Delft University of Technology
Faculty of Electrical Engineering, Mathematics & Computer Science
Delft Institute of Applied Mathematics

Customer segmentation using RFM analysis

Dutch title: Klantsegmentatie met behulp van RFM-analyse

to obtain the degree of

BACHELOR OF SCIENCE
in
APPLIED MATHEMATICS

by

J.M. van Burg

Delft, The Netherlands
July 2020

BSc thesis Applied Mathematics

Customer segmentation using RFM analysis

J.M. van Burg
Student number: 4654730

Delft University of Technology

Defended publicly on July 6th 2020

Supervisor

Dr. J. J. Cai

Other committee member

Dr. J. W. van der Woude

Abstract

This paper is a research on the segmentation of customers. The clustering of customers is done based on the variables recency, frequency and monetary value. Such a clustering is called an RFM-model. The clustering is done using the K-means clustering method. To find the optimal number of clusters the following performance metrics are used: Elbow method, Silhouette Analysis, and Davies-Bouldin Index.The RFM-model is extended by introducing the loyalty variable. This model is called an RFML-model. Lastly further clustering is done within one of the clusters.

Keywords: Segmentation of customers, RFM-model, K-means method, Elbow method, Silhouette Analysis, Davies-Bouldin Index, RFML-model

Preface

This thesis 'Customer segmentation using RFM analysis' is written in order to obtain the degree of Bachelor of Science from the Delft Institute of Applied Mathematics. This research took place in the department of statistics under the supervision of J.J. Cai.

In April I started this project. Since I wanted to look inside the doors of a company I decided to do this project at the company Yoursurprise. This company really wanted to cluster their customers. Hence I started my research into clustering customers using RFM analysis.

I would like to thank my supervisor J.J. Cai for her help and guidance during this project. I would also like to thank J.W. van der Woude for being part of my thesis committee. Furthermore, I would like to thank the company Yoursurpise, and in particular thank Max Hoekman for his help during this project.

Contents

Introduction

A central problem for companies is that they have one marketing strategy for all their customers but want to know how to target specific customers with marketing strategies that are more relevant to their particular behavior. In this way, the rate of response of the customers is higher. This means more sold products. And less time nor money wasted on a never returning customer. It is especially interesting for B2C (Business to Consumer) companies, because they have a lot of different private customers, which makes it harder to see which groups of customers a company has. Instead of analyzing every customer separately, a better way is to look at the segmentation of customers.
In this project, to segment customers, the RFM-model is used (Cheng and Chen, 2009). RFM stands for Recency, Frequency and Monetary. This RFM-model is made for the company Yoursurprise. Yoursuprise is an online company that makes personalised gifts. In Chapter 1 the data that is used is discussed. In Chapter 2 is explained what an RFM-model is and how to determine the optimal number of clusters with the three methods: Elbow method, Silhouette Analyse and David-Bouldin Index. And at the end the RFM-model is composed with the K-means method. The K-means method clusters data based on the Euclidean distance [1]. In Chapter 3 the RFM-model is extended by introducing the loyalty variable. The variable loyalty is added to the model and again the optimal number of clusters is determined. In Chapter 4 we determine which model is best suited for the business. We cluster further with the RFM-model that is made in Chapter 2.

[1] Length of a straight line between two points in a n-dimensional space

1

Data analysis

The data that is used in this project is the data of the customers of the company Yoursurprise. In Figure 1.1 a part of the data is shown. We removed the customers who made there last purchase more that two years ago.

	customer_id	recency	frequency	monetary	loyalty
0	226919	641	10	235.51	3322
1	326431	150	11	236.75	2970
2	691015	330	32	1027.58	2369
3	206822	501	11	202.32	3391
4	373343	264	9	150.44	2826
...	...	...	...	...	...
537374	757098	39	8	318.93	2266
537375	253446	639	8	186.91	3212
537376	936220	503	8	168.02	1989
537377	179653	15	8	156.19	3495
537378	621659	162	8	135.25	2545

Figure 1.1: A part of the data of the customers. In the columns the customer id, recency, frequency, monetary and loyalty of the customers are shown.

First only the recency, frequency en monetary value of a customer is used. Secondly, the loyalty of the customer is taken into consideration. These variables are numeric variables. Lastly, the first purchase channel, the first purchase assortment and the first purchase month are taken into account. The last three variables are categorical.
The precise details of these variables are as follows:

- Recency: the number of days between the last purchase of a customer and the current date [1].

$$receny = current\ date - last\ purchase\ date$$

[1] Last date in the data set

- Frequency: how often a customer made a purchase in the time between the first purchase and the current date.

$$frequency = \frac{total\ times\ of\ purchases}{current\ date - first\ purchase}$$

- Monetary value: the average amount spent per purchase of a customer.

$$monetary\ value = \frac{total\ amount\ spent}{total\ times\ of\ purchases}$$

- Loyalty: how long a customers is active, the number of days between the first purchase and the current date.

$$loyalty = current\ date - first\ purchase\ date$$

- First purchase month: the month of the first purchase.

$$month = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]$$

- First purchase channel[2]: the channel the customers made their first purchase from.
- First purchase assortment: what kind of product they have bought first.

1.1 Distribution of the variables

1.1.1 Recency

In Figure 1.2 the frequency of the recent transaction time (recency) is plotted in a histogram. The most customers had a recency just above zero, which means that they recently made a purchase. Another peak is below a recency of 350 day's, this is approximately one year ago. We can see that the recency variable is not skewed, it does not have a long tail.

1.1.2 Frequency

In Figure 1.3 the frequency of the purchase frequency is plotted in a histogram. We can not see much in this figure because the data is very skewed. We split the data in 80% and 20% and plot this data in two histograms, so that the distribution is more clear in the figures. In Figure 1.4 the first 80%, of the frequency of the purchase frequency, is shown in a histogram. Around frequency 0.0014 there is a peak. The figure shows that the first 80% of the data lies between a frequency of 0.00 and 0.010. In Figure 1.5 the last 20% of the frequency of the purchase frequency is shown in a histogram. We can see that the last 20% of the data is skewed and not much can be seen in this Figure. What we do see is that the most customers have a purchase frequency just above zero. Looking at the data, what can not be clearly seen in this figure, there are two outliers with frequency 5.33 and 5.67, which make the data skewed.

[2] marketing channel

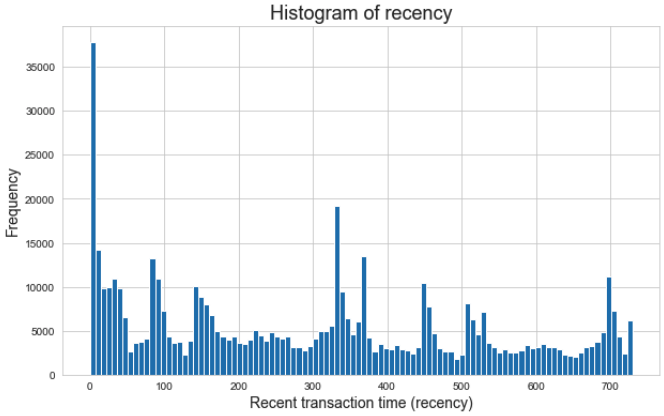

Figure 1.2: **Distribution of the recent transaction time (recency)**. It shows the distribution of the variable recency.

1.1.3 Monetary value

In Figure 1.6 the frequency of the average amount spend per purchase (monetary value) is plotted in a histogram. We can see that, like the variable frequency, the data is skewed. The data is split in the first 80% and the last 20%. In Figure 1.7 the first 80% of the frequency of the average amount spend per purchase (monetary value) is plotted in a histogram. The bar at zero monetary value indicates that the customers have returned their purchase. Local optimums are at monetary value just below 15 and around 18. In Figure 1.8 the last 20% of the frequency of the average amount spend per purchase (monetary value) is plotted in a histogram. The last 20% of the data is skewed. Most of the variable monetary value lies between the value 0.00 and 150.00. And it has a outlier of 1031.17 euros.

1.1.4 Loyalty

In Figure 1.9 the frequency of the loyalty variable is plotted in a histogram. We can again see that the data is skewed.

1.1.5 First purchase month

In Figure 1.10 we can see that most customers made there first purchase in the month May. This might be due to the corona virus. Since the corona virus, people are buying more online, and in combination of Mothers-day in the Netherlands we saw a lot of new customers in the month May. The third biggest slice is the month December, this due Christmas. The months in which the least customers bought for the first time are the July, October, August, January, September and March.

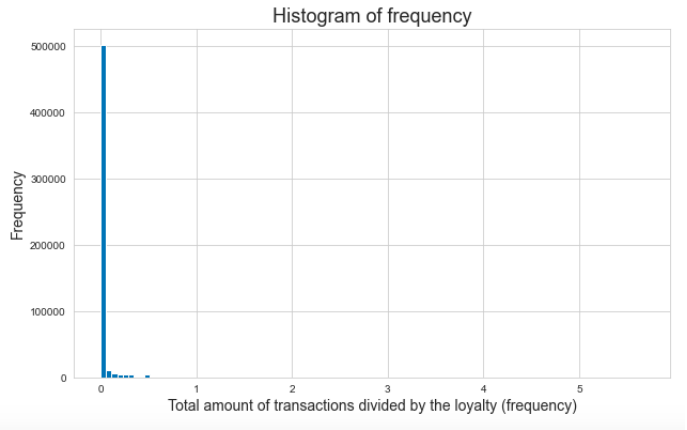

Figure 1.3: **The frequency of the purchase frequency**. It shows the distribution of the variable frequency.

1.1.6 First purchase channel

In Figure 1.11 the distribution of the first purchase channel is plotted as a pie chart. Clearly the channel Generic Paid Search is the channel from which the most customers made there first purchase. For example, a Generic paid search is when someone googles 'mug with picture' Yoursurprise pops up, without mentioning the brand Yoursurprise. The least customers made there first purchase through Branded Paid Search, this is the same as Generic Paid Search but in this case the brand must be in the google search so that Yoursurprise pops up.

1.1.7 First purchase product assortment

In Figure 1.12 the distribution of the first purchase product assortment is plotted as a pie chart. The first product of most customers is Others, these are all the products put together that did not fall in to the top 10. So we will look at the second one, this are the three products Glasses, Books and Chocolate & Food products. Those product groups are compared to the others groups not significant larger.

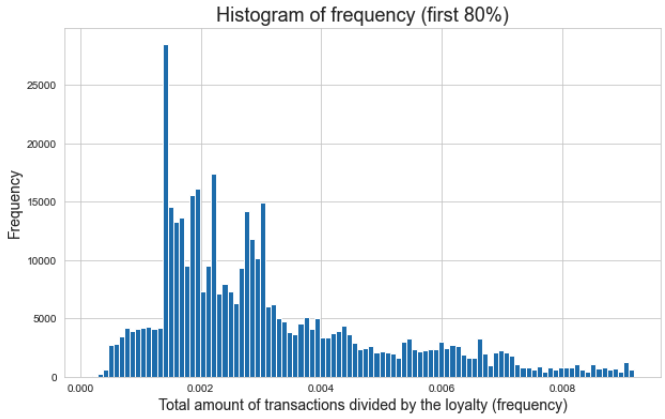

Figure 1.4: **The first 80% of the purchase frequency**. It shows the distribution of the first 80% of the variable frequency.

