Customer segmentation in Retail: An Experiment in Sweden

Nagma Athar Memon

Department of Computer and Systems Sciences

Degree project 30 credits
Computer and Systems Sciences
Degree project at the master level
Spring term 2022
Supervisor: Jaakko Hollmén
Co-supervisor: Alejandro Kuratomi Hernandez
External supervisor: Sten Hallström, Extenda Retail

Stockholm University

Abstract

The retail industry is continuously transforming due to digitalization, globalization, urbanization, and automatization. These factors contribute to new and more transparent retail with new customer behaviors, business models, and international competition. Literature mentions that the customers have more power to influence the retail industry today. Retailers must listen carefully to what their customers say to stay relevant and adopt data-driven strategies to meet evolving customer values. Therefore, it is essential to know customer behavior and value so that the company can target the groups of customers and thus lead their measures toward selected segments. There is a diversified growth in the tastes and likes of customers. To survive in competitive industries, a company needs to be more concerned with Customer Relationship Management. A central problem in Customer Relationship Management (CRM) is clustering customers into meaningful segments. This challenge is typical in retail, where various products are available. When it comes to segmenting customers, segmentation models provide purchasing patterns of customers, whereas clustering is the segmentation of data in several applications by grouping extensive data into groups with similar patterns. This experimental study focuses on customer segmentation, a typical customer analytics technique, and a traditional concept in marketing. The thesis aimed to perform an experiment by efficiently segmenting Swedish pharmacy retail dataset by combining segmentation models and clustering algorithms. Finally, the determined customer segments were profiled and named using a customer profiling technique. To support this aim, the research questions are “Which combination of mentioned segmentation models and clustering algorithms performs best for segmentation of a Swedish pharmacy retail company’s customers?” and “How can extracted customer segments be profiled?” This study was conducted with a Swedish retail software providing company, Extenda Retail, to gain better insight into customer behavior by analyzing experimental dataset of customer transactions of a Swedish pharmacy chain. K-means, Agglomerative and Mean-Shift clustering algorithms were paired with the segmentation models RFM, LRFM and LRFMP to generate the customer segments. This experiment showed that both K-means and Agglomerative clustering algorithms with the LRFMP model are the most suitable solution for customer segmentation for the dataset used in this study. Both combinations: the LRFMP with K-means and the LRFMP with Agglomerative clustering generated two customer segments which are then profiled as: "low-contribution customers" and "high-contributing loyal customers."

Keywords: Customer Segmentation, Customer Relationship Management, Clustering, RFM, LRFMP.

Synopsis

BACKGROUND	The emergence of big data brings a new wave of Customer Relationship Management (CRM)'s strategy in supporting the personalization and customization of sales, services, and customer services. There is a diversified growth in the tastes and likes of customers; hence organizations cannot satisfy every customer fully. To survive in competitive industries, a company needs to be more concerned with Customer Relationship Management. A central problem in Customer Relationship Management (CRM) is to cluster customers into meaningful customer segmentation groups. This challenge is typical in retail, where various products are available.
PROBLEM	Customers have more power to influence the retail industry today. Many studies have used different RFM models and the K-means algorithms to assess the value of customers. However, segmentation in retail using the LRFMP model and applying different clustering algorithms rather than just the K-means have remained sparse so far. Besides, to the best of our knowledge, no previous work in customer segmentation has ever been brought together with the LRFMP model in Swedish pharmacy retail. Thus, the problem addressed is " Which combination of segmentation model and a clustering algorithm performs best for customer segmentation in a Swedish pharmacy retail company dataset?"
RESEARCH QUESTION	This study aims to perform an experiment of segmentation models and clustering algorithms for the Swedish pharmacy retail sector, specifically a pharmacy chain company in Sweden. The clustering algorithms to be executed are K-means, Agglomerative, and Mean-shift clustering algorithms, and they will be paired with the segmentation models RFM, LRFM and LRFMP. To support this aim, the main research question is "Which combination of mentioned segmentation models and clustering algorithms performs best for segmentation of a Swedish pharmacy retail company's customers?" and "How can extracted customer segments be profiled?"
METHOD	This study wanted to gain insights into the behavior of pharmacy company customer by testing segmentation methods. For that reason, an experiment was chosen as the research strategy. The collection of quantitative secondary data (organizational records) is deemed more appropriate to get more insight into customers purchasing behavior. The data was analyzed in Google's BigQuery and Colab using various Python machine learning libraries. A model that offers more information about customer purchasing behavior is the most suitable model to gain more profound and valuable insights into customer purchasing behavior for the Swedish pharmacy dataset. A validation index called Silhouette score was used to measure the performance of clustering algorithms.
RESULT	This thesis showed that both K-means and Agglomerative clustering algorithms with the LRFMP model is the most suitable solution for customer segmentation for the dataset used in this study. Both combinations: the LRFMP with K-means and the LRFMP with Agglomerative clustering generated two customer segments which are then profiled as: "low-contribution customers" and "high-contributing loyal customers."
DISCUSSION	This experimental study delivers essential insights into the customer segmentation in a pharmacy retail company in Sweden. It presents relevant consumer purchasing patterns and behaviors. Additionally, this experiment extends the application of the LRFMP model and clustering algorithms in the B2C setting and the pharmacy retail in Sweden. Retail companies in Sweden can benefit from this experimental study to identify different customer segments and profile them. A retail company can revise its CRM and marketing actions with such a benefit. Thus, services and customer relationships can be enhanced by the company.

Acknowledgement

Throughout the writing of this thesis, I have received good support and assistance. Without this support, the journey to finishing this research would have been more difficult. I would like to sincerely thank and express my gratitude to my supervisor, Jaakko Hollmén, who has supervised my work on this thesis. I would like to thank my co-supervisor, Alejandro Kuratomi Hernandez, for valuable advice, assistance, and guidance throughout this project. His insightful feedback was very valuable and gave a proper direction to my study. Throughout this project, I have received many good suggestions and tricky questions to answer. With his guidance, the project proceeded smoothly. I would also like to thank my reviewer, Petter Karlström, for his feedback.

I would like to acknowledge my host company "Extenda Retail" for allowing me to write a thesis. I am specifically thankful to Sten Hallström, my supervisor at the host company, for first of all taking me in for this thesis, but subsequently for taking time out of busy schedule to mentor and steer me on this journey. Furthermore, for providing me with appropriate resources and guidance. The support and environment at Extenda Retail made this thesis possible, and I am grateful for that.

Additionally, I would like to thank my family for their wise support and for showing faith in me. A special thank you to my husband Mohsin, who has listened to my complaints, and fears with endless support. When things were overwhelming, you reminded me of what is truly important and inspired me to push on.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CRM - Customer Relationship Management
RFM - Recency Frequency Monetary
LRFM – Length Recency Frequency Monetary
LRFMP – Length Recency Frequency Monetary Periodicity

1 Introduction

Companies in different industries face the challenge of efficiently directing marketing efforts to the appropriate clients to offer the right products. Companies are puzzled about which customer segments to target to sell their products (Kansal et al., 2018). To increase their revenue and competitiveness, a tool known as Customer Relationship Management (CRM) provides a system and strategy supporting this marketing process. CRM is the approach of acquiring and collaborating with chosen customers to make premium value for the company and its customers (Soltani and Nivimpour, 2016).

According to one study, there is a diversified growth in the tastes and likes of customers; hence organizations cannot satisfy every customer fully (Peker et al., 2017). This challenge is typical in retail, where various products are available. The relationship between companies and customers has evolved into an undeniable part of businesses, and hence, a mechanism to manage this relation is crucial (Tavakoli et al., 2018). CRM can be an effective way to enhance customer services and build healthy customer relationships (Peker et al., 2017).

CRM system plays a vital role in supporting business strategies and building long-term customer interactions. CRM system facilitates companies to analyze all customer information, sales, marketing strategies, and market trends using technology and helps gain new insight related to customer behavior and value (Pramono et al., 2019). The information about the customers can be utilized in various ways such as designing promotions, product recommendation to the customers, marketing of a new product, and so on (Sokol et al., 2021). Nguyen stated that a central problem in CRM is to cluster customers into meaningful groups that are segmentation of customers (2021).

According to one research, “customer segmentation enables organizations to divide customers into different externally heterogeneous and internally homogeneous groups and interact with each customer segment” (Peker et al., 2017). Segmentation of customers directly connects with customer satisfaction of the companies (Ozan & Iheme, 2019). Customization according to customer preferences allows companies to provide offers that are best suited to the customer and increase customer satisfaction. Many business entities differentiate their customers by members and non-members. Also, many enterprises provide different service levels for different classes of customers (Soudagar, 2012). Therefore, a CRM system should be able to segment customers effectively and optimize marketing programs better, satisfy customers, and increase profits. Customer segmentation maximizes customer satisfaction and hence improves the company’s profit significantly.

When it comes to segmenting customers, segmentation models & clustering has been an effective approach (Kansal et al., 2018; Wedel & Kamakura, 2000 cited in Peker et al., 2017). Segmentation models like Recency, Frequency, and Monetary (RFM) divides a company's customers into groups that have similarities between customers in each group, with respect to customer’s purchasing pattern and behavior. Recency is the time interval from the last purchase which indicates the customer’s purchasing potential. The frequency is the number of transactions in a period, whereas the monetary is the total expended amount in every transaction during a period (Peker et al., 2017). According to Tavakoli et al., RFM analysis is

the most common method used for customer segmentation. The RFM analysis is a behavioral-based data mining technique that extracts customers' profiles using their Recency, Frequency, and Monetary values (Tavakoli et al., 2018).

Clustering is the segmentation of data in several applications by grouping extensive data into groups that have similarities (Pramono et al., 2019). With the rapid rise of data science and machine learning, there have been several algorithms in these two fields to be employed in CRM, and specifically for customer segmentation (Nguyen, 2021). In past studies, clustering algorithms and segmentation models have been widely used to solve customer segmentation (Sun et al., 2021).

