Actividad 4

José Carlos Sánchez Gómez

2024-10-10

```
data =
read.csv("C:\\Users\\jcsg6\\Documentos\\Uni\\SeptimoSemestre\\Estadistica
Avanzada\\corporal.csv")

data = data[, -4]
```

Parte 1

Varianza - covarianza

```
S = cov(data)
R = cor(data)

eigen_s = eigen(S)
eigen_r = eigen(R)

eigen_r$values
```

```
## [1] 3.75749733 0.72585665 0.32032981 0.12461873 0.07169749
```

```
eigen_s$values
```

```
## [1] 359.3980243  80.3757858  27.6229011   4.3074318   0.2343571
```

```
# Proporción de varianza explicada
suma_varianza_explicada_s = 0
for (i in 1 : length(eigen_s$values)) {
  varianza =  eigen_s$values[i] / sum(diag(S))
  suma_varianza_explicada_s = suma_varianza_explicada_s + varianza
  cat("Componente ", i, ": ", varianza, "\n")
}
```

```
## Componente  1 :  0.7615357
## Componente  2 :  0.1703099
## Componente  3 :  0.05853072
## Componente  4 :  0.009127104
## Componente  5 :  0.0004965839
```

```
cat("Suma total de la varianza: ", suma_varianza_explicada_s)
```

```
## Suma total de la varianza:  1
```

```
suma_varianza_explicada_r = 0
for (i in 1 : length(eigen_r$values)) {
  varianza =  eigen_r$values[i] / sum(diag(R))
```

```
  suma_varianza_explicada_r = suma_varianza_explicada_r + varianza
  cat("Componente ", i, ": ", varianza, "\n")
}
```

```
## Componente  1 :  0.7514995
## Componente  2 :  0.1451713
## Componente  3 :  0.06406596
## Componente  4 :  0.02492375
## Componente  5 :  0.0143395
```

```
cat("Suma total de la varianza: ", suma_varianza_explicada_r)
```

```
## Suma total de la varianza:  1
```

De acuerdo con los resultados anteriores podemos entender que los componentes más importantes son los primeros dos, pues estos representan poco más del 90%.

```
CP1_s = eigen_s$vectors[, 1]
CP2_s = eigen_s$vectors[, 2]

CP1_r = eigen_r$vectors[, 1]
CP2_r = eigen_r$vectors[, 2]
```

```
CP1_s
```

```
## [1] -0.34871002 -0.76617586 -0.47632405 -0.05386189 -0.24817367
```

```
CP2_s
```

```
## [1]  0.9075501 -0.1616581 -0.3851755  0.0155423 -0.0402221
```

```
CP1_r
```

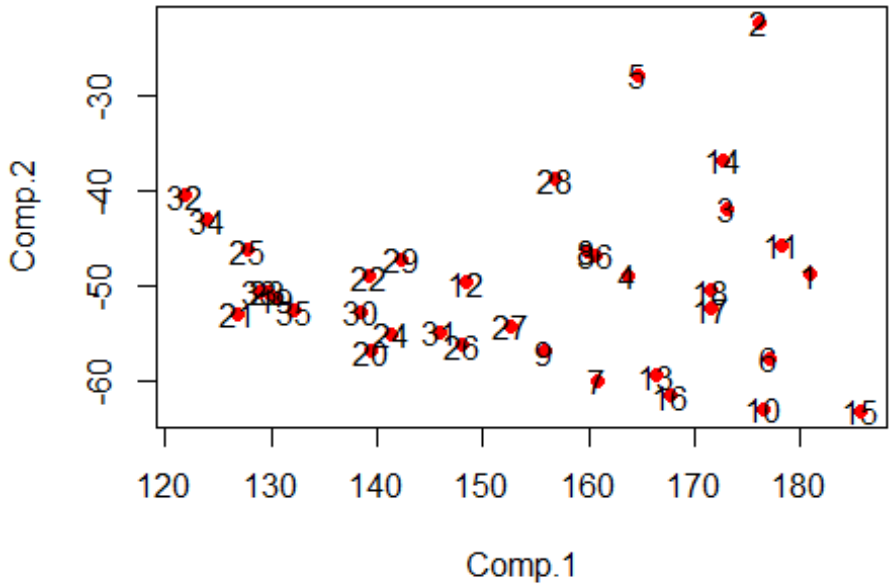
```
## [1] -0.3359310 -0.4927066 -0.4222426 -0.4821923 -0.4833139
```

```
CP2_r
```

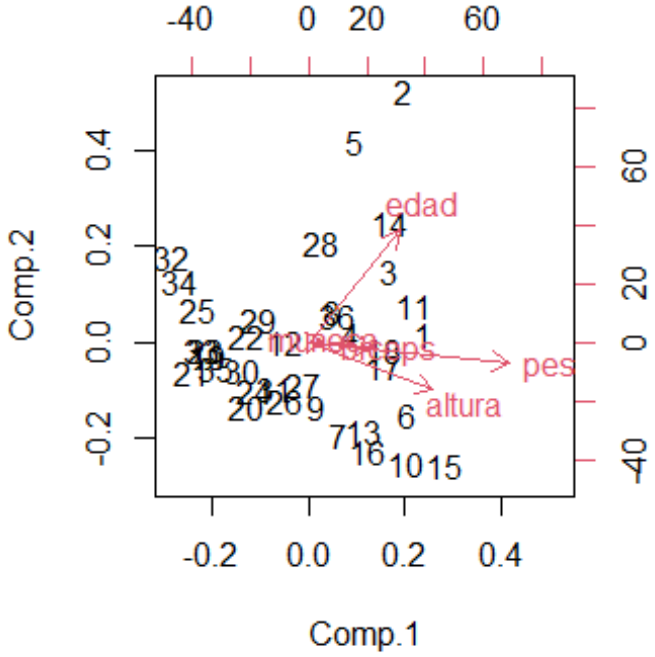
```
## [1]  0.8575601 -0.1647821 -0.4542223  0.1082775 -0.1392684
```

