Practica PCA

Ana Buendia Ruiz-Azuaga

Contents

–

Información sobre el Dataset

El fichero de datos DB_3.sav contiene las variables ZTLIBROP, ZTEJERCI, ZTPOBACT, ZTENERGI, ZPSERVI, ZPAGRICU, ZTMEDICO, ZESPVIDA, ZTMINFAN y ZPOBDENS que respectivamente son los valores para cada país del mundo de:

Variables

- Número de libros publicados (ZTLIBROP).
- Cociente entre número de individuos en ejército de tierra y población total del estado (ZTEJERCI).
- Cociente entre población activa y total (ZTPOBACT).
- Tasa de consumo energético (ZTENERGI).
- Población del sector servicios (ZPSERVI).
- Población del sector agrícola (ZPAGRICU).
- Tasa de médicos por habitante (ZTMEDICO).
- Esperanza de vida (ZESPVIDA).
- Tasa de mortalidad infantil (ZTMINFAN).
- Densidad de población (ZPOBDENS).
- Porcentaje de la población urbana (POBURB).

Otra información de interés

El dataset contiene un total de 34 instancias con 12 variables, contando con la etiqueta del país al que corresponde las 11 variables mencionadas anteriormente.

Primer acercamiento

Cargamos los datos a partir del fichero *DB_3.sav*.

Primera visualización

Echamos un primer vistazo a los datos, observando directamente el dataframe en el que vienen, para ver si los datos efectivamente están estandarizados (datos numéricos, continuos, tanto negativos como positivos y entorno al 0), y para comprobar la existencia de *NA*s.

```
head(datos)
```

```
##     ZPOBDENS   ZTMINFAN   ZESPVIDA    ZPOBURB   ZTMEDICO    ZPAGRICU    ZPSERVI
## 1 -0.8571914  0.9614839 -1.5498548 -0.2148886 -0.7338252 -0.60635548 0.48366865
## 2 -1.0167171  1.2133839 -0.8948879 -0.7652623 -0.8876598  0.05765957 0.05831739
## 3 -0.9918559 -0.3185794  0.4388628  1.1381136  1.2660251 -0.75993719 0.78468647
## 4 -1.0778341 -1.0203009  1.0819213  1.2802934  0.4968519 -1.05806639 1.40635370
## 5 -0.9493848  0.6119084 -0.3232805  0.4409735 -0.3107799  0.04410824 0.11066832
## 6 -1.0726547 -1.0280121  1.1295552  0.8170622  0.5160813 -1.10775460 1.62884513
##     ZTLIBROP   ZTEJERCI   ZTPOBACT   ZTENERGI
## 1 -0.5751040 -0.5604298 -0.7828970  0.1281385
## 2 -0.9696156 -0.2033670 -2.1340940 -0.5261359
## 3 -0.4763677 -0.4830344 -0.4189086 -0.3792220
## 4  1.0594990 -0.6117037  0.6244667  1.1899325
## 5 -0.5320795 -0.7149139 -0.4272645 -0.7252616
```

```
## 6  1.5770836 -0.8658606  0.9660097  2.7497848
```

Vemos tanto la existencia de valores perdidos, y que las variables efectivamente están estandarizadas, ya que variables que intuitivamente tomarían valores enteros y relativamente altos como la esperanza de vida o los libros publicados contienen decimales, valores positivos y negativos y todos se encuentran cercanos a 0.

Tratamiento de los *NA*

Vamos a ver si hay, donde, y cuantos *NA* hay en los datos

```
cbind(apply(is.na(datos),2,sum),apply(is.na(datos),2,sum)/dim(datos)[1])
```

```
##          [,1]       [,2]
## ZPOBDENS    0 0.00000000
## ZTMINFAN    0 0.00000000
## ZESPVIDA    0 0.00000000
## ZPOBURB     0 0.00000000
## ZTMEDICO    0 0.00000000
## ZPAGRICU    0 0.00000000
## ZPSERVI     0 0.00000000
## ZTLIBROP    1 0.02941176
## ZTEJERCI    0 0.00000000
## ZTPOBACT    0 0.00000000
## ZTENERGI    0 0.00000000
```

Vemos que apenas hay variables con datos faltantes, pero una de las variables (*ZTLIBROP*) tiene uno de sus registros faltantes. Como, al ser un único valor perdido en un conjunto de datos con 34 instancias es claro que tenemos menos del 5% de valores perdidos, podemos imputarlo con el valor de la media (ya que es una variable cuantitativa) sin que afecte significamente al resultado del análisis.

```
not_available<-function(data,na.rm=F){
  data[is.na(data)]<-mean(data,na.rm=T)
  data
}

datos_pca<-as.data.frame(apply(datos, 2, not_available))
```

Comprobamos ahora que se ha imputado correctamente el valor perdido:

```
cbind(apply(is.na(datos_pca),2,sum),apply(is.na(datos_pca),2,sum)/dim(datos_pca)[1])
```

```
##          [,1] [,2]
## ZPOBDENS    0    0
## ZTMINFAN    0    0
## ZESPVIDA    0    0
## ZPOBURB     0    0
## ZTMEDICO    0    0
## ZPAGRICU    0    0
## ZPSERVI     0    0
## ZTLIBROP    0    0
## ZTEJERCI    0    0
## ZTPOBACT    0    0
## ZTENERGI    0    0
```

En efecto se ha realizado la imputación de valores perdidos sin ningún problema.

Recodificación

En este caso no es necesaria.

Exploración Univariante

Exploración descriptiva

En este apartado iremos variable por variable obteniendo los resultados de aplicar diferentes medidas descriptivas, clásicas y resistentes, de centralidad, forma y dispersión.

```
#Definimos las medidas resistentes
PMC<-function(x){ return((as.double(quantile(x,0.25))+as.double(quantile(x,0.75)))/2)}

trimedia<-function(x){return((median(x)+PMC(x))/2)}

centrimedia<-function(x){
  indices<-(x>quantile(x,0.25)&x<quantile(x,0.75))
  valores<-x[indices]
  return(sum(valores)/length(valores))
}

RIQ<-function(x){return(quantile(x,0.75)-quantile(x,0.25))}

MEDA<-function(x){return(median(abs(x-median(x))))}

CVc<-function(x){return((quantile(x,0.75)-quantile(x,0.25))/(quantile(x,0.75)+quantile(x,0.25)))}

H1<-function(x){return((quantile(x,0.25)+quantile(x,0.75)-2*median(x))/(2*median(x)))}
H2<-function(x){return(median(x)-(quantile(x,0.1)+quantile(x,0.9))/(2))}
H3<-function(x){return(H2(x)/median(x))}

#Creamos una función que aplique todas estas medidas

descriptivo<-function(x){

  temp<-rbind(PMC(x),trimedia(x),centrimedia(x))
  rownames(temp)<-c("PMC","Trimedia","Centrimedia")
  centralidad<-list(clasica=list(media=mean(x)),resistente=temp)

  temp<-rbind(RIQ(x),MEDA(x),CVc(x))
  rownames(temp)<-c("Rango Inter-Cuartílico","MEDA","CVc")
  dispersion<-list(clasica=list(desviación_típica=sd(x),Coef_varización=sd(x)/mean(x),rango=range(x)),r

  temp<-rbind(H1(x),H2(x),H3(x))
  rownames(temp)<-c("Asimetría de Yule","Asimetría de Kelly","Asimetría de Kelly adimensional")
  forma<-list(clasica=list(skewness=skewness(x),kurtosis=kurtosis(x)),resistente=temp)
  cat(names(x))
  return(list(centralidad=centralidad,dispersion=dispersion,forma=forma))
}
```

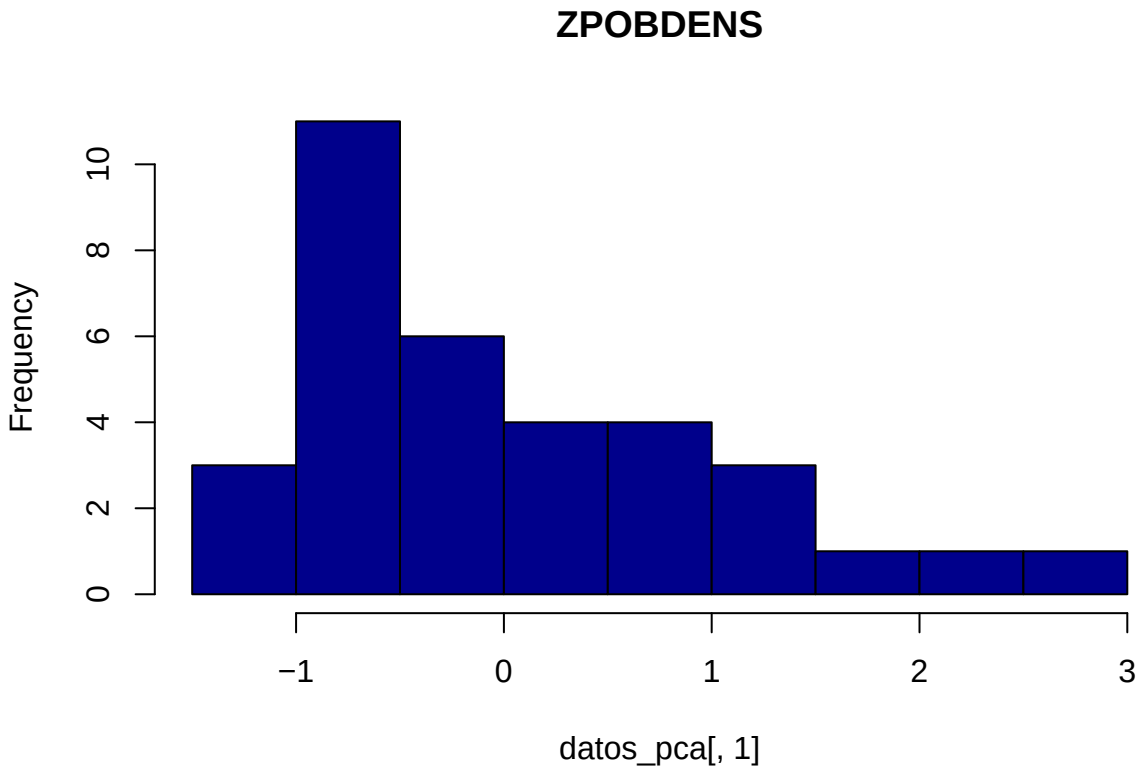
Aplicamos la función para cada una de las variables:

ZPOBDENS

```
descriptivo(datos_pca[,1])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] 1.06373e-16
##
##
## $centralidad$resistente
##                    [,1]
## PMC         -0.1474835
## Trimedia    -0.1545405
## Centrimedia -0.2293829
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] 9.400883e+15
##
## $dispersion$clasica$rango
## [1] -1.077834  2.861621
##
##
## $dispersion$resistente
##                               75%
## Rango Inter-Cuartílico  1.4043955
## MEDA                    0.6924864
## CVc                    -4.7611940
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] 1.11789
##
## $forma$clasica$kurtosis
## [1] 0.8308405
##
##
## $forma$resistente
##                                        25%
## Asimetría de Yule              -0.08733974
## Asimetría de Kelly             -0.37369403
## Asimetría de Kelly adimensional  2.31250000
```

```
hist(col="darkblue",datos_pca[,1],main="ZPOBDENS")
```



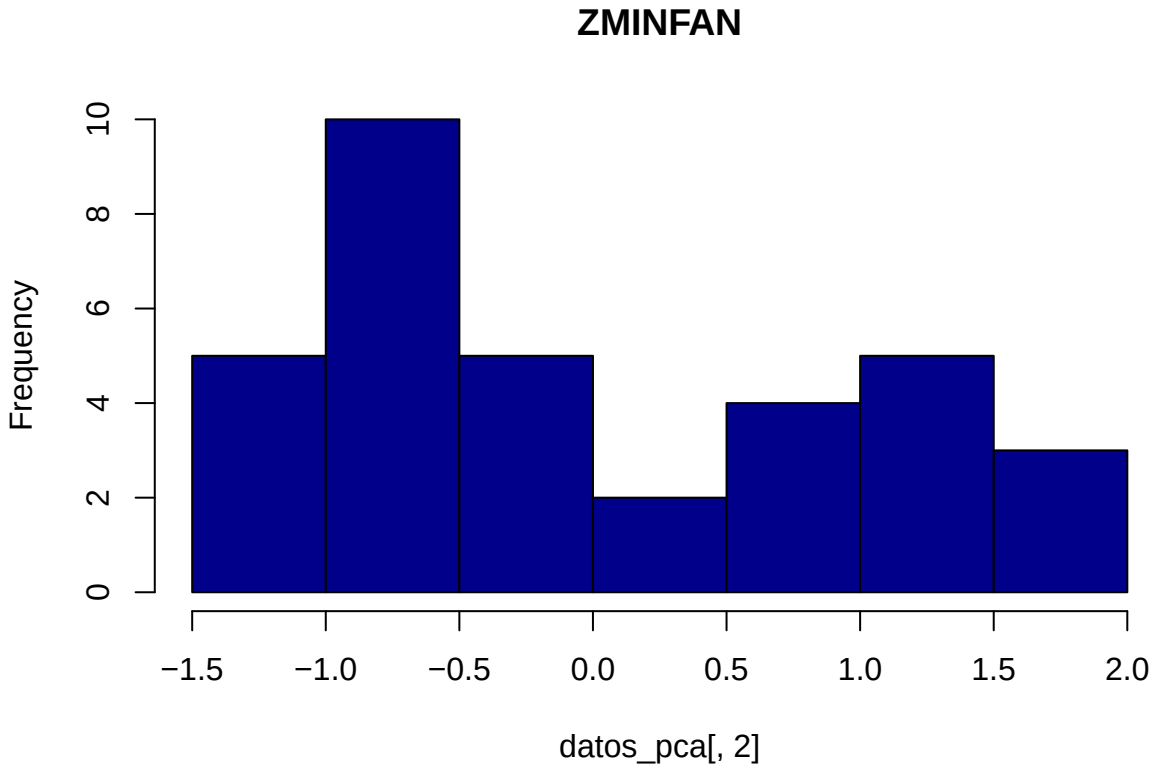
Las medidas resistentes de centralidad están ligeramente desplazadas hacia la izquierda. Tenemos un valor del *MEDA* inferior al de la desviación típica. En los estimadores de simetría, obtenemos que existe cierta asimetría, y el valor de *kurtosis* indica una acumulación de los datos.

ZMINFAN

```
descriptivo(datos_pca[,2])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] 8.982296e-17
##
##
## $centralidad$resistente
##                    [,1]
## PMC         -0.0300511
## Trimedia    -0.2115862
## Centrimedia -0.2268480
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] 1.113301e+16
##
## $dispersion$clasica$rango
## [1] -1.102554  1.904824
##
##
## $dispersion$resistente
##                                75%
## Rango Inter-Cuartílico   1.8571199
## MEDA                     0.6040459
## CVc                    -30.8993711
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] 0.5733687
##
## $forma$clasica$kurtosis
## [1] -1.202237
##
##
## $forma$resistente
##                                      25%
## Asimetría de Yule              -0.9235577
## Asimetría de Kelly             -0.6132994
## Asimetría de Kelly adimensional  1.5600769
```

```
hist(col="darkblue",datos_pca[,2],main="ZMINFAN")
```



Lo primero que llama la atención en el histograma es que a prmera vista parecen dos normales pegadas, pero esto puede deberse a que tenemos muy pocos datos en el conjunto a estudiar y sea una normal con una cola larga.

La media se encuentra ligeramente desviada a la izquierda, y de nuevo la MEDA es menor que la desviación típica.

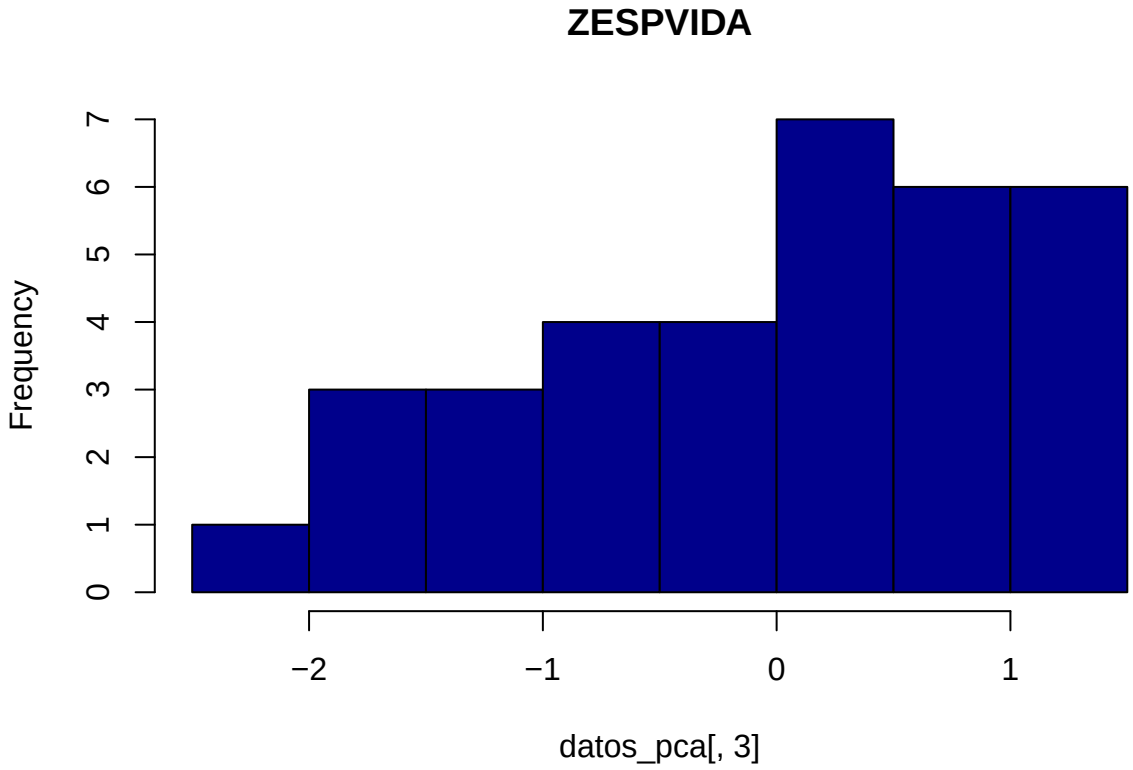
Vemos en el coeficiente de curtosis que la distribución es muy aplanada y asimetría, lo que se corresponde con el histograma.

ZESPVIDA

```
descriptivo(datos_pca[,3])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] -1.72884e-16
##
##
## $centralidad$resistente
##                    [,1]
## PMC         0.07863105
## Trimedia    0.17836465
## Centrimedia 0.17613181
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] -5.784225e+15
##
## $dispersion$clasica$rango
## [1] -2.145279  1.248640
##
##
## $dispersion$resistente
##                                75%
## Rango Inter-Cuartílico  1.6493257
## MEDA                    0.7621433
## CVc                    10.4877506
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] -0.6073997
##
## $forma$clasica$kurtosis
## [1] -0.7827668
##
##
## $forma$resistente
##                                       25%
## Asimetría de Yule               -0.7172544
## Asimetría de Kelly               0.4995611
## Asimetría de Kelly adimensional  1.7963476
```

```
hist(col="darkblue",datos_pca[,3],main="ZESPVIDA")
```



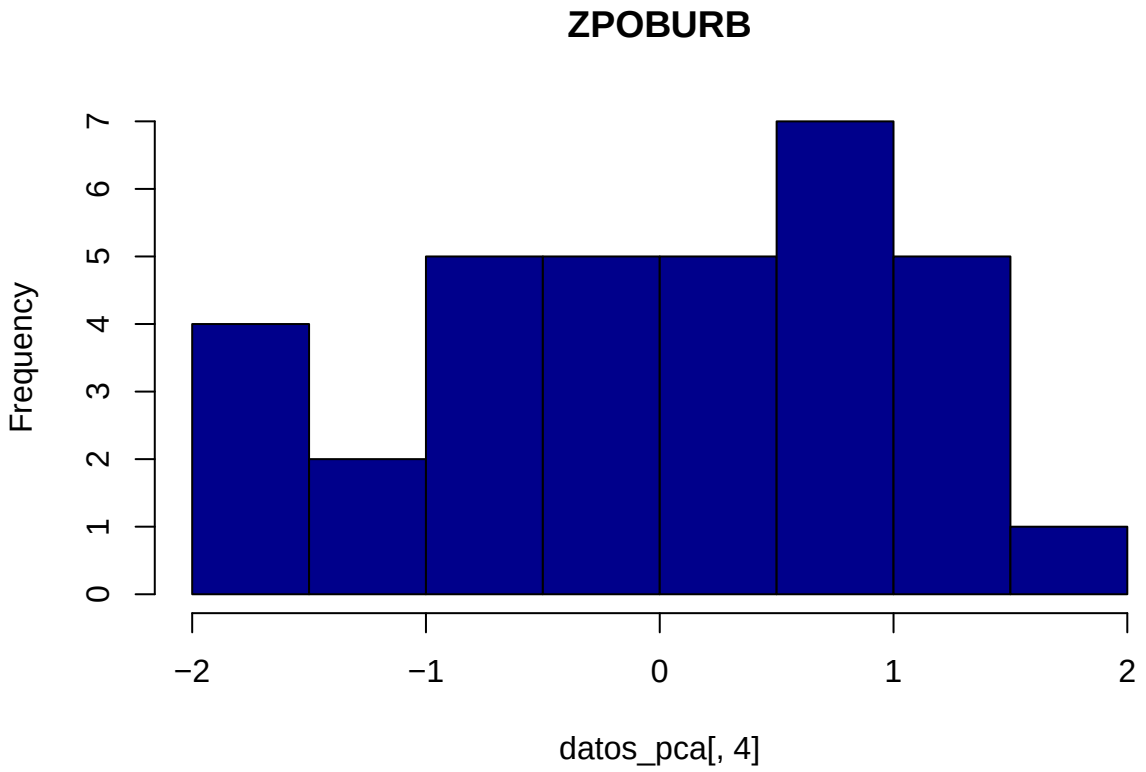
Las medidas de centralidad se encuentran desplazadas a la derecha, y la MEDA vuelve a ser menor que la desviación típica. De nuevo mirando el coeficiente de curtosis vemos una distribución bastante aplanada y con asimetría.

