Tampere University

Antti Ryhänen

DESIGNING CUSTOMER SEGMENTATION MODEL FOR ANALYSING CONSUMER DATA

Case: Consumer segmentation model for retail sales

Master's Thesis in Technology
Management and Business
Samuli Pekkola
Jukka Huhtamäki
May 2023

ABSTRACT

Antti Ryhänen: Designing customer segmentation model for analysing consumer data
Master's Thesis in Technology
Tampere University
Master's Degree Programme in Information and Knowledge Management
May 2023

Customer segmentation has a crucial impact on today's highly competitive business environment, and it is especially affecting to organizations marketing processes. The main idea behind customer segmentation is to divide customers into homogeneous groups based on their various characteristics. This enables organizations to tailor their marketing strategies to the various needs of different customers, without creating distinct plan for each individual customer. Customer segmentation can also help organizations to understand their customers better and to find latent business opportunities. There are many different approaches to customer segmentation in the consumer markets, but the four major approaches are behavioural, demographic, geographic, and psychographic segmentation. To utilize these approaches, it is recommended to use machine calculation and various data scientific methods to be able to process bigger amounts of data.

The objective of this research was to design customer segmentation model, that is appropriate for analysing quantitative sales data in retailer consumer market. The purpose of the model is to show how behavioural based consumer segmentation could be implemented and utilized in the client organization. Research includes the theory and the empirical case study section, which follows the design science research as a strategic framework. The study explores various consumer segmentation approaches and data scientific methods that can be utilized in the segmentation model. The model is developed in case study, in which the identified methods are practically applied by using the client organization's consumer sales data. The study is also investigating data scientific frameworks that can be utilized in the iterative development process of segmentation model along with the design science.

The main result of this research is the consumer behaviour-based segmentation model, that utilizes customer's recency, frequency, and monetary based modelling as a base segmentation method, which is widely used in the behavioural segmentation. The segmentation model divides consumers into seven homogenous segments based on their buying behaviour during the last five years. The used method is easy to understand, and it enables arbitrary tailoring of the limit values and the labels of segments. Some additional geographic and product related attributes were also added to model as an explanatory features. Another segmentation method considered in this research is K-Means clustering. The study found that this unsupervised method would be a proper solution if more than three features were used as a dividing criterion in the segmentation. However, clustering is not completely excluded from the model, as it offers a good comparison for manually created segments and enables several further development opportunities.

Keywords: Case study, Consumer segmentation, CRISP-DM, Customer segmentation, Data science, Design science research, Proof-of-Concept (PoC), K-Means clustering, RFM modelling

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

TIIVISTELMÄ

Antti Ryhänen: Asiakassegmentointimallin suunnittelu kuluttajadatan analysointiin
Diplomityö
Tampereen yliopisto
Tietojohtamisen DI-tutkinto-ohjelma
Toukokuu 2023

Asiakassegmentoinnilla on suuri vaikutus nykypäivän erittäin kilpailullisessa liiketoimintaympäristössä ja se vaikuttaa varsinkin organisaatioiden markkinointiprosesseihin. Asiakassegmentoinnilla tarkoitetaan samankaltaisten asiakkaiden jakamista ryhmiin erilaisten ominaisuuksien perusteella. Tämä mahdollistaa organisaatioiden markkinointistrategioiden räätälöimisen asiakkaiden eri tarpeisiin niin, että jokaiselle yksittäiselle asiakkaalle ei tarvitse luoda erillistä suunnitelmaa. Asiakassegmentointi voi myös auttaa organisaatioita ymmärtämään paremmin asiakkaitaan ja hahmottamaan piileviä liiketoimintamahdollisuuksia. Kuluttajamarkkinoiden asiakassegmentoinnille on olemassa useita erilaisia lähestymistapoja, mutta neljä pääasiallista lähestymistapaa ovat käyttäytymiseen, demografiaan, geografiaan ja psykografiaan perustuva segmentointi. Näiden lähestymistapojen hyödyntämiseen on suositeltavaa käyttää koneellista laskentaa ja datatieteellisiä metodeja, jotta suurempien datamassojen käsittely on mahdollista.

Tämän tutkimuksen tavoitteena oli suunnitella asiakassegmentointimalli, mikä soveltuu kvantitatiivisen myyntidatan analysointiin vähittäismyyntiyrityksen kuluttajamarkkinoilla. Mallin tarkoituksena on osoittaa, miten ostokäyttäytymiseen perustuvaa kuluttajasegmentointia voitaisiin toteuttaa ja hyödyntää kohdeorganisaatiossa. Tutkimus sisältää teoria osuuden ja empiirisen case tutkimuksen, mikä noudattaa suunnittelutiedettä strategisena viitekehyksenään. Tutkimuksessa tarkastellaan erilaisia kuluttajasegmentoinnin lähestymistapoja ja datatieteellisiä metodeja, joita voidaan hyödyntää segmentointimallissa. Mallin kehitys on kuvattu vaiheittain case tutkimuksessa, jossa tunnistettuja menetelmiä sovelletaan käytännön tasolla hyödyntäen asiakasorganisaation kuluttajamyyntidataa. Lisäksi tutkimuksessa tarkastellaan datatieteellisiä viitekehyksiä, joita voidaan hyödyntää suunnittelutieteen ohella segmentointimallin iteratiivisessa kehitysprosessissa.

Tutkimuksen tulos on kuluttajakäyttäytymiseen perustuva segmentointimalli, mikä hyödyntää asiakkaan viimeaikaisuuteen, frekvenssiin ja rahalliseen arvoon perustuvaa mallinnusta segmentoinnin perustana. Kyseinen metodi on laajalti käytetty työkalu asiakaskäyttäytymiseen perustuvassa segmentoinnissa. Segmentointimallissa kuluttajat pisteytetään heidän viimeisen viiden vuoden ostokäyttäytymisensä perusteella seitsemään eri ryhmään. Kyseinen metodi on helposti ymmärrettävä ja se mahdollistaa segmenttien raja-arvojen, sekä nimeämisten mielivaltaisen räätälöinnin. Lisäksi mallissa on hyödynnetty geografisia ja tuotteisiin liittyviä attribuutteja segmenttien selittämiseksi ja ymmärtämiseksi. Toinen tutkimuksessa harkittu segmentointimenetelmä on K-Means klusterointi. Tutkimuksen aikana todettiin, että tämä ohjaamattomaan oppimiseen perustuva menetelmä olisi sopiva, mikäli segmentointikriteereinä hyödynnettäisiin useampaa kuin kolmea ominaisuutta. Klusterointia ei kuitenkaan suljettu kokonaan pois lopullisesta mallista, sillä se tarjoaa hyvän vertailukohteen käsin luoduille segmenteille, sekä mahdollistaa useita jatkokehitysmahdollisuuksia.

Avainsanat: Asiakassegmentointi, CRISP-DM, Datatiede, Kuluttajasegmentointi, K-Means klusterointi, Proof-of-Concept (PoC), RFM mallinnus, Suunnittelutiede, Tapaustutkimus

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

PREFACE

In 2017, I started my Information and Knowledge Management studies at the Tampere University of Technology without a prior knowledge of what those studies might include, and what kind of future work opportunities my studies would open. Now, after six awesome years, I can say that I made the right choice. This thesis will be the last step of my studies, for now, and it feels surreal to finally graduate.

I would like to thank my supervisors, Professor Samuli Pekkola and Senior Research Fellow Jukka Huhtamäki, who offered expert guidance throughout the research process and were always available when the support was needed. I would also like to thank all the people from the client organization, who actively participated in the research process. In addition, I would like to thank my current employer and all the colleagues who sparred, helped, and supported me during the research.

During my studies, I have gotten to know amazing people, with whom I will be in contact even after my studies. Huge thanks to all of you. I would also like to mention that Hiki-Hockey is the most unique hockey team I have ever been a part of. Therefore, I also want to thank all my teammates during these years. Lastly, but not least, I would like to thank my family, closest friends, and Jenni who supported me throughout the journey.

Tampere, May 2023

Antti Ryhänen

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

AHP	Analytical Hierarchical Process
B2B	Business-to-Business
B2C	Business-to-Consumer
CRISP-DM	Cross-industry standard process for Data Mining
KDD	Knowledge Discovery in Databases
KPI	Key Performance Indicator
MVP	Minimum Viable Product
PCA	Principal Components Analysis
PoC	Proof-of-concept
RFM	Recency, Frequency, and Monetary
SQL	Structured Query Language
SSE	Sum of Squared Errors

1. INTRODUCTION

The first chapter of this research is introduction, which starts by familiarizing with the relevant research background. The research objective and questions are described in subsection 1.2, which is followed by defining the structure of thesis.

1.1 Background

In highly competitive business environment of digital era, organizations and the marketing units must understand their customers more precisely than before to remain competitive (Tsiptsis, 2009). The digitization of the world and the related constant increase in the amount of data and information sources are shaping the customer buying behaviour. Partly for this reason, organizations should tailor their marketing strategies to meet the various customer needs. Organizations that can utilize the data effectively in their marketing and align it with the business requirements, can achieve a competitive advantage in various business areas compared to other organizations in their industry. (Blanchard et al., 2019; Collica, 2011; Martínez et al., 2021; Rejeb et al., 2020.) This raises the demand for data science and analytics, that can be used for deriving information and knowledge from raw data.

Data science can be shortly described as a study of data, that combines statistical methods, informatics, analytics, and business understanding for creating value from data. Marketing and sales are typical processes that can benefit from utilizing data science and analytics along with the domain knowledge. (Cao, 2018.) In addition to the fact that the efficient usage of data in marketing is valuable for business, the marketing processes also produces a lot of data about sales and customers for the organization. When used correctly, data science can be utilized to refine data into valuable information and knowledge.

One way to improve the marketing process by using data science and analytics, is customer segmentation, which means dividing customers into homogenous groups by their characteristics (Baig et al., 2021; Sarkar et al., 2018). Customer segmentation can be practiced in both, consumer, and business markets, but this research is focusing on customer segmentation in consumer markets. The four most typical ways to practice consumer segmentation, are to divide consumers into segments by their behavioural, demographic, geographic, and psychographic characteristics (Kotler, 2019, p. 300; Sarkar et al., 2018). This study explores these various customer segmentation approaches and applies the behavioural segmentation in case research using client's sales data.

Behavioural segmentation is popular in retail sales industry, because it mostly utilizes the transactional data, which often has a good availability in retail sales organization's databases (Kotler, 2019, p. 310; Tsiptsis, 2009). One widely used behavioural segmentation method is RFM modelling, which is based on dividing consumers into groups by their recency, frequency, and monetary attributes (Sokol & Holý, 2021). The RFM modelling is considered as a base segmentation method in this Research.

1.2 Research objective and questions

The objective of this research is to design Proof-Of-Concept (PoC) -type of customer segmentation model and select appropriate data scientific framework during the iterative development process. The segmentation model should be suitable for historical consumer retail sales data. Used data is gathered from client organization's database and model will be tested against client organization business requirements. The solution is developed in collaboration with client organization, which is working in the retail sales industry. Goyat (2011) notes that each company requires different type of customer segmentation strategy. This is one key aspect that separates the research from existing segmentation studies and makes this research unique. In this regard the initial point of the study is to familiarize with the available data, business needs, existing scientific material, and recent research. The data availability and client's business needs set some requirements and the limitations of the research.

To find a solution for the research objective described above, the problem is presented in a form of questions. The main research question summarizes the problem into one question:

MRQ: What kind of segmentation model is appropriate for analysing quantitative sales data in retailer consumer markets?

The main research question covers the entire subject of thesis, including limitations and objectives. The question above is filtering the topic by stating that the model will be designed to Business-to-consumer (B2C) -type of business area, research will be using quantitative approaches and the methods are designed for transactional sales data. The research is trying to solve the consumer segmentation related business problem by designing and developing PoC -type of artifact. This research is using higher level term customer segmentation when referring to both, business and consumer markets, and the term consumer segmentation is used when referring to consumer markets only.

However, the main research question still covers a very broad area for research, so the question is sliced into several sub research questions. The first sub research question is focused on finding various consumer segmentation approaches from existing research:

SRQ1: What kind of consumer segmentation approaches have been identified in existing research and what are the strengths and weaknesses of those approaches?

The first sub research questions above provides detailed information about the different approaches that can been used when solving consumer segmentation related business problems. After answering the question, researcher can utilize the approaches that are possible and appropriate in this research. The first sub research question can also raise further development ideas that cannot be included in this research. The second sub research question is for finding the relevant data scientific framework to follow when designing and developing consumer segmentation model:

SRQ2: What are the most suitable data scientific frameworks for developing consumer segmentation model iteratively in collaboration with the client?

The framework should consider all the necessary phases that are required in segmentation model development when examining quantitative data. The framework should also be suitable for iterative development process. The third research question aims to find appropriate analytical tools and methods that can be used when in consumer segmentation:

SRQ3: What kind of analytical methods can be used in segmentation model when analysing consumer data?

Answering the question provides practical guidance of tools and methods that can be used for effective analysis on consumer data. The tools and methods should be applicable to consumer segmentation approaches found by answering the first sub research question. The fourth sub research question is for practical case study research:

SRQ4: How can the identified frameworks and methods be practically applied to a real-world case study of consumer segmentation?

This question gathers the approaches, frameworks and methods collected in the previous questions together and tries to answer how the information can be utilized in practice. The case research is used for answering the last question.

The end result of this research is a consumer segmentation solution that will provide explanatory and descriptive information about the consumers of the client organization.

Since the solution is PoC, it will open a lot of further development possibilities for client organization and scientific research.

1.3 Structure of thesis

The research starts by introduction chapter, which includes familiarizing with research background, introducing the research objective and questions, and structure of thesis. The introduction chapter if followed by theory section, which is divided into two parts.

The first theory section, chapter 2, deals with the consumer segmentation, starting with basics of customer segmentation. After that, various consumer segmentation approaches are explored. The last section of chapter 2 presents the objectives and identified benefits of consumer segmentation.

The second theory section in chapter 3 defines how data science can be applied to consumer segmentation. First, data scientific frameworks for development process are presented under the sub section 3.1, which is followed by explaining the various methods used in this research.

The fourth chapter is presenting the research methodological choices of this research. The research methodological choices are following the Saunders et al. (2019) onion model.

The fifth chapter includes the development process and analysis of the case study. The process is following the structure of design science research methodology (DSRM). The chapter is also reviewing the results and analysis for iteration.

The sixth chapter is for research discussion, which is followed by conclusions. Conclusions chapter is divided into summary, research evaluation, and research limitations and future research possibilities.

2. CONSUMER SEGMENTATION

This chapter is for theoretical background of customer segmentation. Chapter starts by defining the basics of customer segmentation in a B2C context. After the basics, various approaches and viewpoints to customer segmentation are presented. The objectives and the most typical benefits of customer segmentation are explained in the end of this chapter.

2.1 Customer segmentation

Modern data gathering methods and information systems enables companies to collect massive amounts of data from various business processes. Companies that can utilize the collected data may achieve huge business advantage compared to their competitors that does not utilize their data. One way to utilize the data is segmentation, in which aggregated data entities are segregated into multiple separate groups or segments that have similar characteries. This means that the groups are homogenous. (Baig et al., 2021; Sarkar et al., 2018.) As a simple example of segmentation could be a basket full of various fruits which can be grouped by their characteries, like colour, size, and sweetness. After segmentation, the contents of the fruit basket are easier to understand and manage.

Similarly, in customer segmentation process customers are divided into various homogeneous subgroups or segments based on the attributes and characteristics (Sarkar et al., 2018; Tsiptsis 2009). The customer attributes in segmentation could be for example demographical characteristics, geographical information, or purchase behaviour. These attributes can be used for various segmentation methods, which are described more detail in subsection 2.2.

Customer segmentation can be useful in many business processes, but it is best known as a marketing support tool. It can be used to differentiate customers into various homogeneous segments, which enables organizations to understand their customers better and build tailored marketing strategies for each customer group. (Tsiptsis 2009.) Customer segmentation is important for marketing operations, since all the customers and their behaviour are naturally different (Sarkar et al. 2018). This is why the customer segmentation should be considered when planning the marketing strategy. In addition, Zhou et al. (2021) mention that customer segmentation can be also described as market segmentation.

Customer segmentation can be utilized with both, consumer, and business customers, but there are inherent differences between these markets. Business customers are most likely making bigger transactions with longer and more complex sales process. Usually, there are also less business customers than consumer customers. Due to these reasons, same customer segmentation models should not be applied similarly to both of these markets. (Tsiptsis 2009.) This research is focusing on consumer side and all the methods are designed for the consumer markets.

2.2 Approaches to customer segmentation in consumer markets

Consumer segmentation can be approached from multiple different point of views. Tsiptsis (2009) mentions six different customer segmentation approaches for consumer markets, value based, behavioural, propensity based, loyalty based, demographic and need based segmentation. Similarly, Cooil et al. (2008) raises value based, behavioural, lifestyle, lifecycle, and activity-based segmentation. According to Sarkar et al. (2018) and Kotler (2019, p. 300), the four major segmentation approaches for consumer markets are behavioural, demographic, geographic, and psychographic segmentation. In psychographic segmentation consumers are grouped by creating psychographic profiles with psychology and demographics. These profiles could be created by using three variables: activities, interests, and opinions on consumer. (Kotler, 2019, p. 308.) The psychographic segmentation is not considered deeply in this paper, since the research is focusing more on behavioural, demographic, and geographic segmentation. These three factors and some of their typical attributes are defined in the table 1 below.

***Table 1:** Segmentation factors and attributes. (Modified from Cooil et al., 2008; Kotler, 2019; Sarkar et al., 2018; Tsiptsis, 2009)*

Behavioural segmentation	**Demographic segmentation**	**Geographic segmentation**
• Purchase amount • The location of the purchases/transactions • Transaction times/seasonality • Transaction types • Products related to transactions	• Gender • Age • Income • Education • Marital status • Family	• Country • City • Postal address • Geolocation • Climate

Required segmentation attributes mostly depends on the industry and business needs, but the table 1 above opens some of the typical situations. In behavioural segmentation, consumers are grouped by purchase behavioural and product or service usage patterns. It can be used for creating tailored consumer marketing and customized product offering strategies. This type of segmentation approach can often utilize transactional data. The availability of transactional data is good in most retail sales organizations, since it is usually stored. (Kotler, 2019, p. 310; Tsiptsis, 2009.) One popular behavioural segmentation method is to group consumers based on their recency, frequency, and monetary values (Sokol & Holý, 2021). This refers to the RFM method which is explained in subsection 3.2.2.

In demographic segmentation consumers are divided into groups based on demographic characteristics like gender, age, and income. It can be used to identify and understand needs of customer groups with specific characteristics. Tsiptsis (2009) mentions that demographic segmentation is appropriate in life-stage marketing and when promoting specific life-stage-based products. Some examples could be promoting baby products for new parents or offering luxury products for consumers with high-income level. Demographic attributes are often well associated with customer needs and wants and can be relatively easy to measure (Kotler, 2019, p. 302). Yet, although it is easy to measure, the availability of demographic consumer data is quite often limited or outdated in organization’s databases. But for some specific campaigns for example, demographic data can be collected by using market survey or similar methods. (Tsiptsis, 2009.)

Geographic segmentation can be used when dividing consumers by their geographical data like country, city, or postal address. This is useful in marketing because people living in various areas may have different needs. (Kotler, 2019, p. 300) For example, marketing warm winter coats to consumers in Nordic countries is more sensible than marketing them to consumers who live in warm climate countries. Kotler (2019, p. 300) mentions that there are also approaches that combines geographic and demographic data to enrich organization's consumer understanding. This approach is known as geo-demographic segmentation.

