Humboldt-Universität zu Berlin

School of Business and Economics
Ladislaus von Bortkiewicz Chair of Statistics

Data Analysis II
Seminar paper

Comparison of clustering methods on an RFM model

Silvia Ventoruzzo
(592252)

submitted to

Sigbert Klinke

March 27, 2019

Contents

1 Introduction

Understanding its customers is a prime objective for companies. This information aids in both business and marketing decisions. In particular, when it comes to marketing, it is important to be able to distinguish among different classes of customers, in order to better target them. This is where the Recency, Frequency, Monetary (RFM) model and cluster analysis come to help (Birant, 2011).

The RFM model is a frequently used technique to select customers for direct mailings (Bult and Wansbeek, 1995), which describes customers' behavior by three variables relating to their purchase behavior in a specific time frame. The RFM attributes can already be used to segment customers, but a better result can be accomplished using more advanced clustering techniques. Modified versions of this basic model have been developed, like RFM-I (Tkachenko, 2015) and RFMTC (Yeh et al., 2009), but only the main form will be applied in this work.

This study was realized with a similar dataset to the one employed by Chen et al. (2012), differing of one month in the time frame. Therefore, the purpose of this analysis was to challenge the results from Chen et al. (2012) by using two other clustering methods, Agglomerative Hierarchical Clustering and DBSCAN, each of which has its own strengths and weaknesses. Both depend upon some parameters that need to be selected before running the function, but, in contrast with k-Means, they do not require the user to choose the number of clusters in advance (Tan, 2018).

The three RFM variables present some outliers in this dataset, therefore it was believed that DBSCAN would outperform the other methods. However, this technique provided similar results as Agglomerative Hierarchical Clustering, which is the one that achieved a better clustering according to internal validation processes. However, it was k-Means that produced the most clusters, but the customers in each did not share specific features. It is therefore believed that no natural clusters are present in this dataset. In fact, the dataset does not present a good cluster tendency.

The paper is constructed as follows. In section 2 the dataset and the RFM model are described and exploratory data analysis is presented. Further, section 3 introduces the the theory behind the tested clustering methods and explains the cluster validation techniques. Thereafter, section 4 reports how these processes were applied in this study and how outcomes were compared. Finally, section 5 discusses the results from this work and mentions suggestions for future analysis.

2 Data

For this study transactional data from a UK-based online retailer without phisical shops was used. The dataset includes all transactions happened from 01.12.2010 to 09.12.2011. The dataset was provided by Chen et al. (2012) on UCI Machine Learning Repository and the variables contained can be seen in table A.1.

To enable a comparison with Chen et al. (2012), the dataset was restricted to the transactions until 30.11.2011 to have exactly one year worth of transactions.

2.1 RFM Model

RFM (Recency, Frequency, Monetary) is, beside CLV (Customer Lifetime Value), one of the most used approaches to Customer Relationship Management (CRM). Bult and Wansbeek (1995) defines the meaning of the three words as follows:

- Recency: the time since the last order
- Frequency: number of orders in a certain time period
- Monetary: money spent during a certain time period

Customers can then be assigned a value for each RFM variable in a scala which divides the range of the variables' distributions into percentiles, which can have multiple lengths: Birant (2011) uses one from 1 to 5, while Sarvari et al. (2016) one from 1 to 10. As Khajvand et al. (2011) explained, a low value in Recency suggests a higher probability that the customer will buy again, thus earning a higher value in the scala. For Frequency, the opposite is true: higher value, higher probability, hence higher values in the scala. In the case of Monetary a high value means that the customer is profitable for the company, therefore it will receive high values in the scala. The combination of the variables' scores in a 3-digit number can be used to divide the customers into segments for further analysis (Wei et al., 2010). In this case customers' RFM variables were divided into 5 quantiles and the segments are constructed using combinations of the RFM scores, as displayed in table A.2.

This information will be used in subsection 4.4, together with the internal validation index explained in subsection 3.4 to compare and evaluate the different clustering methods.

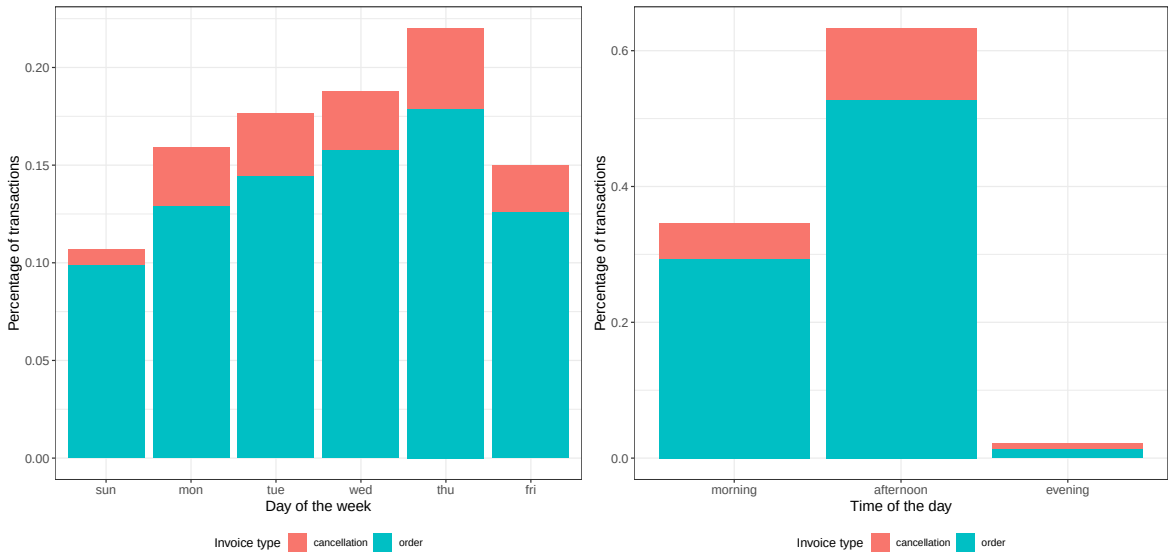
2.2 Exploratory Data Analysis

Transactions consists of both orders and cancellations. The latter are identified by the presence of a "C" at the beginning of the invoice number.

variable	min	1Q	median	3Q	max	iqr	mean	sd
Distinct products	1.00	7.00	15.00	28.00	542.00	21.00	21.43	24.64
Total (£)	0.00	158.24	303.15	473.03	77183.60	314.79	472.64	1148.39

Table 2.1: Descriptive statistics about orders

The orders' total amount and distinct products, whose descriptive statistics are displayed in table 2.1, and the distribution of the transactions across the week and the day, shown in figure 2.1, support the belief expressed by Chen et al. (2012) that this company's customers are mainly businesses.



