# A3-Componentes Principales

Esteban Sierra

2025-10-14

# Parte 0: Análisis Descriptivo

```
D = read.csv("C:/Users/Pansocrates03/Documents/7mo Semestre/ESTADISTICA/corporal.csv")
```

```
summary(D)
```

```
##       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
```

Correlación en las variables numéricas

```
D_numericos = D[, sapply(D, is.numeric)]
cor(D_numericos)
```

```
##             edad      peso    altura    muneca    biceps
## edad   1.0000000 0.5153847 0.3302211 0.6204942 0.4836702
## peso   0.5153847 1.0000000 0.7973737 0.8493361 0.9088813
## altura 0.3302211 0.7973737 1.0000000 0.6595849 0.7086144
## muneca 0.6204942 0.8493361 0.6595849 1.0000000 0.8777369
## biceps 0.4836702 0.9088813 0.7086144 0.8777369 1.0000000
```

# PARTE 1

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.

# 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:

```
#Matriz de varianza covarianza
cat("Covarianza de las variables\n")
```

```
## Covarianza de las variables
```

```
covarianza = cov(D_numericos)
covarianza
```

```
##              edad      peso     altura    muneca    biceps
## edad   111.396825  80.88159  36.666032  7.698095 26.720952
## peso    80.881587 221.08713 124.728698 14.844667 70.738381
## altura  36.666032 124.72870 110.673968  8.156476 39.021048
## muneca   7.698095  14.84467   8.156476  1.381714  5.400571
## biceps  26.720952  70.73838  39.021048  5.400571 27.398857
```

```
cat("\nCorrelación entre las variables\n")
```

```
##
## Correlación entre las variables
```

```
correlacion = cor(D_numericos)
correlacion
```

```
##             edad      peso    altura    muneca    biceps
## edad   1.0000000 0.5153847 0.3302211 0.6204942 0.4836702
## peso   0.5153847 1.0000000 0.7973737 0.8493361 0.9088813
## altura 0.3302211 0.7973737 1.0000000 0.6595849 0.7086144
## muneca 0.6204942 0.8493361 0.6595849 1.0000000 0.8777369
## biceps 0.4836702 0.9088813 0.7086144 0.8777369 1.0000000
```

## 1.1 Valores y Vectores de cada matriz.

```
cat("Valores y vectores propios de la matriz de covarianza\n")
```

```
## Valores y vectores propios de la matriz de covarianza
```

```
vectores_covarianza = eigen(covarianza)
vectores_covarianza
```

```
## eigen() decomposition
## $values
## [1] 359.3980243  80.3757858  27.6229011   4.3074318   0.2343571
##
## $vectors
##             [,1]       [,2]        [,3]         [,4]         [,5]
## [1,] -0.34871002  0.9075501 -0.23248825 -0.001589466  0.026473941
## [2,] -0.76617586 -0.1616581  0.52166894 -0.338508602  0.010707863
## [3,] -0.47632405 -0.3851755 -0.78905759  0.046160807  0.003543154
## [4,] -0.05386189  0.0155423  0.02785902  0.126103480 -0.990039959
## [5,] -0.24817367 -0.0402221  0.22455005  0.931330496  0.137814357
```

```
cat("\nValores y vectores de la matriz de correlación\n")
```

```
##
## Valores y vectores de la matriz de correlación
```

```
vectores_correlacion = eigen(correlacion)
vectores_correlacion
```

```
## eigen() decomposition
## $values
## [1] 3.75749733 0.72585665 0.32032981 0.12461873 0.07169749
##
## $vectors
##            [,1]       [,2]        [,3]       [,4]       [,5]
## [1,] -0.3359310  0.8575601 -0.34913780 -0.1360111  0.1065123
## [2,] -0.4927066 -0.1647821  0.06924561 -0.5249533 -0.6706087
## [3,] -0.4222426 -0.4542223 -0.73394453  0.2070673  0.1839617
## [4,] -0.4821923  0.1082775  0.36690716  0.7551547 -0.2255818
## [5,] -0.4833139 -0.1392684  0.44722747 -0.3046138  0.6739511
```

## 1.2 Proporción de varianza

*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.*

```
cat("Proporción de varianza de la covarianza\n")
```

```
## Proporción de varianza de la covarianza
```

```
lambda_covarianza = c(59.3980243,80.3757858 ,27.6229011,4.3074318,0.2343571)

varianza_total_covarianza = sum(lambda_covarianza)

proporcion_de_varianza_covarianza = lambda_covarianza / varianza_total_covarianza
proporcion_de_varianza_covarianza
```

```
## [1] 0.345460873 0.467468227 0.160655706 0.025052166 0.001363029
```

```
cat("\nProporción de varianza de la correlación\n")
```

```
##
## Proporción de varianza de la correlación
```

```
lambda_correlacion = c(3.75749733, 0.72585665, 0.32032981, 0.12461873, 0.07169749)
varianza_total_correlacion = sum(lambda_correlacion)
proporcion_de_varianza_correlacion = lambda_correlacion / varianza_total_correlacion
proporcion_de_varianza_correlacion
```

```
## [1] 0.75149946 0.14517133 0.06406596 0.02492375 0.01433950
```

## 1.3 Acumule los resultados anteriores (cumsum() puede servirle) para obtener la varianza acumulada en cada componente.

```
cat("Varianza acumulada de covarianza\n")
```

```
## Varianza acumulada de covarianza
```

```
cumsum(proporcion_de_varianza_covarianza)
```

```
## [1] 0.3454609 0.8129291 0.9735848 0.9986370 1.0000000
```

```
cat("\nVarianza acumulada de correlación\n")
```

```
##
## Varianza acumulada de correlación
```

```
cumsum(proporcion_de_varianza_correlacion)
```

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

## 1.4 Según los resultados anteriores, ¿qué componentes son los más importantes?*

Tomando en cuenta los resultados anteriores, podemos concluir que los componentes más importantes respecto a la proporción de varianza de la covarianza es el segundo valor, ya que tiene un valor de 0.46.

Así mismo, el componente más importante según la varianza de correlación es el primer componente, ya que tiene una proporción de 75%.

Los dos primeros componentes principales (CP1 y CP2) son los más importantes, ya que logran explicar casi el 90% de la variabilidad de tus datos.

