Introduction to Data Mining for Business Intelligence

Customer segmentation to define a marketing strategy

PCA AND CLUSTER ANALYSIS

Sofia Gervasoni (100448791)

***(for the R Script please take a look at the R files attached)**

Introduction

I am given a dataset that contains information about the usage of a credit card by 8950 (number of rows) customers. The usage of the credit card is analysed by taking into consideration 17 variables (number of columns), that mainly refers to:

- Amount available in the account to make purchases and how frequently it is updated (e.g. Balance, Balance frequency)
- Information about the purchases made by the costumer (e.g. Purchases, One-off purchases, frequency of purchases and one-off purchases, number of transaction for purchases)
- Information about how the purchases are paid (e.g. Cash advance, installments, and the frequency of these type of payment, percentage of full payments made by the customer)
- Credit limitation for the user (e.g. Credit limit)
- Tenure of credit card service for user (e.g. Tenure)

The aim is to group the credit card holders in clusters, taking into account their characteristics, in order to develop a marketing strategy.

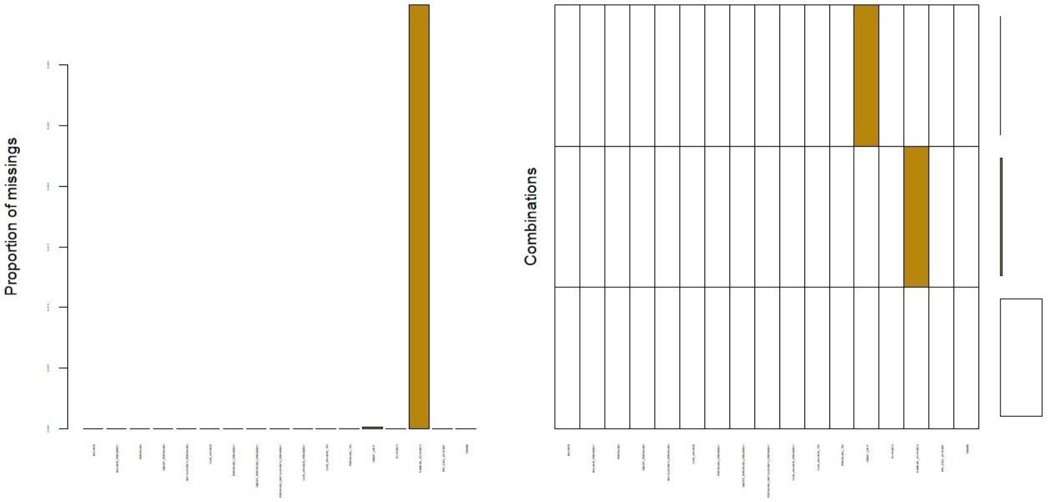
To do so I will apply the *principal component* and *cluster analysis* learned in this course of “Introduction to Data Mining for Business Intelligence”.

Pre-processing

Missing values

This dataset contains 314 missing values, so that, in order to do the analysis it is necessary to solve this problem, substituting the NAs with numbers.

First, it is useful to understand where the missing values are located and if there are any columns (variables) that contain more missing values than others, so that we can treat them in different ways. In order to understand where the missing values are located, I used the libraries *VIM* (for a graphic visualisation of missing values) and *naniar* (to precisely understand the number of missing values for each column).



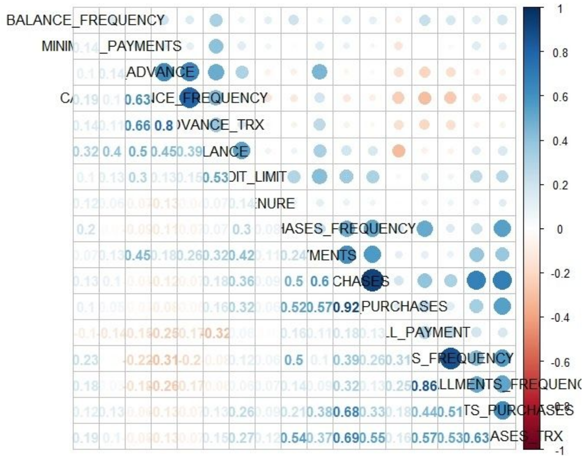
As we can see, in the previous graph, most of the missing values are located in a single column (even if there are some other missing values in another one). More in detail, there are 313 (3.5%) missing values in the variable "Minimum Payments" and just 1 (0.01%) missing value in the column "Credit limit". So that, I approached the problem of missing values in two different ways for the two variables.

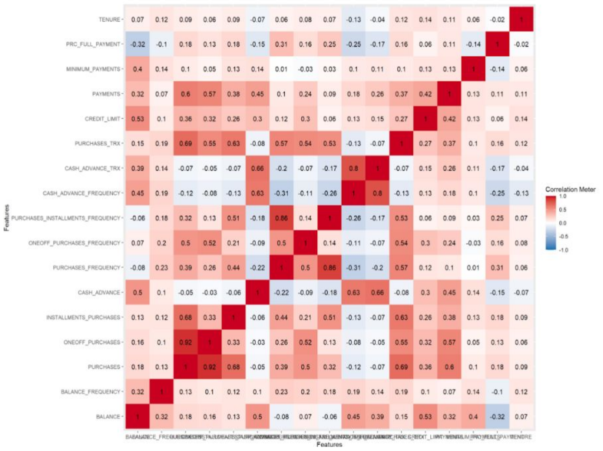
Starting with "Credit limit", the missing value is on the observation 5204, so that we can substitute this NA with the mean or the median (because it is just an observation). So, I tried to substitute first the mean and then the median, and I calculated the standard deviation for the two. From these calculations I found out that substituting the missing value with the mean, the standard deviation is lower. So that, for this variable, I solved the problem of missing value by substituting the NA with the mean.

When it comes to "Minimum payments" the things get harder because there is a larger amount of missing values. In order to impute this large amount of data without losing the correlation structure of the dataset, I used knnImputation() from the library DMwR*.

So now I have no missing values left and I can start my analysis.

Correlation





The strongest correlations are between the variables of the same "category" (as "category" I mean for example "Purchases" and "Cash advance").

