Erasmus
School of
Economics

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

Enhancing Retail Marketing Initiatives through Customer Segmentation and Predictive Analytics

Name student: Ioannis Anagnostelis

Student ID number: 693494ia

Supervisor: Vahe Avagyan

Second assessor: Dr. NM (Nuno) Almeida Camacho

Date final version: 22/10/2024

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Acknowledgements

With the completion of this thesis, my journey in the Data Science and Marketing Analytics Master's program at Erasmus University is also completed. In line with this, this research seeks to capture the knowledge and skills I have gained over the past year on this program and demonstrate my ability to use these tools to solve complex, real-world problems.

First, I would like to thank my thesis supervisor Mr. Vahe Avagyan, for his mentorship and guidance throughout the journey, since his feedback and insights have been essential in shaping both my direction and completion of my thesis.

Secondly, I would like to thank my friends and peers on the program Mike and Kemal, who truly believed in me and encouraged me meanwhile this journey and together spent countless hours of studying. Words cannot describe how grateful and blessed i fell to have these two persons in my life.

Last, I would like to thank and express my unlimited appreciation to my family for their support and love during this difficult and challenging period of my life, even if they live far away from me.

Abstract

In the rapidly evolving retail industry, businesses rely heavily on data-driven techniques, to understand customer behavior, in order to optimize marketing strategies, and minimize attrition rates. This thesis explores in detail the application of data mining methods, particularly RFM analysis, K-Means clustering and Random Forest to enhance customer segmentation and provide actionable personalized marketing insights.

Leveraging a retail dataset, this research identified five groups of households of different purchasing values, based on their Recency, Frequency and Monetary traits. A second round of clustering with K-Means on two out of the five of the identified groups, generated six distinct personas based on detailed purchasing behavior, coupon redemption and degree of engagement with marketing efforts. In addition, the Random Forest model was used to forecast households churn and revealed that income level, household size and campaign engagement, were among the most impactful predictors of attrition.

This study contributes to the academic comprehension of customer segmentation by introducing a dual-layer approach of clustering with K-Means, which integrates RFM metrics and personas for more nuanced and tailored marketing initiatives. By reinforcing the principles of Relationship Marketing Theory, it additionally demonstrates the potential of tailored, data-driven marketing to build loyalty and enduring relationships with customers.

Practically, the findings offer managerial insights by providing specific tailored marketing initiatives for the identified persona, on how to boost engagement of the at risk and low engaged customers, while on focusing on the weaknesses of potential loyalists to convert them into premium customers, ultimately maximizing business profits.

Keywords

Customer Behavior, Purchasing Patterns, RFM, K-Means, Random Forest, Personas, Churn, Marketing Initiatives, Relationship Marketing

Table of contents

1. Introduction

1.1. Background Information

Retail is a key pillar of the economy, including all activities related to the direct sale of products and services to consumers, for personal, professional and household use as noted by Bankim & Vaja, (2015). Operating as the last link in the chain of distribution, retailing connects distributors with the end user via channels, such as brick-and-mortar shops, supermarkets, internet marketplaces, direct sales, mail orders, and vending machines, by purchasing large quantities of products and selling them in smaller quantities to satisfy the constantly changing demand of the consumers. This market consists of a wide array of goods and services, including groceries, apparel, healthcare items, entertainment products and many more.

The expansion and evolution of retail over the years, have brought various difficulties to the surface. More precisely, consumers are getting sometimes confused by the variety of options for shopping, which can restrict their engagement and interest, possibly leading to churn. Further, the shift to self-service and automation by many stores has led to a gap between the service given and the expectations of consumers, who are constantly looking for personal attention. On the other hand, for retailers, high market saturation, which is characterized from the large number of retailers competing for the same clientele, sometimes results in decrease in prices, that further diminishes the margins for profit. Moreover, stores have the tough task of integrating attempts to promote their brands, receive feedback, and enable transactions via modern technologies like websites, social media, and online sales systems among others. Therefore, the retail industry has to keep developing and expanding, to maintain customer engagement, ensure profitability, and meet the evolving demands of a technologically linked market.

1.2. Retail Marketing Transformation

Retail marketing was traditionally focused on the efficient distribution and availability of products, by ensuring that the products were always available, at the right time and in the right place, at the best possible price. However, the last decades it has evolved and transformed from its original format, to incorporate and follow the fast-paced changes both in technological innovations and the ongoing customer purchasing behavior, by adopting a more customer-centric approach. This approach is facilitated by advanced data mining techniques, such as customer segmentation and predictive analytics, by integrating traditional marketing tactics with a deeper understanding of customer behavior. This entails analyzing purchasing habits, demand and needs, forecasting churn, and tailoring marketing efforts to individual preferences, by leveraging large databases. This is mirroring a broader trend in which data-driven information is used to create personalized marketing efforts, as also stated by Mulhern (1997). This broader transition is closely associated with the emergence of retail marketing as an academic field of study.

Although the industry is constantly evolving, there are still quite few obstacles that must be overcome to completely understand consumer behavior across different data sources and environments, as "behavior" is a complex factor that depends on many different other sub-factors, such as demographics, personal preferences and beliefs, psychological factors and financial status. All of these are often very difficult to understand and manage by a business, as they require a large and detailed data, as well as specially designed robust analytical tools. Although retailers have adopted such tools over the years, there is still a need for further in-depth research on different data sources to provide a coherent understanding, also pointed out by Kumar (1997) and Pressey & Mathews (2000).

An insightful empirical evidence paper by Alawadh & Barnawi (2024) presents a thorough framework specifically designed to improve market performance indicators in the retail sector using advanced data analysis techniques such as clustering. This research gives a strong emphasis on the need for powerful and more effective tools that can manage large and complex data to identify not only current consumer behaviors but also to predict future trends, which is vital for retailers aiming to remain competitive in a rapidly evolving market. Moreover, it emphasizes the need of continuous study and development in this subject to more effectively combine and use data from various sources, which is necessary for obtaining a more coherent knowledge of consumer behaviors.

1.3. Research Objectives

The main objective of this thesis is to investigate in depth the integration of data mining and predictive analytics methods within the theoretical framework of Relationship Marketing Theory, in order to uncover hidden insights on how such methods can help marketing initiatives to fulfill the customers' demands and at the same time boost engagement and loyalty levels and reduce churning. More specifically, through customer segmentation with the K-Means algorithm, in combination with the RFM analysis, this research will try to detect and categorize customers into distinct subsets based on similar characteristics, thus enabling the identification of the most valuable and less valuable customers. Based on the insights obtained from the segmentation, the research will create detailed customer profiles or personas, which will actually represent the representative customer within each segment, characterized by certain purchasing behaviors, preferences and needs.

Furthermore, by leveraging predictive analytics and more specific the Random Forest Machine Learning algorithm, this paper seeks to develop a model that can effectively forecast customer churn. Such objective aims to understand and identify key factors influencing the choice for churning and therefore detect at-risk customers. The findings of this model allow proactive engagement strategies aimed at enhancing retention rates, which strengthen long-lasting interaction.

Together, each one of the above-mentioned goals seek to strengthen the main purpose of enhancing retail marketing tactics under the framework of Relationship Marketing Theory. This study employs sophisticated data analytics to provide insights that support businesses to build strong and durable customer relationships, aligning with the principles of this theory, which emphasizes the need of long-term engagement and loyalty. This study not only offers useful insights helping businesses thrive in a competitive market, but also extends the theoretical understanding of Relationship Marketing, by showing how sophisticated data-driven methods could enhance durable interactions in the age of data.

Having established the research objectives, this thesis will actually try to address the following question:

How can advanced analytics enhance the understanding of customer behavior and purchasing patterns across datasets in the retail sector, and how can targeted marketing strategies be employed to optimize customer engagement and retention rates?

To expand the research, the following research sub-questions would also be explored:

RQ1: In what ways can the integration of RFM analysis and K-Means clustering provide a detailed framework for customer segmentation that optimizes purchasing pattern analysis in the retail sector?

RQ2: In what ways do customer personas, developed through dual-layer clustering, enhance the effectiveness of personalized marketing strategies in driving relationship-driven engagement and loyalty?

RQ3: What are the key predictors of customer churn identified through predictive analytics, and how can these insights inform targeted retention strategies in retail?

1.4. Significance of the Study

This thesis, through empirical evidence on the data of a retail store, seeks to help businesses that are focused on recognizing the different characteristics of customers and their unique purchasing behavior, to further enhance their targeting and marketing efforts. Using advanced clustering and predictive analytics methods analyzed in this research, a business can engage customers more efficiently, forming specialized and targeted actions that will perfectly match their needs and wants, thus ensuring greater efficiency, reduction in attrition and potential increasing their profits.

More specifically this thesis demonstrates how the integration of methods such as RFM and K-Means clustering can effectively segment the customer base, enabling more accurate identification and categorization based on three critical metrics: Recency, Frequency and Monetary value.

Furthermore, followed by further clustering and persona construction, from the clusters created through this integration, the study reveals individual distinct customer purchasing patterns within each identified segment. Forming these exclusive profiles is required to fulfill the specific needs and preferences, which, in turn, permits the delivery of customized marketing initiatives and messages through strategic marketing efforts. This customer-centric strategy guarantees that marketing initiatives are well matched with customer demands, therefore promoting the growth of strong relationships in line with the concept of Relationship Marketing.

This study uses also Random Forest, a predictive analytics model, to identify churners and above all to understand the reasons behind their choice to leave. Businesses may therefore deliberately adapt to retain these customers, hence enhancing general engagement and loyalty. By matching marketing strategies with data-driven insights, this study helps businesses achieve sustainable development. Apart from enhancing the accuracy of segmentation and prediction, the strategy in this paper shows the way on which data-driven approaches within Relationship Marketing may result in higher levels of profits and long-term engagement.

On the other hand, by including a two-step clustering technique into a traditional RFM structure, this paper presents a conceptual framework for segmentation and persona construction from an academic perspective. This approach builds on, yet significantly differs from, existing research in consumer segmentation, therefore addressing the apparent simplicity and lack of academic innovation.

Existing literature has mostly examined these techniques individually or using single-layer clustering frameworks. In the present study RFM values for the customers are clustered and illustrated with K-Means and with an additional round of K-Means customer profiles or personas are created, achieving a level of segmentation granularity not typically found in current research. This dual-layer clustering technique expands Relationship Marketing Theory and offers a greater comprehension of customer interaction. This segmentation technique is actually in line with engagement and loyalty as it helps to develops marketing initiatives that are both tailored to the target audience and appealing.

In addition to segmentation, this study uses predictive analytics through the application of Random Forest model, to identify the factors that drive churn. This technique contributes to conceptual theory by revealing key predictors of customer churn, thereby expanding the theoretical understanding of loyalty and retention.

Collectively this thesis is built upon the Relationship Marketing Theory. Serving as a framework for understanding customers behaviors and boosting engagement and loyalty. Through the methodology constructed, this research seeks to address practical challenges in marketing but also expand the theoretical insights of the theory. Achieving such goals can improve customer relationship management both in retail sector but also in different context, ultimately contributing to business practice as well as to academic literature.

The following sections of this paper have been structured as follows:

At first, the review and reference of important scholarly works defining and exploring the theory introduces the major theoretical ideas, notably Relationship Marketing Theory. This establishes the framework that forms the basis for the analysis and results this work presents.

Second, through an investigation of relevant academic publications related to fundamental thesis components like customer segmentation, clustering techniques and predictive analytics, the current study aims to fill in gaps in the existing body of knowledge.

Third, the dataset will be described in detail, as well as the procedure of data manipulation, and justification for the selection of methodologies, including RFM analysis, K-Means clustering, and Random Forest models for attrition prediction,

Fourth, the section of analysis and results will present a thorough review of the customer segmentation, the building of personas, and the creation of a machine learning model for churn prediction. This part presents the thorough study, which clarifies the procedures used throughout the investigation and the method of generating and interpreting the data.

At last, the section of discussion and conclusion will highlight the way these findings connect to the theoretical framework, therefore addressing the implications for Relationship Marketing Theory and offering suggestions for further study areas.

2. Theoretical Background

2.1. Relationship Marketing Theory

Relationship Marketing Theory as denoted by Berry in the paper “Emerging perspectives on services marketing” in 1983, is the practice of “attracting, maintaining and enhancing relationships” between businesses and customers. This approach emphasizes the evolutionary nature of customer interactions, including customer acquisition and retention and fostering of relationships on the long term. Core principles of the theory include, services development, customization of relationships, service augmentation, loyalty-driven pricing and employee engagement, as clearly demonstrated in the papers Berry (1995) and Berry (2002).

Further development of this theoretical framework conducted in the paper of Morgan & Hunt (1994) and Palmatier et al. (2006), demonstrated that customer loyalty and engagement are greater achieved by sustained and mutual beneficial interactions rather than solely transaction made by the customers, pinpointing central extended concepts of trust, commitment and personalized engagement, which collectively build strong and endurable connection with customers. On top of that, Vargo et al. (2004) argue that unique interactions help establish value, thereby stressing the importance of meeting the particular needs and preferences of each consumer in the framework of relationship marketing efforts.

Expanding knowledge of data analytics has changed relationship marketing so much that it prioritizes the need of data-driven insights more and more. While qualitative elements like trust remain vital, modern marketing gains from quantitative techniques that let businesses to segment more precisely their customer base and predict actions that can affect loyalty over time. By using advanced data driven approaches like clustering, marketers successfully segment customers and create highly focused marketing tactics Dolnicar & Leisch (2017); van Doorn et al., (2010). This is in line with the basic concept of Relationship Marketing, which stress the need of ongoing, trust-based customer interactions Berry (1995)

Studies show that predictive analytics, such as churn prediction, enables companies to proactively address customer retention, supporting the theory’s focus on sustaining engagement Ascarza, (2018); Neslin et-al. ,(2006). This thesis uses current advances to illustrate the way current data-driven approaches can improve and extend conventional Relationship Marketing practices, thus ensuring that marketing strategies are both tailored and strategically aligned with the particular needs and preferences of individual customers Payne et al. (2005); Reinartz & Kumar (2000).

2.2. Targeted Marketing and Customer Behavior

Based on the transformative impact of advanced analytics, optimizing marketing strategies involves carefully designed initiatives through data-driven decision-making. In this framework, real time personalization is critical, since it offers marketing messages to consumers based on their personal behavioral patterns and preferences.

This theory is supported by the findings of Guan et al. (2018) who emphasized the importance and effectiveness of tailored marketing, by revealing that households exposed to targeted coupons were more responsive to price reductions and purchased significantly more products than those without exposure, meaning that even a small decrease in price by a coupon, led to an increase in quantity of purchases. More specifically, it showed that purchase rate was increased by 5.73 units per week among the exposed households, in contrast to 0.67 units per week among the unexposed ones. Also, via this paper was provided strong evidence that targeted marketing promotion campaigns in retail, influence category-level food purchases, particularly when promoting less healthy food categories over healthier ones. That indicates that targeted contribution can boost sales affect preferences and interest towards healthier alternatives, by pointing the focus of marketing initiatives. Even if customers do not redeem the coupons they received, it has a significant effect on triggering their interest, the so-called "exposure effect", as mentioned by Venkatesan & Farris (2012) in their paper "Measuring and managing returns from customized coupon campaigns to retailers".

Additionally, the findings from the paper of Kallier Tar & A Wiid (2021), which came out through a survey of 103 participants aging 20 to 40 years, showed that financial initiatives and personalization, are important factors driving customer behaviors. Particularly, 89 % of respondents stated that they are more likely to buy a product if they receive a discount voucher from a retail store at the time of checkout, while 59% said that discount associated with their current purchases would affect their purchasing decisions. In addition to real-time offers, they pointed out that well-timed, personalized campaigns influence behavior, particularly when they include offers based on prior transactions. More specifically, 72 per cent of the respondents said that they would actively engage with a retailer's campaign if it included tailored promotions reflecting their past purchases or preferences.

The findings of these papers demonstrate and describe in detail the impact of real-time personalization and financial incentives on purchasing behavior. This thesis seeks to extend these insights by applying clustering to both identify and refine customer segments into detailed personas, in order to gain greater personalization. Since Relationship Marketing Theory emphasizes the need of engagement and loyalty by gaining a deeper understanding of consumer needs, the present paper aims to link segmentation with tailored marketing incentives that can be more effectively personalized by utilizing data-driven personas.

2.3. Customer Segmentation

Customer segmentation is a cornerstone of marketing practices, which enables businesses to effectively cater for the distinct consumers demands. In the past, segmentation was a method of dividing a broad customer base into smaller, manageable classed as per the customer's common characteristics. Initially, segmentation like that of Smith Wendell (1956), was majorly focused on demographic data-age, income and gender. However, the importance of different other factors such as lifestyle, ethical standards and behaviors has come to the forefront, shifting segmentation from demographic categorization to advanced behavioral clustering approach.

Analyzing the behavior of customer along with purchase history and engagement patterns, offers another level of understanding deeper, customer habits and preferences. This methodology is demonstrated by Gil-Saura & Ruiz-Molina (2009) who classified consumers according to perceived relational advantages like trust, social advantages and special treatment, acquired through engagement with retailers. This study highlights the need of analyzing emotional and social dimensions of consumer groups compared to conventional demographic approaches.

However, additional study is needed to validate these methodologies in different market environments and customer segments. This approach aligns directly with this research's objective to employ advanced analytics for more nuanced customer segmentation by focusing on behavioral data and relational advantages, as it facilitates the creation of marketing strategies that are closely tailored to individual customer needs and behaviors, offering a more sophisticated understanding of consumer dynamics and enrich the academical literature.

2.4. RFM

One of the most well-known and effective methods for behavior-based segmentation is the RFM analysis, which evaluates the customer value based on their transactional behavior and in particular, how recently customers made a purchase (Recency), in what frequency (Frequency), and the monetary value of those purchases (Monetary).

