Práctica 1

Minería de datos

Juan Cantero Jimenez

16 de enero, 2022

Tarea 1: Realiza un análisis univariante numérico del banco de datos.

Primero se cargan los datos y se inspecciona el formato de estos.

```
load("FicheroDatosP1.Rdata")
head(datos)
```

```
##   CodProv NombreProv PobTot2018 PorcVarPob2000_2018 PorcMenores16_2018
## 1       1      Álava     328868                14.8                15.9
## 2       2   Albacete     388786                 7.0                15.5
## 3       3   Alicante    1838819                27.2                15.9
## 4       4    Almería     709340                36.9                18.1
## 5       5      Ávila     158498                -3.9                12.8
## 6       6    Badajoz     676376                 2.2                15.5
##   PorcMayores65_2018 PorcPobExtranjera_2018 EdadMedia2018 TBNatalidad2018
## 1               20.5                    8.6         43.05            7.81
## 2               19.1                    6.1         43.39            7.66
## 3               19.5                   18.3         40.01           10.57
## 4               14.5                   19.7         44.16            8.22
## 5               25.8                    5.8         48.02            5.60
## 6               19.2                    2.8         47.23            6.15
##   TasaBrutaMortalidad TasaMortalidadMenores5anyos EsperanzaVidaH2018
## 1                9.57                       3.95              80.86
## 2                8.80                       2.97              80.35
## 3                7.90                       3.51              78.39
## 4                8.47                       4.47              81.60
## 5               12.93                       2.62              79.63
## 6               12.80                       5.11              81.07
##   EsperanzaVidaM2018 PorcParoAgricultura PorcParoIndustria PorcParoConstruccion
## 1              85.95                 8.3              13.7                 6.9
## 2              85.28                 4.5              15.1                 6.8
## 3              84.19                23.7               4.5                 4.4
## 4              87.05                 1.8              27.0                 5.4
## 5              85.49                 4.2              12.8                 5.9
## 6              85.96                 8.4              11.9                 7.4
##   PorcParoServicios PorcParoOtros
## 1              61.1           9.9
## 2              65.6           8.0
## 3              57.5           9.9
## 4              62.1           3.7
## 5              70.1           6.9
## 6              64.1           8.2
```

Se puede observar que el banco cuenta con 16 variables cuantitativas así como de dos variables cualitativas que permiten identificar el individuo, en este caso "CodProv" y "NombreProv" Para facilitar el trabajo se procederá a separar este banco de datos en variables cualitativas y cuantitativas. La referencia del individuo en el caso del banco de datos cuantitativo se almacenará en el nombre de la fila:

```
all(rownames(datos) == datos$CodProv) # se comprueba que el nombre de las filas

## [1] TRUE

#concuerde con la variable "CovProv"
datos_num <- datos[, which(colnames(datos) != "CodProv" & colnames(datos) != "NombreProv")]
head(datos_num)

##   PobTot2018 PorcVarPob2000_2018 PorcMenores16_2018 PorcMayores65_2018
## 1     328868                14.8               15.9                20.5
## 2     388786                 7.0               15.5                19.1
## 3    1838819                27.2               15.9                19.5
## 4     709340                36.9               18.1                14.5
## 5     158498                -3.9               12.8                25.8
## 6     676376                 2.2               15.5                19.2
##   PorcPobExtranjera_2018 EdadMedia2018 TBNatalidad2018 TasaBrutaMortalidad
## 1                    8.6         43.05            7.81                9.57
## 2                    6.1         43.39            7.66                8.80
## 3                   18.3         40.01           10.57                7.90
## 4                   19.7         44.16            8.22                8.47
## 5                    5.8         48.02            5.60               12.93
## 6                    2.8         47.23            6.15               12.80
##   TasaMortalidadMenores5anyos EsperanzaVidaH2018 EsperanzaVidaM2018
## 1                        3.95              80.86              85.95
## 2                        2.97              80.35              85.28
## 3                        3.51              78.39              84.19
## 4                        4.47              81.60              87.05
## 5                        2.62              79.63              85.49
## 6                        5.11              81.07              85.96
##   PorcParoAgricultura PorcParoIndustria PorcParoConstruccion PorcParoServicios
## 1                 8.3              13.7                  6.9              61.1
## 2                 4.5              15.1                  6.8              65.6
## 3                23.7               4.5                  4.4              57.5
## 4                 1.8              27.0                  5.4              62.1
## 5                 4.2              12.8                  5.9              70.1
## 6                 8.4              11.9                  7.4              64.1
##   PorcParoOtros
## 1           9.9
## 2           8.0
## 3           9.9
## 4           3.7
## 5           6.9
## 6           8.2

codigo_nombre <- datos[, which(colnames(datos) == "CodProv" | colnames(datos) == "NombreProv")]
```

A continuación se crea una función que proporciona algunos estadísticos discriptivos univariantes. Se

ha decidido añadir un test de normalidad Shapiro-Wilks para comprobar si la variable puede ser descrita por una distribución normal. El test Shapiro-Wilks toma como hipótesis nula que los datos provienen de una distribución normal, así p-valores por encima del nivel de significación indican que se acepta la hipótesis nula de normalidad.

```
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)
}
numeric_des <- describe_custom(datos_num)
```

```
## 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           2.474231e+01    44.0188462
## mediana                19.5000000           8.000000e+00    43.3350000
## varianza               17.8764819           1.357293e+04     9.2189555
## des_tipic               4.2280589           1.165029e+02     3.0362733
```

```
## skew                   0.4675219          6.782420e+00    -0.1156349
## kurto                 -0.1559526          4.495095e+01     0.3899858
## maximo                31.2000000          8.480000e+02    50.4900000
## minimo                11.4000000          2.400000e+00    35.2200000
## rango                 19.8000000          8.456000e+02    15.2700000
## 25%                   17.6750000          4.000000e+00    42.1025000
## 50%                   19.5000000          8.000000e+00    43.3350000
## 75%                   22.8000000          1.230000e+01    45.7000000
## shapiro_pvalor         0.3083571          9.117069e-16     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
```

Si se atendemos al vector de medias se podrá observar como los datos se encuentran en escalas muy diferentes, principalmente la variable PobTot2018 que esta en el rango del millón. Para poder comparar las distintas variables entre sí se decide realizar una tipificación de las variables. Esta tipificación debe de entenderse como un cambio de variables.

```
numeric_des["media",]
```

```
##                   PobTot2018         PorcVarPob2000_2018
##                 8.985188e+05                1.115769e+01
##           PorcMenores16_2018          PorcMayores65_2018
##                 1.501538e+01                2.051346e+01
##       PorcPobExtranjera_2018               EdadMedia2018
##                 2.474231e+01                4.401885e+01
##              TBNatalidad2018         TasaBrutaMortalidad
##                 7.663077e+00                1.015635e+01
## TasaMortalidadMenores5anyos          EsperanzaVidaH2018
##                 3.472500e+00                8.034096e+01
##           EsperanzaVidaM2018         PorcParoAgricultura
##                 8.562942e+01                6.796154e+00
##            PorcParoIndustria        PorcParoConstruccion
##                 1.362500e+01                6.176923e+00
##            PorcParoServicios               PorcParoOtros
##                 6.597115e+01                7.421154e+00
```

```
z_value <- function(data){
  mean_sd <- rbind(media=apply(data, 2, mean), sdd = apply(data, 2, sd))
  data_num_z <- t(apply(data, 1, function(x,y){
    (x - y["media", ])/(y["sdd",])
  }, y = mean_sd))
  return(data_num_z)
}
head(z_value(datos_num))
```

```
##      PobTot2018 PorcVarPob2000_2018 PorcMenores16_2018 PorcMayores65_2018
## [1,] -0.4850672           0.2501176          0.3726149       -0.003183858
## [2,] -0.4340460          -0.2855091          0.2041281       -0.334305074
## [3,]  0.8006813           1.1016269          0.3726149       -0.239699012
```

```
## [4,] -0.1610889           1.7677268          1.2992918       -1.422274786
## [5,] -0.6301401          -1.0340132         -0.9331572        1.250346462
## [6,] -0.1891583          -0.6151256          0.2041281       -0.310653559
##      PorcPobExtranjera_2018 EdadMedia2018 TBNatalidad2018 TasaBrutaMortalidad
## [1,]            -0.13855709   -0.31909056     0.085889924          -0.2547093
## [2,]            -0.16001577   -0.20711118    -0.001798742          -0.5891979
## [3,]            -0.05529738   -1.32031796     1.699361384          -0.9801587
## [4,]            -0.04328052    0.04648918     0.325572279          -0.7325502
## [5,]            -0.16259082    1.31778449    -1.206056425           1.2048776
## [6,]            -0.18834124    1.05759711    -0.884531316           1.1484055
##      TasaMortalidadMenores5anyos EsperanzaVidaH2018 EsperanzaVidaM2018
## [1,]                  0.37884107        0.520505846          0.2940763
## [2,]                 -0.39867568        0.009064014         -0.3205378
## [3,]                  0.02975192       -1.956477143         -1.3204324
## [4,]                  0.79140097        1.262597916          1.3031441
## [5,]                 -0.67636023       -0.712971513         -0.1278976
## [6,]                  1.29916701        0.731099542          0.3032496
##      PorcParoAgricultura PorcParoIndustria PorcParoConstruccion
## [1,]           0.2867110        0.01230661            0.5424105
## [2,]          -0.4377658        0.24202992            0.4673963
## [3,]           3.2227486       -1.49730377           -1.3329449
## [4,]          -0.9525257        2.19467812           -0.5828027
## [5,]          -0.4949614       -0.13537267           -0.2077317
## [6,]           0.3057761       -0.28305194            0.9174816
##      PorcParoServicios PorcParoOtros
## [1,]       -0.77723028     0.8682636
## [2,]       -0.05922047     0.2027520
## [3,]       -1.35163813     0.8682636
## [4,]       -0.61767255    -1.3034058
## [5,]        0.65878935    -0.1825442
## [6,]       -0.29855707     0.2728059
```

```
numeric_des_z <- describe_custom(z_value(datos_num))
numeric_des_z["media",]
```

```
##                  PobTot2018         PorcVarPob2000_2018
##                2.029126e-17               -2.391343e-17
##          PorcMenores16_2018          PorcMayores65_2018
##                4.656231e-17                1.267745e-16
##      PorcPobExtranjera_2018                EdadMedia2018
##                1.793938e-17               -5.413755e-16
##             TBNatalidad2018         TasaBrutaMortalidad
##               -2.213607e-16               -3.278627e-16
## TasaMortalidadMenores5anyos          EsperanzaVidaH2018
##               -1.275022e-16               -4.123513e-15
##          EsperanzaVidaM2018         PorcParoAgricultura
##                2.250137e-15                3.951396e-17
##           PorcParoIndustria        PorcParoConstruccion
##               -3.089976e-18               -1.953732e-16
##           PorcParoServicios                PorcParoOtros
##                9.897390e-16                6.316729e-17
```