### 1.5 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.

¡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)

Ecuación para CP1:

$$CP1 = (-0.3487)Edad + (-0.7662)Peso + (-0.4763)Altura + (-0.0539)Muneca + (-0.2482)Biceps$$

Ecuación para CP2:

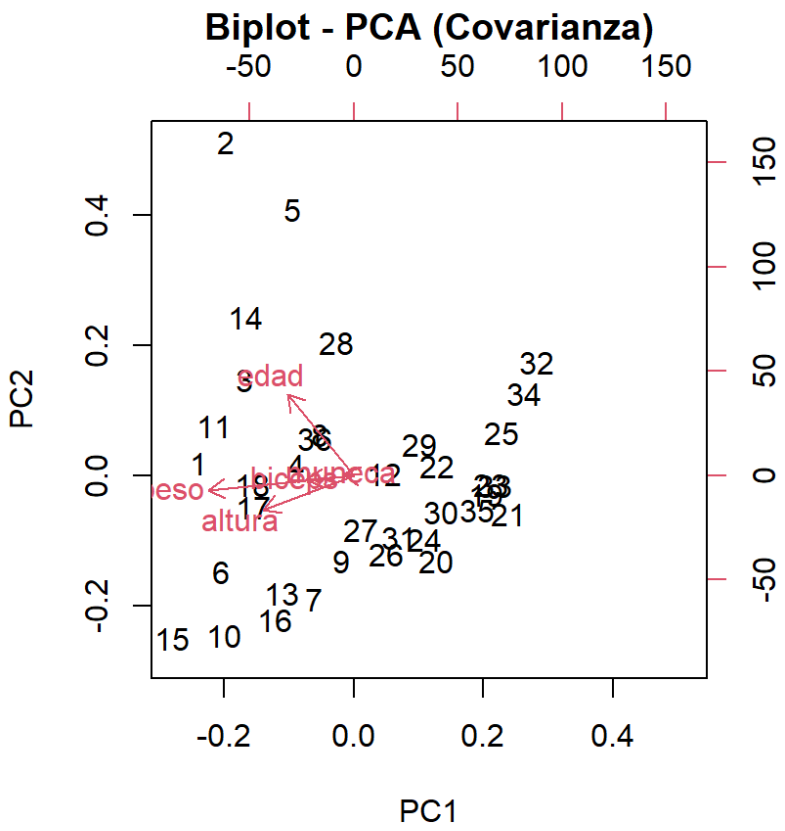
$$CP2 = (0.9076)Edad + (-0.1617)Peso + (-0.3852)Altura + (0.0155)Muneca + (-0.0402)Biceps$$

# Parte 2

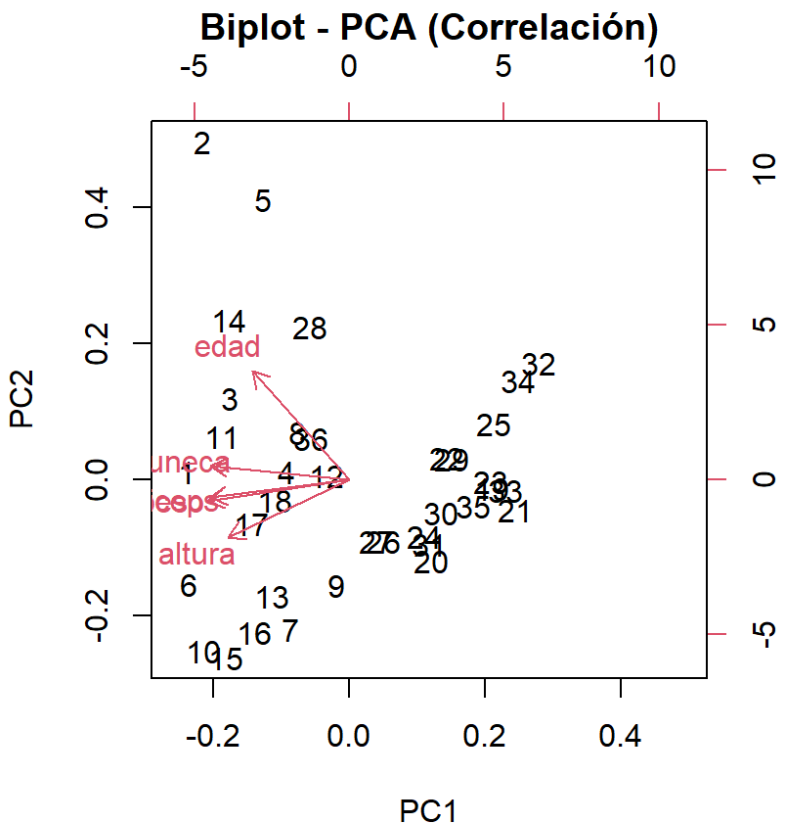
## 1. Obtenga las gráficas respectivas con S (matriz de varianzas-covarianzas) y con R (matriz de correlaciones) de las dos primeras componentes.

```
# PCA con la matriz de covarianza (S)
pca_cov <- prcomp(D_numericos, scale = FALSE)

biplot(pca_cov, choices = 1:2, main = "Biplot - PCA (Covarianza)")
```




```
# PCA con la matriz de correlacion
pca_cor <- prcomp(D_numericos, scale = TRUE)
biplot(pca_cor, choices = 1:2, main = "Biplot - PCA (Correlación)")
```




1.1 Calcule las puntuaciones (scores) de las observaciones para los componentes obtenidos con la matriz de varianzas-covarianzas

```
cat("scores de la matriz de covarianzas\n")
```

```
## scores de la matriz de covarianzas
```

```
summary(pca_cov)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4    PC5
## Standard deviation     18.9578 8.9653 5.25575 2.07544 0.4841
## Proportion of Variance  0.7615 0.1703 0.05853 0.00913 0.0005
## Cumulative Proportion   0.7615 0.9318 0.99038 0.99950 1.0000
```

1.2 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.

```
cat("\nScores de la matriz de correlaciones\n")
```

```
##
## Scores de la matriz de correlaciones
```

```
summary(pca_cor)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5
## Standard deviation     1.9384 0.8520 0.56598 0.35301 0.26776
## Proportion of Variance 0.7515 0.1452 0.06407 0.02492 0.01434
## Cumulative Proportion  0.7515 0.8967 0.96074 0.98566 1.00000
```