(a) Across the week **(b)** Across the day

Figure 2.1: Distribution of the transactions across time

Moreover, outlier detection is an important step in data cleaning. However, since the data represents real transactions, these outliers may also have relevant business information. Nevertheless, they might spoil the results for some clustering methods, like it is the case with k-Means, as explained in subsection 3.1, while others are more adequate in the presence of noise, for example DBSCAN, which will be described in subsection 3.3.

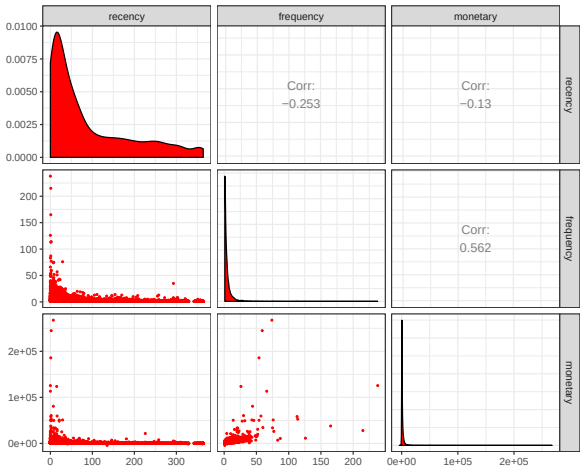


Figure 2.2: Scatterplots of RFM variables

To reveal outliers one can firstly look at the scatterplots of the involved variables, shown in figure 2.2, but, in the multivariate case, one has to take into consideration both the distance to the centroid and the shape of the data, which is represented by the covariance matrix (Filzmoser et al., 2005). Hence, to uncover outliers in the multivariate data, one can use the Mahalanobis distance, which is calculated as follows (Rousseeuw and Van Zomeren, 1990):

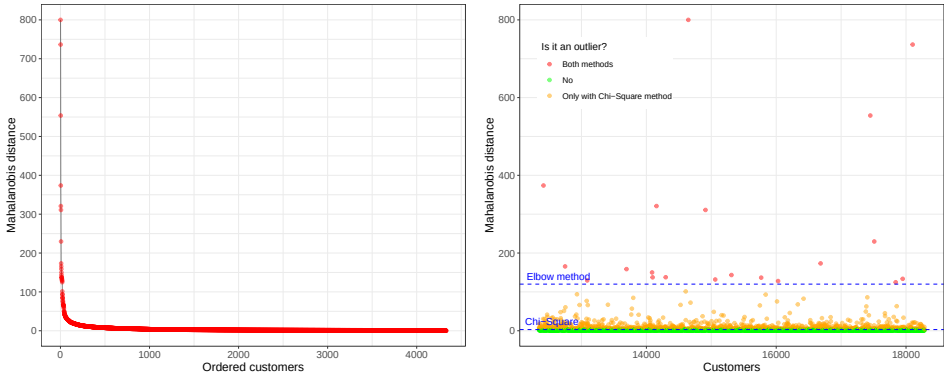
$$MD_i = \sqrt{(x_i - T(X))^T C(X)^{-1}(x_i - T(X))} \tag{1}$$

where $T(X)$, from now on μ, is the mean and $C(X)$, from this point Σ, is the covariance matrix.

One needs to apply robust estimators of μ and Σ, otherwise they will be affected by outliers (Rousseeuw and Van Zomeren, 1990). These will be in fact estimated using the MVE (minimum volume ellipsoid); for details one can read the appendix of Rousseeuw and Van Zomeren (1990).

Outliers are then defined as those points with values of the Mahalanobis distance higher than a threshold. This cutoff value can be set at $\chi^2_{d,0.975}$, i.e. the squared 97.5%-percentile of the Chi-Square distribution with degrees of freedom equal to the number of variables d.

In the dataset used for this study circa 28% of the points are considered outliers according to this methodology, which would however restrict the dataset too much. Therefore an elbow method will be used to remove outliers. The customers ordered by decreasing Mahalanobis Distance will be plotted against their distance, as in figure 2.3a, and one chooses the cutoff point where the distance stops decreasing rapidly, much like in the Elbow Method which will be explained in subsection 3.1. In this way only less than 1% will be removed.



(a) Customers ordered according to distance **(b)** View of thresholds and outliers

Figure 2.3: Plots of Mahalanobis Distance

It is also interesting to look at the distribution of the different variables in the RFM model, whose descriptive statistics are presented in table A.3. From figure 2.2 it is clear that the variables' distributions are skewed and that, in particular Frequency and Monetary, present many outliers. This might be an issue when clustering the dataset, since some methods are sensible to the presence of noise, but this topic will be discussed more in detail in section 3.

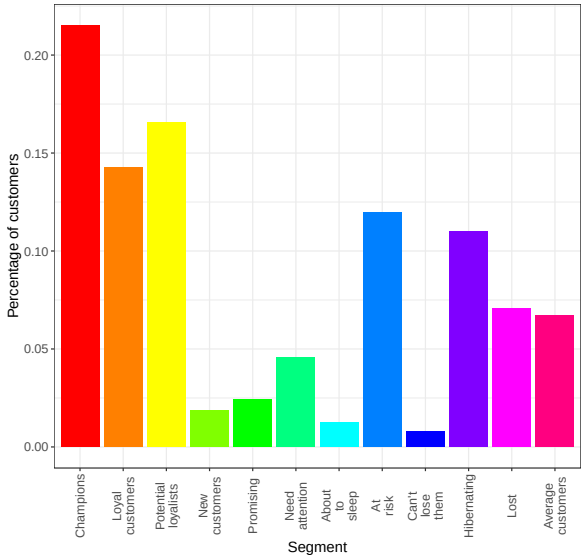


Figure 2.4: Bar plot of RFM Segments

According to the scatterplots in figure 2.2 the points looks quite concentrated with a few distant points. However, by looking at the distribution of the RFM segments, shown in figure 2.4, it seems that customers are quite disperse across the RFM segments. Therefore, it might be interesting to take a deeper interest in clustering.