Como se puede observar todas las variables poseen la misma escala. Ahora si se podrá realizar una correcta comparación entre las distintas variables. Cabe destacar que se omitira la comparación de medias y desviaciones típicas pues al tipificar todas estas pasan a ser 0 y 1 respectivamente.

El coeficiente de asimetria aporta una medida de como se distribuyen los datos en función de la media. Así valores iguales a 0 indican una distribución simétrica, valores superiores indican que existe una mayor proporción de valores con una valor numérico superior a la media e inferiores indican que existe una mayor proporción de valores.

```
numeric_des_z["skew",]
```

```
##                  PobTot2018         PorcVarPob2000_2018
##                3.424991127                0.518095398
##          PorcMenores16_2018          PorcMayores65_2018
##               -0.006526281                0.467521853
##      PorcPobExtranjera_2018               EdadMedia2018
##                6.782419734               -0.115634920
##             TBNatalidad2018         TasaBrutaMortalidad
##                2.144744434                0.551607713
## TasaMortalidadMenores5anyos          EsperanzaVidaH2018
##                1.062593464               -0.131834345
##          EsperanzaVidaM2018         PorcParoAgricultura
##               -1.132546585                1.317765274
##           PorcParoIndustria        PorcParoConstruccion
##                0.119738672               -0.197312885
##           PorcParoServicios               PorcParoOtros
##                0.531269479                0.808073951
```

Las siguientes variables se corresponden con el segundo caso mencionado anteriormente.

```
numeric_des_z["skew",][which(numeric_des_z["skew",] > 0)]
```

```
##                  PobTot2018         PorcVarPob2000_2018
##                   3.4249911                   0.5180954
##          PorcMayores65_2018      PorcPobExtranjera_2018
##                   0.4675219                   6.7824197
##             TBNatalidad2018         TasaBrutaMortalidad
##                   2.1447444                   0.5516077
## TasaMortalidadMenores5anyos         PorcParoAgricultura
##                   1.0625935                   1.3177653
##           PorcParoIndustria           PorcParoServicios
##                   0.1197387                   0.5312695
##               PorcParoOtros
##                   0.8080740
```

Las siguientes variables se corresponden con el tercer caso mencionado anteriormente.

```
numeric_des_z["skew",][which(numeric_des_z["skew",] < 0)]
```

```
##   PorcMenores16_2018        EdadMedia2018   EsperanzaVidaH2018
##         -0.006526281         -0.115634920         -0.131834345
##   EsperanzaVidaM2018 PorcParoConstruccion
##         -1.132546585         -0.197312885
```

Ninguna de las variables posee una distribución totalmente simétrica.

```
numeric_des_z["skew",][which(numeric_des_z["skew",] == 0)]

## named numeric(0)
```

Es necesario mencionar que existen mas variables con valores desplazados hacia la derecha de la media, además, este desplazamiento hacia la derecha es especialmente llamativo, se encuentra por encima de la unidad, en las siguientes variables. En el caso de las variables desplazadas hacia la izquierda de la media, solo existe una, con un coeficiente de asimetría superior a 1 en valor absoluto.

```
numeric_des_z["skew",][which(numeric_des_z["skew",] > 1)]

##                  PobTot2018      PorcPobExtranjera_2018
##                    3.424991                    6.782420
##             TBNatalidad2018 TasaMortalidadMenores5anyos
##                    2.144744                    1.062593
##         PorcParoAgricultura
##                    1.317765

numeric_des_z["skew",][which(numeric_des_z["skew",] < -1)]

## EsperanzaVidaM2018
##          -1.132547
```

A continuación se observara el coeficiente de kurtosis, en general este coeficiente da una medida de la proporción de valores cercanos a la media y en las colas. Un valor igual a 0 indica una proporción similar a una distribución normal, valores superiores indican que existen tanto valores muy cercanos a la media como en los extremos, y valores menores a cero indican que existe mayor proporción de puntos medios.

```
numeric_des_z["kurto",]

##                  PobTot2018         PorcVarPob2000_2018
##                  12.6430811                  -0.2080464
##          PorcMenores16_2018          PorcMayores65_2018
##                   0.4191006                  -0.1559526
##      PorcPobExtranjera_2018               EdadMedia2018
##                  44.9509508                   0.3899858
##             TBNatalidad2018         TasaBrutaMortalidad
##                   8.0893086                  -0.2442150
## TasaMortalidadMenores5anyos          EsperanzaVidaH2018
##                   2.1441768                  -0.6095486
##          EsperanzaVidaM2018         PorcParoAgricultura
##                   1.4968881                   1.8516816
##           PorcParoIndustria        PorcParoConstruccion
##                  -0.6065349                   0.3998551
##           PorcParoServicios               PorcParoOtros
##                  -0.3329146                   0.6088719
```

Así las siguientes variables poseen kurtosis positiva

```
numeric_des_z["kurto",][which(numeric_des_z["kurto",] > 0)]
```

```
##                  PobTot2018          PorcMenores16_2018
##                  12.6430811                   0.4191006
##      PorcPobExtranjera_2018               EdadMedia2018
##                  44.9509508                   0.3899858
##             TBNatalidad2018 TasaMortalidadMenores5anyos
##                   8.0893086                   2.1441768
##          EsperanzaVidaM2018         PorcParoAgricultura
##                   1.4968881                   1.8516816
##        PorcParoConstruccion               PorcParoOtros
##                   0.3998551                   0.6088719
```

Mientras que las siguientes variables poseen kurtosis negativa

```
numeric_des_z["kurto",][which(numeric_des_z["kurto",] < 0)]
```

```
## PorcVarPob2000_2018  PorcMayores65_2018 TasaBrutaMortalidad  EsperanzaVidaH2018
##          -0.2080464          -0.1559526          -0.2442150          -0.6095486
##   PorcParoIndustria   PorcParoServicios
##          -0.6065349          -0.3329146
```

No existe ninguna variable con kurtosis 0

```
numeric_des_z["kurto",][which(numeric_des_z["kurto",] == 0)]
```

```
## named numeric(0)
```

Se puede observar existen mas comunidades con una kurtosis positiva que negativa. Y en general el valor absoluto de las kurtosis positivas es superior al de las kurtosis negativas.
A continuación se analizará el rango entre máximos y mínimos, esto da una medida de el grado de compactación de los valores.

```
numeric_des_z["rango",]
```

```
##                  PobTot2018         PorcVarPob2000_2018
##                    5.528842                    4.669567
##          PorcMenores16_2018          PorcMayores65_2018
##                    5.138845                    4.683000
##      PorcPobExtranjera_2018               EdadMedia2018
##                    7.258186                    5.029192
##             TBNatalidad2018         TasaBrutaMortalidad
##                    6.436348                    4.191968
## TasaMortalidadMenores5anyos          EsperanzaVidaH2018
##                    5.474353                    3.961167
##          EsperanzaVidaM2018         PorcParoAgricultura
##                    4.733446                    4.480317
##           PorcParoIndustria        PorcParoConstruccion
##                    4.184246                    5.025952
##           PorcParoServicios               PorcParoOtros
##                    4.339970                    4.903769
```

Se puede observar la variable más compacta es la EsperanzaVidaH2018 mientras que la más dispersa es la PorcPobExtranjera_2018.