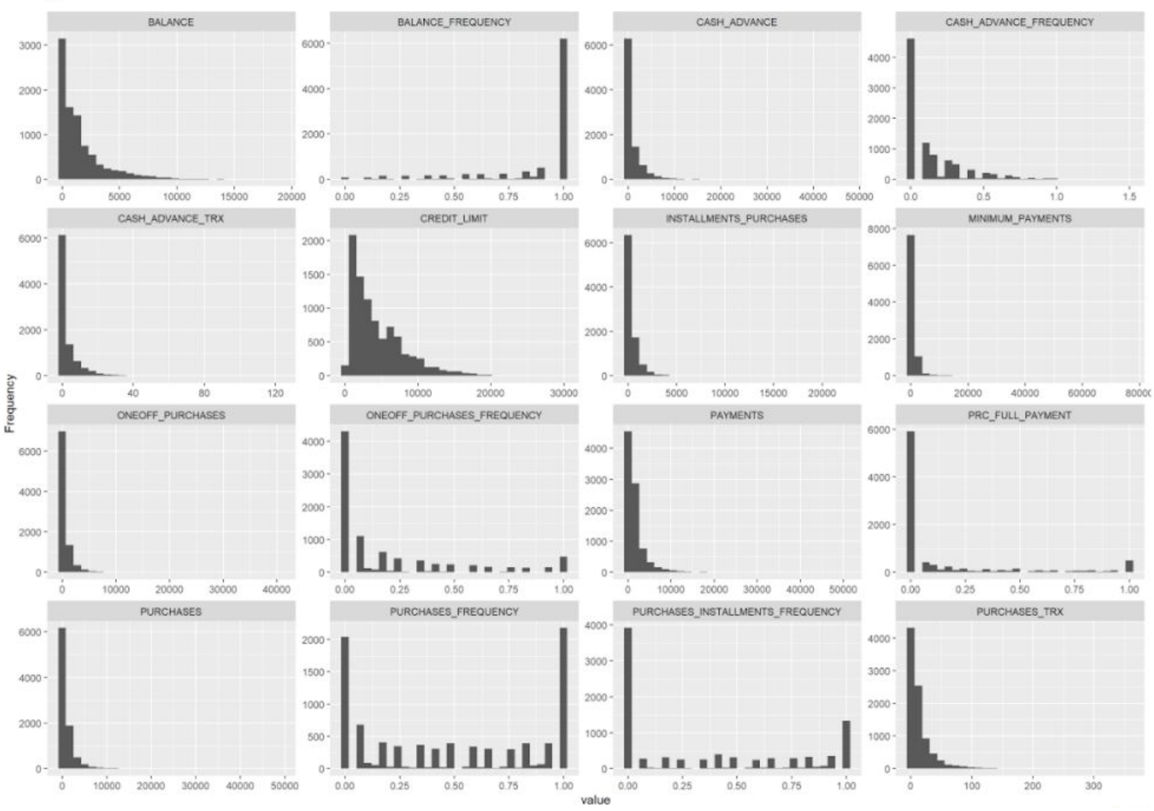
For example, the correlation between:

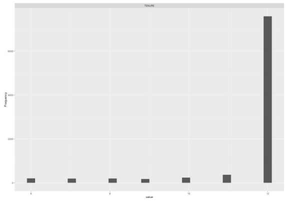
- Cash advance *frequency* and *TRX* is 0.8
- Cash advance and Cash advance frequency is 0.63
- Cash advance and cash advance TRX is 0.66
- Purchases and one-off purchases is 0.92
- Purchases frequency and purchases installment frequency is 0.86 .

There is also a weak negative correlation (-0.26) between Cash advance and the Payment in Installments. This makes sense because if a consumer tends to pay in advance would use less of the payment with installments. Also the correlation between Balance and the percentage of full payments seems to be slightly negative (-0.32).

In addition, the variable Tenure, seems to be the least correlated variable in respect to all the other variables (with values so close to 0).

Transformation

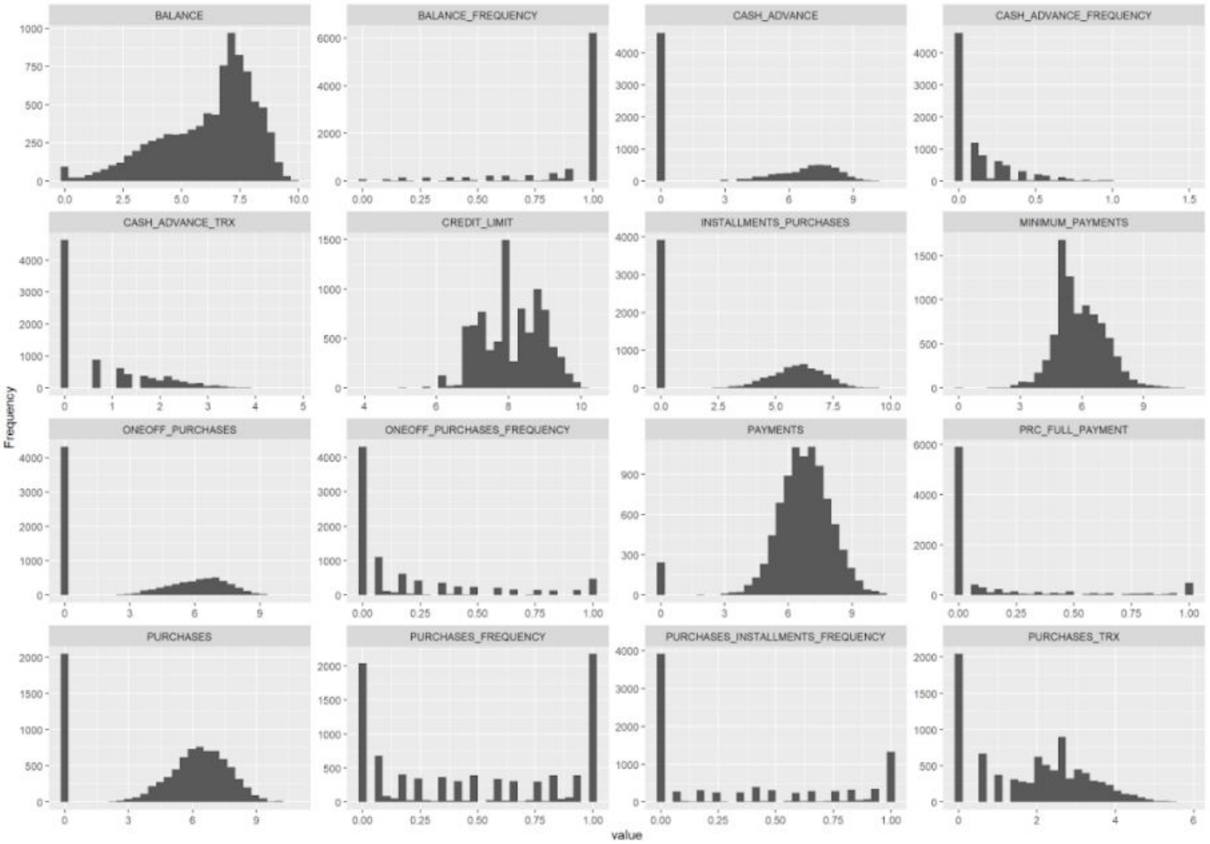




As shown by the previous graphs, the distributions of the variables are not symmetric, so before starting our PCA or cluster analysis we have to transform the variables in order to obtain symmetric distributions (it is not needed a gaussian distribution but at least symmetric).