Parte dos

```
library(stats)
cpS=princomp(data,cor=FALSE) #Para la matriz de correlación usa cor=TRUE
cpaS=as.matrix(data)%*%cpS$loadings  #Calcula las puntuaciones
plot(cpaS[,1:2],type="p", pch = 19, col = "red")
text(cpaS[,1],cpaS[,2],1:nrow(cpaS))
```




```
biplot(cpS)
```



Parece ser que el componente uno tiene una relacion con la variable de altura, debido a los valores del eje x.

Podemos observar en estas graficas que las principales variables que afectan son las de edad, peso y altura; teniendo a la altura y el peso con una mayor relación entre ellas y con el componente uno, pues sus flechas apuntan fuertemente hacia su eje; la edad se ve más balanceada entre las dos.

De igual manera se pueden observar lo que podrian ser algunos datos atipicos, pues estos se alejan considerablemente del grupo central. Estos valores son el 2, 5, 14. Posiblemente estos valores tengan más relación con otras variables.

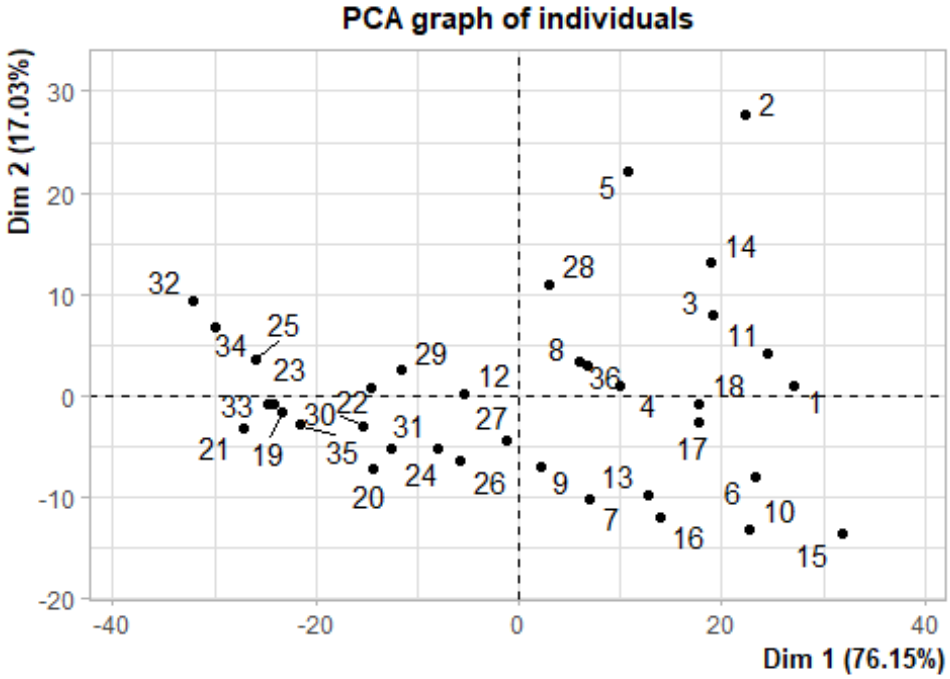
Algunos otros métodos útiles dentro de princomp es el de cor, pues este nos puede decir si el calculo deberia de hacerse con la matriz de correlación, o la de covarianza. Otra es la de scale, pues regresa la escala que fue aplicada a cada variable.

