Analysis_1_CC

2023-05-04

This was my final paper for Matrix Methods and Cluster Analysis subject, in my Specialization (MBA Executivo) in Big Data and Business Analysis. It was used a public database, available at Kaggle. This database provides data on credit card users and their financial behavior through 18 variables, with the objective of creating a customer segmentation based on approximate profiles. It was used the R Markdown setup, developing the analysis in HTML file.

```
#Used packages
#install.packages("tidyverse")
#install.packages("corrplot")
#install.packages("ggplot2")
#install.packages("gridExtra")
#install.packages("rpart")
#install.packages("factoextra")
#rm(list=ls())
#install.packages(readxl)

library(tidyverse)
library(corrplot)
library (ggplot2)
library(gridExtra)
library(rpart)
library(factoextra)
library(readxl)

setwd("C:\\Users\\Marina\\Documents\\Desafio_files\\CC CLuster")
base_original <- read.table("Base.csv", sep=";", header=T)
base <- base_original
head(base)
```

```
##   CUST_ID    BALANCE BALANCE_FREQUENCY PURCHASES ONEOFF_PURCHASES
## 1  C10001   40900749           0.81818     95.40             0.00
## 2  C10002 3202467416           0.90909      0.00             0.00
## 3  C10003 2495148862           1.00000    773.17           773.17
## 4  C10004 1666670542           0.63636   1499.00          1499.00
## 5  C10005  817714335           1.00000     16.00            16.00
## 6  C10006 1809828751           1.00000   1333.28             0.00
##   INSTALLMENTS_PURCHASES CASH_ADVANCE PURCHASES_FREQUENCY
## 1                  95.40            0            0.166667
## 2                   0.00   6442945483            0.000000
## 3                   0.00            0            1.000000
## 4                   0.00    205788017            0.083333
## 5                   0.00            0            0.083333
## 6                1333.28            0            0.666667
##   ONEOFF_PURCHASES_FREQUENCY PURCHASES_INSTALLMENTS_FREQUENCY
```

```
## 1                   0.000000                          0.083333
## 2                   0.000000                          0.000000
## 3                   1.000000                          0.000000
## 4                   0.083333                          0.000000
## 5                   0.083333                          0.000000
## 6                   0.000000                          0.583333
##   CASH_ADVANCE_FREQUENCY CASH_ADVANCE_TRX PURCHASES_TRX CREDIT_LIMIT   PAYMENTS
## 1               0.000000                0             2         1000  201802084
## 2               0.250000                4             0         7000 4103032597
## 3               0.000000                0            12         7500  622066742
## 4               0.083333                1             1         7500          0
## 5               0.000000                0             1         1200  678334763
## 6               0.000000                0             8         1800  140005777
##   MINIMUM_PAYMENTS PRC_FULL_PAYMENT TENURE
## 1        139509787         0.000000     12
## 2       1072340217         0.222222     12
## 3        627284787         0.000000     12
## 4               NA         0.000000     12
## 5        244791237         0.000000     12
## 6       2407246035         0.000000     12
```

We will start by analyzing the number of transactions in the base.

```
nrow(base)
```

```
## [1] 8950
```

The number of samples is significant, so we can continue with the base without needing to get more data.

```
str(base)
```

```
## 'data.frame':    8950 obs. of  18 variables:
##  $ CUST_ID                          : chr  "C10001" "C10002" "C10003" "C10004" ...
##  $ BALANCE                          : num  4.09e+07 3.20e+09 2.50e+09 1.67e+09 8.18e+08 ...
##  $ BALANCE_FREQUENCY                : num  0.818 0.909 1 0.636 1 ...
##  $ PURCHASES                        : num  95.4 0 773.2 1499 16 ...
##  $ ONEOFF_PURCHASES                 : num  0 0 773 1499 16 ...
##  $ INSTALLMENTS_PURCHASES           : num  95.4 0 0 0 0 ...
##  $ CASH_ADVANCE                     : num  0.00 6.44e+09 0.00 2.06e+08 0.00 ...
##  $ PURCHASES_FREQUENCY              : num  0.1667 0 1 0.0833 0.0833 ...
##  $ ONEOFF_PURCHASES_FREQUENCY       : num  0 0 1 0.0833 0.0833 ...
##  $ PURCHASES_INSTALLMENTS_FREQUENCY: num  0.0833 0 0 0 0 ...
##  $ CASH_ADVANCE_FREQUENCY           : num  0 0.25 0 0.0833 0 ...
##  $ CASH_ADVANCE_TRX                 : int  0 4 0 1 0 0 0 0 0 0 ...
##  $ PURCHASES_TRX                    : int  2 0 12 1 1 8 64 12 5 3 ...
##  $ CREDIT_LIMIT                     : num  1000 7000 7500 7500 1200 1800 13500 2300 7000 11000 ...
##  $ PAYMENTS                         : num  2.02e+08 4.10e+09 6.22e+08 0.00 6.78e+08 ...
##  $ MINIMUM_PAYMENTS                 : num  1.40e+08 1.07e+09 6.27e+08 NA 2.45e+08 ...
##  $ PRC_FULL_PAYMENT                 : num  0 0.222 0 0 0 ...
##  $ TENURE                           : int  12 12 12 12 12 12 12 12 12 12 ...
```

All our variables are numeric, except the column **CUST_ID**, which identifies the user. As this information is not relevant for data clustering, we will remove this column, and keep it in case we need to use it later.

We can also already see that there are two columns (**CREDIT_LIMIT** and **MINIMUM_PAYMENTS**) with NAs. As the number of responses with NA is not significant (about 3.5%), we will remove these lines through na.omit.

```
base <- na.omit(base)

base_cust_id <- base %>% select(CUST_ID)
base <- base %>% select(-CUST_ID)
```

Now, let's analyze if there are any duplicate users. For this, we will use the highlighted column **CUST_ID**.

```
n_distintos <- count(distinct(base_cust_id %>% select(CUST_ID)))
n_linhas <- nrow(base)
n <- c(n_distintos, n_linhas)
```

```
df <- data.frame(n)
colnames (df) <- c("Distinct Users", "Number of lines in base")
df
```

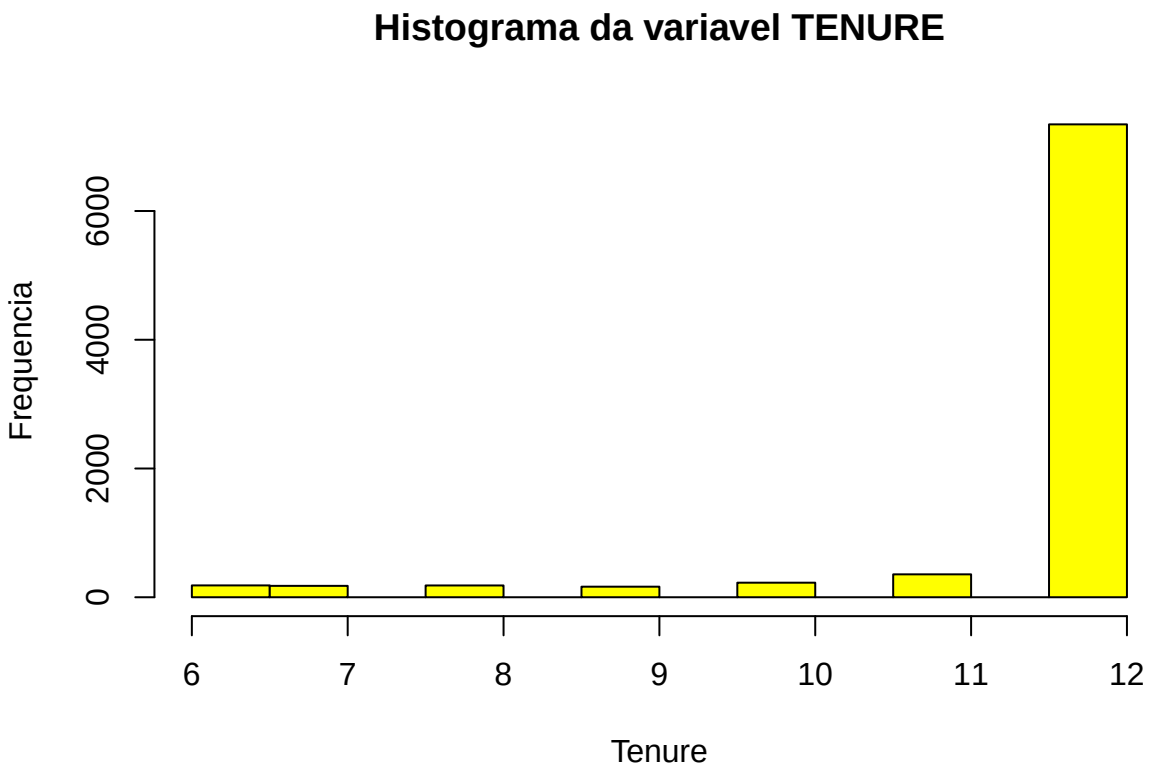
```
##   Distinct Users Number of lines in base
## 1           8636                   8636
```

```
rm(df)
rm(n)
rm(n_distintos)
rm(n_linhas)
```

Therefore, we conclude that there are no repeated transactions.

