Modelling. In addition to *R* built-in functions, code *packages* featuring multiple interrelated functions are created, documented and made available by *R* users. Among those, *partykit* displays functions for regression trees creation, representation and summarizing, based on the beforehand introduced concepts. These tools will be of great importance to create accurate models that help approaching the estimation problem of this project.

3 The Challenge

The classic supply chain model expresses the value chain as a sequential array of processes conducted by different entities. Within the scope of each one of these, *procurement* is traditionally the first stage, followed by *production* and *distribution*, so that goods can be sold to customers (B2C). In the e-commerce context, however, this natural order is reversed, as sales precede physical distribution. Regarding Farfetch, this structure is further transformed. Farfetch main activities take place in the last two steps of the matrix: distribution and sales (Figure 5). Due to these peculiarities, forecasting delivery date also includes predicting the timespan of tasks such as fraud detection and stock availability control, which idiosyncratically follow the actual purchase. This chapter will cover these processes in detail so that methodology can be unraveled afterwards.

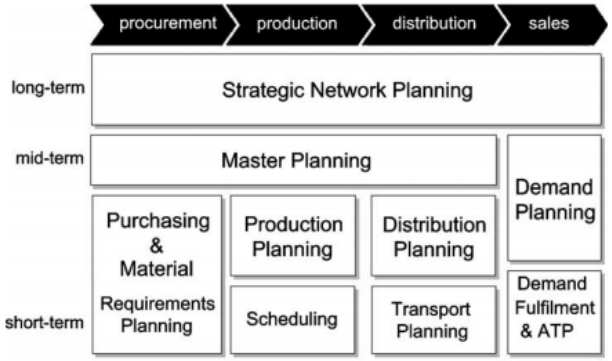


Figure 5 - Supply Chain Planning Matrix

Apart from Farfetch itself, boutiques, carrier companies and customers play the main roles in these processes. While boutiques secure stock availability and customers account for financial survival, courier partners, in between, provide a physical connection between both. In order to coordinate overall activities and interactions, Farfetch Portugal is currently divided in 11 departments. Regarding stores, *Account Management* operates on a strategic level while *Partner Service* safeguards operational efficiency. Customers, in turn, are assisted by *Customer Service*, a department that seeks to reply to any queries concerning products, delivery service, and returns, amid others. Delivery-related issues are solved by the *Operations* team, which is additionally held responsible for supply, fraud detection, payments and continuous improvement. The remaining departments are *Finance*, *People Team*, *Merchandising*, *Office Management*, *Technology*, *Production* (photography) and *Black and White*. Delivery time is dependent on the performance of multiple departments.

3.1 Order Processing

Of all the activities supporting Farfetch business model, those ensuing order request until final delivery are the main focus of this project. This set of tasks, which constitutes *order processing*, can be divided in 6 Steps (Figure 6), which will shape the methodology of this project.

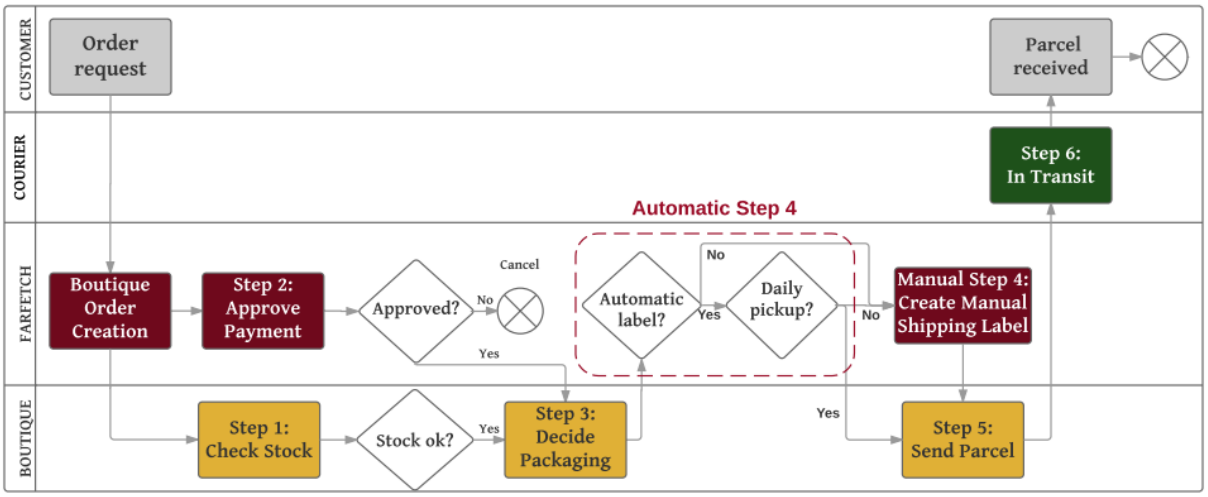


Figure 6 - Order processing Steps

Prior to their purchasing, goods have to become available on the Farfetch platform. To do so, each time a store decides to display one item online, it is sent to one of the Farfetch *Production* offices, where it is photographed according to several requirements. Afterwards, it returns to the origin boutique, where it is also physically displayed for sale. As such, the stock level of the stores should be integrated with Farfetch database, so that only items that are truly available are visible on the website. However, although Farfetch provides an Application Programming Interface to synchronize this data, some stores still disregard it and communication is hardly immediate. Thus, the first activity conducted by the boutique's personnel, following online purchasing, is stock availability validation, an activity internally acknowledged as *Step 1, Check Stock*. In case of a *No Stock*, the order is cancelled and the customer is refunded, or suggested a similar item.

Meanwhile, within Farfetch's domain, fraud recognition processes are executed. Since, on average, almost 3000 orders are requested every day, an automatic tool sorts them in three groups, according to their fraud likelihood. If the client and the corresponding bank information have been previously flagged as trustworthy or blacklisted, the orders are automatically accepted or rejected. Otherwise, payment authenticity is considered dubious, which leads to an investigation process. This stage, defined as *Step 2, Approve Payment*, usually spans a shorter time period that *Step 1*. Hence, as they begin simultaneously, *Step 1* is, in most cases, the bottleneck before *Step 3, Decide Packaging.*

Provided that both *Steps 1* and *2* outcomes are favorable, the boutique is informed that the box in which the item will be shipped can be chosen. The boxes have standard sizes, being provided by Farfetch. Although Farfetch platform suggests the smallest box size available for that item, it is within the store's power to do otherwise or even to add a personalized feature. Box stock level is managed by the stores that ought to request extra boxes whenever they find it necessary.

Like *Step 2*, *Step 4, Create Shipping Label*, is also performed by the *Operations* department and is most often automatic. However, in case customers have not filled shipping-related data accurately during the checkout process, this information has to be corrected manually. Once every mandatory information is properly introduced in the system, an *Air Waybill* (AWB), the document that will serve as an identification of the parcel until the delivery moment, is created. Apart from the previously described delaying issues, orders can also be held on this step while legal matters are solved.

Step 5, Send parcel, comprises the activities following the AWB creation until the courier picks the package in the store. However, boutiques are free to postpone packaging until this stage, which means that both the courier and the store impact on the performance (and *timespan)* of this step. According to the selling volume of the store, pickups can occur daily or can be

scheduled upon necessity, using Farfetch integration tools. Concerning data accuracy on this step, although boutiques mostly log *Step 5* beginning in the store, sometimes this takes place in the warehouse, which leads to erroneously higher timespan measurements. When the carrier finally collects the package, *Step 5, Send parcel,* is completed and *Step 6, In transit*, is performed so that the order can reach to the customer. Regarding all Steps in which boutiques are concerned, their length is typically determined by several variables, among those the store itself. Although some stores deliver satisfactory performances, regardless their order volume, others repeatedly fail evaluation targets. This is due to the highly variable nature of the stores in dimensions such as staff allocation to Farfetch and technological know-how. On the Farfetch side, *Partner Service* department seeks to overcome these issues. In addition to this, according to Farfetch teams, boutiques' performance is generally hampered during sale seasons, other promotional campaigns and weekends. Moreover, they state that boutiques response is strongly delayed when they have a large order *Backlog*.

Although the remaining stages are mostly standardized, the last step, *Step 6*, admits a higher degree of variability, as it strongly depends on the specific features of the order. First of all, origin and destination will mostly define the route complexity and delivery time. On top of this, if this route is international and beyond specific free-circulation markets, such as the European Union, the package will be inspected (and, in some cases, temporarily apprehended) by border control authorities. This problem will be aggravated if the shipped items contain certain materials, such as exotic furs, or if the available import legislation of the shipping country is misinforming or its execution by some means ambivalent. It is also pertinent to mention that not all routes are available due to political issues and that this restriction is variable throughout time. Secondly, customers can choose among different kinds of shipping services, namely, *Standard* (via ground), *Express*, *Same Day* or *Click and Collect,* within the scope of feasibility (for example, transcontinental shipments are not available via *Standard* routes). As the last two services represent a minor share of Farfetch orders, only the remaining two will be covered in this project.

Step 6 is primarily carried out by DHL or UPS, depending on the routes. In case of cross-border shipments, these are always provided by DHL. In addition to those, when the shipping address is isolated, smaller carriers are subcontracted by DHL or UPS to conclude delivery, (this outsourcing is not currently registered in Farfetch database). As a result, exact delivery time in remote areas can be unreliable. Moreover, given a certain combination Origin/Destination, the connecting route is not static, as it depends on the courier partner space availability and logistics organization. As a consequence of both the formerly exposed reasons and the fact that it is by far the longest one (Table 2), *Step 6* will be subject to higher emphasis than the remaining throughout this dissertation.

Table 2 - Main Net Timespan descriptive statistics by Step (in days)

Step	**1**	**2**	**3**	**4**	**5**	**6**
Mean	0.43	0.02	0.14	0.06	0.49	2.33
Median	0.3	0.0	0.0	0.0	0.3	1.9
St. Dev	0.57	0.55	0.55	0.42	0.55	2.17

Estimated delivery date depends on the time spent on each one of these steps. In turn, the outcome of each step is determined by a set of characteristics of the particular order. Some information regarding what frequently impacts on order processing performance was provided by internal teams. Based on these suggestions, several order features were chosen for each step as the factors that potentially influence its timespan. Those will be the inputs of the model, in case their significance is confirmed.

However, total delivery time is not merely the sum of the time spent in each step since *Steps 1 and 2* occur in parallel. Thus, onwards, estimated delivery time will be defined as the maximum

estimation among *Steps 1* and *2* plus the sum of the estimated times of the four remaining steps. Considering that each step is independent from the remaining, this project will be therefore divided in 6 sub-problems. For each one of these, data regarding relevant dependent and independent variables will be collected and treated accordingly.

3.2 Nature of data

In this section, the nature of data relating to each step will be introduced, so that the present problem takes a more understandable form prior to the presentation of the proposed solving method.

Following the previously mentioned suggestions of the Farfetch team, it was critical to discover where relevant information was stored and to develop a methodology to extract it. However, although some of this data was directly available in the Farfetch database, one of the first obstacles of this project was the fact that some variables required additional calculation or classification to be obtained. Furthermore, given the high volume of data, these calculations had not only to be accurate, but also to be performed efficiently, since they would not be concluded otherwise. Within this section, the focus will be on those variables which extraction was straightforward and the remaining will be explained in the next chapter.

Additionally, a very important feature of data is its dimension. Since, with regard to a fair representation of the fashion cycle, data comprises a complete one year of Farfetch selling activity, the total number of orders within this period surpassed one million. As a result of data cleaning processes, however, the total number of registers for each step is slightly variable, surpassing one million records for all steps.

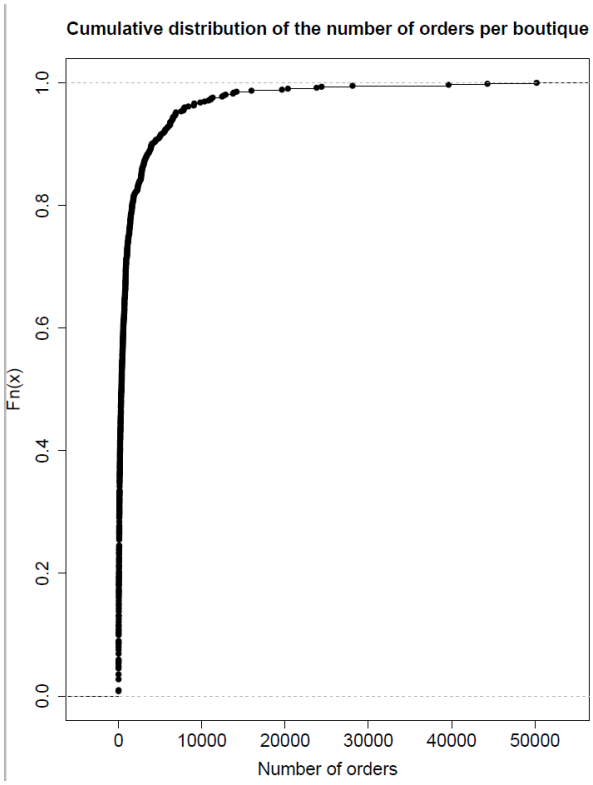
3.2.1 *Step 1 and 3: Check Stock and Decide Packaging*

As both these steps take place in the boutiques and factors that impact on corresponding timespans are expected to be similar, they will be treated jointly for the purpose of selecting the independent variables. Hence, as suggested before, relevant factors concerning these two steps are: *Boutique*, *Weekday*, *Backlog* and *Promotions*. Among those, only the two first can directly be obtained from the database.

3.2.1.1 Boutiques

Within the one-year period of study, 624 *boutiques* (including Brazilian) operated as Farfetch partners, although total sales volume is highly heterogeneous across them. As Figure 7 illustrates, the large majority of the boutiques have sold less than ten thousand orders during the period covered by this analysis. In fact, half of the stores have a corresponding order volume inferior to 400 parcels and nearly half of total order volume is supplied by only 22 stores. This is aligned with the fact that order volume mean is significantly higher than its median. Due to the importance of these stores to Farfetch business and to the present analysis, the corresponding selling volumes are presented bellow in

Figure 8. At first glance, one can perceive that the majority of these are Italian, an inequity that is also extended, although not at the same proportion, to the remaining stores, as more than one hundred of those are also Italian and 54% of the orders are supplied by this country. As a consequence of the dissimilarities between stores, those will be treated differently in the following chapters, according to their prominence.



Descriptive statistics	Value
Minimum	1
First Quartile	83
Median	386
Mean	1809
Third Quartile	1387
Maximum	54770

Figure 7 - Order volume per boutique: main descriptive statistics

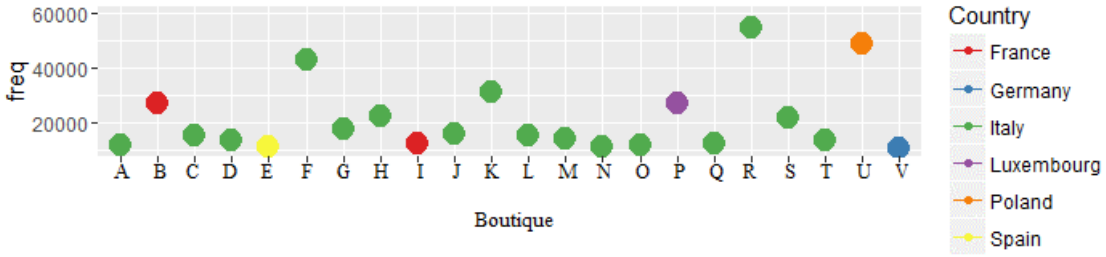


Figure 8 - Sales Volume of the main stores

3.2.1.2 **Weekday**

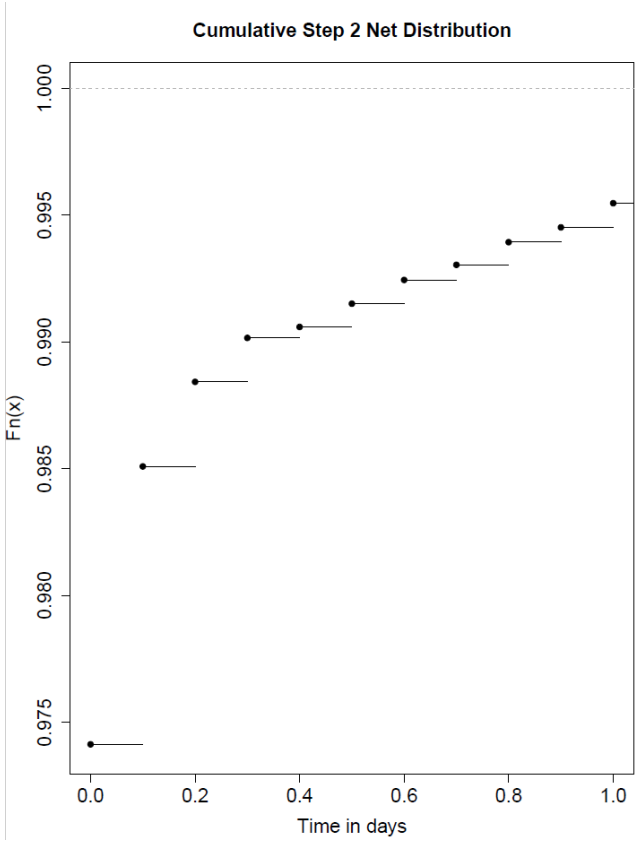
Since boutiques do not operate during weekends and holidays, orders that are placed during the weekend cannot be as promptly satisfied as the remaining. Thus, when stores resume their work on Monday mornings, there is a *backlog* of all the orders that have been requested during the weekend, which further delays their response. As a consequence of this, orders that are placed on Saturdays are expected to have the longest *Step 1* timespans, followed be the orders requested on Sundays. Furthermore, since Monday's workload is also increased due to this *backlog*, orders requested on this day can also suffer delays on this step. With regard to these facts, weekdays will be divided in four categories: Saturdays, Sundays, Mondays and others. Since weekly order request is evenly distributed, the last group will be significantly larger than the others.

3.2.2 ***Step 2: Approve Payment***

According to the *Fraud Team*, *Step 2* does not often impact negatively on delivery time (in addition to occurring in parallel with *Step 1*, it is most of the times executed automatically).

Moreover, factors impacting on this step timespan are not as noticeable as the ones that characterize other steps. These declarations were confirmed by the data collected during this project, as the distribution of the time spent on *Step 2 Net* (excluding weekends and holidays) is the one with the lowest values. As a matter of fact, above 97% of the orders were automatically approved or dismissed, leading to null median and quartile values for *Step 2* Net timespan. Accordingly, on average, *Step 2* is completed in less than 30 minutes. Figure 9 illustrates this information. (Note that cumulative distribution is not continuous, since time was measured with one decimal place.)

Concerning this step, factors that possibly impact on its performance are the following: *Weekday* and *Shipping Location*. The first one was chosen for the reasons that were revealed in the previous section (and will be present across all steps) and the later due to the suspicion that some markets possess a higher level of fraud. Variables related to shipping will be carefully described in the section regarding *Step 6.*



Descriptive statistics	Value
Minimum	0
First Quartile	0
Median	0
Mean	0.017
Third Quartile	0
Maximum	79.1

Figure 9 - *Step 2 Net* main descriptive statistics

3.2.3 *Step 4: Create Shipping Label*

Similarly to *Step 2, Step 4* is also automatic and instantaneous for the vast majority (95 %) of the orders, however the processing time of other orders is significantly larger (average *Step 4 Net* timespan of the remaining 5% orders is 1.24 days). Taking in consideration what has been previously exposed, selected variables for this step are: *Weekday, Shipping Country* and *IsMySwear*. Shipping Country is believed to be relevant since geography, language and culture are expected to influence the likelihood of misfiling shipping information and the time of consequential corrections. *IsMySwear* is a binary variable that was created within the scope of this project. *MySwear* is a brand of shoes that are available on Farfetch selling platform and which delivery time is overextended as a result of them being personalized and not seldom made of exotic materials. *My Swear* orders, nevertheless, were less than 200 in the given period.

3.2.4 *Step 5: Send parcel*

As *Step 5* comprises the bridge between the boutiques and the carriers, it depends on the performance of both entities. As such, the variables designated to explain its performance were those that are expected to define *Steps 1 and 3*: *Boutique*, *Weekday*, *Backlog* and *Promotions*, with the addition of a variable that conveys information concerning the relationship between the stores and the courier. This variable, *IsDailyPickup*, is a binary factor that indicates if a certain store receives a carrier every day. Boutiques with the higher order volumes are expected to be those for which this condition is affirmative, otherwise they would have to manually schedule a pickup on a daily basis. However, Figure 10 is not completely aligned with this idea, which may be due to database flaws or inefficient business models from the boutique side.

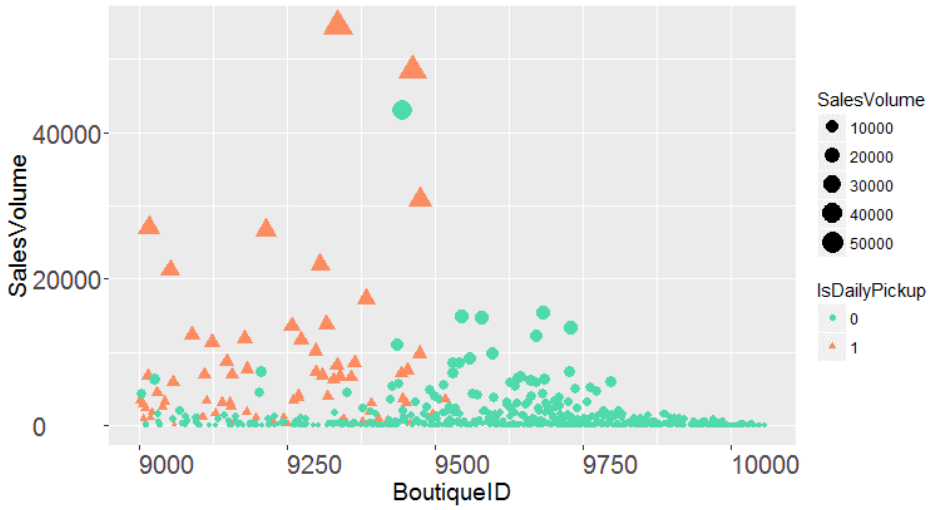


Figure 10 - Sales Volume by Boutique

3.2.5 *Step 6: In transit*

Aligned with its complexity and high variability, *Step 6* is expected to be the one whose timespan depends on a higher number of factors. Therefore, the following variables were selected: *Route* (combination of origin and destination), *Service Type*, *Weekday*, *Is Exotic*, *Border Control* and *Border Trouble Index*. Although distance plays a major role in shipping time, it was not considered a prime variable, as it is determined by the *Route*. However, *Distance* was regarded in the Data Mining process, since it may be a feasible alternative factor to estimate shipping time when data in relation to a certain *Route* is insufficient. Among the variables listed beforehand, only *Route*, *Service Type* and *IsExotic* can be considered direct outputs of Farfetch Database.

Routes

Concerning *Routes*, the first subject of questioning is at what level they should be defined: continent, country, city, etc. On one hand, lower levels lead to more tailored estimations, although, on the other, the larger the number of routes, the smaller the size of data available for each one, which hampers the performance of the estimation model. Furthermore, the quality of the data is also relevant, as the lowest geographical level that the used database stores in a standardized manner is the *country*. Given these pieces of information, the first approach to routes was country-wise. Subsequent contact with data, however, made it evident that some countries were too large both in geographic and buying dimensions not to be subdivided. The main example of this are the United States that accounted for the destination of 28% of Farfetch orders. Due to this fact, the United States were analyzed on a *State* level, which explains why onwards origin and shipping countries and independent territories will also be named as *States*. Still regarding subdivisions within the United States, since the District of Columbia is not a

formal State and related data was not sufficient to produce a separate class, it was treated together with the State of Maryland, since its capital is the nearest to Washington D.C. Other large countries, such as Russia, China, Brazil and Canada, were not divided since there was no automatic way of doing so.

On a continent level, 89% of the orders are shipped from Europe, 6% from North America, 4% from South America and the remaining from Asia, Oceania and Africa. Order destinations are more evenly distributed, as 33% of the orders were sent to Europe, 30% to North America, 24% to Asia, 7% to Oceania, nearly 6% to South America and the other residual orders to Africa. Figure 11 outlines the routes on this level. The inner circle represents origin continents whereas shipping continents can be observed in the outer one.

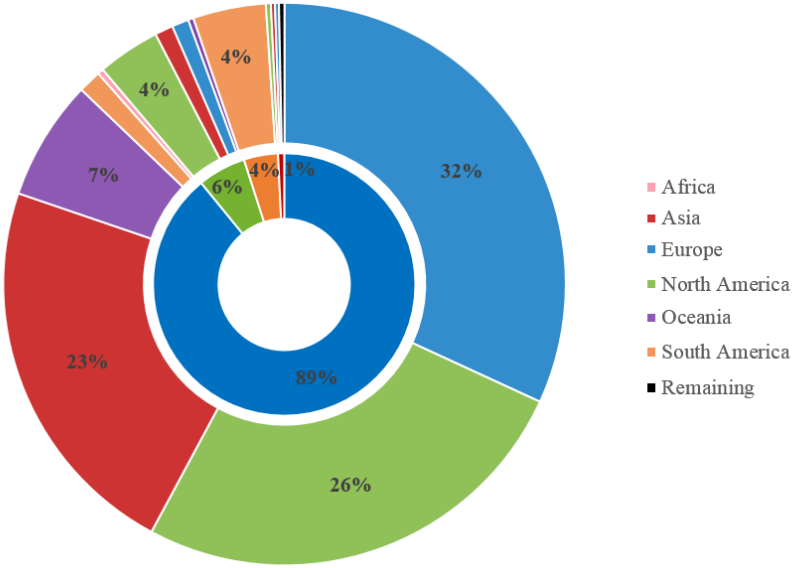


Figure 11 - Continent to continent combinations

During the period covered by this study, boutiques operated in 33 different countries, corresponding to 53 different states, and orders were shipped to 163 different countries, subdivided in 217 states. As a result, 5180 *Routes* can be traced on a state level. These routes, however, have highly disparate levels of usage, as Figure 12 aims to highlight, considering both Country and State as geographical dimentions to build the routes. Similarly, Table 3 and Figure 13 represent the main routes on a country level. An analogous graph can be found in annex, regarding smaller routes.