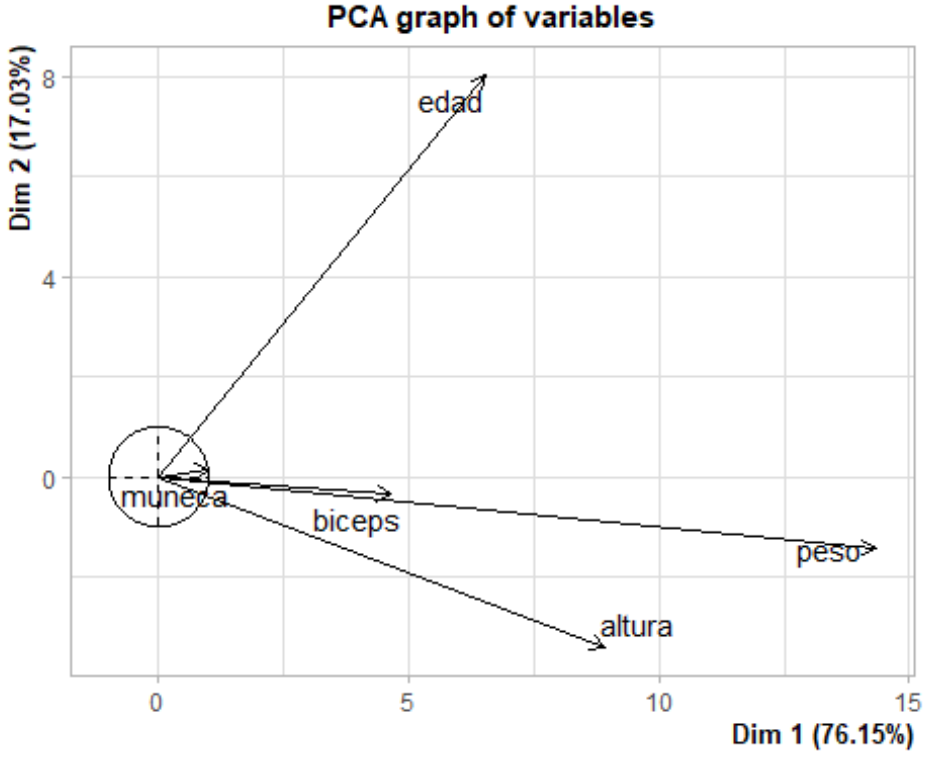
El comando de summary te da un resumen del análisis de componentes principales. Muestra la proporcion de varianza explicada por cada uno de los componentes principales. Loadings regresa la carga de los componentes, que son los coeficientes de la scombinaciones lineales que fomran cada componente. Cada coeficiente muestran cómo cada variable original contribuye a cada componente. Scores proporciona las puntuaciones de cada componente para cada observación en los datos, lo cual indica cómo se proyectan los dtos en el espacio de los componentes principales.

Parte tres

Graficos con matriz de varianza-covarianza

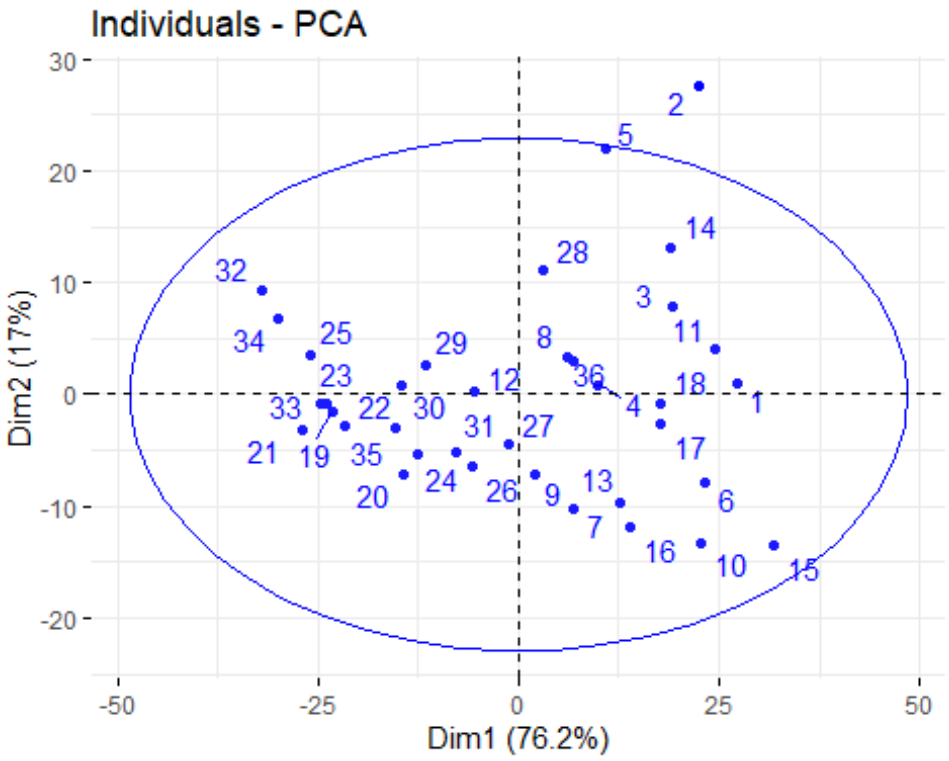
```
library(FactoMineR)
library(ggplot2)

cpS = PCA(data,scale.unit=FALSE, graph = TRUE) #Para matriz de
correlaciones usa scale.unit=TRUE
```



