CASE STUDY IN BUSINESS ANALYTICS
AND QUANTITATIVE MARKETING
Erasmus University Rotterdam
March 2019

UTILIZING CUSTOMER SEGMENTATION METHODS AND IDENTIFYING SEGMENT DYNAMICS

A STUDY OF CUSTOMER SEGMENTATION AT GALL & GALL

Aditya Sharma
Konstantina Mantousi
Linh Dinh
Tatev Karen Aslanyan
Theocharis Asli-Anastasiadis

Abstract

Using data analytics for analysing companies customers and implementing more effective marketing strategies is getting more and more popular nowadays. Customer segmentation based on the transactional data can be performed by simple and popular RFM segmentation based on RFM model which we use as benchmark segmentation method in this paper. However, this method has two significant drawbacks, one regarding the clustering approach itself based on the scores calculated using Pareto rule which does not have any statistical motivation and second regarding the equal or random chosen weights in RFM model assigned to variables Recency, Frequency, and Monetary which also lacks statistical motivation. To overcome these disadvantages, we propose to use K-means clustering in combination with PCA method. Besides, we propose the use of decision tree method to predict the movements of customers between the segments in order to identify potential customers. Finally, we investigate whether taking into account customer heterogeneity leads to different segmentation results than K-means by using Finite Mixture Negative Binomial model.

Keywords: Segmentation, K-means, PCA, Decision Tree, Finite Mixture Negative Binomial

Contents

1 Introduction **2**

2 Data Description **4**

3 Methodology **9**

3.1 KNN Imputation 9

3.2 RFM model 10

3.3 K-means and K-means PCA 13

3.3.1 K-means 13

3.3.2 K-means PCA 15

3.4 Decision Tree 16

3.5 Finite Mixture of Negative Binomial Regression Model 19

3.5.1 FMNB Model Estimation 20

3.5.2 FMNB Model Selection 22

4 Empirical Results **23**

4.1 KNN Imputation Results 23

4.2 RFM results 24

4.3 K-means and K-means PCA Results 27

4.3.1 Principal Component Analysis Results 27

4.3.2 K-means and K-means PCA results 28

4.4 Decision Tree Results 30

4.5 Finite Mixture Model Results 33

5 Robustness Checks **35**

6 Summary and Conclusions **37**

7 Limitations **39**

References **40**

Appendix **43**

1 Introduction

In order to attract and keep highly profitable customers, it is vital that a company can identify and understand their customer value (Chang et al. (2007)). Kotler and Armstrong (2010) mention that for marketing, retaining and growing current customers is as important as acquiring new ones because the former method leads to sustainable customer lifetime value and the latter one brings new sources of values to the company. Furthermore, in this era of big data, each company can obtain an abundance of customer data and thus segmenting customers into clusters that share some similarities is a fundamental task to customize marketing strategies that would satisfy diverse needs of different customers.

According to Kotler and Keller (2011), customers can be segmented based on either their characteristics or their behavior. Characteristics include geographic, demographic, psychographic and firmographic information of a customer while behavioral data contains all the products and services that a customer purchase and attitude or response towards any product or brand. This paper will focus more on the variables that represent the purchase patterns of customers. Combined with the behavioral data, demographic data can also be taken into account to verify any possible shift in the segments or explain the reason for particular segmentation.

It is crucial that the analysis of customer segments is used to optimise the impact of different campaigns on these customers. Even though segmentation of all customers into a few segments is a well-studied topic, this task is often performed with similar clustering techniques without further analysing the type of customers that are in these segments. Moreover, the important subject of individual heterogeneity is not taken into account when using the results of the segmentation for predicting the effect of marketing campaigns on each group of customers. Finally, the dynamics of customers in these segments, whether some part of customers in one segment are likely to move to another segment has been overlooked too.

In this paper, we will explore the above-mentioned aspects of customer segmentation using new approaches which have not been studied yet, to our knowledge. We aim to use different methods to cluster all customers into a few segments and compare the corresponding results. Our goal is to improve these segmentation techniques and make them robust in order to get more accurate results. Additionally, the dynamics of customers in the segments will be studied to predict the possible new segments of these customers. Finally, to account for customer heterogeneity, a new segmentation approach will be proposed. To perform and analyse the tasks as mentioned earlier, we focus on the following research question:

How to utilize customer segmentation methods and identify segment dynamics?

To provide an answer to this research question, we will analyze the following sub-questions:

- How to improve standard customer segmentation methods by using multivariate data technique?
- How to predict the movement of customers between segments and how to target these potential customers?
- Does taking into account personal heterogeneity lead to different segmentation results?

To examine customer segmentation, we employ one of the well-known behavior-based models for customer value segmentation, Recency, Frequency and Monetary (RFM) model by Hughes (1996) which uses these three main factors to measure when, how often and how much customers buy. According to Wei et al. (2010), based on the fact that historical transaction data of customers can be used to predict their future purchases, RFM model can identify which customers belong to the highest and lowest potential groups so that companies can efficiently and effectively allocate their resources. This model has been applied in diverse areas: hospitality (Lee et al. (1998)), banking industry (Hsieh (2004)), online auction (Chan (2005)), public services (King et al. (2007)), online review websites (Li et al. (2010)), and so forth. We will use this method as a benchmark model for classifying all customers into subgroups which we then call RFM-based segmentation. However, FM model makes a lot of assumptions and overlooks many aspects, which will be discussed later, and this over-simplicity of RFM model might lead to inaccurate results. Therefore, other methods of clustering can be employed such as K-means clustering by Sohrabi and Khanlari (2007) or spectral clustering by Chang et al. (2007). Moreover, the dynamics of each segment can be explored by decision tree which is also the case in the researches of Haughton and Oulabi (1993) and Duchessi P. (2013). Last but not least, personal heterogeneity will be studied using the finite mixture model with an appropriate distribution like the negative binomial distribution (Poch and Mannering (1996), Lord and Mannering (2010)).

The rest of the paper is structured as follows. Section 2 provides a detailed description of data and some transformations. Then, the methods and models which are applied in this paper will be discussed in Section 3. Section 4 will display the results of the implementation to answer the research questions. Robustness check will be demonstrated in Section 5 to validate the final clusters. Finally, a conclusion will be drawn based on the analysis in Section 6 and Section 7 will express some limitations of this study.

2 Data Description

In this analysis, we will use the data provided by Gall & Gall, the largest liquor retailer in The Netherlands which has been mainly operating in hard liquor. An initial analysis has been conducted to find five clusters of customers based on what products they buy: Whiskey Lover, New World, Old World, Multi-Buyers and Single Buyers. The goal of this study is to extend this segmentation by focusing on "how" customers of Gall & Gall buy, by using their behavioural data.

The original data set consists of four tables in which we can find detailed information about the customer data and transaction of the company. The Product table describes each product in various subcategories. The Customers table contains information about clients, such as demographics, customer ID, whether a customer allows analysis and other characteristics. The Business to Business (B2B) customer table, where we can match customers ID from the Customer table, and obtain the customers who represent the businesses that make commercial transactions with Gall & Gall. Table 1 summarizes the Customers table, and because there are some missing values in the data set, some of the factors in the table does not sum up to 100%. Ideally, we should only consider customers that allow analysis and commercial emails. However, this would create selection bias since the customers that allow these two options are more likely to be active and responsive to commercial emails. Therefore, to avoid this inaccuracy of the results, we include all customers from the general sales data.

Factor	**TRUE(%)**	**FALSE(%)**
allow analysis	51.60	48.40
allow email	55	45
account holder	33.34	66.67
mailable	42.76	57.24
provide birthdays	92.95	7.05
provide zip codes	89.86	10.13
male	23.75	36.70

Table 1: Summary statistics of customer table

Figure 1 presents the age distribution of active customers over chosen time frame. We observe that the majority of customers are between the age of 36 and 74, with a mean of 51.96 (median of 53).

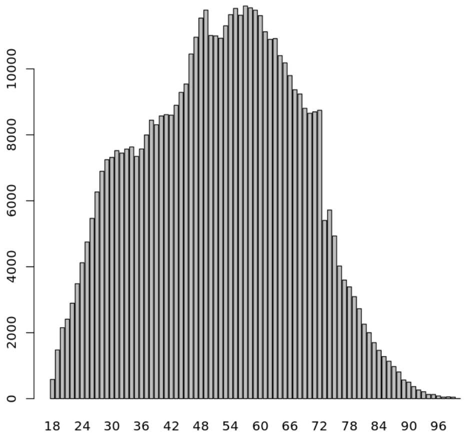


Figure 1: Age distribution of active customers (01/01/2017 - 31/12/2018)

The sales data set consists of online transactions and transactions made by loyalty cardholder at physical shops. The given sales data ranges from 02/10/2016 to 02/01/2019. In order to explore seasonality and have interpretable results, we limit data to two whole years 2017 and 2018.

Figure 2 presents the sales distribution over chosen time-period where the first plot illustrates the two years sales for all customers which shows some seasonality. There are exceptionally high sales in December and some small peaks in April, June, and October. According to this graph, we can conclude that our hypothesis about the seasonality peaks is true. In addition, we notice that the peaks in 2018 are smaller than the peaks in 2017. The second plot in Figure 2 presents the underlying trend of the metrics. The third one represents patterns that repeat with a fixed period. Detrending and deseasonalizing the data before one uses it for the analysis is very important in order to obtain accurate results. There are different approaches for handling seasonality in order to avoid wrong conclusions regarding the number of sales during these periods of picks. The complexity comes from the supplied data which is not in the form of time-series nor the form of panel data; numerous customers have purchased only once over two years. Moreover, the supplied customer data is predominantly time-invariant which makes it impossible the use of known techniques for handling this problem, such as the method of Seasonal Difference. This method cannot be used because we are not dealing with average sales data. However, we can apply the concept of this method to deseasonalize our data. We use the extracted and unused data of 2016 as a leap year, and we use the average amount of orders and spent money per day of 2016 as a benchmark for the number of sold items and spent money for 2017 and 2018. We subtract from these amounts in 2017 and 2018 the benchmark amounts of 2016. The last plot in Figure 2 represents the remainder, the residuals of the original data after the seasonal and trends are removed and we observe that the data contains now much less picks and stationary.

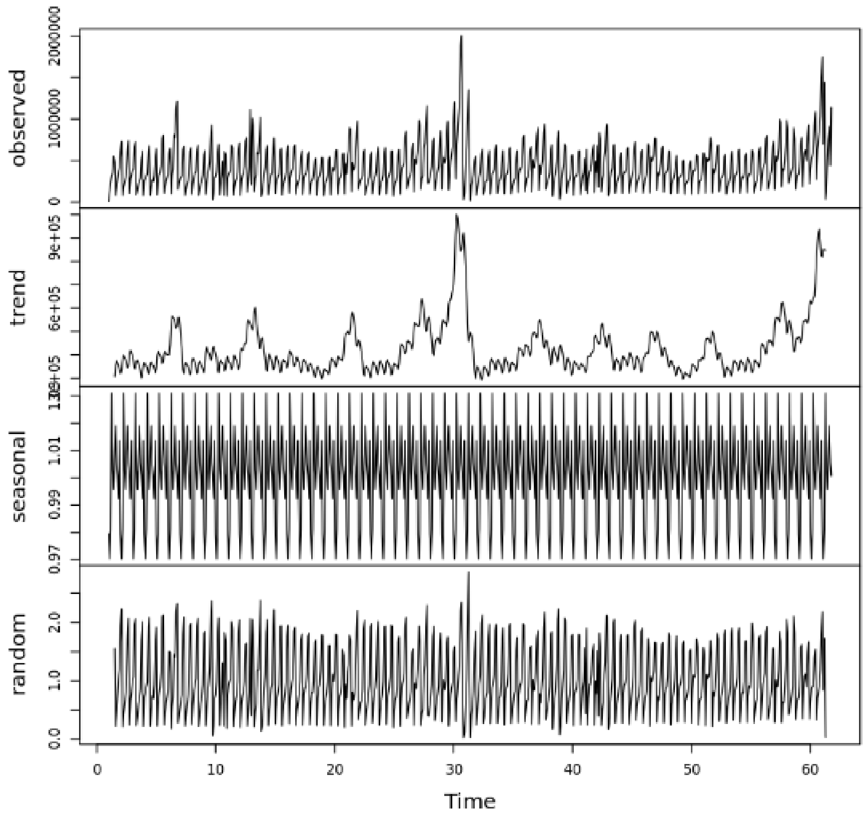


Figure 2: Total sales of all customers (01/01/2017 - 31/12/2018)

Figure 3 presents more insights into the sales table and more specifically, the number of orders per customer. We observe that the number of customers with only one order is the highest and almost doubles the number of customers with two orders. Moreover, we observe that the number of customers decreases as the number of orders increases. Furthermore, Table 17 describes in details the sales per 3 months period according to Frequency. It shows how many customers during each period made a purchased, the total items they bought, the total revenue, the average item per order, the average price per order and the frequency is categorized in seven groups, 1 till 6 and 7 plus. We can see that in the period between October and December for both years, customers tend to be more active and spend more money on their transactions. Thus, because this period includes Sinterklaas and Christmas, two holidays where people tend to celebrate and consume higher quantities of alcohol. The majority of customers tend to purchase once per 3 months and buy on average a little bit more than 2 items, with a total amount per order on average 22 euro.