ZPOBURB

```
descriptivo(datos_pca[,4])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] 4.511206e-16
##
##
## $centralidad$resistente
##                   [,1]
## PMC         0.0413792
## Trimedia    0.0840905
## Centrimedia 0.1147624
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] 2.216702e+15
##
## $dispersion$clasica$rango
## [1] -1.769694  1.509616
##
##
## $dispersion$resistente
##                                75%
## Rango Inter-Cuartílico  1.5467796
## MEDA                    0.7223655
## CVc                    18.6903015
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] -0.3339002
##
## $forma$clasica$kurtosis
## [1] -1.07826
##
##
## $forma$resistente
##                                       25%
## Asimetría de Yule              -0.6736702
## Asimetría de Kelly              0.2958259
## Asimetría de Kelly adimensional  2.3329787
```

```
hist(col="darkblue",datos_pca[,4],main="ZPOBURB")
```



Esta distribución parece bastante plana mirando el histograma, ningún valor parece predominar demasiado sobre los otros, pero de nuevo se necesitarían más datos para confirmar.

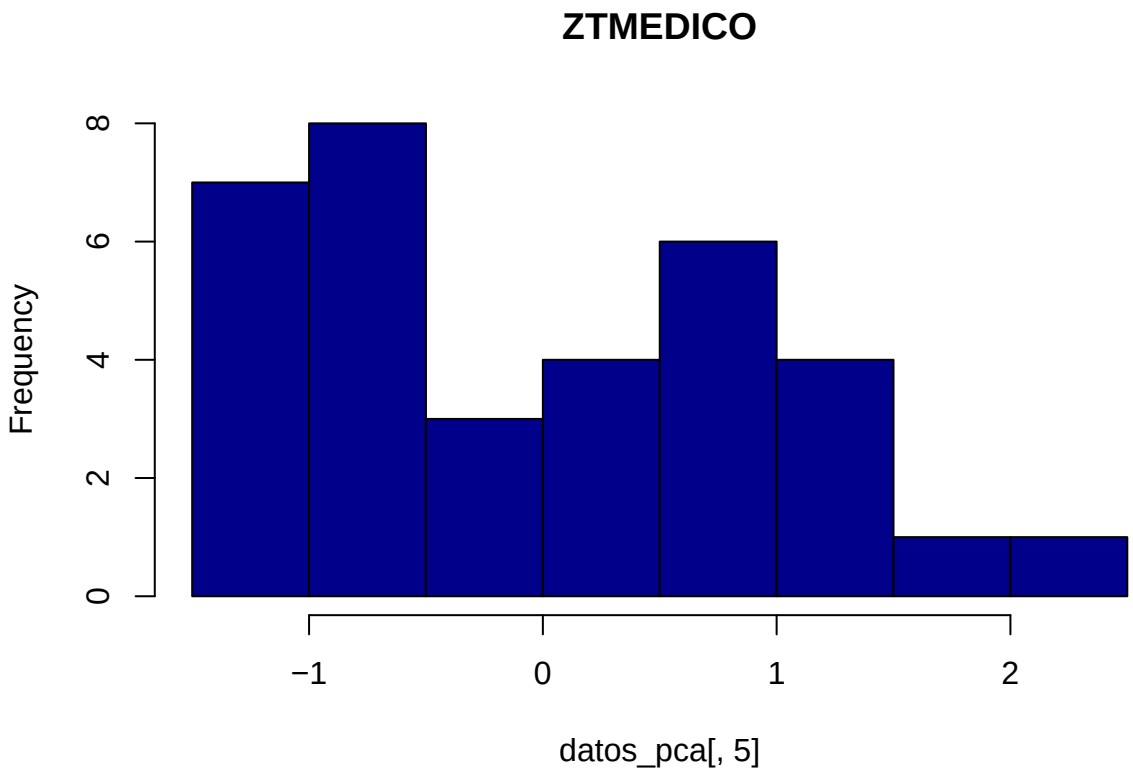
La media y medidas de centralidad están muy cercanas a 0 y la curtosis indica que la distribución es bastante plana.

ZTMEDICO

```
descriptivo(datos_pca[,5])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] 2.416495e-17
##
##
## $centralidad$resistente
##                       [,1]
## PMC         -0.006716126
## Trimedia    -0.149133349
## Centrimedia -0.149734266
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] 4.138224e+16
##
## $dispersion$clasica$rango
## [1] -1.147256  2.371712
##
##
## $dispersion$resistente
##                                  75%
## Rango Inter-Cuartílico    1.7522727
## MEDA                      0.7980172
## CVc                    -130.4526316
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] 0.5139514
##
## $forma$clasica$kurtosis
## [1] -0.8884781
##
##
## $forma$resistente
##                                        25%
## Asimetría de Yule               -0.9769641
## Asimetría de Kelly              -0.3735297
## Asimetría de Kelly adimensional  1.2811833
```

```
hist(col="darkblue",datos_pca[,5],main="ZTMEDICO")
```



Las medidas de centralidad se encuentran desplazadas hacia la izquierda y el CVc tiene un valor muy bajo. La curtosis indica que la distribución es plana

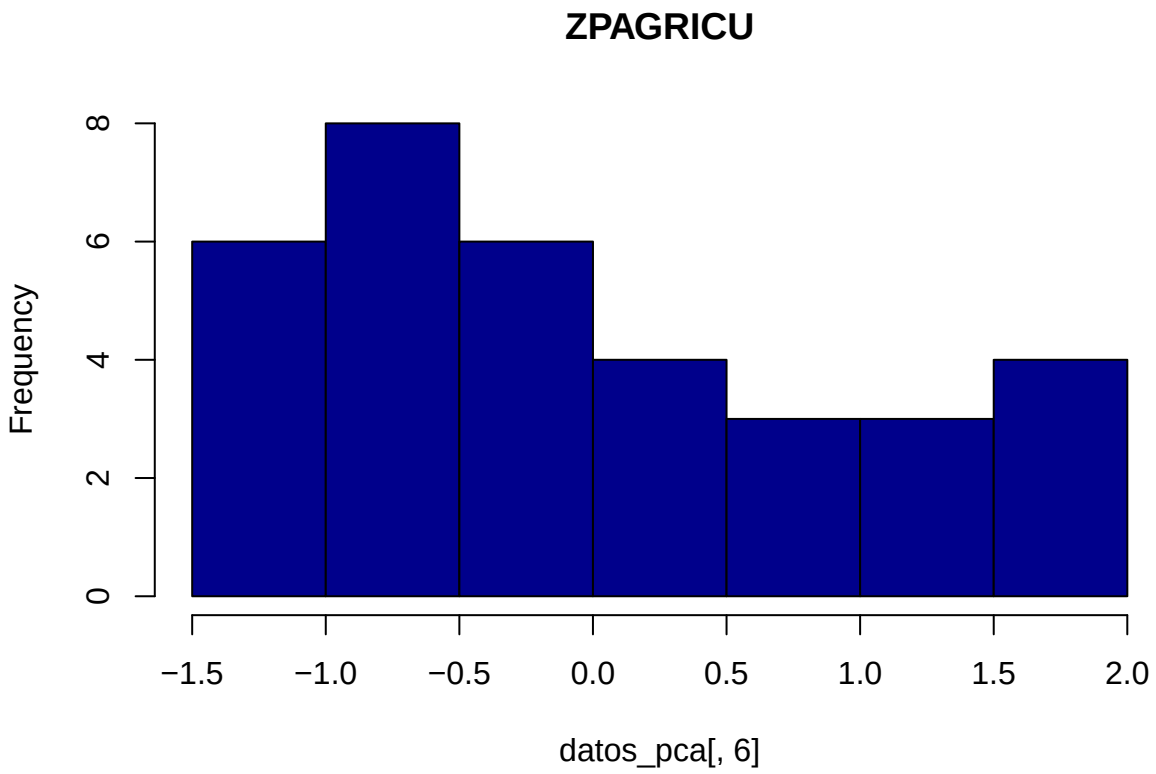
También llama la atención como las dos primeras columnas son muy altas y a partir de ellas el resto del histograma se asemeja más a una normal.

ZPAGRICU

```
descriptivo(datos_pca[,6])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] 6.355338e-17
##
##
## $centralidad$resistente
##                     [,1]
## PMC         -0.06825485
## Trimedia    -0.14081091
## Centrimedia -0.20433276
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] 1.57348e+16
##
## $dispersion$clasica$rango
## [1] -1.234234  1.905157
##
##
## $dispersion$resistente
##                                75%
## Rango Inter-Cuartílico   1.5595319
## MEDA                     0.7069276
## CVc                    -11.4243309
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] 0.6194981
##
## $forma$clasica$kurtosis
## [1] -0.8999068
##
##
## $forma$resistente
##                                      25%
## Asimetría de Yule              -0.6801059
## Asimetría de Kelly             -0.4817497
## Asimetría de Kelly adimensional  2.2578456
```

```
hist(col="darkblue",datos_pca[,6],main="ZPAGRICU")
```



En este caso el histograma se asemeja más al de una normal que en las anteriores variables estudiadas, pese a que se aprecia una asimetría y distribución bastante plana.

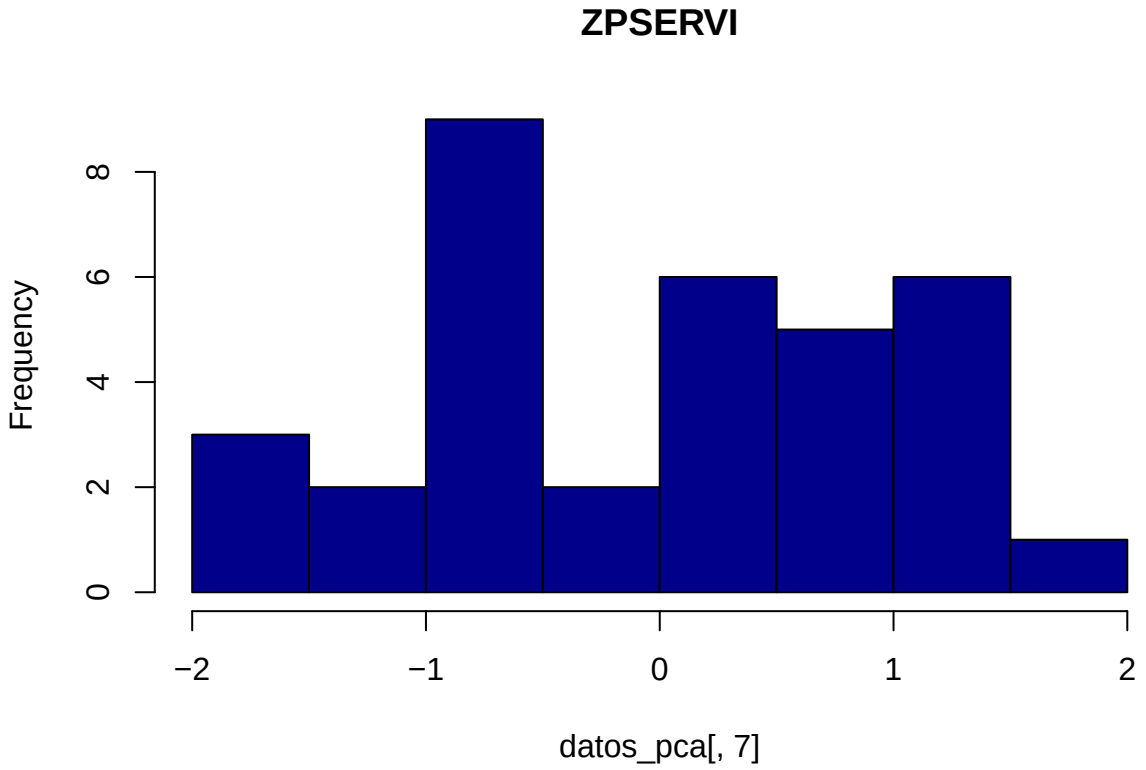
Las medidas de centralidad están desviadas ligeramente a la izquierda y la curtosis indica que la distribución es bastante plana.

ZPSERVI

```
descriptivo(datos_pca[,7])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] -3.529078e-17
##
##
## $centralidad$resistente
##                       [,1]
## PMC         0.0615893232
## Trimedia    0.0485015920
## Centrimedia 0.0006495749
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] -2.833601e+16
##
## $dispersion$clasica$rango
## [1] -1.885211  1.628845
##
##
## $dispersion$resistente
##                                75%
## Rango Inter-Cuartílico  1.5803435
## MEDA                    0.7918077
## CVc                    12.8296875
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] -0.1297167
##
## $forma$clasica$kurtosis
## [1] -1.068007
##
##
## $forma$resistente
##                                        25%
## Asimetría de Yule               0.73913043
## Asimetría de Kelly              0.05071496
## Asimetría de Kelly adimensional 1.43206522
```

```
hist(col="darkblue",datos_pca[,7],main="ZPSERVI")
```



Observamos que las medidas de centralidad están desviadas a la derecha, el MEDA es menor que la desviación típica y la curtosis indica que la distribución es plana.

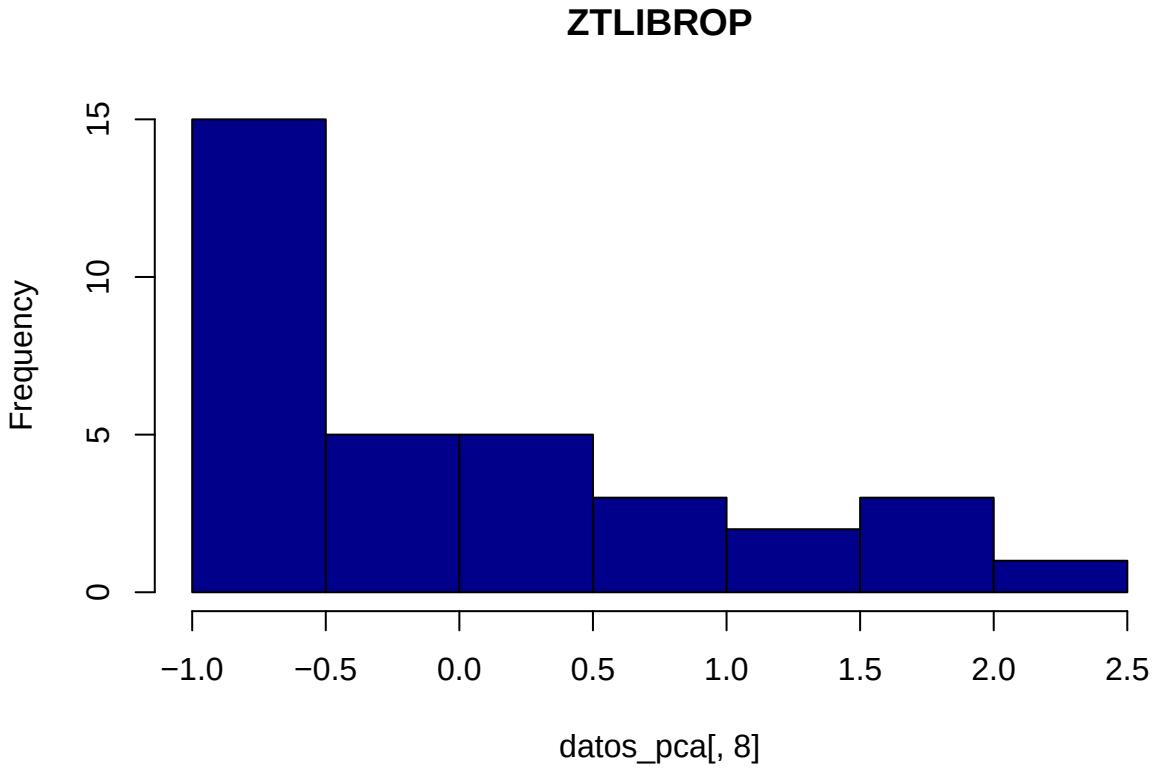
Además, es claro viendo el histograma que la distribución no se asemeja demasiado a la normal, y es posible que la última columna, al tener frecuencia 1, se trae de un outlier.

ZTLIBROP

```
descriptivo(datos_pca[,8])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] 6.729069e-17
##
##
## $centralidad$resistente
##                    [,1]
## PMC         -0.1111009
## Trimedia    -0.1776580
## Centrimedia -0.2615358
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 0.9847319
##
## $dispersion$clasica$Coef_varización
## [1] 1.4634e+16
##
## $dispersion$clasica$rango
## [1] -0.9696156  2.4023604
##
##
## $dispersion$resistente
##                               75%
## Rango Inter-Cuartílico  1.6067690
## MEDA                    0.6849277
## CVc                    -7.2311230
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] 0.8552831
##
## $forma$clasica$kurtosis
## [1] -0.3544396
##
##
## $forma$resistente
##                                       25%
## Asimetría de Yule              -0.5450695
## Asimetría de Kelly             -0.5339910
## Asimetría de Kelly adimensional  2.1865595
```

```
hist(col="darkblue",datos_pca[,8],main="ZTLIBROP")
```



De nuevo ne este caso las medidas de centralidad están desviadas a la izquierda, además se denota una fuerte asimetría en la distribución.

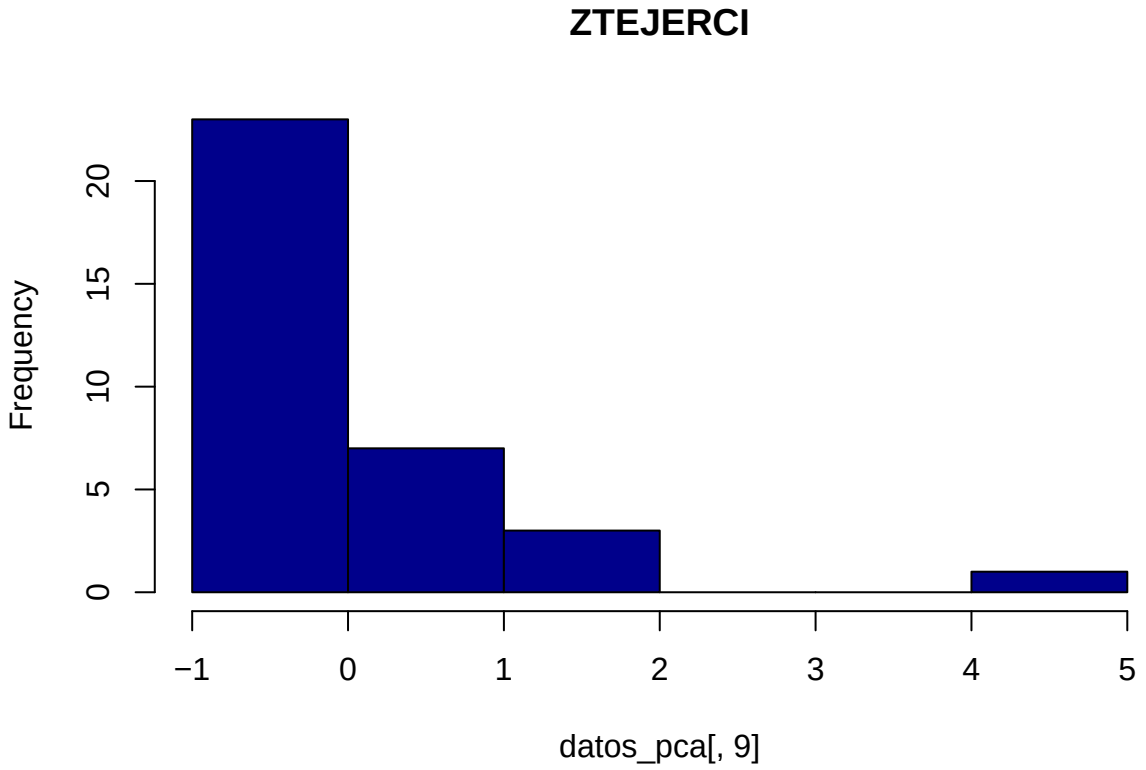
También llama la atención como el rango de esta varuiable es mayor hacia la derecha que el de las demás.