Por último se comprobara el resultad del test Shapiro-Wilks realizado. Valores superiores a 0.05, nivel de significación escogido, indican que no se rechaza la hipótesis nula de normalidad, mientras que valores inferiores indican que se rechaza esta.

```
numeric_des_z["shapiro_pvalor",]
```

```
##                   PobTot2018         PorcVarPob2000_2018
##                 4.625438e-11                2.171123e-01
##           PorcMenores16_2018          PorcMayores65_2018
##                 2.363001e-01                3.083571e-01
##      PorcPobExtranjera_2018              EdadMedia2018
##                 9.117069e-16                2.796724e-01
##              TBNatalidad2018         TasaBrutaMortalidad
##                 2.147524e-06                1.280758e-01
## TasaMortalidadMenores5anyos          EsperanzaVidaH2018
##                 8.399454e-03                5.940608e-01
##           EsperanzaVidaM2018         PorcParoAgricultura
##                 1.228524e-03                7.258436e-05
##           PorcParoIndustria        PorcParoConstruccion
##                 7.224351e-01                5.256164e-01
##           PorcParoServicios               PorcParoOtros
##                 4.406381e-02                3.020272e-02
```

En las siguientes variables no se rechaza la hipótesis nula de normalidad.

```
numeric_des_z["shapiro_pvalor",][which(numeric_des_z["shapiro_pvalor",] > 0.05)]
```

```
##  PorcVarPob2000_2018   PorcMenores16_2018   PorcMayores65_2018
##            0.2171123            0.2363001            0.3083571
##        EdadMedia2018  TasaBrutaMortalidad   EsperanzaVidaH2018
##            0.2796724            0.1280758            0.5940608
##    PorcParoIndustria PorcParoConstruccion
##            0.7224351            0.5256164
```

Mientras que en las siguientes se rechaza la hipótesis nula

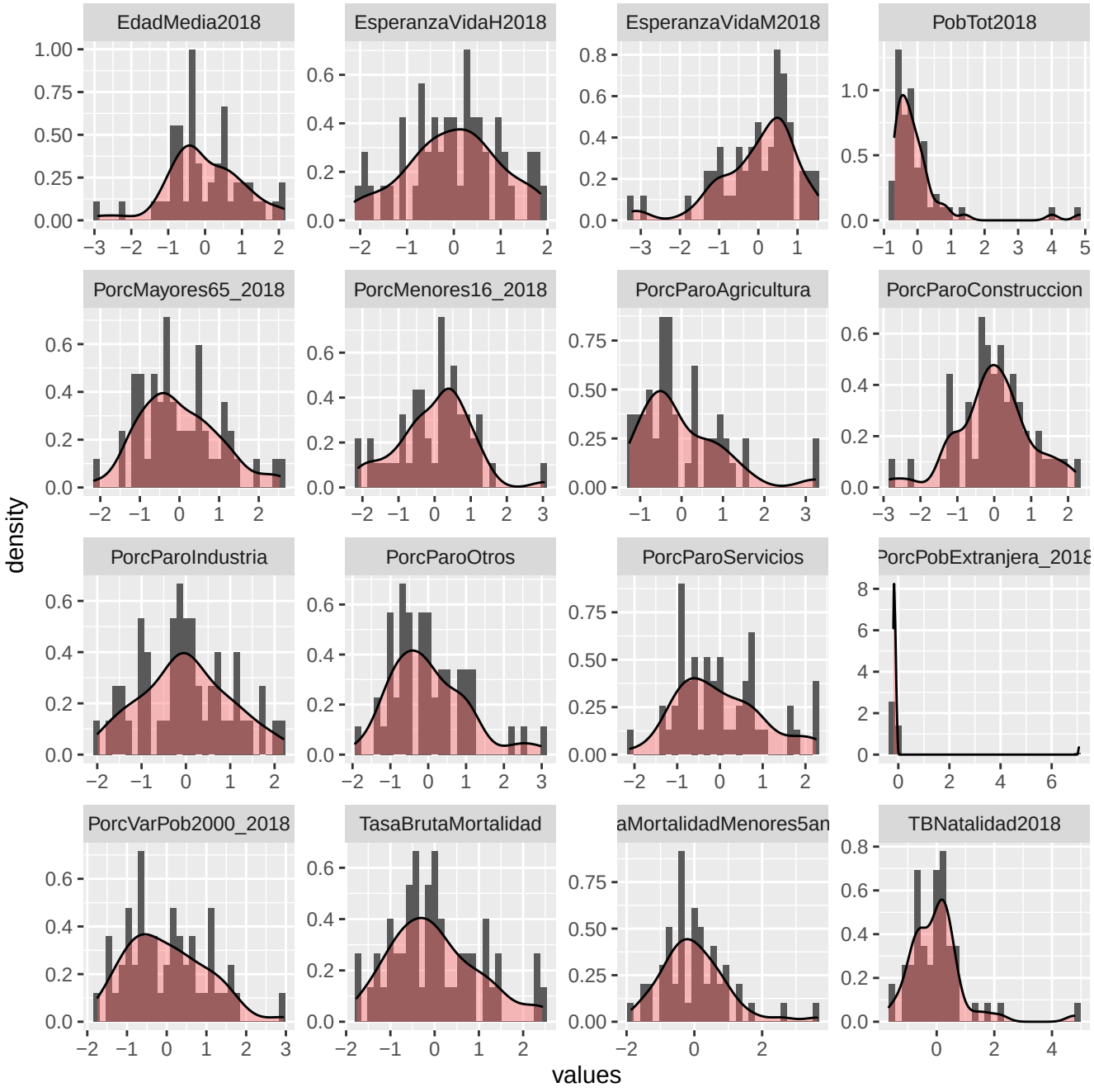
```
numeric_des_z["shapiro_pvalor",][which(numeric_des_z["shapiro_pvalor",] < 0.05)]
```

```
##                   PobTot2018      PorcPobExtranjera_2018
##                 4.625438e-11                9.117069e-16
##              TBNatalidad2018 TasaMortalidadMenores5anyos
##                 2.147524e-06                8.399454e-03
##           EsperanzaVidaM2018         PorcParoAgricultura
##                 1.228524e-03                7.258436e-05
##           PorcParoServicios               PorcParoOtros
##                 4.406381e-02                3.020272e-02
```

