Clustering project

Rafał Rysiejko, Rezart Abbazi

19/10/2019

As for our project for the first block of the Unsupervised Learning classes, we decided to perform cluster analysis on a customer segmentation problem. For that we used dataset containing data about the usage behaviour of about 9000 active credit card holders, available at Kaggle.

For the purposes of our project we also repurposed code from classes as well as taken some inspiration and suggestions in terms of techniques used, from similar projects based on this dataset.

Installing and running the libraries:

```
requiredPackages = c("tidyverse","factoextra","stats","clustertend","flexclust","ggforce"
                     ,"fpc","cluster","ClusterR","knitr","kableExtra","DataExplorer","reshape2",
                     "mclust","dbscan")
for(i in requiredPackages){if(!require(i,character.only = TRUE)) install.packages(i)}
for(i in requiredPackages){library(i,character.only = TRUE) }
```

Loading the data:

```
data_full <- read.csv("Dataset/CC GENERAL.csv",
                      stringsAsFactors = F)
```

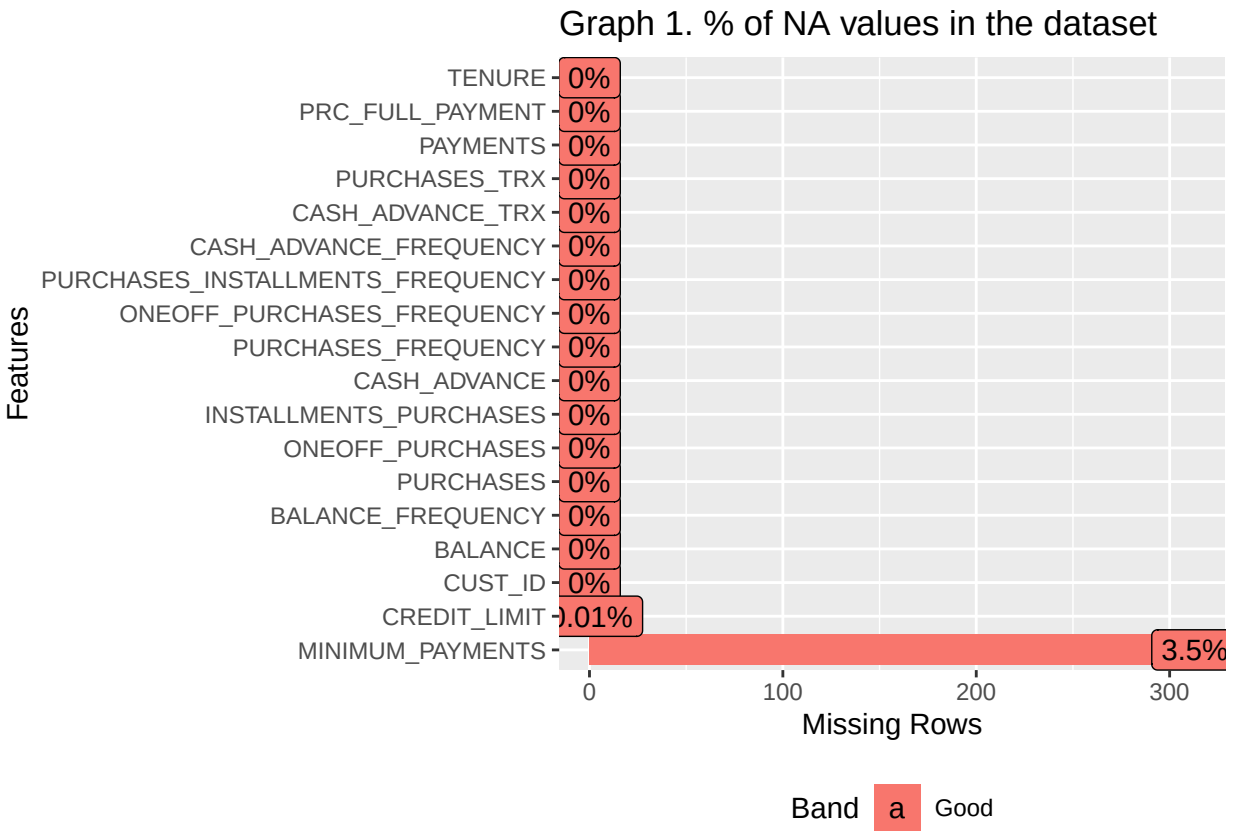
The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables.
1. CUST_ID : Identification of Credit Card holder (Categorical)
2. BALANCE : Balance amount left in their account to make purchases
3. BALANCE_FREQUENCY : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
4. PURCHASES : Amount of purchases made from account
5. ONEOFF_PURCHASES : Maximum purchase amount done in one-go
6. INSTALLMENTS_PURCHASES : Amount of purchase done in installment
7. CASH_ADVANCE : Cash in advance given by the user
8. PURCHASES_FREQUENCY : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
9. ONEOFFPURCHASESFREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
10. PURCHASESINSTALLMENTSFREQUENCY : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
11. CASHADVANCEFREQUENCY : How frequently the cash in advance being paid
12. CASHADVANCETRX : Number of Transactions made with "Cash in Advanced"
13. PURCHASES_TRX : Number of purchase transactions made
14. CREDIT_LIMIT : Limit of Credit Card for user
15. PAYMENTS : Amount of Payment done by user
16. MINIMUM_PAYMENTS : Minimum amount of payments made by user
17. PRCFULLPAYMENT : Percent of full payment paid by user
18. TENURE : Tenure of credit card service for user