ZTEJERCI (variable con conclusiones sesgadas)

```
descriptivo(datos_pca[,9])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] 1.859184e-16
##
##
## $centralidad$resistente
##                    [,1]
## PMC         -0.2594621
## Trimedia    -0.2328606
## Centrimedia -0.2192510
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] 5.378705e+15
##
## $dispersion$clasica$rango
## [1] -0.8658606  4.4262018
##
##
## $dispersion$resistente
##                               75%
## Rango Inter-Cuartílico  0.6788462
## MEDA                    0.3313997
## CVc                    -1.3081800
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] 2.999371
##
## $forma$clasica$kurtosis
## [1] 11.41767
##
##
## $forma$resistente
##                                      25%
## Asimetría de Yule               0.2579423
## Asimetría de Kelly             -0.3448882
## Asimetría de Kelly adimensional  1.6721112
```

```
hist(col="darkblue",datos_pca[,9],main="ZTEJERCI")
```



En este caso tenemos unas medidas de centralidad muy desviadas a la izquierda, con una fuerte asimetría y una concentración de datos muy alta en la primera columna.

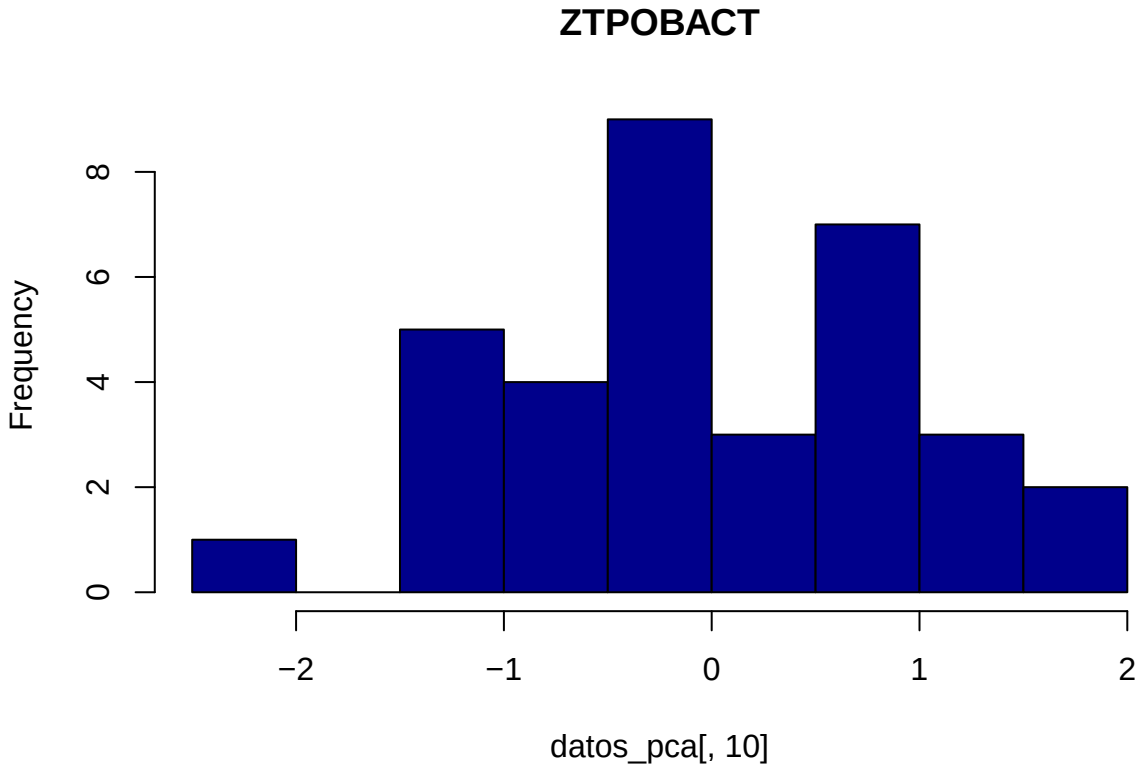
El rango de esta variable es mucho más amplio que el de todas las demás, y resulta llamativo como los últimos valores de la derecha no tienen una frecuencia alta, haciendo sospechar de que sean outliers.

ZTPOBACT

```
descriptivo(datos_pca[,10])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] -2.952027e-16
##
##
## $centralidad$resistente
##                     [,1]
## PMC          0.10268877
## Trimedia    -0.00198850
## Centrimedia  0.02917055
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] -3.387502e+15
##
## $dispersion$clasica$rango
## [1] -2.134094  1.704472
##
##
## $dispersion$resistente
##                               75%
## Rango Inter-Cuartílico 1.5523378
## MEDA                   0.8557514
## CVc                    7.5584589
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] -0.1344579
##
## $forma$clasica$kurtosis
## [1] -0.8348377
##
##
## $forma$resistente
##                                         25%
## Asimetría de Yule               -1.96271531
## Asimetría de Kelly              -0.06440751
## Asimetría de Kelly adimensional  0.60382547
```

```
hist(col="darkblue",datos_pca[,10],main="ZTPOBACT")
```



Esta distribución se parece más a una normal que algunas de las estudiadas anteriormente, está más centrada (sus medidas de centralidad son más cercanas a 0) y su curtosis es baja, lo que indica una distribución aplanada.

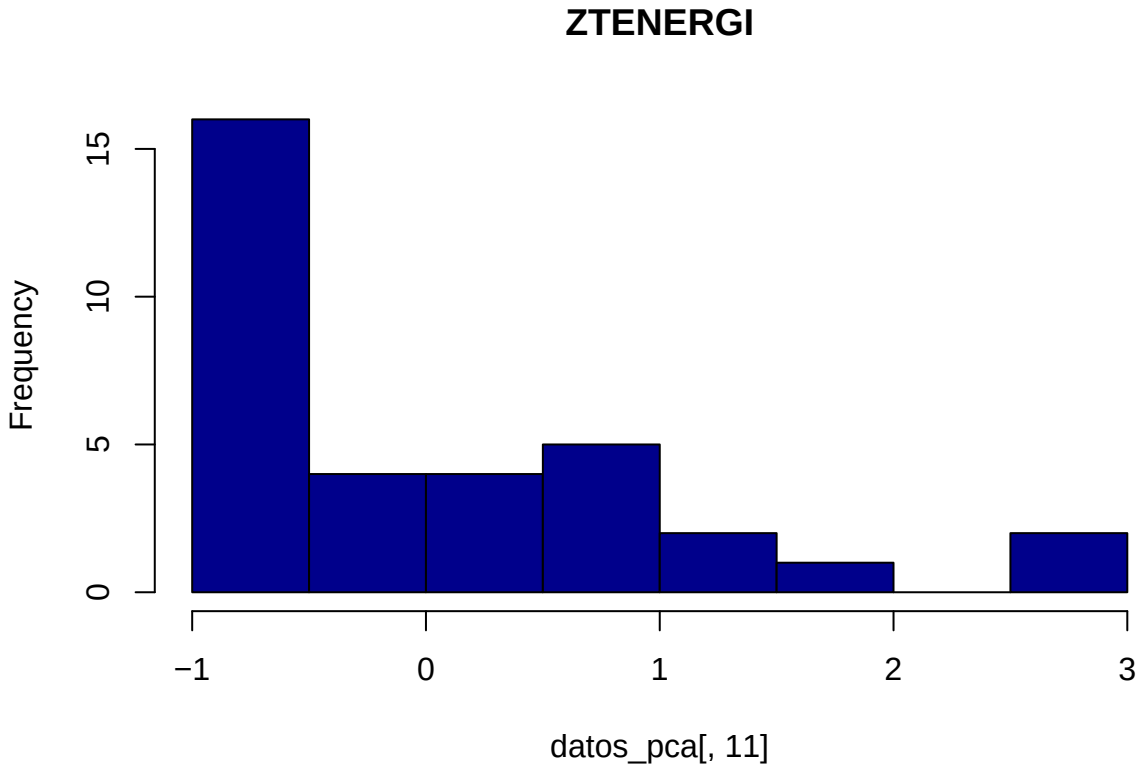
Mirando el histograma apreciamos algunos valores cerca de -2 que pueden parecer outliers por su separación con el resto de los valores representados.

ZTENERGI

```
descriptivo(datos_pca[,11])
```

```
## $centralidad
## $centralidad$clasica
## $centralidad$clasica$media
## [1] 3.586955e-17
##
##
## $centralidad$resistente
##                     [,1]
## PMC         -0.09924855
## Trimedia    -0.24461072
## Centrimedia -0.26271176
##
##
## $dispersion
## $dispersion$clasica
## $dispersion$clasica$desviación_típica
## [1] 1
##
## $dispersion$clasica$Coef_varización
## [1] 2.78788e+16
##
## $dispersion$clasica$rango
## [1] -0.950661  2.749785
##
##
## $dispersion$resistente
##                               75%
## Rango Inter-Cuartílico  1.364062
## MEDA                    0.520457
## CVc                    -6.871948
##
##
## $forma
## $forma$clasica
## $forma$clasica$skewness
## [1] 1.284408
##
## $forma$clasica$kurtosis
## [1] 1.262001
##
##
## $forma$resistente
##                                       25%
## Asimetría de Yule              -0.7454988
## Asimetría de Kelly             -0.5056537
## Asimetría de Kelly adimensional  1.2966382
```

```
hist(col="darkblue",datos_pca[,11],main="ZTENERGI")
```



Al igual que en uno de los casos anteriores tenemos las medidas de centralidad desviadas a la izquierda, una curtosis bastante alta y un rango de valores amplio, con algunos separados de los demás en los extremos que parecen outliers.

NOTAR: el hecho de que algunas de las variables estén desplazadas hacia la derecha, o que los outliers sean en esta dirección, es debido a que estas son estandarizaciones de variables positivas, en una variable positiva el valor mas extremo inferior como mucho es 0.

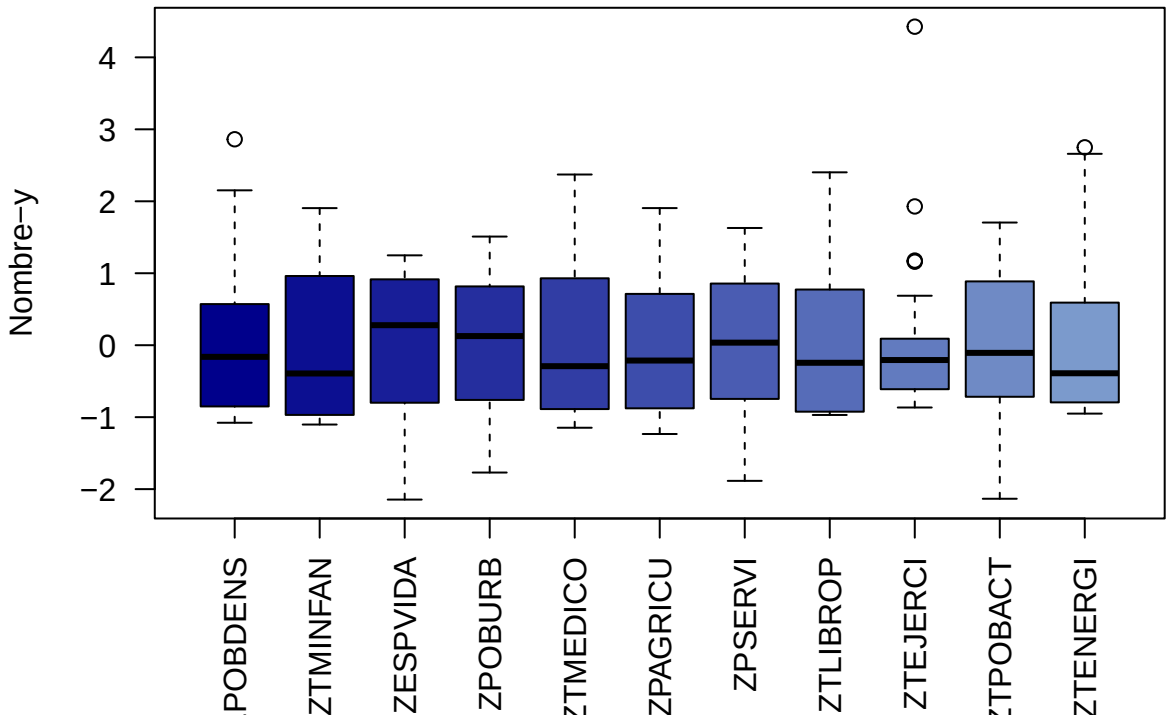
Las que si tienen extensión hacia la izquierda, es debido a que no hay un gran porcentaje de la población que se acerque al extremo inferior, entonces el 0 realmente si acaba siendo un valor "extremo" para estas variables.

También notar que el CVc toma valores bastante "sin sentido" debido a que esta medida no tiene mucho sentido en variables que toman valores positivos y negativos, en el caso de la variable 15 ese valor es devido a que $Q_1 \approx -Q_3$

Exploración Gráfica

Procedemos con los diagramas de cajas y bigotes, para una detección primaria de outliers univariantes.

```
colfunc<-colorRampPalette(c("darkblue","lightblue"))
boxplot(datos_pca,
        xlab=NULL,
        ylab="Nombre-y",
        col=colfunc(15),
        las=2)
```

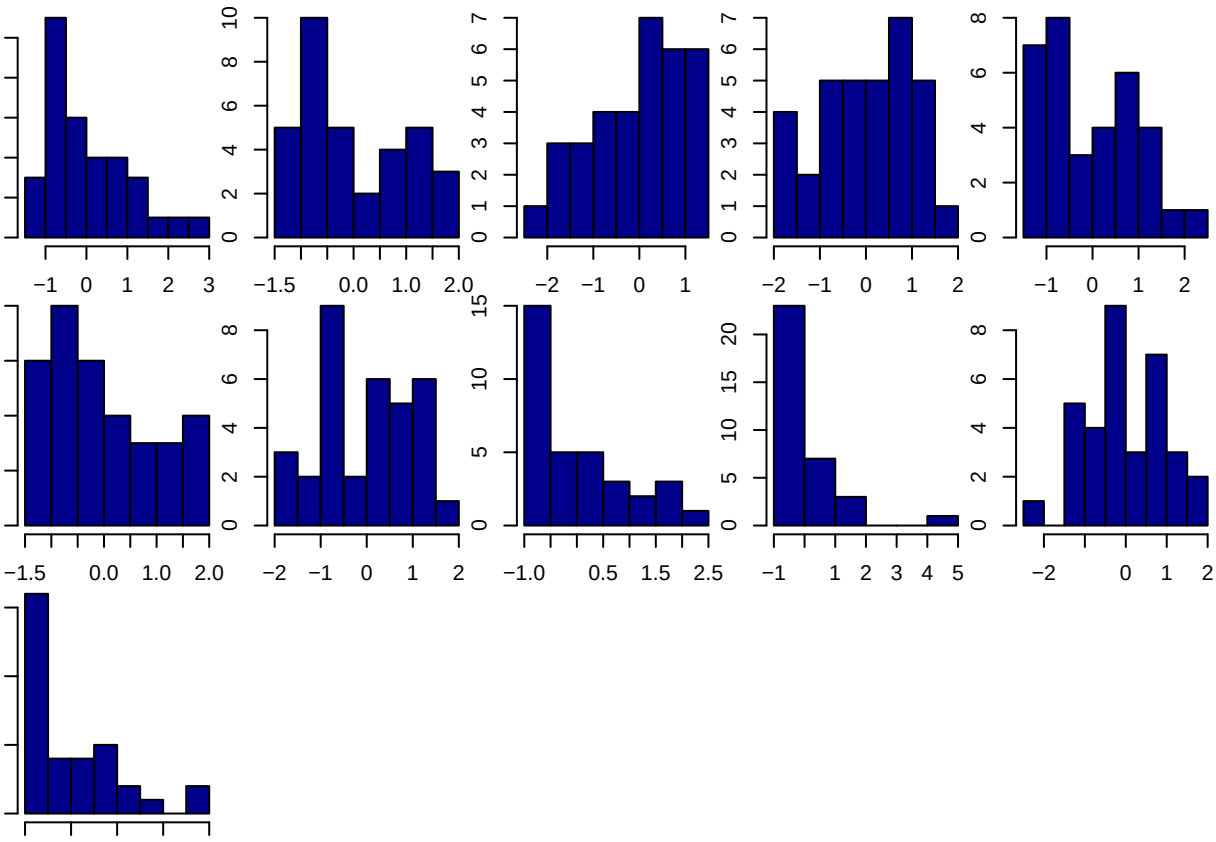


Se puede observar que la mayoría de gráficos se asemejan en cierto modo a una normal,estando todos ligeramente desplazados, y destacando que POBDENS y ZTEJERCI tienen una mayor concentración de valores. La mayoría de outliers se encuentran en ZTEJERCI como cola superior.

Pese a que algunas de las distribuciones de las variables que presentan outliers no se parecen demasiado a una normal, se ha tomado la decisión de eliminarlos, a pesar de que pueda no ser lo más correcto, ya que boxplot asume que la distribución es normal y en algunos casos esto parece dudoso.

A continuación vamos a observar (de nuevo) la forma de las distribuciones de las variables mediante sus histogramas

```
par(mar=c(1,1,1,1))
par(mfrow=c(3,5))
invisible(apply(datos_pca, 2,function(x){hist(x,main=NULL,col="darkblue",xlab=NULL,ylab=NULL)}))
```



Observamos que realmente pocas de las variables realmente se podrían considerar normales con los datos que tenemos, pero son muy pocos, por lo que quizá al recoger una muestra mayor de observaciones se podrían sacar mejores conclusiones sobre las distribuciones de las variables.

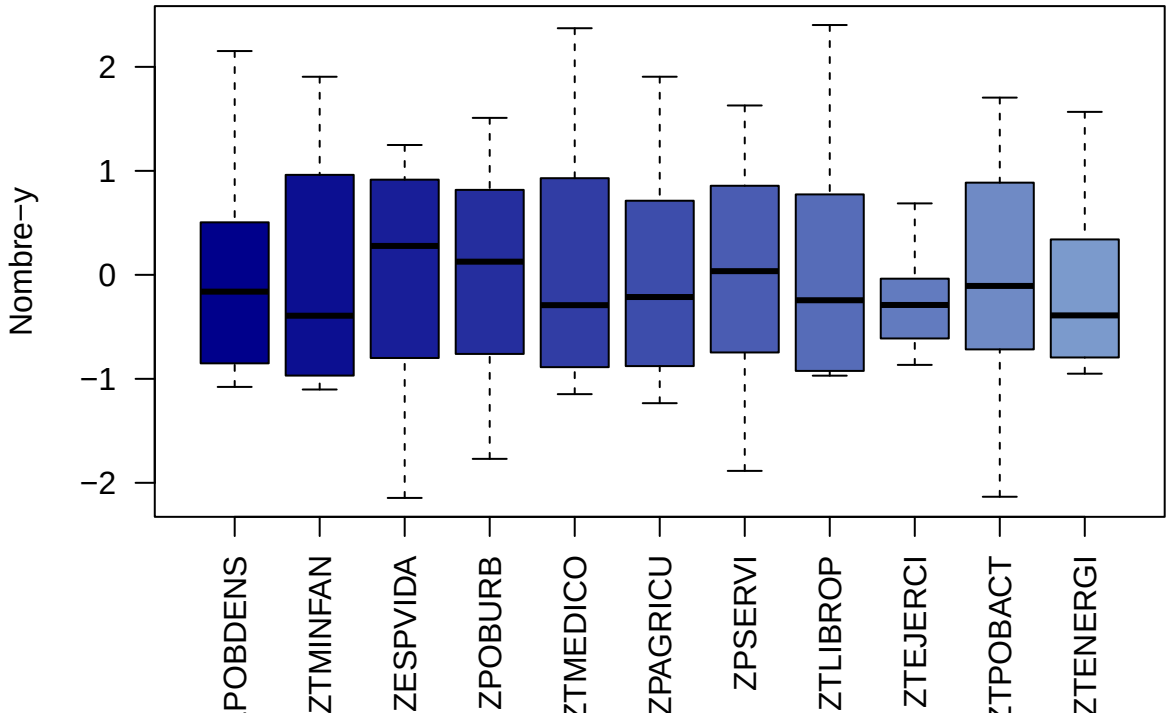
Tratamiento de outliers

Los valores outlier comentados anteriormente (según si el diagrama boxplot los consideraba outliers o no) serán intercambiados por la media.

```
outlier<-function(data,na.rm=T){
  H<-1.5*IQR(data)

  if(any(data<=(quantile(data,0.25,na.rm = T)-H))){
    data[data<=quantile(data,0.25,na.rm = T)-H]<-NA
    data[is.na(data)]<-mean(data,na.rm=T)
    data<-outlier(data)}

  if(any(data>=(quantile(data,0.75, na.rm = T)+H))){
    data[data>=quantile(data,0.75, na.rm = T)+H]<-NA
    data[is.na(data)]<-mean(data,na.rm=T)
    data<-outlier(data)
  }
  return(data)
}

datos_pca[,-12:-13]<-apply(datos_pca[,-12:-13], 2, outlier)
```