By using the behavioural, demographic, geographic or psychographic segmentation can result deeper customer understanding for the organizations that have not utilized consumer segmentation before. However, Goyat (2011) mentions that using only these traditional customer segmentation approaches is not enough in today's competitive marketing, and other factors like benefit sought and ethnocentric approach could also be considered. In addition, he suggests that the traditional approaches like behavioural and demographic approaches could be combined to get even more detailed customer segmentation results.

2.3 The objectives and benefits of consumer segmentation

Consumer segmentation has become an essential practice for companies seeking to create better customer experiences and increase sales. Goyat (2011) mentions that companies who segment their customers, can match their strengths and offerings with the customer groups that are most likely to respond them.

The objectives and benefits of consumer segmentation are diverse and can greatly impact organizations' success in the market. Some of the main objectives of consumer segmentation are to gain deeper understanding of organization's customer base, to improve marketing by targeting specific customer segments, optimize product placement, offering and design to better serve customers' needs, and to identify new business opportunities. (Goyat, 2011; Sarkar et al., 2018; Tsiptsis, 2009) When these objectives are realized, they can also be seen as direct benefits for the organization.

By dividing a customer base into distinct groups based on specific characteristics, organizations can achieve several benefits. In addition to benefits mentioned in earlier section, successful consumer segmentation may also lead to finding latent customer segments, gaining competitive advantage in certain customer groups, and finally earning increased revenue (Goyat, 2011; Sarkar et al., 2018; Tsiptsis, 2009). These identified benefits are described in the table 2 below.

***Table 2:** Benefits of consumer segmentation.*

Benefit	Description
Deeper customer understanding.	Understanding of an organization's consumers, their needs, behaviour, and various characteristics (Sarkar et al., 2018; Tsitsipis, 2009).
Targeted marketing.	Ability to focus marketing efforts more effectively and efficiently by identifying different segments of customer base (Sarkar et al., 2018).
Optimal product placement and design.	Ability to offer and design new products or optimize the placement of existing ones to better serve each segment's needs (Sarkar et al., 2018; Tsitsipis, 2009).
Finding latent customer segments.	Ability to identify new customer segments and opportunities that might be missing (Sarkar et al., 2018).
Competitive advantage.	Ability to concentrate marketing energy and force on subdividing to gain a competitive advantage within the segment (Goyat, 2011).
Higher revenue.	Ability to provide more personalized and targeted experiences for customers, which can create stronger customer loyalty and lead to increased sales and revenue (Sarkar et al., 2018).

Successful customer segmentation can provide a deeper understanding of an organization's consumers, their needs, behaviour, and various characteristics. By understanding consumer needs, preferences, and buying habits, companies can tailor and target their marketing campaigns more effectively and efficiently for each identified customer segment. This may also help organizations to find latent customer segments and opportunities they might be missing. (Goyat, 2011; Sarkar, 2018; Tsiptsis, 2009.) For example clothing retailer that uses consumer segmentation might find out that there is clear segment of customers who prefer trendy and fashionable clothing, and another customer segment that prefers classic and timeless styles. Retailer could use this information when creating tailored marketing campaigns with specific emphasis for each identified segment.

Organizations with detailed customer segments can also develop new products, product offerings, and services to better serve their customers in certain segments. Additionally, customer segmentation can help organizations to optimize product placement and design, which can lead to better customer experience and increased customer loyalty (Goyat, 2011; Sarkar, 2018; Tsiptsis, 2009.) For example grocery store in the city central has identified customer segment of busy businesspeople arriving in lunch time and looking for a healthy premade lunch option. Grocery store could stock readymade healthy meal packages near the store entrance.

These benefits of successful consumer segmentation can lead organization to increased competitive advantage in certain customer groups, by differentiating their offerings from their competitors in the same market area. Finally, successful customer segmentation can be seen as increased sales and revenue. (Goyat, 2011; Sarkar, 2018; Tsiptsis, 2009.)

3. DATA SCIENCE IN CONSUMER SEGMENTATION

The previous chapter included the basics of customer segmentation in B2C business. This chapter is focusing on how the data science can be utilized to meet the objectives and get the benefits presented in previous chapter. Subsection 3.1 presents data scientific approaches that can be used when designing and developing the segmentation model. The methods that can be used in model development are presented in subsection 3.2.

3.1 Data scientific approaches and selection

When designing data-oriented solutions for business problems, like consumer segmentation, it is recommended to use some kind of process model that considers all the necessary steps, at least at a higher level. Sarkar et al. (2018) mentions that in data science and analytics, it is important to match technical aspects to business value. There are multiple useful process models that can be used as a pillar when designing data-oriented solutions for business problems. Some well-known and widely used models in a field of data science are Knowledge Discovery in Databases (KDD) and Cross-Industry Standard Process for Data Mining (CRISP-DM). There are also more modern tools like some lean and agile methods. The KDD and CRISP-DM models are presented in subsections below.

3.1.1 Knowledge Discovery in Databases

The traditional KDD process model was introduced by the Fayyad et al. (1996). The model consists of five steps, which are data selection, data pre-processing, data transformation, data mining and evaluation. The steps are performed to transform data from one stage to another. The states of data in model are source data, target data, preprocessed data, transformed data, patterns and knowledge. (Collier et al., 1998; el Sheikh & Alnoukari, 2012; Fayyad et al., 1996.) The traditional KDD process model is presented in figure 1 below.

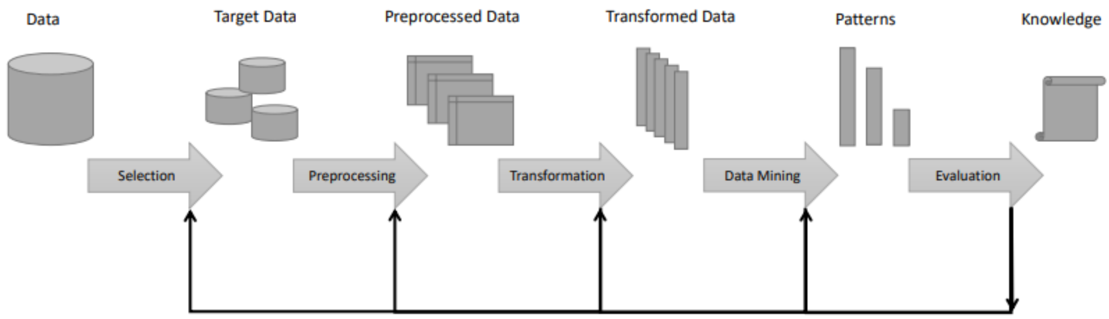


Figure 1: *Traditional KDD Process Model (Collier et al., 1998; El Sheikh & Al-noukari, 2012; Fayyad et al., 1996).*

As the figure 1 shows, the process starts from source data, which is processed into information or knowledge in multiple steps. The data selection step is for extracting only the relevant data batches from the larger data source. After the relevant data is extracted, the data is preprocessed. Pre-processing consists of data preparation and data cleansing, like missing and noisy data detection. These operations ensures that the data is correct for the analysis.

The KDD process model is good tool for defining the data science development process from data point of view, but it lacks with the business perspective and ignores the involvement of human resources. However, the process model can be considered as a cornerstone for the later and more advanced KDD models, such as SEMMA and CRISP-DM. (el Sheikh & Alnoukari, 2012.) Although there are several KDD related process models that could be successfully utilized in customer segmentation model development, this research is focused on CRISP-DM process model, which is presented in following sub section.

3.1.2 CRISP-DM

CRISP-DM differs a little bit from the KDD process model, and it is approaching data related problems more over from the business point of view (Vleugel et al., 2010). CRISP-DM can be seen as elaborated and extended version of original KDD process (Martinez-Plumed et al., 2021). According to el Sheikh & Alnoukari (2012), CRISP-DM was the first KDD related process model which considered steps of business understanding and data understanding. It was found at the turn of the 21st century, but it is still the most widely used and the de-facto standard process for data mining and data science (Martinez-Plumed et al., 2021; Schröer et al., 2021). One reason that explains the popularity of CRISP-DM is because it is an industry-independent process, that can be applied for multiple purposes and is not restricted by any technological choices (Schröer et al., 2021). The process model is iterative and includes six steps, which are connected to each other with arrows. CRISP-DM model is usually presented as a circular figure with the data in the middle.

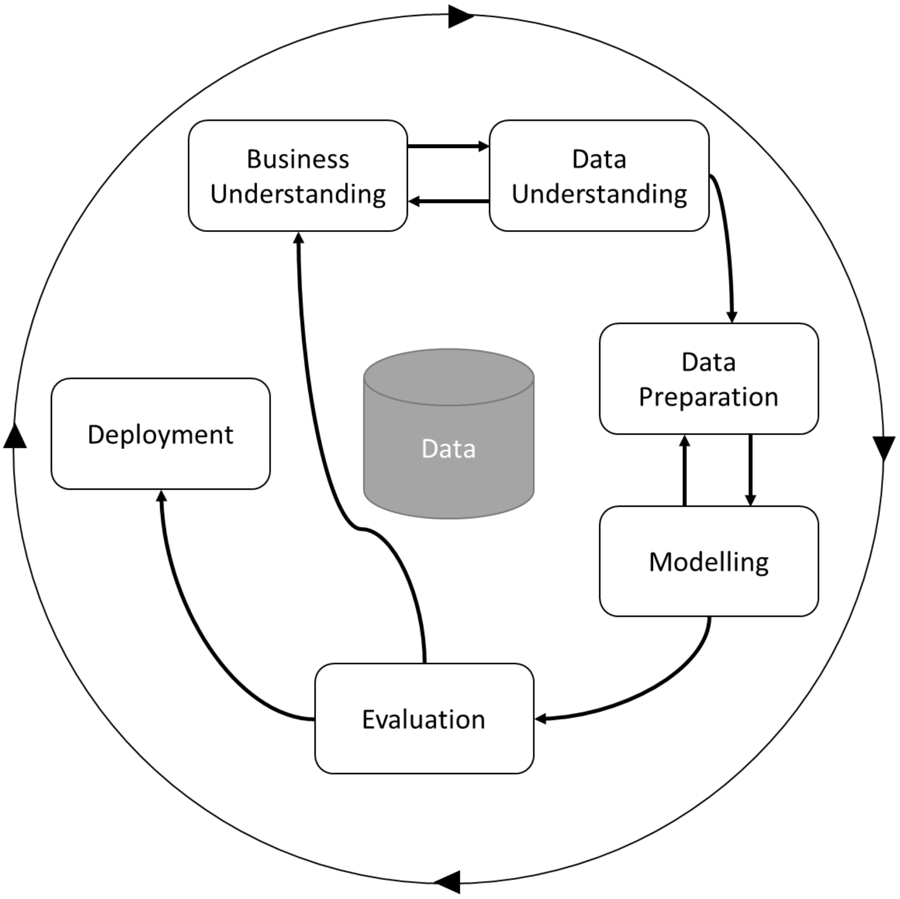


Figure 2: *CRISP-DM Process Model (El Sheikh & Alnoukari, 2012; Martinez-Plumed et al., 2021; Schröer et al., 2021).*

CRISP-DM process model cycle is presented in the figure 2 above. The six high level phases of the model are business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Each phase can include multiple deeper level processes, which makes the cycle more like a framework (El Sheikh & Alnoukari, 2012; Sarkar et al., 2018). The outer ring of the cycle describes the direction of the flow, but there are also arrows between each step.

The initial phase of CRISP-DM is business understanding which regards the business requirements and objectives of the project. Understanding the business requirements and objectives is essential when designing the data science project plan and defining the problem. (el Sheikh & Alnoukari, 2012; Schröer et al., 2021.) In addition, the phase of business understanding can be divided into four more detailed activities: Determination of business objectives, assessment of the situation, determination of goals and generation of a project plan (El Sheikh & Alnoukari, 2012).

The second phase is data understanding. Like in the case of business understanding, it is important to understand what kind of data available and what kind of data is relevant for the research. Data understanding is closely linked to business understanding because these activities support and delimits each other. In concrete level data understanding means collecting relevant data from various data sources, checking, investigating, and describing data and the quality of it (Schröer et al., 2021). Similarly, according to the el Sheikh & Alnoukari (2012) research, more detailed activities for data understanding are collection of initial data, description of data, exploration of data, and verification of data quality.

The third phase is data preparation, in which focus is on preparing the data for analysis. This include selecting the data that is relevant for the segmentation project, data cleaning, data construction, data integration, and transforming the data into correct format for analysis. Data preparation is a step that covers all the necessary activities required to build the final dataset for the modelling phase. It is also depended on the select model in first two phases. (el Sheikh & Alnoukari, 2012; Schröer et al., 2021.) Data preparation is a critical step in the process, as the quality of the data will directly impact the quality of the segmentation results.

The fourth phase is modelling. The focus of this phase is on designing appropriate modelling technique, creating the test case, developing the models and assessment of the generated models (El Sheikh & Alnoukari, 2012; Schröer et al., 2021). Appropriate modelling techniques in consumer segmentation can be such as clustering, dimensional reduction, or RFM -modelling presented in subsection 3.2.2. As in the previous phases, the selection of the modelling techniques also depends on the choices of the business model (Schröer et al., 2021). In addition, according to el Sheikh & Alnoukari (2012), reiteration into the previous phase can be often required, because different modelling techniques may require specific data formats. In a case of consumer segmentation, the goal of this phase is to find model that can be used identify distinct customer segments based on various characteristics like behaviour or demographics.

In the fifth phase, evaluation, the focus is on evaluating the results of the model against the earlier defined business objectives, assessing the accuracy, and reviewing the results of the developed model. At this phase, it is also important to determinate the objectives of the next iteration or further actions. (el Sheikh & Alnoukari, 2012; Schröer et al., 2021.) If the model is found to be accurate and reliable in evaluation, the process can move towards the deployment. Otherwise, the business understanding phase of the next iteration will be refined.

The final phase is deployment. The focus of this phase is on deploying the model, like segmentation model, into the organization's business processes. This involves deployment planning, monitoring, and maintenance planning, building integrations to other systems, creation of final report and user guide, and other steps required in product deployment. (el Sheikh & Alnoukari, 2012; Schröer et al., 2021.)

3.2 Methods for customer segmentation analysis

Recent studies have recognized multiple various tools and methods for different customer segmentation approaches. This research is focused on behavioral segmentation and the considered modelling methods are RFM modelling and K-means clustering. Also, some of the popular data preprocessing and feature engineering techniques are presented in this subchapter.

3.2.1 Data processing and feature engineering

Data processing and feature engineering are crucial steps in data driven solutions like customer segmentation model. It is important to have data in correct form before inputting it to various data modelling algorithms, but these processes are also hard and time consuming to implement. According to Sarkar et al. (2018), data processing and feature engineering are often some of the toughest tasks to implement when building machine learning solutions. They also suggest that data scientists spend about 70-80 percent of their time on data processing, wrangling, and feature engineering, when implementing machine learning models (Sarkar et al., 2018). This underlines the importance of data processing and feature engineering.

Feature engineering can be defined as a process of taking data, cleaning it, preparing the data for analysis, and transforming it for use in models (Blanchard et al., 2019; Nguyen, 2021). The concept feature can be described as a numeric representation of source data that can be used in various calculations (Sarkar et al., 2018; Zheng & Casari, 2018). The following figure 3 represents data preparation pipeline for data preprocessing and feature engineering.

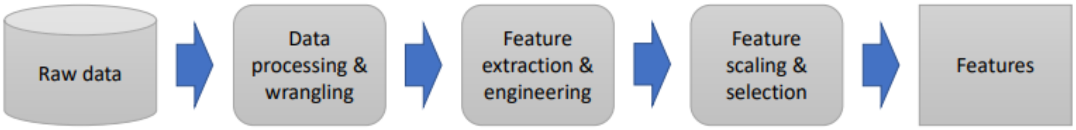

Raw data
Data processing & wrangling
Feature extraction & engineering
Feature scaling & selection
Features

Figure 3: *Data preparation pipeline for data preprocessing and feature engineering. (Modified from Sarkar et al. (2018))*

The data preparation pipeline starts by collecting the raw data. After the data is collected, it is processed and wrangled to correct form. The next step is feature extraction and engineering, which is followed by feature scaling and selection. Finally, the input of the pipeline is features that can be used as a measurable attribute in statistical analysis and machine learning. (Sarkar et al., 2018; Zheng & Casari, 2018) All of these above steps are not always necessary, but the data preparation requirements depend on the algorithms used.

Most machine learning models and algorithms cannot be built by using raw data, because in those data is required to be processed and wrangled into specific format before they can work with it effectively. Data wrangling can be defined as a process of cleaning, transforming, and mapping raw data to utilizable form. (Sarkar et al., 2018) The definitions and steps of data wrangling and feature engineering are overlapping a bit, but in this research data wrangling is considered as a process of data cleaning and transforming, including steps like data filtering, handling missing values and duplicates, outlier detection, changing data types, data merging and other relevant steps to improve data usability.

The next step in data preparation pipeline is feature extraction and feature engineering, which includes manually made techniques or mathematical transformations to improve the performance of the model and data representation. Feature engineering is also typically emphasizing the business and domain knowledge. (Sarkar et al., 2018) Since this study is focusing on the quantitative data analysis, statistical transformations like logarithmic transformations, reciprocal transformations or square root transformations are important for getting better results. Reciprocal transformation is often used with ratios that are obtained by dividing two variables, like population density. Logarithmic and square root transformations are effective methods when data is highly skewed on the right side of graph. Generally, logarithmic transformation is more effective for highly right skewed data, because it compresses the range of large values more effectively, but the advantage of square root transformation is that it can also be used for zero values. (Galli,

2022) This research is preferring log transformations, since zero values are filtered earlier in the process. Function for logarithmic transformation is presented below.

$$y = \log_b(X), \tag{1}$$

The function represents the logarithm of X to the base b, which is equal to y. Logarithmic transformations are often applied to skewed distributions, because they can help to normalize the data by expanding the values in the lower magnitude range and compressing the values in the higher magnitude range. This process can make the distribution more like a normal distribution. (Galli, 2022; Sarkar et al., 2018; Zheng & Casari, 2018.)

The last step in data preparation pipeline is feature scaling and selection. Data scaling is important especially for machine learning because many algorithms are sensitive to the scale the features have. For example, in clustering, features with large value range can influence the results much more efficiently than the features with smaller value range. (Galli, 2022.) If clustering algorithm uses customer age and sales for segmentation, and the value range for age is from 18 to 80 and value range for sales is from 0 to 10 000, the results may be distorted. This research is considering two different scaling methods, Min-Max scaling, and standardization. Function for Min-Max scaling is presented below.

$$MMS(X_i) = \frac{x_i - x_{Min}}{x_{Max} - x_{Min}}, \tag{2}$$

When using Min-Max scaling, the features are scaled in range 0 to 1. In formula 2, each value in feature X is scaled by subtracting it from the minimum value and dividing the resultant by the difference between the maximum and minimum values of it. (Sarkar et al., 2018.) Another scaling method is standardizing, which is presented below.

$$SS(X_i) = \frac{x_i - \mu_X}{\sigma_X}, \tag{3}$$

In standardizing, features are scaled to have mean close to 0 and variance close to 1. Each value in X is subtracted by the mean and the result is divided by the standard deviation. (Sarkar et al., 2018.)

Lastly, the feature selection is the process of selecting a subset of the most relevant and informative features from a larger set of features in a dataset. The goal of feature selection is to improve the performance and efficiency of machine learning models by reducing the dimensionality of the data. There are few major techniques for feature selection, including filter methods, wrapper methods, and embedded methods. Filter methods use statistical measures such as correlation or mutual information to rank features based on their relevance to the target variable. Wrapper methods use the performance of a machine learning model with different subsets of features to evaluate their importance. These methods are very effective, but computationally expensive. Embedded methods incorporate feature selection directly into the training of the machine learning model. (Sarkar et al., 2018; Zheng & Casari, 2018.)

