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))

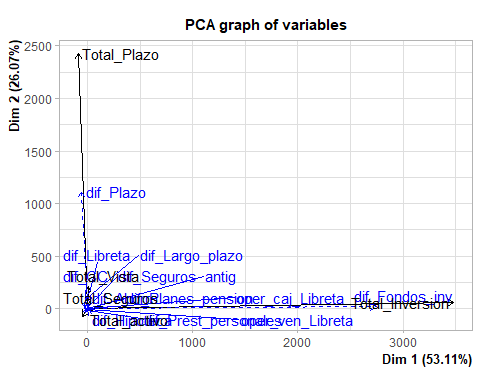
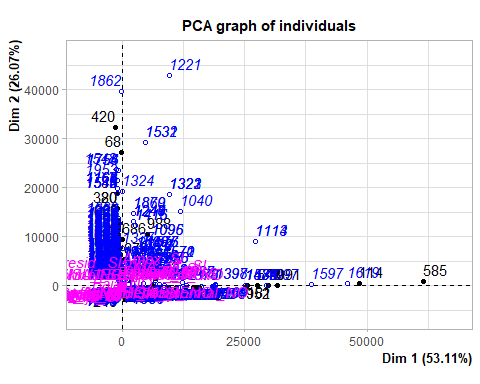
##   
## iter imp variable  
## 1 1 antig Nomina Pension Debito\_aff VISA MCard  
## 2 1 antig Nomina Pension Debito\_aff VISA MCard  
## 3 1 antig Nomina Pension Debito\_aff VISA MCard  
## 4 1 antig Nomina Pension Debito\_aff VISA MCard  
## 5 1 antig Nomina Pension Debito\_aff VISA MCard

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","Debito\_normal\_Rem","Debito\_normal\_Rem2","Debito\_aff\_Rem","Debito\_aff\_Rem2","VISA\_Rem","VISA\_aff\_Rem","VISA\_aff\_Rem2","MCard\_Rem","Amex\_Rem","dif\_resid\_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)  
  
#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)

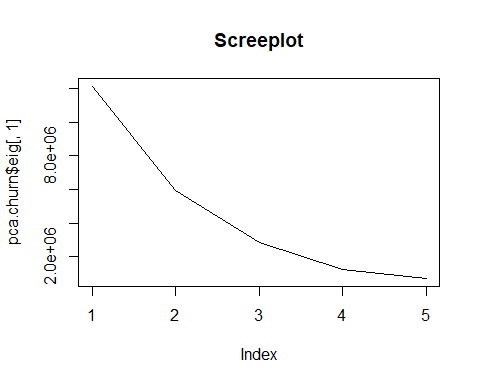


## 3. Obtenga la representación gráfica del ”Screeplot” (diagrama de los “eigenvalues”) y a la vista de las correlaciones entre las variables originales y las componentes principales, decida el número de dimensiones significativas. ¿Cuál es el porcentaje de variancia retenido?

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

Resp.: Los factores latentes para cada dimensión son los siguientes: Dimensión 1 -> Total Inversion Dimensión 2 -> Total Plazo Dimensión 3 -> Total activo Es decir, los factores tanto en relación a Inversión que tienen los individuos y sus Depositos (totales de activos y totales plazo) son significativos y estan 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?

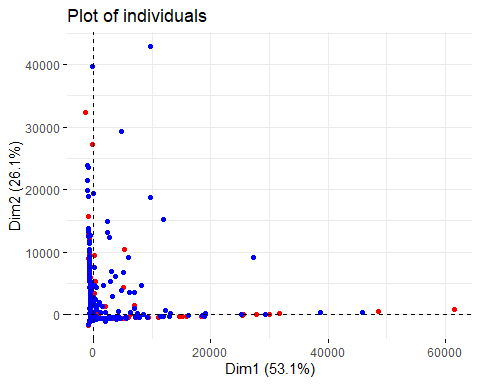
library(factoextra)

## 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)

 Resp.: 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

Resp.: Teniendo en cuenta las dimensiones encontradas en enunciado 4 (abajo): Dimensión 1 -> Total Inversion Dimensión 2 -> Total Plazo Dimensión 3 -> Total activo Y basandonos 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 mas 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 demas dimensiones. Con esto, en nuestra analisis 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.