# INTRODUCTION

Data Analysis not so explored within organizations is an indispensable tool that is indispensable within these companies. In this scenario of uncertainty of degrees and instability, the use of Statistics and its techniques for the understanding of a large amount of data that is generated by these organizations becomes indispensable.

Technological development, arising from scientific discoveries, has supported scientific development itself, expanding, in several orders of magnitude, the ability to obtain information on events and phenomena that are being analyzed. A large mass of information must be processed before being transformed into knowledge. Therefore, there is an increasing need for statistical tools that present a more global view of the phenomenon than that possible in a univariate approach. The denomination “Multivariate Analysis” corresponds to a large number of methods and techniques that use, simultaneously, all the variables in the theoretical interpretation of the data set obtained. (NETO, 2004).

Statistical methods to analyze variables are arranged in two groups: one that deals with statistics, which looks at variables in an isolated way – univariate statistics, and another that looks at variables together – multivariate statistics.

Multivariate Statistics includes methods for analyzing the relationships of multiple dependent variables and/or multiple independent variables, whether or not cause/effect relationships are established between these two groups. Also included in Multivariate Statistics are methods for analyzing relationships between individuals characterized by two or more variables.

Only Multivariate Statistics methods allow exploring the joint performance of the variables and determining the influence or importance of each one, with the remaining ones being presents.

## GOALS

The objective of this present work is to show the application of multivariate statistical techniques in the processing of the database (churn_missing). The database contains 10,000 observations and 15 variables that are:

**CustomerId:** Customer ID ,
**Surname:** Surname ,
**CreditScore:** Actual credit score,
**Geography:** Location, Gender: Gender, Age: Age,
**Tenure:** Number of months the customer stayed with the scompany,
**Balance:** Credit card debit balance , NumOfProducts: Number of products consumed ,
**HasCrCard:** Has credit ,
**IsActiveMember:** If the customer is activated in the bank,
**EstimatedSalary:** Estimated salary,
**Exited:** if the client stopped subscribing to the service or not (1 or 0 respectively).

METHODOLOGY

The software used in this work was R Studio, a free software integrated development environment for R, a programming language for graphs and statistical calculations.
The techniques used in this work for a multivariate analysis were PCA (principal component analysis) and **HIERARCHICAL CLUSTER**.

Principal component analysis **ACP** or **PCA** is the exploratory analysis technique that we use when we have a set of qualitative (numerical) data and we have not yet found the dependent and independent variables. If the variables in our set are categorical it is better to use MCA or Multiple Correspondence Analysis.

In machine learning, **PCA** is a technique that belongs to the "unsupervised" set. Through the PCA technique, it is possible to reduce the number of data without significant loss of information and, consequently, facilitate the interpretation of data.

Cluster analysis intends to group data into groups in order to form groups in which their elements are the most similar to each other. the groups are the most different from each other, it allows creating a centroid of each group that characterizes the average element of each group. This allows characterizing the typical element of a group and the typical differences between groups.

# PCA

## CLEANING THE MEMORY AND FIXING THE DIRECTORY

```
rm(list = ls(all = TRUE)) setwd("~/R")
```

## INSTALLING THE NECESSARY PACKAGES

```
library(FactoMineR)
library(factoextra) library(ggplot2)
library(factoextra)
```

## READING DATA SET

```
churn_missing <- read.csv("churn_missing.csv")
View(churn_missing)
```

## REMOVING NULL DATA

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

## CREATING SUBASSEMBLY

```
churn_missing.active<- churn_missing[1:10000,1:15]
head(churn_missing.active[1:15],4)
```

```
X RowNumber CustomerId  Surname CreditScore Geography Gender Age Tenure  Balance NumOfProducts HasCrCard
1         1   15634602 Hargrave         619    France Female  42      2      0.0             1         1
3         3   15619304     Onio         502    France Female  42      8 159660.8             3         1
4         4   15701354     Boni         699    France Female  39      1      0.0             2         0
5         5   15737888 Mitchell         850     Spain Female  43      2 125510.8             1         1
IsActiveMember EstimatedSalary Exited
             1       101348.88      1
             0       113931.57      1
             0        93826.63      0
             1        79084.10      0
```

## LOADING PACKAGES

```
library(missMDA)
```

## REPLACING MISSING DATA

The database had some missing data which was replaced using the command below.

```
imputePCA(churn_missing)
```