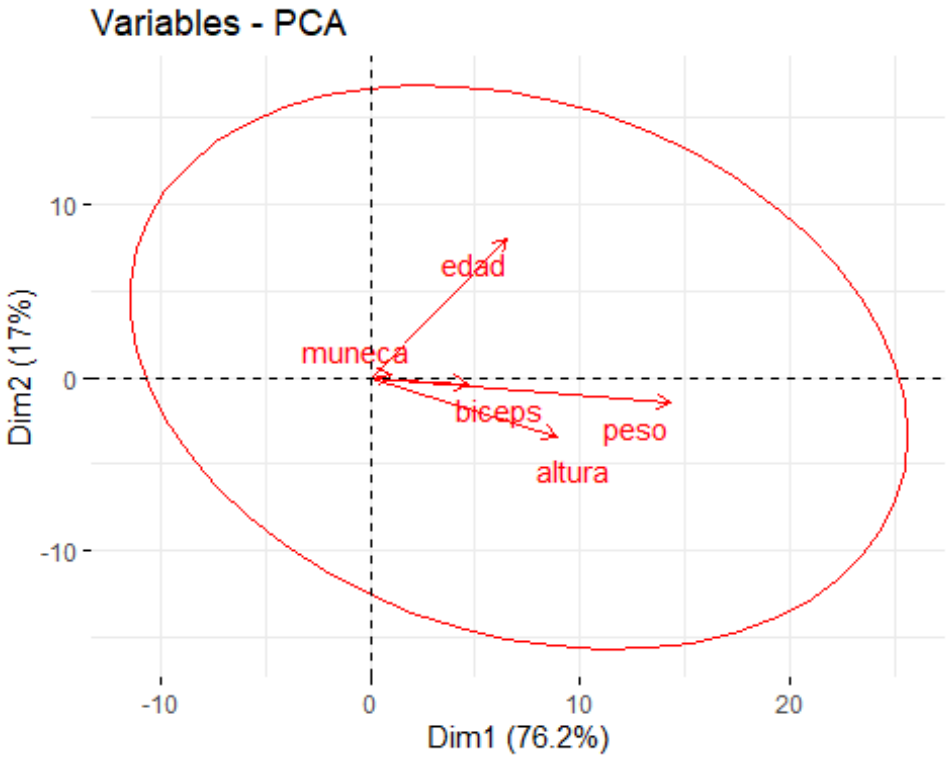



```
library(factoextra)

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

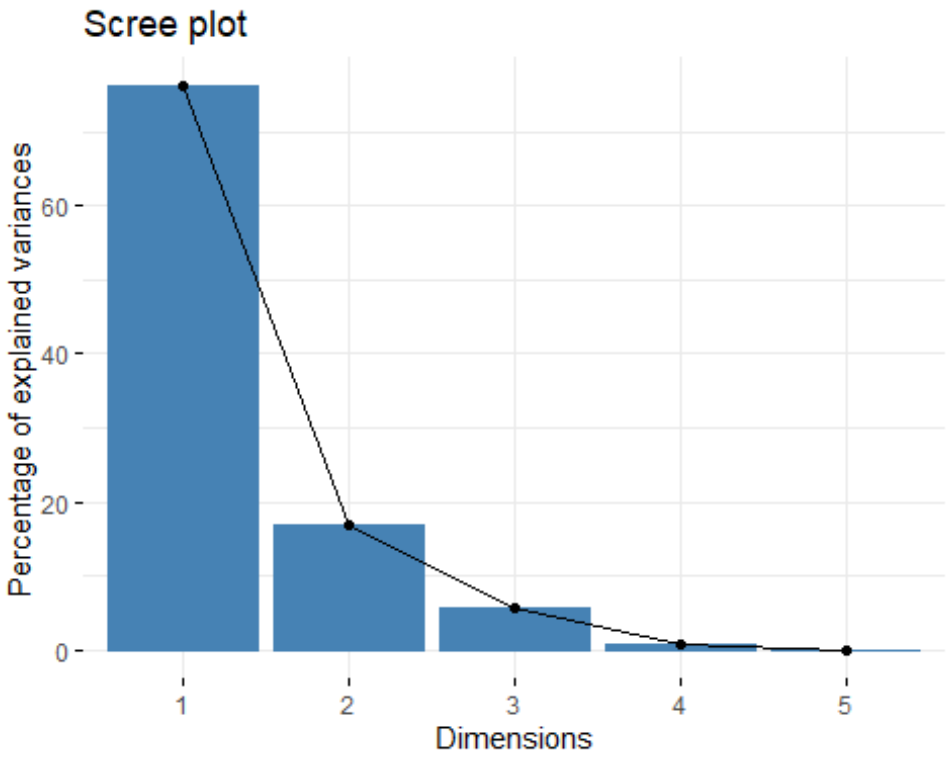
```
fviz_pca_ind(cpS, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```



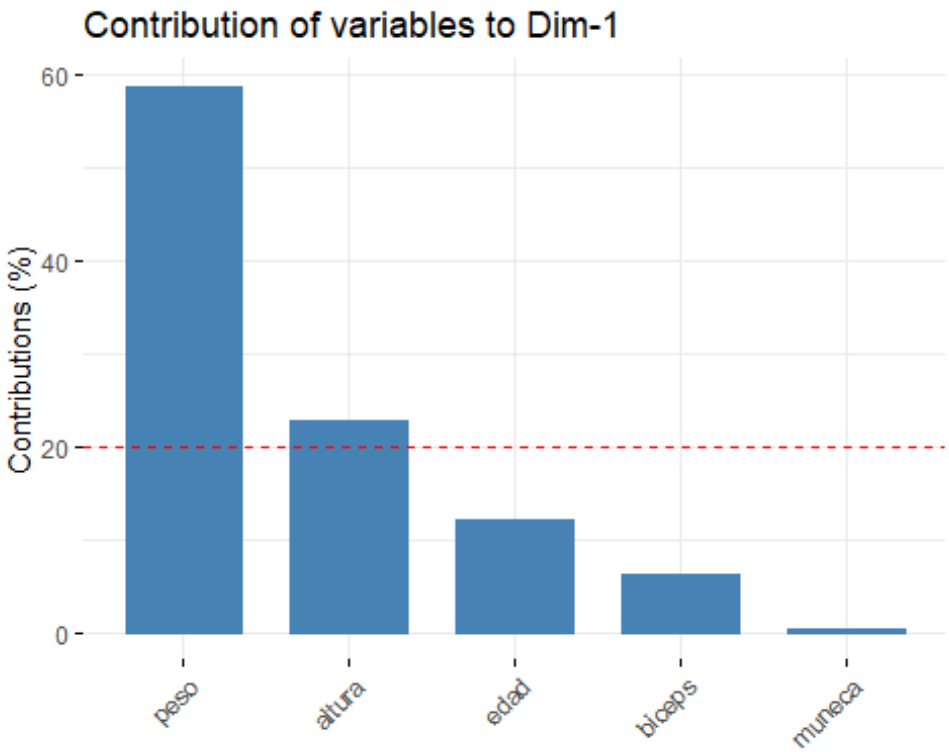

```
fviz_pca_var(cpS, col.var = "red", addEllipses = TRUE, repel = TRUE)
```