Una vez tratados los outliers, vemos de nuevo los gráficos boxplot.

```
boxplot(datos_pca,
        xlab=NULL,
        ylab="Nombre-y",
        col=colfunc(15),
        las=2)
```

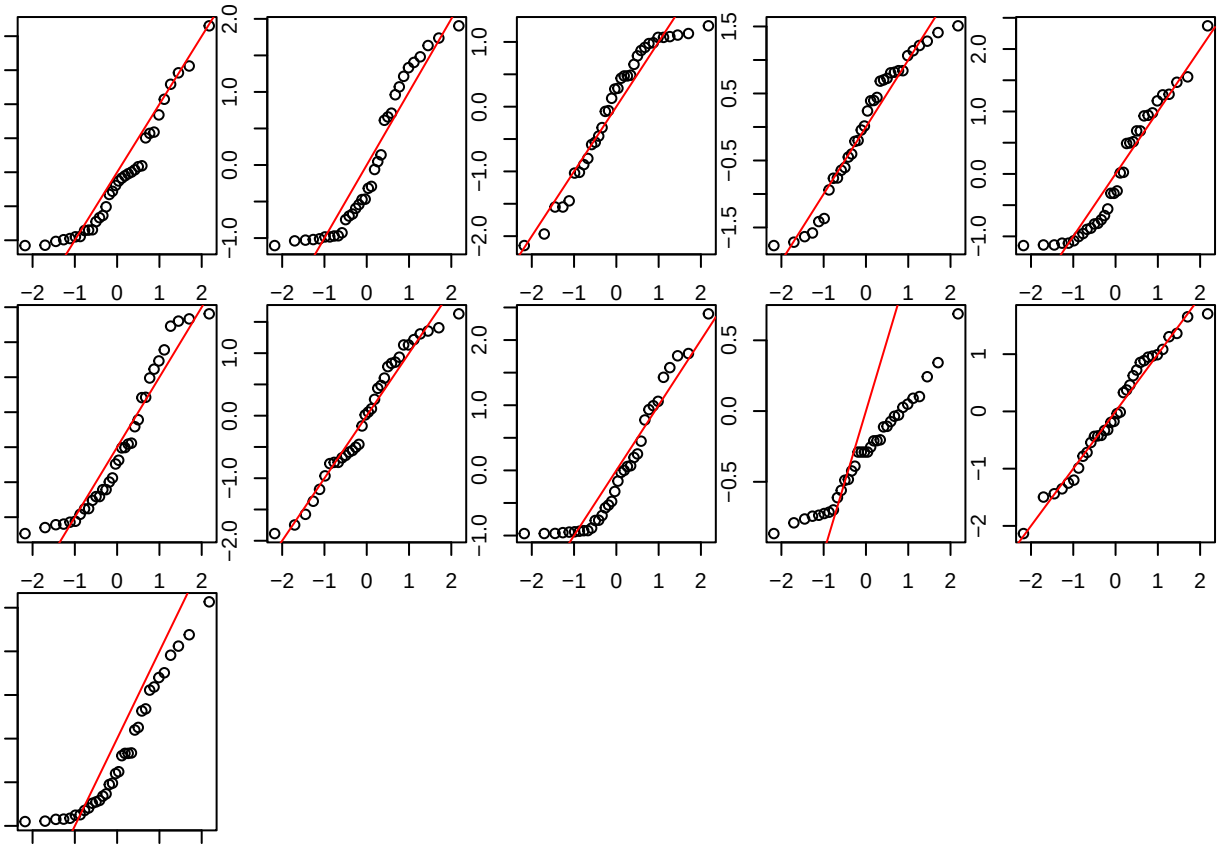


Así comprobamos que hemos eliminado los outliers.

Normalidad

Para poder aplicar ciertas técnicas estadísticas, es importante saber si estamos tratando con variables normales, para ello usaremos el método gráfico *qqplot*.

```
par(mar=c(1,1,1,1))
par(mfrow=c(3,5))
invisible(apply(datos_pca, 2, function(x){
  qqnorm(x,main=NULL)
  abline(a=0,b=1,col="red")
}))
```



Vemos como casi todas las variables salvo la novena se ajustan bastante bien a la normalidad, siendo las cuarta y décima las que mejor se acercan a ella.

En este caso no tomaremos medidas para obtener normalidad en los datos.

Homocedasticidad

La homocedasticidad se debe comparar dentro de una misma variable, para dos o mas grupos diferenciados; en el caso de este dataset, podemos comprobar si existe homocedasticidad entre los grupos definidos por sus continentes.

Comenzamos añadiendo a los datos completos la variable continente:

```
datos_enteros$continente <- c(
  "africa", "africa", "america", "oceania",
  "america", "america", "america", "asia",
  "asia", "africa", "europa", "asia",
  "europa", "europa", "asia", "asia",
  "asia", "asia", "europa", "asia", "asia",
  "africa", "america", "africa", "asia", "europa",
  "europa", "europa", "europa", "europa",
  "europa", "asia", "america", "asia")
```

Comprobamos ahora la homocedasticidad:

```
ind<-which(datos_enteros$continente=="europa"|datos_enteros$continente=="asia"|datos_enteros$continente=
factores<-datos_enteros$continente[ind]
#Como se han eliminado los valores outlier, usamos con centro la media en vez de la mediana
#H0:homocedasticidad
apply(datos_pca[ind,], 2, function(x){
  if(leveneTest(x,as.factor(factores),center=median)$"Pr(>F)"[1]>0.05){
    "Existe homocedasticidad entre los grupos"
  }
  else{"No existe homocedasticidad entre los grupos"}
  })
```

```
##                                       ZPOBDENS
##    "Existe homocedasticidad entre los grupos"
##                                       ZTMINFAN
##    "Existe homocedasticidad entre los grupos"
##                                       ZESPVIDA
## "No existe homocedasticidad entre los grupos"
##                                        ZPOBURB
## "No existe homocedasticidad entre los grupos"
##                                       ZTMEDICO
##    "Existe homocedasticidad entre los grupos"
##                                       ZPAGRICU
##    "Existe homocedasticidad entre los grupos"
##                                        ZPSERVI
##    "Existe homocedasticidad entre los grupos"
##                                       ZTLIBROP
##    "Existe homocedasticidad entre los grupos"
##                                       ZTEJERCI
##    "Existe homocedasticidad entre los grupos"
##                                       ZTPOBACT
##    "Existe homocedasticidad entre los grupos"
##                                       ZTENERGI
##    "Existe homocedasticidad entre los grupos"
```

Vemos como para la mayoría de variables, seccionando los paises según continente, se tiene la misma varianza. Donde no se cumple es para el índice de ZESPVIDA y ZPOBURB.

Exploración Descriptiva

A continuación vamos a sacar los principales estadísticos descriptivos ahora que tenemos los datos transformados.

Originales

```
summary(datos)
```

```
##     ZPOBDENS          ZTMINFAN          ZESPVIDA          ZPOBURB
##  Min.   :-1.0778   Min.   :-1.1026   Min.   :-2.1453   Min.   :-1.7697
##  1st Qu.:-0.8497   1st Qu.:-0.9586   1st Qu.:-0.7460   1st Qu.:-0.7320
##  Median :-0.1616   Median :-0.3931   Median : 0.2781   Median : 0.1268
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.5547   3rd Qu.: 0.8985   3rd Qu.: 0.9033   3rd Qu.: 0.8148
##  Max.   : 2.8616   Max.   : 1.9048   Max.   : 1.2486   Max.   : 1.5096
##
##     ZTMEDICO          ZPAGRICU          ZPSERVI           ZTLIBROP
##  Min.   :-1.1473   Min.   :-1.2342   Min.   :-1.88521   Min.   :-0.9696
##  1st Qu.:-0.8829   1st Qu.:-0.8480   1st Qu.:-0.72858   1st Qu.:-0.9240
##  Median :-0.2916   Median :-0.2134   Median : 0.03541   Median :-0.3237
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.8694   3rd Qu.: 0.7115   3rd Qu.: 0.85176   3rd Qu.: 0.7736
##  Max.   : 2.3717   Max.   : 1.9052   Max.   : 1.62885   Max.   : 2.4024
##                                                         NA's   :1
##     ZTEJERCI          ZTPOBACT          ZTENERGI
##  Min.   :-0.86586   Min.   :-2.1341   Min.   :-0.9507
##  1st Qu.:-0.59889   1st Qu.:-0.6735   1st Qu.:-0.7813
##  Median :-0.20626   Median :-0.1067   Median :-0.3900
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.07996   3rd Qu.: 0.8789   3rd Qu.: 0.5828
##  Max.   : 4.42620   Max.   : 1.7045   Max.   : 2.7498
##
```

Tratados

```
summary(datos_pca)
```

```
##     ZPOBDENS          ZTMINFAN          ZESPVIDA          ZPOBURB
##  Min.   :-1.07783   Min.   :-1.1026   Min.   :-2.1453   Min.   :-1.7697
##  1st Qu.:-0.84968   1st Qu.:-0.9586   1st Qu.:-0.7460   1st Qu.:-0.7320
##  Median :-0.16160   Median :-0.3931   Median : 0.2781   Median : 0.1268
##  Mean   :-0.08672   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.40244   3rd Qu.: 0.8985   3rd Qu.: 0.9033   3rd Qu.: 0.8148
##  Max.   : 2.15204   Max.   : 1.9048   Max.   : 1.2486   Max.   : 1.5096
##     ZTMEDICO          ZPAGRICU          ZPSERVI           ZTLIBROP
##  Min.   :-1.1473   Min.   :-1.2342   Min.   :-1.88521   Min.   :-0.9696
##  1st Qu.:-0.8829   1st Qu.:-0.8480   1st Qu.:-0.72858   1st Qu.:-0.9145
##  Median :-0.2916   Median :-0.2134   Median : 0.03541   Median :-0.2442
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.8694   3rd Qu.: 0.7115   3rd Qu.: 0.85176   3rd Qu.: 0.6923
##  Max.   : 2.3717   Max.   : 1.9052   Max.   : 1.62885   Max.   : 2.4024
##     ZTEJERCI          ZTPOBACT          ZTENERGI
##  Min.   :-0.86586   Min.   :-2.1341   Min.   :-0.9507
##  1st Qu.:-0.59889   1st Qu.:-0.6735   1st Qu.:-0.7813
##  Median :-0.28969   Median :-0.1067   Median :-0.3900
##  Mean   :-0.28969   Mean   : 0.0000   Mean   :-0.1690
```

```
##  3rd Qu.:-0.04621   3rd Qu.: 0.8789   3rd Qu.: 0.3346
##  Max.   : 0.68708   Max.   : 1.7045   Max.   : 1.5672
```

Análisis exploratorio multivariante

Estudiando los datos

Supuestos de correlación

Comenzamos comprobando si existe correlación entre las variables:

```
cor(datos_pca)
```

```
##             ZPOBDENS   ZTMINFAN   ZESPVIDA     ZPOBURB   ZTMEDICO     ZPAGRICU
## ZPOBDENS  1.00000000 -0.2109635  0.1735014  0.05132529  0.0525974 -0.018865672
## ZTMINFAN -0.21096354  1.0000000 -0.9668834 -0.75749544 -0.7509189  0.752161039
## ZESPVIDA  0.17350137 -0.9668834  1.0000000  0.78714604  0.7361282 -0.753537421
## ZPOBURB   0.05132529 -0.7574954  0.7871460  1.00000000  0.6353621 -0.938022359
## ZTMEDICO  0.05259740 -0.7509189  0.7361282  0.63536205  1.0000000 -0.674513147
## ZPAGRICU -0.01886567  0.7521610 -0.7535374 -0.93802236 -0.6745131  1.000000000
## ZPSERVI  -0.05804281 -0.5901653  0.6121270  0.89001519  0.4446255 -0.907218449
## ZTLIBROP  0.23071577 -0.7157201  0.7004692  0.66106622  0.6093143 -0.666850181
## ZTEJERCI  0.05154113 -0.0254503  0.1017843  0.03566091  0.1513290 -0.007128371
## ZTPOBACT  0.23749261 -0.6032470  0.5411995  0.15488847  0.5336484 -0.147150671
## ZTENERGI  0.17123880 -0.7155678  0.6651297  0.61549363  0.7507638 -0.688730826
##              ZPSERVI   ZTLIBROP     ZTEJERCI    ZTPOBACT   ZTENERGI
## ZPOBDENS -0.05804281  0.2307158  0.051541133  0.23749261  0.1712388
## ZTMINFAN -0.59016525 -0.7157201 -0.025450301 -0.60324698 -0.7155678
## ZESPVIDA  0.61212697  0.7004692  0.101784303  0.54119951  0.6651297
## ZPOBURB   0.89001519  0.6610662  0.035660910  0.15488847  0.6154936
## ZTMEDICO  0.44462549  0.6093143  0.151329048  0.53364837  0.7507638
## ZPAGRICU -0.90721845 -0.6668502 -0.007128371 -0.14715067 -0.6887308
## ZPSERVI   1.00000000  0.5044833 -0.173017490 -0.05616141  0.4159030
## ZTLIBROP  0.50448334  1.0000000  0.113918652  0.41591469  0.6602394
## ZTEJERCI -0.17301749  0.1139187  1.000000000 -0.07216152  0.1076248
## ZTPOBACT -0.05616141  0.4159147 -0.072161520  1.00000000  0.6004197
## ZTENERGI  0.41590296  0.6602394  0.107624821  0.60041968  1.0000000
```

El contraste de esfericidad de Bartlett permite comprobar si las correlaciones son distintas de 0 de modo significativo. La hipótesis nula es que det(R)=1. La función "cortest.bartlett" del paquete "pysch" realiza este test. Esta función trabaja con datos normalizados.

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     outlier
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
# Se normalizan los datos
datos_normalizados<-scale(datos_pca)
```

```
# Se hace el test de esfericidad
cortest.bartlett(cor(datos_normalizados))
```

```
## Warning in cortest.bartlett(cor(datos_normalizados)): n not specified, 100 used
```

```
## $chisq
## [1] 1285.077
##
## $p.value
## [1] 1.141378e-232
##
## $df
## [1] 55
```

Para estos datos se obtiene un test significativo de modo que se rechaza la hipótesis nula y por tanto los datos no están incorrelados. Aún así, seguiremos adelante y si algo no funciona puede deberse a esto.

Análisis exploratorio de los datos

Como ya hemos imputado los valores perdidos y eliminado los outliers no hace falta repetir el procedimiento.

Estudio de posibilidad de reducción de la dimensión

Vamos a realizar un análisis de componentes principales (PCA). La función "prcomp" del paquete base de R realiza este análisis.

Pasamos los parámetros "scale" y "center" a TRUE para considerar los datos originales normalizados. Además, el campo "rotation" del objeto PCA es una matriz cuyas columnas son los coeficientes de las componentes principales. Finalmente, en el campo "sdev" del mismo objeto obtenemos la información sobre desviaciones típicas de cada componente principal, proporción de varianza explicada y acumulada.

```
PCA<-prcomp(datos_pca, scale=T, center = T)
PCA$rotation
```

```
##                  PC1         PC2         PC3          PC4         PC5
## ZPOBDENS  0.06796131  0.39036079  0.07578571 -0.873056137  0.05981546
## ZTMINFAN -0.37529798 -0.11573027 -0.08212636 -0.014741875  0.41788865
## ZESPVIDA  0.37165513  0.07270402 -0.00407411  0.012970861 -0.53281544
## ZPOBURB   0.35808785 -0.29124050 -0.06094609 -0.113795410 -0.05152603
## ZTMEDICO  0.33374417  0.13909259 -0.09587119  0.280930962  0.15782496
## ZPAGRICU -0.36224935  0.29940891  0.04348437  0.061235615 -0.15998056
## ZPSERVI   0.29607632 -0.48878963  0.09183567 -0.148420697 -0.03064997
## ZTLIBROP  0.32594252  0.08892433 -0.07078859 -0.135187927  0.28062939
## ZTEJERCI  0.02397325  0.17948476 -0.93319975 -0.006241743 -0.12518016
## ZTPOBACT  0.20289865  0.56167354  0.29659750  0.281932179 -0.11796713
## ZTENERGI  0.33155119  0.20152127 -0.02054058  0.148110844  0.61275026
##                  PC6         PC7         PC8          PC9         PC10
## ZPOBDENS -0.24337351  0.09619809 -0.04353317  0.022846323 -0.01646372
## ZTMINFAN -0.01756878  0.05245475 -0.39097670 -0.157260993 -0.47091956
## ZESPVIDA  0.03028763 -0.04105430  0.29289664 -0.091799724 -0.54354628
## ZPOBURB  -0.07065324 -0.10651593 -0.29868358 -0.792450091  0.16228759
## ZTMEDICO -0.36060911  0.78403566 -0.06560314  0.001929436 -0.06125296
## ZPAGRICU  0.13248858  0.09343760 -0.01200167 -0.325592955 -0.48343682
## ZPSERVI  -0.07352269 -0.09280074 -0.37422057  0.449245272 -0.39418278
## ZTLIBROP  0.85707761  0.21913126 -0.01229527  0.015266047 -0.02755826
```

```
## ZTEJERCI -0.02870986 -0.16141551 -0.18218984  0.127931321  0.01085377
## ZTPOBACT  0.02948798 -0.25103160 -0.61023870  0.097328148  0.08409493
## ZTENERGI -0.21278600 -0.45367225  0.34643517 -0.066593066 -0.23535405
##                  PC11
## ZPOBDENS -0.016686681
## ZTMINFAN -0.511610948
## ZESPVIDA -0.424556791
## ZPOBURB   0.088144761
## ZTMEDICO  0.058614024
## ZPAGRICU  0.617855535
## ZPSERVI   0.361609762
## ZTLIBROP  0.008370356
## ZTEJERCI  0.063899183
## ZTPOBACT -0.095710436
## ZTENERGI  0.143880098
```

```
PCA$sdev
```

```
##  [1] 2.4818030 1.2805818 1.0311152 0.9515420 0.6448055 0.6027181 0.4818699
##  [8] 0.3354192 0.2491626 0.1641974 0.1390789
```

```
summary(PCA)
```

```
## Importance of components:
##                            PC1    PC2     PC3     PC4    PC5     PC6     PC7
## Standard deviation     2.4818 1.2806 1.03112 0.95154 0.6448 0.60272 0.48187
## Proportion of Variance 0.5599 0.1491 0.09665 0.08231 0.0378 0.03302 0.02111
## Cumulative Proportion  0.5599 0.7090 0.80568 0.88799 0.9258 0.95881 0.97992
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.33542 0.24916 0.16420 0.13908
## Proportion of Variance 0.01023 0.00564 0.00245 0.00176
## Cumulative Proportion  0.99015 0.99579 0.99824 1.00000
```

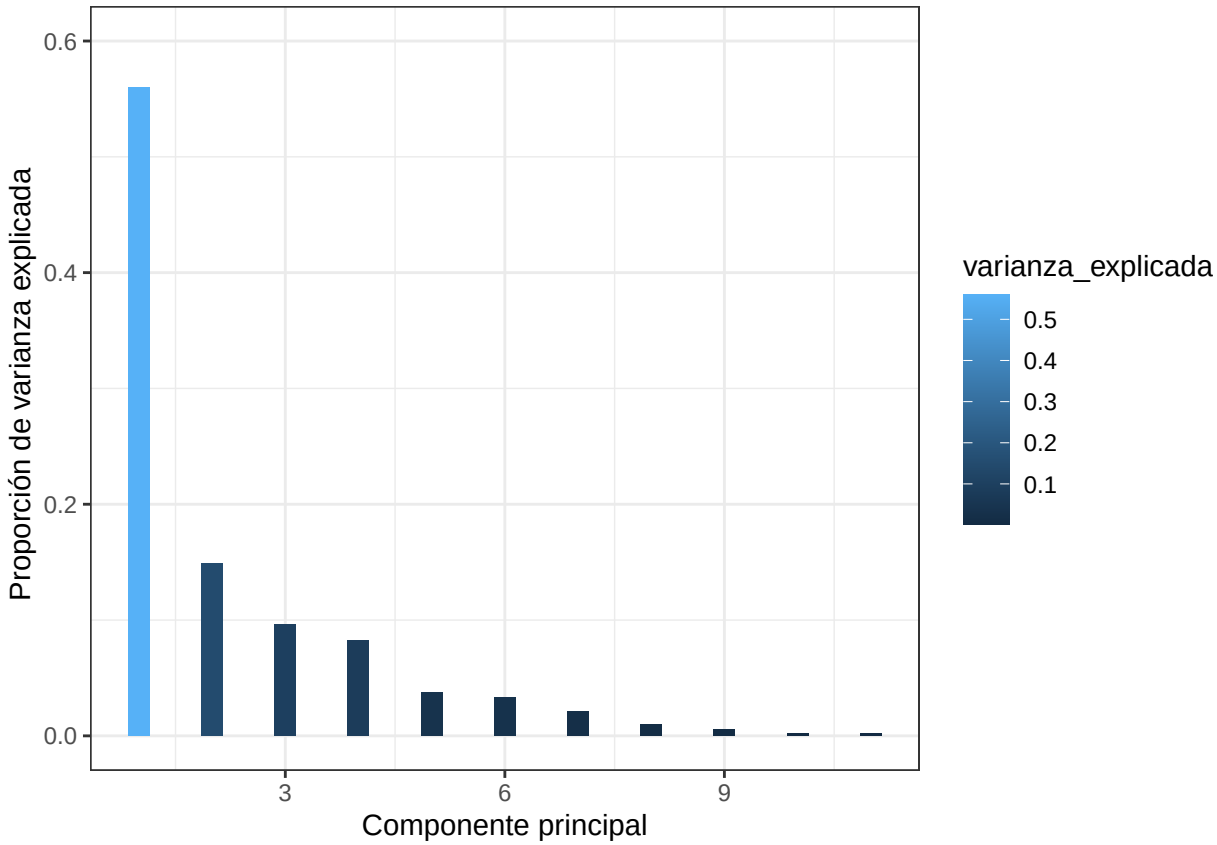
Representamos ahora la proporción de varianza explicada y acumulada:

