Exercise PCA Session 13-01-2020

Daniel Ferreira Zanchetta and Lais Silva Almeida Zanchetta

1. Lea el fichero “churn.txt”. Razone si realizar un ACP con datos estandarizados o sin estandarizar.

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)
library(mice)

## Warning: package 'mice' was built under R version 3.6.2

## Loading required package: lattice

##
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':
##
##     complete

## The following objects are masked from 'package:base':
##
##     cbind, rbind

setwd("C:/Users/Daniel/Documents/Certificados & Faculdade/UPC Master Big
Data/Data Analytics/Aula 8 - 08-01/exer_cart")
churn <- read.table(file = "churn.txt",header = TRUE,sep = "")
churn$antig[churn$antig==99] <- NA
churn <- mice::complete(mice(churn, m=1))

##
##  iter imp variable
##   1   1  antig  Nomina  Pension  Debito_aff  VISA  MCard
##   2   1  antig  Nomina  Pension  Debito_aff  VISA  MCard
##   3   1  antig  Nomina  Pension  Debito_aff  VISA  MCard
```

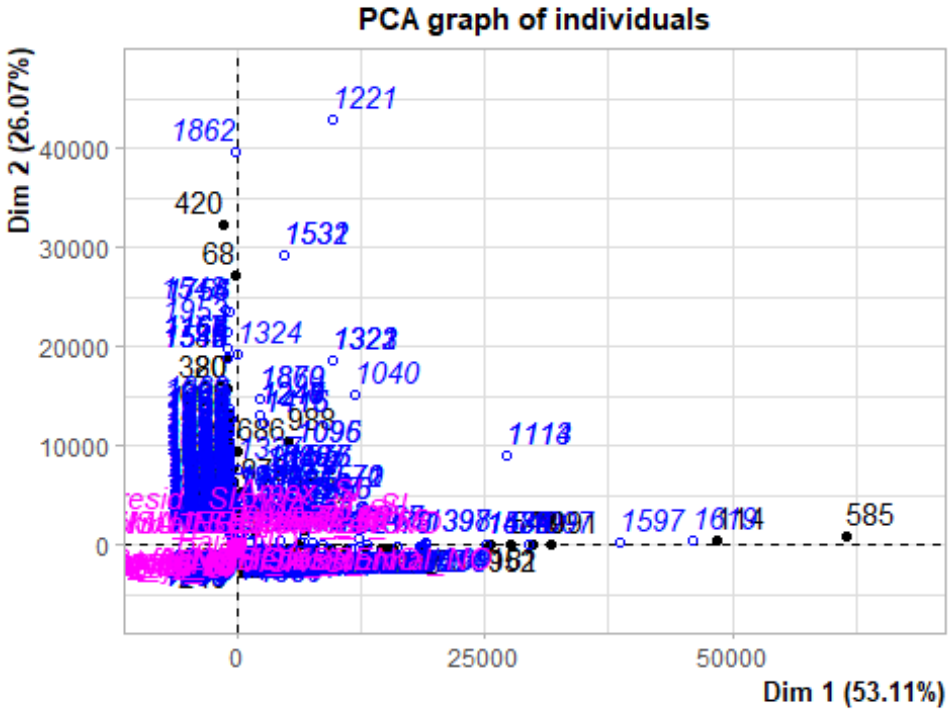
```
##   4   1  antig  Nomina  Pension  Debito_aff  VISA  MCard
##   5   1  antig  Nomina  Pension  Debito_aff  VISA  MCard