3 Clustering methods

Clustering is an unsupervised learning procedure, since it concerns itself with finding groups in data that is not categorized (Madhulatha, 2012). For this reason is one useful step in exploratory data analysis (Jain et al., 1999).

During the years many clustering several algorithms have been developed, from easier ones like k-Means to more complicated ones like OPTICS. Nonetheless, they all share the steps listed by Jain et al. (1999):

1. Represent data patterns and, optionally, select variables
2. Choose the distance measure suitable to the data
3. Cluster the data
4. Extract cluster-representing data (for example, prototypes or centroid), if required
5. Assess the output, if required

However, the numerous techniques diverge with respect to other factors. In particular, the differences between the methods used in this study, are summarized in 3.1.

	k-means	Hierarchical Clustering	DBSCAN
Hierarchical vs. Partitional	Partitional	Hierarchical	Partitional
Exclusive vs. Overlapping	Exclusive	Overlapping	Exclusive
Complete vs. Partial	Complete	Complete	Partial
Cluster type	Prototype-based	Can be both	Density-based

Table 3.1: Characteristics of employed clustering methods

As explained by Tan (2018) a partitional clustering is one which separates the data points in non-intersecting clusters where each point is in only one cluster. The difference between complete and partial clustering methods is that the former allocates all points to a cluster, while the latter does not, which is sometimes the case for outliers and/or noise. Lastly, prototype-based clustering techniques assign the points to the cluster whose prototype, often the centroid, is closer to them then the prototype of any other cluster, while density-based ones unite points in a high-density area surrounded by a low-density area into a cluster.

Subsections 3.1, 3.2 and 3.4 will explain the theory about the implemented clustering methods and the reason for the parameters decisions taken, while subsection 3.4 illustrates how different clustering results will be compared.

3.1 k-Means

For a complete comparison of the clustering results, also the k-means algorithm was implemented, analogously to Chen et al. (2012).

This technique starts from k centroids and assigns each other point to the nearest centroid. Afterwards, it calculates the new centroid of each cluster according to an objective function and it repeats this process until centroids stop moving or a set amount of repetitions has been reached (Tan, 2018). However, letting the process select the initial centroids at random delivers only poor and not-reproducible results. Hence, it is suggested to choose them properly

One therefore needs to define four parameters first: the number of clusters k, the initial centroids, the distance measure and the objective function. In the case of data in the Euclidean space, one should use the Euclidean distance paired with Sum of Squared Error (SSE) as objective function, which is calculated in the following way (Tan, 2018):

$$SSE = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2 \tag{2}$$

where $dist(\mathbf{c}_i, \mathbf{x})$ is the Euclidean distance between the centroid of the i^{th} cluster and a point $\mathbf{x}$.

To find k many approaches have been proposed, but in this case we made use of the Silhouette Index, which will be explained in detail in subsection 3.4, and the Elbow Method. This is a visual rule of thumb where one clusters the data multiple times with different values of k, calculates the percentage of total variance explained (TVE) for each run, orders these according to their TVE. One then plot the total variance explained against the number of clusters and chooses the value of k where adding another cluster does not add sufficient information (Madhulatha, 2012).

For selecting the appropriate initial centroids there are also multiple approaches. One approach which works well, but may be computationally expensive, is choosing these points using hierarchical clustering. One takes a sample of the dataset and clusters it using hierarchical clustering, then draws out the solution with k clusters and uses the centroids of these as initial points for k-Means (Tan, 2018).

k-Means is often the clustering method of choice because of its simplicity and computa-

tional efficiency, but that comes at the cost of performance. In fact, k-means is effective at finding spherical clusters of similar sizes and densities, but may fail at identifying natural clusters if they do not satisfy the mentioned conditions (Tan, 2018). Moreover, k-means does not work well with data containing outliers, which were therefore removed before clustering in order to improve the calculations.

3.2 Hierarchical Clustering

With this term one groups a variety of clustering methods, which can be either agglomerative, that is starting from the single points and clustering them together, or divisive, i.e. going from one big cluster to single points by continuously splitting clusters (Tan, 2018). In this study the agglomerative approach will be applied, since it is the most common.

At each stage of the process the two closes points get clustered and the distance matrix is adjourned according to the Lance-Williams Formula (Lance and Williams, 1967):

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}| \tag{3}$$

where i and j are the two groups previous groups which merge into h and k is another group.

The values d_{hk} therefore depends on how d_{hi}, d_{hj} and d_{hj} are calculated, i.e. the distance measure, and the values of the coefficients α_i, α_j, β and γ, that is the linkage measure.

The difficulty relies in the fact that one can only calculate the distance between two points, while clusters usually contain more. Hence we need a linkage measure to allow us to determine the distance between clusters (Yim and Ramdeen, 2015). This choice is of utmost importance, since it affects the clustering procedure and merging process (Mazzocchi, 2008), according to equation 3.

With regard to distance, Soler et al. (2013) explained why the Mahalanobis distance is a suitable choice where the data does not present too many dimensions. This distance measure is calculated in the following way (Xiang et al., 2008):

$$d_A(i,j) = \sqrt{(x_i - x_j)^T A^{-1} (x_i - x_j)} \tag{4}$$

where A is the covariance matrix.

Concerning linkage, Punj and Stewart (1983) tested multiple linkage measures and found Group Average Linkage (UPGMA) and Ward's Minimum Variance to perform the best. However, the problems posed by outliers are much stronger when using Ward's method than group average (Tan, 2018). Since the target dataset contains only three variables, does not present natural groups of points but some noise, firstly outliers were removed and then Agglomerative

Clustering was implemented with the Mahalanobis function as distance measure and Group Average (UPGMA) as linkage measure. This measure implicates that distance between two clusters is considered the "average pairwise proximity among all pairs of points in the different clusters" (Tan, 2018) and, as a result, places the following coefficients in the Lance-Williams formula: $\alpha_i = \frac{n_i}{n_k}$, $\alpha_j = \frac{n_j}{n_k}$, $\beta = 0$ and $\gamma = 0$ (Lance and Williams, 1967).