```
library("ggplot2")
```

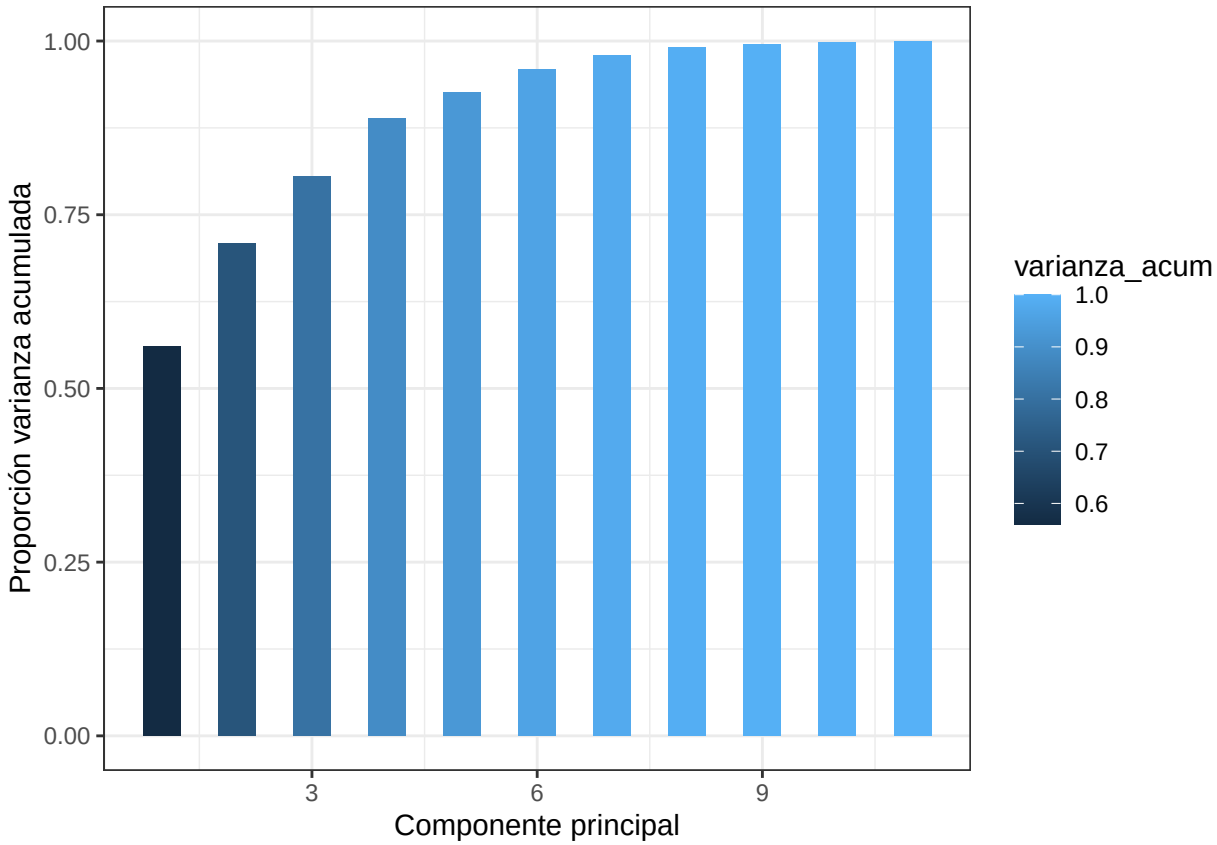
```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
# El siguiente gráfico muestra la proporción de varianza explicada
varianza_explicada <- PCA$sdev^2 / sum(PCA$sdev^2)
ggplot(data = data.frame(varianza_explicada, pc = 1:11),
       aes(x = pc, y = varianza_explicada, fill=varianza_explicada )) +
  geom_col(width = 0.3) +
  scale_y_continuous(limits = c(0,0.6)) + theme_bw() +
  labs(x = "Componente principal", y= " Proporción de varianza explicada")
```




```
# El siguiente gráfico muestra la proporción de varianza explicada
varianza_acum<-cumsum(varianza_explicada)
ggplot( data = data.frame(varianza_acum, pc = 1:11),
        aes(x = pc, y = varianza_acum ,fill=varianza_acum )) +
  geom_col(width = 0.5) +
  scale_y_continuous(limits = c(0,1)) +
  theme_bw() +
  labs(x = "Componente principal",
       y = "Proporción varianza acumulada")
```



Buscamos ahora el número de componentes principales óptimo. En este caso vamos a utilizar la regla de Abdi et al. (2010). Se promedian las varianzas explicadas por la componentes principales y se seleccionan aquellas cuya proporción de varianza explicada supera la media. # En este caso se eligen tan solo cuatro direcciones principales tal y como se puede ver # que acumulan casi un 80% de varianza explicada

```
PCA$sdev^2
```

```
##  [1] 6.15934602 1.63988972 1.06319852 0.90543213 0.41577420 0.36326910
##  [7] 0.23219856 0.11250601 0.06208202 0.02696079 0.01934294
```

```
mean(PCA$sdev^2)
```

```
## [1] 1
```

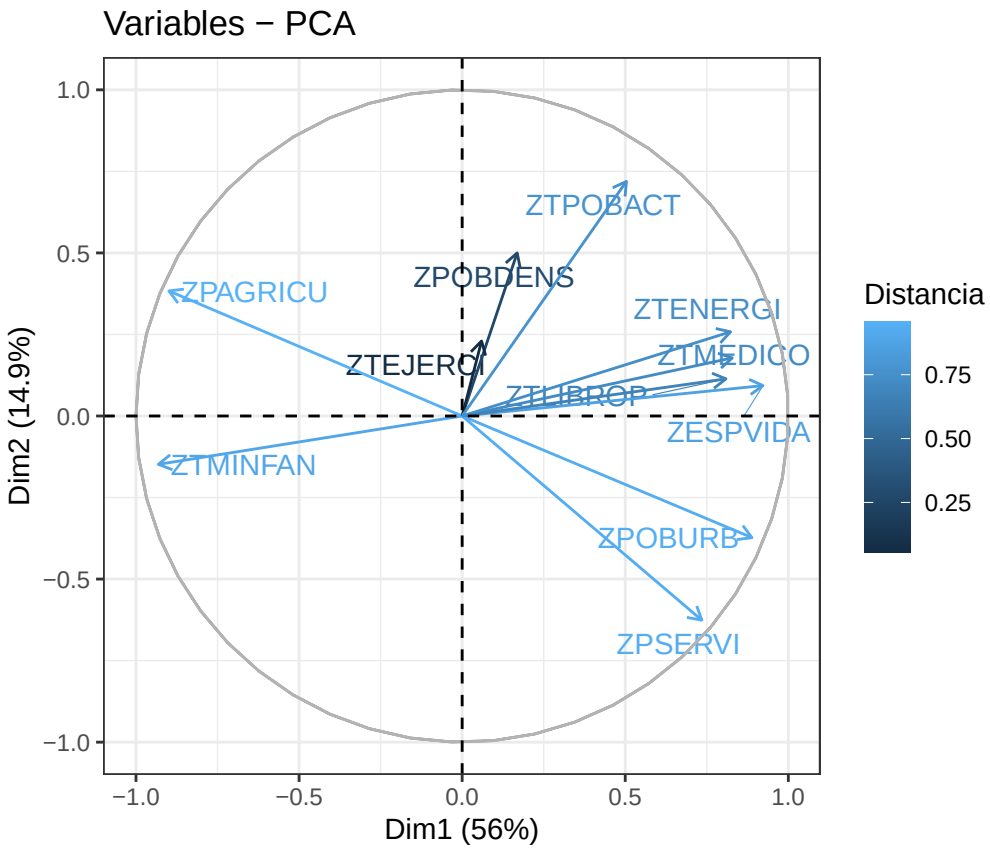
En este caso la media es 1, luego se escogerían la primera, segunda y tercera. Esto son, 3 componentes principales.

Ahora vamos a representar las componentes principales, realizando una comparativa entre la primera y segunda componente principal analizando qué variables tienen más peso para la definición de cada componente principal, después entre la primera y la tercera y finalmente entre la segunda y tercera.

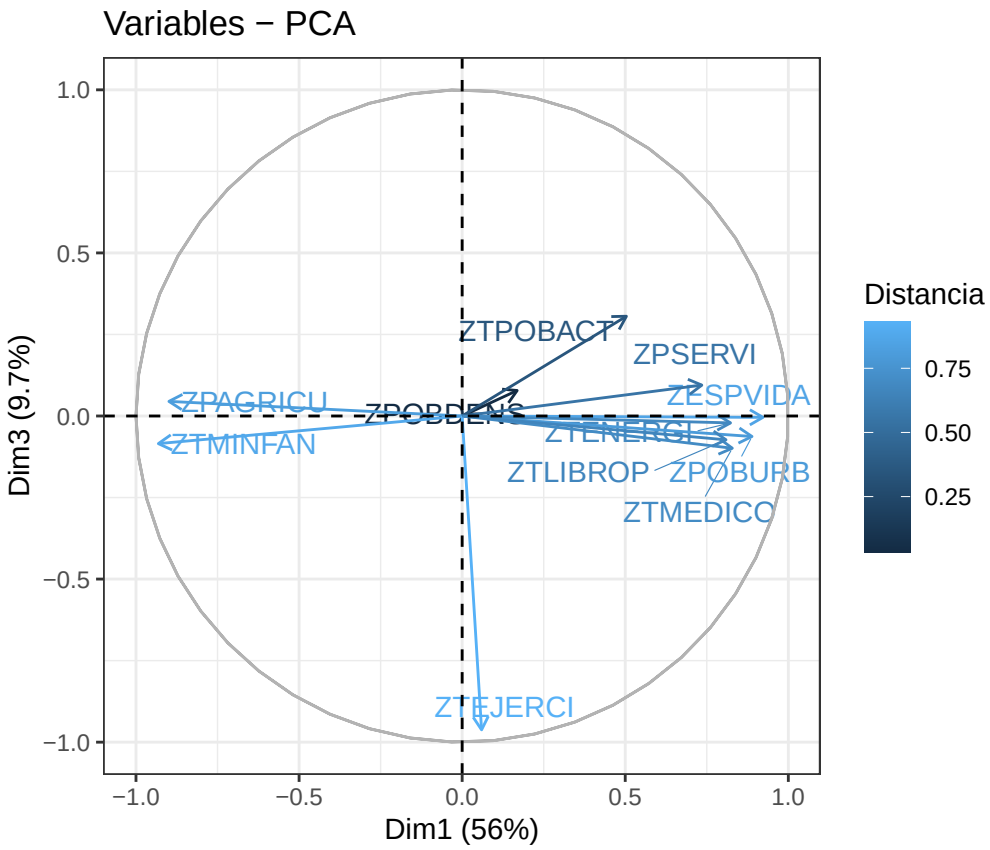
```
library("factoextra")
```

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

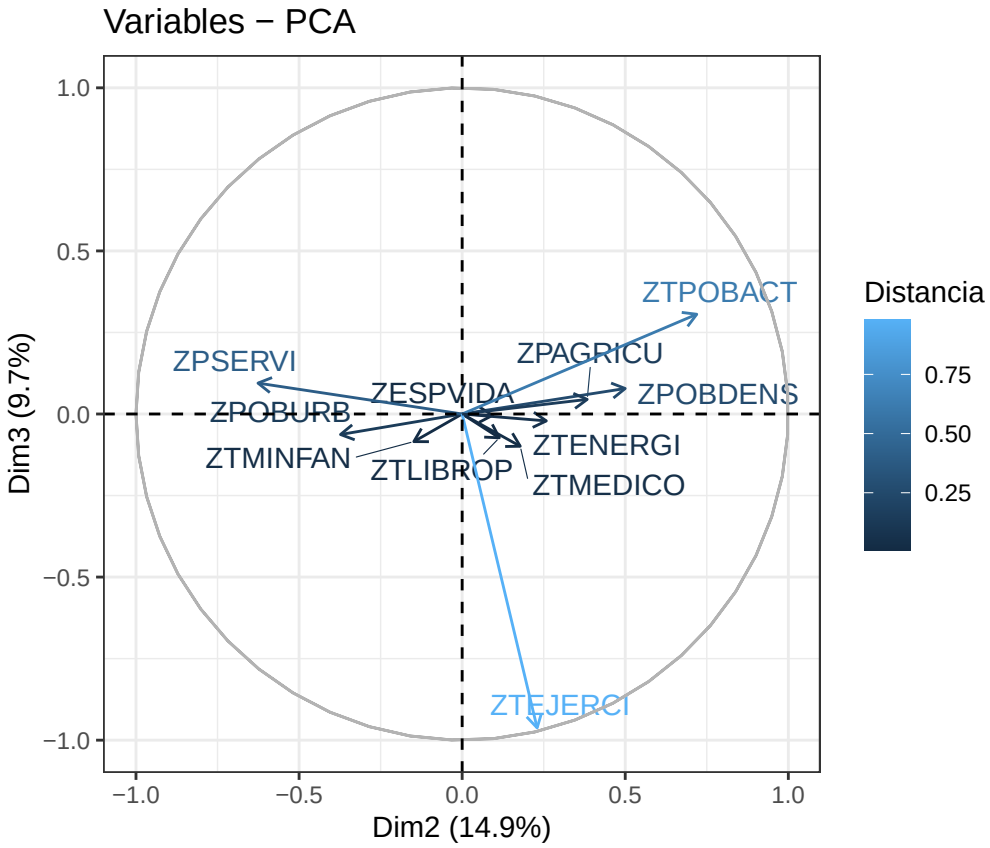
```
fviz_pca_var(PCA,
             repel=TRUE,col.var="cos2",
             legend.title="Distancia")+theme_bw()
```




```
fviz_pca_var(PCA,axes=c(1,3),
             repel=TRUE,col.var="cos2",
             legend.title="Distancia")+theme_bw()
```

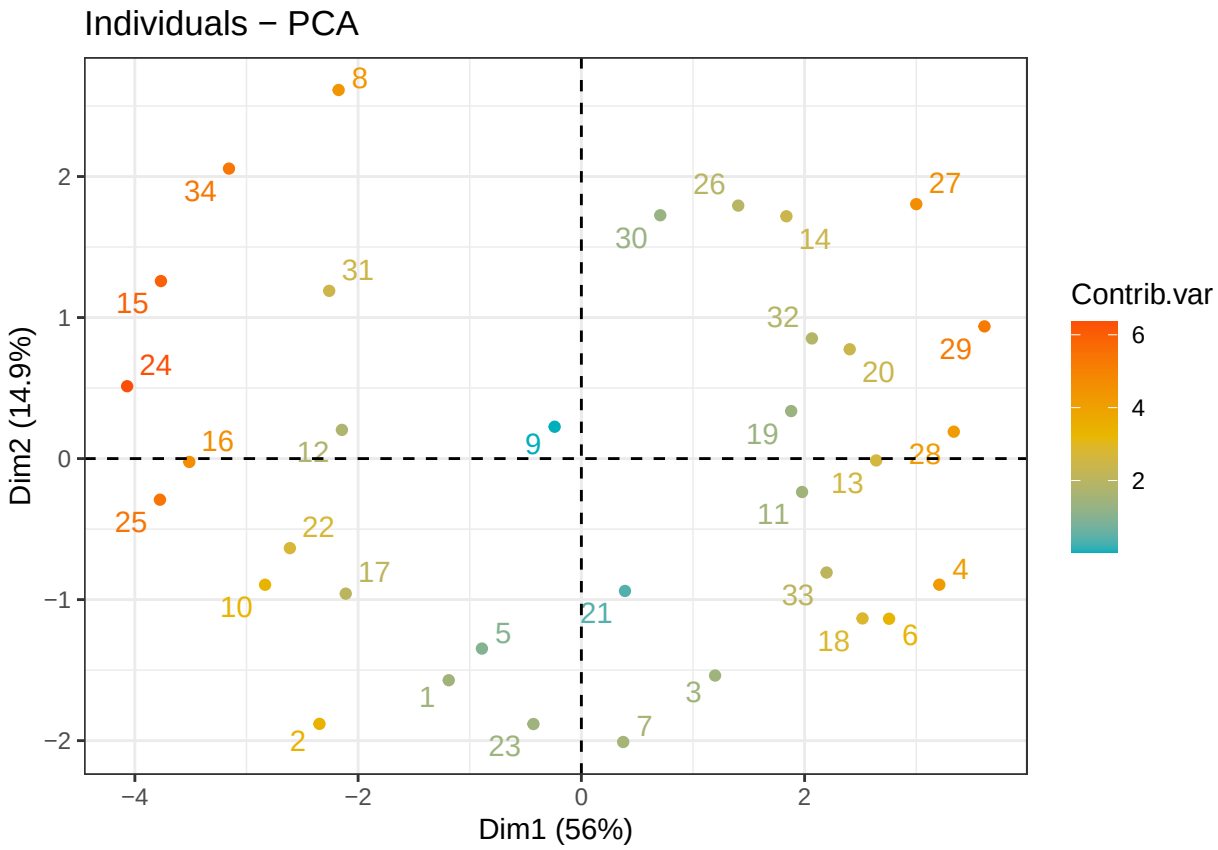



