R Notebook

Componentes principales

Francisco Mestizo Hernández A01731549

Comenzamos cargando los datos y las librerias que necesitamos

```
install.packages("factoextra")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(stats)
library(factoextra)
```

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

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

```
library(FactoMineR)
library(ggplot2)

#Leemos los datos del csv (No los imprimo porque ocupan mucho espacio en la pantalla)
X = read.csv('paises_mundo.csv')
head(X)
```

```
##   CrecPobl MortInf PorcMujeres  PNB95 ProdElec LinTelf ConsAgua PropBosq
## 1      1.0      30          41   2199     3903      12       94       53
## 2      3.0     124          46   4422      955       6       57       19
## 3      4.3      21          13 133540    91019      96      497        1
## 4      2.5      34          24  44609    19883      42      180        2
## 5      1.3      22          31 278431    65962     160     1043       22
## 6      1.4       6          43 337909   167155     510      933       19
##   PropDefor ConsEner EmisCO2
## 1       0.0      341     1.2
## 2       0.7       89     0.5
## 3       0.0     4566    13.1
## 4       0.8      906     3.0
## 5       0.1     1504     3.5
## 6       0.0     5341    15.3
```

```
#Hay que recordar poner el csv
```

Parte 1

Para las siguientes partes necesitaremos la matriz de covarianza (que llamaremos S) y la de correlacion (que llamaremos R)

```r
S = cov(X)
S
```

```
##                 CrecPobl       MortInf   PorcMujeres         PNB95
## CrecPobl     1.538298e+00  2.195026e+01 -6.078026e+00 -8.933379e+04
## MortInf      2.195026e+01  1.032859e+03 -9.249342e+00 -2.269332e+06
## PorcMujeres -6.078026e+00 -9.249342e+00  7.698322e+01  2.813114e+05
## PNB95       -8.933379e+04 -2.269332e+06  2.813114e+05  4.999786e+10
## ProdElec    -4.973964e+04 -1.043435e+06  2.260248e+05  2.247791e+10
## LinTelf     -1.369079e+02 -4.381366e+03  4.499750e+02  2.039550e+07
## ConsAgua    -4.827092e+01 -1.288211e+03 -1.568313e+03  1.097481e+07
## PropBosq    -3.887018e+00 -1.466316e+01  6.517895e+01  2.474311e+05
## PropDefor    3.361974e-01  1.276296e+01  2.680592e-01 -5.806203e+04
## ConsEner    -8.384169e+02 -4.442568e+04  2.855207e+02  1.415628e+08
## EmisCO2     -1.137877e+00 -9.485500e+01 -2.150132e+00  2.501673e+05
##                  ProdElec       LinTelf      ConsAgua      PropBosq
## CrecPobl    -4.973964e+04 -1.369079e+02 -4.827092e+01     -3.887018
## MortInf     -1.043435e+06 -4.381366e+03 -1.288211e+03    -14.663158
## PorcMujeres  2.260248e+05  4.499750e+02 -1.568313e+03     65.178947
## PNB95        2.247791e+10  2.039550e+07  1.097481e+07 247431.122807
## ProdElec     1.821909e+10  7.583050e+06  1.399817e+07  70359.785965
## LinTelf      7.583050e+06  3.841247e+04  1.193110e+04    248.715789
## ConsAgua     1.399817e+07  1.193110e+04  3.301981e+05  -2220.757895
## PropBosq     7.035979e+04  2.487158e+02 -2.220758e+03    401.003509
## PropDefor   -3.180340e+04 -9.940461e+01 -6.743793e+01      2.625263
## ConsEner     6.801296e+07  3.426262e+05  2.092242e+05  -5153.438596
## EmisCO2      1.392779e+05  6.385700e+02  4.869328e+02    -12.897193
##                 PropDefor      ConsEner       EmisCO2
## CrecPobl     3.361974e-01 -8.384169e+02     -1.137877
## MortInf      1.276296e+01 -4.442568e+04    -94.855000
## PorcMujeres  2.680592e-01  2.855207e+02     -2.150132
## PNB95       -5.806203e+04  1.415628e+08 250167.323509
## ProdElec    -3.180340e+04  6.801296e+07 139277.888640
## LinTelf     -9.940461e+01  3.426262e+05    638.570000
## ConsAgua    -6.743793e+01  2.092242e+05    486.932763
## PropBosq     2.625263e+00 -5.153439e+03    -12.897193
## PropDefor    1.817253e+00 -1.051522e+03     -2.632487
## ConsEner    -1.051522e+03  5.014395e+06  10286.159781
## EmisCO2     -2.632487e+00  1.028616e+04     27.268614
```

```r
R = cor(X)
R
```

```
##                CrecPobl     MortInf PorcMujeres       PNB95    ProdElec
## CrecPobl     1.00000000  0.55067948 -0.55852711 -0.32212154 -0.29711119
## MortInf      0.55067948  1.00000000 -0.03280139 -0.31579250 -0.24053689
## PorcMujeres -0.55852711 -0.03280139  1.00000000  0.14338826  0.19085114
## PNB95       -0.32212154 -0.31579250  0.14338826  1.00000000  0.74476081
## ProdElec    -0.29711119 -0.24053689  0.19085114  0.74476081  1.00000000
## LinTelf     -0.56321228 -0.69558922  0.26167018  0.46539599  0.28664508
## ConsAgua    -0.06772953 -0.06975563 -0.31106243  0.08541500  0.18047653
## PropBosq    -0.15650281 -0.02278415  0.37096694  0.05525919  0.02603078
## PropDefor    0.20107881  0.29459348  0.02266339 -0.19262327 -0.17478434
## ConsEner    -0.30187731 -0.61731132  0.01453216  0.28272492  0.22501894
## EmisCO2     -0.17568860 -0.56520778 -0.04692837  0.21425123  0.19760017
```

```
##                 LinTelf    ConsAgua    PropBosq   PropDefor    ConsEner
## CrecPobl    -0.56321228 -0.06772953 -0.15650281  0.20107881 -0.30187731
## MortInf     -0.69558922 -0.06975563 -0.02278415  0.29459348 -0.61731132
## PorcMujeres  0.26167018 -0.31106243  0.37096694  0.02266339  0.01453216
## PNB95        0.46539599  0.08541500  0.05525919 -0.19262327  0.28272492
## ProdElec     0.28664508  0.18047653  0.02603078 -0.17478434  0.22501894
## LinTelf      1.00000000  0.10593934  0.06337138 -0.37623801  0.78068385
## ConsAgua     0.10593934  1.00000000 -0.19299225 -0.08705811  0.16259804
## PropBosq     0.06337138 -0.19299225  1.00000000  0.09725032 -0.11492480
## PropDefor   -0.37623801 -0.08705811  0.09725032  1.00000000 -0.34833836
## ConsEner     0.78068385  0.16259804 -0.11492480 -0.34833836  1.00000000
## EmisCO2      0.62393719  0.16227447 -0.12333592 -0.37396154  0.87965517
##                EmisCO2
## CrecPobl    -0.17568860
## MortInf     -0.56520778
## PorcMujeres -0.04692837
## PNB95        0.21425123
## ProdElec     0.19760017
## LinTelf      0.62393719
## ConsAgua     0.16227447
## PropBosq    -0.12333592
## PropDefor   -0.37396154
## ConsEner     0.87965517
## EmisCO2      1.00000000
```