Por último se realizará una inspección visual de la distribución de cada variable. Para esto se hace uso de la función multiplehist().

```r
multiplehist <- function(data, densiti=FALSE){
  require(ggplot2)
  data <- as.data.frame(data)
  var_name <- names(data)

  gathered <- data.frame(list(variables=rep(var_name, rep(nrow(data),ncol(data))) ,values=do.call(



  if (densiti){
    h <- ggplot(data=gathered) + aes(x = values) + geom_histogram(aes(y=..density..)) + geom_densi
    print(h)
  }else{
    h2 <- ggplot(data=gathered) + aes(x = values) + geom_histogram()  + facet_wrap( ~ variables, s
    print(h2)
  }


}
multiplehist(z_value(datos_num), densiti = TRUE)
```

```
## Loading required package:  ggplot2
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```

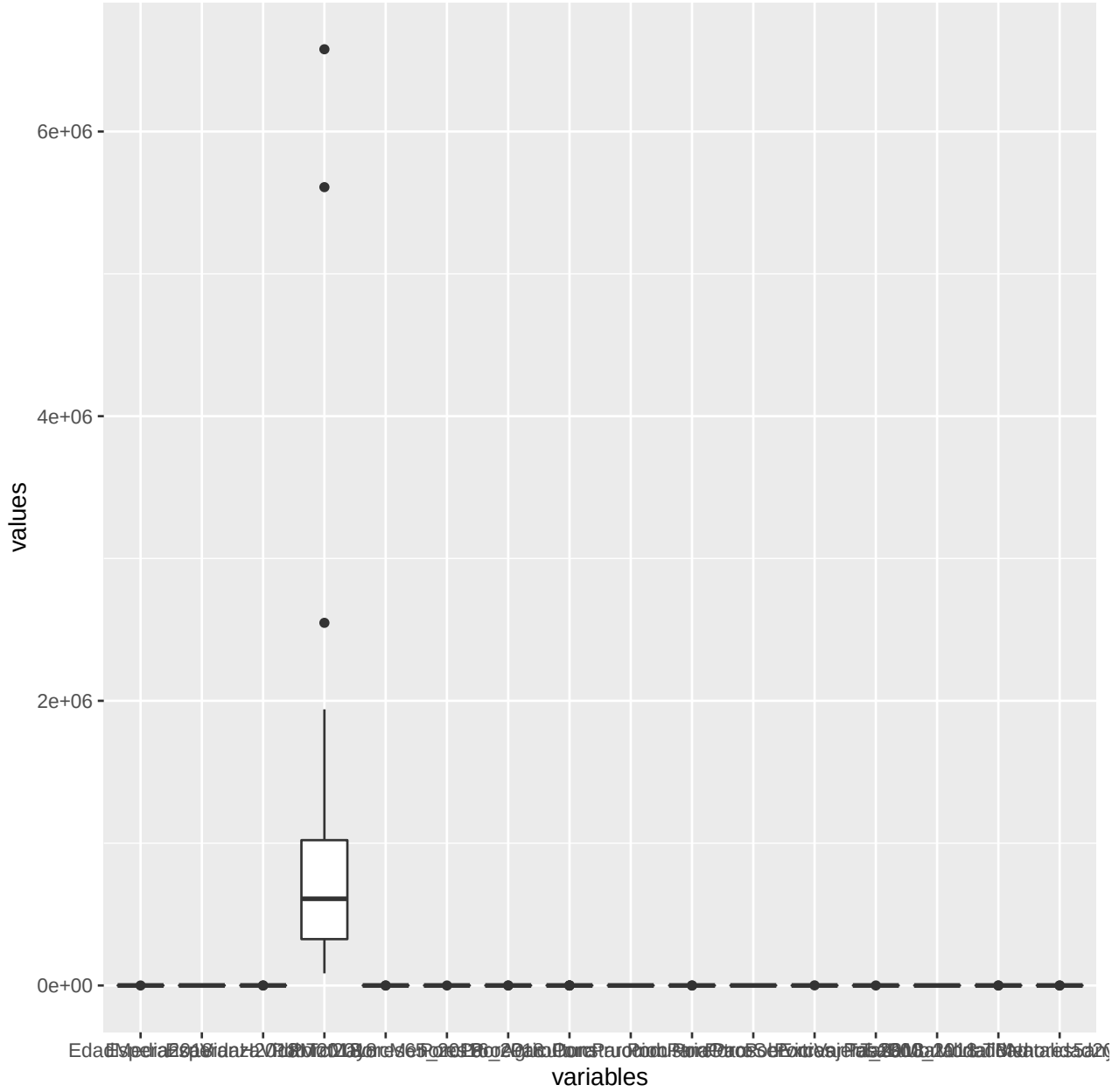


Se puede observar como la variable PorcPobExtranjera_2018 posee una gran cola hacia la derecha, algo que concuerda con el coeficiente de asimetria encontrado, esta conclusión se puede obtener para las distintas variables. Si atendemos también a la variable PorcPobExtranjera_2018 se puede entender el coeficiente de kurtosis encontrado, como se puede observar existen valores muy cercanos a la media, así como varios valores extremos.

Tarea 2: Realiza un análisis de datos anómalos (outliers)

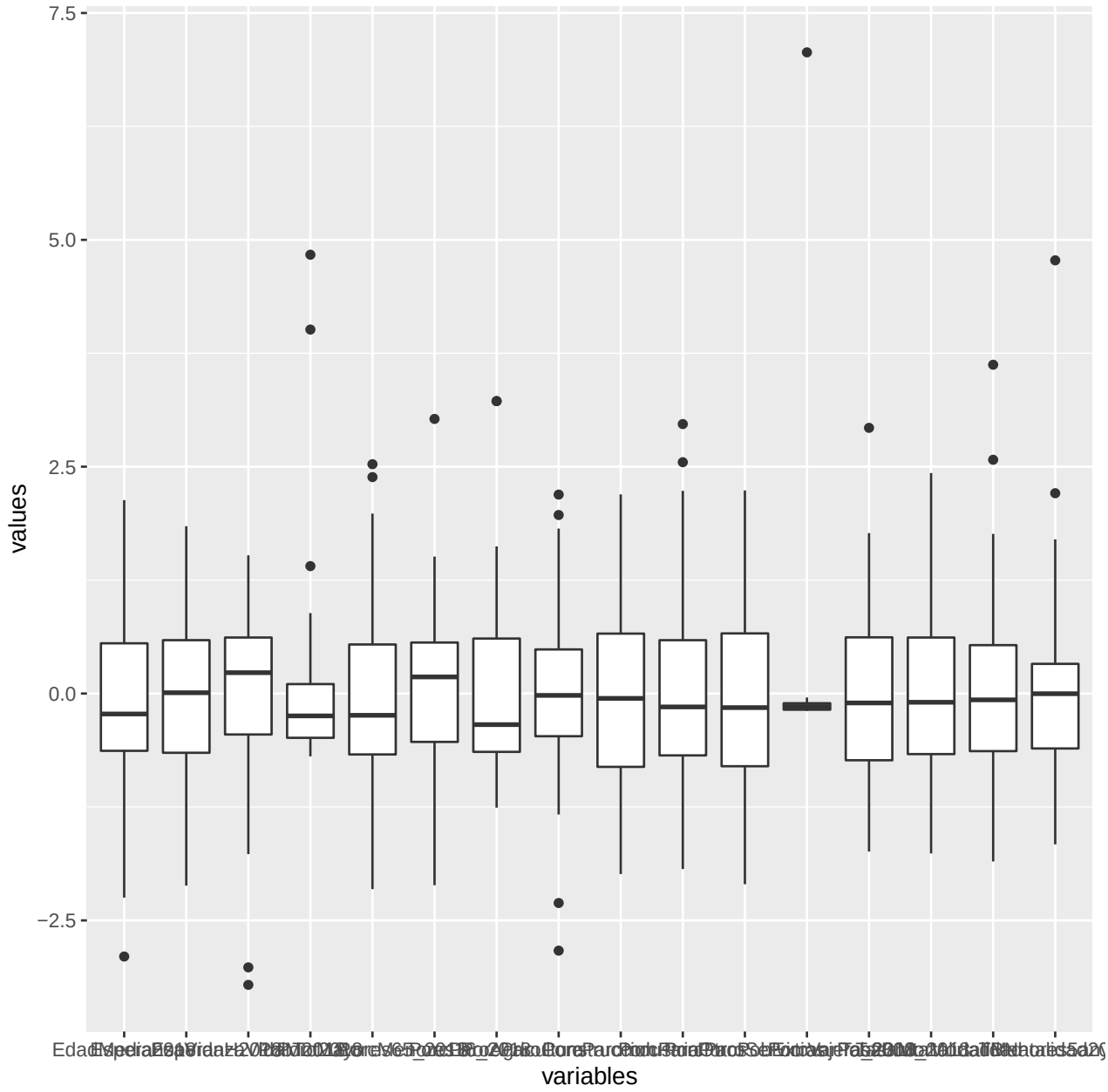
1. Análisis univariante de datos anómalos.
   Primero se realizará una inspección ocular de los distintos boxplot de cada variable mediante la función multipleboxplot().

```r
multipleboxplot <- function(data){
  require(ggplot2)
  data <- as.data.frame(data)
  var_name <- names(data)

  gathered <- data.frame(list(variables=rep(var_name, rep(nrow(data),ncol(data))) ,values=do.
  h <- ggplot(data=gathered) + aes(x = variables, y =values) + geom_boxplot()
  print(h)