A group (or segment) is viewed as a collection of individuals who share a common interest or goal. Like individual profiles, group profiles aim to provide a clear picture of the group's personality, emotions, behavior, and effectiveness as a segment. To fully understand the segment, what they are capable of, how effective they are, and how much of a threat they pose, group profiling can be performed. Therefore, after obtaining the customer segments, it is also essential to profile customers in segments to know the specific customer behavior in the respected segments. One research suggested a group profiling technique (Ha and Park, 1998). This group profiling technique was employed in various studies with different RFM models (Chang and Tsay, 2004; Peker et al., 2017).

This study focuses on customer segmentation, a typical customer analytics technique, and a traditional concept in marketing (Griva et al., 2021). This study is conducted with a Swedish retail software company, Extenda Retail, with the aim of gaining better insight into customer behavior. With the advancement in customer segmentation, the company can uncover different customer behaviors related to purchases to find appropriate marketing strategies that are susceptible to target specific groups.

1.1 The Company

Extenda Retail AB is a Swedish IT company providing retail software solutions. Extenda Retail was established during the fall of 2018, but its roots are way deeper. The new company is the result of the merger of Visma Retail and Extenda. Two leading retail technology firms with almost 40 years each in retail, together had prerequisites to become one united software provider to leading retailers. Extenda Retail has multiple retail software solution such as POS & Checkout Services, Customer Engagement, Retail Cloud Solution, Warehouse Management, Retail ERP, CRM platform, etc.

Extenda Retail AB has a CRM platform called Relevate, designed to build customer loyalty by personalization and data-driven insights. With this product, Extenda Retail AB enables retailers to offer their customers a unique and tailored shopping experience daily. They want to have a tool to optimize customer segmentation and construction of promotional campaigns for retailers. Extenda Retail aims to do advanced segmentation; however, they did not disclose current methods they are utilizing for customer segmentation in their product.

1.2 Research Problem

With the purpose of improving customer segmentation, Extenda Retail can effectively satisfy customer needs and preferences by segmenting customers effectively (Pramono et al., 2019). Providing personalized services and a successful tailored marketing strategy is essential for companies to improve customer satisfaction and long-term customer loyalty (Peker et al., 2017).

Retail is one of Sweden's largest and most important industries. Today, the retail industry in Sweden is continuously transforming due to digitalization, globalization, urbanization, and automatization within the industry (Swedish Trade Federation, 2019a cited in Sieradzki & Sollbe, 2020). These factors contribute to new and more transparent retail with new customer behaviors, new business models, and international competition. An essential structural change in the ongoing transformation is the fact that e-commerce has been taking sales shares from the traditional physical retailing for a long period of time (Swedish Trade Federation, 2019b cited in Sieradzki & Sollbe, 2020). They found out that the customers have more power to influence the retail industry. According to their study, retailers must listen carefully to what their customers say to stay relevant and adopt data-driven strategies to meet evolving customer values (Sieradzki & Sollbe, 2020). Therefore, it is essential to know customer preferences and likes so that decision-makers can target the customers' groups and execute their actions toward specified segments. Customer segmentation allows firms to understand customers' behavior and preferences.

As mentioned in the introduction section, segmentation models like RFM and clustering algorithms are effectively used for customer segmentation. RFM values provide purchasing patterns & frequency of a customer and customer's contribution to the company's revenue. Based on the RFM variable values, cluster analysis can be performed to cluster the customers into distinct groups using clustering algorithms. Hence, RFM model variables and clustering have been used to segment customers by many studies (Parvaneh et al., 2014; Peker et al., 2017; Kansal et al., 2018;).

RFM model has evolved during the past two decades (Peker et al.,2017). Previous studies have developed new RFM models by changing the model's variables such as GRFM model, LRFM model, LRFMP model, RFMC model (Chang et al., 2011; Chang et al., 2004; Parvaneh et al., 2014; Huang et al., 2020). The LRFM model with a variable Length L as days between first and last purchase has received attention in multiple studies (Hosseini et al., 2010; Kao et al., 2011; Li et al., 2011; Wei et al., 2012 cited in Peker et al.,2017).

One recent study introduced the periodicity of customer visits (P) into the original LRFM model to measure the regularity of customers (Peker et al., 2017). The study successfully performed segmentation and provided valuable insights about different customer behaviors in the Turkish grocery industry. Their study successfully identified different customer profiles, which gave valuable insights into different customer profiles. Peker et al. suggested applying their proposed model for the data from different countries and domains (Peker et al., 2017).

In 2019, Sheikh et al. applied the LRFMP model proposed by Peker et al. with the K-means algorithm in the Iranian Fintech Industry (Sheikh et al., 2019). Many studies have used different RFM models and the K-means algorithms to assess the value of customers. However, segmentation in retail using the LRFMP model and applying different clustering

algorithms rather than just the K-means have remained sparse so far (Parvaneh et al., 2014; Peker et al., 2017; Kansal et al., 2018; Sheshasaayee et al., 2018). There is a significant knowledge gap in customer segmentation in pharmacy retail in Sweden. Besides, to the best of our knowledge, no previous work in customer segmentation has ever been brought together with the LRFMP model in Swedish pharmacy retail.

Thus, the purpose of this thesis is to do an experiment for efficiently segmenting Swedish pharmacy retail customers by combining segmentation models and clustering algorithms. Data used in this study comes from a client company of Extenda Retail, a well-known pharmacy chain company operating in Sweden. This study implements customer segmentation with an extended RFM model: length, recency, frequency, monetary, and periodicity (LRFMP) for the pharmacy dataset as well as other segmentation models such as RFM and LRFM and compare them. After getting values for model variables for each customer, this experimental study would choose a clustering algorithm that performs well to group customers. As this study compares different algorithms, a well- known validation index called Silhouette index is used to evaluate the performance of algorithms with varying parameters. Silhouette index is used for final clustering validation and evaluation to measure the general quality of clustering (Dudek, 2020).

1.3 Research question

This study aims to perform an empirical evaluation of clustering algorithms for the pharmacy retail in Sweden. The pharmacy retail dataset from a Swedish pharmacy chain, a client of the company in this study, was analysed by combining the L, R, F, M, and P attributes and clustering algorithms. The clustering algorithms to be executed are K-means, Agglomerative, and Mean-shift clustering algorithms, and they will be paired with the segmentation models RFM, LRFM and LRFMP. To measure the performance of the clustering algorithms and evaluate the general quality of clustering, a validation index called the Silhouette score was used. To support this aim, the following research question is proposed.

- Which combination of the above-mentioned segmentation model and a clustering algorithm performs best for segmentation of a Swedish pharmacy retail company's customer?
- How can extracted customer segments be profiled?

1.4 Thesis Structure

The study first examines prior studies in the research literature for customer segmentation in the retail industry (Chapter 2). Then it presents the research strategy, data collection, and analysis method used (Chapter 3). Next, it describes and analyzes the findings from the study to address the research question (Chapter 4). Finally, it concludes with a discussion of the thesis work in terms of its originality and significance, limitations, and future research (Chapter 5).

2 Extended Background

2.1 Customer Relationship Management

In the 1990s, in the business domain, the concept of CRM gradually emerged. From the beginning the CRM succeeded, achieved prominence as an area of academic exploration, and encouraged the global business and research community. (Soltani and Nivimpour, 2016). According to Ling and Yen, CRM is a set of methods to help a business build profitable customer relationships (Ling and Yeng, 2001). CRM is a tool and strategy for managing customer interactions using technology to automate business processes. CRM consists of sales, marketing, and customer service activities (Anshari et al., 2018). The front-office parts facilitate the flow of information with customers. Organizations that implement CRM aim to enable the seamless dissemination of customer knowledge throughout the organization. The back-office parts help with data mining and thus identify and analyze customers' needs and actions (Soltani and Navimipour, 2016).

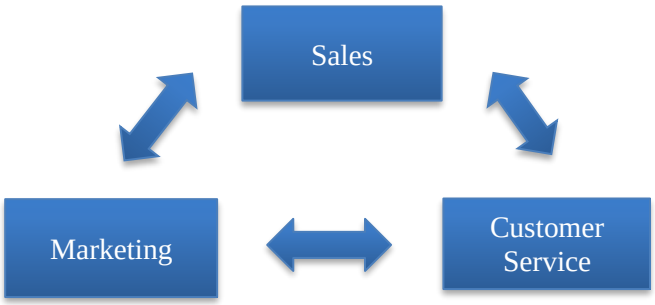

Figure 1 CRM scope & module (M. Anshari et al., 2019)

2.2 Customer Segmentation

Customer segmentation divides a company's customers into groups that have similarities between customers in each group. Companies can focus on valuable customer clusters and execute measures toward the individual clusters. One advantage of customer segmentation is to comprehend customers' behavior and preferences. By implementing customer segmentation, companies can utilize information about different customer behavior and preferences (Dibb, 1998).

Hence, customer segmentation is used to determine customer potential (Peker et al., 2017). Researchers have segmented customers in retail using different datasets reflecting either customer characteristics or customer behavior. Figure 1 shows how researchers used different data for customer segmentation (Griva et al., 2021). With plenty of data sources, including social networks, data-driven marketing, a major driving force behind customer segmentation, is becoming increasingly important (Nguyen, 2021). RFM segmentation attributes are well accomplished characteristics for customer segmentation (Peker et al., 2017).

Customer Characteristics	Customer Behavior
Demographics	Content of a basket
Geographic	Basket characteristics
Psychographic	Visit characteristics
Attitudinal	

Table 1: Customer Segmentation data in Retail (Griva et al., 2021)

2.3 RFM & LRFMP model

RFM is worked out by Arthur Hughes of the US Database Marketing laboratory to analyze and predict customer behavior (Hughes, 1998; Zhou et al., 2011; Peker et al., 2017). RFM analysis depicts and separates customers by three behavior variables, i.e., Recency (R), Frequency(F), and Monetary(M) (Peker et al., 2017). If the recency is low, the likelihood of repurchasing is high. If the frequency is higher, it indicates higher loyalty of customers to the company. If the monetary variable is high, it indicates the customer's trust company that they will take preference to respond to the production and service of the company (Zhou et al., 2011; Peker et al., 2017).