Descriptive statistics for the dataset:

To better investigate the possible problem of missing observation, we visualize them on the graph number 1.

Table 1: Summary statistics of the dataset

CUST_ID	Length:8950	Class :character	Mode :character	NA	NA	NA	NA
BALANCE	Min. : 0.0	1st Qu.: 128.3	Median : 873.4	Mean : 1564.5	3rd Qu.: 2054.1	Max. :19043.1	NA
BALANCE_FREQUENCY	Min. :0.0000	1st Qu.:0.8889	Median :1.0000	Mean :0.8773	3rd Qu.:1.0000	Max. :1.0000	NA
PURCHASES	Min. : 0.00	1st Qu.: 39.63	Median : 361.28	Mean : 1003.20	3rd Qu.: 1110.13	Max. :49039.57	NA
ONEOFF_PURCHASES	Min. : 0.0	1st Qu.: 0.0	Median : 38.0	Mean : 592.4	3rd Qu.: 577.4	Max. :40761.2	NA
INSTALLMENTS_PURCHASES	Min. : 0.0	1st Qu.: 0.0	Median : 89.0	Mean : 411.1	3rd Qu.: 468.6	Max. :22500.0	NA
CASH_ADVANCE	Min. : 0.0	1st Qu.: 0.0	Median : 0.0	Mean : 978.9	3rd Qu.: 1113.8	Max. :47137.2	NA
PURCHASES_FREQUENCY	Min. :0.00000	1st Qu.:0.08333	Median :0.50000	Mean :0.49035	3rd Qu.:0.91667	Max. :1.00000	NA
ONEOFF_PURCHASES_FREQUENCY	Min. :0.00000	1st Qu.:0.00000	Median :0.08333	Mean :0.20246	3rd Qu.:0.30000	Max. :1.00000	NA
PURCHASES_INSTALLMENTS_FREQUENCY	Min. :0.0000	1st Qu.:0.0000	Median :0.1667	Mean :0.3644	3rd Qu.:0.7500	Max. :1.0000	NA
CASH_ADVANCE_FREQUENCY	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.1351	3rd Qu.:0.2222	Max. :1.5000	NA
CASH_ADVANCE_TRX	Min. : 0.000	1st Qu.: 0.000	Median : 0.000	Mean : 3.249	3rd Qu.: 4.000	Max. :123.000	NA
PURCHASES_TRX	Min. : 0.00	1st Qu.: 1.00	Median : 7.00	Mean : 14.71	3rd Qu.: 17.00	Max. :358.00	NA
CREDIT_LIMIT	Min. : 50	1st Qu.: 1600	Median : 3000	Mean : 4494	3rd Qu.: 6500	Max. :30000	NA's :1
PAYMENTS	Min. : 0.0	1st Qu.: 383.3	Median : 856.9	Mean : 1733.1	3rd Qu.: 1901.1	Max. :50721.5	NA
MINIMUM_PAYMENTS	Min. : 0.02	1st Qu.: 169.12	Median : 312.34	Mean : 864.21	3rd Qu.: 825.49	Max. :76406.21	NA's :313
PRC_FULL_PAYMENT	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.1537	3rd Qu.:0.1429	Max. :1.0000	NA
TENURE	Min. : 6.00	1st Qu.:12.00	Median :12.00	Mean :11.52	3rd Qu.:12.00	Max. :12.00	NA