In 1998, Claudio Marcus on his paper shifted the traditional RFM analysis (frequency, frequency and monetary) into a customer value matrix, thus introducing an innovation in segmentation. By streamlining the RFM analysis, this matrix focuses on two critical variables: the average purchase amount and the quantity of purchases. This approach facilitates the implementation of personalisation and simplifies the process by classifying customers into four actionable segments: Frequent high spenders, Frequent low spenders, Infrequent high spenders, and Infrequent low spenders. The findings of this study indicate that such an approach offers small businesses with the capacity to adopt data driven and efficient strategies that improve marketing and retention outcomes even with limited resources. The feasibility of this approach is particularly beneficial for businesses with restricted analytical capacities.

Another study by Chen et al. (2009), which used a dataset from a Taiwanese supermarket, developed an RFM-Apriori algorithm and found interesting trends in consumer purchasing behaviors that are frequently overlooked by traditional methods. The findings showed that incorporating RFM criteria into sequential pattern mining, can effectively filter out low-value patterns, thus focusing on high-value customer groups. The firm was able to more accurately target the behaviors of key consumer segments and adjust marketing efforts. As a result of this approach and as Miglautsch (2000) explained in his study, the RFM in general aids in identifying the values, for example the high-value customer groups that are most likely to respond positively to targeted marketing initiatives.

RFM analysis is directly align and further reinforces the conceptual theory this thesis is built on, since the identification and determination of customers value, as highlighted also on these papers, tends to be extremely useful for understanding the level of engagement of customers and the relationship with the business.

2.5. Integration of RFM with K-Means

To improve the accuracy and the efficacy of customer segmentation, earlier research has successfully combined the RFM analysis with the clustering techniques, such as the K-Means clustering. Gustriansyah et al. (2019) for instance, in a study on retail environment found three distinct customer segments: Recent Low Spenders, Moderate Shoppers, and High-Value Regular Purchasers by means of such integration. More specifically, in relation to the transaction data of a pharmacy, by concentrating on particularly valuable customer groups, the effective implementation of the optimal clustering, valuated by the Silhouette Score, showcased the stability of the method in enhancing segmentation, therefore allowing precise marketing strategies and efficient inventory management.

In order to investigate customer purchasing behavior, Anitha & Patil (2022) also utilized K-Means clustering in combination with RFM scores, therefore strengthening the validity of clustering again with the Silhouette Score. The study identified three main customer subsets: low-value consumers, characterized by infrequent purchases and lower spending; moderate-value consumers, indicated by medium expenditure and buying frequency; and high-value customers, characterized by frequent purchases and higher spending. By applying RFM and K-Means clustering, businesses gained the ability to tailor their marketing initiatives on the segments with the greatest potential profits.

Integration of RFM with K-Means based on the papers above, enables a more precise segmentation approach, which is essential for the theoretical framework of this paper, since it allows the identification of customer groups based on purchasing behavior. By clustering these behavioral insights, the framework ensures that marketing efforts are aligned with better recognized and targeted customer needs, thus enhancing the ability to build long-term customer relationships through data-driven engagement.

2.6. Data-Driven Personas

The research undertaken by Arian et al. (2021) implemented a comparable clustering methodology, within the transportation sector, employing K-Means, hierarchical clustering, and decision tree analysis to categorize users into personas with respect to their social and demographic characteristics and travel behaviors. This research showed that the efficiency of customizing treatments depending on these personas on behavior modification supports even more the use of clustering methods in the formulation of tailored tactics.

Similarly, by employing non-negative matrix factorization (NMF), An et al. (2018) remodeled the process of persona building by dividing social media data into several demographic and behavioral groups. The research revealed that this approach could effectively create groups of customers with certain habits and preferences, also called data-driven personas, therefore enabling the personalizing process and helping businesses to enhance engagement, for example by focusing on the regular high spenders or infrequent low spenders.

Furthermore, in 2017 Miaskiewicz & Luxmoore demonstrated how the combination of qualitative observations with quantitative data may improve the evolution of personas in corporate environments. Their approach greatly improved the complexity and usefulness of personas by tying them to actual stories and actions, therefore making these personas not only statistically relevant but also rather practical.

Building on current approaches, this thesis offers a novel conceptual framework for customer segmentation and persona construction by using a dual-layer clustering within a standard RFM structure. This study differs from other studies by adding a K-Means clustering step for certain groups, therefore creating comprehensive personas with a deeper understanding of customer behavior. This technique streamlines past models by offering a more precise methodology suitable for many sectors, therefore eliminating sector-specific modifications and improving their clarity.

For instance, Pramono et al. (2019) employed a multi-step clustering process with RFM analysis and Customer Lifetime Value (CLV) for a deeper financial evaluation. In this study the customers are initially segmented into subsets based on their RFM value with Hierarchical Clustering (Ward's Method). This approach focuses on forming large, general segments based on the recency, frequency and monetary scores, while minimizing variance within each subset. The segmentation has then been refined by the addition of CLV metric, which projects future value of customer. Finally, the K-Means clustering algorithm is applied to further break down the larger clusters into smaller ones, enabling more actionable segments based on both past behavior (RFM) and future potential (CLV).

While Pramono et al. (2019) restrict the analysis to the refining of customer groups using CLV, the framework in the present thesis takes a step further by applying K-Means clustering twice: first on the original RFM scores and subsequently within certain segments. The second K-Means round helps to create specific personas from the broad segments, therefore offering more significant behavioral insights. Ultimately, this results in a more granular and successful approach of creating engagement and loyalty than Pramono et al.'s concentration on profitability as it helps the creation of more focused and customized marketing tactics. Beyond simple segmentation to persona-driven marketing solutions, this innovative approach provides academic value as well as practical application.

Lastly, the study of (Sheikh et al., 2019) is particularly closer to the conceptual framework of this thesis study, since it employes a two-step clustering methodology of K-Means in combination with LRFMP (Length, Recency, Frequency, Monetary, and Periodicity). The key differences though, lie in the narrow industry focus and the introduction of additional variables of Length and Periodicity , which makes the model less generalizable to broader sectors. Also, while the dual-layer approach allows to refine high-value customer segments, their methodology is highly focused and specialized for Fintech customers, limiting it capability for application across diverse environments. This present thesis research takes a broader, sector-independent approach by using a traditional RFM model paired with two rounds of K-Means clustering, The focus on persona creation after an initial RFM-based segmentation makes it more versatile for application, providing a deeper understanding and more tailored marketing strategies. This method advances the theoretical foundation of Relationship Marketing by merging advanced data analytics with the customizing of relationships, therefore enabling businesses to create more solid, data-driven client engagement plans.

2.7. Predictive Analytics and Churn

Churn prediction, which refers to forecasting the rate at which customers stop using a service or a product, stands as a critical aspect of predictive analytics. This is significantly affecting the profitability and growth, since retention is often costlier than acquisition. The necessity of such predictive models is a result of the complexity of consumer behavior, which is actually influenced by different factors, like the changes in purchasing patterns, competition, and the level of satisfaction of consumers.

Through an extensive case study with a telecommunications company, conducted by Lejeune (2001) showed that Machine Larning techniques are quite useful in estimating attrition as they have great capacity to manage large datasets and find complex, nonlinear patterns in data. As proven by Vafeiadis et al. (2015) and Lalwani et al. (2022), highly advanced machine learning models achieve accuracy rates greater than the traditional approaches like Logistic Regression and Naïve Bayes. In particular, AdaBoost, Decision Trees, and Random Forests regularly achieve rates greater than 80 %, indicating more accurate predictions.

On top of that, the integration of Machine Learning models with other analytical techniques can identify early indicators of customer dissatisfaction. Bojei et al. (2013) underscored that this synergy enables businesses to effectively leverage consumer feedback insights, thereby increasing the overall customer experience and enhancing customer loyalty. More specifically, research done by Larivière & Van Den Poel (2005) which provides empirical data of the application of machine learning to enhance customer retention and profitability forecasts in financial services, make use of regression forests for continuous variables like profitability and Random Forest for binary outcomes like purchase or churn. Using data of 100,000 customers, it investigated the traits related to consumer behavior, demographics, and intermediate effects. Based on significant improvement in expected accuracy, the results show that these sophisticated techniques outperform traditional models, and the variables such as the function of intermediaries and past customer contacts significantly impact purchase selections and churning rates.

Based on the results of these research, this thesis demonstrates the effectiveness of using machine learning techniques, specifically Random Forest, to get outstanding accuracy in churn prediction compared to traditional models. By accurately spotting at-risk customers, the recommended approach helps to control customer attrition proactively. This supports the thesis's goal of improving customer retention within the Relationship Marketing Theory. This study helps create targeted retention efforts by identifying attrition causes and enhancing customer engagement and loyalty.

3. Data

3.1. Data Description

This paper uses a dataset called "The Complete Journey" offered by a worldwide leader in customer data science. It was obtained via Kaggle[1], a well-known data science competition online platform, where users exchange datasets and compete to solve data driven challenges, by using mainly advanced data mining techniques. This dataset contains information for 2.500 shoppers at a certain anonymous retailer, for a period of two years. Among this information are details of each household's transactions in different product categories, making it a rich source for analyzing purchasing behavior and patterns.

Additionally, it includes a range of tables including demographic and transaction data (e.g., purchase events, items bought, discounts applied), information on marketing campaigns and the coupons redeemed by every household. The dataset contains 35 variables including. transactional data (e.g., total spending, frequency of transactions) and demographic factors (e.g., household size, income). Churn is the response variable for this study, indicating if the household stopped interacting with the retailer throughout the observed time.

3.2. Dataset Cleaning and Manipulation

To prepare the dataset for further study, extensive data cleansing and modification were carried out. Since the data was obtained in the form of several tables, as was already noted, the goal was to create a cohesive and integrated dataset from many sources. This initial stage of enabling future analysis of the included information is thus the merging of several tables.

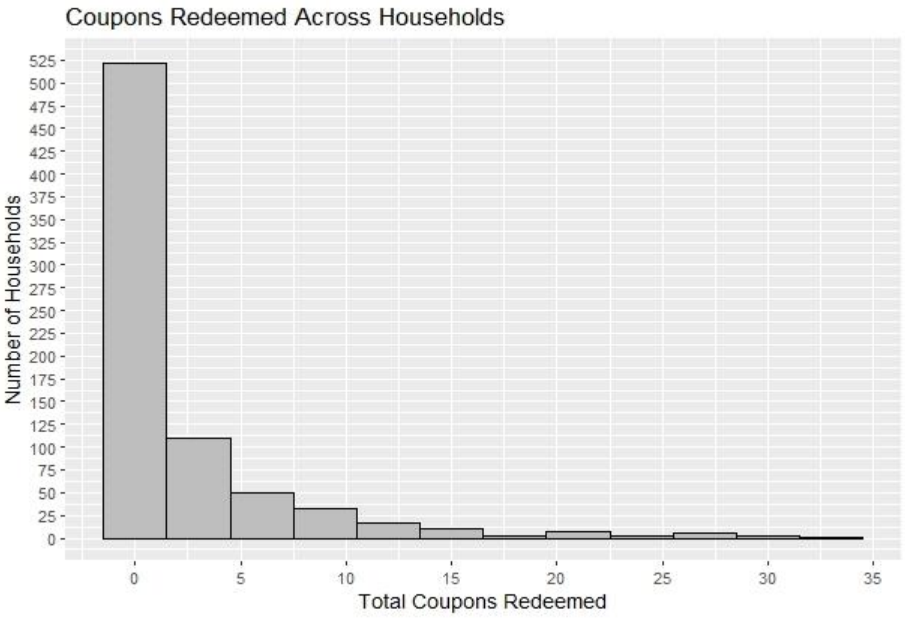
The first stage entailed combining the tables by using common key identifiers. This procedure was performed with considerable attention and detail, into six different sub-merging phases before reaching the final merged dataset, in order to avoid any loss of important information for each unique households across all transactions, always seeking to properly identify and interpret every operation and result emerged.

During the combination of the first data tables, namely campaign and coupon data with demographics dataset, several further actions implemented to capture more intricate aspects of the information provided. More specifically, new features were created such as, the total number of campaigns per household, the duration of each campaign, the total number of unique coupons per product and the total coupons redeemed, with the purpose to offer a comprehensive understanding of the manner on which households engage with marketing initiatives, and the way these initiatives actually affect their

[1] Dunnhumby - The Complete Journey (kaggle.com)

purchasing behavior. The total coupons redeemed for instance, can be used to give insights on the efficiency of different promotional efforts, enabling a targeted marketing approach. As it is shown in the *Figure 1* below to be more specific, a right-skewed plot is illustrated, meaning that the most households even did not redeemed coupons at all, or they redeem maximum five coupons within two years. This indicates that the majority did not engage with the retailer's coupon-based marketing efforts, supporting the research of this thesis to investigate the purchasing patterns ang engagement in individual level, through the creation of personas, in order to provide to retailer practical suggestions, in the format of marketing initiatives on how to enhance their efforts, tailored to personal expectations.

Figure 1: Coupons Redemption



Also, the analysis effectively handled missing values in the data, where variables with gap in information in the corresponding linked data tables, such as the total number of campaigns households participated in, the duration of different types of campaigns, and the total number of coupons per product, were replaced with zeros for the numerical variables and "None" for the categorical variables. An example table (*Table 1)* is created below, to show the changes made.

This choice has been made due to the fact that these missing values were not random, but frequently represented important or relevant exclusions. For instance, a unique household that was not targeted from any campaign or did not redeem any coupons. Following such approach guaranteed that following analysis would neither overestimate nor incorrectly interpret household engagement as a result of missing information

Table 1: Missing Values Replacement

Variable	Missing Count (NAs)	Replacement
Campaign ID	41	None
Campaign type	41	None
Total campaigns	41	0
Duration of campaigns	41	0
Coupons redeemed per product	490	0
Total Coupons Redeemed	490	0

In order to ensure further that the data used in the analysis is efficient and clean, unnecessary columns were removed, such as START_DAY, END_DAY and PRODUCT_ID. These columns, while necessary for the initial merging processes and calculation of important features, like the total number products purchased or the duration of each campaign, they didn't provide any further information, thus they were omitted to reduce the noise and computational time during analysis.

Then, by further exploring the dataset, some gaps were found, in particular in the "DEPARTMENT" variable, where some categories appeared without description and others as "unknown". To manage these categories, they were categorized under a unified label, "DEPARTMENT_UNKNOWN".

Also, the "Grocery" department, which contained an excessively high number of entries, was divided into sub-departments, for instance, Dairy, Beverages, and Snacks. This segmentation was essential, since the department "Grocery" was dominating in total purchases (31.2%), including many different types of products, which can actually belong in different departments, while all the other departments had a share lower than 18 %, see *Table 2*. This segmentation helped to avoid dominance and offer a more complete view on consumer preferences and behavior, allowing further deeper study of purchase trends across a range of product categories.

Table 2: Share of Departments

No.	Department	Total Purchases	Share
1	Grocery	215.411	31.2 %
2	Drug GM	118.195	17.1 %
3	Produce	89.026	12.9 %
4	Meat-PCKGD	58.466	8.47 %
5	Meat	53.401	7.73 %
6	Deli	35.639	5.16 %
…	…	…	…
44	Housewares	1	0.000145%

The next step in the data manipulation section involved the creation of dummy variables, the categorical variables were actually converted to numerical, for every department type. This approach allows a wide range of statistical, clustering and machine learning techniques to be implemented, since the most of them benefit from numerical input.

Additionally, every value in department level from the dataset was then converted into fraction of the total number of products purchased in order to facilitate fair comparison among households of different classes and purchasing frequencies. This type of normalization (standardization) stands as crucial in the analysis, since reflects the relative engagement of customers (households) with different product categories, thereby helping to focus on behavioral purchasing patterns rather than purchasing volumes.

Building upon the previous steps occurred in the manipulation, the next transformation involved the combination of all the department categories into refined groups created. More precisely, departments were categorized into four main retail stores areas (departments), classified based of their common or related category of products. This approach tends to make the dataset more manageable and easier to be analyzed and further enhance the ability to identify trends and purchasing patterns in individual level.

Now the new grouped departments consist of:

- **Grocery Essentials** (DEPART_GROCERY_ESSENTIALS). This group combines basic and routine grocery items, typically covering everyday necessities from food to hygiene for daily household needs.

- **Perishables & Fresh Foods** (DEPART_PERISHABLES_FRESH_FOODS)**.** This grouped department focuses on fresh and frequently purchased foods essential for immediate consumption, providing insights into health habits, dietary preferences and the premium consumers that are willing to pay for foods based on their quality and condition, and more specific foods that are being recently made or harvested.

- **Health & Wellness** (DEPART_HEALTH_WELLNESS). This group shows consumers' health awareness and wellness trends, since it includes products bought for health and personal care.

- **Specialty & Miscellaneous** (DEPART_SPECIALTY_MISC)**.** This diverse group includes items such as floristry and garden supplies, gourmet foods from the salad bar, items from gas stations and entertainment categories, as well as a broad assortment of miscellaneous items, capturing insights into discretionary and niche market expenditures.

Next, in the process of the following mergers, several columns were renamed, to provide clarity to the reader and making it easier to understand the data fields. For example, the variable MARITAL_STATUS_CODE, was first renamed to MARITAL, and then the levels included were converted from “A”, “B”, “U” to “Married”, “Single” and “Unknown”.

Additionally, while merging the tables it was noticed that although the transactions table contained detailed information on 2,500 households, about the products they bought, the amount they spent and the distinct transactions they made, the information about demographics in the corresponding table was missing. This means that the information on demographics corresponded to only 801 households, creating a problem in aggregating and creating a single data set. To tackle this issue, the households with gap in the demographic table were filtered out keeping only the ones had complete information across all linked data sources. By focusing only on unique households with complete demographic profiles, the study could ensure a higher quality of data analysis, especially in the cluster analysis, centered around fully characterized observations.

Finally, after all the subprocesses were taken, as mentioned above, the final dataset was ready to be formed, after merging all the sub tables. The final dataset, was consolidated by removing all redundant columns, properly handling missing values and creating new, relevant attributes. Including a wide spectrum of elements vital for understanding consumer behavior, response to marketing efforts, and demographic consequences, this last version is clean and detailed fragmented data into a potent analytical tool ready to assist the sophisticated analysis and models needed for this thesis. The dataset is illustrated in *Appendix [1]* and is labeled *Table 3*.