After having selected these parameters, it is possible to procede with the hierarchical clustering, which can be graphically displayed in a dendogramm. This plot is a tree-form graph showing which clusters were merged at each step (Tan, 2018). An example will be displayed with the application of this method in subsection 4.2. The x-axis represent the different points in an order that makes sense for the merging process, while the y-axis represente the distance between clusters at the cluster merging. It is therefore conventional to select the appropriate number of clusters by counting the longest lines (Forina et al., 2002).

The main advantage of hierarchical clustering with respect to k-Means is that it can bring out hierarchical structures (Tan, 2018). However, it has the disadvantage that it does not have global criterion, just local one to merge the different clusters. Furthermore, each merging of the clusters is final, which can be a problem in the presence of high-dimensional data with many outliers (Tan, 2018).

3.3 DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) was designed by Ester et al. (1996) with the purpose of revealing clusters and noise in spatial data.

This method splits the dataset into different types of points (Tan, 2018):

- Core points: The ones inside the cluster, that is, the ones where, inside a space of length *Eps*, there are no less than *MinPts* points
- Border points: These fall within the neighbourhood of ones or many core points.
- Noise points: They are neither of the above.

All noise points are eliminated and not assigned to any cluster. Core points within a *Eps* distance are grouped in a cluster, while border points are assigned to one or more clusters if they are near enough to a core point. In case they are allocated to multiple clusters, one needs to decide to which one to put it (Tan, 2018).

The two parameters to imputed are therefore *Eps* and *MinPts*. The latter influences the distinction between core, border and noise points, while the former regulate the size of the

clusters (Verma et al., 2012).

The heuristic method to select these parameters has been proposed by Ester et al. (1996). For a certain value of k, one calculates the distance of each point to its k^{th} closest point, arranges the point in descending order of distance and plots them against their distance. Lastly, one choose the value that is the "threshold point is the first point in the first “valley” of the sorted k-dist graph" (Ester et al., 1996). k is then the value of *MinPts* and the threshold point is the value of *Eps*.

Ester et al. (1996) suggests that 4 is the optimal value for k in the case of 2-dimensional data. Instead, Sander et al. (1998) proposes to set it at 2 times the data dimensions. However, Schubert et al. (2017) advises to increase this value if the dataset has more dimensions, more outliers and more points.

In this study 3-dimensional data containing outliers was clustered, therefore different values of k have been tested and compared using the Silhouette Index. A more in depth explanation will be provided in subsection 3.4.

The main advantage of this clustering method is that it is better at dealing with noise and clusters of different sizes and shapes. However, it has issues in the case of high-dimensional data and when cluster have differing densities (Tan, 2018).

3.4 Cluster validation

Evaluating the results from a clustering process is especially important because often clusters will be found even though no such structure is present in the data (Tan, 2018).

In general, there are two kinds of cluster validation techniques: external and internal ones. External methods make use of external knowledge about the data, while internal ones rely only on the information present in the data (Rendón et al., 2011).

In the former case, one can assess the goodness of the clustering algorithm by comparing the results to the actual groups in a confusion matrix. In the latter case, when this is not possible because of the absense of a priori knowledge about the data, one can calculate the measure in which points in a cluster are related to each other and how different are the clusters from each other (Tan, 2018).

Clustering methods were evaluated externally, comparing their results with the RFM Segments explained in subsection 2.1, and internally using the Silhouette Coefficient. This index was chosen, since it has been proven to perform well also in the presence of noise (Handl et al., 2005). In particular, the Overall Average Silhouette Width was considered, which is the

average of the Silhouette values $s(i)$ for all the points in the dataset. This value is calculated in the following way (Rousseeuw, 1987):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5}$$

where $a(i)$ is the average distance between point i and all other points in the same cluster, while $b(i)$ is the minimum of the average distance of i with all points in cluster where i is not included. $s(i)$ can assume values between -1, signifying that the point i is missclassified and 1, which indicates that the point i is well-clustered (Rousseeuw, 1987). Therefore, one wishes to choose the parameters or the clustering methods in order to maximize the Overall Average Silhouette Width, as explained by Rousseeuw (1987).

Still, to have a broader picture, one can also plot all the Silhouette Widths in a Silhouette plot, an example of which is present in subsection 4.3. The heigher the bars, the higher the Width $s(i)$. Thus, one wishes for the Silhouette of all clusters to be as high as possible (Rousseeuw, 1987). Beside being used as a comparison method between clustering techniques, this plot can also be imployed in improving clustering results by transfering a point with negative Width $s(i)$ to its neighbouring cluster (Rousseeuw, 1987).

Lastly, a measure for cluster validity will also be computed to check if the data contains non-random structures. In particular, the Hopkins Statistic will be used, which is calculated in the following way (Tan, 2018):

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i} \tag{6}$$

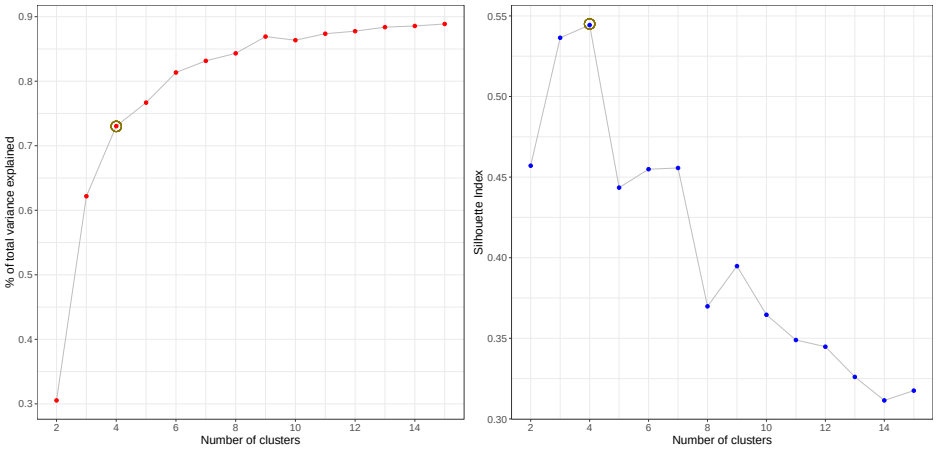
where p points are randomly sampled from the dataset and other p points generated from the uniform distribution with the same variation as the dataset. w_i and u_i then represent the distances to the nearest neighbour of the sample and the artificial points respectively (Tan, 2018). A sign of the presence of a cluster structure is that $u_i > w_i$, which implies that values of H close to 0 indicate that there might be natural groups in the data (Tan, 2018).