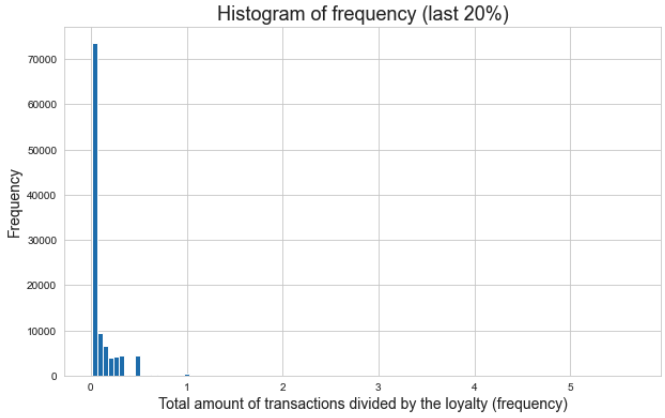

Figure 1.5: **The last 20% of the purchase frequency**. It shows the distribution of the last 20% of the variable frequency.

1.2 Transformed data

Before we cluster the data, the data is log-transformed and standardized. First some data is being log-transformed. Log transformation is used to deal with skewed data, it removes or reduce skewness (Feng et al., 2014). Since the variables monetary value and frequency are skewed, those are being log-transformed
After the log-transformations the data is being standardized.

Data is standardized to make it better compared by putting it on the same scale (Mohamad and Usman, 2013). Hence all the numeric variables are being standardized so that the variables can be compared, because all the data has now a mean of zero and a standard deviation of one. The data sets are approximately in the same range after

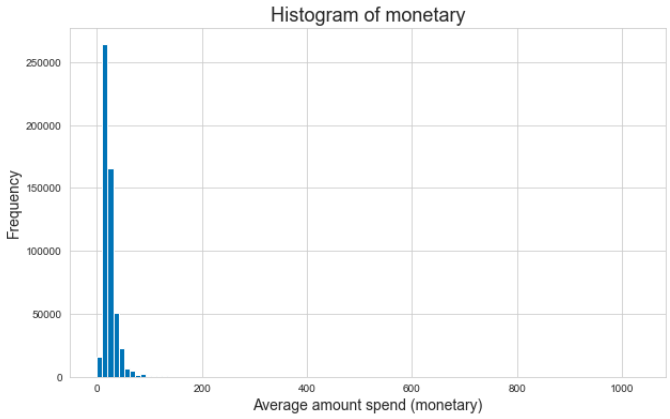

Figure 1.6: **The frequency of the average amount spend per purchase (monetary value)**. It shows the distribution of the variable monetary value.

standardizing. The variables are being standardized using the following formula:

$$z = \frac{x - \mu}{\sigma}$$

Here z is the standardized data point, x is the data point that is being standardized, μ is the mean of all the data points and σ is the standard deviation of the data points. The distribution of the variables recency, frequency and monetary after scaling can be seen in Figure 1.13 , Figure 1.14, and Figure 1.15 respectively.

In Figure 1.13 shows that the distribution of the transformed variable recency looks the same, but now scaled to the range [-1.3538, 1.9324].

In Figure 1.14 can be seen that the data is less skewed, compared to the data in Figure 1.3, where the data was not transformed. The transformed data does not look normal, we can say that the log-transformation not really helped. This is probably because when the original data follows a log-normal distribution log-transformation makes it normal, the original purchase frequency does not follow a log-normal distribution and thus the log-transformation not works (Feng et al., 2014). The range after scaling is [-0.3148, 32.7341].

In Figure 1.15 the transformed monetary value can be seen. We see that the transformed data follows now a log-normal distributed. The range after scaling is [-6.8340, 8.4819].

After transforming the data we can use it to make the RFM-model.

In Figure 1.16 the distribution of the transformed loyalty variable is shown. Here we see that the data is less skewed.

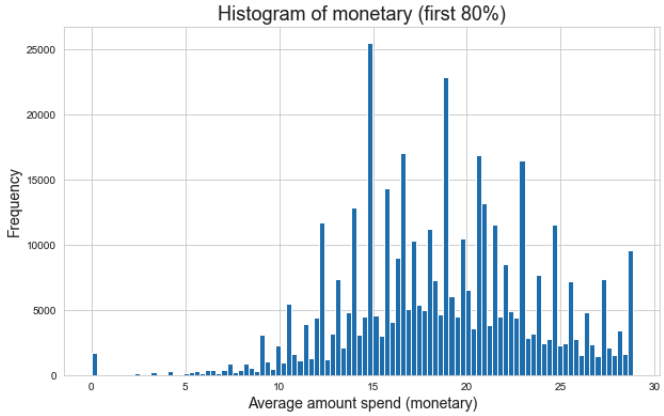

Figure 1.7: **The first 80% of the frequency of the average amount spend per purchase (monetary value)**. It shows the distribution of the first 80% of the variable monetary value.

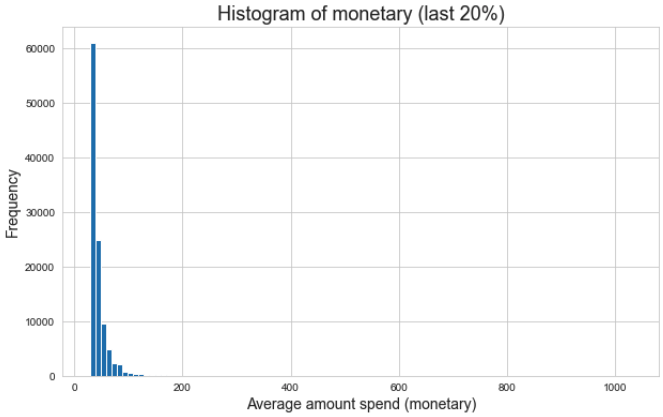

Figure 1.8: **The last 20% of the frequency of the average amount spend per purchase**. It shows the distribution of the last 20% of the variable monetary value.

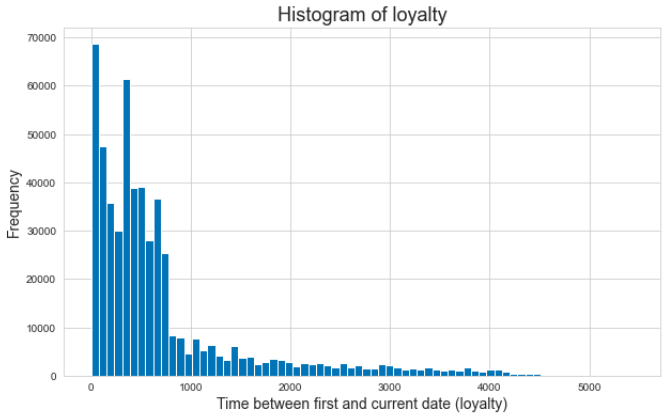

Figure 1.9: **The frequency of the loyalty. It shows the distribution of the loyalty.**

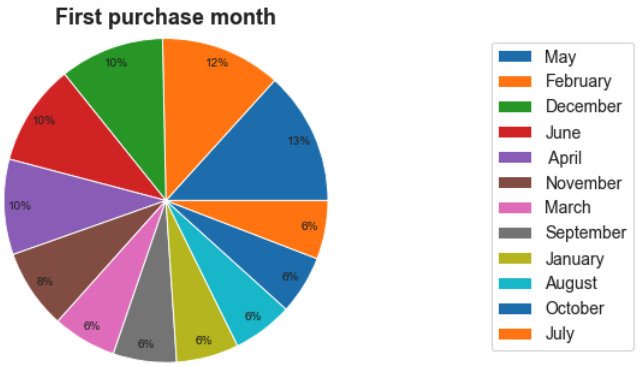

Figure 1.10: **The distribution of the month of the first purchase of a customer.**

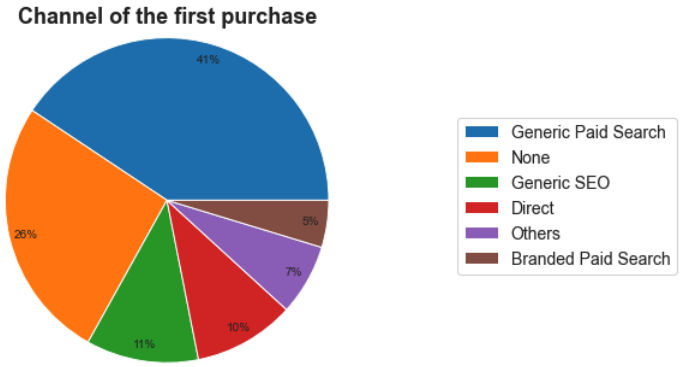

Figure 1.11: **The distribution of the first purchase channel.**

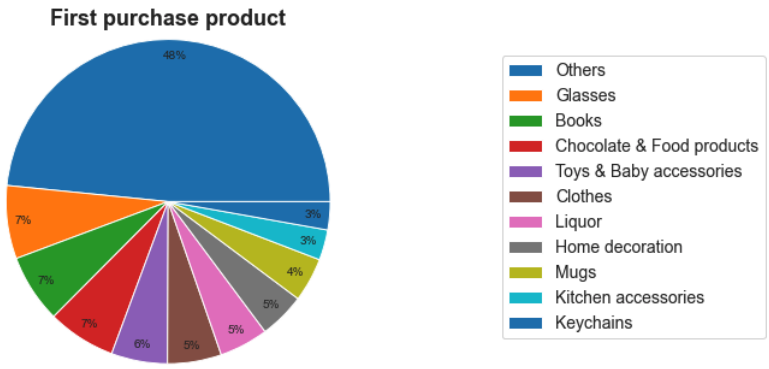

Figure 1.12: **The distribution of the first purchase product assortment.**

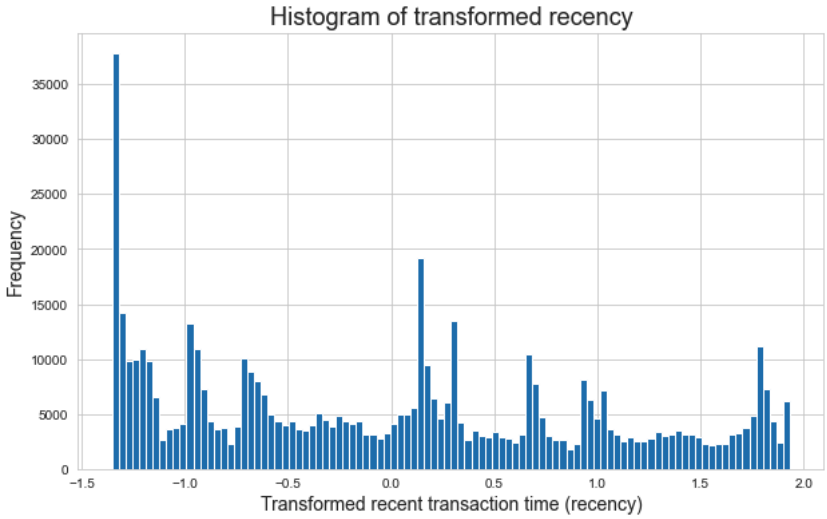

Figure 1.13: **The distribution of the transformed recency.**

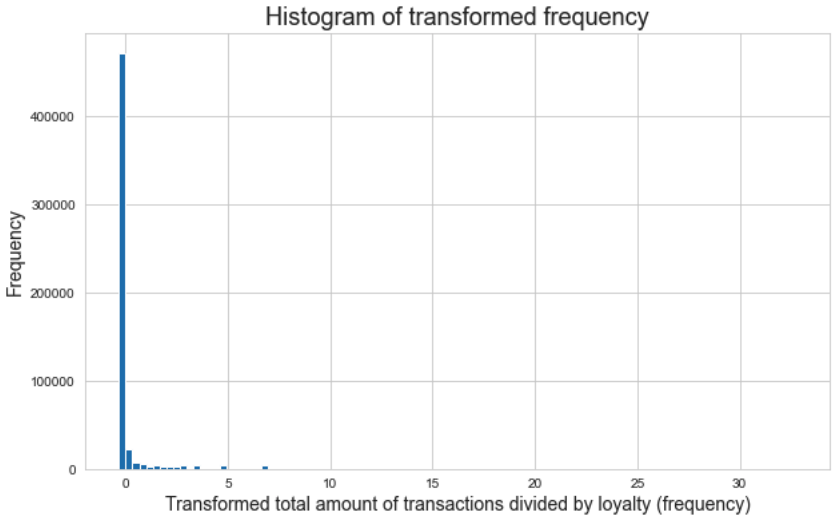

Figure 1.14: **The distribution of the transformed frequency**.