Next, we will analyze the variable **Tenure**, which brings information about how long the user has had the credit card. We will plot the variable's histogram for this.

```
hist(base$TENURE, main = "Histograma da variavel TENURE", xlab = "Tenure", ylab = "Frequencia",col = c(
```



We can see from the histogram that a part of the data has **TENURE** < 12. The amount has been filtered and counted below:

```
count(base %>% filter(base$TENURE < 12)) %>%
  as.data.frame()
```

```
##      n
## 1 1290
```

This corresponds to 15% of the lines. Thus, we will remove these cases and then remove the variable **Tenure**. We do this with the aim of shrinking the database, to facilitate mining.

```
base_tenure <- base %>% select(TENURE)

base <- base %>%
  filter(TENURE == 12)

base <- base %>% select(-TENURE)

head(base)
```

```
##      BALANCE BALANCE_FREQUENCY PURCHASES ONEOFF_PURCHASES
## 1   40900749           0.81818     95.40              0.00
## 2 3202467416           0.90909      0.00              0.00
## 3 2495148862           1.00000    773.17            773.17
```

```
## 4  817714335           1.00000     16.00            16.00
## 5 1809828751           1.00000   1333.28             0.00
## 6  627260806           1.00000   7091.01          6402.63
##   INSTALLMENTS_PURCHASES CASH_ADVANCE PURCHASES_FREQUENCY
## 1                  95.40            0            0.166667
## 2                   0.00   6442945483            0.000000
## 3                   0.00            0            1.000000
## 4                   0.00            0            0.083333
## 5                1333.28            0            0.666667
## 6                 688.38            0            1.000000
##   ONEOFF_PURCHASES_FREQUENCY PURCHASES_INSTALLMENTS_FREQUENCY
## 1                   0.000000                         0.083333
## 2                   0.000000                         0.000000
## 3                   1.000000                         0.000000
## 4                   0.083333                         0.000000
## 5                   0.000000                         0.583333
## 6                   1.000000                         1.000000
##   CASH_ADVANCE_FREQUENCY CASH_ADVANCE_TRX PURCHASES_TRX CREDIT_LIMIT   PAYMENTS
## 1                   0.00                0             2         1000  201802084
## 2                   0.25                4             0         7000 4103032597
## 3                   0.00                0            12         7500  622066742
## 4                   0.00                0             1         1200  678334763
## 5                   0.00                0             8         1800  140005777
## 6                   0.00                0            64        13500 6354314328
##   MINIMUM_PAYMENTS PRC_FULL_PAYMENT
## 1        139509787         0.000000
## 2       1072340217         0.222222
## 3        627284787         0.000000
## 4        244791237         0.000000
## 5       2407246035         0.000000
## 6        198065894         1.000000
```

The next step will be the analysis of the correlation matrix between the variables, so that we can assess how they relate to each other and if there is a need to continue with all columns. Collinearity >80% was considered. For space reasons, the matrix will not be shown here.

The collinearity matrix indicates that the variables below are correlated:

Purchase - **OneOffPurchase** (91,7%)

Purchase Frequency - **Purchase Installment Frequency** (85,7%)

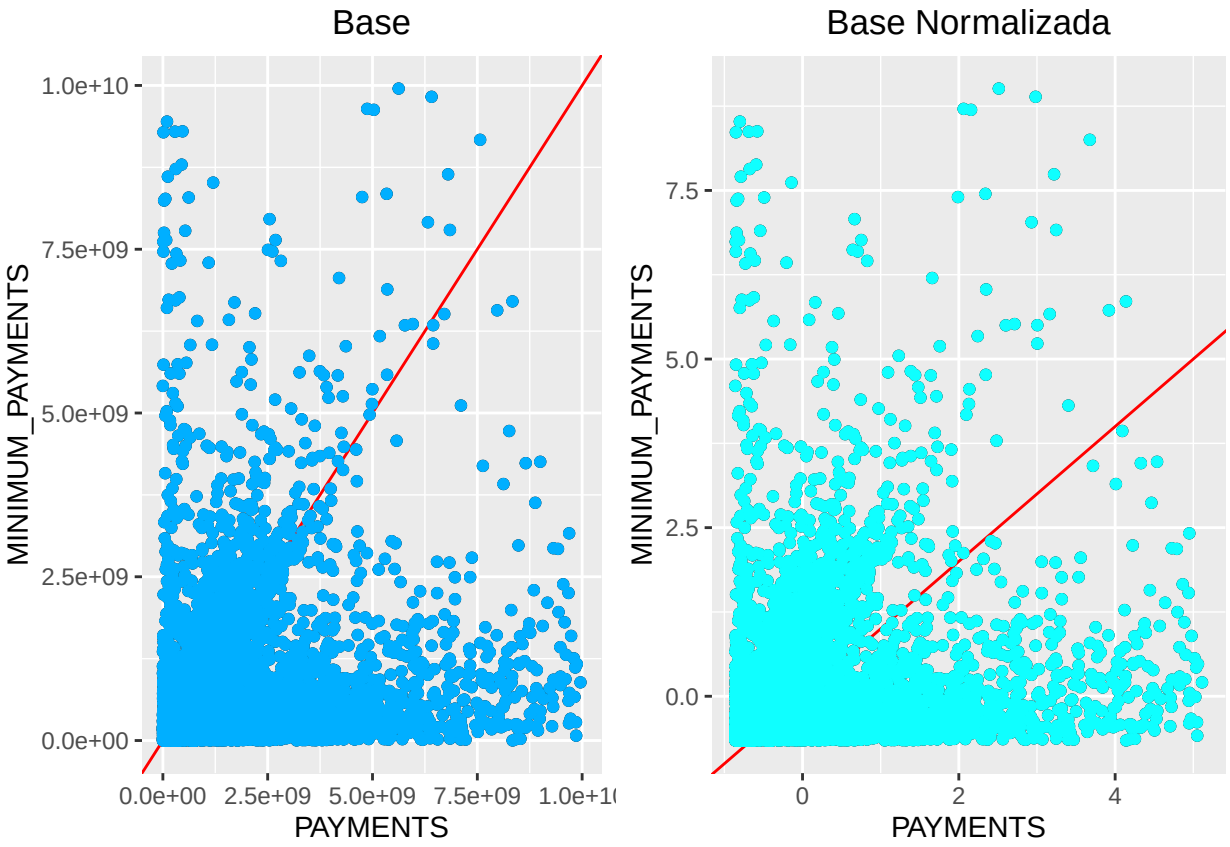
Cash Advcance Frequency - **Cash Advance Trx** (82,7%)

So, we will remove 3 correlated variables. The variables below were chosen because they have greater collinearity with the other variables. They are: **One Off Purchases**, **Purchases Frequency** and **Cash Advance Trx**. This way, we will have 13 columns.

```
base_oneoff_purchases <- base %>% select(ONEOFF_PURCHASES)
base_purchases_frequency <- base %>% select(PURCHASES_FREQUENCY)
base_cashadvance_trx <- base %>% select(CASH_ADVANCE_TRX)

base <- base %>% select(-ONEOFF_PURCHASES, -PURCHASES_FREQUENCY, -CASH_ADVANCE_TRX)
```

From now on, we will normalize the data so that there are no scaling problems.

```
base_normal <- as.data.frame(scale(base))

b_orig <- ggplot(base, aes(x=PAYMENTS, y=MINIMUM_PAYMENTS)) +
  geom_point() +
  labs(title="Base") +
  geom_abline(color = "red")+
  geom_point(color = "#00AFFF")+
  theme(plot.title = element_text(hjust = 0.5))


b_norm <- ggplot(base_normal, aes(x=PAYMENTS, y=MINIMUM_PAYMENTS)) +
  geom_point() +
  labs(title="Base Normalizada") +
  geom_abline(color = "red")+
  geom_point(color = "#0FFFFF")+
  theme(plot.title = element_text(hjust = 0.5))


grid.arrange(b_orig, b_norm, ncol=2)
```