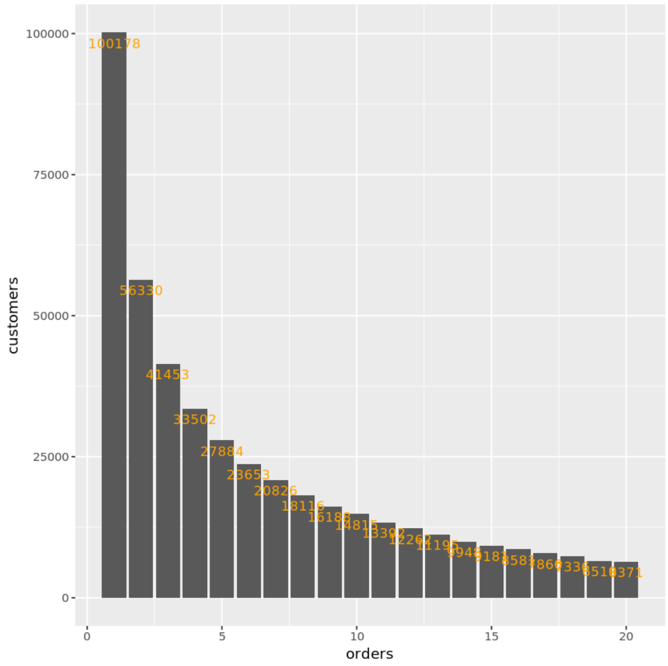


Figure 3: The first 20 most frequent number of purchases

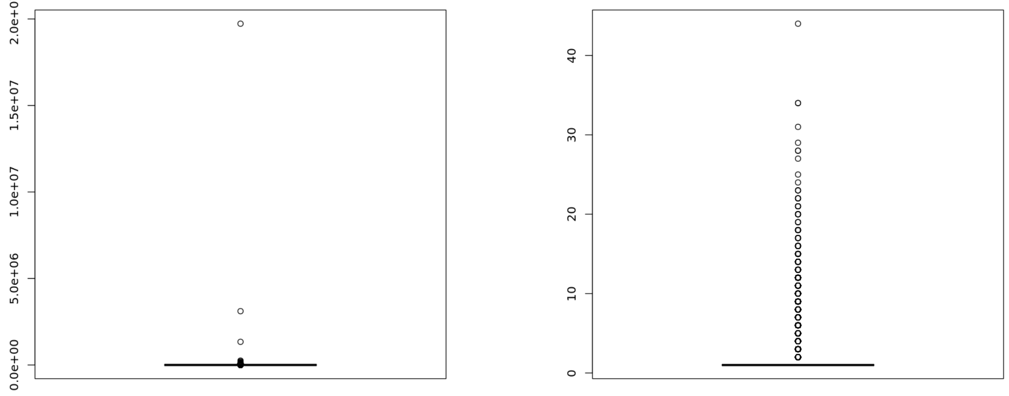
In addition, Table 2 summarizes sales per 3 months period, which shows how many customers during each period had purchased, the total items they bought, the total of orders and revenue, and the average items purchased from each customer per order and the average price of them. During two years, there were in total [redacted] orders with total revenue [redacted] [1]. Furthermore, we can see that in the period between October and December in both years, the number of customers raises, and they tend to be more active and spend more money on their transactions. Thus, because the period includes Sinterklaas and Christmas, two holidays where people tend to celebrate and consume higher quantities of alcohol. The rest of the periods, Gall & Gall has a similar number of customers and tend to spend also almost the same money for the transactions.

[1] According to the Non-Disclosure Agreement, sensitive information that may reflect revenue of the company should be censored.

Period	**Customers**	**Total Items**	**Total Orders**	**Revenue**	**Items/order**	**Price/Order**
10-12/2018	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
7-9/2018	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
4-6/2018	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
1-3/2018	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
10-12/2017	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
7-9/2017	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
4-6/2017	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
1-3/2017	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Total	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 2: Sales table summary [1]

There are three variables for RFM model: `Recency` is the time difference (in terms of days) between the date of analysis and the date of last purchase for each customer based on the benchmark of analysis date 31 December 2018. `Frequency` is the total number of transactions that a customer has made over the entire period of two years. `Monetary` is the total amount that each customer has spent over the defined period. In order to adjust for seasonality, we calculate the daily average spend in the year 2016 and then take the difference between transaction amount and the daily amount in 2016. Since the sales of 2016 are incomplete, we only do this for December sales. After the RFM method is performed, we will exclude the customers that have the lowest scores on all three factors as we aim to base our analysis and conclusions on active and not inactive customers. The scope of the report is to study different approaches for customer segmentation, so we remove B2B customers in order to reduce some outliers in our analysis. However, since the B2B customers' list does not include all the customers that have large and regular purchases, we need to leave out customers with have the highest total spend and a high number of purchases per day. There are two main reasons: a real B2B customer may forget to tick the B2B box in the online registration form, or a cashier in a physical store may scan the store's loyalty card when their customers forget their cards. Therefore, given Figure 4, we will exclude customers with threshold value more than 10,000 euro spend over two years and customers with more than 3 purchases per day.



(a) Monetary value

(b) Number of purchases per day

Figure 4: Boxplot of the distribution of two variables

[1] According to the Non-Disclosure Agreement, sensitive information that may reflect revenue of the company should be censored.

To perform our analysis, we take into account some demographic variables: `Gender`, `Age_category`, `Allow-analysis`, and `Opt-in`. Variable `Gender` is transformed such that it takes value of 1 if gender is female, 0 if gender is male, and NA if unknown or not available. Table 3 explains how the variable `Age_category` is computed from `Age`. Variables `Allow-analysis` and `Opt-in` take value TRUE if a customer chooses any option and FALSE otherwise. We also take into account another behaviour variable that is `Loyal-time` which represents the amount of time that a customer has stayed with the company. Table 16 in the Appendix summarizes all the variables' definitions.

`Age`	18-24	25-34	35-44	45-54	55-64	65+	NA
`Age_category`	1	2	3	4	5	6	NA

Table 3: Age and Age-category variables

Furthermore, we investigate the existence of missing data in Figure 5 in order to analyse whether the amount of missing observations might affect the results of the analysis. From the histogram in the left-hand panel of Figure 5 we observe that the proportions of missing values in each of the customer table variables are: 7.0682% of `Age_category`, 26.3659% of `Gender`, 0.0002% of `Allow_Analysis` and `Opt_in`. Moreover, the right-hand panel of Figure 5 shows all existing combinations of missing (orange) and non-missing (light grey) values in the observations. We observe that whenever `Allow_Analysis` and `Opt_in` are missing, `Age_category` and `Gender` are also missing. The frequencies of the combinations are visualized by small horizontal bars. Given this missing patterns, we implement kNN imputation on variables `Age_category`, `Allow_Analysis` and `Opt_in` because they have reasonably low proportions of missing. Since there is not a lot of variation in the values of `Gender` and very high proportion of missing values, we would not impute for this variable.

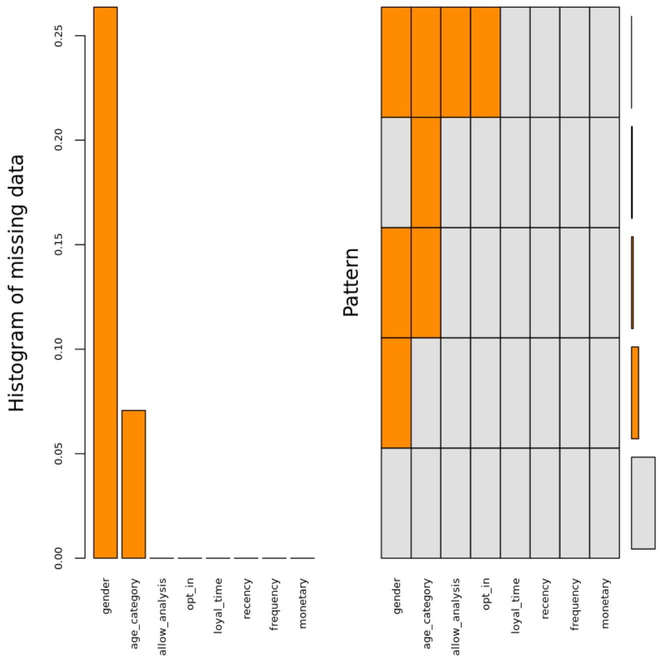


Figure 5: Aggregation plot of the customer table variables

3 Methodology

In this section, we will discuss in details all the models and methods in the following order: KNN imputation, RFM model, Kmeans and Kmeans with Principal Component Analysis, decision tree, Finite-mixture model of Negative Binomial distribution. RFM model is the benchmark model in our analysis.

3.1 KNN Imputation

Missing data is a widely-known problem since most of the statistical methods assume that the data used in the analysis is complete while often this assumption does not hold. Even a small amount of missing values in the data can cause serious problems leading to wrong estimation results and conclusions. There exist numerous methods for solving this problem. These techniques for imputation, estimation of missing values, are divided into univariate and multivariate methods. Univariate imputation is usually used because of its simplicity. More specifically, for a continuous variable many researchers replace the missing data with the mean or median of the observed values, and for a categorical variable, they replace missing values with the mode of the observed values. However, univariate imputation often destroys the multivariate structure of the data, which introduces bias. In addition, it leads to eliminating the variability in the imputed values. Therefore, multivariate approaches perform better and reflect sampling variability.

One single imputation approach for handling missing values is K-Nearest Neighbors(KNN) which leads to more accurate results since it estimates the missing values by their closest neighbors and not all observations. KNN has been introduced by Troyanskaya et al. (2001) which aims to find for missing observation their k most similar neighbors(donors) in the observed data and use them to estimate the value of missing point. This method falls in the category of donor-based methods since it uses K most similar donors to estimate the missing value. This method can be used for the data which contains continuous, discrete, ordinal and categorical variables which makes it particularly useful for dealing with all kind of missing data. The idea behind KNN is that the missing value of a point can be approximated by the values of the points that are closest to it, assuming that if these observations are similar so should be their corresponding values. It is a simple method that can predict both quantitative and qualitative attributes and as a nonparametric method, it does not make an assumption of any particular model. Besides, it is not necessary to create a predictive model for each feature with missing data. However, KNN has also a few disadvantages. The method is quite sensitive to outliers, and it requires to pre-specify the distance and the number of neighbors (k). A small k produces deterioration in the performance of classifier after imputation in the estimation process. Moreover, KNN does not take into account the possible negative correlations between variables. Finally, the most accurate way of imputing data is multiple imputation (MI) firstly introduced by Rubin (1975) which instead performing imputation once it imputes the same data for multiple times. This imputation method takes into account the uncertainty of imputed missing values by leading to more accurate estimates for the standard error which is not the case for single imputation. However, MI's greatest disadvantage is its complexity especially when it is applied to a large amount of data like this analysis. Therefore, we will use KNN imputation approach because it can handle data of large size while providing more accurate results than for instance mean or mode univariate imputation.

When implementing KNN imputation, we need to consider following parameters:

- The number of neighbors: Taking a low k will increase the influence of noise, and the results might be biased. On the other hand, taking a high k will tend to blur local impacts which are what we are looking for. It is also recommended to take an odd k for binary classes to avoid ties.

- The aggregation method: We allow for the arithmetic mean, median and mode for numeric variables (continuous) and mode for categorical ones (both ordinal and nominal).

- Normalizing the data: This will give every attribute the same influence in identifying neighbors when computing certain distances like the Euclidean one. The algorithm normalizes the data when both numeric and categorical variable are provided.

- Similarity distance: among the various distance metrics available, we will focus on the main ones, Euclidean, Hamming, and Manhattan.

  - Euclidean distance $d(x_k, x_l) = \sqrt{(\sum_{j=1}^{n}(x_{kj} - x_{lj})^2)}$ is good to use if the input variables are continuous.

  - Hamming distance $d(x_k, x_l) = \sum_{j=1}^{n} I(x_{kj} \neq x_{lj})$ can be used when there are nominal variables.

  - Manhattan distance $d(x_k, x_l) = \sum_{j=1}^{n} |x_{kj} - x_{lj}|$ is a good measure if the input variables are ordinal.

Algorithm 1 KNN basic algorithm

```
Input: Data that contain missing values
Output: Imputed data ($x_{ij}$)
  • For every $x_j$ where $j = 1, ..., n$
      – Compute pairwise distances between observations, leaving out variable $x_j$ and using only
        pairwise complete information
      – For each observation $x_i$ , $i = 1, ..., n$, with $x_{ij}$ missing:
          ∗ Find k donors with observed value in variable $x_j$ that have the smallest distances
            from unit $x_i$
          ∗ Set $x_{ij}$ to aggregated value of variable $x_j$ from donors
  • Return imputed data matrix $(x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$
```

3.2 RFM model

We use the RFM model as a benchmark segmentation method for clustering customers into different clusters. There are several reasons for the popularity of RFM model as an approach for ranking, comparing and classifying customers. Firstly, this method is straightforward and contains only three variables making it easy to understand. Secondly, the transaction data used in the method is usually stored in a database system which makes it easy to access and extract Lumsden et al. (2008). Finally, RFM is very valuable in predicting response and can boost a companys profits in the short term Khajvand and Tarokh (2011). However, this model also has some disadvantages. Wei et al. (2010) argue that the simplicity of RFM model has been overvalued as the model can only use a limited number of variables, while other characteristics can influence a customer's response and purchase. Moreover, it ignores the analysis of new firms setting up in a short period but also customers that only purchase once and placed small orders. Finally, as most customers do not buy recently and regularly and spend little on their purchase, the model may also underestimate this large proportion of customers Miglautsch (2002). Using the transactions data of the customers as an input, RFM model determines and gives as an output R-score, F-score, and M-score based on those variables described in Section 3.2. The score variables are determined based on the Pareto Principle (80-20 rule) which implies that 80% of the company's revenue come from 20% of customers, we divide each variable into 5 groups using the 20% threshold.

Recency	**R score**	**Frequency**	**F score**	**Monetary**	**M score**
quantile-1 (20%)	5	quantile-1 (20%)	5	quantile-1 (20%)	5
quantile-2 (20%)	4	quantile-2 (20%)	4	quantile-2 (20%)	4
quantile-3 (20%)	3	quantile-3 (20%)	3	quantile-3 (20%)	3
quantile-4 (20%)	2	quantile-4 (20%)	2	quantile-4 (20%)	2
quantile-5 (20%)	1	quantile-5 (20%)	1	quantile-5 (20%)	1

Table 4: RFM scores quantiles

Table 4 presents this transformation procedure from Recency, Frequency and Monetary numerical variables to categorical variables of scores. The first quantile of variable Recency includes 20% of most recent customers (R-score 5) while quantile-5 includes 20% of least recent customers who have made their last purchase in the early stages of the two years (R-score 1). Quantile-1 of Frequency includes the 20% of most frequent buyers (F-score 5) and quantile-5 represents the 20% of least frequent customers (F-score 1). Finally, the customers who are in quantile-1 of Monetary variable are the 20% highest spenders (M-score 5), and the last quantile of Monetary variable represent the least 20% spenders (M-score 1). The higher the R, F, and M scores are, the greater the potential value a customer can possess.

