Practica 2

Juan Cantero Jimenez

1/30/2022

Tarea 3

Recuperamos el banco de datos de la práctica 1 que encontrarás en el fichero datos___prac1___ok.RData. Recuerda que las dos primeras variables simplemente identifican cada provincia con su código y su nombre.

a. Calcula las desviaciones típicas de las variables cuantitativas del banco de datos. Responde, a la vista del significado de las variables y del resultado anterior, si consideras que se debería realizar el análisis de componentes principales con la matriz de varianzas-covarianzas o con la de correlaciones.

```
describe_custom <- function(data){
  require(e1071)
  result <- apply(data, 2, function(x){


    c(media=mean(x),
      mediana=median(x),
      varianza = var(x),
      des_tipic = sd(x),
      skew = e1071::skewness(x),
      kurto = e1071::kurtosis(x),
      maximo = max(x),
      minimo = min(x),
      rango = max(x)- min(x),
      quantile(x , 0.25 ),
      quantile(x, 0.50),
      quantile(x, 0.75),
      shapiro_pvalor = shapiro.test(x)$p.value)

  })
  return(result)
}

load("datos_prac1_ok.RData")
cod_nombre <- datos.p1[, 1:2]
datos.p1 <- datos.p1[,-(1:2)]
numeric_des <- describe_custom(datos.p1)
```

```
## Loading required package: e1071
```

```
numeric_des
```

```
##                 PobTot2018 PorcVarPob2000_2018 PorcMenores16_2018
```

```
## media          8.985188e+05          11.1576923       15.015384615
## mediana        6.091640e+05           9.6500000       15.450000000
## varianza       1.379157e+12         212.0628808        5.636229261
## des_tipic      1.174375e+06          14.5623790        2.374074401
## skew           3.424991e+00           0.5180954       -0.006526281
## kurto          1.264308e+01          -0.2080464        0.419100591
## maximo         6.578079e+06          53.8000000       22.200000000
## minimo         8.514400e+04         -14.2000000       10.000000000
## rango          6.492935e+06          68.0000000       12.200000000
## 25%            3.255698e+05           0.4500000       13.750000000
## 50%            6.091640e+05           9.6500000       15.450000000
## 75%            1.020944e+06          20.1750000       16.350000000
## shapiro_pvalor 4.625438e-11           0.2171123        0.236300056
##                PorcMayores65_2018 PorcPobExtranjera_2018 EdadMedia2018
## media                  20.5134615            8.597692308    44.0188462
## mediana                19.5000000            8.000000000    43.3350000
## varianza               17.8764819           23.081453394     9.2189555
## des_tipic               4.2280589            4.804316121     3.0362733
## skew                    0.4675219            0.586304437    -0.1156349
## kurto                  -0.1559526           -0.699257688     0.3899858
## maximo                 31.2000000           19.700000000    50.4900000
## minimo                 11.4000000            2.400000000    35.2200000
## rango                  19.8000000           17.300000000    15.2700000
## 25%                    17.6750000            4.000000000    42.1025000
## 50%                    19.5000000            8.000000000    43.3350000
## 75%                    22.8000000           11.975000000    45.7000000
## shapiro_pvalor          0.3083571            0.004143023     0.2796724
##                TBNatalidad2018 TasaBrutaMortalidad TasaMortalidadMenores5anyos
## media             7.663077e+00          10.1563462                3.472500000
## mediana           7.660000e+00           9.9350000                3.385000000
## varianza          2.926143e+00           5.2993021                1.588666176
## des_tipic         1.710597e+00           2.3020213                1.260423015
## skew              2.144744e+00           0.5516077                1.062593464
## kurto             8.089309e+00          -0.2442150                2.144176781
## maximo            1.583000e+01          15.7500000                8.040000000
## minimo            4.820000e+00           6.1000000                1.140000000
## rango             1.101000e+01           9.6500000                6.900000000
## 25%               6.627500e+00           8.6200000                2.672500000
## 50%               7.660000e+00           9.9350000                3.385000000
## 75%               8.222500e+00          11.5775000                4.145000000
## shapiro_pvalor    2.147524e-06           0.1280758                0.008399454
##                EsperanzaVidaH2018 EsperanzaVidaM2018 PorcParoAgricultura
## media                  80.3409615       85.629423077        6.796154e+00
## mediana                80.3500000       85.880000000        5.000000e+00
## varianza                0.9943696        1.188350641        2.751175e+01
## des_tipic               0.9971808        1.090114967        5.245164e+00
## skew                   -0.1318343       -1.132546585        1.317765e+00
## kurto                  -0.6095486        1.496888083        1.851682e+00
## maximo                 82.1800000       87.290000000        2.370000e+01
## minimo                 78.2300000       82.130000000        2.000000e-01
## rango                   3.9500000        5.160000000        2.350000e+01
## 25%                    79.6900000       85.137500000        3.425000e+00
## 50%                    80.3500000       85.880000000        5.000000e+00
## 75%                    80.9275000       86.302500000        9.975000e+00
```

```
## shapiro_pvalor          0.5940608        0.001228524        7.258436e-05
##                PorcParoIndustria PorcParoConstruccion PorcParoServicios
## media                 13.6250000            6.1769231       65.97115385
## mediana               13.3000000            6.1500000       65.00000000
## varianza              37.1403431            1.7771041       39.27934766
## des_tipic              6.0942877            1.3330807        6.26732380
## skew                   0.1197387           -0.1973129        0.53126948
## kurto                 -0.6065349            0.3998551       -0.33291461
## maximo                27.0000000            9.1000000       80.00000000
## minimo                 1.5000000            2.4000000       52.80000000
## rango                 25.5000000            6.7000000       27.20000000
## 25%                    8.7000000            5.5500000       60.95000000
## 50%                   13.3000000            6.1500000       65.00000000
## 75%                   17.6500000            6.8250000       70.12500000
## shapiro_pvalor         0.7224351            0.5256164        0.04406381
##                PorcParoOtros
## media             7.42115385
## mediana           7.00000000
## varianza          8.15072021
## des_tipic         2.85494662
## skew              0.80807395
## kurto             0.60887189
## maximo           15.90000000
## minimo            1.90000000
## rango            14.00000000
## 25%               5.47500000
## 50%               7.00000000
## 75%               9.10000000
## shapiro_pvalor    0.03020272
```