churn_tidy <- churn %>%
  separate(Baja, into = c("Baja_Rem","Baja"), sep = " ", extra = "merge",
fill = "left") %>%
  separate(edatcat, into = c("edatcat_Rem",
"edatcat","edatcat_Rem2","edatcat_Rem3"), sep = "([\\ \\.\\.])", extra =
"merge", fill = "right") %>%
  separate(Nomina, into = c("Nomina_Rem", "Nomina"), sep = " ", extra =
"merge", fill = "left") %>%
  separate(Pension, into = c("Pension_Rem", "Pension"), sep = " ", extra
= "merge", fill = "left") %>%
  separate(Debito_normal, into =
c("Debito_normal_Rem","Debito_normal_Rem2", "Debito_normal"), sep = "([\\
\\ ])", extra = "merge", fill = "left") %>%
  separate(Debito_aff, into = c("Debito_aff_Rem","Debito_aff_Rem2",
"Debito_aff"), sep = "([\\ \\. ])", extra = "merge", fill = "left") %>%
  separate(VISA, into = c("VISA_Rem", "VISA"), sep = " ", extra =
"merge", fill = "left") %>%
  separate(VISA_aff, into = c("VISA_aff_Rem","VISA_aff_Rem2",
"VISA_aff"), sep = "([\\ \\. ])", extra = "merge", fill = "left") %>%
  separate(MCard, into = c("MCard_Rem", "MCard"), sep = " ", extra =
"merge", fill = "left") %>%
  separate(Amex, into = c("Amex_Rem", "Amex"), sep = " ", extra =
"merge", fill = "left") %>%
  separate(dif_resid, into = c("dif_resid_Rem","dif_resid_Rem2",
"dif_resid"), sep = "([\\ \\. ])", extra = "merge", fill = "left") %>%
  transform(sexo = ifelse(.$sexo == "No informado", "MUJER", "HOMBRE"))
%>%
  select(-c("Baja_Rem",
"edatcat_Rem","edatcat_Rem2","edatcat_Rem3","Nomina_Rem","Pension_Rem","D
ebito_normal_Rem","Debito_normal_Rem2","Debito_aff_Rem","Debito_aff_Rem2"
,"VISA_Rem","VISA_aff_Rem","VISA_aff_Rem2","MCard_Rem","Amex_Rem","dif_re
sid_Rem","dif_resid_Rem2"))
```

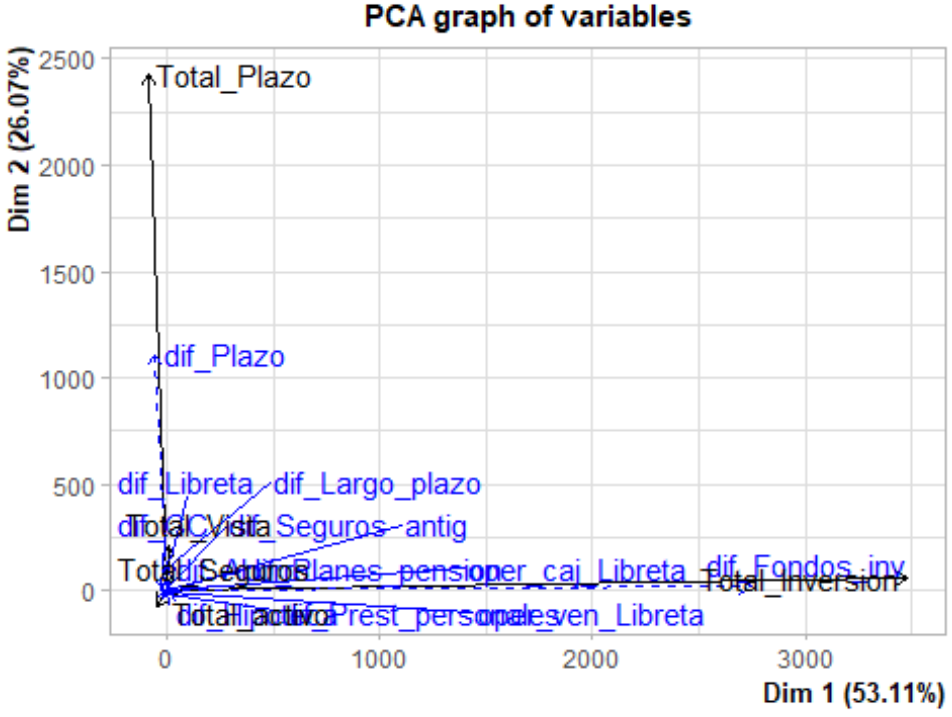
Resp.: Vamos a optar por no realizar la estandarización de los datos de las 5 variables activas (Total_activo, Total_plazo, Total_inversion, Total_Seguros y Total_Vista) pues nuestra asunción es de que estas variables estan medidas en las mismas unidades, que en este caso es unidad monetaria.

2. Efectúe un Análisis de Componentes Principales tomando como variables activas los productos bancarios antes especificados. Declare como ilustrativos los clientes que se dieron de baja (de esta forma la configuración obtenida reflejará la de los clientes "normales").

