A4-Componentes Principales_fer

Fernanda Pérez

2024-10-08

```r
if (!require(ggplot2)) {
  install.packages("ggplot2")
  library(ggplot2)
}
```

```
## Cargando paquete requerido: ggplot2
```

```r
if (!require(psych)) {
  install.packages("psych")
  library(psych)
}
```

```
## Cargando paquete requerido: psych
```

```
##
## Adjuntando el paquete: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
if (!require(FactoMineR)) {
  install.packages("FactoMineR")
  library(FactoMineR)
}
```

```
## Cargando paquete requerido: FactoMineR
```

```r
if (!require(corrplot)) {
  install.packages("corrplot")
  library(corrplot)
}
```

```
## Cargando paquete requerido: corrplot
```

```
## corrplot 0.94 loaded
```

```r
library(ggplot2)
library(psych)
library(FactoMineR)
library(corrplot)

data <- read.csv("D:/Downloads/corporal.csv", header = TRUE)
```

```r
head(data)
```

```
##   edad peso altura   sexo muneca biceps
## 1   43 87.3  188.0 Hombre   12.2   35.8
## 2   65 80.0  174.0 Hombre   12.0   35.0
## 3   45 82.3  176.5 Hombre   11.2   38.5
## 4   37 73.6  180.3 Hombre   11.2   32.2
## 5   55 74.1  167.6 Hombre   11.8   32.9
## 6   33 85.9  188.0 Hombre   12.4   38.5
```

Se despliegan las 5 primeras filas de cada columna con la que estamos trabajando de nuestro dataset

```r
#solo columnas numéricas
numeric_data <- data[, sapply(data, is.numeric)]

# Verifica las columnas numéricas seleccionadas
head(numeric_data)
```

```
##   edad peso altura muneca biceps
## 1   43 87.3  188.0   12.2   35.8
## 2   65 80.0  174.0   12.0   35.0
## 3   45 82.3  176.5   11.2   38.5
## 4   37 73.6  180.3   11.2   32.2
## 5   55 74.1  167.6   11.8   32.9
## 6   33 85.9  188.0   12.4   38.5
```

Aquí filtramos para solo trabajar con valores numericos

```r
#summary solo columnas numéricas
summary(numeric_data)
```

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

Vemos el summary de cada variable numerica por columna.

```r
#desviación estándar de cada variable
desviacion_estandar <- sapply(numeric_data, sd)
desviacion_estandar
```

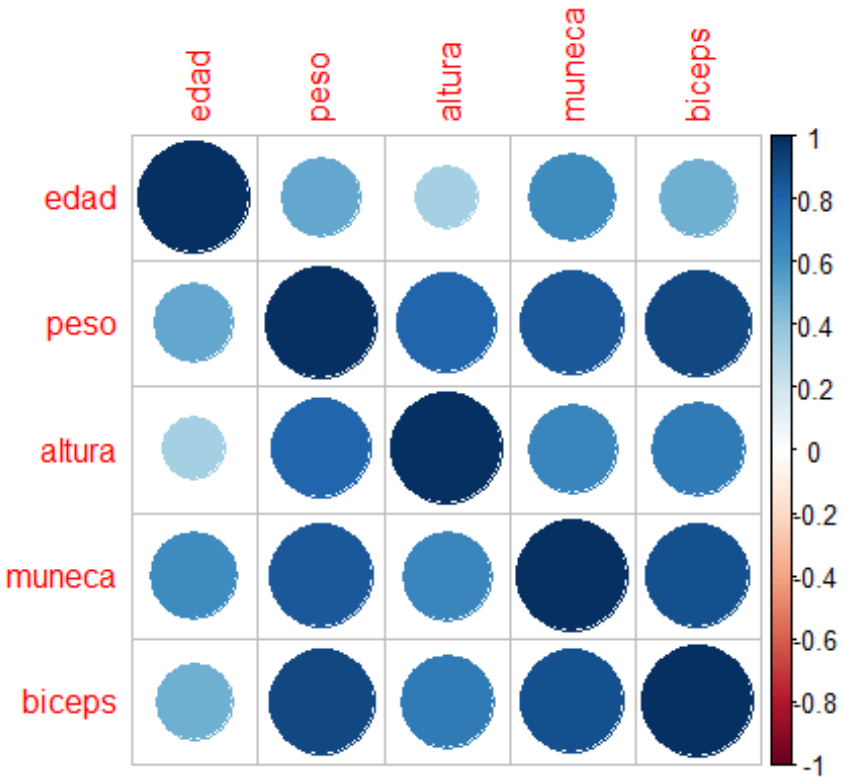
```
##      edad      peso    altura    muneca    biceps
## 10.554469 14.868999 10.520170  1.175463  5.234392
```

Vemos la desviación estandar de cada variable. Las variables con mayor dispersión son el peso y la edad, lo que nos refleja mayor diversidad en esas medidas dentro de la muestra de estudiantes. Y por otro lado la circunferencia de la muñeca es la medida más consistente.

```r
#matriz de correlación
cor_matrix <- cor(numeric_data)

#matriz de correlación
cor_matrix
```

```
##             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
```

```r
#matriz de correlación
library(corrplot)
corrplot(cor_matrix, method = "circle")
```



Las relaciones más fuertes:

Peso y bíceps (0.908): Hay una correlación muy fuerte y positiva entre el peso y el tamaño del bíceps. O sea que a medida que el peso aumenta, el tamaño del bíceps también va a aumentar significativamente.

Peso y muñeca (0.849): Hay una correlación fuerte y positiva entre el peso y la circunferencia de la muñeca. O sea que a medida que el peso aumenta, el tamaño de la circunferencia de muñeca también va a aumentar significativamente.

Bíceps y muñeca (0.877): La relación entre el tamaño del bíceps y la circunferencia de la muñeca también es muy fuerte y positiva.

Relaciones intermedias:

Altura y peso (0.797): Vemos una correlación moderadamente alta entre la altura y el peso, lo que sugiere que a mayor altura, generalmente el peso también aumenta.

Altura y bíceps (0.708): Vemos una relación positiva moderada entre la altura y el tamaño del bíceps, lo que indica que las personas más altas tienden a tener bíceps más grandes.

Altura y muñeca (0.659): Vemos una correlacion moderadamente de altura con la circunferencia de la muñeca.

Edad y muñeca (0.620): La edad tiene una correlación moderada con la circunferencia de la muñeca.

Edad y peso (0.515): La edad también tiene una correlación moderada con el peso, lo que indica que a mayor edad, el peso tiende a aumentar, pero no de manera tan significativa.

Relaciones bajas:

Edad y bíceps (0.483): La correlación que vemos entre la edad y el tamaño del bíceps es positiva, pero se destaca que es más débil en comparación con otras relaciones.

Edad y altura (0.330): Vemos una correlación débil entre la edad y la altura, es la más baja de todas. Destaca que la altura no varía considerablemente con la edad en este conjunto de datos.

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:

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.