Overall, data processing and feature engineering are important steps in any data scientific model, including customer segmentation. These steps help in identifying patterns and relationships between customer attributes and is important for getting accurate segmentation results. According to Sarkar et al. (2018), data processing and feature engineering should be used as an iterative process, as the whole segmentation process. When considering the iterative CRISP-DM process model presented in earlier chapters, data preparation step is in between data understanding and modelling.

3.2.2 RFM Model

One of the most common customer segmentation methods for analysing transactional data is Recency, Frequency and Monetary (RFM) model. It is a statistical method used for determining a customer's value based on their buying behaviour. The basic RFM model uses transaction data to calculate three important and informational attributes for each individual customer. (Christy et al., 2021; Sarkar et al., 2018; Sarvari et al., 2016.) According to Wu et al. (2020), the model can be used for customer segmentation, customer behaviour analysis, and measuring customer lifetime value. The three attributes, recency, frequency and monetary are presented in figure 4 below.

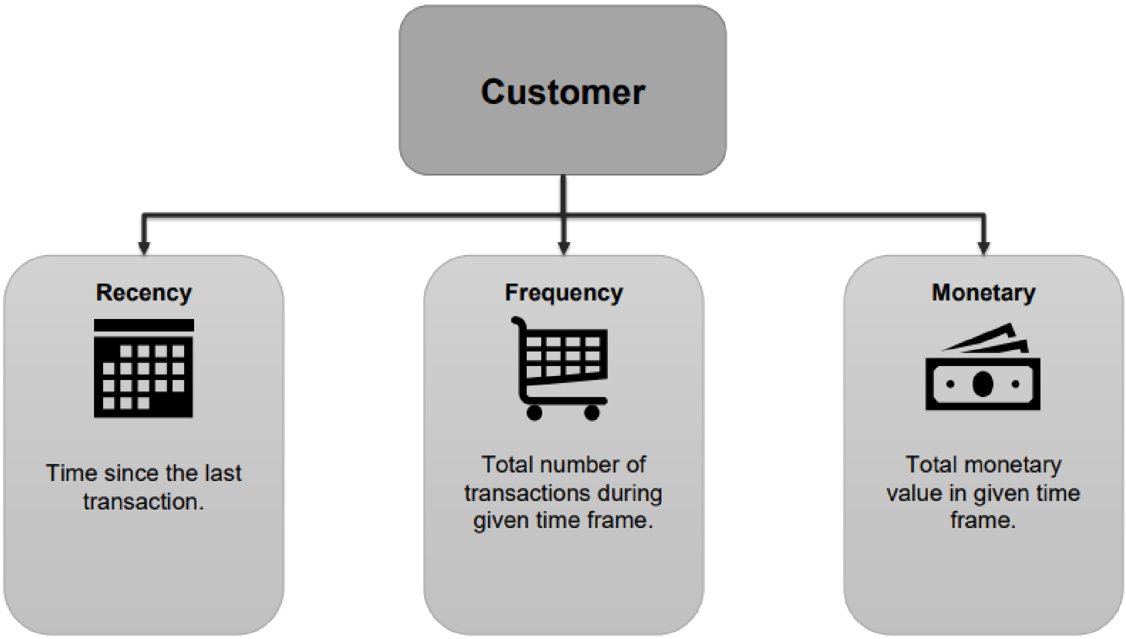

Customer
Recency
Time since the last transaction.
Frequency
Total number of transactions during given time frame.
Monetary
Total monetary value in given time frame.

Figure 4: *The basic RFM Model.*

Recency value in RFM model represents on how recently individual customer has made a transaction. The value can be calculated by taking time interval between the customer's last purchase date and the statistical time period end, which is more often the latest transaction among the transaction of whole dataset. (Sarkar et al., 2018; Wu et al., 2020.) Knowing the recency of customer can be important information for the sales organization because more recent customers can be considered as more valuable than those who have not used services or purchased anything for a longer period. Sokol & Holý (2021) points out that, for example, a retailer can use different marketing techniques when attracting customers who have visited the store past week, versus those customers who have not visited for months.

Another attribute in RFM model, frequency, represents the number of transactions customer has made during the statistical period. The value of frequency can be calculated by taking count of distinct purchases made in given time frame by individual customer. (Sarkar et al., 2018.) According to Wu et al. (2020), customers with higher frequency are also more likely to be loyal customers with the stronger intention to purchase or use company's services again. Loyal customers are most likely more valuable than those who visits rarely. Sokol & Holý (2021) adds that frequency is one of the most measured key performance indicators (KPI) along with the average value of a shopping basket.

Third attribute in RFM model is monetary. It represents the monetary value of all the transactions that individual customer has made during the statistical period. (Sarkar et al., 2018; Wu et al., 2020.) Often, the monetary value is calculated by summing all the transactions made in given time frame, but the measuring can also be done alternative ways like calculating average monetary value of transactions in time interval (Sokol & Holý, 2021).

Referring to Sokol & Holý (2021), the formulas for calculating recency, frequency and monetary values from transactional data can be written follows:

$$R(C_i) = \min\{Days\ since\ a\ purchase\ by\ customer\ i\}, \tag{4}$$

$$F(C_i) = \frac{Number\ of\ baskets\ purchased\ by\ customer\ i}{Time\ frame}, \tag{5}$$

$$M(C_i) = \frac{Total\ value\ of\ products\ purchased\ by\ customer\ i}{Time\ frame}. \tag{6}$$

In formulas 4, 5 and 6, *R(C$_i$)*, *F(Ci)* and *M(Ci)* represents the recency, frequency, and monetary values for customer *i*. After collecting recency, frequency, and monetary values from transactional data, researcher can give RFM scores to customers by dividing customer into categories. This means that each RFM attribute is divided into appropriate number of categories. According to Stromi et al. (2020), suitable number of categories is typically somewhere between 3 and 5. For example, if customers monetary values are in range 10-900 €, researcher could use limit values 600 € and 300 € for categorizing. If researcher is using 3 categories, customers that has spent more than 600 €, would be in category 3 and customers that has spent less than 300 € would be in category 1. After the same operation is done to each of three attributes, $3^3 = 27$ different customer segments are identified and the most valuable customers can be found from segment, where all the RFM values are 3. These values can be represented in a form of matrix which can be used for identifying the most valuable customer and the one that needs more attention (Stormi et al., 2020.)

Dividing customer into categories using predefined limit values is easy to implement technique for quick results. Another RFM scoring method is to divide customer attribute values into equal groups or quantiles. This can be done by sorting the customers by one attribute at time and dividing them into equal groups. (Miglautsch, 2000.) For example, if the study is using 4 categories, the researcher can start by sorting the data in descending order based on recency values, and giving the first quarter four points for recency, the second quarter three, and so on. Similarly, the researcher can sort customers in ascending order by frequency and monetary values and follow the same scoring steps. After scoring, $4^3 = 64$ various segments have been obtained, and the best of which is [R=4, F=4, M=4].

Sarvari et al. (2016) also suggests that calculated each of the RFM values should be rescaled. For rescaling R, F, and M values, they suggest equation called normalization, which can be presented as follows:

$$X(C_i) = \frac{x_i - x_{Min}}{x_{Max} - x_{Min}}, \tag{7}$$

where *X* describes R, F, or M values, and *C_i* presents individual customer. Normalization is used to scale the data to a standard scale or range (Sarkar et al., 2018). The normalization formula above scales values to range between 0 and 1. Rescaling is important step especially for various machine learning algorithms, because they assume that the input data is standardized. Unscaled data might affect negatively to the calculations and distort the results. (Sarkar, 2018.)

Although the RFM model is excellent tool for behavioural customer segmentation, it also has some disadvantages. The basic RFM model includes only three factors, recency, frequency and monetary, and does not consider other attributes like customer demographics or geographics. According to Sarvari et al. (2016), these attributes can be used for describing and explaining customers buying behaviour but are not enough for predictive analytics like exploratory and forecasting analysis. They also point out that the

RFM model does not consider the impact of customer life stages or life cycle transitions. In addition, Zhou et al. (2021) states that lacks information on customer loyalty related to long-term dynamic changes that result from a customer's shopping activity. Lastly, while RFM model is mainly used for finding various customer segments, the users may utilize only the most attractive ones when targeting markets. This can result in several profitable segments being missed. (Sarvari et al., 2016.)

Some of the issues above can be solved by modifying the RFM model by adding new attributes, weighting variables, or combining other methods to it. Some studies, like Heldt et al. (2021) and Martínez et al. (2021) have combined product information to RFM model to get more detailed about the products various customers are buying. Heldt et al. (2021) names to RFM model with product information as RFM/P model. Peker et al. (2017) enhanced RFM model into LRFMP model where (L) stands for customer relation length and (P) describes periodicity. The (L) in can be calculated by taking time interval between customer's first and last transaction, and the (P) is the standard deviation of customer's inter-purchase times, which means time between two consecutive purchases. Zhou et al. (2021) suggest using similar inter-purchase time (T) in their RFMT model. There are also many more expanded RFM models to various purposes for designing more detailed and informal customer segmentation using transactional data.

The results of the RFM analysis can be further refined by emphasizing certain attributes. Sometimes organization's business department may for example highlight the importance of frequency attribute in customer segmentation. There are also some tools and algorithms that can be used in weighting RFM attributes. One method that can be used when determining the of RFM weights is a Principal Component Analysis (PCA) method (Wu et al., 2020). Another method, Analytical Hierarchical Process (AHP) helped Martínez et al. (2021) to calculate the different RFM score weights by product category.

Overall, the RFM method is useful tool in multiple database marketing and customer segmentation situations, especially when there is a good availability to transactional sales data (Christy et al., 2021; Sarkar et al., 2018; Sarvari et al., 2016; Wu et al., 2020). Researcher or user of the segmentation model can use the matrix of RFM attributes in decision making or divide customers into various segments manually by defining the characteristics of segment. For example, segment "Top customers" would require high RFM score for each attribute and segment "Loyal customers" would require high frequency and recency scores. Another widely used method for creating customer segments is to use clustering methods, like K-means with RFM modelling. (Christy et al., 2021; Sarkar et al., 2018; Sarvari et al., 2016; Sokol & Holý, 2021.)

3.2.3 Unsupervised learning and K-Means Clustering

Unsupervised learning is a machine learning technique that can group data points that has similar characteristics without knowing these groups beforehand in contrast to supervised learning. This is why the unsupervised learning is excellent method for customer segmentation too (Baig et al., 2021). One of the well-known unsupervised learning techniques is clustering, which is the process of dividing various objects into homogenous groups with similar attributes (Wu et al., 2020). Arunachalam & Kumar (2018) presents that clustering can be used for both, exploratory and descriptive analytics.

One of the simplest and most popular clustering algorithms is K-means (Sarkar et al., 2018; Wu et al., 2020). It is a standard algorithm that takes number of clusters and parameters as input and partitions the dataset into previously defined number of groups or clusters. The algorithm is trying to form the cluster to be as homogenous as possible. (Christy et al., 2021; Tabianan et al., 2022.) The K-means algorithm work iteratively by computing the values of cluster centroids before each iteration round. In each iteration, after the values of the cluster centroids are calculated, each data point is assigned to the closest cluster centroid. The process is repeated until the location of cluster centroids does not change or the maximum number of iterations has been reached. (Christy et al., 2021; Sarkar et al., 2018.) The four basic steps of K-means algorithm are presented below:

1. The K number of cluster centroids are randomly initialized.
2. Each data point in dataset is assigned to the closest cluster centroid. K-means clustering uses normal Euclidian distance as a distance metric between data point and cluster centroid.
3. Recalculating the cluster centroid values by calculating the average distance of each data point belonging to the cluster.
4. The process is repeated until the location of cluster centroids does not change and the assignment becomes stable, or the maximum number of iterations has been reached.

The ”K” in K-means clustering represents the number of clusters and centroids in model. The optimal number of clusters can be found by utilizing the widely used elbow method. The idea of elbow method is to calculate sum of square distances between the cluster centroid and each data point in cluster. This metric can be called as sum of squared errors (SSE) or inertia, and it is presented in formula 8. SSE/inertia is calculated for different number of clusters and the results are generally plotted as two-dimensional graph.

The graph usually forms an elbow type of shape, where the SSE/inertia rapidly decreases till the certain point of “K”, after which the decreasing slows down. (Baig et al., 2021; Tabianan et al., 2022; Wu et al., 2020.)

$$SSE = \sum_k \sum_{x_i \in k} (x_i - C_k)^2. \tag{8}$$

In the formula 8 above, C_k describes the location value of the cluster *k* centroid and x_i represents the data point that is assigned to cluster *k* (Baig et al., 2021; Wu et al., 2020). Sometimes, when determining the optimal number of clusters by using the elbow method, the “elbow” can be unclear, and the optimal number of clusters is hard to find. In that case, the researcher should consider additional metrics to determine the optimal number of clusters. Christy et al. (2021) and Sharaf Addin et al. (2022) are using silhouette score in their study, which is used for calculating similarity of data point in its own distance compared to the other clusters. The silhouette score for each data point *i* can be calculated as follows:

$$S(i) = \frac{b(i) - a(i)}{\max\,(a(i), b(i))}. \tag{9}$$

The variable *a(i)* in formula 9 above represents the intra-cluster distance, which means the average distance from *i* to every other data point within the same cluster. Respectively, the variable *b(i)* represents the inter-cluster distance, which means the average distance to closest neighbouring cluster, that *i* is not part of. Output values of silhouette score varies in range from -1 to 1. The overall silhouette score can be calculated as a mean value of each *S(i)* and higher silhouette score means better formed clusters (Christy et al., 2021; Haroon, 2017; Sharaf Addin et al., 2022.)

According to Wu et al. (2020) study, customer segmentation is one of the major applications of K-means clustering, because it can be very effective when identifying customer value and developing relevant marketing strategies. K-means clustering is also widely used together with RFM modelling presented in subsection 3.2.2. For example, (Christy et al., 2021; Sarvari et al., 2016; Sharaf Addin et al., 2022; Wu et al., 2020) are using K-means clustering together with RFM modelling in their studies.

Even though K-means algorithm was first published in 1955, over 65 years ago, it is still dominating the marketing industry and is one of the most popular clustering algorithms (Arunachalam & Kumar, 2018; Jain, 2010). Baig et al. (2021) suggests that one of the reasons why K-means is still so effective tool is because it is conceptually relatively simple, it scales easily to very large datasets, and it works well in practice. (Arunachalam & Kumar, 2018) adds that most of the marketing practitioners are still relying on relatively simple clustering algorithms in market segmentation, like K-means.

There are also some disadvantages when using clustering and K-means. Since clustering creates customer groups without prior knowledge of data, these groups may be difficult to interpret and the way that clusters are found may be unclear without more detailed research (Baig et al., 2021).

4. RESEARCH METHODOLOGY

This chapter discusses the methodological choices of the research. The research methodological choices are presented with Saunders et al. (2019) onion model, which is planned for the business-related research. The following figure 5 is modified from Saunders et al. (2019) onion model and it contains five methodological choices relevant to this research.

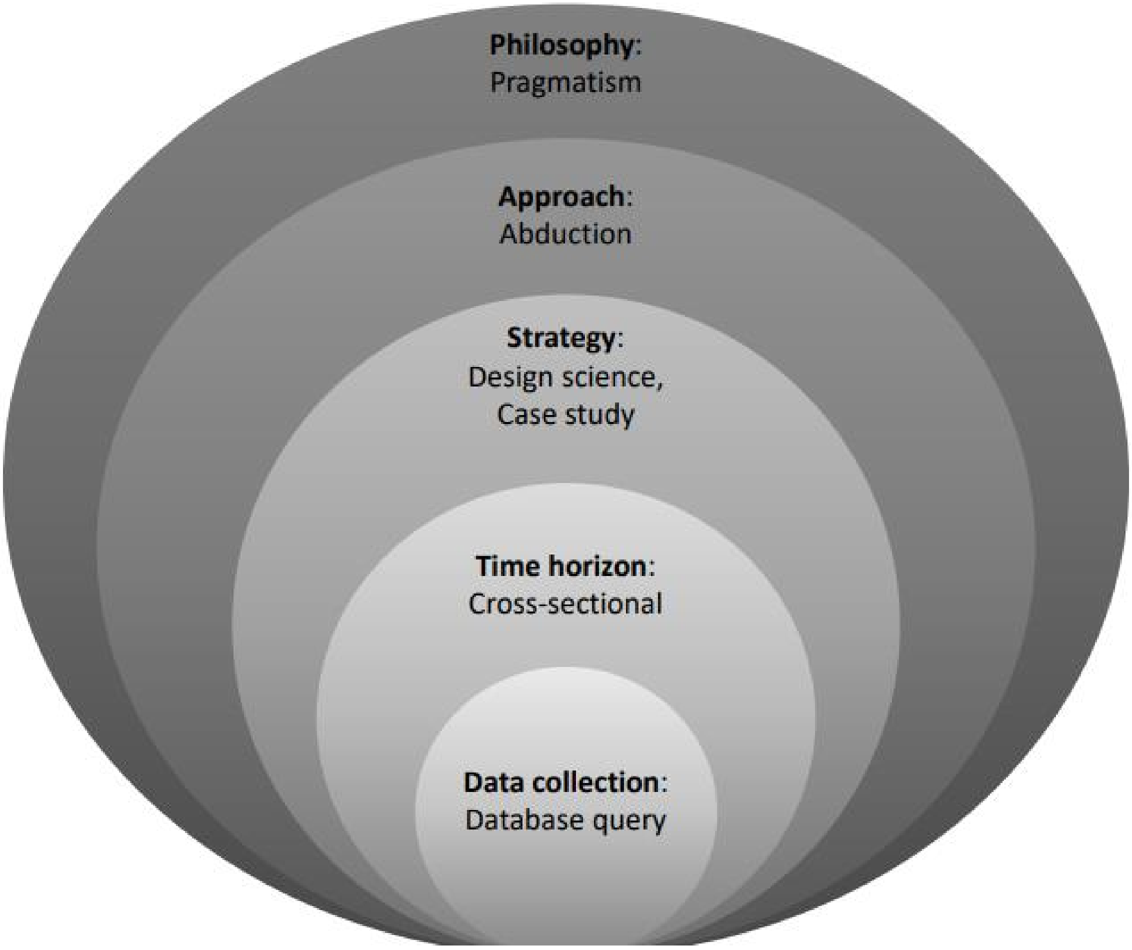


Figure 5: *Research methodological choices. (Modified from Saunders et al. (2019))*

Research methodological approaches used in this research are presented in a figure 5 above. These choices are explained more deeply in following subsections.

4.1 Research Philosophy

The outermost layer of the research onion in figure 5 presents the research philosophy. Saunders et al. (2019, p. 130) describes research philosophy as a system of beliefs and assumptions about the development of knowledge. Those belief and assumptions shape the researcher thinking thorough the research process and that is why research philosophy should be one of the first steps when planning the research. (Saunders et al. 2019, p. 130.)

Saunders et al. (2019) identify five main research philosophies in their book: positivism, critical realism, interpretivism, post-modernism, and pragmatism. The objective of this research is to create Proof-Of-Concept -type of consumer segmentation solution, which will solve case organization's real-life business problem. An appropriate research philosophy for this research is pragmatism because it emphasizes the practical application of knowledge and the usefulness of the research outcomes. Pragmatist researcher believes that knowledge is valuable only if it can be used to solve practical problems, like dividing organization's consumers into segments. (Saunders et al. 2019, p. 151.)

Other suitable research philosophy option for this research would be positivism, which emphasizes on finding objective facts about the world. Positivists believe that these facts can be discovered through accurate observation and measurement. (Saunders et al. 2019, p. 144.) The positivism would be appropriate research philosophy option if the research were mainly interested in identifying repeating patterns in consumer behaviours and other characteristics.