}
multipleboxplot(datos_num)
```



El hecho de que exista una gran diferencia en las escalas dificulta la interpretación. Así se usará el banco de datos tipificado.

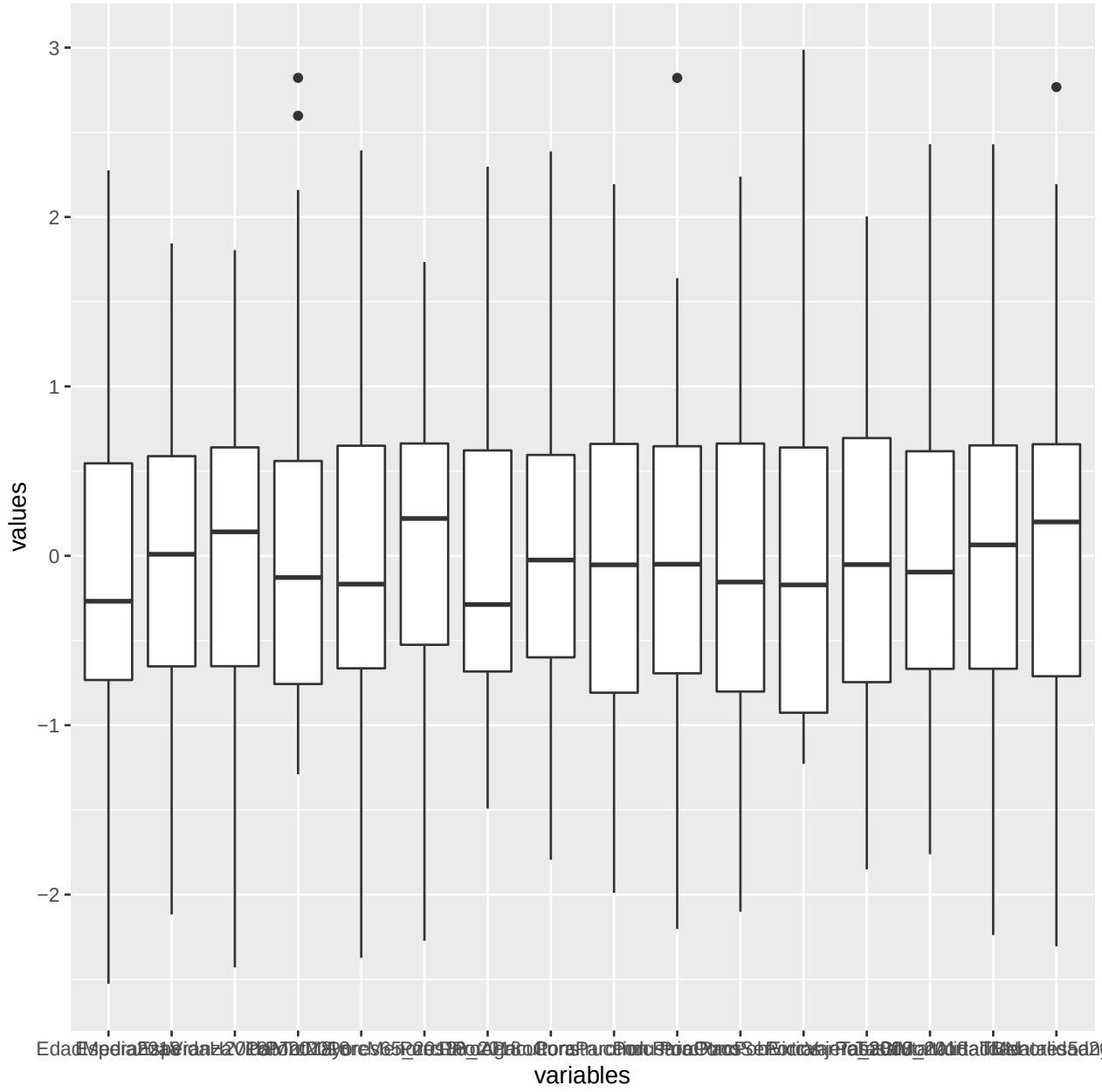
```
multipleboxplot(z_value(datos_num))
```



Todas las variables poseen valores extremos, usando como regla para determinación de esta, la misma que en un boxplot. Esto es valores superiores al cuartil superior o inferior mas o menos 3/2 de la distancia entre el cuartil superior o inferior. Estos valores se descartaran y serán sustituidos, imputación, por los valores medios de la variable antes de eliminar estos valores extremos. Para ello se usará la siguiente función.

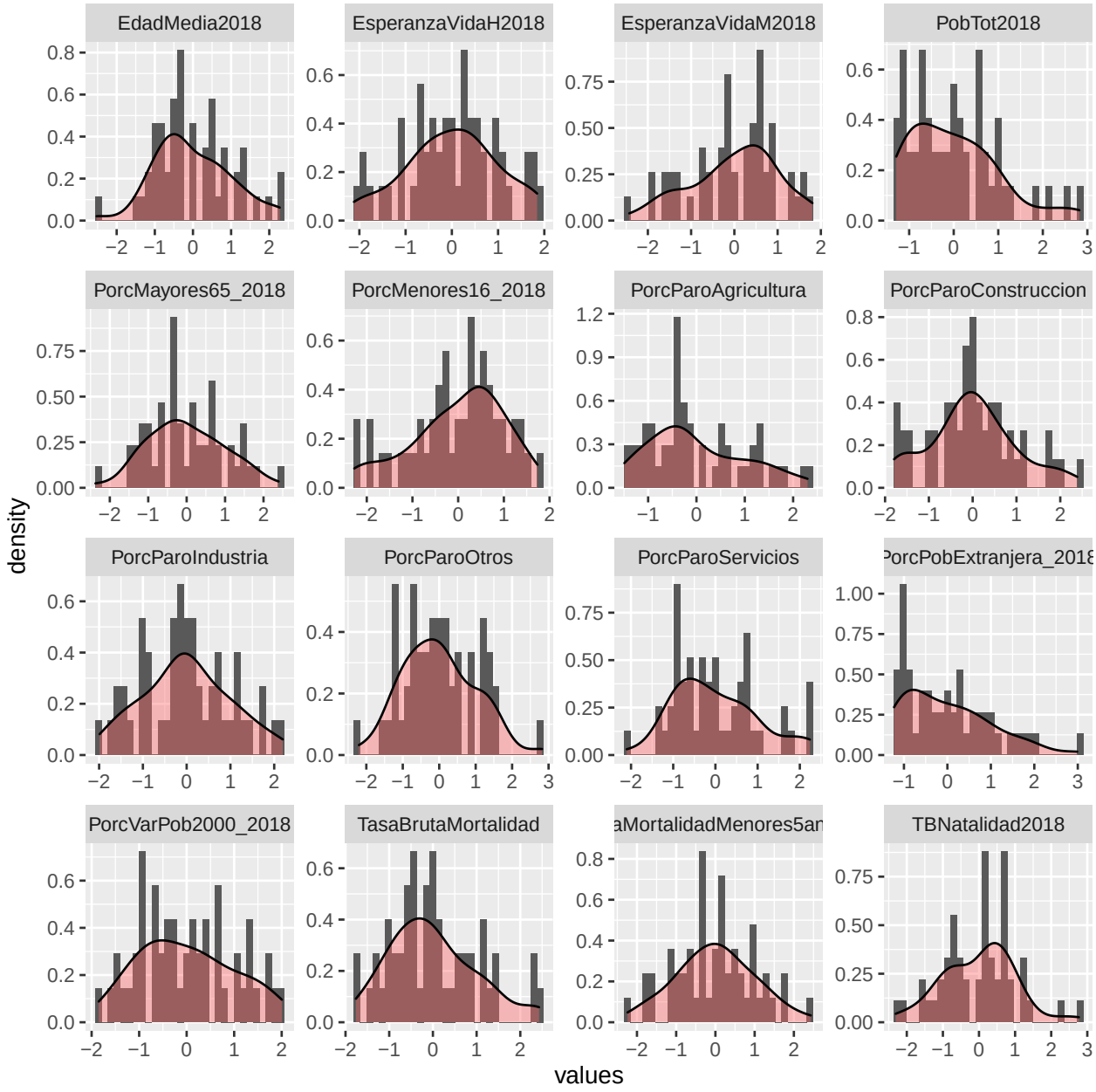
```
extremos_mean <- function(data, constant){
 extremos_media <- apply(data, 2, function(x){
   d = sum(quantile(x, c(0.25, 0.75))*c(-1, 1))
   c(
     superior = quantile(x, c(0.75),names=FALSE)+d*constant,
     inferior = quantile(x, c(0.25),names=FALSE)-d*constant,
     media = mean(x))
 })
```