Concerning the routes on a state level, 4397 of these, 85% of the total set, were used less than 100 times, representing only 6% of the orders of the given period. On the other hand, the top 10 most significant routes, listed in Table 4, account for 33% of the orders.

Table 3 - Cumulative distribution of route usage in Country and State levels – part 1

Level	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Country	1	2	10	398.5	59	51040
State	1	2	7	198.2	40	51040

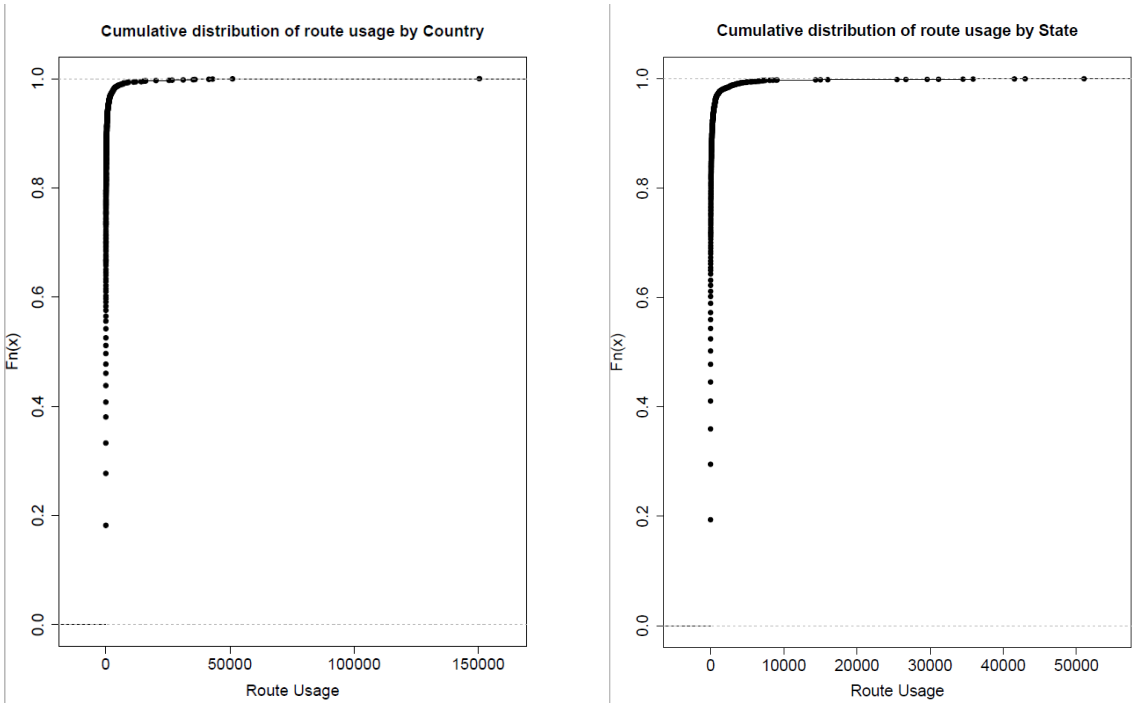


Figure 12 - Cumulative distribution of route usage in Country and State levels - part 2

Table 4 - Top 10 Routes

From	To	Frequency
Italy	United Kingdom	5.0%
Brazil	Brazil	4.2%
Italy	Australia	4.1%
	Hong Kong	3.5%
	California	3.4%
	Russian Federation	3.0%
	New York	2.9%
	Germany	2.6%
	Korea, Republic of	2.5%
	China	1.6%

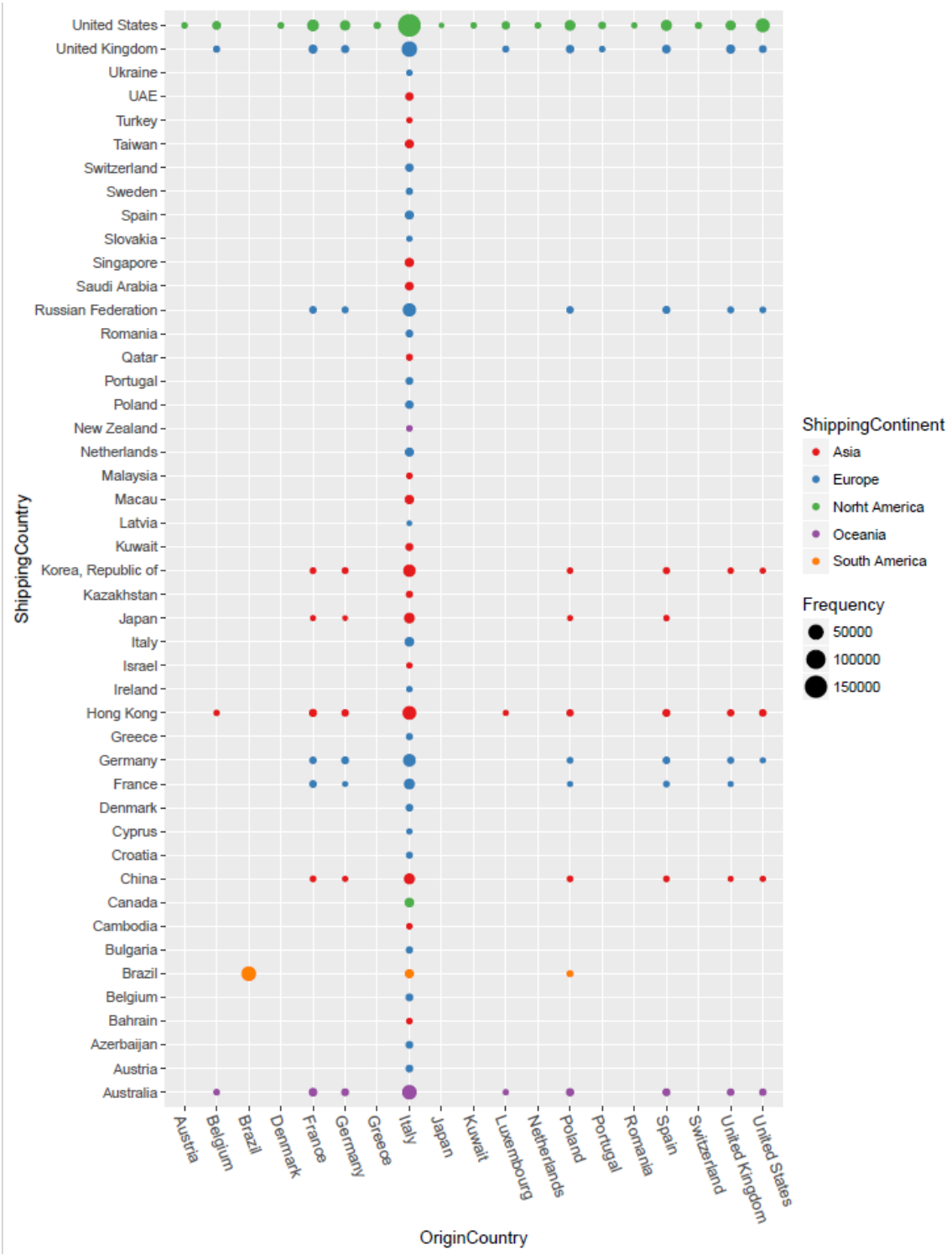


Figure 13 - Main country to country routes (frequency>=1000)

Service

Due to the fact that the majority of the shipments are transcontinental, most of the orders are sent by *Express* services. Furthermore, as the client is at liberty to request this type of service even when it is not strictly necessary, only 19% of the orders are shipped via *Standard* service.

Is Exotic

An item is considered Exotic when it is made of exotic furs or materials, a feature that may lead to delays in customs, hampering *Step 6* performance. One order, however, may include several

items, all shipped together in one package. As such, an order is considered exotic when at least one of these items is exotic, which was determined by searching the substring "exotic" in the materials description in the database. In the period of this study, 40% of the orders included exotic items, supplied by the majority of the boutiques.

3.2.6 *Data summary*

In conclusion, Table 5 summarizes the factors selected to estimate the timespan of each step. Factors displayed in italic are those that will be introduced in the next chapter, since they required deeper analysis.

Table 5 - Independent variables per Step

	Steps				
Variables	**1 + 3**	**2**	**4**	**5**	**6**
Backlog	X			X	
BorderControl					X
BorderTroubleIndex					X
Boutique	X			X	
Daily Pickup				X	
Distance					(X)
IsExotic					X
IsMySwear			X		
Promotions	X			X	
Route					X
Service Type					X
Shipping State		X	X		
Weekday	X	X	X	X	X

3.3 Current solution

Among the factors described in the previous section, only quite a narrow set is considered by the current delivery time estimation tools. These provide information both to the customers and Farfetch itself.

Upon checkout, the customer is given an estimation of the total shipping time, although this information is presented as "Delivery time", misguiding the client. Moreover, there are only three possible estimation values: two regarding *Express* services, according to destination group (Europe & USA or Rest of the world), and one for *Standard* routes. Afterwards, once *Step 6* begins, the customer receives an e-mail containing a link for the courier (DHL or UPS) website, in which an updated delivery time estimation and ongoing tracking information can be found. This solution has two main drawbacks: foremost, as the information provided at first is too generic, the customer only receives reliable feedback once the courier e-mail is sent, secondly, Farfetch has no control over the accuracy of the information conveyed in the courier's website.

As for the estimation created for internal purposes, this merely covers *Step 6*, based on the continent level. Similarly to the values provided to the customer, this estimation does not take into consideration any other factors, such as the ones presented in the previous section. The new solution, which will be presented in the next chapter, attempts to include these up to now neglected categories.

4 The Project

Generically, project success is determined by the quality of planning and execution stages (Globerson and Zwikael 2002). Hence, the current project also demanded a settled methodology to be developed. Since its purpose was straightforward, the main resolutions were theoretical and operational-related. Considerations about what data to obtain and how to collect it were the first to arise, followed by doubts about how to treat it subsequently. The project is divided in three main stages: *Data Collection*, *Data Preparation* (which includes data cleaning and classification) and *Modelling*. Prior to *Data Preparation*, a brief data analysis was required in order tailor the subsequent activities to the specific outlines of the data.

4.1 Data collection

Concerning the first group of resolutions, it was indisputable that data would have to embody the orders of a certain time period (on behalf of representativeness of the annual fashion cycle, it ranges one year of data, from March of 2015 until February of 2016) and the corresponding independent and dependent variables. As a result, the main data pieces created to foster this project are matrices where each line represents an order. These were extracted from the Farfetch database using *SQL* language. *Data collection* was the most time consuming activity of this project and was even required after the beginning of ensuing phases, such as *Data preparation*. This was due to the fact that database structure was initially hard to tackle and the volume of data to be collected was large (6 data frames with more than 1 million registers and several columns). Data was later restructured using *RStudio* and *Excel*, where parallel analysis were also conducted. Since *Excel* specifications were not suitable for this kind of analysis, *RStudio* was the main tool used in the last stages.

Following the data presented in the previous chapter, variables that were not explained, so far, since they were only created within the scope of this project will be clarified in the following sections. In addition, this chapter will introduce the dependent variables of the model, the timespans of each step.

4.1.1 *New variables*

As stated beforehand, some of the desired input variables of the model could not be directly extracted from the database. Hence, these were alternatively calculated based on the data that was available.

4.1.1.1 Boutique-related Steps

Concerning *Steps 1*, *3* and *5*, two new variables were created. Most importantly, store *Backlog* was indicated by Farfetch Analysts as the most relevant predictor of boutique performance and promptness. Moreover, *Promotions* were also included.

Backlog

Despite its importance, the Farfetch's database did not store any kind of information regarding *Backlog*. In turn, it stored the times that marked the beginning of each step for each order. This piece of information could provide historical *Backlog,* once reshaped, by counting all orders found between *Steps 1* and *5* in a given moment for a single store.

First decision regarding this variable was the format in which it would be stored. In order for *Backlog* to be a fair indicator of boutiques level of unfulfilled workload (and therefore performance) this should be measured daily, for each store. The time chosen was 7 am, since it portrays the moment before boutique opening with a slight margin to tackle possible schedule fluctuations and other irregularities. Implementation, however, revealed unexpected complexity, due to two reasons. Firstly, the database did not collect all this information in local times. While order request time (*Step 1* beginning) was stored in system time (*GMT-0*), *Step 5* ending was stored in local time. Secondly, boutique time zones were not correctly stored in the database, as nearly 30% of the stores were associated to inaccurate time zones. This was spotted by comparing the boutique country to the corresponding time zones. It was particularly significant in North American and Brazilian stores, since Partner Service departments in those markets were not aware of that data field, since it did not impact on their work. Therefore, most of these stores were associated to the default time zone (*GMT Standard Time).* This was corrected during this project.

The main idea behind *Backlog* estimation is to count how many orders were requested to a certain store from 7 am of one day until 7 am of the next, and then subtract how many were dispatched to the carrier (excluding those that were cancelled). This value is then added to the *backlog* of the previous day. As this cumulative approach may lead to the propagation of errors, a more accurate estimation was performed for the first day of each month of the covered period. This consisted on counting the orders that were requested in the previous 30 days that were not sent at that point. Annex B1. Monthly estimation further clarifies this procedure. The reason why this could not be executed for all days is because it is a time and resource-consuming practice.

For the remaining days, two values were necessary (per day and store) for this calculation: the total number of requested orders (*In flow*) and the total number of orders picked by the courier (*Out flow*). As stated before, both flows should be calculated, not based on the standard day (0-24h), but on the *day* definition that results from the decision of calculating *Backlog* at 7 am (7am-7am). Hence, from now on, *D* will stand for the standard day, while *d* will symbolize the 7am to 7am day used in this analysis.

Concerning *In flow*, since this information is stored in GMT time, calculations to determine it depend on the store country or, more precisely, on the time zone. For a Portuguese boutique, for example, *In flow* on day *d* would be the total number of orders requested on day *D* after 7 am plus the orders requested on *D+1* until 7 am. For a boutique on a *+7h Country Offset (COff),* time zone, this would simply be the number of orders requested on day *D* (on system time). Accordingly, Figure 14 and Equations (1) and (2) portray the method used to calculate *In Flow*. The query used to collect this information can be found in Annex B2. In flow*Out Flow* estimation followed a similar, yet simpler logic, due to the fact that Time Zones did not have to be considered (Annex B3. Out Flow).

In the end, a table was created to summarize the information generated so that Net Flow could be obtained (Annex C). From this table, *Backlog* was estimated cumulatively with monthly corrections, as stated before. In order to test the accuracy of this estimation, one store was chosen and *Backlog* was calculated for each day of the covered period. Values were later compared with the estimated ones. As shown in Annex D, this estimation fitted reality.

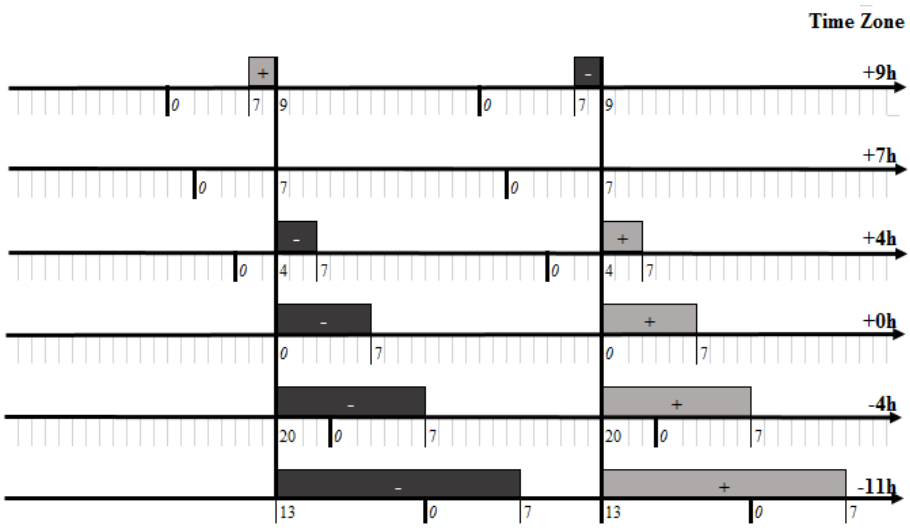


Figure 14 - In flow calculation by Time Zone

$$
\begin{aligned}
if\ COff < 7: In\ Flow_d &= In\ Flow_{D\ GMT} + In\ Flow_{D+1\ GMT\ (h<7-COff)} - In\ Flow_{D\ GMT\ (h<7-COff)} \\
= 7: In\ Flow_d &= In\ Flow_{D\ GMT} \\
> 7: In\ Flow_d &= In\ Flow_{D\ GMT} + In\ Flow_{D-1\ GMT\ (h>24+7-COff)} - In\ Flow_{D\ GMT\ (h>24+7-COff)}
\end{aligned}
\quad (1)
$$

$$
Out\ Flow_d = OutFlow_D - In\ Flow_{D\ (h<7)} + In\ Flow_{D+1\ (h<7)} \quad (2)
$$

By analyzing *backlog* values that resulted from this process, it is perceivable that they range from 0 until 3742, which has to do with both store order number disparities and promotional occurrences. As Figure 15 portrays, the majority of the values are concentrated bellow 500.

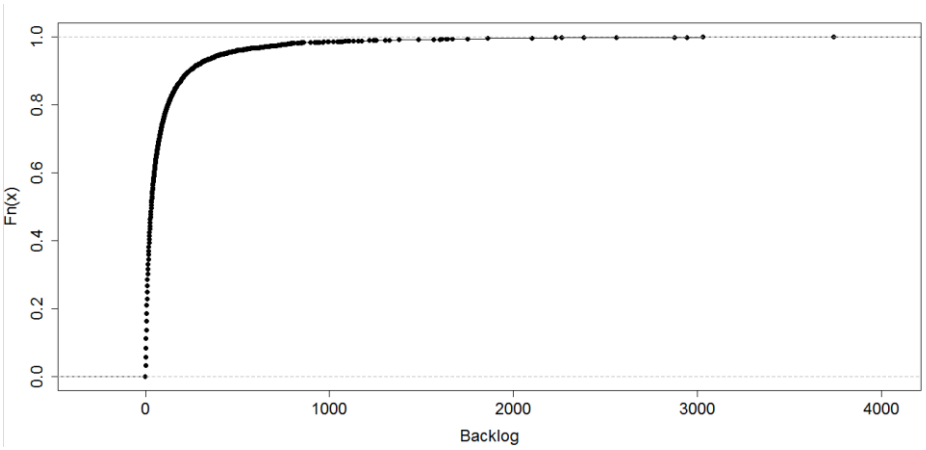


Figure 15 - *Backlog* estimation: cumulative distribution

Accordingly, upon grouping this continuous variable in intervals, unequal ranges were chosen, so that classes were evenly distributed. The 10 final *backlog* classes are displayed in Table 6.

Table 6 - *Backlog* classes

	A	B	C	D	E	F	G	H	I	J
From	0	10	30	50	100	200	300	500	1000	2000
To	10	30	50	100	200	300	500	1000	2000	-

Promotions

Upon the beginning of the present project, Farfetch promotions were recorded in a non-database table, featuring the promotion type, the applicable dates and the targeted markets. Basic promotions types are displayed in Table 7, although combinations of those can also be found

when they overlap. As the purpose of this analysis is to understand the influence of promotions in boutique performance, the specific customer markets for which those are available are not particularly relevant. What is significant, however, is the weight of those markets in overall sales volume. This was calculated for each promotional occurrence, and values widely ranged from 1 (one country only) to 100% (global promotions). Given this fact, a *coverage index* was created and combined with the promotion basic type. This classifies promotions in 3 different levels: *level 1*, when coverage is smaller than 40 %, *level 3*, when more than 80% of customer segment is targeted and *level 2*, for intermediate values.

Table 7 - Promotion types

PromotionID	Promotion	Description
FS	*Free Shipping*	No shipping costs
S	*Sale*	Sale season. Wide-ranging discounts
X10	*X10*	10% discount over sale price
X20	*X20*	20% discount over sale price
X30	*X30*	30% discount over sale price
SP	*Sale Preview*	Earlier sale season for a narrow set of clients
VIPSP	*VIP Sale Preview*	Earlier sale season for VIP customers
X20P	*X20 Preview*	Earlier X20 promotions for a narrow set of clients
PX30	*Private X30*	X30 promotions for a special set of clients

In the end, 15 Promotion Categories resulted from this process (among them, *No Promotion).* These were applicable in 60% of the days.

4.1.1.2 *Step 6*

Two main variables were created for *Step 6: Border Control* and *Border Trouble Index*. In addition to this, a supplementary variable, *Distance*, was calculated.

Border control

According to Farfetch workers, customs are accountable for higher unpredictability and longer delivery times. On top of this, the presence of exotic items can magnify their negative impact. Following this pieces of information, two indexes were created: *Border Control* and *Border Trouble Index*. The first takes the form of a Boolean variable that flags orders that cross borders outside free trade markets, like the European Union[1]. The second is also Boolean and pinpoints those orders that not only undergo border control, but also include exotic items. In the scope of this study, 67% of the orders were subjected to border control, 41% of which included exotic items.

Distance

Despite not being one of the main variables of the model, *Distance* was used to complement the analysis. Since *Step 6* was analyzed on a state level, distances were calculated between the capitals of origin and shipping states. This was conducted using the corresponding capital coordinates and the *Haversine* formula.

4.1.2 *Step timespans*

Step timespans are the dependent variables of each sub-problem and the most important piece of data, since they define total delivery time, as Equation (3) displays.

[1] The EU was considered an overall free trade market in this study, although there are some exceptions, in reality.

$$if\ Step1\ \geq\ Step2\ then\ Delivery\ Time = Step_1 + \sum_{s=3}^{6} Step_s$$
$$else: Delivery\ Time = \sum_{s=2}^{6} Step_s \quad (3)$$

In this project, however, these timespans were not simply determined as the chronological differences between the beginning and ending moments of the corresponding steps. In fact, this would be erroneous since neither the boutiques nor the couriers operate during weekends and holidays, leading to magnified delivery times for those orders that are requested before these days. As a result, this would induce the model to suppose that there was a low performance level associated to these orders. To overcome this problem, timespans were measured excluding weekends and holidays, leading to the variable denomination of *Step x Net*. Nevertheless, new concerns arose from this decision, as weekend and holiday definitions depend on the countries involved, a piece of information that was not stored in Farfetch Database. Moreover, *Step 6* can take place in several countries and the country-crossing moments are not available either.
Following this, the first approach was to study the distribution of delivered parcels throughout the week, on a country level, in order to assess the corresponding weekend patterns. However, conclusions were unattainable in some cases, due to the fact that some samples were too small or data was incoherent. Hence, to support this analysis, *DHL* official working days were compared to delivery data. As a result, country weekends (given by *DHL*) were assumed to be the operating ones, except in those cases where, in fact, more than 10% of the parcels were delivered on those days, in a sample of at least 50 orders. As Table 8 displays, six weekend types resulted from this approach: *Saturday and Sunday* (for the majority of the countries), *Friday and Saturday* (Afghanistan, Israel, Jordan, Oman, etc.), *Friday* (Iraq, Kuwait, Qatar, Saudi Arabia, UAE, Yemen, etc.), *Saturday* (Nepal), *Sunday* (Hong Kong, Lebanon, Macau, Thailand, Uzbekistan, etc.), and *None* (Singapore). Annex E better illustrates prior results and methodology.

Table 8 - Weekday distribution delivery (after corrections)

			Delivered on						
n	**n (%)**	**Weekend**	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
910889	83.38%	Sat + Sun	25%	15%	19%	20%	18%	1%	1%
3200	0.29%	Fri + Sat	13%	15%	21%	20%	3%	1%	28%
127391	11.66%	Sun	20%	8%	18%	20%	19%	13%	1%
4	0.00%	Sat	50%	0%	25%	0%	0%	0%	25%
35227	3.22%	Fri	10%	4%	20%	20%	1%	22%	23%
15746	1.44%	None	14%	3%	24%	21%	17%	10%	11%

This classification was extended to all steps. Regarding the first five steps, Net timespan calculation was made using boutique country holidays and weekend days, whereas *Step 6* took into consideration the shipping country alone. This information was later stored in two tables in Farfetch database. An SQL function that converts a time interval to its corresponding net value, for a given country was initially created. However, since including functions in SQL queries strongly damages their performance, this was later adapted and included directly in the main code. Annex F comprises the final code that was used alternatively.

Step 6 was also used to determine how many decimal places could accurately be used to define the timespan. To do so, DHL logs that indicate the beginning of the step were compared with the first delivery tracking dates for a sample of one tenth of the orders of the covered period. As timespan was measured in days, one decimal place roughly corresponds to 2 hours.

Therefore, since 90% of the values of the sample diverged less than that, step timespans were stored with one decimal place.