4 Cluster analysis

In contrast to Chen et al. (2012), different clustering methods were tested against each other. In subsection 4.1 it is illustrated how k-Means was implemented in this study, while subsection 4.2 explains how DBSCAN was applied to the dataset and subsection 4.3 deals with Agglomerative Hierarchical Clustering on the same data. Finally, subsection 4.4 will show how the methods were compared.

4.1 Implementation of k-Means

To select the value of the parameter k, as explained in subsection 3.1, the Silhouette Index and the percentage of total variance explained were calculated for k ranging from 2 to 15 and then displayed in figure 4.1. It is clear from both perspective that the appropriate number of clusters in this case is $k = 4$.



(a) Percentage of total variance explained **(b)** Silhouette Index

Figure 4.1: Plots for the choice of the number of clusters k

Then the initial centroids for the algorithm have been calculated using Agglomerative Hierarchical Clustering, cutting the tree at 4 clusters and computing the mean of each variable for every cluster.

Finally, the k-Means algorithm is implemented on the data looking for 4 clusters starting from these points, which are displayed in table A.4.

4.2 Implementation of Agglomerative Hierarchical Clustering

The first step of Hierarchical Clustering is choosing the distance and linkage measures. As described in subsection 3.2, the Mahalanobis distance and the Group Average Linkage (UPGMA) were selected. This leads to the Lance-Williams Formula looking as follows:

$$d_{hk} = \frac{n_i}{n_k}\sqrt{(x_h - x_i)^T A_{hi}^{-1}(x_h - x_i)} + \frac{n_j}{n_k}\sqrt{(x_h - x_j)^T A_{hj}^{-1}(x_h - x_j)} \tag{7}$$

After deciding on this method, one can run the agglomerative hierarchical clustering function, which, differently from the other methods discussed in this paper, does not directly deliver the cluster of membership of the points. Instead, this produces details about each merging step: which clusters were merged and the distance between them, in this cased called *Height*. This information can then be displayed graphically using a dendogramm. The one produced from this dataset is presented in figure 4.2, from which we can deduce, according to the reasoning explained in subsection 3.2, that the appropriate can be either 3 (Variation A) or 5 (Variation B), depending on how long we wish to keep the lines.

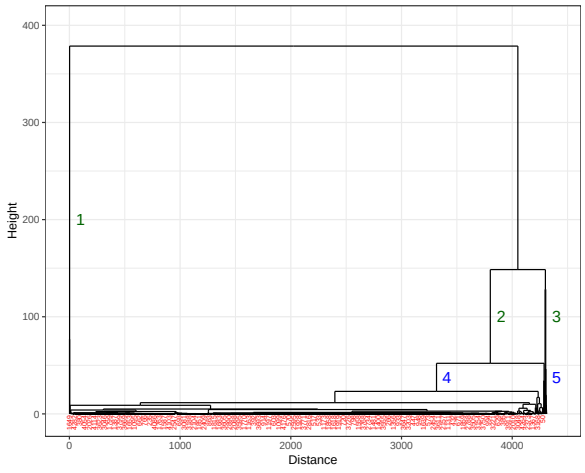
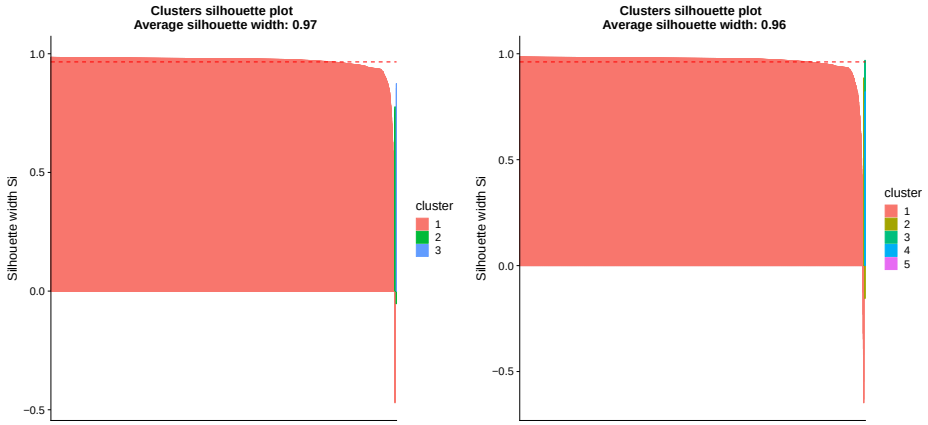


Figure 4.2: Dendogramm

Therefore, using the Silhouette method, the two variations were tested. As one can also see from figure 4.3 that in both cases more than 99% of the points were clustered together. In fact, 100% of the points in Cluster 1 in Variation B belong to the same cluster in Variation A. Thus, since Variation A's Overall Average Silhouette Width is higher than that of Variation B, even if by just a little, only Variation A will compared to the other clustering methods in subsection 4.4.



(a) Variation A **(b)** Variation B

Figure 4.3: Silhouette Plots

4.3 Implementation of DBSCAN

As described in subsection 3.3, one needs to firstly select the two marameters *MinPts* and*Eps*. This was done running the DBSCAN function multiple times and comparing the results by the Silhouette Index, whose interpretation is described in subsection 3.4 and by the k-dist method described by Ester et al. (1996).

Many values were tested, starting with 4 (Ester et al., 1996) and 6, i.e. two times the data dimension (Sander et al., 1998), for *MinPts* and values between 0.05 and 0.20 for *Eps*, which is the area where the "valley" is for those values of k, like it can be seen in figure 4.4.