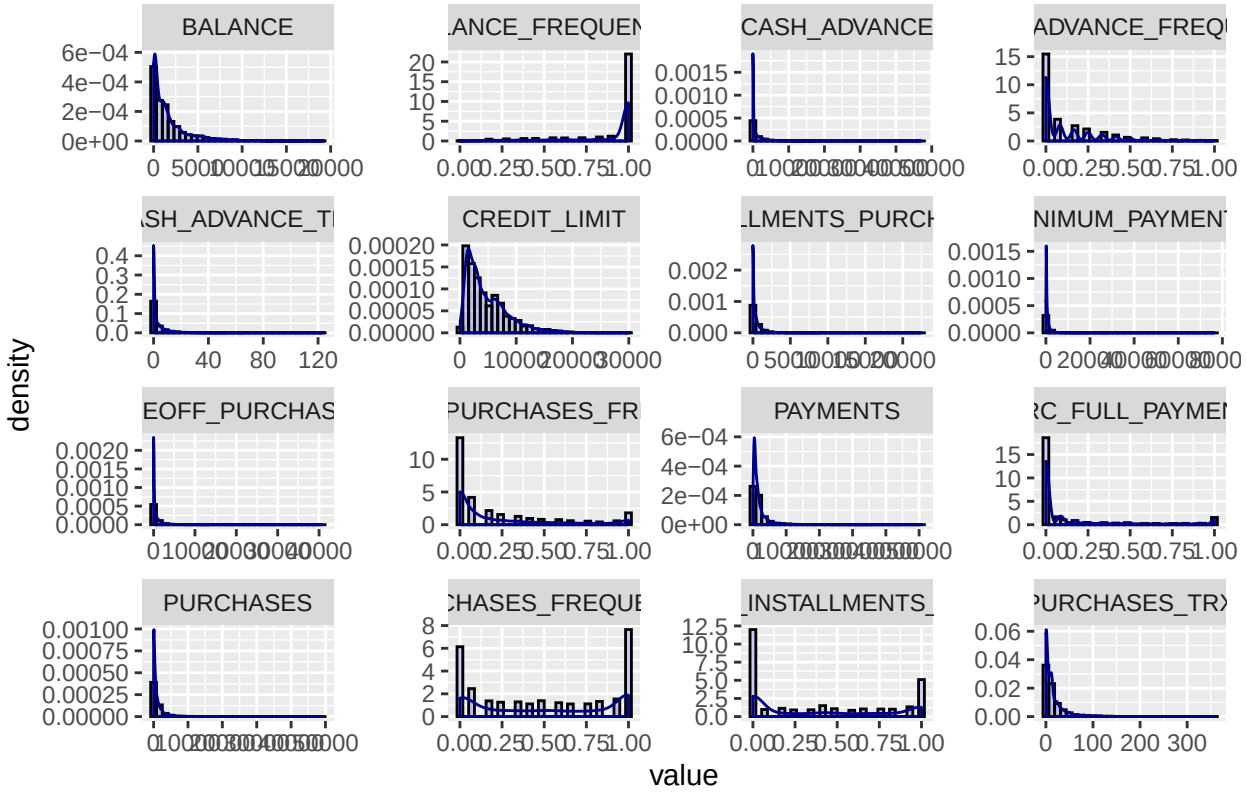
As the number of missing observations is fairly small compared to the dataset volume. During the data transformation we droped them.

`TENURE` variable has 1366 values that are not `12`. We are going to remove those rows for customers with `TENURE` < 12 so that all data are based on 1 year's worth of customer behavior. Then remove this variable from the dataset. We also dropped `CUST_ID` variable as it is of no use it this particular problem.

```
data = data_full %>%
  select(-CUST_ID) %>%
  drop_na() %>%
  filter(., TENURE==12) %>%
  select(-TENURE)
```

We then created histograms to analyze the distribution of the variables.

