Actividad 4: Componentes Principales

Daniela Jiménez Téllez

2024-10-08

Problema

En la base de datos Corporal las medidas corporarles de 36 estudiantes de la universidad. Haz una análisis de Componentes principales con la matriz de varianzas-covarianzas y la matriz de correlaciones. Compara los resultados y argumenta cuál es mejor según los resultados obtenidos.

Primero se realiza un análisis descriptivo para conocer las variables. Incluye las medidas que vienen en el summary() y la desviación estándar. Describe las correlaciones que se establecen entre las variables.

Importación de librerías

```
library(stats)
library(FactoMineR)
library(ggplot2)
library(factoextra)
```

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

Importación de datos

```
datos <- read.csv("Corporal.csv")
```

Parte I

Realiza el análisis de los valores y vectores propios con la matriz de covarianzas y con la de correlación. Analiza la varianza explicada por cada componente en cada caso e interpreta dentro del contexto del problema.

```
summary(datos)
```

```
##       edad            peso           altura          sexo
##  Min.   :19.00   Min.   :42.00   Min.   :147.2   Length:36
##  1st Qu.:24.75   1st Qu.:54.95   1st Qu.:164.8   Class :character
##  Median :28.00   Median :71.50   Median :172.7   Mode  :character
##  Mean   :31.44   Mean   :68.95   Mean   :171.6
##  3rd Qu.:37.00   3rd Qu.:82.40   3rd Qu.:179.4
##  Max.   :65.00   Max.   :98.20   Max.   :190.5
##      muneca           biceps
##  Min.   : 8.300   Min.   :23.50
##  1st Qu.: 9.475   1st Qu.:25.98
##  Median :10.650   Median :32.15
##  Mean   :10.467   Mean   :31.17
##  3rd Qu.:11.500   3rd Qu.:35.05
##  Max.   :12.400   Max.   :40.40
```

```
sapply(datos, sd)
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introducidos por coerción
```

```
##      edad      peso    altura      sexo    muneca    biceps
## 10.554469 14.868999 10.520170        NA  1.175463  5.234392
```

La matriz de varianza-covarianza muestra cómo varían las variables con respecto a la media y entre sí mismas. Se usa en PCA cuando las variables están en la misma escala o cuando no es necesario estandarizar. Por otro lado, la matriz de correlación estandariza las variables. Esto sirve cuando las variables tienen diferentes escalas como peso y altura.

Igualmente, en este caso se usaron los eigenvalores y vectores para entender qué proporción de la varianza total del conjunto de datos es capturada por cada componente, y muestran los componentes.

1. Calcule las matrices de varianza-covarianza S con cov(X) y la matriz de correlaciones R con cor(X) y realice los siguientes pasos con cada una:

- Calcule los valores y vectores propios de cada matriz.La función en R es: eigen().
- Calcule la proporción de varianza explicada por cada componente en ambas matrices. Se sugiere dividir cada lambda entre la varianza total (las lambdas están en eigen(S)$values). La varianza total es la suma de las varianzas de la diagonal de S. Una forma es sum(diag(S)). La varianza total de los componentes es la suma de los valores propios (es decir, la suma de la varianza de cada componente), sin embargo, si sumas la diagonal de S (es decir, la varianza de cada x), te da el mismo valor (¡comprúebalo!). Recuerda que las combinaciones lineales buscan reproducir la varianza de X.
- Acumule los resultados anteriores (cumsum() puede servirle) para obtener la varianza acumulada en cada componente.

```
# Matriz de covarianzas
S <- cov(datos[, -which(names(datos) == "sexo")])
valores_propios_cov <- eigen(S)$values
vectores_propios_cov <- eigen(S)$vectors

# Matriz de correlaciones
R <- cor(datos[, -which(names(datos) == "sexo")])
valores_propios_cor <- eigen(R)$values
vectores_propios_cor <- eigen(R)$vectors

# Varianza total (covarianza)
varianza_total_cov <- sum(diag(S))

# Varianza explicada (covarianza)
varianza_explicada_cov <- valores_propios_cov / varianza_total_cov

# Varianza total (correlación)
varianza_total_cor <- sum(diag(R))

# Varianza explicada (correlación)
varianza_explicada_cor <- valores_propios_cor / varianza_total_cor

# Varianza acumulada
varianza_acumulada_cov <- cumsum(varianza_explicada_cov)
varianza_acumulada_cor <- cumsum(varianza_explicada_cor)

print("Varianza explicada por los componentes (Covarianza):")
```

```
## [1] "Varianza explicada por los componentes (Covarianza):"
```

```
print(varianza_explicada_cov)
```

```
## [1] 0.7615357176 0.1703098726 0.0585307219 0.0091271040 0.0004965839
```

```
print("Varianza acumulada (Covarianza):")
```

```
## [1] "Varianza acumulada (Covarianza):"
```

```
print(varianza_acumulada_cov)
```

```
## [1] 0.7615357 0.9318456 0.9903763 0.9995034 1.0000000
```

```
print("Varianza explicada por los componentes (Correlación):")
```

```
## [1] "Varianza explicada por los componentes (Correlación):"
```

```
print(varianza_explicada_cor)
```

```
## [1] 0.75149947 0.14517133 0.06406596 0.02492375 0.01433950
```

```
print("Varianza acumulada (Correlación):")
```

```
## [1] "Varianza acumulada (Correlación):"
```

```
print(varianza_acumulada_cor)
```

```
## [1] 0.7514995 0.8966708 0.9607368 0.9856605 1.0000000
```

- Según los resultados anteriores, ¿qué componentes son los más importantes?

Los dos primeros componentes explican una cantidad importante de la varianza gracias a sus altos valores. Esto nos dice que son los más importantes en la reducción de la dimensionalidad de los datos.

- Escriba la ecuación de la combinación lineal de los Componentes principales CP1 y CP2 (eiX, donde ei está en eigen(S)$vectors[1], e2X para obtener CP2, donde X = c(X1, X2, …)) ¿qué variables son las que más contribuyen a la primera y segunda componentes principales? (observe los coeficientes en valor absoluto de las combinaciones lineales). Justifique su respuesta.