Será necesario el uso de la matriz de correlaciones debido a que existe una gran diferencia de escala en los datos, tanto en valor absoluto como de varianzas.

b. Realiza un Análisis de Componentes Principales sobre este banco de datos en el modo en el que hayas justificado en el primer apartado y responde las siguientes preguntas:

```
pca <- princomp(datos.p1, cor=TRUE)
summary(pca)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4    Comp.5
## Standard deviation     2.2607359 1.8096841 1.4611593 1.18603695 1.1243074
## Proportion of Variance 0.3194329 0.2046848 0.1334367 0.08791773 0.0790042
## Cumulative Proportion  0.3194329 0.5241177 0.6575544 0.74547212 0.8244763
##                           Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
## Standard deviation     0.84422875 0.79495756 0.66104100 0.58047880 0.49955116
## Proportion of Variance 0.04454514 0.03949735 0.02731095 0.02105973 0.01559696
## Cumulative Proportion  0.86902145 0.90851880 0.93582975 0.95688948 0.97248644
##                           Comp.11     Comp.12     Comp.13     Comp.14
## Standard deviation     0.46840774 0.327498050 0.256113938 0.194823512
## Proportion of Variance 0.01371286 0.006703436 0.004099647 0.002372263
## Cumulative Proportion  0.98619930 0.992902734 0.997002381 0.999374643
##                           Comp.15      Comp.16
## Standard deviation     0.0998918915 5.226489e-03
```