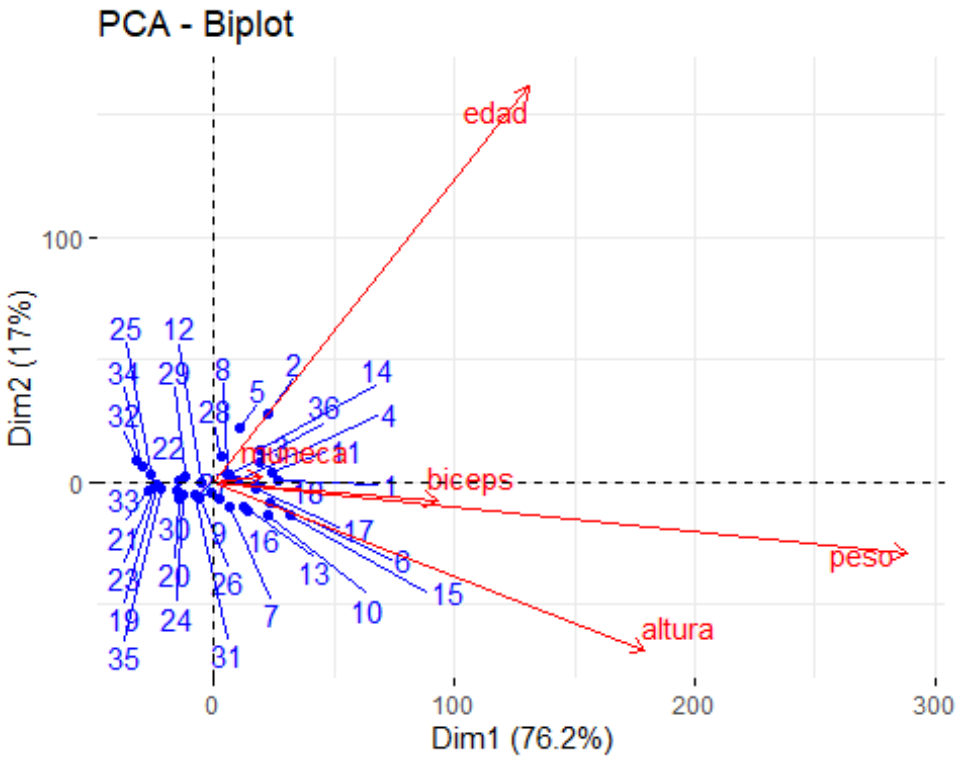
```r
fviz_screeplot(cpS)
```




```r
fviz_contrib(cpS, choice = "var")
```