For the variables balance, purchases, one-off purchases, installment purchases, cash advance, cash advance TRX, purchases TRX, credit limit, payments and minimum payments I used a logarithmic transformation in order to obtain distributions that are more symmetric than before.

Whereas when it comes to the variables that refers to the frequencies and the percentage of full payments they are close to logic variables, so I started the analysis without transforming them (there is no transformation that makes them symmetric) and I treated them in two different ways: first I did the *analysis maintaining these variables* and then I did the *analysis dropping these variables*.

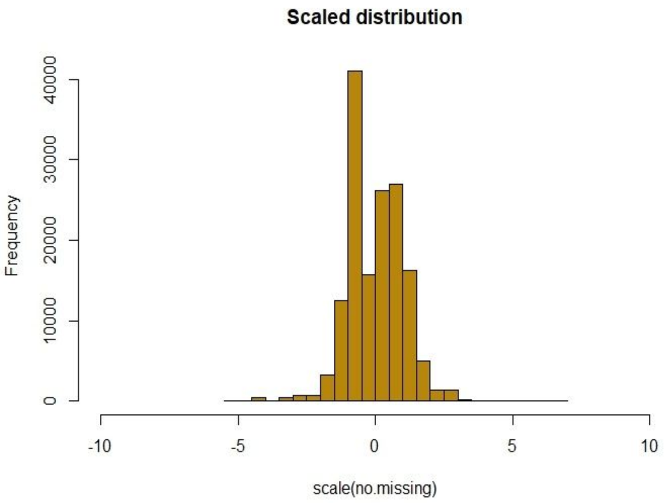


As we can see, after the transformation, the variables appear more symmetrical (except the variables that refer to frequencies and the percentage of full payments on which I did not apply any transformation).

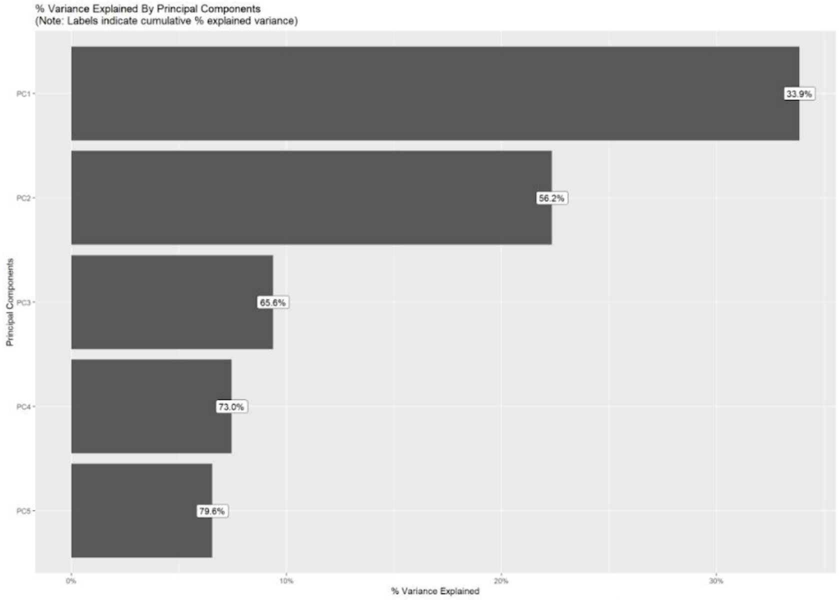
Analysis with all the variables

Principal component analysis

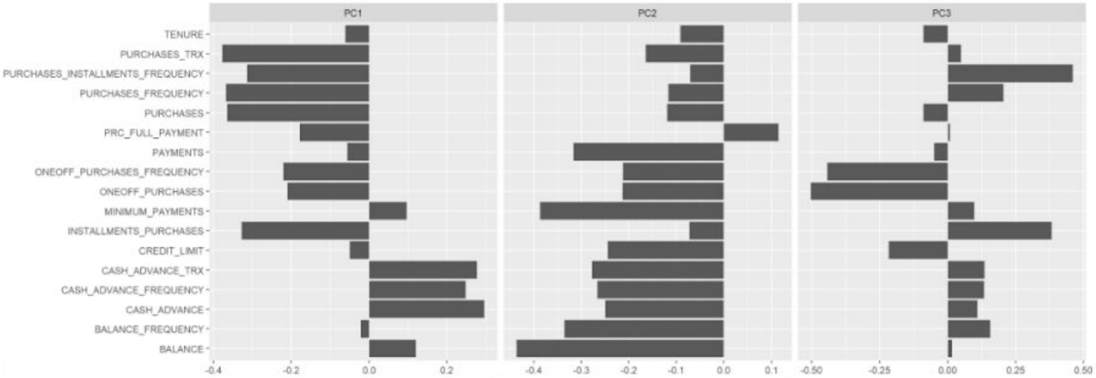
The aim of Principal component analysis is to visualize high dimensional projects in lower dimensions. While reducing the number of variables, we have to make sure that principal components maximize the explanation of the variance (in order to lose the lowest number of information as possible). It is important to scale the datas before starting the analysis, so that it is possible to obtain a symmetric distribution (normalized).