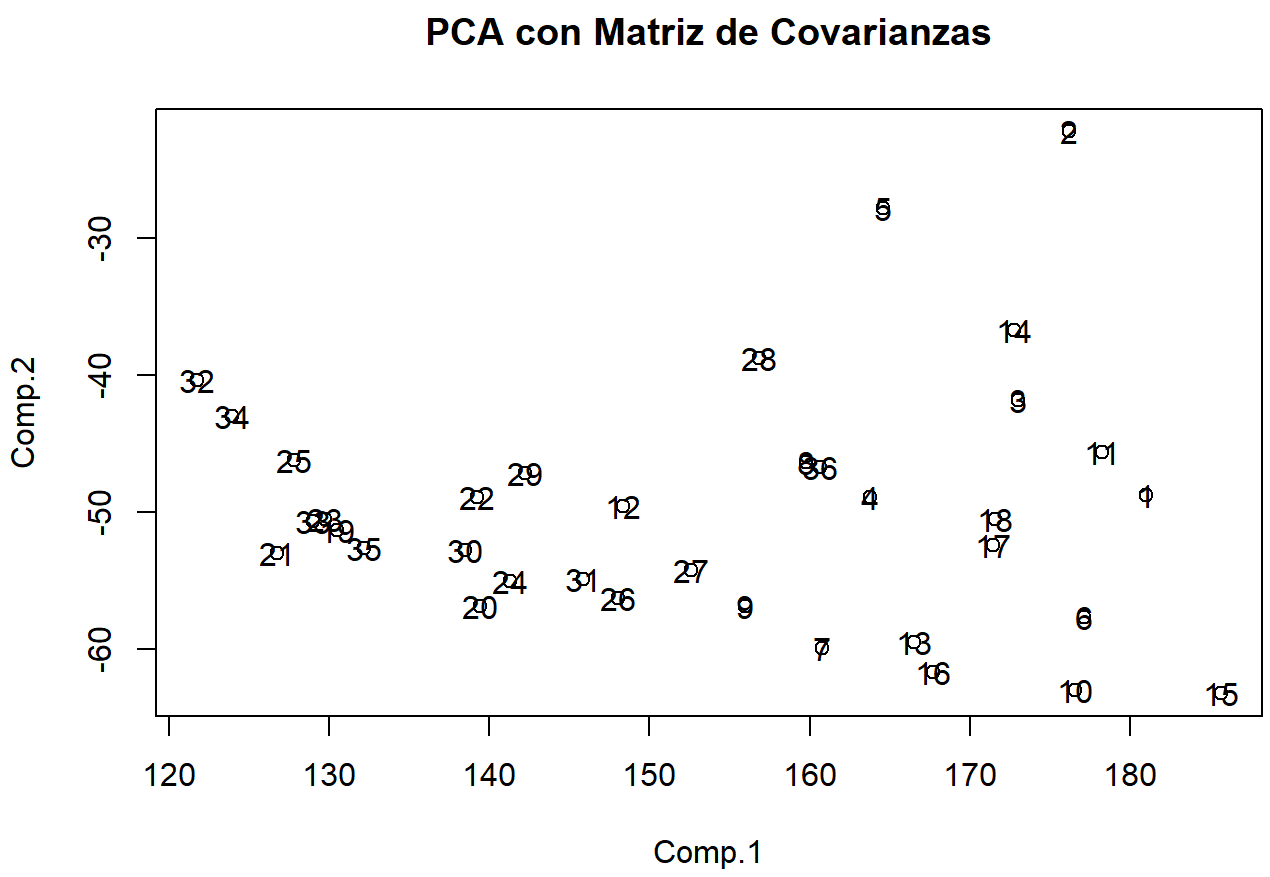
2. ¡No te olvides de seguir los mismos pasos con la matriz de correlaciones (se obtiene con cor(x) si x está compuesto por variables numéricas)

Parte II

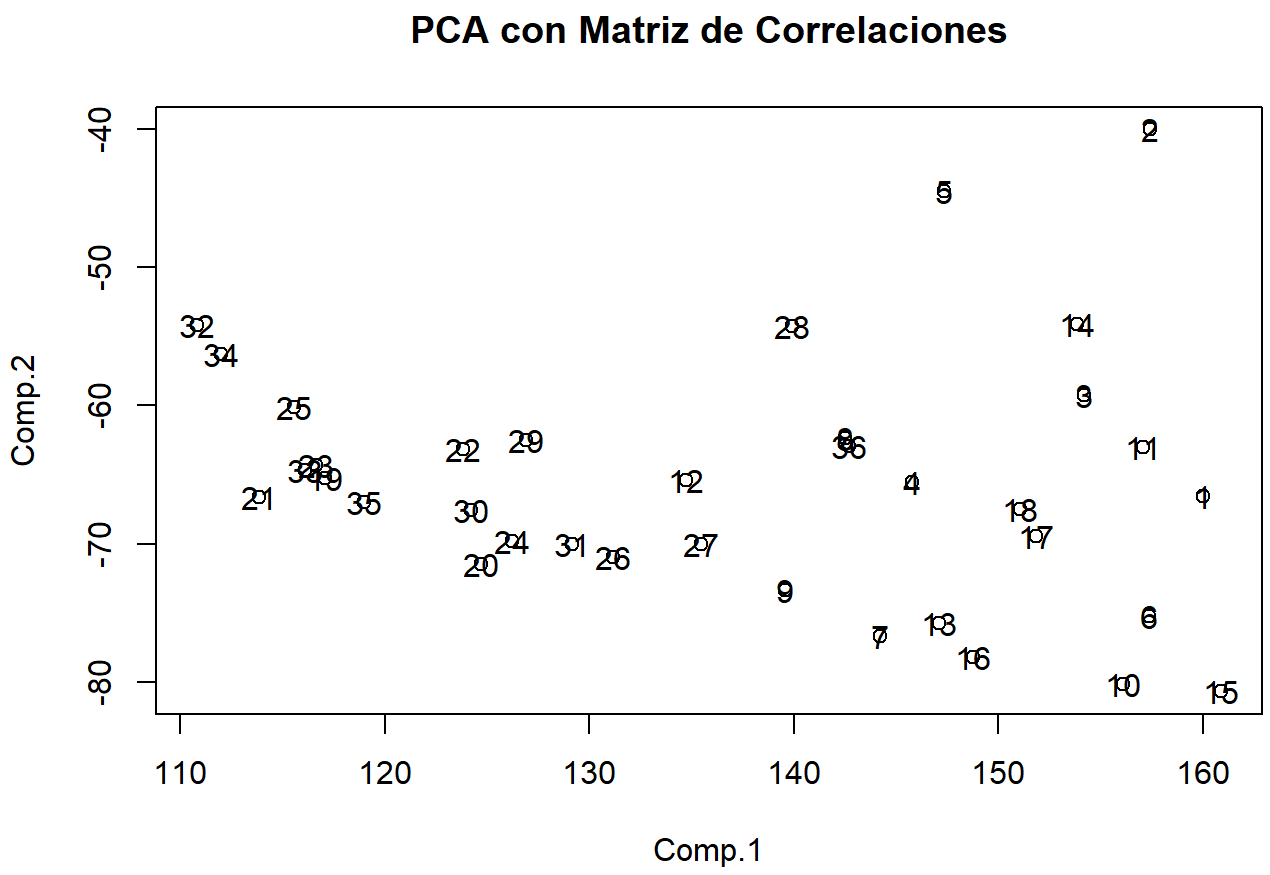
1. Obtenga las gráficas respectivas con S (matriz de varianzas-covarianzas) y con R (matriz de correlaciones) de las dos primeras componentes.

- Calcule las puntuaciones (scores) de las observaciones para los componentes obtenidos con la matriz de varianzas-covarianzas.
- Calcule las puntuaciones (scores) de las observaciones para los componentes obtenidos con la matriz de correlaciones. Recuerde que en la matriz de correlaciones las variables tienen que estar estandarizadas.

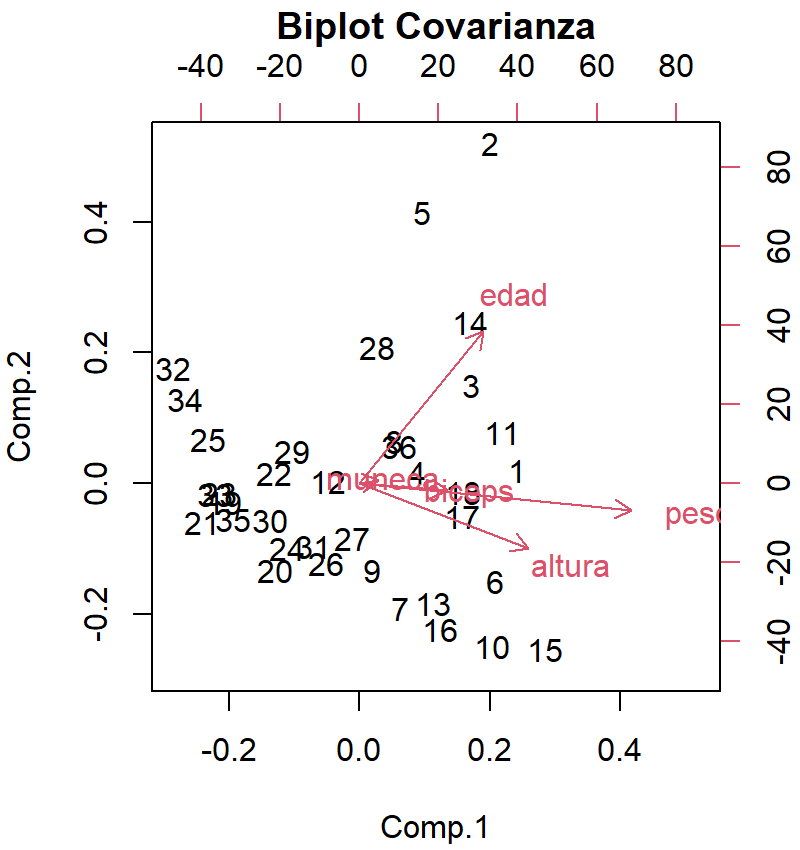
```
# Análisis de Componentes Principales (PCA) con covarianza
cp_cov <- princomp(datos[, -which(names(datos) == "sexo")], cor = FALSE)
puntuaciones_cov <- as.matrix(datos[, -which(names(datos) == "sexo")]) %*% cp_cov$loadings

# Gráfico PCA con covarianza
plot(puntuaciones_cov[,1:2], type = "p", main = "PCA con Matriz de Covarianzas")
text(puntuaciones_cov[,1], puntuaciones_cov[,2], labels=1:nrow(datos))
```