4.2 Approach to theory development

The second layer of Saunders et al. (2019) onion model is approach to theory development, which defines how the theory is used in research. Usually, there is two main approach options, deductive and inductive. Study with deductive approach starts from the general theory or hypothesis and the theory is tested against empirical study, observations, or data collection. On the contrary, inductive approach starts with observations or data collection, and the new theory or hypothesis are generated by utilizing the data. (Saunders et al. 2019, pp 152-156.) Pure deductive or inductive approach might be difficult to follow in this research because study is going back and forth between data and theory.

The appropriate research approach for this study is abduction, which can be described as a mix of deductive and inductive approaches. In abduction, data is used for exploring the phenomenon of the subject and identifying themes and patterns. New or modified existing theory is created and tested with data collection. (Saunders et al. 2019, pp 152-156.) In addition, Saunders et al. (2019, p 156) mentions that abductive approach is often used with pragmatistic research philosophy.

4.3 Research strategy

The next layer of the research onion is research strategy. Saunders et al. (2019, p. 189) describes research strategy as a methodological link between selected research philosophy and data analysis method choices. Overall, the research strategy is a well-defined plan of how the research is going to meet the requirements set by the research problem and how the research goals will be achieved. (Saunders et al. 2019, p. 189.)

The research strategies Saunders et al. (2019, p. 190) discuss are experiment, survey, archival and documentary research, case study, ethnography, action research, grounded theory, and narrative inquiry. These strategies can be associated with research philosophy and approach to theory development, but more often, the boundaries between these methodological choices are open. It is also possible to use multiple research strategies together because the choices are not mutually exclusive. Most importantly, the choice of research strategies should be in line with the research objectives and questions. (Saunders et al., 2019, pp. 189-190.)

The appropriate research strategy of this study is design science along with the case study. Design science can be seen as a problem-solving paradigm that aims to create new knowledge by producing practical solutions, like customer segmentation model. In design science, these practical solutions are often called as artifacts, which can vary between constructs, models, methods, and instantiations. The business needs of design science research should be relevant to real-world context, like case organization with a specific business problem. The rigor in form of applicable knowledge, quality, and validity of the artifact in design science research are achieved by using the existing literature, foundations, and methodologies. (Hevner et al., 2004.) The objectives and limitations of this research are also defined by the business environment, available data, and client organizations together with the researcher. The rigor of this research is achieved by applying methods and methodologies from systematic literature review.

The design science research is iterative process that consists of six phases. The first phase is problem identification and motivation. The second phase is for defining the objectives and research questions of the study. The third phase consists of the design and development of the artifact to achieve the earlier defined research objectives. In the fourth phase the earlier created artifact prototype is demonstrated against the business requirements. The fifth phase is for evaluation of the artifact effectiveness which can be followed by production or iteration back to design and development phase. The final phase is communication, where the research outcomes are communicated to relevant stakeholders, clients, and potential users. (Peffers et al., 2007.) In this research, the first two phases are covered in chapter 1, but also partly in chapter 5. The third, fourth and fifth phase are covered in case study chapter 5, and the communication is maintained throughout the iterative research process. Six phases of the design science research methodology process model are presented in figure 6 below.

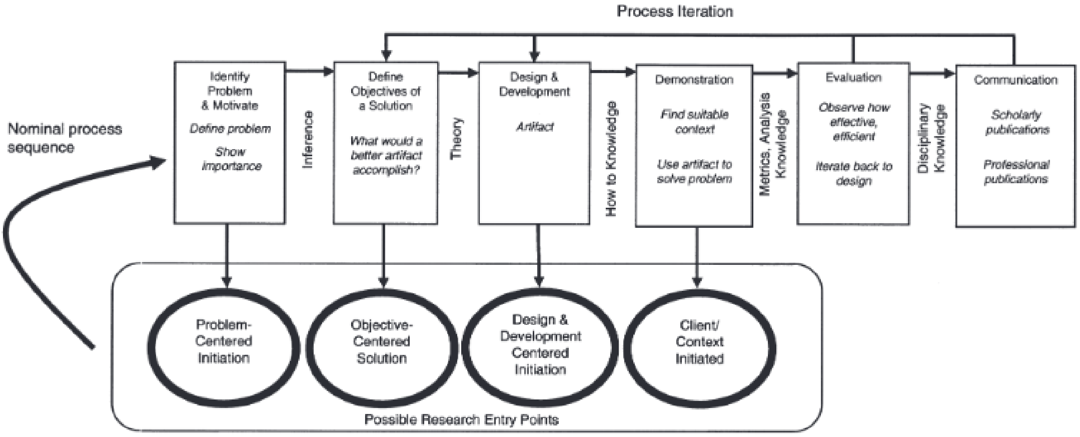

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

***Figure 6:** Design science research methodology process model* (Peffers et al., 2007).

Another research strategy used in this study is case study, which can be used to generate context-specific understanding of how the frameworks, methods, and models can be practically applied to real-world business case, which in this case is consumer segmentation (Saunders et al., 2019, pp. 196-198). Since design science research is trying to create artifacts that can be used to solve practical problems, case study supports the design science objectives well. These two research strategies have been used together often. (Hevner et al., 2004; Offermann et al., 2009; Peffers et al., 2007.)

The segmentation model developed in case study is using client organization's quantitative sales data as a data source. Saunders et al. (2019) describes the characteristics of quantitative research as examining the relationships between variables, that are numerically measured and analysed by utilizing various statistical and graphical techniques. These descriptions also apply the design of this research. The quantitative research can be either mono method quantitative study or multi-method quantitative study. These two describes how the data is collected for the study. Mono method means that study is using only one data collection technique, and multi-method study is using multiple data collection methods. (Saunders et al. 2019, pp. 176-178.) Since the case study is using only the data from client organization's database, the case study can be defined as a mono method quantitative study.

4.4 Time Horizon

The time horizon in research represents the period of time when the data is collected and analysed. Saunders et al. (2019, p. 212) specifies two main time horizons for the research, cross-sectional and longitudinal time horizons. They describe cross-sectional as a “snapshot” of particular time or timeframe, and longitudinal as a series of snapshots. (Saunders et al., 2019.)

Even though this research is using historical data collected during multiple recent years, the time horizon is cross-sectional. The data is collected as a large snapshot, and it is used to generate customer segments based on customer behaviour during one particular timeframe. If the research would study for example customer behaviour year by year and compare the years against each other, the research would more likely be longitudinal.

4.5 Data collection

Saunders et al. (2019) mentions that while research design is like an overall plan of the research project, data collection and analysis is more like a tactics and detailed information about the research process. This research is using quantitative research design and structural database as a data source.

The quantitative data used in case study is collected by querying Structured Query Language (SQL) database. The publish layer of the database, which is queried, is in dimensional form. This means that the database includes fact and dimension entities. Fact entities are measurements, metrics, or facts of a business processes, like for example transactional sales data. Dimension, on the other hand, are generally slowly changing descriptive attributes for facts. (Linstedt, 2016.)

5. CASE STUDY

This chapter includes the real-world case study on how the identified customer segmentation frameworks and methods can be utilized in practice. The chapter is following the guidelines of design science research model presented in subsection 4.3, starting by defining the problem and objectives of a solution. Subsection 5.2 includes design and development of the solution, which is followed by demonstration and evaluation of each iteration. While the design science research model is used as a higher-level framework in this study, CRISP-DM model is used for dividing the process level tasks into smaller pieces.

5.1 Defining the problem and objectives of a solution

The first two phases of design science research model are identifying the problem and motivate and defining the objectives of a solution (Peffers et al., 2007). These two phases are already mostly covered in chapter 1, but this section elaborates on the case study specific problems and objectives.

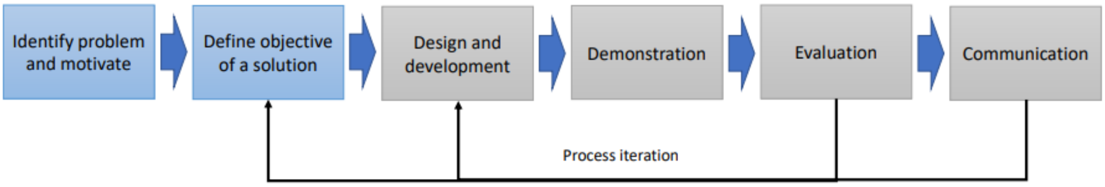


Figure 7: *First two phases of design science research process model.*

The design science research process flow used in case study is presented in figure 7 above. In order to identify the problem relevant to the business and define the objectives of a solution, it is important to understand the business and available data. These two steps are in line with the CRISP-DM model and will be presented in the following subsections 5.11 and 5.12. The number of iterations used in the development process of this research was agreed to be 3 and the length of each iteration was approximately 3-4 weeks. The development process started by kick-off meeting which included discussion about the research possibilities and objectives.

5.1.1 Business understanding

The client is retail sales company that operates in numerous countries, but mainly in Nordic countries. They want to utilize their existing data to improve their customer understanding. One of their primary objectives regarding to this research is to divide their customers into segments that can be utilized in various marketing processes. In addition to retail sales, the company provides some industry specific services which are also considered in this study when comparing various segments.

The client organization has practiced a bit of consumer segmentation in the past, but now there is a need to implement even more precise segmentation model that uses data scientific approaches to make data processing more effective. The final consumer segmentation model is planned to be a part of a larger consumer data project.

As mentioned in chapter 1, the segmentation model implemented in this research is PoC that illustrates what kind of consumer segmentation can be developed using existing data, but also what additional actions are required to improve the segmentation results. So, the objective and solution of the case study is to utilize the existing research and findings to implement a segmentation model that offers various opportunities for further development but can also offer valuable information about the customer behaviour right away.

5.1.2 Data understanding

Another important factor when determining the solution objectives is to understand the available data. The solution developed in this case study primarily utilized transactional data, specifically consumer sales data. In addition to transactional data, product and geographic data are used to create additional attributes that can be used to enrich the segmentation process.

Because of transactional data, the solution is built on RFM model, which is presented in subsection 3.2.2. In this study, RFM model is used for creating customer segments based on their buying behaviour. Geographic and product data are utilized to explain the RFM segmentation results in the later stages of the research.

The data used in this research is stored in dimensional relational database, which means that there are fact and dimension tables that are connected to each other by keys that have generally one-to-many relationships with each other. This means that each row in fact table has one corresponding key value in dimension table, but one row in dimension table may have multiple corresponding key values in fact table. The sales data is stored in fact table that includes relations to dimensional customer and product tables.

Sales data from the last five years were selected for the study. It includes more than 44 million transaction rows, but the results are filtered in database level before fetching. The quality of data is mainly good, but there were some minor issues like outliers from purchase entries for cash customers. Data filtering and outlier detection processes are explained in the data preparation section.

5.2 Design and development

The third phase of design science research process is design and development. In this phase, the business objectives set earlier are concretized by creating designing and developing the artifact or solution.

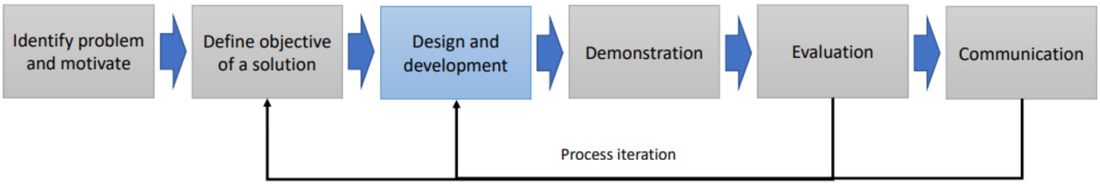


Figure 8: *Design and development phase of design science research process model.*

In this research, design and development phase is divided into data preparation and modelling in accordance with the CRISP-DM model. Additionally, relatively simple Kanban board is used to divide the project into smaller tasks, and to prioritize the tasks for each iteration round. The Kanban board used in this research included lists of to do, backlog, in progress, testing and done. These steps are explained in table 3 below.

Table 3: *Kanban board.*

To Do	Backlog	In progress	Testing	Done
Unprioritized list of tasks to be done.	Prioritized list of tasks for each iteration.	List of tasks that are under the development.	List of tasks that are already developed but requires testing.	List of tasks that are tested and done.

The Kanban board To Do -list includes all the identified tasks that are not yet prioritized for ongoing iteration. Tasks in kanban board are flowing from left to right. The kanban approach is generally known as a simple workflow management tool for teams but it can also be used in individual development projects like this research. (Hammarberg & Sundén, 2014.)

The segmentation model is developed by using the Python programming language in interactive web based Jupyter Notebooks environment, which is widely used environment in a field of data science, because the code can be divided into short shells that can be run separately. Jupyter Notebooks cells also accepts various text content. These features enable the creation of a clear and consistent story around the code.

A few open-source programming libraries have also been used in the implementation, such as pandas, NumPy and scikit-learn. Pandas is popular python programming library which is mainly used for data analysis and data manipulation. It provides two main data structures, which are one-dimensional array-like series and two-dimensional table-like structure dataframe. (Pandas, 2023.) This study is using dataframe structures for data transformation and analysis. NumPy is widely used python library for data analysis and scientific computing. It provides various mathematical functions, data types and operations. (NumPy, 2023.) Scikit-learn is a python machine learning library that provides wide range of supervised and unsupervised learning tools (Scikit-learn, 2023). This study is using scikit-learn when building K-means clustering model.

5.2.1 Data preparation

Data preparation starts by loading required data from the client database. Data loads are done by using SQL queries with basic select, from, join, and where structures. Using the SQL language enables partial filtering of the results already while querying the database. The SQL query which is used for fetching sales data from database is presented in program 1 below.

```
SELECT
    f.CUSTOMER_SITE_BILL_TO_ID,
    cs.CUSTOMER_CODE,
    cc.COST_CENTER_CODE,
    f.TRANSACTION_DAY_ID,
    f.INVOICE_NUMBER,
    f.INVOICE_LINE,
    f.COST_CENTER_CODE,
    f.TOTAL_PRICE,
    f.PRODUCT_ID
FROM PUBLISH.F_SALES f
LEFT OUTER JOIN PUBLISH.D_COST_CENTER cc
    ON f.COST_CENTER_CODE = cc.COST_CENTER_CODE
INNER JOIN PUBLISH.D_CUSTOMER_SITE cs
    ON f.CUSTOMER_SITE_BILL_TO_WID = cs.CUSTOMER_SITE_WID
WHERE cc.ORGANISATION_WID = 18
AND f.TRANSACTION_DAY_WID >= '2018-01-01'
AND cs.CUSTOMER_TYPE = 'B2C'
```

Program 1: *SQL query for fetching sales data.*

The query filters the data by selecting only rows with transaction date greater than 2018-01-01 and where customer type is B2C. The organization filter is used for selecting only the customers of Finnish organization. Query is also fetching only the sales that are made by customer that exists in the database. After filtering, the SQL query results 4 340 106 transaction rows which included 1 100 496 distinct invoice numbers and 284 517 distinct customers. In addition to fact table, some dimension tables like D_PRODUCT, D_CUSTOMER and D_COST_CENTER are fetched by separate queries for later analysis.

After the data is collected from the database, the tables are transformed into pandas dataframe. Before the RFM dataset can be created, certain columns must be selected from the source data based on which the recency, frequency and monetary attributes are formed. In this case, recency attribute can be created by using the transaction day. Frequency values can be generated by using invoice numbers, and monetary values can be derived from the total price of transaction. In addition, the individual customer must be identified by unique key, which in this case is customer code. Creation of RFM dataset is shown in program 2 below.

```
# Time parameters
end_date = F_SALES['transaction_day_wid'].max()
snapshot_date = end_date + dt.timedelta(days=1)

# Selecting columns
df = F_SALES[['customer_code', 'transaction_day_wid', 'invoice_number', 'total_price_local']]

# Changing attribute naming
df.columns = ['customer_code', 'recency', 'frequency', 'monetary']

# Grouping by customer id to get recency, frequency and monetary for
each customer
df_rfm = df_rfm.groupby(['customer_id']).agg({
    'recency': lambda x: (snapshot_date - x.max()).days,
    'frequency': 'nunique',
    'monetary': 'sum'
}).reset_index()
```

Program 2: *Creating RFM dataframe.*

The program above takes customer code, transaction day, invoice number and total transaction price from the sales table, changes the attribute names and groups the dataframe by each customer. The recency values are calculated by subtracting the customer's latest transaction day from the snapshot date that is the latest transaction day of the whole dataset plus one day. The frequency values are calculated by counting every distinct invoice number for each customer and monetary values are calculated by summing every transaction prices together for each customer. These aggregation functions work similarly as formulas 4, 5 and 6 in subsection 3.2.2, but the time frames for frequency and monetary values does not need to be considered since the data is filtered by transaction day earlier. Alternatively, the group by function could also be implemented in the SQL query when fetching the data from database.

One important operation of data pre-processing is outlier detection. Outliers are data points that significantly differ from the from the majority of the data in dataset. Outliers may be result of various errors, like human data entry error. In this type of research, the consumer dataset might also include business customers that may be difficult to distinguish from the consumer customers and have exceptionally high frequency and monetary values. Also, since the client works in the retail sales industry, there might be multiple transactions, made by for example cash customer, recorded for the general customer account. One way to identify outliers is to visualize the data. According to Sarkar et al. (2019), typical ways to visualize data for outlier detection are scatter- and boxplots. The figure 9 below presents how the outliers can be found by using scatterplot and boxplots.

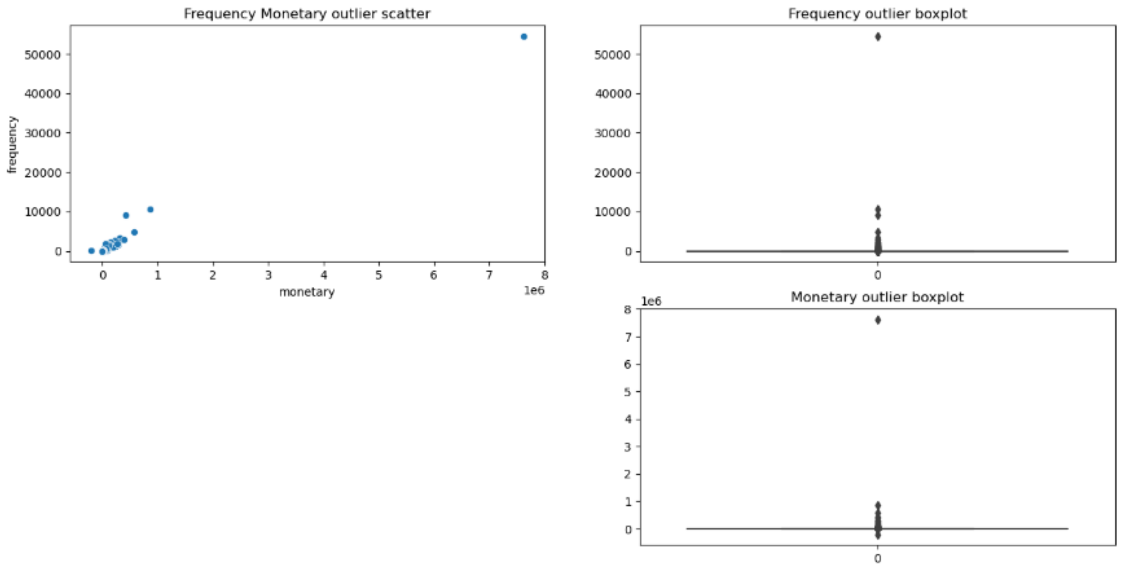


Figure 9: *Frequency and recency outliers.*

The scatterplot on left side of the figure above presents customers by frequency and monetary values in two-dimensional graph. The boxplots for frequency and monetary values are presented on the right side of the figure. As the figure shows, there are values with multiple times higher frequency and monetary values compared to majority of data points. The furthest outlier frequency value is 54 516, while the mean frequency value before outlier removal is around 2,9. Removing these outliers is necessary because they can falsify the results of analysis significantly. One way to remove the outliers from big datasets is to use quantiles. The following program 3 shows how the outliers can be removed from the RFM dataset by using quantiles.

```
def remove_outliers_rfm(df_rfm, q_high):
    q_high_f = df_rfm['frequency'].quantile(q_high)
    q_high_m = df_rfm['monetary'].quantile(q_high)
    df_outliers = df_rfm[(df_rfm['frequency'] > q_high_f) &
                         (df_rfm['monetary'] > q_high_m)]
    df_rfm = df_rfm[(df_rfm['frequency'] < q_high_f) &
                    (df_rfm['monetary'] < q_high_m) &
                    (df_rfm['monetary'] > 0)]
    return df_rfm, df_outliers
```

Program 3: *Detecting outliers from RFM dataset.*

The function in program 3 takes dataframe and highest quantile as a parameter and finds the limit values or frequency and monetary that meets the quantile. The quantile used in this research is 0,995. After that the function filters results that has higher frequency or monetary values than the limit values. In this case outliers are removed only by the frequency and monetary values. Recency does not have outliers because the values are calculated within the specific time frame. The outlier removal function returns the filtered dataset and the dataset of outliers that has been removed. After the outlier customers are removed, the recency, frequency and monetary values can be reviewed as in figure 10 below.