## 2. Interprete los gráficos en términos de:

2.1 Las relaciones que se establecen entre las variables y los componentes principales

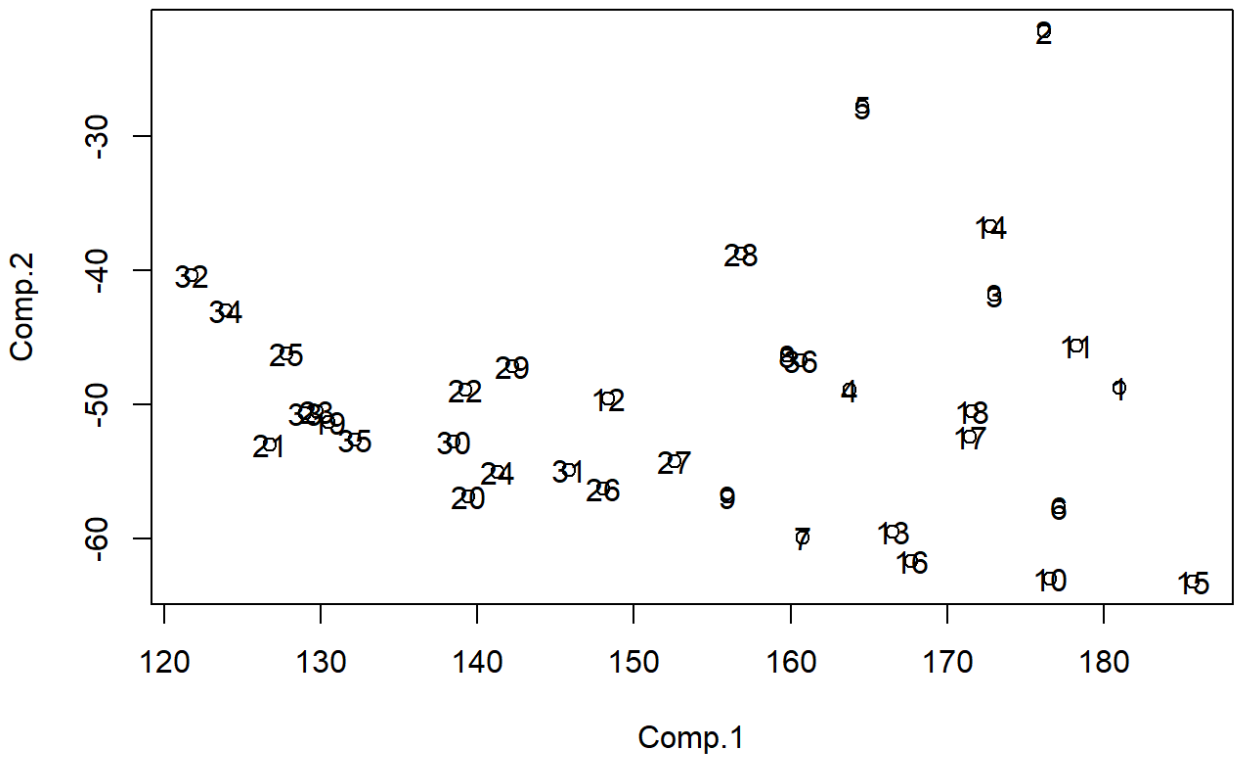
2.2 La relación entre las puntuaciones de las observaciones y los valores de las variables

2.3 Detecte posibles datos atípicos

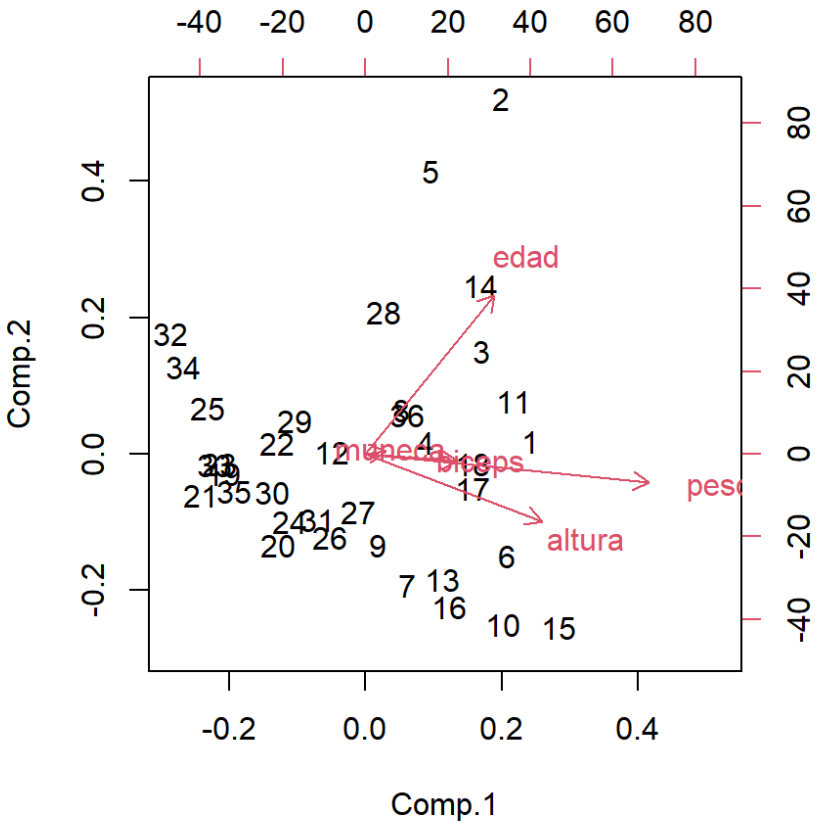
## 3. Explora el: princomp() en library(stats). Puedes poner help(princomp) en la consola o buscarlo en la ventana de ayuda. Indaga: ¿qué otras opciones tiene para facilitarte el análisis? En particular, explora los comandos y subcomandos: summary(cpS), cpaS$loading$, $cpaS$scores. ¿Cómo se interpreta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