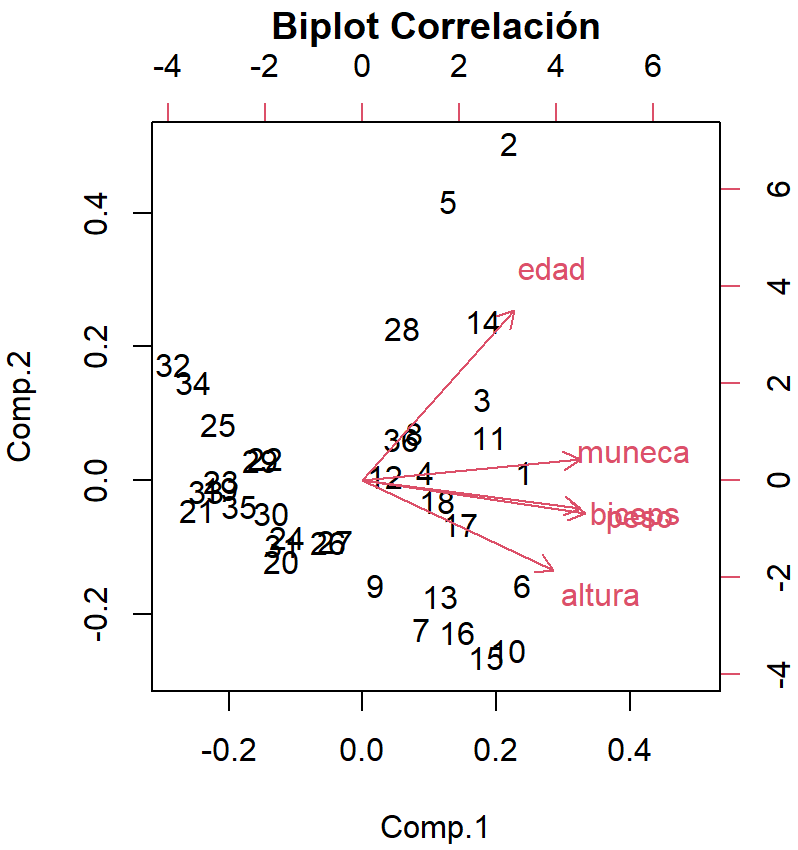

```
# Análisis de Componentes Principales (PCA) con correlación
cp_cor <- princomp(datos[, -which(names(datos) == "sexo")], cor = TRUE)
puntuaciones_cor <- as.matrix(datos[, -which(names(datos) == "sexo")]) %*% cp_cor$loadings

# Gráfico PCA con correlación
plot(puntuaciones_cor[,1:2], type = "p", main = "PCA con Matriz de Correlaciones")
text(puntuaciones_cor[,1], puntuaciones_cor[,2], labels=1:nrow(datos))
```




```
# Biplots
biplot(cp_cov, main = "Biplot Covarianza")
```




```
biplot(cp_cor, main = "Biplot Correlación")
```



2. Interprete los gráficos en términos de:

- Las relaciones que se establecen entre las variables y los componentes principales.
- La relación entre las puntuaciones de las observaciones y los valores de las variables.
- Detecte posibles datos atípicos

3. Explora el: princomp() en library(stats). Puedes poner help(princomp) en la consola o buscarlo en la ventana de ayuda. Indaga: ¿qué otras opciones tiene para facilitarte el análisis? En particular, explora los comandos y subcomandos: summary(cpS), cpaS$loading$, $cpaS$scores. ¿Cómo se interpreta el resultado?

Con comandos como summary(), loadings, y scores, se puede interpretar la importancia de los componentes principales y cómo contribuyen las variables en el espacio de los componentes principales. Estos comandos pueden ayudar a generar conclusiones sobre la manera en la que están creados en los datos o su origen, y a tomar decisiones basadas en la varianza.

Parte III

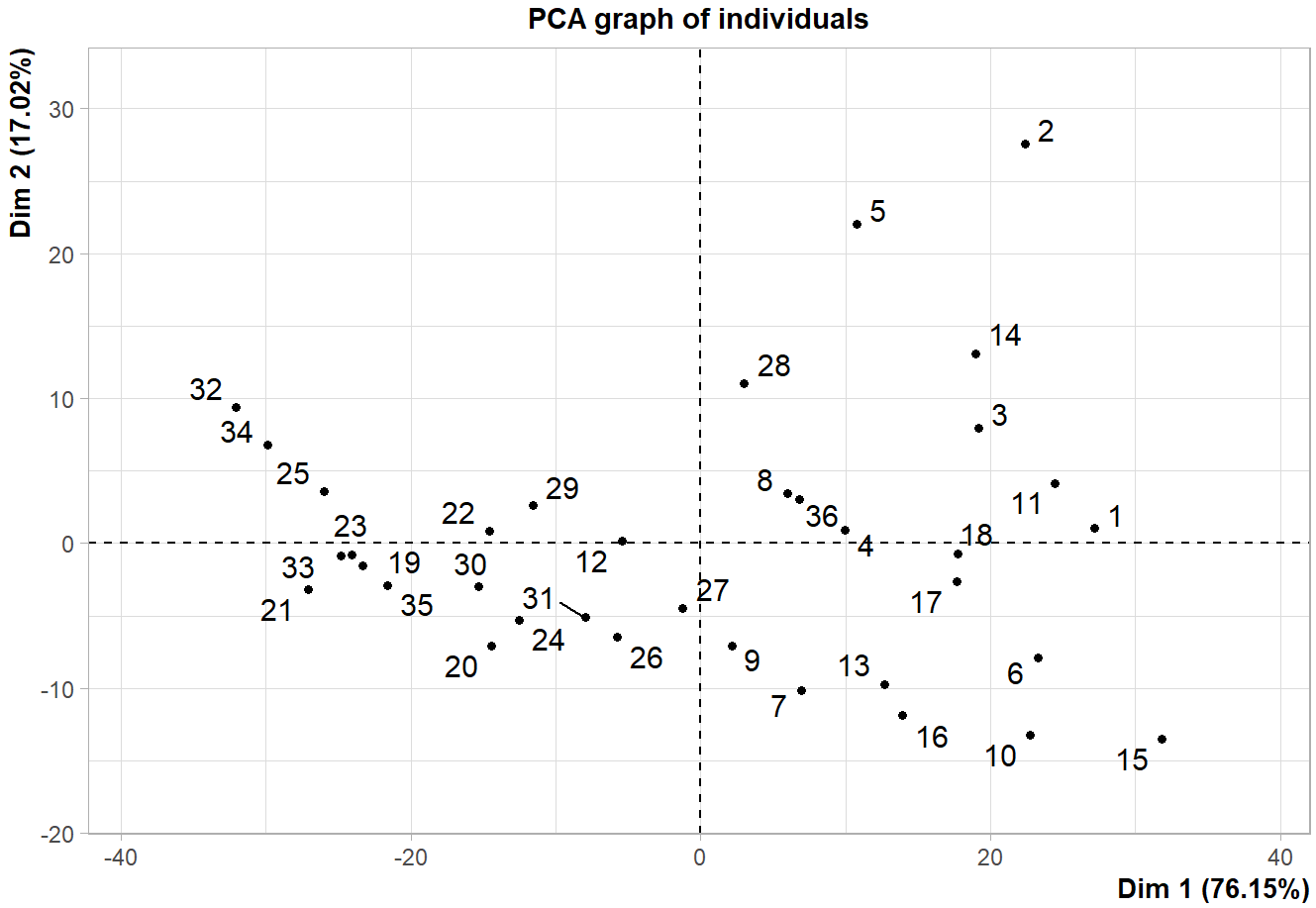
1. Explore los siguientes gráficos relativos a Componentes Principales.