Prior to the study and creation of the six models that will deliver the estimated length of each step, it was crucial to grasp the main characteristics of the corresponding datasets, in order to have an overview of the problem and corresponding requirements. Table 9 displays the main descriptive statistics by step. As stated beforehand, *Step 6* is the longest step, which explains the emphasis it is given in this dissertation. Furthermore, as portrayed in the previous chapter, both *Steps 2* and *4* are automatic in the majority of the cases. Across all steps, means expressively higher than medians and large maximum values suggest that data cleaning is crucial in this dataset.

Table 9 - Net Timespan descriptive statistics by Step

Step	Min	1st Q	Median	Mean	3rd Qu	Max	SD	VC
1	0.0	0.0	0.3	0.43	0.6	68.2	0.57	1.32
2	0.0	0.0	0.0	0.02	0.0	79.1	0.55	32.33
3	0.0	0.0	0.0	0.14	0.1	61.0	0.55	3.96
4	0.0	0.0	0.0	0.06	0.0	40.1	0.42	6.71
5	0.0	0.2	0.3	0.49	0.7	28.2	0.55	1.11
6	0.0	1.5	1.9	2.33	2.8	182.6	2.17	0.93

4.2 Preliminary data preparation

As discussed in *State of the Art*, datasets often include records that are not accurate. Hence, outlier detection was conducted, a subject that will be covered afterwards. In addition to that, Farfetch employees from the *Operations* department stated that tuples in which some specific conditions were fulfilled would be undoubtedly incorrect or unsuitable for modelling. There were three main situations covered by this statements, related to different steps. Concerning *Step 2*, sometimes orders have to be investigated, which significantly delays the process. As these occurrences are flagged in the database, it is possible to disregard these values. Once in *Step 5*, the log that defines the end of this step should be introduced by the store. In cases where it is done otherwise (by Farfetch workers, for example), both *Step 5* and *6* timespans should be discarded. At last, in *Step 6*, approximately 8% of the orders are not delivered in the first attempt, by customers' fault, a delay that should not be incorporated in the model. Hence, these records were also removed.

4.3 Univariate analysis and classification

Each one of the 6 datasets included the Net timespan for that specific step and the corresponding values of the independent variables (selected and introduced beforehand) for each order. Due to preliminary record removal and the occurrence of invalid data attributes, matrix size varied according to the step under analysis. It surpassed one million registers in all cases.

Prior to model creation, a univariate analysis was conducted to explore which independent variables significantly impacted on Net timespan and quantify these effects. This was performed for several reasons. Firstly, to check if chosen factors were appropriate and, in case of showing to be irrelevant, consider discarding them from the model. Secondly, concerning variables grouped in classes, explore the possibility of regrouping them, if they display too similar characteristics.

To perform this analysis, both descriptive statistics and hypothesis tests were conducted for each one of the factor levels using *RStudio*. Concerning the hypothesis tests, two unilateral tests on group timespan, *t*, mean were made. For both of them, the null hypothesis was defined as: "Group mean is equal to overall mean", while the alternative hypothesis is either "Group mean

is higher than overall mean" or "Group mean is lower than overall mean". Since standard deviation of the population is unknown, the test statistic, *X*, is assumed to follow a Student's T distribution and is calculated as follows, where *n* is the sample size, *s* the sample standard deviation, $\mu_{t,k}$ the sample *k* mean and μ_t the population mean:

$$X = \sqrt{n} \cdot \frac{\mu_{t,k} - \mu_t}{s} \quad (4)$$

4.3.1 *Steps 1 and 3*

As explained in the previous chapters, independent variables for these two steps are: *Backlog*, *Weekday*, *Promotions* and *Boutique* (Table 5 - Independent variables per Step). Table 10 illustrates the results of univariate analysis for each factor, excluding *Boutique*, which is displayed in Annex H, for the bestselling boutiques. In both tables, extreme *p-values* can be found, which means that the null hypothesis can most of the times be significantly rejected by one of the unilateral tests. Although these results suggest that all factors significantly impact on step timespan, it is relevant to notice that sample sizes are very large, leading to high *x* values, even when the difference of means is not particularly high. On the other hand, when samples are extremely large, their means should tend to the same values, in case factors are not significant. Moreover, the fact that mean timespans of these steps are also small increases the significance of mean differences between groups. These considerations are also applicable for the remaining steps.

As expected, by observing Table 10, *Step 1 Net* increased along with *Backlog* level; this is not particularly significant on *Step 3*. Also aligned with prognosis is the fact that orders requested on Sundays and Mondays spent, on average, more time on *Step 1*, although Saturday orders do not follow the expected pattern. Concerning *Promotions*, although results were not strongly conclusive, this factor was later divided in 2 levels: "Higher than average" and "Lower than average", based on Hypothesis testing results.

Since more than 600 boutiques were covered by this study, regression trees, the method that will be applied to the final estimation, could not be created in *RStudio* without decreasing the number of levels of this factor. Unlike other variables whose values were divided in groups beforehand, *Boutique* is a categorical variable without an implicit numerical order. Therefore, grouping would have to be made according to other *Boutique* characteristics, such as *State*, *Country*, *Order Volume*, etc. In order to explore these possibilities, boutiques were grouped both by *Country* and *Order Volume* and univariate analysis were applied to both categorizations (see Annex I). Since this lead to the conclusion that, for both categorizations, the majority of group levels impacted significantly on Net timespan[2], a classification based on more than one factor was considered. As the purpose of this classification was to estimate Net Timespan, mean *Steps 1 Net* and *Step 3 Net* by store were also found essential to include in this analysis. Due to the fact that variable *Boutique* is also present in *Step 5*, *Step 5 Net* was also included, and resultant classes were used as factor levels for the three steps. *Country* factor, however, had to be discarded because this classification was to be made based on numerical variables only. Lastly, given the 4 final factors (*3 Net Timespans* plus *Order Volume*), Euclidian distance was calculated between each pair of *Boutiques* and those were classified accordingly. Since *Order Volume* and *Net Timespans* have different magnitudes, these factors were converted to standard distributions, which were given the same weight for classification purposes. Corresponding *dendrogram* suggested the ideal number of classes, based on which a non-hierarchical

[2] Please note that, although mean value differences between the groups only have one decimal place, *Steps 1* and *3* are, on average, completed in 10 and 3 hours, respectively

classification was performed, since this category of methods leads to locally optimal solutions. This was made using four different number of classes: 10, 11, 12 and 13.

Table 10 - *Steps 1* and *3* univariate factor analysis

				Hypothesis testing				Step 1 Net			Hypothesis testing				Step 3 Net		
			n	> than mean	p value	< than mean	p value	Median	Mean	sd	> than mean	p value	< than mean	p value	Median	Mean	sd
	ALL		**1128555**					**0,3**	**0,43**	**0,57**					**0,0**	**0,14**	**0,55**
Weekday	*(by order day)*																
	1	*Sunday*	**134181**	Y	1,00	N	0,00	0,4	0,55	0,56	Y	1,00	N	0,00	0,0	0,15	0,56
	2	*Monday*	**161257**	Y	1,00	N	0,00	0,4	0,52	0,64	Y	1,00	N	0,00	0,0	0,16	0,60
	3	*Others*	**696509**	N	0,00	Y	1,00	0,3	0,40	0,53	N	0,00	Y	1,00	0,0	0,13	0,52
	7	*Saturday*	**136598**	N	0,00	Y	1,00	0,1	0,37	0,60	Y	1,00	N	0,00	0,0	0,16	0,59
Backlog		**Intervals**															
	A	*0-10*	**280534**	N	0,00	Y	1,00	0,2	0,30	0,48	N	0,00	Y	1,00	0,0	0,12	0,54
	B	*10-30*	**268208**	N	0,00	Y	1,00	0,2	0,32	0,47	N	0,00	Y	1,00	0,0	0,13	0,63
	C	*30-50*	**139120**	N	0,00	Y	1,00	0,3	0,38	0,51	Y	1,00	N	0,00	0,0	0,15	0,59
	D	*50-100*	**171479**	Y	1,00	N	0,00	0,4	0,44	0,52	Y	1,00	N	0,00	0,0	0,15	0,62
	E	*100-200*	**130093**	Y	1,00	N	0,00	0,4	0,53	0,55	Y	1,00	N	0,00	0,0	0,14	0,37
	F	*200-300*	**51097**	Y	1,00	N	0,00	0,5	0,63	0,63	Y	1,00	N	0,00	0,0	0,14	0,33
	G	*300-500*	**42405**	Y	1,00	N	0,00	0,6	0,74	0,71	Y	1,00	N	0,00	0,0	0,15	0,36
	H	*500-1000*	**29017**	Y	1,00	N	0,00	0,7	0,84	0,71	Y	1,00	N	0,00	0,0	0,18	0,38
	I	*1000-2000*	**11133**	Y	1,00	N	0,00	1,4	1,45	0,74	Y	1,00	N	0,00	0,0	0,19	0,33
	J	*>2000*	**5438**	Y	1,00	N	0,00	2,2	2,24	0,85	Y	1,00	N	0,00	0,0	0,15	0,21
Promotions	**NoPromo**		**457147**	N	0,00	Y	1,00	0,3	0,41	0,52	N	0,00	Y	1,00	0,0	0,13	0,62
	S-1		**73029**	N	0,00	Y	1,00	0,3	0,42	0,54	N	0,84	N	0,16	0,0	0,14	0,46
	S-2		**38754**	N	0,00	Y	1,00	0,3	0,42	0,53	Y	1,00	N	0,00	0,0	0,15	0,33
	S-3		**62913**	N	0,00	Y	1,00	0,3	0,40	0,53	N	0,00	Y	1,00	0,0	0,13	0,41
	X10-2		**3768**	N	0,00	Y	1,00	0,3	0,38	0,53	N	0,01	Y	0,99	0,0	0,13	0,33
	X10-3		**66317**	Y	1,00	N	0,00	0,3	0,56	0,80	Y	1,00	N	0,00	0,0	0,18	0,45
	X20-3		**122211**	Y	1,00	N	0,00	0,3	0,48	0,60	Y	1,00	N	0,00	0,0	0,14	0,40
	X30-3		**40706**	Y	1,00	N	0,00	0,3	0,48	0,59	N	0,00	Y	1,00	0,0	0,13	0,38
	SP-1		**11722**	N	0,00	Y	1,00	0,4	0,37	0,49	N	0,00	Y	1,00	0,0	0,12	0,39
	SP-2		**19944**	N	0,00	Y	1,00	0,2	0,38	0,57	Y	1,00	N	0,00	0,0	0,15	0,32
	SP-3		**38296**	Y	1,00	N	0,00	0,4	0,52	0,59	N	0,00	Y	1,00	0,0	0,12	0,44
	FS-3		**143572**	N	0,00	Y	1,00	0,3	0,40	0,56	N	0,18	N	0,82	0,0	0,14	0,71
	PX30-3		**28917**	Y	1,00	N	0,00	0,3	0,47	0,62	N	0,00	Y	1,00	0,0	0,12	0,39
	S+X20P-3		**13164**	N	0,00	Y	1,00	0,2	0,35	0,52	Y	1,00	N	0,00	0,0	0,15	0,42
	VIPSP-3		**8095**	N	0,00	Y	1,00	0,3	0,35	0,47	N	0,01	Y	0,99	0,0	0,13	0,51

Afterwards, an ANOVA analysis was executed in order to choose the most suitable number, which was proven to be 10, as this was the option with highest F-values[3] (Table 11). It is relevant to mention that results validation was a concern across the whole classification process.

Table 11 - F-values of ANOVA analysis for different number of Boutique Clusters

	10 Classes	**11 Classes**	**12 Classes**	**13 Classes**
Step 1 Net	81069	73980	67773	62512
Step 3 Net	7719,3	7289,5	6631,8	6352
Step 5 Net	97290	89172	82417	75905

Table 12 displays the regression coefficients for each one of the 10 classes, by step. Except for *Step 3*, classes have rather dissimilar characteristics, which supports this method. More information concerning this classification intermediate and final results can be found in Annex J. One detail worth mentioning is the fact that 4 of the 10 classes only include one store. These are Farfetch best-selling stores.

[3] F-value is given by the divion of the variance between the groups by the variance within them.

Table 12 - Regression coefficients by Boutique Class

Step	1			3			5		
Coefficients	**Estimate**	**Std. Error**	**t value**	**Estimate**	**Std. Error**	**t value**	**Estimate**	**Std. Error**	**t value**
Class A	0.61	0.002	263.80	0.05	0.002	21.98	0.67	0.002	293.15
Class B	0.70	0.002	286.10	0.09	0.002	35.31	0.18	0.002	72.08
Class C	0.99	0.003	379.80	0.14	0.003	52.05	0.40	0.003	154.17
Class D	0.46	0.003	148.90	0.06	0.003	19.36	0.43	0.003	138.72
Class E	0.74	0.002	315.10	0.13	0.002	53.88	0.42	0.002	181.16
Class F	0.35	0.001	314.80	0.19	0.001	168.58	0.43	0.001	378.23
Class G	0.43	0.003	164.30	0.15	0.003	58.82	0.60	0.003	230.82
Class H	0.34	0.001	318.00	0.13	0.001	118.81	0.47	0.001	437.11
Class I	0.41	0.001	292.20	0.14	0.001	101.42	0.58	0.001	419.43
Class J	0.33	0.001	284.50	0.13	0.001	114.86	0.57	0.001	492.56

4.3.2 *Steps 2 and 4*

As presented in Chapter 3, both *Steps 2* and *4* are the shortest and, on average, completed in less than 1 and 2 hours, respectively. Factors that were selected to explain *Step 2* Net timespan were: *Shipping State* and *Weekday*. Besides those, *Step 4* was also hypothetically delayed when the order brand was *MySwear*. As Table 13 depicts, *Weekday* effect on time spent on both steps is not particularly relevant for the purpose of this analysis (since it is not distinguishable on the mean values with 2 decimal places, it is smaller than 1 hour), although it is statistically significant, due to the high dimension of the samples (as explained in the previous section).

Concerning shipping states, the only one that indeed impacted on timespans was *Brazil*, which mean values are 0.29 for *Step 2* and 1.09 for *Step 4*. Hence, a Boolean factor, *ToBrazil* was created to flag orders sent to this country. At last, as Table 14 shows, *IsMySwear* impact on *Step 4* is perceivable, although it is not statistically significant due to the small number of *MySwear* orders. Although the creation of these two target variables enriches the current model, they may not be suitable in the long run. This may imply a more exhaustive reassessment of the model afterwards.

Table 13 - *Steps 2* and *4* univariate factor analysis: Weekday

				Hypothesis testing							Hypothesis testing						
				> than mean	p value	< than mean	p value	Step 2 Net			> than mean	p value	< than mean	p value	Step 4 Net		
			n					Median	Mean	sd					Median	Mean	sd
	ALL		**1118219**					0,0	0,02	0,55					0,0	0,06	0,42
Weekday	*(by order day)*																
1	*Sunday*		126507	N	0,771	N	0,23	0,0	0,02	0,17	N	0,25	N	0,75	0,0	0,06	0,42
2	*Monday*		151948	Y	0,977	N	0,02	0,0	0,02	0,28	Y	1,00	N	0,00	0,0	0,07	0,45
3	*Tuesday*		170481	N	0,212	N	0,79	0,0	0,02	0,20	N	0,05	Y	0,95	0,0	0,06	0,41
4	*Wednesday*		171788	N	0,054	N	0,95	0,0	0,02	0,19	N	0,00	Y	1,00	0,0	0,06	0,41
5	*Thursday*		154126	N	0,35	N	0,65	0,0	0,02	0,20	N	0,88	N	0,12	0,0	0,06	0,42
6	*Friday*		158989	N	0,02	Y	0,98	0,0	0,02	0,18	Y	0,99	N	0,01	0,0	0,06	0,42
7	*Saturday*		129796	Y	0,97	N	0,03	0,0	0,02	0,17	N	0,00	Y	1,00	0,0	0,06	0,39

Table 14 - *Steps 2* and *4* univariate factor analysis: *IsMySwear*

			Hypothesis testing						
			> than mean	p value	< than mean	p value	Step 4		
		n					Median	Mean	sd
	ALL	**1118219**					0,0	0,02	0,55
Is My Swear?	**No**	1123748	N	0,382	N	0,618	0,0	0,06	0,42
	Yes	166	N	0,869	N	0,131	0,0	0,11	0,56

4.3.3 *Step 5*

Step 5 is defined by the factors that were chosen to *Steps 1* and *3* plus *Daily Pickup* (Table 5). Concerning *Weekday* and *Backlog*, conclusions are similar to those of Step 1: weekend proximity and *backlog* accumulation lead to slightly longer timespans. As such, weekday was reshaped in 3 levels: *Friday plus Saturday*, *Sunday plus Monday* and *Others*. *Promotions*, however, impact more significantly on this step than in the remaining ones. Hence, as described for those, this factor was divided in two levels for following analysis. Factor *Daily Pickup* was kept, although it does not embody the most remarkable results. As mentioned before, *Boutique* factor was clustered in classes regarding this step as well. Table 15 and Table 16 depict the relationship between these factors and *Step 5 Net.*

Table 15 – *Step 5* univariate factor analysis I

				Hypothesis testing				Step 5 Net		
			n	> than mean	p value	< than mean	p value	Median	Mean	sd
	ALL							0,3	0,49	0,547
Weekday	*(by order day)*									
	1	*Sunday*	**133049**	Y	0,97	N	0,03	0,3	0,49	0,52
	2	*Monday*	**160050**	Y	1,00	N	0,00	0,3	0,50	0,55
	3	*Tuesday*	**180277**	N	0,00	Y	1,00	0,3	0,46	0,52
	4	*Wednesday*	**181206**	N	0,00	Y	1,00	0,3	0,44	0,50
	5	*Thursday*	**162066**	N	0,00	Y	1,00	0,3	0,43	0,53
	6	*Friday*	**166228**	Y	1,00	N	0,00	0,3	0,56	0,63
	7	*Saturday*	**135343**	Y	1,00	N	0,00	0,4	0,59	0,55
			1118219							
Backlog		**Intervals**								
	A	*0-10*	**276888**	N	0,00	Y	1,00	0,3	0,43	0,52
	B	*10-30*	**265298**	N	0,00	Y	1,00	0,3	0,46	0,52
	C	*30-50*	**137826**	Y	0,98	N	0,02	0,3	0,49	0,54
	D	*50-100*	**170339**	Y	1,00	N	0,00	0,3	0,52	0,57
	E	*100-200*	**129348**	Y	1,00	N	0,00	0,3	0,54	0,58
	F	*200-300*	**50752**	Y	1,00	N	0,00	0,3	0,55	0,58
	G	*300-500*	**42245**	Y	1,00	N	0,00	0,3	0,57	0,60
	H	*500-1000*	**28986**	Y	1,00	N	0,00	0,4	0,67	0,64
	I	*1000-2000*	**11114**	Y	1,00	N	0,00	0,4	0,55	0,46
	J	*>2000*	**5423**	Y	1,00	N	0,00	0,4	0,60	0,43

Table 16 - *Step 5* univariate factor analysis II

		Hypothesis testing					Step 5 Net		
		n	> than mean	p value	< than mean	p value	Median	Mean	sd
	ALL						0,3	0,49	0,547
Promotions	**NoPromo**	**452872**	N	0,00	Y	1,00	0,3	0,44	0,50
	S-1	**72515**	N	0,03	Y	0,97	0,3	0,49	0,55
	S-2	**38624**	Y	1,00	N	0,00	0,3	0,51	0,54
	S-3	**62596**	Y	1,00	N	0,00	0,3	0,50	0,55
	X10-2	**3744**	N	0,00	Y	1,00	0,3	0,44	0,50
	X10-3	**65791**	Y	1,00	N	0,00	0,5	0,66	0,64
	X20-3	**121493**	Y	1,00	N	0,00	0,3	0,55	0,60
	X30-3	**39380**	N	0,00	Y	1,00	0,3	0,48	0,54
	SP-1	**11682**	Y	1,00	N	0,00	0,3	0,52	0,57
	SP-2	**19866**	Y	1,00	N	0,00	0,4	0,60	0,58
	SP-3	**38048**	Y	1,00	N	0,00	0,3	0,54	0,55
	FS-3	**142801**	N	0,00	Y	1,00	0,3	0,48	0,55
	PX30-3	**27751**	Y	1,00	N	0,00	0,3	0,53	0,60
	S+X20P-3	**13015**	Y	1,00	N	0,00	0,4	0,60	0,59
	VIPSP-3	**8041**	Y	1,00	N	0,00	0,3	0,52	0,52
Is Daily Pickup	**0**	**529409**	Y	1,00	N	0,00	0,3	0,51	0,54
	1	**588810**	N	0,00	Y	1,00	0,3	0,48	0,55

4.3.4 *Step 6*

Step 6 is the most complex and wide one. As such, factors selected to explain its behavior are: *Border Control, Border Trouble Index, Is Exotic, Route, Service Type* and *Weekday.* Moreover, *Distance* was also considered as an additional factor.

Table 17 - *Step 6* univariate factor analysis I

				Hypothesis testing						
				> than mean	p value	< than mean	p value	Step 6Net		
			n					Median	Mean	sd
	ALL		**1118219**					1,9	2,33	2,168
Weekday	*(by order day)*									
	1	*Sunday*	121794	Y	1,00	N	0,00	2,0	2,42	2,16
	2	*Monday*	146556	Y	1,00	N	0,00	2,0	2,41	2,27
	3	*Tuesday*	164906	Y	1,00	N	0,00	2,0	2,35	2,28
	4	*Wednesday*	165721	N	0,00	Y	1,00	1,9	2,28	2,32
	5	*Thursday*	148782	N	0,00	Y	1,00	1,8	2,16	1,98
	6	*Friday*	153678	N	0,00	Y	1,00	1,9	2,28	2,10
	7	*Saturday*	125273	Y	1,00	N	0,00	2,0	2,43	1,98
Service	**Express**		831995	N	0,000	Y	1,000	1,9	2,23	2,166
	Standard		194719	Y	1,000	N	0,000	2,7	2,77	2,120
IsExotic	**Yes**		415036	N	0,708	N	0,292	1,9	2,33	2,019
	No		611679	N	0,290	N	0,710	1,9	2,33	2,264
BorderControl	**Yes**		**690089**	Y	1,00	N	0,00	1,9	2,40	2,25
	No		**336626**	N	0,00	Y	1,00	1,8	2,18	1,97
BorderTroubleIndex	**Yes**		**283098**	Y	1,00	N	0,00	2,0	2,44	2,08
	No		**743617**	N	0,00	Y	1,00	1,9	2,29	2,20

As Table 17 displays, orders that are requested during the weekend spend, on average, more time on *Step 6*. Difference, however, as in other steps, is not remarkable and is not straightforwardly sustained by common sense, as orders are supposed to be shipped during the weekdays, regardless the order day. A possible explanation, however, is an overload of shipments to courier partners in the beginning of the week, given the fact that they cease activity during weekends.

Service, on the other hand, affords more assertive conclusions, as *Express* shipments are, on average, performed in approximately less 12 hours than *Standard* ones. Concerning customs issues, as expected, although being exotic *per se* does not significantly impact on shipping time, once associated with border control it slightly intensifies its delaying effect.

With *Route* types ranging from *intracity* to *transcontinental*, this factor was expected to be highly significant. This was confirmed, as information concerning Farfetch 20 most used routes (Table 18) is significantly heterogeneous. This shows that *Route* strikingly impacts on timespan. Moreover, as also confirmed by *Distance* analysis, shipment time is not only explained by distance. For example, shipments within Italy take, on average, approximately as much time as those made between this country and the state of New York. However, mean timespans reveal an overall increasing trend, as distance grows. Exceptions to this proportion are groups *C, D* and *F*. Group *C* incoherently long shipping times can be explain by some important *Routes* within Europe, such as Italy to United Kingdom, Spain, Poland or Portugal, whose means are high. In group *D*, shipping countries like Saudi Arabia, Russia, Kazakhstan and Azerbaijan are responsible for the long transit times. On the other hand, group *F* displays abnormally fast delivery times (given the distance) due to the fact that corresponding *Routes* connect European stores with customers on the east coast of the United States, which are highly accessible by plane.