Graph 2. Histograms of variables

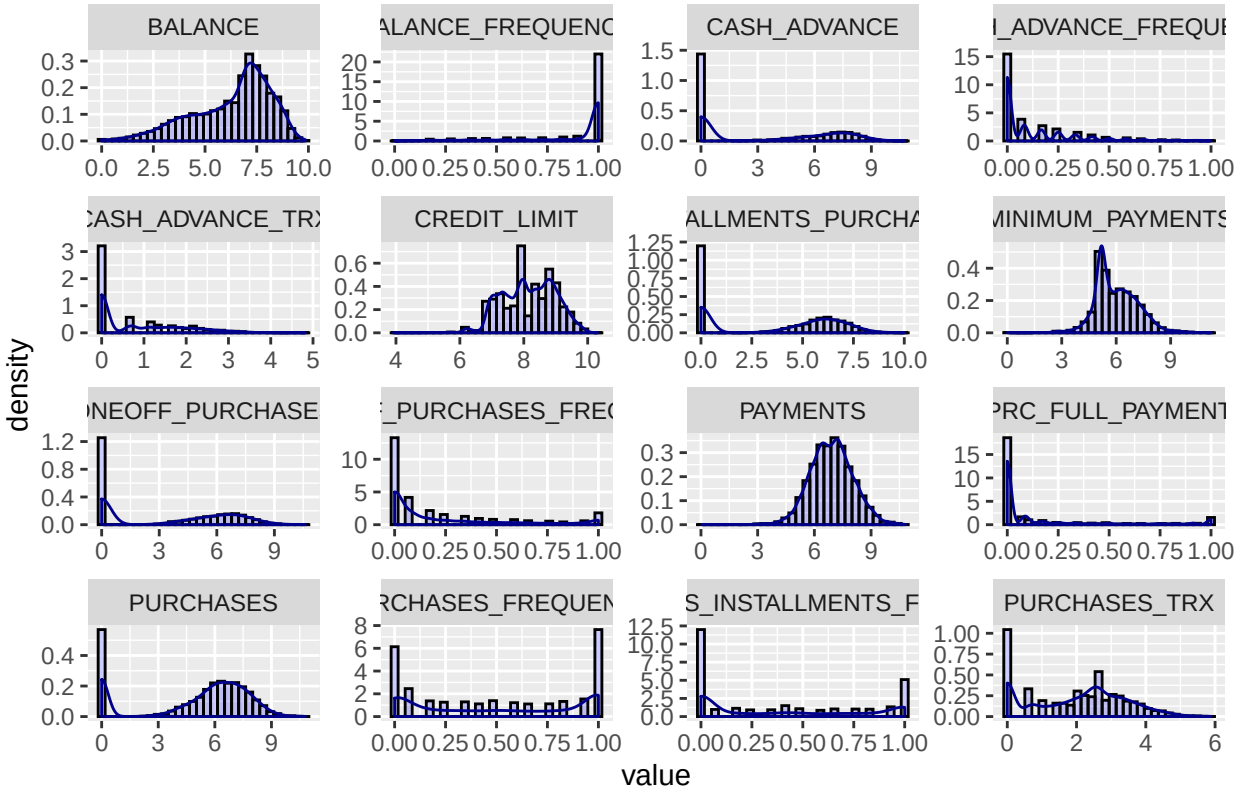


Variables: BALANCE, CASH_ADVANCE,CASH_ADVANCE_TRX, CREDIT_LIMIT, INSTALLMENTS_PURCHASES, MINIMUM_PAYMENTS, ONEOFF_PURCHASES, PAYMENTS,PURCHASES, PURCHASES_TRX are right - skewed (Positive Skewnees). We are using a log transformation to reduce the negative impact of skewnees on the model performance.

```
transformed_var <- c("BALANCE", "CASH_ADVANCE","CASH_ADVANCE_TRX", "CREDIT_LIMIT",
                     "INSTALLMENTS_PURCHASES", "MINIMUM_PAYMENTS", "ONEOFF_PURCHASES",
                     "PAYMENTS","PURCHASES", "PURCHASES_TRX")
data <- data %>% mutate_at(vars(transformed_var), funs(log(1 + .)))
```