Realizamos el calculo de vectores propios para S

```
eigen = eigen(S)
```

Sacamos los valores de las varianzas acumuladas

```
sum <- 0
for (varianza in eigen[1]) {
  print(varianza/sum(diag(S)))
  sum <- sum + varianza/sum(diag(S))*100
}
```

```
##  [1] 9.034543e-01 9.647298e-02 6.795804e-05 4.554567e-06 1.782429e-07
##  [6] 7.530917e-09 5.317738e-09 6.657763e-10 8.502887e-11 2.107843e-11
## [11] 6.989035e-12
```

Y desplegamos las varianzas acumuladas

```
cumsum(sum)
```

```
##  [1]  90.34543  99.99273  99.99953  99.99998 100.00000 100.00000 100.00000
##  [8] 100.00000 100.00000 100.00000 100.00000
```

Ahora hacemos el mismo procedimiento pero para la matriz de correlacion R

```
eigen = eigen(R)

sum <- 0
for (varianza in eigen[1]) {
  print(varianza/sum(diag(R)))
  sum <- sum + varianza/sum(diag(R))*100
}
```

```
##  [1] 0.366352638 0.175453813 0.124582832 0.078592361 0.072194597 0.066290906
```

```
##  [7] 0.051936828 0.029709178 0.015278951 0.013302563 0.006305332
```