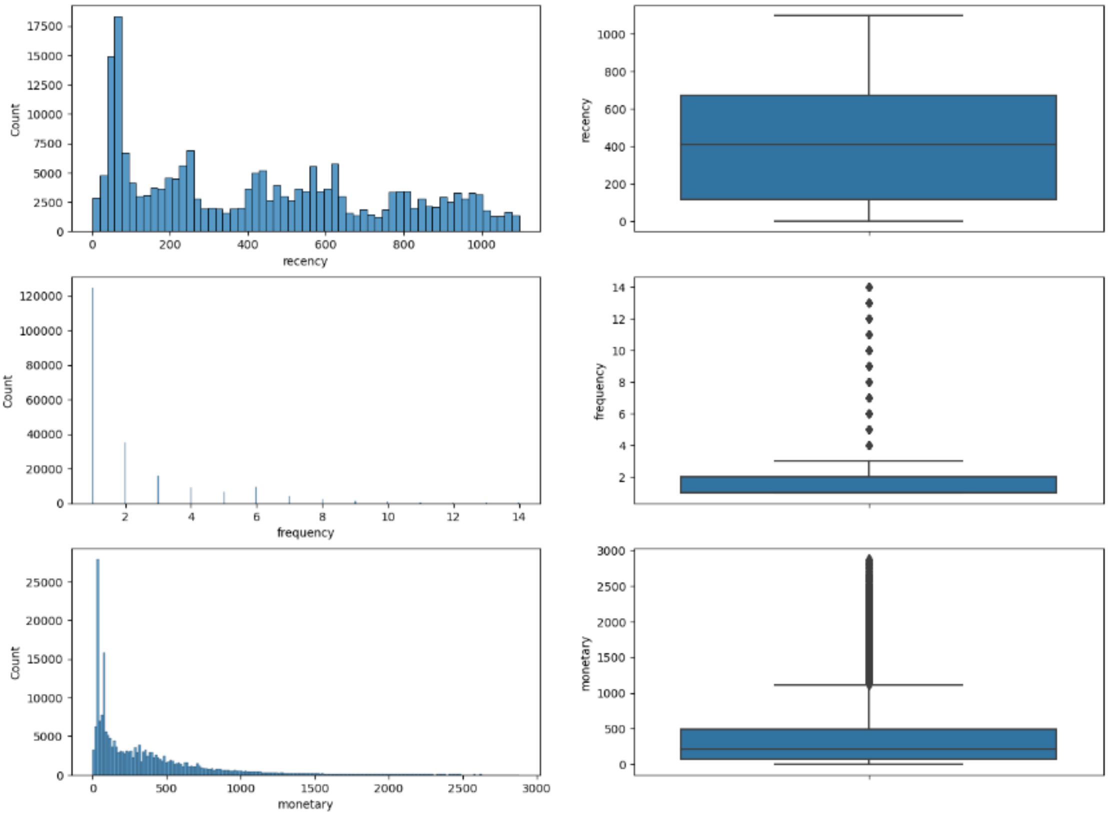


Figure 10: *RFM data after outlier removal.*

As the frequency and monetary graphs on the right side of the figure 10 shows, there is still some values outside of the boxplot, but the values are such that they are suitable for the consumer of retail organization. The graphs on the left side of figure 10 presents recency, frequency, and monetary values in relation to the count of customers with exact R, F, or M value. For example, the majority of customers has frequency value 1, so they have visited only once in given time frame. The figure above also shows that especially monetary values of the data is highly skewed. Data skewness in this study is reduced by using log transformation.

Another important data pre-processing process is data scaling. Data can be scaled by using normalization or standardization. Normalization scales data into range 0 to 1, and standardization scales data to have a standard deviation of 1 and mean of 0. Both of these scaling operations were considered in this study, but after testing standardization gave better results. StandardScaler for standardizing and MinMaxScaler for normalization from scikit-learn pre-processing library were used. The log transformation and scaling functions are presented in program 4 below.

```
def scaler_minmax(unscaled_df):
    scaler = MinMaxScaler()
    scaled_data = pd.DataFrame(scaler.fit_transform(unscaled_df))
    return scaled_data

def scaler_standard(unscaled_df):
    scaler = StandardScaler()
    scaled_data = pd.DataFrame(scaler.fit_transform(unscaled_df))
    return scaled_data

df_log = np.log(df_rfm[['recency', 'frequency', 'monetary']]+1)
scaled_data_std = scaler_standard(df_log)
```

***Program 4:** Data scaling before clustering.*

Functions in program 4 above takes unscaled RFM dataframe as input and returns scaled values of recency, frequency and monetary. Scaling process is important for K-means clustering in the modelling phase of research.

5.2.2 Modelling

The next phase of the research process is modelling, which includes building RFM model, scoring the RFM attributes, creating segments, and building the K-means clustering model. The program 5 below presents the functionality behind scoring RFM values.

```
R_Labels = range(4, 0, -1)
F_Labels = range(1, 5)
M_Labels = range(1, 5)

R_Groups, rbins = pd.qcut(df_rfm['recency'], q=4, labels=R_Labels,
duplicates='raise', retbins=True)
F_Groups, fbins = pd.cut(df_rfm['frequency'], bins=[0, 1,3,6,
df_rfm['frequency'].max()], labels=F_Labels, retbins=True)
M_Groups, mbins = pd.qcut(df_rfm['monetary'], q=4, labels=M_Labels,
duplicates='raise', retbins=True)

df_rfm_pre = df_rfm.assign(R = R_Groups.values, F = F_Groups.values, M
= M_Groups.values)
df_rfm_pre['rfm_segment_concat'] =
df_rfm_pre['R'].astype(str)+df_rfm_pre['F'].astype(str)+df_rfm_pre['M'
].astype(str)
df_rfm_pre['rfm_sum'] = df_rfm_pre['R'].astype(int) +
df_rfm_pre['F'].astype(int) + df_rfm_pre['M'].astype(int)
```

***Program 5:** Scoring the RFM values.*

The labels for recency, frequency and monetary values are in range 1 to 5, which means that the model is creating four categories for each attribute. The recency labels go in reverse order, that is from the largest to the smallest value, because smaller recency values are better. Recency and monetary values are grouped by using pandas library qcut functionality that divides the values into four equal groups. Frequency values are grouped differently, since the values cannot be divided into four equal groups, because there are too many one-time visitors in data. For solving the problem, program is using bins 0, 1, 3, 6, and maximum value of frequency. This means that the customers with frequency value 1 are in category 1, customers with frequency values 2 and 3 are in category 2, customers with frequency values 4, 5 and 6 are in category 3 and customers that have higher frequency than 6 are in category 4.

After the RFM attributes are categorized, the scores are assigned to each customer. In addition, based on the RFM scores, a concatenated segment and the sum of the R, F and M scores are added to each customer. The concatenated segment can be immediately used in the segmentation analysis and the sum of scores is helpful in further segmentation analysis.

The RFM scores and sum values can be used to divide customer into the various segments. For example, in this research, customers are divided into 8 different segments by using the RFM sum and scores. The segments and the rules of how different segments are formed in this study are presented in the table 4 below.

Table 4: *RFM segments and the rules.*

Segment	**Rule**	**Description**
Top customers	RFM sum >= 11	Customers with highest score. [444, 443, 434, 344].
Loyal customers	RFM sum >= 7 AND Frequency = 4	Customers with highest frequency score, but not in top customers.
Big spenders	RFM sum >= 7 AND Monetary = 4	Customers with highest monetary score, but not in top customers.
Good customers	RFM sum >= 7	Customers with good RFM score, but not in groups above.
Needs attention	RFM sum >= 5 AND Recency = 1	Customers with lowest recency score and with moderately low score.
One-time visitor	RFM sum >= 5 AND Frequency = 1	Customers who have visited only once and has moderately low score.
Potential customer	RFM sum >= 5	Customers with score higher than 5, but not in groups above.
Lost customers.	RFM sum < 5	Customers with lowest score. [111, 211, 121, 112]

The segments above are mainly outlined by using the RFM sum, and the more detailed divisions are made by utilizing R, F, or M scores. Customers with RFM sum 11 or 12 are segmented in the top customers group and customers whose RFM sum is less than 11 and greater or equal to seven, are either loyal customers, big spenders, or good customers, depending on the individual R, F, or M scores. Similarly, “Needs attention” customers, one-time visitors, and potential customers are the ones with RFM sum in between 5 and 7. The customers with lowest score can be described as lost customers.

Customer segments can also be created by using clustering. This research is using the K-means clustering algorithm which was described in subsection 3.2.3. Before running the K-means algorithm, it is important that data is normalized and not skewed. Also, the number of clusters need to be determined. As mentioned in theory, two ways to find the optimal number of clusters are elbow method and silhouette score. Program 6 below includes functions for both, elbow method and silhouette score. Both methods are implemented by using the scikit-learn machine learning library and visualized by using the matplotlib library.

```
def elbow_method(df, low, high):
    inertia_list = []
    for k in range(low, high):
        kmeans = KMeans(n_clusters=k)
        kmeans.fit(df)
        inertia_list.append(kmeans.inertia_)
    plt.figure(figsize=(15,10))
    plt.plot(range(low, high), inertia_list, marker='o',
markerfacecolor='red', linestyle='dashed')
    plt.title('Elbow method')
    plt.xlabel('K value')
    plt.ylabel('Inertia')

def silhouette_score_rfm(df, low, high):
    silhouette_scores = []
    for k in range(low, high):
        kmeans = KMeans(n_clusters=k)
        kmeans.fit(df)
        cluster_labels = kmeans.predict(df)
        silhouette = silhouette_score(df, cluster_labels)
        silhouette_scores.append(silhouette)
        print(k)
    plt.figure(figsize=(15,10))
    plt.plot(range(low, high), silhouette_scores, marker='o', marker-
facecolor='red', linestyle='dashed')
    plt.title('Elbow method')
    plt.xlabel('K value')
    plt.ylabel('Silhouette score')
```

***Program 6:** Finding the optimal number of clusters by using elbow method and silhouette score.*

Both functions in program 6 above takes the dataframe, and lowest and highest possible number of clusters as input. The elbow method calculates inertia values for each cluster and appends them to list. The equations for calculating inertia values (formula 5) and silhouette scores (formula 6) for each cluster are presented in subsection 3.2.3. The results of elbow method are presented in figure 11 below.

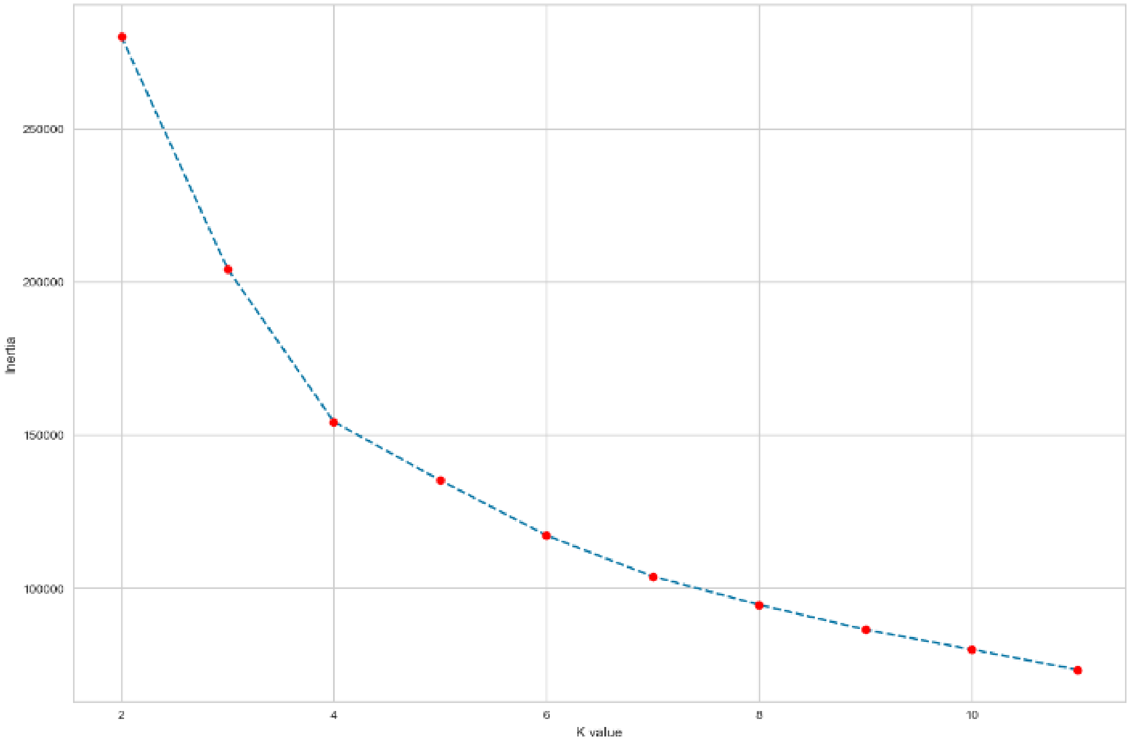


Figure 11: *The elbow method.*

The optimal number of clusters according to elbow method is 4, because that’s when the curve is at its greatest. For comparison, based on the elbow method, the study is also considering clustering with three and five clusters. However, the elbow method does not give fully clear answer of the optimal number of clusters, so the study is also using the silhouette score to confirm the result. The similar graph for silhouette score is presented in figure 12 below.

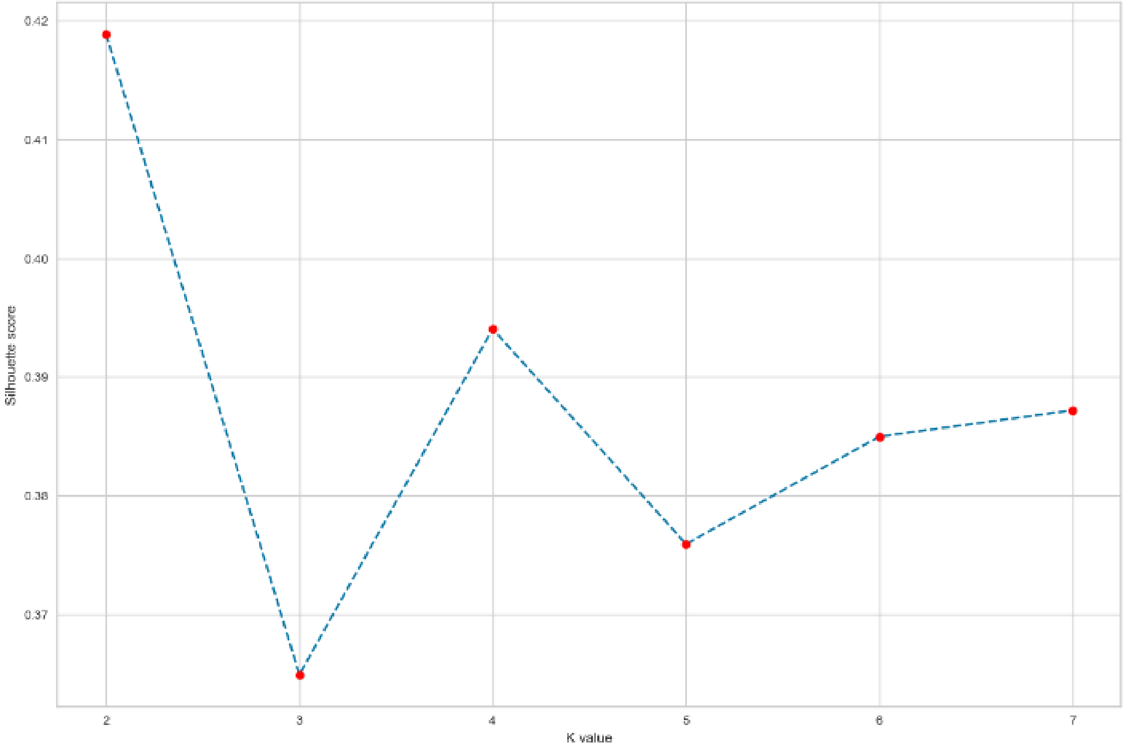


Figure 12: *The silhouette score method.*

Also, the silhouette score method gives four as an optimal number of clusters. The K-means model can be built by using the function in program 7.

```
def kmeans(df_scaled, k, original_df):

    # Running K-Means algorithm
    kmeans = KMeans(n_clusters = k, init='k-means++', random_state=1)
    model = kmeans.fit(df_scaled)
    cluster_labels = kmeans.labels_
    centers = kmeans.cluster_centers_
    new_df = original_df.assign(Cluster = cluster_labels)

    # Visualizing the results
    fig = plt.figure(figsize=(15,10))
    ax = fig.add_subplot(111, projection='3d')
    ax.scatter(df_scaled ['recency'], df_scaled ['frequency'],
df_scaled ['monetary'], cmap = "brg", c=model.predict(df_scaled))
    ax.scatter(centers[:, 0], centers[:, 1], c='black')
    ax.set_xlabel('recency')
    ax.set_ylabel('frequency')
    ax.set_zlabel('monetary')
    df_scaled = df_scaled.assign(Cluster = cluster_labels)
    return new_df, centers, df_scaled
```

Program 7: *K-means algorithm.*

The K-means function above takes normalized dataframe, number of clusters, and the original dataframe as an input. Generated cluster labels are also assigned to the original dataframe that includes the original RFM data. This can help understanding the clustering results and the features of each cluster.

5.3 Demonstration

The design and development phase are followed by demonstration of the results. The results were demonstrated after each iteration, but this section represents how the final results of the research were demonstrated. After this section, the results after each iteration round are evaluated and analysed.

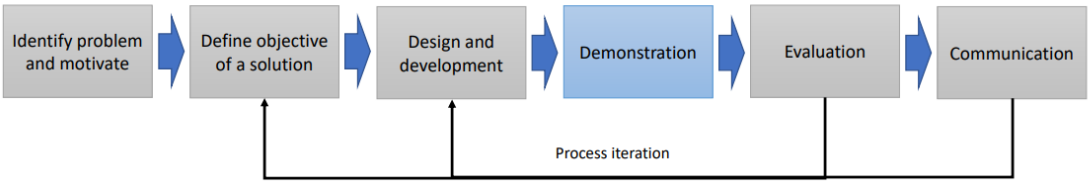


***Figure 13:** Demonstration phase of design science research process model.*

RFM segmentation was implemented in the case study in two different ways. First, customers were divided into segments by using manually made rules that are presented in table 4. After that, RFM data was applied to K-means clustering algorithm. The table 5 below presents the statistical features of each segment created by manually formed rules.