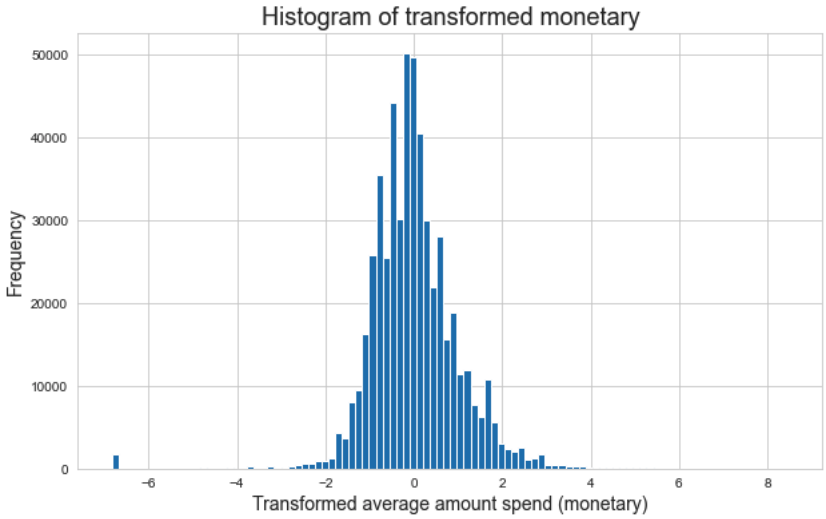

Figure 1.15: **The distribution of the transformed monetary value**.

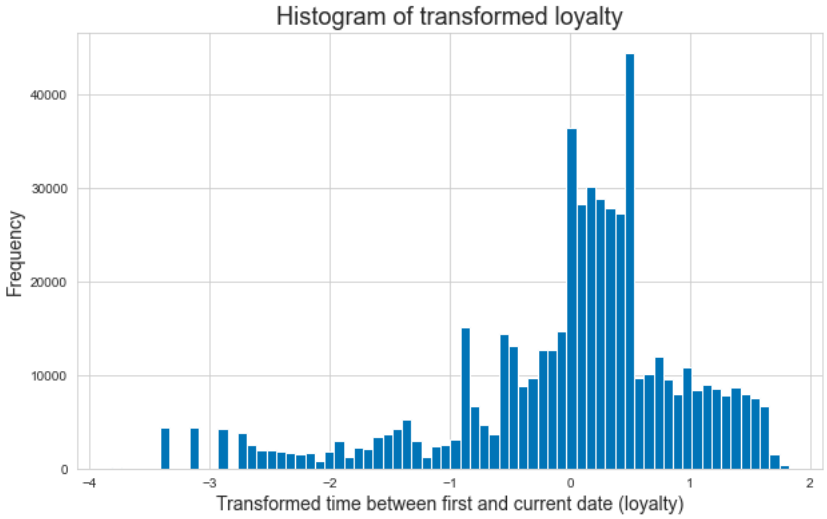

Figure 1.16: **The distribution of the transformed loyalty variable**.

2

RFM model

2.1 What is an RFM model

The RFM analytic model is proposed by Hughes (1994), it is a model that divides large data by three variables, namely, recency, frequency and monetary.

Each customer is given a score for the recency, frequency and monetary variable. After giving each customer a recency score, frequency score, and a monetary score, the customers can be segmented into clusters (*RFM segmentation*). A cluster is a collection of data objects that are similar to each other, and differ with objects from other clusters. The cluster with a high frequency score can be seen as a loyal customer for example. For each cluster a specific marketing strategy can be applied (Cheng and Chen, 2009).

2.2 K-means algorithm

The goal is to find subgroups of the data points such that the points in the subgroups are as similar as possible, this similarity is measured based on the euclidean distance [1] (*K-means clustering: algorithm, applications, evaluation methods, and drawbacks*). In this report the data points are customers, so the goal is to find subgroups with customers who are as similar as possible. This is done using the K-means clustering method. The K-means method is considered as one of the most used clustering algorithms due to its simplicity.
The K-means method finds k non-overlapping clusters. It makes the clusters such that the euclidean distance between a data point and the centroid of that cluster is as small as possible. The centroid of a cluster is the center of that cluster. The K-means method step by step:

[1] Length of a straight line between two points in a n-dimensional space

1. Define the data points as $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}\}$, each data point is a n-dimensional real vector. In total we have N data points.

2. Define a number of k clusters.

3. Select k centroids arbitrary, define the set of centroinds as $C = \{\boldsymbol{c_1}, \boldsymbol{c_2}, .., \boldsymbol{c_k}\}$.

4. Assign each data point to the cluster where the euclidean distance between the centroid of that cluster and the data point is the smallest.

$$\boldsymbol{x_i^j} = min\{\boldsymbol{x_i} - \boldsymbol{c_j}\}$$

   When data point $\boldsymbol{x_i}$ is assignt to centroid $\boldsymbol{c_j}$, we denote it as $\boldsymbol{x_i^j}$.

5. Recalculate the centroids, the centroids become the mean of the data points in that cluster.

$$\boldsymbol{c_j} = \frac{1}{M} \sum_{1}^{M} \boldsymbol{x_i}$$

   With M the number of data points in a cluster.

6. Keep doing steps four and five until there is no change in the centroids. Then the clustering is optimilized.

(*Understanding K-means clustering in machine learning*)

2.3 Finding the optimal number of clusters

2.3.1 The Elbow method

To find the optimal number of clusters the Elbow method can be used. The Elbow method is based on the mean squared error between a data point and its centroid. When the number of clusters increases the error will become smaller. But you do not want to much clusters because to much clusters is difficult to apply for practical purposes (Bholowalia and Kumar, 2014) (Syakur et al., 2018). The Elbow method step by step:

1. Compute the clusters, using the K-means method, for different number of clusters k. (In this report k goes from 2 too 8 [2]).

[2]We choose these numbers because for the RFM-model these are the most used number of clusters

2. For every value k compute the mean squared error from every data point ($\boldsymbol{x_i}$) to its centroid ($\boldsymbol{c_j}$).

$$MSE = \frac{1}{N}\sum_{1}^{N}(\boldsymbol{x_i} - \boldsymbol{c_j})^2$$

With N as the number of data points in a cluster.

3. Add the mean squared errors of the clusters together. This we call the total mean squared error.

4. Plot the number of clusters k against the total mean squared error.

5. Where the graph bends the most is the optimal number of clusters.

Because the data is very large it will take a lot of time to calculate the optimal number of clusters with the Elbow method with all of the data. Hence a random sample of 10000 data points is used. To see if the random sample is representative, the Elbow method is calculated 10 times with different random samples. In Figure 2.1 the average mean squared error of the 10 samples and the minimum and maximum values is plotted against the number of clusters. In this Figure we can see that the 10 samples give approximately the same Elbow graph. And thus the random sample represents the data well.

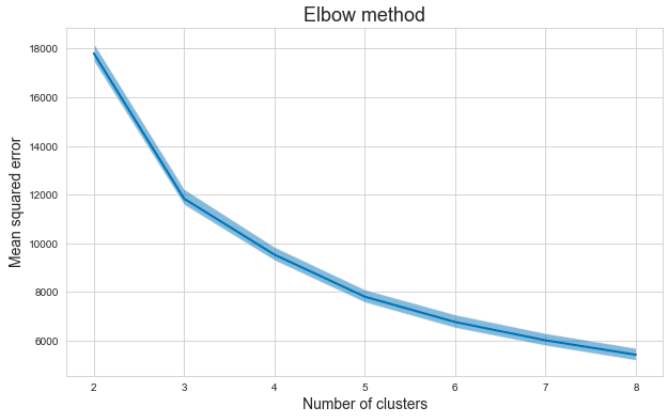

Figure 2.1: The mean, minimum and maximum of the mean squared error plotted against the number of clusters.

At three clusters a clear bent can be seen in the graph, hence the optimal number of clusters, calculated by the Elbow method, is three. Looking at the the number of clusters larger than three clusters the mean squared error becomes smaller.

2.3.2 The Silhouette Analyse

Another way of finding the optimal number of clusters is the Silhouette Analysis. This method is based on the separation distance between clusters. It measures the distance

between data points from one cluster to data points from another cluster. The Silhouette scores lie between the values -1 and 1. Here a value close to -1 indicates that the clustering of the date points is not optimal. A value close to 1 indicates that the clustering of the data points is optimal (Ogbuabor and Ugwoke, 2018). The Silhouette Analysis step by step:

1. Let X= $\{\boldsymbol{x_1}, \boldsymbol{x_2}, .., \boldsymbol{x_N}\}$ be data points and C=$\{c_1, c_2, .., c_k\}$ be the k clusters sets. Note that here C is not the centroids, but sets of data points in a cluster.

2. Compute for every data point $\boldsymbol{x_i^j}$ in a cluster set c_j the average distance between that data point and all the other data points in that cluster ($\boldsymbol{x_l} \in c_j$) and where the distance between to point is defined as $d(\boldsymbol{x_i}, \boldsymbol{x_l})$.

$$z_i^j = \frac{1}{M-1} \sum_{l=1, l \neq i}^{M} d(\boldsymbol{x_i^j}, \boldsymbol{x_l^j})$$

   With M the number of data points in cluster c_j.

3. Compute the distances between that data point $\boldsymbol{x_i^j}$ and all the other data points that are not in the same cluster ($\boldsymbol{x_l} \notin c_j$).

$$y_i^j = \min \frac{1}{K} \sum_{l=1}^{K} d(\boldsymbol{x_i^j}, \boldsymbol{x_l})$$

   With K the number of data point that are not in the cluster c_j.

4. The Silhouette score of data point $\boldsymbol{x_i^j}$ in cluster c_j can now be defined as:

$$s_i^j = \frac{y_i^j - z_i^j}{\max\{z_i^j, y_i^j\}}$$

   And with $s_i^j = 0$ when $y_i^j = z_i^j$.