| | X | RowNumber | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 15634602 | Hargrave | 619 | France | Female | 42 | 2 | 0.00 | 1 |
| 3 | 3 | 3 | 15619304 | Onio | 502 | France | Female | 42 | 8 | 159660.80 | 3 |
| 4 | 4 | 4 | 15701354 | Boni | 699 | France | Female | 39 | 1 | 0.00 | 2 |
| 5 | 5 | 5 | 15737888 | Mitchell | 850 | Spain | Female | 43 | 2 | 125510.82 | 1 |
| 8 | 8 | 8 | 15656148 | Obinna | 376 | Germany | Female | 29 | 4 | 115046.74 | 4 |
| 12 | 12 | 12 | 15737173 | Andrews | 497 | Spain | Male | 24 | 3 | 0.00 | 2 |
| 13 | 13 | 13 | 15632264 | Kay | 476 | France | Female | 34 | 10 | 0.00 | 2 |
| 14 | 14 | 14 | 15691483 | Chin | 549 | France | Female | 25 | 5 | 0.00 | 2 |
| 15 | 15 | 15 | 15600882 | Scott | 635 | Spain | Female | 35 | 7 | 0.00 | 2 |
| 16 | 16 | 16 | 15643966 | Goforth | 616 | Germany | Male | 45 | 3 | 143129.41 | 2 |
| 18 | 18 | 18 | 15788218 | Henderson | 549 | Spain | Female | 24 | 9 | 0.00 | 2 |
| 20 | 20 | 20 | 15568982 | Hao | 726 | France | Female | 24 | 6 | 0.00 | 2 |
| 21 | 21 | 21 | 15577657 | McDonald | 732 | France | Male | 41 | 8 | 0.00 | 2 |
| 25 | 25 | 25 | 15625047 | Yen | 846 | France | Female | 38 | 5 | 0.00 | 1 |
| 26 | 26 | 26 | 15738191 | Maclean | 577 | France | Male | 25 | 3 | 0.00 | 2 |
| 27 | 27 | 27 | 15736816 | Young | 756 | Germany | Male | 36 | 2 | 136815.64 | 1 |
| 28 | 28 | 28 | 15700772 | Nebechi | 571 | France | Male | 44 | 9 | 0.00 | 2 |
| 30 | 30 | 30 | 15656300 | Lucciano | 411 | France | Male | 29 | 0 | 59697.17 | 2 |
| 31 | 31 | 31 | 15589475 | Azikiwe | 591 | Spain | Female | 39 | 3 | 0.00 | 3 |
| 32 | 32 | 32 | 15706552 | Odinakachukwu | 533 | France | Male | 36 | 7 | 85311.70 | 1 |
| 34 | 34 | 34 | 15659428 | Maggard | 520 | Spain | Female | 42 | 6 | 0.00 | 2 |
| 35 | 35 | 35 | 15732963 | Clements | 722 | Spain | Female | 29 | 9 | 0.00 | 2 |
| 36 | 36 | 36 | 15794171 | Lombardo | 475 | France | Female | 45 | 0 | 134264.04 | 1 |
| 38 | 38 | 38 | 15729599 | Lorenzo | 804 | Spain | Male | 33 | 7 | 76548.60 | 1 |
| 44 | 44 | 44 | 15755196 | Lavine | 834 | France | Female | 49 | 2 | 131394.56 | 1 |
| 46 | 46 | 46 | 15754849 | Tyler | 776 | Germany | Female | 32 | 4 | 109421.13 | 2 |
| 47 | 47 | 47 | 15602280 | Martin | 829 | Germany | Female | 27 | 9 | 112045.67 | 1 |
| 48 | 48 | 48 | 15771573 | Okagbue | 637 | Germany | Female | 39 | 9 | 137843.80 | 1 |
| 49 | 49 | 49 | 15766205 | Yin | 550 | Germany | Male | 38 | 2 | 103391.38 | 1 |
| 50 | 50 | 50 | 15771873 | Buccho | 776 | Germany | Female | 37 | 2 | 103769.22 | 2 |
| 51 | 51 | 51 | 15616550 | Chidiebele | 698 | Germany | Male | 44 | 10 | 116363.37 | 2 |
| 52 | 52 | 52 | 15768193 | Trevisani | 585 | Germany | Male | 36 | 5 | 146050.97 | 2 |
| 54 | 54 | 54 | 15702298 | Parkhill | 655 | Germany | Male | 41 | 8 | 125561.97 | 1 |
| 56 | 56 | 56 | 15760861 | Phillipps | 619 | France | Male | 43 | 1 | 125211.92 | 1 |
| 57 | 57 | 57 | 15630053 | Tsao | 656 | France | Male | 45 | 5 | 127864.40 | 1 |
| 58 | 58 | 58 | 15647091 | Endrizzi | 725 | Germany | Male | 19 | 0 | 75888.20 | 1 |
| 59 | 59 | 59 | 15623944 | T'ien | 511 | Spain | Female | 66 | 4 | 0.00 | 1 |
| 60 | 60 | 60 | 15804771 | Velazquez | 614 | France | Male | 51 | 4 | 40685.92 | 1 |
| 61 | 61 | 61 | 15651280 | Hunter | 742 | Germany | Male | 35 | 5 | 136857.00 | 1 |
| 63 | 63 | 63 | 15702014 | Jeffrey | 555 | Spain | Male | 33 | 1 | 56084.69 | 2 |
| 64 | 64 | 64 | 15751208 | Pirozzi | 684 | Spain | Male | 56 | 8 | 78707.16 | 1 |
| 65 | 65 | 65 | 15592461 | Jackson | 603 | Germany | Male | 26 | 4 | 109166.37 | 1 |
| 66 | 66 | 66 | 15789484 | Hammond | 751 | Germany | Female | 36 | 6 | 169831.46 | 2 |
| 68 | 68 | 68 | 15641582 | Chibugo | 735 | Germany | Male | 43 | 10 | 123180.01 | 2 |
| 69 | 69 | 69 | 15638424 | Glauert | 661 | Germany | Female | 35 | 5 | 150725.53 | 2 |
| 70 | 70 | 70 | 15755648 | Pisano | 675 | France | Female | 21 | 8 | 98373.26 | 1 |
| 72 | 72 | 72 | 15620344 | McKee | 813 | France | Male | 29 | 6 | 0.00 | 1 |
| 73 | 73 | 73 | 15812518 | Palermo | 657 | Spain | Female | 37 | 0 | 163607.18 | 1 |
| 76 | 76 | 76 | 15780961 | Cavenagh | 735 | France | Female | 21 | 1 | 178718.19 | 2 |
| 77 | 77 | 77 | 15614049 | Hu | 664 | France | Male | 55 | 8 | 0.00 | 2 |
| 79 | 79 | 79 | 15575185 | Bushell | 757 | Spain | Male | 33 | 5 | 77253.22 | 1 |
| 80 | 80 | 80 | 15803136 | Postle | 416 | Germany | Female | 41 | 10 | 122189.66 | 2 |
| 82 | 82 | 82 | 15663706 | Leonard | 777 | France | Female | 32 | 2 | 0.00 | 1 |
| 83 | 83 | 83 | 15641732 | Mills | 543 | France | Female | 36 | 3 | 0.00 | 2 |
| 86 | 86 | 86 | 15805254 | Ndukaku | 652 | Spain | Female | 75 | 10 | 0.00 | 2 |
| 89 | 89 | 89 | 15622897 | Sharpe | 646 | France | Female | 46 | 4 | 0.00 | 3 |
| 94 | 94 | 94 | 15640635 | Capon | 769 | France | Male | 29 | 8 | 0.00 | 2 |
| 97 | 97 | 97 | 15738721 | Graham | 773 | Spain | Male | 41 | 9 | 102827.44 | 1 |
| 98 | 98 | 98 | 15693683 | Yuille | 814 | Germany | Male | 29 | 8 | 97086.40 | 2 |
| 100 | 100 | 100 | 15633059 | Fanucci | 413 | France | Male | 34 | 9 | 0.00 | 2 |
| 101 | 101 | 101 | 15808582 | Fu | 665 | France | Female | 40 | 6 | 0.00 | 1 |
| 104 | 104 | 104 | 15776605 | Bradley | 528 | Spain | Male | 36 | 7 | 0.00 | 2 |
| 105 | 105 | 105 | 15804919 | Dunbabin | 670 | Spain | Female | 65 | 1 | 0.00 | 1 |
| 108 | 108 | 108 | 15812878 | Parsons | 785 | Germany | Female | 36 | 2 | 99806.85 | 1 |
| 109 | 109 | 109 | 15602312 | Walkom | 605 | Spain | Male | 33 | 5 | 150092.80 | 1 |
| 110 | 110 | 110 | 15744689 | T'ang | 479 | Germany | Male | 35 | 9 | 92833.89 | 1 |

| | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|
| 1 | 1 | 1 | 101348.88 | 1 |
| 3 | 1 | 0 | 113931.57 | 1 |
| 4 | 0 | 0 | 93826.63 | 0 |
| 5 | 1 | 1 | 79084.10 | 0 |
| 8 | 1 | 0 | 119346.88 | 1 |
| 12 | 1 | 0 | 76390.01 | 0 |
| 13 | 1 | 0 | 26260.98 | 0 |
| 14 | 0 | 0 | 190857.79 | 0 |
| 15 | 1 | 1 | 65951.65 | 0 |
| 16 | 0 | 1 | 64327.26 | 0 |
| 18 | 1 | 1 | 14406.41 | 0 |
| 20 | 1 | 1 | 54724.03 | 0 |
| 21 | 1 | 1 | 170886.17 | 0 |
| 25 | 1 | 1 | 187616.16 | 0 |
| 26 | 0 | 1 | 124508.29 | 0 |
| 27 | 1 | 1 | 170041.95 | 0 |
| 28 | 0 | 0 | 38433.35 | 0 |
| 30 | 1 | 1 | 53483.21 | 0 |
| 31 | 1 | 0 | 140469.38 | 1 |
| 32 | 0 | 1 | 156731.91 | 0 |
| 34 | 1 | 1 | 34410.55 | 0 |
| 35 | 1 | 1 | 142033.07 | 0 |
| 36 | 1 | 0 | 27822.99 | 1 |
| 38 | 0 | 1 | 98453.45 | 0 |
| 44 | 0 | 0 | 194365.76 | 1 |
| 46 | 1 | 1 | 126517.46 | 0 |
| 47 | 1 | 1 | 119708.21 | 1 |
| 48 | 1 | 1 | 117622.80 | 1 |
| 49 | 0 | 1 | 90878.13 | 0 |
| 50 | 1 | 0 | 194099.12 | 0 |
| 51 | 1 | 0 | 198059.16 | 0 |
| 52 | 0 | 0 | 86424.57 | 0 |
| 54 | 0 | 0 | 164040.94 | 1 |
| 56 | 1 | 1 | 113410.49 | 0 |
| 57 | 1 | 0 | 87107.57 | 0 |
| 58 | 0 | 0 | 45613.75 | 0 |
| 59 | 1 | 0 | 1643.11 | 1 |
| 60 | 1 | 1 | 46775.28 | 0 |
| 61 | 0 | 0 | 84509.57 | 0 |
| 63 | 0 | 0 | 178798.13 | 0 |
| 64 | 1 | 1 | 99398.36 | 0 |
| 65 | 1 | 1 | 92840.67 | 0 |
| 66 | 1 | 1 | 27758.36 | 0 |
| 68 | 1 | 1 | 196673.28 | 0 |
| 69 | 0 | 1 | 113656.85 | 0 |
| 70 | 1 | 0 | 18203.00 | 0 |
| 72 | 1 | 0 | 33953.87 | 0 |
| 73 | 0 | 1 | 44203.55 | 0 |
| 76 | 1 | 0 | 22388.00 | 0 |
| 77 | 1 | 1 | 139161.64 | 0 |
| 79 | 0 | 1 | 194239.63 | 0 |
| 80 | 1 | 0 | 98301.61 | 0 |
| 82 | 1 | 0 | 136458.19 | 1 |
| 83 | 0 | 0 | 26019.59 | 0 |
| 86 | 1 | 1 | 114675.75 | 0 |
| 89 | 1 | 0 | 93251.42 | 1 |
| 94 | 1 | 1 | 172290.61 | 0 |
| 97 | 0 | 1 | 64595.25 | 0 |
| 98 | 1 | 1 | 197276.13 | 0 |
| 100 | 0 | 0 | 6534.18 | 0 |
| 101 | 1 | 1 | 161848.03 | 0 |
| 104 | 1 | 0 | 60536.56 | 0 |
| 105 | 1 | 1 | 177655.68 | 1 |
| 108 | 0 | 1 | 36976.52 | 0 |
| 109 | 0 | 0 | 71862.79 | 0 |
| 110 | 1 | 0 | 99449.86 | 1 |