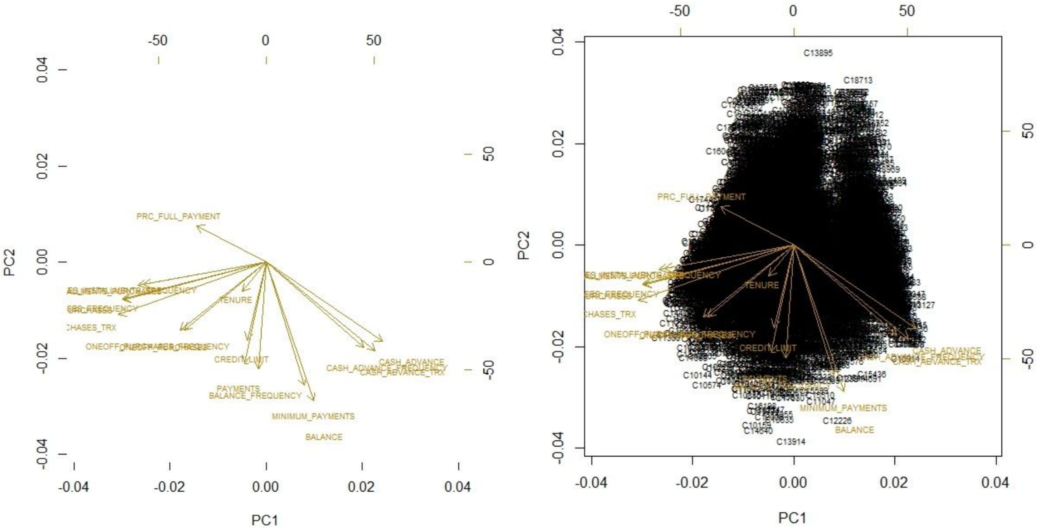
Once the data has been scaled, it is possible to apply the command *prcomp()* that returns the principal components.



In order to obtain a satisfactory explanation of the variance (56.2% - 65.6%) , in this case, we have to keep 2 or 3 principal components, where the first principal component explains the 33.9% of the variance, the second PC explains the 22.3% of the variance and the third PC explain the 9.4% of the variance (as we can see from the scree plot). The third component is not so relevant as the first two in explaining the variance.



The first principal component is "purchases TRX", the second principal component is "Balance", whereas the third principal component is "one-off purchases".



As we can see from the biplot, the first two principal components (purchases TRX and Balance) are orthogonal, actually two variables that are uncorrelated create an angle of 90°, because:

If I standardize: $x = \frac{xi - \bar{x}}{\sigma}$

Cor (x, y) $= \frac{\sum(Xi-\overline{X})\,(Yi-\overline{Y})}{\sigma_x\ \sigma_y} = \frac{\sum Xi\,Yi}{\sigma x\ \sigma y}$, *if* $\bar{x} = 0$ *and* $\sigma_x = 1$ *(because of the standardization)*

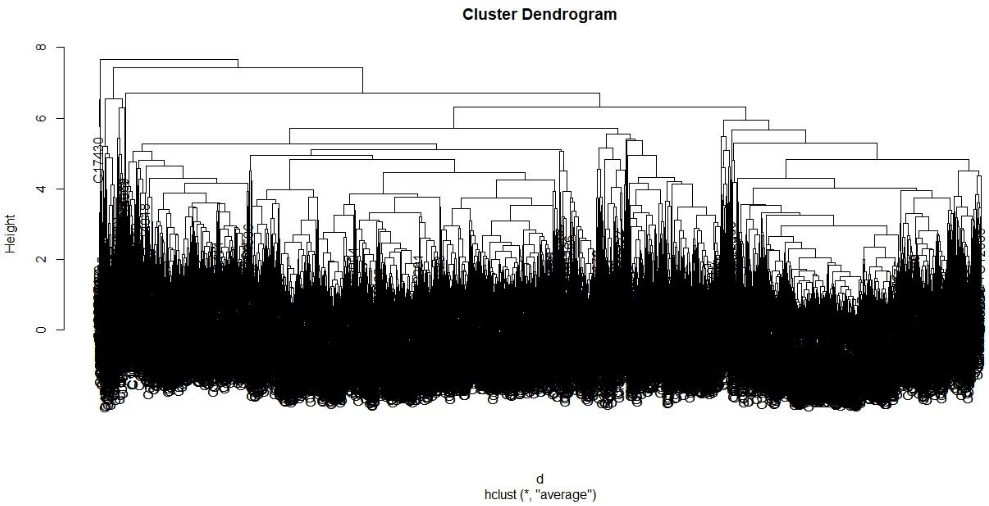
So, Cor (x, y) = 0 if and only if $\sum x_i y_i = 0$*, that means that the variables are uncorrelated just if they are orthogonal (that means that between them there is an angle of 90°).*

$$\cos\vartheta = \frac{x^T y}{|x||y|} \rightarrow \cos\vartheta = 0\ (90°) \rightarrow x^T y = \sum x_i y_i = 0$$

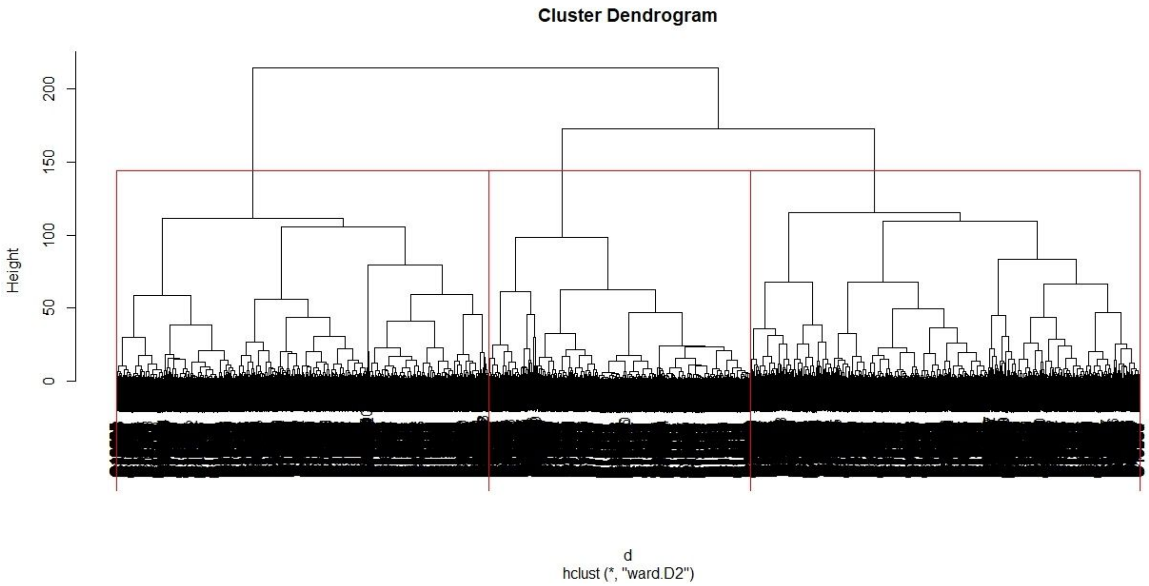
Cluster Analysis

Hierarchical analysis

Before starting, we have conducted analysis to understand which is the best method (between ward, single, complete and average) to apply to hclust(). In order to do so, we applied the cophenetic distance and we found out that the higher correlation between the euclidean distance and cophenetic distance is obtained with the "average method". Although, applying the average method we obtained a chaining problem.