Y desplegamos los acumulados

```
cumsum(sum)
```

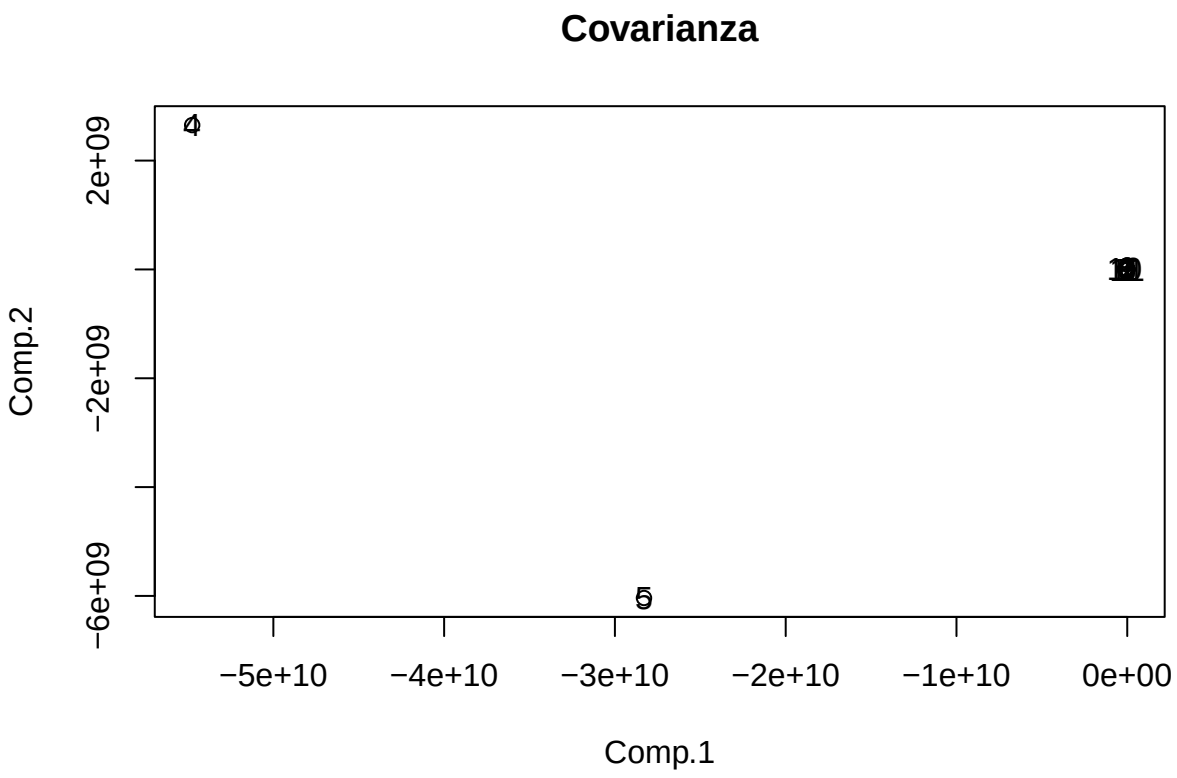
```
##  [1]  36.63526  54.18065  66.63893  74.49816  81.71762  88.34671  93.54040
##  [8]  96.51132  98.03921  99.36947 100.00000
```

Podemos ver que los valores de las varianzas azumuladas tienen más sentido, ya que se ve como va aumentando el porcentaje. Esto se debe a que con la correlación los datos están estandarizados, pero cuando hacemos el análisis con la matriz de covarianza no lo están. Por lo tanto, la escala de los datos afecta el resultado.

Viendo la varianza, los componentes principales más importantes serían el primero y el segundo, por eso para la siguientes partes se usarán esos. Más adelante se verán las variables que más afectan a estas componentes principales

Parte 2

```
datos=S
cpS=princomp(datos,cor=FALSE)
cpaS=as.matrix(datos)%*%cpS$loadings
plot(cpaS[,1:2],type="p", main = "Covarianza")
text(cpaS[,1],cpaS[,2],1:nrow(cpaS))
```