4. Methodology and Motivation for Methods Selection

This thesis uses advanced analytical techniques like unsupervised and supervised learning methods to apply segmentation and predict customer behavior within the retail store. The primary goal is to draw actionable insights that can actually enhance marketing tactics and improve customer engagement and retention. The RFM model is first used specifically to calculate important metrics on the dataset and based on these metrics, the K-Means clustering technique is used to cluster the data. Following the initial clustering, certain clusters were identified as particularly relevant or insightful and selected for further analysis to explore their distinct characteristics in more depth, which then led to the creation of detailed customer personas. Finally, the Random Forest, machine learning algorithm is employed on the initial dataset to identify the most important variables contributing to churn prediction. Though they are not necessarily the main factors directly driving churn, these factors are very important in helping to pinpoint customers with higher likelihood to churn. The flowchart of the methodology is illustrated as *Figure 2* in Appendix 2.

The RFM model (Recency, Frequency, Monetary) is widely used for customer value analysis because of the simplicity and ability it offers over useful insights based on customer behavior. RFM's strength lies in its focus on behavioral data and enables easy segmentation based on past purchasing patterns, which is ideal for retail contexts, where recent purchases and spending volume and frequency are key indicators of future engagement. For marketers that want to quickly grasp and react to consumer behavior, this model is absolutely essential as it measures involvement and loyalty without requiring sophisticated financial models. Chen et al. (2009) and Gustriansyah et al. (2019) highlighted the importance of RFM in enhancing segmentation precision, as it focuses on historical data patterns that reflect customer commitment. Focusing on transactional data, the RFM model is adaptable for K-Means clustering as shown in the theoretical background discussed below. Thus, in this theory, it is the first stage of analysis.

K-Means was chosen due to its efficiency and ability to handle large datasets. For environments like retail, K-Means is computationally scalable and as Aggarwal et al. (2014) points out it is especially useful when dealing with numerical data like RFM scores. Its interpretability makes it ideal for customer segmentation, where stakeholders need actionable groupings. It offers clear, distinct clusters that are easy to interpret and implement, making it a more practical choice for this study's goal.

The novelty of the methodology lies in the application of two rounds of K-Means clustering, where in the first-round customers are grouped into macro segments while the second-round dives deeper into selected clusters to create more granular personas. These steps ensure that the broad segments produced by RFM clustering are refined into useful profiles for targeted marketing. By means of this dual-layer method, over-segmentation, a risk connected with hierarchical models, is prevented , and stable, interpretable clusters are created allowing insightful analysis of customers behavior. Unlike other models like GMM or DBSCAN, which can handle noise and complicated data frames, K-Means strikes a compromise between computing efficiency and understandable results, therefore enabling their application in a practical marketing environment.

The Random Forest algorithm was applied to forecast attrition, because of its ability to handle high-dimensional data and make reliable predictions with inherent feature relevance rankings. Although logistic regression and other traditional models are sometimes used in attrition analysis, they often lacking to fairly capture the nonlinear interactions inherent in retail customer behavior. Achieving accuracy rates of over 80%, Vafeiadis et al. showed in their 2015 study that Random Forest consistently outperforms conventional models in the prediction of customer retention. It is also rather helpful for determining the main causes of turnover as it helps companies to prioritize feature significance and therefore target at-risk customers. Unlike other machine learning models, Random Forest also offers out-of-bag (OOB) error rates, which give an unbiased estimate of the model's performance, increasing confidence in its predictive power.

In conclusion RFM, two-step of K-Means clustering and the Random Forest model in this thesis are motivated by their complementary strengths. Together, they form a scalable, interpretable, and highly accurate framework for understanding customer behaviors and predicting churn, making them ideal for the different sectors.

4.1. Methods

4.1.1. RFM

The methodology begins with the application of RFM (Recency, Frequency, Monetary) analysis, a widely recognized technique in customer value assessment. This method comes from the practice of direct marketing in sales companies catalogs in the 1960s but explored in more detail in the paper "Strategic Database Marketing" of Arthur Hughes. This method was actually developed and used originally in the industry and then further refined within the academic community. This analysis predicts a segmentation of customers in the company's database based on past behavior. RFM analysis, is defined as:

- **Recency (R)**: "how recently a purchase was made by the customer", with a shorter time since the last purchase indicates higher engagement.
- **Frequency (F):** "how often purchases are made by the customer", with a higher frequency indicating greater customer loyalty and casual purchasing behavior.
- **Monetary value (M):** "what is the total amount of money a customer has spent during a period", with higher monetary values suggesting higher customer value.

After calculating these parameters for each distinct customer, traditionally the classic approach segments customers into quantiles based on scores from 1 to 5 for each RFM metric. This thesis though, adopts a different strategy. Instead of pre-determined scoring, it leverages the raw RFM values directly for further analysis. The K-means algorithm groups customers into clusters based on Recency, Frequency, Monetary and then analyzes the average values of these metrics within each cluster. This approach offers a flexible and adaptive framework for comprehending consumer behavior and value, therefore facilitating the dynamic labeling and identification of subsets.

4.1.2. K-Means

K-Means clustering is an unsupervised algorithm, which partitions dataset into K distinct non overlapping clusters, by minimizing the variance within each cluster while optimizing the distances between data points and their representative cluster centers or centroids. Hence, it tries to solve that problem: $minimize_{C_1,..,C_k}\left\{\sum_{k=1}^{K} W(C_k)\right\}, where$

- $\boldsymbol{K}$ represents the total number of clusters the dataset is divided to, and it is a predefined number ranging from $\boldsymbol{k = 1,2,\ldots,K}$
- $\boldsymbol{C_k}$ denotes the k_{th} cluster, which is the subset of the dataset and contains a group of data points that are closer to each other, than to the points in other clusters
- $\boldsymbol{W(C_k)}$ is the within cluster sum of squares for cluster $\boldsymbol{C_k}$ which measures the variance within the cluster, by calculating the sum of squared distances between each data point in the cluster and the cluster's centroid.

For the K-Means based on the squared Euclidean distance, the cluster center consists of the column-wise mean values across all the data points of the cluster. The column-wise mean thus, represents the average response sequence for all clustering variables among all data points of the cluster. The elegance of this model lies in its simplicity and efficiency, since data points can be divided into subsets, in that way where each subset consists of data points similar as possible to each other, and dissimilar to data points belonging the other subsets.

The process of clustering a data set based on k-means is actual easy, as long as the number ($\boldsymbol{k}$) of the resulting clusters is predetermined. The main process of this algorithm is the following:

1. Firstly, the initial $\boldsymbol{k}$ centroids should be determined, one for each cluster. These initial centroids must be chosen carefully, because different starting points for the centroids, might give different results.
2. The next step is selecting each data point from the dataset and correlating it with the centroid closest to it. When this is done for all data points in the dataset, the first step is complete and a first and “rough” clustering has already been performed.

3. Then $\boldsymbol{k}$ new centroids are required to be recalculated, which will be the center for each cluster resulting from the previous step. $\boldsymbol{\mu_K} = \frac{\mathbf{1}}{|\mathrm{C_k}|} \sum_{i\in C_k} \boldsymbol{x_i}$, $where$ $|\boldsymbol{C_k}|$ the data points in cluster *K*
4. After the new centroids are defined, the same process of assigning each of the data points to the nearest centroid, is repeated, but now for the new centroid. The result of this repetition is that, at each step, the centroids are repositioned, and the data points are assigned to the appropriate cluster, each time based on the nearest centroid.
5. When in some iteration no permutations are noted data points, then the execution of the algorithm terminates.

The algorithm aims to minimize an objective function, the so-called squared error defined as:

$$minimize \sum_{k=1}^{K} . \sum_{i\in C_k} . [\![\boldsymbol{x_i} - \boldsymbol{\mu_k}]\!]^2 \ , where$$

$[\![\boldsymbol{x_i} - \boldsymbol{\mu_k}]\!]^2$ is the squared Euclidean distance used to measure the distance of each element $\boldsymbol{x_i}$ in cluster $\boldsymbol{C_k}$ from the centroid $\boldsymbol{\mu_k}$. The index $\boldsymbol{i}$ runs across all the data points that have been assigned to each cluster implicitly, and the sum of $\boldsymbol{k}$ accumulates this distance for all clusters from 1 to $\boldsymbol{K}$.

Although it can be shown that the algorithm always terminates, it is worth emphasizing that it does not successfully always find the optimal solution, since the algorithm is significantly affected by the initial centroids. For this reason, it is often recommended to execute it several times until this effect is reduced. To formalize this, the within cluster variation for each cluster $\boldsymbol{C_k}$ is defined as:

$$W(C_k) = \frac{1}{|C_k|} \sum i \in c_k \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \ , where$$

$\boldsymbol{\bar{x}_{kj}}$ represents the mean of the ***j***th feature within **k,** and $|\boldsymbol{C_k}|$ denotes the number observations in cluster $\boldsymbol{k}$, and $\boldsymbol{p}$ is the number of features. The overall objective though, is to minimize the summed variance across all clusters: $minimize \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i\in C_k}. \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$

This is verifying that each cluster is compact as possible, and the centroids represent the center point of their respective clusters effectively. The process is repeated until the cluster assignments no longer change, indicating that a local optimum has been reached.

4.1.3. Cluster Validation

Cluster validation is one of the most important steps when diving datasets into distinct segments, since it is serves as a quality measure, by verifying that the segments were not formed due to chance. Its importance lies in the ability to provide metrics that help to check the quality of the clustering, showing how well data points within each cluster relate to each other compared to those in different clusters. Based on the clustering algorithm and the nature of the data the appropriate indices can be selected for validation, calculate these indices, and interpreting the results.

For instance, indices like the Elbow method, Silhouette index, Hubert index, and Dunn index assess the compactness and separation of the clusters.

As mentioned by Shi et al., (2021) in the paper "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm" the **Elbow method** is an approach which estimates the optimal number of clusters in the dataset by plotting the sum of squared errors (SSE) against the number of clusters and finds the point where the curve makes an "elbow". This point is actually indicating where the gain in the variance explained drops off, suggesting that while adding more clusters the clustering outcome does not improve.

The elbow is typically visualized by plotting: $SSE(k) = \sum_{i=1}^{k}. \sum_{x \in C_i} \| x - \mu_i \|^2 \; , where$

- $\boldsymbol{k}$ is the number of clusters $C_i$is the set of points in cluster $\boldsymbol{i}$
- $\boldsymbol{x}$ is a data point
- $\boldsymbol{\mu_i}$ is the centroid of cluster $\boldsymbol{i}$

Furthermore, based on the research paper of Rousseeuw, (1987)**,** the **Silhouette index**, measures how close each point in one cluster is to points in neighboring clusters, providing a value that quantifies how well each object has been classified, resulting into efficient interpretation of cluster analysis. The Silhouette Index formula is:

$$s(i) = \frac{b(i) - a(i)}{\max\,(a(i), b(i))}, where$$

- ***a(i)*** is the average dissimilarity of point ***i*** to all the other points within the same cluster and
- ***b(i)*** is the minimum average dissimilarity of point ***i*** to points in a different cluster.

Objects with a high Silhouette value are considered to be well clustered and those with low values can be outliers or misclassified.

The **Hubert index** *Γ* therefore as provided by (Hubert & Arabic, 1985) is typically evaluated by using a comparison of the silhouette scores with a measure of the distance between clusters. The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. The calculation of the Hubert index can be formulized as follows: $\Gamma = \sum_{i<j}(x_i x_j s(i) s(j))$ *, where:*

- x_i *and* x_j are indicators that are equal to 1 if the respective objects *i* and *j* belonging to the same cluster and 0 otherwise.

This equation assess the correlation of the silhouette scores of all pairs of objects belonging in the same cluster.

Meanwhile, the **Dunn index** as provided by Dunn, (1974)**,** focuses on identifying the smallest distance between observations in different clusters relative to the largest internal cluster distance. This index is especially useful for identifying compact and well separated clusters, where a higher Dunn index value indicates better clustering quality, and it is calculated using the following formula:

$$DI = \frac{min_{1\leq i\neq j\leq k}\delta\left(C_{i,}, C_j\right)}{max_{1\leq i\leq k}\,\Delta(C_i)}\,, where:$$

- $\delta\left(C_{i,}, C_j\right)$ is the distance between cluster C_i and C_j, which can be the minimum distance between any two points in the two different clusters.
- $\Delta(C_i)$ is the maximum internal cluster distance within the cluster C_i, also known as the diameter of cluster C_i

The goal is to maximize this index in the evaluation of the cluster solutions, aiming for high separation between different clusters, and tightness within the clusters.

4.1.4. Random Forest

The Random Forest is well-known machine learning algorithm, introduced by Breiman, (2001), which integrates the power of multiple Decision Trees, to enhance accuracy and prevent overfitting.
The algorithm starts by making several subsets from the training data by using the bootstrapping method. This method is actually randomly selecting samples from the training data with randomized replacement, as consequence, some observations to be duplicated in each subset, while others may be excluded. Each of these different subsets operate as the training set for a separate individual decision tree.

To dive in more detail in the Bootstrapping method, let a dataset $\boldsymbol{D}$ of size $\boldsymbol{N}$ including both features and the target variable: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$.

For each of the trees in the forest, a bootstrap sample D_i^* is created by sampling N instances with replacements from D. Each element (x_i, y_i) in D has a $\frac{1}{N}$ probability to be selected for each N draws in D_i^*. Mathematically is represented as: $P\left((x_i^*, y_i^*) = (x_j, y_j)\right) = \frac{1}{N} \;\; for\ all\ j = 1,2,\ldots,N$

During the growth of each tree, Random Forest employs a random selection of features instead of using all features to calculate the optimal split at each node. This number of features, commonly represented as $\boldsymbol{mtry = M}$, a model parameter that changes depending on the total amount of features accessible. Typical options for $\boldsymbol{mtry}$ are defined as the square root $\sqrt{M}$ or one-third $\frac{M}{3}$ of the overall number of features. Next, the decision tree uses these features to determine the optimal separation according to specific criteria such as, the ***Gini index*** for classification tasks, and **variance reduction** for regression task. ***Gini index*** is actually a measure that quantifies the impurity or purity of a set of elements, and it is mainly used in decision trees in order to determine the best point to split the data at each node. The Gini index is calculated as:

$$Gini(D) = 1 - \sum_{i=1}^{J} p_i^2 \text{ ,} where$$

$\boldsymbol{D}$ is the dataset at a node**,** $\boldsymbol{J}$ is the number of classes, and $\boldsymbol{p_i}$ is the proportion of class $\boldsymbol{i}$ in the dataset at the node. A Gini Index equals to 0 denotes perfect purity, indicating all elements in the node belong to a single class. The higher Gini Index, up to 0.5 in a binary classification, it means higher purity or in other words evenly split among different classes.

The previously mentioned procedure is executed iteratively until certain conditions are met, such as achieving a maximum depth of the tree or a minimum size of the nodes. The procedure is iterated in order to generate a predetermined quantity of trees, each distinct as a result of the irregularity in both the data subsets and the selected features for divisions. After building the forest of trees, the model proceeds to generate predictions on the new data. In classification tasks more specifically, every tree in the forest submits a vote for a class, and the class that receives the highest number of votes is selected as the prediction, while in regression tasks, the model estimates the mean of the outputs generated by all the trees.

An inherent characteristic of Random Forest is the ability to assess its accuracy by utilizing the **out-of-bag (OOB)** error technique. Every tree is evaluated using the data, which is not included in its bootstrap sample, referred to as **out-of-bag (OOB)** data. The evaluation of the prediction error for each observation is based exclusively on the trees on which the particular observation was out-of-bag (OOB). The out-of-bag (OOB) error, is defined as the average of these errors across all observations, offers an unbiased evaluation of the model's performance. The OOB error provides an unbiased estimate of the classification error rate as:

$$OOB\ Error = \frac{1}{N} \sum_{i=1}^{N} 1\,(\hat{y}_i \neq y_i) \text{ ,} where$$

$\hat{y}_i$ is the predicted class for the ith OOB sample, y_i is the actual class, and 1 is the indicator function that is equal to 1 if $\hat{y}_i \neq y_i$ and 0 otherwise.

Additionally, Random Forest offers critical insights into the importance of each feature. The present approach calculates the level of the prediction inaccuracy that emerges from the permutation of feature values across the out-of-bag (OOB) samples. An increase in OOB error signals the significance of the feature, and these statistical measurements of relevance are aggregated across all trees to furnish a comprehensive viewpoint.

4.1.5. Data Sampling

To ensure the effectiveness of supervised classifiers, training sets should precisely reflect the events they are meant to replicate. This entails the creation of separate, homogenous groups within the data and include all relevant characteristics. In this manner though, in order to prevent skewed analysis and biased learning, balancing the distribution of the data stands to be vital. Imbalanced datasets compromise the ability of the classifier to precisely identify minority classes by preferring majority classes. For this reason, it is crucial to analyze the sensitivity of the classifier with respect to the distribution of sample training. Diverse methodologies for altering distributions are suggested in the literature of the respective classes such as the ***Oversampling method***, in which extra data is added training for all classes until they have an equal number of observations to the majority class. Specifically, ROSE method generates synthetic instances by sampling from a kernel density estimate of the feature space for every class using a smoothed bootstrap technique. These synthetic samples, which are not perfect replicas, are produced using the features of the existing data points. This is usually achieved by random sampling or more sophisticated techniques such bootstrapping or kernel density estimations. This helps in creating more realistic synthetic data points that help the model learn better by not overfitting on repeated instances as denoted in the paper by Menardi & Torelli (2014).

4.1.6. Evaluation metrics for predictions

Churn prediction is a binary classification problem where customers are classified as either "Churned" or "Not Churned." This makes the application of a confusion matrix of great relevance since it helps the model to be accurately evaluated in respect to these two categories.

The evaluation metrics for predictions include precision, recall, F1 score, and accuracy which can be generated in the confusion matrix, as also suggested by Miao & Zhu (2022). An example of the confusion matrix and an explanation of the terms, see *Table 4*, it consists of would be provided in order to characterize the metrics used.