***Table 5:** Statistics for each RFM segment.*

Segment	Recency (avg)	Frequency (avg)	Monetary (avg)	Monetary (sum)	Customer count
Lost customers	1242.92	1.02	64.10	3 658 910 €	57080
Potential	785.43	2.20	109.11	970 768 €	8897
One-time visitor	548.43	1.00	132.78	7 696 595 €	57963
Good customers	365.48	2.29	288.46	17 735 290 €	61481
Needs attention	1435.76	1.50	367.84	9 783 448 €	26597
Big spenders	529.50	2.96	919.33	41 469 240 €	45108
Loyal customers	488.62	8.81	1116.79	8 134 761 €	7284
Top customers	122.11	10.15	1428.36	23 432 300 €	16405

The table 5 above includes average of recency, frequency, and monetary values, sum of monetary values, and the number of customers in each segment. The table is ordered by average monetary values. The biggest customer segment by customer count is good customers. The big spenders spend by far the most money in total compared to other segments and top customers are the best customer based on the average values of all RFM attributes. Figure 14 below demonstrates the sizes of each segment compared to others.

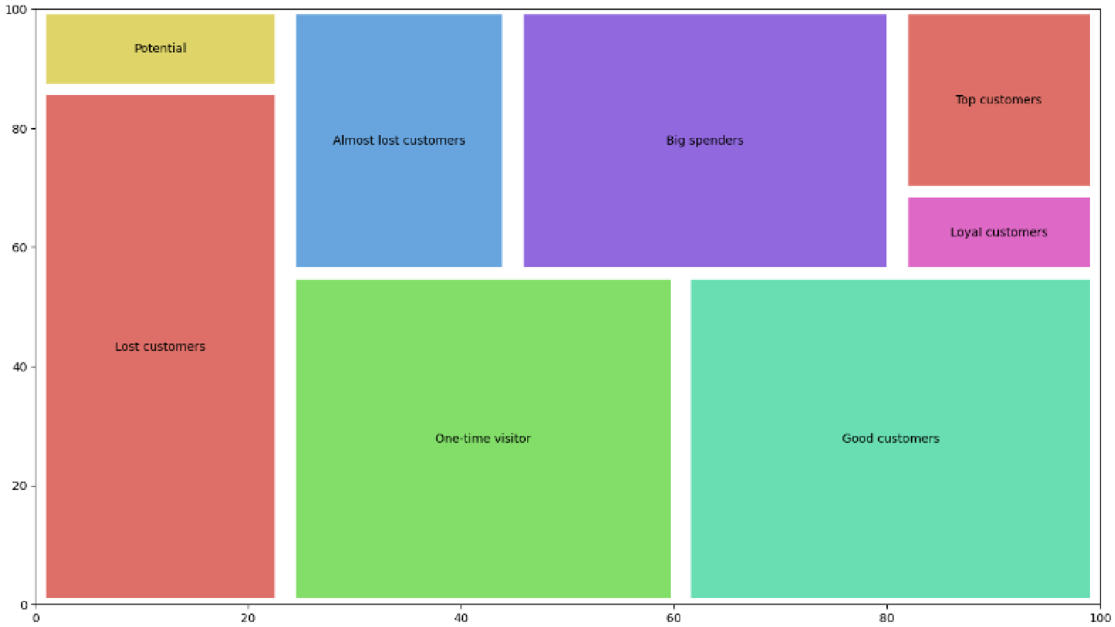


Figure 14: *Visual presentation of RFM segment size compared to others.*

For comparison, RFM data is also segmented by using K-means clustering, of which algorithm is presented in program 7. Figures below represents the results of K-means clustering in three-dimensional graphs and snake plots. The snake plot is a visual presentation of various features that clusters have.

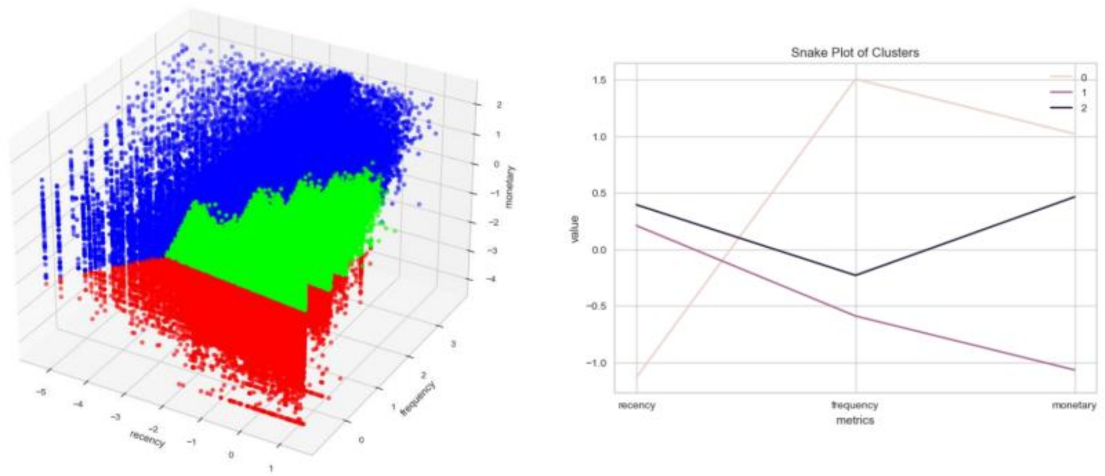


Figure 15: *Clustering results when K=3.*

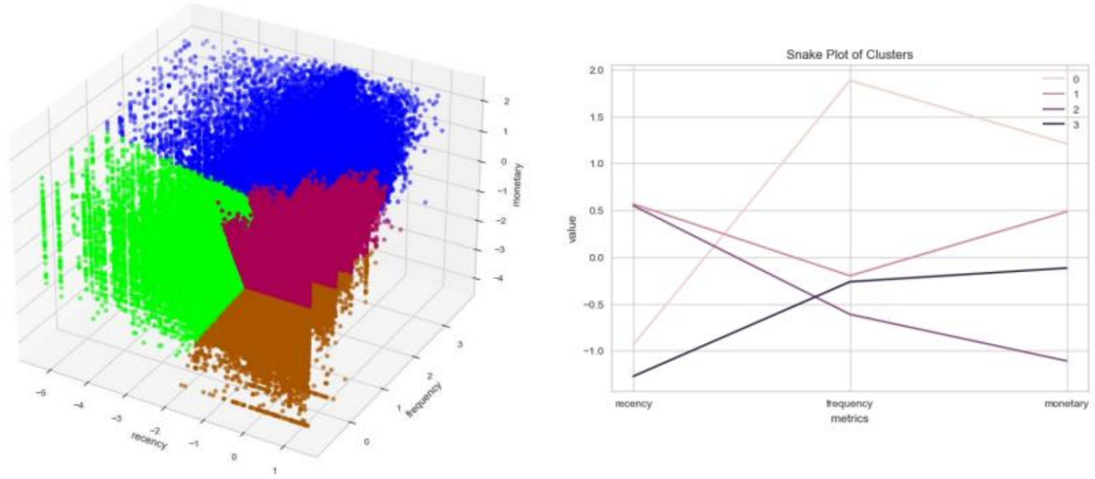


Figure 16: *Clustering results when K=4.*

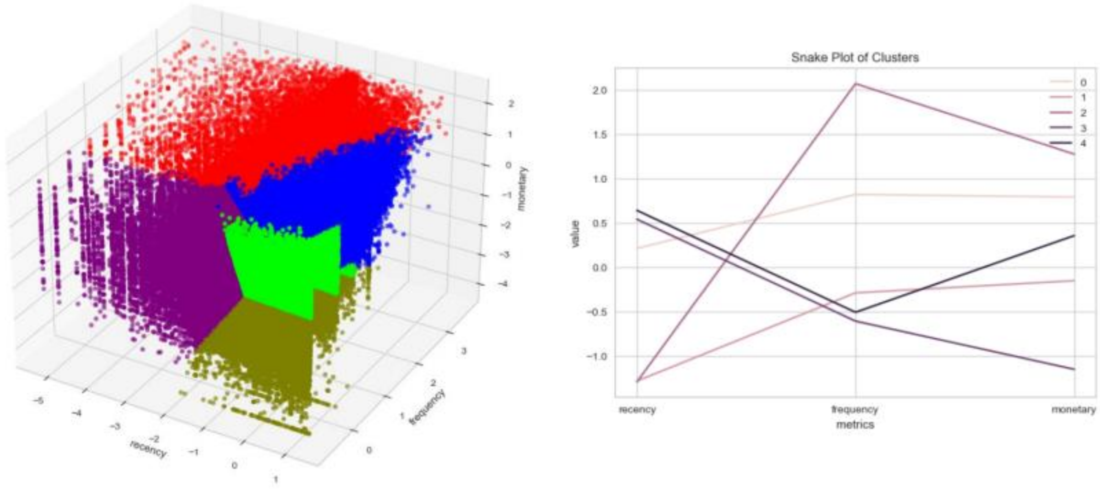


Figure 17: *Clustering results when K=5.*

Figure 15 represents clustering results when K=3, similarly in figure 16 K=4, and in figure 17 K=5. On the left side graphs, the colour of the data point represents the cluster data point is assigned to. For example, when K=3, the green and red clusters are clearly differentiated by monetary values, and when K=4, green and blue clusters have differences in frequency values, and so on. However, even though these three-dimensional graphs can be used for rough demonstration of how the data points are divided into various clusters, it is not enough to clearly understand the clustering results. The left side graphs also do not show how the data points are distributed under the mass, such as in the back corner of the graph when recency and monetary values are at a minimum and frequency is at a maximum. To achieve deeper understanding, the snake plots on the right side can be used to demonstrate the various characters of each cluster.

To obtain even more detailed information about the clusters, the statistics are presented similar as manually formed segments earlier. The table 6 below presents the statistical information of how the data is divided into clusters, when K=4.

Table 6: *Statistics for each cluster.*

Cluster	Recency (avg)	Frequency (avg)	Monetary (avg)	Monetary (sum)	RFM sum (avg)	Customer count
0	262.59	7.311	1124.60	51 326 880 €	10.80	45 640
1	963.45	1.73	441.73	43 578 500 €	6.46	98 652
2	966.49	1.11	56.73	4 993 224 €	4.28	88 004
3	159.54	1.61	267.57	12 982 710 €	7.58	48 519

For comparison, the average RFM sum values are added to table. These values demonstrate immediately which groups are the most valuable ones regarding to RFM values. The results also shows that customers are divided into two larger and two smaller clusters.

In addition to general segmentation, some industry-specific attributes are used explain the segmentation results and generated RFM values. To demonstrate how the various attributes correlates with each other, this research is using correlation matrix, which is presented as a heatmap in figure 18 below.

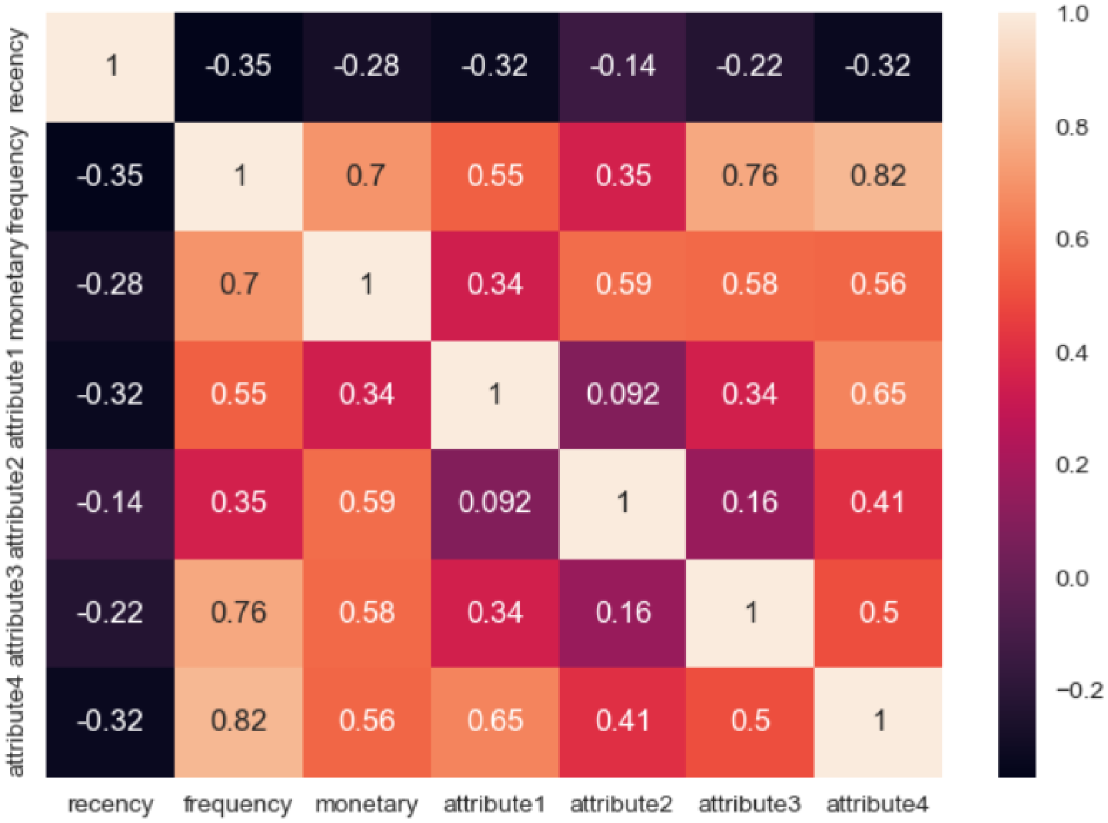


Figure 18: *Correlation matrix heatmap of RFM values and additional attributes.*

In correlation matrix, the correlation value between two attributes demonstrates how well those attributes correlates with each other. If value is 1, two values correlate perfectly, as shown in the centre line of the graph, and if the value is 0, there is no correlation between attributes. Similarly, negative values represent negative correlations between attributes. For example, recency values correlate negatively with other attributes, because smaller recency values are better than higher ones.

The same industry-specific attributes are also used to explain the RFM segments by calculating the percentage of how many customers fulfils the attribute within segment. These percentages are presented as horizontal bar charts in figure 19 below.

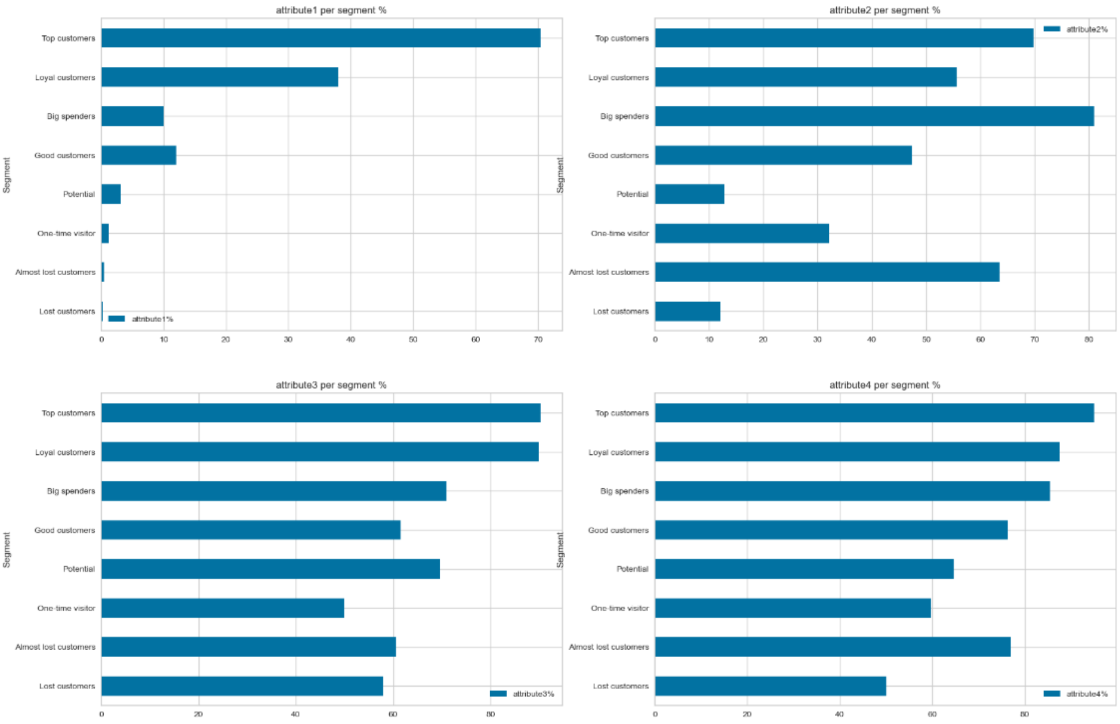


Figure 19: *Explaining segment by additional attributes.*

The bar charts in figure 19 above presents how the various attributes are divided in each consumer segment. These segmentation results are evaluated in next chapter.

5.4 Evaluation

This section presents how the model was evaluated after each iteration. The evaluation was partly integrated to the same client meeting where the current results were demonstrated. Two to four representatives of the client's organization were present in each meeting, so that the state of the model and potential business applications could be discussed from several perspectives. After the meetings, the current results were also sent to client organization for even closer examination. This enabled not only direct verbal feedback but also more carefully considered written feedback. The collected feedback was used in model evaluation and to create objectives for the next iteration.

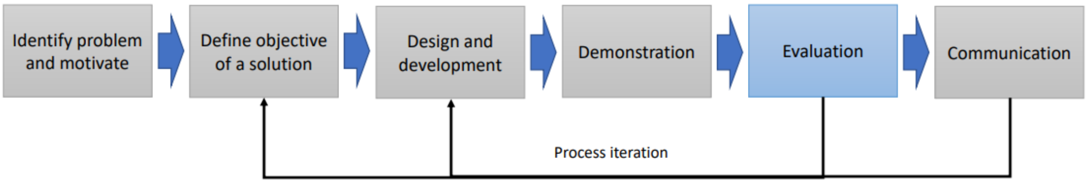


Figure 20: *Evaluation phase of design science research process model.*

The model development process started by initial meeting where the objectives for the first iteration were set. These objectives included data collection, preparation, and developing the basic RFM segmentation model. The objectives for the next iteration were set based on the previous iteration's feedback and evaluation. The feedback was mainly collected after demonstrating the results, but after third iteration client organization was also able to test the results by using interactive report. The evaluations for each iteration are presented in next three subsections, starting from the main results of current state, and then analysing the collected feedback and results.

5.4.1 Results and analysis: Iteration 1

Objectives of the first iteration were to investigate available data and possible segmentation solutions, fetching the data by writing SQL queries, prepare the data for RFM model, and building the MVP of RFM segmentation model. The first version of segmentation model used limited dataset for testing purposes. It included consumers from one Finnish city only and the time frame included only last two years. The number of customers in pre-processed RFM dataset was 4102. Smaller dataset was easier to manage in early phases of research.

The most time-consuming process during first iteration was data investigation and preparation, which included implementing SQL queries and filtering the results, handling various attributes, such as time parameters, detecting and removing outliers, and creating RFM dataset with recency, frequency, and monetary features. Logarithmic transformations and scaling were not implemented in the first iteration.

The main results of the MVP segmentation model were recency, monetary and frequency scores created by using bins in range [1, 4], 64 concatenated RFM segments, and manually made segments using RFM sum and scores. Table 7 represents how the data is divided in manually made segment.

Table 7: *RFM segments after first iteration.*

Segment	Recency (avg)	Frequency (avg)	Monetary (avg)	Monetary (sum)	Customer count
Requires activation	450.53	1.02	90.82	84917.32	935
Potential	260.52	2.03	150.84	16743.07	111
One-time visitor	171.65	1.00	203.69	117327.97	576
Good customers	71.38	2.88	302.55	284702.31	941
Needs attention	515.25	1.25	488.55	169040.13	346
Top customers	37.97	4.36	762.26	620484.55	814
Big spenders	238.69	1.77	811.16	307429.71	379

In the first iteration, top customers were customers with RFM sum 10 or higher and the group with worst score was named as requires activation. Otherwise, the rules were similar as in table 4. The group of loyal customers does not show in the table above, because there was none, since all the customers with frequency score 4, were most likely included in top customers. The sizes of each customer segment are visualized in figure 21 below.