**DELETE COLUMNS**

Eleven columns were removed from the database because they are not significant for this study.

```
churn_missing<-churn_missing[,-c(1,2,3,4,5,6,7,8,10,14,15)]
head(churn_missing)
```

The 4 variables that will be worked on in this analysis will be the Number of products that the customer has in the bank, if he is an active customer or not, if he has a credit card and the Number of months that the customer stayed at the bank.

```
Tenure NumOfProducts HasCrCard IsActiveMember
1      2             1         1              1
2      1             1         0              1
3      8             3         1              0
4      1             2         0              0
5      2             1         1              1
6      8             2         1              0
```

**STANDARDIZATION**

```
churn_missing.active <-scale(churn_missing)
```

**GENERATING PCA**

```
res.pca<-PCA(churn_missing, graph = F)
View(res.pca)
```

**EXTRACT SELF-VALUES**

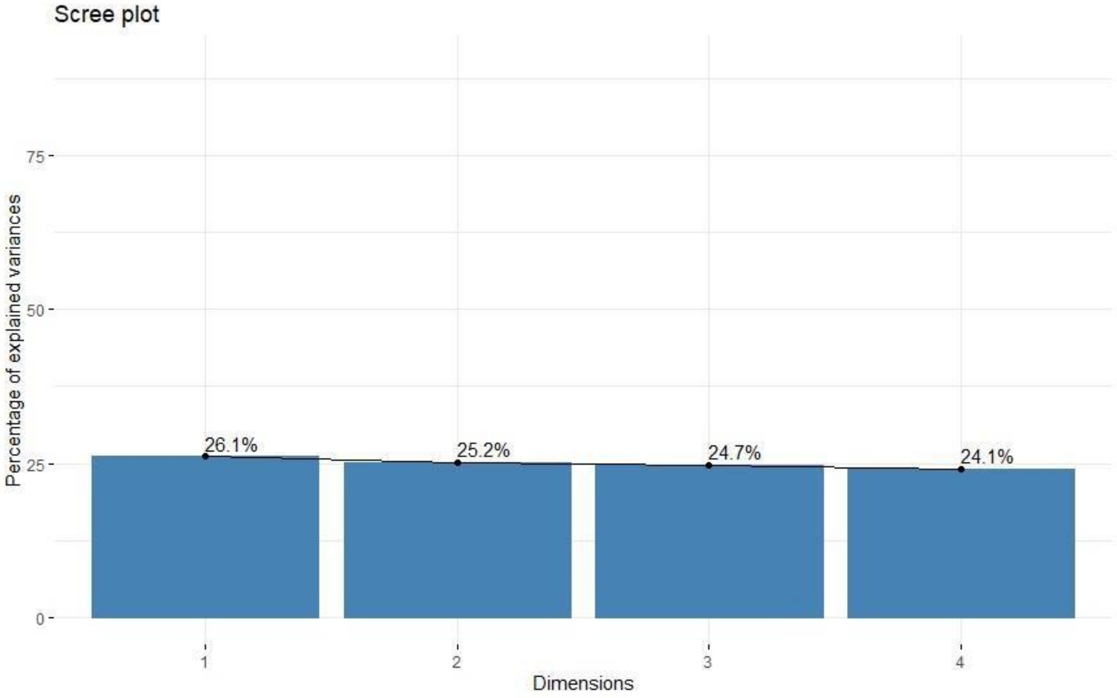
```
eig.val<- get_eigenvalue(res.pca) eig.val
 res.pca <- prcomp(churn_missing,  scale = TRUE)
```

| | eigenvalue | variance.percent | cumulative.variance.percent |
|---|---|---|---|
| Dim.1 | 1.0431586 | 26.07896 | 26.07896 |
| Dim.2 | 1.0074298 | 25.18574 | 51.26471 |
| Dim.3 | 0.9865259 | 24.66315 | 75.92786 |
| Dim.4 | 0.9628857 | 24.07214 | 100.00000 |

We can use the size of the eigenvalues to determine the number of principal components. using the Kaiser criterion, we will only use principal components with eigenvalues that are greater than 1.