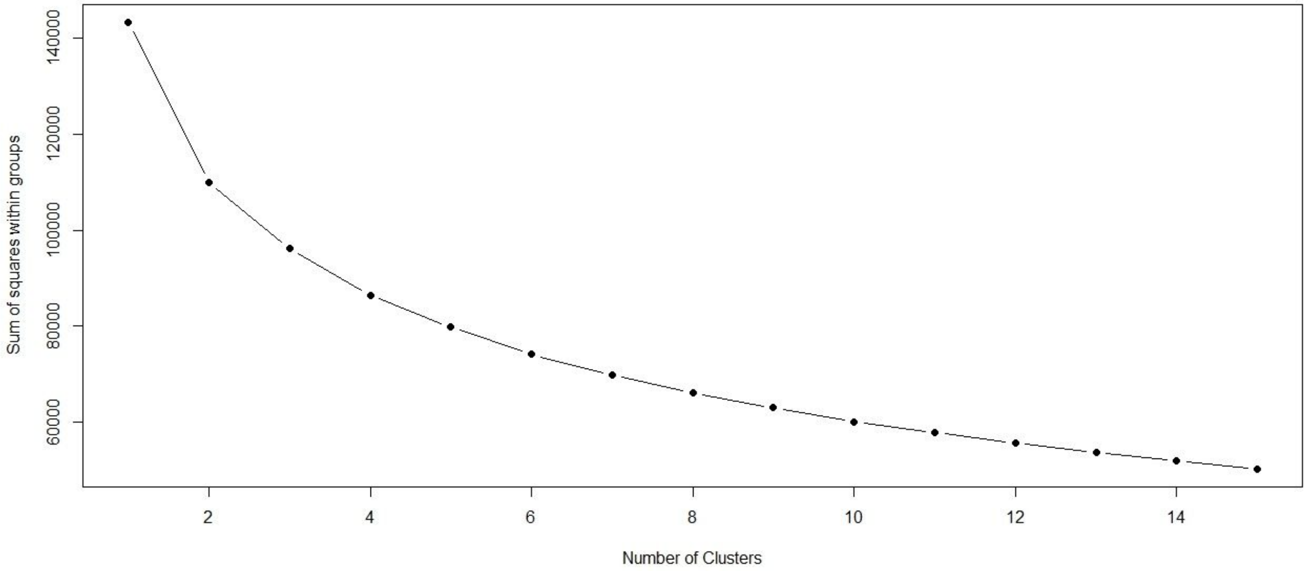


The second method suggested by the cophenetic distance is the complete method, but even in this case the results are not clear. So that, I decided to apply the "ward" method obtaining the following dendrogram, where in my opinion the ideal number of clusters is 3.



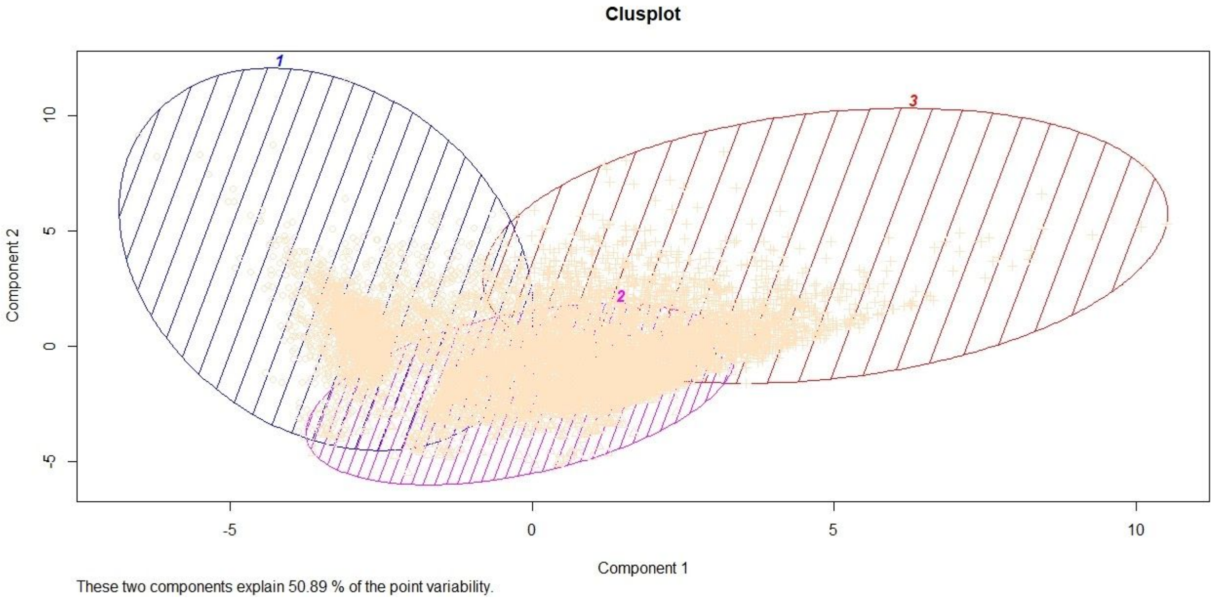
K-Mean

By applying this method, it is necessary to choose the number of clusters in advance, and in order to do so it possible to use the "elbow chart".



This "elbow chart" suggests 3 as the ideal number of clusters, actually 3 is the value located in the "elbow".

So that it is possible to apply the command kmeans() to find out how the 3 clusters are composed and graph them with a clusplot.



Partial conclusions

Once found that the ideal number of clusters (both for Hierarchical and K-mean methods) is 3, it is time to understand the characteristics of the individuals that make up these 3 clusters.

Thanks to the command cutree(), it is possible to obtain the individuals that belong to the different clusters. So I created 3 vectors that contain the number of rows of the different observations and, applying the command summary(), I obtained information about range, mean, median and quantiles for each variable and each cluster. Thanks to these indicators I was able to understand the common characteristics of the individuals in a cluster and the differences between the 3 clusters.

The first cluster contains 3405 individuals, the second cluster contains 2294 individuals and the third contains 3251 individuals.

Starting from the variable *"Balance"*, the first group seems to have the lowest values for this variable, whereas the second cluster has intermediate values and the third group has the highest values. That means that the customers that belong to the third group are the ones that have the larger amount available to make purchases ("the richest"), whereas the ones that belong to the first group are the ones that have the lower amount available ("the poorest"). Additionally, the minimum value of balance for the third group is higher than 0 and the maximum value for this group is the highest value between the 3 clusters. The results obtained for "balance" are confirmed by the results obtained for "*Balance frequency*", that is the variable that refers to how many times the balance is updated. Also in this case the third cluster is the one with the highest values, the second has intermediate values, whereas the third is the one with lower values.