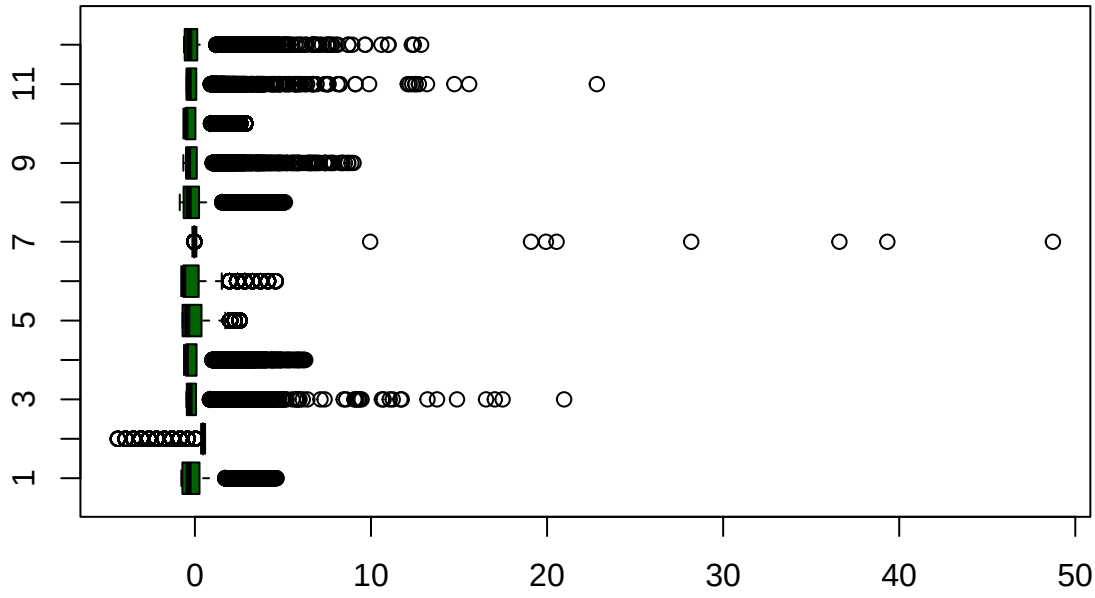



```
rm(b_orig)
rm(b_norm)
```

As we can see, the base was scaled down, with the variables in line. On the straight line, we can see a flattening of the scale of values, although there was no change in shape.

Now, with our base already quite standardized, we will be dealing with outliers. For this, we will start by plotting all dimensions, so that we can identify which ones have outliers. For reasons of space, we will only plot the dimensions that had outliers.

```
boxplot(base_normal$BALANCE, base_normal$BALANCE_FREQUENCY, base_normal$PURCHASES, base_normal$CASH_ADV
```

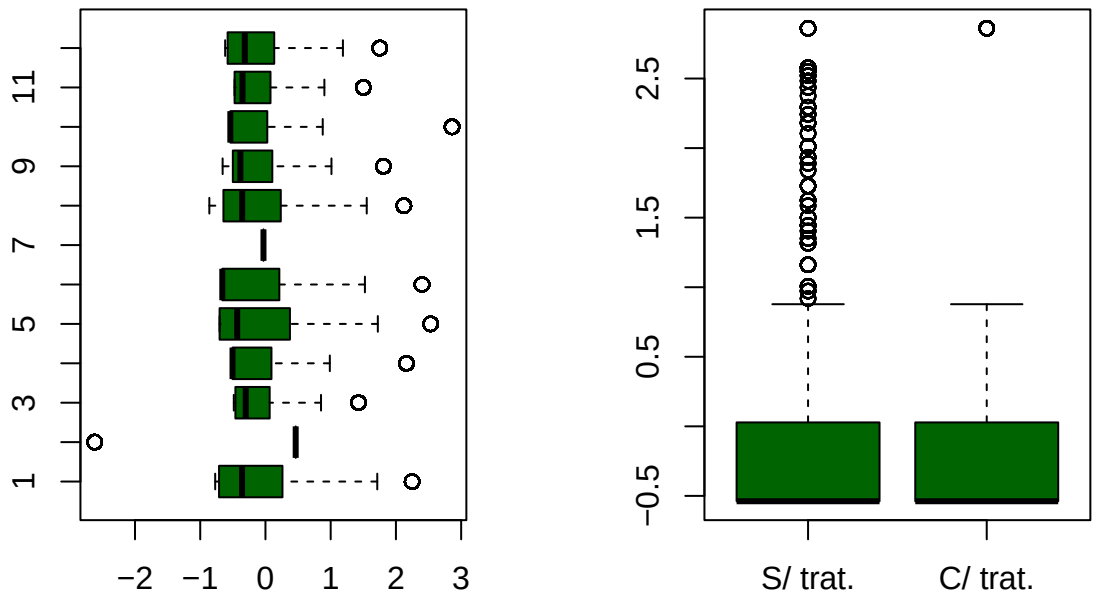


It was considered to remove lines that had outliers. However, this option proved to be unfeasible, given the large number of outliers. Thus, it was chosen to replace the outliers with values from the quartiles, as follows:

a) negative outliers will be replaced by 5% of the quartile;
b) positive outliers will be replaced by 95% of the quartile.

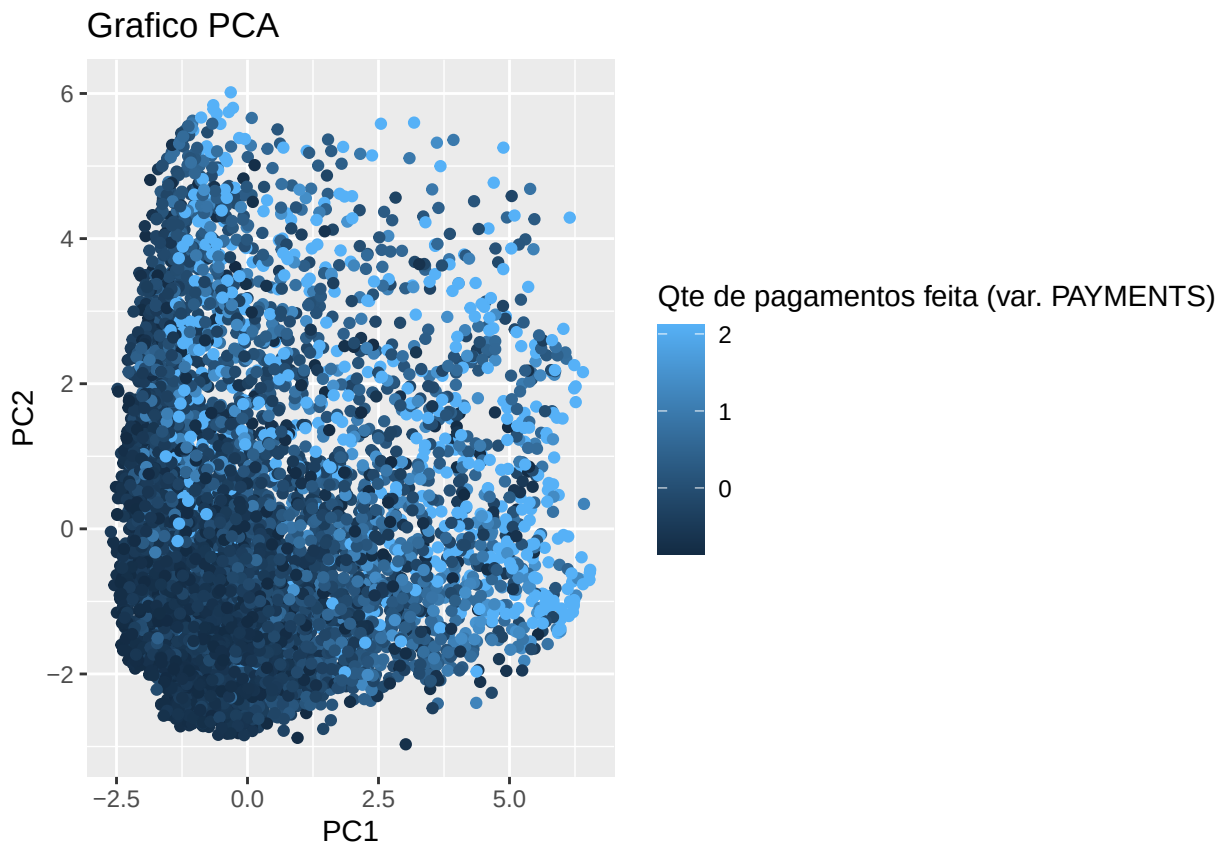
With the outliers treated, our dimensions were as follows:

```
par(mfrow=c(1,2))

boxplot(base_normal$BALANCE, base_normal$BALANCE_FREQUENCY, base_normal$PURCHASES, base_normal$CASH_ADV

boxplot (base_safepoint2$PRC_FULL_PAYMENT, base_normal$PRC_FULL_PAYMENT, col = "darkgreen", names = c("
```