Sugerencias en R

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



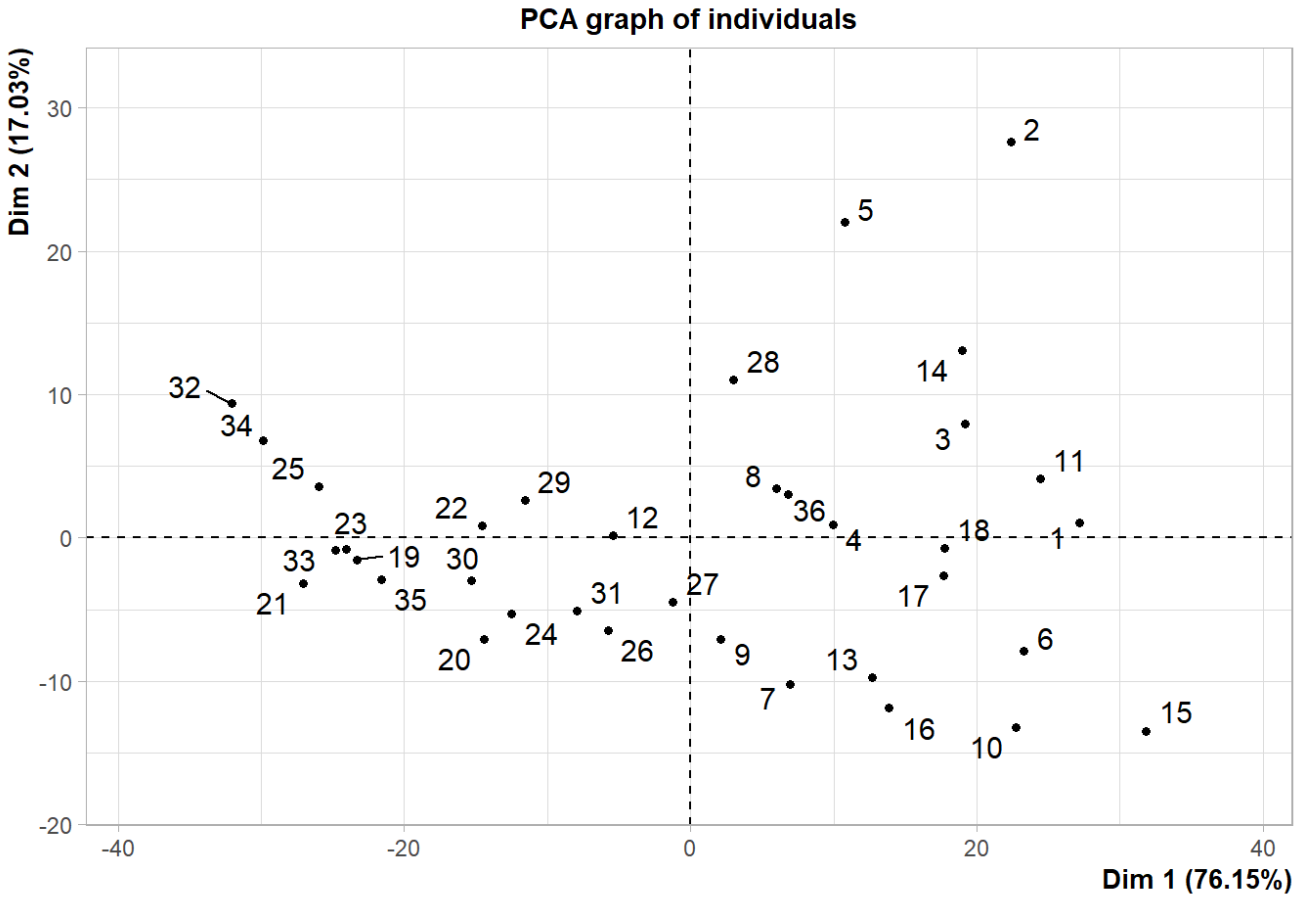

```
biplot(cpS)
```

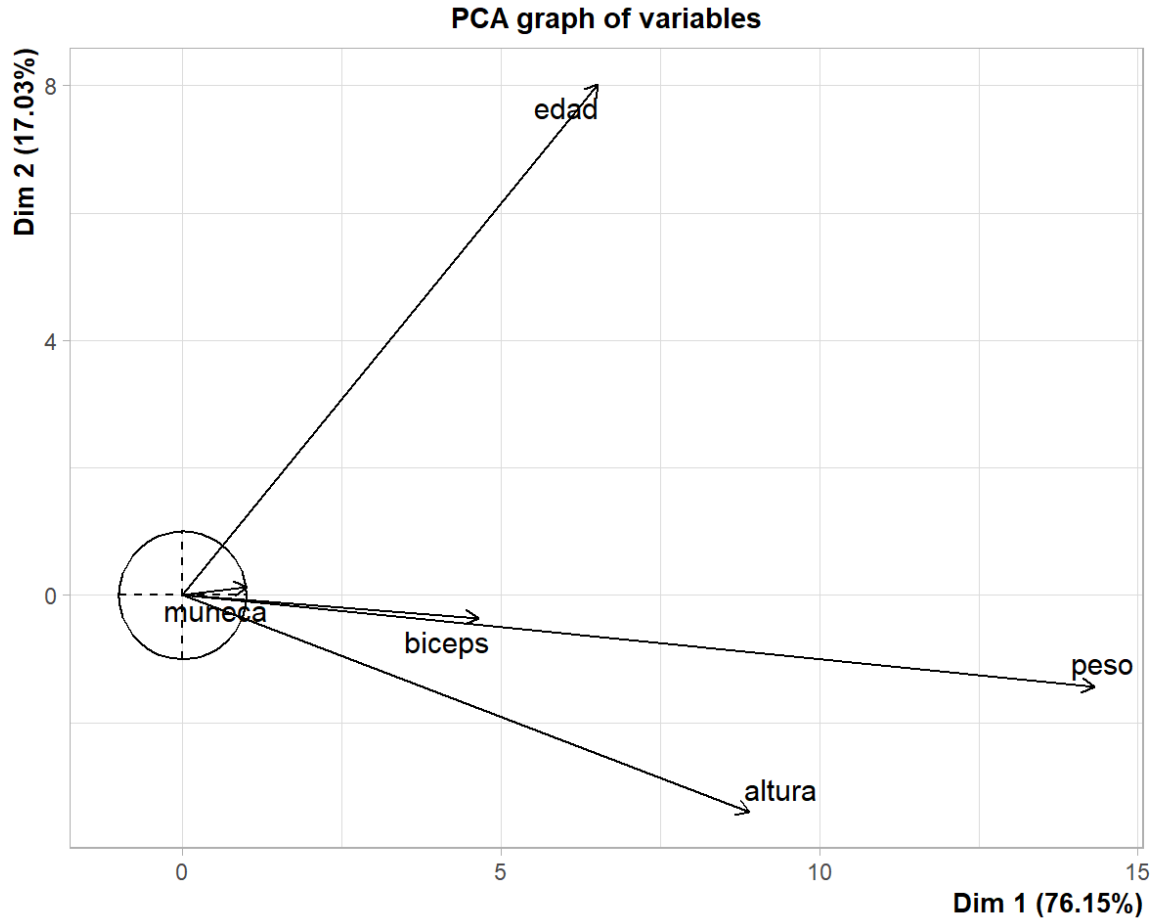



# Parte 3

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=D_numericos
cpS = PCA(datos,scale.unit=FALSE)
```



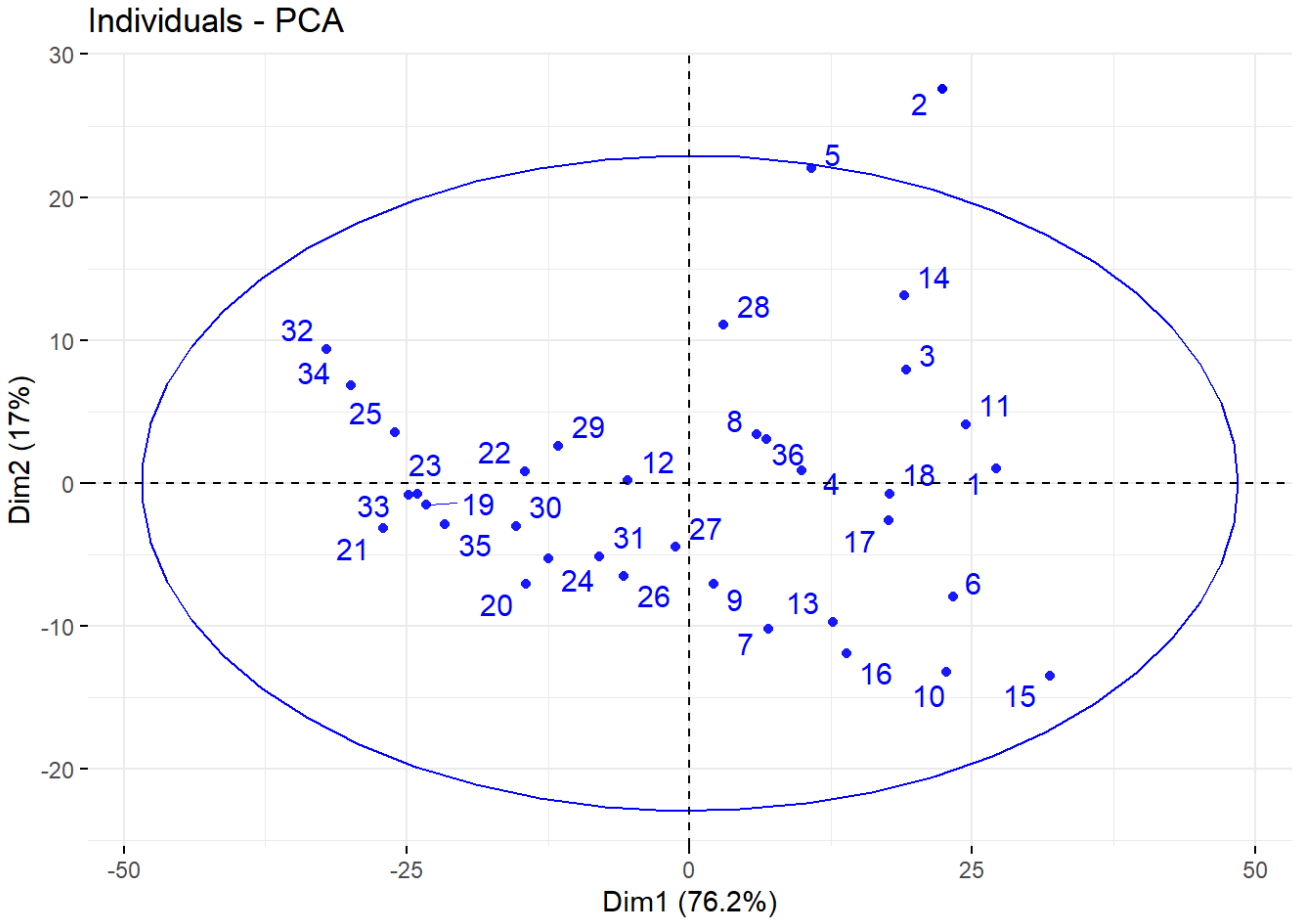

Estas dos gráficas muestran las gráficas y las variables donde cada variable (edad, peso, altura, etc.) se representa como un punto o un vector. Estos datos se pueden interpretar de varias formas. La distancia del punto al círculo de

correlación indica qué tan bien la variable es representada por el plano CP1/CP2.Al mismo tiempo el ángulo entre los vectores indica la correlación entre las variables: ángulos pequeños (vectores en la misma dirección) significan correlación positiva, ángulos de 90◦ significan no correlación, y ángulos de 180◦ significan correlación negativa. En este caso no tenemos ninguna correlación negativa.