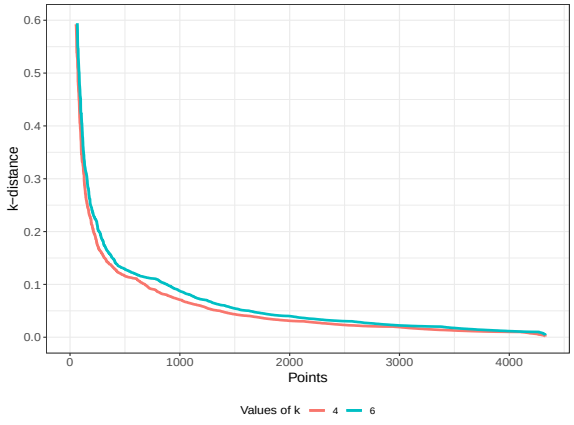


Figure 4.4: k-dist Graph

MinPts	Eps	Clusters	Noise Points	Silhouette Index
6	0.20	2	4.23	0.510
6	0.18	2	4.64	0.506
6	0.19	2	4.36	0.505
6	0.17	3	4.85	0.438
6	0.16	4	5.22	0.419
4	0.17	3	4.13	0.352
4	0.18	3	3.99	0.341
4	0.20	2	3.69	0.334
4	0.19	2	3.79	0.334
6	0.13	2	7.37	0.293

Table 4.1: First 10 tested parameter combinations by Silhouette Index

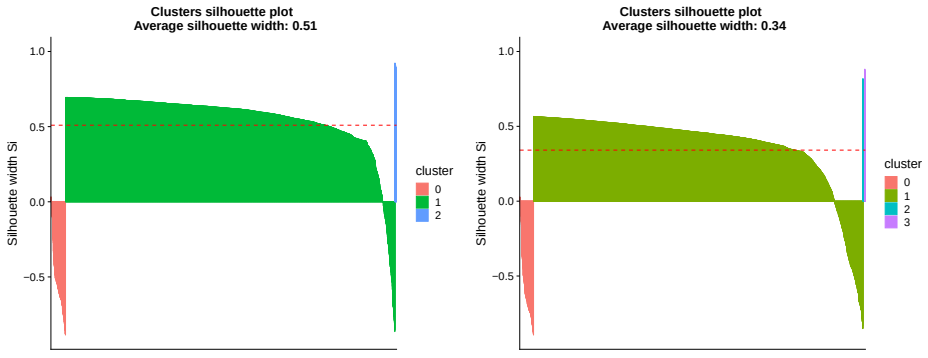
Even though from figure 4.4 it looks like the best value for *Eps*, with respect to both values

of k, is around 0.11, table 4.1 showing the combinations of parameters ordered by decreasing Silhouette Index show another picture. In fact, according to this value the best combination of tested parameters is $MinPts = 6$ and $Eps = 0.20$.

One can in addition see that this combination produces only two clusters and that 4.23% of the points are being identified as noise point and thus not clustered. This might however not be optimal from the business point of view, since one wishes to know the group of membership of all points. For these reasons we will firstly compare the following parameter combinations:

- Variation A: $MinPts = 6$, $Eps = 0.20$
- Variation B: $MinPts = 4$, $Eps = 0.18$

From the Silhouette Plots for these two variations, presented in figure 4.5, it becomes clear how variation B does not present an added value to variation A. In fact, almost all of the noise points from variation A is also considered noise in variation B, while the rest gets split up among the clusters. Moreover, the same number of customers is in Cluster 1 in both variations, out of which 99.87% is shared. Therefore, since variation A's Overall Average Silhouette Width is also 17 percentage points higher than that of Variation B, only Variation A will be considered in the cluster evaluation in subsection 4.4.



(a) Variation A **(b)** Variation B

Figure 4.5: Silhouette Plots

4.4 Cluster results

Each clustering method tested produced different results, therefore it is important to compare them.

Firstly of all, it is meaningful to point out that 60% of the customers have been placed in the biggest cluster by all three analyzed techniques. However, this group contains customers

of different segments, almost 40% of which being "Champions" or "Loyal customers", that is customers with above-average values of all three variables.

It is further interesting to note that over 95% of the customers, without the outliers, have been placed in the same cluster by both Agglomerative Hierarchical Clustering and DBSCAN. In fact, one can see no great different between the descriptive statistics from table A.3 to the ones in table 4.2.

variable	min	1Q	median	3Q	max	iqr	mean	sd
recency	0.00	16.00	49.00	147.00	364.00	131.00	92.13	99.56
frequency	1.00	1.00	2.00	5.00	25.00	4.00	3.79	3.80
monetary	-1592.49	277.28	600.02	1355.97	9114.94	1078.69	1069.02	1267.20

Table 4.2: Descriptive statistics of customers in biggest cluster from Hierarchical Clustering and DBSCAN

Additionaly, DBSCAN was able to find all outliers that were removed for the other two clustering methods. In fact, it indetified as noise points more customers than those excluded with the Elbow Method, but 15% of the points tagged as outliers using the Mahalanobis Distance and the Chi-Square value.

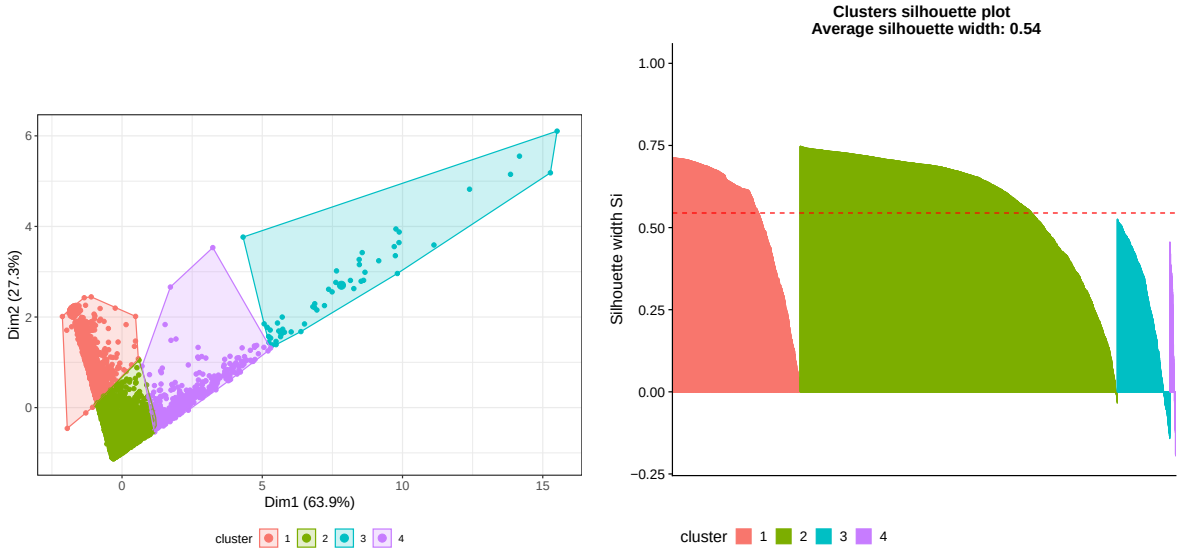
However, according to the Silhouette Index, the most effective clustering method was Agglomerative Hierarchical Clustering with a value of 0.966, compare to 0.544 of k-Means and 0.510 of DBSCAN. These results, coupled with a value of the Hopkins Statistic for cluster tendency being almost 0, suggest that there are no natural clusters in the data. Nonetheless, we will look into greater detail at the results provided by k-Means, which is the technique which split the data into more clusters.