RFM model has evolved during the past two decades (Peker et al.,2017). RFM models have several advantages. For instance, they are easy to implement, and management can easily interpret the results of models (Peker et al., 2017). Hence, RFM analysis is commonly utilized in customer segmentation. Previous studies have developed new RFM models by changing the model's variables. One similar study has extended the RFM model to the GRFM model by adding product category group information (Chang et al., 2011).

Chang et al. developed the LRFM model by including a new feature L, customer relation length, into the RFM model (Chang et al., 2004). Another variant is the LRFMP model introduced by Peker et al. with P as the periodicity of customer visits (P) into the original LRFM model to measure the regularity of customers. According to the LRFM model, customers with similar profiles may have different visiting patterns. Thus, Peker et al. introduced the periodicity of customers' visits (P) into the original LRFM model to describe customers' behaviors and regularity. The definitions of the LRFMP model variables (Peker et al., 2017) are shown in Table 2.

In 2019, Sheikh et al. applied the LRFMP model proposed by Peker et al. in the Iranian Fintech Industry. They implemented two-stage customer segmentation with the LRFMP model and the K-means algorithm. They examined behavior of the business customers in Iran’s financial technology industry (Sheikh et al., 2019). There is currently no application of the LRFMP model in other domains and countries, so we aim to apply it to the Swedish pharmacy retail industry and benchmark it against previous methodologies for customer segmentation. Peker et al. described Periodicity as “the standard deviation of the customer’s inter-visit times” (Peker et al., 2017). Periodicity demonstrates customer's purchase trends. A lower periodicity value of the customers indicates that they can be considered regular customers as they make transactions at relatively fixed intervals (Peker et al., 2017).

Parameter	Definition	Parameter Explanation
Length	Length of relationship between customer and the company	Interval between the first transaction and the last transaction
Recency	Last purchase date in an observation period	Interval between the last transaction and the end of the observation period
Frequency	Number of purchases in a particular period	Number of transactions which occurred in the observation period
Monetary	Value of purchases in a particular period	Average amount of money spent during the observation period.
Periodicity	Regularity of customers	The average interval between days which transaction(s) occurred

Table 2: Definition of LRFMP model parameters

2.4 Clustering

Clustering is a successful method in customer segmentation (Kansal et al., 2018; Peker et al., 2017). A process of producing clusters of objects so that the same type of object remains in the same cluster is known as Clustering (Hossain, 2017, Peker et al., 2017). Clustering is an essential technique of data mining for performing data analysis (Hossain, 2017). Clustering techniques are subsets of unsupervised learning (Parvaneh et al., 2014; Kansal et al., 2018). A clustering algorithm allows identifying groups of customers sharing similar purchasing habits in their customer base (Aktas et al., 2021).

Many clustering methods have been developed to form clusters in data mining. Dealing with clustering algorithms has challenges when it comes to finding the most suitable clustering models for the actual data and finding the optimum parameters (Aktas et al., 2021). K-means algorithm is the most accepted clustering algorithm (Chugh, 2020; Peker et al., 2017). K-means requires a pre-defined number of clusters (k). For a given k value, it begins with the randomly developed k central points and afterwards the distance between each sample and each central point is calculated. After creating the clusters, the mean of an individual cluster is recomputed for the existing objects in the cluster. The method is iterated until convergence occurs (Peker et al., 2017). Many previous studies have successfully implemented K-means algorithms for customer segmentation (Parvaneh et al., 2014; Peker et al., 2017; Kansal et al., 2018; Sheshasaayee et al., 2018). K-means has been used widely because of its ability to fast processing of massive amounts of data (Parvaneh et al., 2014). Peker et al. employed K-means algorithm to cluster grocery chain customers based on LRFMP variables (Peker et al., 2017).

Comparing different algorithms leads the way to understanding the best-suited methods and parameters for the studied domain (Aktas et al., 2021). Therefore, one study compared different clustering algorithms for customer segmentationKansal et al. executed three clustering algorithms (K-means, Agglomerative, and Meanshift) with RFM mode for customer segmentation and compared the outcomes. The mean-shift clustering algorithm is an algorithm that assumes all the data points in the feature space as an empirical probability density function. The algorithm clusters each data point by allowing data point to converge to a region of local maxima which is achieved by fixing a window around each data point finding the mean and then shifting the window to the mean and repeat the steps until all the data point converges forming the clusters (Kansal et al., 2018).

Agglomerative Clustering is about forming a hierarchy represented by dendrograms. The dendrogram acts as a memory for the algorithm to talk about how the clusters are being formed. The clustering starts with forming N clusters for N data points and then merging along with the closest data points together in each step such that the current step contains one cluster less than the previous one. They found out that there is not much significant difference in K-means and Agglomerative clustering, and both were able to cluster data well than Mean shift algorithm (Kansal et al., 2018). Ozan & Iheme suggested a machine learning model, namely Multi-layer Perceptron (MLP) which successfully generalized the company's data segmentation intuition.

2.5 Group Profiling

To interpret the results of customer segmentation, descriptive profiles of customers can be created so that CRM and marketing strategies can be implemented according to different customer profiles. One research suggested a group profiling technique (Ha and Park, 1998). Ha and Park's research was based on the RFM model and for positioning customer segments, each cluster's average R, F, M values were compared with the total average R, F, M values of all clusters. If the average value is greater than the total average, an upward arrow ↑ is given to that value. A downward arrow is given if the contrasting case occurs (Ha and Park, 1998).

Another study proposed a dsegments based on frequency and monetary and constructed four customer types, "best customers, spender customers, uncertain customers, and frequent customers" (Marcus, 1998). Another study profiled customers by using L and R attributes: "close relationship, potential relationship, lost relationship, and a new relationship" (Chang and Tsay, 2004). One study created five customer segments: "High-contribution loyal customers, Low-contribution loyal customers, Uncertain customers, Uncertain customers, and Uncertain customers" (Peker et al., 2017).

This study utilizes group profiling technique suggested by Ha and Park. So, the final determined customer segments were profiled by registering the up arrow, if the average LRFMP feature value of the cluster is greater than the aggregate average, contrarily utilizes, the down arrow. In this way, we can fully utilize the knowledge resulting from customer segmentation, which would benefit the Swedish pharmacy retail sector to improve services for various segments.

3 Method

This chapter aims to present the research strategy & method, the data collection and data analysis method chosen for this study, and the alternative methods and techniques that are considered.

3.1 Research Strategy

Johannesson & Perjons defined research strategy as an approach to answering a research question (Johannesson & Perjons, 2012). Denscombe stated that no single strategy could be recommended as the 'best' in all circumstances (Denscombe, 2010).

As this study focuses on experimenting with the segmentation methods with the dataset in an organization, the chosen method for performing this research is an experiment. An experiment is an empirical study under controlled conditions designed to examine the properties of specific factors (Denscombe, 2010). This study wanted to gain insight into how different combinations of customer behaviours and algorithms can be utilized to segment customers. This experimental study allows testing customers' purchasing & visiting patterns by analyzing customer transaction receipts with several machine learning algorithms, which would help create better segments of customers. Hence, an experiment was selected as an ideal research strategy. Experiment determines the cause of changes to the variable of interest being studied, and it is not usually enough to show that two things that occur are linked. Hence, experimental analysis of the dataset has been conducted to determine helpful customer profiles and how transaction history and different machine learning algorithms can be combined for better results.

According to Denscombe, there are six other research strategies: surveys, experiments, ethnography, phenomenology, grounded theory, and action research. These strategies can be blended into mixed methods. A researcher employs surveys they want to acquire information correlating to groups of people: what they do, what they think, who they are (Denscombe, 2010). Surveys are inappropriate for this study as the study will not require collecting primary data from large amounts of informants and may require secondary data to facilitate richer understanding (Johannesson & Perjons, 2012).

Another research method is Ethnography which refers to studying cultures & groups and aims at understanding and observing them through spending extensive time in the field of study. It depends on primary data through direct observations, and the data collected may be both in-depth and detailed (Denscombe, 2010). Ethnography was rejected as this study utilizes secondary data for customer segmentation. Grounded theory is a research strategy that focuses on creating theories based on empirical fieldwork. However, the approach relies on the iterative addition of new data and utilizes a comparative method to analyze the data. On the other hand, action research is a practical approach where the researcher takes an active part in creating best practice guidelines to solve some practical problems (Denscombe, 2010). Both grounded theory and action research have been excluded due to time constraints.

Another research method is case study which is the study of one or just a few instances of a phenomenon to provide in-depth knowledge of events, relationships, experiences, or processes occurring in that instance. Denscombe stated that the defining characteristic of the case study strategy is to focus on just one instance which is to be investigated (Denscombe, 2010). If experimental analysis delivers promising results, Extenda Retail implements this study techniques to the Relevate (a CRM platform) and applies promotional strategies to segmented customers of a client; a case study on the actual performance would be interesting to see the effects of new promotional strategies designed using experimented methods in this study. However, it would take much time, effort, and client approvals to implement in the production environment. Hence, as of now, it has been excluded as inappropriate due to time constraints.

3.2 Data Collection

To answer the stated research question, the collection of quantitative secondary data (organizational records) is deemed most appropriate.

Johannsson & Perjons stated that regardless of the type of data, there are five widely used data collection methods: questionnaires, interviews, focus groups, observation studies, and document studies. For this study, documents have been chosen as the main data collection method. Several other methods were rejected due to the inefficiency to attend the research problem. Firstly, questionnaires were excluded as questionnaires are typically used to gather primary data that is brief and unambiguous, which may be about simple facts (Johannesson & Perjons, 2012). Interviews are suitable when a researcher needs to get insight into the population's feelings, emotions, and experiences (Denscombe, 2010). A focus group is an interview in which a group of respondents participate and discuss a specific topic (Johannesson & Perjons, 2012). As this study focuses on the customer's purchasing behavior by focusing on retail transactions, interviews and focus groups were rejected. Observations were also excluded as they are better suited when studying fieldwork in its natural setting (Denscombe, 2010).

Documentary research uses outside sources and documents to support academic work, basically secondary data (Denscombe, 2010). Johannesson & Perjons stated some common types of documents: Government Publications, Organizational Records, Academic Publications, Newspapers & Magazines, Personal Communications, and social media Streams (Johannesson & Perjons, 2012).