When it comes to the *"Purchases"*, the third group is the one with higher purchases, and this is normal since they have the higher amount of money available for purchases. But when it comes to the other two groups something strange happens. Actually, the individuals that belong to the second cluster, make lower purchases than the individuals that belong to the third cluster, even if they have higher financial means. These results are also confirmed by the variable that refers to the *one-off purchases*, that contains information about the maximum amount spent in a single purchase. Also the variables *"purchases frequency"* and *"one-off purchase frequency"*, confirms our results, since it is a measure of how frequent the purchases (and one-off purchases) are made. Also the variable purchase TRX, that gives the number of purchases made by each individual, confirms what I have said until now.

As we have said for the purchases, even if the individuals of the first group are the ones that have the lower financial means, they spent more than what the individuals of the second group do and in order to do so they pay in installments. Infact, compared to the other two groups they are the ones who rely on installment payments the most. The individuals of the second group do not rely on *installment payments* (very low values). The individuals of the third group rely on the installment payments, even if on average they do it less than the individual in the first group. These results are confirmed by the variable “*purchases in installments frequency*”, that is a measure of how frequently the purchases are paid in installments.

As it is expected, when it comes to the *cash advance* payment, the individuals that use more this type of payment are the ones that belong to the second cluster (as they use less the installment payment), whereas the individual of the first group tend not to pay in advance (as they have the tend to use installment payments). The values for the third group are on average, but lower than the values for the second group. These results are confirmed by the variable “*cash advance frequency*” (that is a measure of how frequently the purchases are paid in advance) and “*cash advance TRX*” (that is a measure of how many transactions are paid in advance).

When it comes to the *credit limit*, the first cluster is the one with lower values, and this is normal since they have the lower balances and tend to pay with installments payments instead of paying in advance. The individuals of the second cluster, on average, are the one that have the highest credit limit, and this is also normal since they have high balances and they tend to pay in advance (and not in installment). The individuals of the third group, even though they have the highest balance, have (on average) a lower credit limit than the individuals of the second group, and this is most likely due to the fact that they tend to pay on installment instead of paying in advance.

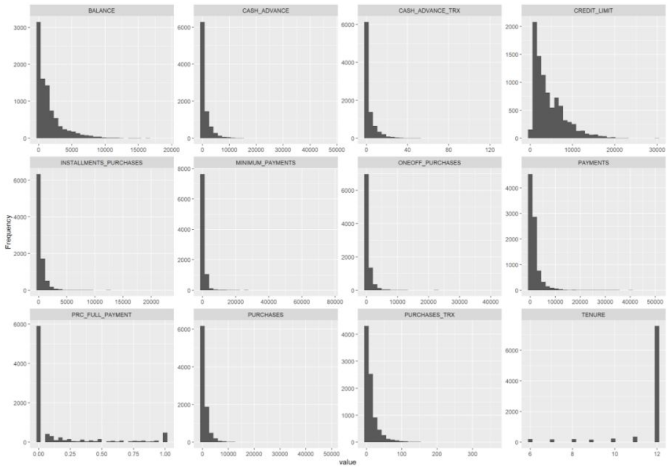
Something strange happened when it came to analyse the percentage of full payments. In this case, the first cluster has the highest values, whereas the second has the lowest values. This is probably due to the fact that the bank (or in general the institution that provides the credit), wants to be sure that the individuals that belong to the first group pay back the debit, so they put a clause for these clients that have to pay the debit within a certain period. Infact, the customers of the first group are the most risky for the bank, since they are the ones that could have problems in repaying the debt.

The variable *Tenure* does not provide any relevant information to our analysis, since the values for the three groups are more or less the same.

Analysis without Frequences

As said before, the frequencies do not add information to our analysis (they just confirm the results of the corresponding variables), moreover they have an asymmetric distribution that does not become symmetric with any transformation. This is due to the fact that these variables assume values that range between 0 and 1, and many of them assume 0 or 1, so that they can be compared to logic variables. So I tried to do the analysis dropping these variables to understand if I get different results compared to the previous analysis.

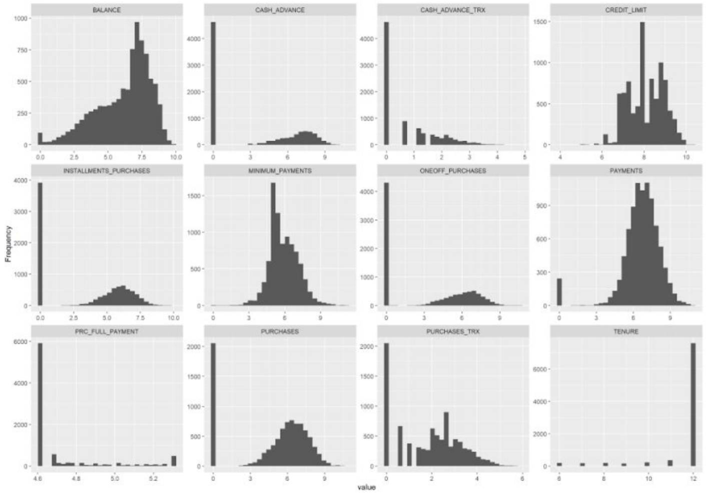
Transformation



As before, the distributions are not symmetrical, so that I applied some transformations.

As all these distributions (except tenure) are right skewed the ideal transformation is the logarithmic one.

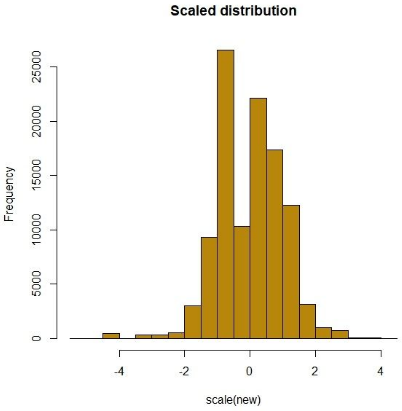
As before, I did not apply any transformation for the variable Tenure, since mean and median are so close and and this canu many observations assume value 12, so that there is no transformation that could be useful to get a symmetric distribution for this variable.