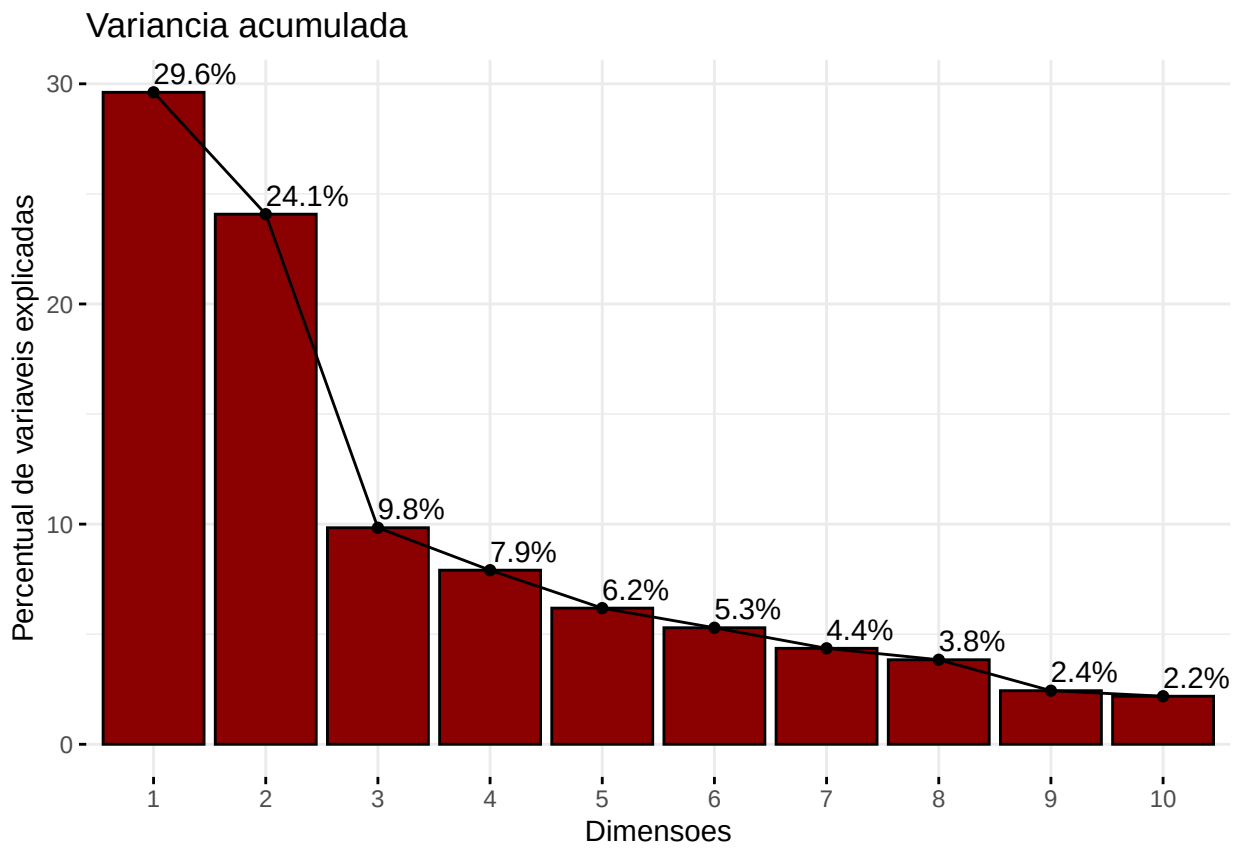


Then, the dimensionality reduction will be done. We will plot the directions and the percentage of variance explained by each dimension. The PCA method was used.

```
pca <- prcomp(base_normal, scale=TRUE, center=TRUE)
pca_df <- data.frame(x=pca$x[,"PC1"], y=pca$x[,"PC2"])

ggplot(data = pca_df, aes(x,y, color=base_normal$PAYMENTS)) +
  geom_point() + xlab("PC1") + ylab("PC2")+
  labs(color = "Qte de pagamentos feita (var. PAYMENTS)")+labs(title="Grafico PCA")
```




```
fviz_screeplot(pca,
               addlabels = TRUE,
               main = "Variancia acumulada",
               xlab = "Dimensoes",
               ylab = "Percentual de variaveis explicadas",
               barfill = "darkred",
               barcolor = "black")
```

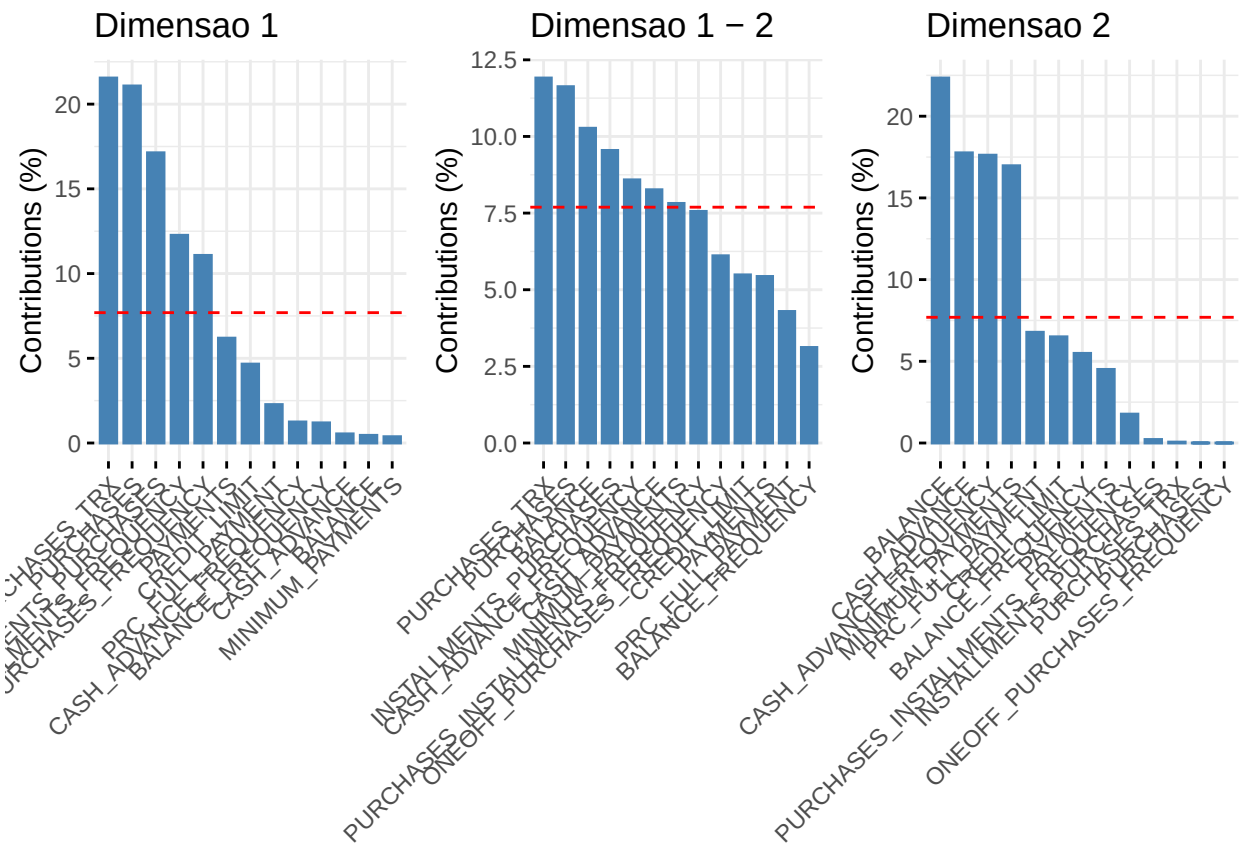


According to the graph, 5 dimensions account for 80% of the base variables. So, we know that we can go with 5 dimensions for this case. Therefore, we will plot below which are these 5 dimensions that represent this percentage. The rest will not be used.

```
"Contribuiçao das variaveis"
```

```
## [1] "Contribuiçao das variaveis"
```