Table 18 - *Step 6* univariate factor analysis II

			Hypothesis testing				Step 6Net		
		n	> than mean	p value	< than mean	p value	Median	Mean	sd
	ALL	**1118219**					1,9	2,33	2,168
Routes	Italy-UK	**51043**	Y	1,00	N	0,00	2,8	2,93	1,88
	Brazil-Brazil	**42997**	Y	1,00	N	0,00	2,2	2,79	3,32
	Italy-Australia	**41524**	Y	1,00	N	0,00	2,8	2,77	1,50
	Italy-Hong Kong	**35886**	N	0,00	Y	1,00	1,9	2,30	1,27
	Italy-California	**34503**	N	0,00	Y	1,00	1,9	2,05	1,59
	Italy-Russian	**31162**	Y	1,00	N	0,00	2,9	3,47	2,63
	Italy-New York	**29606**	N	0,00	Y	1,00	1,0	1,35	1,41
	Italy-Germany	**26695**	N	0,00	Y	1,00	1,9	2,05	1,34
	Italy-South Republic	**25475**	N	0,00	Y	1,00	2,0	2,28	1,11
	Italy-China	**16022**	Y	1,00	N	0,00	2,1	2,93	2,59
	Italy-France	**15014**	N	0,00	Y	1,00	1,9	2,25	1,54
	Italy-Japan	**14354**	Y	1,00	N	0,00	2,8	2,99	1,55
	Italy-Italy	**9091**	N	0,00	Y	1,00	0,9	1,24	1,02
	Italy-Canada	**8969**	Y	0,98	N	0,02	2,0	2,37	1,92
	Italy-Macau	**8517**	Y	1,00	N	0,00	2,8	3,24	2,19
	Italy-Singapore	**8080**	Y	1,00	N	0,00	2,0	2,52	1,27
	Italy-Florida	**7471**	N	0,00	Y	1,00	1,9	2,01	1,31
	Italy-Brazil	**7257**	Y	1,00	N	0,00	2,0	3,13	4,46
	UK-UK	**7089**	N	0,00	Y	1,00	0,8	0,99	0,89
	Italy-Taiwan	**7061**	N	0,00	Y	1,00	1,9	2,07	1,34
Distance		Intervals (km)							
	A	<750	**114751**	N	0,00	Y	1,00	1,1	1,86	2,70
	B	<1250	**98356**	N	0,00	Y	1,00	1,8	1,91	1,69
	C	<2000	**116241**	Y	1,00	N	0,00	2,7	2,57	1,91
	D	<4000	**94090**	Y	1,00	N	0,00	2,8	3,11	2,54
	E	<6500	**71002**	N	0,00	Y	1,00	1,8	2,27	2,79
	F	<8000	**125910**	N	0,00	Y	1,00	1,8	1,89	1,75
	G	<9000	**96825**	Y	1,00	N	0,00	1,9	2,36	2,21
	H	<10000	**158549**	Y	1,00	N	0,00	2,0	2,40	1,67
	I	<15000	**77168**	Y	1,00	N	0,00	2,0	2,41	2,22
	J	>15000	**73599**	Y	1,00	N	0,00	2,8	2,77	2,01

As factor *Boutique* in other steps, also *Route* has too many levels to be included in a regression tree. Hence, classification was also used in order to cluster this variable. To do so, each route was associated to its usage (number of shipments), *Distance* and mean *Step 6* timespan. Three different solutions were delivered by this method, with 3 different numbers of clusters: 20, 25 and 30. As for *Boutique* clusters, an ANOVA analysis was executed in order to choose the best one. As Table 19 suggests, the chosen number was 20.

Table 19 - F-values of ANOVA analysis for different number of Route Clusters

Number of classes	20	25	30
F-values	39359	33034	27267

Route classes are introduced in Table 20 and more information concerning this solution can be found in Annex L.

Table 20 - Regression coefficients by Route Class

Class	Estimate	Std. Error	t value	Pr(>|t|)
1	2.19	0.01	232.70	<2.2E-16
2	2.09	0.01	206.17	<2.2E-16
3	2.78	0.01	231.90	<2.2E-16
4	2.29	0.01	138.40	<2.2E-16
5	1.73	0.01	187.53	<2.2E-16
6	2.06	0.02	135.62	<2.2E-16
7	2.36	0.01	276.15	<2.2E-16
8	2.96	0.01	206.72	<2.2E-16
9	2.00	0.01	197.61	<2.2E-16
10	1.72	0.01	116.59	<2.2E-16
11	2.85	0.03	109.85	<2.2E-16
12	2.77	0.01	226.28	<2.2E-16
13	2.79	0.02	160.79	<2.2E-16
14	2.85	0.01	204.98	<2.2E-16
15	2.93	0.01	265.51	<2.2E-16
16	1.35	0.01	93.12	<2.2E-16
17	2.20	0.01	236.79	<2.2E-16
18	2.78	0.02	160.98	<2.2E-16
19	2.25	0.02	110.77	<2.2E-16
20	2.30	0.01	2.34	<2.2E-16

4.3.5 *Overall observations*

Univariate analysis has revealed that the majority of the factors are significant, from a statistical point of view. The vast majority of the hypothesis tests were conclusive, although, as stated before, large samples allow conclusive results to be delivered even when mean differences are not outstanding. In the case of the first 5 steps, many factors that are mathematically significant may not be particularly relevant for the purpose of this project, since estimated delivery date value is to be shared with the client in days. For example, although *Weekday* significantly impacts on *Step 5* timespan, the biggest difference between group and overall mean is approximately 2 hours and 20 minutes. Focusing on *Steps 2* and *4*, further irrelevant dissimilarities can be found.

Another fact worth highlighting is the proximity between standard deviation and mean in the majority of the considered groups. In fact, coefficient of variance is often higher than one, implying that there is a high level of dispersion within the groups, which may impact negatively on model quality. This may also be explained by the interaction between variables, which would be better understood in case a multivariate analysis was performed. Nevertheless, the regression model will take the combination of variables into consideration.

4.4 Data cleaning

Following the preliminary elimination of the inaccurate tuples, each one of the 6 datasets was divided in 2 parts: a *training set* (70%), containing the information that would feed the models, and a *testing set* (30%), which would be used to assess their estimation quality. Due to this split, the following sections will describe data analysis that were conducted based only on training data, as the purpose of the testing set is to represent unknown information.

Prior to the creation of the model, data was cleaned so that erroneous values did not damage its performance. As the *testing set* timespans represent what in practice would be the real transit times of those orders for which an estimated delivery date was provided by the model, outlier

removal was conducted in *training sets* only. This was assumed to better fit reality as upon model construction only available data could be treated.

As only one independent variable is present in each data frame (Net timespan), outlier detection has to be based on its value. However, as explained in Chapter 2, it would be biased to calculate the overall mean of this variable for each one of the 6 steps and classify as outliers those tuples for which net timespan is further than a certain distance from this centroid. An extreme example of this misconception would take place in *Step 6*, as longest routes highest timespans would erroneously increase the probability of those being considered outliers. Thus, outlier detection was executed within groups that were formed according to combinations of the relevant factors described in the previous section. For each one of these groups, descriptive statistics were calculated. The general criteria used to cast outliers was the boxplot rule. Those net timespan values that were more than 3 interquartile ranges smaller than the first quartile or larger than the third (severe outliers) would be removed. As expected, with the exception of *Step 6*, outliers were only found on the highest side of the spectrum. In fact, data concerning *Steps 2* and *4* was so asymmetrical that the general criteria had to be reshaped in order to be applicable.

4.4.1 *Step 1*

Following the univariate analysis described beforehand, all *Step 1* factors were combined into groups to detect outliers. Due to the initial high number of combinations, *Backlog* was regrouped in 3 classes (under 100, over 1000 and in between). As such, the final maximum number of classes was expected to be 240 (4 *Weekdays* x 3 *Backlog* x 2 *Promotions* x 10 *Boutique Classes*), a number that was, in reality, reduced to 168 as not all combinations were present in the data.

By executing the steps detailed in Annex M, 1.6% of the values were classified as outliers, a rate that was considered reasonable. Also, as final mean decreased in the direction of the median and maximum value became more moderate (Table 21), this method was assumed to be valid.

Table 21 - *Step 1 Net* distribution before and after data cleaning

	Min	**1st Q**	**Median**	**Mean**	**3rd Q**	**Max**
Before	0.0	0.0	0.3	0.429	0.6	24.6
After	0.0	0.0	0.3	0.394	0.6	5.5

4.4.2 *Step 2*

As *Weekday* did not impact very significantly on *Step 2 Net*, outlier removal for this group was only based on the factor *To Brazil*, which flagged the orders that were shipped to this country. Hence, only two groups were considered in this analysis. However, general outlier detection was found unsuitable for the group of the orders that were not shipped to Brazil, as both first and third quartile values were equal to zero, which led to a null interquartile range. Hence, if this criteria was to be applied, all those values that were different from zero would be discarded.

To overcome this issue, *Step 2 Net* statistics were recalculated without the null values for both groups and resultant quartile statistics were to be used to identify outliers. New *quartile* values concerning orders not shipped to Brazil, however, again led to a null interquartile range, since both first and third quartile values were 0.1. Hence, for this group, interquartile range was considered to be the smallest positive value given the decimal accuracy of the data collected for Net Timespans (0.1). As this would embody a more permissive outlier detection method, the casting criteria was changed, and mild outliers, instead of extreme ones, were discarded. Although this technique is not academically-based, it was a conservative manner of performing data cleaning, as it led to the removal of less than 1% (742) of the group values and no value smaller than 0.25 days was removed. Regarding shipments to Brazil, 3.3% (1245) of those were

considered outliers, although the highest remaining net timespan was nevertheless 2.7 days. Table 22 better depicts this information.

Table 22 - *Step 2 Net* distribution before and after data cleaning

To Brazil			Min	1st Q	Median	Mean	3rd Q	Max
No	**Before**	**With 0**	0.0	0.0	0.0	0.002	0.0	24.3
		Without 0	0.1	0.1	0.1	0.180	0.1	24.3
	After		0.0	0.0	0.0	0.001	0.0	0.2
Yes	**Before**	**With 0**	0.0	0.0	0.0	0.293	0.1	79.1
		Without 0	0.1	0.2	0.6	0.994	1.2	79.1
	After		0.0	0.0	0.0	0.184	0.0	2.7

4.4.3 *Step 3*

As the factors associated with *Steps 1* and *3* are the same, *Step 3* was also divided in 168 groups for outlier detection. However, as *Step 3* values are generally closer to zero than those of *Step 1* (Section 4.3.1), null quartile values and inter quartile ranges were found in 68 of those groups, a problem that was explained in the previous section. Moreover, since these groups comprehend 54.7% of the total data and one fifth of the corresponding values are not null, 11% of data would be automatically rejected in case this criteria was applied. Hence, similarly to the treatment given to *Step 2* data, quartile values were recalculated without the null values and consequent mild outliers were rejected. Overall, less than 1% of data was rejected.

Table 23 - *Step 3 Net* distribution before and after data cleaning

	Min	1st Q	Median	Mean	3rd Q	Max
Before	0.0	0.0	0.0	0.138	0.1	61.0
After	0.0	0.0	0.0	0.125	0.1	5.2

4.4.4 *Step 4*

Similarly to *Step 2* data cleaning, *Weekday* was not considered significant for this phase in *Step 4*. Thus, factors selected to form data groups were: *To Brazil* and *Is MySwear*, which led to the creation of 3 groups, as no *MySwear* items were shipped to Brazil in the covered period. Alike the previously described outlier detection activities, operational obstacles were found in this step, with the exception of the groups concerning shipments to Brazil. As such, for both groups featuring orders shipped to the remaining countries, quartile values were recalculated and outliers were detected accordingly (Table 24). Overall, less than 1% of the values were considered outliers.

Table 24 - *Step 4 Net* distribution before and after data cleaning

To Brazil	*MySwear*			Min	1st Q	Median	Mean	3rd Q	Max
Yes	**No**	**Before**	**With 0**	0.0	0.1	1.0	1.087	1.6	40.1
		After		0.0	0.1	1.0	1.016	1.5	6.1
No	**Yes**	**Before**	**With 0**	0.0	0.0	0.0	0.103	0.0	6.1
			Without 0	0.1	0.1	0.3	0.712	0.7	6.1
		After		0.0	0.0	0.0	0.051	0.0	0.8
	No	**Before**	**With 0**	0.0	0.0	0.0	0.008	0.0	28.7
			Without 0	0.1	0.1	0.2	0.656	0.5	28.7
		After		0.0	0.0	0.0	0.003	0.0	1.1

4.4.5 *Step 5*

As *Step 5* factors are the same as those of *Steps 1 and 3* plus *Daily Pickup*, which was proven not remarkably significant in section 4.3.3, 168 groups were created. With the exception of two groups, interquartile range was positive and standard criteria was applicable. For the remaining two groups, interquartile range was considered to be 0.1, as it has been done before given the same situation. As a result, approximately 3% of the values were considered outliers, and *Step 5 Net* distribution was reshaped as displayed in Table 25.

Table 25 - *Step 5 Net* distribution before and after data cleaning

	Min	1st Q	Median	Mean	3rd Q	Max
Before	0.0	0.2	0.3	0.491	0.7	28.2
After	0.0	0.2	0.3	0.451	0.7	4.3

4.4.6 *Step 6*

As mentioned before, *Step 6* is the one with the highest variability. It is also the one for which operational performance is given less focus, as multiple carriers may be involved in the same shipment and, unlike stores, those are not evaluated nor compensated for the service they provide. Hence, outliers are expected to have a larger weight in data that portrays this step.

All *Step 6* factors were used to create the groups for this step. However, as 5180 *Routes* can be found and, among those, 4399 were used less than 100 times, the number of levels of this factor had to be reduced so that groups were large enough to create reliable quartile values. Hence, *Routes* for which frequency was smaller than 100 were grouped together in a residual group, reducing the total number of *Route* categories to 781. Although the 2873 final groups led to the removal of 4% of the values, a final maximum shipping time of 21 days was still not detected. Null shipping times correspond to *Click & Collect* orders, a service that allows the customer to collect the package in the store.

Table 26 - *Step 6 Net* distribution before and after data cleaning

	Min	1st Q	Median	Mean	3rd Q	Max
Before	0.0	1.5	1.9	2.327	2.8	180.6
After	0.0	1.5	1.9	2.152	2.8	20.7

4.5 Modelling

Data cleaning led to data reshaping as distinctively high timespan values were casted out and final mean values decreased towards median ones. Although, given the volume of data, no automatic procedure was (or can be) executed to verify that no accurate values were eliminated and no erroneous ones remained in the dataset, this evidence supported the assumption that data became more suitable to perform modelling activities.

As described in Chapter 2, regression trees are appropriate structures to support predictions given the nature of the present data, as no underlying theoretical distribution is required for the variables[4] and multiple factors are allowed. As *R* provides both theoretical and operational support to implement this methodology, this step was majorly conducted in *RStudio. Partykit* package function *ctree* creates regression trees by recursively partitioning independent variables. (Hothorn and Zeileis 2015). *Ctree* conditional inference trees are built according to the bellow described pseudo-code (Hothorn, Hornik, and Zeileis 2006):

Do while tree is not complete

For all pairs of input and response variables

Test null hypothesis of independence between them

If the null hypothesis is not rejected

Create a binary split in the selected variable

Exit For

End if

Next

If all hypothesis were rejected within the **For** structure

Select input variable with the strongest association with response

Create a binary split in the selected variable

End if

End while

4.5.1 *First model*

For each one of the previously described six steps, a regression tree was created using these tools. As a result, an expected (mean) value was determined for each terminal node. Afterwards, these trees were used to predict the values of the dependent variables of the Testing data frames. These estimations were then compared to the real values in order to obtain several mean error measures: *Mean Error (ME), Mean Absolute Error (MAE), Mean Delay (MDelay), Mean Squared Error (MSE), Mean Percentual Error (MPE)* and *Mean Absolute Percentual Error (MAPE).* Annex N displays the code used to perform this methodology, spanning all these stages. Table 27 includes the average results by step and, as an example, Annex O displays the *Step 1* regression tree.

Non-absolute error measures can lead to inaccurate conclusions regarding the prediction quality of the method since positive and negative error values offset each other. However, null mean error values are a proxy for a non-biased model, since errors expected value should be zero. Concerning other mean error measures, *Steps 2* and *4* are those for which absolute prediction

[4] Both independent and dependent variables distributions are irregular, including both categorical and numerical variables.

accuracy is more favorable, which can be explained by their data homogeneity. In the scope of these two steps, *mean percentual* and *absolute mean percentual errors* are extremely biased statistics, as they exclude records with null values, which, as described beforehand, are the majority of the data.

Table 27 - First regression tree results in days (D) and hours (H)

	Step													
	1		**2**		**3**		**4**		**5**		**6**		**Total**	
	D	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**
ME	0.0	-0.9	0.0	0.2	0.0	0.5	0.0	0.3	0,0	1,0	0.2	4.1	0.2	5.2
MAE	0.3	7.4	0.0	0.4	0.2	4.6	0.0	1.1	0,3	8,2	0.8	19.5	1.6	41.2
Mdelay	0.2	4.1	0.0	0.3	0.1	2.6	0.0	0.7	0,2	4,6	0.5	11.8	1	24.1
MSE	0.3	6.8	0.0	0.7	0.3	7.4	0.1	2.9	0,3	6,7	4.3	102.6	5.3	127.1
MPE	-21%		61%		41%		-21%		-51%		-20%			
MAPE	69%		83%		61%		98%		84%		41%			

Focusing on boutique-related steps, results interpretation is not as straightforward. At first glance, errors can be perceived as non-relevant, especially by considering values in days. Hence, in order to better evaluate model quality, mean and median values of timespan distributions should be considered (Table 28). By analyzing errors and by comparing their magnitudes to the ones of corresponding mean timespan values, one can realize that especially when *Step 3* is concerned, *mean absolute errors* are quite substantial. *Mean squared error* emphasize this idea.

Table 28 - Main Net Timespan descriptive statistics by Step

Step	**1**	**2**	**3**	**4**	**5**	**6**
Mean	0.43	0.02	0.14	0.06	0.49	2.33
Median	0.3	0.0	0.0	0.0	0.3	1.9
St. Dev	0.57	0.55	0.55	0.42	0.55	2.17

As expected, *Step 6* is the one with highest error measures, with the exception of percentual measures (which is, in a relative manner, a positive indicator and covers almost all the data in this step). Moreover, an average 40% error in estimating transit times does not express a positive evaluation of the model and discrepancies between *mean absolute* and *mean squared errors* indicate that there are very large absolute error values.

Globally, conclusions regarding the quality of the prediction model are hard to asses based on these pieces of information only. This is due to many reasons. Firstly, large error values can be due to genuinely unpredictable data. Secondly, wrong factors or the wrong kind of model may have been chosen. At last, as testing sets were not subjected to data cleaning procedures, these still include outliers, which largely increase mean error measures, as means are strongly impacted by extreme values.

In order to have a clearer notion of the quality of the current prediction model, another model was created disregarding all factors. As such, the timespan estimation of each tuple would be equal to that step timespan mean value. In case this *control* model results (Table 29) are not worse than those displayed in Table 27, either the factors are meaningless or the methodology is not suitable for this problem.

Although Table 29 depicts overall worse results, this difference is not remarkable enough to support the assumption that the first model is fully developed and does not need further improvements. On the other hand, at this point, there is no evidence that better results can be achieved. On the verge of a set of unknown possibilities and outcomes, prior assumptions and

decisions were reanalyzed and reconsidered for *Steps 1, 3, 5* and *6*, since the results of the remaining two steps were considered satisfactory

Table 29 - Control regression tree results in days (D) and hours (H)

	Step													
	1		**2**		**3**		**4**		**5**		**6**		**Total**	
	D	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**
ME	0.0	-0.9	0.0	0.0	0.0	0.0	0,0	0,0	0,0	0,0	0,0	-0,1	0	-1
MAE	0.3	7.4	0.0	0.8	0.2	5.1	0,1	2,8	0,4	9,2	1,0	24,9	2	50,2
Mdelay	0.2	4.1	0.0	0.4	0.1	2.5	0,1	1,4	0,2	4,6	0,5	12,4	1,1	25,4
MSE	0.3	6.8	0.0	0.8	0.3	7.4	0,2	4,2	0,3	7,1	459,7	11031,7	460,8	11058
MPE	-27%		91%		41%		84%		-78%		-43%			
MAPE	74%		91%		61%		84%		111%		64%			

4.5.2 *Second Model*

One of the main characteristics of this project is the heterogeneity of data, a fact that is also conveyed by the high number of levels of certain variables. Due to this fact, categorical variables *Boutique* and *Route* were classified in clusters (c.f. Section 4.3), so that a tree model could be created.

Concerning *Boutique*, a classification was made based on the number of orders requested from each store and the average net timespans of the related three steps. This classification was then used as a factor in those steps. Although the selection of order frequency as an input allowed the creation of single classes for the best-selling boutiques, creating a unique classification targeting the three steps simultaneously might have led to a misemployment of timespan information. Instead, a classification could have been performed for each step based on the corresponding timespan data, which was attempted for this second modelling phase. Table 30 depicts the new classification criteria and the corresponding cluster sizes for each step.

Table 30 - New *Boutique* classification criteria and cluster size

			Step		
From	**To**	**Cluster**	**1**	**3**	**5**
0.0	0.1	**A**	6	340	143
0.1	0.2	**B**	91	106	21
0.2	0.3	**C**	174	46	35
0.3	0.4	**D**	100	26	96
0.4	0.5	**E**	90	22	122
0.5	0.6	**F**	60	18	69
0.6	0.7	**G**	30	7	46
0.7	0.8	**H**	23	15	37
0.8	1.0	**I**	24	16	26
1.0	-	**J**	25	27	28
		Total	623	623	623

Routes classification was executed taking *Frequency*, *Distance* and *Average Step 6 Net* in consideration. Hence, similarly to what was described concerning *Boutique*, too many variables might have been used. As such, *Routes* were reclassified solely based on *Average Step 6 Net* values (Table 31).

Table 31 - New *Route* classification criteria

From	To	Cluster	n
0	1	**A**	370
1	1.5	**B**	489
1.5	2	**C**	1050
2	2.5	**D**	782
2.5	3	**E**	639
3	3.5	**F**	442
3.5	4	**G**	323
4	5	**H**	307
5	6	**I**	118
6	10	**J**	176
10	-	**K**	113

Following the reclassification of *Boutiques* and *Routes*, the same modelling methods that were applied for the first model were conducted. Table 32 depicts new regression results. Although *Step 3* new prediction accuracy decreased and *Step 5* error data does not support conclusive statements, both *Steps 1* and *6* overall error values decreased slightly.

Table 32 - Second regression tree results in days (D) and hours (H)

	Step									
	1		3		5		6		Total	
	D	H	D	H	D	H	D	H	D	H
ME	0.0	0.8	0.0	0.0	0.1	2.7	0.2	4.1	0.3	7.6
MAE	0.3	6.9	0.2	5.3	0.3	7.9	0.7	17.8	1.5	37.9
Mdelay	0.2	3.8	0.1	2.7	0.2	5.3	0.5	11.0	1	22.8
MSE	0.3	6.1	0.3	8.3	0.3	6.9	4.0	96.9	4.9	118.2
MPE	-20%		45%		-22%		-15%			
MAPE	67%		73%		70%		36%			

4.5.3 *Third Model*

After the conclusion of the second model, a less theoretical approach was considered. This consisted in discarding classification of both *Boutiques* and *Routes*, using this factor directly at the *Boutiques* or *Route* level. Since the number of factor combinations would make it unfeasible for *RStudio* to create conditional decision trees, descriptive statistics were simply calculated for each group and mean net timespan values were assumed to be the expected timespan values for that combination of independent variables. This was calculated using the *describeBy* function from package *psych.* Table 33 portrays the results of this methodology, the one that delivered smaller error statistics up to the moment (with the exception of *Step 3*).

Table 33 - Third model results in days (D) and hours (H)

	Step									
	1		3		5		6		Total	
	D	H	D	H	D	H	D	H	D	H
ME	0.0	0.8	0.0	0.3	0.0	1.0	0.2	4.1	0.2	6.2
MAE	0.3	6.6	0.2	3.8	0.3	7.0	0.7	17.2	1.5	34.6
Mdelay	0.2	3.7	0.1	2.0	0.2	4.0	0.4	10.7	0.9	20.4
MSE	0.2	5.9	0.3	6.1	0.2	5.7	3.9	94.6	4.6	112.3
MPE	-19%		17%		47%		-13%			
MAPE	65%		76%		80%		33%			

4.5.4 *Forth Model*

Across all the previously introduced models, *Step 6* error measurements are significantly higher than the remaining. Although this is comprehensible due to the heterogeneous nature of both dependent and independent variables in this step, an alternative approach to *Route* was tested, by replacing *Route* by the corresponding distances. The main goal of this approach was to avoid the existence of very small groups that lead to non-significant mean values. This was implemented based on two approaches: distance as a categorical variable and distance as a continuous variable. As displayed in Table 34, both models deliver worse results than the model introduced in the previous section.

Table 34 - Forth model results (*Step 6*)

	Distance			
	Categorical		**Continuous**	
	Days	**Hours**	**Days**	**Hours**
ME	0.2	4.1	0.2	4.2
MAE	0.8	20.1	0.8	19.0
Mdelay	0.5	12.1	0.5	11.6
MSE	4.3	103.5	4.2	100.9
MPE	-21%		-18%	
MAPE	43%		39%	

4.5.5 *Fifth Model*

The last method explored in the scope of this project was a mix of the third and fourth models described beforehand. As such, *Step 6* timespans were estimated using factor *Route* for those groups whose size was larger than 50 and *Distance* for the remaining. This way, too small and non-significant groups were avoided, while important routes were expected to benefit from a better accuracy level.

Furthermore, two other enhancements were made to the model. Firstly, *Border Control* and *Border Trouble Index* were joined in a three level variable. Secondly, median group values were also tested to create group estimations. Overall results (Table 35) were slightly better than those delivered by the third model, especially those provided by model b). Model a), on the other hand, delivers better results for *MSE* and *Mdelay*, since mean group values are generally higher than median ones (outliers are mostly too high values).

Table 35 - Fifth model results (*Step 6*)

	a) Mean		**b) Median**	
	Days	**Hours**	**Days**	**Hours**
ME	-0.2	-4.4	-0.4	-8.5
MAE	0.7	17.6	0.7	16.9
Mdelay	0.5	11.0	0.5	12.7
MSE	4.0	96.4	4.1	99.3
MPE	13%		4%	
MAPE	34%		30%	

As results' improvements were not remarkable, error statistics were recalculated on a *Route* level, in order to understand which classes undermined the model performance. Annex O displays that information concerning routes that were used at least 1000 times. Results are highly heterogeneous, as can be observed in the first two lines. *Italy-Kazakhstan* is the *Route* for which the model delivered worse results, followed be *Italy-New Zealand.* The first can be explained by the fact that shipments to Kazakhstan require the participation of clients to be completed, which increases the variability of the process. The high number of courier partners

involved in long and transcontinental shipments can be an explanation for the second. Overall shipments to Brazil and China are also poorly predicted by the model. This highlights the necessity of subdividing these countries in smaller regions. On the best side of the spectrum, the majority of shipments to Germany, New York and Hong Kong exhibited the smallest error values.