Cluster	Frequency	Proportion
1	1101	25.42%
2	2708	62.53%
3	51	1.18%
4	451	10.41%

Table 4.3: Frequency and proportion of different clusters from k-Means

K-Means produced four clusters of different sizes, as it is shown in table 4.3, which also seem to be quite distinct according to figure 4.6a, where the points are plotting according to

the first two principal components. However, according to the Silhouette plot in figure 4.6b, they do not deliver good results, since two clusters have below average Silhouette Scores and there is a big fluctuation of the Silhouette values inside each cluster.



(a) Cluster plot

(b) Silhouette plot

Figure 4.6: Plots for k-Means clustering

However, it is also essential to look how the different variables are distributed in each cluster, as is shown in table A.6. Firstly, it is necessary to understand if the employed clustering technique caught natural groups or if these were not present to begin with. Secondly, it is useful when drawing business and marketing decisions according to the group in which each customer has been clustered.

Looking at the information present in table 4.3 and A.5, that is, the contingency table of clusters against RFM segments, it becomes even more explicit how there is no clear distinctions between the clusters. They have no noticeably different levels of the RFM variables and the RFM segments are quite evenly distributed across the clusters (see table A.5).

All things considered, we can draw the conclusion that there is no group structure in the data. One can therefore take better marketing decisions looking at the RFM scores and segments.

5 Conclusions

The objective of this work was firstly to compare clustering methods using an RFM model, that is producing groups to be used for business or marketing decisions, and secondly to test the results from Chen et al. (2012), achieved using k-Means, against two other clustering methods, Agglomerative Hierarchical Clustering and DBSCAN, to see if there are indeed groups of customers and if one can find better ones.

The data contained transactions for a UK-based online company, for which Recency, Frequency and Monetary (RFM) have been calculated according to Bult and Wansbeek (1995). Outliers in the RFM model have been then identified and removed when running k-Means and Agglomerative Hierarchical Clustering. DBSCAN is in fact able to identify outliers independently. The data has been standardized, to avoid the problem of different scales for the variables, and the parameters for each clustering method have been selected according to relevant literature (Madhulatha, 2012; Tan, 2018; Soler et al., 2013; Punj and Stewart, 1983; Forina et al., 2002; Ester et al., 1996).

The three techniques delivered different results, which have been compared looking at their size, Silhouette Index and at how they reflect the RFM Segments. According to the Silhouette Index, the most effective method is Agglomerative Hierarchical Clustering, which however places most of the customers in the same cluster. K-Means was the procedure which produced more clusters, which however did not show relevant differences in the RFM variables.

All in all, the findings show that there no innate groups in the data, which is in fact better split using the segments based on the RFM scores.

Unfortunately, an exact comparison with Chen et al. (2012) was not achievable because of the different dataset, which indeed includes transactions for the same company but for a time period differing of one month. Moreover, it was not possible to verify exactly which customers were clustered in Chen et al. (2012). However, the distribution of the RFM variables was very similar to the ones displayed in Chen et al. (2012), which might be a good indicator of the comparability of the projects.

Further research could examine different variables, maybe expanding the basic RFM model to one of its extended versions, and undertake Principal Component Analysis as basis of the clustering as a process of removing noise (Husson et al., 2010). Finally, one could also test different parameters for the employed clustering methods or even experiment with others.

References

Birant, D. (2011): "Data mining using RFM analysis," in *Knowledge-oriented applications in data mining*, InTech.

Bult, J. R. and T. Wansbeek (1995): "Optimal selection for direct mail," *Marketing Science*, 14, 378–394.

Chen, D., S. L. Sain, and K. Guo (2012): "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *Journal of Database Marketing & Customer Strategy Management*, 19, 197–208.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996): "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, 226–231.

Filzmoser, P., R. G. Garrett, and C. Reimann (2005): "Multivariate outlier detection in exploration geochemistry," *Computers & geosciences*, 31, 579–587.

Forina, M., C. Armanino, and V. Raggio (2002): "Clustering with dendrograms on interpretation variables," *Analytica Chimica Acta*, 454, 13–19.

Handl, J., J. Knowles, and D. B. Kell (2005): "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, 21, 3201–3212.

Hebbali, A. (2019): "RFM - Customer Level Data," `https://cran.r-project.org/web/packages/rfm/vignettes/rfm-customer-level-data.html`, last accessed on 23. March 2019.

Husson, F., J. Josse, and J. Pages (2010): "Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data," *Applied Mathematics Department*.

Jain, A. K., M. N. Murty, and P. J. Flynn (1999): "Data clustering: a review," *ACM computing surveys (CSUR)*, 31, 264–323.

Khajvand, M., K. Zolfaghar, S. Ashoori, and S. Alizadeh (2011): "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study," *Procedia Computer Science*, 3, 57–63.

Lance, G. N. and W. T. Williams (1967): "A general theory of classificatory sorting strategies: 1. Hierarchical systems," *The computer journal*, 9, 373–380.

Madhulatha, T. S. (2012): "An overview on clustering methods," *arXiv preprint arXiv:1205.1117*.

Mazzocchi, M. (2008): *Statistics for marketing and consumer research*, Sage.

Punj, G. and D. W. Stewart (1983): "Cluster analysis in marketing research: Review and suggestions for application," *Journal of marketing research*, 20, 134–148.

Rendón, E., I. Abundez, A. Arizmendi, and E. M. Quiroz (2011): "Internal versus external cluster validation indexes," *International Journal of computers and communications*, 5, 27–34.

Rousseeuw, P. J. (1987): "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, 20, 53–65.

Rousseeuw, P. J. and B. C. Van Zomeren (1990): "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical association*, 85, 633–639.

Sander, J., M. Ester, H.-P. Kriegel, and X. Xu (1998): "Density-based clustering in spatial databases: The algorithm gdbscan and its applications," *Data mining and knowledge discovery*, 2, 169–194.