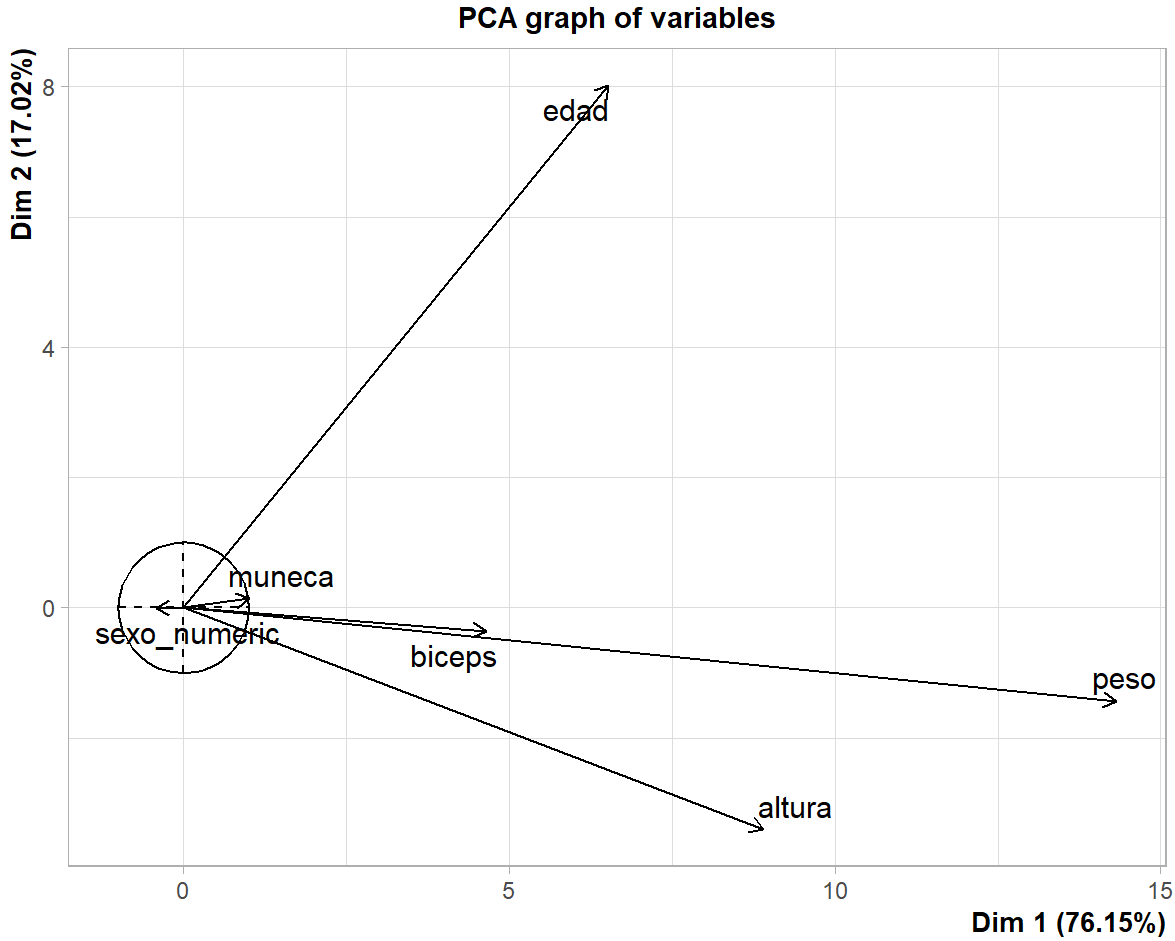
2. Interprete cada gráfico e identifica qué es lo que se está graficando en cada uno. Realiza el análisis con la matriz de varianzas y covarianzas y correlación.

library(FactoMineR) library(ggplot2) datos=matriz de datos cpS = PCA(datos,scale.unit=FALSE) #Para matriz de correlaciones usa scale.unit=TRUE 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)

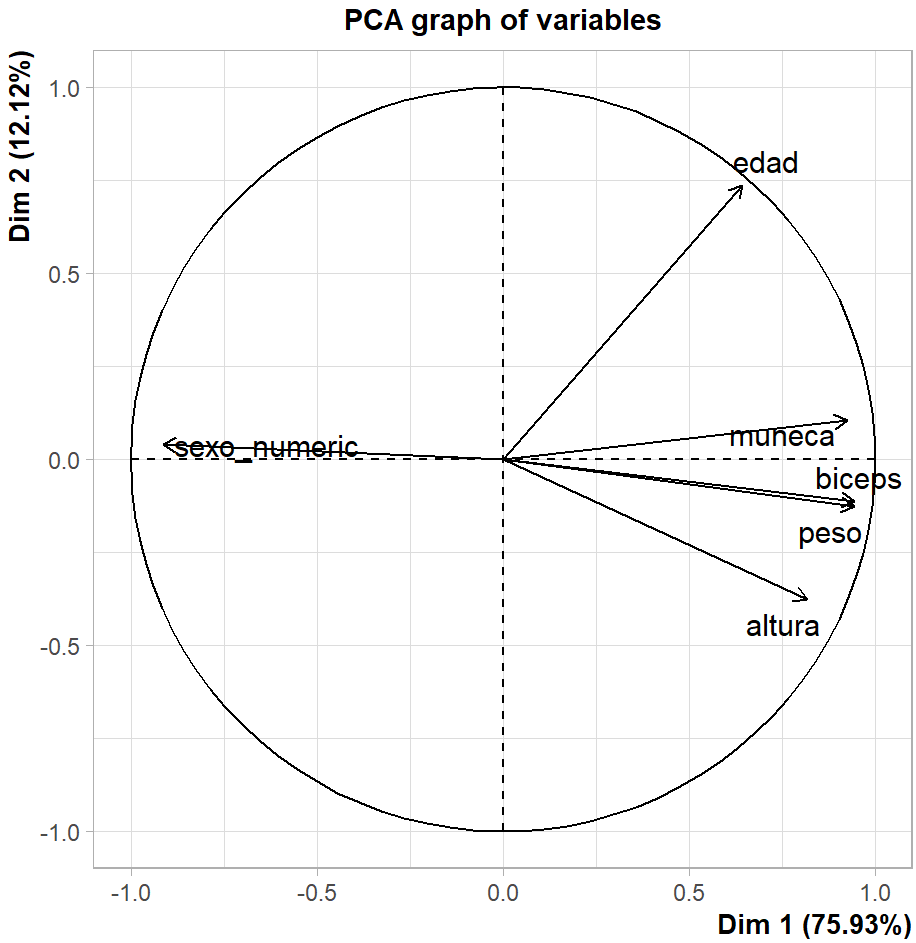
fviz_contrib(cpS, choice = c(“var”) fviz_pca_biplot(cpS, repel=TRUE, col.var=“red”, col.ind=“blue”)

```
datos$sexo <- as.factor(datos$sexo)
datos$sexo_numeric <- as.numeric(datos$sexo)

cpS_var_cov <- PCA(datos[, -which(names(datos) == "sexo")], scale.unit = FALSE)
```