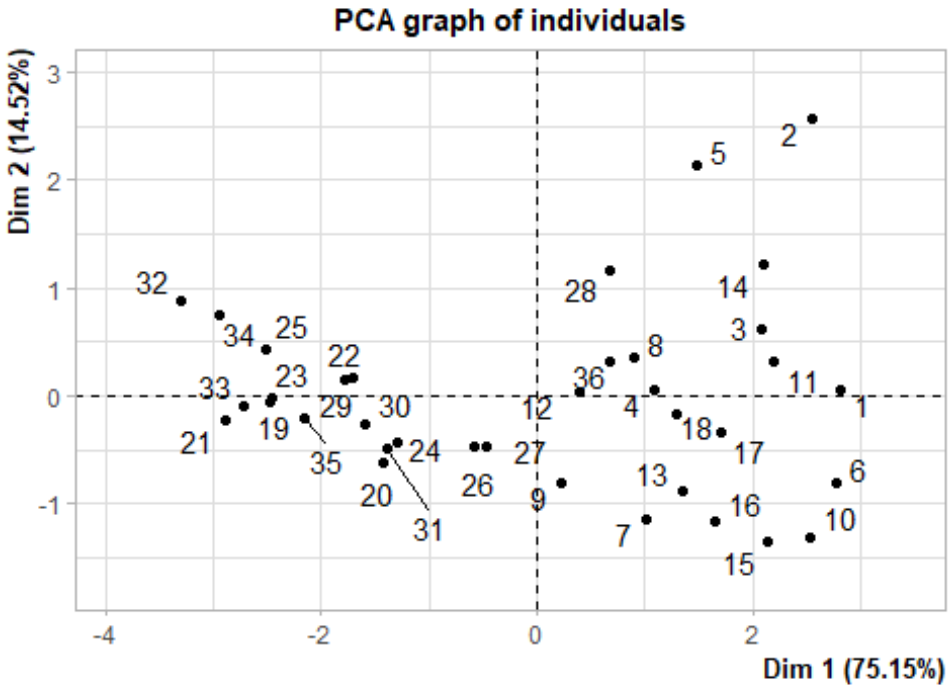


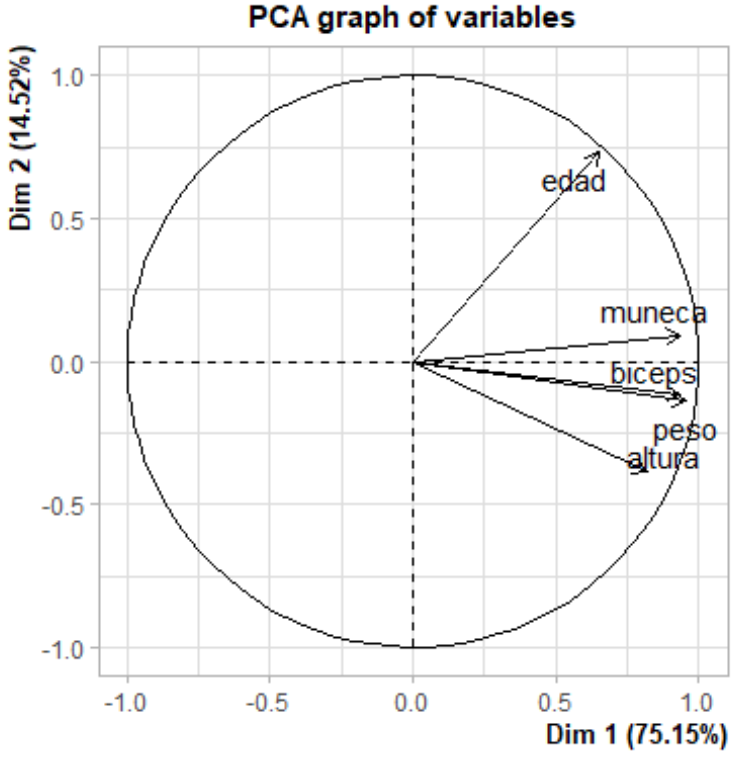
```r
fviz_pca_biplot(cpS, repel=TRUE, col.var="red", col.ind="blue")
```



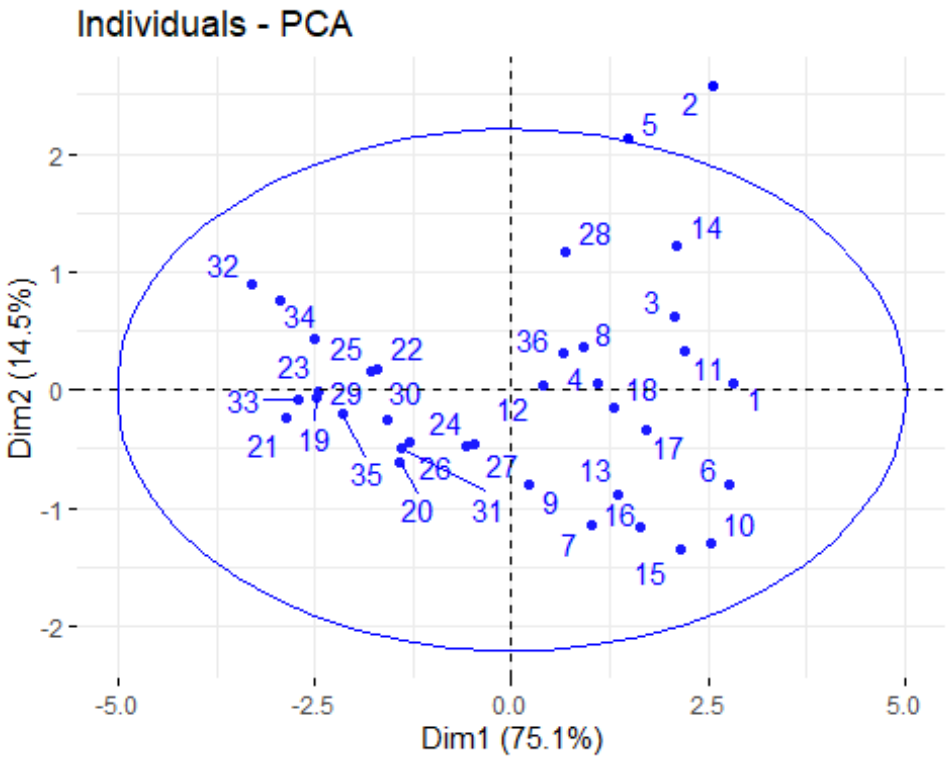
Graficas con matriz de correlación

```r
library(FactoMineR)
library(ggplot2)

cpS = PCA(data,scale.unit=TRUE, graph = TRUE) #Para matriz de correlaciones usa scale.unit=TRUE
```