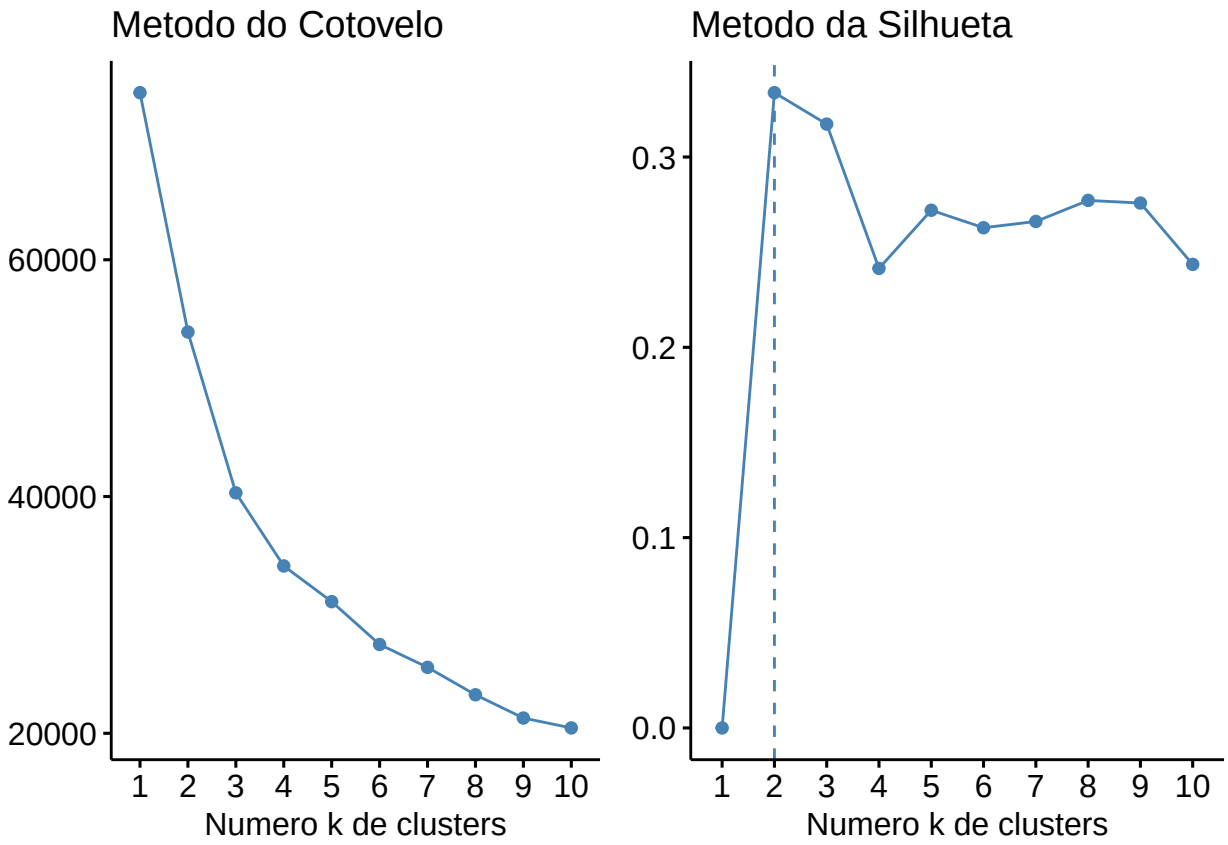
```
a <- fviz_contrib(pca, choice = "var", axes = 1, title = "Dimensao 1")
b <- fviz_contrib(pca, choice = "var", axes = 1:2, title = "Dimensao 1 - 2")
c <- fviz_contrib(pca, choice = "var", axes = 2, title = "Dimensao 2")
grid.arrange(a,b, c, ncol = 3)
```



So, as we will continue with 5 dimensions, we will choose the number of clusters using the base with dimensionality reduction.

We're going to do this in two ways: using the elbow method, and using the silhouette method. The two methods were plotted below, indicating the optimal number of clusters.

```
pca_6 <- pca$x[,1:5]

a <- fviz_nbclust(pca_6, kmeans, method = "wss") + ggtitle("Metodo do Cotovelo") + xlab("Numero k de clu

b <- fviz_nbclust(pca_6, kmeans, method = "silhouette") + ggtitle("Metodo da Silhueta") + xlab("Numero

grid.arrange(a,b, ncol = 2)
```



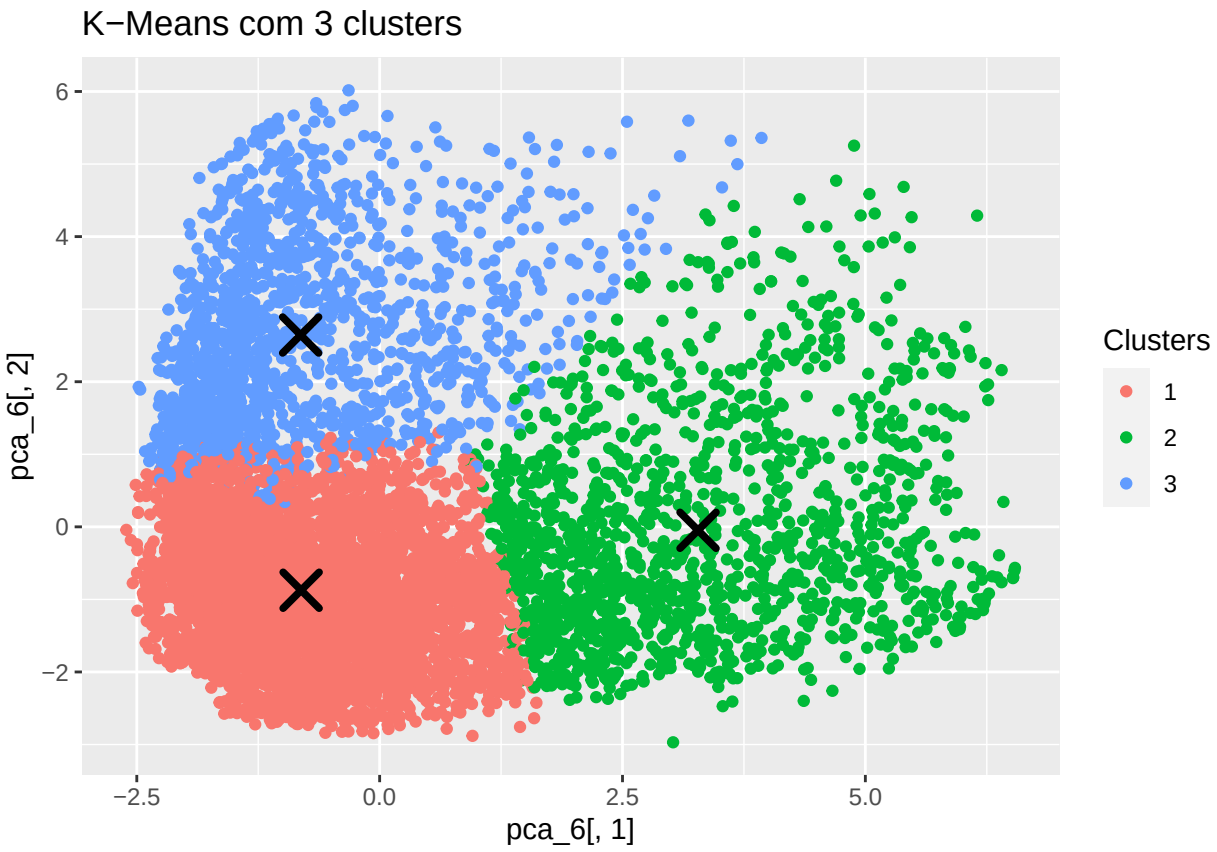

```
clust <- 3
```

The elbow method shows us a sharp curve at k = 3, and slight changes in slope thereafter. The silhouette method, on the other hand, indicates k = 2 as the optimal number, with a sharp drop at k = 3. It was chosen to continue with k = 3, because it is the point that presents considerable inflection by both methods.

So, we will follow with 5 dimensions and 3 clusters. We will do the clustering in two ways: by hierarchical clustering and k-means method.

We will start with k-means clustering, considering 3 clusters.

```
km.res <- kmeans(pca_6, centers = clust, iter.max = 100, nstart = 100)

ggplot() +
  geom_point(aes(x=pca_6[, 1], y=pca_6[, 2], color=factor(km.res$cluster))) +
  geom_point(aes(x=km.res$centers[, 1], y=km.res$centers[, 2]), color="black", size=5, shape=4, stroke=
  scale_color_discrete(name = "Clusters")+labs(title="K-Means com 3 clusters")
```



By clustering by K-Means, we can see that the 2nd cluster is well spaced, especially when compared to the 1st cluster.

Next, we'll run hierarchical clustering.

```
res.dist <- dist(pca_6, method = "euclidean")
clust.hq <- hclust(d = res.dist, method = "ward.D2")

#fviz_dend(clust.hq, k = clust, cex = 0.5, color_labels_by_k = TRUE, rect = TRUE) + ggtitle('Dendograma
```

For the analysis between the two clusters, we will plot the silhouettes for the two clusters:

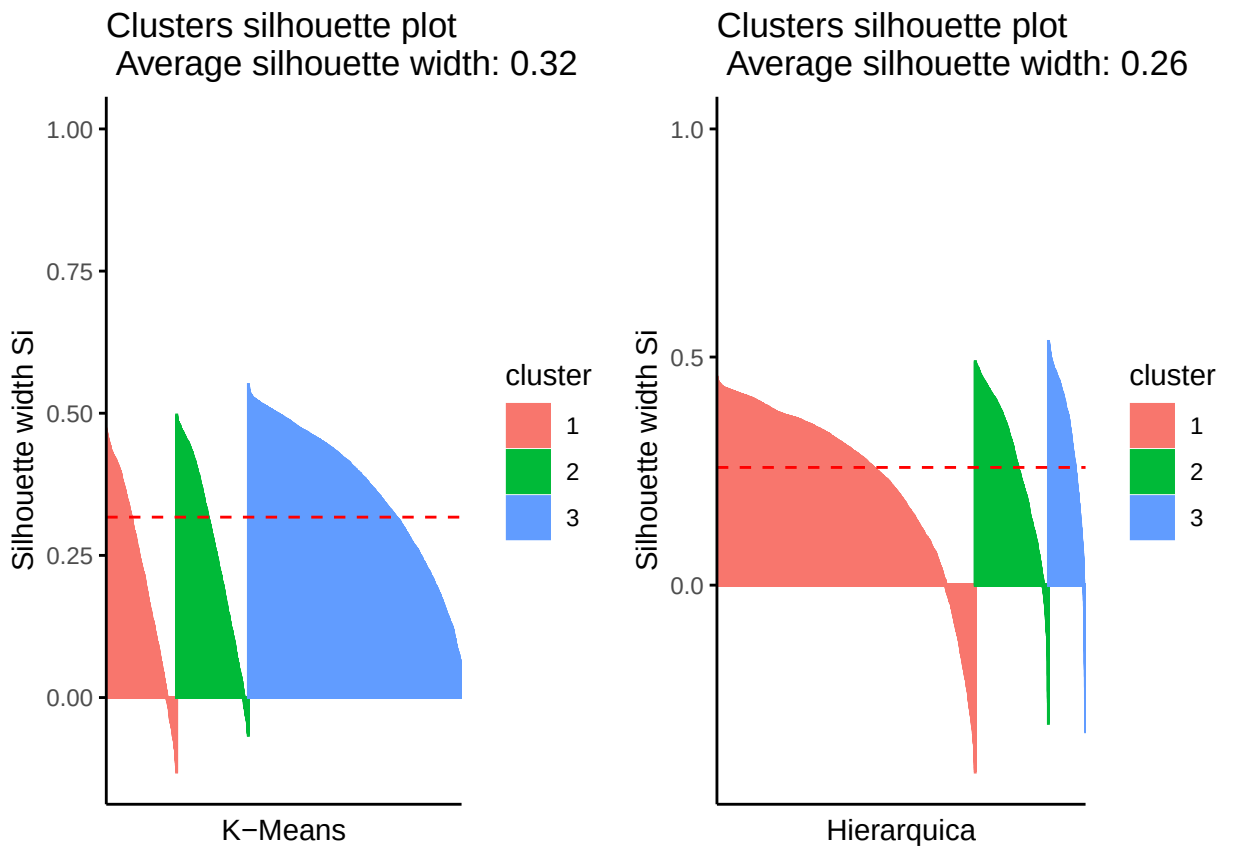
```
km_silh <- eclust(pca_6, "kmeans", k = 3, graph = FALSE, stand=FALSE, iter.max = 100,
                  nstart = 100)
hc_silh <- eclust(pca_6, "hclust", k = 3, graph = FALSE, stand=FALSE, iter.max = 100,
                  nstart = 100)

a <- fviz_silhouette(km_silh, ggtheme = theme_classic(), xlab = "K-Means")
```

```
##   cluster size ave.sil.width
## 1       1 1455          0.22
## 2       2 1483          0.26
## 3       3 4408          0.37
```

```
b <- fviz_silhouette(hc_silh, ggtheme = theme_classic(), xlab = "Hierarquica")
```

```
##   cluster size ave.sil.width
## 1       1 5148          0.24
## 2       2 1458          0.27
## 3       3  740          0.33
```

```
grid.arrange(a,b, ncol = 2)
```



From the graphics, it is clear that clustering by K-Means brings more gains. In addition to better balancing the 3 clusters, it brings less negative values and a higher clustering index.

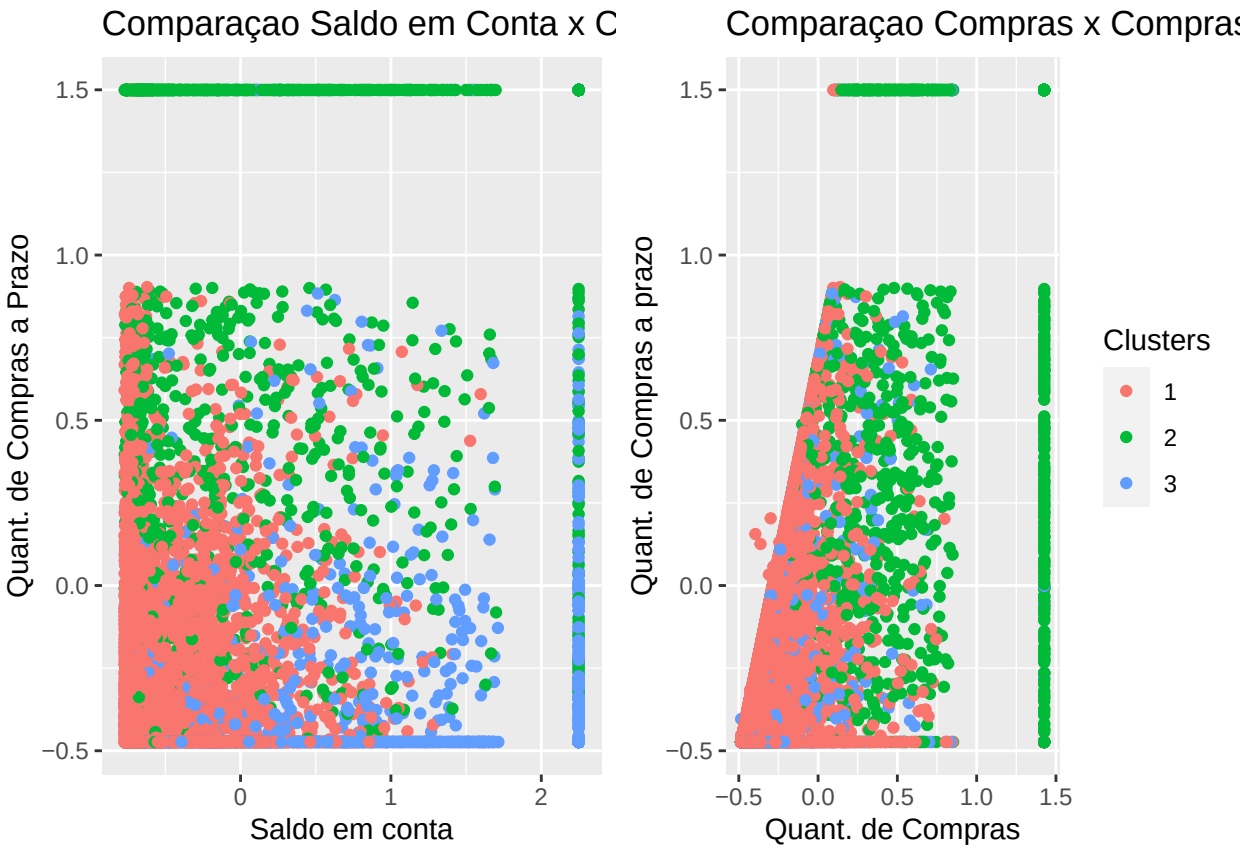
So here we end the clustering. Five dimensions were used, divided into 3 clusters using the K-Means method.

Below is the interpretation of the 3 groups created.

Interpretation of Clustering

```
nova_base <- base_normal
nova_base$Cluster <- km.res$cluster

#nova_base <- nb
nova_base$Cluster2 <- as.character(nova_base$Cluster)

#Comparision BALANCE x INSTALLMENTS_PURCHASES
a <-   ggplot() +
```

```
  geom_point(aes(x=nova_base$BALANCE, y=nova_base$INSTALLMENTS_PURCHASES, color=factor(nova_base$Cluste
  labs(title="Comparaçao Saldo em Conta x Compras a Prazo", x = "Saldo em conta", y = "Quant. de Compras

#Comparision PURCHASES x INSTALLMENTS_PURCHASES
b <-   ggplot() +
  geom_point(aes(x=nova_base$PURCHASES, y=nova_base$INSTALLMENTS_PURCHASES, color=factor(nova_base$Clus
  scale_color_discrete(name = "Clusters")

#Comparision BALANCE x CREDIT_LIMIT
c  <-  ggplot() +
  geom_point(aes(x=nova_base$BALANCE, y=nova_base$CREDIT_LIMIT, color=factor(nova_base$Cluster)))  +
  scale_color_discrete(name = "Clusters")+labs(title="Comparaçao Saldo em Conta x Limite do Cartao de C

grid.arrange(a,b, ncol = 2)
```