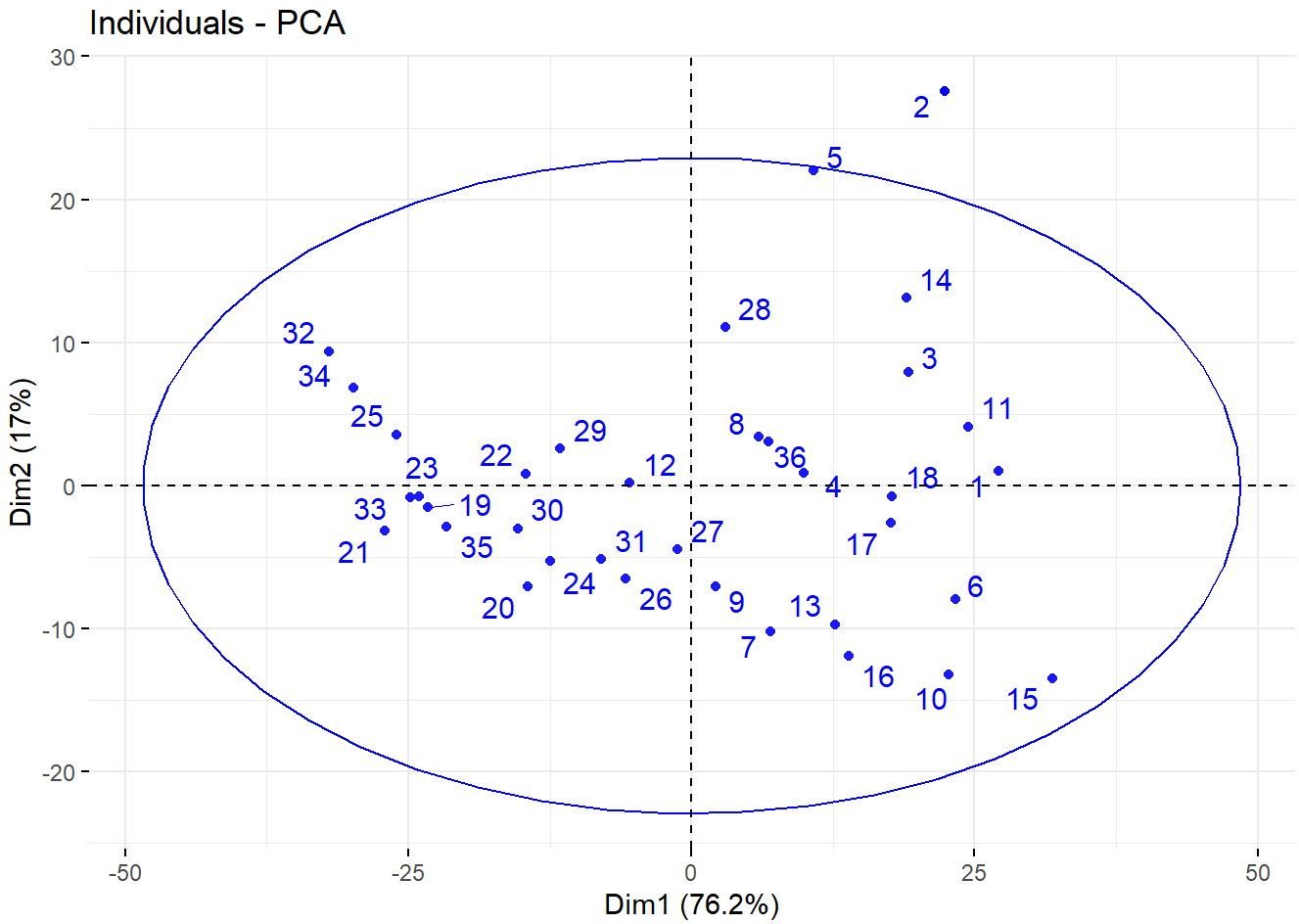

PCA graph of individuals
Dim 2 (17.02%)
Dim 1 (76.15%)
30
20
10
0
-10
-20
-40
-20
0
20
40
2
5
14
28
32
3
34
25
8
29
11
22
36
1
23
18
4
33
19
30
12
27
17
21
35
31
24
26
20
9
13
6
7
16
10
15


PCA graph of variables
Dim 2 (17.02%)
Dim 1 (76.15%)
8
4
0
0
5
10
15
edad
muneca
sexo_numeric
biceps
peso
altura

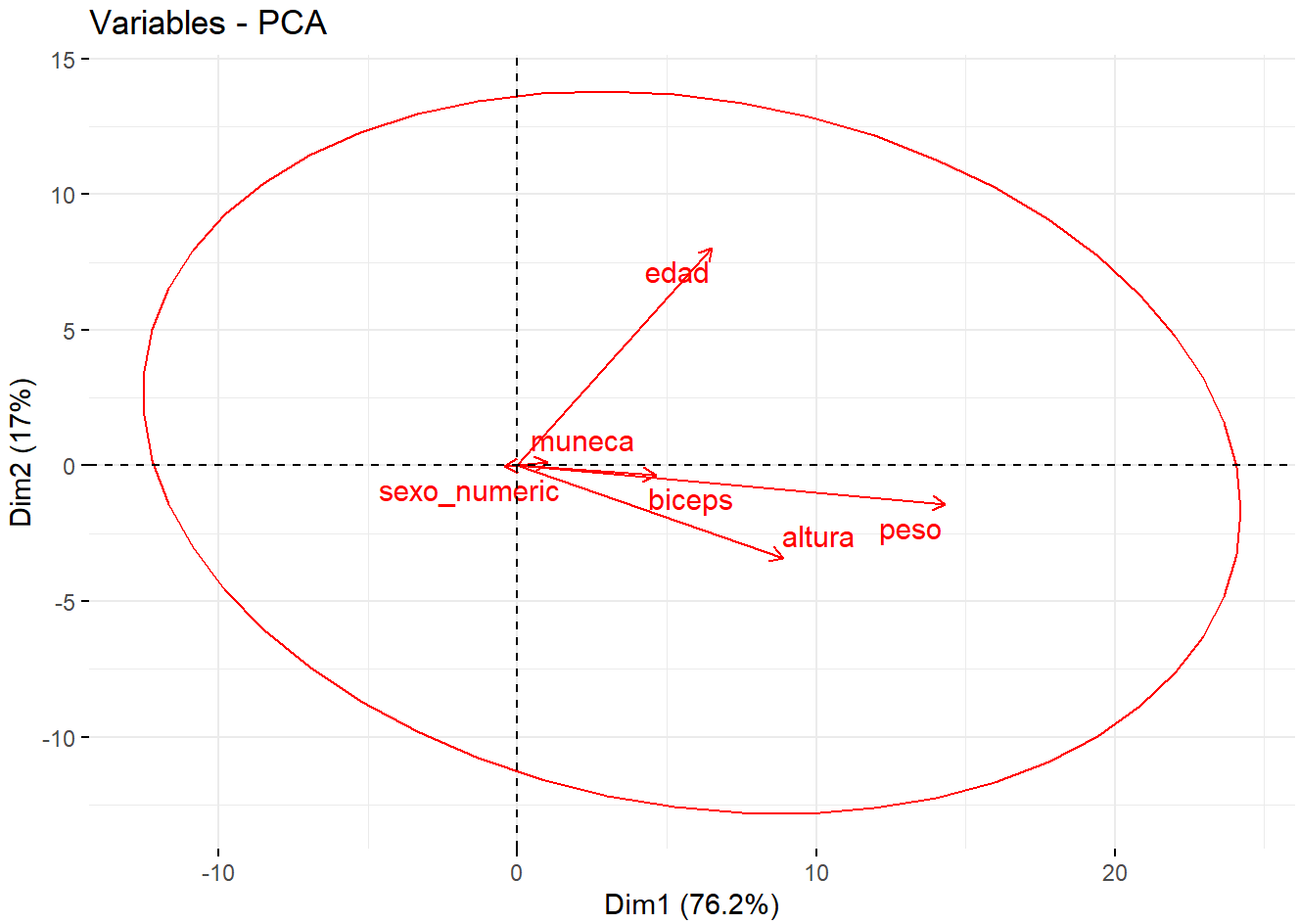
```
cpS_cor <- PCA(datos[, -which(names(datos) == "sexo")], scale.unit = TRUE)
```