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

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)

Matrices de covarianza y correlación

```
#la matriz de covarianza (S) y la matriz de correlación (R)
S <- cov(numeric_data)
```

```r
R <- cor(numeric_data)

S
```

```
##              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
```

```r
R
```

```
##             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
```

Valores y vectores propios

Para la matriz de covarianza (S)

```r
#valores y vectores propios de la matriz de covarianza
eigen_S <- eigen(S)

# Valores propios (autovalores)
eigen_S$values
```

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

```r
# Vectores propios (autovectores)
eigen_S$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
```

Matriz de correlación (R)

```r
#valores y vectores propios de la matriz de correlación
eigen_R <- eigen(R)

#Valores propios (autovalores)
eigen_R$values
```

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

```r
#Vectores propios (autovectores)
eigen_R$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
```

Proporción de varianza explicada por cada componente

matriz de covarianza (S)

```r
varianza_explicada_S <- eigen_S$values / sum(eigen_S$values)
varianza_explicada_S
```

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

```r
# Varianza acumulada
varianza_acumulada_S <- cumsum(varianza_explicada_S)
varianza_acumulada_S
```

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

matriz de correlación (R):

```r
varianza_explicada_R <- eigen_R$values / sum(eigen_R$values)
varianza_explicada_R
```

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

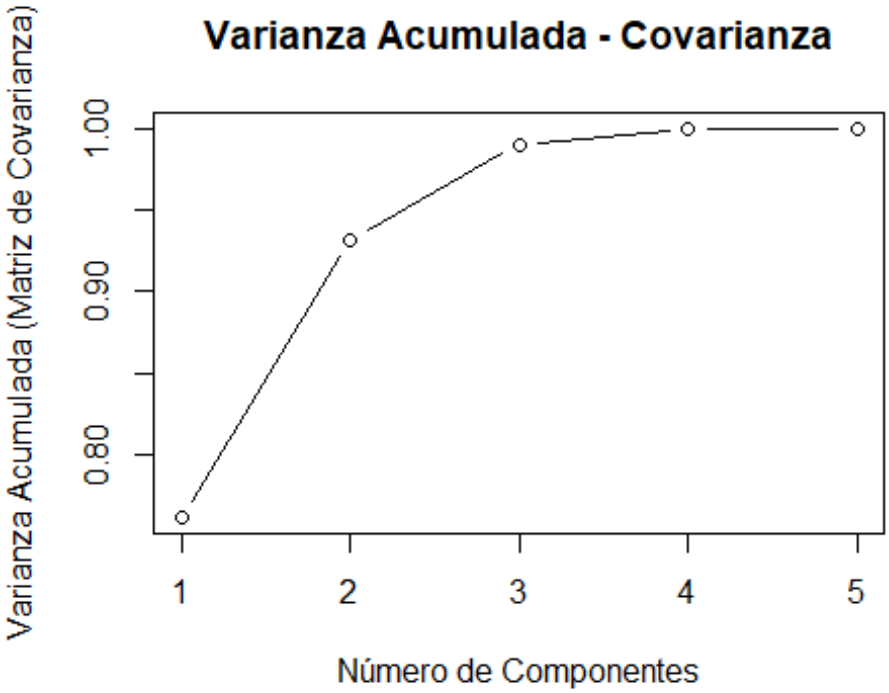
```r
# Varianza acumulada
varianza_acumulada_R <- cumsum(varianza_explicada_R)
varianza_acumulada_R
```

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

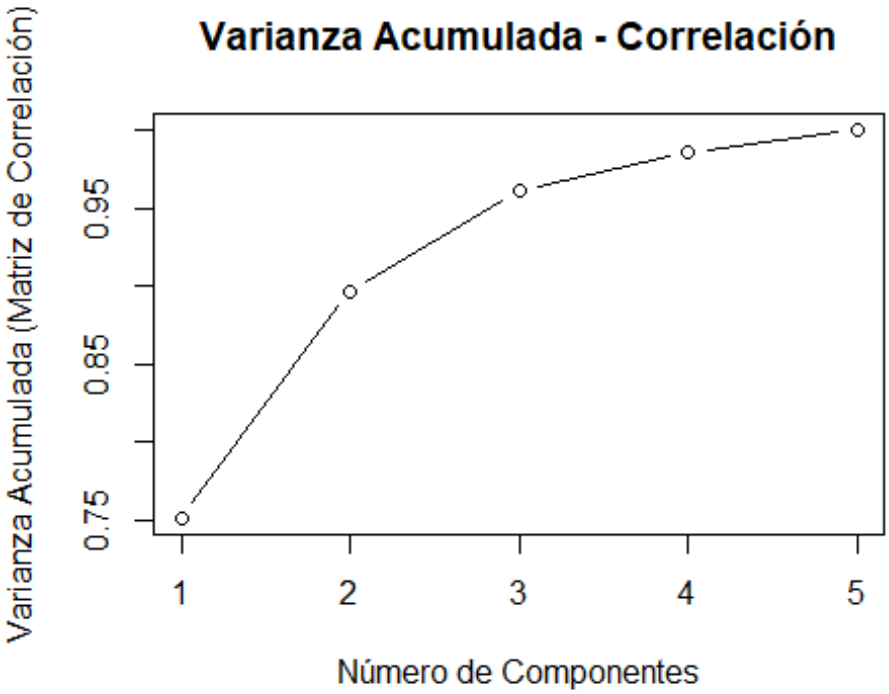
Identificamos los componentes más importantes

Los componentes más importantes van a ser los que explican la mayor proporción de la varianza.

```r
plot(varianza_acumulada_S, type = "b", xlab = "Número de Componentes",
ylab = "Varianza Acumulada (Matriz de Covarianza)", main = "Varianza
Acumulada - Covarianza")
```




```r
plot(varianza_acumulada_R, type = "b", xlab = "Número de Componentes",
ylab = "Varianza Acumulada (Matriz de Correlación)", main = "Varianza
Acumulada - Correlación")
```



Calculamos los valores y vectores propios utilizando la matriz de covarianza y la matriz de correlación.

En la matriz de covarianza, los resultados nos indican que el primer componente principal (CP1) explica el 76.15% de la varianza total, y que el segundo componente aporta un 17.03%, o sea que con solo los dos primeros componentes se explica aproximadamente el 93.15% de la varianza. Sugieriendonos que la mayor parte de la variabilidad en los datos puede resumirse con solo dos componentes.

Y en el caso de la matriz de correlación, vemos resultados similares, con el CP1 que explica el 75.15% de la varianza y el segundo componente un 14.52%, nos da un total del 89.67% con los dos primeros componentes. Confirmandonos que el comportamiento de los datos puede resumirse también con pocos componentes en este caso.

El análisis nos muestra que en los dos casos (covarianza y correlación), los dos primeros componentes principales nos explican la gran mayor parte de la variabilidad de los datos.

Y en en las 2 gráficas, el primer componente principal (CP1) explica la mayor parte de la varianza, y tambien cabe destacar que los dos primeros componentes acumulados nos explican más del 93% de la varianza total en la matriz de covarianza y aproximadamente del 90% en la matriz de

Parte 2

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.

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

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), cpa$S loading$, $cpaS$scores. ¿Cómo se interpreta el resultado?

1.Calcular las puntuaciones (scores) y graficarlas

1.1matriz de varianza-covarianza (S)

```
library(stats)
```

```r
datos <- numeric_data

cpS <- princomp(datos, cor=FALSE)

summary(cpS)
```

```
## Importance of components:
##                             Comp.1    Comp.2     Comp.3      Comp.4
Comp.5
## Standard deviation     18.6926388 8.8398600 5.18223874 2.046406827
0.4773333561
## Proportion of Variance  0.7615357 0.1703099 0.05853072 0.009127104
0.0004965839
## Cumulative Proportion   0.7615357 0.9318456 0.99037631 0.999503416
1.0000000000
```

```r
cpS$loadings
```