```
fviz_pca_var(PCA,axes=c(2,3),
             repel=TRUE,col.var="cos2",
             legend.title="Distancia")+theme_bw()
```

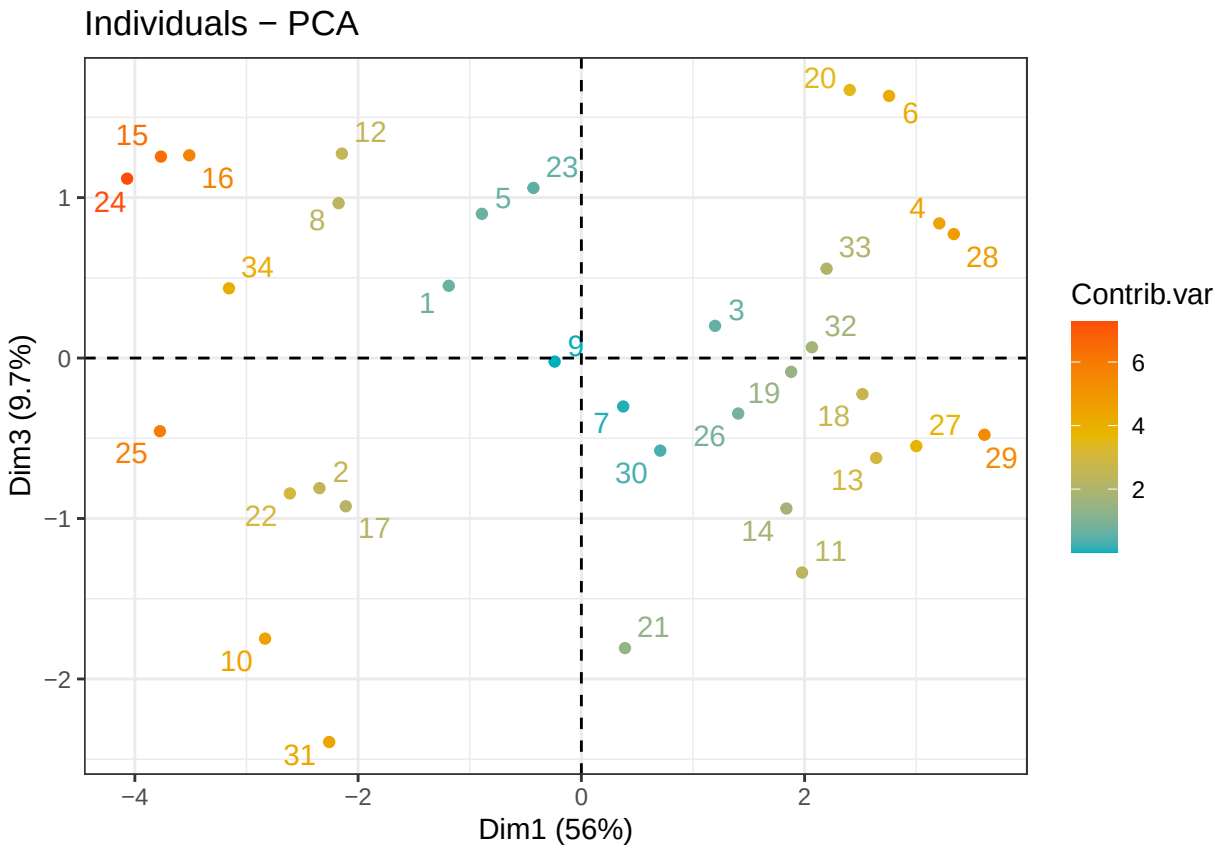


Es posible también representar las observaciones de los objetos junto con las componentes principales mediante la orden "contrib" de la función "fviz_pca_ind", así como identificar con colores aquellas observaciones que mayor varianza explican de las componentes principales.

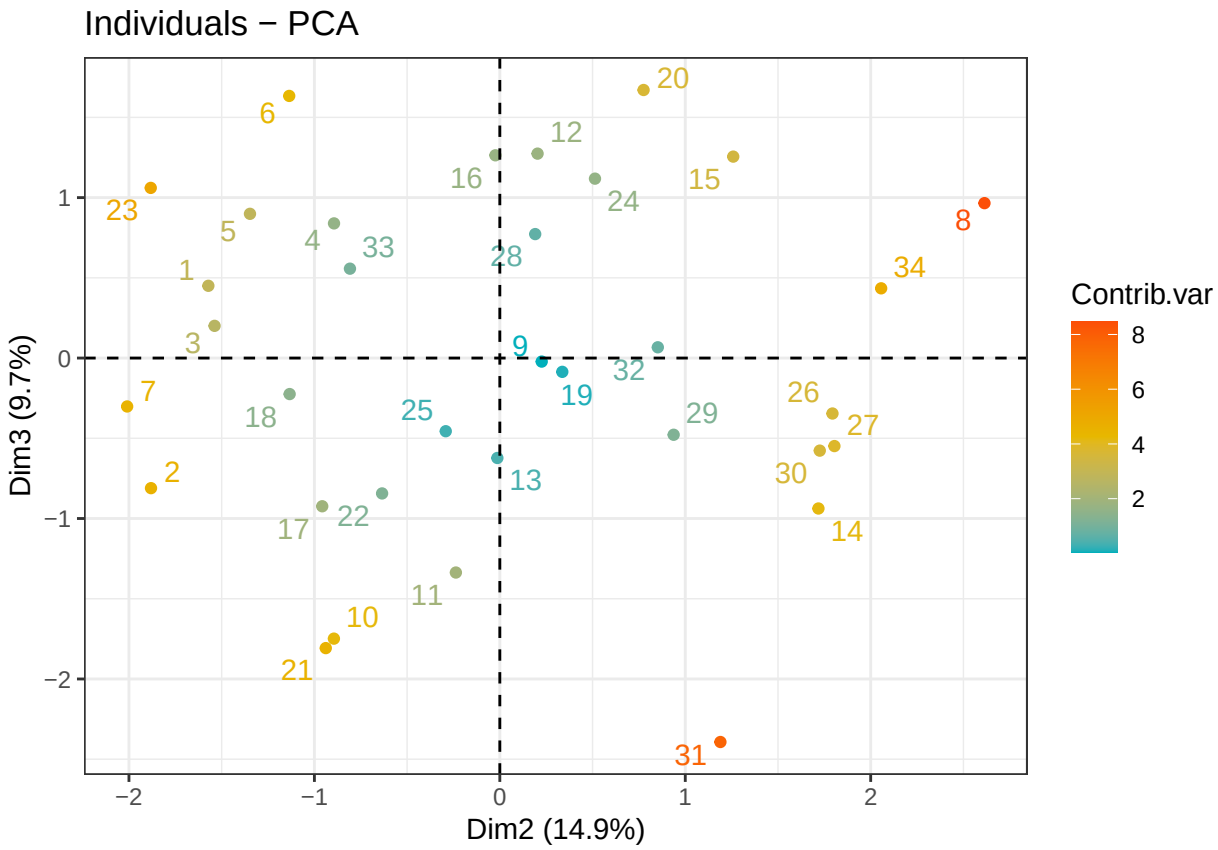
```
# Observaciones en la primera y segunda componente principal
fviz_pca_ind(PCA,col.ind = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel=TRUE,legend.title="Contrib.var")+theme_bw()
```




```
# Observaciones en la primera y tercera componente principal
fviz_pca_ind(PCA,axes=c(1,3),col.ind = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel=TRUE,legend.title="Contrib.var")+theme_bw()
```

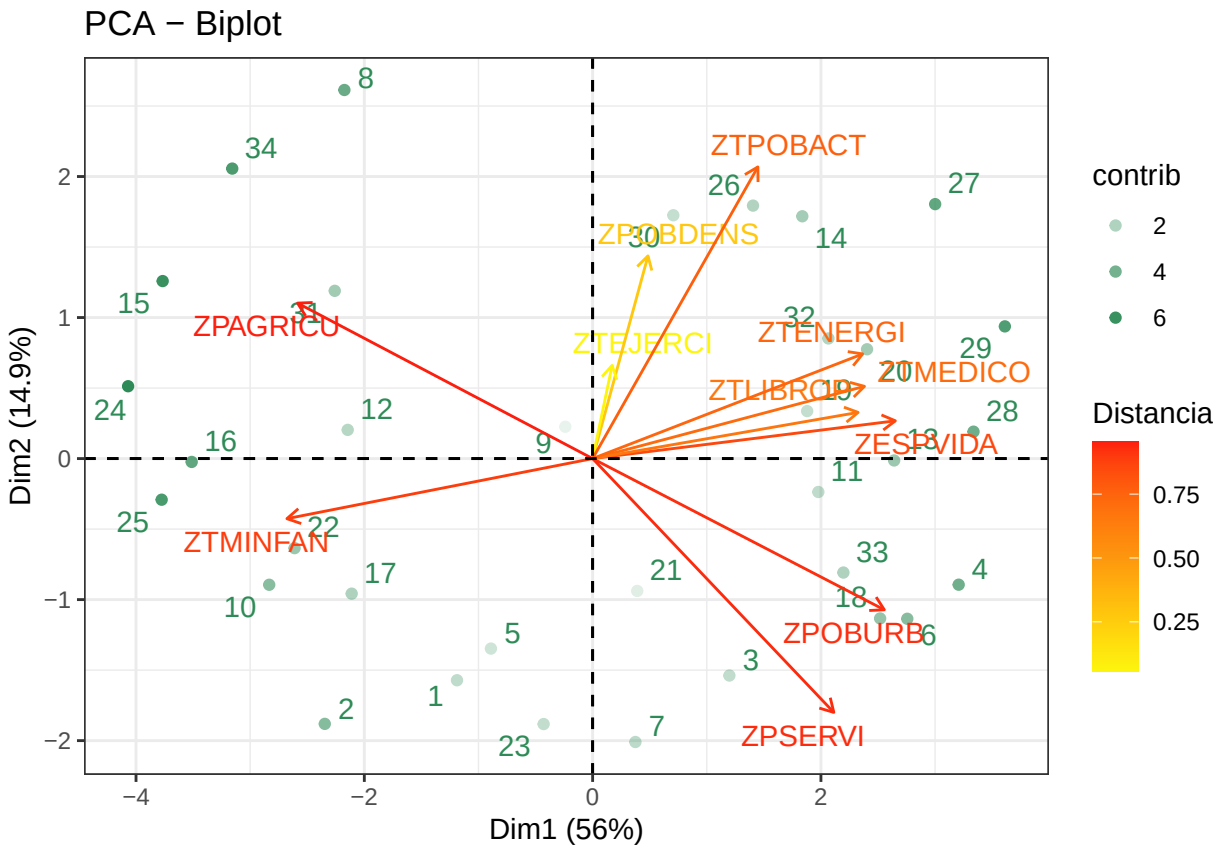



```
# Observaciones en la segunda y tercera componente principal
fviz_pca_ind(PCA,axes=c(2,3),col.ind = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel=TRUE,legend.title="Contrib.var")+theme_bw()
```

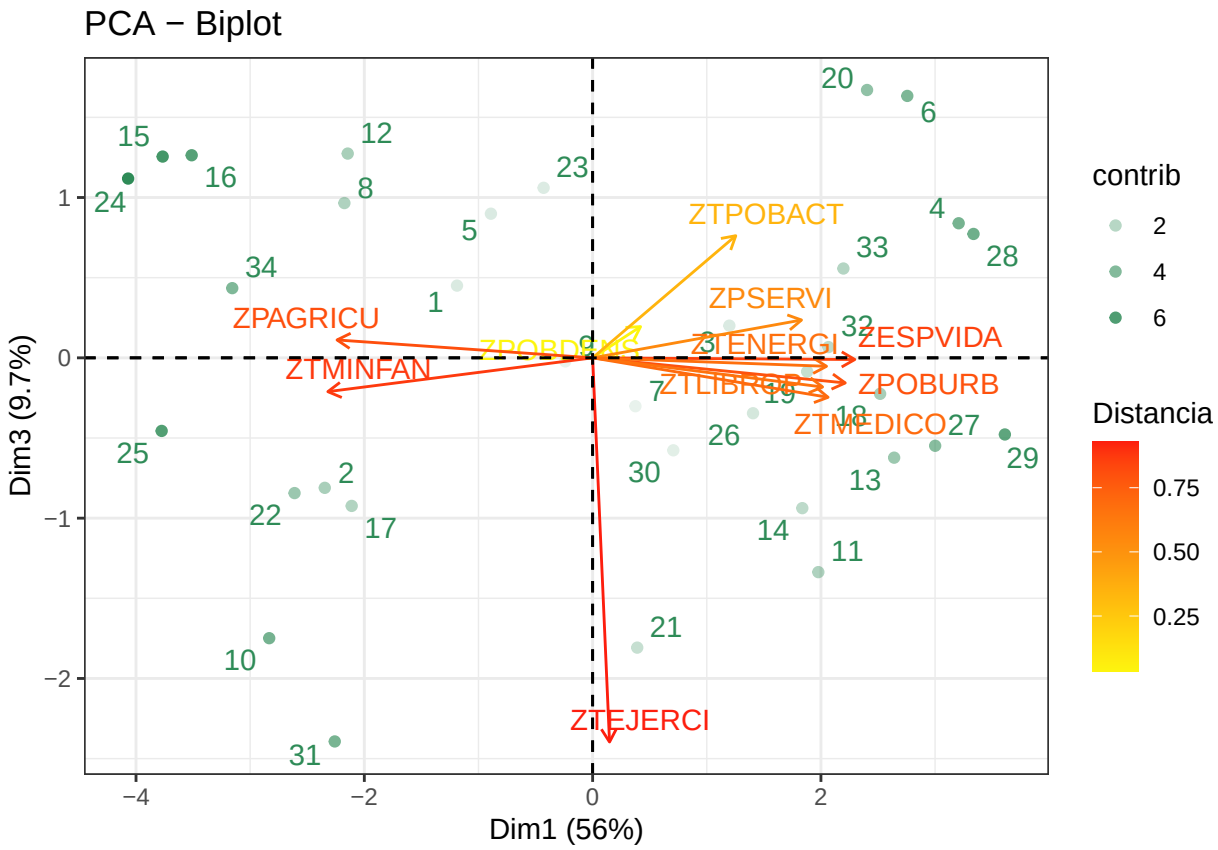


Además haremos una representación conjunta de variables y observaciones que relaciona visualmente las posibles relaciones entre las observaciones, las contribuciones de los individuos a las varianzas de las componentes y el peso de las variables en cada componentes principal.

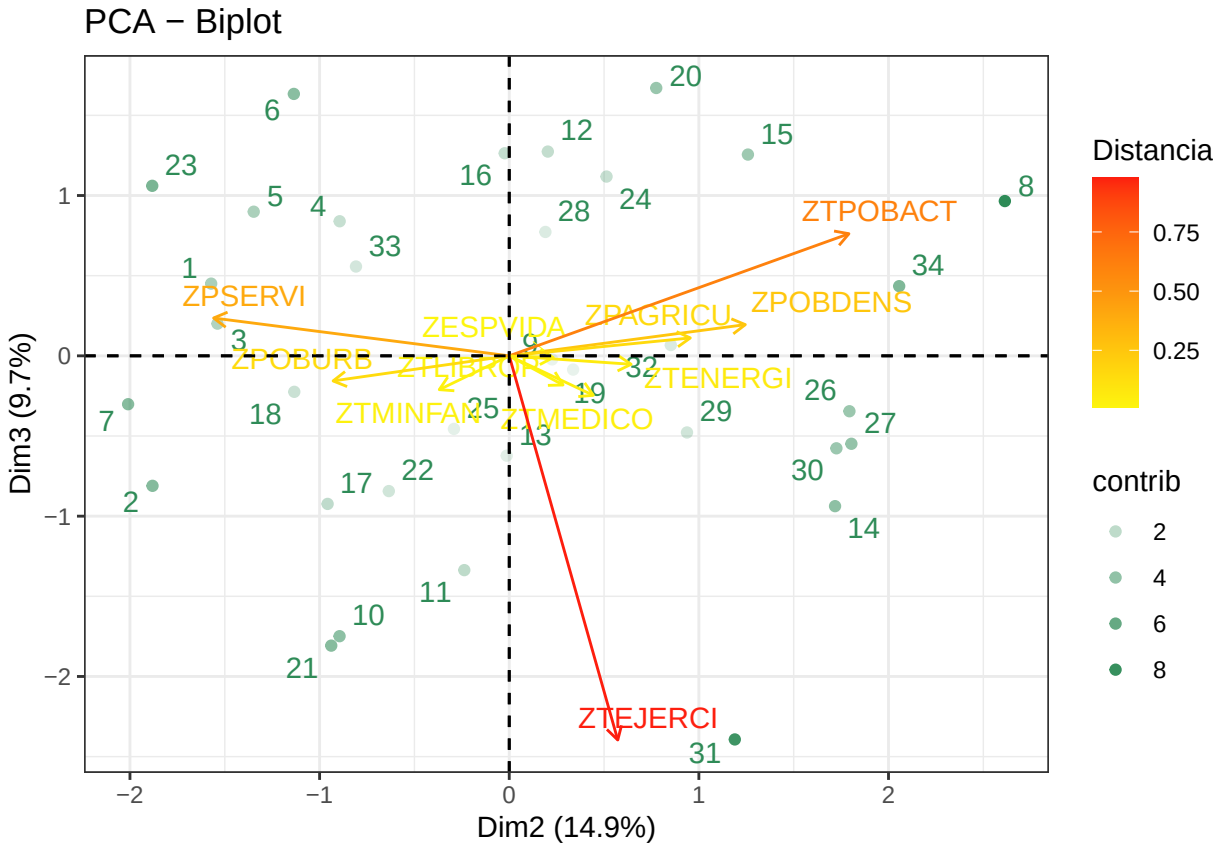
```
# Variables y observaciones en las 1?  y 2? componente principal
fviz_pca(PCA,
         alpha.ind ="contrib", col.var = "cos2",col.ind="seagreen",
         gradient.cols = c("#FDF50E", "#FD960E", "#FD1E0E"),
         repel=TRUE,
         legend.title="Distancia")+theme_bw()
```




```
# Variables y observaciones en las 1?  y 3? componente principal
fviz_pca(PCA,axes=c(1,3),
         alpha.ind ="contrib", col.var = "cos2",col.ind="seagreen",
         gradient.cols = c("#FDF50E", "#FD960E", "#FD1E0E"),
         repel=TRUE,
         legend.title="Distancia")+theme_bw()
```




```
# Variables y observaciones en las 2  y 3? componente principal
fviz_pca(PCA,axes=c(2,3),
         alpha.ind ="contrib", col.var = "cos2",col.ind="seagreen",
         gradient.cols = c("#FDF50E", "#FD960E", "#FD1E0E"),
         repel=TRUE,
         legend.title="Distancia")+theme_bw()
```



Por último, ya que el objeto de este estudio es reducir la dimensión de las variables utilizadas, es posible obtener las coordenadas de los datos originales tipificados en el nuevo sistema de referencia. De hecho lo tenemos almacenado desde que utilizamos la función prcomp para crear la variable PCA

```
head(PCA$x,n=3)
```

```
##            PC1       PC2        PC3       PC4        PC5        PC6        PC7
## [1,] -1.187951 -1.571749  0.4504449 0.3781493  1.4228328 -0.2719916 -0.6235956
## [2,] -2.346246 -1.881066 -0.8108192 0.2014503  0.4528028 -0.2176741 -0.1130183
## [3,]  1.197924 -1.538337  0.2007784 0.8972738 -0.3885080 -0.7714447  0.8072288
##              PC8        PC9       PC10        PC11
## [1,] -0.09144981  0.3475349  0.3629623  0.0009993838
## [2,]  0.67613362  0.3189091 -0.2221354 -0.1452388044
## [3,] -0.15129705 -0.4099810  0.1343895 -0.0562911808
```

Reducción de dimensión mediante variables latentes

Como ya hemos comprobado las correlaciones y test de esfericidad anteriormente, nos saltaremos ese paso.

Vamos a comparar las salidas con el método del factor principal y con el de máxima verosimilitud.

```
library("polycor")
```

```
##
## Attaching package: 'polycor'
```

```
## The following object is masked from 'package:psych':
##
##     polyserial
```

```
poly_cor<-hetcor(datos_pca)$correlations

modelo1<-fa(poly_cor,
            nfactors = 3,
            rotate = "none",
            fm="mle") # modelo máxima verosimilitud

modelo2<-fa(poly_cor,
            nfactors = 3,
            rotate = "none",
            fm="minres") # modelo minimo residuo
```

Comenzamos comparando las comunalidades:

```
# Comparacion comunalidades
sort(modelo1$communality,decreasing = T)->c1
sort(modelo2$communality,decreasing = T)->c2
head(cbind(c1,c2))
```

```
##                 c1        c2
## ZPAGRICU 0.9896762 0.9965633
## ZTMINFAN 0.9768934 0.9635737
## ZESPVIDA 0.9632209 0.9521015
## ZTENERGI 0.9577687 0.9267156
## ZPSERVI  0.9286045 0.8955316
## ZPOBURB  0.9120204 0.8448246
```

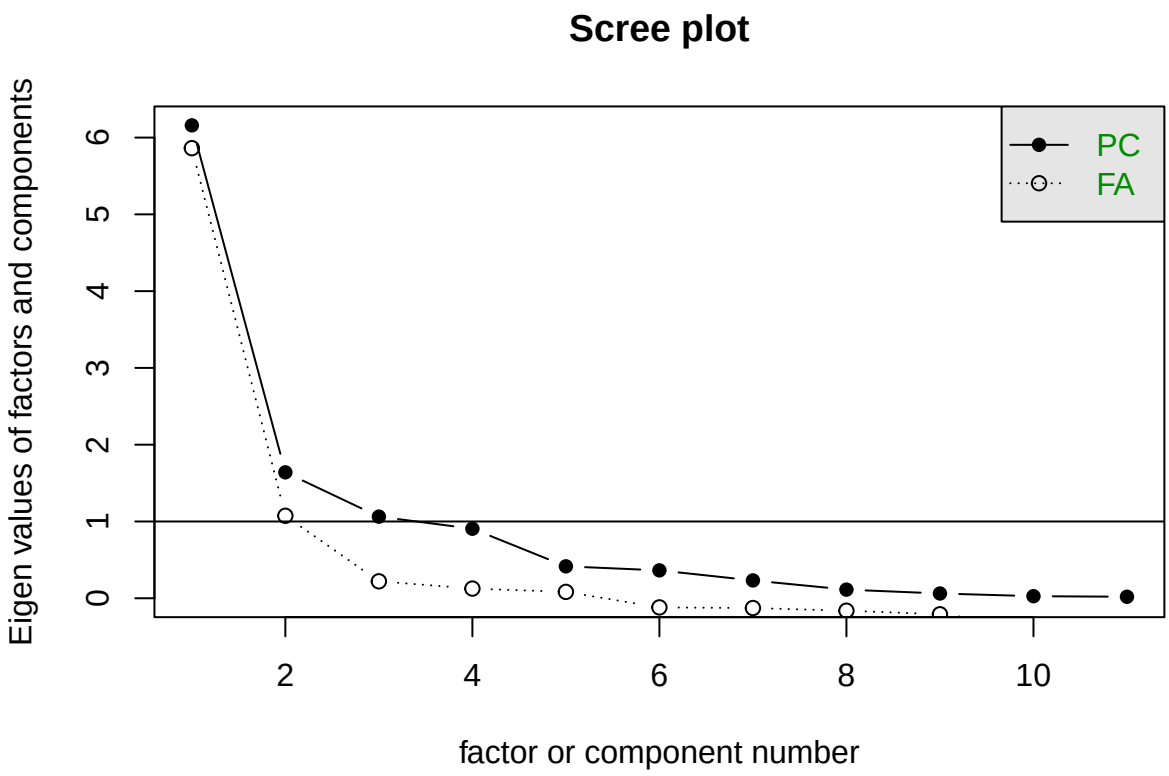
Ahora comparamos las unicidades, es decir la proporción de varianza que no ha sido explicada por el factor (1-comunalidad)