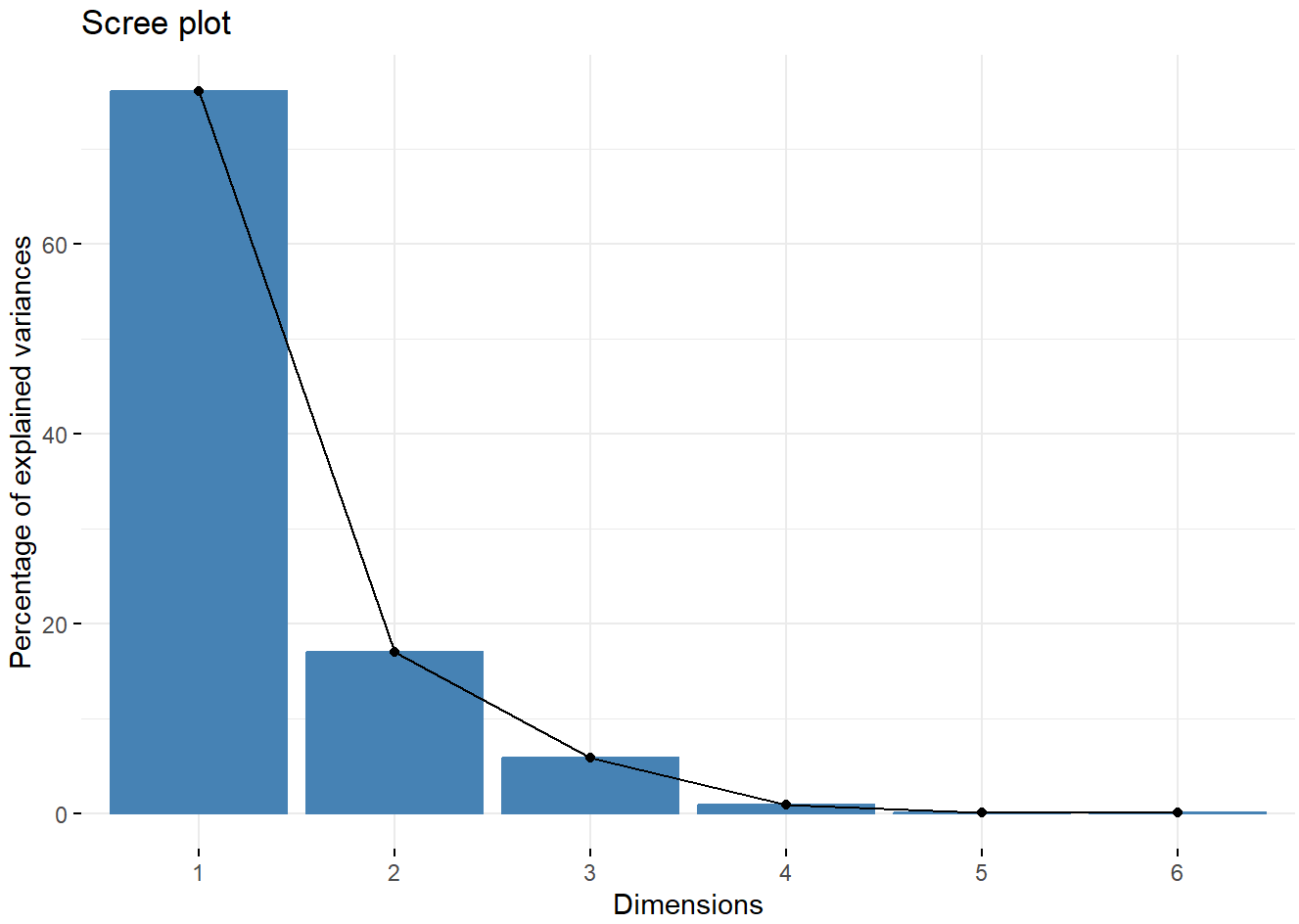

```
fviz_pca_ind(cpS_var_cov, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```



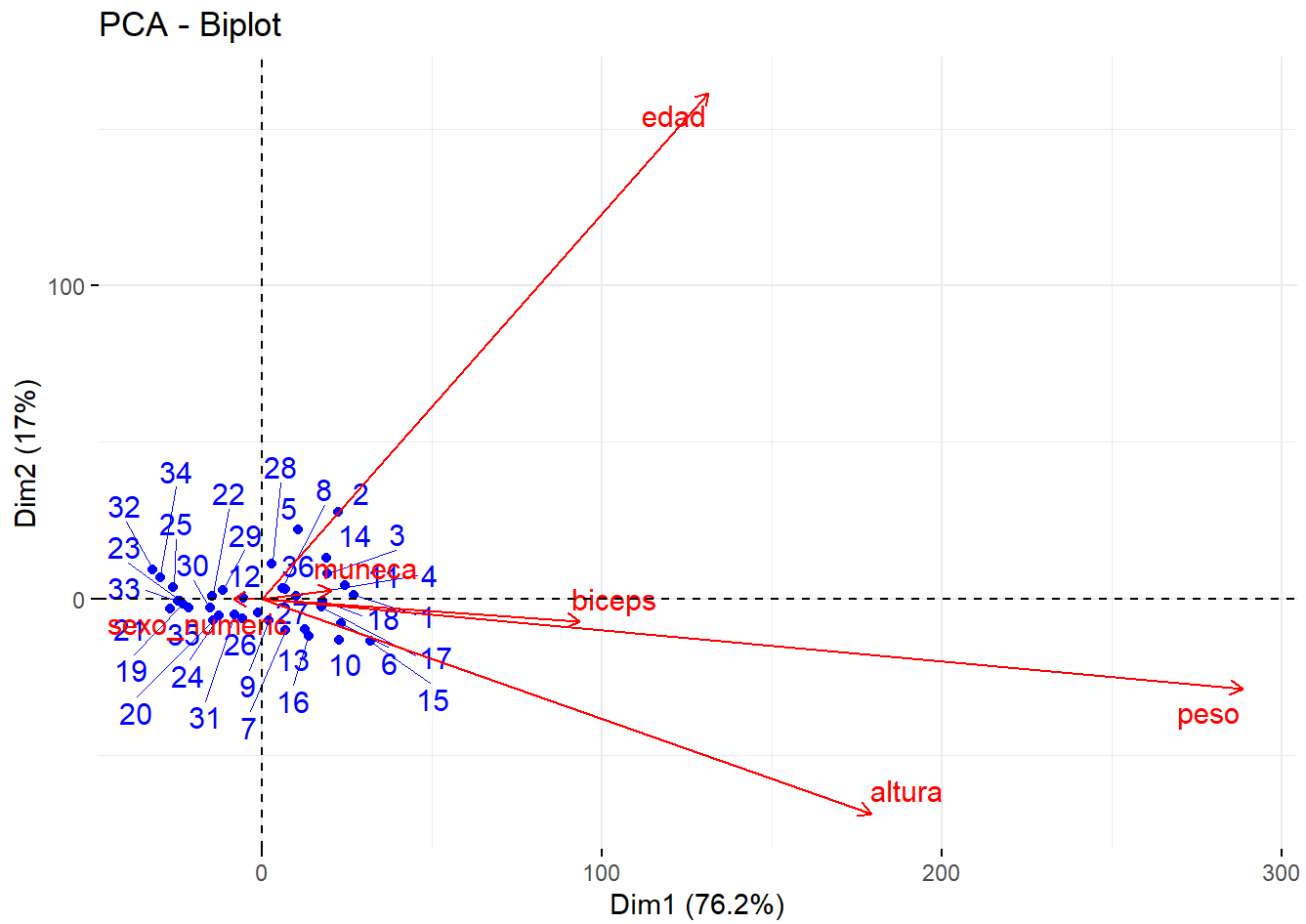