RFM-based segmentation

A simple and widely-known approach for customer segmentation is combining R, F, and M scores in different ways to obtain some meaningful segments. Table 5 presents one possible setting defining the segments. The segment "Best customers" is the customer segment with the highest scores for R, F and M. These customers are the most frequent and recent buyers but also spend the most money compared with the rest of the customers. The "Loyal customer" segment contains customers with high spending and low recency but slightly less than "Best customers". However, the frequency of purchases of these type of customers highly depends upon the promotion period. "Potential loyalist" is the group of customers having high recency (R-score 3-5) along with spending a fair amount of money on purchases (M-score 1-3). The frequency of this category's customers is expected to be average (F-score 1-3), so they have the potential to become "Loyal customer". "New customers" are those customers involved in most recent purchases (R-score 4-5) but the frequency and spendings are on the low level (F- and M-scores 1-2). The "Need attention" segment refers to customers having above average recency along with frequency and spendings on purchases being above average as well (R-, F- and M-scores 2-3). This segment of customers needs to carefully examined as they are the ones who are most likely to churn and go from Better to Good if they are not provided proper attention. The "At risk" segment contains customers who have ordered frequently and spent a good amount of money on purchases, but that happened a long time ago. This segment might represent the customers who are not any more happy with the present deals and discounts offered by the company and their loyalty towards the brand is fading. The "Lost customers" is the one with mediocre recency, frequency and monetary amount (R-, F- and M-scores 1-2).

	Segment	**Description**	**R**	**F**	**M**
1	Best Customers	Have ordered recently,buy often and spent the most	4-5	4-5	4-5
2	Loyal Customers	Spent lot of money,are responsive to promotions	3-5	3-5	3-4
3	Potential Loyalist	Recent customers that order more than once and spent large amount of money	3-5	1-3	1-3
4	New Customers	Bought more recently but not often	4-5	1-1	1-1
5	Promising	Recently ordered but didn't spent too much money	3-4	1-2	1-2
6	Need Attention	Above average recency,frequency,monetary value	2-3	2-3	2-3
7	About to sleep	Below average recency,frequency,monetary value	2-3	1-2	1-2
8	At Risk	Spent lot of money and ordered often but long time ago	1-2	2-5	2-5
9	Can't be lost	Have spent lot of money and often but very long time ago	1-2	4-5	4-5
10	Lost Customers	With lowest frequency,monetary and recency scores	1-1	1-1	1-1
11	Others	Remaining customers that don't fall under any of the above defined categories			

Table 5: Segment definitions

This segmentation method, which we use as a benchmark method, has two significant disadvantages. Firstly, RFM-based segmentation is based on R, F, and M scores by applying the Pareto assumption(20% rule) but one could also use 10% or 30% decision rules for defining the scores, and there is no universal decision rule for this. All these rules lead to different results, and one is not able to state which one is better. Secondly, this clustering approach assigns equal weights to Recency, Frequency and Monetary value as an indicator of importance or allows to assign random weights which again leads to inaccurate results. Secondly, RFM-based segmentation assumes that all three model variables Recency, Frequency, and Monetary have equal weights. According to Hughes (2000), this should also be the case, and all three variables should have the same weight. However, there are many ways of assigning different weights to R, F and M value variables. Miglautsch (2000) argues that the weights need to reflect the order of importance in the model such that Recency has weight 3, Frequency has 2, and Monetary has 1. Khajvand and Tarokh (2011) applied the Analytic Hierarchy Process method based on experts' views to get a relative weight for each variable. So, very regularly users assign different weight to these variables based on their perception or just by their assumption without providing the reason behind their choice of weights while which lead to inaccurate estimation and segmentation results.

3.3 K-means and K-means PCA

To overcome two major issues of RFM-based segmentation mentioned in the previous section, we use a widely-known clustering technique, K-means, which overcomes the first major drawback of the benchmark method. Moreover, we further elaborate K-means by combining it with multivariate technique PCA in order to determine the importance weights statistically to overcome also the second major drawback of RFM-based segmentation. We use K-means and K-means PCA to cluster all customers into three clusters: Good(1), Better(2) and Best(3). It is a common practice to either cluster all customers into two clusters, Good and Bad, like Hand and Henley (1997) or three clusters, Good, Better and Best, like Rigdon et al. (2011). The latter is the reason for choosing three as the number of clusters in K-means and K-means PCA in this analysis. So, K-means is executed to handle the first drawback while we believe that K-means combined with PCA approach can handle both first and the second issues since with PCA the importance weights will be statistically determined based on the supplied data. To determine which of these two methods leads to more accurate results we will apply them both on the same data set with the same settings and present both results to draw a conclusion which of this two methods is better in terms of clustering performance.

3.3.1 K-means

K-means is one of the most well-known algorithms for cluster analysis, which is originally known as Forgy's method, Forgy (1965). Sohrabi and Khanlari (2007) apply this clustering technique to model Customer Lifetime Value based on RFM attributes. Cheng and Chen (2009) combine RFM measures and K-means algorithm to improve classification accuracy and derive some classification rules. One of the main benefits of K-means clustering is the ease of implementation, its efficiency and the short running time compared to other clustering methods, for instance, the hierarchical clustering approach, especially if k is small. The simplicity of this approach makes it easy to explain the results in contrast to support vector machines or artificial neural networks while the flexibility of this method allows for easy adjust if problems occur. Moreover, K-means becomes a good solution for pre-clustering, reducing the space into disjoint smaller subspaces, where other clustering approaches can be applied. However, there are also some disadvantages of this clustering method. One of them is the requirement to pre-specify the number of clusters K. Moreover, the K-means approach is sensitive to outliers and determines the local optimum rather than global.

Hierarchical clustering is an alternative approach that does not require a particular choice of K which results in an attractive tree-based representation of the observations. However, with a large number of variables, K-means clustering may be computationally faster than hierarchical method especially if k is small. Additionally, hierarchical clustering is very sensitive to outliers, and it is not suitable for large data sets. The main idea behind the method is to assign points that are close to each other into the same cluster, checking the distance of each point from the center of the cluster that it belongs to and add all these distances. K-means clustering requires the number of clusters K to be pre-specified, and then the algorithm will assign each observation to precisely one of the K clusters. Specifically, the K-means clustering procedure results from an intuitive mathematical problem. Let $C_1, ..., C_K$ denote the sets that contain the indices of the observations in each cluster. These sets should satisfy the following properties:

1. $C_1 \cup C_2 \cup ... \cup C_K = 1, ..., n$.

2. $C_k \cap C_{k'} = \emptyset$, for all k $\neq k'$.

In other words, the clusters should be non-overlapping, and each observation should belong to at least one of the K clusters. Moreover, the clustering is optimal which has the smallest possible within-cluster variation. The within variation of C_k cluster is a measure $W(C_k)$ of the amount by which the observations in a cluster differ from each other. Therefore, the following optimization problem should be solved:

$$\min_{C_1,..,C_K} \sum_{k=1}^{K} W(C_k) \tag{3.1}$$

The within-cluster variation is defined using the squared Euclidean distance as follows:

$$W(C_k) = \frac{1}{\mid C_k \mid} \sum_{i,i \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{3.2}$$

The number of observations in kth cluster is denoted by $\mid C_k \mid$. Thus, the optimization problem for K-means can be described as follows:

$$\min_{C_1,..,C_K} \sum_{k=1}^{K} \frac{1}{\mid C_k \mid} \sum_{i,i \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{3.3}$$

The pseudocode of the K-means is presented in Algorithm 2.

Algorithm 2 K-means algorithm

```
Step 1: Assign each data point to a random cluster
$z^0_{i\bar{k}} = 1$ for $\bar{k} \in \{1, 2, ...., K\}$ and $z^0_{ik} = 0$ for $k \neq \bar{k}$ and k $\in \{1, 2, ...., K\}$
t=1

Step 2: while clusters change do
solve $\mu^t = \arg\min_\mu \sum_{k=1}^{K} \sum_{i=1}^{n} \mid\mid x_i - \mu_k \mid\mid^2 \underbrace{z^{t-1}_{ik}}_{\text{fixed}}$
where $\mu^t_k$ is the centroid of cluster k at iteration t-1.
solve $z^t = \arg\min_z \sum_{k=1}^{K} \sum_{i=1}^{n} \mid\mid x_i - \underbrace{\mu^t_k}_{\text{fixed}} \mid\mid^2 z_{ik}$
subject to $\sum_{k=1}^{K} z_{ik} = 1, i = 1, 2, ...., n$

where $z_{ik} \in \{0, 1\}, i = 1, 2, ...., n; k = 1, 2, ...., K$. Point i is assigned to the cluster with the
closest $\mu$.
t$\leftarrow t + 1$
```

3.3.2 K-means PCA

As mentioned earlier, there is a need to obtain the importance weights of three attributes Recency, Frequency and Monetary statistically. For this purpose, we use the widely-known multivariate data technique Principal Component Analysis to determine the importance weights of all three variables in RFM model to improve the accuracy of the predictions.

We define by Σ is the [3x3] variance-covariance matrix of a random vector X = $(R,F,M)^T$ and the pairs of eigenvalue and eigenvector of Σ as $(\lambda_1,e_1),(\lambda_2,e_2),(\lambda_3,e_3)$ such that $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Then the j^{th} principal component is defined as follows:

$$PC_j = \alpha_{j1}R + \alpha_{j2}F + \alpha_{j3}M \qquad j = 1, 2, 3$$

where R represents [1xN] vector containing recency values , F is [1xN] vector of frequency values, and M is [1xN] vector of monetary values for all N customers. Moreover, α_{j1}, α_{j2} and α_{j3} are the first, second and third loadings of j^{th} principal component(PC) respectively.

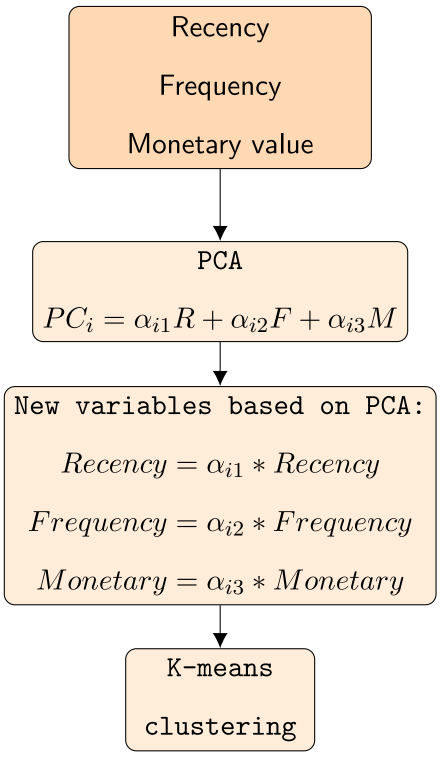
In order to find the PC's that can summarize the data by not losing too much information, we look at the proportion of total variation that is explained by each PC to decide whether it is beneficial to include or exclude that particular PC. We calculate the proportion of variation explained by j^{th} component using the corresponding eigenvalue as follows:

$$PRTV_j = \frac{\lambda_j}{\sum_{k=1}^{3} \lambda_k} = \frac{\lambda_j}{tr(\Sigma)}$$

The $PRTV_j$ represents the proportion of total variation explained by the j^{th} component which measures the information retained when reducing from third to j^{th} component.

For deciding the optimal number of PC's required for this data we will use the Elbow Rule and the Kaiser Rule. Finally, once the PC's with corresponding loadings of these three variables Recency, Frequency, and Monetary are obtained we will use them as weights of importance for these variables. The idea behind this approach is that we calculate choose the optimal number of required PC's based on the variance, the information, explained by them and we use the corresponding loadings as importance weights for variables Recency, Frequency, and Monetary. Since three variables in the model have different measurements, it is necessary to scale all variables to avoid that the monetary value will take away all variation leading to wrong predictions. We will test the necessity of applying PCA on scaled data by using biplots.

After obtaining the optimal number of components and their corresponding loadings by PCA described above we use these loadings as importance weights for the variables Recency, Frequency, and Monetary. So, using these weights we create new, weighted Recency, Frequency, and Monetary variables. So, these three new variables are created such that each of them is the product of the first principal components corresponding loading and the value of each variable as it is illustrated in the diagram below. Finally, these new variables based on principal component analysis are used in K-means clustering to define the clusters. In the remaining parts of this analysis we will call this method as K-means PCA which is demonstrated in the following figure:


Recency
Frequency
Monetary value
PCA
$PC_i = \alpha_{i1}R + \alpha_{i2}F + \alpha_{i3}M$
New variables based on PCA:
$Recency = \alpha_{i1} * Recency$
$Frequency = \alpha_{i2} * Frequency$
$Monetary = \alpha_{i3} * Monetary$
K-means
clustering

3.4 Decision Tree

One of the questions that we have been focused on in this paper is the question regarding the prediction of possible movements between segments. In the previous section, we clustered all customers into three segments: Good, Better and Best with K-Means and K-Means PCA methods. The next step in our analysis to characterise the type of customers in each of these three groups and to investigate which customers are likely to move from one group to another. For instance, we aim to predict which customers have a large likelihood of moving from Good segment to Better, and so on. So, we need a model that will estimate these transitional likelihoods of each customer from one segment to another. For this purpose, we propose the use of decision tree.

We aim to cluster all customers into three groups as it has been done with K-means, but now we add demographic variables to the classification to analyse whether there are customers in one segment who are more similar to the customers of another segment than to their own Kotler and Armstrong (2010). In this way, we will be able to identify customers with the highest potential that are most likely to move from less favorite segment to more favorite one utilizing transition matrix containing transitional probabilities for each customer and each segment. Besides, using the same idea, we can identify the potential droppers who are likely to drop from a more favorite segment to less favorite one.

Haughton and Oulabi (1993) has used decision tree for analysing customers buying behaviour. Moreover, Duchessi P. (2013) have used decision tree to profile the online and mobile technologies and services that ski resorts use for their promotional and advertising strategies for two important segments. Furthermore, Shui Hua Han (2012) have used decision tree method to extract important parameters related to long-term value, credit, and loyalty. They applied this model to telecom operators in China achieving high prediction accuracy. This unsupervised learning method is very popular because of its higher interpretability compared to other classification methods. Another advantage of decision tree is that it can handle easily both missing values in data and irrelevant attributes. Finally the method is very fast and gives compact results in the form of pruned tree. However, decision tree has also disadvantages such as instability of the results once a small change in the input data occurs. Finally, preparing decision tree, especially if they are large with many branches, can be complex and time-consuming challenge.