5. The Silhouette score of cluster c_j is:

$$S_j = \frac{1}{M} \sum_{i=1}^{M} s_i^j$$

   with M the number of data point in cluster c_j.

6. The global Silhouette index is:

$$S = \frac{1}{k} \sum_{j=1}^{k} S_j$$

   where k is the number of clusters.

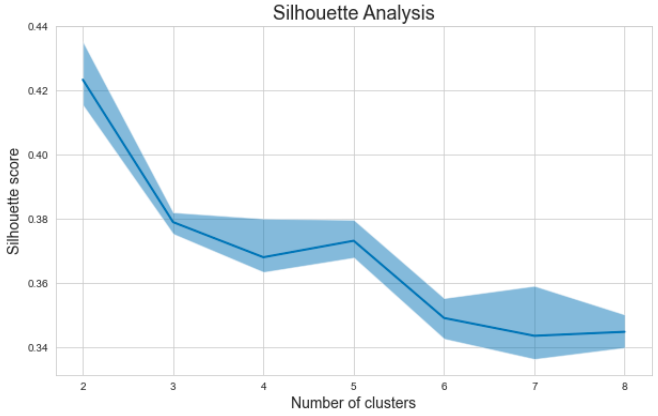

Figure 2.2: The mean, maximum and minimum Silhouette score plotted against the number of clusters.

Since the data is so large the Silhouette score is calculated with 10 random samples. In Figure 2.2 the average, minimum and maximum Silhouette score is plotted against the number of clusters.

The closer the Silhouette score is to 1 the better, the highest score is at two clusters. Hence the optimal number of clusters is two. And the second best score is at three clusters.

2.3.3 Davies-Bouldin Index

The Davies-Bouldin Index is based on almost the same measurements as the Silhouette Analysis. Namely the separation distance between clusters. The distance between other clusters must be as large as possible and the distance in a cluster as small as possible. The Davies-bouldin Index step by step:

1. Let X= $\{\boldsymbol{x_1}, \boldsymbol{x_2}, .., \boldsymbol{x_N}\}$ be data points, C=$\{c_1, c_2, .., c_k\}$ be the k clusters sets and let V=$\{v_1, v_2, .., v_k\}$ be the centroids of the clusters.

2. Let s_i be the average distance between a data point in a cluster and the centroid of that cluster.

3. Let d_{ij} be the distance between cluster centroids v_i and v_j.

4. Calculate
$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

5. The Davies-Bouldin Index is:
$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

with k the number of clusters.

(Petrovic, 2006), (*How to measure clustering performances when there are no ground truth*), (Xiao, Lu, and Li, 2017)

The Davies-Bouldin score is plotted against the number of clusters in Figure 2.3.

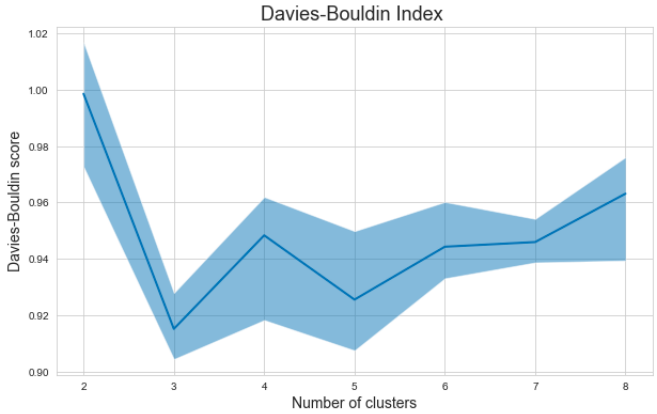

Figure 2.3: The mean, minimum and maximum Davies-Bouldin score plotted against the number of clusters.

The lower the Davies-Bouldin score the better, at three clusters the score is the lowest. Hence the optimal number of clusters is three. An other minimum in the graph can be at five clusters, this because the minimum Davies-Bouldin score is almost as low as the score at three clusters, but the maximum score at five clusters is much higher.

2.3.4 Results

The result of the Elbow method is shown in Figure 2.1. At three clusters there is a clear bend. Thus the optimal number of clusters from the Elbow method would be three clusters.
The result of the Silhouette Analyse in Figure 2.2 is two cluster because the graph is at its highest at two clusters. Since two clusters is a not a good number of clusters for the RFM-model [3] , we look at the second highest point in the graph, that is three clusters. Four and five clusters do also give a relatively good Silhouette score. After five clusters the Silhouette score drops.
The result of the Davies-Bouldin Index in Figure 2.3 is three cluster because at three clusters the graph is at its lowest. The second best number of clusters would be five clusters because there the graph has a minimum as well.

Looking at those three methods the optimal number of clusters is three. The second best number of clusters would be five looking at these methods. Hence first we cluster the RFM-model in three clusters en secondly in five clusters.

[3]because two clusters is not workable for marketing purposes

2.4 RFM-model clustered in three clusters

With the K-means method the data is clustered in three clusters, the clusters can be seen in Figure 2.4. In this figure we see the clusters numbered by zero, one and two. The columns show the variables recency, frequency and monetary. For every cluster the mean value of the variable is shown.
From this figure we can label the three clusters as follows:

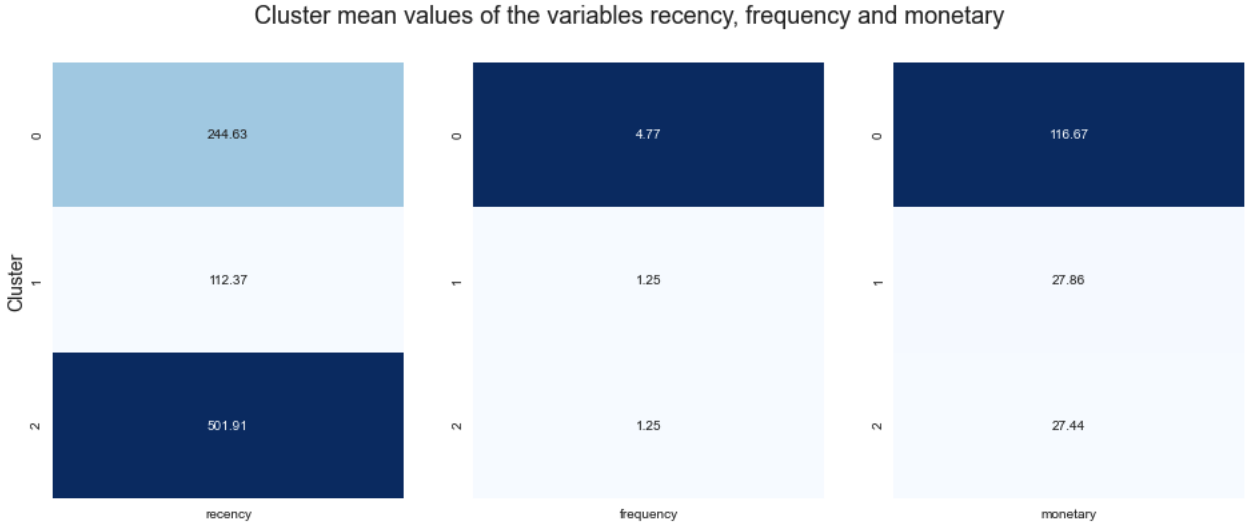

Figure 2.4: Customers divided in three clusters based on the variables recency, frequency and monetary value.

- Cluster 0: Has an average recency of 244.63 day's, an average frequency of 4.77 and an average monetary value of 116.67 euros. So the customers in this clusters have spend much and frequently made a purchase, we can classify the customers in this cluster as 'loyal customers'.
- Cluster 1: Has an average recency of 112.37 day's, an average frequency of 1.25 and an average monetary value of 27.86 euros. The customers in this cluster recently bought something, did not spend much and haven't made a lot of purchases. Customers in cluster 1 can be called the 'new customers'.
- Cluster 2: Has an average recency of 501.91 day's, an average frequency of 1.25 and an average monetary value of 27.44 euros. The customers in this cluster did not made much purchases, have not spend much and there last purchase was more then a year ago. We can say that customers in cluster 2 are 'lost costumers' who probably not come back.

Three clusters is for marketing purposes within Yoursurprise not the best number of clusters. They want to see how the groups are divided with more clusters and if they

can work with the kind of clusters that appear. Since five clusters is the second best number of cluster we will look at the RFM-model divided in five clusters.

2.5 RFM-model clustered in five clusters

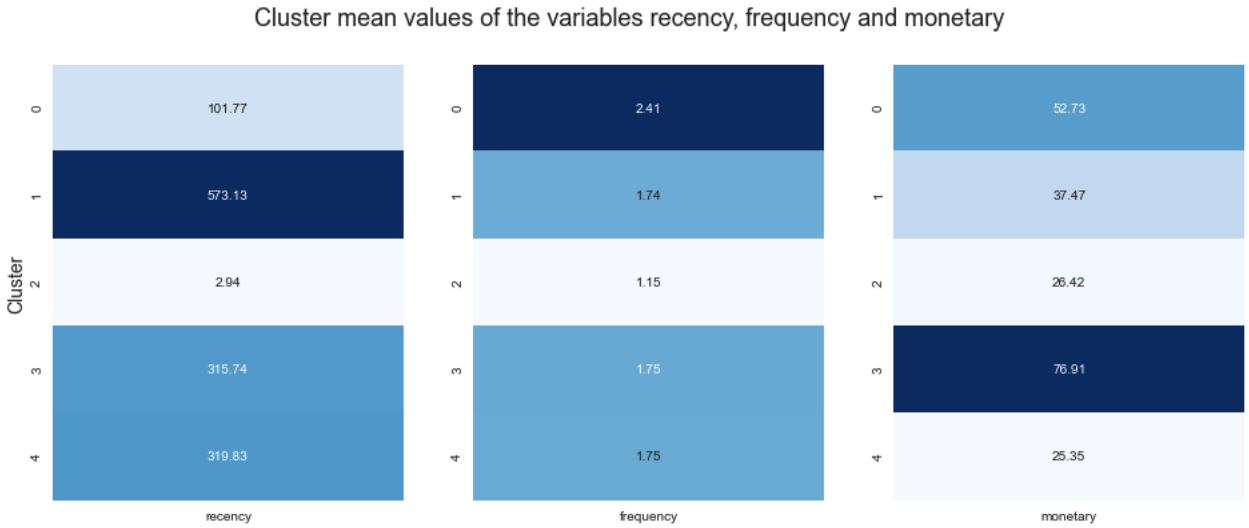

Figure 2.5: Customers divided in five clusters based on the variables recency, frequency and monetary value.

In Figure 2.5 the customers are clustered in five clusters. The clusters are numbered by zero, one, two, three and four. The columns are the variables recency, frequency and monetary. We can label the five clusters as follows:

- Cluster 0: Has an average recency of 101.77 day's, an average frequency of 2.41 and an average monetary value of 52.73 euros. We label this cluster as 'loyal customers'.
- Cluster 1: Has an average recency of 573.13 day's, an average frequency of 1.74 and an average monetary value of 37.47 euros. We label this cluster as 'lost customers'.
- Cluster 2: Has an average recency of 2.94 day's, an average frequency of 1.15 and an average monetary value of 26.42 euros. We can label this cluster as 'new customers'.
- Cluster 3: Has an average recency of 315.74 day's, an average frequency of 1.75 and an average monetary value of 76.91 euros. We can label this cluster as 'potential loyal customers'.