4.5.6 *Best results*

In order to estimate the final delivery date, the best model was chosen for each step. Table 36 displays the final results and a total mean absolute error value of 1.5 days was obtained globally. *Step 6* particularly damages the performance of the model, which is expected to be overcome as explained in the previous section.

Table 36 - Best model results, in days (D) and hours (H)

	Step													
	1		**2**		**3**		**4**		**5**		**6**		**Total**	
	D	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**
Model	***3***		***1***		***3***		***1***		***3***		***5b)***			
ME	0.0	0.8	0.0	0.2	0.0	0.3	0.0	0.3	0.0	1.0	-0.4	-8.5	-0.4	-8.5
MAE	0.3	6.6	0.0	0.4	0.2	3.8	0.0	1.1	0.3	7.0	0.7	16.9	1.5	35.8
Mdelay	0.2	3.7	0.0	0.3	0.1	2.0	0.0	0.7	0.2	4.0	0.5	12.7	1	23.4
MSE	0.2	5.9	0.0	0.7	0.3	6.1	0.1	2.9	0.2	5.7	4.1	99.3	4.9	120.6
MPE	-19%		61%		17%		-21%		-47%		4%			
MAPE	65%		83%		76%		98%		80%		30%			

5 Conclusion and future work

One of the major challenges of the e-commerce sector is to conquer the trust of potential and existing customers. This can be achieved by providing trustworthy information concerning their shopping experience, including delivery. As such, the goal of the present project was to create a tool to estimate the delivery dates of the orders of a luxury fashion e-seller, Farfetch. The project was shaped by the fact that order process involved multiple agents and activities, increasing its complexity. Fortunately, due to the success of Farfetch, a high volume of data was available to create an estimating model. Hence, Data Mining was the adopted approach, performed according to its established major stages.

Business Understanding was conducted amongst several teams as to thoroughly acquire know-how about order processing. This was a major support for the following stage, *Data Understanding*, due to the nature of Farfetch database. The sophisticated structure of this database and its scarce documentation reflected the recent exponential growth of the company. Hence, understanding and interpreting its structure was not straightforward. Moreover, due to the high volume of data to be gathered, extracting information was very time-consuming. Several obstacles emerged, which were surpassed by restructuring queries towards an increasing efficiency level. This was an iterative and considerably enriching stage.

At the end of this phase, the majority of the variables were proven significant in the scope of this analysis and many assumptions were corroborated. This was an uplifting finding that supported the continuance of the project towards *Data Preparation* activities.

Once data was extracted and validated, it was necessary to reshape it to the desired form of model inputs. However, at this point, data was stored in several spreadsheets and *Excel* was not a suitable tool to conduct all the *Data Preparation* stage. Therefore, *RStudio* was used for this purpose. This phase embraced several activities, namely *Data Cleaning* and *Classification.* The first covered many tasks, from variable standardization to outlier detection. Identifying outliers was challenging, since a standard procedure could not be applied to all datasets, due to their asymmetrical nature. Tailored criteria had to be created for each step, according to the specific features of its data. Classifying variables in clusters was also demanding, since concerning data distributions were highly heterogeneous. Moreover, the factors to use in order to create these clusters were not a straightforward choice. As such, several possibilities were explored.

At last, *Modelling* stage began. The first model was created following the decisions and assumptions taken beforehand. Once it was completed, interpreting results was found to be a rather controversial activity. Firstly, model quality was heterogeneous across all steps. Secondly, for the steps where results were not as promising as expected, identifying the causes of underperformance was rather puzzling. By critically evaluating all the phases of this project, this could be due to a misguided selection of factors, inadequate decisions upon classification and outlier removal or even an unfitting model choice. However, this could also be due to the high variable and therefore unpredictable nature of data. As no absolute conclusions could be taken, other classification and modelling approaches were explored. This was conducted iteratively based on the results of each model. As displayed beforehand (Table 36), global mean absolute error is 1.5 days, mainly due to the last step. This step was given more prominence

since, as expected, it displayed the highest variability. By further analyzing it, it was clear that some countries need to be divided geographically in order to obtain more accurate results.

Despite being concluded from the curricular point of view, this project should not be considered finished as many paths can still be explored. As stated before, further exploring geographical dimensions may be critical to improving the quality of *Step 6's* model. Also, other clustering techniques could be developed, as this activity majorly impacts on modeling through its factors. Furthermore, ideally, more recent records should have a bigger weight, although the existence of an annual fashion cycle should not be disregarded. Moreover, outlier detection could be performed in the light of the new classification criteria and some atypical patterns (and flaws) should be studied in order to prevent inaccurate data from undermining the model. On top of this, the resolution of not removing outliers in testing samples should be reconsidered, since extreme values are undermining the performance of the model. This could lead to smaller mean error measures, impacting most significantly on mean squared errors.

In order to implement the models created in the scope of this project, these have to be updated. To do so, data has to be collected and treated for the months that were not covered by this project (since March up to the moment). Given this input, factors that were proven significant may lose relevance in the upcoming models, while new factors may be found. Hence the model should be re-evaluated and controlled periodically. This analysis should be conducted taking into consideration the business understanding shared by Farfetch teams.

In addition to the model, conclusions regarding factors and their influence on timespans are extremely valuable in the business context. Understanding (and quantifying) the standard performance of Farfetch partners constitutes an objective tool to evaluate and compare them. This information empowers Farfetch to negotiate in order to enhance delivery service. Furthermore, this is an important contribution to the company since these patterns have not been thoroughly analyzed beforehand and hence may result in valuable insights to the organization.

References

Akter, Shahriar, and Samuel Fosso Wamba. 2016. "Big Data Analytics in E-Commerce: A Systematic Review and Agenda for Future Research." *Electronic Markets* 26 (2): 173–94. doi:10.1007/s12525-016-0219-0.

Alden, Dana L., Jan-Benedict E.M. Steenkamp, and Rajeev Batra. 2006. "Consumer Attitudes toward Marketplace Globalization: Structure, Antecedents and Consequences." *International Journal of Research in Marketing* 23 (3): 227–39. doi:10.1016/j.ijresmar.2006.01.010.

Al-maghrabi, T, and C Dennis. 2009. "Driving Online Shopping: Spending and Behavioral Differences among Women in Saudi Arabia." *International Journal of Business Science and Applied Managment* 44 (0): 1–46. http://bura.brunel.ac.uk/handle/2438/3825.

Azar, Sana, Shamila Nabi Khan, Teaching Fellow, and Junaid Shavaid. 2015. "Logistic Support and Familiarity with Online Retailing." *The Journal of Developing Areas*, 428–37.

Basu, Susanto, Gary Becker, Kathy Bradbury, Kerwin Charles, Raj Chetty, Steve Davis, Jordi Galí, et al. 2006. "Measuring Trends in Leisure : The." *National Bureau of Economic Research*.

Bhaskar, Phani, and Prasanna Kumar. 2015. "E-Loyalty and E-Satisfaction of E- Commerce." *International Journal in Management and Social Science* 03 (11): 489–96.

Bulut, Zeki Atıl. 2015. "Determinants of Repurchase Intention in Online Shopping : A Turkish Consumer ' S Perspective." *International Journal of Business and Social Science* 6 (10): 55–63.

Cao, X., and P. L. Mokhtarian. 2009. "The Intended and Actual Adoption of Online Purchasing: A Brief Review of Recent Literature." *Institute of Transportation Studies* Issues in (530): 57–66. doi:10.1007/s11116-007-9132-x.

Chen, Fen, and Yan Zhang. 2011. "Online Marketing of Luxury Goods - Take Chinese Market as Example." In *2011 International Conference on Management and Service Science*, 1–4. IEEE. doi:10.1109/ICMSS.2011.5998973.

Davenport, T.H., 2012. 2012. "The Human Side of Big Data and High-Performance Analytics." *International Institute for Analytics*, 1–13.

Davenport, Thomas H, Jeanne Harris, and Jeremy Shapiro. 2010. "Competing on Talent Analytics."

Emerson, Carol J., and Curtis M. Grimm. 1996. "Logistics and Marketing Components of Customer Service: An Empirical Test of the Mentzer, Gomes and Krapfel Model." *International Journal of Physical Distribution & Logistics Management* 26 (8): 29–42. doi:10.1108/09600039610128258.

Enderlein, G. 1987. "Hawkins, D. M.: Identification of Outliers. Chapman and Hall, London – New York 1980, 188 S., £ 14, 50." *Biometrical Journal* 29 (2). WILEY- VCH Verlag: 198–198. doi:10.1002/bimj.4710290215.

Eurostat. 2015. "E-Commerce Statistics for Individuals." http://ec.europa.eu/eurostat/statistics-explained/index.php/E-commerce_statistics_for_individuals.

Falk, Martin, and Eva Hagsten. 2015. "E-Commerce Trends and Impacts across Europe." *International Journal of Production Economics* 170 (220): 357–69. doi:10.1016/j.ijpe.2015.10.003.

Gefen, David. 2002. "Customer Loyalty in E-Commerce." *Journal of the Association for Information Systems* 3: 27–51.

Globerson, Shlomo, and Ofer Zwikael. 2002. “The Impact of the Project Manager on Project Management Planning Processes.” *Project Management Journal* 33 (3): 58–64.

Guercini, S, and A Runfola. 2015. “Internationalization through E-Commerce. the Case of Multibrand Luxury Retailers in the Fashion Industry.” *Advances in International Marketing* 26: 15–31. doi:10.1108/S1474-797920150000026002.

Ha, Hong- Youl, Swinder Janda, and Siva K. Muthaly. 2010. “A New Understanding of Satisfaction Model in E- re- purchase Situation.” *European Journal of Marketing* 44 (7/8): 997–1016. doi:10.1108/03090561011047490.

Han, Jiawei, and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques*. Vol. 54. doi:10.5860/CHOICE.49-3305.

Hawkins, Simon, Hongxing He, Graham Williams, and Rohan Baxter. 2002. “Outlier Detection Using Replicator Neural Networks.” In , 170–80. Springer Berlin Heidelberg. doi:10.1007/3-540-46145-0_17.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. “Unbiased Recursive Partitioning: A Conditional Inference Framework.” *Journal of Computational and Graphical Statistics* 15 (July): 651–74. doi:10.1198/106186006X133933.

Hothorn, Torsten, and Achim Zeileis. 2015. “Partykit: A Modular Toolkit for Recursive Partytioning in R.” *Journal of Machine Learning Research* 16: 3905–9. http://jmlr.org/papers/v16/hothorn15a.html.

Hull, Gordon. 2015. “Successful Failure: What Foucault Can Teach Us about Privacy Self-Management in a World of Facebook and Big Data.” *Ethics and Information Technology* 17 (2). Springer Netherlands: 89–101. doi:10.1007/s10676-015-9363-z.

Khalifa, Mohamed, and Vanessa Liu. 2007. “Online Consumer Retention: Contingent Effects of Online Shopping Habit and Online Shopping Experience.” *European Journal of Information Systems* 16: 780–92. doi:10.1057/palgrave.ejis.3000711.

Khan, Shahzad Ahmad, Yan Liang, and Sumaira Shahzad. 2015. “An Empirical Study of Perceived Factors Affecting Customer Satisfaction to Re-Purchase Intention in Online Stores in China.” *Journal of Service Science and Management* 8 (June): 291–305. doi:10.4236/jssm.2015.83032.

Kong, Hong, Kwok Kee, and Northern Ireland. 2003. “Repurchase Intention in B2C E-Commerce-A Relationship Quality Perspective.” *Information & Management* 48 (6): 1–12.

Kumawat, Alka, and J. K. Tandon. 2014. “Factors Influencing Customer’s Satisfaction Level Towards Online Shopping in Jaipur and Gurgaon.” *International Journal of Innovative Research and Development || ISSN 2278 – 0211* 0 (0).

Labajos, Neus Soler, and Ana Jimenez-Zarco. 2016. “E-Commerce: The Effect of the Internet and Marketing Evolution.”

Li, Dahui, Glenn Browne, and James Wetherbe. 2006. “Why Do Internet Users Stick with a Specific Web Site? A Relationship Perspective.” *International Journal of Electronic Commerce* 10 (4): 105–41. doi:10.2753/JEC1086-4415100404.

Lin, Chun-Chun, Hsueh-Ying Wu, and Yong-Fu Chang. 2011. “The Critical Factors Impact on Online Customer Satisfaction.” *Procedia Computer Science* 3. Elsevier: 276–81. doi:10.1016/j.procs.2010.12.047.

Lin, Yong, Jing Luo, Shuqin Cai, Shihua Ma, and Ke Rong. 2016. “Exploring the Service Quality in the E-Commerce Context: A Triadic View.” *Industrial Management & Data Systems* 116 (3): 388–415. doi:10.1108/IMDS-04-2015-0116.

Machado, a. 2005. “Drivers of Shopping Online: A Literature Review.” *Proceedings of IADIS International Conference E Commerce*, 236–42.

Maimon, O., and L. Rokach. 2010. *Data Mining and Knowledge Discovery Handbook*. Edited by Oded Maimon and Lior Rokach. Boston, MA: Springer US. doi:10.1007/978-0-387-09823-4.

Martens, Bertin. 2013. “What Does Economic Research Tell Us About Cross-Border E-Commerce in the EU Digital Single Market?” *SSRN Electronic Journal*. doi:10.2139/ssrn.2265305.

Mayor, Eric. 2015. *Learning Predictive Analytics with R*.

McAfee, A., and E. Brynjolfsson. 2012. “Big Data : The Management Revolution.” *Harvard Business Review* 90 (10): 60–66.

Mentzer, John T., Roger Gomes, and Robert E. Krapfel. 1989. “Physical Distribution Service: A Fundamental Marketing Concept?” *Journal of the Academy of Marketing Science* 17 (1). Springer-Verlag: 53–62. doi:10.1007/BF02726354.

Ng, Ee Hong, Gupta Sumeet, and Hee-woong Kim. 2007. “Online Customer Retention : The Resistance to Change Perspective.” *Icis*, 1–19.

Nuseir, Mohammed T. 2016. “Exploring the Use of Online Marketing Strategies and Digital Media to Improve the Brand Loyalty and Customer Retention.” *International Journal of Business and Management* 11 (4): 228. doi:10.5539/ijbm.v11n4p228.

Papadimitriou, Spiros, Hiroyuki Kitagawa, Phillip B Gibbons, and Christos Faloutsos. 2003. “LOCI: Fast Outlier Detection Using the Local Correlation Integral.” *Proceedings - International Conference on Data Engineering*.

Polites, Greta L., Clay K. Williams, Elena Karahanna, and Larry Seligman. 2012. “A Theoretical Framework for Consumer E-Satisfaction and Site Stickiness: An Evaluation in the Context of Online Hotel Reservations.” *Journal of Organizational Computing and Electronic Commerce* 22 (1): 1–37. doi:10.1080/10919392.2012.642242.

Reichheld, Frederick F, and Phil Schefter. 2000. “E-Loyalty: Your Secret on the Web.” *Harvard Business Review* 78 (4): 105–13. doi:10.1007/PL00012187.

Rose, Susan, Moira Clark, Phillip Samouel, and Neil Hair. 2012. “Online Customer Experience in E-Retailing: An Empirical Model of Antecedents and Outcomes.” *Journal of Retailing*. Vol. 88. doi:10.1016/j.jretai.2012.03.001.

Santos, Renata Carneiro, Kavita Miadaira Hamza, and Vitor Koki da Costa Nogami. 2016. “E-Commerce de Artigos de Moda: Análise Da Influência Dos Atributos Da Compra Online.” *Revista Interdisciplinar de Marketing* 5 (1): 64–80.

Schroeck, M., R. Shockley, J. Smart, D. Romero-Morales, and P.P Tufano. 2012. “Analytics: The Real-World Use of Big Data. IBM Institute for Business Value.”

Seo, Songwon. 2006. “A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets.”

Sousa, Fernanda, and Fernando Nicolau. 2001. “Validação Em Classificação Hierárquica Ascendente – Alguns Resultados.” In *Novos Rumos Em Estatística*, 403–13.

Statista. 2016. “Global Retail E-Commerce Sales Volume from 2013 to 2018 (in Billion U.S. Dollars).” http://www.statista.com/statistics/222128/global-e-commerce-sales-volume-forecast/.

Terzi, Nuray. 2011. “The Impact of E-Commerce on International Trade and Employment.” *Procedia - Social and Behavioral Sciences* 24. Elsevier: 745–53.

doi:10.1016/j.sbspro.2011.09.010.

Valvi, Aikaterini C, and Douglas C West. 2013. “E-Loyalty Is Not All about Trust, Price Also Matters: Extending Expectation-Confirmation Theory in Bookselling Websites.” *Journal of Electronic Commerce Research* 14 (1): 99–123.

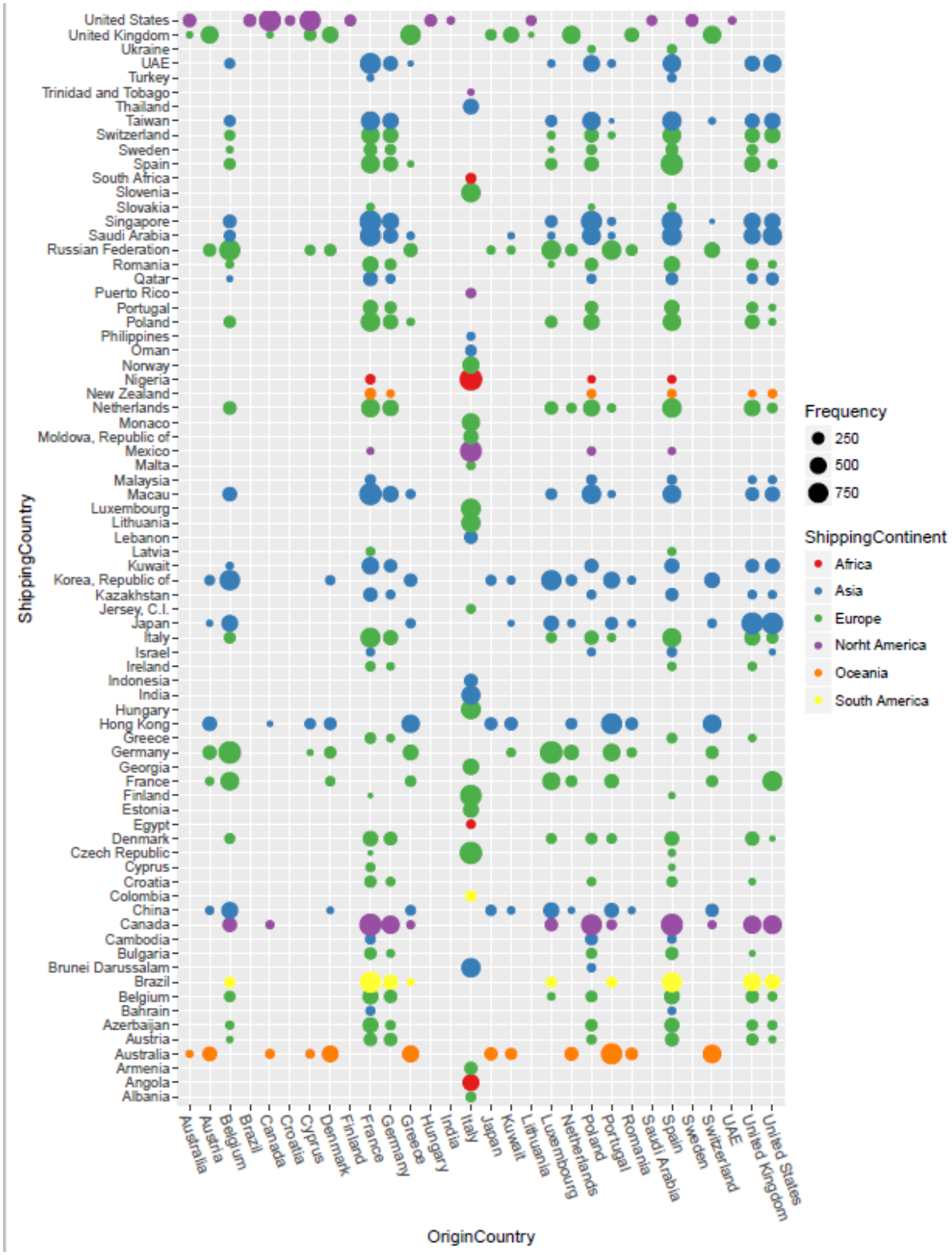
Wang, John. 2014. *Encyclopedia of Business Analytics and Optimization*. IGI Global.

White, M. 2012. “Digital Workplaces: Vision and Reality.” *Business Information Review* 29 (4): 205–14. doi:10.1177/0266382112470412.

Wulf, Kristof De, Gaby Odekerken-Schröder, and Dawn Iacobucci. 2001. “Investments in Consumer Relationships: A Cross-Country and Cross-Industry Exploration.” *Journal of Marketing* 65 (4): 33–50. doi:10.1509/jmkg.65.4.33.18386.

Zhou, Lina, Liwei Dai, and Dongsong Zhang. 2007. “Online Shopping Acceptance Model: Critical Survey of Consumer Factors in Online Shopping.” *Journal of Electronic Commerce Research* 8 (1): 41–62. doi:10.1086/209376,.

ANNEX A: Main country to country routes (frequency<1000)



ANNEX B: *Backlog* Queries Extracts

B1. Monthly estimation

```
SET TRANSACTION isolation level READ uncommitted

DECLARE @1fEV DATE= '20160201 00:06:00:000'
DECLARE @mes INT= 30
DECLARE @zero INT= 0
DECLARE @85 INT= 85
DECLARE @79 INT= 79
DECLARE @97 INT= 97

SELECT G1.siteid,
       @1fev,
       Count(*) AS backlog
FROM   glborders g1
       LEFT JOIN [BI_SYNC].[dbo].farorderlog ff6 (nolock)
              ON g1.siteid = ff6.siteid
                 AND g1.orderid = ff6.orderid
                 AND ff6.logtype = @85 -- Status Change: Sent
       LEFT JOIN [BI_SYNC].[dbo].farorderlog ff7 (nolock)
              ON g1.siteid = ff7.siteid
                 AND g1.orderid = ff7.orderid
                 AND ff7.logtype = @79 -- Status Change: Received
       LEFT JOIN [BI_SYNC].[dbo].[farrefordersa] ref
              ON ref.orderid = g1.orderid
                 AND ref.siteid = g1.siteid
       INNER JOIN farsitesinfo sinfo
               ON sinfo.localid = g1.siteid
       LEFT JOIN [BI_ETL].[dbo].[bi_dimgeographyglobaltzoffset] timezone
              ON timezone.timezoneid = sinfo.timezone
WHERE  Datediff(dd, g1.datacriado, @1fEV) <= @mes
       AND Dateadd(hh, timezone.countrytzoffset, g1.datacriado) < @1fEV
       AND ( Datediff(hh, @1fEV, ff6.date) > @zero
              OR ( ( ff6.date IS NULL )
                   AND ( Ff7.date IS NULL ) ) )
       AND ( ref.reallevelid IS NULL
              OR ref.reallevelid <> @97 )
GROUP  BY G1.siteid
ORDER  BY g1.siteid
```

B2. In flow

```
SET TRANSACTION isolation level READ uncommitted

DECLARE @inicio DATETIME= '20150222'
DECLARE @fim DATETIME='20150301'
DECLARE @seven INT=7
DECLARE @85 INT= 85
DECLARE @96 INT= 96
DECLARE @97 INT= 97

SELECT g.siteid,
       Datepart(d, g.datacriado),
       Datepart(m, g.datacriado),
       Datepart(year, g.datacriado),
       timezone.countrytzoffset,
       Count(*)
FROM   glborders g
       INNER JOIN [BI_SYNC].[dbo].farorderlog f6 (nolock)
               ON g.siteid = f6.siteid
                  AND g.orderid = f6.orderid
                  AND f6.logtype = @85 -- Status Change: Sent
       LEFT JOIN farrefordersa ref
              ON ref.siteid = g.siteid
                 AND g.orderid = ref.orderid
                 AND ref.reallevelid = @97
       LEFT JOIN farrefordersa ref1
              ON ref1.siteid = g.siteid
                 AND g.orderid = ref1.orderid
                 AND ref1.reallevelid = @96
       INNER JOIN farsitesinfo sinfo
               ON sinfo.localid = g.siteid
       LEFT JOIN [BI_ETL].[dbo].[bi_dimgeographyglobaltzoffset] timezone
              ON timezone.timezoneid = sinfo.timezone
```

```
WHERE  ( g.datacriado >= @inicio
       AND g.datacriado <= @fim )
       AND ref.who IS NULL
       AND ref1.who IS NULL --nao foi cancelada
       AND Datepart(hh, g.datacriado) >= 24 +
           ( @seven - timezone.countrytzoffset ) --CountryOffSet<=7
--and datepart(hh,g.DataCriado)>=@seven-timezone.CountryTZOffSet --CountryOffSet>7
GROUP  BY g.siteid,
          Datepart(d, g.datacriado),
          Datepart(m, g.datacriado),
          Datepart(year, g.datacriado),
          timezone.countrytzoffset
ORDER  BY g.siteid,
          Datepart(year, g.datacriado),
          Datepart(m, g.datacriado),
          Datepart(d, g.datacriado),
          timezone.countrytzoffset
```