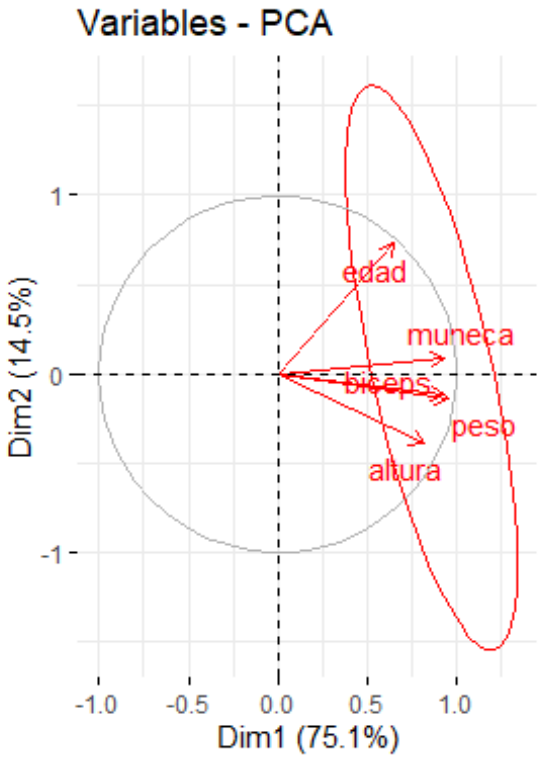




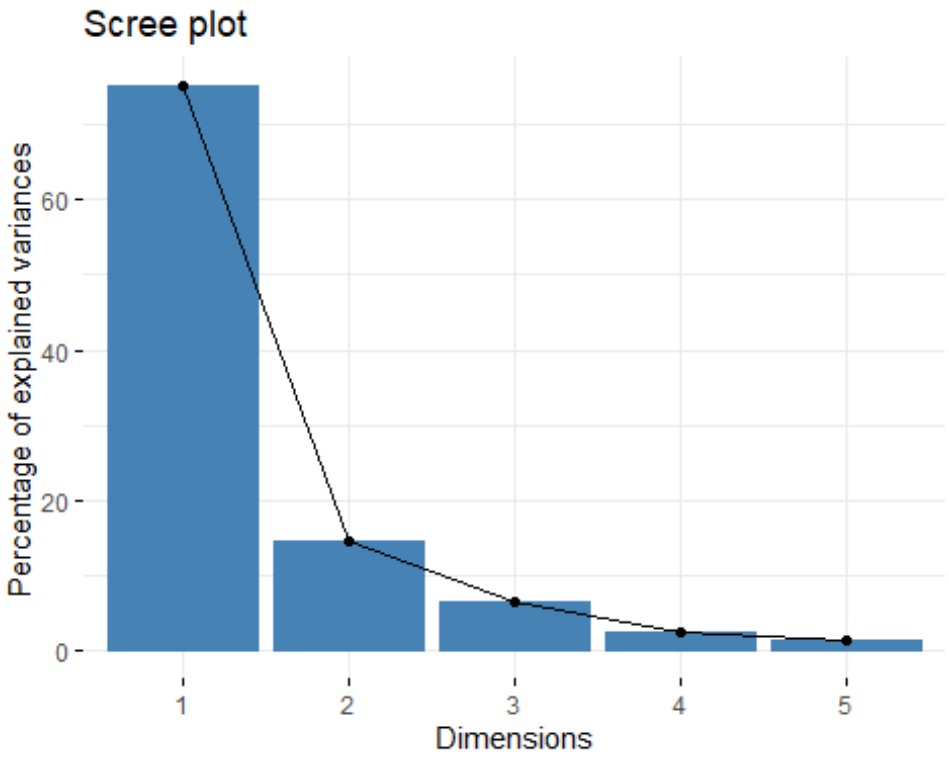

```r
library(factoextra)

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



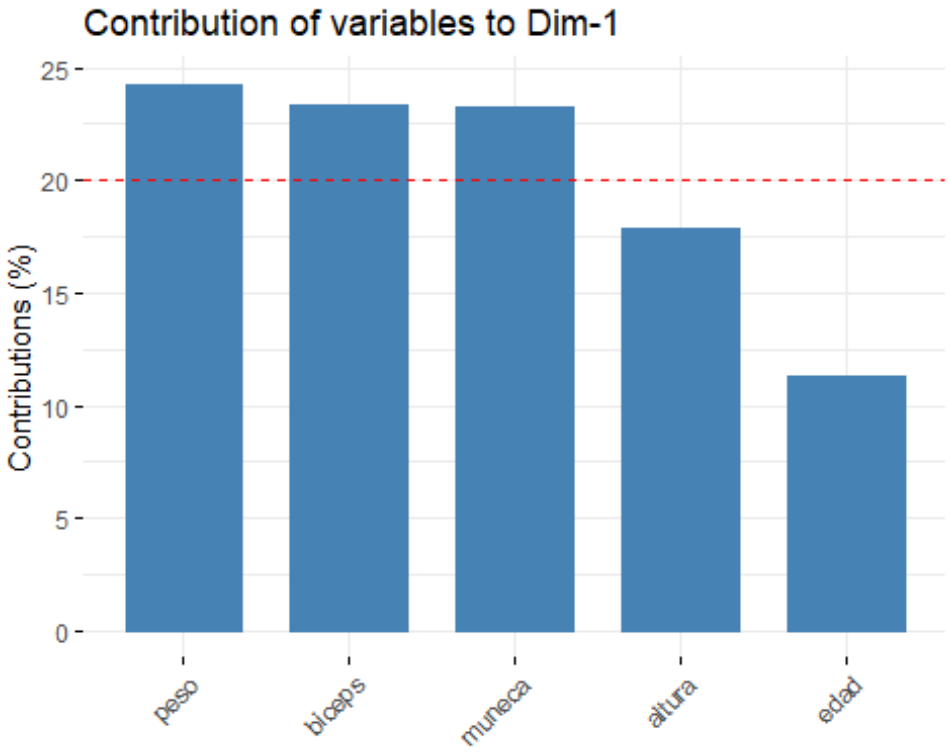