The idea behind decision tree is in line with the tree analogy where tree starts from the top, and the trees are drawn upside down, moving from the top to the internal nodes and consequently to the terminal nodes, also called leaves. The segments of the trees that connect the nodes are called branches. The classification tree predicts that each observation belongs to the most commonly occurring class of all observations in the space to which it belongs. The process of building a Classification Tree can be described as follows:

Step 1: Construct the regions R_i,...,R_J such that total classification error E = 1-$\max_j(\hat{p}_{mj})$, the fraction of observations in that region that do not belong to the most common class, is minimized:

$$\textbf{arg}\ \ \textbf{min}_{R_j} \sum_{j=1}^{J}(1-\max_j(\hat{p}_{mj})) \tag{3.4}$$

where $\hat{p}_{mj}$ is the proportion of observations in mth that are from the jth class.

Step 2: Divide predictor space, the set of possible values for X_i i = 1,...,p; into R_j j = 1,...,J distinct and non-overlapping spaces(regions).

Step 3: Make the same prediction for each observation from the region R_j, where the prediction is simply the mean of corresponding values of the response variable.

The limitation of this algorithm for constructing the tree is that it is computationally infeasible to consider all possible partitions of feature space into J regions. Moreover, it has been found that classification error is not sufficiently sensitive for tree-growing. This task of growing a classification tree relies on the chosen method for splitting the tree and each of these measures, also called impurity measures, leads to different tree structures. Two widely-known impurity measures are Gini-index and Entropy. Gini-index is a measure of node purity where the small value indicates that the node contains mostly observations from a single class and it is defined as follows:

$$G = \sum_{j=1}^{J}\hat{p}_{mj}(1-\hat{p}_{mj}) \tag{3.5}$$

Therefore, because of these insights, instead of considering all possible split combinations the recursive binary splitting technique can be used to optimally split the predictor space where each step is via two new nodes on the tree by using Gini-index and above algorithm can be extended as follows:

Step 4: Select the predictor X_j and cut-point s such that the splitting of predictor space into regions R_1(j,s) = {X|X_j<s},the region of predictor space in which X_j takes on a value less than s, and R_2(j,s) = {X|X_j ≥s} leads to largest possible information gain which is equivalent to minimizing Gini-index.

Step 5: Repeat Step 4 for finding best predictor j and best cutpoint s to split the tree further until the criterion is reached.

As mentioned earlier, Gini-index is a measure for impurity of the tree, that is when members of one class are in another class where they do not belong. Consequently, we aim to build a tree which has leaves that are as pure as possible and since lower values of Gini-index indicate more homogeneous and purer leaves we aim to minimize it in Step 4. The purpose of using decision tree is in this analysis twofold: building a pure tree and predicting the transitional probabilities of all customers per segment, we use both classification error rate E and Gini-index G for these purposes respectively. So, we use Gini-index for evaluating the quality of each split in the tree, that is pruning process of the tree, and we use classification error rate in the predicting process applied on the pruned tree which leads to higher prediction accuracy.

Coussement et al. (2014) introduced the connection between decision tree and RFM segmentation. We use the variables from RFM model; Recency, Frequency, and Monetary; as features for decision tree but we also add demographic variables such as Age-category, Gender, Allow-analysis, Opt-in and Loyal-time variables described in section 2.

Using the prediction results from the pruned tree, we calculate the probability's of each cluster in the final nodes, such that each customer has certain probability in being in one of the tree classes and these three probabilities sum up to one. These probabilities can be seen as the transition rates of the customers. We then use algorithm 3 to determine the new clusters of all customers based on the transition rates.

Algorithm 3 Assigning new classes to all customers

```
Input: transition rates ($p_{i,k}$) for customer $i = 1, ..., N$ and cluster $k = 1, ..., K$; and "real" cluster found by K-means algorithm ($\tilde{\kappa}_i$)
Output: new cluster ($\hat{k}_i$) for customer $i = 1, ..., N$

• For each $i = 1, ..., N$
    – Get $max.position_i$ that contains all the positions that have $max(p_{i,k})$
    – If $max.position_i$ contains only one position $c$, then set $\hat{k}_i = c$
    – Else
        ∗ If $max.position_i$ contain $\tilde{\kappa}_i$, then set $\hat{k}_i = \tilde{\kappa}_i$
        ∗ Else, then set $\hat{k}_i = max.position_i$
```

The idea behind Algorithm 3 is that knowing the actual class labels from K-means(PCA) we compare this with the class probabilities per customer and if the maximum probability, the probability which is the largest among three probabilities, of being in particular class differs from the actual class from K-means(PCA) we define that particular class as the new class of that customer. For instance, if the customer has been classified to Class 1(Good segment) in K-means and he has p_1 likelihood for being in Class 1, p_2 likelihood for being in Class 2 and p_3 likelihood for being in Class 3 with p_2 being the largest among all probabilities, then we define the new class of this customer as Class 2 (Better segment).

3.5 Finite Mixture of Negative Binomial Regression Model

Customers are different in terms of their demographic characteristic, their responsiveness to marketing actions and their preferences. According to Smith (1956), the heterogeneity of customer needs is the driving force behind the segmentation. In the K-means clustering, the aim was to cluster customers into "Good", "Better" and "Best" segment thus with k = 3, but this was based on our assumption that there are only three subgroups of customers. There arises a question of whether it would be more optimal to cluster customers into more than those three groups to obtain more homogeneous clusters. We can analyse whether taking into account customer heterogeneity will justify the use of only 3 clusters or to cluster customers into more than three subgroups is preferable. Hence, the last research sub-question that we aim to answer is whether the clustering approach for segmenting all customers into sub-groups which also takes into account individual heterogeneity will lead to different results than what we obtained earlier. We aim to analyse whether this method verifies the usage of the same number of segments used in K-means and K-means PCA with corresponding characteristics or it will suggest the usage of another setting for customer classification. For this purpose, we use the Finite Mixture of Negative Binomial Regression. The finite mixture model is a popular method for detecting and handling unobserved subgroups within the entire customer group.

Most of the standard statistical models can be problematic in answering our research questions because they assume the same distribution of the products and no heterogeneity between customers. However, there are differences across individuals based on consumer characteristics such as age and gender make them act differently. Even though individual differences are almost impossible to take into account when the data set consists of thousands of customers, one can create clusters which have approximately the same heterogeneity. So, the assumption behind this method is to cluster observations into segments of customers such that all customers from one segment have the same responsiveness level concerning different marketing instruments. This motivates the use of finite mixture model, which is able to take into account the heterogeneity of customers and leads to accurate and efficient results. It is known that finite mixture model assumes homogeneous attitudes within customer segments and heterogeneous perceptions across segments. Due to the heterogeneity of individuals and the contagion, the data is usually over-dispersed in that the variance is greater than the mean, making the Poisson assumption restrictive. Therefore, to overcome this issue, we will use the Negative Binomial distribution which allows an extra parameter to model the variance making this distribution appropriate for this study. Moreover, Negative Binomial distribution is often preferable to use in combination of finite mixture models since they are the fundamental building blocks of all discrete random variables.

Suppose, the customer group consists of S components or subgroups with probabilities λ_1,...,λ_S and p(y|θ_s) represents the corresponding probability density function of component s. Therefore, joint probability density function is defined as p(y,s) = p(y|θ_s)λ_s with λ_s = p(s) and $\sum_{s=1}^{S} \lambda_s = 1$. The idea behind finite mixture model is to assume that there are S segments where the individuals within a segment have the same heterogeneity. When the component identifier is unobserved, the unconditional probability density function of y is defined as follows:

$$p(y) = \sum_{s=1}^{S} p(y, s) = \sum_{s=1}^{S} p(y \mid \theta_s)\lambda_s = \lambda_1 p(y \mid \theta_1) + ... + \lambda_S p(y \mid \theta_S) \tag{3.6}$$

As it was mentioned earlier, Poisson distribution doesn't solve the problem of over-dispersion which is the result of Bernoulli trials with unequal probability of Poisson trials. To overcome this problem, a variety of ways have been proposed within the framework of negative binomial modeling Poch and Mannering (1996); Hauer et al. (2001); Lord and Mannering (2010).

Suppose, there are N observations with i = 1,...,N, T time periods with t = 1,...,T$_i$ and J independent variables with j = 1,...,J. The negative binomial model for ith observation at time t,y$_{it}$ is defined as follows:

$$P(y_{it} \mid X_i, \beta) = \frac{\Gamma(d + y_{it})}{\Gamma(d)\Gamma(1 + y_{it})} \left(\frac{d}{d + exp(\sum_{j=1}^{J} x_{ijt}\beta_j)} \right)^{d} \left(\frac{exp(\sum_{j=1}^{J} x_{ijt}\beta_j)}{d + exp(\sum_{j=1}^{J} x_{ijt}\beta_j)} \right)^{y_{it}} \tag{3.7}$$

Where x$_{ijt}$ is the value of j$_{th}$ independent variable at time t for observation i, d is the dispersion parameter and $\Gamma()$ is cumulative distribution function of gamma distribution. The expectation and variance of y$_{it}$ is defined as follows:

$$E(y_{it}) = exp(\sum_{j=1}^{J} x_{ijt}\beta_j) \qquad Var(y_{it}) = E(y_{it})\left(1 + \frac{E(y_{it})}{d}\right) \tag{3.8}$$

As previously mentioned, the finite mixture model allows for individual heterogeneity. However, from Equation 3.7 we observe that the coefficient parameter β and d which are constant over all observations. Therefore in order to adjust Negative Binomial distribution to Finite Mixture model we transform the coefficient parameter β and d which are constant over all observations to β_s and d$_s$, such that each segment s will have a segment-specific coefficient estimate and dispersion parameter, assuming that all observations within each component response with the same way to independent variables Zou et al. (2014). So, when we combine the finite mixture model with the negative binomial distribution, and we obtain the following model when conditioning on all S components:

$$\begin{aligned} P(y_i) &= \sum_{s=1}^{S} \lambda_s P(y_i \mid X_i, \beta_s, d_s) \\ &= \sum_{s=1}^{S} \lambda_s \prod_{t=1}^{T_i} \frac{\Gamma(y_{it} + d_s)}{\Gamma(d_s)\Gamma(1 + y_{it})} \left(\frac{exp(\sum_{j=1}^{J} x_{ijt}\beta_{js})}{d_s + exp(\sum_{j=1}^{J} x_{ijt}\beta_{js})} \right)^{y_{it}} \left(\frac{d_s}{d_s + exp(\sum_{j=1}^{J} x_{ijt}\beta_{js})} \right)^{d_s} \end{aligned} \tag{3.9}$$

3.5.1 FMNB Model Estimation

In order to estimate the parameters of the finite mixture model, the method of maximum likelihood(ML) can be used. However, the log-likelihood function in ML has, in general, multiple local maxima and gives poor results. ML estimation based on numerical optimization algorithm does not work smoothly, and this is mainly because the number of components is unknown. Another, better approach for estimating the parameters in finite mixture model is the Expectation-Maximization algorithm (EM) described in Cameron and Trivedi (2005). The finite mixture regression model is given by:

$$y_{it} = \alpha_{s_i} + x'_{it}\beta_{s_i} + \varepsilon_{it} \tag{3.10}$$

where t = 1,...,T$_i$ and i = 1,...,N. The segment of ith observation is s$_i$∈{ 1,...,S}. Suppose, P(s$_i$ = s) =λ_s such that $\sum_{s=1}^{S}\lambda_s = 1$ with p$_s \geq 0$ where λ_s represents the weight of component s. So, α and β can take S different values α_1,...,α_S and β_1,...,$\beta_S$with corresponding weights p$_1$,...,p$_S$ respectively. The likelihood function is described as follows:

$$L = \prod_{i=1}^{N}\sum_{i=1}^{S}\lambda_s\prod_{t=1}^{T_i}p(y_{it} \mid x_{it}, \theta_s) \tag{3.11}$$

where p(y$_{it}$ | x_{it},θ_s) represents the conditional distribution of y$_{it}$. Suppose, w$_{is}$ indicates whether i belongs to segment s such that it takes value 1 if ith observation belongs to group s:

$$w_{is} = \begin{cases} 1 & \text{if i} \in s \\ 0 & \text{if else} \end{cases} \tag{3.12}$$

The complete data likelihood function and complete log-likelihood functions are defined as follows:

$$\begin{aligned} L = p(y, s \mid x, \theta) &= \prod_{i=1}^{N}\prod_{s=1}^{S}\Big(\lambda_s P(y_i \mid X_i, \theta_s)\Big)^{z_{is}} = \prod_{i=1}^{N}\prod_{s=1}^{S}\Big(\lambda_s\prod_{t=1}^{T_i}P(y_{it} \mid x_{it}, \theta_s)\Big)^{z_{is}} \\ l = \ln\, p(y, s \mid x, \theta) &= \sum_{i=1}^{N}\sum_{s=1}^{S}w_{is}\Big(\ln(\lambda_s) + \sum_{t=1}^{T_i}\ln(P(y_{it} \mid x_{it}, \theta_s))\Big) \\ &= \sum_{i=1}^{N}\sum_{s=1}^{S}w_{is}\ln(\lambda_s) + \sum_{i=1}^{N}\sum_{s=1}^{S}w_{is}\sum_{t=1}^{T_i}\ln(P(y_{it} \mid x_{it}, \theta_s)) \end{aligned} \tag{3.13}$$

The EM algorithm is iterative process consisting of two-steps:

E-step: Determine the expectation of the complete log-likelihood with respect to s|y with current estimate of θ ($\hat{\theta}$), that is: E$_{s|y}$[ln(P(y,s| x,$\hat{\theta}$))].

M-step: Maximize the expected value for parameter θ to update new estimator $\hat{\theta}^*$, that is:
arg max$_\theta$ E$_{s|y}$[ln(P(y,s| x,θ))] where different initialization estimates can be used.
In order to adjust the EM algorithm to the negative binomial distribution described earlier we adjust the weights defined in Equation 3.12. Suppose, w$_i$ = [w$_{i1}$,...,w$_{iS}$]$'$ is S dimensional vector containing all weights assigned to observation and W is the matrix containing w$_i$ vectors for all N observations.