B3. Out Flow

```
SET TRANSACTION isolation level READ uncommitted

--declare @inicio datetime= '20150301'
--declare @fim datetime='20160229'
SELECT g.siteid,
       Datepart(d, f6.date),
       Datepart(m, f6.date),
       Datepart(year, f6.date),
       Count(*)
FROM   glborders g
       INNER JOIN [BI_SYNC].[dbo].farorderlog f6 (nolock)
               ON g.siteid = f6.siteid
                  AND g.orderid = f6.orderid
                  AND f6.logtype = 85 -- Status Change: Sent
       LEFT JOIN farrefordersa ref
              ON ref.siteid = g.siteid
                 AND g.orderid = ref.orderid
                 AND ref.reallevelid = 97
       LEFT JOIN farrefordersa ref1
              ON ref1.siteid = g.siteid
                 AND g.orderid = ref1.orderid
                 AND ref1.reallevelid = 96
       INNER JOIN farsitesinfo sinfo
               ON sinfo.localid = g.siteid
WHERE  --(f6.date>=@inicio and f6.date<=@fim)
  Datepart(month, F6.date) >= 2
  AND Datepart(year, F6.date) = 2015

  AND ref.who IS NULL
  AND ref1.who IS NULL
--and datepart(hh,f6.date)< 7
GROUP  BY g.siteid,
          Datepart(dd, f6.date),
          Datepart(m, f6.date),
          Datepart(year, f6.date)
ORDER  BY g.siteid,
          Datepart(year, f6.date),
          Datepart(m, f6.date),
          Datepart(dd, f6.date)
```

ANNEX C: *Backlog* final calculation

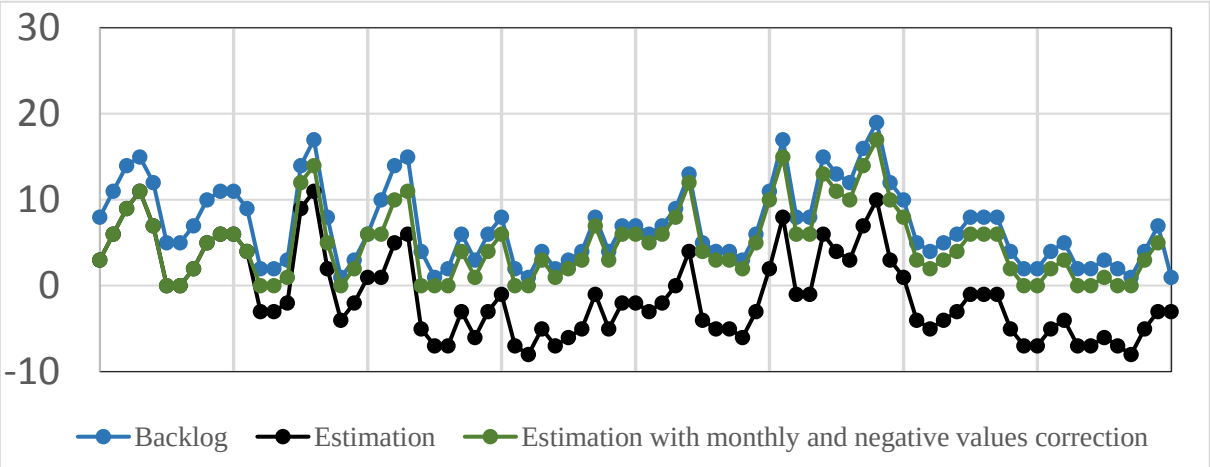
Given input and output flows described in chapter 4, *Backlog* calculation for every store and day is conducted as follows:

Net Flow Calculation

			In Flow			**Out Flow**				**Net Flow**
Store	Day	D	D (h<7am)	D+1 (h<7am)	d	D	D (h<7am)	D+1 (h<7am)	d	d
x	n	a	b	c	**i=a-b+c**	e	f	g	**h=e-f+g**	**i-h**

ANNEX D: *Backlog* accuracy

Comparison between estimated and real *Backlog* for the period in which the disparity was higher is displayed bellow. In conclusion, corrected *backlog* estimation accurately represents *Backlog* level.



Backlog accuracy test

ANNEX E: Weekend control

Weekday distribution delivery (before corrections)

			Delivered						
n	**n (%)**	**Weekends**	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
1092457	100,00%		23,49%	13,87%	19,26%	20,04%	17,93%	3,35%	2,07%
1060778	97,10%	Sat +Sun	24%	14%	19%	20%	18%	3%	1%
30856	2,82%	Fri + Sat	10%	5%	21%	20%	1%	18%	24%
524	0,05%	Sun	20%	6%	20%	21%	19%	14%	0%
4	0,00%	Sat	50%	0%	25%	0%	0%	0%	25%
295	0,03%	Sun *	27%	16%	19%	16%	18%	5%	0%

Weekday control (extract)

		Weekend Type	Delivered									Weekend Type
Country	**N**	***Before adjustments***	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**	*Suspicious?*	*Significant n?*	***After adjustments***
Afghanistan	**5**	2	0%	0%	20%	40%	20%	0%	20%	*Yes*	*No*	*2*
Albania	**293**	1	13%	16%	24%	24%	21%	2%	0%	*No*	*Yes*	*1*
Algeria	**5**	2	0%	40%	0%	0%	20%	20%	20%	*Yes*	*No*	*2*
Andorra	**8**	1	13%	0%	25%	38%	25%	0%	0%	*No*	*No*	*1*
Angola	**847**	1	27%	18%	18%	16%	15%	5%	0%	*No*	*Yes*	*1*
Argentina	**9**	1	33%	0%	22%	22%	22%	0%	0%	*No*	*No*	*1*
Armenia	**460**	1	20%	30%	12%	21%	17%	1%	1%	*No*	*Yes*	*1*
Aruba	**2**	1	0%	0%	50%	50%	0%	0%	0%	*No*	*No*	*1*
Australia	**91287**	1	36%	13%	5%	21%	19%	2%	3%	*No*	*Yes*	*1*
Austria	**5447**	1	20%	21%	22%	20%	17%	0%	0%	*No*	*Yes*	*1*
Azerbaijan	**5481**	1	21%	22%	16%	17%	14%	10%	0%	*Yes*	*Yes*	*3*
Bahamas	**11**	1	9%	45%	18%	9%	18%	0%	0%	*No*	*No*	*1*
Bahrain	**2357**	2	3%	2%	27%	22%	0%	26%	20%	*Yes*	*Yes*	*6*
Bangladesh	**1**	2	0%	100%	0%	0%	0%	0%	0%	*No*	*No*	*2*
Barbados	**3**	1	0%	0%	0%	33%	67%	0%	0%	*No*	*No*	*1*
Belarus	**125**	1	5%	39%	15%	19%	17%	3%	2%	*No*	*Yes*	*1*
Belgium	**6415**	1	17%	21%	22%	21%	18%	0%	0%	*No*	*Yes*	*1*
Benin	**23**	1	48%	4%	9%	30%	0%	9%	0%	*No*	*No*	*1*
Bermuda	**38**	1	11%	18%	29%	8%	32%	3%	0%	*No*	*No*	*1*

ANNEX F: SQL for timespan values collection

```
SET TRANSACTION isolation level READ uncommitted

DECLARE @USid INT= 216
DECLARE @ITid INT= 101
DECLARE @inicio DATE= '20150301'
DECLARE @fim DATE='20160229'
DECLARE @SITE1 INT = 9295
DECLARE @SITE2 INT = 9177
DECLARE @SITE3 INT =9588
DECLARE @SITE4 INT =9214
DECLARE @SITE5 INT =9022
DECLARE @SITE6 INT =9030
DECLARE @SITE7 INT =9140
DECLARE @SITE8 INT =9164
DECLARE @SITE9 INT =9859
DECLARE @SITE10 INT =9943
DECLARE @SITE11 INT =9003
DECLARE @SITE12 INT =9939
DECLARE @Order INT =2709669

SELECT g.orderid                                                    AS OrderID,
       g.siteid  AS SiteiD,
       ( lo.sigla + Cast(g.orderid AS VARCHAR(10)) ) AS             OrderNumber,
       g.datacriado                                                 AS I_Step1,
       f2.date                                  AS                  I_Step2,
       f3.date AS I_Step3,
       f4.date AS I_Step4,
       f5.date AS I_Step5,
       f6.date AS                                                   I_Step6,
       f7.date AS                                                   ReceivedDate
-----Step 1-------------------------------------------------------------------------
       ,
       Datediff(dd, g.datacriado, f2.date) AS Step1,
       ( CASE
           WHEN ( CONVERT(VARCHAR(10), f2.date, 120) ) =
                CONVERT(VARCHAR(10), g.datacriado, 120)
                --same day
                AND ( (SELECT Count(*)
                       FROM   analysts.dbo.[auxsosexceptiondates] ed
                       WHERE  ed.paisid = bop2.paisid
                              AND CONVERT(VARCHAR(10), ed.exceptionday, 120) =
                                  CONVERT(VARCHAR(10), g.datacriado, 120)) = 0 )
                ---not                          a                       holiday
                AND NOT ( ( wd.weekendid = 1
                            AND (( Datepart(dw, g.datacriado) = 1
                                    OR Datepart(dw, g.datacriado) = 7 )) )
                          --not                                          a weekend
                          OR ( wd.weekendid = 2
                               AND (( Datepart(dw, g.datacriado) = 6
                                       OR Datepart(dw, g.datacriado) = 7 )) )
                          OR ( wd.weekendid = 3
                               AND Datepart(dw, g.datacriado) = 6 )
                          OR ( wd.weekendid = 4
                               AND Datepart(dw, g.datacriado) = 7 )
                          OR ( wd.weekendid = 5
                               AND Datepart(dw, g.datacriado) = 1 ) ) THEN
           Round(Cast((SELECT Datediff(hh, g.datacriado, f2.date)) / 24.00 AS
                      FLOAT), 1
           )
           ELSE ( Round(Cast((SELECT ( CASE
                                         WHEN (SELECT Count(*)
                                               FROM
                                                analysts.dbo.[auxsosexceptiondates] ed
                                               WHERE  ed.paisid = bop2.paisid
                                                 AND CONVERT(VARCHAR(10),ed.exceptionday,120
) =
                                                     CONVERT(VARCHAR(10),g.datacriado, 120 ))

= 0
                                               AND NOT
                                             ( ( wd.weekendid = 1
                                                 AND ((Datepart(dw, g.datacriado) = 1
                                                   OR  Datepart(dw, g.datacriado) = 7 )) )
                                                OR ( wd.weekendid = 2
                                                                                AND (
                                    ( Datepart(dw, g.datacriado) = 6
                                      OR Datepart(dw, g.datacriado) = 7 )) )
                                     OR ( wd.weekendid = 3
AND Datepart(dw, g.datacriado) = 6 )
```

```
                                    OR ( wd.weekendid = 4 AND Datepart(dw, g.datacriado) = 7
 )
                                    OR ( wd.weekendid = 5 AND Datepart(dw, g.datacriado) = 1
 ))
                                                THEN
                                                (SELECT 24 - (
                                                Datepart(hh, g.datacriado) ))
                                                ELSE 0
                                            END ) + ( CASE
                                                        WHEN (SELECT Count(*)
                                                              FROM
          analysts.dbo.[auxsosexceptiondates] ed
                WHERE  ed.paisid = bop2.paisid
                       AND CONVERT(VARCHAR(10) ,ed.exceptionday, 120) =
                           CONVERT(VARCHAR(10),f2.date,120))
                  = 0
               AND NOT
          ( ( wd.weekendid = 1
              AND (
                   ( Datepart(dw, f2.date) = 1
                      OR
              Datepart(dw, f2.date) = 7 )) )
               OR
          ( wd.weekendid = 2
          AND (
          ( Datepart(dw, f2.date) = 6 OR Datepart(dw, f2.date) = 7 )) )
          OR
          ( wd.weekendid = 3 AND Datepart(dw, f2.date) = 6 )
             OR
          ( wd.weekendid = 4 AND Datepart(dw, f2.date) = 7 )
             OR ( wd.weekendid = 5 AND Datepart(dw, f2.date) = 1 ) )
          THEN (
          SELECT
          ( Datepart(hh, f2.date) ))
          ELSE 0
          END ) + 24 * Count(*)
          FROM   [BI_SYNC].[dbo].[z_bi_dimdate] dd
          WHERE  [fulldate] > CONVERT(VARCHAR(10), g.datacriado, 120)
          AND [fulldate] < CONVERT(VARCHAR(10), f2.date, 120  )
          AND [fulldate] NOT IN (SELECT
          CONVERT(VARCHAR(10), ed.exceptionday,120)
                   FROM analysts.dbo.[auxsosexceptiondates] ed
                   WHERE ed.paisid = bop2.paisid
          AND CONVERT(VARCHAR(10), ed.exceptionday, 120) >
          CONVERT(VARCHAR(10), g.datacriado, 120 )
          AND CONVERT(VARCHAR(10), ed.exceptionday, 120) <
          CONVERT(VARCHAR(10), f2.date, 120 ))
          AND NOT ( ( wd.weekendid = 1
          AND (( Datepart(dw, [fulldate]) = 1
                OR Datepart(dw, [fulldate]) = 7
             )) )
          OR ( wd.weekendid = 2
            AND (( Datepart(dw, [fulldate]) = 6 OR Datepart(dw, [fulldate] )
              = 7 ))
          )
          OR ( wd.weekendid = 3 AND Datepart(dw, [fulldate]) = 6 )
          OR ( wd.weekendid = 4 AND Datepart(dw, [fulldate]) = 7 )
          OR ( wd.weekendid = 5 AND Datepart(dw, [fulldate]) = 1 )
          )
          ) / 24.00 AS FLOAT), 1) )
          END ) AS Step1Net,

               -----END Step 1-----------------------------------------------------------------
```

(…)(Similar code for *Steps 2* to *6*)

```
FROM    [BI_SYNC].[dbo].glborders g (nolock)
        INNER JOIN [BI_SYNC].[dbo].faroderlog f2 (nolock)
                ON g.siteid = f2.siteid
                   AND g.orderid = f2.orderid
                   AND f2.logtype = 87 -- Status change:Stock OK
        INNER JOIN [BI_SYNC].[dbo].faroderlog f3 (nolock)
                ON g.siteid = f3.siteid
                   AND g.orderid = f3.orderid
                   AND f3.logtype = 33 -- Status Change: Payment OK
        INNER JOIN [BI_SYNC].[dbo].faroderlog f4 (nolock)
                ON g.siteid = f4.siteid
```

```
                AND g.orderid = f4.orderid
                AND f4.logtype = 77 -- status Change: Package OK
        INNER JOIN [BI_SYNC].[dbo].farorderlog f5 (nolock)
                ON g.siteid = f5.siteid
                AND g.orderid = f5.orderid
                AND f5.logtype = 34 -- Status Change: Ready to Send
        INNER JOIN [BI_SYNC].[dbo].farorderlog f6 (nolock)
                ON g.siteid = f6.siteid
                AND g.orderid = f6.orderid
                AND f6.logtype = 85 -- Status Change: Sent
        INNER JOIN [BI_SYNC].[dbo].farorderlog f7 (nolock)
                ON g.siteid = f7.siteid
                AND g.orderid = f7.orderid
                AND f7.logtype = 79 -- Status Change: Received
        --Weekend type
        INNER JOIN [ANALYSTS].[FARFETCH\ines.carvalho].[weekenddelivery] wd
                ON wd.countryid = g.scountryid
        --Shipping country
        INNER JOIN [BI_SYNC].[dbo].bopaises bop1
                ON bop1.paisid = g.scountryid
        --Boutique country
        INNER JOIN [BI_SYNC].[dbo].bolocais lo
                ON lo.localid = g.siteid
        INNER JOIN [BI_SYNC].[dbo].bopaises bop2
                ON bop2.paisid = lo.paisid
left outer join [BI_SYNC].[dbo].[FarOrderStock] st (nolock) on st.OrderID = g.OrderID and st.S
iteID = g.SiteID
WHERE  f7.date IS NOT NULL --delivered
       AND ( f6.date >= @inicio
             AND f6.date <= @fim )
       AND f7.date >= f6.date
       --STORES with stock far
       AND bop1.paisid <> @USid --USA must be on state level
       AND g.siteid <> @SITE1
       AND g.siteid <> @SITE2
       AND g.siteid <> @SITE3
       AND g.siteid <> @SITE4
       AND g.siteid <> @SITE5
       AND g.siteid <> @SITE6
       AND g.siteid <> @SITE7
       AND g.siteid <> @SITE8
       AND g.siteid <> @SITE9
       AND g.siteid <> @SITE10
       AND g.siteid <> @SITE11
       AND g.siteid <> @SITE12
ORDER  BY g.orderid,
          g.siteid
```

ANNEX H: Boutique impact on *Steps 1* and *3*

Factor Boutique impact on *Steps 1* and *3* Net Timespan

α= 0,05			Hypothesis testing				Step 1 Net			Hypothesis testing				Step 3 Net		
		n	> than mean	p value	< than mean	p value	Median	Mean	sd	> than mean	p value	< than mean	p value	Median	Mean	sd
	ALL	1128555					0,30	0,43	0,57					0,00	0,14	0,55
by boutique	9600	3746	N	0,00	Y	1,00	0,10	0,19	0,25	N	0,00	Y	1,00	0,00	0,01	0,07
	9710	3782	N	0,00	Y	1,00	0,10	0,21	0,26	N	0,00	Y	1,00	0,00	0,13	0,26
	9636	5364	N	0,00	Y	1,00	0,20	0,31	0,36	N	0,00	Y	1,00	0,00	0,04	0,14
	9429	5401	N	0,00	Y	1,00	0,20	0,27	0,30	N	0,00	Y	1,00	0,00	0,10	0,20
	9514	5547	N	0,00	Y	1,00	0,20	0,31	0,39	N	0,00	Y	1,00	0,00	0,09	0,37
	9439	5761	N	0,00	Y	1,00	0,20	0,27	0,38	Y	1,00	N	0,00	0,00	0,19	0,40
	9628	5911	Y	1,00	N	0,00	0,50	0,56	0,48	Y	1,00	N	0,00	0,00	0,34	0,57
	9672	5937	N	0,00	Y	1,00	0,00	0,13	0,23	N	0,00	Y	1,00	0,00	0,07	0,27
	9058	5963	Y	1,00	N	0,00	0,60	0,56	0,45	N	0,00	Y	1,00	0,00	0,07	0,21
	9796	6030	Y	1,00	N	0,00	0,40	0,54	0,51	N	0,00	Y	1,00	0,00	0,04	0,15
	9661	6208	Y	1,00	N	0,00	0,40	0,50	0,49	N	0,00	Y	1,00	0,00	0,08	0,19
	9684	6305	N	0,00	Y	1,00	0,10	0,25	0,37	N	0,00	Y	1,00	0,00	0,07	0,44
	9329	6317	N	0,00	Y	1,00	0,10	0,24	0,42	N	0,00	Y	1,00	0,00	0,02	0,11
	9026	6447	N	0,00	Y	1,00	0,20	0,38	0,55	N	0,00	Y	1,00	0,00	0,04	0,24
	9640	6545	N	0,00	Y	1,00	0,20	0,30	0,43	N	0,00	Y	1,00	0,00	0,02	0,11
	9359	6625	N	0,02	Y	0,98	0,30	0,41	0,59	Y	1,00	N	0,00	0,10	0,28	0,57
	9339	6669	N	0,00	Y	1,00	0,20	0,28	0,33	N	0,00	Y	1,00	0,00	0,07	0,21
	9016	6733	N	0,00	Y	1,00	0,00	0,16	0,26	N	0,00	Y	1,00	0,00	0,02	0,25
	9644	6807	N	0,00	Y	1,00	0,30	0,36	0,43	N	0,00	Y	1,00	0,00	0,13	0,28
	9309	6829	Y	0,99	N	0,01	0,30	0,44	0,53	N	0,00	Y	1,00	0,00	0,03	0,14
	9158	6955	N	0,00	Y	1,00	0,10	0,19	0,26	N	0,00	Y	1,00	0,00	0,03	0,13
	9111	6983	N	0,00	Y	1,00	0,10	0,21	0,26	Y	1,00	N	0,00	0,30	0,37	0,39
	9442	7045	N	0,00	Y	1,00	0,10	0,22	0,32	Y	1,00	N	0,00	0,10	0,30	0,41
	9530	7288	N	0,00	Y	1,00	0,20	0,37	0,48	N	0,00	Y	1,00	0,00	0,08	0,23
	9206	7324	Y	1,00	N	0,00	0,50	0,59	0,46	N	0,00	Y	1,00	0,00	0,04	0,15
	9727	7471	Y	1,00	N	0,00	0,60	0,89	0,82	N	0,00	Y	1,00	0,00	0,04	0,14
	9453	7519	N	0,00	Y	1,00	0,10	0,18	0,25	N	0,00	Y	1,00	0,00	0,07	0,19
	9300	7605	N	0,00	Y	1,00	0,10	0,24	0,34	N	0,00	Y	1,00	0,00	0,03	0,13
	9183	7684	N	0,00	Y	1,00	0,20	0,30	0,38	Y	1,00	N	0,00	0,10	0,16	0,29
	9334	8377	N	0,00	Y	1,00	0,10	0,17	0,22	Y	1,00	N	0,00	0,10	0,17	0,24
	9364	8519	N	0,00	Y	1,00	0,10	0,22	0,27	Y	1,00	N	0,00	0,30	0,34	0,39
	9541	8535	N	0,00	Y	1,00	0,20	0,26	0,29	N	0,00	Y	1,00	0,00	0,03	0,14
	9529	8579	N	0,00	Y	1,00	0,30	0,39	0,35	N	0,00	Y	1,00	0,00	0,07	0,23
	9148	8756	N	0,00	Y	1,00	0,00	0,15	0,21	Y	1,00	N	0,00	0,00	0,18	0,30
	9560	9114	N	0,00	Y	1,00	0,30	0,33	0,31	N	0,00	Y	1,00	0,00	0,07	0,20
	9474	9806	N	0,00	Y	1,00	0,10	0,23	0,28	Y	1,00	N	0,00	0,20	0,38	0,44
	9597	9833	N	0,00	Y	1,00	0,30	0,31	0,46	N	0,00	Y	1,00	0,00	0,02	0,12
	9298	10202	Y	1,00	N	0,00	0,80	0,93	0,75	N	0,00	Y	1,00	0,00	0,04	0,18
	9436	11013	N	0,00	Y	1,00	0,10	0,20	0,33	Y	1,00	N	0,00	0,40	0,36	0,41
	9124	11290	N	0,00	Y	1,00	0,10	0,22	0,30	N	0,00	Y	1,00	0,00	0,11	0,27
	9274	11766	N	0,00	Y	1,00	0,10	0,26	0,42	Y	1,00	N	0,00	0,00	0,17	0,29
	9178	11866	Y	1,00	N	0,00	0,60	0,69	0,58	N	0,00	Y	1,00	0,00	0,10	0,38
	9671	12209	N	0,00	Y	1,00	0,20	0,32	0,43	N	0,00	Y	1,00	0,00	0,04	0,17
	9089	12409	Y	1,00	N	0,00	0,50	0,57	0,38	N	0,00	Y	1,00	0,00	0,08	0,19
	9728	13392	N	0,00	Y	1,00	0,00	0,18	0,24	N	0,00	Y	1,00	0,00	0,02	0,10
	9258	13633	N	0,00	Y	1,00	0,20	0,27	0,46	N	0,00	Y	1,00	0,00	0,03	0,23
	9317	13823	Y	1,00	N	0,00	0,60	0,71	0,64	N	0,00	Y	1,00	0,00	0,12	0,26
	9579	14857	N	0,00	Y	1,00	0,30	0,36	0,38	N	0,00	Y	1,00	0,00	0,06	0,17
	9544	15029	Y	1,00	N	0,00	0,30	0,45	0,51	N	0,00	Y	1,00	0,00	0,09	0,21
	9681	15547	N	0,00	Y	1,00	0,10	0,21	0,27	Y	1,00	N	0,00	0,40	0,48	0,47
	9383	17330	Y	1,00	N	0,00	0,30	0,47	0,63	Y	1,00	N	0,00	0,00	0,28	0,47
	9053	21352	Y	1,00	N	0,00	0,50	0,60	0,55	N	0,00	Y	1,00	0,00	0,07	0,22
	9306	22038	N	0,00	Y	1,00	0,20	0,26	0,35	Y	1,00	N	0,00	0,00	0,24	0,36
	9214	26689	Y	1,00	N	0,00	0,50	0,58	0,51	Y	1,00	N	0,00	0,00	0,17	0,42
	9017	27091	Y	1,00	N	0,00	0,70	0,89	0,73	N	0,00	Y	1,00	0,00	0,08	0,26
	9475	30877	Y	1,00	N	0,00	0,40	0,46	0,53	N	0,00	Y	1,00	0,00	0,06	0,18
	9446	43084	Y	1,00	N	0,00	0,80	0,99	0,83	N	0,14	N	0,86	0,00	0,14	0,27
	9462	48731	Y	1,00	N	0,00	0,60	0,70	0,56	N	0,00	Y	1,00	0,00	0,09	0,25
	9336	54773	Y	1,00	N	0,00	0,50	0,61	0,64	N	0,00	Y	1,00	0,00	0,05	0,16