```
 result <- apply(data, 1, function(x, y){
   idx<-which(x < y["inferior",] | x > y["superior",])
   x[idx] <- y["media",idx]
   x
 }, y = extremos_media)
 return(as.data.frame(t(result)))
}
datos_num_smooth <- extremos_mean(datos_num, constant = 3/2)
multipleboxplot(z_value(datos_num_smooth))
```



Se puede observar que se ha reducido el número de valores extremos, cabe destacar que la representación sigue mostrando valores extremos, esto es debido a que se han cambiado los valores extremos por la media, lo que centra la distribución de los valores, haciendo que otros valores antes no extremos puedan considerarse ahora extremos. También es util observar los histogramas de la variable sin valores extremos.

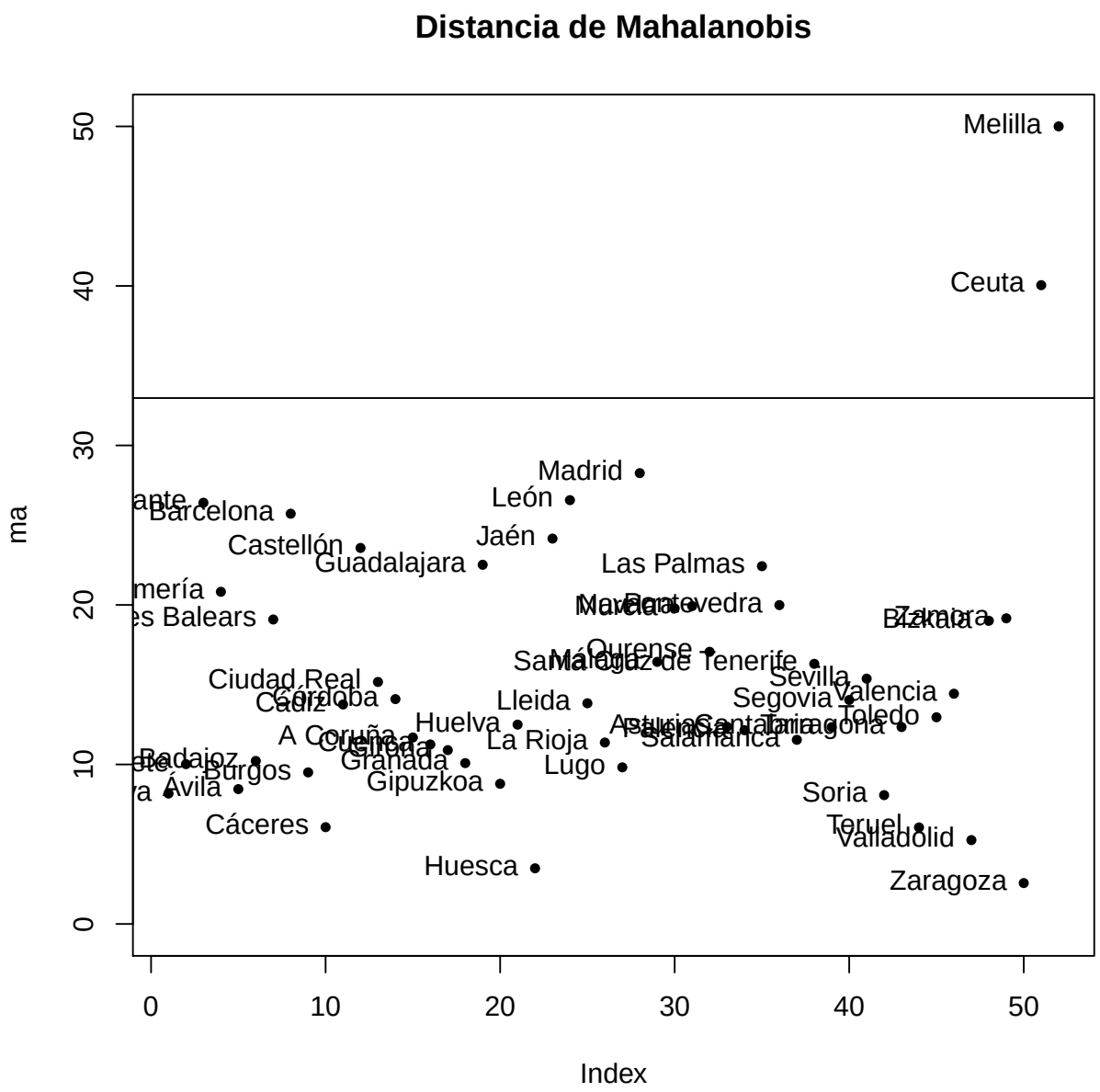
```
multiplehist(z_value(datos_num_smooth), densiti = TRUE)

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



2. Datos anómalos multivariantes
   Para la descripción de datos anómalos multivariantes se hará uso de la distancia de Mahalanobis y se utilizara como valor de discriminación k + 3 sqrt(2k) siendo k el número de variables.

```
x <- datos_num
rownames(x) <- codigo_nombre$NombreProv
ma <- mahalanobis(x, apply(x, 2, mean), cov(x))
k <- dim(x)[2]
Lim <- k + 3 * sqrt(k * 2)
plot(ma, pch = 20, ylim = c(0, max(ma, Lim, na.rm = TRUE)))
text(ma, rownames(x), pos = 2)
```

```
abline(h = Lim)
title("Distancia de Mahalanobis")
```



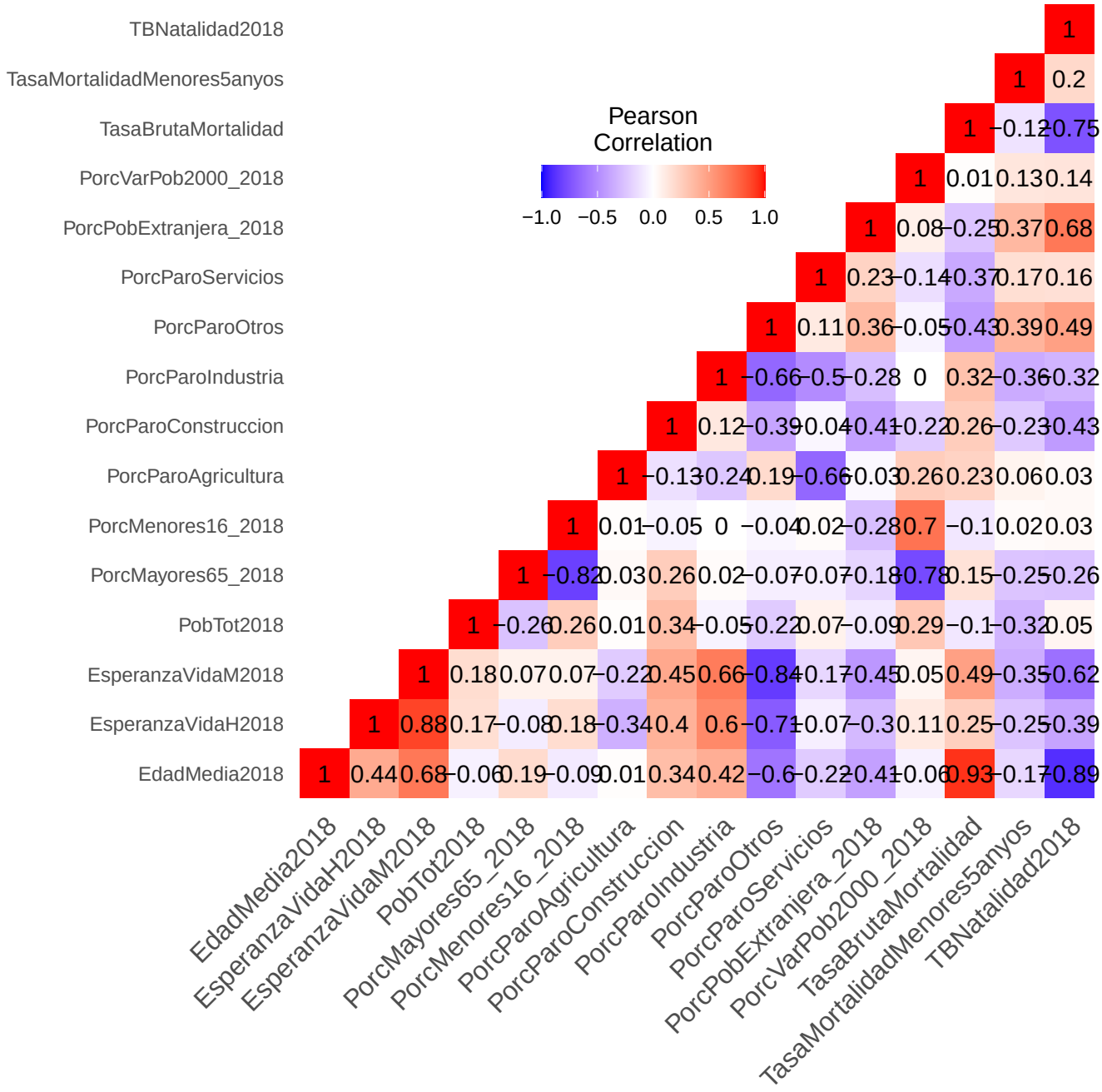
Se puede observar como las ciudades autónomas de Ceuta y Melilla son consideradas como datos anómalos. Esto es lógico si se tienen en cuenta la idiosincrasia de estas regiones del territorio español.