Table 4: Confusion Matrix

		Actual	
		Not churn	**Churn**
Predicted	**Not churn**	True Negative (TN)	False Negative (FN)
	Churn	False Positive (FP)	True Positive (TP)

A true positive (TP) is a customer that the algorithm correctly predicts that will fall in churning category and then actually churns. On the other hand, a true negative (TN) result from a customer being faithfully projected to be a retained customer. False positive (FP) is the situation when a customer is expected to churn but actually does not leave. Lastly, a false negative (FN) is a customer who is expected to stay while existing in the churning group.

Accuracy is among the most often used measures of a machine learning method and it stands as the proportion of correct predictions among the whole sample. The formula provided below is utilized for its calculation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Nevertheless, the retailer is likely to incur a greater expense if it attempts to predict that a customer will not churn when the customer actually does (false negative). The retailer could develop anti-churn strategies in advance to retain customers if the model has a low false negative and a high recall, thus recall is also incorporated as an evaluation metric. Recall is the ratio of the total number of actual churn instances to the proportion of churn instances that were correctly predicted (TP). The formula for recall is shown below:

$$Recall = \frac{TP}{TP + FN}$$

The percentage of customers expected to leave out of the overall count is known as precision. Reducing false positives by using of precision aids to lower the unnecessary expenses and resources capitalized to churning prevention for customers who are incorrectly identified to belong in the churning group. Precision is computed with the formula below:

$$Precision = \frac{TP}{TP + FP}$$

Last, the F1 score, could be actually a better indicator than accuracy, since the dataset shows an imbalanced mix of churning and non-churning customers. For researchers seeking to decrease the number of false positives and false negatives, F1 is the harmonic mean of accuracy and recall. The formula for F1 score is:

$$F1\ score = 2 \times \frac{Recall\ \times Precision}{Recall + Precision} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

5. Analysis & Results

5.1. RFM Clustering

As mentioned in the methodology section of this thesis, the first phase of the analysis involved the generation of the RFM metrics for each household in the dataset, serving as indicators of their engagement and loyalty with the retail store. The RFM analysis has proved to be a really simple but at the same time significant marketing tool in establishing the value of the customer, since it evaluates three crucial metrics: Recency, Frequency and Monetary. However, before the evaluation of these metrics, all the transactions in the dataset were filtered, retaining only those with a positive sales value, which indicates transactions had actually occurred. Additionally, since the dataset obtained consists of demographics exclusively for 801 households, the rows with missing information (such as age, marital status and income) where removed, to maintain a clean and consistent dataset for the analysis.

Moving to the computation of the metrics, the Recency value was initially assessed by calculating the difference between the last purchase and the final day of data's record, for each household. This metric is actually determining the level of engagement of the household with the retail store. Next, the Frequency value was calculated, by counting the total number of unique transactions occurred within the record period (up to 712 days) for each household, showing the consistency of purchasing and the level of loyalty. Last, the Monetary was computed through the aggregation of the total amount spent in every unique transaction for each distinct household, which demonstrates the capacity and willingness to pay, see *Table 5* in *Appendix [1]*.

5.2. K-Means Clustering

After calculating and establishing the RFM metrics, the next phase of the methodology was the utilization of the K-Means clustering algorithm, based on the Recency, Frequency and Monetary values of each household. This approach is actually partitioning the 801 unique households in the dataset into subsets, based on their similar RFM characteristics, thereby every subset reflects a group of households with shared value.

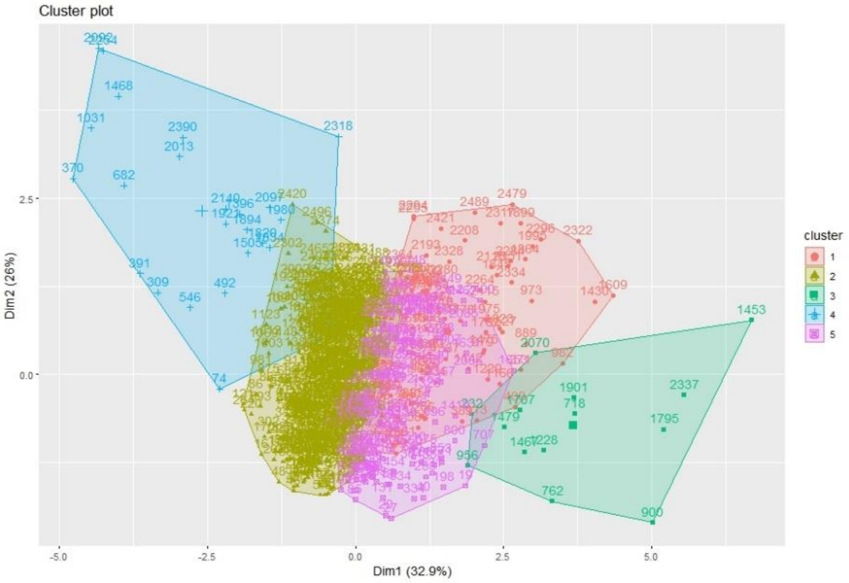
Before applying the K-Means algorithm, several important steps were undertaken, involving the creation and preparation of the dataset for clustering in order to ensure the validity and correct application of the method. After computing the RFM metrics, the relevant columns were extracted to form a new dataset and further scaled, to normalize the data and control potential bias in clustering due to scaling discrepancies, since the K-Means operates based on the Euclidean distances among data points and generates clusters accordingly.

Next, to determine the ideal number of clusters centers to be used, the K-Means model was employed with different values of cluster numbers, ranging from 1 to 20 and an Elbow Plot was created. This approach, as shown in the *Figure 3* (*Appendix [2])*, suggested that five was the best suited number of clusters for this algorithm, since after this point the curve starts to get flatten and is forming the "elbow". This result demonstrates a balance between the compactness and separation, as the Sum of Squared Errors (SSE), which implies as the total variance explained inside the data, does not significantly increase, by the inclusion of more clusters after this point.

Further assessing the validity and robustness of the clustering analysis, the NbClust tool in R was utilized. This powerful tool includes a range of indices among them the Dunn index, Hubert index and Silhouette score, which are helping to identify a significant "knee" or peak, indicating the ideal clustering structure, by analyzing variations in intra-cluster and inter-cluster distances.

After the calculations and the results implying for each index, the tool uses a majority voting system to indicate the optimal number of clusters to be used, ensuring that no single index dominates the decision, and then generates two plots for Dunn and Huber Index, see *Figure 4 & Figure 5*, in *Appendix [2]*. Both indices pointed out the 5 clusters being the optimal solution, since the peaks in both second differences highlight that this point best compromise between tight intra-cluster distances and well-separated clusters.

Figure 6: K-Means Clusters

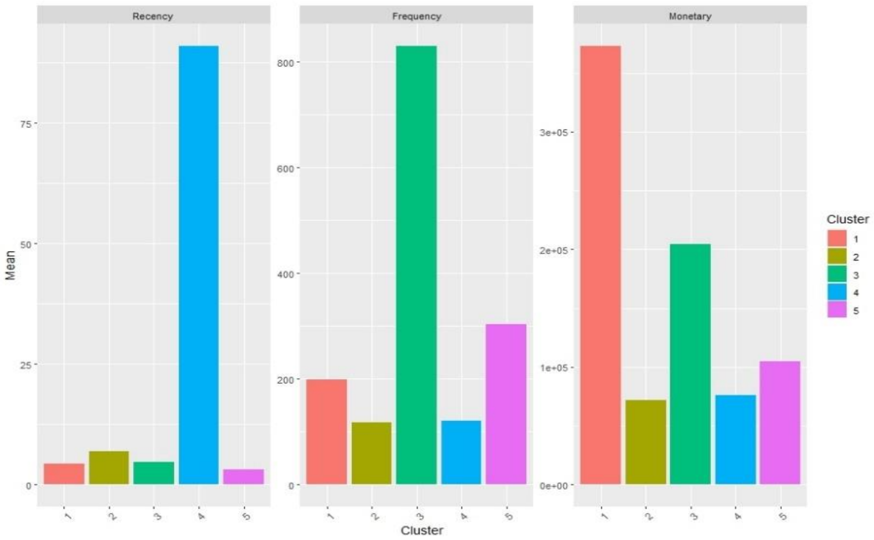


Upon the implementation of the algorithm the clusters were successfully generated, capturing 69.5 % of the variance explained within the data, see *Figure 6*. This portion shows the ability the algorithm has to effectively capture the basic patterns and correlations between the variables, therefore implying a really strong clustering. In the default visualization of the K-Means the clusters appeared to overlap when plotted against the original variables.

This overlap happens due to the high dimensionality in the data, where two-dimensional plots do not accurately represent the distances and separations between clusters. Thus, to assess the results and verify the integrity of the clusters, a three-dimensional plot was conducted, *Figure 7 (Appendix [2])*, to reduce the data from higher dimensions. This approach provides a clearer perspective of the segmentation and the spatial distribution of each cluster.

Furthermore, each household was assigned to one of the five clusters created. With a special focus on the RFM metrics, this basic classification helps the individual identification of households based on their cluster affiliation, therefore enabling the thorough study of the features defining each cluster. Bar plots were used to graphically show the average values of these measures for every cluster therefore enhancing the clarity of cluster profiles. The visual examination by *Figure 8* reveals significant distinct patterns that underscore the diversity within the cluster base, therefore enabling the manual labeling process into the dataset.

Figure 8: Means of clusters



More specifically, the households belonging to Cluster 1 can be called **"Big Spenders"** since, as shown in the plot, they appeared to spend a significant amount of money in the last two years at the retail store, in a relatively very low number of visits, and with fairly recent activity. On the contrary, households belonging to Cluster 2 seem to have quite low scores among all measurements, which makes them "**Low Engagers**".

The households included in Cluster 3 are therefore characterized by the high frequency with which they visit the store, spend a great amount of money during their purchases and have made their last purchase very recently, which made them "**Loyal Enthusiasts**".

Cluster 4, on the other hand, can be characterized as "**Dormant Shoppers**" as they appear to be disengaged with the store by showing abstinence greater than 80 days. Also, during the 2 years of record they had made minimal transactions with a very low number regarding their spending.

Finally, Cluster 5 is labeled as "**Potential Loyalists**", since it is a group with moderate frequency and monetary expenditure. Recent engagement suggests that these customers are set to become more valuable and require proper marketing practices that encourage them to make transactions more frequent and with a higher monetary value.

5.3. Further K-Means Clustering – Creation of Personas

After the formation of the initial clusters, the analysis of the cluster distribution, shown to the *Table 6* in *Appendix [1],* led to the decision of additional deeper segmentation effort, by using second round of K-Means, specifically on the segments of "Low Engagers" and "Potential Loyalists". This decision was made due to the fact that these segments were identified as the most interesting and informative among the others, since together they comprise the 83.4% of the dataset. Therefore, applying segmentation to these groups, can provide distinct profiles or personas, enabling retailer to personalize marketing initiatives to boost engagement and loyalty on the low-engaged identified personas, while gaining profitability out of the high value ones.

The clustering analysis for creating representative household personas was entailed with great attention to detail, in two separate phases, with identical procedure for both of the groups.

First, the dataset for "Low Engagers" was isolated, merged with the broader initial dataset **(***Table 1-Appendix [1]*), in order to add extra information about the households, and then adjusted to concentrate on the numerical factors essential for clustering with K-Means. This process ensured that each segment is examined with precision, allowing for tailored actions based on unique behavioral insights. Similar steps occurred for the "Potential Loyalists," emphasizing the need for a segmented approach in customer relationship management.

Afterwards, the datasets were refined by eliminating irrelevant or unnecessary variables, followed by a detailed analysis of correlations among the remaining variables. This analysis conducted a correlation plot (corrplot), see *Figure 9.1 & Figure 9.2 Appendix [2],* to visualize interdependencies and the magnitude of correlations among numerical variables. This process helped to ensure that the analysis would be based on relevant and independent variables. The correlations obtained from the corrplot models show the variables most likely to affect the clustering result, therefore guiding the choice of the most relevant variables for the analysis.

Moreover, both datasets underwent an outlier detection process with the interquartile range (IQR) technique, since the outliers might skew clustering results by affecting centroid computations. By means of identification and elimination of outliers the resulting clusters were assured to be robust and representative, avoiding extreme values that can affect the clustering. In particular eleven households appeared to be outliers in the "Low Engagers" group, and four in the "Potential Loyalists".

Last, both datasets were prepared for clustering to develop distinct customer personas by scaling the numerical variables, to normalize the data range across the datasets, as exactly occurs also in the previous clustering phase in the RFM with K-Means clustering. The K-Means clustering algorithm was applied to each group of households, both on "Low Engagers" and "Potential Loyalists", and the Elbow Method was used to determine the optimal number of clusters, revealing that three clusters were most suitable for both segments, see *Figure 10 & Figure 11* from *Appendix [2].*

To ensure the validity of these clusters, the NbClust package was again utilized in both phases, assessing the robustness and stability of the cluster structures. By verifying that the clusters accurately reflected the underlying data patterns with the three clusters, the analysis proceed to the final vital step of creation of detailed household profiles or personas. This was accomplished through carrying out an in-depth investigation of the average values of the key variables within each cluster, which allowed for precise identification of the unique preferences and behaviors of each segment, see *Figure 12 & Figure 13* from *Appendix [2].*

In order enhance the profiling process, radial plots were utilized to visually present these distinct representative personas, for both "Low Engagers" and "Potential Loyalists". Known also as spider or radar charts, these plots offer an easy comparison of the performance of every persona across several important criteria and help demonstrate clearly the strengths and possible areas for improvement of every group. In addition, the modes or key demographic features were estimated and provided below in the *Table 7* below to offer a clearer and more representative nature to the households, participated in the clustering and profiling process.

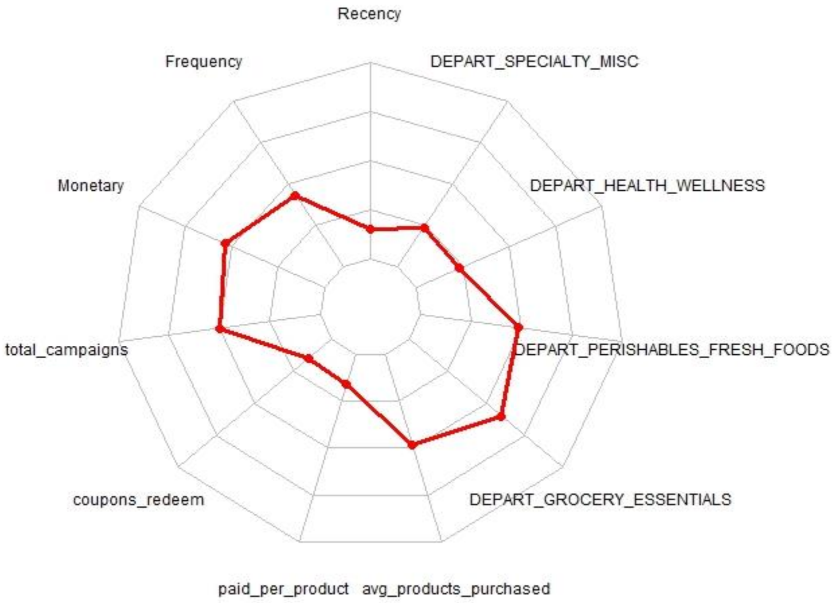
The results of clustering and interpretation of the personas are provided below[2]:

1) **From the cluster of Low Engagers:**

Persona 1: This persona represents a household of a married couple, around 45-54 years old, exhibiting low engagement with the store, as shown in *Figure 14*, since it has only 110 transactions over 712 days and recency of 7.6, suggesting selective infrequent shopping visits. This household tend to spend approximately $150,000, mainly on Grocery Essentials (44%) and Perishables & Fresh Foods (43%), revealing a clear preference for better quality, everyday products. Each transaction consisted of 16 products, spending approximately 2.5 dollars per item. Even if it was targeted by six marketing campaigns on average over the two years, this household persona shows low engagement with promotional efforts, since redeemed only three coupons on average. Thereby, this persona can be labeled as "**Infrequent Essentials**".