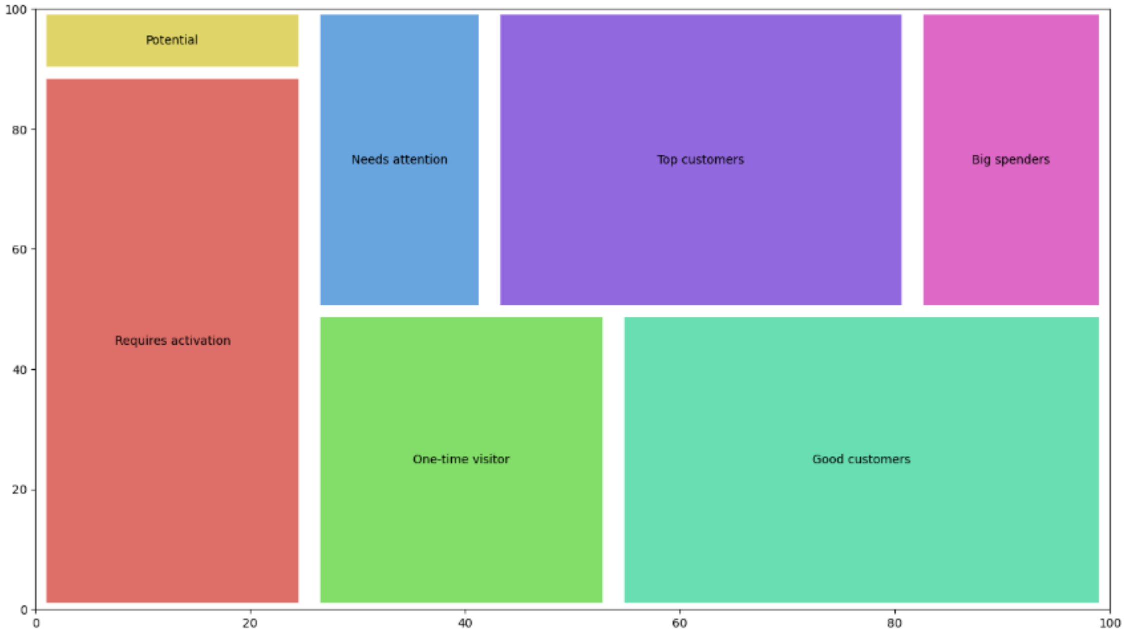


Figure 21: *Visual presentation of RFM segments after first iteration.*

The first iteration was about understanding the case environment, data and developing the MVP artifact that can be used for demonstrating the results. The objectives of the first iteration were fulfilled and a lot of development ideas were raised in discussions and evaluation after first demonstration. It was found that some interesting consumer profiles can be already detected using the model created in first iteration, for example big spenders are representing a group with low number of transactions, but with the relatively high monetary value. In addition, the results showed that good and top customers were the most frequent ones. This raised discussions about how the model could better explain these customer groups by using various additional attributes for example.

Based on the evaluation discussions and gathered feedback it became clear that the time frame of test data used in model was too short to obtain appropriate results in this type of retail industry. The problem could be seen as a low frequency averages and lack of loyal customers. City-specific delineation in the selection of test data proved to be a good choice, as it was possible to examine an entire part of the population in exact area. If the customers had been limited by random selection, for example, the results could have been further from the truth.

Overall, the evaluation of the model proved that RFM modelling can be appropriate consumer segmentation method for client organization, even though the results were not yet useful from business point of view at this early stage. The MVP model raised several development directions, such as focusing on features of some certain service users, using some outer data sources like open data, and improving the current model overall. However, since the model was still in early phases, it was agreed that the next iteration should focus on improving the RFM model by expanding the dataset, utilizing machine learning, and adding some explanatory attributes that can be used to explain the customer groups.

5.4.2 Results and analysis: Iteration 2

The second iteration was for improving the segmentation model implemented in the first iteration. The range of data was extended to include all the transactions made in Finland after 2020-01-01. Now the dataset included 210 299 distinct customers. Also, the additional explanatory attributes, customer’s geometric information and K-means clustering functionality were included to the model. K-means modelling included finding the optimal number of clusters by elbow and silhouette score methods, adding scaling, and running the clustering algorithm which is presented in program 7. The scaling was implemented by using the Min-Max scaler, but the logarithmic transformation was not implemented at this point yet. The clustering results with 4 clusters are presented in figure 22 below.

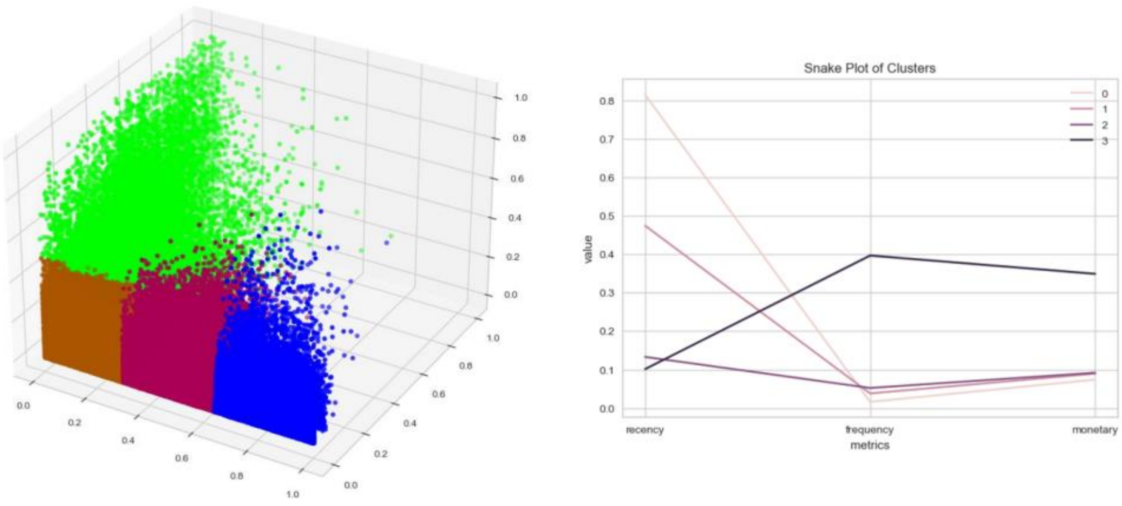


Figure 22: *Clustering results iteration 2, when K=4.*

As the figure above shows, the clusters 0, 1 and 2 were mainly differentiated by recency attribute, while frequency and monetary values are almost the same. Only cluster 3 differs from the other by frequency and monetary attributes. This happens because the attributes have big differences with their value ranges, as explained in subsection 3.2.1. Even though the Min Max scaler scales each attribute in range 0 to 1, the value differences inside the attribute remains the same.

The correlation matrix demonstrated in figure 18 is used to present the correlations between RFM attributes and explanatory attributes. Matrix shows that frequency and monetary values have high correlation, which basically means that the customers who visits frequently, brings also more monetary value. Frequency also correlates well with attributes 3 and 4, which are both industry-specific service visits. Customers with attribute 1 are likely to have also attribute 4, but the correlation between attribute 1 and attribute 2 is near zero. This kind of correlation matrix is very useful when adding new attributes and analysing if those attributes effects to the results.

Another use case for explanatory attributes is presented in figure 19. It explains how high percentage of customers in each RFM segments has the exact attribute. According to figure 20, more than 70 percent of top customers has an attribute 1, which is an industry specific service. Also, loyal customers are using this service fairly frequently. Most of the customers in big spenders’ segment has attribute 2, which means that customer has bought product from industry specific product group. Attributes 3 and 4 also describes industry specific service usage. These services are used moderately in each customer segment.

In the evaluation of second iteration, the focus was on the general quality of the data and model, the value added by the new attributes, the necessity of clustering, and the presentation and comprehensibility of the model. According to the collected feedback, the overall state of the RFM based model was already good and it created new information about the segments and various attributes. The first identified data and model related issue was that the segments were still unevenly distributed, especially the number of loyal customers was very small. The problem was found to be the fact that the time frame was still too small. There were also some minor data-distorting issues with merge functionality when adding the new attributes to the RFM dataframe. However, the new attributes were perceived as appropriate for the model, as they were able to explain certain segments better.

At this stage of the study, it was found that the K-means clustering did not add any extra value to the segmentation results, since the basic RFM model was perceived to be more accurate and easier to customize for various business requirements. It was still agreed that the clustering could be appropriate segmentation solution in the future, but it requires improvements. The last focus area in second evaluation was the presentation and comprehensibility of the model. The current segmentation results were presented as a tables and visuals, but it was agreed that various attributes, segments, and visuals would require more detailed explanations, that would improve the usability of the results.

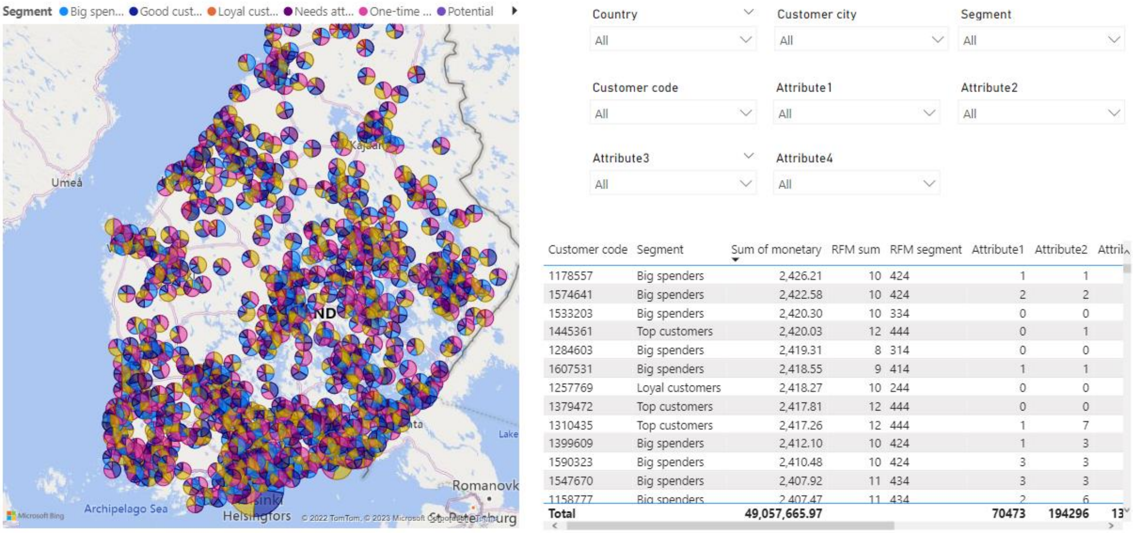
Overall, the RFM segmentation results after second iteration were already useful and promising when trying to understand organization’s consumer buying behaviour. The next steps for third iteration based on evaluation were to grow the time frame to last five years, develop the data preparation to avoid data-distorting issues and to improve clustering results, and focus on the data presentation.

5.4.3 Results and analysis: Iteration 3

Objectives for third iteration were to improve the results of analysis by developing the data preparation process, create interactive visualizations, and increase the date limit to five years, beginning from 2018-01-01. In addition, the segmentation model required multiple small fixes.

The logarithmic operation and scaling were added to data preparation by using the program 4. The difference when using scaled and unscaled data in K-means clustering algorithm can be seen in figures 22 (unscaled) and 16 (scaled). Even though the results of creating segments by using clustering were promising, at this point it was decided along with the client organization that the study will focus more on the traditional RFM segmentation, but the clustering results can be used as a comparison. In addition, because the research is PoC by the nature, the clustering can be useful in future development. If product would include additional, like demographic, geographic and psychographic attributes as a segmentation criterion in future, it would be hard and time consuming to score the segments manually without using machine learning functionalities.

The interactive visualizations were created by using Microsoft Power BI reporting tool and the report can be used in organization’s Power BI service environment. Power BI is business intelligence tool that can be used for multiple purposes such as data modelling, visualizing, analysing, and sharing information across the organization (Arnold, 2022). By using the Power Bi, segmentation results can be used as a support in decision making directly by organization’s business and marketing departments. In this research, the segmentation results for each consumer are saved to file, which is imported to Power BI as a dataset. The report is using segmentation results as a dimension table, which unique identifier is customer code. This way the results can be used together with full sales and other data imported from database. The following figure 23 represents how the RFM segmentation results could be visualized in Power BI map by each city.



Country: All
Customer city: All
Segment: All
Customer code: All
Attribute1: All
Attribute2: All
Attribute3: All
Attribute4: All

Customer code	Segment	Sum of monetary	RFM sum	RFM segment	Attribute1	Attribute2	Attril
1178557	Big spenders	2,426.21	10	424	1	1	
1574641	Big spenders	2,422.58	10	424	2	2	
1533203	Big spenders	2,420.30	10	334	0	0	
1445361	Top customers	2,420.03	12	444	0	1	
1284603	Big spenders	2,419.31	8	314	0	0	
1607531	Big spenders	2,418.55	9	414	1	1	
1257769	Loyal customers	2,418.27	10	244	0	0	
1379472	Top customers	2,417.81	12	444	0	0	
1310435	Top customers	2,417.26	12	444	1	7	
1399609	Big spenders	2,412.10	10	424	1	3	
1590323	Big spenders	2,410.48	10	424	3	3	
1547670	Big spenders	2,407.92	11	434	3	3	
1158777	Big spenders	2,407.47	11	434	2	6	
Total		49,057,665.97			70473	194296	13

Figure 23: *Example Power BI map visualization.*

The figure above shows how consumers are divided by each city in Finland. The size of pie charts presents the number of consumers, and the colors show how the consumers are divided into various segments. The user can filter the results by using various slicers or by clicking the segment. The tables can be used for detailed information.

Based on the third evaluation, the data related issues were successfully fixed and the results were closer to the assumed values. Extending the time frame to five years also improved the segmentation results, but it was also discussed that the RFM scoring could be still improved and the sizes of different customer segments equalized. A possible solution to the problem could be to extend the RFM scoring range from 4 to 5. This would raise the maximum RFM score to 15, allowing for an even more detailed definition of segment creation rules.

The interactive data presentation allowed client organization to test the results with various filters and selections in a real business environment. This enabled an even more business-oriented evaluation of the results, and raised new concrete use cases and further development ideas for consumer segmentation.

Based on the collected feedback the overall state of the segmentation model after three rounds meets the objectives of PoC. The results of the solution proved that the current model can be used as a reference or base for future consumer segmentation solutions in client organization. It was also found that the current segmentation model can be used for detecting various customer groups with different needs. This information can be used for example in marketing when contacting different customer groups.

6. DISCUSSION

The theory of this research started by identifying various customer segmentation approaches that can be utilized in consumer markets. According to Sarkar et al. (2018) and Kotler (2019, p. 300), the most typical approaches are behavioural, demographic, geographic, and psychographic segmentation. This research focused mostly on behavioural segmentation, since it can be utilized by using transactional data, which was well available in the case organization's database and the data was also moderately good quality. Although the behavioural data was mainly used in the segmentation process, geographical and client's product-related data were also utilized in the analysis of the results. To improve customer understanding and other segmentation results, it would be crucial to also use the demographic customer data along with the buying behavior data (Goyat, 2011; Nguyen, 2021). Sarvari et al. (2016) also considered in their research that RFM modelling together with demographic data are one of the best segmentation approaches. However, to collect the demographic data, the study should have used additional data collection methods, like mass surveys. In addition, the number of customers in this research was so high that it would be impossible to reach enough of them for the analysis. In the future, demographic segmentation could be implemented, for example, separately for certain target groups or RFM segments.

The case research structure followed design science research methodology as a strategic framework. Additionally, the CRISP-DM process model was aligned to the DSRM framework as a detailed process level framework. Using CRISP-DM provided more data-oriented viewpoints to the process. Defining the business and data understanding, as well as data preparation and modelling provided some additional and necessary focus on those important steps during the development process. DSRM and CRISP-DM both follow an iterative development approach, which is essential for the design and development of customer segmentation models (Arunachalam & Kumar, 2018).

The behavioral customer segmentation model developed in case research allows users to divide their consumers into segments by using their purchase recency, frequency, and monetary features. The RFM modelling is relatively simple and one of the most common behavioral customer segmentation methods when analyzing transactional data. (Christy et al., 2021; Sarkar et al., 2018; Sarvari et al., 2016.) The analysis shows that a large part of the consumers is not visiting very frequently or are one-time visitors, which also explains that there is a lot of competition for customers in a client's industry. The discussions with the client also revealed that some customers may appear more than once in the data, with different codes, which may have a minor impact on the results and the number of one-time visitors. However, there are also numerous amounts of relatively frequent visitors that according to Wu et al. (2020) can be considered as the most valuable ones. These frequent visitors are most likely placed in either loyal customers or top customers' segment. Another monetarily valuable customer segment is big spenders of which customers are not that frequent visitors but spend a high amount of money when visiting. This customer segment behavior can be explained by figure 19, which shows that more than 80 % of big spenders have purchased an industry specific product. Similarly, the figure shows that the top and loyal customers are most likely using the various services provided by the client organization. Defining the behavior of various customer segments can be very useful when planning for example the tailored marketing campaigns.

To make the segmentation results more informative, data is presented by using various visualizations, like heatmaps, bar charts, and maps for presenting geographical data. In addition to visualizations made in Jupyter notebooks, the segmentation results were also imported to Microsoft Power BI that allows creating interactive visualizations, like maps. The study is using geographical data to divide customers by each city. This map is enriched by RFM segmentation results and provides information about what kind of customer segments are in each location and how many customers these segments include. There are also other tools for creating this kind of visual, but Power BI was appropriate for this case because it enables sharing the report and the results across the organization.

The results were provided by using the basic RFM segmentation with quantiles and manually created rules. Study also considered segmentation by using the K-means clustering, which is useful when user does not know how the segments should be divided and what kind of labels there should be. In a situation where more attributes than recency, frequency, and monetary would be used for segmentation, clustering would become a very practical tool. This is why also the clustering algorithm was developed in this research. For example, some typical additional demographic attributes could be age, gender, and income level. Additional geographical attributes could be for example population density of the area or various distance metrics. Some studies also suggest other computed attributes like customer relation length, periodicity, and inter-purchase time to be added to RFM model (Martínez et al., 2021; Peker et al., 2017; Zhou et al., 2021). Creating segmentation rules and labels manually for data with more than three attributes would be time consuming and inaccurate. At this point, using machine learning algorithms would be appropriate.

Correlation matrix presented in figure 18 can be used to describe how the various attributes correlates with each other. Using correlation matrix is also one way to reduce dimensionality by detecting attributes which are appropriate for the analysis. This could be useful when selecting additional attributes for clustering. In the marketing, correlation matrix can also be used determine how valuable various features are in different cases.

The traditional RFM model was chosen as the main segmentation method of the case study because it was possible to obtain moderately accurate and marketing-usable results easily from available data. Even though multiple recent studies, like Christy et al. (2021), Sarvari et al. (2016), Sharaf Addin et al. (2022) and Wu et al. (2020), have used K-means clustering to segment RFM attributes, this study found out that the traditional RFM scoring with only three features tends to be a more accurate and modifiable segmentation method in this situation. In addition, since the three iterations of case project was relatively short timeframe, good results were obtained faster using traditional RFM segmentation.