**PLOT THE GRAPH**

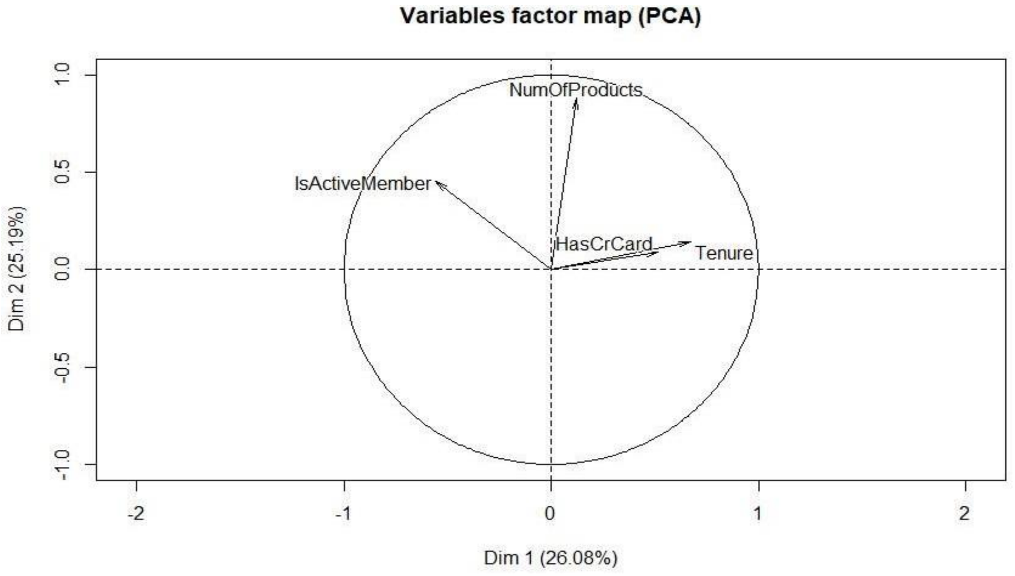
```
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 90))
```




This graph shows the percentage of estimated variances. We can see that almost 100% of the information (variations) contained in the data are retained by the first four main components.

Other graphs that show the position of the variables and the concentration of the data.

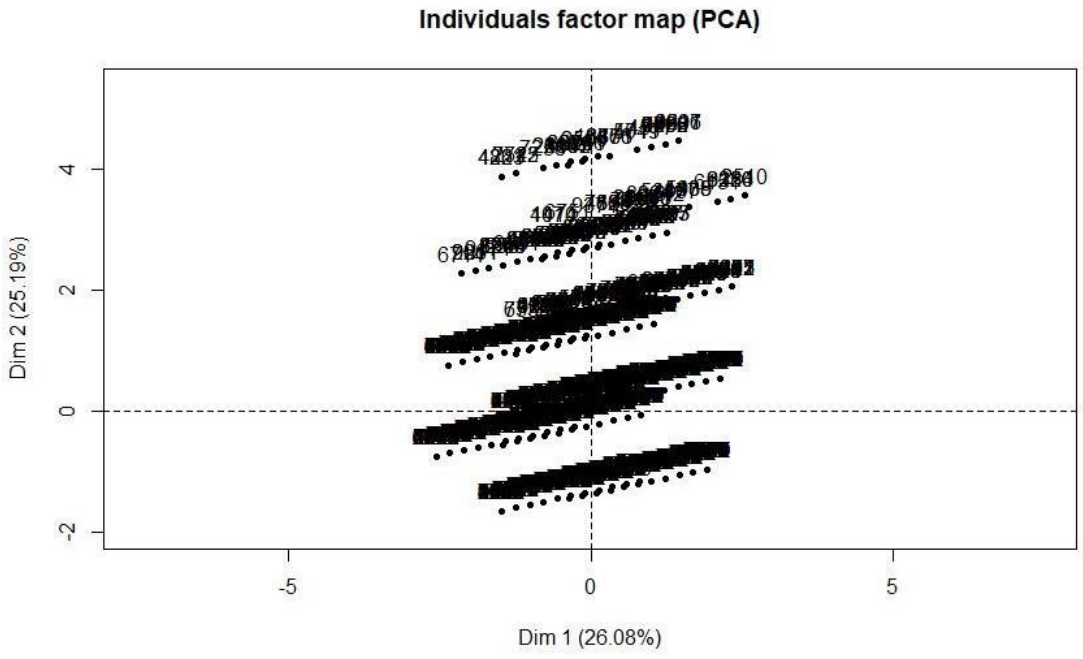
```
plot(PCA(churn_missing))
```




This other variable correlation graph shows the relationship of all variables where positively correlated variables are grouped while negatively correlated variables are positioned on opposite sides of the origin of the graphs (opposite quadrants).

In the graph above, we have 3 positively correlated variables on the upper right side, which are the number of products the customer has in the bank, whether or not they have a credit card and the number of months the customer stayed at the bank.

The ISACTIVEMENBER variable, which says whether the customer is active or not, has a negative correlation with the other variables, that is, being an active customer or not, does not imply that the customer has or does not have any financial products.