E-step

In E-step the expectation of complete log-likelihood (l) function conditional on data which is given and the parameters that have to be estimated. We denote by $\hat{\Lambda}$ = ($\hat{\lambda}_1$,...,$\hat{\lambda}_S$)',$\hat{B}$ = ($\hat{\beta}_1$,...,$\hat{\beta}_S$)' and $\hat{D}$ = ($\hat{d}_1$,...,$\hat{d}_S$)' the estimates of corresponding parameters of the model where β_s represents the slope coefficient vector of sth segment with dimension J and d$_s$ is the dispersion parameter of segment s. Consequently, the expectation of l is defined as follows:

$$E_s[l \mid \hat{\Lambda}, \hat{B}, \hat{D}] = \sum_{i=1}^{N}\sum_{s=1}^{S}E[w_{is} \mid y_i]\ln(\hat{\lambda}_s) + \sum_{i=1}^{N}\sum_{s=1}^{S}E[w_{is} \mid y_i]\sum_{t=1}^{T_i}\ln(P(y_{it} \mid x_{it}, \hat{\beta}_s, \hat{d}_s)) \tag{3.14}$$

Conditional distribution of weight w$_{is}$ based Bayes theorem is defined as follows:

$$P[w_{is} \mid y_i] = \frac{\prod_{s=1}^{S} \hat{\lambda}_s (P(y_i \mid X_i, \hat{\beta}_s, \hat{\beta}_s))^{w_{is}}}{\sum_{i=1}^{N} \prod_{s=1}^{S} \hat{\lambda}_s P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)} \tag{3.15}$$

Therefore,

$$E[w_{is} \mid y_i] = \frac{\hat{\lambda}_s P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)}{\sum_{s=1}^{S} \hat{\lambda}_s P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)} \tag{3.16}$$

M-step

After performing E-step the E$_s[l \mid \hat{\Lambda}, \hat{B}, \hat{D}]$ should be maximized such that $\lambda_s \in$[0,1] and λ_s = 1-$\sum_{s=1}^{S-1} \lambda_s$. We perform this step by using Lagrangian optimization method and obtain the following complete likelihood function and the corresponding first order derivatives as follows:

$$\begin{aligned}
&L = \sum_{i=1}^{N}\sum_{s=1}^{S} E[w_{is} \mid y_i] \ln\hat{\lambda}_s + \sum_{i=1}^{N}\sum_{s=1}^{S} E[w_{is} \mid y_i] \ln P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s) - k(\sum_{s=1}^{S} \hat{\lambda}_s - 1) \\
&\frac{\partial L}{\partial \hat{\lambda}_s} = \sum_{i=1}^{N} \frac{E[w_{is} \mid y_i]}{\hat{\lambda}_s} = 0 \quad \Leftrightarrow \quad k = 0 \\
&\frac{\partial L}{\partial \hat{\beta}_{js}} = \sum_{i=1}^{N} E[w_{is} \mid y_i] \frac{\partial \ln P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)}{\partial \hat{\beta}_{js}} = 0 \\
&\frac{\partial L}{\partial \hat{d}_s} = \sum_{i=1}^{N} E[w_{is} \mid y_i] \frac{\partial \ln P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)}{\partial \hat{d}_s} = 0
\end{aligned} \tag{3.17}$$

where k is Lagrange multiplier. Only one for the three equation from Equation 3.17 has an analytical solution, namely the first equation, last two equation cannot be solved analytically. Therefore, the Broyden - Fletcher - Goldfarb - Shanno(BFGS) algorithm will be used to solve the last two equations.

3.5.2 FMNB Model Selection

We use Finite Mixture model in combination with negative binomial distribution as a classification method, as we mentioned earlier. In K-means we, in advance, have assumed that there should be 3 clusters of customers: Good, Better and Best. However, with this model, we aim to find the optimal number of clusters using the data itself. However, it is impossible to select the number of segments in RFM model by using the standard t-test. The LR test for $_s = 0$ has no standard deviation and consequently, parameters α_s and β_s are not identified in this case. This problem is known as Devis problem and therefore, to avoid it, we use Bayesian Information Criteria(BIC) to evaluate the performance of the model for the different number of segments. The consistency property of BIC means that it is guaranteed to select the true model as the sample size grows infinitely large. BIC overcomes AIC's problem by making the parameter inclusion threshold more stringent as the sample size grows. More specifically, according to Vrieze (2012), the Type I error of BIC goes to zero whereas the AIC's does not. It is known that AIC will fail to select the true model with non-vanishing probability as N becomes large, even in the case that the true model is under consideration. Thus, the consistency of BIC makes it quite attractive. However, the situation changes, in the case that the number of parameters in the true model is infinite. Then the AIC is asymptotically efficient in MSE of estimation, whereas BIC is not. In our case, there are few parameters in the model, and the number of observations is quite large which indicates to use the BIC information criterion for model selection. The BIC is defined as follows:

$$\begin{aligned} BIC &= -2\ln L + C\ln\sum_{i=1}^{N} T_i \\ C &= S - 1 + (J+1)S \end{aligned} \tag{3.18}$$

4 Empirical Results

4.1 KNN Imputation Results

As it was mentioned earlier, we use KNN imputation to estimate missing observations in the data to get complete data in order to perform the analysis. The figure 6 shows for each variable both the original (grey) and imputed data points (orange). In each diagonal panel, a density plot is drawn with two lines: the grey one represents the density of the observed values, and the orange one represents the density of the imputed values of the chosen highlighted variable. We observe that there is no particular pattern into imputing variable. More specifically, the variable `Age_category` has comparable amounts of imputed values for all age categories. The figure verifies that the likelihood of missing values is unrelated to both on the missing values themselves and observed values. This indicates that the data is not Missing Not At Random (MNAR) so we can be sure that imputation does not lead to biased results.

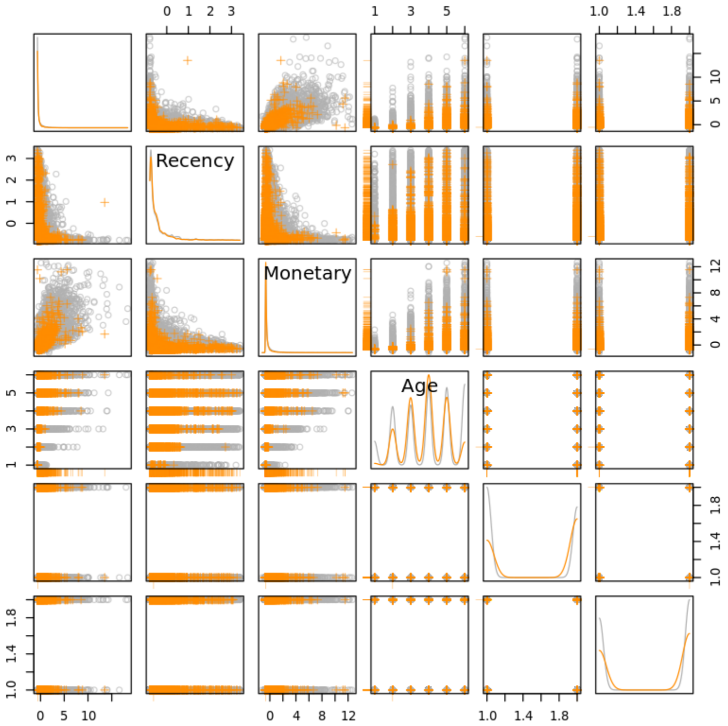


Figure 6: Scatterplot Matrix of all variables with imputed values in `Age` are highlighted as orange

4.2 RFM results

As it was mentioned earlier, by using transactional data RFM model determines the Recency, Frequency and Monetary variables. Table 6 represents the correlation matrix of these three variables as described in Section 3.2 from which we observe that the variables Frequency and Monetary are highly and positively corrected which means that customers who usually come to the store or make purchases are also the customers who spent the most. Moreover, we observe that Monetary is not very strongly but negatively correlated which might suggest that customers who have made a purchase recently have not spent much money. Finally, we observe that there is some negative correlation between Frequency and Recency which indicates that customers who have recently made purchases are not the frequent customers of the company.

	Recency	**Frequency**	**Monetary**
Recency	1.000	-0.324	-0.327
Frequency	-0.324	1.000	0.795
Monetary	-0.327	0.795	1.000

Table 6: Correlation matrix of R,F and M

Table 7 illustrates the RFM estimation results for each of the variables of RFM model and R, F, and M scores. We observe that the Recency variable has mean 169.95 which is quite low (the lower, the better) with min and max value 0 and 729 respectively. We also observe that the average customer has a total number of transactions [redacted] and revenue [redacted]. The maximum number of transactions that any of the customers has made is [redacted] while the minimum value is only 1. With regards to the amount that Gall & Gall's customers spend, we can see significant differences between maximum and minimum value. More specifically, the maximum amount is [redacted] and the minimum is -[redacted] due to product returns . Last but not least, the mean of Recency, Frequency, and Monetary scores are all close to 3.[1]

	Rec. days	**Trans. count**	**Revenue**	**R-score**	**F-score**	**M-score**
Min.	0.00	[redacted]	[redacted]	1.00	1.00	1.00
Median.	12.00	[redacted]	[redacted]	3.00	3.00	3.00
Mean.	169.95	[redacted]	[redacted]	3.01	2.86	2.99
Max.	729.00	[redacted]	[redacted]	5.00	5.00	5.00
Customers	[redacted]					
Analysis date	31-12-2018					

Table 7: RFM summary results [1]

Figure 7 plots the relationship between monetary and frequency values of the customers. The plot shows clear positive(increasing) relation between these two measures suggesting that customers who more often make a purchase spent more money compared to the less frequent comers.

As it was mentioned in Section 3.1 higher values of Recency indicates that customer made the last shopping a long time ago. Figure 8 represents the relation between Monetary and Recency values

[1] According to the Non-Disclosure Agreement, sensitive information that may reflect revenue of the company should be censored.

which looks very similar to the relation between Recency and Frequency from Figure 9. We observe that larger values of Recency are characterized by low Monetary value. This relation suggests that customers who have brought the most monetary amount to the company are the ones who have purchased in recent past, while the customers who have shopped for the last time in the distant past did not bring a lot of monetary value to the company. From Figure 9 we observe that for the smaller values of Frequency the Recency is very large which indicates that customers with low frequency have purchased in the distant past while those with high frequency have made a purchase in the recent past. So, the customers who visited in the recent past are more likely to return compared to those who visited a long time ago. As such, a higher frequency would be associated with the most recent visits.

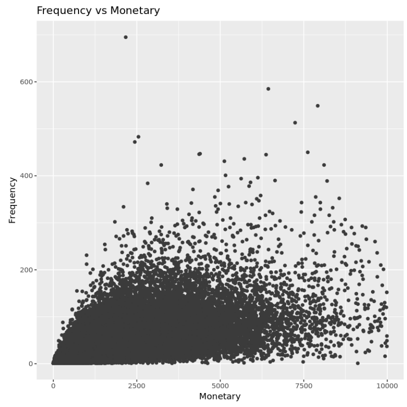


Figure 7: Relation between F and M

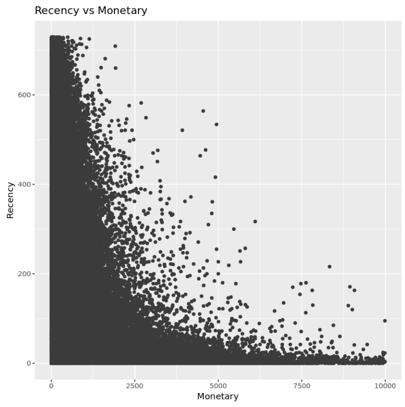


Figure 8: Relation between R and M

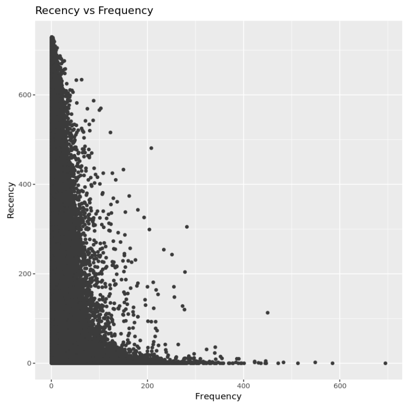


Figure 9: Relation between R and F

RFM based segmentation

Table 8 represents the segmentation based on R, F and M score obtained in RFM model using the principle described in Section 3.3. The table also presents the characteristics of customers falling in each of these 11 segments. The Loyal Customers segment contains approximately 18% of all customers. Approximately 56% of people in this segment have opted for allowing analysis. Similarly, approximately 59% of people have opted for commercial email option (Opt-in) so that they get news and different promotion activities information via emails. It is also worth mentioning that this segment mainly consists of customers who fall in the 65 and older age category. Finally, we observe that in the gender section of this entire table females mainly dominate in all segments with the female variable of every segment being close to approximately 60%. At Risk Customers segment contains approximately 16% of all customers with 55% of customers who have opted for allowing analysis and 53% of people who have opted for commercial email option. This segment again consists of customers who are 65 and older. The Best Customers segment contains approximately 14% of the total customer population. Approximately 58% of them have opted for allowing analysis and 60% of people have opted for commercial email option. This segment again consists of customers who are 65 and older. The About to Sleep segment contains approximately 12% of the total size. Approximately 41% of people in this segment have opted for allowing analysis and 52% of people have opted for commercial email option. This segment mainly consists of customers who are in the 45-54 age category. The Potential Loyalist segment contains approximately 6% of the total size. Approximately 43% of people in this segment have opted for allowing analysis and 35% of people have opted for commercial email option. This segment again consists of customers who fall in the 45-54 age category. The Lost segment contains approximately 8% of the total size. Approximately 45% of people in this segment have opted for allowing analysis and 45% of people have opted for commercial email option. This segment mainly again consists of customers who are between 25 and 34 which shows that the young generation is not effectively targeted in order to turn them into Best or Loyal customers. The Need Attention segment contains approximately 4% of the total size. Approximately 55 % of people in this segment have opted for allowing analysis and 59 % of people have opted for commercial email option. This segment again consists mainly of customers who are 65 and older. The Promising customers' segment forms the smallest segment of all, containing approximately 2% of the total size. Approximately 36% of people in this segment have opted for allowing analysis and 50% of people have opted for commercial email option. This segment again mainly again consists of customers who fall in the 25-34 age category. The New customers' segment forms approximately 5% of the total size. Approximately 48% of people in this segment have opted for allowing analysis and 54% of people have opted for commercial email option. This segment also consists of customers who fall in the same 25-34 age category. The Others segment consists of 12% of size. Approximately 50% of people in this segment have opted for allowing analysis and 54% of people have opted for commercial email option. This segment consists of customers who are 65 and older.