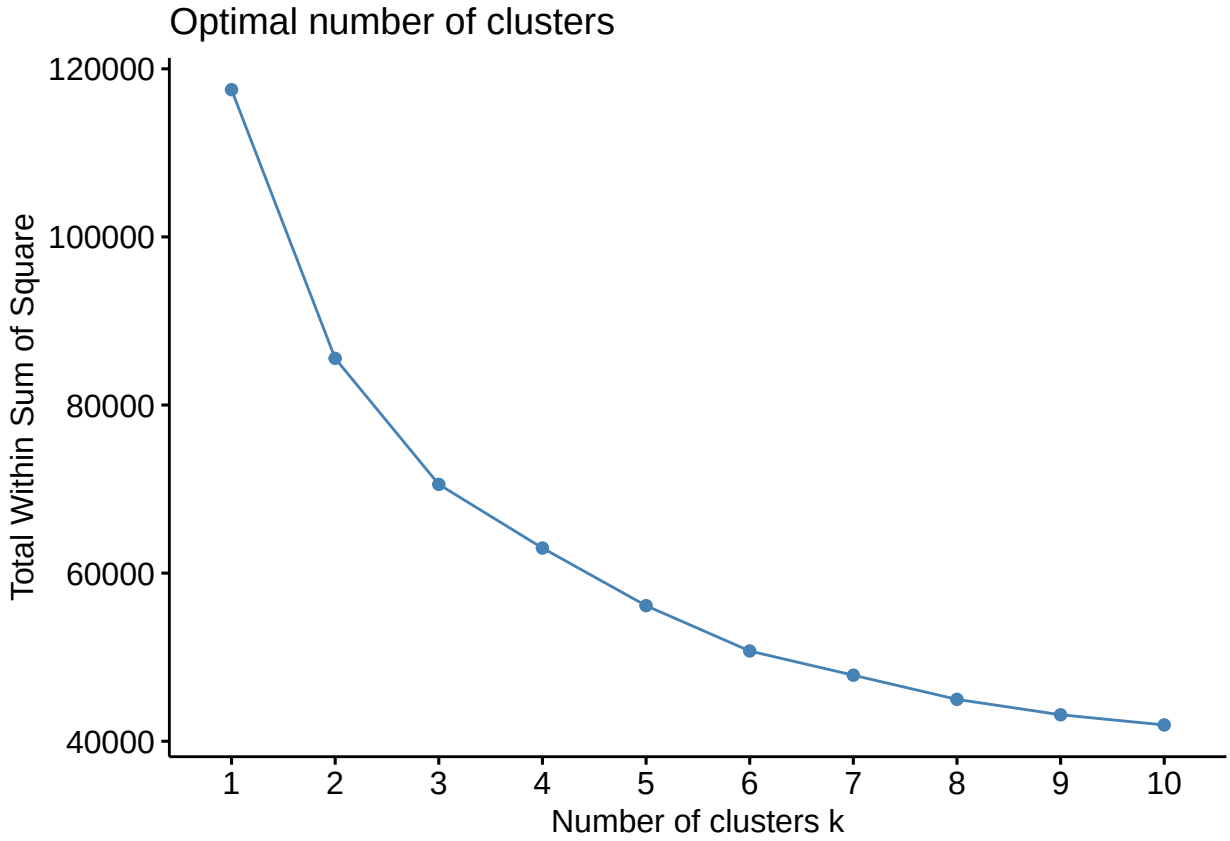
Histograms after the transformation.

Graph 3. Histograms of variables after log transformation



In the pre-diagnostic phase of the research we had to determine the optimal number of clusters using K-means method: where arbitrary data is chosen to be the centroids of this clusters. Once all the element of our dataset will be assigned to a cluster recalculate the positions of the centroids and reassign each data item to the closest cluster based on the mean value of the items on the cluster. This makes K-Means clustering algorithm very sensitive to outliers and noise, thereby reducing its performance too. K-means is also does not work quite well in discovering clusters that have non-convex shapes or very different size.