For this study, the organizational records document type was chosen. Data for quantitative study comes from a pharmacy chain that is a client of Extenda Retail. Dataset used in this study comes from the client company, a well-known pharmacy chain company operating in Sweden with a market share of around 28 percent. The dataset contains purchase transactions of more than one million customers and consists of receipts from all their stores and online purchase in Sweden. Receipts contain all items bought, amount, customer's membership number, quantity purchased, promotions and a purchase DateTime.

3.3 Data Analysis

Raw data needs to be prepared, interpreted, analyzed, and presented before concluding the results. Thus, a researcher needs to transform large volumes of data into meaningful pieces of information (Johannsson & Perjons, 2012). The anonymized dataset was extracted from a database to BigQuery in the Google Cloud Platform (GCP). GCP is a secure, reliable, high-performance infrastructure for cloud computing, data analytics & machine learning offered by Google. BigQuery is cloud data warehouse that enables scalable analysis of billions of rows using a SQL-like syntax (Google Cloud, 2022).

BigQuery contains purchase transactions of more than one million customers between January 1, 2015, and January 31, 2022. Deleting irrelevant features, transactions with missing values, and aggregating transaction records on the same day for each customer, pre-processing the data takes a significant part to obtain a proper dataset before analysis. In order to meet the requirements for customer segmentation, the information of some dimensions must be merged to obtain information about customer demographics and transactions.

For data pre-processing and implementing customer segmentation (segmentation models and clustering algorithms), Google Colaboratory(or Colab) was used. Colab is an online Jupyter Notebooks environment from Google. It is in the cloud, so no software installation is necessary, and it is available from any internet-connected computer (Google Colab, 2022). In the first step, the features of RFM, LRFM, LRFMP models are computed for each customer. Then the researcher calculates the descriptive statistics of those attributes. After that, all variables of segmentation models are standardized before clustering using the StandardScaler to rescale with the mean of zero and a standard deviation of one. K-means, Agglomerative, and Mean-shift clustering algorithms are executed on calculated segmentation variables, clustering results with the determined number of clusters are examined, and segments are finally profiled.

3.4 Research Ethics

Denscombe stated that the four fundamental principles for research practices are "to protect the interests of the participants, ensure that participation is voluntary and based on informed consent, avoid deception and operate with scientific integrity and comply with the laws of the land"(Denscombe, 2010, p. 331). For this study, the first principle and last principle were fulfilled. There were no ethical risks to participants because the customer data used in this study is anonymized. No personal or other information that could identify the customers was used. Additionally, the company in this study follows local Swedish laws for storing and handling the data. The researcher signed a Non-Disclosure Agreement with Extenda Retail. This agreement permits the researcher to publish the results of this study. Still, it does not allow the publication of the client company dataset used.

4 Result

In this section the results of the study are shown and discussed.

4.1 Data pre-processing

First, secondary data of customer transactions in pharmacy retail were extracted from a database in the BigQuery data warehouse of Google Cloud. The dataset contains 202753 transactions between the period January 1, 2015, and January 31, 2022. Data preprocessing is an essential step because it enhances the accuracy and efficiency of modeling (Parvaneh et al., 2014).

Data preprocessing in this study involves data aggregation, data cleaning, and data reduction to improve data quality for customer segmentation. Extracted data was in the lowest granularity, where each item sold has a separate row. However, the study was interested in each customer basket. Hence, data aggregation was performed to have a row for each customer receipt (or basket) with the total amount and the total quantity purchased at a DateTime. Data cleansing takes place in the Google BigQuery data warehouse by extracting relevant features, eliminating duplicates, removing unreasonable records such as customers who have zero amount of monetary and transaction records, and putting the data into the desired format.

Several customers with only one transaction within the observation period. Customers who have one transaction are considered useless. The customers who have purchased less than two times were excluded. The final dataset includes 6955 instances with eight features. These features are shown in Table 4. These features are as follows:

CustId is the unique number assigned to every customer. ReceiptNumber is the unique number assigned to each transaction. TotalQuantity is the total number of items per transaction in units. PurchaseDateTime is the purchase date and time for when the transaction has happened. TotalAmount is the total amount in SEK per transaction. Age is the age of customer. Gender is the gender of customer. PurchaseCounter is the counter for the transactions by each customer.

CustId	Receipt Number	Total Quantity	Total Amount	Purchase DateTime	Age	Gender	Purchase Counter
C1	01	9	700.02	2018-07-24 09:04:49	88	Man	3
C2	02	7	1848	2016-10-21 12:43:35	90	Woman	7
C3	03	3	281	2021-09-06 12:40:38	102	Woman	3
C4	04	3	285	2019-05-09 13:18:51	103	Woman	37

Table 3: Sample of the dataset

4.2 Feature Engineering

4.2.1 RFM & LRFM model

The selected attributes from the dataset were transformed into LRFM & RFM model variables in the feature engineering step. Length, Recency, Frequency, and Monetary were extracted from transaction data to obtain a new dataset with the LRFM model variables for each customer. On the other hand, Recency, Frequency, and Monetary were generated to obtain a new dataset with the RFM model variables for each customer. These newly developed datasets were then stored in a CSV file for clustering later.

4.2.2 LRFMP model

Before calculating the LRFMP variables, customers who have few transactions in the observation period are considered useless because of the periodicity of customers. Hence, we excluded customers who have less than three transactions. Therefore, the dataset used for the LRFMP model is left with purchase records of 1338 customers.

Since the segmentation is on the LRFMP model, the selected attributes from the original dataset were transformed to Length, Recency, Frequency, Monetary, and Periodicity from transaction data to obtain a new dataset with the LRFMP model variables for each customer. The features of the LRFMP model were generated for each customer, and this newly developed dataset was then stored in a CSV file. The descriptive statistics of these attributes are shown in Table 5.

	Maximum	**Minimum**	**Average**	**Standard Deviation**
Length	2260.0	0.0	446.79	452.52
Recency	2325.0	0.0	781.54	363.63
Frequency	111.0	3.0	4.08	4.28
Monetary	88461.87	55.52	1451.02	3520.65
Periodicity	1409.26	0.0	141.75	192.53

Table 4: Descriptive statistics of the LRFMP variables

4.3 Cluster Analysis

4.3.1 Clustering for LRFMP model metrics

First, earlier generated, the CSV file of the LRFMP model was imported into the Google Colab, which is an online Jupyter Notebook environment. The LRFMP attributes can be different in scale. They are standardized to have a common scale before building a machine learning model. Standardization solves the problem by decreasing the potential effects of variable differences (Peker et al., 2017). Python sklearn library offers StandardScaler() function to standardize the data values into a standard format. Thus, before clustering, the

LRFMP model variables are standardized using StandardScaler(), which standardizes features by removing the mean and scaling to unit variance. Then feature distribution has a mean value of 0 and a standard deviation of 1. Three clustering algorithms (K-means, Agglomerative & MeanShift) have been utilized for customer segmentation.

K-means Clustering

An essential step for any unsupervised algorithm is determining the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k. For each of the K values, we calculate average distances to the centroid across all data points. Plot these points, and if the line chart resembles an arm, then a point of a curve at which a change in the direction occurs indicates that the model fits best at that point.

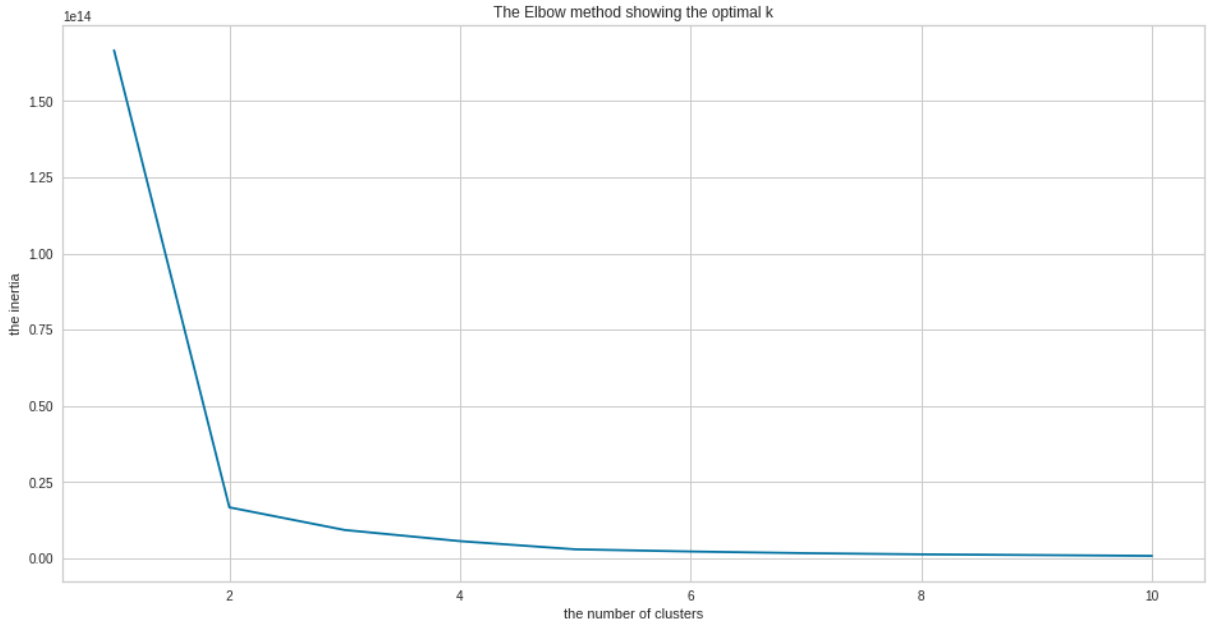

Figure 2: Elbow method for LRFMP

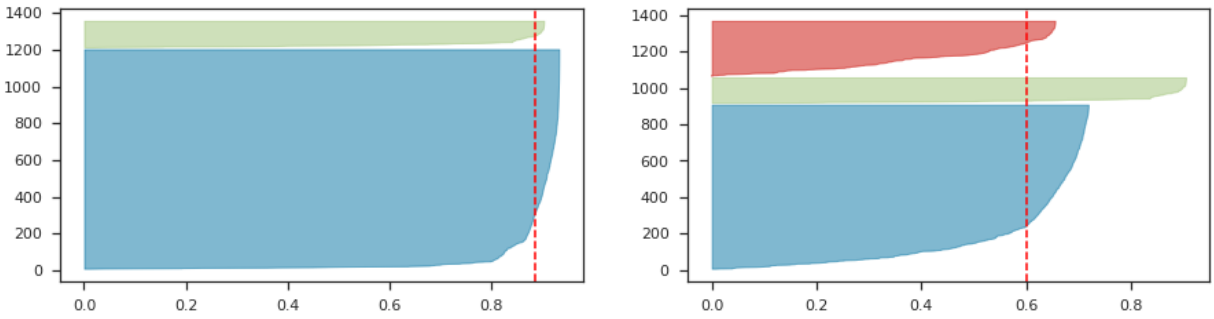

Figure 3: Silhouette Visualizer for K = 2 & K = 3(LRFMP)