[2] All the values on the interpretation of personas represent mean values

Figure 14: Persona Infrequent Essentials



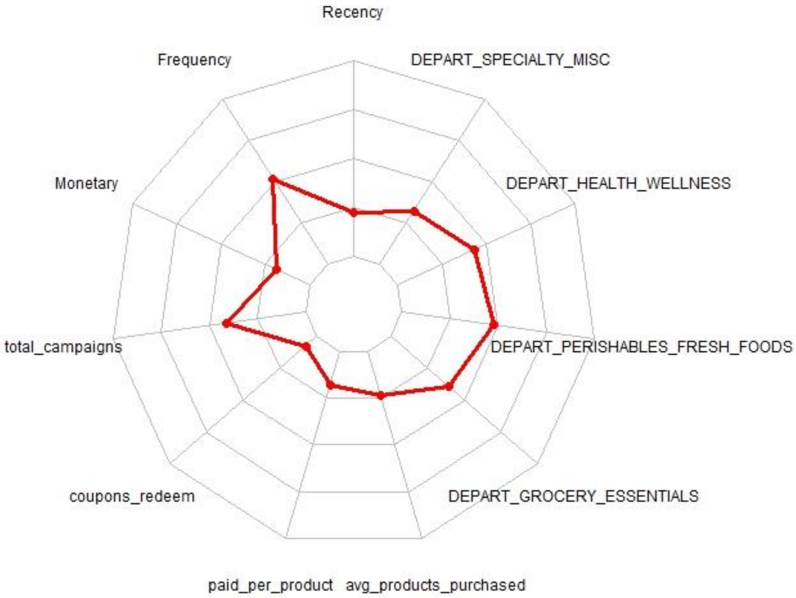
Persona 2: This household consists of two members, aged 45-54 years old and they are marked by an infrequent interaction with the store with 87 transactions and a recency of 8.73, while spending about $46,000 mostly on the departments of Perishables & Fresh Foods (49%) and Grocery Essentials (39%), indicating a taste for fresh food. They purchased a norm of 11 items at $2.25 apiece on every visit. The low average on coupon (0.6) and limited participation in 2.4 campaigns, might shows a reduced awareness. Overall, based on these mentioned this persona can be named as **"Budget-Conscious Shoppers"**, see *Figure 15 - Appendix [2]*

Persona 3: Given a recency score of 5 and 150 transactions, this persona's engagement is less frequent, and the $50,000 expenditure puts it in a lower monetary bracket. This persona represents a household consists of two members aged 45-54 years old, who spend a balanced amount on Grocery Essentials (47%) and Perishables & Fresh Foods (37%), with a notable preference also on Health and Wellness (18%), approximately $2.25 per product, and usually buys 7.3 products per visit, mirroring a restrained shopping pattern, looking for quality and health-oriented items without overspending. Participation in promotional activities (4.3) and seldom used coupons (0.5) indicate modest marketing interest. Thus, this household can be described as **"The Selective Shopper".** The radial plot can be shown in *Appendix 2 - Figure 16.*

2) From the cluster of Potential Loyalists:

Persona 1: This persona, whose age is ranging from 45 to 54 years, visits regularly the store with 380 transactions over 2 years and recency of 4.2 yet spends just $53,000. The price per purchase is $1.75, with 5 goods per visit, mostly from Grocery Essentials (35%) and Perishables & Fresh Foods (33%) departments. Also holds moderate interest in Health and Wellness (22.5%), indicating a health-conscious selection. In addition, participates in 6.25 campaigns with a 2.15 coupon redemption, indicating moderate marketing incentive receptivity. Thus, can be characterized as **“Routine Essentials Shopper”,** see *Figure 17.*

Figure 16: Routine Essentials Shopper



Persona 2: Such household profile consists of two members, aged 35-44 years, they make 303 transactions and spends on average $162,500 highlighting a potential high-value consumer. It participates in in 8.79 campaigns and use coupons (6 on average), indicating good marketing response. Also, this persona buys 10.8 goods each visit for $2.5 each, preferring high-quality Grocery Essentials (45%) and Perishables & Fresh Foods (49%). The low spending on Health and Wellness (4%) and Specialty Miscellaneous (1.26%) shows selective spending. As result of the findings this persona is named **“The Valued Frequent Shopper”,** the radial plot of this persona is illustrated in *Appendix [2]* and is the *Figure 18***.**

Persona 3: With a recency score of 3.2 and 254 transactions, this persona which included two members, aged 45-54 years, spends $75,000, positioning it in the mid-range in terms of monetary contribution. It shows a balanced shopping, dividing budget almost equally between Grocery Essentials (44%) and Perishables & Fresh Foods (42.5%), while buying on average 7.3 products per visit, giving $2.2 per product, with a focus on both value and quality, especially evident in their notable 12% spending on Health and Wellness, though minimal on Specialty Miscellaneous (1.2%). The moderate participation in campaigns (6.23) and low coupon redemption rate (1.2) indicates a cautious engagement with promotional activities. Together all these purchasing habits illustrate a **“Balanced Occasional Shopper”.** The radial plot of this persona is illustrated in *Appendix [2]* and is the *Figure 19***.**

Table 7: Modes of Demographics of Personas

Persona	Age	Size of Household	Marital
Infrequent Essentials	45-54	2	MARRIED
Budget-Conscious Regulars	45-54	2	UNKNOWN
The Selective Shopper	45-54	2	UNKNOWN
Routine Essentials Shopper	45-54	1	UNKNOWN
The Valued Frequent Shopper	35-44	2	MARRIED
Balanced Occasional Shopper	45-54	2	UNKNOWN

5.4. Predictions

The next step of the methodology migrates focus towards predictive analytics modeling to address a critical aspect of marketing, the customer attrition. Since until this point the mainly focus was on low or high engagers, through this step the research wanted to analyze additionally the churners, and the factors influenced this decision. Such approach enables a deeper understanding not just for engagement and satisfaction, but also for proactive retention strategies.

To predict households churn with Random Forest, a structured approach was taken starting with the preparation of the designed dataset for predictions, generated from the initial dataset. Within this approach irrelevant or unnecessary columns were removed, in order to focus only on the relevant variables needed. Then, a new variable for churning was created, which included the households had already churned, if they didn’t have any transaction within the last three weeks, which actually means that the had Recency larger than 21 days on the recorded period (712 days).

Furthermore, a really important procedure, known as encoding, for prediction modeling purposes was implemented. This encoding process targeted the categorical variables of the dataset, among them key demographic and campaign related features like Age, Marital, Income, Household Size and Campaign Type, and converted them into a set of binary columns, each one of them representing a category of the variable, thus facilitating the use of these variables in statistical models that require numeric input. In that way the prediction model can be further be assisted, since it ensures that all inputs are in a numeric format, which Random Forest requires to efficiently compute and evaluate splits in the data during the model training.

The dataset after this process, combines original variables, 40 in total, not subjected to encoding with the newly formed dummy variables, ensuring it is comprehensively prepared for predictive modeling.

Next, before modeling, the dataset was oversampled and shuffled, to correct imbalances and guarantee minority class presence. In accordance with the binary classification character of the model, the target variable was converted into a two-level factor, “churned” and “not churned”. The data then was split into testing and training sets, serving as fundamental step in every prediction model since it allows the model to learn and generalize from one subset (training) and then validates its predictive accuracy and robustness on an unseen subset (testing), ensuring that the model performs well in real-world scenarios. The model worked on the test data, while 70% of the actual data was selected as training data and the unselected observations as test data.

Once all primary steps completed, the Random Forest algorithm was implemented and trained on the training subset with 500 trees considering four variables at each split, see *Table 8*. This initial model uses also importance and proximity parameters to estimate variables importance and generate proximity matrices. Proximity matrices are used to measure the similarity between pairs of instances based on how frequently they end up in the same terminal node across the trees in the forest, providing insights into how instances cluster together based on their features.

The model achieves an out-of-bag (OOB) error rate equal to 16.6 %, indicating model’s prediction error rate when tested on the portion of the data not used in the training of each tree within the algorithm. This serves as an unbiased estimate, suggesting that the model correctly predicts the target variable about 83.4 % on new, unseen data.

Table 8: First Random Forest Model

Classification Random Forest		
Number of Trees	500	
No. Variables at each split	4	
OOB estimate error rate	16.6 %	
	Churned	**Not Churned**
Churned	546	323
Not Churned	158	15

In order to determine the optimal number of trees though, an out-of-bag (OOB) error rate plot was generated. This visual illustration helped to identify the point where the error rate stabilizes or reduces with an increasing number of trees, pinpoint in this model 230 tress as the optimal number, see *Figure 20*. Next, an attempt to find the optimal mtry values carried out, which represents the number of predictors at each split by iterating the Random Forest model across values from 1 to 17.

Each iteration of the OOB error was recorded and plotted, see *Figure 21*, and revealed that mtry value of 14 leads to the minimum OOB error, establishing it as the best parameter for balancing model’s complexity with predictive accuracy.

Utilizing the best Random Forest model's optimal parameters (mtry=14, ntree=230), a final model with an enhanced OOB error rate of 9.88% applied, thereby indicating better model performance. Therefore, predictions were generated on the new, unseen test data and a confusion matrix was generated to evaluate the model's performance, through accuracy, specificity and sensitivity.

Figure 20: OBB bootstrap (ntree)

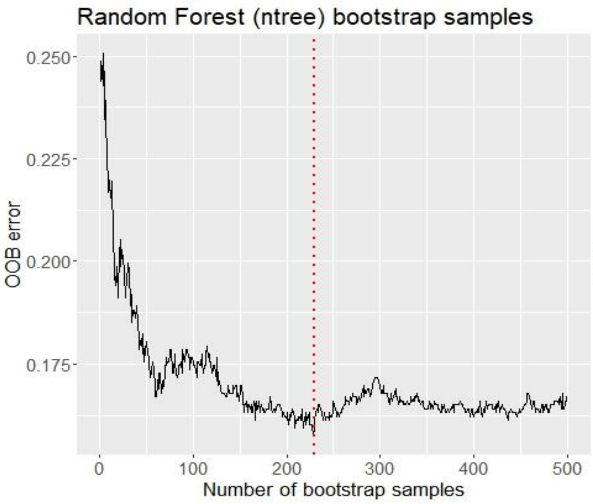
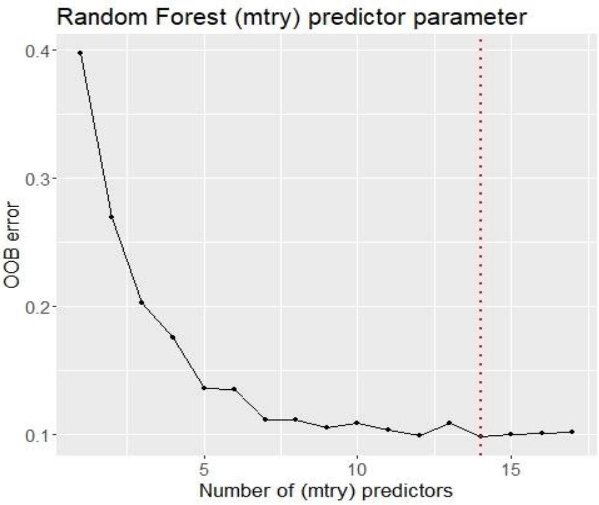


Figure 21: OBB bootstrap (mtry)



With an overall accuracy of 91.25%, the confusion matrix verifies the validity of the parameters on the chosen model, in categorizing churn. This level of accuracy shows that, in about 91 out of 100 occasions the algorithm can correctly predict whether someone will churn. Furthermore, the model's high sensitivity level of 98.67 % shows that it is very successful in spotting real churn instances, thereby helping detect most of the households who at the risk of leaving. Last, although less than the sensitivity, the specificity of 82.63 % is significant and reveals that the model around 82% of the time properly detects non-churn instances. The F1 score is approximately 0.924, indicating good balance between precision and sensitivity, while the Balanced Accuracy is 90.65%, showing that the model is consistent in its performance across both classes, see *Table 9* below.

Table 9: Confusion Matrix Results

Prediction	**Churned**	**Not Churned**
Churned	297	45
Not Churned	4	214
Accuracy	0.9125	
Sensitivity	0.9867	
Specificity	0.8263	
F1 score	0.924	
Balanced Accuracy	0.9065	

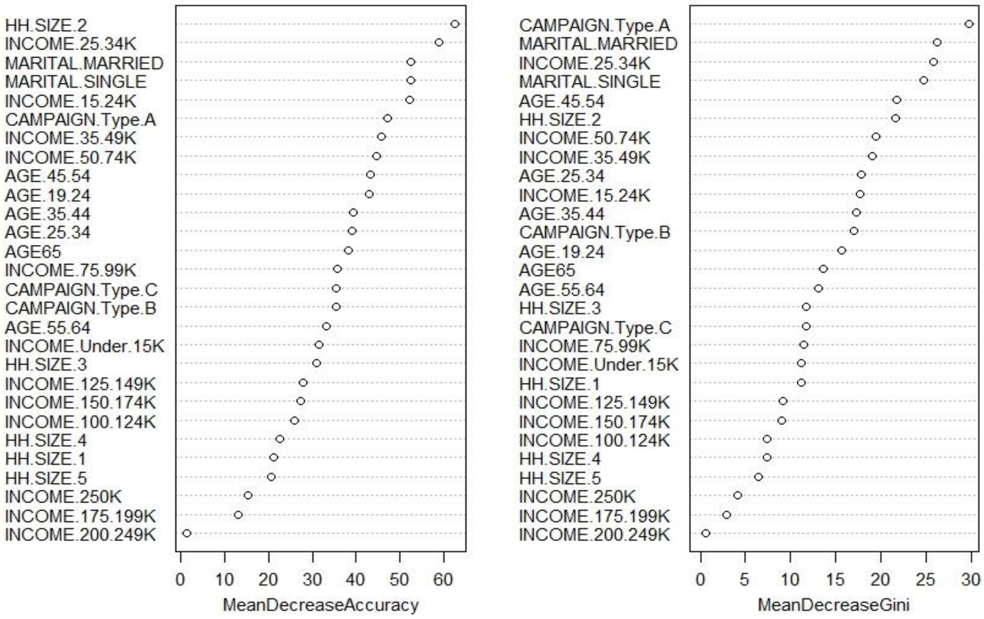
These three indicators taken together establish that the model is an effective instrument for focusing retention initiatives as it is trustworthy in differentiating between non-churning and churning households.

The analysis concludes with the illustration of the variable's importance plot, *Figure 22*. This plot is generated within the model of the Random Forest as a parameter, as previously stated, and it helps to illustrate and identify the variables that contribute the most and the least to the prediction accuracy of the model regarding churn.

Based on the plot illustrated. the variables used for prediction in the Random Forest model were ranked descending by Mean Decrease in Accuracy, indicating the degree of accuracy lost when the feature is removed from the model, and by Mean Decrease in Gini, which reflects the level of contribution of each feature to the model's ability to split the data, related to the overall predictive power. Together these two measures underline the importance and influence of features on predictions.

Diving into more details on the most influential factors of churn, can provide clear picture on how the model of Random forest can be utilized and integrated with the previously used approach of creating personas, to identify in real world scenarios, the reasons behind making such decisions. Except of only identify and interpret the variables, tailored suggestion can be further given.

Figure 22: Feature Importance Plot



Firstly, Campaign Type A appears to be among the most significant factors in prediction of churning, since as it seems in the plot households engaged with this specific campaign type show a lower likelihood to churn, while those who did not participate are more likely to leave. Targeting households with low engagement with this campaign, by creating personalized efforts like exclusive discounts, points or perks, can effectively enhance re-engagement. Such approach can be particularly useful with personas like the Infrequent Essentials or Routine Essentials Shoppers.

Furthermore, the Marital status, and especially Married and Single, plays crucial role in predictions of churn in this model, since as it looks married households might have different purchasing habits, probably being more consistent in their expenditures due to family-related needs. On the other hand, single households may be more price conscious or selective, and probably be more likely to leave in cases they did not receive offering meeting with their personal needs.

Additionally, the income brackets of 25.000$ up to 74.000$, as illustrated in the plot reflecting variables that increase the likelihood of households to stop using services or purchasing products from the store, based on the predictions of Random Forest. This finding signifying that the household belong to that segments are likely price sensitive and might churn in cases of perceiving expensive products or if there are insufficient savings opportunities. Such households are probably highly interested to discounts, promos and affordable pricing strategies.

Moreover, a household size of two people, like couples or small families, is also a notable prediction indicator for churn. This variable is actually signaling such households have more selective shopping habits in comparison to households' high bigger size, thus their needs are generally more focused, and they probably turn towards churn in cases where promotions or offerings don't actually meet their personal needs and preferences.

Last, among the most important indicators of churning based on the model is the age of the households. More specifically, households aging 25-34, 35-44 and 45-54 appear to prone more easily to churn, mirroring a different purchasing priority compared to younger ones. Older households look possibly for stability, convenience and long-term loyalty incentives in their shopping habits, rather than fresh new or experimental products.

On the other hand, the variables affecting the least the model's prediction regarding churn are mostly about household with income levels or even bigger number of members within the household. This is actually quite normal, as it shows that households that tend to earn more in general are more likely to remain engaged with the retailer, as they do not have specific expectations and are able to allocate money in order to meet their needs.

In the same manner, households including more than three members, namely households of four or five, tend to be more loyal and engaged with the store. The larger household the greater steady need for standard commodities, like groceries and household essentials, which ultimately resulting in more frequent visits and consistent purchases. This need strengthens their relationship with the store while encouraging a more lasting degree of engagement and loyalty.

6. Discussion and Conclusion

The main purpose of this thesis was to explore in detail the integration of data mining and predictive analytics techniques, to uncover hidden insights on how can be used from businesses and researchers, to understand customer behavior and ultimately enhance marketing efforts, to meet individual needs and demands. More precisely, the main research question to be addressed was:

How can advanced analytics enhance the understanding of customer behavior and purchasing patterns across datasets in the retail sector, and how can targeted marketing strategies be employed to optimize customer engagement and retention rates?

To tackle the primary and subsidiary research questions, the research through the analysis of a retail dataset demonstrated that, by segmenting customers into unique groups based on their purchasing value, effectively creating distinct personas, and further using machine learning model to identify the factors affecting the churn decision, businesses can create tailored engagement initiatives, aiming at retaining high value customers, while addressing the needs of at risks and low engagers, prior to disengage. Thus, these findings indicate the way targeted marketing strategies, drawn from analytical insights, can optimize customer engagement and retention, showcasing the practical applications and value of advanced analytics in different context.

6.1. Discussion on Findings

The systematic analysis carried out in the previous chapters led to the main findings of this research:

RQ1: *In what ways can the integration of RFM analysis and K-Means clustering provide a detailed framework for customer segmentation that optimizes purchasing pattern analysis in the retail sector?*

Integrating RFM with K-Means clustering allowed the formation of five distinct household segments based on their Recency, Frequency and Monetary measures. These segments created, such as Big Spenders, Low Engagers, Loyal Enthusiasts, Dormant Shoppers and Potential Loyalists, revealed useful and clear insights on the behavioral patterns of the households, therefore determining their actual value to the business. Each label given to these five segments, reflects the purchasing pattern and value to the retail store, over two years of transactions, making easier to identify the degree of engagement and loyalty.

The identification of both high and low value segments is essential, as it helps businesses to develop more personalized and tailored marketing initiatives and in the long term increase engagement and profitability. Such empirical findings directly support the conceptual theory designed in this thesis, by exemplifying and reinforcing the principles of Relationship Marketing theory, since the integration of these two methods proves to be effective in fostering long-term customer relationships by offering actionable insights to businesses into customer value.