The identified benefits of consumer segmentation were presented in table 2 in theory chapter 2.3. The first benefit in the table was the ability to understand consumers, their needs, behaviour, like buying habits, and various characteristics (Sarkar et al., 2018; Tsiptsis, 2009). The segmentation model implemented in case study can help client organization to understand their consumers behaviour and buying habits better by defining the various segments. The behaviour of various customers can be explained by using various attributes, like in figure 19. Also, the geographical location of customer can affect their needs. The second benefit in table 2 was targeted marketing, that can be achieved by creating tailored marketing plans for various segments. The other benefits of table 2 like optimal product placement and design, finding latent customer segments, competitive advantage and higher revenue may be realized by actually using the segmentation model. However, even though the results are already promising, the model is in the PoC phase, and the further development is required.

7. CONCLUSIONS

This chapter concludes the results and findings of this research. The summary section examines how the research results and findings corresponded to the research objectives, and how the research questions were answered. The summary is followed by research evaluation that covers the entire study. Lastly, the limitations and the future research possibilities of this study will be reviewed.

7.1 Summary

The objective of this research was to design Proof-Of-Concept (PoC) -type of customer segmentation model for consumer markets and use appropriate data scientific framework during the iterative development process. The segmentation model was designed and developed in collaboration with client organization to meet their business needs. Design science research strategy was used as a higher-level iterative framework during the case study research process. Additionally, CRISP-DM process model was utilized in development process with design science as a data scientific framework. Overall, the PoC -type of segmentation model was successfully implemented and the discussions about the further development and deploying the product into production are also ongoing.

This research had one main research question that summarizes the research objectives into one question, and four sub research questions that divides the main research question into smaller parts that are easier to manage. The research questions were defined in chapter 1.2. The brief answers based on this research are presented below in order starting from first sub research question and ending to main research question that summarizes the whole study.

SRQ1: What kind of consumer segmentation approaches have been identified in existing research and what are the strengths and weaknesses of those approaches?

The first sub research question above was answered in theory chapter 2 by using the existing research. It was recognized that customer segmentation can be implemented in both business and consumer markets, which are inherently different market environments and require different approaches. This study focused more on consumer segmentation. The study identified four main consumer segmentation approaches, which are demographic, behavioural, geographic, and psychographic segmentation.

In demographic segmentation, consumers are divided into segments by their demographic information, such as age, marital status, and gender. This approach can be used to understand customer needs and it is appropriate when marketing decisions are made based on the customer's life situation. The weakness of this approach is that the retail sales organizations rarely keep the customer demographic information in their databases, or the information is out of date. In another approach, behavioural segmentation, consumers are divided into segments based on their buying behaviour. This approach can utilize the transactional sales data, which usually stored in databases with a good availability. The third considered approach is geographic segmentation, that utilizes the geographic data, like country, city, or address. Using this approach can result understanding on how customer needs vary between different areas. In addition, the study also noticed the fact that using only one segmentation approach is not necessarily enough and approaches can also be combined to obtain better results.

SRQ2: What are the most suitable data scientific frameworks for developing consumer segmentation model iteratively in collaboration with the client?

The second sub research question was answered in subchapter 3.1, and it was for finding appropriate iterative data scientific framework that can be used as a supportive guideline in design and development process of segmentation model. Multiple frameworks were considered but the CRISP-DM process model was selected for this research. It is the de-facto standard process and the most widely used framework in data mining and data science. CRISP-DM process model is also applicable to design science research methodology which was used as a research strategy along with the case study in this research. The business and data understanding phases of CRISP-DM were used in case research problem identification and objective definition. Data preparation and modelling phases were utilized in DSRM design and development stage. Even though the CRISP-DM model is widely used and appropriate framework in customer segmentation development process, the study did not consider whether the CRISP-DM is the absolute best framework to implement the customer segmentation model. However, since the research was done as an independent work, the CRISP-DM framework was sufficient. Other good alternatives could have been, for example, several other frameworks derived from the KDD process model, or frameworks based on agile development, which could be more appropriate when the development process is done in teams.

SRQ3: What kind of analytical tools and methods can be used in segmentation model when analysing consumer data?

The third sub research question focused on investigating analytical tools and methods for consumer segmentation. Found tools and methods are defined in subchapter 3.2. According to findings, one of the most important steps in data science project is data preprocessing, which was presented in subchapter 3.2.1. This study utilized multiple preprocessing methods such as data filtering, outlier detection scaling and merging. The base segmentation method used in this study was RFM modelling, which is presented in subchapter 3.2.2. The method stands for segmenting consumers by their buying behaviour using recency, frequency, and monetary features. The selection was made based on the availability and the quality of client organization's transactional data. Another utilized method was K-means clustering, that can be used to divide customers into segments without prior knowledge about the data. Overall, the study found that various analytical tools and methods can be used for solving different consumer segmentation related problems, and the most suitable ones depends on the use case. The clustering would be an appropriate solution when segmentation is based on multiple different attributes, like in demographic or mixed approach segmentation.

SRQ4: How can the identified frameworks and methods be practically applied to a real-world case study of consumer segmentation?

The fourth sub research question was for applying the frameworks and methods practically, which is presented in case study chapter 5. In this research, consumer segmentation model was designed and developed iteratively by utilizing CRISP-DM model and design science frameworks. Additionally, basic kanban board was used for prioritizing the individual development work. The PoC customer segmentation model used B2C consumer sales data, so the behavioural segmentation was a clear choice as the main approach in case study. RFM modelling and clustering were implemented for segmenting customers by their purchasing behaviour. The results of segmentation were also visualized by using the python libraries and MS Power BI reporting tool. The various steps of the model need to be aligned together as a data flow from source data to results. In addition to understanding the various methods, applying them requires also base understanding of client organization's business and the data that is used in model. It is also essential to have permissions to client organization's database or dataset. These requirements could not be achieved without sufficient planning and good mutual communication with client organization.

MRQ: What kind of segmentation model is appropriate for analysing quantitative sales data in retailer consumer markets?

The four sub research questions were used for answering the most important subtopics of the research problem. Overall, the study managed to answer the research questions well and the objectives of research were fulfilled. The behavioural segmentation approach and the RFM modelling were recognized as an appropriate basis for the segmentation model, which was later enriched by using explanatory geographical and product related attributes. The study found that the behavioural segmentation and RFM analysis is an appropriate starting point in developing segmentation model when the availability and quality of transactional data is good. The RFM model is relatively easy to understand and the rules for segment creation can be tailored for organization's needs. In addition, the new features are easily applicable to the RFM based model.

For the development part of segmentation model, following DSRM and CRISP-DM enabled the efficient iterative development process, considering all the relevant steps including domain understanding, data processing, modelling, evaluation, and communication. In addition to these, the case study also focused on data presentation, which supports the efficient testing, evaluation, and usage of the model.

7.2 Research evaluation

The structure of the research was divided into two main parts. The case study was for implementing practical segmentation model that can be used for testing and evaluation of the solution regarding the business needs of the client organization. The theory part in chapters 2 and 3 was for finding appropriate methods that can be used in case study. The research emphasized a pragmatistic philosophy, where the main purpose was to implement an artifact that creates value in practical business environment. The development process of consumer segmentation model went back and forth between theory and empirical case study.

Design science research and case study were used as a research strategy in this study. The success of DSR is evaluated in this section by using Hevner et al. (2004) design science research guidelines and the case study is evaluated after that by using the Yin's (2018) case study criteria. The DSR research guidelines and evaluations are presented in table 8 below:

Table 8: *Design science research guidelines and evaluation.*

Guideline	Evaluation
Design as an artifact	The research produced a consumer segmentation model as an artifact. The model used various methods and algorithms that were appropriate in solving the business problem.
Problem relevance	The objective of this research was to develop a practical PoC segmentation model for relevant business needs. The developed model can already be utilized by the client, and the further development is considered.
Design evaluation	The utility, quality, and efficacy of design artifact has been demonstrated by presenting and discussing the case research results after each iteration.
Research contributions	The research identified how various methods can be utilized in segmentation in consumer markets, and how those methods can be applied in practice. The research also recognized multiple future research possibilities for client organization and research.
Research rigor	The research followed rigorously the guidelines of DSRM and CRISP-DM process models when developing the artifact. Various methods like feature engineering, RFM modelling and K-means clustering were utilized step-by-step.
Design as a search process	The business environment and available resources were essential throughout the research process because they defined the direction of the research.
Communication of research	The results of the research were constantly communicated with the client organization and University. During the research, there were also several discussions with experts in the field of data science. In addition, the results are published in a form of master's thesis report.

The first design science guideline according to Hevner et al. (2004) is that the research needs to produce the artifact that solve the important organizational problem. They point out that the artifact can be for example a construct, a model, a method, or an instantiation, but it rarely is a fully developed ready information system. In this study the artifact is PoC -type of segmentation model with a lot of further development opportunities.

The second design science guideline is the relevance of the research problem. In design science research the problem needs to be relevant for the organization that will be utilizing the artifact. (Hevner et al., 2004.) The research problem of this study is in line with the objectives of the client organization, and they can utilize it when planning the development process of their marketing processes.

The third design science guideline is the design evaluation, which refers that the artifact needs to be tested and evaluated in the business environment with well-executed evaluation methods. (Hevner et al., 2004.) This research used a case study, in which the artifact was demonstrated and evaluated after each iteration with the real business data. Also, the various use cases of model were discussed during the evaluation sessions. However, consistent evaluation techniques were not used during the case research, but the discussions during the sessions were open and conductive to research. The model was also tested and evaluated by using the interactive report as a testing environment.

The research contributions are presented as a fourth guideline of design science research. Hevner et al. (2004) states that designed artifact should fulfil at least of the three contributions, which are the design artifact, foundations, and methodologies. The segmentation model was developed by combining various methods and approaches from existing research and applied to the client organization’s business environment. The model also provided multiple further research and development opportunities for the client organization, and for marketing and data science.

The fifth step, research rigor, means that the artifact should be constructed and evaluated by following strictly the appropriate methodologies and guidelines from knowledge base (Hevner et al., 2004). In this study, the rigor was derived by rigorously following the various methodologies, frameworks, and methods that are presented in the theory section of this research.

The sixth guideline of design science is design as a search process. This means that to find an effective artifact for the business problem, the study needs to utilize the available means and satisfy the laws of the problem environment (Hevner et al., 2004). In the designed model, the used methods and approaches were selected based on the business environment and possibilities of the available data. Additionally, as Goyat (2011) states, each company and industry have different type of requirements when it comes to customer segmentation model designing, the model developed in this research is also using some industry specific attributes.

The last design science guideline is communication of research, which means that the study needs to be presented to both technology- and management-oriented audiences (Hevner et al., 2004). Experts from both audience groups were involved in the research process and evaluation of the results. The process and the results were also discussed with the thesis supervisors during the research. In addition, the results are published in the form of a master's thesis report.

In addition to design science research, this study also included the case study. Yin (2018) mentions several criteria that can be used in evaluation of case study research. To measure the quality of research design in case study, Yin recommends four tests that are commonly used in social science research: construct validity, internal validity, external validity, and reliability. In addition, Yin mentions that research should also be generalizable and versatile, research questions and hypothesis should be well defined, and data collection and analyzes should be justified. (Yin, 2018.)

The construct validity refers that how well the concepts of the research are validated and corresponds with reality (Yin, 2018, pp. 41-46). In this research, the formed customer behavior-based segments are created by using the real-world sales data. The RFM segmentation rules and codes used in research are defined in case study and the results are tested, discussed, and evaluated with the client organization's representatives. The internal validity refers that how well the research can establish casual relationships between attributes (Yin, 2018, pp. 41-46). This study found that there are clear relationships between various attributes that can explain customer behavior in retail business. For example, correlation matrix in figure 18 and bar charts in figure 19 were used to demonstrate how different segments and attributes affect each other. External validity refers to research ability to generalize findings with other contexts. In a single-case study, this can be achieved by linking the case study with the theory. (Yin, 2018, pp. 41-46.) In this study, the used methods are created by utilizing existing research and the model could be mostly applied also in other contexts with similar behavioral data. The case study results are linked to the theory in several parts of the research, but especially in the discussion chapter. The fourth criterion is reliability, which refers that the various methods used in the research are reliable and can be repeated with the similar results (Yin, 2018, pp. 41-46). This study is using widely tested methods such as RFM modelling and K-means clustering. In addition, the model has been tested multiple times during the research process.

According to the evaluation criterions and descriptions above, the research successfully fulfilled the guidelines of design science research and the case study. However, as the magnitude of the research and resources were relatively small, all the possible solutions and approaches were not considered. The study still managed to achieve the objectives and appropriate results for the client organization and the science. The PoC segmentation model allows the investigation of even more detailed customer segmentation models for the various business problems related to consumer marketing.

7.3 Research limitations and future research possibilities

This research has a several limitations that may affect the results. Firstly, the study utilized only the behavioral approach when creating customer segments. Using additional demographic, geographic, or psychographic attributes would make even better segmentation results. Demographic and psychographic data could have been collected by using various methods, but also the time and resources of this research were limited. If the additional attributes are included in the segmentation model in the future, the data preparation and clustering processes should be reviewed. There were also some minor issues with the data, since even after the data preparation, the data included some outliers. Consumer sales data also included some business customers that were hard to distinguish from consumer customers.

The PoC segmentation model in case study included customers only from Finland, but in the future, also other countries could be considered. This would make geographic segmentation more relevant. Another future development step for the segmentation model is to build a full data pipeline from database to Power BI report. Pipeline would include all the relevant steps and it would update the report automatically to keep the results current.

As future research, it would be interesting to study how the customer segmentation could be effectively utilized in retail business environment and how the segmentation affects and creates value in various areas of business. This kind of would require longer timeframe measurements and interviews. Another future research ideas would be possibilities of artificial intelligence in consumer market segmentation, building model for analyzing customer lifetime value (CLV), or analyzing segmentation overall from consumer point of view by considering for example ethical aspects of customer segmentation.

REFERENCES

Arnold, J. (2022). Learning Microsoft Power BI. O'Reilly Media, Inc.

Arunachalam, D., & Kumar, N. (2018). Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making. Expert Systems with Applications, 111, 11–34. https://doi.org/10.1016/j.eswa.2018.03.007

Baig, M. R., Govindan, G., & Shrimali, V. R. (2021). Data Science for Marketing Analytics. Packt Publishing, Limited.

Blanchard, T., Behera, D., & Bhatnagar, P. (2019). Data Science for Marketing Analytics. Packt Publishing, Limited.

Cao, L. (2018). Data Science. ACM Computing Surveys, 50(3), 1–42. https://doi.org/10.1145/3076253

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. Journal of King Saud University. Computer and Information Sciences, 33(10), 1251–1257. https://doi.org/10.1016/j.jksuci.2018.09.004

Collica, R. S. (2011). Customer segmentation and clustering using SAS Enterprise Miner (Second edition). SAS.

Collier, K., Carey, B., Grusy, E., Marjaniemi, C., & Sautter, D. (1998). A Perspective on Data Mining.

Cooil, B., Aksoy, L., & Keiningham, T. L. (2008). Approaches to Customer Segmentation. Journal of Relationship Marketing (Binghamton, N.Y.), 6(3–4), 9–39. https://doi.org/10.1300/J366v06n03_02

El Sheikh, A. A. Rahman., & Alnoukari, M. (2012). Business intelligence and agile methodologies for knowledge-based organizations: cross-disciplinary applications. Business Science Reference.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. The AI Magazine, 17(3), 37–54.

Galli, S. (2022). Python feature engineering cookbook (Second edition). Packt Publishing, Limited.

Goyat, S. (2011). The basis of market segmentation: a critical review of literature. In European Journal of Business and Management www.iiste.org ISSN (Vol. 3, Issue 9).

Hammarberg, M., & Sundén, J. (2014). Kanban in action (First edition). Manning Publications.

Haroon, D. (2017). Python Machine Learning Case Studies. Apress L. P.

Heldt, R., Silveira, C. S., & Luce, F. B. (2021). Predicting customer value per product: From RFM to RFM/P. Journal of Business Research, 127, 444–453. https://doi.org/10.1016/j.jbusres.2019.05.001

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kotler, P. (2019). Marketing management (Fourth European edition). Pearson.

Linstedt, D. (2016). Building a scalable data warehouse with data vault 2.0 (First edition). Morgan Kaufmann.

Martínez, R. G., Carrasco, R. A., Sanchez-Figueroa, C., & Gavilan, D. (2021). An RFM Model Customizable to Product Catalogues and Marketing Criteria Using Fuzzy Linguistic Models: Case Study of a Retail Business. Mathematics (Basel), 9(16), 1836. https://doi.org/10.3390/math9161836

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Miglautsch, J. R. (2000). Thoughts on RFM scoring. Journal of Database Marketing & Customer Strategy Management, 8(1), 67–72. https://doi.org/10.1057/palgrave.jdm.3240019

Nguyen, S. P. (2021). Deep customer segmentation with applications to a Vietnamese supermarkets' data. Soft Computing (Berlin, Germany), 25(12), 7785–7793. https://doi.org/10.1007/s00500-021-05796-0

NumPy. (2023). NumPy user guide. https://numpy.org/doc/stable/user/index.html

Offermann, P., Levina, O., Schönherr, M., & Bub, U. (2009). Outline of a design science research process. Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, DESRIST '09. https://doi.org/10.1145/1555619.1555629

Pandas. (2023). User Guide. https://pandas.pydata.org/docs/user_guide/index.html

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. Marketing Intelligence & Planning, 35(4), 544–559. https://doi.org/10.1108/MIP-11-2016-0210

Rejeb, A., Rejeb, K., & Keogh, J. G. (2020). Potential of big data for marketing: A literature review. Management Research and Practice, 12(3), 60–73.

Sarkar, D., Bali, R., & Sharma, T. (2018). Practical Machine Learning with Python A Problem-Solver’s Guide to Building Real-World Intelligent Systems (First edition). Apress. https://doi.org/10.1007/978-1-4842-3207-1

Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes, 45(7), 1129–1157. https://doi.org/10.1108/K-07-2015-0180

Saunders, M., Lewis, P., & Thornhill, A. (2019). Research Methods for Business Students. Pearson Education, Limited.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. International conference on enterprise information systems / International conference on project management / International conference on health and social care information systems and technologies 2020 (centeris/projman/hcist 2020), 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Scikit-learn. (2023). User Guide. https://scikit-learn.org/stable/user_guide.html

Sharaf Addin, E. H., Admodisastro, N., Mohd Ashri, S. N. S., Kamaruddin, A., & Chong, Y. C. (2022). Customer Mobile Behavioral Segmentation and Analysis in Telecom Using Machine Learning. Applied Artificial Intelligence, 36(1). https://doi.org/10.1080/08839514.2021.2009223

Sokol, O., & Holý, V. (2021). The role of shopping mission in retail customer segmentation. International Journal of Market Research, 63(4), 454–470. https://doi.org/10.1177/1470785320921011

Stormi, K., Lindholm, A., Laine, T., & Korhonen, T. (2020). RFM customer analysis for product-oriented services and service business development: an interventionist case study of two machinery manufacturers. Journal of Management and Governance, 24(3), 623–653. https://doi.org/10.1007/s10997-018-9447-3

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. Sustainability (Basel, Switzerland), 14(12), 7243. https://doi.org/10.3390/su14127243

Tsiptsis, Konstantinos. (2009). Data mining techniques in CRM inside customer segmentation (First edition). Wiley.

Vleugel, A., Spruit, M., & van Daal, A. (2010). Historical Data Analysis through Data Mining from an Outsourcing Perspective. International Journal of Business Intelligence Research, 1(3), 42–65. https://doi.org/10.4018/jbir.2010070104

Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. Mathematical Problems in Engineering, 2020, 1–7. https://doi.org/10.1155/2020/8884227

Yin, R. K. (2018). Case Study Research and Applications: Design and Methods (Sixth edition). SAGE Publications, Inc.

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists (First edition). O’Reilly.

Zhou, J., Wei, J., & Xu, B. (2021). Customer segmentation by web content mining. Journal of Retailing and Consumer Services, 61, 102588. https://doi.org/10.1016/j.jretconser.2021.102588