Solely relying on the Elbow method for selecting K would lead to choosing a solution in which two clusters exist, which is motivated by the values in Figure 2. Then, silhouette plots were obtained using Silhouette Visualizer for two and three clusters in Figure 3 to provide a general idea of how clusters looked for the various K values. In this case, K = 2 was the most reasonable solution because it shows a horizontal line in the elbow. By fitting the model, we got clusters where each data belongs. Besides that, segments obtained by K-means were analyzed using a snake plot, as shown in Figure 4. By using this plot, we can have a good visualization of the data on how the cluster differs from each other.

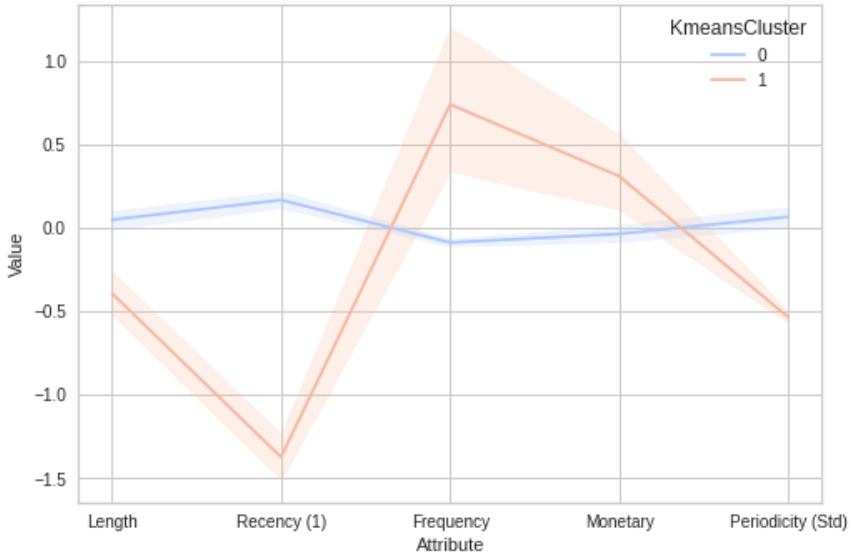

Figure 4: Snake plot of K-means algorithm (LRFMP)

Agglomerative Clustering

A dendrogram is being used to find the optimal number of clusters for Agglomerative clustering. Fig 5 shows the dendrogram of the new LRFMP dataset. The dendrogram is the hierarchical representation of an object. In a dendrogram, one should look for the longest vertical line, which is not cut by any of the horizontal line extended virtually over the complete width of the graph. By drawing a hypothetical horizontal line cutting through the longest vertical line, we get the horizontal line cutting two vertical blue lines providing the optimal number of clusters (Kansal et al., 2018). Based on the observation, the n_cluster = 2 is the suitable hyperparameter for our agglomerative model. By fitting the model, we got four customer segments. For the Agglomerative clustering, the snake plot is presented in Figure 6.

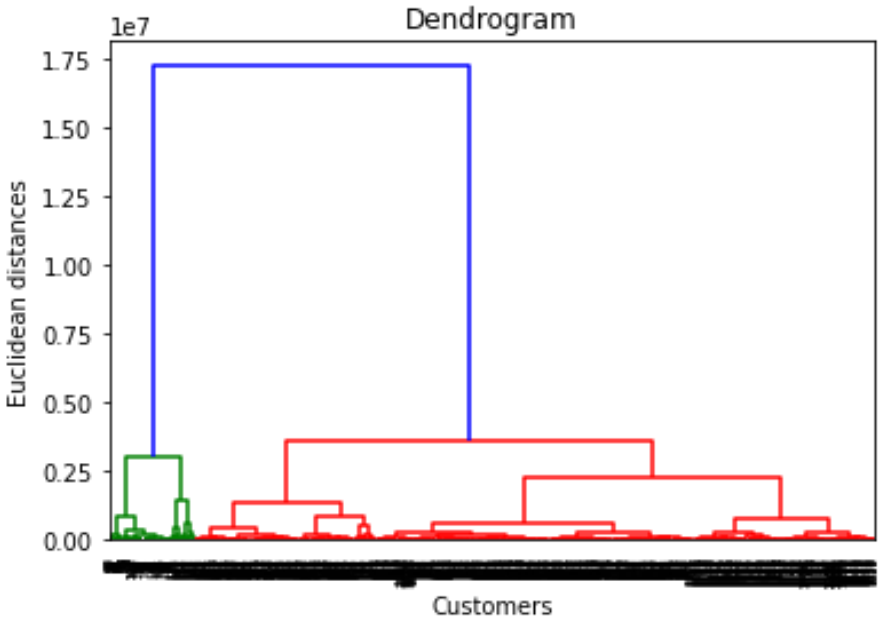

Figure 5: Dendrogram Structure for LRFMP

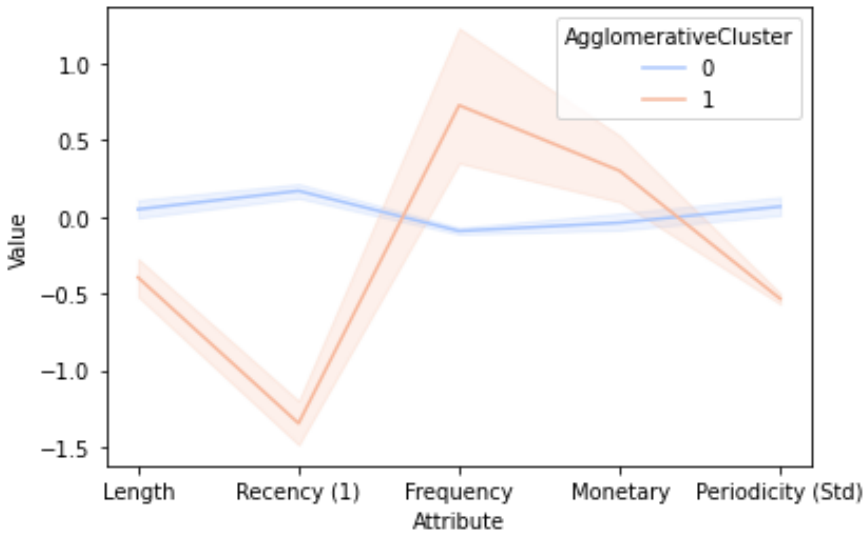

Figure 6: Snake plot for Agglomerative clustering (LRFMP)

Mean-Shift Clustering

The Mean-Shift algorithm is a centroid-based algorithm that updates candidates for centroids to be the mean of the points within a given region (scikit-learn, 2022). This algorithm clusters each data point by allowing the data point to converge to a region of local maxima. This is accomplished by specifying a window around each data point, finding the mean, then shifting the window to the mean, and repeating the steps until all the data points converge, forming the clusters. The Mean-Shift technique is used for real data analysis in which the initial shape of the data cluster is not presumed (Kansal et al., 2018).

The only input requisite for the Mean-shift algorithm is the bandwidth which is the radius of the circle (kernel) defining how much the data points should be in the cluster (Kansal et al.,

2018). The estimate_bandwidth() function from scikit-learn was utilized to estimate the bandwidth with quantile=0.4 for the median of all pairwise distances. Then the estimated bandwidth was used to do the clustering. The Mean-Shift algorithm gave four customer segments presented below (Figure 7).

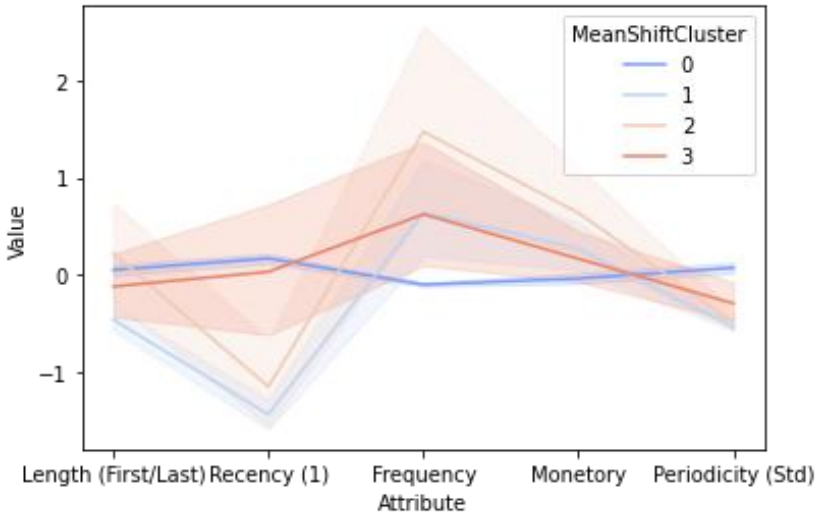

Figure 7: Snake plot for Mean-Shift clustering (LRFMP)

	K-means	**Agglomerative Clustering**	**Mean-Shift Clustering**
Clusters	**Total customers**		
0	1194	1188	1182
1	144	150	121
2	-	-	18
3			27

Table 5: Number of objects in K-means, Agglomerative & Mean-Shift Clusters (LRFMP)

Table 6 illustrates the silhouette score for the three algorithms applied for the LRFMP model. The table indicates that the silhouette values of K-means and Agglomerative clustering with 2 clusters are same. On the other hand, the Mean-Shift algorithm has lowest silhouette score as compared to other two algorithms for four clusters.