- Cluster 4: Has an average recency of 319.83 day's, an average frequency of 1.75 and an average monetary value of 1.75 euros. We can label this cluster as 'customers at risk'.

Those five clusters are more applicable for the company Yoursurprise. Because these are the groups that categorize types of behaviors that they would like to act upon.

The number of customers per cluster can be seen in Figure 2.6. Cluster zero, the 'loyal customers', is the largest cluster and cluster four, the 'customers at risk', the smallest.

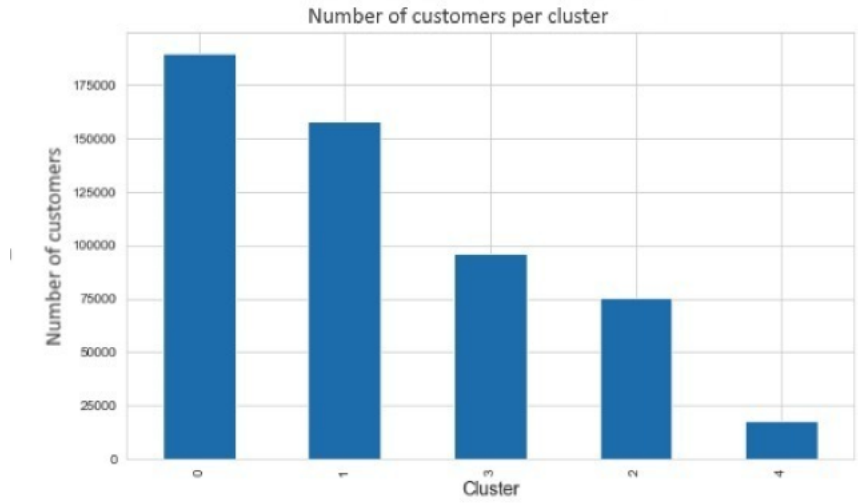

Figure 2.6: The number of customers per cluster for the RFM-model clustered in five clusters.

To get a better view on the clusters the distributions of the variables are plotted in boxplots. A boxplot shows how the data is distributed, every piece of the box is 25% of the data, the dots are the outliers. In Figure 2.7 the variable recency is plotted for the five clusters. The boxplot for the recency in cluster three is very spread out. The recency in cluster two has a squeezed distribution, with a few outliers. The recency in cluster four is spread out and has a lot of outliers.

In Figure 2.8 the distribution of the variable frequency is plotted per cluster without the outliers. In this figure we can see that the frequency in clusters one, three and four look similar and the distribution of the frequency in cluster two is squeezed en the distribution of the frequency in cluster zero is spread out. The medians look the same for every cluster. That those distributions are a so similar is not what we want. We want that the variables differ per cluster.

In Figure 2.9 the distribution of the variable monetary value is plotted per cluster without the outliers. Here we see that the monetary value in clusters zero and three has a spread out distribution and the monetary value in clusters one, two and three are less spread out. In the RFM-model cluster three has monetary value 76.91, in the boxplot

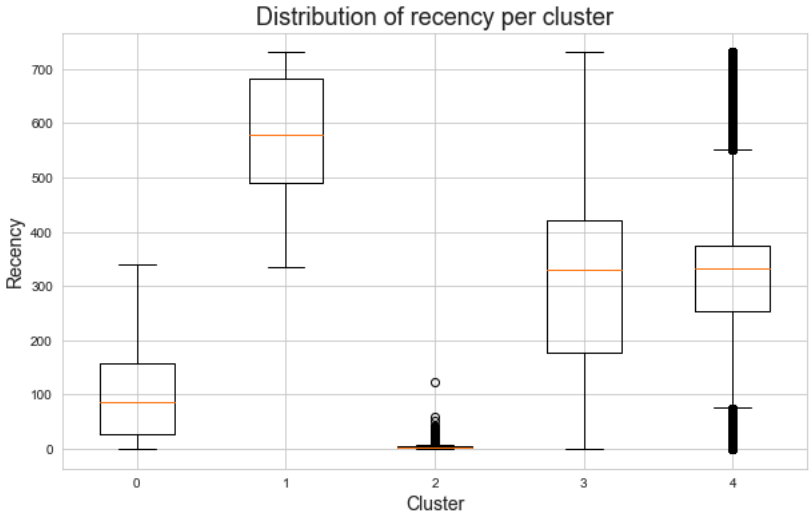

Figure 2.7: The distribution of the variable recency plotted per cluster in a boxplot.

can be seen that 75% of the customers in that clusters had spend on average less that 82 euros. Because of the skewness the mean is that high.

It seems that the frequency en monetary variable do have a lot of overlap in the RFM-model and that the clustering is mostly based on the recency variable.

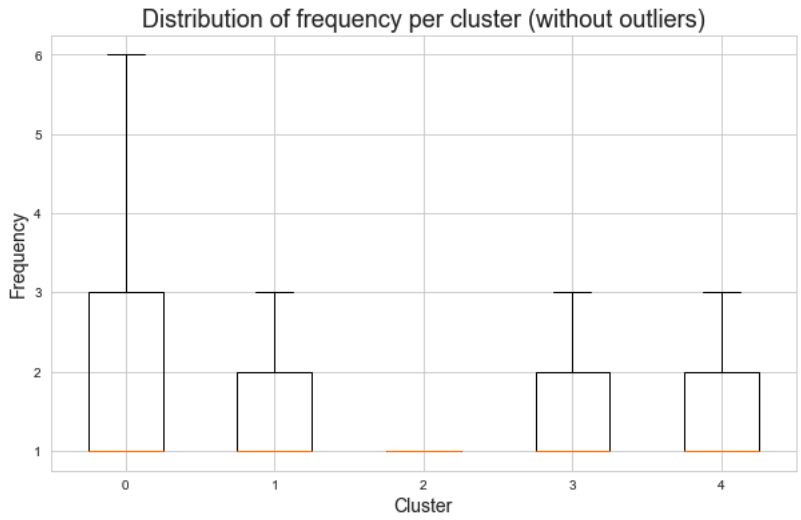

Figure 2.8: The distribution of the variable frequency plotted per cluster in a boxplot.

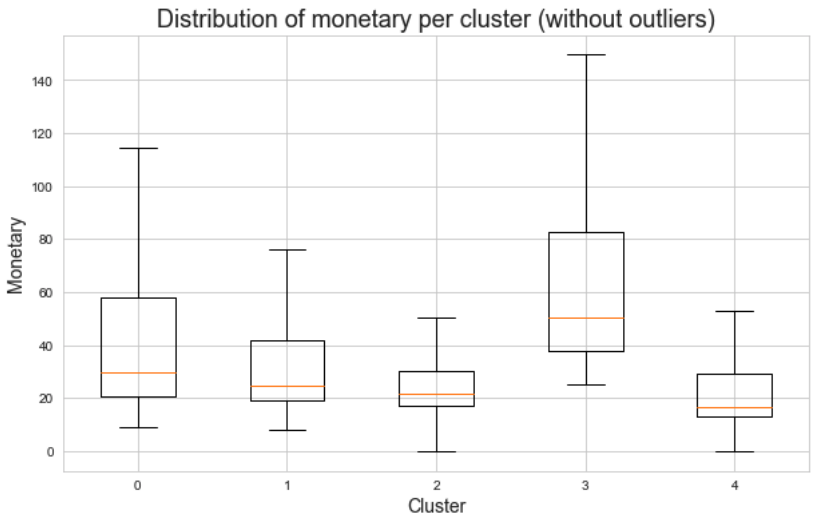

Figure 2.9: The distribution of the variable monetary value plotted per cluster in a boxplot.

3

Extension RFM model

In this Chapter the variable loyalty is added, this variable refers to the relationship between customer and company. And with this extra variable the customers are clustered again. First the optimal number of clusters is calculated with the following methods, Elbow method, Silhouette Analysis and Davies-Bouldin Index. Currently when a customer made a purchase once or twice a year and the last purchase is almost a year ago, the customer can be seen as a lost customer. But it is a customer for a long time. For the company this customers can also be seen as a loyal customers. From now the extension of the RFM-model will be called RFML-model (recency, frequnecy, monetary and loyalty) (Alizadeh-Zoeram and Karimi, 2018).

3.1 Finding the optimal number of clusters

For the extension of the RFM-model the K-means algorithm works the same. Now it is in a four dimensional space instead of a three dimensional space.

Finding the optimal number of clusters is done with the same methods as for the RFM-model.

The result of the Elbow method of the RFML-model in Figure 3.1 shows that the optimal number of clusters is three clusters.

The result of the Silhouette Analyse of the RFML-model can be seen in Figure 3.2. We see a clear drop in the Silhouette score from two to three clusters. From three to eight clusters the silhouette score stays low, hence from three clusters the number of clusters can be seen as optimal number of clusters. The optimal number of clusters would be five.

The result of the Davies-Bouldin Index of the RFML-model can be seen in Figure 3.3. The optimal number of clusters is two, because at two clusters the graph has an optimum. The second best number of clusters is three clusters.

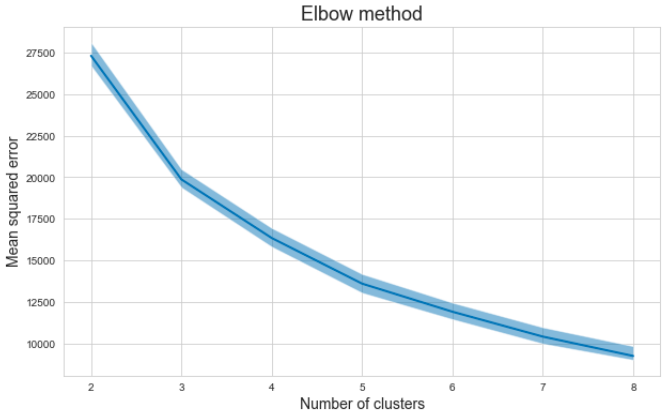

Figure 3.1: Elbow method for the RFML-model

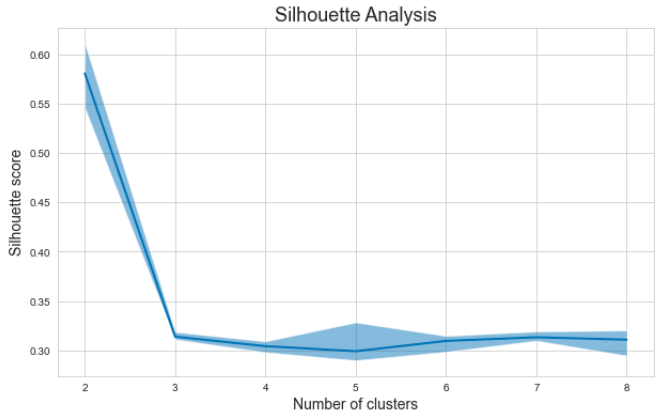

Figure 3.2: Silhouette Analysis for the RFML-model

3.1.1 Results

Looking at all the three methods the optimal number of clusters is three clusters. But since the company wants more than three clusters we are only going to look at the RFML-model clustered in five clusters. So that we can compare the RFM-model and RFML-model.