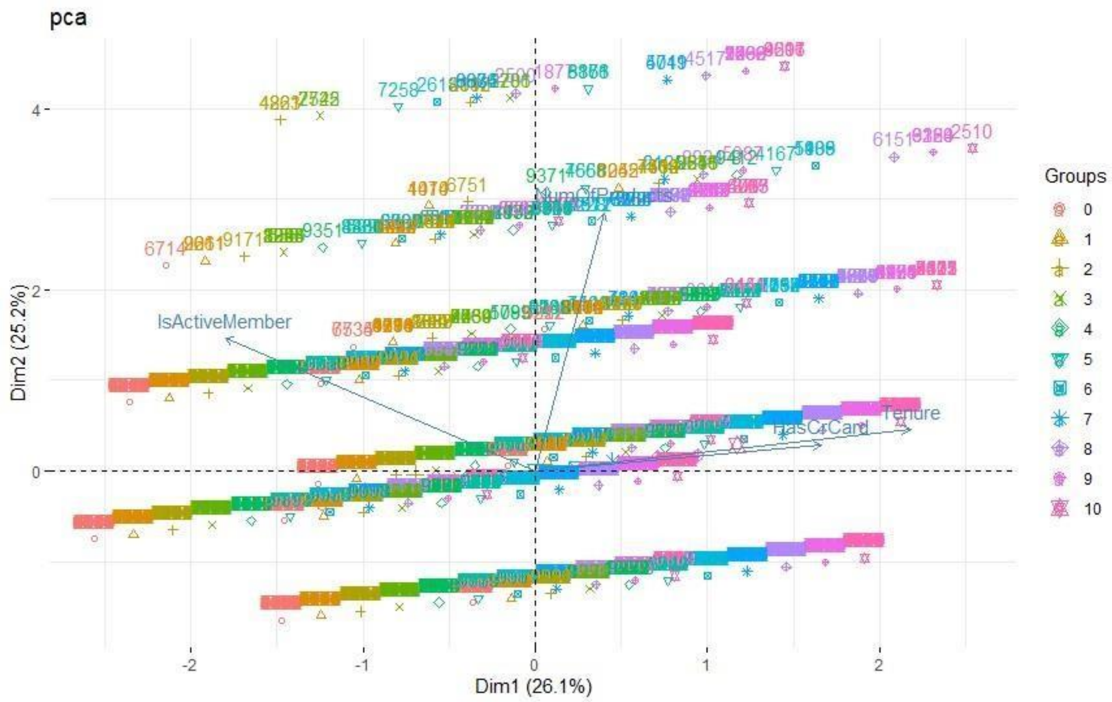

# CLUSTER

## CREATE CLUSTER GROUP

```
grupo<- as.factor(churn_missing [,1])
```

## PLOT GRAPHIC BIPLOT

```
fviz_pca_biplot(res.pca , habillage = grupo , title = " GRUPO")
```

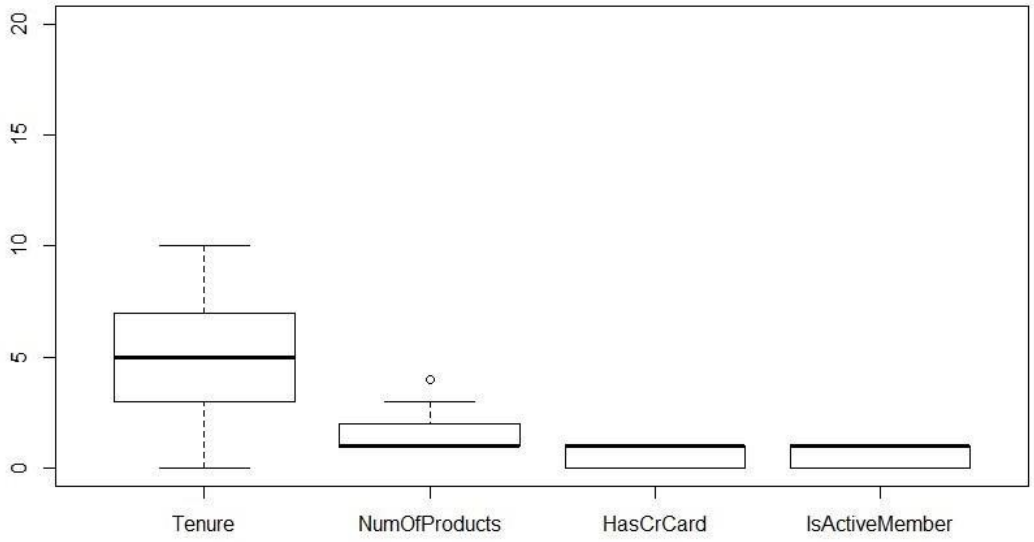



This graph shows the cluster groups.

**PLOTTING THE BOXPLOT GRAPHIC**

The Boxplot plot evaluates various points such as data symmetry, outliers, concentration of values.

```
boxplot(anadados, ylim= c
(0,20))
```




We can see in the graph above that the variable TENURE presents its data concentrated between the first and the second quartile, whereas the variable NUMOFPRODUCTS presented discrepant data (data that do not seem to be part of the data set).

It is observed that most customers have a single product.

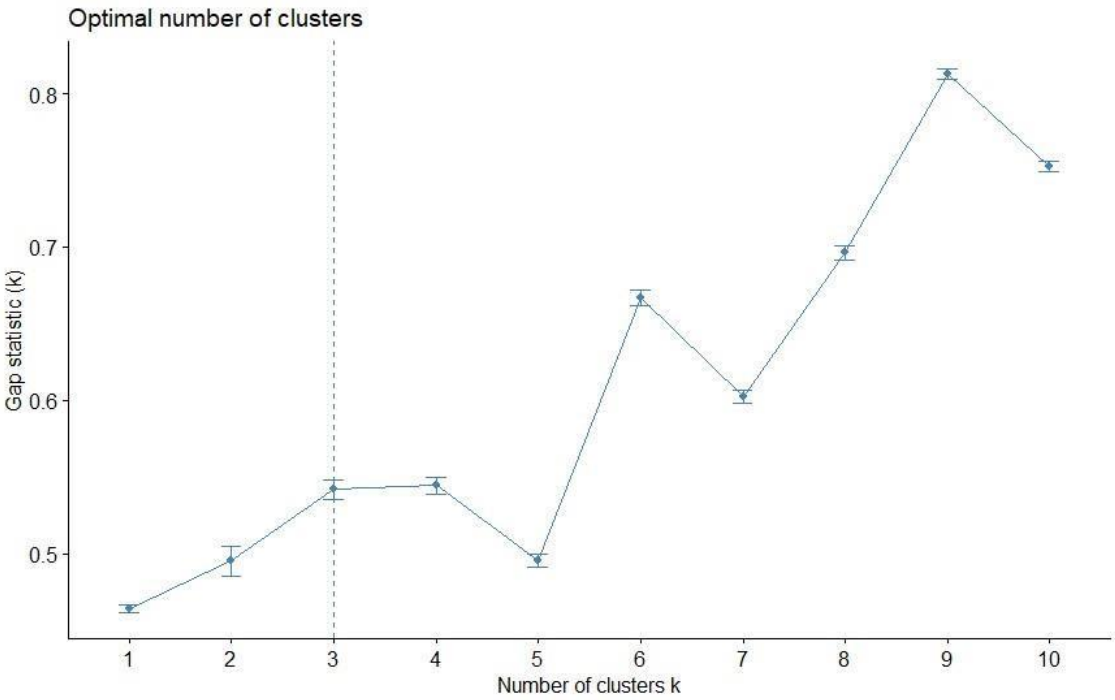
**TRANSFORM TO SCALE**

```
dados<- scale(anadados)
```