Silhouette Score (LRFMP)

K-means (clusters = 2)	0.88
Agglomerative (clusters = 2)	0.88
Mean-Shift (clusters = 4)	0.72

Table 6: Silhouette scores for LRFMP dataset

4.3.2 Clustering for RFM & LRFM metrics

The CSV files of the RFM and the LRFM model were imported into two separate Google Colab. Before clustering, RFM & LRFM model variables were standardized using StandardScaler() function, which standardizes features by removing the mean and scaling to unit variance. All the three clustering algorithms (K-means, Agglomerative & MeanShift) were performed on both the model metrics, same as explained in the above section.

Table 7 depicts the silhouette scores of the RFM model for three clustering algorithms implemented in this study. The table shows that there is not much difference in the silhouette score of K-means and the Agglomerative clustering. K-means algorithm has the highest value of silhouette score. However, Mean-Shift clustering has the lowest silhouette score. Hence, we can say that the K-means algorithm performs well for RFM model metrics. Table 7 also displays the silhouette scores of the LRFM dataset. For the LRFM, Agglomerative clustering algorithm gave better silhouette score than other two algorithms.

	Silhouette Score (RFM)	**Silhouette Score (LRFM)**
K-means Clustering	0.89 (cluster = 2)	0.64 (cluster = 4)
Agglomerative Clustering	0.88 (cluster = 2)	0.88 (cluster = 2)
Mean-Shift Clustering	0.83 (cluster = 3)	0.63 (cluster = 4)

Table 7: Silhouette scores for RFM & LRFM datasets

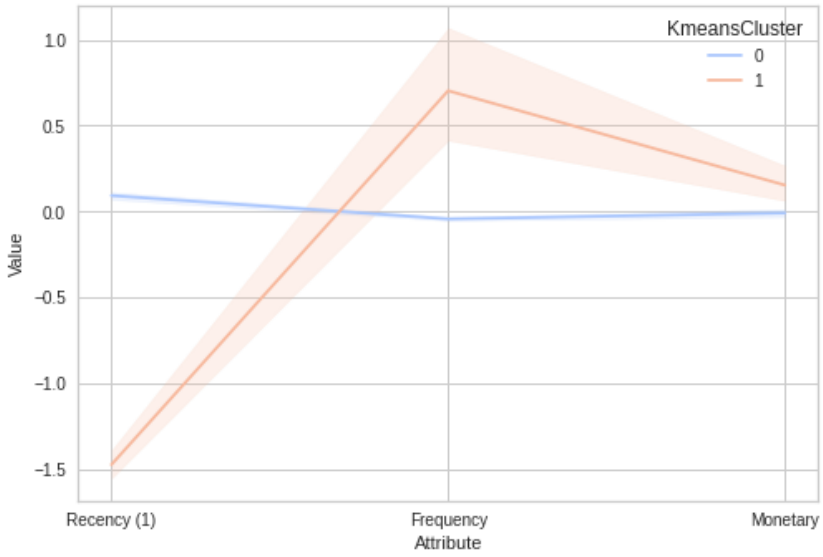

Figure 8: Snake plot for K-means algorithm (RFM)

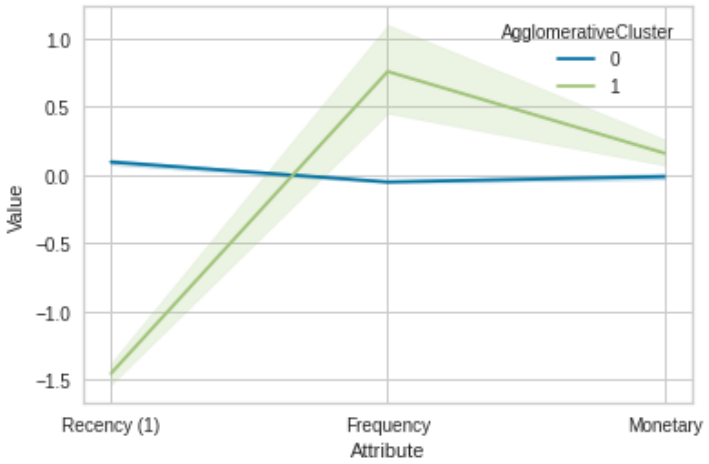

Figure 9: Snake plot for Agglomerative clustering (RFM)

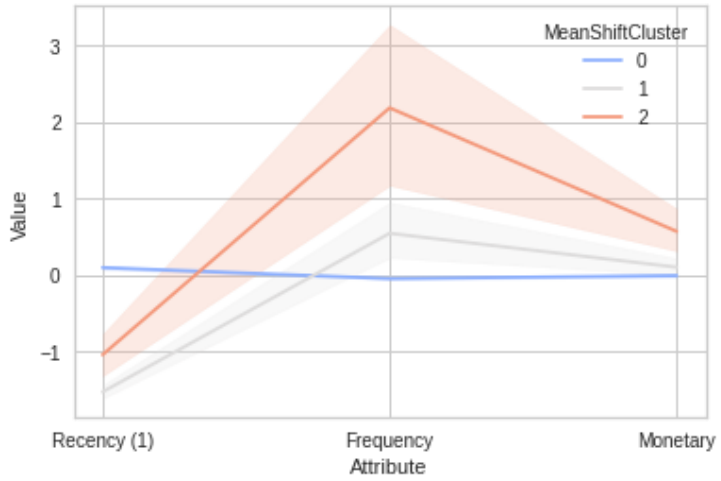

Figure 10: Snake plot for Mean-Shift clustering (RFM)

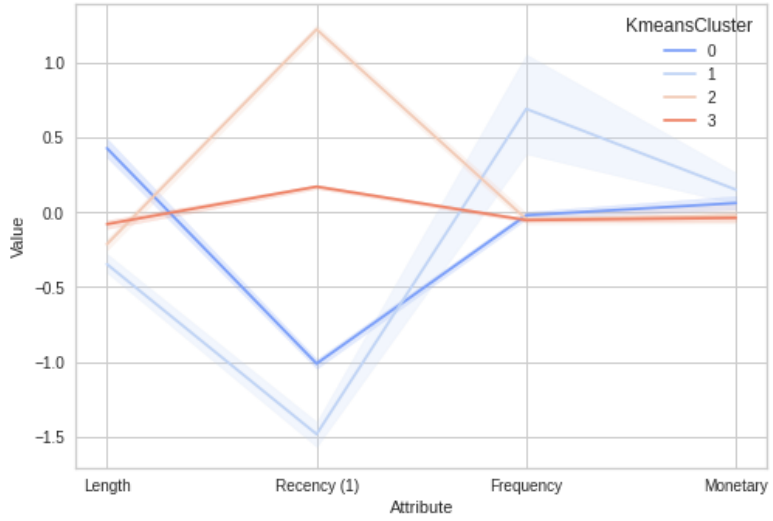

Figure 11: Snake plot for K-means algorithm (LRFM)

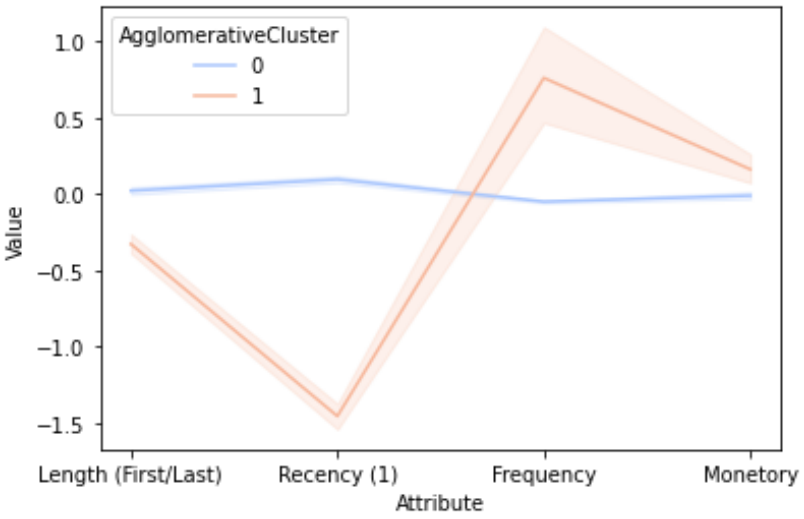

Figure 12: Snake plot of Agglomerative clustering (LRFM)

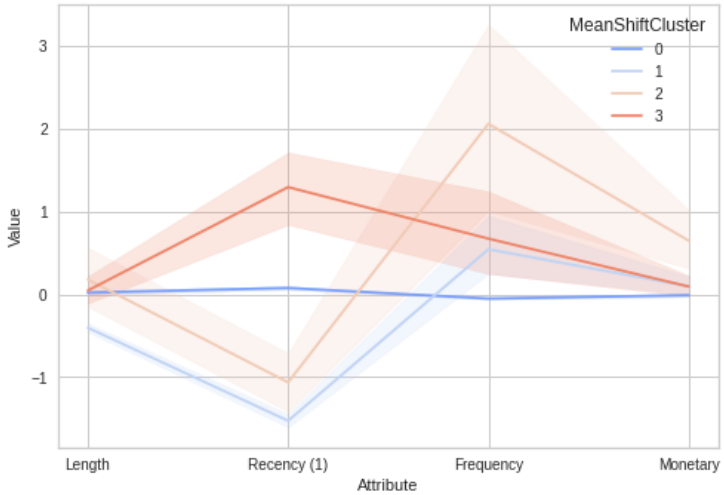

Figure 13: Snake plot of Mean-Shift clustering (LRFM)

Segmentation models like LRFMP, LRFM, and RFM show the customer purchasing behavior. According to the company in this study, a segmentation model that provides more information about customers' purchasing behavior would be ideal for implementing it in the CRM platform. The Length, Recency, Frequency, Monetary, and Periodicity model delivers more details about customer behavior than other models. On the other hand, three clustering algorithms are explored and compared using a validation matrix to identify different customer groups based on LRFMP, LRFM & RFM analysis. Table 6 and 7 shows the silhouette scores for all three clustering algorithms performed for each segmentation model.