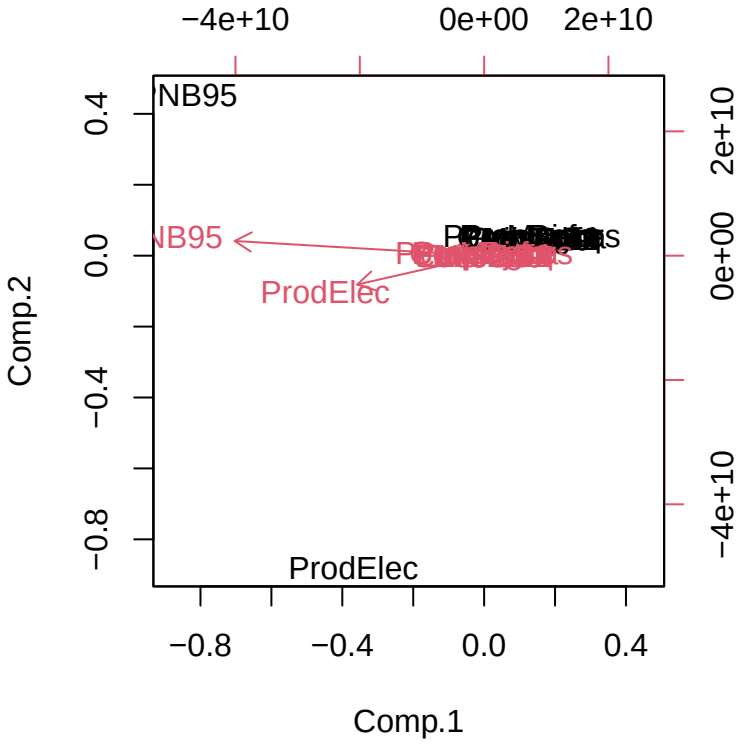



```
biplot(cpS)
```

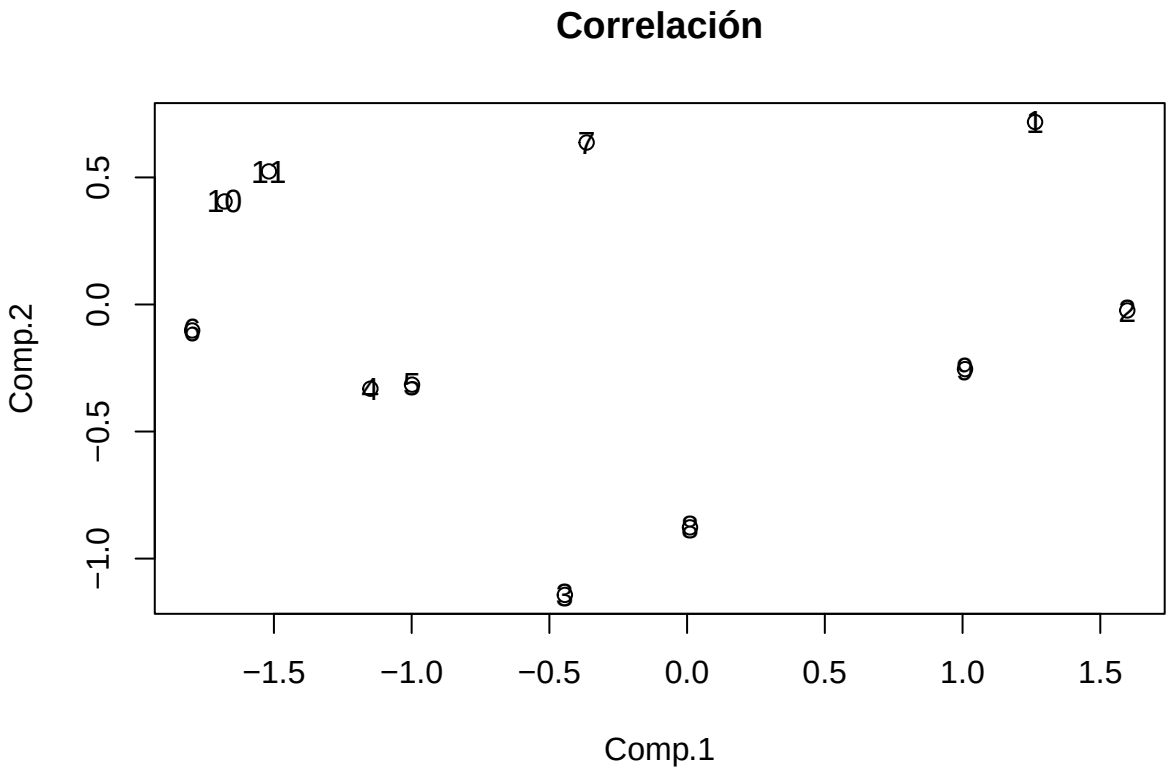
```
## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped
```