**DEFINE NUMBER OF CLUSTERS**

```
fviz_nbclust(dados, kmeans, method =  "gap_stat", print.summary = TRUE
)
```

```
Clustering k = 1,2,..., K.max (= 10): .. done
Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
.................................................. 50
.................................................. 100
```




As a result of the graph above, it shows that the ideal number of clusters to work with in this dataset is 3. Since the ISACTIVEMERBER variable already showed a negative correlation, it was removed in our model because it does not represent the problem in question.

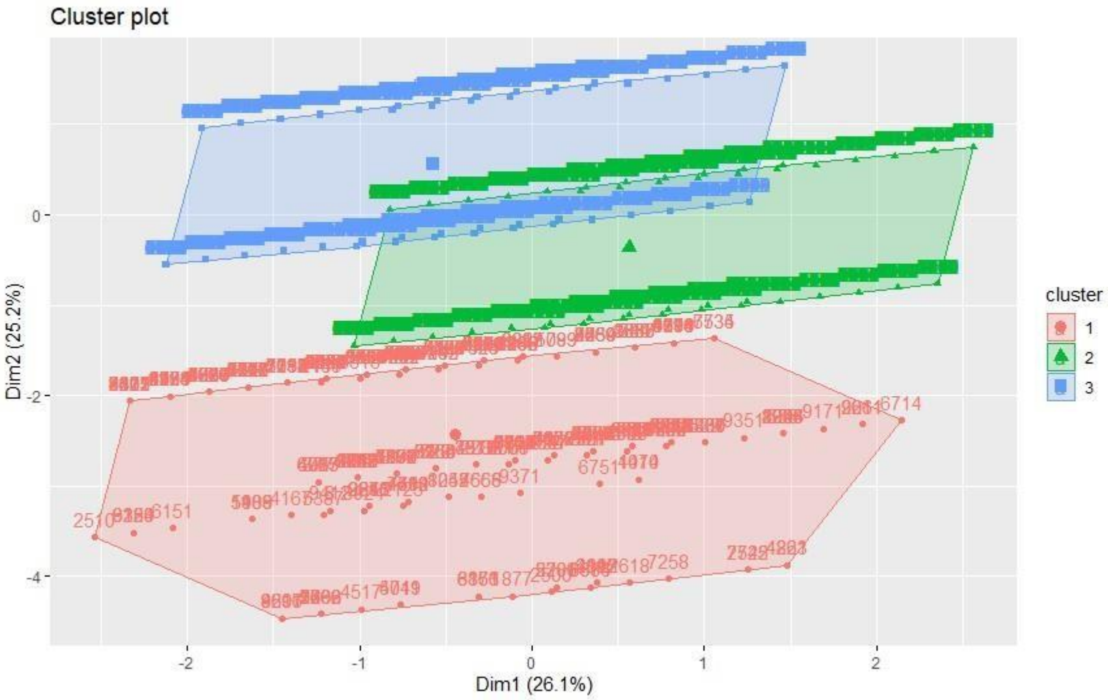
Through the 3 clusters pointed out by the model we can verify the correlation of these variables and through them we can verify whether or not there is a problem and its possible solutions.