3.2 RFML-model clustered in five clusters

In Figure 3.4 the RFML-model is clustered in five clusters. The clustered are numered by zero, one, two, three and four. The columns are the variables recency, frequency and monetary.

- Cluster 0: Has a recency of 205.82 day's, a frequency of 2.42, a monetary value of 45.19 euros and a high loyalty of 810.45 day's. We label this cluster as 'loyal

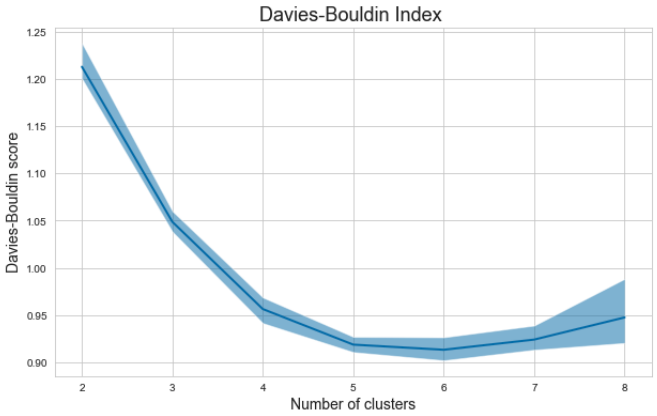

Figure 3.3: Davies-Bouldin Index for the RFML-model

customers'

- Cluster 1: Has the highest recency of 504.84 day's, a frequency of 1.81 , a monetary value of 36.32 and the highest loyalty of 996.45 day's. We label this cluster as 'lost customers'.
- Cluster 2: Has a recency of 48.54 day's, a frequency of 1.17, a monetary value of 26.27 euros and a loyalty of 53.56 day's. We label this cluster as 'active customers'.
- Cluster 3: Has a recency of 308.51 day's, a frequency of 2, a monetary value of 78.94 euros and a loyalty of 748.55 day's. We label this cluster as 'potential loyal customers'.
- Cluster 4: Has a low recency of 2.55 day's, a frequency of 1.14, a monetary value of 26.33 euros and a low loyalty of 2.76. We label this cluster as 'new customers'.

Cluster two and four are very similar, this is not what we like to see.
The number of customers per clusters can be seen in Figure 3.5. Here we see that cluster two, the 'active customer', is the largest group and that cluster three, the 'potential loyal customer', is the smallest group.

To see if the variable are distributed well per cluster, we look at the plotted boxplots.

In Figure 3.6 the distribution of the variable recency is plotted per cluster in boxplots. Clusters zero and three have a distribution of the recency variable that is spread out. The recency variable is squeezed in cluster four.

In Figure 3.7 the distribution of the variable frequency is plotted per cluster in boxplots. It is plotted without the outliers, because the outliers made the figure unclear. We see that the distribution of the frequency in clusters one and three look the same and the cluster two and four look the same.

In Figure 3.8 the distribution of the variable monetary value is plotted per cluster in boxplots. This is plotted without the outliers, because the outliers made the picture

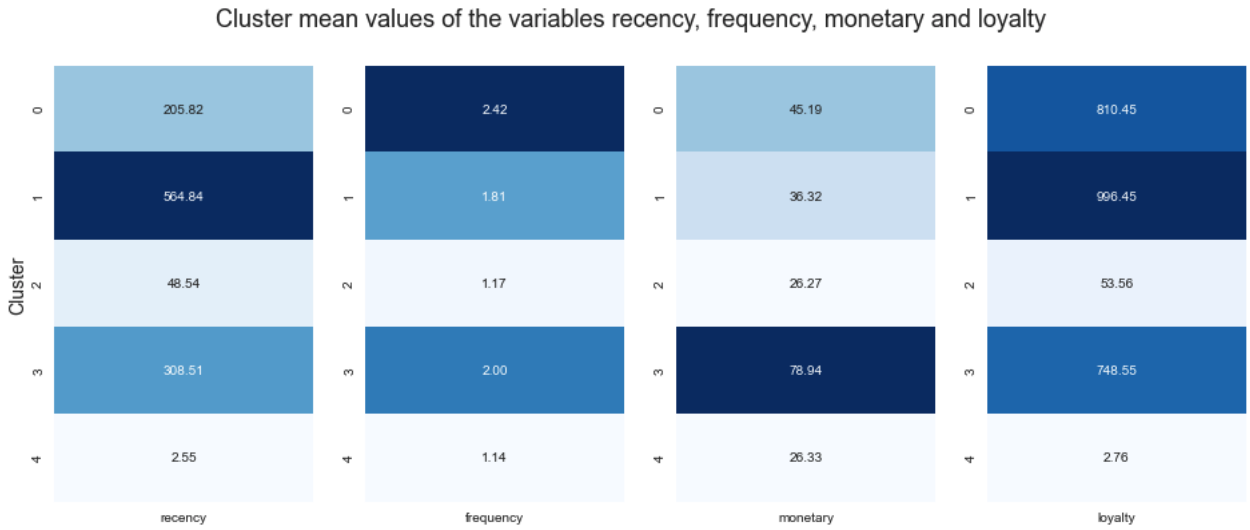

Figure 3.4: Customers divided in five clusters based on the variables recency, frequency, monetary value and loyalty.

unclear. Her we see that the distribution is spread out in cluster three. And that the distribution in clusters two and three are very similar.

In Figure 3.9 the distribution of the variable loyalty is plotted per cluster in box-plots. Here we see that the distribution in cluster zero is very spread out and that the distribution in the clusters two and four are squeezed.

The biggest difference between the RFM-model and the RFML-model is that they have are the same groups of customers. The RFML-model has a group labeled by 'active customer' and the RFM-model has not. The RFM-model has a groups labeled by 'customers at risk' and the RFML-model has not. The distributions of the variables look similar.

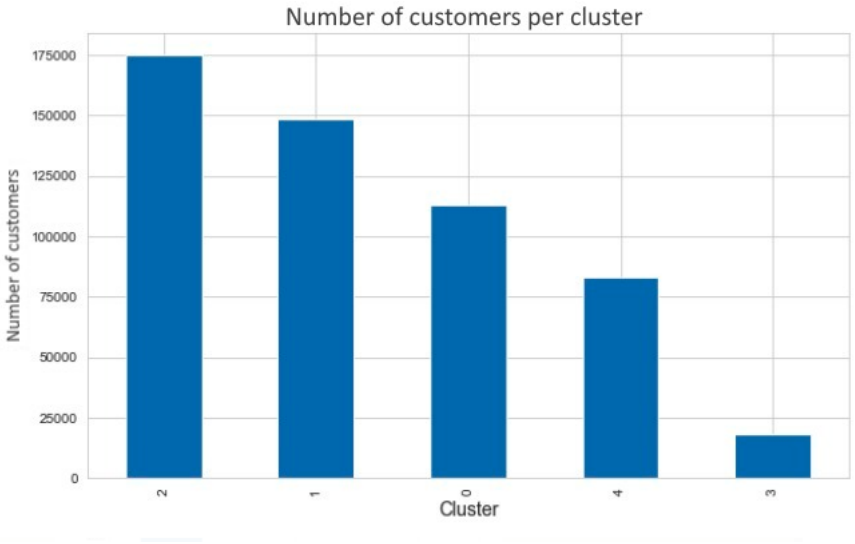

Figure 3.5: The number of customers per cluster for the RFML-model clustered in five clusters.

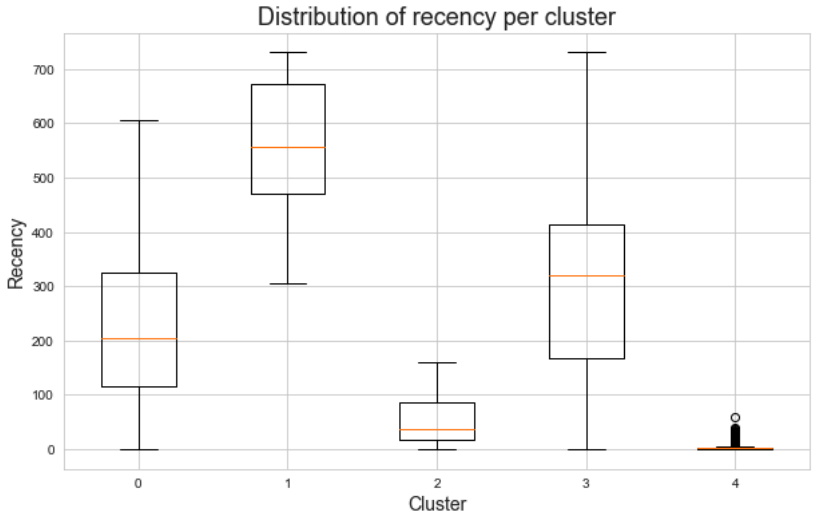

Figure 3.6: The distribution of the variable recency plotted per cluster in a boxplot.

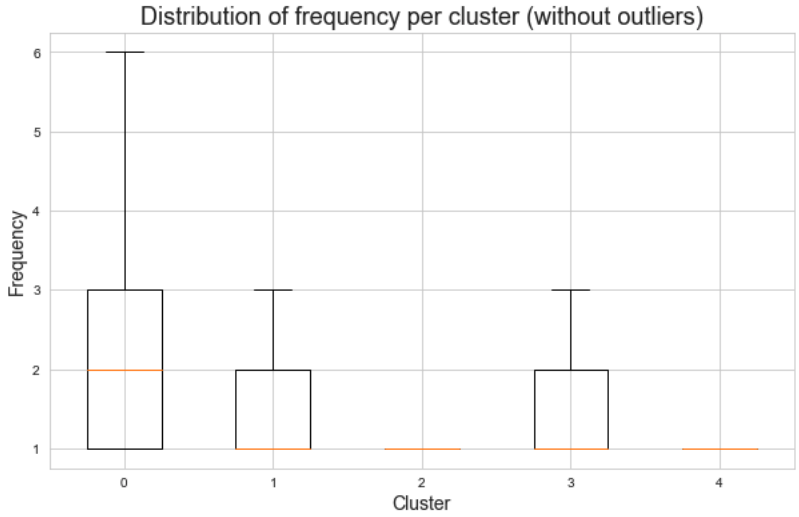

Figure 3.7: The distribution of the variable frequency plotted per cluster in a boxplot.

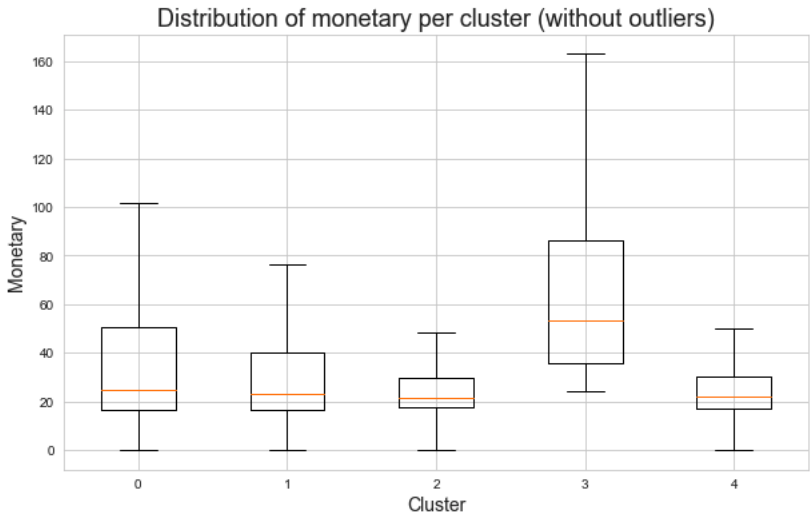

Figure 3.8: The distribution of the variable monetary value plotted per cluster in a boxplot.