Graph 4.



From the output we can notice that the total within sum of squares is decreasing rapidly till the fourth cluster as so this would be the optimal number for us. This means that 4 might be a good number of centroids for clustering on this data.

After assesing the optimal number of clusters we performed the Duda-Hart test for whether a data set should be split into two clusters for kmeans class with hypothesis that state:
H0: homogeneity of cluster (data within cluster as similar)
H1: heterogeneity of cluster (one can easily split the cluster)

```
test1 <- kmeans(data,4)
dudahart2(data,test1$cluster)
```

```
## $p.value
## [1] 0
##
## $dh
## [1] 0.1877885
##
## $compare
## [1] 0.947791
##
## $cluster1
## [1] FALSE
##
## $alpha
## [1] 0.001
##
```

```
## $z
## [1] 3.090232
```

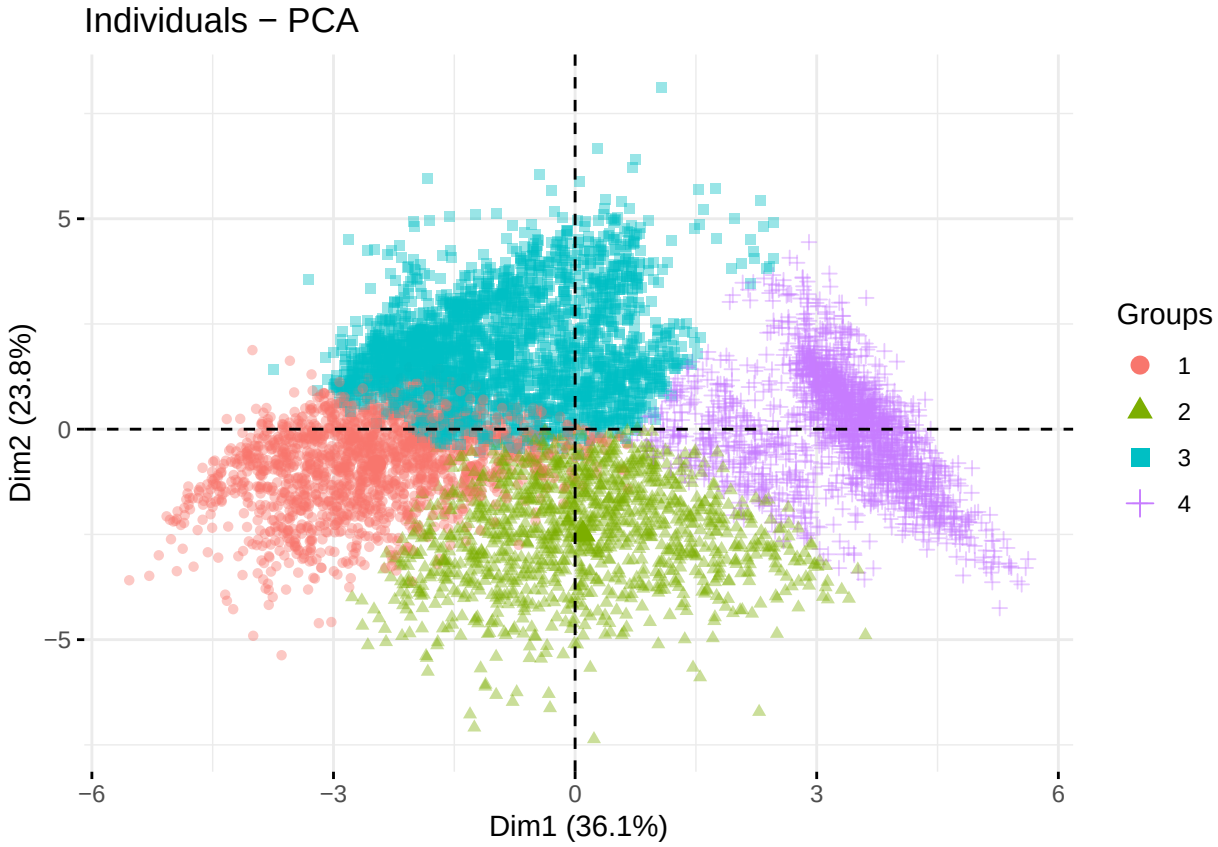
As the *cluster1=FALSE* we have basis to reject *H0* of homogeneity in favour of accepting *H1*. The dataset is not homgenic and should be splited into two clusters.