```
fviz_pca_var(cpS, col.var = "red", addEllipses = TRUE, repel = TRUE)
```



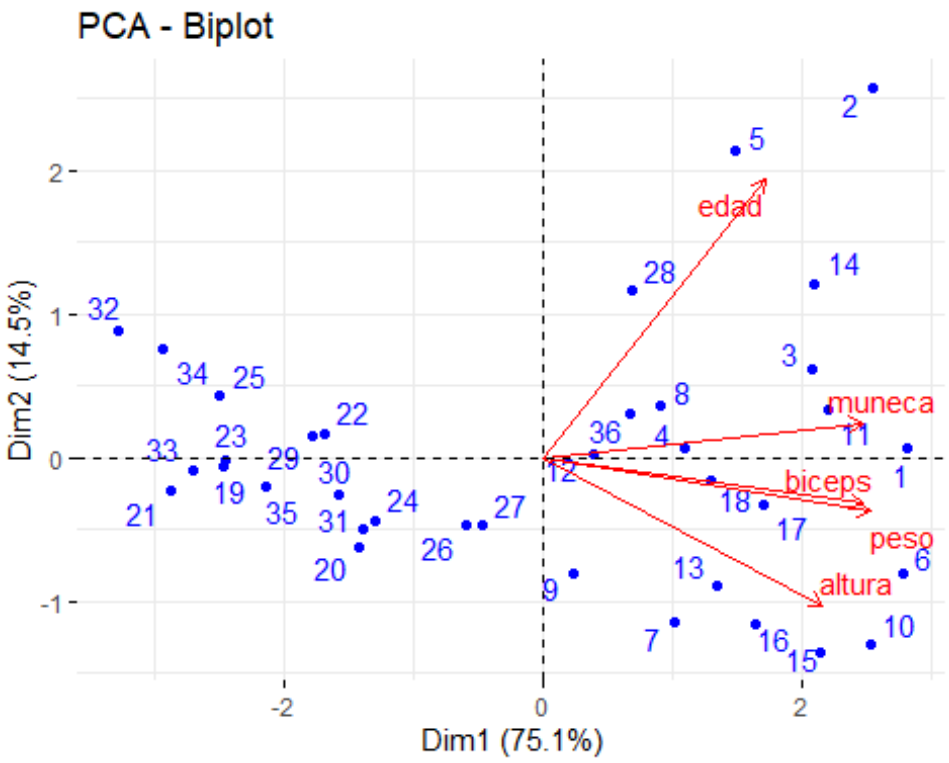

```
fviz_screeplot(cpS)
```




```
fviz_contrib(cpS, choice = "var")
```




```
fviz_pca_biplot(cpS, repel=TRUE, col.var="red", col.ind="blue")
```



Con estos gráficos podemos confirmar lo que hemos estado estipulando durante todo el análisis. Las gráficas de varianza-covarianza nos dicen que hay dos principales datos atipicos (2, 5), y que las variables que más peso tiene son las de peso, altura y un poco de edad. Teniendo la edad un poco más de relación entre los dos principales componentes que las dos anteriores. El peso y la altura parecen tener más relación con el primer componente principalmente.

En las gráficas de correlación, podemos encontrar los mismos dos valores atípicos (2, 5). Sin embargo, las gráficas de correlación parecen darnos a entender que hay tres variables que contribuyen por igual al primer componente (peso, biceps y altura). Además de que las flechas de las variables apuntan más hacia el primer componente prinicipalmente, a excepción de la edad, que cómo en las gráficas anteriores, nos da a entender que tiene relación con el componente dos también.

Un comando interesante de PCA, que nos podría ayudar es el de var, pues este nos regresa la lista de matrices con los valores de las variables activas.

Parte cuatro

- Compare los resultados obtenidos con la matriz de varianza-covarianza y con la correlación. ¿Qué concluye? ¿Cuál de los dos procedimientos aporta componentes con de mayor interés?

– Considero que el mejor procedimiento es el de correlación pues al tener todos los datos bajo una misma escala, provee un mejor análisis de los datos, ya que los valores de varias variables no se encuentran bajo el mismo rango. Esto lo podemos observar

en la gráfica de contribución de las variables a la dimension uno, dónde casi todas las variables aportan la misma cantidad.

- Indique cuál de los dos análisis (a partir de la matriz de varianza y covarianza o de correlación) resulta mejor para los datos indicadores económicos y sociales del 96 países en el mundo. Comparar los resultados y argumentar cuál es mejor según los resultados obtenidos.

– Considero que es mejor el análisis de correlación, pues al tener los datos escalados, las diferencias físicas regionales de cada país no debería de afectar a los resultados obtenidos.

- ¿Qué variables son las que más contribuyen a la primera y segunda componentes principales del método seleccionado? (observa los coeficientes en valor absoluto de las combinaciones lineales, auxíliate también de los gráficos)

– Las variables que más afectan al modelo son las de altura y edad.

- Escriba las combinaciones finales que se recomiendan para hacer el análisis de componentes principales.

– La mejor combinación sería las que tienen los componentes que más peso tienen. En este caso es el primero y el segundo, los cuales son contribuidos en gran parte por la altura, edad y peso.

- Interpreta los resultados en término de agrupación de variables (puede ayudar “índice de riqueza”, “índice de ruralidad”, etc)

– Las variables las podriamos agrupar de tal forma en que entren a un contexto de indice de riqueza y de pobreza. – Las variables que pueden entrar en el indice de riqueza son las de altura, bicep, peso y edad. – Las variables que pueden entrar en el indice de ruralidad son las de muñeca, y edad.