Tarea 3: Relación entre las variables

El análisis de las relaciones entre las variables se hará mediante la representación de la matriz de correlación, para esto se hace uso de la siguiente función:

```r
correlation_ggplot <- function(data){ # Se crea una función que genera<< gráficos
  #de correlación para la variables numéricas en el argumento data, usando
  #el motor gráfico ggplot2
  tipos <- sapply(data, function(x){
    is.numeric(x)
  })#Se obtiene la posición de las columnas de tipo numerico en el data.frame

  numeric_names <- sort(names(data)[tipos])#Se seleccionan los nombres de las
  #variables de tipo factor. Notese que se han ordenado los nombres, esto es
  #necesario debido a que ggplot2 ordenará posteriormente las variables a
  #representar.
  data_new <- data[,numeric_names] #Se crea un data.frame adicional que
  #facilitará la representación con ggplot2
  correlation_mat <- round(cor(data_new),2)#Se crea la matriz de correlación
  #en este caso con la función cor

  correlation_mat[upper.tri(correlation_mat)] <- NA #Se elimina la información
  # de la diagonal superior, pues esta repetida
  correlation_mat <- t(correlation_mat) # Por conveniencia se transpone la
  #matriz



  melt_corr <- data.frame(list(Var1=rep(numeric_names, length(numeric_names)),
                               Var2=rep(numeric_names,
                                        rep(length(numeric_names),
                                            length(numeric_names))),
                               value=rep(NA, length(numeric_names)^2)))
  #Se crea un data.frame auxiliar con toda la informacióna representar. Este
  #se encuentra vacio y se rellenará con el bucle for que se encuentra más abajo.
  for (x in 1:nrow(melt_corr)){
    if(!is.na(correlation_mat[melt_corr[x, 2], melt_corr[x,1]])){
      melt_corr[x, 3] <- correlation_mat[melt_corr[x, 2], melt_corr[x,1]]
    }
  }#Se rellena el campo value del dataframe creado con la información presente
  #en correlation_mat gracias a la información aportada por las variables
  #Var1 y Var2

  melt_corr$Var1 <- as.factor(melt_corr$Var1)
  melt_corr$Var2 <- as.factor(melt_corr$Var2)
  melt_corr$value <- as.numeric(melt_corr$value)#Se convierten las variables
  #al tipo neceario
  melted_cormat <- melt_corr[!is.na(melt_corr$value),]#Se eliminan los NA

  h <- ggplot(data = melted_cormat, aes(Var1, Var2, fill = value))+ #Se definen
    #el data.set de donde saldrán los datos, así como que hacer con cada
    #variable
    geom_tile() + #Se genera un geom de tipo raster, una imagen.
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1,1), space = "Lab",
                         name="Pearson\nCorrelation") + #Se crea una escala
```

```
    #de color para usar en la imagen creada con geom_raster.
    theme_minimal()+
    theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                     size = 12, hjust = 1))+#Se define como se
    #representará la etiqueta del eje x.
    coord_fixed()#Se fijan las coordenadas del plot creado.
  j <-  h +
    geom_text(aes(Var1, Var2, label = value), color = "black", size = 4) +
    theme(
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      panel.grid.major = element_blank(),
      panel.border = element_blank(),
      panel.background = element_blank(),
      axis.ticks = element_blank(),
      legend.justification = c(1, 0),
      legend.position = c(0.6, 0.7),
      legend.direction = "horizontal")+
    guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                                 title.position = "top", title.hjust = 0.5))
  #Se añaden los valores del coeficiente de correlación en la posición adecuada.
  print(j)#Se imprime el gráfico creado.

}
correlation_ggplot(datos_num)
```

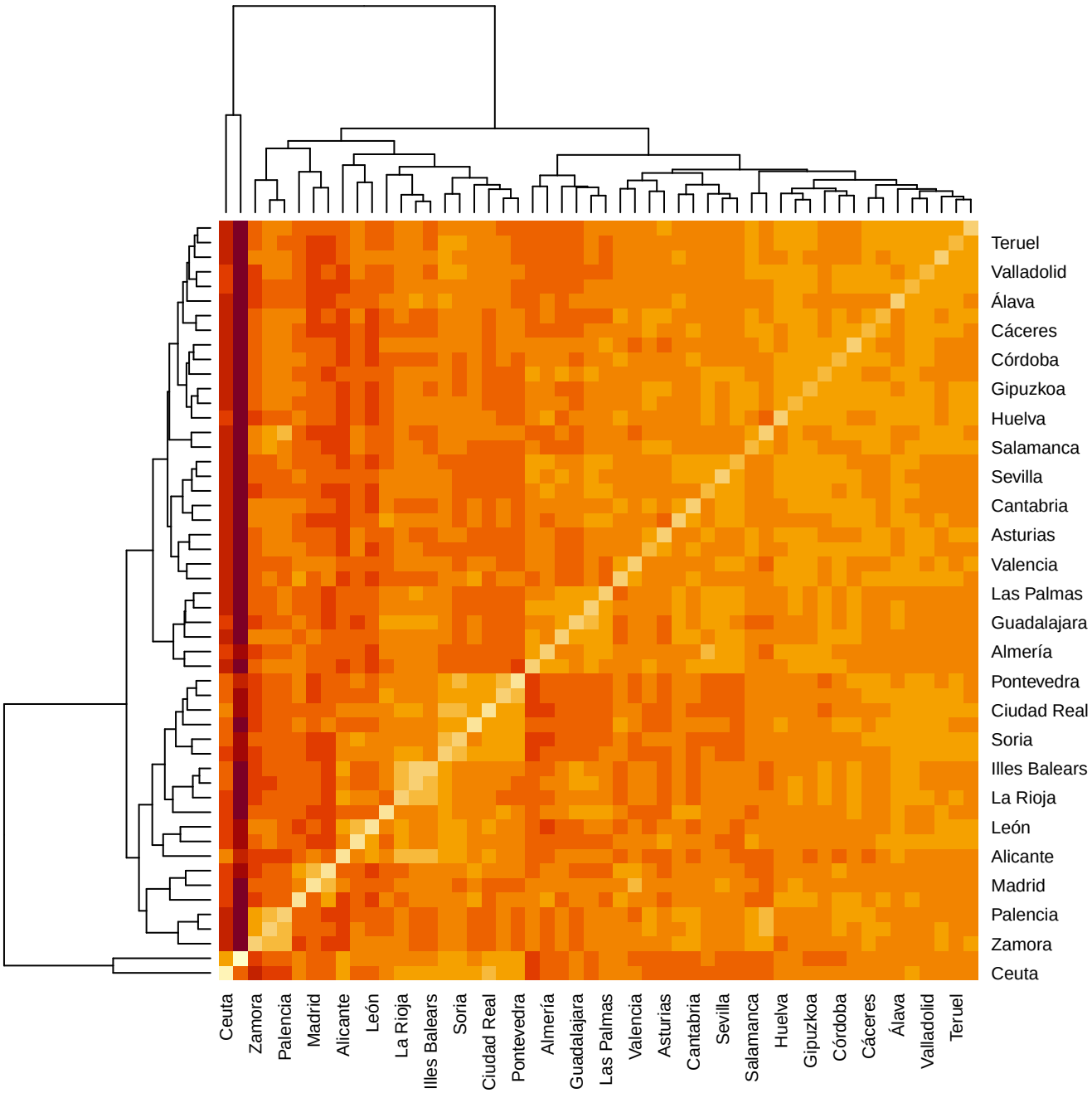


Es necesario destacar que el coeficiente de correlación de Pearson solo es capaz de medir las relaciones lineales entre las variables.
existe una fuerte relación lineal positiva entre las variables EdadMedia2018 y TasaBrutaMortalidad, esto es lógico pues una población mas envejecida tendrá mas decesos que una menos envejecida, asumiendo que la mayor fuente de decesos son las personas de mayor edad. Además la mayor correlación negativa la podemos encontrar entre las variables EdadMedia2018 y TBNatalidad2018 algo lógico pues a mayor edad media en la población menos individuos capaces de reproducirse y generar nuevos individuos. Tambien es llamativa la correlación negativa entre el PorcMayores65_2018 y el PorcMenores16_2018 que ejemplifica de forma clara el envejecimiento demográfico de la población española.

Tarea 4: Relación entre los individuos.

Para el analisis de las relaciones entre individuos se realizará un analisis de las distancias entre individuos.

```
x <- z_value(datos_num)
rownames(x) <- codigo_nombre$NombreProv
distancias <- dist(x)
heatmap(as.matrix(distancias))
```



Se puede observar que las ciudades autonomas de Ceuta y Melilla son las que poseen la mayor distancias con el resto de las provincias. Las diferencias entre las demás provincias no son tan palpables en esta representación.