For the LRFMP model, the K-means and Agglomerative clustering algorithm provided a better silhouette score than the Mean-Shift algorithm. Figures 4, 6, and 7 show the snake plots for the LRFMP model. For the LRFMP model, both K-means and Agglomerative clustering generated two customer segments, while the Mean-Shift algorithm developed four customer segments.

For the RFM model, the K-means algorithm shows a higher silhouette score (0.88) than other algorithms. Figures 8, 9, and 10 show the three algorithms' snake plots for the RFM model. We can see in those figures that the K-means and the Agglomerative clustering delivered two customer segments, whereas the Mean-Shift clustering provided three customer segments with various purchasing behaviors. For the LRFM model, the agglomerative clustering algorithm delivered a higher silhouette score with two clusters. While the K-means and Mean-Shift clustering gave four clusters and a lower silhouette score as compared to Agglomerative clustering.

As the LRFMP model offers more information about customer purchasing behavior, it is more suitable for customer segmentation. For the LRFMP model variables, three clustering algorithms were performed. K-means & Agglomerative clustering delivered a 0.88 silhouette score for two clusters and Mean-Shift clustering gave a 0.72 silhouette score for four clusters. Table 8 shows the calculated LRFMP score for segments generated by the LRFMP model and K-means algorithm. Table 9 shows the calculated LRFMP score for segments by the LRFMP model and Agglomerative clustering.

K-means clustering, and Agglomerative clustering generated two customer segments with the highest silhouette score of 0.88; hence we can say that both performs well for the dataset used in this study. Later, the LRFMP scores for the results of both algorithms were created to be utilized in profiling segments (Table 8 & 9).

Clusters	Count	Average Length	Average Recency	Average Frequency	Average Monetary	Average Periodicity	LRFMP Scores
0	1194	468.26	841.82	3.69	1320	154.23	L↑ R↑ F↓ M↓ P↑
1	144	268.79	281.67	7.23	2537	38.30	L↓ R↓ F↑ M↑ P↓
Aggregated Average		446.79	781.53	4.07	1450.97	141.75	

Table 8: Results by the LRFMP model and K-means clustering

Clusters	Count	Average Length	Average Recency	Average Frequency	Average Monetary	Average Periodicity	LRFMP Scores
0	1188	469.36	843.37	3.67	1317.24	154.74	L↑ R↑ F↓ M↓ P↑
1	150	268.09	291.79	7.19	2510.52	38.94	L↓ R↓ F↑ M↑ P↓
Aggregated Average		446.79	781.53	4.06	1451.01	141.75	

Table 9: Results by the LRFMP model and Agglomerative clustering

4.3.3 Group profiling

Based on clustering analysis results, profiles for segments of customers are constructed in this section. The strategic positioning of customer clusters is described in Tables 10 & 11, which illustrate why segments are unique from the LRFMP point of view. The average LRFMP values of each segment are compared with the total average LRFMP values of all clusters to generate the LRFMP scores shown in Tables 8 & 9.

Based on the segmentation results by K-means and Agglomerative clustering using the LRFMP scores (Table 10 & 11), we obtained two segments profiled as “Low-contribution customers” and "High-contribution customers”. Segmentation results obtained by K-means and Agglomerative clustering generated almost same results with two segments 0 & 1. Segment 0 has average F value lower than the aggregate averages and average L & R values greater than the aggregate average. This segment has the highest R & P values. These attributes indicate that they have not visited the store recently, and their purchases are irregular. Segment 0 includes customers with the lowest contribution to the company's revenue. Thus, the customers in this segment can be classified as "Low-contribution customers." Thus, converting such lost customers into loyal ones is a challenge for the company to improve its profitability.

On the other hand, segment 1 has average F & M values greater than the average and R & P values lower than the average. Additionally, customers in segment 1 purchase more recently, frequently, and regularly. Although segment 1 is the smallest, customers in this segment have the highest contribution to the company's revenue during the observation period. Hence, segment 1 can be described as "high-contribution customers,"

Segments	LRFMP Scores	Segment name
0	L↑R↑F↓M↓P↑	Low contribution cutomers
1	L↓R↓F↑M↑P↓	High contribution customers

Table 10: Group profiling (LRFMP and K-means clustering)

Segments	LRFMP Scores	Segment name
0	L↑R↑F↓M↓P↑	Low contribution cutomers
1	L↓R↓F↑M↑P↓	High contribution customers

Table 11: Group profiling (Agglomerative clustering & the LRFMP)

5 Conclusion & Discussion

The thesis aimed to efficiently segment a Swedish pharmacy retail company's customer by combining segmentation models and clustering algorithms. To answer the first research question, *"Which combination of the mentioned segmentation model and a clustering algorithm performs well for segmentation of a Swedish pharmacy retail company's customer?"* the pharmacy retail dataset from a Swedish pharmacy chain, a client of the company in this study, was analysed by combining the L, R, F, M, and P criteria and machine learning algorithms. First transactional data is transformed into three segmentation model metrics RFM, LRFM, and LRFMP. Then K-means, Agglomerative clustering, and Mean-Shift clustering were performed on those metrics and compared using a validation index called the Silhouette score.

As the LRFMP model offers more information about customer purchasing behavior, it is more suitable than the RFM and the LRFM model. The LRFMP was chosen as the most suitable model to gain deeper and more valuable insights into customer purchasing behavior for the Swedish pharmacy dataset. For the LRFMP model, three clustering algorithms were executed. K-means & Agglomerative clustering performed well for the customer segmentation and gave a 0.88 silhouette score for two clusters. On the other hand, Mean-Shift clustering gave four clusters with a 0.72 silhouette score. Hence, both algorithms (K-means and Agglomerative clustering) with the LRFMP model are most appropriate solution for customer segmentation of the Swedish pharmacy retail company's dataset used in this study.

Moreover, to answer the second research question, *"How can extracted customer segments be profiled?"* segmented customers are then profiled so that the company could ultimately improve customer services and customer relationships. Customers are segmented into two clusters. The LRFMP Scores were determined and based on those scores, these two segments of customers are then profiled as: "Low-contribution customers," and "High-contribution customers." Personalized promotions and other marketing strategies can be delivered to these segments by using this information.

Previous studies have created more customer segments as compared to this study. One study in Turkish grocery retail created five customer segments using L, R, F, M, and P customer attributes (Peker et al., 2017). Another study profiled customers and proposed four relationship types: close relationship, potential relationship, lost relationship, and a new relationship (Chang and Tsay, 2004). This study utilized a pharmacy dataset in Sweden, which created two customer segments. The first segment represents customers who have not purchased recently, less contributed to revenue, and are irregular. This might be because of the fact that customers from the younger age group do not visit the pharmacy store repeatedly, and hence they are not frequent and contribute less to the company's revenue. The second segment represents customers who purchase more recently, frequently, and regularly. After looking into the age of this customer segment, it revealed that the customers in this segment are from an older age group who visits the pharmacy store more often. Hence, they have a high contribution to the company's revenue.

5.1 Originality and Significance

This study contributes to prior literature by delivering essential insights into the customer segmentation in pharmacy retail in Sweden. It indicates how a pharmacy company can divide its customers into meaningful segments. Moreover, which customer segments are essential for them, which segments would require more focus, and which segments contribute most to their business. Additionally, retail pharmacy companies in Sweden can benefit from the methodology employed in this study to define various customer segments and can revise their strategies for better customer satisfaction. Refinements in strategies can improve customer service quality and increase customer loyalty, ultimately increasing profits.

The contributions of this study are first, the literature on the RFM approach towards customer segmentation has further developed by this research by executing and comparing segmentation models and clustering algorithms for the pharmacy retail dataset. The P variable has recently been introduced in the segmentation domain by the research of Peker, Kocyigit, and Eren in the grocery retail industry in Turkey. Additionally, this study extends the application of the LRFMP model in the B2C setting and the pharmacy retail in Sweden. The three-step approach, which includes the execution of the three segmentation models, three clustering algorithms for segmenting customers, and finally group profiling of generated clusters adopted in the study, adds further improvement to the clustering of pharmacy retail customers.

5.2 Limitations

Despite the benefits and contributions of this study, there are still limitations to this study. Only customers of a pharmacy chain functioning in Sweden are assessed in this study. The buyers from other countries and other domains might reveal distinct attributes. Hence, the chosen dataset for this study may not represent the entire retail domain. Only one cluster validation index was used; different cluster validation indices could have been used. This study was limited to the chosen observation period from January 2015 to January 2022. The report also serves as a reproducibility tool that future studies in different countries and domains can use as a roadmap for customer segmentation. However, in this study, limited attributes of customer behaviors are utilized for segmentation. Other attributes such as product type purchased, and product category purchased could be introduced to the segmentation model to interpret the customer’s behavior more soundly. Only three clustering algorithms have been investigated in this study, and more clustering algorithms could also be of great interest to optimize the result.

5.3 Future research

This study only considered the customers of a pharmacy chain that operates in Sweden. Future research might extend this study to international companies to validate the results. Since the study was conducted in Sweden, a study on datasets from other countries would establish whether the results are consistent throughout different countries. As mentioned in the research strategy section, future research can do a case study by implementing results obtained from this study, and promotional strategies can be applied to the segmented

customers. For future work, other clustering algorithms could be executed and compared. Further studies can also consider the importance of marketing strategies in each segment and their behavior. Advancement of this study can be accomplished by adding features related to purchasing patterns, namely products purchased, and product category purchased could be introduced to the segmentation model. Finally, further studies can perform three-step customer segmentation explored in this study to other domains such as the fashion retail industry, which might give a distinct insight into customer behaviors.

References

Aktas¸A.A., Okan Tunalı, O., Bayrak, A.T. (2021) “Comparative Unsupervised Clustering Approaches for Customer Segmentation”, 2nd International Conference on Computing and Data Science (CDS), pp. 530-535, DOI: 10.1109/CDS52072.2021.00097.