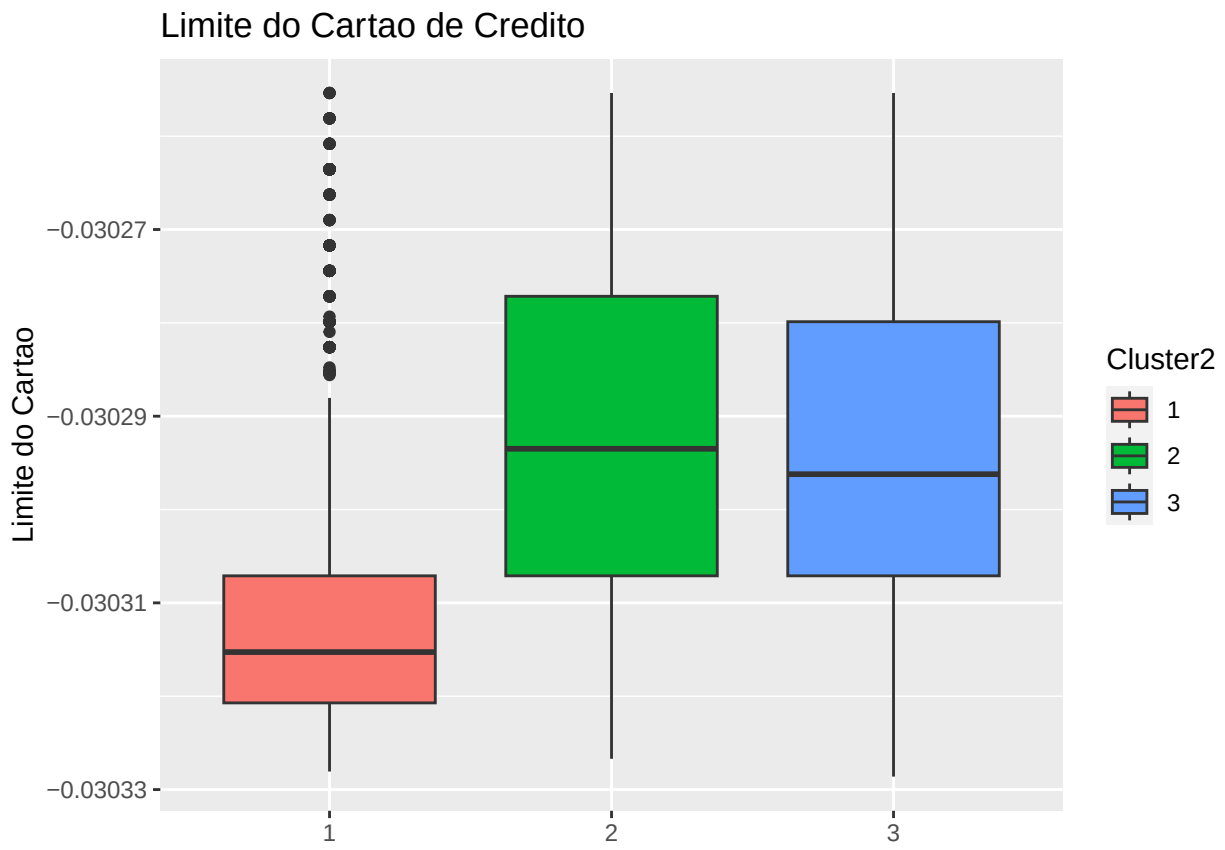


Through the graphs, it is possible to start tracing the profile of each group. The first graphic separates the blue from the pink cluster. The blue group has more account balance, while the pink group is farther to the left. The pink group is distributed in relation to the balance, but with a greater amount of installment purchases than the blue group.

The second graph, comparing the amount of installment and spot purchases, shows that the pink group, once again, differs from the blue group in terms of a greater amount of purchases - both in cash and in installments.

```
nova_base %>%
ggplot(aes(x = Cluster2, y = CREDIT_LIMIT, fill = Cluster2)) +
```

```
    geom_boxplot()+
  labs(title= "Limite do Cartao de Credito", x = NULL, y = "Limite do Cartao")+
  scale_color_discrete(name = "Clusters")
```



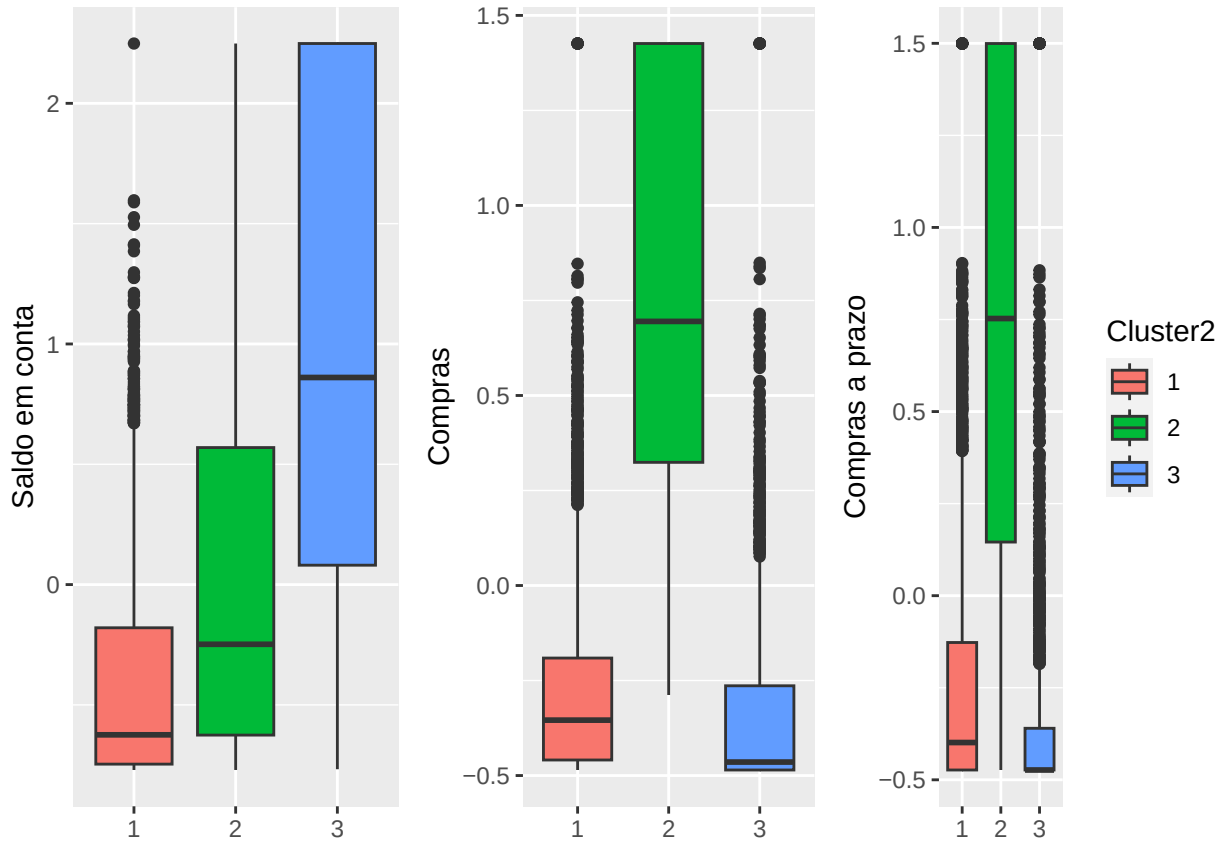
The third graphic shows the credit card limit. The pink group has the lowest values. When the blue and pink groups are compared, an interesting piece of information appears: the blue group has a higher account balance, but a slightly lower credit card limit.

Therefore, we can believe that:

1. The pink cluster has the lowest purchasing power. These are people who have a lower account balance and a lower credit card limit. Your purchases tend to be term purchases.
2. The green group has people with intermediate purchasing power between the two groups, but with higher consumption. They buy more, cash or installments, and therefore tend to have a high card limit.
3. The blue group does not have high credit card consumption. They are people with higher account balances, but with low purchase activity and average credit card limit. They do not usually make installment purchases.

```
#cor <- c("#F8766D", "#00BA38", "#619CFF")

a <- nova_base %>%
ggplot(aes(x = Cluster2, y = BALANCE, fill = Cluster2)) +
    geom_boxplot()+
```

```
  labs(title= NULL, x = NULL, y = "Saldo em conta")+ theme(legend.position="none")

b <- nova_base %>%
ggplot(aes(x = Cluster2, y = PURCHASES, fill = Cluster2)) +
    geom_boxplot()+
  labs(title= NULL, x = NULL, y = "Compras")+ theme(legend.position="none")

c <- nova_base %>%
ggplot(aes(x = Cluster2, y = INSTALLMENTS_PURCHASES, fill = Cluster2)) +
    geom_boxplot()+
  labs(title= NULL, x = NULL, y = "Compras a prazo")+
  scale_color_discrete(name = "Clusters")

grid.arrange(a,b, c, ncol = 3)
```