```
library(factoextra)
```

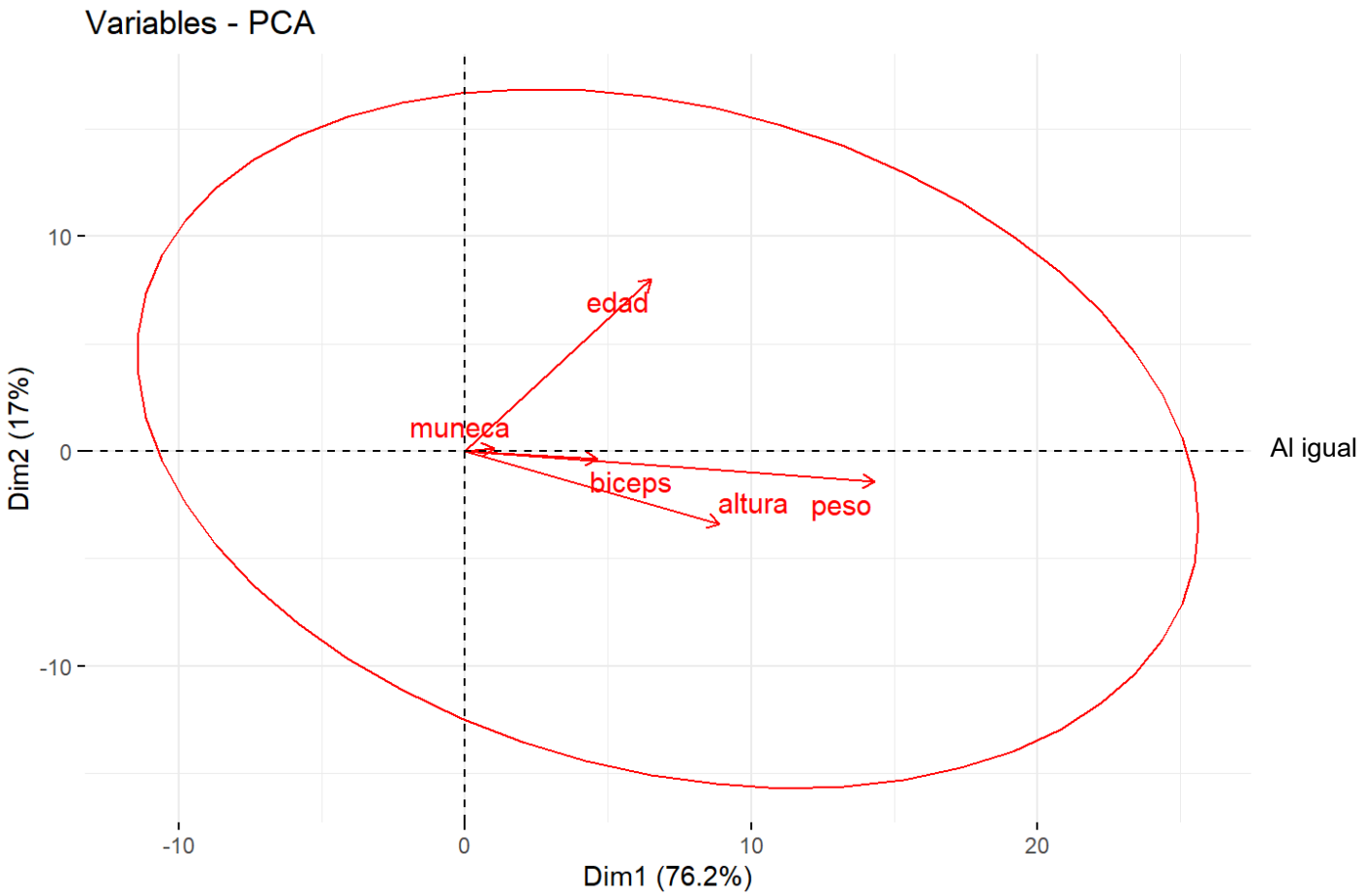
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

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



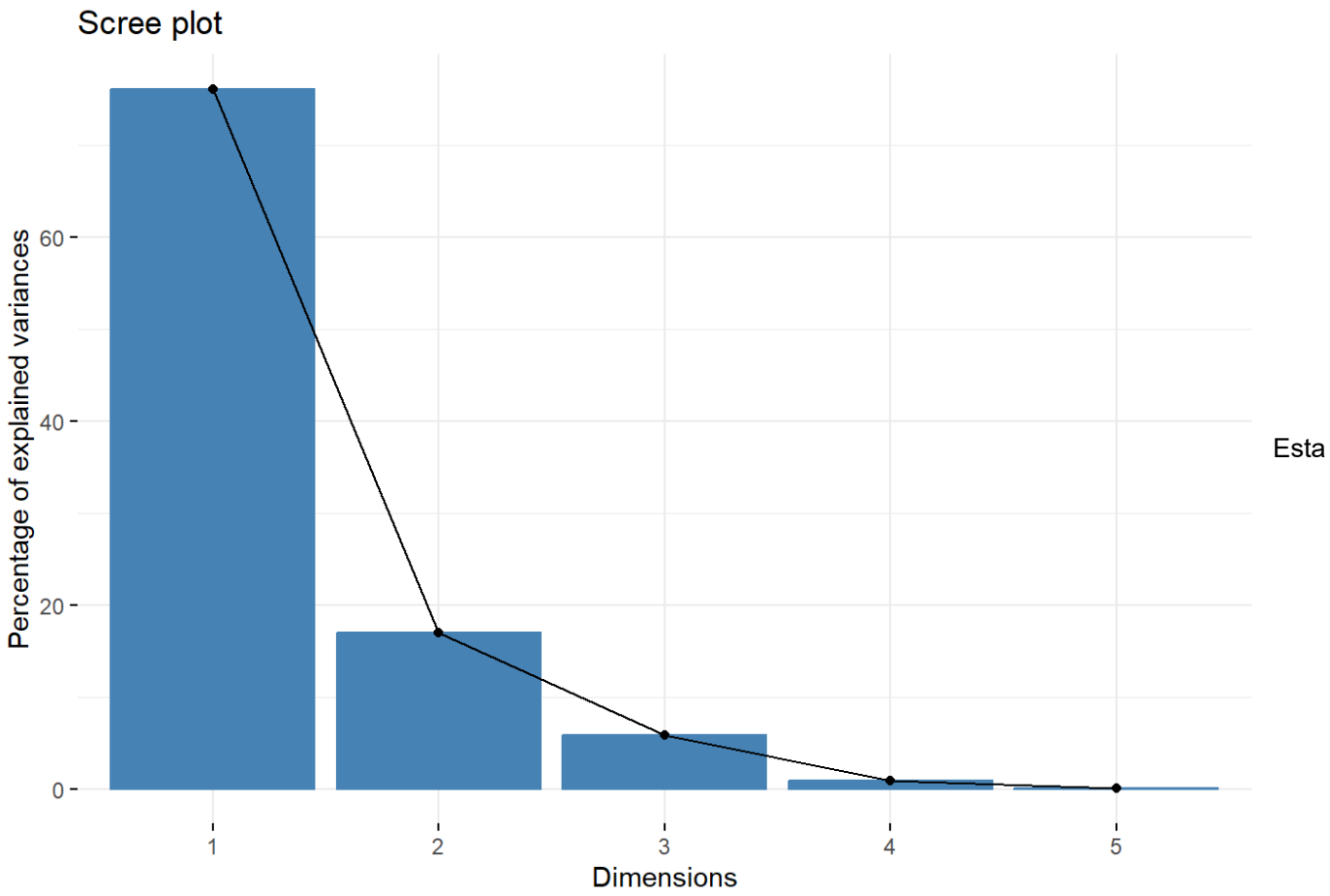

Este es un gráfico de dispersión centrado únicamente en la proyección de los individuos sobre el plano formado por las