***RQ2**: In what ways do customer personas, developed through dual-layer clustering, enhance the effectiveness of personalized marketing strategies in driving relationship-driven engagement and loyalty?*
In order to understand better the customer behavior and derive personalized insights on the behavioral patterns of households, the creation of personas was implemented, by running a second round of K-Means clustering in two specific groups generated from the integration of RFM and K-Means in the preceding step. The selected groups for this segmentation were the Low Engagers and Potential Loyalists, since these groups were containing more information in terms of engagement and potential future loyalty, among the others. Hence, under the umbrella of these two segments, unique different from each other personas were created, by including additional information beyond the RFM metrics, like purchasing patterns, coupon redemption and campaign engagement. The six personas created were the Infrequent Essentials, Budget-Conscious Shoppers, The Selective Shopper, Routine Essentials Shopper, The Valued Frequent Shopper and Balanced Occasional Shopper, and they represent unique representative household types.

The findings from these personas help businesses to understand in detail unique purchasing patterns and behaviors, therefore exposing both strengths and weaknesses over time. This allows the development of marketing initiatives meant to satisfy needs and preferences of the target audience, therefore fostering stronger interactions, boosting engagement or even influencing behaviors more effectively. For example, personas seen to show inconsistent purchasing behavior or poor engagement, could be targeted by special offers or incentives, meant to boost interaction. Loyalty programs, on the other hand, can help to boost the interaction between the company and its high-value personas offering profitability. This empirical evidence is directly linked with the conceptual framework of this thesis aligning personalized initiatives with the principles of Relationship Marketing theory, reinforcing the emphasis on sustained engagement and loyalty.

***RQ3:** What are the key predictors of customer churn identified through predictive analytics, and how can these insights inform targeted retention strategies in retail?*
Sine the prior steps were successfully divided the households into insightful segments and further partitioned two of these segments to create personas, the next phase of the analysis entailed the employment of predictive analytics to identify churners and the most influential factors pushing households towards this decision. This was tackled though Random Forest model, which successfully identified that, **Age** (45-54, 25-34, 35-44), **Marital Status** (Single, Married), **Household Size** (Household including 2 members), **Income** (25-34K, 35-49K, 50-74K) and **Campaign engagement** (Campaign Type A), were among the most significant key predictors. The model achieved an overall accuracy of 88.57%, which indicates its effectiveness in predicting which households were most at risk of churning.

These results confirm the conclusions of (Vafeiadis et al., 2015), (Larivière & Van Den Poel, 2005), and (Lalwani et al., 2022), thereby stressing the need of Machine Learning algorithms like Random Forest in the prediction of churn compared to traditional model, since these algorithms show great accuracy rates, often exceeding 80%. Also, the findings are in line with the conceptual framework designed in this thesis as they underscore the importance of identifying the key predictors of churn, in order to proactively act with retention strategies to address specific factors drives customers in disengagement. Businesses that utilize quarterly findings can design tailored initiatives in at risk customers to prevent churning or even to build stronger relationship in the future. For example, in this analysis, households with poor engagement may be incentivized with personalized offers or loyalty programs thereby increasing the possibility to remain loyal and not churn. The predictive ability of the Random Forest model supports the theoretical case for data-driven client retention techniques anticipating attrition and fostering long-term loyalty.

6.2. Main Contributions of the Study

This thesis contributes to the academic field of Relationship Marketing by presenting an innovative approach for customer segmentation and creation of personas. More precisely, the integration of traditional RFM with a twostep K-Means clustering offers a novel way to identify value-based customer groups and segment these groups into distinct personas, each one characterized from specific needs and behaviors. This nuanced demonstration of the value and degrees of engagement of consumers helps to create precise, tailored initiatives, therefore filling a gap in traditional segmentation techniques within the conceptual framework. Along with the fundamental concept of the theory, the empirical evidence acquired from this study expands Relationship Marketing theory by demonstrating the way in which modern data analytics enhance relationships with customers by means of customized engagement. Also, the implementation of Random Forest for churn prediction enriches the theoretical discussion, by offering a nuanced way to improve retention strategies within the Relationship Marketing framework, further strengthening customer loyalty over time.

On the other hand, the framework presented in this study is especially useful to businesses that usually have a more customer-centric approach. It offers techniques meant to enhance marketing initiatives, thus increasing engagement levels and profitability. Through segmentation of customers data, based on transactional traits and creation of actionable personas out of these segments, marketers can create tailored marketing actions matching audience needs. Predictive modeling for attrition also gives businesses the means to detect at-risk consumers and make proactive retention plans, thereby improving relationships and level of profitability.

In summary, the contribution of this research is twofold since it adds value to the body of knowledge as well as to the practical use as it creates new paths for scholarly research and motivates more study of how such approaches may be modified and used in different environments.

6.3. Practical Implications

This research offers four key marketing initiatives as suggestions for the retail store, in order to further enhance the applicability and support the effectiveness of the findings. By aligning such initiatives retailers and marketers can actually create more customized and efficient strategies for households.

Initiative 1: Loyalty Program

This initiative seeks to boost loyalty by offering rewards, benefits, discounts and personalized engagement. Personas are tailored to receive benefits (perks) meeting their purchasing habits and preferences.

- **Infrequent Essentials**:
  Offer an **"Exclusive Perks"** program which reward high value but infrequent personas with benefits, for instance exclusive product access, to increase store visits.

- **Routine Essentials Shopper**:
  Give discounts or loyalty gifts, for repeat purchases, to keep shoppers engaged in the long term.

- **The Valued Frequent Shopper**:
  Offer an **"Insider Consultation"** service With personalized recommendations and VIP access to new products, reinforcing their loyalty and high-value status.

Initiative 2: Savings and Discounts

This initiative focuses on price-sensitive household persona and encourages broadly purchasing behavior through targeted savings and cross-category incentives.

- **Budget-Conscious Shoppers**:
  Launch a "**Smart Saver Program",** which will reward shoppers with exclusive discounts or cashback on future purchases, for the total points collected from their transactions, to encourage engagement focusing on their budget-conscious behavior.

- **Balanced Occasional Shopper**:
  Offer a **"Cross-Category Explorer" bonus**, rewarding for purchases from different departments, encouraging exploration of new products and diversify their shopping patterns.

Initiative 3: Exclusive Offers

This initiative is customized for low-engagement, quality-conscious personas who tend to respond to minimal, personalized incentives.

- **The Selective Shopper**:
  Give limited-time discounts on products they usually buy, focusing on offers to boost engagement to boost engagement without overwhelming them.

Initiative 4: Churn Proactive Plans

This initiative is made to addresses general attrition threats based on income, household size, age and the campaign engagement, which were identified as impactful on churn from prediction analysis.

- **Low-Income Households** (25-34K, 35-49K, 50-74K):
  Run a **Reactivation Campaign** with discounts on essentials to re-engage those with lower income.
- **Non-Engaged Campaign Participants**:
  Offer **exclusive loyalty offers** to the households that did not respond on previous campaigns (e.g. Campaign Type A) in in order to encourage them to participate in the future.
- **Households with Two or More Members**:
  Through a **Family-Centric Retention Program**, motivate larger households to engage more frequent, by offering family-size discounts and promotions
- **Households Age** (45-54 and 25-34):
  Give **special designed content and exclusive discounts** to meet life-stage demands, cater specific age groups and reduce the possibility of attrition.

6.4. Limitations of the Study

While the results give valuable insights, several limitations should be acknowledged. First, the dataset used in this research consisted of demographics exclusively for 801 households out of 2.500 included in the table of transactions. This resulted in a problem of finding the best way to segment the households into groups, while choosing the most suitable variables for the model of K-Means clustering.

In addition, while the demographics table included the variable “Income”, which is a crucial feature for determining the potential purchasing capacity, as it represents the income bracket of the household, there was a gap in the description of the dataset, on whether this income level is annual or monthly. Thus, the implementation of this variable in clustering or even in the description of personas was not feasible. Nevertheless, the inclusion of the “Income” in prediction analysis was vital, since Machine Learning models evaluate the impact of the independent variables on the dependent one, meaning that the research was actually seeking to identify if this feature was important on the final decision of household to churn, and therefore provide marketing initiatives to prevent such occasions.

Furthermore, after the initial clustering (integration of RFM with K-Means), the resulted clusters were skewed, since the distribution of two out of five clusters was dominating. More specifically “Low Engagers” and “Potential Loyalists”, comprised the 83.4% of the dataset. This led to the decision of using only these two groups for further clustering and create of personas out of them since the included the majority of observations (information). Nevertheless, these two segments, didn’t affect the analysis at all, since the main objective of enhance engagement was still feasible.

Lastly, the original objective was to apply the Random Forest model specifically to the two out of five personas created (Low Engagers and Potential Loyalists) to predict churn and identify the factors driving the decision withing these groups. However, due to the number of households and imbalance in the cluster sizes of these two personas, the model was overfitting and was poor generalizing. To address this issue, the Random Forest model was applied by using the entire dataset of 801 households, to identify broader the factors contributing to churn. The oversampling model was used to prevent additional overfitting, therefore facilitating data balance and improving forecasting capacity for attrition.

6.5. Suggestions for Future Research

Future research could explore the effect of marketing efforts, over the long term on different environments, expanding and testing the implementation of the models by incorporating real time customer data (e.g. online engagement, social media interactions), to provide a more nuanced and up to date view of customer’s behavior, leading probably to more precise marketing strategies.

Based on the findings of this thesis attention has gained in the area of customer journey mapping and attribution analysis. Researchers can explore the way on which personas, generated from the integration of RFM with K-Means clustering, can be further enhanced by tracking touchpoints across different channels (e.g. online, in-store, social media, mobile apps). In that way businesses can target customers more efficiently; by making sure their marketing efforts are tailored on important points in the customer’s buying journey. Also, attribution analysis can help in understanding which touchpoints or channels contribute the most to conversion and retention, especially when linked with predictions of churning.

Moreover, the implementation of different clustering algorithms, such as the K-Prototypes, could effectively help future study and provide different insights, since this model incorporates both numerical and categorical variables in the segmentation process. Also, in the context of clustering, in the persona creation process, instead of applying the second round of K-Means, future studies can apply alternative options, such as hierarchical clustering or NMF (Non-negative Matrix Factorization), like the study of An et al. (2018), to offer more differentiated customer profiles, which better reflect different patterns and preferences.

6.6. Final Remark

To conclude with, this thesis study has demonstrated and underlined how important advanced data analytics is on changing and improving retail marketing potentials. By employing advanced segmentation and predictive models to grasp behaviors and properly respond to customer expectation may greatly increase marketing efficacy, thus establishing a more dynamic and customer-centric environment, while contributing to the academic and practical frameworks of Relationship Marketing.

7. References

1. *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*. (2019). IEEE.
2. Alawadh, M., & Barnawi, A. (2024). A Consumer Behavior Analysis Framework toward Improving Market Performance Indicators: Saudi’s Retail Sector as a Case Study. *Journal of Theoretical and Applied Electronic Commerce Research*, *19*(1), 152–171. https://doi.org/10.3390/jtaer19010009
3. An, J., Kwak, H., Jung, S. gyo, Salminen, J., & Jansen, B. J. (2018). Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, *8*(1). https://doi.org/10.1007/s13278-018-0531-0
4. Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785–1792. https://doi.org/10.1016/j.jksuci.2019.12.011
5. Arian, A., Pan, M. M., & Chiu, Y. C. (2021). Personas: A market segmentation approach for transportation behavior change. In *Transportation Research Record* (Vol. 2675, Issue 11, pp. 172–185). SAGE Publications Ltd. https://doi.org/10.1177/03611981211028623
6. Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, *55*(1), 80–98. https://doi.org/10.1509/jmr.16.0163
7. Bankim, M., & Vaja, R. (n.d.). *RETAIL MANAGEMENT. 2*.
8. Berry, L. L. (1995). *Relationship Marketing of Services-Growing Interest, Emerging Perspectives*.
9. Berry, L. L. (2002). Relationship Marketing of Services Perspectives from 1983 and 2000. *Journal of Relationship Marketing*, *1*(1), 59–77. https://doi.org/10.1300/J366v01n01_05
10. Bojei, J., Julian, C. C., Wel, C. A. B. C., & Ahmed, Z. U. (2013). The empirical link between relationship marketing tools and consumer retention in retail marketing. *Journal of Consumer Behaviour*, *12*(3), 171–181. https://doi.org/10.1002/cb.1408
11. Breiman, L. (2001). *Random Forests* (Vol. 45).
12. Chen, Y. L., Kuo, M. H., Wu, S. Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers’ purchasing data. *Electronic Commerce Research and Applications*, *8*(5), 241–251. https://doi.org/10.1016/j.elerap.2009.03.002
13. *Customer Segmentation*. (n.d.).
14. Dolnicar, S., & Leisch, F. (2017). Using segment level stability to select target segments in data-driven market segmentation studies. *Marketing Letters*, *28*(3), 423–436. https://doi.org/10.1007/s11002-017-9423-8

15. Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*(1), 95–104. https://doi.org/10.1080/01969727408546059
16. Gil-Saura, I., & Ruiz-Molina, M. E. (2009). Retail customer segmentation based on relational benefits. *Journal of Relationship Marketing*, *8*(3), 253–266. https://doi.org/10.1080/15332660902991197
17. Guan, X., Atlas, S. A., & Vadiveloo, M. (2018). Targeted retail coupons influence category-level food purchases over 2-years. *International Journal of Behavioral Nutrition and Physical Activity*, *15*(1). https://doi.org/10.1186/s12966-018-0744-7
18. Gustriansyah, R., Suhandi, N., & Antony, F. (2019). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470–477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477
19. Hubert, L., & Arabic, P. (1985). Comparing Partitions. In *Journal of Classification* (Vol. 2).
20. Kallier Tar, S. M., & A Wiid, J. (2021). Consumer perceptions of real-time marketing used in campaigns for retail businesses. *International Journal of Research in Business and Social Science (2147- 4478)*, *10*(2), 86–105. https://doi.org/10.20525/ijrbs.v10i2.1075
21. Kumar, N. (1997). *The Revolution in Retailing: from Market Driven to Market Driving*.
22. Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271–294. https://doi.org/10.1007/s00607-021-00908-y
23. Larivière, B., & Van Den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, *29*(2), 472–484. https://doi.org/10.1016/j.eswa.2005.04.043
24. Lejeune, M. A. P. M. (2001). Measuring the impact of data mining on churn management. *Internet Research*, *11*(5), 375–387. https://doi.org/10.1108/10662240110410183
25. Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5
26. Miao, J., & Zhu, W. (2022). Precision–recall curve (PRC) classification trees. *Evolutionary Intelligence*, *15*(3), 1545–1569. https://doi.org/10.1007/s12065-021-00565-2
27. Miaskiewicz, T., & Luxmoore, C. (2017). The Use of Data-Driven Personas to Facilitate Organizational Adoption–A Case Study. *Design Journal*, *20*(3), 357–374. https://doi.org/10.1080/14606925.2017.1301160
28. Miglautsch, J. R. (2000). Thoughts on RFM scoring. In *Journal of Database Marketing* (Vol. 8).
29. Morgan, R. M., & Hunt, S. D. (n.d.). *The Commitment-Trust Theory of Relationship Marketing*.

30. Mulhern, F. J. (1997). Research in Marketing Retail marketing: From distribution to integration. In *Intern. J. of Research in Marketing* (Vol. 14).
31. *neslin-et-al-2006-defection-detection-measuring-and-understanding-the-predictive-accuracy-of-customer-churn-models*. (n.d.).
32. Palmatier, R. W., Dant, R. P., Grewal, D., & Evans, K. R. (2006). Factors Influencing the Effectiveness of Relationship Marketing: A Meta-Analysis. *Journal of Marketing*, *70*, 136–153. http://www.marketingpower.com/jmblog.
33. Payne, A., Frow, P., & Marketing, R. (2005). *A Strategic Framework for Customer Relationship Management*.
34. Pressey, A. D., & Mathews, B. P. (n.d.). *Barriers to relationship marketing in consumer retailing*. http://www.emerald-library.com
35. Reinartz, W. J., & Kumar, V. (2000). On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. In *Journal of Marketing* (Vol. 64).
36. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In *Journal of Computational and Applied Mathematics* (Vol. 20).
37. 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. https://doi.org/10.1080/1051712X.2019.1603420
38. Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip Journal on Wireless Communications and Networking*, *2021*(1). https://doi.org/10.1186/s13638-021-01910-w
39. Smith, W. R., & Sessions, A. &. (n.d.). *PRODUCT DIFFERENTIATION AND MARKET SEGMENTATION AS ALTERNATIVE MARKETING STRATEGIES*.
40. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003
41. van Doorn, J., Lemon, K. N., Mittal, V., Nass, S., Pick, D., Pirner, P., & Verhoef, P. C. (2010). Customer engagement behavior: Theoretical foundations and research directions. *Journal of Service Research*, *13*(3), 253–266. https://doi.org/10.1177/1094670510375599
42. Vargo, S. L., Lusch, R. F., Vargo Is Visiting Professor Of Marketing, S. L., Smith, R. H., Hunt, S., Laczniak, G., Malter, A., Morgan, F., & O'brien, M. (2004). A New Dominant Logic / 1 Evolving to a New Dominant Logic for Marketing. In *Journal of Marketing* (Vol. 68).