Next we perfomed an actual clusterin using k - means clustering an scaled dataset.

```
km_fitted <- kmeans(scale(data), centers = 4)
```

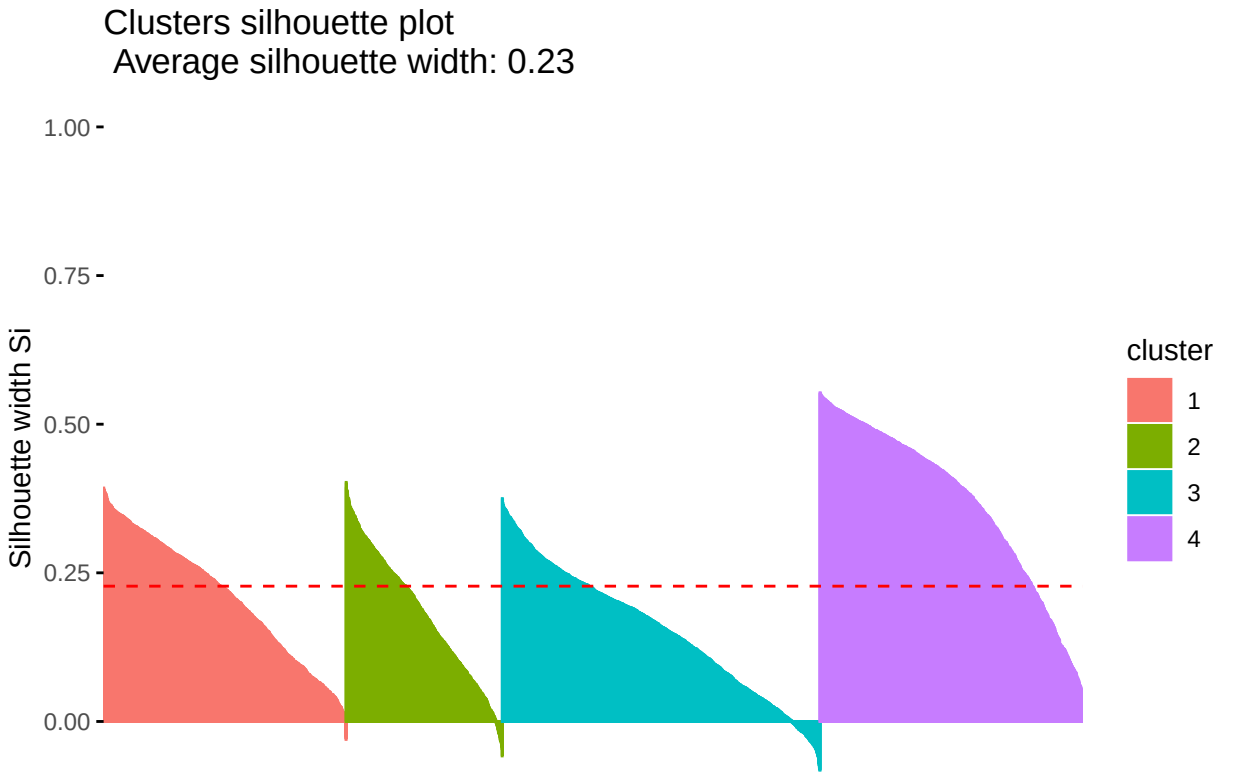
We then visualized the performed clustering by applying Principal component analysis (PCA), which reduces the dimensionality of multivariate data, to in this case two that can be visualized graphically with minimal loss of information.

Graph 5.



As a next step we used silhouette coefficient method to determine the quiality of the clustering, which combines both cohesion and separation. The value of the silhouette coefficient can vary between -1 and 1. A negative value is undesirable because this corresponds to a case in which ai, the average distance to points in the cluster. We want the silhouette coefficient to be positive ($a_i < b_i$) and for ai to be as close to 0 as possible.

Graph 6.