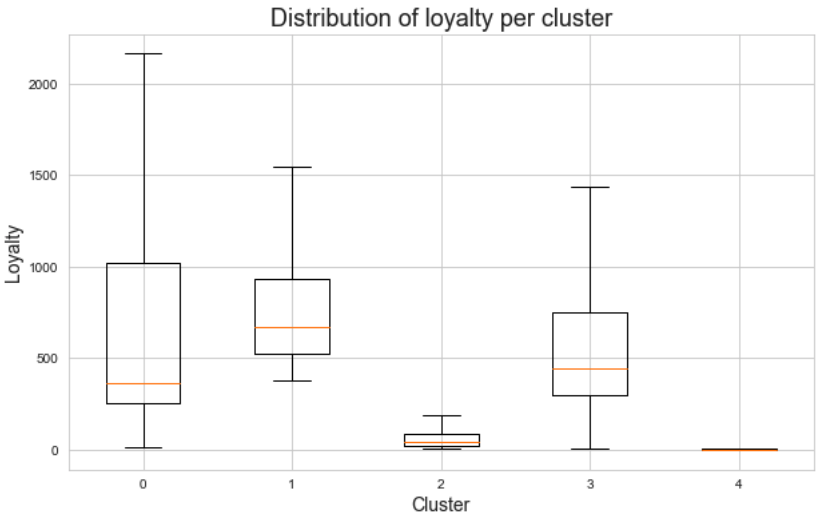

Figure 3.9: The distribution of the variable loyalty plotted per cluster in a boxplot.

4

Further subdividing a single group

Since the company does not value the variable loyalty and since the frequency is already depending on the loyalty, we decided to look further into the RFM-model and not look into the RFML-model. We are going to look further into the RFM-model with five clusters, because this is the model the company can work with the best. Cluster zero is the cluster with the most customers in it. The label of this cluster is 'loyal customers'. The company is really interested in this group because they do made the most purchases and they value loyal customers. To find out more about this cluster, this cluster is being further clustered.

4.1 Finding the optimal number of clusters in the 'loyal customers' group

To find the optimal number of clusters we use the methods Elbow method, Silhouette Analysis and Davies-Bouldin Index.

From the Elbow method we can see in Figure 4.1 that three is the optimal number of clusters. The Silhouette Analysis tells us that also three clusters is the optimal number, this can be seen in Figure 4.2. In Figure 4.3 we can see that the Davies-Bouldin score drops at three clusters, from there the score stays low. The minimum score is at five clusters. But at three clusters it is also low. That is why, when we look at all the methods, we can say that three clusters is the optimal number of clusters.

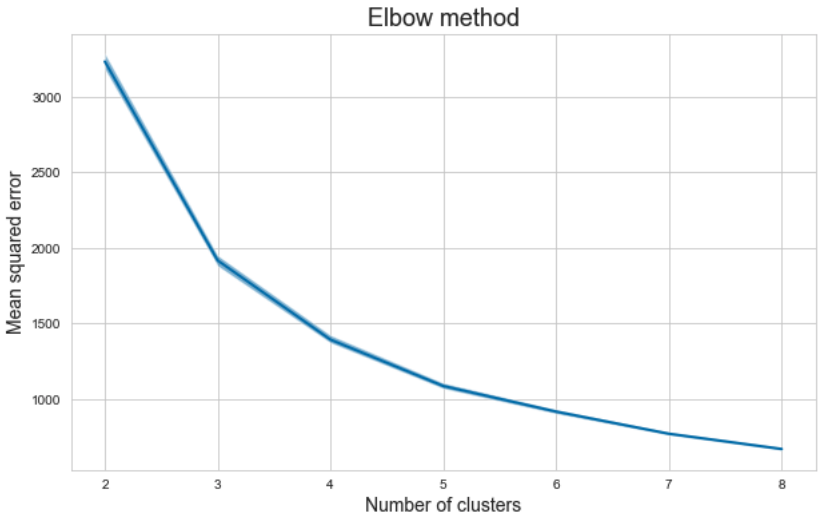

Figure 4.1: Elbow method for finding the optimal number of clusters in the 'loyal customers' group.

4.2 'Loyal customers' clustered in three clusters

With the K-means clustering method the 'loyal customers' from the RFM-model is being further clustered. The clustering of the cluster is shown in Figure 4.4. Here the clusters are labeled by zero, one and two. The columns are the variables recency, frequency and monetary value. The clusters are different from each other. This is because in the RFM-model the distributions of the variables are spread out. Here we can see that cluster zero is actually belongs more to the cluster 'new customer'. And now we can make a distinction between 'loyal customers who spend much' and 'loyal customers'. To get an better view on the distributions, the distribution of the variables are plotted per cluster.

In Figure 4.5 the number of customers per clusters is plotted. Here we see that cluster one is the biggest, this are the loyal customers who spend much. Cluster zero is the smallest this is the groups that not really belongs to the loyal customers.

In Figure 4.6 the distribution of the variable recency is plotted per cluster. Here we see that indeed cluster zero has a different distribution that the distribution in clusters one and two.

In Figure 4.7 the distribution of the variable frequency is plotted per cluster, without the outliers. Here again we see that cluster zero has a different distribution.

In Figure 4.8 the distribution of the variable monetary value is plotted per cluster. Again cluster zero has, compared to the others clusters, a different distribution. But also the distribution of clusters one and two are not as similar as the distributions of variables recency and frequency.

Overall we can see that the clusters one and to are look a like expect for the variable monetary value. It looks like cluster zero did not belong here.

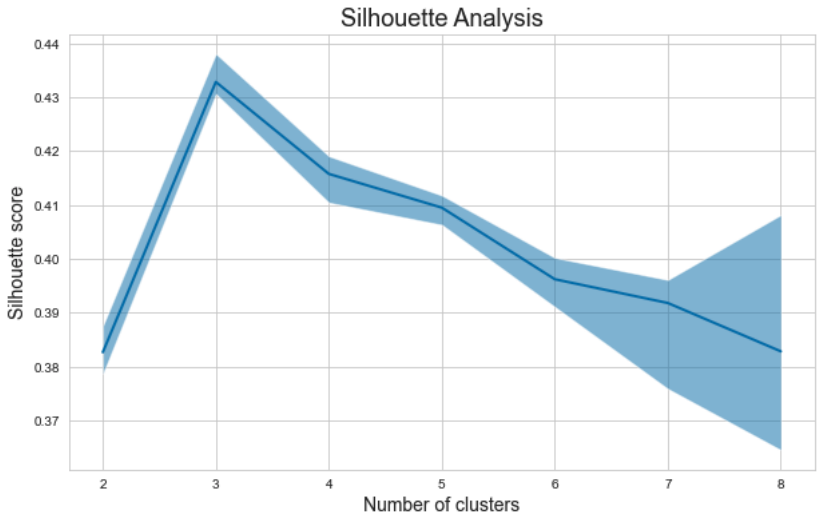

Figure 4.2: Silhouette Analysis for finding the optimal number of clusters in the 'loyal customers' group.

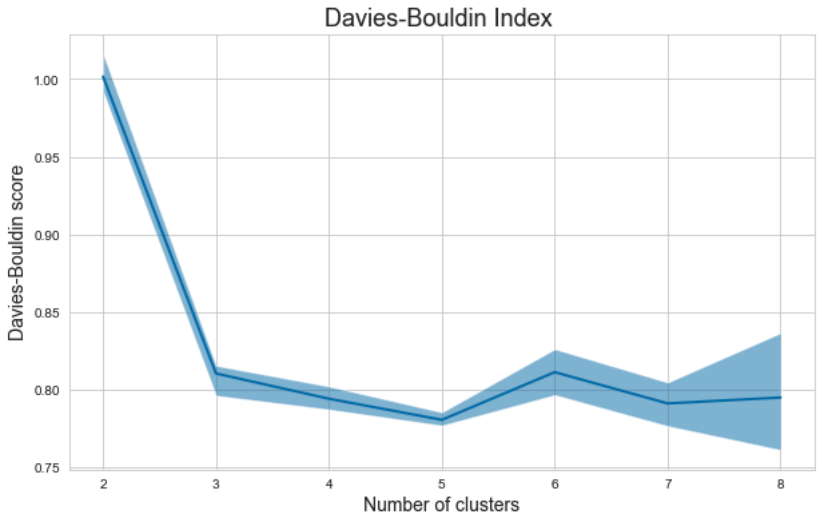

Figure 4.3: Davies-Bouldin Index for finding the optimal number of clusters in the 'loyal customers' group.

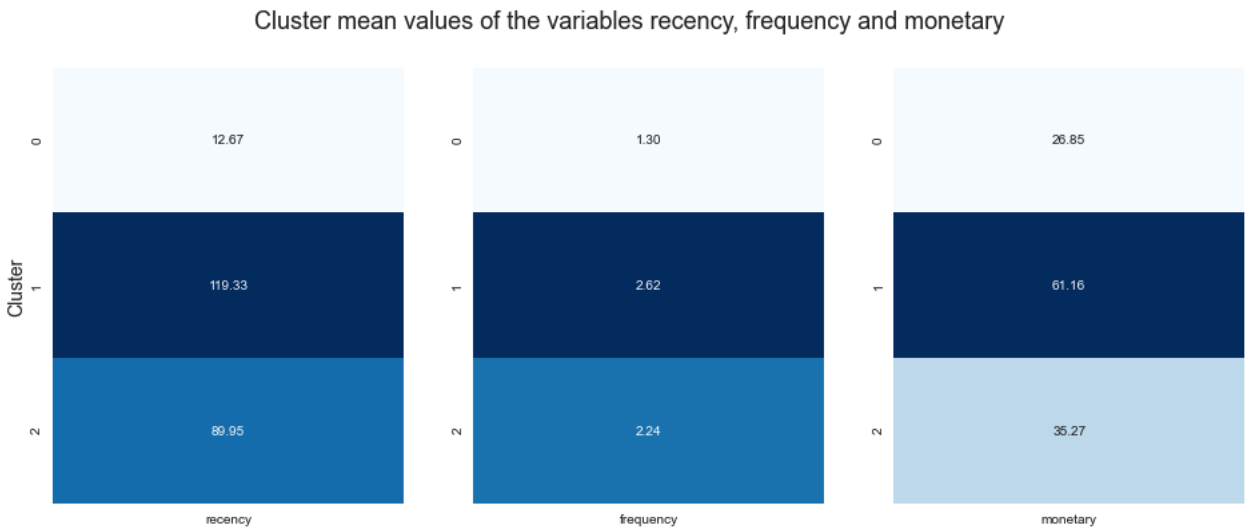

Figure 4.4: Customers divided in three clusters based on the variables recency, frequency and monetary.