```
library(FactoMineR)
```

```r
#Función PCA, utilizando el scale.unit = FALSE pues hemos elegido no estandarizar las variables activas. Tambien estamos utilizando como suplementarios para test los individuos con Baja SI
pca.churn <- PCA(churn_tidy,quanti.sup = c(4,19:30),quali.sup = c(1:3,5:12,18),ind.sup = (1001:2000),scale.unit = FALSE)
```


PCA graph of individuals
Dim 2 (26.07%)
Dim 1 (53.11%)
1221
1862
420
68
380
1040
585
0
25000
50000
10000
20000
30000
40000

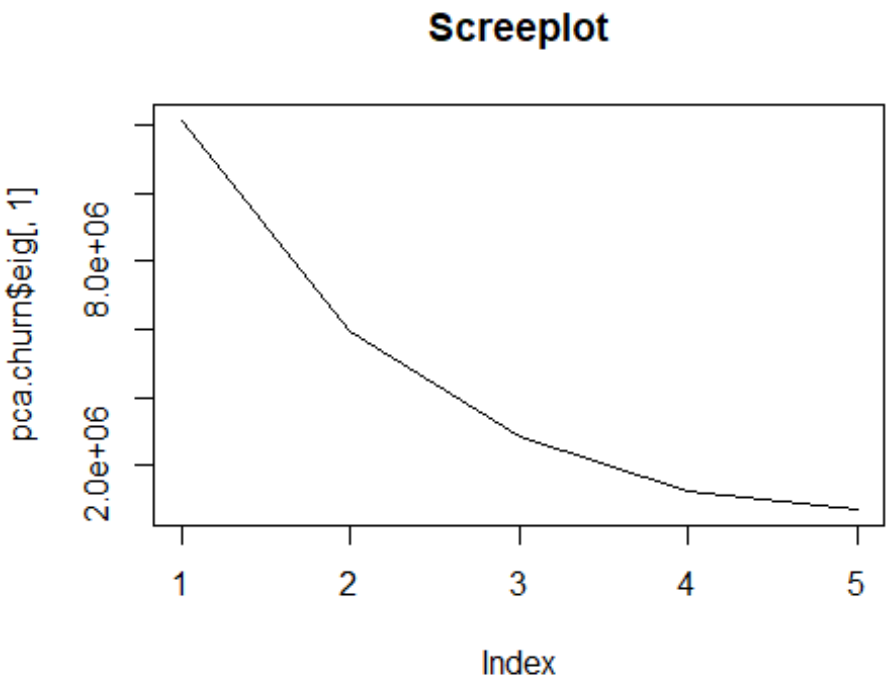

PCA graph of variables
Dim 2 (26.07%)
Dim 1 (53.11%)
Total_Plazo
dif_Plazo
dif_Libreta
dif_Largo_plazo
Total_Inversion
dif_Fondos_inv
0
1000
2000
3000
500
1500
2500

3. Obtenga la representación gráfica del ”Screeplot” (diagrama de los “eigenvalues”) y a la vista de las correlaciones entre las variables originales y las componentes principales, decida el número de dimensiones significativas. ¿Cuál es el porcentaje de variancia retenido?

```
pca.churn$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of
variance
## comp 1 12148095.6              53.105708
53.10571
## comp 2  5962949.5              26.067185
79.17289
## comp 3  2847991.0              12.450065
91.62296
## comp 4  1219523.9               5.331180
96.95414
## comp 5   696750.7               3.045863
100.00000
```

```
#Grafico Screeplot
plot(pca.churn$eig[,1],type="l",main="Screeplot")
```




```
#Proyección de individuos con los componentes principales
pca.churn$var$cor[,1:3]
```

```
##                         Dim.1       Dim.2        Dim.3
## Total_activo    -0.027974404 -0.04529893  0.998518858
## Total_Plazo     -0.036334447  0.99864018  0.021831305
## Total_Inversion  0.999824817  0.01731661  0.006920594
## Total_Seguros   -0.012830626 -0.00487857 -0.024866937
## Total_Vista      0.005318147  0.25202229 -0.009219268
```

R.: Numero de Dimensiones es igual a 3. Porcentaje de variancia retenido entre los dos componentes es de 91.62% de la nube original.

4. Efectúe una rotación “varimax” para hacer más evidente los factores latentes (intangibles) presentes en sus datos activos. ¿Cuáles son en este caso estos factores latentes?. (La rotación tendrá más o menos sentido en función de si se ha optado por realizar un ACP con los datos estandarizados o no).

```
nd <- 3
pca.rotation <- varimax(pca.churn$var$cor[,1:nd])
pca.rotation$loadings