Bhattacherjee, A. (2012) “Social Science Research: Principles, Methods and Practices 2nd ed”., Tampa, FL: Global Text Project.

Bhardwaj, A. (2020) “The Importance of CRM Software Development In Retail” Available at The Importance of CRM Software Development In Retail (oodles.io) (Accessed: February 2022).

Bostrom, G.O. and Wilson, T.L. (2009) “Swedish retail banking: a competitive update”, Competitiveness Review: An International Business Journal, 19 (5), pp. 377-390, DOI: 10.1108/10595420910995993

Chang, H.H. and Tsay, S.F. (2004) “Integrating of SOM and K-mean in data mining clustering: an empirical study of CRM and profitability evaluation”, Journal of Information Management, 11 (4), pp. 161-203.

Chang, H.-C. and Tsai, H.-P. (2011) “Group RFM analysis as a novel framework to discover better customer consumption behavior”, Expert Systems with Applications, 38(12), pp. 14499-14513.

Chugh, S., Baweja, V.R. (2020) “Data Mining Application in Segmenting Customers with Clustering”, 2020 International Conference on Emerging Trends in Information Technology and Engineering.

Denscombe, M. (2010) “The good research guide for small-scale social research projects”, 4th ed. England: Berkshire: Open University Press.

Dibb, S. (1998), “Market segmentation: strategies for success”, Marketing Intelligence & Planning, 16 (7), pp. 394-406.

Dudek, A. (2020) “Silhouette Index as Clustering Evaluation Tool”, Springer International Publishing (Studies in Classification, Data Analysis, and Knowledge Organization), DOI: 10.1007/978-3-030-52348-0.

Griva1, A., Bardaki, C., Pramatari, K., Doukidis, G. (2021) “Factors Affecting Customer Analytics: Evidence from Three Retail Cases”, Information Systems Frontiers: A Journal of Research and Innovation, DOI: https://doi.org/10.1007/s10796-020-10098-1

Google Colab (2022) “Welcome To Colaboratory”.

Available at: Welcome To Colaboratory - Colaboratory (google.com)

Google Cloud (2020) “Google Cloud Platform (GCP) - Trusted by leading companies”. Available at: Cloud Computing, Hosting Services, and APIs | Google Cloud (Accessed: March, 2022)

Granov, A. (2021). “Customer loyalty, return and churn prediction through machine learning methods: for a Swedish fashion and e-commerce company”, Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-184709.

Giri, C. and Johansson, U. (2021) “Data-driven Business Understanding in the Fashion and Apparel Industry”. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=edsswe&AN=edsswe.oai.DiVA.org.hb.26470&site=eds-live&scope=site.

Ha, S.H. and Park, S.C. (1998), “Application of data mining tools to hotel data mart on the intranet for database marketing”, Expert Systems with Applications, 15(1), pp. 1-31

Hossain, A.S.M.S. (2017) “Customer Segmentation using Centroid Based and Density Based Clustering Algorithm”, 3rd International Conference on Electrical Information and Communication Technology (EICT), pp.1-6, DOI: 10.1109/EICT.2017.8275249

Huang, Y., Zhang, M., He, Y. (2020) “Research on improved RFM customer segmentation model based on K-Means algorithm”, 5th International Conference on Computational Intelligence and Applications (ICCIA), pp. 24-27, DOI: 10.1109/ICCIA49625.2020.00012.

Johannesson, P., Perjons, E. (2014) An Introduction to Design Science. Springer International Publishing, New York.

Kansal, T., Bahuguna, S., Singh, V., Choudhury, T. (2018) “Customer Segmentation using K-means Clustering”, *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pp. 135-139, DOI: 10.1109/CTEMS.2018.8769171.

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

Martyn Denscombe, 2010, The good research guide for small-scale social research projects, 4th ed. England: Berkshire: Open University Press.

Marcus, C. (1998), “A practical yet meaningful approach to customer segmentation”, Journal of Consumer Marketing, 15(5), pp. 494-504.

Nguyen, S. (2021), “S.P. Deep customer segmentation with applications to a Vietnamese supermarkets’ data”, Soft Comput 25, 7785–7793. https://doi.org/10.1007/s00500-021-05796-0.

Ozan, S., Iheme, L.O.(2019) “Artificial Neural Networks in Customer Segmentation”, 27th Signal Processing and Communications Applications Conference (SIU), pp 1-4, DOI: 10.1109/SIU.2019.8806558

Parvaneh, A., Tarokh, M. J. & Abbasimehr, H. (2014) “Combining Data Mining and Group Decision Making in Retailer Segmentation Based on LRFMP Variables”, International Journal of Industrial Engineering & Production Research, 25(3), pp. 197-206.

Peker, S., Kocyigit, A., and Eren, P.E.(2017) “LRFMP model for customer segmentation in the grocery retail industry: a case study”, Marketing Intelligence & Planning, 35(4), pp. 544-559.

Pradnya Paramita Pramono, P.P, Surjandari, I, Laoh, E. (2019) “Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method”, 16th

International Conference on Service Systems and Service Management, ICSSSM 2019. DOI: 10.1109/ICSSSM.2019.8887704.

Soltani, Z., Navimipour, N. J. (2016) "Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research", Computers in Human Behavior 61, pp. 667-688.

Sokol, O. and Holý V. (2021) "The role of shopping mission in retail customer segmentation", International Journal of Market Research, 63(4), pp. 454–470.

Soudagar, R. (2012) "Customer Segmentation and Strategy Definition in Segments: Case Study: An Internet Service Provider in Iran" Available at http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-44406

Sun, Z.H., , Zuo, T.Y., Liang, D., Ming, X., , Chen, Z., Qiu, S. (2021) "GPHC: A heuristic clustering method to customer segmentation", Applied Soft Computing Journal, 1568-4946, DOI: 10.1016/j.asoc.2021.107677

Sheikh, A., Ghanbarpour, T., & Gholamiangonabadi, D. (2019) "A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting", Journal of Business-to-Business Marketing, 26(2), 197-207, DOI: 10.1080/1051712X.2019.1603420

Swedish Trade Federation (Svensk Handel) (2019) "Läget i handeln: 2019 års rapport om branschens ekonomiska utveckling" , pp. 6, 11 - 12, Available at: https://www.svenskhandel.se/globalassets/dokument/aktuellt-ochopinion/rapporter-och-foldrar/e-handelsrapporter/laget-ihandeln_svensk-handel.pdf [Accessed 2020-03-15]

Swedish Trade Federation (Svensk Handel). (2019b) "Den fysiska handelsplatsen i en digital värld", pp. 5 - 7, 41. Available at: https://www.svenskhandel.se/globalassets/dokument/aktuellt-ochopinion/rapporter-och-foldrar/ovriga-rapporter/den-fysiskahandelsplatsen-i-en-digital-varld.pdf [Accessed 2020-03-15].

Soltani, Z., Nima Jafari Navimipour, N.J. (2016) "Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research", Computers in Human Behavior, 667-688

Sheshasaayee, A., Logeshwari, L. (2018) "Implementation of clustering technique based RFM analysis for customer behaviour in online transactions", Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, pp. 1166-1170, DOI: 10.1109/ICOEI.2018.8553873.

Tavakoli,M., Molavi,M., Masoumi, V., Mobini, M., Etemad S. ,and Rahmani, R. (2018) "Customer Segmentation and Strategy Development based on User Behavior Analysis, RFM model and Data Mining Techniques: A Case Study", IEEE 15th International Conference on e-Business Engineering, pp. 119–126.

Wedel, M., Kamakura, W.A. (2000) "Market Segmentation: Conceptual and Methodological Foundations", Kluwer, Boston, MA.

Zhou, X., Zhang, Z., Lu, Y. (2011) “Review of Customer Segmentation method in CRM”, International Conference on Computer Science and Service System, pp. 4033 – 4035. DOI: 10.1109/CSSS.2011.5974617.

Appendix A - Reflection Document

The goal of the thesis was to perform customer segmentation, a customer analytics technique. The thesis has aimed to do an experiment to segment Swedish pharmacy retail customers by combining segmentation models and clustering algorithms and then finally profile segments using a customer profiling technique. The thesis conducted has been a prolonged process where a lot of new knowledge and insights have been learned. After completing the thesis, my contribution to computer and system sciences lies within the field of data science, which fills a gap in the research field. This thesis has collected, analyzed, and presented studies from other authors, including previous literature studies in similar research areas and studies on implementing customer segmentation and clustering algorithms.

The thesis has analyzed the collected data, which was organizational records according to its chosen method, although this step contained some puzzles. I obtained experience working on Google's BigQuery, Google Cloud Platform and worked with Python, although more programming experience would have been more beneficial. This thesis faced conflict of interests between the company and the university. The company wanted to do a Proof of Concept (POC) to discover a certain method. In contrast, the university wanted to contribute to the computer and system science research area by developing a new research framework that was built upon relevant existing scientific research. It took a lot of effort to align the requirements of the university and the company, but after a few conversations this was resolved. Because this study has been carried out and documented by only one person, planning has been simplified, and deadlines and planned goals have only had to be adjusted for me. Of course, another author could have facilitated the work because thesis writing is a prolonged process. But after completing the thesis, I want to express that lone work is beneficial in achieving planned goals.

Many of my previous courses have been contributing to the work, and for the writing and research part, I would say that my two method courses (Scientific Communication and Research Methodology and Research Methodology for Computer and Systems Sciences) have contributed significantly to the thesis. I got interested in the data science and machine learning area from a course called "Data Mining" during my master's at the Department of Computer and Systems Sciences. That course taught me data science and machine learning on a deeper academic level.

After working on this for about five months and at the same time educating myself in a subject that I previously did not master, I can hand it over with pridefulness. The result has provided self-discipline and knowledge about Data Science and Customer Segmentation by choosing a path outside my comfort zone. My thesis and the master's degree will be the ultimate aid to a career in information technology. This thesis has been an incredibly worthwhile undertaking. I have immensely enhanced my English skills and ability to process large volumes of information such as academic writing and research papers. Additionally, I now have a better understanding of scientific writing.