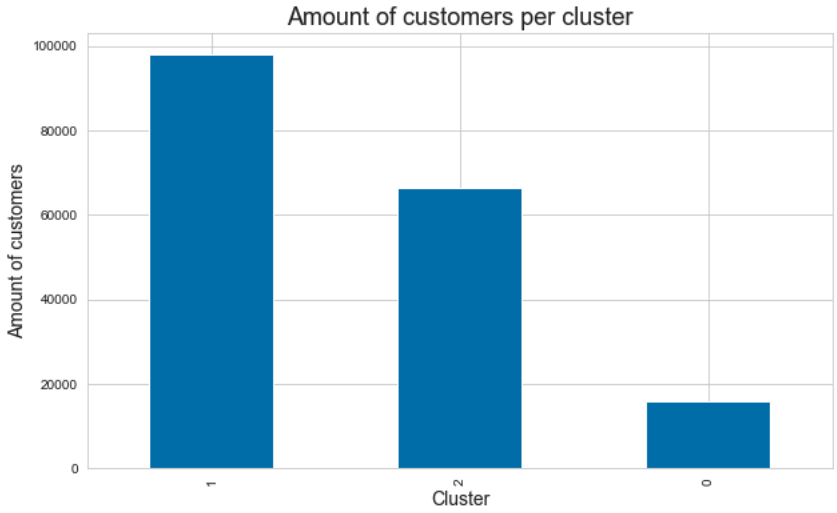

Figure 4.5: The number of customers per cluster for the further clustering of the 'loyal customers' group.

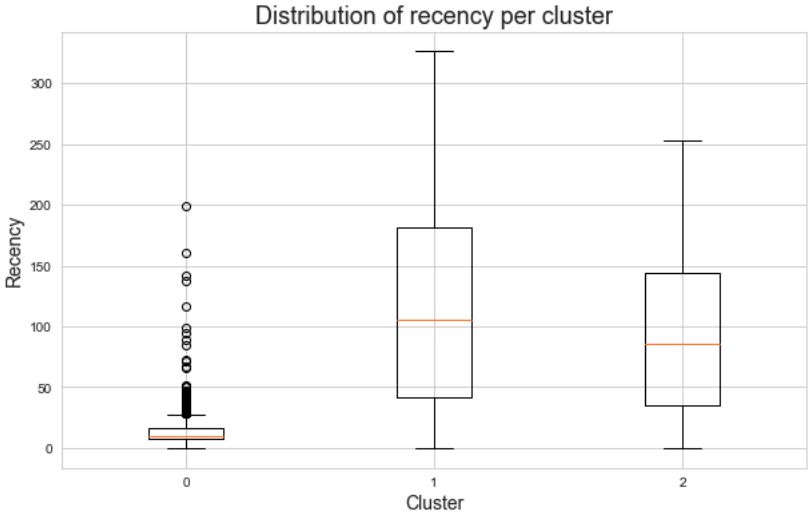

Figure 4.6: The distribution of the variable recency plotted per cluster in a boxplot.

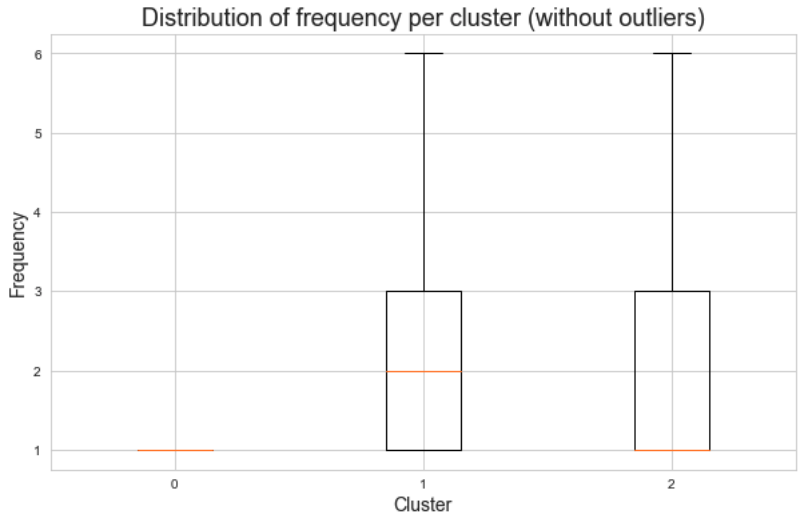

Figure 4.7: The distribution of the variable frequency plotted per cluster in a boxplot.

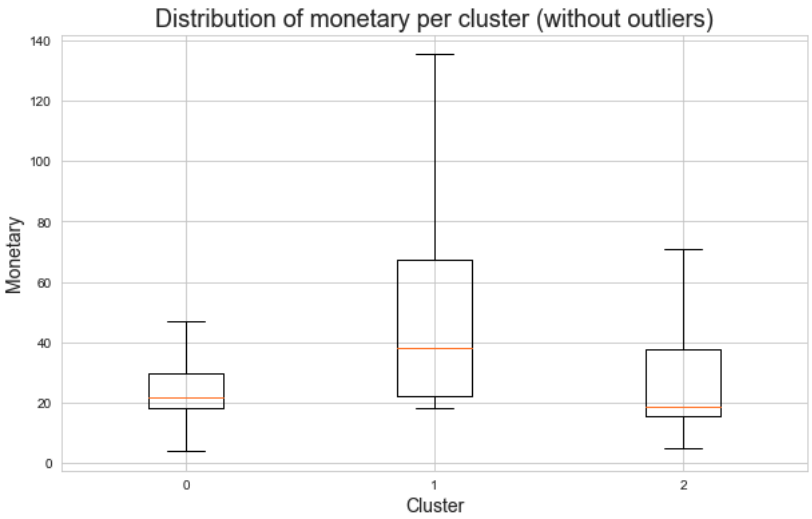

Figure 4.8: The distribution of the variable monetary value plotted per cluster in a boxplot.

5

Conclusion and Next steps

5.1 Conclusion

The aim of this research is to analyse the segmentation of the customers.
In Chapter 1 the data that is being analysed. Here we saw that the variables frequency and monetary value where very skewed. To fix this we log-transformed those variables. And after that standardized the data.
In Chapter 2 the RFM-model is created using the K-means clustering method. The conclusion was that the RFM-model divided into in five clusters was the best model. The resulting groups are 'loyal customers', 'lost customers', 'potential loyal customers', 'new customers' and 'customers at risk'. For the company these are the customer segments they deem applicable for marketing purposes. After comparing the distributions of the variables per cluster we saw that the variable recency did not have a lot of overlap. The variables frequency and monetary value did have a lot of overlapping values.
In Chapter 3 the RFM-model was extended. The variable loyalty was added. The five groups of customers that came out of this model where 'loyal customers', 'lost customers', 'active customers', 'potential loyal customer' and 'new customer'. In this RFML-model the clusters two and four are very similar.
In Chapter 4 we compared the RFM and RFML-model. Since the company did not see the added value of the loyalty variable, the RFML-model has not the customer groups they would like to see, and the cluster two and four from the RFML-model look similar. We concluded that the RFM-model is a better model. Cluster zero from the RFM-model are the 'loyal customers', this is an interesting groups for the company and is also the biggest group. We split up this group further in order to discover subcategories of customers, we found that the optimal number of cluster was three. after clustering we found that a small group, cluster zero, does not seems to fit in the group 'loyal customers' but rather in the group 'new customer'.

5.2 Next steps

To goal after this research is to implement this model in the marketing strategies of the company. The clusters of the RFM-model will be used especially for email-marketing. They make different emails per cluster. After making such a marketing strategy they will test it to check if it works.

To do so first there has to be further research in the other clusters of the RFM-model. Now we only looked further into the group of the 'loyal customers'. Here we observed that a part of this clusters did not seem to belong in this cluster. Maybe in the other clusters there are customers that are also not in the right cluster.

Since the company is growing it's customers rapidly they get a lot of new customers every year. So the distribution of the variables will change, the RFM-model is made with the distribution of customers at a certain point. But you might eed to change it when the distribution shifts. Not only the new customers make the distribution change, the distribution is different each month. Especially at the peak moments, like Mothers-day and Christmas, the distribution is different. Right after Christmas the recency of most customers will be low. So the RFM-model will give a big group of customers with a low recency. But when creating clusters using the RFM-model a month before Christmas the recency of most customers will be high. The cluster of the RFM-model have to be redetermined more times a year and then look to the differences.

Another step that can be taken is to add another variable that is important to the company. We tried the variable loyalty, but that did not provide the insights we wanted. Another variable that can be added to the RFM-model is a variable that looks if the customers is a Holiday shopper, one that makes their purchases during specific events or with a certain interval.

Bibliography

[1] Ali Alizadeh-Zoeram and Ahmadreza Karimi. "A New Approach for Customer Clustering by Integrating the LRFM Model and Fuzzy Inference System". In: *Iranian Journal of Management Studies* 11 (Jan. 2018), pp. 351–378. DOI: 10.22059/ijms.2018.242528.672839.

[2] Purnima Bholowalia and Ashok Kumar. "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN". In: *International Journal of Computer Applications* 105 (2014), pp. 17–24.

[3] Ching-Hsue Cheng and You-Shyang Chen. "Classifying the segmentation of customer value via RFM model and RS theory". In: *Expert Systems with Applications* 36 (Apr. 2009), pp. 4176–4184. DOI: 10.1016/j.eswa.2008.04.003.

[4] I. Dabburo. *K-means clustering: algorithm, applications, evaluation methods, and drawbacks*. URL: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. (accessed: Sept 17,2018).

[5] Changyong Feng et al. "Log-transformation and its implications for data analysis". In: *Shanghai archives of psychiatry* 26 (Apr. 2014), pp. 105–9. DOI: 10.3969/j.issn.1002-0829.2014.02.009.

[6] Dr. Micheal J. Garbade. *Understanding K-means clustering in machine learning*. URL: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. (accessed: Sept 17,2018).

[7] Ismail Mohamad and Dauda Usman. "Standardization and Its Effects on K-Means Clustering Algorithm". In: *Research Journal of Applied Sciences, Engineering and Technology* 6 (Sept. 2013), pp. 3299–3303. DOI: 10.19026/rjaset.6.3638.

[8] Godwin Ogbuabor and Ugwoke. "HEALTHCARE DATASET USING SILHOUETTE SCORE VALUE". In: 2018.

[9] Slobodan Petrovic. "A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters". In: 2006.

[10] *RFM segmentation*. URL: https://www.optimove.com/resources/learning-center/rfm-segmentation. (accessed: May 2020).

[11] M Syakur et al. “Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster”. In: *IOP Conference Series: Materials Science and Engineering* 336 (Apr. 2018), p. 012017. DOI: 10.1088/1757-899X/336/1/012017.

[12] H. Wei. *How to measure clustering performances when there are no ground truth.* URL: https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c. (accessed: Jan 2).

[13] Junwei Xiao, Jianfeng Lu, and Xiangyu Li. “Davies Bouldin Index based hierarchical initialization K-means”. In: *Intelligent Data Analysis* 21 (Nov. 2017), pp. 1327–1338. DOI: 10.3233/IDA-163129.