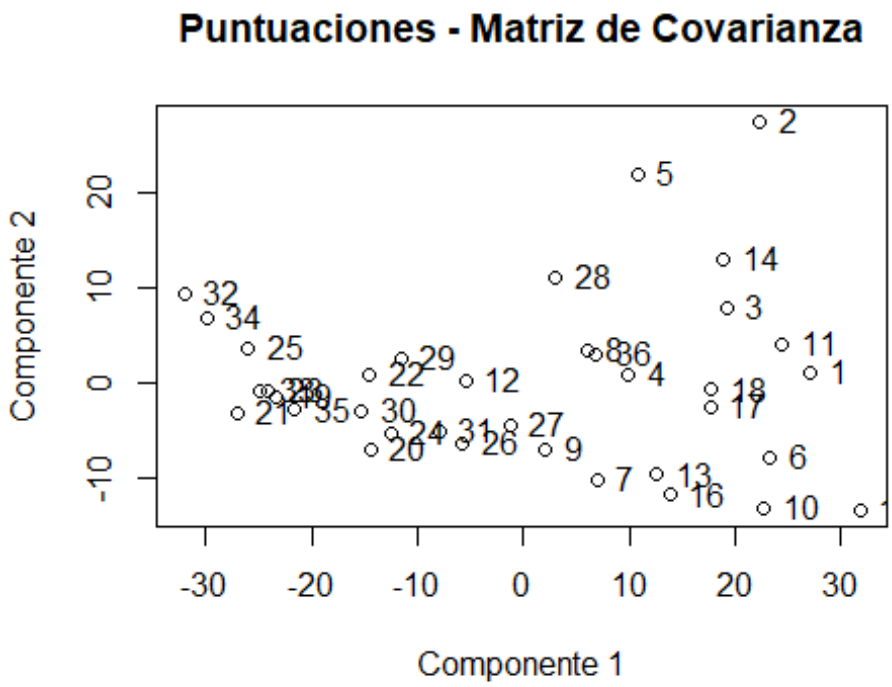
```
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## edad    0.349  0.908  0.232
## peso    0.766 -0.162 -0.522  0.339
## altura  0.476 -0.385  0.789
## muneca                      -0.126 -0.990
## biceps  0.248        -0.225 -0.931  0.138
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

```r
cpS$scores
```

```
##           Comp.1      Comp.2      Comp.3       Comp.4      Comp.5
##  [1,]  27.162853   1.0278492   5.0022646  0.936226898 -0.51688356
##  [2,]  22.363542  27.5955807   3.0635949 -0.083381259  0.02552809
##  [3,]  19.167874   7.9566157  -1.5770026 -2.610776762  0.80391745
##  [4,]   9.959001   0.8923731   5.5146952  0.123453725 -0.35579895
##  [5,]  10.775593  22.0203437  -0.7562826  0.179967226 -0.41646606
##  [6,]  23.283948  -7.9268214   2.7958617 -2.093392841 -0.62252321
##  [7,]   6.949553 -10.1882447   1.5804639 -5.636477243  0.75692216
##  [8,]   5.981213   3.4214568  -7.0113449 -0.999845471 -0.13795746
##  [9,]   2.128453  -7.0823040   9.6199213 -2.402765355  0.30931008
## [10,]  22.742222 -13.2447241  -5.8006902 -1.900258608 -0.11415400
## [11,]  24.427931   4.1227827  -3.0914640  1.417935347  0.45836253
## [12,]  -5.438123   0.1807499   1.3551969 -5.147087631 -0.71928452
## [13,]  12.665261  -9.7148314  -4.4445147  0.469977365 -0.44199755
## [14,]  18.962350  13.1080907   4.5325770  0.310839551 -0.27648044
## [15,]  31.842783 -13.4784052  -1.4672915  5.610391303  0.61177438
## [16,]  13.884278 -11.8930081  -6.4032979 -2.225813208 -0.01138562
```

```
## [17,]  17.653813  -2.6451319  -0.8986274 -0.529020358  0.37187295
## [18,]  17.723299  -0.7428241   0.1219847  1.785013852  0.68809035
## [19,] -23.293603  -1.5208783   0.2627514  1.143811767 -0.16480880
## [20,] -14.414169  -7.0887516   0.1030611  0.006854239 -0.32687435
## [21,] -27.078917  -3.1933468  -0.4483831  0.722326288 -0.02028518
## [22,] -14.579228   0.8324474  -9.1400445  1.717699742  0.23470254
## [23,] -24.042246  -0.7779288  -5.8550300 -0.340341079  0.26832127
## [24,] -12.494468  -5.2751971   3.0622990  1.094339917 -0.51675730
## [25,] -26.002609   3.5759758   1.6616974  0.054118319 -0.33475598
## [26,]  -5.766003  -6.4856729  -6.5862305  2.330421808 -0.76268815
## [27,]  -1.211876  -4.4901315   4.4920764  1.153351801  0.26364518
## [28,]   3.020501  11.0467489 -10.8052957  0.255974364 -0.43453383
## [29,] -11.574038   2.5907341   9.5304169  1.466717121  0.84144772
## [30,] -15.335150  -2.9912143   6.9968010  0.493427421 -0.36660212
## [31,]  -7.926087  -5.1312097   4.1467185  2.808113699  0.29328661
## [32,] -32.046176   9.3863372   0.8359798 -1.341797979  0.73976836
## [33,] -24.800765  -0.8616289  -0.1246471 -0.477476584  0.58698947
## [34,] -29.884003   6.8137270  -9.5237493 -0.372525171  0.27802711
## [35,] -21.626441  -2.8831824   7.4391447  0.704477945 -0.64549912
## [36,]   6.819433   3.0436244   1.8163894  1.375519851 -0.34623005
```

```
plot(cpS$scores[, 1:2], type = "p", main = "Puntuaciones - Matriz de
Covarianza", xlab = "Componente 1", ylab = "Componente 2")
text(cpS$scores[, 1], cpS$scores[, 2], labels = 1:nrow(datos), pos = 4)
```



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

Para la matriz de covarianza vemos que el primer componente principal (Componente 1) agarra una parte grande de la varianza del conjunto de datos, como lo vemos en la dispersión significativa de los puntos a lo largo del eje X. Esto nos indica que las variables con mayor varianza van a ser las que más influyen en este componente. Y para el segundo componente principal (Componente 2), que se representa en el eje Y, agarra menos varianza en comparación con el Componente 1, sin embargo nos ofrece información útil sobre diferencias entre las observaciones.

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

Las puntuaciones de las observaciones se relacionan con los valores originales de las variables, o sea que las observaciones que están más alejadas del origen en el eje X, posiblemente van a ser las que tienen altos valores en variables que dominan el primer componente.

Detecte posibles datos atípicos:

Podemos ver algunos posibles datos atípicos, como por ejemplo, las observaciones 2 y 5 destacan en la gráfica porque vemos que estan muy alejadas del centro de la distribución de puntuaciones.

1.2 matriz de correlación (R)

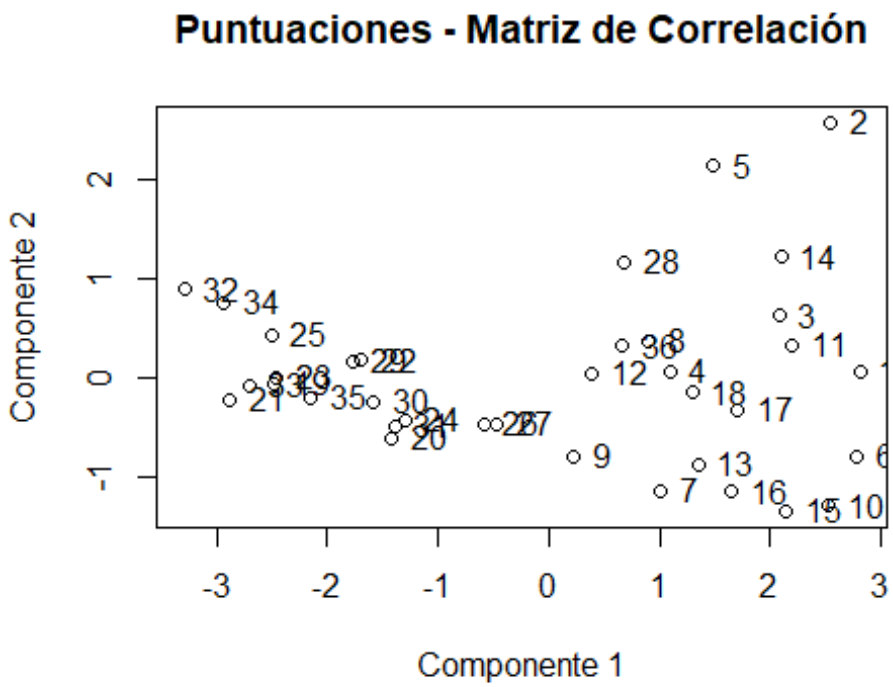
```
cpR <- princomp(datos, cor=TRUE)

summary(cpR)

## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
Comp.5
## Standard deviation     1.9384265 0.8519722 0.56597686 0.35301378
0.2677639
## Proportion of Variance 0.7514995 0.1451713 0.06406596 0.02492375
0.0143395
## Cumulative Proportion  0.7514995 0.8966708 0.96073676 0.98566050
1.0000000

cpR$loadings

##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## edad    0.336  0.858  0.349  0.136  0.107
## peso    0.493 -0.165         0.525 -0.671
## altura  0.422 -0.454  0.734 -0.207  0.184
## muneca  0.482  0.108 -0.367 -0.755 -0.226
## biceps  0.483 -0.139 -0.447  0.305  0.674
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
```

```
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

```
cpR$scores
```

```
##           Comp.1      Comp.2      Comp.3      Comp.4       Comp.5
##  [1,]  2.8139915  0.06282760  0.51434516 -0.37618363 -0.161649397
##  [2,]  2.5508161  2.57369731  0.42896223  0.01252075  0.083602262
##  [3,]  2.0792069  0.62112516 -0.12602006  0.51138786  0.430775853
##  [4,]  1.0933160  0.06328171  0.46145821 -0.35236278 -0.008424496
##  [5,]  1.4893629  2.13420572 -0.08620983 -0.19530483 -0.097669770
##  [6,]  2.7801900 -0.79964368 -0.11180511 -0.52796031  0.113681564
##  [7,]  1.0141243 -1.14171806 -0.27787746  0.22743193  0.800375496
##  [8,]  0.9063369  0.35803327 -0.79126430  0.07179533 -0.031461084
##  [9,]  0.2285350 -0.80075813  0.71215644 -0.15394896  0.481123407
## [10,]  2.5302453 -1.30235901 -0.76205083  0.03215070  0.050616130
## [11,]  2.2033222  0.32934887  0.10037610  0.49363388 -0.135246631
## [12,]  0.3885728  0.02978904 -0.70291329 -0.72426251  0.460456523
## [13,]  1.3480354 -0.88888844 -0.48237353 -0.13878866 -0.248233214
## [14,]  2.0994018  1.21514134  0.47434543 -0.23319402 -0.019726560
## [15,]  2.1447355 -1.35354752  0.76511713  0.71259130 -0.587575667
## [16,]  1.6489148 -1.16117562 -0.85070099  0.08586963  0.111234627
## [17,]  1.7030809 -0.33209829  0.01673614  0.27827557  0.099895723
## [18,]  1.2932746 -0.15858301  0.48173868  0.55369253 -0.076249945
## [19,] -2.4795617 -0.06280633  0.02839564 -0.11803106 -0.136704692
## [20,] -1.4200084 -0.61570309 -0.15277478 -0.25447677 -0.063137788
## [21,] -2.8791600 -0.22853227 -0.06023367 -0.03148088 -0.068564803
## [22,] -1.6992789  0.16837324 -0.63755548  0.43611800 -0.277172176
## [23,] -2.4625686 -0.01072936 -0.59031600  0.26691381  0.024784946
## [24,] -1.3015384 -0.43354360  0.20575074 -0.40705451 -0.177314913
## [25,] -2.5058729  0.42780280 -0.01308499 -0.30917018 -0.015086855
## [26,] -0.5896282 -0.46963951 -0.61738513 -0.25029697 -0.536163469
## [27,] -0.4747287 -0.46682854  0.62201914  0.09167385 -0.007586913
## [28,]  0.6816507  1.16291258 -1.08391248  0.03253793 -0.282947483
## [29,] -1.7786024  0.15640801  1.29302710  0.33642964  0.183446578
## [30,] -1.5894735 -0.25254138  0.54948615 -0.44020946 -0.006577363
## [31,] -1.3903223 -0.49360911  0.76675148  0.17233872 -0.188151664
## [32,] -3.2962547  0.88748511  0.06759476  0.35410490  0.371715392
## [33,] -2.7100620 -0.08340844  0.02833828  0.31628667  0.201732879
## [34,] -2.9371073  0.75312128 -0.93702305  0.36683866 -0.011037680
## [35,] -2.1514986 -0.20099407  0.51126095 -0.63846467 -0.074866432
## [36,]  0.6685529  0.31355440  0.25564126 -0.20140147 -0.201892385
```

```
plot(cpR$scores[, 1:2], type = "p", main = "Puntuaciones - Matriz de
Correlación", xlab = "Componente 1", ylab = "Componente 2")
text(cpR$scores[, 1], cpR$scores[, 2], labels = 1:nrow(datos), pos = 4)
```



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

Y para la matriz de correlaciín las variables se estandarizan, y esto genera que las relaciones que captura el Componente 1 reflejan combinaciones lineales de las variables que tienen un peso parecido en términos de correlación. La dispersión de los puntos que vemos es menor en magnitud en comparación con la matriz de covarianza, dado que se ajusta por la escala de las variables.

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

Podemos ver que las puntuaciones indican cómo las observaciones se sitúan respecto a las variables estandarizadas. Las observaciones que están más alejadas en el eje X, nos indican que son observaciones con valores altos en variables relacionadas positivamente con el Componente 1.

Detecte posibles datos atípicos:

Aquí también vemos que las observaciones 2 y 5 estan alejadas del grupo principal de datos, siendo datos atipicos.

En los dos casos, los dos primeros componentes toman la mayor parte de la variabilidad en los datos. Y tambien en los dos las observacciones 2 y 5 pareciera que son datos atípicos.

```r
cpS <- princomp(numeric_data, cor=TRUE)
summary(cpS)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
Comp.5
## Standard deviation     1.9384265 0.8519722 0.56597686 0.35301378
0.2677639
## Proportion of Variance 0.7514995 0.1451713 0.06406596 0.02492375
0.0143395
## Cumulative Proportion  0.7514995 0.8966708 0.96073676 0.98566050
1.0000000
```

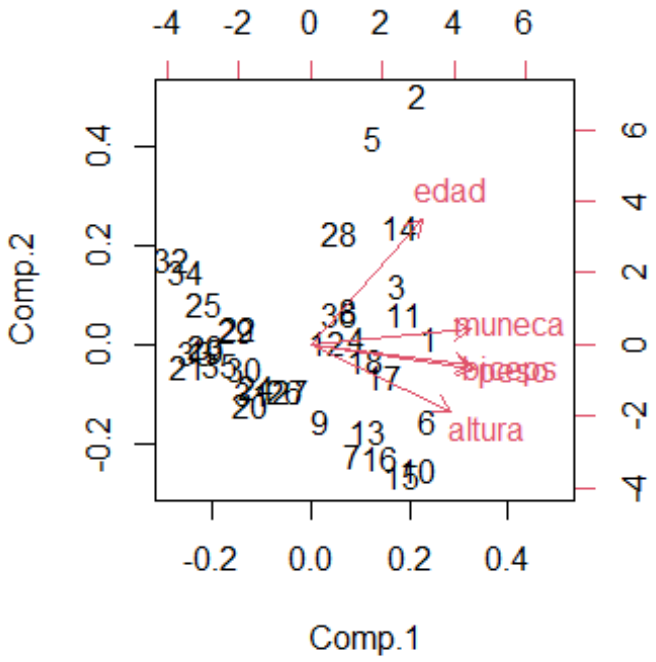
```r
cpS$loadings
```

```
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## edad    0.336  0.858  0.349  0.136  0.107
## peso    0.493 -0.165         0.525 -0.671
## altura  0.422 -0.454  0.734 -0.207  0.184
## muneca  0.482  0.108 -0.367 -0.755 -0.226
## biceps  0.483 -0.139 -0.447  0.305  0.674
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