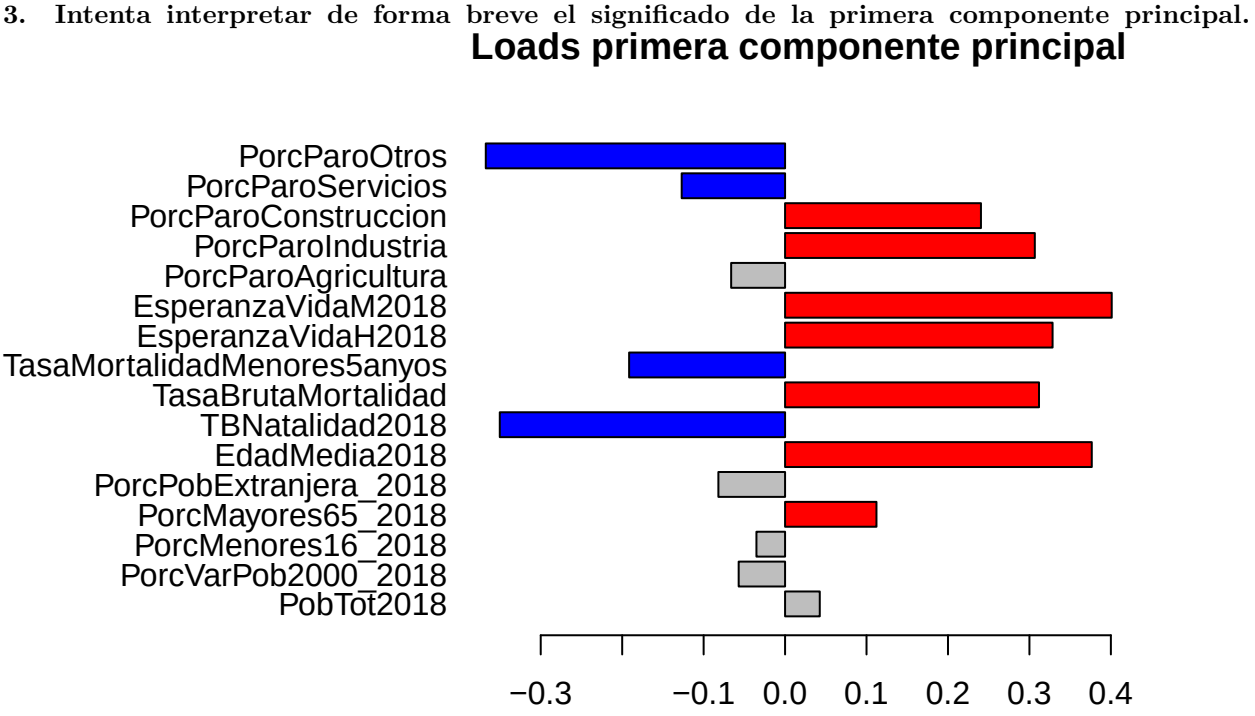
```
## Proportion of Variance 0.0006236494 1.707262e-06
## Cumulative Proportion  0.9999982927 1.000000e+00
```

1. ¿Qué porcentaje de varianza del banco de datos original explica la primera componente principal? ¿Y la segunda?. Respectivamente el 0.319432 y 0.20468448.

2. ¿Con cuántas componentes principales nos deberíamos quedar si queremos mantener al menos el 90% de la varianza del banco de datos orignal? Serán necesarias las 7 primeras componentes principales para recoger un 90 % de la variabilidad.

```
loads <- as.matrix(pca$loadings)
colors <- rep("grey", nrow(loads))
colors[loads[,1] > 0.1] <- "red"
colors[loads[,1] < -0.1] <- "blue"
par(mar= c(5.1, 14.1, 4.1, 2.1))
barplot(loads[,1],horiz = T,las=1,main = "Loads primera componente principal", col=colors)
```