dos primeras Componentes Principales (CP1 y CP2), usando los resultados del PCA basado en la Covarianza

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



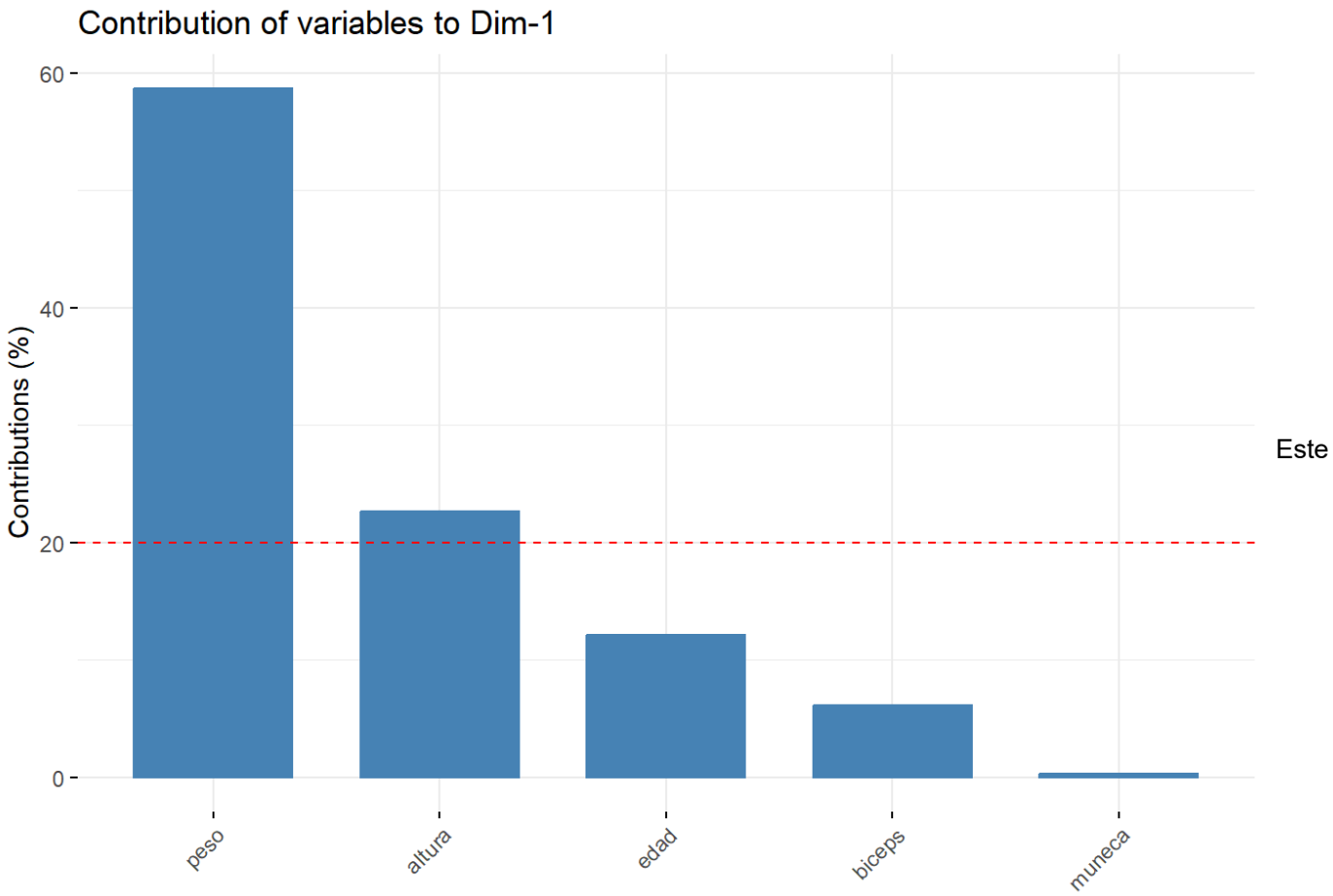

que el gráfico anterior, este muestra el mismo plano pero a través de los vectores de las variables en lugar de los puntos individuales.