Segments	**Size (%)**	**RFM values**	**Gender(%)**	**Age(%)**	**Allow-analysis (%)**	**Opt-in(%)**
Loyal Customers	17.9	R $\downarrow F \uparrow M \uparrow$	63.19	65+(27.7)	56.38	59.02
At Risk	15.7	R $\downarrow F \downarrow M \downarrow$	63.67	65+(22.1)	55.24	53.49
Best Customers	14.4	R $\uparrow F \uparrow M \uparrow$	66.92	65+(33.7)	58.13	60.27
About To Sleep	11.7	R $\downarrow F \downarrow M \downarrow$	62.86	45-54(20.6)	41.02	51.58
Potential Loyalist	6.4	R $\uparrow F \downarrow M \downarrow$	64.49	45-54(23.9)	43.64	35.49
Lost	7.9	R $\downarrow F \downarrow M \downarrow$	61.20	25-34(20.2)	45.08	45.16
Need Attention	3.8	R $\downarrow F \downarrow M \downarrow$	63.48	65+(22.2)	55.22	58.98
Can not be lost	3.5	R $\downarrow F \uparrow M \uparrow$	66.17	65+(28.2)	53.84	51.12
Promising Customers	2.1	R $\uparrow F \downarrow M \downarrow$	59.74	25-34(21.7)	35.77	49.63
New Customers	4.6	R $\uparrow F \downarrow M \downarrow$	63.59	25-34(22.4)	47.54	53.74
Others	11.6	R $\downarrow F \downarrow M \downarrow$	66.12	65+(23.6)	50.33	53.98

Table 8: RFM based segmentation

Table **??** in the Appendix presents the results of RFM-based segmentation for four holidays in over two years. As a general comment, we can tell that the highest percentage of customers cannot categorize in a group (Others). On the other hand, we notice that the majority of active customers during these holidays belong in the segments of Potential Loyalist, About to Sleep and Loyal Customers. Potential Loyalists are customers with middle to high recency score (3-5), and low to middle frequency and monetary score (1-3). About to Sleep are the customers with low to mid recency (2-3) and low frequency and monetary score (1-2). Loyal Customers are a step before Best customers, with mid to high recency and frequency score (3-5) but they spend less money than the Best Customers, that is why they have a middle monetary score (4-5). Best Customers, tend to be less active in these periods and having less than 9% contribution to the company's sales.

4.3 K-means and K-means PCA Results

In this section, we will present the results of Principal Component Analysis. Consequently, we use these PCA results for K-means PCA method and we present the clustering results of K-means and K-means combined with PCA method (K-means PCA) side by side for comparison purposes.

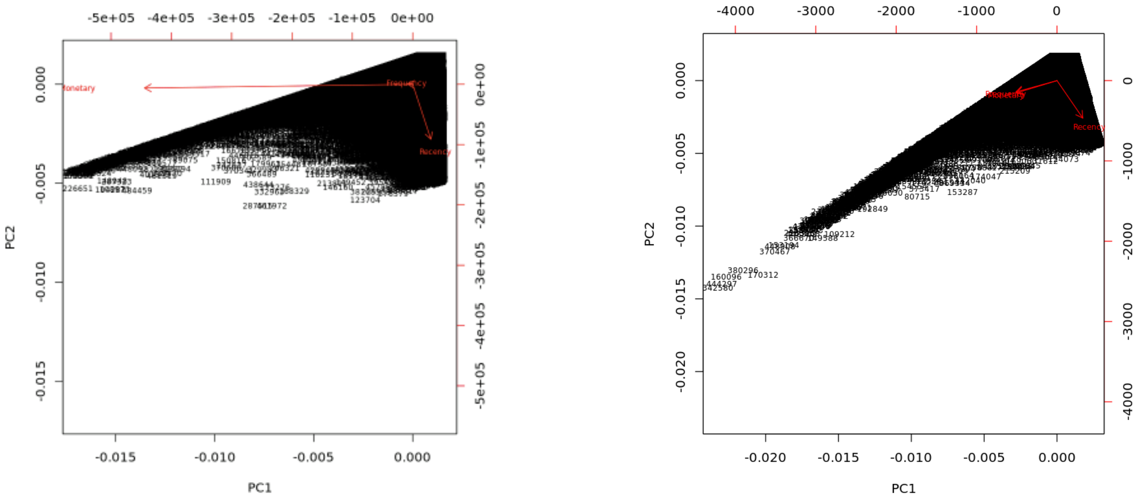
4.3.1 Principal Component Analysis Results

Like it was mentioned before, one of the difficulties in Principal Component Analysis is that PCA is sensitive to the measurements of the variables. Figure 10 presents two biplots of PCA results based on unscaled and scaled data. The first figure which corresponds to the PCA based on the original unscaled data shows that the pile corresponding to Monetary variable is very large compared to the Frequency and Recency, leading to the conclusion that this variable has the most substantial variance. This will break down the effect of the Recency and Frequency in the model. Meanwhile, in the second PCA biplot based on the scaled data all three variables have almost equal piles(variances). Moreover, we observe that the piles corresponding to Monetary and Frequency variables are following the same direction while the Recency pile follows an entirely different direction. Finally, all piles have comparable margins. Thus, from this comparison, we can infer that different measurements do indeed affect the PCA results in this case and we should scale the data before performing PCA to avoid inaccurate estimation results.

	PC1	**PC2**	**PC3**
Recency	+0.4023	-0.9154	+0.0012
Frequency	-0.6472	-0.2854	-0.7068
Monetary	-0.6474	-0.2835	+0.7073
Eigenvalue	1.3914	0.9024	0.4995
PVE	0.6453	0.2715	0.0832

Table 9: PCA results

The Table 9 illustrates the principal component loadings, eigenvalues and the variation that is explained by each of these components. In order to determine how many principal components we should use we consider two different criteria, the Elbow rule, and the Kaisers rule. As regards the Elbow rule, we select the optimal number of principal components by analysing the scatterplot based on the PCA results. In this case, Figure 11 suggests to use only one PC since the elbow in the graph is positioned before the number of components approaches to 2. Another way of determining the optimal number of components is by Kaiser's rule, according to which we have to retain those dimensions that have eigenvalues greater than one. From Table 9 we observe that only the first PC has larger than one eigenvalue. So, Kaiser rule also suggests the use of only one principal component, which explains the 64.53% of the total variation. The first two PCs together explain 90.24% of variation. In our case, we are going to use only one principal component as the criteria suggest as it is enough to avoid loss of information meanwhile making them more interpretable.



(a) unscaled

(b) scaled

Figure 10: PCA biplots

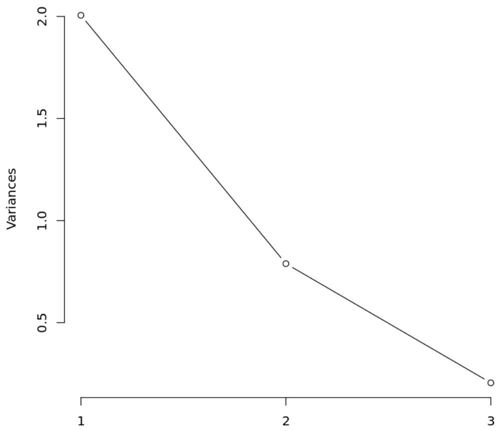


Figure 11: Scatterplot based on scaled PCA

4.3.2 K-means and K-means PCA results

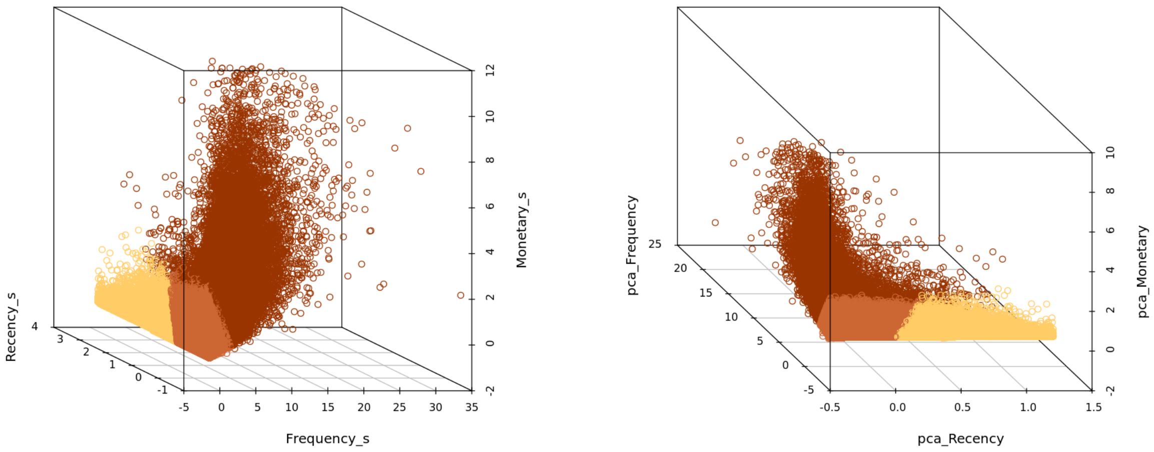
Table 10 represents the results of both standard K-means and K-means PCA. Two different approaches are performed in order to determine the clusters. In the first approach, clusters are calculated using standard K-means clustering technique based on customer's RFM variables. On the other hand, the second approach uses PCA weighted RFM variables. From this procedure three clusters are created: Good(1), Better(2) and Best(3).

The upper part of Table 10 illustrates the results based on standard K-means where we see that Best customers are defined as the customers with the least Recency and the largest Frequency and Monetary values corresponding to the third row of the table. Better customers follow, with quite low Recency and quite high value in Frequency and Monetary variables which corresponds to the second row of the same sub-table. Last but not least, Good customers' segment is specified by the lowest Frequency and Monetary and the highest Recency described in the first row of the upper table. Best customers represent the smallest segment with the size of 9.35% of all customers, while the better segment is the largest one with 67.19%. Moreover, in clustering approaches, the goal is usually to get high similarity within each group, and low similarity between each group. High similarity within a group means low variance within the cluster. Here, it is clear that the within sum of squares for each cluster is quite low, which indicates that there is high similarity within each group. The next step is to determine the similarity between the groups. Between-cluster sum of squares illustrates these similarities. In this case, the low similarity between the groups means a high variance between the clusters. In our case, the between-cluster sum of squared is quite high, which suggests that the similarity between the clusters is low. Besides, it is known, that in optimal clustering, since the clusters are different from each other, then most of the total variance is explained by the variance between the groups. Thus, since the variance within each group is small, it would explain only a small fraction of the total variance in the data. In general, the total sum of squares is established as the sum of the between-cluster and within-cluster sum of squares. The lower part of Table 10 represents the results of K-means PCA clustering where we use the same procedure to define the clusters. It is evident that the sizes of the clusters remain relatively the same with very small deviations, which indicates that the results are quite accurate. However, using different weights for each of the variables leads to smaller variance within the clusters which is an improvement over standard K-means. More specifically, the within-cluster sum of squares using PCA is approximately half of the within-cluster sum of squares in K-means without PCA. In this case, the similarity within each group is high, which means that the clusters are defined much better. The less variation we have within clusters, the more homogeneous the data points are within the same cluster. Finally, we observe that the between-cluster sum of squares is higher compared to the total variance within the clusters and most of the total variance is explained by the variance between the clusters.

K-means						
Cluster	Recency	Frequency	Monetary	Size	Within sum of squares	Label
1	1.5913	-0.4674	-0.4528	23.46 %	279154.61	Good
2	-0.4628	-0.1483	-0.1696	67.19%	194341.11	Better
3	-0.6657	2.2369	2.3535	9.35%	77952.09	Best
Between cluster:	698784.64	**Total sum squares:**	1250232.44	**Total within cluster:**	551447.80	
K-means PCA						
Cluster	Recency	Frequency	Monetary	Size	Within sum of squares	Label
1	0.6053	-0.3138	-0.3050	25.13%	16607.07	Good
2	-0.1944	-0.0837	-0.0978	65.74%	70842.94	Better
3	-0.2660	1.4670	1.5443	9.13%	114629.38	Best
Between cluster:	951144.19	**Total sum squares:**	1153223.58	**Total within cluster:**	202079.39	

Table 10: K-means and K-means PCA results

Figure 12 illustrates the clusters of K-means and K-means PCA, respectively. Specifically, in Figure 12(a), three groups of customers are represented with different colors. Best customers are illustrated using the orange color, and as it can be seen from the graph that they have the smallest Recency while at the same time their Frequency and Monetary value is quite high. The largest group which is the Better customers is represented using the brown color. Here, the Recency is larger than the previous group of customers, and the Frequency and Monetary value are high. Good customers belong in the yellow cluster with high Recency and small Frequency and Monetary value.



(a) K-means Standard

(b) K-means PCA

Figure 12: Scatterplots of K-means and K-means PCA clusters

In Figure 12(b) the clusters of customers are defined as follows. The orange color corresponds to Best customers, the Recency of which is very low and the Frequency and Monetary value is very high. The brown color represents Better customers where Recency remains low but compared to the previous group is still higher. The Frequency and Monetary values are again high. Finally, Good customers represented by yellow color have the highest Recency and the smallest Frequency and Monetary values. It is obvious that the graphs of the clusters are quite similar in both approaches as regards the size and the definitions of the groups. It is noticeable that the approach K-means with PCA produces the clusters that are better visualized, and the differences between the groups are more distinct. So, we can conclude that K-means PCA which combines PCA with K-means clustering and determines the importance weights statistically, provides more accurate results. We have come to this conclusion based on the Between and Within cluster sum of squares. For this reason, in the remaining of the research will be used as the primary method to examine the rest research questions.