3. Intenta interpretar de forma breve el significado de la primera componente principal.

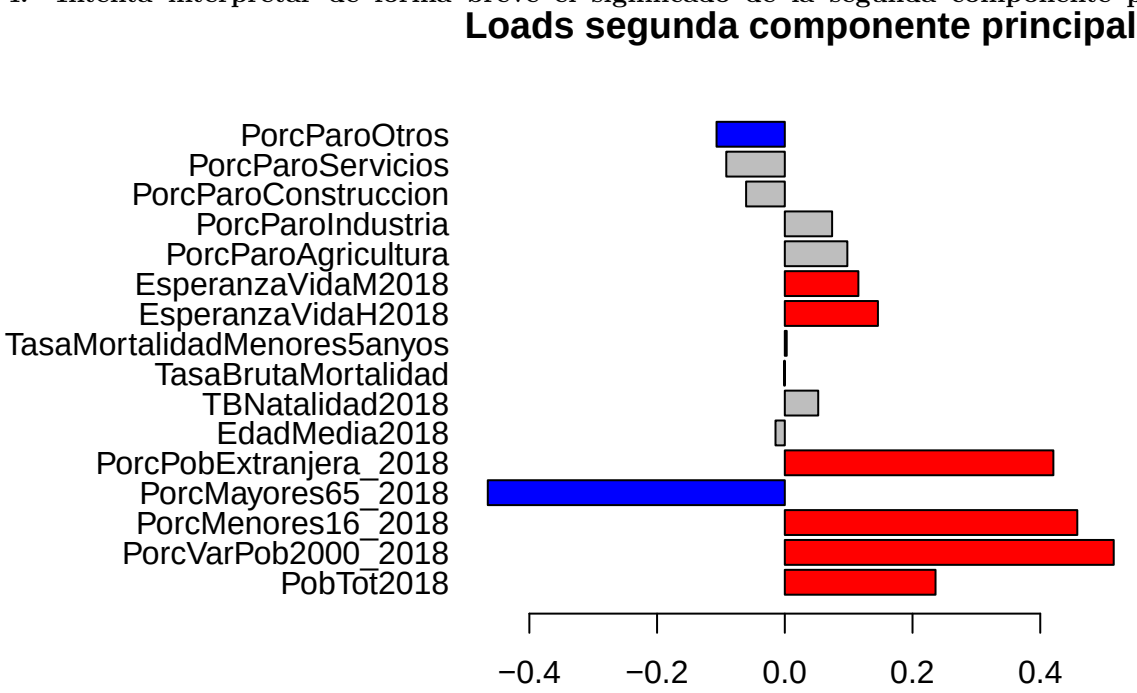


En el gráfico anterior se han representado aquellas variables que superan una carga de 0.1 en rojo, las que son menores a -0.1 en azul y las demás en gris. Aquellas observaciones que posean un valor absoluto alto en las variables PorcParoConstruccion, PorcParoIndustria, EsperanzaVidaM2018, EsperanzaVidaH2018, TasaBrutaMortalidad, EdadMedia2018 y PorcMayores65_2018 poseeran un valor positivo en esta componente. A su vez, aquellas que posean un valor absoluto alto en PorcParoOtros, PorParoServicios, TasaMortalidad-Menores5anyos y TBNatalidad2018 podran recibir un valor negativo en esta primera componente. Debido a la heterogeneidad de las variables, no veo conveniente aportar una interpretación más holística a la componente.

```
colors <- rep("grey", nrow(loads))
colors[loads[,2] > 0.1] <- "red"
colors[loads[,2] < -0.1] <- "blue"
```

```
par(mar= c(5.1, 14.1, 4.1, 2.1))
barplot(loads[,2],horiz = T,las=1,main = "Loads segunda componente principal", col=colors)
```