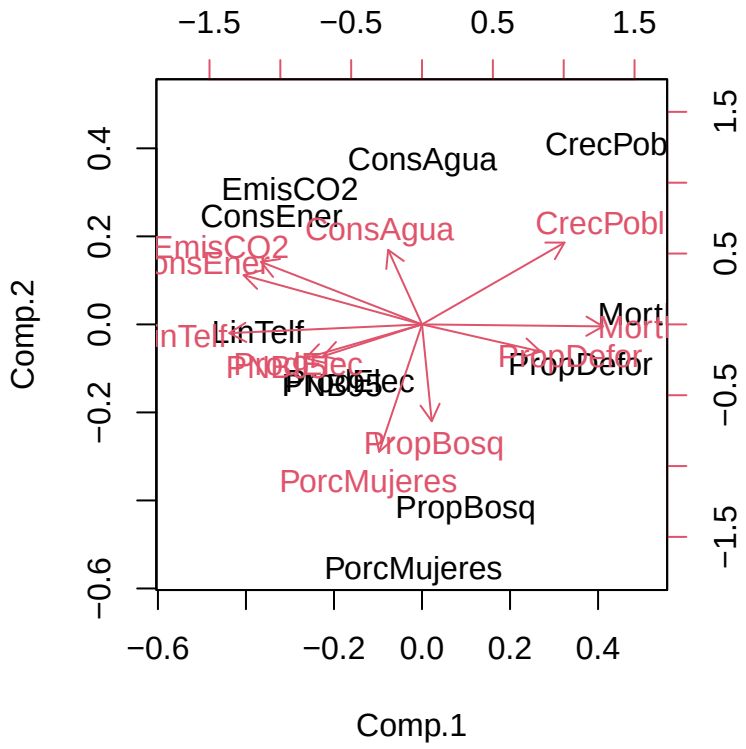
```
## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped
```



```r
datos=R
cpS=princomp(datos,cor=FALSE)
cpaS=as.matrix(datos)%*%cpS$loadings
plot(cpaS[,1:2],type="p", main = "Correlación")
text(cpaS[,1],cpaS[,2],1:nrow(cpaS))
```




```
biplot(cpS)
```