```
sort(modelo1$uniquenesses,decreasing = T)->u1
sort(modelo2$uniquenesses,decreasing = T)->u2
head(cbind(u1,u2))
```

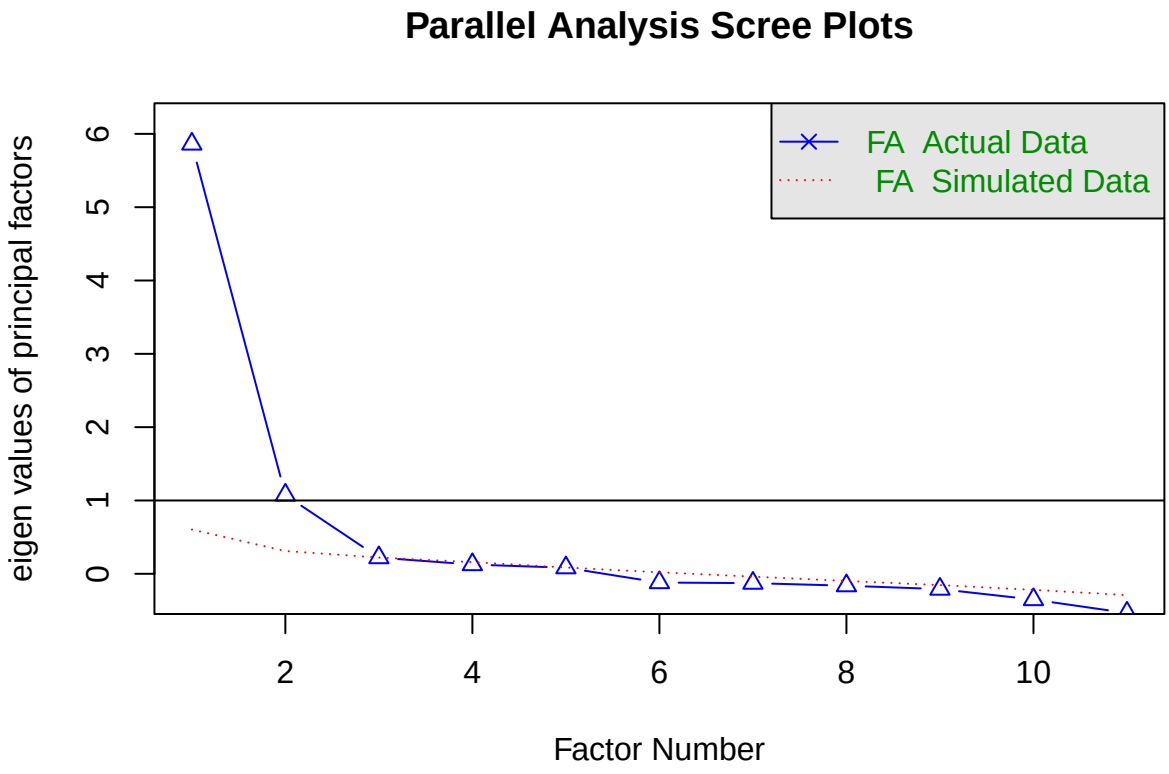
```
##                  u1        u2
## ZTEJERCI 0.97398011 0.9233303
## ZPOBDENS 0.90544264 0.5051294
## ZTLIBROP 0.41684425 0.3942481
## ZTMEDICO 0.31611448 0.3174441
## ZTPOBACT 0.23640549 0.3106714
## ZPOBURB  0.08797956 0.1551754
```

Determinemos ahora el número óptimo de factores. Hay diferentes criterios, entre los que destacan el Scree plot (Cattel 1966) y el análisis paralelo (Horn 1965).

```
scree(poly_cor)
```




```
fa.parallel(poly_cor,n.obs=200,fa="fa",fm="minres")
```




```
## Parallel analysis suggests that the number of factors =  2  and the number of components =  NA
```

Se deduce que el número óptimo de factores es 3, como ya vimos en el apartado anterior.

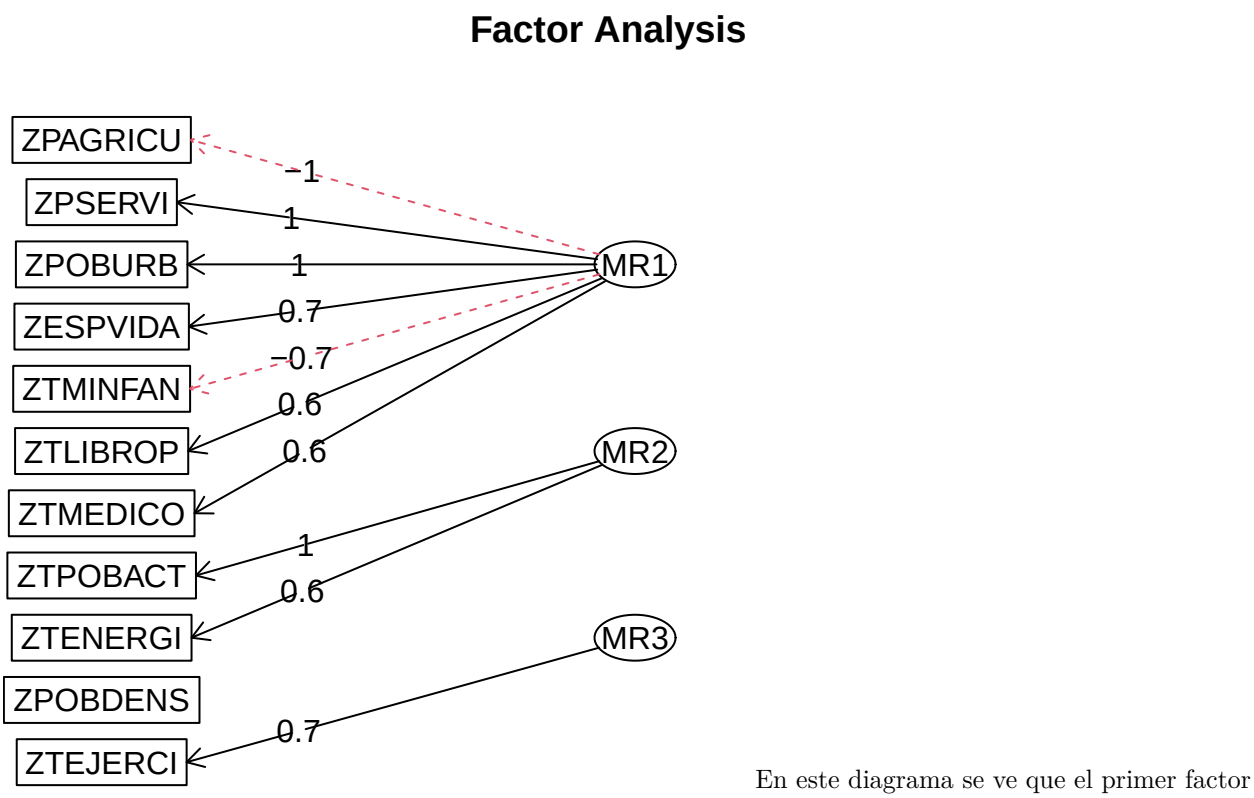
Estimamos el modelo factorial con 3 factores implementando una rotacion tipo varimax para buscar una interpretación más simple y mostramos la matriz de pesos factorial rotada.

```
modelo_varimax<-fa(poly_cor,nfactors = 3,rotate = "varimax",
                   fa="mle")
print(modelo_varimax$loadings,cut=0)
```

```
##
## Loadings:
##          MR1    MR2    MR3
## ZPOBDENS  0.013  0.270  0.060
## ZTMINFAN -0.725 -0.608 -0.009
## ZESPVIDA  0.742  0.535  0.084
## ZPOBURB   0.955  0.112  0.041
## ZTMEDICO  0.607  0.531  0.179
## ZPAGRICU -0.975 -0.110 -0.023
## ZPSERVI   0.956 -0.159 -0.242
## ZTLIBROP  0.628  0.436  0.145
## ZTEJERCI  0.000  0.039  0.702
## ZTPOBACT  0.055  0.962 -0.153
## ZTENERGI  0.575  0.584  0.135
##
##                  MR1   MR2   MR3
## SS loadings    4.951 2.518 0.659
## Proportion Var 0.450 0.229 0.060
## Cumulative Var 0.450 0.679 0.739
```

Para verlo más claro vamos a representarlo:

```
fa.diagram(modelo_varimax)
```



En este diagrama se ve que el primer factor

está asociado con ZPSERVI, ZPAGRICU, ZPOBURB, ZESPVIDA, ZTMINFAN, ZTLIBROP y ZTMEDICO. mientras la segunda es con ZTPOBACT y ZTENERGI y la última con ZTEJERCI.

Finalmente comprobamos que en efecto el número de factores elegido es suficiente.

```
library(stats)
factanal(datos_pca,factors=3, rotation="none")
```

```
##
## Call:
## factanal(x = datos_pca, factors = 3, rotation = "none")
##
## Uniquenesses:
## ZPOBDENS ZTMINFAN ZESPVIDA  ZPOBURB ZTMEDICO ZPAGRICU  ZPSERVI ZTLIBROP
##    0.905    0.023    0.037    0.088    0.316    0.010    0.071    0.417
## ZTEJERCI ZTPOBACT ZTENERGI
##    0.974    0.236    0.042
##
## Loadings:
##          Factor1 Factor2 Factor3
## ZPOBDENS          0.291
## ZTMINFAN -0.893  -0.395   0.154
## ZESPVIDA  0.888   0.354  -0.221
## ZPOBURB   0.930  -0.199
## ZTMEDICO  0.760   0.269   0.185
## ZPAGRICU -0.961   0.255
## ZPSERVI   0.831  -0.455  -0.178
## ZTLIBROP  0.738   0.180
## ZTEJERCI          0.119   0.105
## ZTPOBACT  0.357   0.775   0.186
## ZTENERGI  0.770   0.281   0.535
##
##                Factor1 Factor2 Factor3
## SS loadings      5.915   1.477   0.487
## Proportion Var   0.538   0.134   0.044
## Cumulative Var   0.538   0.672   0.716
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 26.21 on 25 degrees of freedom.
## The p-value is 0.397
```

Con ese p-valor para cierto nivel de confianza se acepta la hipótesis y por tanto el número de factores es suficiente.

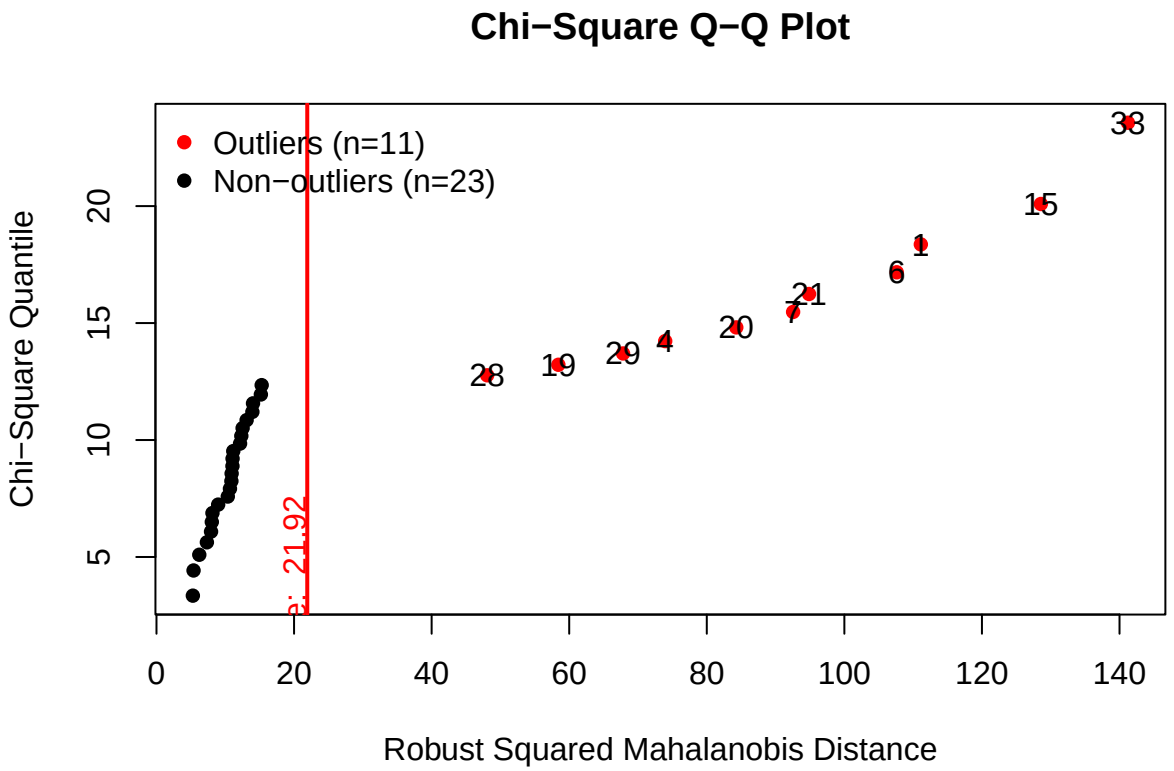
Análisis de la normalidad multivariante

A continuación hacemos una exploración gráfica de la normalidad de las distribuciones individuales de nuestros predictores representando los histogramas y los gráficos qqplot.

Como en el apartado univariante ya hemos comprobado la normalidad de cada una de las variables no vamos a repetirlo.

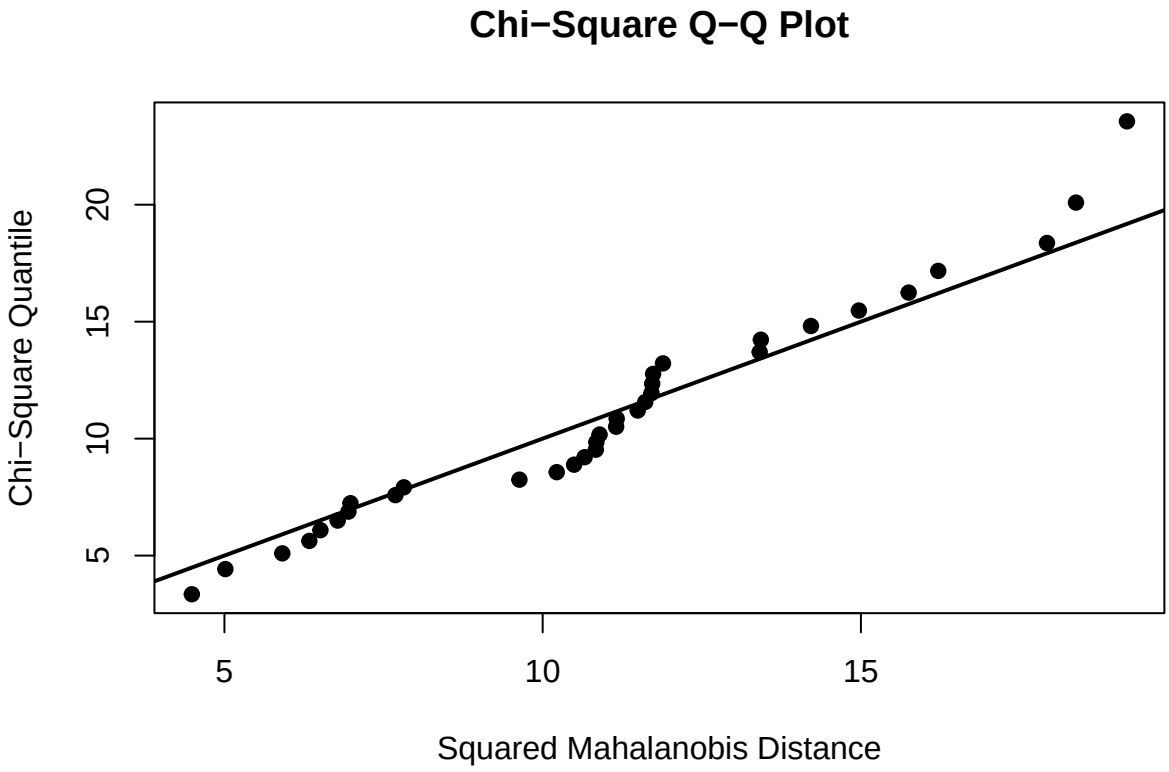
El paquete MVN contiene funciones que permiten realizar los tres test que se utilizan habitualmente para contrastar la normalidad multivariante. Esta normalidad multivariante puede verse afectada por la presencia de outliers. En este paquete también encontramos funciones para el análisis de outliers.

```
library(MVN)
outliers <- mvn(data = datos_pca, mvnTest = "hz", multivariateOutlierMethod = "quan")
```



Se detectan 11 outliers en las observaciones. Aunque considerar outliers en 11 observaciones de 33 puede deberse a la falta de datos, y más cuando los outliers ya se eliminaron varios en el análisis univariante. Sin embargo solo de los dos test realizados a continuación encuentran evidencias al 5% de significación de falta de normalidad multivariante.

```
royston_test <- mvn(data = datos_pca, mvnTest = "royston", multivariatePlot = "qq")
```




```
royston_test$multivariateNormality
```

```
##      Test        H      p value MVN
## 1 Royston 24.07257 7.374655e-05  NO
```

```
hz_test <- mvn(data = datos_pca, mvnTest = "hz")
hz_test$multivariateNormality
```

```
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.9888665 0.0761232 YES
```

Pese a que uno de los test ha rechazado que siga una normal multivariante, vamos a hacerle caso al otro test y seguir, ya que probablemente esto se deba al bajo tamaño de observaciones con el que contamos y en el gráfico los valores no se alejan demasiado de la recta.

Clasificación

Como acabamos de comprobar la noramlidad multivariante y ya habíamos comprobado la univariante, solo falta la homogeneidad de la varianza.

Homogeneidad de la varianza

Usaremos el test de Box M, que es una extensión del de Barttlet para escenarios multivariantes. Hay que tener en cuenta que es muy sensible a que los datos efectivamente se distribuyan según una normal multivariante. Por este motivo se recomienda utilizar una significación (p-value) <0.001.

La hipóstesis nula a contrastar es la de igualdad de matrices de covarianzas en todos los grupos.

```
library(biotools)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:EnvStats':
##
##     boxcox
```

```
## ---
## biotools version 4.2
```

```
#datos_enteros<-read.spss("DB_3.sav", to.data.frame = TRUE)
datos_enteros[15:34,1]<-"china"
datos_enteros[1:15,1]<-"aa"
datos_enteros$PAIS <- as.factor(datos_enteros$PAIS)
col_names <- names(datos_enteros[,2:12])
# to do it for some names in a vector named 'col_names'
datos_enteros[col_names] <- lapply(datos_enteros[col_names] , as.double)
datos_enteros[,1]
```

```
##  [1] aa    aa    aa    aa    aa    aa    aa    aa    aa    aa    aa    aa
## [13] aa    aa    aa    china china china china china china china china china
## [25] china china china china china china china china china china
## Levels: aa china
```

```
boxM(data = datos_enteros[, 2:12], grouping = datos_enteros[, 1])
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  datos_enteros[, 2:12]
## Chi-Sq (approx.) = NA, df = 66, p-value = NA
```

En este caso no rechazamos la hipótesis nula ya que **p-value = 0.132 > 0.001** y por tanto **asumimos la homogeneidad de varianzas**. Es importante recordar que para que esta **conclusión sea fiable** debe darse el supuesto de **normalidad multivariante**.

Mediante análisis discriminante lineal

La función lda del paquete MASS realiza la clasificación.