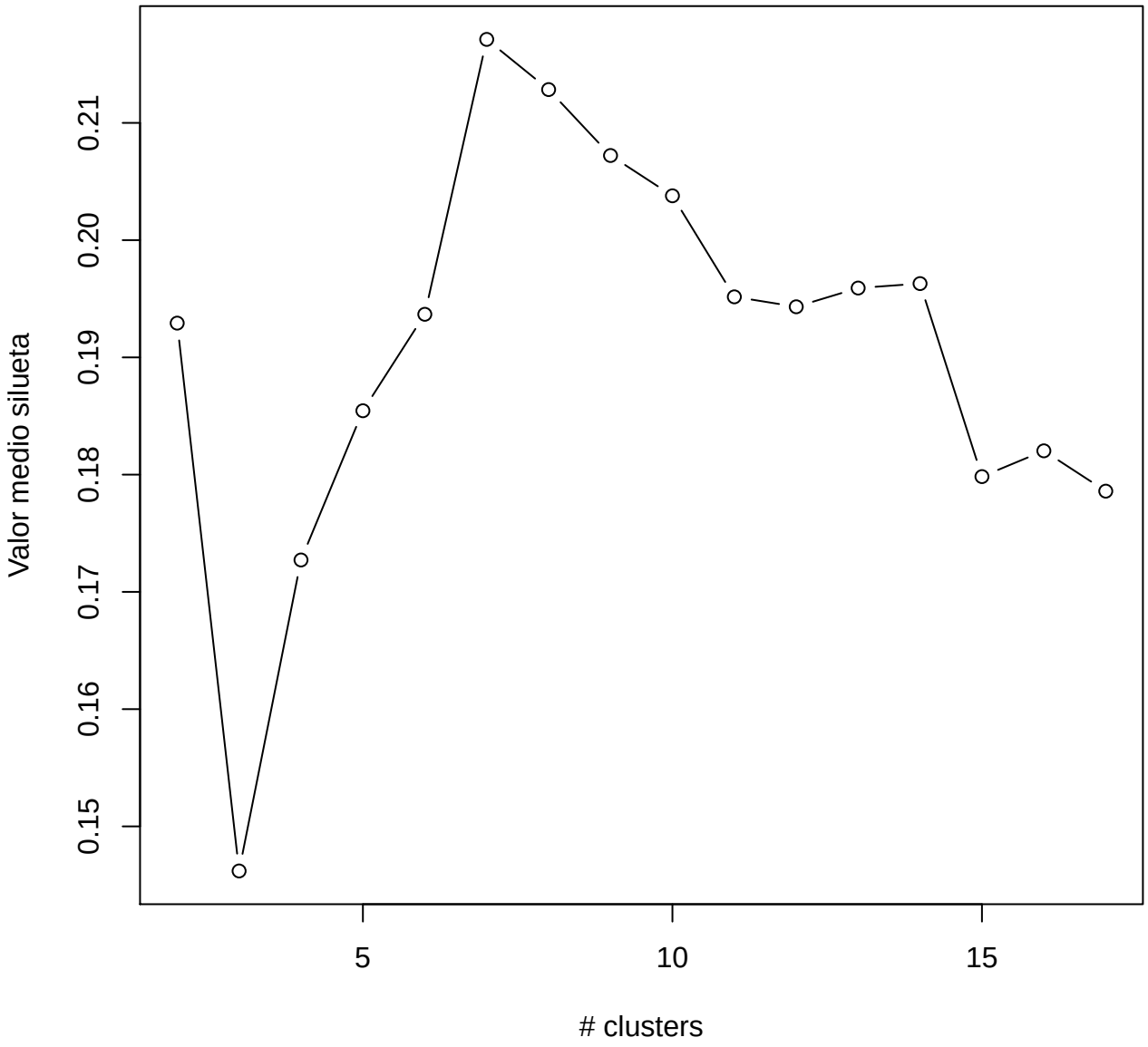
Además se desea realizar un análisis de grupos mediante el método de K-means, a grande rasgos este método intenta situar en el mismo grupo a aquellos individuos mas similares. Para verificar que se ha utilizado un valor de k correcto, el número de clusters, se utilizará el método de la silueta. Este método

mide la similaridad de los individuos dentro de un grupo, a mayor este valor, mayor similaridad dentro del cluster. Puesto que el metodo de agrupación K-means es sensible a volores extremos, se eliminarán las ciudades autonomas de Ceuta y Melilla, los outlaiers obtenidos en el análisis multivariante, así como con el banco de datos con la inputación por medias

```
set.seed(1)
x <- z_value(datos_num[c(-51,-52),])
#x <- z_value(datos_num)
rownames(x) <- codigo_nombre$NombreProv[c(-51,-52)]
valor_silueta <- function(k){
  km.result <- kmeans(x, centers = k, nstart = 1000)
  ss <- cluster::silhouette(km.result$cluster, dist(x))
  mean(ss[, 3])
}
k.values <- 2:17
avg_sil_values <- sapply(k.values, valor_silueta)
avg_sil_values
```

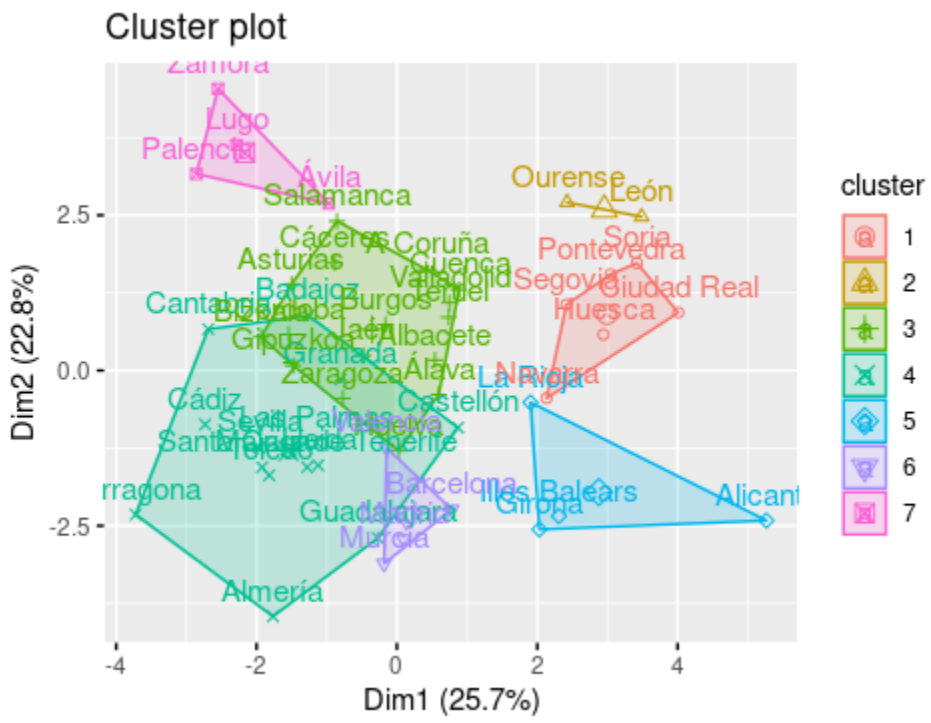
```
##  [1] 0.1929245 0.1462008 0.1727241 0.1854491 0.1936763 0.2171226 0.2128418
##  [8] 0.2072199 0.2037815 0.1951652 0.1943180 0.1959155 0.1962949 0.1798336
## [15] 0.1820313 0.1785938
```

```
plot(k.values, avg_sil_values, xlab="# clusters", ylab = "Valor medio silueta",type="b")
```




```
### Como se puede observar el valor que ofrece un mayor valor de silueta es con un k = 7.
cluster <- kmeans(x, centers = k.values[which.max(avg_sil_values)], nstart = 25)
#library(factoextra)
#fviz_cluster(cluster, data=x)
```

Es necesario destacar que la función fviz_cluster realiza un análisis de componentes principales para representar valores multidimensionales en dos dimensiones. En este caso, las dos componentes principales recogen el 25.7 % y 22.8 % de la variabilidad total de los datos.




```
l_cluster <- list()
for (x in 1:k.values[which.max(avg_sil_values)]){
  l_cluster[[x]] <- names(cluster$cluster)[which(cluster$cluster == x)]
}
l_cluster
```

```
## [[1]]
## [1] "Ciudad Real" "Huesca"      "Navarra"     "Pontevedra"  "Segovia"
## [6] "Soria"
##
## [[2]]
## [1] "León"    "Ourense"
##
## [[3]]
##  [1] "Álava"      "Albacete"   "Burgos"     "Cáceres"    "Córdoba"
##  [6] "A Coruña"   "Cuenca"     "Gipuzkoa"   "Huelva"     "Jaén"
## [11] "Asturias"   "Salamanca"  "Teruel"     "Valladolid" "Bizkaia"
## [16] "Zaragoza"
##
## [[4]]
##  [1] "Almería"                "Badajoz"                "Cádiz"
##  [4] "Castellón"              "Granada"                "Guadalajara"
##  [7] "Lleida"                 "Málaga"                 "Las Palmas"
## [10] "Santa Cruz de Tenerife" "Cantabria"              "Sevilla"
## [13] "Tarragona"              "Toledo"
##
## [[5]]
## [1] "Alicante"      "Illes Balears" "Girona"        "La Rioja"
```

```
##
## [[6]]
## [1] "Barcelona" "Madrid"    "Murcia"    "Valencia"
##
## [[7]]
## [1] "Ávila"    "Lugo"     "Palencia" "Zamora"
```

Como se puede observar, en virtud de los datos presentados, se pueden diferenciar 7 grupos de pronvincias. Cabe destacar que dado que existen variables dispares como PobTotal2018 e PorcParoIndustria la interpretación de estos grupos es de excasa relevancia pues mezclan distintos atributos. Solo se puede concluir que las comunidades presentes en los mismos cluster son similares en función de los datos aportados.