```r
cpS$scores
```

```
##           Comp.1      Comp.2      Comp.3      Comp.4       Comp.5
##  [1,]  2.8139915  0.06282760  0.51434516 -0.37618363 -0.161649397
##  [2,]  2.5508161  2.57369731  0.42896223  0.01252075  0.083602262
##  [3,]  2.0792069  0.62112516 -0.12602006  0.51138786  0.430775853
##  [4,]  1.0933160  0.06328171  0.46145821 -0.35236278 -0.008424496
##  [5,]  1.4893629  2.13420572 -0.08620983 -0.19530483 -0.097669770
##  [6,]  2.7801900 -0.79964368 -0.11180511 -0.52796031  0.113681564
##  [7,]  1.0141243 -1.14171806 -0.27787746  0.22743193  0.800375496
##  [8,]  0.9063369  0.35803327 -0.79126430  0.07179533 -0.031461084
##  [9,]  0.2285350 -0.80075813  0.71215644 -0.15394896  0.481123407
## [10,]  2.5302453 -1.30235901 -0.76205083  0.03215070  0.050616130
## [11,]  2.2033222  0.32934887  0.10037610  0.49363388 -0.135246631
## [12,]  0.3885728  0.02978904 -0.70291329 -0.72426251  0.460456523
## [13,]  1.3480354 -0.88888844 -0.48237353 -0.13878866 -0.248233214
## [14,]  2.0994018  1.21514134  0.47434543 -0.23319402 -0.019726560
## [15,]  2.1447355 -1.35354752  0.76511713  0.71259130 -0.587575667
## [16,]  1.6489148 -1.16117562 -0.85070099  0.08586963  0.111234627
## [17,]  1.7030809 -0.33209829  0.01673614  0.27827557  0.099895723
## [18,]  1.2932746 -0.15858301  0.48173868  0.55369253 -0.076249945
## [19,] -2.4795617 -0.06280633  0.02839564 -0.11803106 -0.136704692
```

```
## [20,] -1.4200084 -0.61570309 -0.15277478 -0.25447677 -0.063137788
## [21,] -2.8791600 -0.22853227 -0.06023367 -0.03148088 -0.068564803
## [22,] -1.6992789  0.16837324 -0.63755548  0.43611800 -0.277172176
## [23,] -2.4625686 -0.01072936 -0.59031600  0.26691381  0.024784946
## [24,] -1.3015384 -0.43354360  0.20575074 -0.40705451 -0.177314913
## [25,] -2.5058729  0.42780280 -0.01308499 -0.30917018 -0.015086855
## [26,] -0.5896282 -0.46963951 -0.61738513 -0.25029697 -0.536163469
## [27,] -0.4747287 -0.46682854  0.62201914  0.09167385 -0.007586913
## [28,]  0.6816507  1.16291258 -1.08391248  0.03253793 -0.282947483
## [29,] -1.7786024  0.15640801  1.29302710  0.33642964  0.183446578
## [30,] -1.5894735 -0.25254138  0.54948615 -0.44020946 -0.006577363
## [31,] -1.3903223 -0.49360911  0.76675148  0.17233872 -0.188151664
## [32,] -3.2962547  0.88748511  0.06759476  0.35410490  0.371715392
## [33,] -2.7100620 -0.08340844  0.02833828  0.31628667  0.201732879
## [34,] -2.9371073  0.75312128 -0.93702305  0.36683866 -0.011037680
## [35,] -2.1514986 -0.20099407  0.51126095 -0.63846467 -0.074866432
## [36,]  0.6685529  0.31355440  0.25564126 -0.20140147 -0.201892385

biplot(cpS)
```



En el análisis de componentes principales, el Componente 1 se ecuentra influenciado en particular por las variables muñeca, bíceps, y altura, mientras que el Componente 2 está relacionado con edad.

Las observaciones 14, 1, y 11 están fuertemente influenciadas por las variables físicas, mientras que las observaciones 2 y 5 son posibles datos atípicos, ya que están alejadas del grupo principal y son más afectadas por la edad. El biplot permite visualizar cómo

las variables y las observaciones se relacionan en el espacio de los componentes principales.

¿qué otras opciones tiene para facilitarte el análisis?

con la función princomp() podemos ver que hay diversas opciones que nos facilitan el análisis de componentes principales:

cor = TRUE: Nos indica que se debe usar la matriz de correlación para el PCA. Nos es útil cuando las variables están en diferentes escalas.

scores = TRUE: Calcula las puntuaciones de las observaciones en los nuevos componentes principales. Esto nos permite entender cómo se proyectan las observaciones en el nuevo espacio de los componentes.

biplot(): Esta opción permite visualizar simultáneamente las puntuaciones de las observaciones y las cargas de las variables en el mismo gráfico, haciendo más facil la interpretación conjunta.

summary(cpS) Este comando nos muestra un resumen de los componentes principales, incluyendo los valores propios y la varianza explicada por cada componente.

Interpretación: Los valores propios (autovalores) indican la cantidad de varianza explicada por cada componente. La proporción de varianza explicada muestra qué porcentaje de la variabilidad total de los datos es capturada por cada componente.

cpaS$loading Este subcomando nos va a devolver las cargas (loadings), o sea los coeficientes que muestran qué tanto contribuye cada variable original a los componentes principales.

Interpretación: Las cargas altas nos van a indicar que una variable tiene un gran peso en la construcción del componente.

cpaS$scores. Con este comando podemose ver las puntuaciones (scores) de las observaciones en el espacio de los componentes principales.

Interpretación: Las puntuaciones nos van a indicar qué tan bien una observación se proyecta en los componentes principales.

Parte 3

```
if (!require(factoextra)) {
  install.packages("factoextra")
  library(factoextra)
}
```

```
## Cargando paquete requerido: factoextra
```

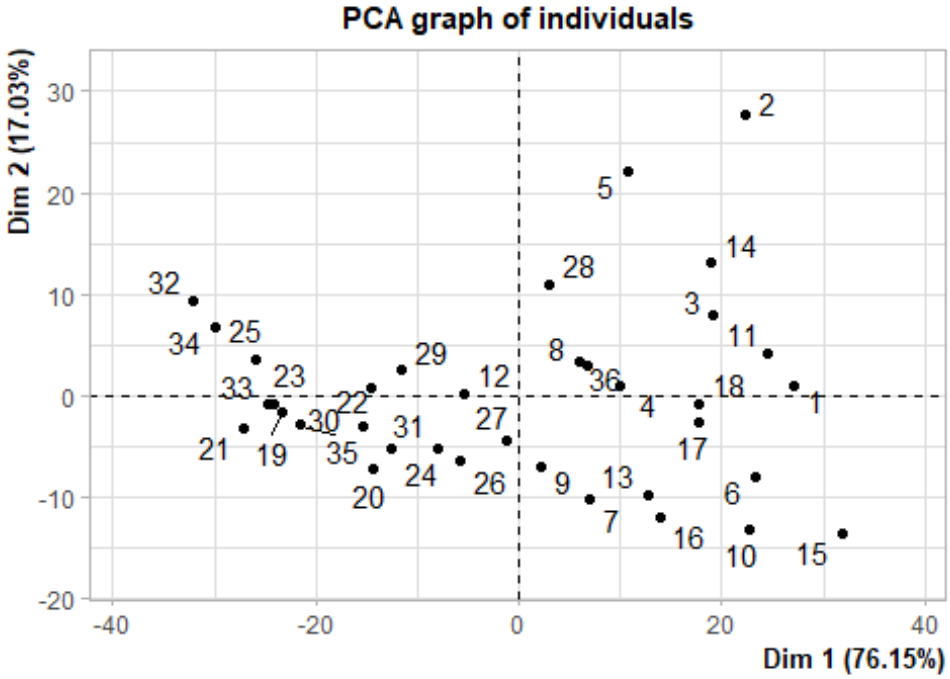
```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