```
fviz_pca_var(cpS_var_cov, col.var = "red", addEllipses = TRUE, repel = TRUE)
```



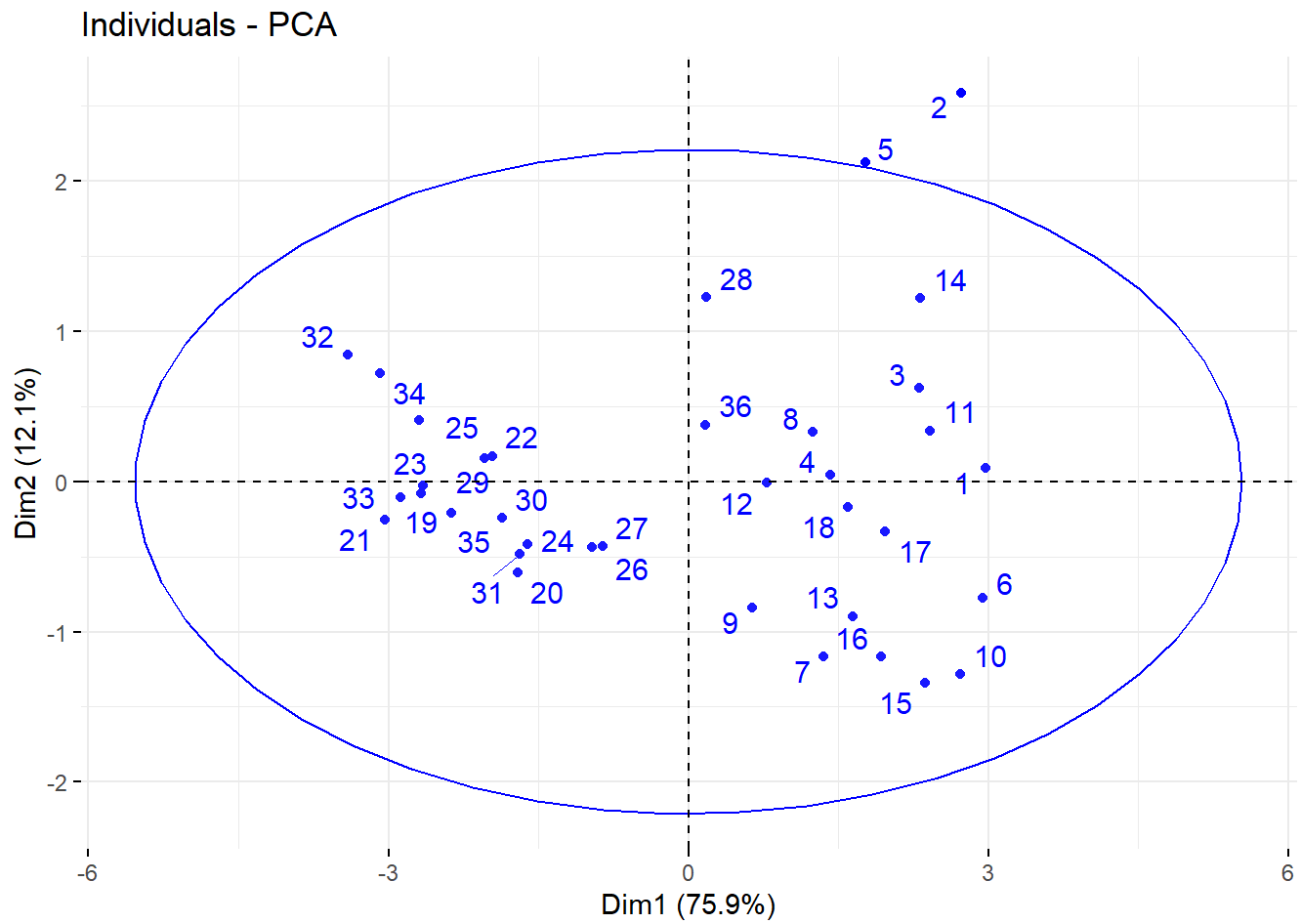