```
fviz_screeplot(cpS)
```



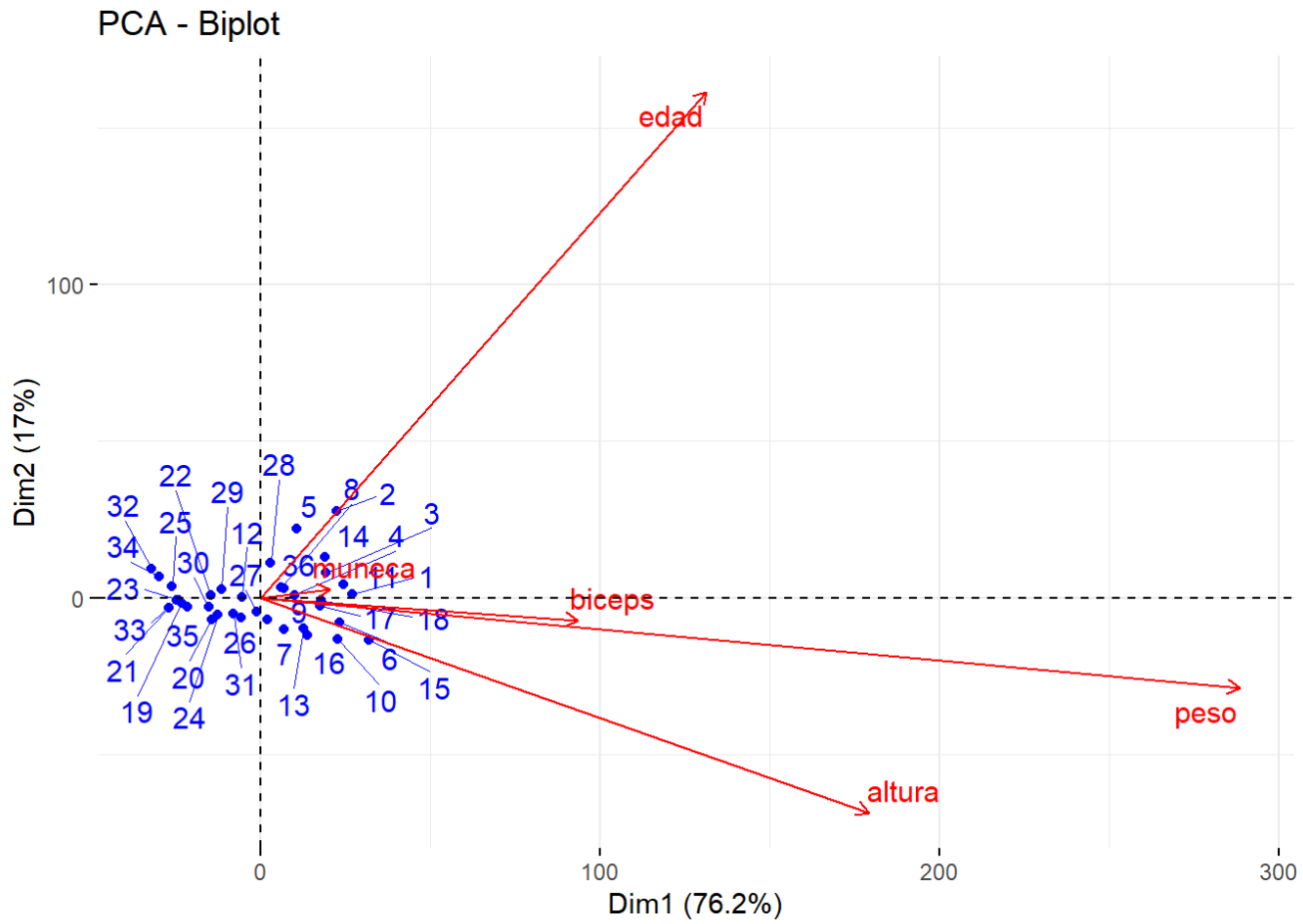

Esta gráfica muestra el valor de los valores propios (eigenvalues) o el porcentaje de varianza explicado en el eje Y para cada Componente Principal consecutiva (CP1, CP2, CP3, etc.) en el eje X. Su propósito es demostrar la importancia relativa de cada componente

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




Este gráfico demuestra la contribución absoluta de las variables al primer componente

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




demuestra simultáneamente la relación entre los individuos y las variables en el espacio de las dos primeras Componentes Principales (CP1 y CP2), usando los resultados del PCA de la matriz de Covarianza (cpS). Los puntos azules (col.ind="blue") muestran la posición de cada individuo, indicando si tienen valores altos o bajos en las variables que definen los componentes, mientras que las flechas rojas (col.var="red") representan las variables, permitiendo visualizar la correlación entre las variables y la influencia de las variables sobre los individuos. De esta forma, el gráfico permite interpretar en qué medida los individuos son descritos por ciertas variables (por ejemplo, los individuos en la dirección de la flecha de "peso" tienen un peso superior al promedio).

# Parte 4

Finalmente: Concluye sobre el análisis de componentes principales realizado e interprete los resultados.

## 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?

El Análisis de Componentes Principales (PCA) se realizó utilizando tanto la matriz de Covarianza (S) como la de Correlación (R). El PCA con Correlación (R) demostró ser más eficiente y más interpretable, ya que concentró el 75.15% de la varianza total en la primera componente (CP1) y alcanzó cerca del 90% de varianza acumulada con solo dos componentes. Esto es crucial porque las variables (edad, peso, altura, etc.) están en diferentes escalas, y estandarizarlas previene que la variable con mayor dispersión (como el peso) domine artificialmente el análisis.

El PCA con Correlación define el CP1 como un factor claro de "Tamaño Corporal General" (al cargar fuertemente en peso, altura, bíceps y muñeca), y el CP2 como un factor que aísla la "Edad". En contraste, el PCA con Covarianza resultó inusual, con el CP2 (dominado por la edad) explicando más varianza que el CP1 (dominado por el peso). Por

lo tanto, el PCA con Correlación (scale=TRUE) es el procedimiento que aporta componentes de mayor interés y más robustos, siendo la opción recomendada para este tipo de datos multivariados con escalas mixtas.

## Indique cuál de los dos análisis (a partir de la matriz de varianza y covarianza o de correlación) resulta mejor para los datos. Comparar los resultados y argumentar cuál es mejor según los resultados obtenidos.

El análisis de Componentes Principales (PCA) basado en la matriz de Correlación (R) resulta ser la opción claramente mejor para este conjunto de datos. Esto se debe a que las variables se miden en diferentes escalas (edad en años, peso en kg, medidas corporales en cm), y el PCA basado en la Correlación primero estandariza los datos, asegurando que todas las variables contribuyan equitativamente al análisis. Al comparar los resultados, el análisis de Correlación es más eficiente, logrando concentrar un 75.15% de la varianza en la primera componente (CP1) y casi el 90% en las dos primeras, con el CP1 ofreciendo una interpretación robusta como un factor de “Tamaño Corporal General”.

Por el contrario, el PCA basado en la matriz de Covarianza (S) permite que la variable con mayor varianza, como el Peso, domine desproporcionadamente los componentes, haciendo que la distribución de la varianza sea menos eficiente y la interpretación menos clara. Dado que el objetivo del PCA es reducir la dimensionalidad de una manera significativa e interpretable, el procedimiento basado en la Correlación aporta componentes con mayor interés y es el método estándar cuando se trabaja con variables de escalas heterogéneas.

## ¿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)

El método seleccionado como mejor es el basado en la matriz de Correlación (R), ya que estandariza las variables y ofrece una interpretación más robusta. Observando los coeficientes de los vectores propios (las combinaciones lineales) y el biplot de la correlación: a la primera componente principal (CP1), la que explica el 75.15% de la varianza, contribuyen fuertemente las variables de Peso, Altura, Muñeca y Bíceps, lo que define a este componente como el factor de “Tamaño Corporal General”. Por su parte, la segunda componente principal (CP2), que explica el 14.52% de la varianza, está dominada casi por completo por la variable Edad, aislando la dimensión de la variabilidad relacionada con el envejecimiento.