43. Venkatesan, R., & Farris, P. W. (2012). *Measuring and Managing Returns from Retailer-Customized Coupon Campaigns*.

8. Appendix 1: [Tables]

Table 3: Final Dataset

Variable Name	Type	Description
household_key	int	Unique identifier for each household
BASKET_ID	num	Unique identifier for each shopping basket or transaction
DAY	int	Day of the year when the transaction occurred
QUANTITY	Int	Quantity of products purchased in the transaction
SALES_VALUE	num	Sales value of the products purchased (in dollars)
RETAIL_DISC	num	Retail discount applied to the purchase (in dollars)
TRANS_TIME	int	Time of day when the transaction occurred (in military time)
WEEK_NO	int	Week number of the year when the transaction occurred
COUPON_DISC	num	Discount from manufacturer coupons (in dollars)
COUPON_MATCH_DISC	num	Discount from retailer matching manufacturer coupons
AGE	chr	Age range of the household head
MARITAL	chr	Marital status of the household head
INCOME_DESC,	chr	Income range of the household
HOMEOWNER_DESC	chr	Homeownership status of the household
HH_COMP_DESC	chr	Household composition
HOUSEHOLD_SIZE_DESC	chr	Size of the household
KID_CATEGORY_DESC	chr	Presence and age group of children in the household
CAMPAIGN_type	chr	Type of marketing campaign the household was exposed
total_campaigns	int	Total number of campaigns the household participated in
duration_of_campaign	int	Duration of the campaign (in days)
total_coupons_prod	int	Total number of coupons applicable to products purchased
total_coupons	int	Total number of coupons redeemed by the household
total_products_purchased	int	Total number of products purchased by the household
paid_per_product	num	Amount paid per product after applying discounts (in dollars)
total_paid_per_trans	num	Total amount paid per transaction (in dollars)
total_discount_per_household	num	Total discount received by the household (in dollars)
avg_paid_per_trans	num	Average amount paid per transaction by the household
avg_disc_per_household	num	Average discount per transaction for the household (in dollars)
total_transactions	int	Total number of transactions made by the household
avg_products_purchased	num	Average of products purchased per transaction by the household
days_since_last_purchase	num	Number of days since the household's last purchase

DEPART_GROCERY_ESSENTIALS	num	Proportion of grocery essentials purchased out of total products
DEPART_PERISHABLES_FRESH_FOODS	num	Proportion of perishable and fresh foods purchased out of total products
DEPART_HEALTH_WELLNESS	num	Proportion of health and wellness products purchased out of total products
DEPART_SPECIALTY_MISC	num	Proportion of specialty and miscellaneous items purchased out of total products

Table 5: RFM Metrics Per Household

No.	**Household_key**	**Recency**	**Frequency**	**Monetary**
1	1	6	85	123213.382
2	7	3	59	103147.972
3	8	6	113	143827.353
4	13	3	275	167260.934
5	16	22	98	10288.087
…	…	…	…	…
…	…	…	…	…
801	2499	3	90	137010.933

Table 6: Clusters Distribution

Cluster	**Big Spenders**	**Low Engagers**	**Loyal Enthusiasts**	**Dormant Shoppers**	**Potential Loyalists**
Share	11.9 %	65 %	1.7%	3%	18.4%
No. Households	95	521	14	24	147

9. Appendix 2: [Plots]

Figure 2: Methodology Flowchart

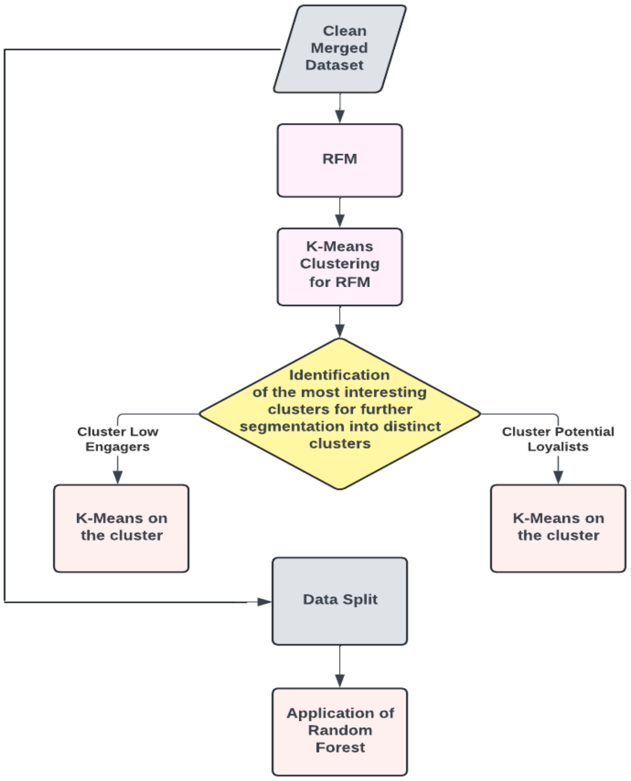

Clean Merged Dataset
RFM
K-Means Clustering for RFM
Identification of the most interesting clusters for further segmentation into distinct clusters
Cluster Low Engagers
Cluster Potential Loyalists
K-Means on the cluster
K-Means on the cluster
Data Split
Application of Random Forest

Figure 3: Elbow Plot

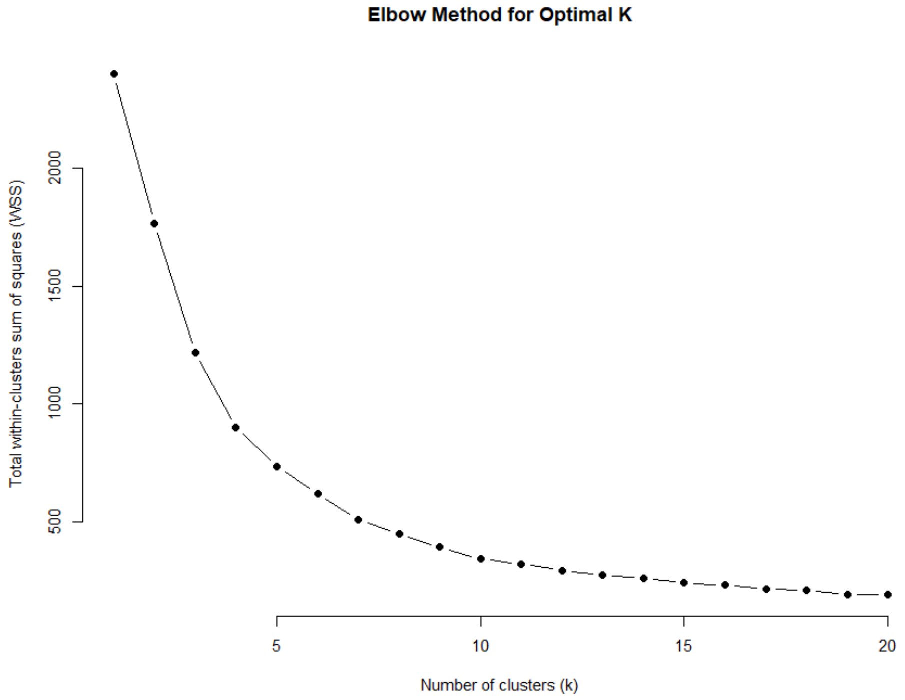


Figure 4: DUNN index (NBClust)

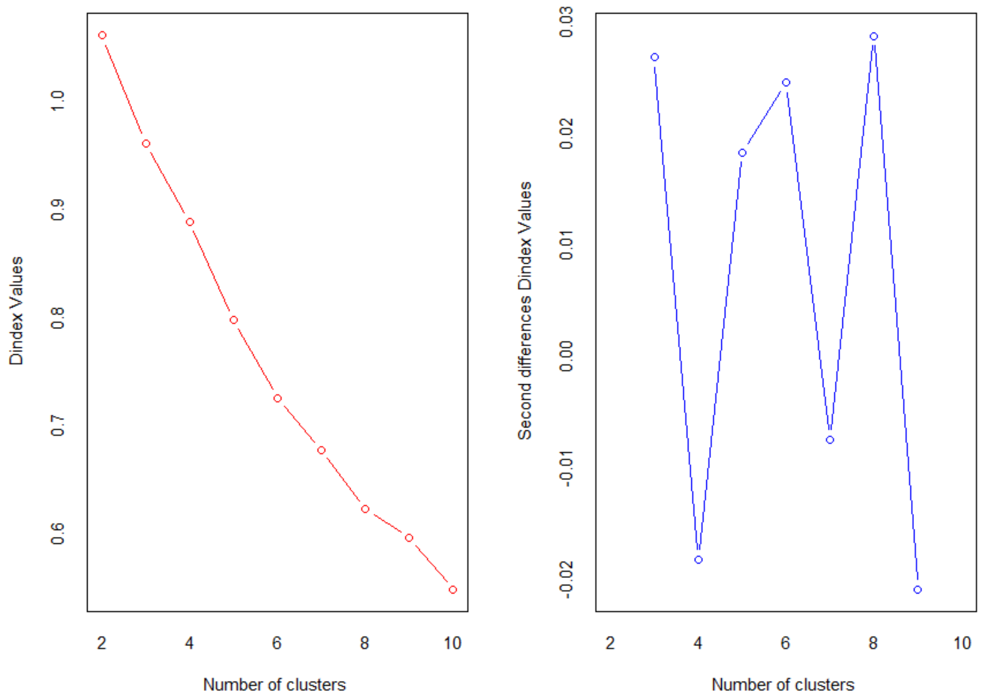


Figure 5: Hubert Index (NBClust)

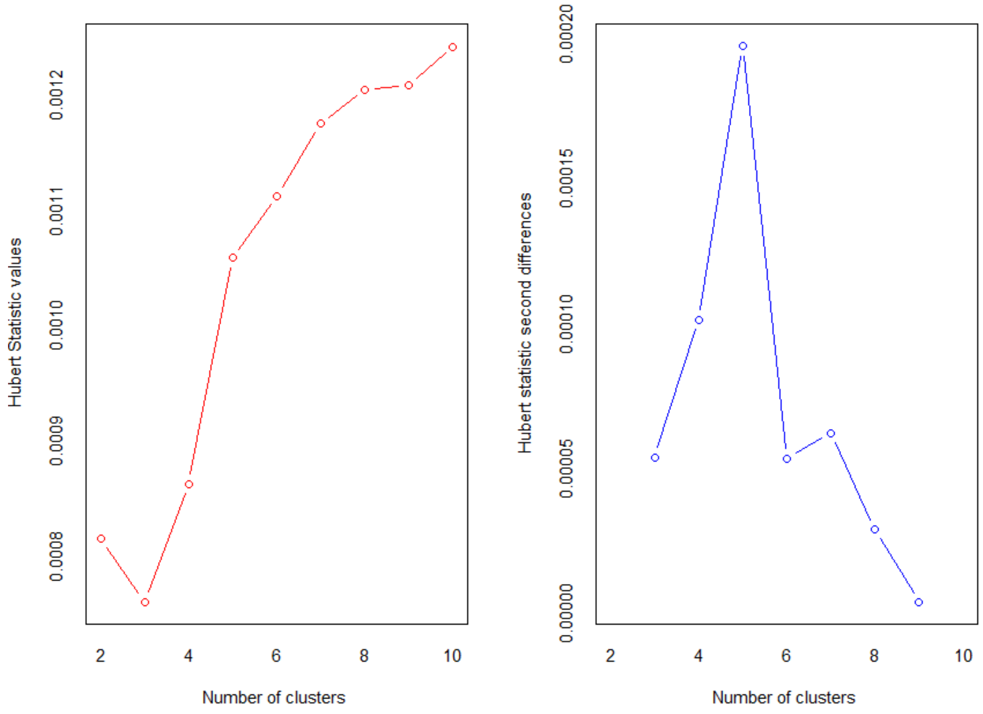


Figure 7: 3D Clusters Plot

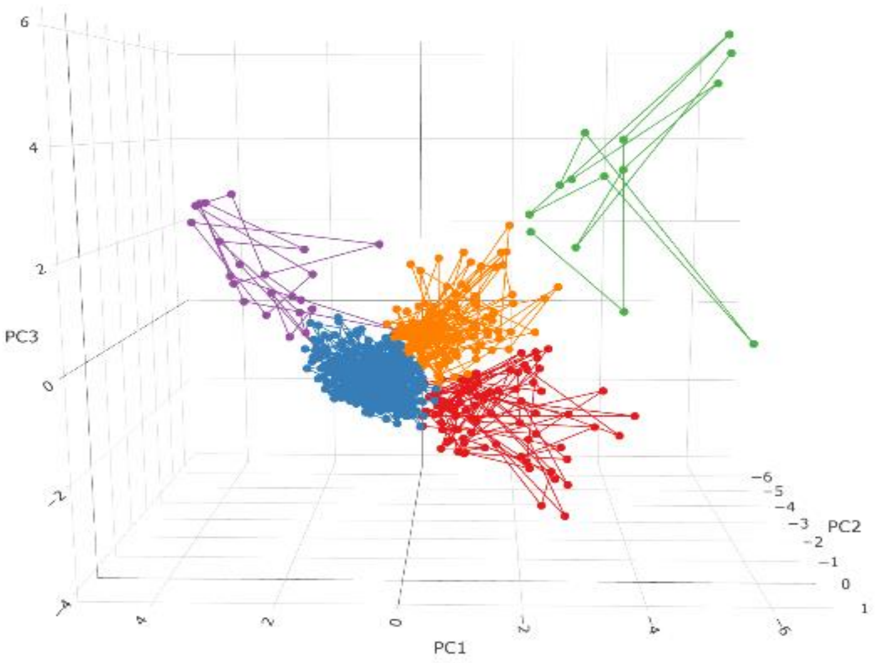


Figure 9.1: Correlation Plot

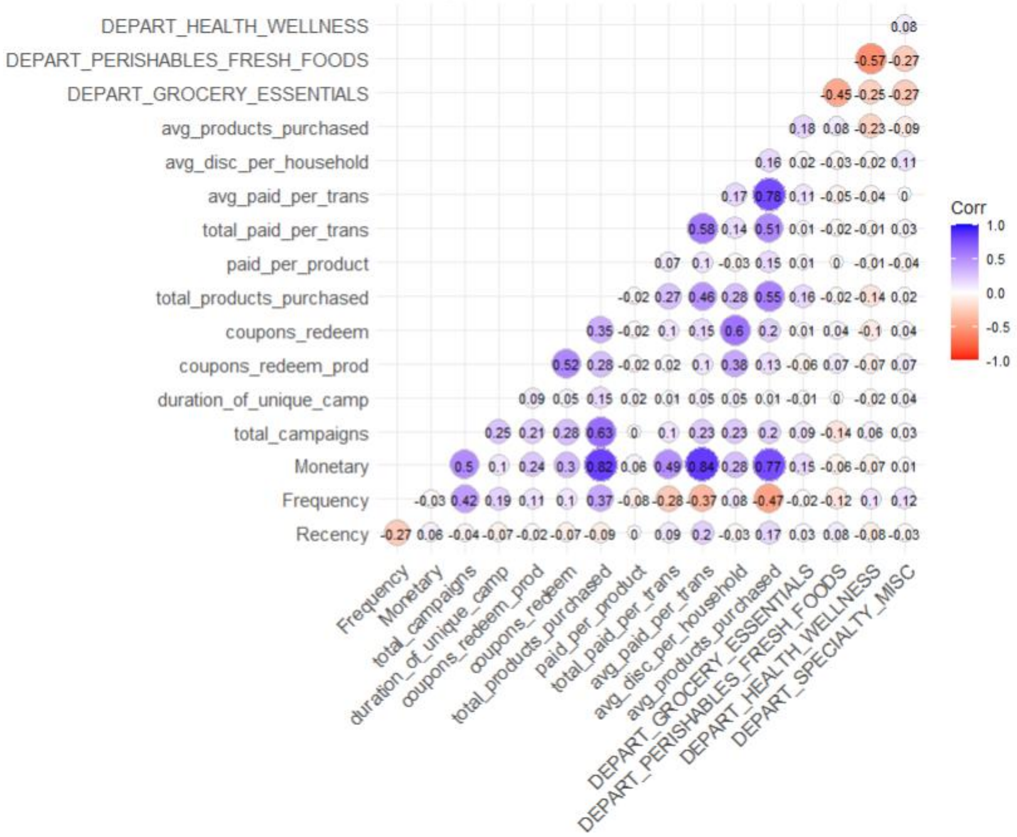


Figure 9.2: Correlation Plot

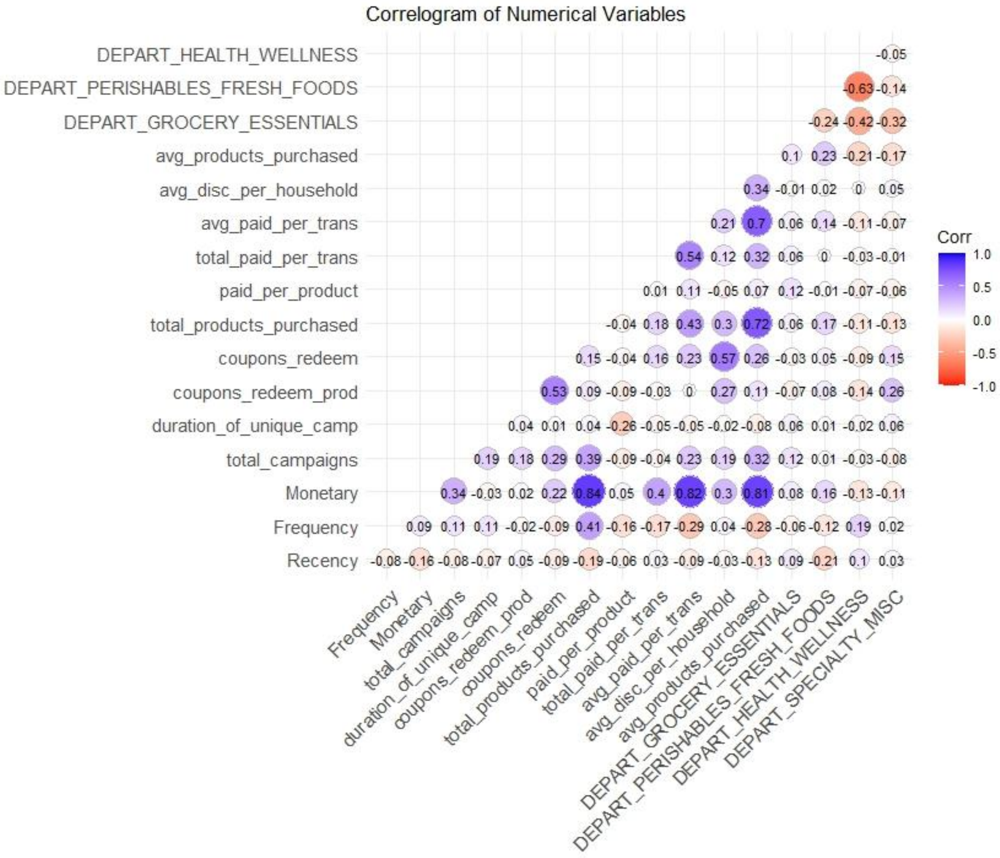


Figure 10: K-Means Clusters (Low Engagers)