4. Intenta interpretar de forma breve el significado de la segunda componente principal.



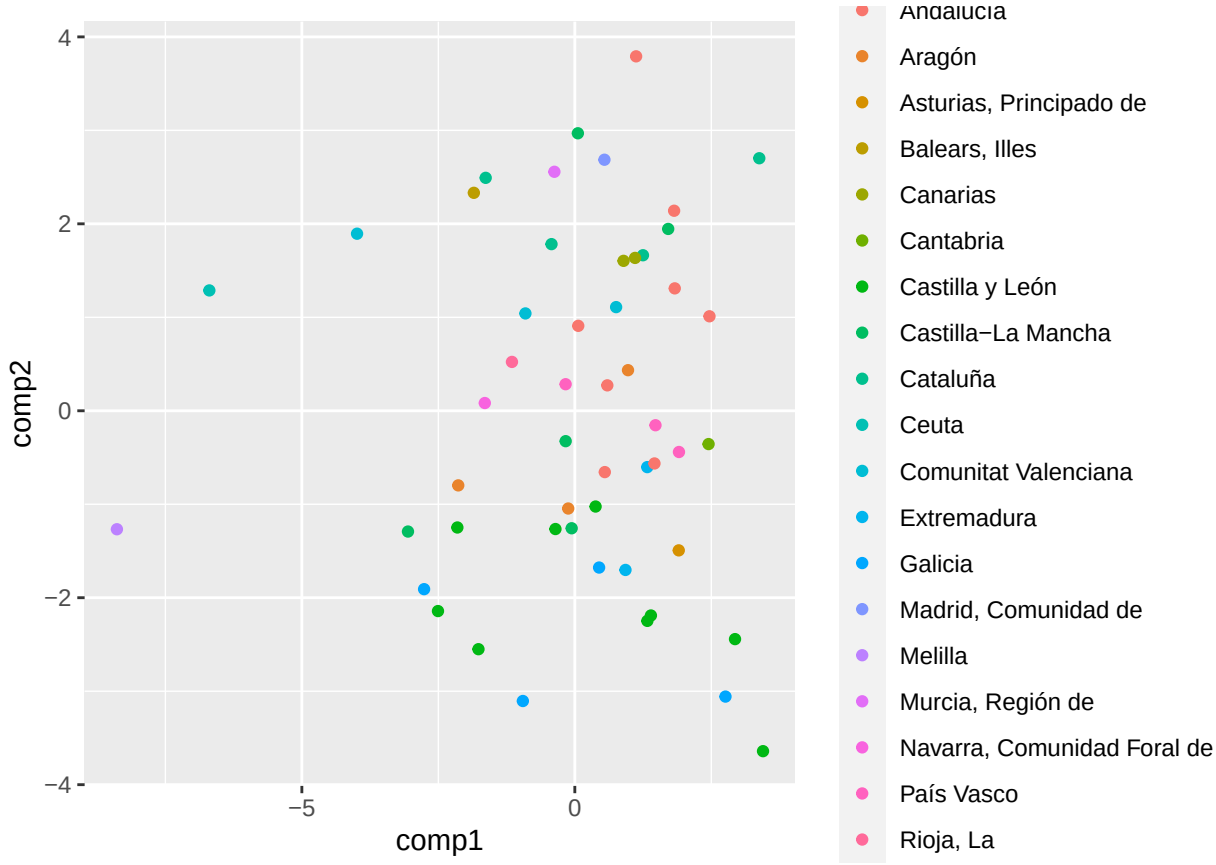
Como se puede ver en la figura anterior, esta componente principal dará valores positivos altos a aquellas observaciones que posean una población no envejecida, con un alto porcentaje de población extranjera. Además esto se vera acrecentado si la observación posee una alta esperanza de vida y una alta población total.

```
comunidades_provincias <- read.csv("ComunidadesProvincias.csv",sep = ";",encoding="latin1")

cod_nombre_color <- t(apply(cod_nombre, 1, function(x, y){
  c(CodProv=as.numeric(x[1]), x[2], CCAA=y[, "CCAA"][y[,"Cprov"]==as.numeric(x[1])])
}, y = comunidades_provincias))

cod_nombre_color <- as.data.frame(cod_nombre_color)
gather <- data.frame(comp1 = pca$scores[,1], comp2 = pca$scores[,2],
                     coloring = factor(cod_nombre_color$CCAA))
library(ggplot2)
ggplot(data=gather, aes(x=comp1, y=comp2, col=coloring))+geom_point()
```

5. Representa todas las provincias en un gráfico según las dos primeras componentes principales y comenta el resultado, resaltando las que tengan algún comportamiento que te llame la atención (por ejemplo, las más extremas en alguna de las dos primeras componentes). Puedes utilizar las Comunidades Autónomas a las que pertencen (la relación está disponible entre los ficheros de la práctica 1) para marcar en diferentes colores, por ejemplo, las provincias de cada comunidad, y así comprobar si aparecen cercanas en el gráfico.



En la imagen anterior se puede observar como las ciudades autónomas de Ceuta y Melilla se encuentran alejadas del resto de comunidades autónomas. Esto se da princpalmente debido al valor que poseen en la primera componente principal en la que poseen un alto valor negativo, el valor de la segunda componente es similar al resto de comunidades autónomas. Este comportamiento puede deberse a que posean un alto valor en las variables, PorcParoOtros, PorcParoServicios TasaMortalidadMenores5anyos y TBNatalidad2018.