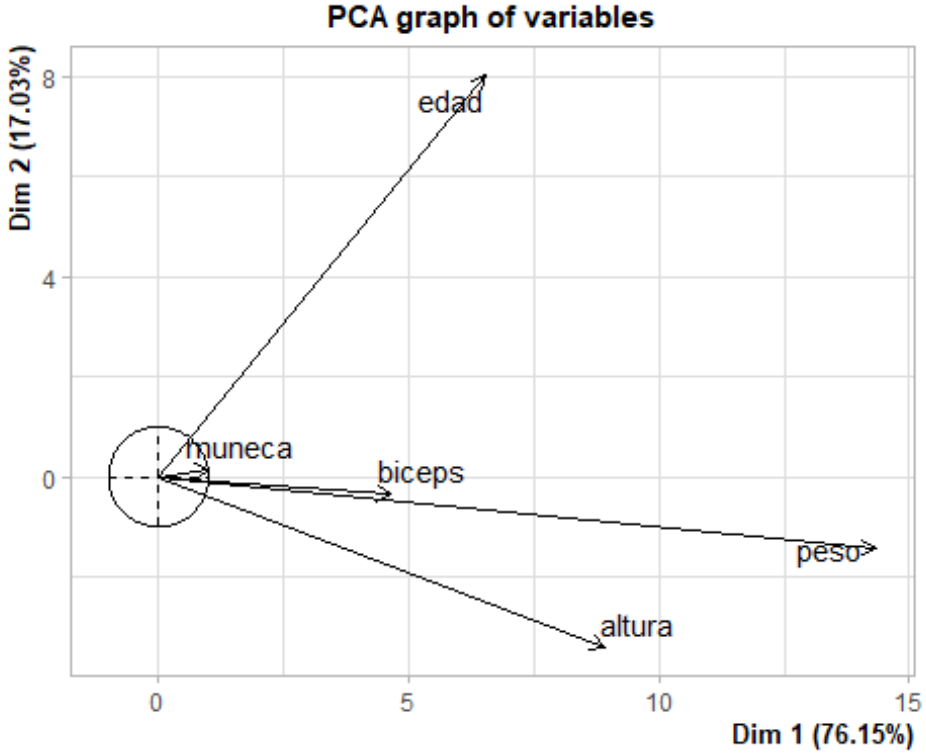
```r
if (!require(ggplot2)) {
  install.packages("ggplot2")
  library(ggplot2)
}

if (!require(FactoMineR)) {
  install.packages("FactoMineR")
  library(FactoMineR)
}
```

Realizar el Análisis de Componentes Principales (PCA)

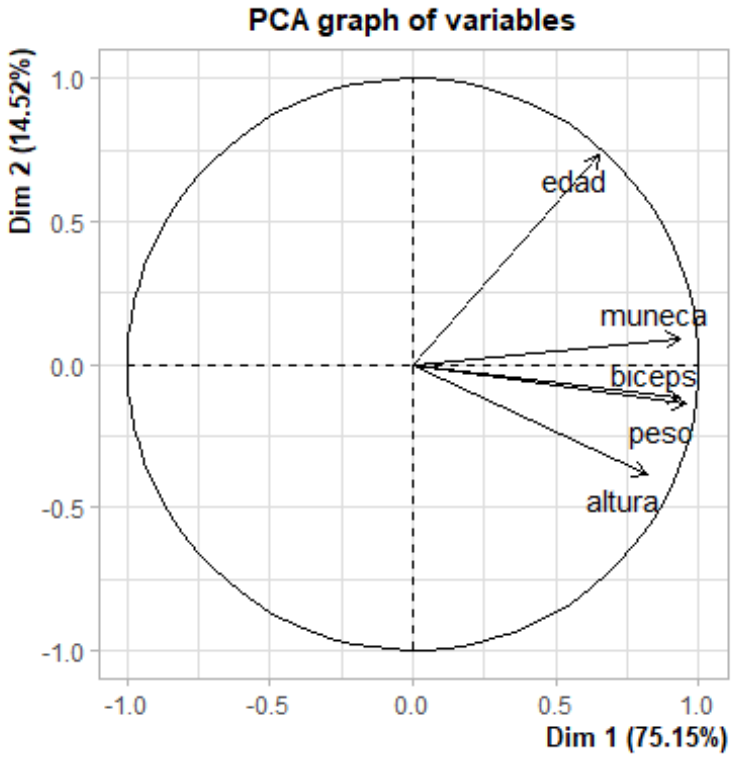
Se usa el comando PCA() perteneciente a la librería FactoMineR que se utilizará para hacer el análisis.

```r
cpS <- PCA(numeric_data, scale.unit = FALSE)
```






```r
cpS_cor <- PCA(numeric_data, scale.unit = TRUE)
```



Interpretaciones:

Gráfico de individuos (matriz de covarianza): Vemos que el componente 1 logra explicar el 76.15% de la varianza y está influenciado principalmente por variables físicas como lo son elpeso y la altura. Los individuos que destacan son el 2 y el 5 ya que como ya habiamos visto anterirmente están lejos del centro, indicandonos que son potenciales outliers.

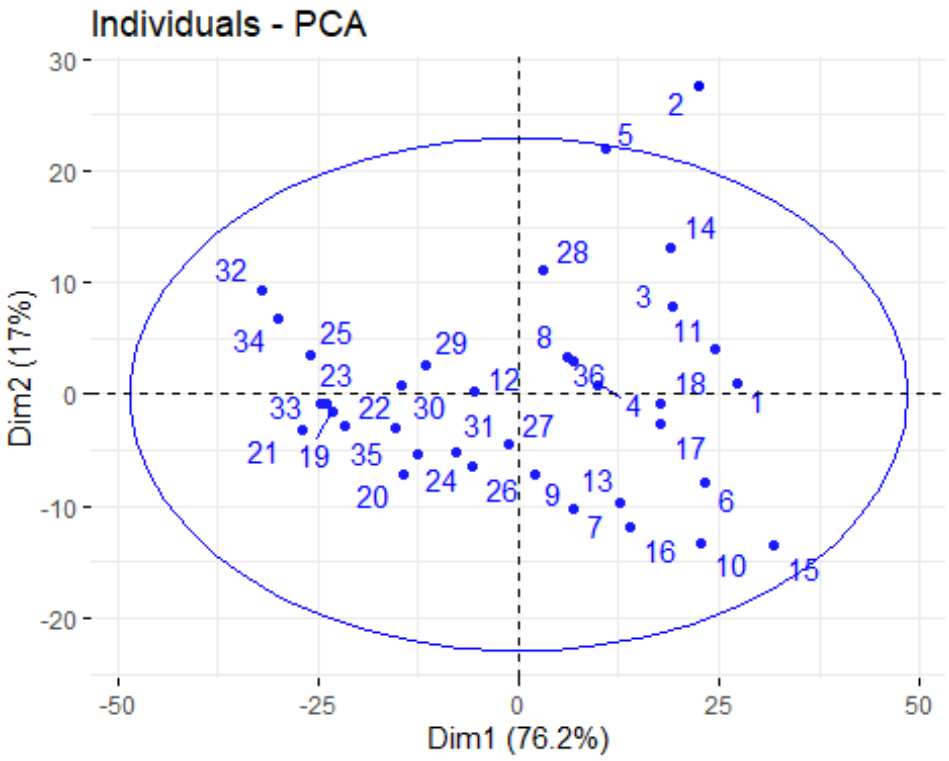
Gráfico de variables (matriz de covarianza): Vemos que peso, bíceps y altura tienen una influencia muy fuerte en el componente 1, y por otro lado edad afecta mayormente al componente 2.