ANNEX I: Boutique Sales Volume and Country impact on *Steps 1* and *3*

Factor Boutique Sales Volume impact on *Steps 1* and *3* Net Timespan

α= 0			Hypothesis testing				Step 1 Net			Hypothesis testing				Step 3 Net		
		n	> than mean	p value	< than mean	p value	Median	Mean	sd	> than mean	p value	< than mean	p value	Median	Mean	sd
ALL		1E+06					0,30	0,43	0,57					0,00	0,14	0,55
by Sales volume																
A	>10000	5E+05	Y	1,00	N	0,00	0,40	0,56	0,62	N	0,00	Y	1,00	0,00	0,12	0,30
B	>7500	94327	N	0,00	Y	1,00	0,20	0,26	0,32	Y	1,00	N	0,00	0,00	0,14	0,29
C	>5000	2E+05	N	0,00	Y	1,00	0,30	0,38	0,49	N	0,00	Y	1,00	0,00	0,10	0,31
D	>2500	2E+05	N	0,00	Y	1,00	0,20	0,33	0,47	Y	1,00	N	0,00	0,00	0,14	0,43
E	>1000	1E+05	N	0,00	Y	1,00	0,20	0,33	0,43	N	0,00	Y	1,00	0,00	0,13	0,39
F	>750	44006	N	0,00	Y	1,00	0,20	0,29	0,40	N	0,00	Y	1,00	0,00	0,09	0,31
G	>500	27173	N	0,00	Y	1,00	0,30	0,39	0,70	Y	1,00	N	0,00	0,00	0,24	1,49
H	>250	26948	N	0,00	Y	1,00	0,30	0,37	0,57	Y	1,00	N	0,00	0,00	0,31	1,75
I	>100	15398	Y	1,00	N	0,00	0,30	0,55	1,34	Y	1,00	N	0,00	0,00	0,51	2,08
J	>50	4055	Y	1,00	N	0,00	0,30	0,47	0,75	Y	1,00	N	0,00	0,00	0,30	0,80
K	>25	1545	Y	1,00	N	0,00	0,40	0,52	0,92	Y	1,00	N	0,00	0,00	0,31	1,04
L	>=1	708	Y	1,00	N	0,00	0,30	0,51	0,72	Y	1,00	N	0,00	0,00	0,32	0,92

Factor Boutique Country impact on *Steps 1* and *3* Net Timespan

α= 0		Hypothesis testing				Step 1 Net			Hypothesis testing				Step 3 Net		
	n	> than mean	p value	< than mean	p value	Median	Mean	sd	> than mean	p value	< than mean	p value	Median	Mean	sd
ALL	1E+06					0,30	0,43	0,57					0,00	0,14	0,55
by country															
Italy	6E+05	N	0,00	Y	1,00	0,30	0,41	0,54	N	0,00	Y	1,00	0,00	0,11	0,28
France	74940	Y	1,00	N	0,00	0,50	0,56	0,60	N	0,00	Y	1,00	0,00	0,09	0,27
Spain	65542	N	0,00	Y	1,00	0,10	0,25	0,37	N	0,00	Y	1,00	0,00	0,12	0,29
United States	63499	N	0,00	Y	1,00	0,30	0,39	0,51	N	0,00	Y	1,00	0,00	0,09	0,28
Poland	54487	Y	1,00	N	0,00	0,50	0,67	0,58	N	0,00	Y	1,00	0,00	0,08	0,24
Germany	47973	Y	1,00	N	0,00	0,40	0,45	0,56	N	0,00	Y	1,00	0,00	0,11	0,36
United Kingdom	46800	N	0,00	Y	1,00	0,20	0,28	0,45	Y	1,00	N	0,00	0,00	0,14	0,36
Brazil	43770	Y	1,00	N	0,00	0,50	0,58	0,85	Y	1,00	N	0,00	0,30	0,82	1,96
Belgium	22985	N	0,01	Y	0,99	0,40	0,42	0,42	N	0,00	Y	1,00	0,00	0,07	0,23
Luxembourg	19511	Y	1,00	N	0,00	0,50	0,59	0,53	Y	1,00	N	0,00	0,00	0,21	0,45
Portugal	13742	Y	1,00	N	0,00	0,40	0,56	1,26	Y	1,00	N	0,00	0,00	0,29	2,01
Greece	9040	N	0,00	Y	1,00	0,30	0,28	0,33	N	0,00	Y	1,00	0,00	0,05	0,22
Switzerland	8449	N	0,44	N	0,56	0,40	0,43	0,53	N	0,00	Y	1,00	0,00	0,04	0,15
Austria	5987	Y	1,00	N	0,00	0,40	0,46	0,54	N	0,00	Y	1,00	0,00	0,08	0,28
Netherlands	5924	N	0,00	Y	1,00	0,20	0,29	0,37	N	0,00	Y	1,00	0,00	0,07	0,27
Denmark	5741	N	0,00	Y	1,00	0,40	0,40	0,40	N	0,00	Y	1,00	0,00	0,06	0,18
Kuwait	4511	Y	1,00	N	0,00	0,50	0,61	0,82	N	0,00	Y	1,00	0,00	0,06	0,25
Romania	4407	N	0,00	Y	1,00	0,20	0,25	0,34	N	0,00	Y	1,00	0,00	0,10	0,27
Japan	3410	Y	1,00	N	0,00	0,40	0,53	0,61	N	0,00	Y	1,00	0,00	0,06	0,20
Cyprus	3065	N	0,00	Y	1,00	0,20	0,25	0,34	N	0,83	N	0,17	0,00	0,14	0,31
Canada	2211	N	0,19	N	0,81	0,30	0,42	0,53	Y	1,00	N	0,00	0,20	0,38	0,50
Andorra	1639	N	0,00	Y	1,00	0,00	0,16	0,23	N	0,00	Y	1,00	0,00	0,09	0,32
Australia	1095	N	0,37	N	0,63	0,30	0,42	0,58	Y	0,99	N	0,01	0,00	0,16	0,36
Sweden	1030	Y	0,98	N	0,02	0,50	0,45	0,39	N	0,00	Y	1,00	0,00	0,06	0,26
Lithuania	979	N	0,00	Y	1,00	0,00	0,13	0,26	N	0,55	N	0,45	0,00	0,14	0,29
Finland	883	N	0,00	Y	1,00	0,10	0,23	0,26	N	0,00	Y	1,00	0,00	0,06	0,21
Hungary	780	N	0,00	Y	1,00	0,30	0,38	0,49	N	0,00	Y	1,00	0,00	0,10	0,22
Saudi Arabia	645	N	0,00	Y	1,00	0,00	0,21	0,35	N	0,00	Y	1,00	0,00	0,11	0,26
Croatia	643	N	0,00	Y	1,00	0,20	0,26	0,34	N	0,00	Y	1,00	0,00	0,04	0,14
India	573	N	0,72	N	0,28	0,30	0,44	0,45	N	0,00	Y	1,00	0,00	0,07	0,28
UAE	358	N	0,00	Y	1,00	0,20	0,32	0,42	N	0,00	Y	1,00	0,00	0,02	0,11
Bulgaria	194	N	0,00	Y	1,00	0,00	0,17	0,25	N	0,00	Y	1,00	0,00	0,02	0,11
Morocco	185	N	0,04	Y	0,96	0,30	0,37	0,46	Y	0,98	N	0,02	0,00	0,24	0,70
NULL	88	N	0,02	Y	0,98	0,10	0,34	0,40	Y	0,99	N	0,01	0,00	0,26	0,51
Singapore	64	N	0,00	Y	1,00	0,10	0,16	0,20	N	0,02	Y	0,98	0,00	0,08	0,21

ANNEX J: Classification of boutiques

Non Hierarchical Classification (NH-Means) of "Boutiques"

Initial Cluster Centers

	Cluster									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sales	54773	48731	43084	30877	27091	1	22038	13392	17330	15547
Step 1	0.610	0.702	0.991	0.459	0.893	2.000	0.260	0.183	0.473	0.210
Step 3	0.051	0.087	0.136	0.060	0.079	0.800	0.236	0.021	0.276	0.483
Step 5	0.675	0.176	0.400	0.425	0.524	0.000	0.338	0.469	0.441	0.497

Final Cluster Centers

	Cluster									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sales	54773	48731	43084	30877	26890	454	21695	3840	13806	8020
Step 1	0.610	0.702	0.991	0.459	0.734	0.445	0.430	0.333	0.408	0.331
Step 3	0.051	0.087	0.136	0.060	0.127	0.298	0.152	0.139	0.133	0.130
Step 5	0.675	0.176	0.400	0.425	0.420	0.395	0.602	0.460	0.584	0.557

Distances between Final Cluster Centers

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1		6042.00	11689.00	23896.00	27883.00	54318.928	33078.00	50932.53	40967.46	46752.52
2	6042.00		5647.00	17854.00	21841.00	48276.928	27036.00	44890.53	34925.46	40710.52
3	11689.00	5647.00		12207.00	16194.00	42629.928	21389.00	39243.53	29278.46	35063.52
4	23896.00	17854.00	12207.00		3987.00	30422.928	9182.00	27036.53	17071.46	22856.52
5	27883.00	21841.00	16194.00	3987.00		26435.928	5195.00	23049.53	13084.46	18869.52
6	54318.93	48276.93	42629.93	30422.93	26435.93		21240.93	3386.40	13351.47	7566.41
7	33078.00	27036.00	21389.00	9182.00	5195.00	21240.928		17854.53	7889.46	13674.52
8	50932.53	44890.53	39243.53	27036.53	23049.53	3386.397	17854.53		9965.08	4180.01
9	40967.46	34925.46	29278.46	17071.46	13084.46	13351.473	7889.46	9965.08		5785.06
10	46752.52	40710.52	35063.52	22856.52	18869.52	7566.409	13674.52	4180.01	5785.06	

Number of Cases in each Cluster

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Valid**	**Missing**
Cases	54773	48731	43084	30877	26890	454	21695	3840	13806	8020	623	0

ANOVA

Step		Df	Sum Sq	Mean Sq	F value	Pr(>F)
1	**Class**	10	237623	23762	81069	<2.20E-16
	Residuals	1E+06	330788	0.3		
3	**Class**	10	22845	2284.5	7719.3	<2.20E-16
	Residuals	112854	333987	0.3		
5	**Class**	10	281484	323525	97290	<2.20E-16
	Residuals	1E+06	333987	0.3		

Residuals

Step	Min	1st Q	Median	3rd Q	Max
1	-0,991	-0,331	-0,11	0,158	6,78E+01
3	-0,19	-0,142	-0,128	-0,051	6,08E+01
5	-0,675	-0,3255	-0,1689	0,2286	2,78E+01

ANNEX K: Best and worst Routes

20 Best and worst routes

			Hypothesis testing				Step 6Net		
		n	> than mean	p value	< than mean	p value	Median	Mean	sd
	ALL	1118219					1,900	2,329	2,168
20 best Routes (n>100)	New York-New Jersey	715	N	0,00	Y	1,00	0,80	0,81	0,56
	New York-Connecticut	283	N	0,00	Y	1,00	0,80	0,82	0,42
	New York-New York	3072	N	0,00	Y	1,00	0,80	0,86	0,63
	New York-Massachuset	669	N	0,00	Y	1,00	0,80	0,86	0,50
	Florida-Florida	182	N	0,00	Y	1,00	0,80	0,86	1,03
	Germany-Germany	3536	N	0,00	Y	1,00	0,80	0,86	0,73
	Romania-Germany	170	N	0,00	Y	1,00	0,80	0,87	0,21
	Spain-Slovakia	137	N	0,00	Y	1,00	0,80	0,88	0,37
	Pennsylvania-New York	120	N	0,00	Y	1,00	0,80	0,89	0,89
	Belgium-Belgium	204	N	0,00	Y	1,00	0,80	0,89	0,62
	Poland-Slovakia	108	N	0,00	Y	1,00	0,80	0,89	0,28
	France-Latvia	149	N	0,00	Y	1,00	0,80	0,90	0,32
	Germany-Belgium	283	N	0,00	Y	1,00	0,80	0,90	0,71
	Belgium-Poland	241	N	0,00	Y	1,00	0,80	0,91	0,32
	Germany-Poland	381	N	0,00	Y	1,00	0,80	0,91	0,39
	New York-Maryland	254	N	0,00	Y	1,00	0,80	0,91	0,65
	Luxembourg-Germany	984	N	0,00	Y	1,00	0,80	0,91	0,85
	Belgium-Austria	111	N	0,00	Y	1,00	0,80	0,91	0,45
	Spain-Latvia	137	N	0,00	Y	1,00	0,90	0,92	0,32
	Luxembourg-Poland	244	N	0,00	Y	1,00	0,80	0,93	0,51
20 worst Routes (n>100)	Canada-UK	114	Y	0,97	N	0,03	1,80	7,92	31,90
	Canada-Australia	157	Y	0,97	N	0,03	2,70	5,37	20,61
	Greece-Russia	321	Y	1,00	N	0,00	4,70	5,27	3,07
	Italy-Indonesia	307	Y	1,00	N	0,00	3,50	5,03	5,37
	Italy-Kazakhstan	2472	Y	1,00	N	0,00	4,00	4,92	3,32
	Canada-New York	177	N	0,91	N	0,09	0,90	4,88	25,57
	UK-Kazakhstan	149	Y	1,00	N	0,00	3,90	4,82	3,82
	Spain-Kazakhstan	276	Y	1,00	N	0,00	4,00	4,75	3,01
	Greece-China	180	Y	1,00	N	0,00	3,50	4,69	3,93
	France-Kazakhstan	335	Y	1,00	N	0,00	3,80	4,68	3,39
	Germany-Kazakhstan	155	Y	1,00	N	0,00	3,90	4,63	2,98
	Cyprus-Russia	187	Y	1,00	N	0,00	3,90	4,63	3,20
	Spain-Turkey	150	Y	1,00	N	0,00	3,00	4,61	5,31
	Poland-Kazakhstan	159	Y	1,00	N	0,00	4,00	4,55	2,32
	Florida-Russia	132	Y	1,00	N	0,00	3,75	4,52	3,26
	Italy-South Africa	187	Y	1,00	N	0,00	3,60	4,51	4,15
	France-New Zealand	207	Y	1,00	N	0,00	3,50	4,51	4,53
	Italy-Finland	867	Y	1,00	N	0,00	4,10	4,40	2,16
	New York-Russia	563	Y	1,00	N	0,00	3,80	4,36	2,90
	France-Turkey	115	Y	1,00	N	0,00	2,90	4,34	5,85

ANNEX L: Classification of Routes

Non Hierarchical Classification (NH-Means) of "Routes"

Initial Cluster Centers

Cluster	n	Step6Mean	Distance
1	25154	2.31	9311.61
2	2903	1.20	5728.41
3	30234	2.78	0.00
4	5631	2.52	10029.80
5	71	.92	0.00
6	18828	2.06	1221.08
7	2	3.80	9540.15
8	11191	2.93	8224.40
9	3396	2.69	1451.32
10	6440	1.23	0.00
11	109	4.13	19754.78
12	28989	2.77	16185.02
13	4029	2.55	16925.77
14	3	1.30	3601.60
15	35656	2.93	1465.04
16	20791	1.35	6819.10
17	6248	2.38	6707.64
18	17812	2.27	9059.39
19	3	4.60	14303.07
20	10564	2.25	1127.57

Final Cluster Centers

Cluster	n	Step6Mean	Distance
1	24659	2.19	9661.96
2	50	2.84	6702.77
3	30234	2.78	0.00
4	3742	2.20	9243.45
5	55	2.06	1218.73
6	20340	2.77	1810.61
7	76	3.25	8822.07
8	10640	2.96	9100.55
9	2061	2.17	1262.55
10	4940	1.65	704.84
11	35	4.26	14684.49
12	28989	2.77	16185.02
13	238	4.53	17197.20
14	40	3.36	3641.40
15	35656	2.93	1465.04
16	20791	1.35	6819.10
17	3066	2.13	6145.71
18	17812	2.27	9059.39
19	30	3.46	11449.77
20	10564	2.25	1127.57

Number of Cases in each Cluster I

Cluster	1	2	3	4	5	6	7	8	9	10
Cases	2	1037	1	15	1180	2	1005	2	31	6

Number of Cases in each Cluster II

Cluster	11	12	13	14	15	16	17	18	19	20	Valid	Missing
Cases	190	1	73	659	1	1	17	1	584	1	4809	0

ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Class	19	3255415	171338	39359	<2,2E-16
Residuals	596504	2596705	4,0		

Residuals

Min	1st Q	Median	3rd Q	Max
-2.957	-0,823	-0,290	0,4	1,78E+02

ANNEX M: R code for outliers classification (*Step 1*)

```
(# Table with average Net Timespan values per group)
x<-tapply(M1$Step1Net,list(M1$WD,M1$BL,M1$BC,M1$PC),mean,table=T)

#Create groups within Training matrix
        Train1$Group<-paste(Train1$PC,Train1$BC,Train1$BL,Train1$WD,sep="")
        colnames(Train1)[8]<-"Group"

# Create table with descriptive statistics
install.packages("plyr")
library(plyr)
library(dplyr)
grp <- group_by(Train1,Group=Group)
Q3<-summarise(grp, Q3=quantile(Step1Net,0.75))
Q1<-summarise(grp, Q1=quantile(Step1Net,0.25))
Q<-merge(Q1,Q3,by="Group")
rm(Q1)
rm(Q3)

#Merge quarters to Train dataset
TrainT<-merge(TrainT,Q,by="Group")
TrainT$IQR<-TrainT$Q3-TrainT$Q1

#Classify as mild or extreme outliers
TrainT$MOutlier<-ifelse(TrainT$Step1Net<TrainT$Q1-1.5*TrainT$IQR |
TrainT$Step1Net>TrainT$Q3+1.5*TrainT$IQR,1,0)
TrainT$EOutlier<-ifelse(TrainT$Step1Net<TrainT$Q1-3*TrainT$IQR |
TrainT$Step1Net>TrainT$Q3+3*TrainT$IQR,1,0)
TrainT$MOutlier<-as.factor(TrainT$MOutlier)
TrainT$EOutlier<-as.factor(TrainT$EOutlier)

(#remove extra information)
TrainT<-cbind(TrainT[,1:11],TrainT[,13:14])
```

ANNEX N: R code for regression trees (*Step 1*)

```
install.packages("partykit")
library(partykit)
library(Formula)
#Create tree
fit <- partykit::ctree(Step1Net~WD + BL + PC +
BC,data=Train[which(Train$EOutlier==0),])

# Print tree in tree-like format
print(fit)

#Get tree node characteristics
id<-nodeids(fit)
id<-nodeids(fit,terminal=T)
nodeprune(fit,id)
depth(fit)
width(fit)

# Get predictions for Test data
pred1<-predict(fit,newdata=Test)

#Calculate errors
Errors<-cbind(pred1,Test$Step1Net)
colnames(Errors)[1]<-"Estimation"
colnames(Errors)[2]<-"RealValue"
Errors <-cbind(Errors, round(Errors[,1]- Errors [,2],2))
colnames(Errors)[3]<-"Error"
Errors <-cbind(Errors, round(abs(Errors [,3]),2))
colnames(Errors)[4]<-"AbsoluteError"
Erros<-cbind(Errors,ifelse(Errors[,3]<0,0,Errors[,3]))
colnames(Errors)[5]<-"Delay"
Errors <-cbind(Errors, round(Errors [,3] x Errors [,3],2))
colnames(Erros)[6]<-"SqError"
Erros<-cbind(Erros,100xErros[,3]/Erros[,2])
colnames(Erros)[7]<-"PercError"
Erros<-cbind(Erros,abs(Erros[,6]))
colnames(Erros)[8]<-"AbsPercError"

#Get error statistics
mean(Erros[,3])
```

```
summary(Erros[,3])

mean(Erros[which(Erros[,2]!=0),7])

summary(Erros[which(Erros[,2]!=0),7])


#plot Errors and Delay

plot(ecdf(Erros[,3]),main="Step 1 Error Distribution",xlab="Error
(days)",cex.lab=1.5, cex.axis=1.5, cex.main=1.5)

plot(ecdf(Erros[,4]),main="Step 1 Absolute Error Distribution",xlab="Error
(days)",cex.lab=1.5, cex.axis=1.5, cex.main=1.5,)

plot(ecdf(Erros[,5]),main="Step 1 Delay",xlab="Error (days)",cex.lab=1.5,
cex.axis=1.5, cex.main=1.5)

plot(ecdf(Erros[,6]),main="Step 1 Squared Error Distribution",xlab="Error
(days^2)",cex.lab=1.5, cex.axis=1.5,cey.axis=1.5,cex.main=1.5)

plot(ecdf(Erros[which(Erros[,2]!=0),7]),main="Step 1 Percentual Error
Distribution",xlab="Error",cex.lab=1.5, cex.axis=1.5, cex.main=1.5)

plot(ecdf(Erros[which(Erros[,2]!=0),8]),main="Step 1 Absolute Percentual Error
Distribution",xlab="Error",cex.lab=1.5,cex.axis=1.5 ,cex.main=1.5,cex.main=1.5)
```