4.4 Decision Tree Results

The previous section was shown that K-means with PCA is a preferable approach for clustering customers compared to standard K-means since it also overcomes the drawback of RFM-based segmentation with regards to the importance weights. So, as a class variable for our decision tree, we use the cluster-variable created by K-means PCA model which takes value 1 if the customer is from the Good segment, 2 from Better segment and 3 from the Best segment. Moreover, for building the tree we use as attributes the variables produced by RFM model, namely the variables of Recency, Frequency, and Monetary variables but unlike K-means PCA, we also add demographic variables of Age, Gender, Allow-analysis, Opt-in and Loyal-time.

Figure 13 presents the tree based on Gini-index impurity measure. From Figure 13, we see that three out of seven variables have been considered as necessary for the splits which are the variables of Recency, Frequency, and Monetary value. Customers who have Recency larger than 202 ends up in class 1 and what is surprising is that the Recency variable is the most reliable identifier for group membership in class 1, which we defined earlier as the Good customers. 26% of all customers are assigned to this class. What we observe is that the classification of customers into class 2 and 3 is only based on their Monetary and Frequency values. From the remaining 74% of customer population 65% is classified as class 2 customers (Better customers) and 8% as class 3 customers (Best customers).

As it was mentioned in Section 4.5 pruning of the tree is the optimal way of obtaining less complicated and purer tree by removing the sections that provide little power to classification instances. Figure 14 represents the pruned standard tree which is much smaller and with fewer leaves than the earlier unpruned tree from figure 13. We also observe that now only variables Recency and Monetary value are used for the splits. We also observe that 26% of all customers with larger than 202 Recency value are assigned to class 1 while from the remaining 74% of customers, the ones with lower than 1394 Monetary value are assigned to class 2(65%) and remaining 8% with larger Monetary value are assigned to the class 3. Unlike the unpruned tree, the pruned tree only employs the variables Recency and Monetary for the tree splits.

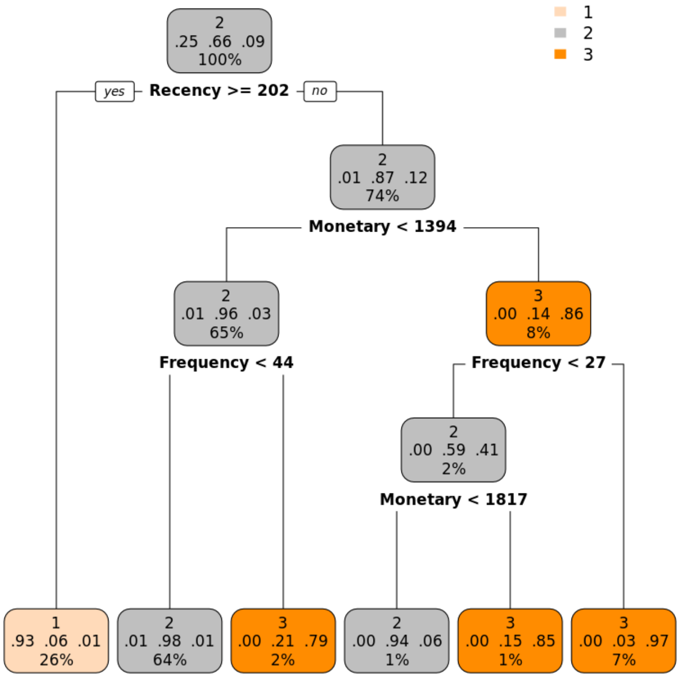


Figure 13: PCA decision tree

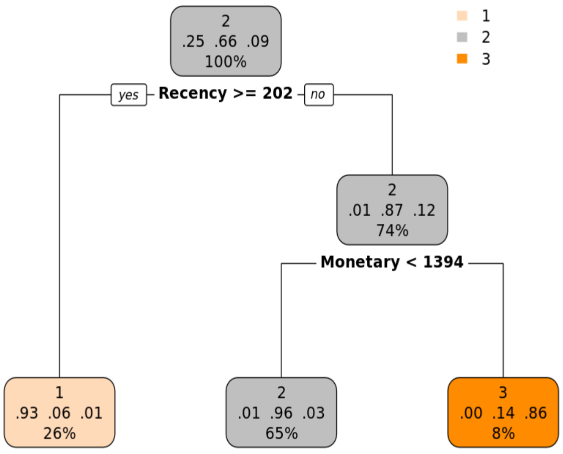


Figure 14: PCA Pruned decision tree

Given the pruned tree, the transition probabilities are estimated for each customer has a particular likelihood of staying in the same class or moving to the other two classes. Using the algorithm 3 we update new classes for all customer and analyse the number of customers that are likely to move from one particular class to another. Table 11 displays the size of the type of customers with similar characteristics that are likely to move from one class to another or stay in the same class. We observe that in case of standard prediction, the Good customer segment is the most homogeneous group where 97.3% of customers are likely to stay in this group. Customers in Good segment who are likely to stay, are mostly men, the largest age-group is 45-54, and the majority of these customers allow both to analyse their data and send commercial emails. Only, 2.7% of customers in Good segment are likely to move to Better segment. We observe that only 39.4% of them allow to analyse their data, but 52.9% does allow to send them commercial emails. Finally, there are no customers in this class who are likely to move from Good segment to Best segment. Unlike Good segment, Better one is less homogeneous as 97% of Better customers are likely to stay in the same segment. This segment contains mostly men, and the major age-group is 65 and older. Compared to the Good segment, Better customers almost equally often allow to analyse their data and more likely to permit to send them commercial emails. 1.9% of Better customers are likely to move to Best class. Moreover, we observe that compared to the stable Better group, these customers are younger, 45-54 age-group and more often allow to analyse their data and send them commercial emails. Finally, 1.1% customers in Better segment are likely to drop to Good segment who are predominantly women compared to the ones who are likely to move to Best segment less often allow to analyse their data and send commercial emails, 53.1% compared to 55.8% and 56.6% compared to 59.2%, respectively. Best customers' segment is the most inharmonious segment. 79.2% of Best customers are likely to stay in this segment, and these customers are predominantly female. We observe that these customers are mostly people over 65 and they often don't allow to analyse their data but allow to send them commercial emails (35.1% and 56.8%). We observe that a large amount of Best customers, 20.3% are likely to move from this segment, and most of them are male(60%). Only 60% of these customers allow to both analyse their data and send them emails which is much higher than customers in Best group willing to stay in Best group or to move to Better one. Finally, there are no customers who are likely to move from Best to Good segment.

Prediction PCA-adjusted					
Transition	**Size**(%)	**Gender**(%)	**Age**(%)	**Allow analysis**(%)	**Opt-in**(%)
Good → Good	97.259	44.926	45-54(21.143)	52.978	51.281
Good → Better	2.741	49.428	55-64(21.915)	39.379	52.859
Good → Best	0.000	0.000	0.000	0.000	0.000
Better → Better	97.006	43.668	65+(23.252)	51.495	57.493
Better → Best	1.853	40.489	45-54(22.046)	55.752	59.194
Better → Good	1.141	51.945	65+(25.921)	53.052	56.616
Best → Best	79.204	62.162	65+(50.000)	35.135	56.757
Best → Better	20.796	40.000	65+(47.368)	60.000	60.000
Best → Good	0.000	0.000	0.000	0.000	0.000

Table 11: Perdition results based on transition probabilities

4.5 Finite Mixture Model Results

Table 12 presents the results of Finite Mixture model using Negative Binomial distribution. For the model selection, four different number of components are used k = 3,...,6. The number of purchases per customers over 2 years has been used as a dependent variable in the model, the Monetary value and Recency has been used as independent variables. For each mixture regression, all components with their corresponding coefficient estimates, sizes and BIC are presented. We observe that in all cases all coefficients estimates are highly significant. However, the margins of intercept estimates are much larger than the ones of the independent variables monetary value and Recency. Based on the BIC values for different k's we observe that the model with 6 components is the most optimal model. Therefore, we focus on this model results for further analysis.

In case of finite mixture model combined with Negative Binomial density, the estimated coefficient can be interpreted in the following way: 1% increase in independent variable x_{ij} leads to $\hat{\beta}_j$ % increase in the dependent variable, ceteris paribus. In the case of k = 6, we observe that cluster 2 is the largest one and cluster 5 is the smallest. We observe that for cluster 2, the coefficient of monetary value is large positive(0.056) compared to the other clusters indicating that 1% increase in monetary leads to an increase 0.056% in Frequency,keeping the other variables fixed while the coefficient of Recency is -0.077 indicating a negative correlation between Recency and Frequency. These results suggest that customers from this cluster tend to purchase more often if they spent more but also lower Recency (more lately customer has purchased a product) more frequent he buys. This group can be considered as the Better group of customers based on the above conclusions. The reason for defining cluster 2 as Better but not the Best segment is that in terms of both Monetary and Recency cluster 5 perform better, meaning that customers in cluster 5 purchase more often once they spent more and once they have purchased something lately. Therefore, we can refer to cluster 5 as the Best cluster. In case of cluster 3, we observe that even though the relation between Monetary and Frequency is significant indicating that more they spent more often they come, but this relation is not very strong since the coefficient is equal to 0.009. The margin of the Recency coefficient is also very small but still negative, namely -0.003. We can define this as the Good customer class. Cluster 1 we can define as the new-customer class since the coefficient of Monetary is low(0.005) but the coefficient of Recency is substantial. The reason for this is that even though we observe that more lately they order more frequently they buy but the low coefficient is an indication for a low level of spending pattern related to this frequent purchases. Cluster 4 we can classify as At Risk segment which contains customers who spent a lot of money and often shopped but a long time ago. his is evidenced by the positive correlation between Recency and Frequency and the high value of the coefficient of Monetary variable. Finally, the last cluster which is obtained is cluster 6, that contain the remaining customers.

Table 13 describes in detail the characteristics of customers per segment for the selected model with 6 components described earlier. What we observe is that the largest segment which we have called Better segment with the size 31.80% consists mostly of males, 58% men versus 42% female. This is

	Component	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
	Coef. Estim.						
k = 3	**BIC:** 3542046						
	Intercept	2.148***	1.304***	3.718***			
	Monetary	0.003***	0.032***	0.019***			
	Recency	-0.023***	-0.045***	0.008***			
	Cluster Size(%)	38.70	45.90	15.40			
k = 4	**BIC:** 3511813						
	Intercept	3.169***	2.597***	3.805***	1.007***		
	Monetary	0.011***	0.002***	0.0018***	0.049***		
	Recency	-0.006***	-0.054***	0.012***	-0.063***		
	Cluster Size(%)	23.10	12.00	21.60	43.30		
k = 5	**BIC:** 3372539						
	Intercept	3.4801***	1.780***	3.359***	0.953***	2.913***	
	Monetary	0.037***	0.008***	0.005***	0.081***	0.016***	
	Recency	-0.061***	-0.074***	-0.002***	-0.093***	0.011***	
	Cluster Size(%)	31.60	11.50	28.10	11.00	17.80	
k = 6	**BIC:** 3356910						
	Intercept	3.777***	2.796***	0.852***	3.256***	2.037***	2.429***
	Monetary	0.005***	0.056***	0.009***	0.023***	0.099***	0.000***
	Recency	-0.085***	-0.077***	-0.003***	0.015***	-0.114***	-0.003***
	Cluster Size(%)	9.87	31.80	19.70	15.53	9.70	13.40

Table 12: Estimation Results of Finite-Mixture Model

much higher than all the other cluster's average male percentage. Also, compared to the remaining clusters this cluster's most frequent age-group is 45-54 while for the other clusters it holds that the most frequent age-group is 65+. Finally, this cluster contains the customers who mostly do not allow to analyse their data (48%) and compared to other clusters they also less often allow to send them commercial emails(54%).

Finite Mixture Model Clusters					
Cluster	**N(%)**	**Gender(%)**	**Age(%)**	**Allow analysis(%)**	**Opt-in(%)**
1	9.870	51.258	65+(30.060)	59.330	55.432
2	31.800	41.694	45-54(20.262)	47.847	54.288
3	19.700	52.643	65+(26.941)	56.507	57.368
4	15.530	52.551	65+(25.427)	56.398	57.920
5	9.700	55.527	65+(33.164)	55.676	62.083
6	13.400	51.813	55-64(23.895)	56.158	60.042

Table 13: Finite mixture clusters characteristics

5 Robustness Checks

In this section, we aim to verify the results obtained earlier and to check how these results change when there is a change in the model settings. Firstly, we aim to compare the performance of RFM-based segmentation to the K-means results. One way of doing so is in terms of comparing Best segment customers. Namely, both methods create a Best customer segment, and consequently, the customers who belong to the Best segment in the RFM-based segmentation should also belong to the Best segment of K-means. For this purpose, we examine the customers of Best segment from RFM-based segmentation and K-means respectively. Moreover, we do the same for Best segments from RFM-based segmentation and K-means PCA. We find that 64.45% Best customers from RFM-based segmentation are also Best customers in K-means clustering and 63.11% Best customers from RFM-based segmentation are Best customers in K-means PCA clustering. These results suggest that RFM-based segmentation leads to comparable results to K-means and K-means PCA clustering. Besides, we observe that the results are more similar in K-means case than in K-means PCA case. The 35% difference between these clustering results might be caused by the fact that RFM-based segmentation creates segments based on the assumption created by the user, which leads to less accurate results than the K-means or K-means PCA.

As it was mentioned earlier, beside numerous advantages of K-means, there are also a few disadvantages. Namely, K-means results are highly dependent on the initialization and the chosen distance measure. According to Singh et al. (2013), the selection of distance metric is critical in clustering and could be a way to test the reliability of the results. For this purpose, we perform K-means clustering using different distances and different centroids. More specifically, five different cases are going to be examined in this section. Firstly, we perform K-means using the Manhattan distance and random initial centroid. The next case is to consider is using Manhattan distance and pre-specified centroid (starting centroid equal to 25). Then we apply the K-means clustering using Euclidean distance with random initial centroid and in the next case, we pre-specify the starting point to be equal to 25. Finally, we compare all these cases with K-means PCA in order to check whether the results are accurate or not.