Sarvari, P. A., A. Ustundag, and H. Takci (2016): "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis," *Kybernetes*, 45, 1129–1157.

Schubert, E., J. Sander, M. Ester, H. P. Kriegel, and X. Xu (2017): "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems (TODS)*, 42, 19.

Soler, J., F. Tencé, L. Gaubert, and C. Buche (2013): "Data clustering and similarity," in *The Twenty-Sixth International FLAIRS Conference*.

Tan, P.-N. (2018): *Introduction to data mining*, Pearson Education India.

Tkachenko, Y. (2015): "Autonomous CRM control via CLV approximation with deep reinforcement learning in discrete and continuous action space," *arXiv preprint arXiv:1504.01840*.

Verma, M., M. Srivastava, N. Chack, A. K. Diswar, and N. Gupta (2012): "A comparative study of various clustering algorithms in data mining," *International Journal of Engineering Research and Applications (IJERA)*, 2, 1379–1384.

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

Xiang, S., F. Nie, and C. Zhang (2008): "Learning a Mahalanobis distance metric for data clustering and classification," *Pattern recognition*, 41, 3600–3612.

Yeh, I.-C., K.-J. Yang, and T.-M. Ting (2009): "Knowledge discovery on RFM model using Bernoulli sequence," *Expert Systems with Applications*, 36, 5866–5871.

Yim, O. and K. T. Ramdeen (2015): "Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data," *The quantitative methods for psychology*, 11, 8–21.

A Tables

Variable name	Data type	Description
InvoiceNo	Nominal	Invoice number
StockCode	Nominal	Product code
Description	Nominal	Product name
Quantity	Numeric	Quantity of the specific product in the specific transaction
InvoiceDate	Nominal	Date and time of each transaction
UnitPrice	Numeric	Product price per unit in sterling
CustomerID	Nominal	Customer number
Country	Nominal	Country name: where the relative customer resides

Table A.1: Variables in the used dataset (Source: Chen et al. (2012))

Segment	Description	R	F	M
Champions	Bought recently, buy often and spend the most	4 - 5	4 - 5	4 - 5
Loyal Customers	Spend good money. Responsive to promotions	3 - 5	3 - 5	3 - 5
Potential Loyalist	Recent customers, spent good amount, bought more than once	4 - 5	1 - 3	1 - 3
New Customers	Bought more recently, but not often	4 - 5	1	1
Promising	Recent shoppers, but haven't spent much	3 - 4	1	1
Need Attention	Above average recency, frequency and monetary values	2 - 3	3	3
About To Sleep	Below average recency, frequency and monetary values	2	2	2
At Risk	Spent big money, purchased often but long time ago	1 - 2	3 - 4	3 - 4
Can't Lose Them	Made big purchases and often, but long time ago	1	5	5
Hibernating	Low spenders, low frequency, purchased long time ago	2	2	2
Lost	Lowest recency, frequency and monetary scores	1	1	1

Table A.2: Customer segmentation according to RFM Scores (Source: Hebbali (2019))

	recency	frequency	monetary
min	0.00	1.00	-4287.63
1Q	15.00	1.00	288.51
median	48.00	3.00	629.31
3Q	145.00	5.00	1532.06
max	364.00	126.00	34514.18
mean	90.6641	4.5746	1430.83
sd	99.4549	6.3443	2604.91

Table A.3: Descriptive statistics RFM variables after removal of outliers

	1	2	3	4
recency	-0.00	-0.86	-0.86	-0.01
frequency	-0.11	4.28	11.26	1.38
monetary	-0.10	4.50	10.95	9.29

Table A.4: Initial centroids to use in k-Means algorithm

RFM Segment	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Champions	19.62%	21.49%	25.49%	22.39%
Loyal customers	15.35%	14.07%	9.8%	13.97%
Potential loyalists	15.26%	17.17%	9.8%	17.52%
New customers	1.09%	1.96%	3.92%	2.88%
Promising	2.54%	2.18%	5.88%	3.1%
Need attention	5.18%	4.28%	7.84%	4.66%
About to sleep	1.73%	1.07%	1.96%	1.11%
At risk	11.44%	12.52%	1.96%	11.75%
Can't lose them	1.27%	0.66%	0%	0.67%
Hibernating	11.99%	10.6%	15.69%	10.86%
Lost	7.18%	7.24%	7.84%	6.21%
Average customers	7.36%	6.76%	9.8%	4.88%

Table A.5: Contingency tables of clusters with RFM Segments

	min	1Q	median	3Q	max	iqr	mean	sd
Cluster 1								
recency	0.00	16.00	49.00	148.00	364.00	132.00	94.74	103.34
frequency	1.00	1.00	3.00	5.00	87.00	4.00	4.54	6.23
monetary	-1592.49	294.29	626.99	1562.12	25079.60	1267.83	1466.20	2549.21
first_purchase	0.00	4.00	8.20	10.70	12.00	6.70	7.29	3.77
Cluster 2								
recency	0.00	15.00	48.00	146.00	364.00	131.00	90.10	98.52
frequency	1.00	1.00	3.00	5.00	126.00	4.00	4.56	6.44
monetary	-4287.63	287.54	627.52	1501.62	34514.18	1214.08	1384.87	2547.36
first_purchase	0.00	3.60	8.00	10.53	12.00	6.92	7.14	3.84
Cluster 3								
recency	0.00	12.00	44.00	116.00	363.00	104.00	87.51	103.73
frequency	1.00	1.00	2.00	6.00	33.00	5.00	5.37	7.60
monetary	35.40	205.16	497.61	1499.68	33350.76	1294.52	2032.91	5160.64
first_purchase	0.30	3.45	7.50	10.20	11.90	6.75	6.86	3.97
Cluster 4								
recency	0.00	14.00	43.00	132.00	364.00	118.00	84.46	94.70
frequency	1.00	1.00	2.00	5.00	48.00	4.00	4.65	5.88
monetary	-1192.20	282.96	662.59	1712.12	26318.05	1429.16	1552.36	2653.17
first_purchase	0.10	3.20	7.80	10.80	12.00	7.60	7.02	3.98

Table A.6: Descriptive statistics of RFM variables in the different clusters

Declaration of Authorship

I hereby confirm that I have authored this Seminar Paper independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, March 27, 2019

Silvia Ventoruzzo