Gráfico de individuos (matriz de correlación): En este grafico al trabajar con los datos estandarizados, el componente 1 logra explicar el 75.15% de la varianza. Y seguimos viendo que 2 y 5 siguen siendo atípicas.

Gráfico de variables (matriz de correlación): Podemos ver como peso y altura siguen siendo aun las variables que más influyen en el componente 1, pero las variables están estandarizadas, reduciendo el efecto de su escala original.

Gráfico de individuos (puntuaciones de las observaciones)

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



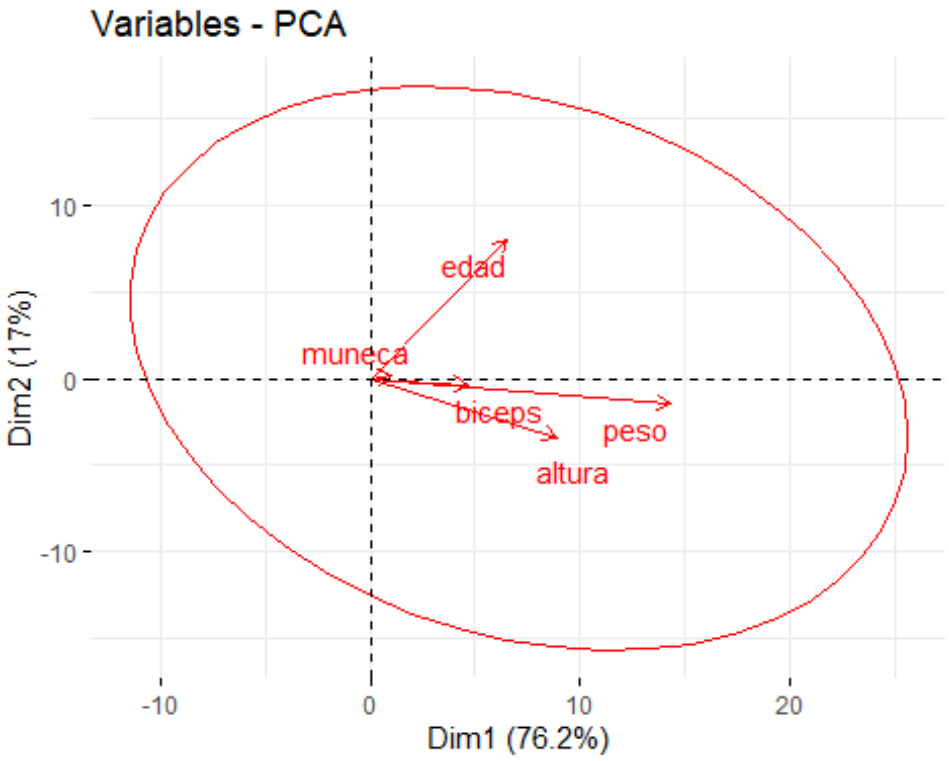
Lo que representa: En este gráfico podemos ver cómo se proyectan las observaciones (individuos) en el espacio de los primeros dos componentes principales, en este caso.

Su interpretación nterpretación: Cada uno de los puntos representa una observación, y su ubicación representa cómo esque se relaciona con los componentes principales. Y las elipses indican agrupaciones o clusters de observaciones similares.

Interpretación: En Dim 1 (76.2%): Aqui se explica la gran mayor parte de la variabilidad en los datos y las observaciones que estan alejadas de cero en este eje, como lo son 2 y 5, están más influidas por las variables que contribuyen fuertemente a este componente como lo son el peso, la altura, etc. En Dim 2 (17%): Esta explica una menor proporción de la variabilidad y está influido por variables menos dominantes.

Gráfico de variables (cargas de las variables)

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

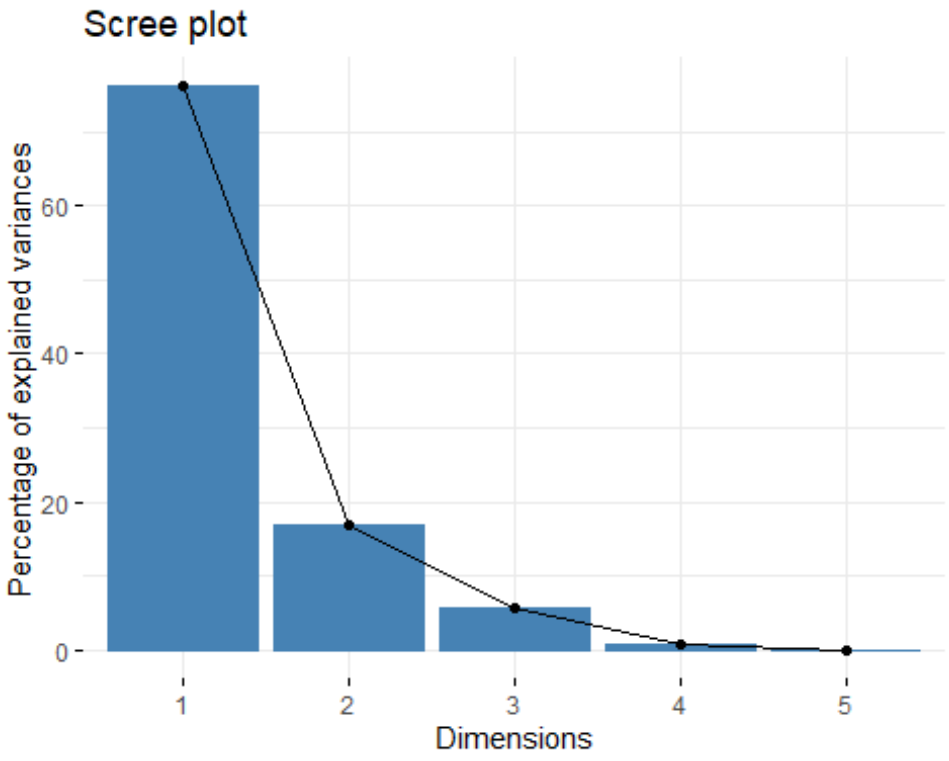


Lo que representa: En este gráfico podemos ver cómo se proyectan las variables originales en el espacio de los componentes principales. Los vectores nios van a indicar cuales son las variables que contribuyen más a cada componente. Su interpretación: Las variables que tienen vectores largos y sean cercanos a los ejes principales tienen una mayor influencia en esos componentes.

Interpretación: En Dim 1 (76.2%): Vemos como está influido principalmente por las variables de peso, bíceps, y la altura, ya que sus vectores son largos en la dirección de este componente. En Dim 2 (17%): Como vemos está más influido por la variable edad, ya que como podemos observar tiene un vector largo hacia arriba.