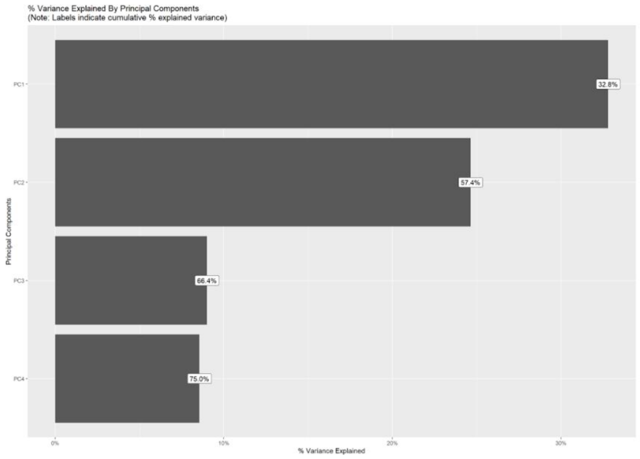
As before, thanks to the transformations applied, I obtained more symmetrical distributions, so now I can start my analysis.

Principal component analysis

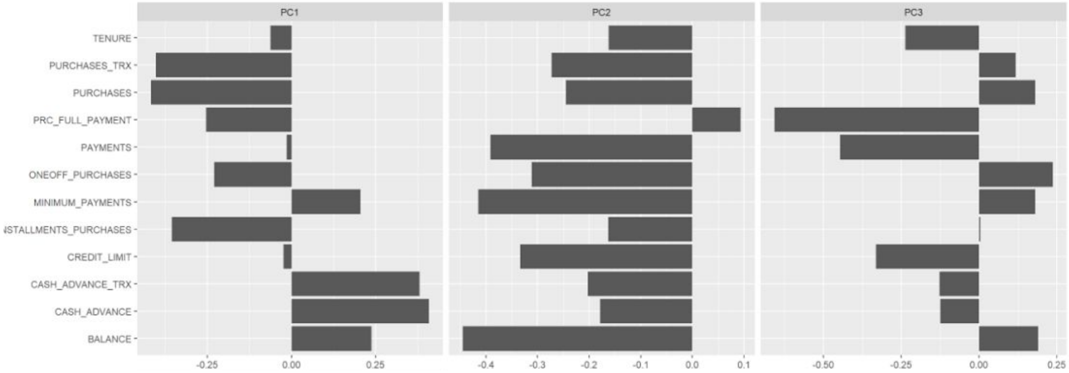
Even in this case before starting the analysis I scale my data, obtaining a more symmetric distribution.



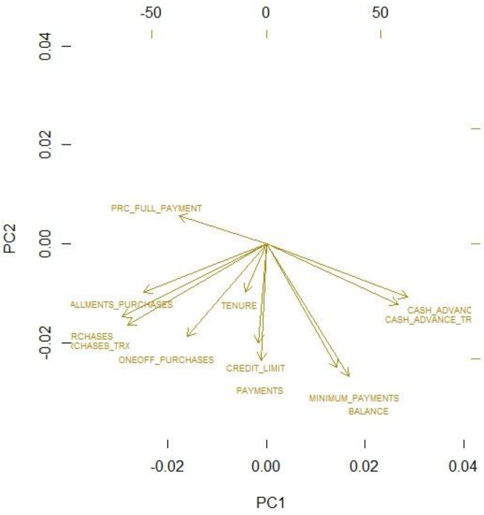
I applied the command prcomp() and as before the first two principal components refer respectively to purchases and balance. If in the first analysis (the one with the frequences), the first component was “purchases TRX” in this case the first principal component is “purchases”, but this does not bring relevant changes to our analysis since as we have seen before “purchases TRX” lead to the same results (they both gives information about purchases). In this case the third component changes (from one-off purchase to the % of full payments), but as seen in the previous analysis the contribution of the third principal component is not so relevant as the one of the first two.

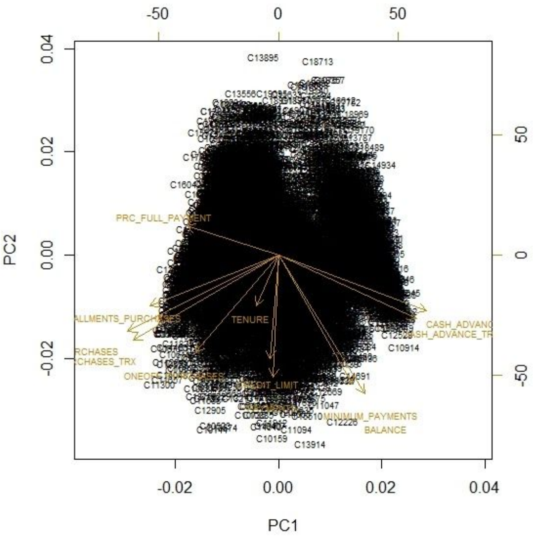


The percentage of variance explained by the first 2 - 3 principal components (57.4% - 66.4%), is more or less the same as in the previous analysis, and even better, since they explain a little bit more than before. In this case the first principal component explains the 32.8% of the total variance, whereas the second variable explains the 24.6%, and the third PC explains the 9% of the total variance.



The results are confirmed by the biplot, because the vectors of the variables Purchases and Balance create an angle of 90° (they are orthogonal), so that they are the principal components.