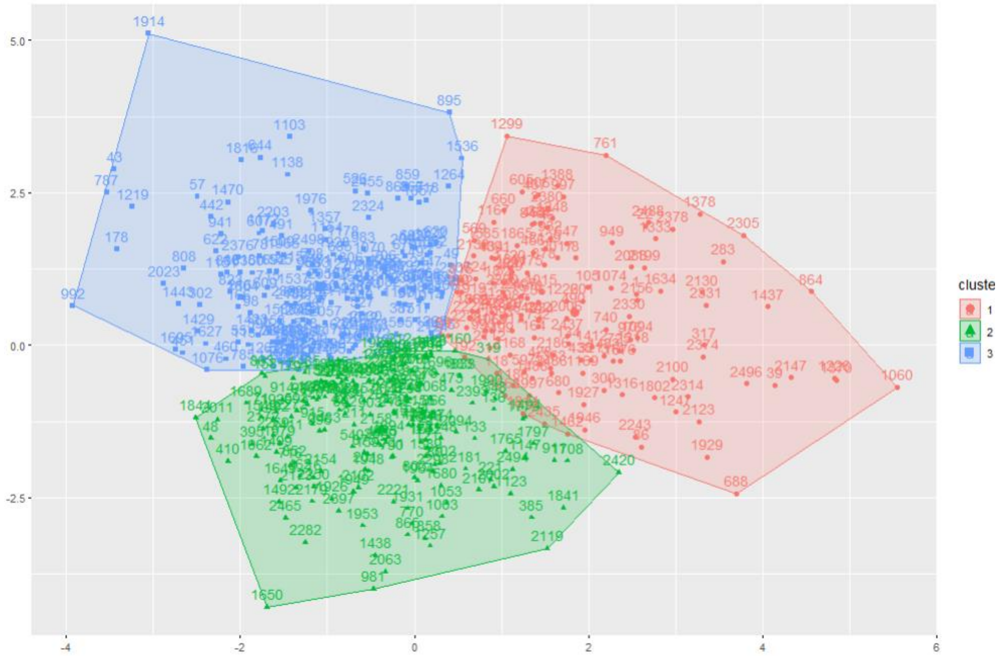


Figure 11: K-Means (Potential Loyalists)

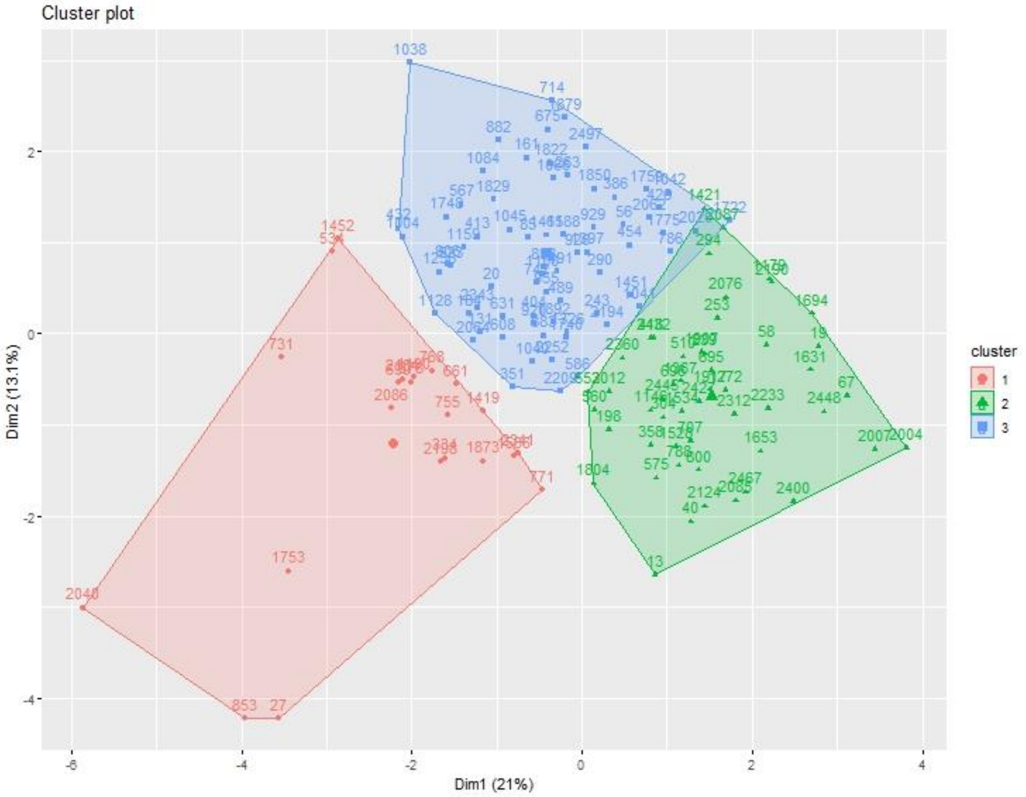


Figure 12: Means Low Engagers

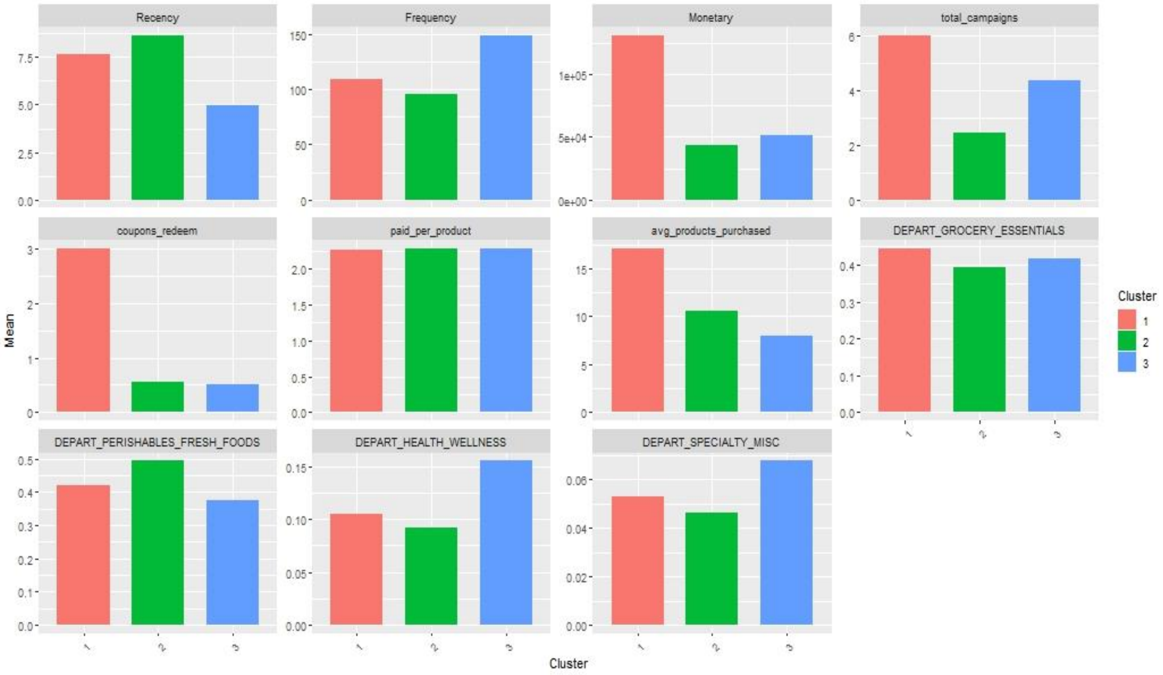


Figure 13: Means Potential Loyalists

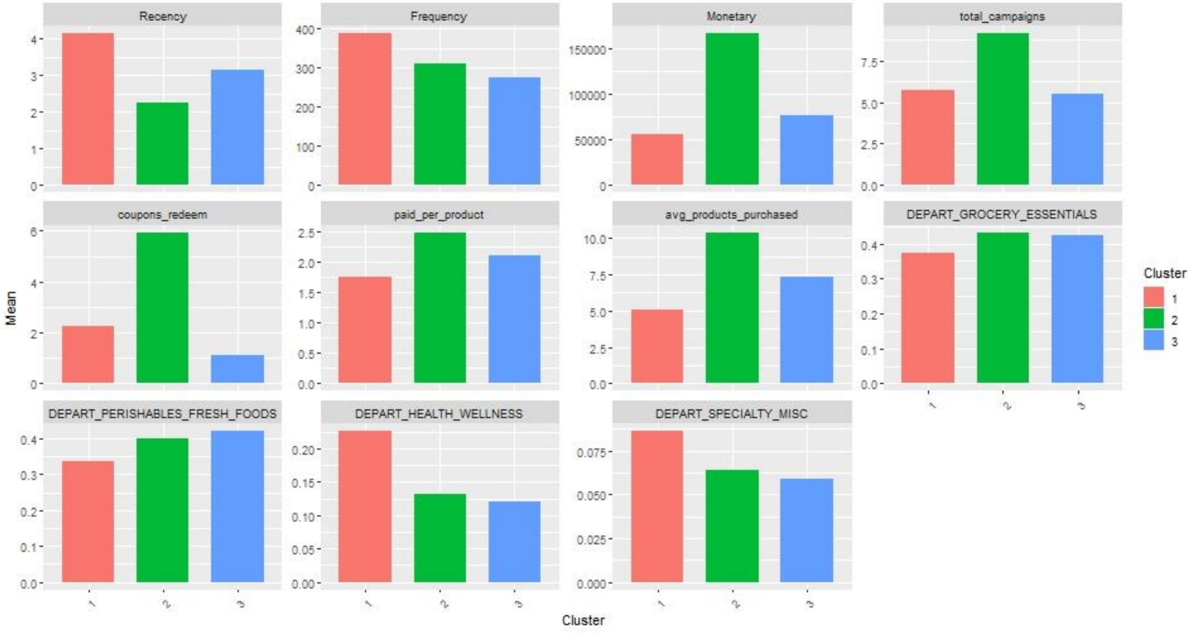


Figure 15: Budget-Conscious Shoppers

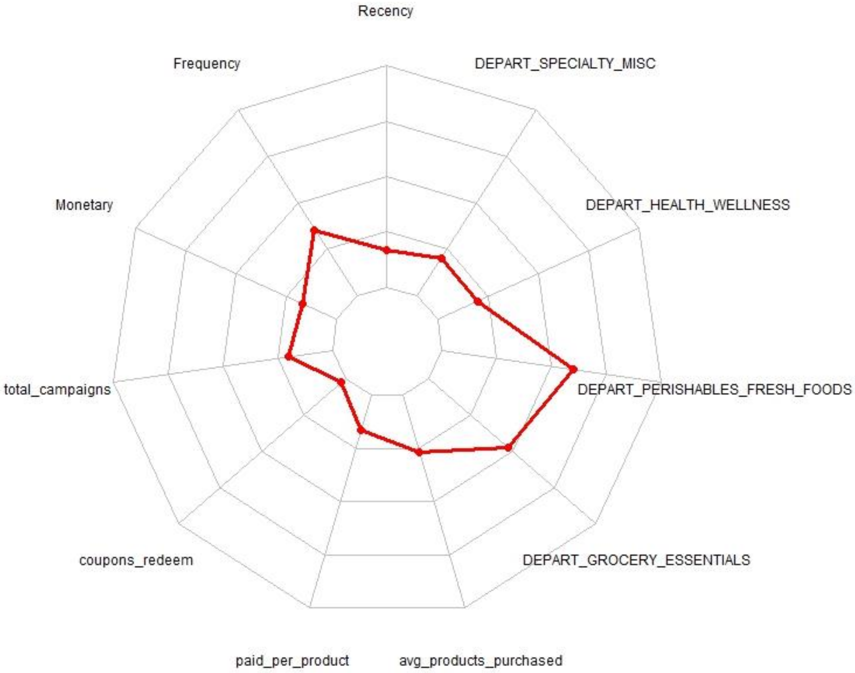


Figure 16: The Selective Shopper

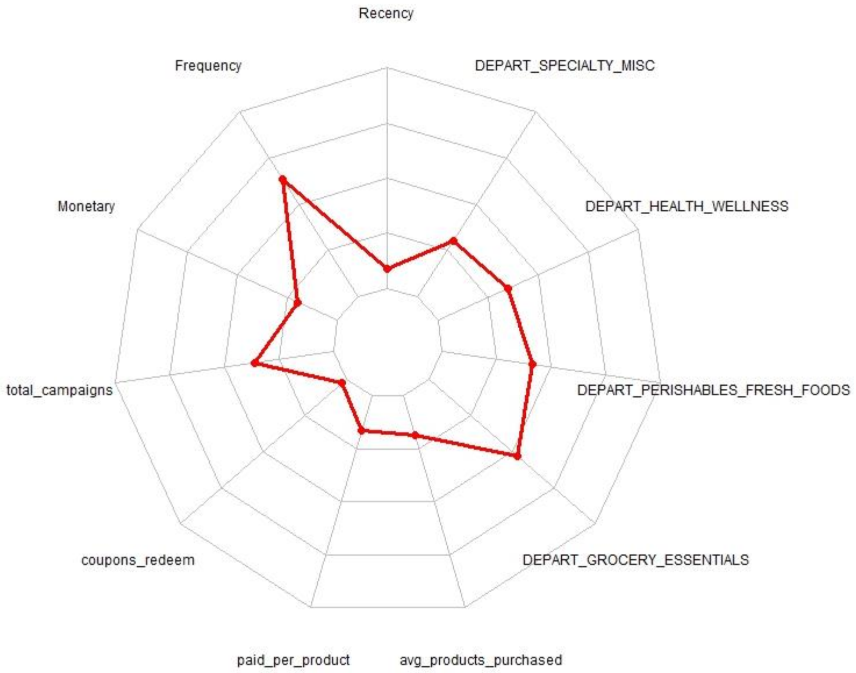


Figure 18: The Valued Frequent Shopper

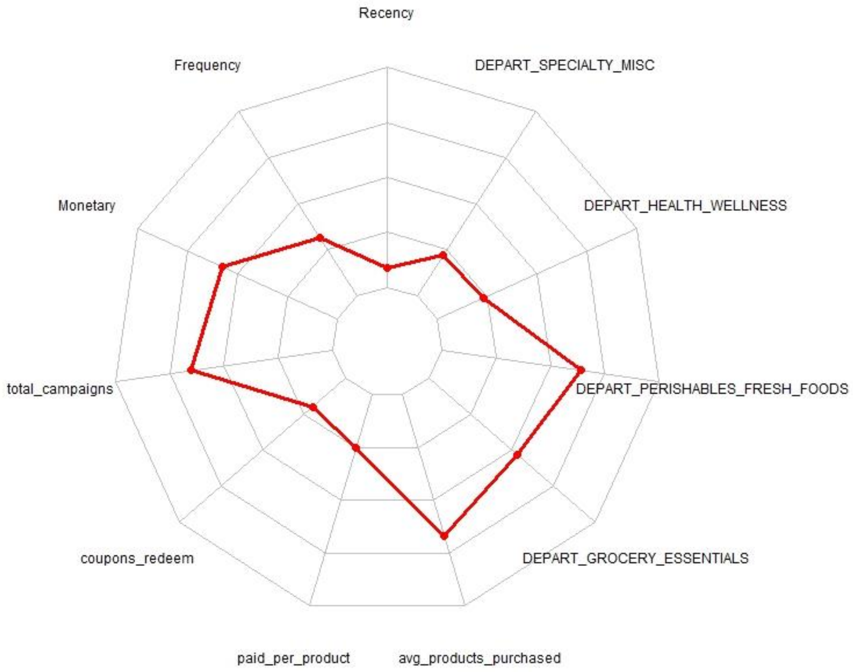
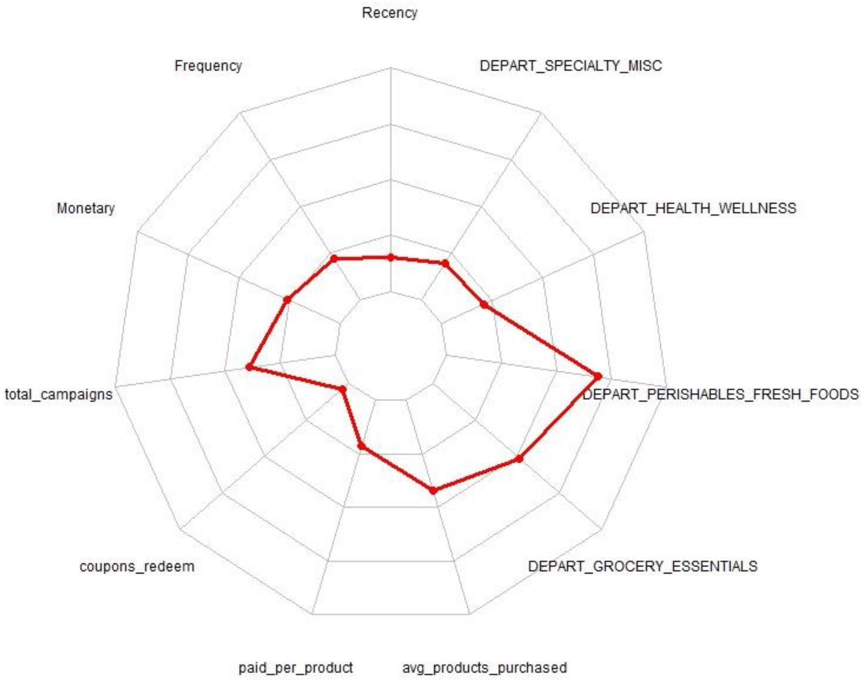


Figure 19: Balanced Occasional Shopper


Recency
DEPART_SPECIALTY_MISC
DEPART_HEALTH_WELLNESS
DEPART_PERISHABLES_FRESH_FOODS
DEPART_GROCERY_ESSENTIALS
avg_products_purchased
paid_per_product
coupons_redeem
total_campaigns
Monetary
Frequency