```
fviz_screeplot(cpS_var_cov)
```



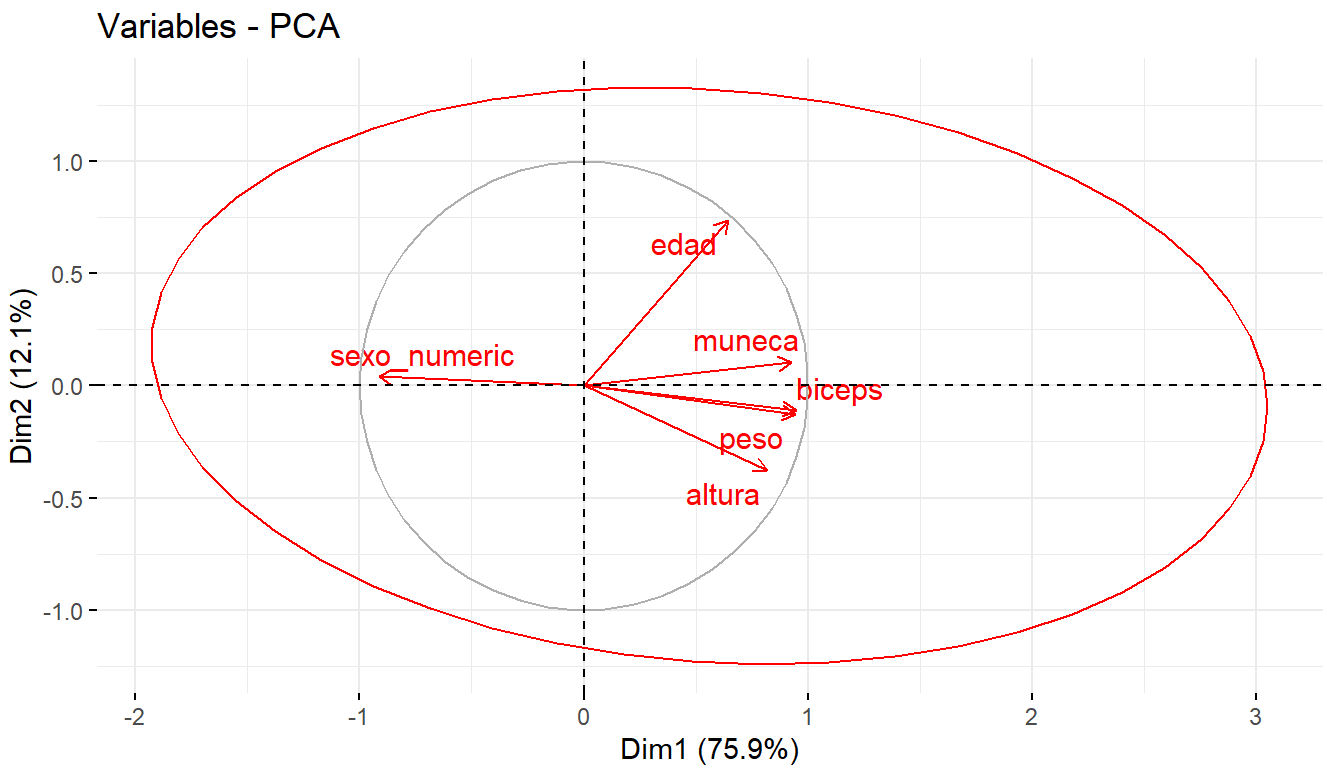

```
fviz_pca_biplot(cpS_var_cov, repel = TRUE, col.var = "red", col.ind = "blue")
```



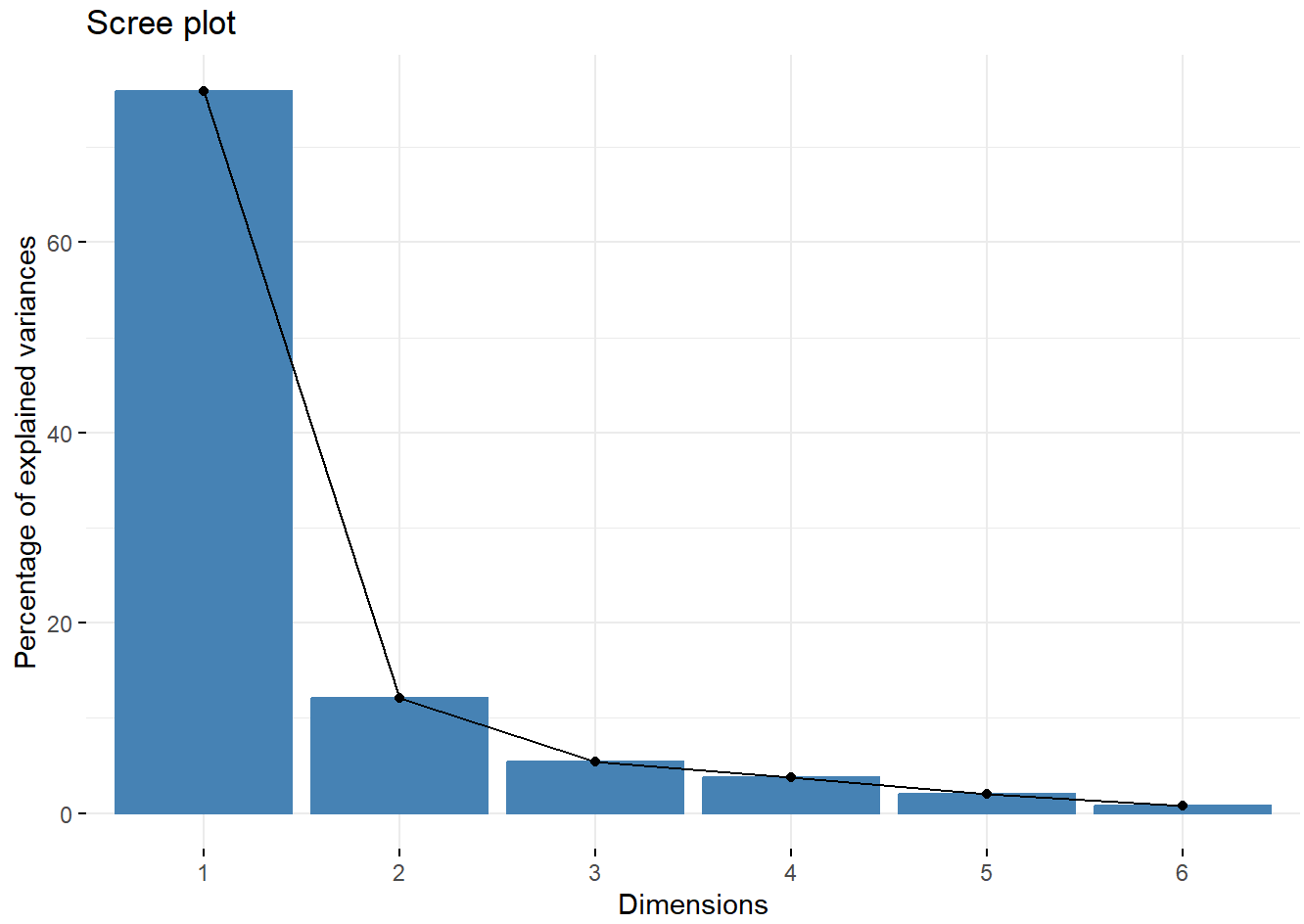

```
fviz_pca_ind(cpS_cor, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```



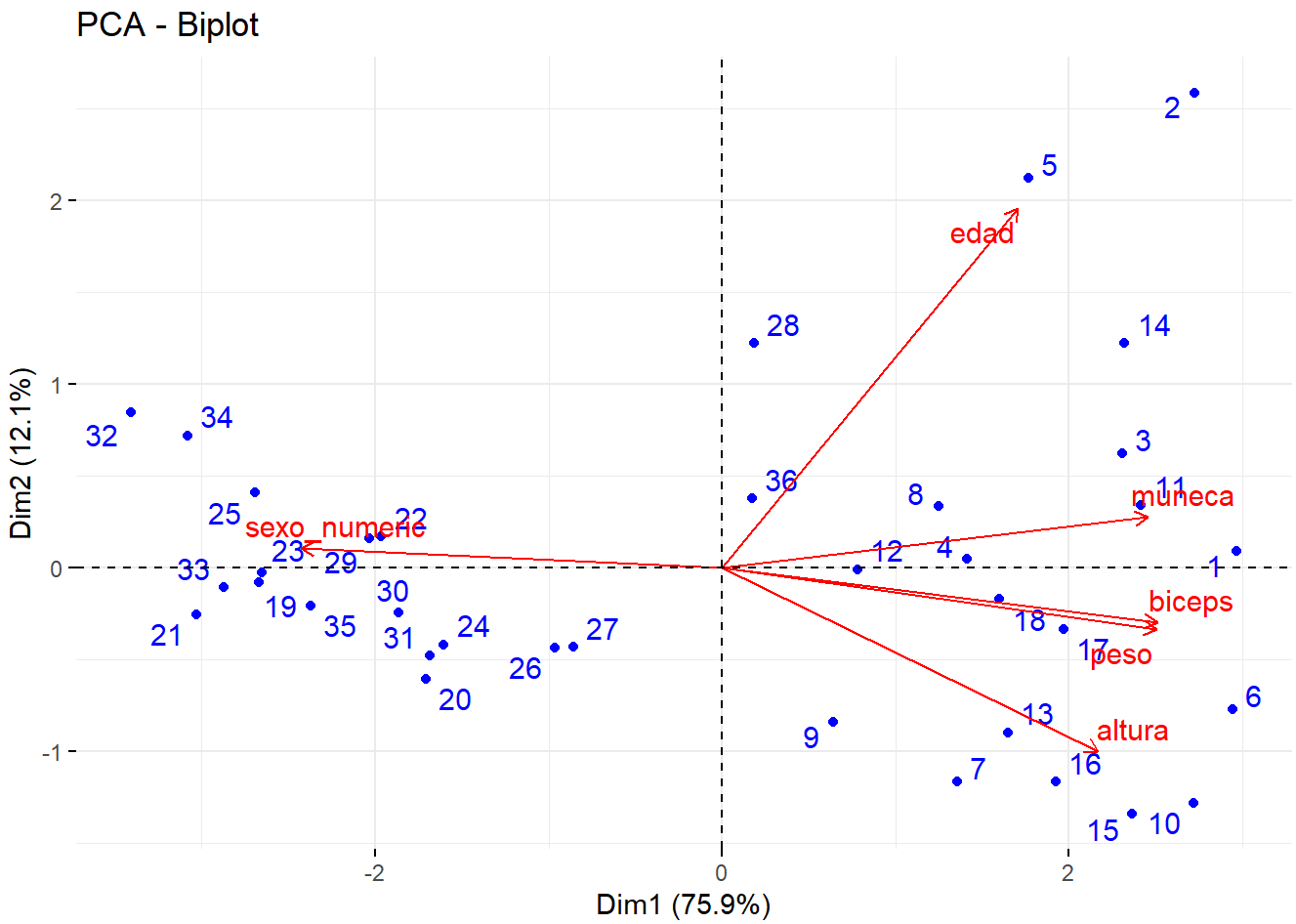

```
fviz_pca_var(cpS_cor, col.var = "red", addEllipses = TRUE, repel = TRUE)
```




```
fviz_screeplot(cpS_cor)
```




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



3. Explora el comando PCA, (puedes poner help(PCA) en la consola o buscarlo en la ventana de ayuda) ¿qué otras opciones tiene para facilitarte el análisis?

El comando PCA mejora el análisis de componentes principales. Hay varias opciones como la estandarización de las variables (scale.unit), el control sobre cuántos componentes calcular (ncp) que permiten un análisis flexible y adaptado a las necesidades específicas de los datos con los que trabajo.