Table 14 presents the results of each of the cases. It is visible that the total within-cluster sum of squares is largest in comparison with the other 3 cases. Also, in the cases using Manhattan distance and Euclidean distance with pre-specified centroid the Within-cluster variation is also high. The smallest variation within clusters is obtained using Euclidean distance with a random initial centroid in case of K-means PCA, as illustrated in Table 10, which indicates that there is high similarity within each group. Moreover, as regards the variance between the clusters, the highest between-cluster sum of squares is obtained again in K-means PCA using Euclidean distance with random initial centroid, which means in this case that the similarity between the groups for this case is the lowest. So, we observe that as expected different distance measure leads to different k-means results especially regarding the between-cluster variance and Within-cluster variance. Moreover, K-means PCA which is implemented using Euclidean distance and random initial centroid gives better results than K-means based on Manhattan distance and Euclidean distances with random or pre-specified initial centroids. Finally, we observe that changing the distance measure leads to larger differences between different K-means results while the changes in centroid lead to only small changes in the results.

K-means **Manhattan distance and Random centroid**						
Cluster	Recency	Frequency	Monetary	Size%	Within sum of squares	Label
1	1.3945	-0.4756	-0.4614	27.74	337618.12	Good
2	-0.5153	-0.1207	-0.1406	63.09	287841.24	Better
3	-0.6734	2.2706	2.3647	9.17	81925.32	Best
Between cluster:	871456.17	**Total sum of squares:**	1578840.85	**Total within cluster:**	707384.68	
K-means **Manhattan distance and pre-specified initial centroid**						
Cluster	Recency	Frequency	Monetary	Size%	Within sum of squares	Label
1	1.4557	-0.4766	-0.4619	26.31	328099.22	Good
2	-0.4951	-0.1531	-0.1738	63.10	295618.32	Better
3	-0.6686	2.0985	2.1856	10.59	82231.87	Best
Between cluster:	816744.39	**Total sum of squares:**	1522693.8	**Total within cluster:**	705949.41	
K-means **Euclidean distance and pre-specified initial centroid**						
Cluster	Recency	Frequency	Monetary	Size%	Within sum of squares	Label
1	1.6111	-0.4676	-0.4529	23.02	271736.34	Good
2	-0.4531	-0.1715	-0.1938	66.37	212596.73	Better
3	-0.6611	2.0884	2.1961	10.61	74264.18	Best
Between cluster:	726779.21	**Total sum of squares:**	1285376.46	**Total within cluster:**	558597.25	

Table 14: K-means results using different distances and initial centroids

The Calinski and Harabasz (CH) index Caliński and Harabasz (1974) attempts to measure the effectiveness of the clustering. This index is the ratio between the dissimilarity and tightness of an average cluster. The term dissimilarity measures how a cluster is different from other separate clusters, while the term tightness measures how members within a cluster are similar to each other. Thus, a high CH index would indicate high dissimilarity between clusters and very strong tightness within a cluster, in other words, low distances among the members, suggesting robust clustering.

CH index	
K-means	431946.57
K-means PCA	427197.17

Table 15: CH-index

Table 15 represents the results of the CH index in K-means clustering with and without PCA using Euclidean distance. It is obvious that, the value of the index is very high and similar in both cases. This indicates that the dissimilarity between clusters is very high and the variation within clusters low, which suggests robust clustering.

6 Summary and Conclusions

In this analysis, we have examined different methods for customer segmentation and techniques that can improve the performance of these methods. Firstly, we have used the scores created by RFM model and made a segmentation based on these scores for different possible combinations of R, F and M scores. Secondly, we have used multivariate data method PCA to determine the weights of Recency, Frequency, and Monetary variables in order to drop the assumption of equal weights and used this in the combination of K-means. We compare K-means and K-means PCA results from which we conclude that Best segment has the smallest size, Good one the medium size, and Better the largest size. Based on Within sum of squares and Between sum of squares K-means PCA produces more concentrated clusters than standard K-means. Therefore, we can infer that by assigning different weights for each variable while considering the proportion of variation captured by all variables, the multivariate statistics technique PCA enhances customer segmentation method.

Furthermore, we used decision tree to predict the movement of customers between segments and analyze the characteristics of these customers for all possible scenarios. We concluded that the largest transition is likely to happen in the Best segment where 20.8 % of customers are likely to move from Best to Better segment. Furthermore, we observed that there 2.7% Good segment customers who are likely to move to the Better segment, 1.9% customers likely to move from Better to Best and 1.1% from Better to Good segment. The prediction of transition probabilities between the segments is an essential way of identifying the possible turners and potential customers in order to prevent turners for the particular movement or encourage the potential customers to move from less preferable class to preferable one. These results can directly be used in the marketing campaign for the purposes mentioned above. So, instead of targeting all customers with the same way it would be more optimal if the customer group who are likely to move from Best to Better will get the most attention. Customers from Good to Better and Better to Best also need special targeting approach. Therefore, instead of targeting all these customers in the same way, one can better target them differently. Firstly, the loyalty card should be different per segment such as the card types Loyal, Golden and Premium for the segments Good, Better and Best, respectively. This will make sure that they stay in their segments or encourage them to move to a preferable segment in case of Good and Better segments. However, as we mentioned earlier, unstable customers should get something else besides these segment-specific loyalty cards or the context of commercial emails should be differentiated too. More specifically, the customers from Best to Better group who are predominantly older than 65 males who allow both to analyse their data and to send them commercial emails with new arriving products such that they will be the first ones with the opportunity to buy these products.

In order to encourage customers from Better segment move to the Best segment who are predominantly male with most frequent age-group of 45-54 they can be targeted, for instance, in terms of birthday gift-card of some monetary amount that they can spend on the next purchase. Finally, the group of Good customers who are likely to move to Better segment can save up points with their loyalty card for each spent. Once they achieve some specified maximum amount of point they can get gift-card which can be spent on their next purchase.

When taking into account heterogeneity by implementing the Finite Mixture model, the analysis suggests that there should be another type of segmentation with more clusters of customers. Finite mixture models provide six clusters instead of the three clusters that are pre-specified in the K-means clustering method, namely the Best, Better, Good, New, At Risk and Other customer segments. From this procedure, customers with homogeneous attitudes are grouped together when at the same time there are heterogeneous perceptions across segments. In addition, this supervised method, apart from the heterogeneity of consumers, it also takes into account the relationship between the variables which is not the case in the other clustering methods. Furthermore, all clusters have 65 and older as the most frequent age-group. Moreover, the Better cluster stands out with a higher percentage of males compared to the remaining clusters and customers in this segment allow less often to analyze their data.

Based on all the results obtained and mentioned earlier, we can conclude that PCA is an appropriate multivariate method to improve the standard clustering methods like K-means and Classification trees. Moreover, we find that decision tree can be used to predict the movement of customers between segments in order to analyze these customers and prevent or encourage particular movement from one segment to another. Finally, taking into account personal heterogeneity, finite mixture models lead to different segmentation results than the previous approaches, and it suggests the use of larger number of segments, clustering all the customers into smaller subgroups with homogeneous attributes.

7 Limitations

Throughout this analysis, we came across different problems and limitations which have affected our analysis. First of there was a limitation in terms of marketing instruments, for instance, instruments like display or feature, which could be used to analyze the effect of different campaigns on different segments. This would undoubtedly lead to more accurate predictions. Moreover, because of the limited amount of periods in the data, not all holidays have been taken into account during the process of detrending and deseasonalization. Only the months October, November, and December have been considered since the data of the leap year 2016 was incomplete, only the data of above mentioned months have been supplied. For imputation, due to the incapacity of the program and the large data sets, more robust imputation methods for handling missing observations could not be applied, for instance, the multiple imputation method or Detect Deviating Cells (DDC) algorithm for handling both missing values but also outliers. Finally, we encounter limitations in terms of supplied demographic variables such as zip-codes or income which could be included in the clustering methods to produce more accurate results.

References

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Chan, C.-C. H. (2005). Online auction customer segmentation using a neural network model. *International Journal of Applied Science and Engineering*, 3(2):101–109.

Chang, E.-C., Huang, S.-C., Wu, H.-H., and Lo, C.-F. (2007). A case study of applying spectral clustering technique in the value analysis of an outfitters customer database. In *2007 IEEE International Conference on Industrial Engineering and Engineering Management*, pages 1743–1746. IEEE.

Cheng, C.-H. and Chen, Y.-S. (2009). Classifying the segmentation of customer value via rfm model and rs theory. *Expert systems with applications*, 36(3):4176–4184.

Coussement, K., Van den Bossche, F. A., and De Bock, K. W. (2014). Data accuracy's impact on segmentation performance: Benchmarking rfm analysis, logistic regression, and decision trees. *Journal of Business Research*, 67(1):2751–2758.

Duchessi P., K. E. (2013). Decision tree models for profiling ski resorts' promotional and advertising strategies and the impact on sales. 40(15):5822–5829.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541.

Hauer, K., Rost, B., Rütschle, K., Opitz, H., Specht, N., Bärtsch, P., Oster, P., and Schlierf, G. (2001). Exercise training for rehabilitation and secondary prevention of falls in geriatric patients with a history of injurious falls. *Journal of the American Geriatrics Society*, 49(1):10–20.

Haughton, D. and Oulabi, S. (1993). Direct marketing modeling with cart and chaid. *Journal of direct marketing*, 7(3):16–26.

Hsieh, N.-C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert systems with applications*, 27(4):623–633.

Hughes, A. M. (1996). Boosting response with rfm. *Marketing Tools*, pages 4–8.

Hughes, A. M. (2000). *Strategic database marketing: the masterplan for starting and managing a profitable, customer-based marketing program*, volume 12. McGraw-Hill New York.

Khajvand, M. and Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted rfm model in retail banking context. *Procedia computer science*, 3:1327–1332.

King, G., Gakidou, E., Ravishankar, N., Moore, R. T., Lakin, J., Vargas, M., Téllez-Rojo, M. M., Hernández Ávila, J. E., Ávila, M. H., and Llamas, H. H. (2007). A politically robust experimental design for public policy evaluation, with application to the mexican universal health insurance program. *Journal of Policy Analysis and Management*, 26(3):479–506.

Kotler, P. and Armstrong, G. (2010). Principles of marketing: Upper saddle river.

Kotler, P. and Keller, K. (2011). *Marketing management 14th edition*. Prentice Hall.

Lee, E., Lim, N., and Park, H. (1998). Nursing medical research and statistical analysis. *Seoul: Soomoonsa*.

Li, Y.-M., Lin, C.-H., and Lai, C.-Y. (2010). Identifying influential reviewers for word-of-mouth marketing. *Electronic Commerce Research and Applications*, 9(4):294–304.

Lord, D. and Mannering, F. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation research part A: policy and practice*, 44(5):291–305.

Lumsden, S.-A., Beldona, S., and Morrison, A. M. (2008). Customer value in an all-inclusive travel vacation club: An application of the rfm framework. *Journal of Hospitality & Leisure Marketing*, 16(3):270–285.

Miglautsch, J. (2002). Application of rfm principles: What to do with 1–1–1 customers? *Journal of Database Marketing & Customer Strategy Management*, 9(4):319–324.

Miglautsch, J. R. (2000). Thoughts on rfm scoring. *Journal of Database Marketing & Customer Strategy Management*, 8(1):67–72.

Poch, M. and Mannering, F. (1996). Negative binomial analysis of intersection-accident frequencies. *Journal of transportation engineering*, 122(2):105–113.

Rigdon, E. E., Ringle, C. M., Sarstedt, M., and Gudergan, S. P. (2011). Assessing heterogeneity in customer satisfaction studies: across industry similarities and within industry differences. In *Measurement and Research Methods in International Marketing*, pages 169–194. Emerald Group Publishing Limited.

Rubin, D. B. (1975). Inference and missing data. *ETS Research Bulletin Series*, 1975(1):i–19.

Shui Hua Han, Shui Xiu Lu, S. C. L. (2012). A study on customer segmentation of telecom customers based on customer value by decision tree model. 39(4):3964–3973.

Singh, A., Yadav, A., and Rana, A. (2013). K-means with three different distance metrics. *International Journal of Computer Applications*, 67(10).

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing*, 21(1):3–8.

Sohrabi, B. and Khanlari, A. (2007). Customer lifetime value (clv) measurement based on rfm model.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological methods*, 17(2):228.

Wei, J.-T., Lin, S.-Y., and Wu, H.-H. (2010). A review of the application of rfm model. *African Journal of Business Management*, 4(19):4199–4206.

Zou, Y., Zhang, Y., and Lord, D. (2014). Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. *Analytic methods in accident research*, 1:39–52.

Appendix

List of Tables

Variable	Explanation
`Customer-ID`	The unique ID of each customer
`Recency`	The time difference (in terms of days) between the date of analysis and the date of last purchase for each customer
`Frequency`	The total amount of transactions that the customer has made over the entire specified time period
`Monetary`	The total amount that each customer has spent over the entire specified time period
`Gender`	Dummy variable that takes value 1 if gender is female and 0 if gender is male
`Age-category`	Categorical variable that takes value 1,2,3,4,5,6 for the age groups 18-24,25-34 35-44,45-54,55-64,65+ respectively
`Allow-analysis`	Dummy variable that takes value 1 if customer allows to analyse his/her data
`Opt-in`	Dummy variable that takes value 1 if customer allows to send him/her commercial emails
`Loyal-time`	The amount of time that a customer has stayed with the company

Table 16: Variables description

Period	**Frequency**	**Customers**	**Total Itm.**	**Total Ord.**	**Revenue**	**Items**	**Prices**
10-12/2018	1x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	2x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	3x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	4x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	5x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	6x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	7x+	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Subtotal		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
7-9/2018	1x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	2x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	3x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	4x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	5x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	6x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	7x+	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Subtotal		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
4-6/2018	1x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	2x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	3x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	4x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	5x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	6x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	7x+	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Subtotal		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
1-3/2018	1x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]]	[redacted]
	2x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	3x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

	4x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	5x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	6x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	7x+	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Subtotal		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
10-12/2017	1x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	2x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	3x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	4x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	5x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	6x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	7x+	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Subtotal		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
7-9/2017	1x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	2x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	3x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	4x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	5x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	6x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	7x+	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Subtotal		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
4-6/2017	1x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	2x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	3x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	4x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	5x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	6x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	7x+	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Subtotal		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
1-3/2017	1x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	2x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	3x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	4x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	5x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	6x	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	7x+	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Subtotal		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 17: Summary sales ordered by recency and frequency [1]

[1] According to the Non-Disclosure Agreement, sensitive information that may reflect revenue of the company should be censored.