##
## Loadings:
##                 Dim.1  Dim.2  Dim.3
## Total_activo    -0.112          0.991
## Total_Plazo             0.998
## Total_Inversion  0.996
## Total_Seguros
## Total_Vista             0.252
##
##                Dim.1 Dim.2 Dim.3
## SS loadings    1.005 1.065 0.993
## Proportion Var 0.201 0.213 0.199
## Cumulative Var 0.201 0.414 0.613
```

R.: Los factores latentes para cada dimensión son los siguientes: Dimensión 1 -> Total Inversión Dimensión 2 -> Total Plazo Dimensión 3 -> Total activo Es decir, los factores tanto en relación a Inversión que tienen los individuos y sus Depósitos (totales de activos y totales plazo) son significativos y están correlacionados con las bajas de estos individuos.

5. Represente gráficamente la nube de puntos individuo activos. Sobre esta nube proyecte los individuos suplementarios. ¿Piensa Ud. que la configuración de los clientes que han sido baja es distinta de la de los clientes que no han sido baja?

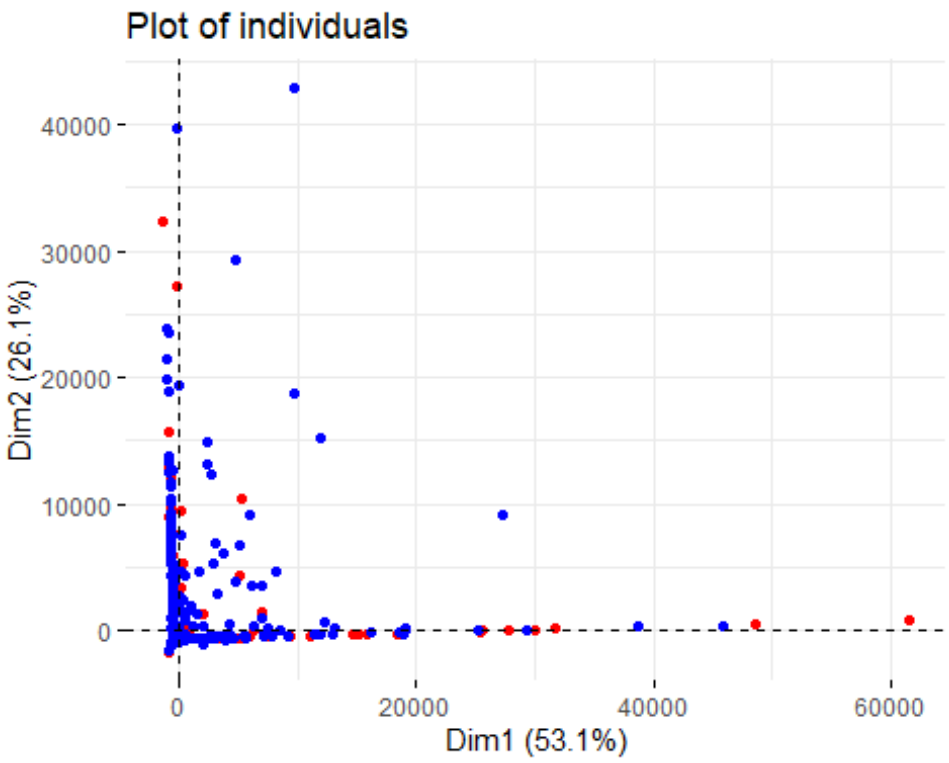
```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 3.6.2
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

```
fviz_pca_ind(pca.churn, geom.ind = "point",
             col.ind = "red",
             col.ind.sup = "blue",
             axes = c(1, 2),
             title = "Plot of individuals",
             label = "quali",
             pointsize = 1.5)
```



R.: Nuestra conclusión es que no es diferente su configuración.

6. Calcule el centroide (=punto medio) de los clientes que han sido baja en las componentes principales retenidas en la pregunta 3. A partir de la fórmula, calcule el v.test (= test value) del centroide de los clientes que han sido baja y diga si este centroide ocupa una posición significativamente distinta de los clientes "no baja", en cada una de las componentes principales retenidas.

```
n <- nrow(churn_tidy)
nsup <- nrow(pca.churn$ind.sup$coord)

psisupdim1 <- mean(pca.churn$ind.sup$coord[,1])
psisupdim2 <- mean(pca.churn$ind.sup$coord[,2])
psisupdim3 <- mean(pca.churn$ind.sup$coord[,3])

lambda1 <- pca.churn$eig[1,1]
lambda2 <- pca.churn$eig[2,1]
lambda3 <- pca.churn$eig[3,1]

v.testdim1 <- (psisupdim1)/sqrt((1-(nsup/n))*(lambda1/nsup))
v.testdim1

## [1] 8.926452

vtestdim2 <-  (psisupdim2)/sqrt((1-(nsup/n))*(lambda2/nsup))
vtestdim2

## [1] 24.76995

vtestdim3 <-  (psisupdim3)/sqrt((1-(nsup/n))*(lambda3/nsup))
vtestdim3

## [1] 17.70009
```

R.: Teniendo en cuenta las dimensiones encontradas en enunciado 4 (abajo):

- Dimensión 1 -> Total Inversión
- Dimensión 2 -> Total Plazo
- Dimensión 3 -> Total activo

Y basándonos por el v.test resultado, hemos percibido que todas las dimensiones representan una diferencia significativa entre los individuos que han sido Baja, de los activos (que no han sido baja), siendo lo de Dimensión 2 (Total Plazo) la que es más representativa, seguida por el Total activo. En relación a total de inversión, aunque hemos notado un valor de v.test alto de los Baja para lo de no Bajas, no ha sido tan expresivo cuanto las demás dimensiones. Con esto, en nuestra análisis vimos que los individuos que son baja NO se comportan de igual forma que los individuos que no son baja, cuanto es considerado los productos bancario en cuestión.