```
library(MASS)
continente <- datos_enteros$continente
continente <- as.factor(continente)

modelo_lda <- lda(formula = continente ~ ZPOBDENS + ZTMINFAN, data = datos_enteros[,-1])
modelo_lda
```

```
## Call:
## lda(continente ~ ZPOBDENS + ZTMINFAN, data = datos_enteros[,
##     -1])
##
## Prior probabilities of groups:
##     africa    america       asia     europa    oceania
## 0.14705882 0.17647059 0.35294118 0.29411765 0.02941176
##
## Group means:
##           ZPOBDENS   ZTMINFAN
## africa  -0.6742549  1.3449888
```

```
## america -0.9219340 -0.3391427
## asia     0.5598936  0.2636180
## europa   0.3261989 -0.6833203
## oceania -1.0778341 -1.0203009
##
## Coefficients of linear discriminants:
##                LD1        LD2
## ZPOBDENS 0.6469424 -1.0959151
## ZTMINFAN 1.3424019  0.3166111
##
## Proportion of trace:
##    LD1    LD2
## 0.6223 0.3777
```

La salida de este objeto, nos muestra las probabilidades a priori de cada grupo.

Ahora podemos realizar predicciones:

```
nuevas_observaciones <- data.frame(ZPOBDENS = -0.5, ZTMINFAN = 0.7)
predict(object = modelo_lda, newdata = nuevas_observaciones)
```

```
## $class
## [1] asia
## Levels: africa america asia europa oceania
##
## $posterior
##      africa   america      asia    europa     oceania
## 1 0.2948288 0.1252754 0.4343744 0.1432166 0.002304865
##
## $x
##         LD1       LD2
## 1 0.6162101 0.7695853
```

Validación

```
library(biotools)

pred <- predict(modelo_lda, dimen = 1)
confusionmatrix(datos_enteros$continente, pred$class)
```

```
##         new africa new america new asia new europa new oceania
## africa           0           0        5          0           0
## america          0           0        1          5           0
## asia             2           0        7          3           0
## europa           0           0        1          9           0
## oceania          0           0        0          1           0
```

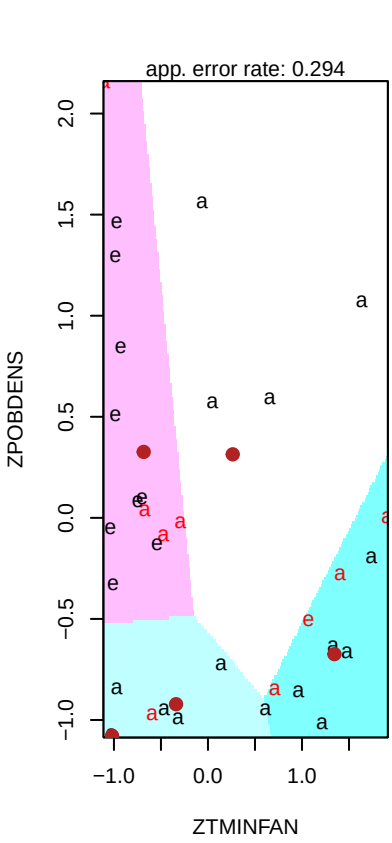
```
# Porcentaje de errores de clasificación
trainig_error <- mean(datos_enteros$continente != pred$class) * 100
paste("trainig_error=", trainig_error, "%")
```

```
## [1] "trainig_error= 52.9411764705882 %"
```

El error con los datos de training es del 52%, lo cuál es bastante alto.

Visualización

```
library(klaR)
continente <- datos_enteros$continente
continente <- as.factor(continente)
partimat(continente ~ ZPOBDENS + ZTMINFAN,
         data = datos_pca, method = "lda", prec = 200,
         col.mean = "firebrick",nplots.vert =1, nplots.hor=3)
```



Mediante análisis discriminante cuadrático

Realizamos ahora análogamente para el modelo cuadrático.

```
library(MASS)

#modelo_qda <- qda(formula = continente ~ ZPOBDENS + ZTMINFAN, data = datos_enteros[,-1])
#modelo_qda
```

La salida de este objeto, nos muestra las probabilidades a priori de cada grupo.

Clustering

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
```

```
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ----------------------------------------- tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x dplyr::recode()  masks car::recode()
## x dplyr::select()  masks MASS::select()
## x purrr::some()    masks car::some()
```

```
library(cluster)
library(factoextra)
```

Vamos a realizar clustering aplicando kmeans, un método no jerárquico de agrupamiento bastante robusto.

Como ya hemos tratado los valores perdidos y outliers vamos simplemente a normalizar los datos:

```
# Para evitar que el análisis cluster se vea influido por cualquier variable arbitraria se estandarizan
datos_pca<-scale(datos_pca)
head(datos_pca)
```

```
##        ZPOBDENS   ZTMINFAN   ZESPVIDA    ZPOBURB   ZTMEDICO    ZPAGRICU
## [1,] -0.8930492  0.9614839 -1.5498548 -0.2148886 -0.7338252 -0.60635548
## [2,] -1.0779535  1.2133839 -0.8948879 -0.7652623 -0.8876598  0.05765957
## [3,] -1.0491372 -0.3185794  0.4388628  1.1381136  1.2660251 -0.75993719
## [4,] -1.1487934 -1.0203009  1.0819213  1.2802934  0.4968519 -1.05806639
## [5,] -0.9999095  0.6119084 -0.3232805  0.4409735 -0.3107799  0.04410824
## [6,] -1.1427900 -1.0280121  1.1295552  0.8170622  0.5160813 -1.10775460
##         ZPSERVI   ZTLIBROP   ZTEJERCI   ZTPOBACT   ZTENERGI
## [1,] 0.48366865 -0.5840208 -0.7378177 -0.7828970  0.4086678
## [2,] 0.05831739 -0.9846493  0.2352615 -2.1340940 -0.4910548
## [3,] 0.78468647 -0.4837537 -0.5268974 -0.4189086 -0.2890268
## [4,] 1.40635370  1.0759263 -0.8775511  0.6244667  1.8687892
## [5,] 0.11066832 -0.5403293 -1.1588230 -0.4272645 -0.7648816
## [6,] 1.62884513  1.6015360 -1.5701880  0.9660097  0.0000000
```

Aplicamos ahora clustering con 4 clusters:

```
k4 <- kmeans(datos_pca, centers = 4, nstart = 25)
str(k4)
```

```
## List of 9
##  $ cluster     : int [1:34] 1 2 1 4 1 4 1 3 1 2 ...
##  $ centers     : num [1:4, 1:11] -0.395 -0.61 0.436 0.254 0.056 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4] "1" "2" "3" "4"
##   .. ..$ : chr [1:11] "ZPOBDENS" "ZTMINFAN" "ZESPVIDA" "ZPOBURB" ...
##  $ totss       : num 363
##  $ withinss    : num [1:4] 30.1 14.7 17.1 74.8
##  $ tot.withinss: num 137
##  $ betweenss   : num 226
##  $ size        : int [1:4] 7 6 6 15
##  $ iter        : int 2
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

La salida que proporciona la función kmeans es una lista de información, de la que destacan las siguientes:

- *cluster*: es un vector de enteros, de 1 a K, que indica el cluster en el que ha sido ubicado cada observación.
- *centers*: una matriz con los sucesivos centros de los clusters.
- *totss*: la suma total de cuadrados.
- *withinss*: vector de suma de cuadrados dentro del cluster (un componente por cluster).
- *tot.withinss*: suma total de cuadrados dentro de conglomerados, i.e. sum(withinss).
- *betweens*: suma de cuadrados entre grupos, i.e. totss-tot.withinss.
- *size*: el número de observaciones en cada cluster.

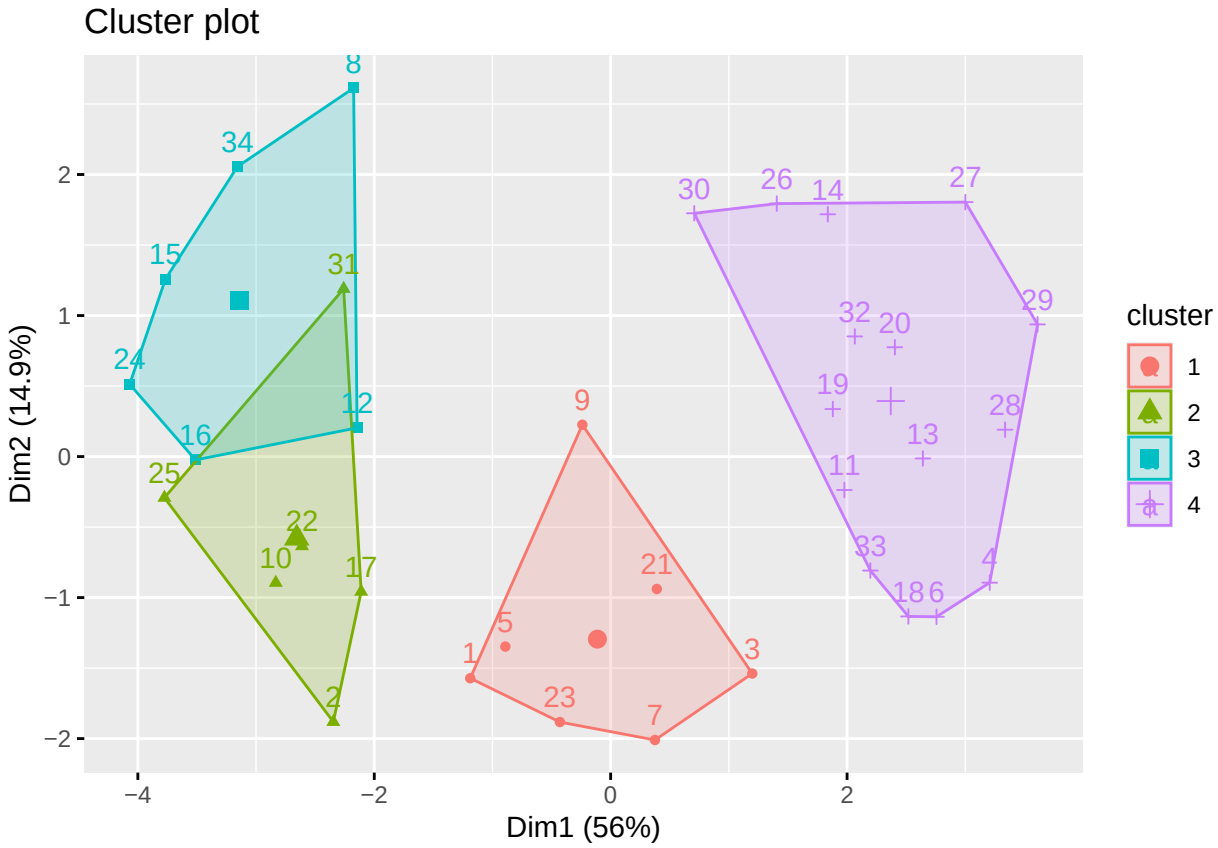
```
k4
```

```
## K-means clustering with 4 clusters of sizes 7, 6, 6, 15
##
## Cluster means:
##     ZPOBDENS    ZTMINFAN   ZESPVIDA    ZPOBURB   ZTMEDICO   ZPAGRICU    ZPSERVI
## 1 -0.3954906  0.05596578 -0.1378483  0.5117358 -0.2599595 -0.3469444  0.5678041
## 2 -0.6098895  1.28492690 -0.9901558 -0.7576182 -0.9533600  0.7126404 -0.6276979
## 3  0.4361005  0.86466516 -1.0377898 -1.5097957 -1.0110480  1.5814311 -1.3137131
## 4  0.2540779 -0.88595418  0.8755075  0.6681555  0.9070776 -0.7557212  0.5115891
##     ZTLIBROP    ZTEJERCI   ZTPOBACT   ZTENERGI
## 1 -0.3785607 -0.26416335 -0.7068722 -0.3922256
## 2 -0.7374936  0.93968111 -1.1844611 -0.8221749
## 3 -0.9566336 -0.84620246  0.2377966 -1.0047631
## 4  0.8543126  0.08588477  0.7085395  0.9138138
##
## Clustering vector:
##  [1] 1 2 1 4 1 4 1 3 1 2 4 3 4 4 3 3 2 4 4 4 1 2 1 3 2 4 4 4 4 4 2 4 4 3
##
## Within cluster sum of squares by cluster:
## [1] 30.05747 14.65619 17.10100 74.77098
##  (between_SS / total_SS =  62.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Al mostrar la variable k4 se ve como las agrupaciones dieron como resultado tamaños realtivamente equilibrados (6, 15, 7 y 6). También se ven los centros de cada grupo (medias) en todas las variables. Y por último la asignación de grupo para cada observación por ejemplo, africasu se asignó al 3, Argelia al 1, etc.

Una forma visual de resumir los resultados de forma elegante y con una interpretación directa es mediante el uso de la función *fviz_cluster*.

```
fviz_cluster(k4,data=datos_pca)
```



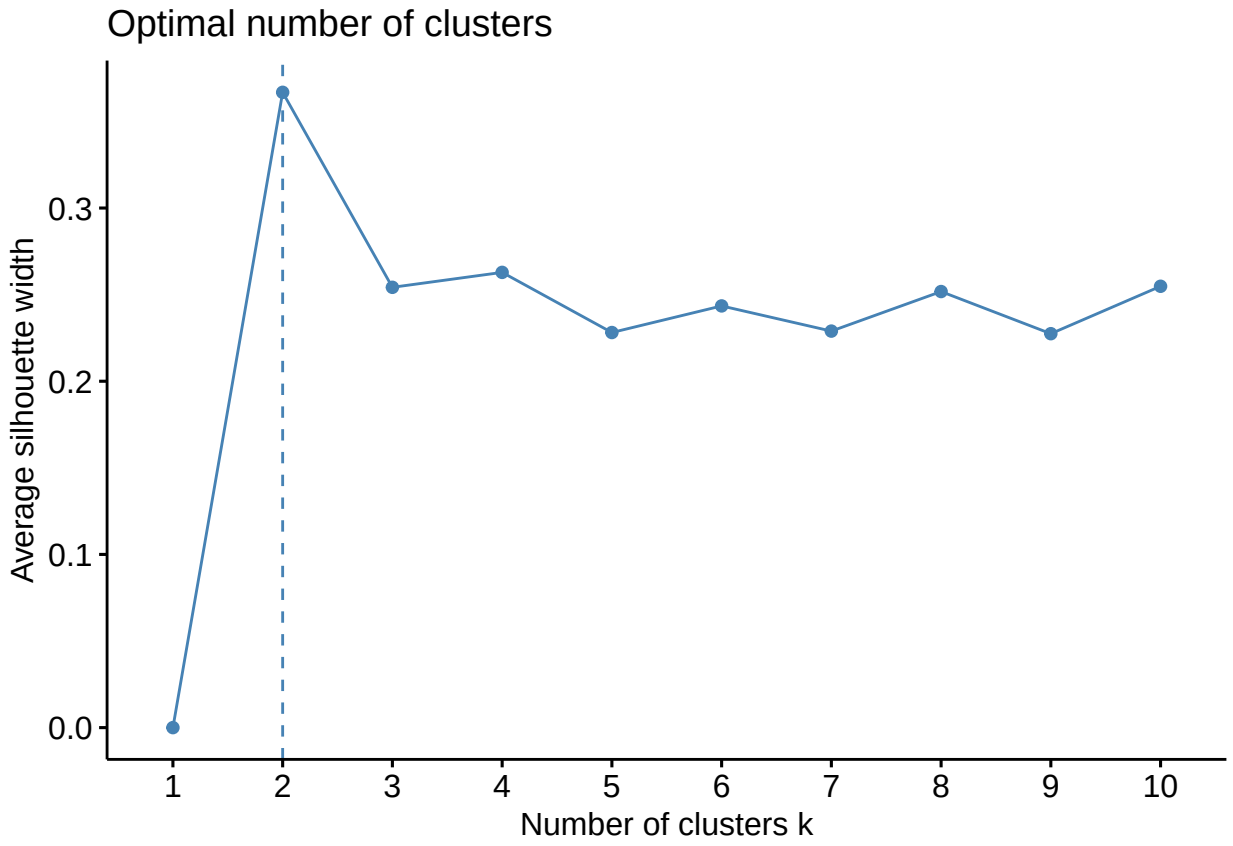
Apreciamos que los clusters 1 y 4 se solapan entre ellos ligeramente, aunque puede ser culpa de la proyección. También parece que el cluster 4 está formado por varios outliers, mientras los otros parecen más compactos.

Buscando el número óptimo de clusters

Buscamos el número de clusters que maximice la similaridad intracluster y minimice la similaridad intercluster. Como kmeans recibe como argumento el número de clusters a realizar, debemos buscar nosotros cuál será el mejor número de clusters.

Para ello vamos a usar el Método de Silhouette, ya que se basa en la medida de calidad de agrupamiento Silhouette, que es ampliamente utilizada. El número óptimo de clusters según este enfoque es, de entre un rango de valores posibles para k, aquel que maximiza la silueta promedio.

```
set.seed(123)

fviz_nbclust(datos_pca, kmeans, method = "silhouette")
```



Vemos un pico considerable en el valor 2, lo que parece razonable ya que contamos con pocos datos. Repetimos ahora el clustering realizado antes fijando como número de clusters 2.

```
k2 <- kmeans(datos_pca, centers = 2, nstart = 123)
str(k2)
```

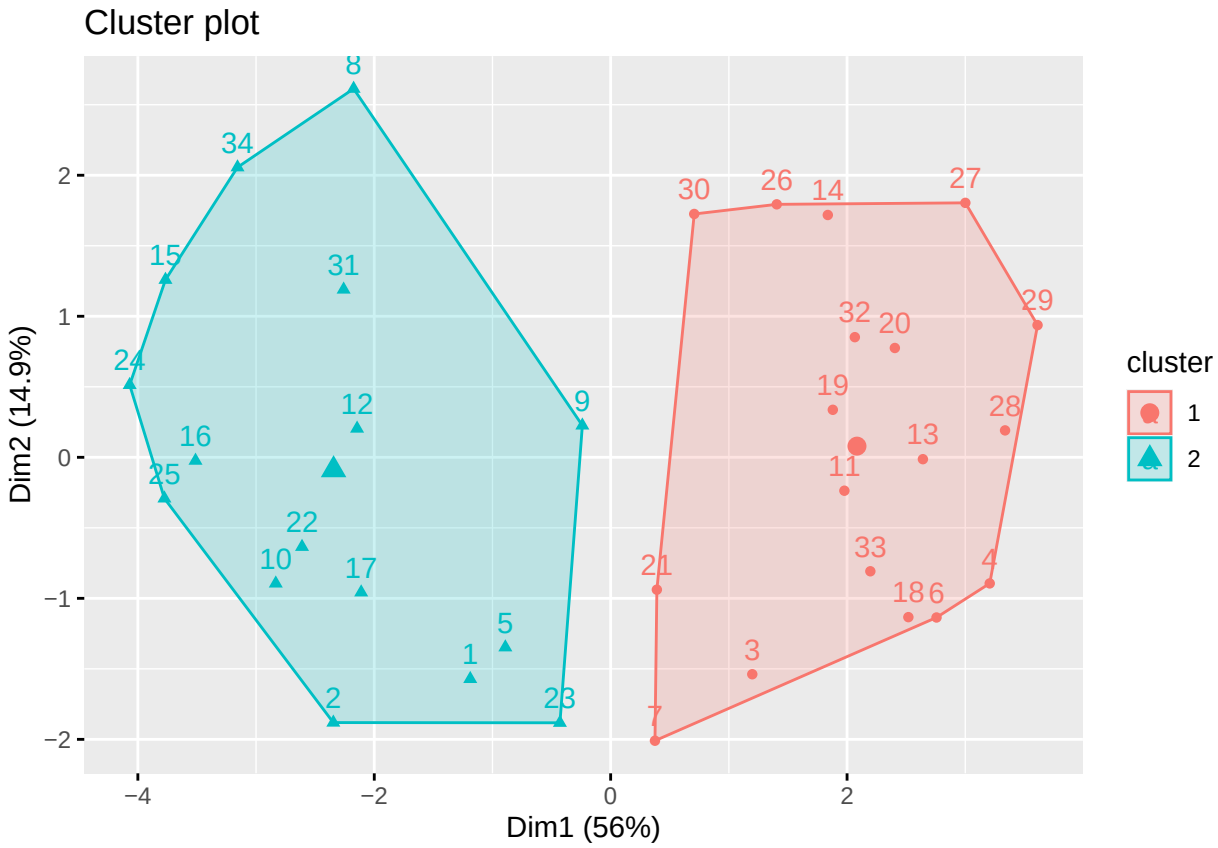
```
## List of 9
##  $ cluster     : int [1:34] 2 2 1 1 2 1 1 2 2 2 ...
##  $ centers     : num [1:2, 1:11] 0.204 -0.23 -0.785 0.884 0.777 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:11] "ZPOBDENS" "ZTMINFAN" "ZESPVIDA" "ZPOBURB" ...
##  $ totss       : num 363
##  $ withinss    : num [1:2] 107.3 88.7
##  $ tot.withinss: num 196
##  $ betweenss   : num 167
##  $ size        : int [1:2] 18 16
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
k2
```

```
## K-means clustering with 2 clusters of sizes 18, 16
##
## Cluster means:
##     ZPOBDENS   ZTMINFAN   ZESPVIDA    ZPOBURB   ZTMEDICO   ZPAGRICU    ZPSERVI
## 1  0.2043031 -0.7853941  0.7769316  0.7240592  0.7671309 -0.7448802  0.5956415
## 2 -0.2298410  0.8835684 -0.8740481 -0.8145666 -0.8630222  0.8379902 -0.6700966
```

```
##     ZTLIBROP   ZTEJERCI   ZTPOBACT   ZTENERGI
## 1  0.6366902  0.1499519  0.4474332  0.6725565
## 2 -0.7162765 -0.1686959 -0.5033623 -0.7566261
##
## Clustering vector:
##  [1] 2 2 1 1 2 1 1 2 2 2 1 2 1 1 2 2 2 1 1 1 1 2 2 2 2 1 1 1 1 1 2 1 1 2
##
## Within cluster sum of squares by cluster:
## [1] 107.32705  88.71239
##  (between_SS / total_SS =  46.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(k2,data=datos_pca)
```



Observamos al realizar esta agrupación dos clusters bien diferenciados, que no se solapan y de tamaños casi idénticos (18 y 16).

Además los centros no parecen demasiado influenciados por algún valor concreto (outlier). También destaca como las medias de las variables en cada cluster están muy diferenciadas, ya que en todas ellas una tiene media postiva y la otra negativa, de forma que se comprueba que en efecto los valores de cada variable (y más en conjunto) son muy diferentes entre los clusters.