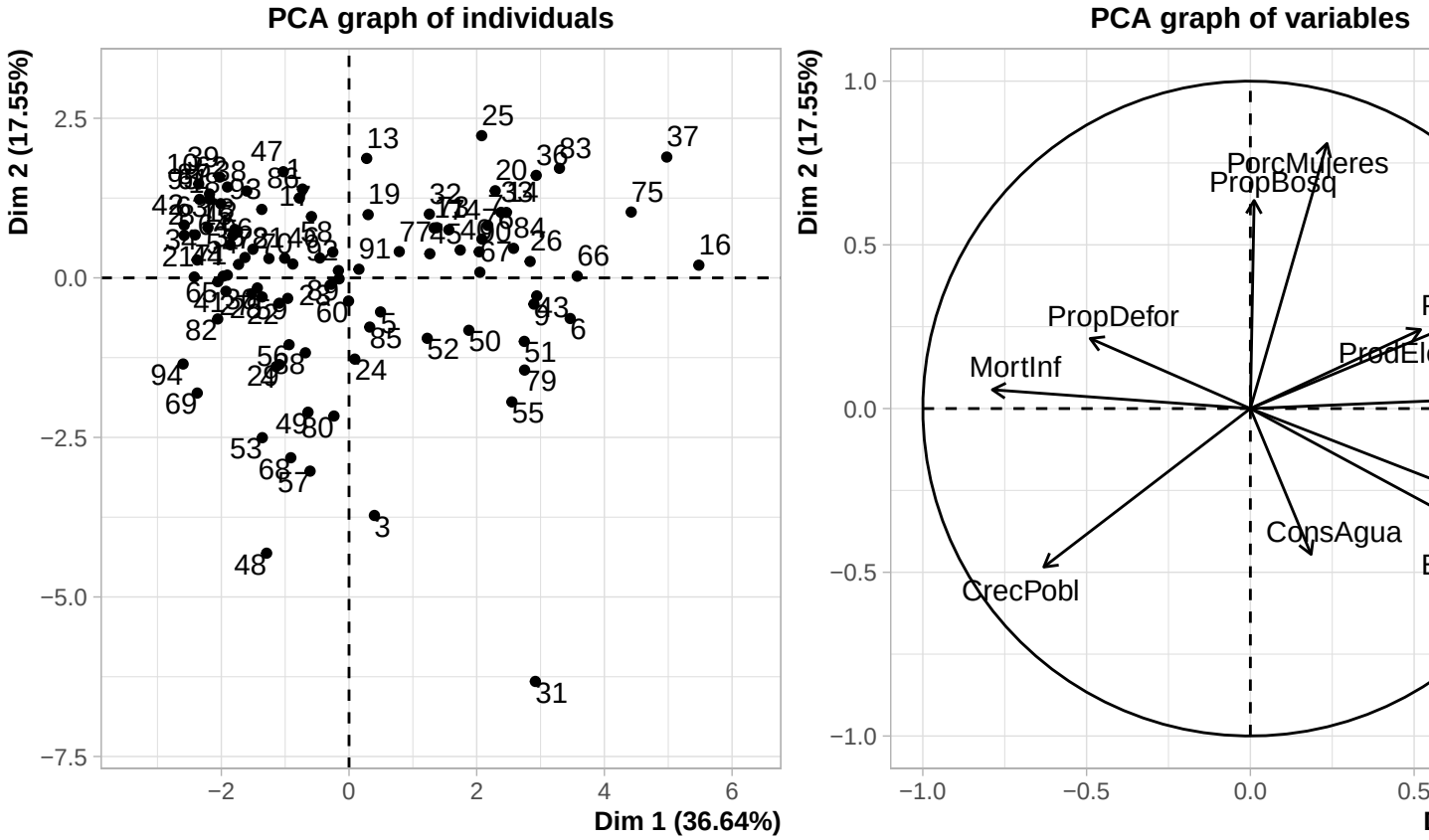


El primer gráfico nos indica las puntuaciones que se obtienen por cada componente para cada individuo. Esto se hace para los dos componentes que tienen mas variacion.

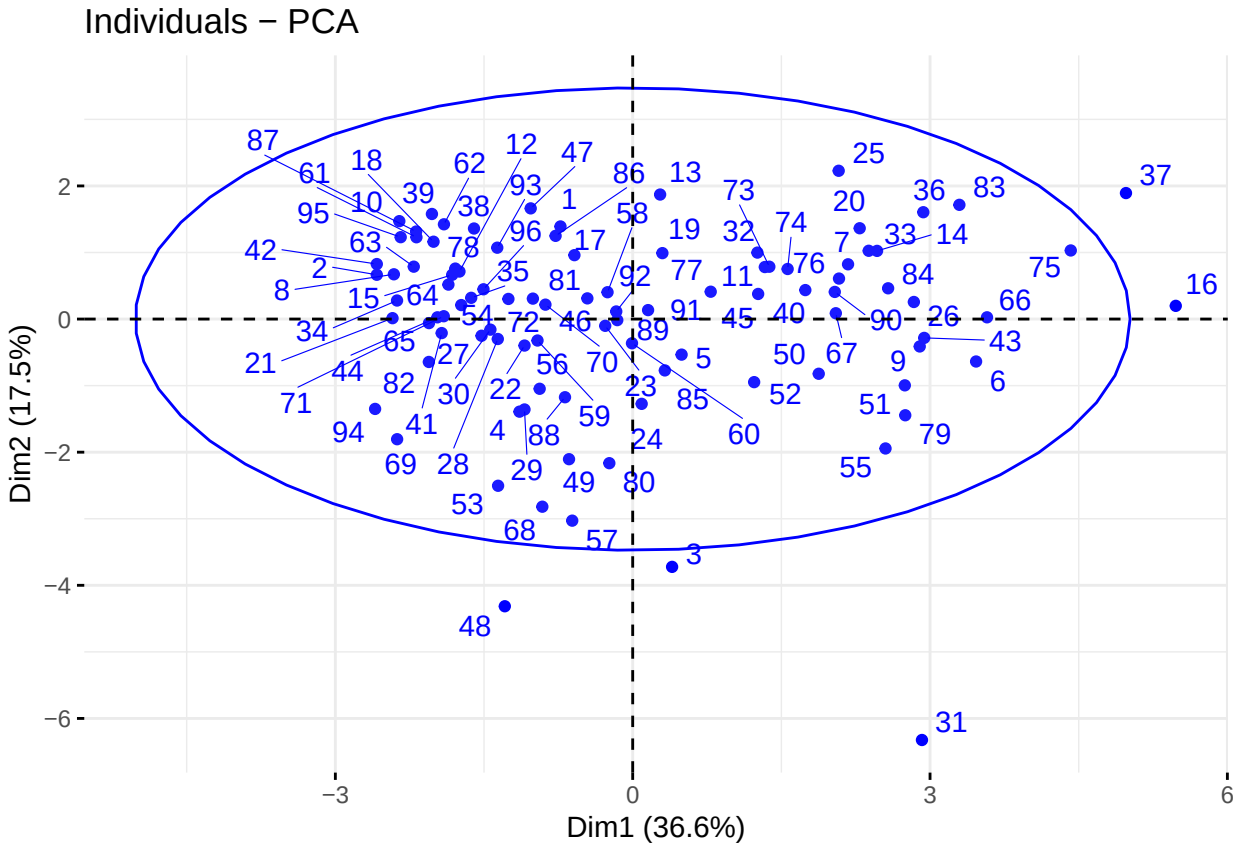
La segunda gráfica nos vincula las variables originales con los componentes y se puede apreciar el peso que tiene cada variable sobre el componente