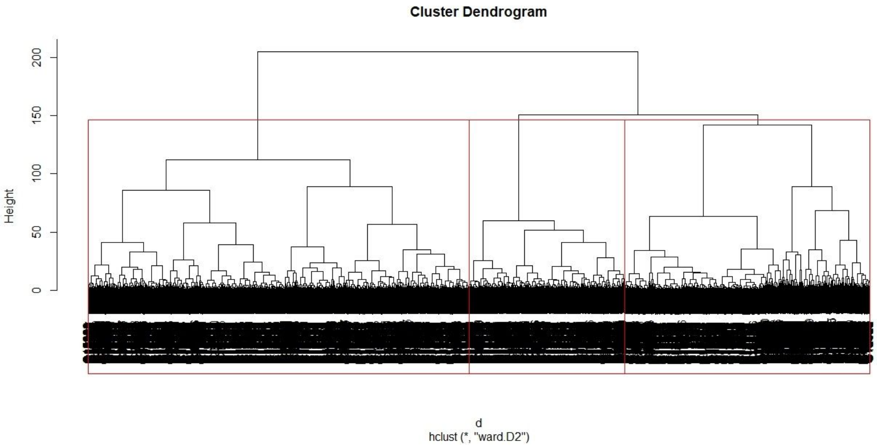




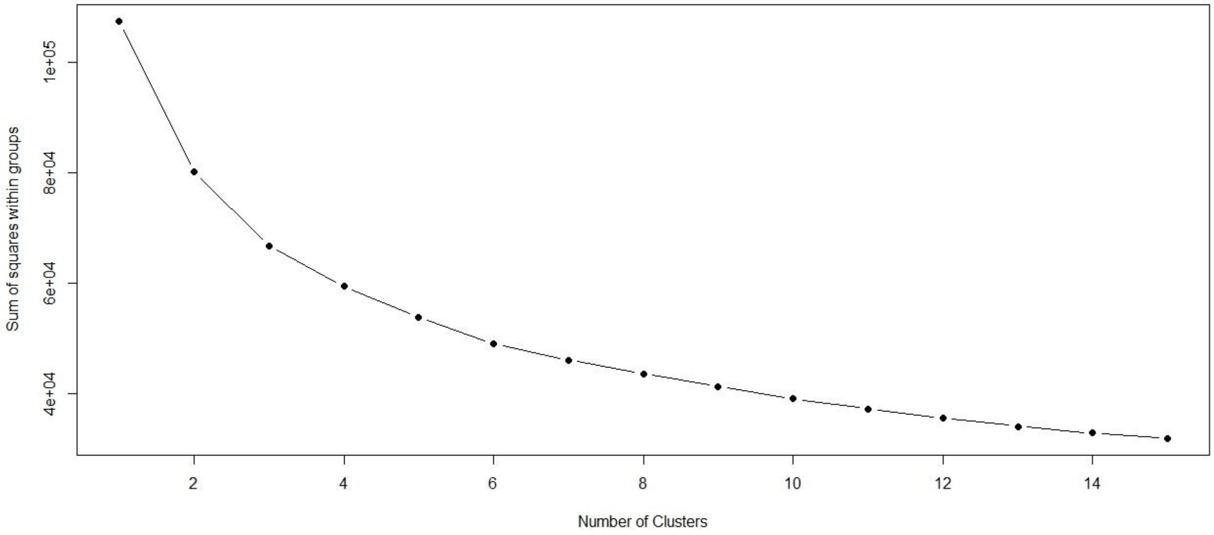
Cluster analysis

Hierarchical analysis

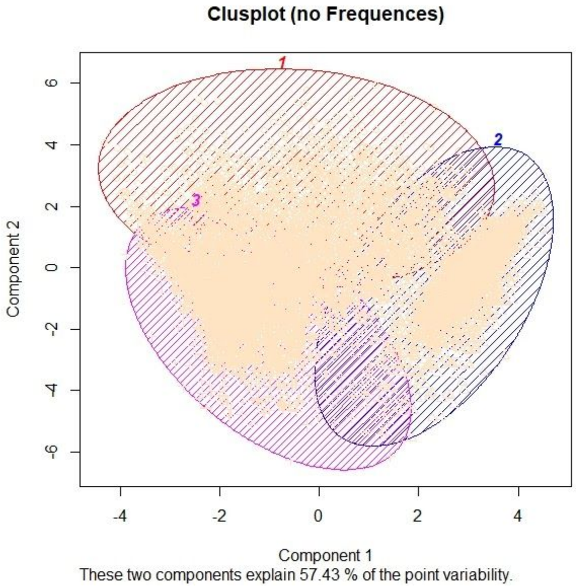
As before, in order to choose the right method to apply to hclust() we use the cophenetic distance, and we calculate the correlation with the euclidean distance. Even in this case the method with the high correlation is the "average" method, but it presents chaining problems. So the second method suggested is the "ward" method, that provides a clear representation of the clusters, and even in this case I decided to cluster in 3 groups.



K-Mean



As in the previous analysis, the suggested number of clusters (from the elbow chart) is three, so that 3 is the number of clusters from which we start with the k-mean analysis.



As we can see from the graph above the results do not present relevant changes from what obtained in the previous analysis, moreover in this case the first two components explain a higher proportion of the variance (57.43% vs 50.89%).

Conclusions

As before, now it is time to understand the characteristics of the different groups. The main difference with the previous analysis lies within the numbers of variables, in particular in the first group there are 4359 individuals (vs 3405), in the second 2808 (vs 2294) and in the third there are 1783 (vs 3251).

Even if the number of individuals for each group changes, when it comes to the variable “Balance” the three groups maintain the same characteristics obtained before: the group 1 present the lowest values, the second group presents intermediate values, whereas the third group presents the highest values.

The same happens for the variable “Purchases”: the second group still is the one that presents the lowest values, whereas the individuals that belong to the groups 1 and 3 are the ones that make the highest purchases. Once again the results are confirmed by the variable “one-off purchases” and “purchases TRX”.

When it comes to the “installments payments”, as before, the individuals of the first group tend to pay in installment more than what the individuals of the other two groups do. The individuals of the second group are the ones that use less of the installment payment.

For the variable “Cash advance” a slight difference appears. As before, the individuals of the first group tend not to pay in advance (as they mostly pay in installment), whereas the individuals of the second group tend to pay in advance (as they do not tend to pay in installment). The difference appears when it comes to the third group: in the previous analysis they had intermediate values for this variable, whereas in this case they have the highest values. And these results are confirmed by the variable “Cash advance TRX”.

Also referring to the variable “Credit limit” there are some changes. The third group in this case is the one that has the highest values, and this is probably due to the fact that they tend to pay more in advance than before. Whereas for the other two groups we have we have more or less the same results.

When it comes to the variables “Payments” (number of payments) and “Minimum payment” (lower payment done), the third group is the one with highest values, whereas the first and the second group have more or less the same values, with a slight difference that consist in the fact that the individuals in the first group make more payments but with a lower values, whereas the individuals of the second group tend to make a number lower payments but with a higher minimum value.

The results for the percentage of full payments do not present changes from the previous analysis.

Even in this case, the variable Tenure does not provide any relevant information to our analysis, since the values for the three groups are more or less the same.

In conclusion, even if there are small changes, I could say that the two analyses lead to similar conclusions:

- The ***first group*** is the one that has the lower amount available for purchases (Balance values), although they tend to do many purchases that are paid mostly in installments (the cash advance is so rare for these individuals). The institution that provides the credit should be aware of the fact that these are the most "risky" customers, since they tend to spend a lot of money even if their balances are not that high as for the other individuals.
- The individuals of the ***second group*** have intermediate values for the variable balance, but they do not seem to spend more than what they have. They are also on time when it comes to pay their purchases, since they seem to prefer paying in advance than paying in installments. The credit institutions should rely on this type of customers, providing them advantageous conditions and gaining their loyalty.
- The individuals of the ***third group*** are the "richest", because they have the higher values for the variable "balance", so that they can afford making a high number of purchases, regardless of the form of payment. Even in this case the credit institution should rely on these customers but keeping under control their purchases (in the case that they exceed the value of the amount available).