In our cluster silhouette plot the average silhouette width is 0.23. The first group (red) has the lowest number of values under the x-axes means negative numbers. The same goes for the second group (green) where the range is smaller and the number of negative elements is insignificant. The third group (blue) has a similar distribution with the first group, but the only difference is that no negative elements are more evident and in the longer range. The fourth group (purple) is the group with the highest number of elements where the limit elements go over 0.50. What we can notice in this group is also the absence of negative elements. We can notice from this clustering that first, the third and fourth group have a significant number of values above the silhouette width coefficient, where in the first group are close to 0,5.

Last but not least we combined the dataset with the information. This allowed us to investigate the descriptive statistics of particular clusters and try to characterize the cluster members.

Table 2: Mean

cluster	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT
1	6.61	0.98	7.52	6.62	5.38	6.51	0.96	0.37	0.59	0.01	0.08	3.34	8.32	7.51	5.99	0.26
2	7.66	0.96	6.82	6.05	4.66	7.28	0.71	0.30	0.53	0.32	1.64	2.85	8.45	7.59	6.92	0.01
3	8.55	0.78	5.69	2.13	4.39	6.52	0.55	0.07	0.47	0.01	0.05	2.05	7.83	6.15	5.19	0.24
4	7.26	0.94	1.02	0.93	0.13	6.67	0.00	0.02	0.00	0.36	1.66	0.32	8.03	6.87	6.38	0.03

Table 3: Median

cluster	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT
1	6.74	1	7.54	6.67	6.35	0.00	1.00	0.58	0.67	0.00	0.00	3.33	8.70	7.48	5.62	0
2	7.88	1	6.83	6.04	5.82	7.42	0.75	0.25	0.58	0.25	1.95	2.77	8.52	7.51	6.92	0
3	8.39	1	5.83	0.69	5.35	0.00	0.50	0.00	0.42	0.00	0.00	2.20	7.82	6.17	5.13	0
4	7.37	1	0.00	0.00	0.00	7.01	0.00	0.00	0.00	0.25	1.61	0.00	8.01	6.77	6.38	0

Table 4: Standard Deviation

cluster	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT
1	1.36	0.07	0.93	1.82	2.77	1.62	0.21	0.33	0.39	0.03	0.26	0.78	0.72	0.87	1.02	0.37
2	0.93	0.07	0.99	2.99	2.56	1.37	0.28	0.29	0.37	0.21	0.78	0.95	0.75	0.94	0.93	0.11
3	1.79	0.26	1.15	2.71	2.56	1.39	0.34	0.12	0.38	0.03	0.23	0.79	0.74	0.95	1.02	0.34
4	1.15	0.17	1.99	1.91	0.75	1.63	0.06	0.05	0.03	0.10	0.78	0.45	0.80	1.06	1.03	0.11

Observing the three tables, respectively, the mean, median and standard deviation we can notice the following:
* On the 'Mean' table we can notice clusters with the similarities about the balance, balance frequency, credit limit and payments but significant differences on purchases, where the third group has lower mean value, compared to the others. The same we can confirm about instalments purchases and purchases trx.
* On the 'Median' table, we can confirm by taking a look at the values the similarity with the mean in the four clusters. Here aswell we similarities between clusters on payments, credit limit, minimum payments and very significant differences in purchase frequency, purchases and I the most interesting is cash advance where individuals of cluster one and four have no elements.
* On the 'Standard Deviation', we can notice that what we mentioned above about the other two tables (mean, median) is not very accurate for this table, and here we can see why: If in the other two examples, we saw differences in purchases here the values are very similar and the same we can say about cash advance and instalments purchases.

Based on this statistical value found on these three tables, we can assume:
1. In the first cluster we have frequent user from the table purchase frequency (0.86), with (probably) lower-income that spends his money mostly on consumer goods.
2. In the second cluster we have frequent user, with (probably) higher income that spends his money mostly on consumer goods.
3. In the third cluster we have users, with (probably) higher income than average, which spends his money more for higher-priced products with longterm use we can notice the one-off purchase is 0.90.
4. In the fourth cluster we have users with a low frequency of usage, with (probably) mid to low income which spends his money more on consumer goods of basic needs.