**GENERATE KMENS**

```
dados_kmeans<- kmeans(dados, 3)
```

**VIEW THE GRAPHIC**

```
fviz_cluster(dados_kmeans, data =
dados)
```




In this graph we can observe the 3 main groups of clusters separately. In pink we have the group of customers and the length of stay at the bank, in green we have the group of customers who have a financial product and in blue we have the group of customers who have a credit card.

```
dados_kmeans$cluster
```

```
  [1] 2 2 1 3 2 3 2 1 2 2 3 3 3 3 2 2 3 2 3 2 2 3 3 2 2 2 3 2 2 1 2 3 2 2 3 2 2 2 2 3 3 2 3 2 2 2 2 3 3 3 3 3 3 2 3
 [58] 3 3 2 3 3 3 2 2 2 3 2 2 3 1 3 2 2 2 3 2 2 2 3 3 3 3 2 3 2 3 3 1 2 1 3 3 2 2 2 2 2 3 3 2 3 3 3 2 2 2 2 3 3 2 3 3 2
115] 2 2 3 3 3 3 3 3 3 2 3 3 2 3 3 3 2 2 2 2 2 2 2 2 3 2 3 2 2 3 3 2 3 3 3 2 3 3 3 3 2 3 3 2 2 3 2 3 3 2 3 2 3 3 3 3 3
172] 3 3 3 2 2 2 3 2 3 2 2 3 2 2 3 3 2 2 3 3 3 3 3 3 2 2 2 2 3 2 2 2 2 3 2 3 2 3 2 3 2 3 2 3 2 3 3 3 2 2 2 2 3 2 3 2 1
229] 3 3 2 3 2 2 2 3 2 1 1 2 3 3 2 2 3 3 3 2 2 3 3 2 2 2 3 2 2 3 2 2 2 3 3 2 2 3 2 3 3 3 2 1 2 3 2 2 2 3 2 3 2 2 2 2 2
286] 3 3 3 2 3 2 3 3 3 2 3 2 1 2 2 3 3 3 2 3 2 2 3 3 3 2 3 3 3 2 3 3 2 3 3 2 2 3 3 3 3 3 3 2 2 2 3 2 3 3 3 2 3 3 3 2 3
343] 2 3 3 3 3 2 2 2 2 2 3 2 2 3 2 3 2 2 2 2 2 3 2 2 3 2 2 3 2 2 2 2 2 2 1 3 3 1 2 3 2 3 3 2 3 2 2 3 2 2 3 3 2 2 2 3 2
400] 3 3 2 2 2 2 3 3 3 2 3 3 2 2 1 3 3 3 2 3 3 3 3 3 3 3 3 3 2 3 3 2 2 3 3 3 2 3 3 3 2 3 3 2 2 3 2 3 2 3 3 3 2 3 3 2 3
457] 2 3 2 2 2 3 3 2 2 3 2 1 3 2 2 3 3 3 3 3 2 3 2 2 2 3 1 3 3 2 3 2 2 2 2 2 2 3 3 2 3 2 2 2 2 2 3 2 3 3 3 3 2 2 2 3 3
514] 2 2 3 1 3 2 2 3 2 2 2 3 2 2 3 3 2 2 2 2 3 3 3 3 3 1 3 2 2 2 3 1 3 3 3 2 1 3 2 3 3 3 1 3 3 2 2 2 2 2 3 2 2 3 2 2 1
571] 2 3 3 2 2 2 3 2 3 2 3 2 2 3 3 3 3 2 3 3 2 3 2 3 2 3 3 2 3 2 1 2 2 3 2 2 2 3 2 3 3 3 2 2 2 3 3 3 2 2 3 2 2 2 2 3 2
628] 3 2 3 3 3 3 2 3 3 3 2 2 3 3 3 2 3 3 3 2 3 3 2 2 2 3 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 3 2 3 2 2 2 3 2 2 3 2 2 2 3
685] 3 3 2 3 3 3 3 3 3 2 2 2 2 2 2 2 3 3 2 2 2 1 2 3 2 3 3 3 3 3 1 3 3 3 3 2 2 2 3 3 2 2 2 1 3 3 2 3 3 3 2 2 2 2 3 3 2
742] 2 2 1 2 2 3 3 2 2 2 3 2 3 3 3 3 2 3 2 3 3 3 3 1 3 2 3 3 2 3 3 2 3 3 2 3 3 2 3 3 2 2 3 2 2 2 2 3 2 2 2 3 2 3 3 2 3
799] 2 2 2 2 2 2 2 3 3 2 2 3 3 2 2 1 3 3 3 3 2 3 3 3 3 2 3 2 2 2 2 3 2 2 2 2 3 3 2 2 2 3 3 3 2 3 3 3 3 3 2 2 3 3 2 2 2
856] 3 3 2 2 3 3 3 2 2 3 2 2 3 3 2 2 1 2 3 2 3 2 2 3 2 3 1 3 3 2 2 3 2 2 2 3 2 3 1 3 2 3 3 3 3 3 2 2 3 3 2 2 3 2 3 2 2
913] 3 3 2 3 2 2 3 3 2 3 1 2 2 2 2 2 2 3 3 2 3 2 1 2 2 2 2 3 2 3 3 3 3 2 3 2 2 3 2 3 2 2 2 3 2 2 3 2 2 3 2 1 2 2 2 3 3
970] 2 2 2 3 2 3 2 3 3 3 3 3 3 2 3 2 2 3 3 2 3 3 2 2 3 3 1 2 2 2 3
 reached getOption("max.print") -- omitted 9000 entries ]
```

CONCLUSION

Multivariate analysis was extremely important in the analysis of this database, which had a large number of variables to be analyzed simultaneously. Because only through multivariate analysis can we define the groups of variables, which variables had the greatest significance to be analyzed.

The PCA showed some missing data in the database after we corrected these data, we also took out the qualitative variables that cannot be analyzed by this kind of statistical method and then we eliminated some variables that were not relevant for the analysis, The PCA also showed the direction of these variables to their dimensions, if these variables had a positive or negative significance, as for the outliers it also showed us which variable had some discrepant data.

Cluster analysis allowed us to see these variables in a clearer way, showing the groups of clusters and their positioning in the graph on the graph, as well as being able to separate the numbers of clusters that would be of greater importance for the study. The ideal number of clusters was 3, as shown in the last graph, we can see each group properly sorted and separated and their proportions.

In the last graph, second result of the ideal number of clusters to be studied, we related the time that the customer stayed in the bank with the number of products they had and if among these products any was a credit card.

According to the cluster graph, we can see that the variable length of stay of the customer at the bank (TENURE) is not related to whether or not the customer has a financial product and whether it is a credit card. The other two variables, which are the group of people who have a financial product and the people who have a credit card, are related even in a small way.

The Multivariate Statistics is used to cross variables and through this crossing to analyze some type of problem within a group, in this case it was used to identify the group of people who still don't have a credit card because the group appeared with a large percentage of people without this service, which could give direction to the company to analyze why these people still do not have this financial product and intensify the offer of this product within the company.

ANA PAULA DE SOUZA VANDERLEY

SALVADOR - 2019