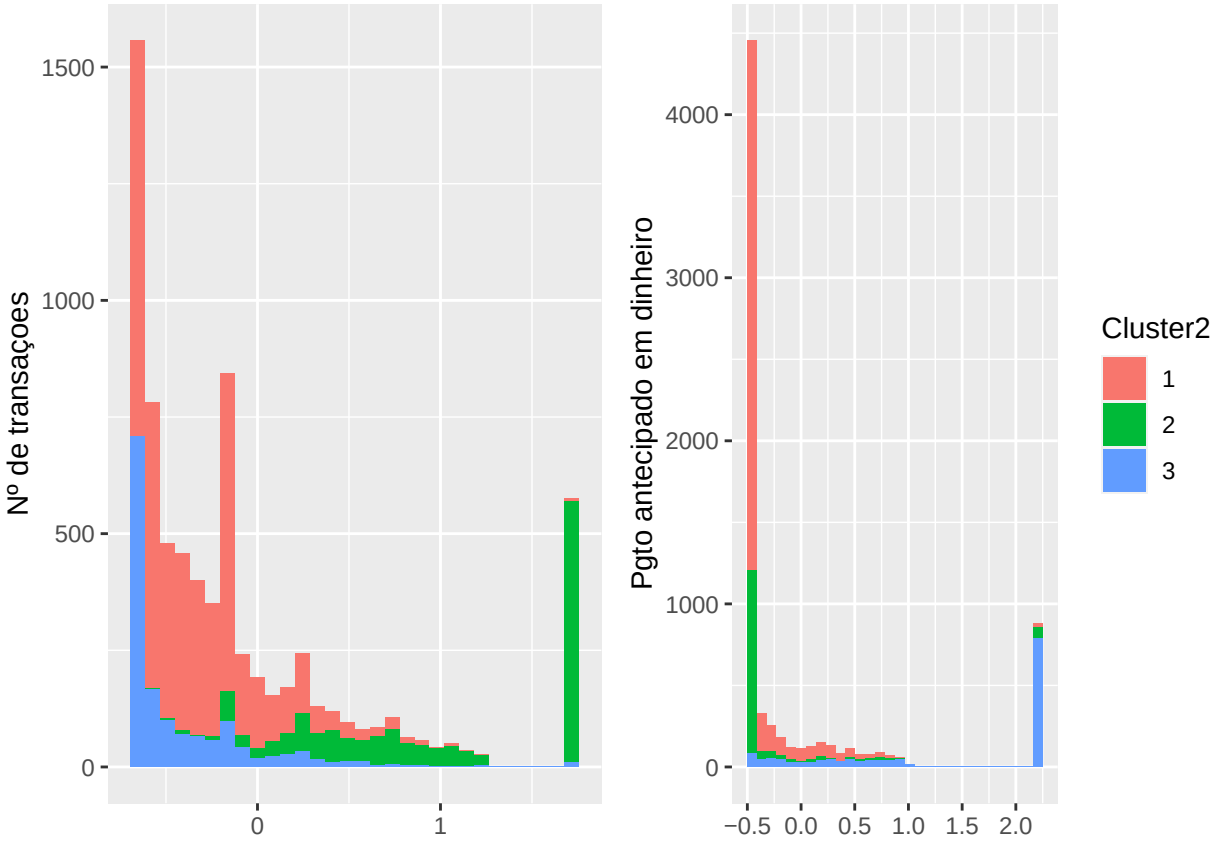


As expected, the boxplot charts above confirmed the assumptions about each cluster.

```
#cor <- c("#F8766D", "#00BA38", "#619CFF")

a <- nova_base %>%
ggplot(aes(x = PURCHASES_TRX, fill = Cluster2)) +
    geom_histogram()+
  labs(title= NULL, x = NULL, y = "Nº de transaçoes")+ theme(legend.position="none")

b <- nova_base %>%
ggplot(aes(x = CASH_ADVANCE, fill = Cluster2)) +
```

```
    geom_histogram()+
  labs(title= NULL, x = NULL, y = "Pgto antecipado em dinheiro")+
  scale_color_discrete(name = "Clusters")


grid.arrange(a,b, ncol = 2)
```

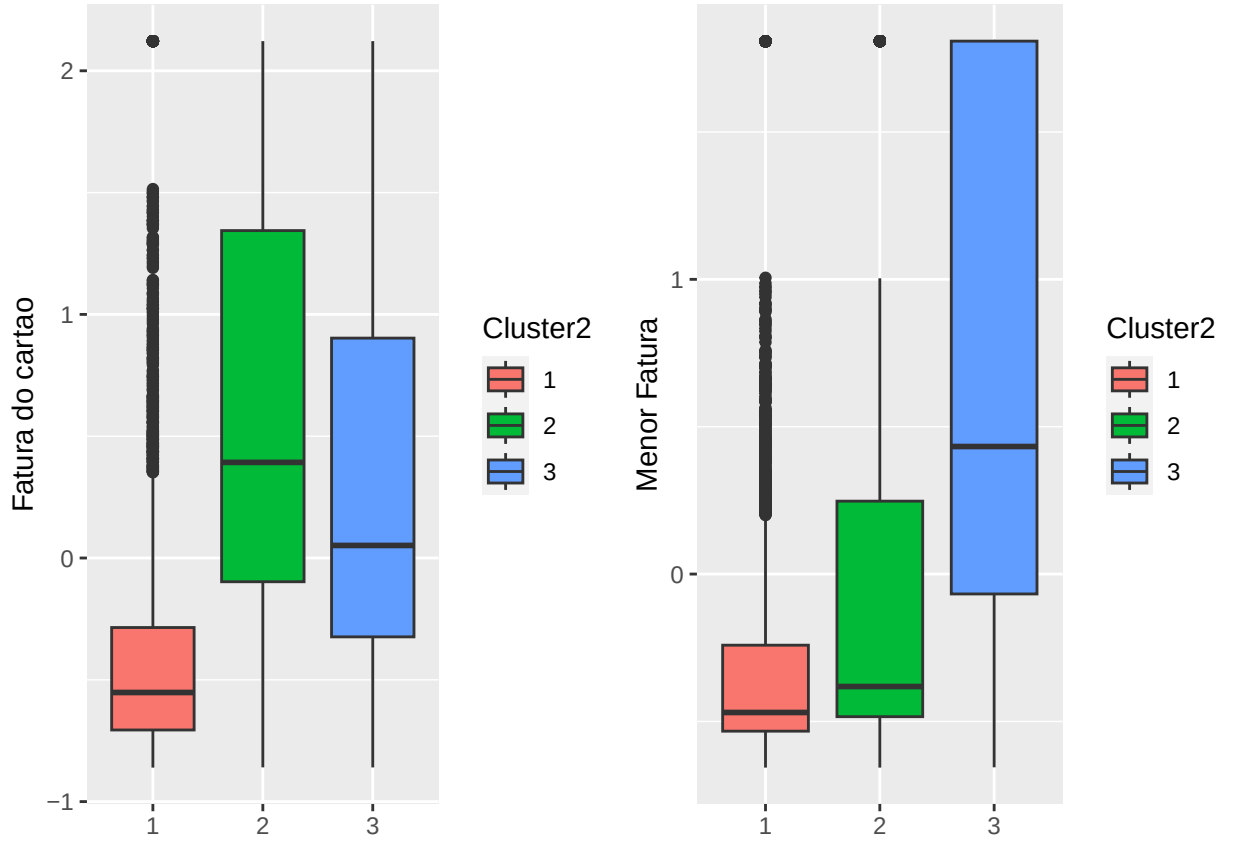
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The graphics above complement the clusters profile. They prove that the blue group has low credit card activity, often preferring to pay in cash. Like the pink group, they have a low number of transactions, unlike the green group, which has higher card consumption.

```
a <- nova_base %>%
ggplot(aes(x = Cluster2, y = PAYMENTS, fill = Cluster2)) +
    geom_boxplot()+
  labs(title= NULL, x = NULL, y = "Fatura do cartao")+
  scale_color_discrete(name = "Clusters")

b <- nova_base %>%
ggplot(aes(x = Cluster2, y = MINIMUM_PAYMENTS, fill = Cluster2)) +
    geom_boxplot()+
  labs(title= NULL, x = NULL, y = "Menor Fatura")+
  scale_color_discrete(name = "Clusters")
```

```
grid.arrange(a,b, ncol = 2)
```



Finally, we analyze the payouts of the 3 groups. We can conclude here that the pink group, as expected, has the lowest base invoice. The green group, due to high consumption, has the largest bill. However, when we analyze the blue group, we see that they are people who do not usually have the highest bill, but have a higher average ticket than the other groups - this is interpreted as having the highest minimum bill, so their expenses usually be high but less frequent.

Conclusion

The work reaches its conclusion with the presentation of the three clustered groups based on credit card consumption. The analysis of the groups indicates that there is a specific group (green) that is more conducive to consumption on the card, and may have more targeted actions. This group, according to the cluster analysis, is more consumerist, even if they do not have such a high account balance or have to resort to installment purchases. Another direction, starting from the cluster analysis, is the implementation of actions for the blue group. This has a balance but does not have as much consumption, being a market with good potential to be explored. For that, first it is necessary to understand why this group does not have so much consumption on the card, since they have the means to do so. Finally, the pink group has the least potential to be developed, as it is made up of people who do not have much balance in their account and a lower limit on their card. This group can be better explored through installment payment actions. The analysis of purchases versus installment purchases showed a small difference between them, and together with the low bill, it shows that this group does not consume because they are unable to pay in cash.

To improve the work, I recommend the application of other collinearity techniques and other k analysis for clustering.