ANNEX O: Step 1 Regression tree (Step 1, first model)

```
Model formula:
Step1Net ~ WD + BL + PC + BC

Fitted party:
[1] root
|   [2] PC in N
|   |   [3] BL in A
|   |   |   [4] BC in 1, 4, 6, 7, 8, 9, 10
|   |   |   |   [5] WD in 1, 2
|   |   |   |   |   [6] WD in 1
|   |   |   |   |   |   [7] BC in 1, 4, 6, 9, 10
|   |   |   |   |   |   |   [8] BC in 1, 9: 0.458 (n = 8270, err = 830.7)
|   |   |   |   |   |   |   [9] BC in 4, 6, 10
|   |   |   |   |   |   |   |   [10] BC in 10: 0.414 (n = 13397, err = 1425.2)
|   |   |   |   |   |   |   |   [11] BC in 4, 6: 0.425 (n = 15673, err = 1410.2)
|   |   |   |   |   |   [12] BC in 7, 8
|   |   |   |   |   |   |   [13] BC in 8: 0.386 (n = 15689, err = 889.4)
|   |   |   |   |   |   |   [14] BC in 7: 0.407 (n = 1336, err = 171.7)
|   |   |   |   |   [15] WD in 2
|   |   |   |   |   |   [16] BC in 1, 4, 9: 0.427 (n = 6757, err = 843.0)
|   |   |   |   |   |   [17] BC in 6, 7, 8, 10
|   |   |   |   |   |   |   [18] BC in 6, 7, 10: 0.349 (n = 29912, err = 3489.5)
|   |   |   |   |   |   |   [19] BC in 8: 0.334 (n = 17060, err = 1825.7)
|   |   |   |   [20] WD in 3, 7
|   |   |   |   |   [21] BC in 1, 4, 7, 9
|   |   |   |   |   |   [22] WD in 7
|   |   |   |   |   |   |   [23] BC in 1, 4, 7
|   |   |   |   |   |   |   |   [24] BC in 1: 0.408 (n = 1503, err = 272.7)
|   |   |   |   |   |   |   |   [25] BC in 4, 7: 0.301 (n = 3801, err = 619.4)
|   |   |   |   |   |   |   [26] BC in 9: 0.221 (n = 8564, err = 905.7)
|   |   |   |   |   |   [27] WD in 3
|   |   |   |   |   |   |   [28] BC in 1, 7, 9
|   |   |   |   |   |   |   |   [29] BC in 1: 0.301 (n = 4656, err = 331.6)
|   |   |   |   |   |   |   |   [30] BC in 7, 9: 0.321 (n = 50323, err = 5078.0)
|   |   |   |   |   |   |   [31] BC in 4: 0.388 (n = 6981, err = 930.4)
|   |   |   |   |   [32] BC in 6, 8, 10
|   |   |   |   |   |   [33] BC in 10
|   |   |   |   |   |   |   [34] WD in 7: 0.194 (n = 13465, err = 1295.7)
|   |   |   |   |   |   |   [35] WD in 3: 0.252 (n = 65567, err = 4947.4)
|   |   |   |   |   |   [36] BC in 6, 8
|   |   |   |   |   |   |   [37] WD in 3
|   |   |   |   |   |   |   |   [38] BC in 8: 0.277 (n = 78091, err = 7614.8)
|   |   |   |   |   |   |   |   [39] BC in 6: 0.281 (n = 72250, err = 7702.1)
|   |   |   |   |   |   |   [40] WD in 7
|   |   |   |   |   |   |   |   [41] BC in 8: 0.238 (n = 15762, err = 1664.0)
|   |   |   |   |   |   |   |   [42] BC in 6: 0.284 (n = 14869, err = 2299.9)
|   |   |   [43] BC in 2, 3, 5
|   |   |   |   [44] WD in 1, 2, 3
|   |   |   |   |   [45] BC in 3
|   |   |   |   |   |   [46] WD in 1, 2: 0.673 (n = 472, err = 67.6)
|   |   |   |   |   |   [47] WD in 3: 0.466 (n = 5251, err = 660.0)
|   |   |   |   |   [48] BC in 2, 5
|   |   |   |   |   |   [49] WD in 1
|   |   |   |   |   |   |   [50] BC in 5: 0.483 (n = 1155, err = 5.2)
|   |   |   |   |   |   |   [51] BC in 2: 0.534 (n = 255, err = 10.1)
|   |   |   |   |   |   [52] WD in 2, 3: 0.524 (n = 15616, err = 2643.7)
|   |   |   |   [53] WD in 7
|   |   |   |   |   [54] BC in 2, 5
|   |   |   |   |   |   [55] BC in 5: 0.482 (n = 1433, err = 8.8)
|   |   |   |   |   |   [56] BC in 2: 0.456 (n = 1214, err = 8.8)
|   |   |   |   [57] BC in 3: 0.394 (n = 1631, err = 352.5)
|   |   [58] BL in B, C
|   |   |   [59] BL in C
|   |   |   |   [60] BC in 1, 2
|   |   |   |   |   [61] BC in 1: 1.481 (n = 958, err = 323.3)
|   |   |   |   |   [62] BC in 2: 1.276 (n = 170, err = 43.4)
|   |   |   |   [63] BC in 3
|   |   |   |   |   [64] WD in 1, 2: 1.655 (n = 817, err = 436.7)
|   |   |   |   |   [65] WD in 3: 2.029 (n = 1580, err = 1225.8)
|   |   |   [66] BL in B
```

```
|   |   |   |   [67] BC in 1, 4, 7, 9, 10
|   |   |   |   |   [68] WD in 1, 2
|   |   |   |   |   |   [69] BC in 1: 0.536 (n = 5973, err = 391.5)
|   |   |   |   |   |   [70] BC in 4, 7, 9, 10
|   |   |   |   |   |   |   [71] WD in 1
|   |   |   |   |   |   |   |   [72] BC in 4, 7, 9
|   |   |   |   |   |   |   |   |   [73] BC in 7, 9: 0.499 (n = 3019, err = 421.1)
|   |   |   |   |   |   |   |   |   [74] BC in 4: 0.433 (n = 1152, err = 7.0)
|   |   |   |   |   |   |   |   [75] BC in 10: 0.672 (n = 722, err = 142.1)
|   |   |   |   |   |   |   [76] WD in 2
|   |   |   |   |   |   |   |   [77] BC in 4, 7, 9: 0.481 (n = 7917, err = 1276.5)
|   |   |   |   |   |   |   |   [78] BC in 10: 0.416 (n = 2248, err = 367.8)
|   |   |   |   |   [79] WD in 3, 7
|   |   |   |   |   |   [80] BC in 1, 4, 9
|   |   |   |   |   |   |   [81] WD in 7
|   |   |   |   |   |   |   |   [82] BC in 1, 9
|   |   |   |   |   |   |   |   |   [83] BC in 1: 0.411 (n = 1902, err = 331.4)
|   |   |   |   |   |   |   |   |   [84] BC in 9: 0.343 (n = 655, err = 167.2)
|   |   |   |   |   |   |   |   [85] BC in 4: 0.140 (n = 364, err = 16.4)
|   |   |   |   |   |   |   [86] WD in 3
|   |   |   |   |   |   |   |   [87] BC in 1, 9: 0.433 (n = 17058, err = 1941.3)
|   |   |   |   |   |   |   |   [88] BC in 4: 0.412 (n = 2825, err = 325.0)
|   |   |   |   |   |   [89] BC in 7, 10
|   |   |   |   |   |   |   [90] WD in 3: 0.493 (n = 7670, err = 1399.7)
|   |   |   |   |   |   |   [91] WD in 7: 0.394 (n = 913, err = 211.9)
|   |   |   |   [92] BC in 2, 3, 5, 6, 8
|   |   |   |   |   [93] WD in 1, 3
|   |   |   |   |   |   [94] BC in 2, 8
|   |   |   |   |   |   |   [95] WD in 1: 0.500 (n = 2468, err = 32.0)
|   |   |   |   |   |   |   [96] WD in 3: 0.670 (n = 11409, err = 2734.5)
|   |   |   |   |   |   [97] BC in 3, 5, 6
|   |   |   |   |   |   |   [98] BC in 3
|   |   |   |   |   |   |   |   [99] WD in 1: 0.854 (n = 1766, err = 481.3)
|   |   |   |   |   |   |   |   [100] WD in 3: 0.672 (n = 6274, err = 1360.7)
|   |   |   |   |   |   |   [101] BC in 5, 6
|   |   |   |   |   |   |   |   [102] WD in 1: 0.529 (n = 1678, err = 43.2)
|   |   |   |   |   |   |   |   [103] WD in 3: 0.748 (n = 8879, err = 2210.0)
|   |   |   |   |   [104] WD in 2, 7
|   |   |   |   |   |   [105] BC in 2, 5, 6, 8
|   |   |   |   |   |   |   [106] BC in 2, 5, 6
|   |   |   |   |   |   |   |   [107] WD in 7: 0.771 (n = 2824, err = 1074.6)
|   |   |   |   |   |   |   |   [108] WD in 2: 0.724 (n = 6606, err = 1178.7)
|   |   |   |   |   |   |   [109] BC in 8
|   |   |   |   |   |   |   |   [110] WD in 7: 0.756 (n = 101, err = 58.3)
|   |   |   |   |   |   |   |   [111] WD in 2: 0.569 (n = 875, err = 263.6)
|   |   |   |   |   |   [112] BC in 3: 0.869 (n = 3153, err = 882.4)
|   [113] PC in Y
|   |   [114] BL in A, B
|   |   |   [115] BC in 1, 4, 6, 7, 8, 9, 10
|   |   |   |   [116] WD in 1, 2
|   |   |   |   |   [117] BL in A
|   |   |   |   |   |   [118] WD in 1
|   |   |   |   |   |   |   [119] BC in 10: 0.452 (n = 1842, err = 262.4)
|   |   |   |   |   |   |   [120] BC in 6, 7, 8, 9: 0.424 (n = 7980, err = 990.3)
|   |   |   |   |   |   [121] WD in 2
|   |   |   |   |   |   |   [122] BC in 7, 8, 10
|   |   |   |   |   |   |   |   [123] BC in 10: 0.361 (n = 2939, err = 403.8)
|   |   |   |   |   |   |   |   [124] BC in 7, 8: 0.366 (n = 5548, err = 833.8)
|   |   |   |   |   |   |   [125] BC in 6, 9: 0.342 (n = 6960, err = 922.7)
|   |   |   |   |   [126] BL in B
|   |   |   |   |   |   [127] BC in 1, 6, 7, 9
|   |   |   |   |   |   |   [128] BC in 1, 7, 9
|   |   |   |   |   |   |   |   [129] BC in 1, 9: 0.571 (n = 7832, err = 2420.9)
|   |   |   |   |   |   |   |   [130] BC in 7
|   |   |   |   |   |   |   |   |   [131] WD in 1: 0.553 (n = 660, err = 221.0)
|   |   |   |   |   |   |   |   |   [132] WD in 2: 0.650 (n = 1314, err = 469.7)
|   |   |   |   |   |   |   [133] BC in 6
|   |   |   |   |   |   |   |   [134] WD in 1: 1.497 (n = 65, err = 6.0)
|   |   |   |   |   |   |   |   [135] WD in 2: 0.771 (n = 174, err = 81.1)
|   |   |   |   |   |   [136] BC in 4, 8, 10: 0.494 (n = 9112, err = 2285.7)
|   |   |   |   [137] WD in 3, 7
|   |   |   |   |   [138] BL in A
```

```
|   |   |   |   |   |   [139] BC in 6, 8, 10
|   |   |   |   |   |   |   [140] WD in 7
|   |   |   |   |   |   |   |   [141] BC in 6, 8: 0.272 (n = 8463, err = 1360.5)
|   |   |   |   |   |   |   |   [142] BC in 10: 0.205 (n = 2737, err = 283.9)
|   |   |   |   |   |   |   [143] WD in 3
|   |   |   |   |   |   |   |   [144] BC in 6, 10
|   |   |   |   |   |   |   |   |   [145] BC in 6: 0.288 (n = 26861, err = 3172.3)
|   |   |   |   |   |   |   |   |   [146] BC in 10: 0.279 (n = 17694, err = 1891.0)
|   |   |   |   |   |   |   |   [147] BC in 8: 0.297 (n = 27227, err = 3259.1)
|   |   |   |   |   |   [148] BC in 4, 7, 9
|   |   |   |   |   |   |   [149] BC in 9
|   |   |   |   |   |   |   |   [150] WD in 3: 0.386 (n = 7686, err = 1216.1)
|   |   |   |   |   |   |   |   [151] WD in 7: 0.289 (n = 1676, err = 293.9)
|   |   |   |   |   |   |   [152] BC in 4, 7
|   |   |   |   |   |   |   |   [153] WD in 3
|   |   |   |   |   |   |   |   |   [154] BC in 4: 0.458 (n = 1035, err = 176.4)
|   |   |   |   |   |   |   |   |   [155] BC in 7: 0.369 (n = 1169, err = 131.5)
|   |   |   |   |   |   |   |   [156] WD in 7
|   |   |   |   |   |   |   |   |   [157] BC in 7: 0.667 (n = 356, err = 164.4)
|   |   |   |   |   |   |   |   |   [158] BC in 4: 0.243 (n = 286, err = 43.2)
|   |   |   |   |   [159] BL in B
|   |   |   |   |   |   [160] WD in 7
|   |   |   |   |   |   |   [161] BC in 1, 4, 7, 10
|   |   |   |   |   |   |   |   [162] BC in 1: 0.211 (n = 1284, err = 185.2)
|   |   |   |   |   |   |   |   [163] BC in 4, 7, 10: 0.169 (n = 2091, err = 266.0)
|   |   |   |   |   |   |   [164] BC in 6, 8, 9
|   |   |   |   |   |   |   |   [165] BC in 9: 0.309 (n = 1454, err = 305.7)
|   |   |   |   |   |   |   |   [166] BC in 6, 8: 0.420 (n = 582, err = 295.9)
|   |   |   |   |   |   [167] WD in 3
|   |   |   |   |   |   |   [168] BC in 1, 8
|   |   |   |   |   |   |   |   [169] BC in 1: 0.551 (n = 5529, err = 679.0)
|   |   |   |   |   |   |   |   [170] BC in 8: 0.646 (n = 2037, err = 647.8)
|   |   |   |   |   |   |   [171] BC in 4, 6, 7, 9, 10
|   |   |   |   |   |   |   |   [172] BC in 6, 7, 9: 0.424 (n = 13775, err = 2439.5)
|   |   |   |   |   |   |   |   [173] BC in 4, 10: 0.388 (n = 8099, err = 1305.8)
|   |   |   [174] BC in 2, 3, 5
|   |   |   |   [175] BL in A
|   |   |   |   |   [176] BC in 2, 5: 0.615 (n = 1179, err = 322.4)
|   |   |   |   |   [177] BC in 3: 0.509 (n = 455, err = 114.4)
|   |   |   |   [178] BL in B
|   |   |   |   |   [179] BC in 2, 5
|   |   |   |   |   |   [180] WD in 1, 2
|   |   |   |   |   |   |   [181] BC in 5
|   |   |   |   |   |   |   |   [182] WD in 1: 1.074 (n = 894, err = 460.8)
|   |   |   |   |   |   |   |   [183] WD in 2: 0.984 (n = 1224, err = 504.3)
|   |   |   |   |   |   |   [184] BC in 2
|   |   |   |   |   |   |   |   [185] WD in 1: 0.898 (n = 639, err = 236.5)
|   |   |   |   |   |   |   |   [186] WD in 2: 0.750 (n = 946, err = 154.7)
|   |   |   |   |   |   [187] WD in 3, 7
|   |   |   |   |   |   |   [188] BC in 5
|   |   |   |   |   |   |   |   [189] WD in 7: 0.741 (n = 780, err = 300.8)
|   |   |   |   |   |   |   |   [190] WD in 3: 0.830 (n = 4657, err = 1526.8)
|   |   |   |   |   |   |   [191] BC in 2
|   |   |   |   |   |   |   |   [192] WD in 3: 0.690 (n = 5426, err = 1161.0)
|   |   |   |   |   |   |   |   [193] WD in 7: 0.874 (n = 555, err = 275.1)
|   |   |   |   |   [194] BC in 3
|   |   |   |   |   |   [195] WD in 1, 2, 3
|   |   |   |   |   |   |   [196] WD in 1, 2
|   |   |   |   |   |   |   |   [197] WD in 1: 1.440 (n = 516, err = 273.7)
|   |   |   |   |   |   |   |   [198] WD in 2: 1.196 (n = 1560, err = 600.1)
|   |   |   |   |   |   |   [199] WD in 3: 1.008 (n = 2512, err = 1194.5)
|   |   |   |   |   |   [200] WD in 7: 0.496 (n = 540, err = 184.9)
|   |   [201] BL in C
|   |   |   [202] BC in 1, 2, 4, 5
|   |   |   |   [203] BC in 1, 2
|   |   |   |   |   [204] WD in 1, 2, 7
|   |   |   |   |   |   [205] BC in 1
|   |   |   |   |   |   |   [206] WD in 1, 7
|   |   |   |   |   |   |   |   [207] WD in 1: 1.536 (n = 1135, err = 446.0)
|   |   |   |   |   |   |   |   [208] WD in 7: 1.437 (n = 620, err = 281.0)
|   |   |   |   |   |   |   [209] WD in 2: 1.734 (n = 882, err = 836.1)
|   |   |   |   |   |   [210] BC in 2: 1.220 (n = 261, err = 98.1)
|   |   |   |   |   [211] WD in 3: 1.066 (n = 948, err = 109.3)
```

```
|   |   |   |   [212] BC in 4, 5
|   |   |   |   |   [213] BC in 5: 0.792 (n = 190, err = 57.4)
|   |   |   |   |   [214] BC in 4
|   |   |   |   |   |   [215] WD in 1: 0.380 (n = 207, err = 24.1)
|   |   |   |   |   |   [216] WD in 2: 0.515 (n = 184, err = 22.5)
|   |   |   [217] BC in 3, 9
|   |   |   |   [218] WD in 1, 2
|   |   |   |   |   [219] BC in 3
|   |   |   |   |   |   [220] WD in 1: 2.817 (n = 270, err = 53.0)
|   |   |   |   |   |   [221] WD in 2: 3.313 (n = 528, err = 66.4)
|   |   |   |   |   [222] BC in 9: 2.329 (n = 160, err = 90.9)
|   |   |   |   [223] WD in 3, 7
|   |   |   |   |   [224] WD in 7: 2.069 (n = 232, err = 65.6)
|   |   |   |   |   [225] WD in 3: 1.818 (n = 2346, err = 865.5)

Number of inner nodes:    112
Number of terminal nodes: 113
```

ANNEX P: Main Routes Error Measures

		Mean				Median			
Route	**n**	**ME**	**MAE**	***MDelay***	**MSE**	**ME**	**MAE**	***MDelay***	**MSE**
Italy-UK	35656	-0,13	0,65	0,39	2,94	-0,37	0,60	0,49	3,07
Brazil-Brazil	30234	-0,36	1,56	0,96	11,67	-0,64	1,54	1,09	11,96
Italy-Australia	28989	-0,08	0,74	0,41	2,22	-0,28	0,73	0,50	2,29
Italy-Hong Kong	25154	-0,08	0,64	0,36	1,69	-0,43	0,56	0,49	1,87
Italy-California	24163	-0,13	0,54	0,34	2,42	-0,18	0,52	0,35	2,45
Italy-New York	20791	-0,08	0,58	0,33	2,47	-0,34	0,50	0,42	2,58
Italy-Germany	18828	-0,06	0,51	0,28	1,12	-0,25	0,46	0,35	1,18
Italy-China	11191	-0,41	1,25	0,83	7,11	-0,71	1,23	0,97	7,47
Italy-France	10564	-0,15	0,74	0,44	2,29	-0,50	0,68	0,59	2,53
Italy-Japan	10088	-0,08	0,95	0,52	2,17	-0,35	0,93	0,64	2,30
Italy-Italy	6440	-0,09	0,54	0,31	1,29	-0,34	0,49	0,41	1,41
Italy-Canada	6248	-0,14	0,77	0,46	2,72	-0,37	0,73	0,55	2,85
Italy-Macau	5992	-0,24	1,16	0,70	7,51	-0,48	1,13	0,80	7,68
Italy-Singapore	5631	-0,08	0,76	0,42	1,77	-0,50	0,68	0,59	2,01
Italy-Florida	5239	-0,08	0,48	0,28	1,24	-0,14	0,46	0,30	1,26
Italy-Brazil	5044	-0,73	1,47	1,10	21,42	-1,10	1,44	1,27	22,09
Italy-Netherlands	4955	-0,06	0,48	0,27	0,63	-0,28	0,42	0,35	0,70
Italy-Taiwan	4924	-0,08	0,58	0,33	1,60	-0,19	0,55	0,37	1,65
Italy-Spain	4707	-0,05	0,79	0,42	1,66	-0,07	0,78	0,42	1,68
Italy-Massachusetts	4561	-0,06	0,51	0,29	1,64	-0,12	0,49	0,30	1,67
Italy-New Jersey	4456	-0,05	0,50	0,28	0,84	-0,28	0,44	0,36	0,92
Italy-Saudi Arabia	4301	-0,20	1,40	0,80	5,57	-0,65	1,36	1,01	5,97
Italy-Texas	4270	-0,13	0,51	0,32	1,40	-0,17	0,49	0,33	1,43
France-Australia	4029	-0,20	0,68	0,44	6,45	-0,29	0,68	0,49	6,49
Italy-Illinois	4009	-0,11	0,48	0,29	1,47	-0,13	0,47	0,30	1,48
Italy-Pennsylvania	3870	-0,10	0,53	0,32	1,23	-0,13	0,51	0,32	1,25
Spain-UK	3834	-0,14	0,26	0,20	0,85	-0,18	0,25	0,22	0,86
Italy-UAE	3591	-0,11	0,71	0,41	1,48	-0,35	0,66	0,50	1,59
France-California	3469	-0,12	0,42	0,27	0,84	-0,02	0,40	0,21	0,85
Italy-Poland	3398	-0,06	0,45	0,25	2,98	-0,25	0,40	0,32	3,03
Italy-Switzerland	3339	-0,09	0,50	0,30	0,99	-0,31	0,44	0,38	1,08
Italy-Washington	3116	-0,02	0,52	0,27	1,62	-0,10	0,48	0,29	1,64
Poland-Australia	3085	-0,16	0,63	0,40	2,21	-0,27	0,62	0,45	2,25
France-Hong Kong	3014	-0,09	0,29	0,19	0,31	-0,15	0,27	0,21	0,33
France-New York	2903	-0,26	0,40	0,33	2,12	-0,32	0,40	0,36	2,15
Italy-Kuwait	2894	-0,06	0,69	0,37	0,80	-0,39	0,67	0,53	0,95
New York-California	2832	-0,06	0,57	0,31	0,61	-0,29	0,56	0,42	0,71
Italy-Maryland	2786	-0,09	0,53	0,31	0,76	-0,15	0,52	0,33	0,79
Italy-Ohio	2770	-0,07	0,51	0,29	1,34	-0,05	0,50	0,27	1,37
Spain-Hong Kong	2741	-0,23	0,37	0,30	1,42	-0,25	0,36	0,31	1,43
Spain-Australia	2568	-0,16	0,66	0,41	1,97	-0,27	0,66	0,46	2,03
Italy-Michigan	2527	-0,08	0,51	0,29	1,08	-0,11	0,48	0,29	1,09
Germany-Germany	2507	-0,10	0,23	0,16	0,36	-0,06	0,22	0,14	0,35
Germany-Australia	2486	-0,03	0,61	0,32	0,88	-0,08	0,60	0,34	0,88
Spain-California	2483	-0,18	0,49	0,33	2,07	-0,07	0,47	0,27	2,05
Italy-Belgium	2403	-0,13	0,58	0,36	1,29	-0,37	0,52	0,45	1,40
Italy-Portugal	2389	-0,02	0,69	0,36	1,19	0,00	0,67	0,34	1,20
Italy-Romania	2341	-0,07	0,58	0,33	0,56	-0,41	0,56	0,48	0,72
Italy-Austria	2323	-0,10	0,53	0,31	1,15	-0,32	0,50	0,41	1,24
Spain-New York	2273	-0,33	0,57	0,45	5,94	-0,45	0,56	0,50	6,04
Italy-Azerbaijan	2247	-0,40	1,44	0,92	16,08	-0,68	1,42	1,05	16,46
Poland-California	2212	-0,24	0,61	0,43	10,61	-0,16	0,59	0,37	10,61
Italy-Denmark	2170	-0,27	0,74	0,50	2,34	-0,54	0,70	0,62	2,55
New York-New York	2145	-0,10	0,31	0,21	0,44	-0,11	0,31	0,21	0,44
France-France	2082	-0,14	0,62	0,38	1,03	-0,44	0,57	0,51	1,23
Germany-Hong Kong	2014	-0,26	0,35	0,31	1,61	-0,25	0,35	0,30	1,61
Spain-Germany	1974	-0,19	0,30	0,24	2,19	-0,23	0,29	0,26	2,20
Italy-Oregon	1965	-0,16	0,48	0,32	0,56	-0,15	0,46	0,30	0,56
Italy-Virginia	1953	-0,20	0,64	0,42	3,57	-0,21	0,61	0,41	3,56
France-Germany	1922	-0,12	0,23	0,17	0,45	-0,12	0,22	0,17	0,45
Italy-Georgia US	1911	-0,17	0,55	0,36	3,28	-0,21	0,54	0,37	3,28
Germany-California	1872	-0,08	0,42	0,25	0,85	-0,01	0,39	0,20	0,85
Germany-New York	1858	-0,26	0,41	0,33	3,19	-0,28	0,40	0,34	3,19
Poland-Hong Kong	1850	-0,18	0,43	0,30	0,63	-0,31	0,41	0,36	0,69
Italy-Indiana	1819	-0,08	0,49	0,28	1,00	-0,14	0,46	0,30	1,02
Italy-Bulgaria	1757	0,03	0,56	0,27	0,45	-0,38	0,49	0,43	0,59
Italy-Kazakhstan	1754	-0,19	1,89	1,04	8,84	-0,81	1,78	1,30	9,48
Poland-New York	1752	-0,25	0,42	0,33	2,35	-0,26	0,41	0,33	2,35
Italy-Sweden	1702	-0,20	0,91	0,55	2,36	-0,54	0,91	0,73	2,64
Italy-Connecticut	1603	-0,01	0,60	0,31	0,60	-0,04	0,59	0,31	0,67

Italy-Arizona	1556	-0,26	0,63	0,44	4,46	-0,27	0,61	0,44	4,50
Italy-Qatar	1493	-0,20	0,85	0,53	4,65	-0,47	0,80	0,63	4,83
Poland-Brazil	1486	-0,51	1,23	0,87	8,61	-0,83	1,18	1,01	9,01
Italy-Greece	1370	-0,16	0,88	0,52	1,95	-0,27	0,88	0,58	2,06
Italy-Croatia	1332	-0,05	0,61	0,33	0,70	-0,38	0,59	0,49	0,84
California-New York	1270	-0,23	0,51	0,37	4,58	-0,28	0,50	0,39	4,61
Italy-Colorado	1183	-0,09	0,56	0,32	0,83	-0,19	0,52	0,35	0,86
France-China	1179	-0,43	1,15	0,79	5,36	-0,80	1,11	0,95	5,81
Spain-France	1179	-0,33	0,44	0,38	2,52	-0,33	0,44	0,38	2,52
Spain-China	1173	-0,51	1,27	0,89	5,11	-0,96	1,23	1,10	5,77
Poland-Germany	1147	-0,10	0,19	0,15	0,56	-0,07	0,18	0,13	0,55
Poland-China	1129	-0,69	1,40	1,05	8,45	-1,02	1,40	1,21	9,02
Italy-Malaysia	1109	-0,41	1,29	0,85	7,43	-0,76	1,25	1,01	7,82
Italy-Bahrain	1108	-0,06	0,70	0,38	1,16	-0,32	0,61	0,46	1,24
Belgium-Australia	1097	-0,16	0,59	0,37	1,01	-0,24	0,59	0,41	1,05
Italy-Cambodia	1064	-0,12	0,65	0,38	0,82	-0,48	0,61	0,55	1,04
Italy-New Zealand	1041	-0,58	1,73	1,15	21,42	-0,91	1,74	1,32	22,25
Belgium-Hong Kong	989	-0,15	0,31	0,23	0,66	-0,17	0,29	0,23	0,67
Spain-Japan	965	-0,15	0,84	0,49	1,51	-0,29	0,82	0,56	1,55
Belgium-California	953	-0,11	0,40	0,25	0,77	-0,02	0,37	0,19	0,76
New York-Hong Kong	940	-0,06	0,61	0,34	2,80	-0,26	0,60	0,43	2,90
Italy-Israel	939	-0,47	1,26	0,86	8,58	-0,63	1,24	0,93	8,73
Italy-Ireland	926	-0,04	0,72	0,38	1,20	-0,36	0,66	0,51	1,33
Italy-Ukraine	914	-0,13	1,54	0,84	5,87	-0,92	1,42	1,17	6,69
Italy-North Carolina	892	-0,09	0,46	0,27	0,55	-0,11	0,46	0,29	0,59
Luxembourg-Australia	868	-0,03	0,51	0,27	0,60	-0,06	0,50	0,28	0,61
Luxembourg-Hong Kong	858	-0,28	0,46	0,37	3,22	-0,34	0,44	0,39	3,25
Germany-China	856	-0,36	1,14	0,75	4,64	-0,67	1,08	0,87	4,95
Italy-Slovakia	856	-0,02	0,62	0,32	0,77	-0,05	0,61	0,33	0,79
New York-Australia	846	-0,06	0,59	0,32	1,09	-0,18	0,59	0,38	1,12
Italy-Turkey	837	-0,65	2,04	1,34	28,86	-1,12	1,95	1,54	29,81
Italy-Cyprus	835	-0,10	0,63	0,37	0,89	-0,17	0,62	0,40	0,90
Italy-Iowa	818	-0,21	0,56	0,39	6,31	-0,19	0,53	0,36	6,33
Belgium-New York	805	-0,22	0,37	0,29	1,04	-0,26	0,36	0,31	1,07
Italy-Minnesota	802	-0,09	0,51	0,30	0,72	-0,13	0,50	0,32	0,75