Parte 3

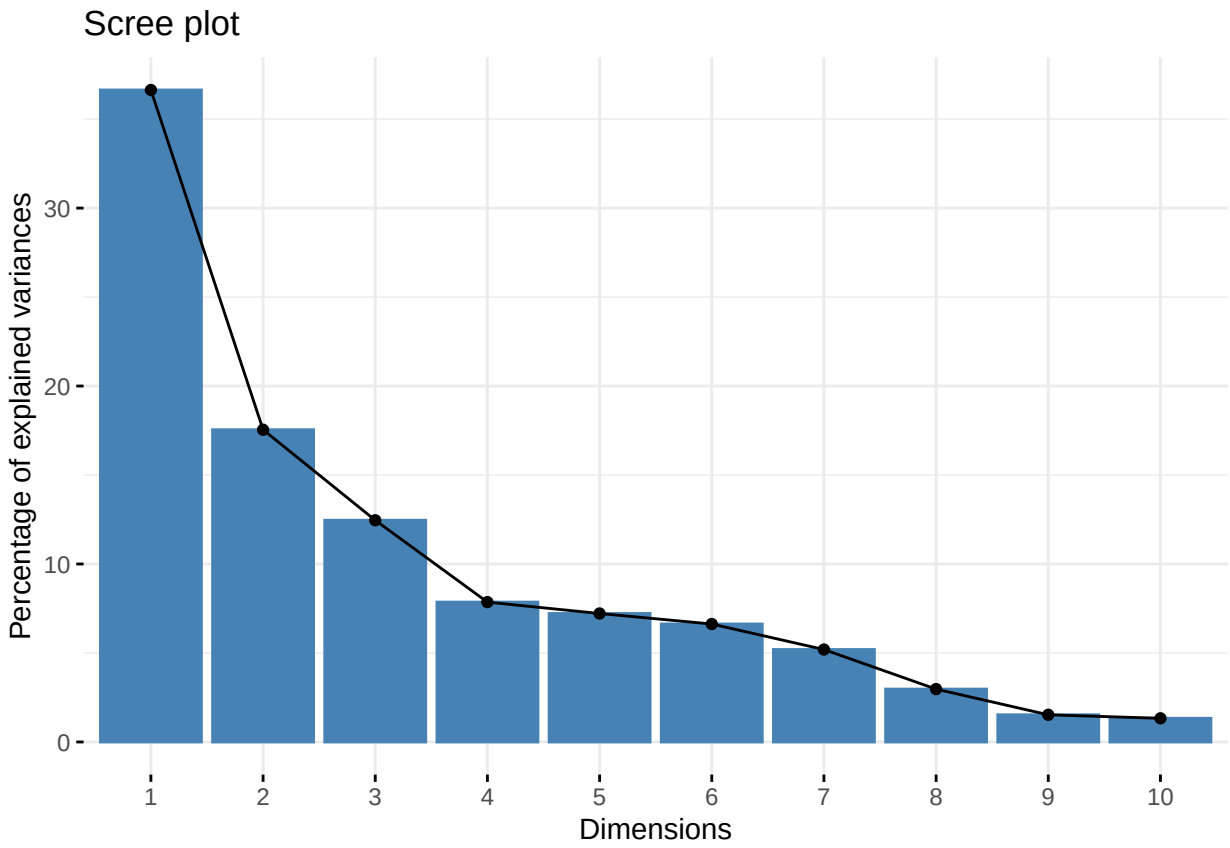
```
datos=X
cp3 = PCA(datos)
```



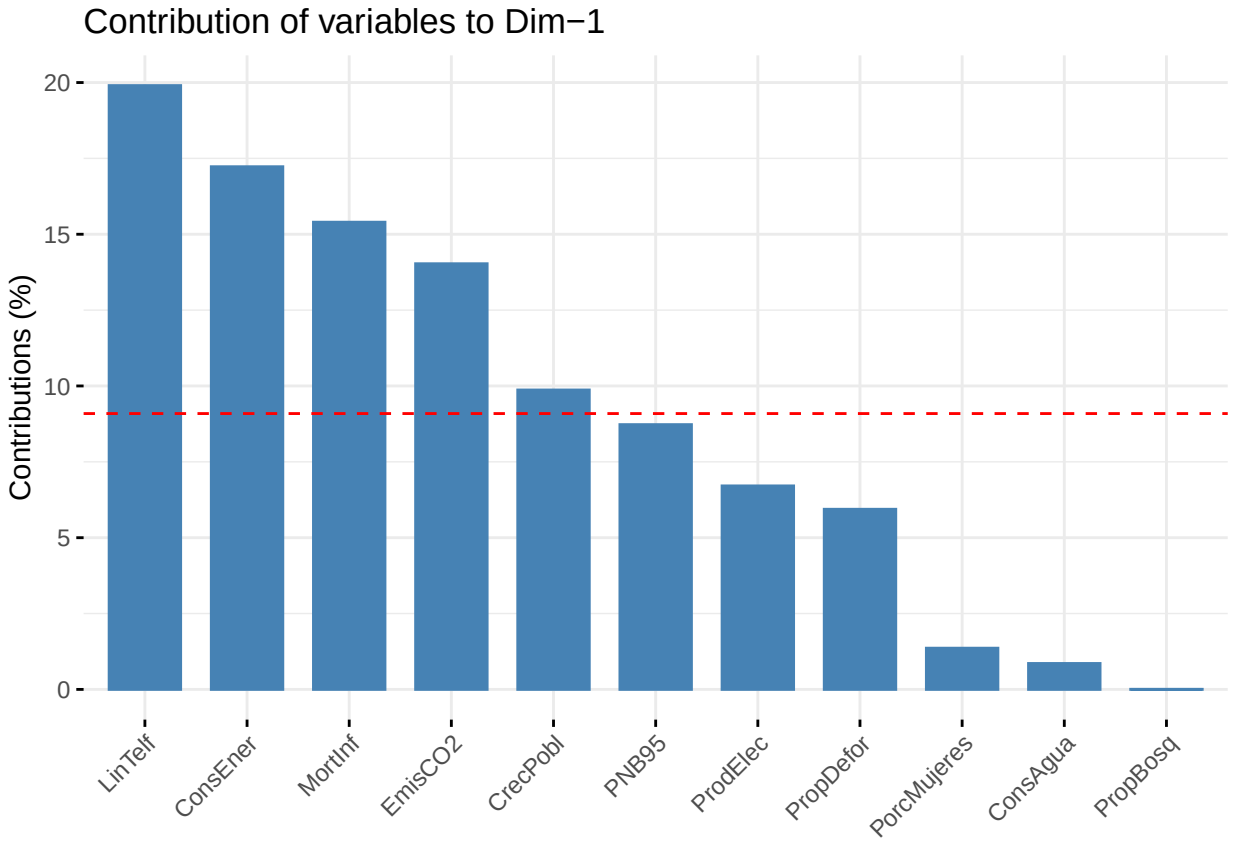

```
fviz_pca_ind(cp3, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```




```
fviz_screeplot(cp3)
```




```
fviz_contrib(cp3, choice=c("var"))
```



En la gráfica donde aparece el elipse, podemos ver el 95% de los datos contenidos en la elipse. Por lo tanto, podemos ver que hay datos atipicos, aproximadamente 5.

Después podemos ver el gráfico de la varianza explicada para cada componente y se puede ver que lo mejor es elegir dos componentes.

Y por ultimo, en la tercera gráfica vemos cuales variables son las que contribuyen más al primer componente principal. Ya podiamos ver más o menos esta misma información en las gráficas de la parte dos pero en esta nos enfocamos en el primer componente principal. Podemos ver que la mitad de svariables contribuyen más que el promedio a este componente.