Gráfico de screeplot (varianza explicada por componente)

```
fviz_screeplot(cpS)
```



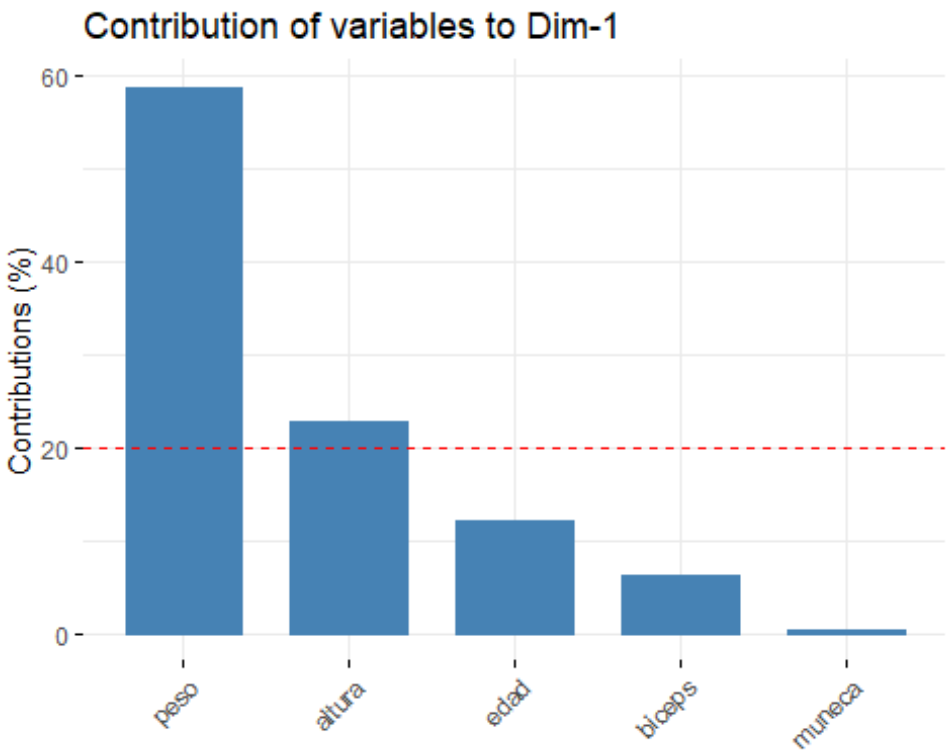
Lo que representa: En este gráfico vemos la proporción de varianza explicada por cada componente principal.

Su interpretación: Nos ayuda a identificar cuántos componentes principales van a ser suficientes para explicar la mayor parte de la varianza.

Interpretación: El componente 1 explica aproximadamente el 76% de la varianza. Y el componente 2 explica alrededor del 17% de la varianza. Y como vemos ya a partir del tercer componente, la cantidad de varianza explicada es muy muy baja en comparación de los primeros dos componentes, lo cual nos indica que la mayor parte de la información se captura con los primeros dos componentes.

Gráfico de contribución de variables a los componentes

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



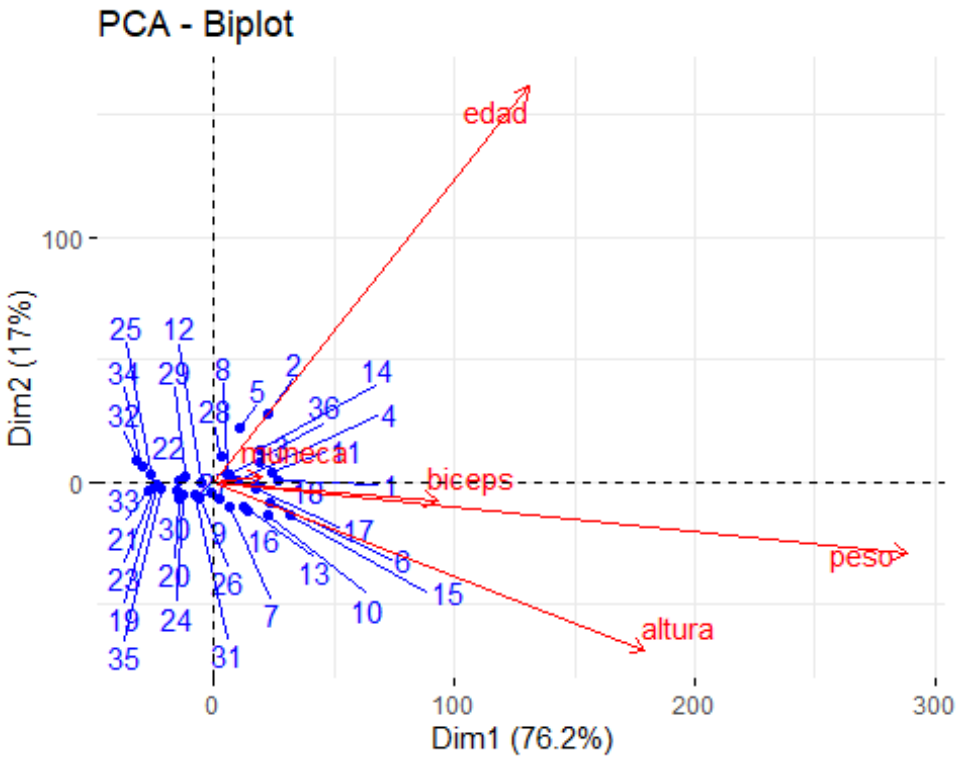
Lo que representa: En este grafico vemos qué tanto es que contribuye cada variable a la construcción de los componentes principales.

Su interpretación: Las variables que tienen contribuciones más altas van a ser más importantes en la construcción de los componentes.

Interpretación: Como vemos peso es la variable que más contribuye, y logra explicar casi el 60% del Componente 1. La altura y la edad también contribuyen significativamente despues de peso. Y vemos que bíceps y muñeca tienen contribuciones mucho más pequeñas.

Biplot (proyección de individuos y variables en un solo gráfico)

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



Lo que representa: El biplot hace la combinación sobre la proyección de los individuos (observaciones) y las variables en el mismo gráfico.

Su interpretación: Nos permite ver cómo esque las variables influyen en las observaciones.

Interpretación: Vemos que nuevamente peso y altura están fuertemente asociados con el componente 1, que explica el 76.2% de la varianza. Y edad por otro lado influye principalmente en el componente 2, explicando el 17% de la varianza. Nuevamente se destaca como 2 y 5 están alejadas, lo que indica que son muy influidas por estas variables.

Explora el comando PCA

```
help(PCA)

## starting httpd help server ... done
```

¿qué otras opciones tiene para facilitarte el análisis? ncp: Número de componentes principales a calcular (por default 5). graph: Indica si se desea visualizar gráficos automáticamente al ejecutar PCA (boolean). quali.sup: Permite agregar variables cualitativas al análisis. ind.sup: Permite agregar individuos suplementarios al análisis.

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?

Al momento de hacer la comparación de los resultados obtenidos con la matriz de varianza-covarianza y la matriz de correlación podemos observar en nuestro análisis que los dos metodos nos proporcionan resultados similares en términos de los primeros componentes principales. Sin embrago continuando con el análisis la matriz de correlación es más adecuada cuando las variables tienen diferentes escalas, ya que estandariza las variables ya que se encarga de que tenga media cero y desviación estandar uno. En este particular caso apesar de que las dos matrices nos brindaron información importante y crucial la matriz de correlación nos permitió hacer una comparación más justa entre variables con diferentes escalas como lo fueron altura y peso y evita que las variables con mayor rango tengan un impacto desproporcionado, por esa razón ese es el procedimiento que nos aporta componentes con de mayor interés

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.

Para este caso que que se nos indica un análisis a datos económicos y sociales de 96 países en el mundo, la matriz de correlación va a ser la más apropiada. Ya que al trabajar con el tipo de datos como el PIB, esperanza de vida, tasa de desempleo, etc. como vemos estan en diferentes escalas y diferentes unidades, lo cual dificulta la situación. Asi que al usar la matriz de correlación nos permitirá una mejor comparación entre variables al estandarizarlas ya que evita que las que tengan un rango mayor o mayor varianza sean más dominantes por l aescala, asi que al buscar tener resultados más equilibrados y que representen mejor la mejor estructura de datos se elige esa.

¿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)

En el análisis que se realizó como vimos las variables que más contribuyen al componente 1 son peso y altura, dado que como previamente analisamos sus vectores tienen mayor longitud en los gráficos de variables y son responsables de gran parte de la varianza explicada por este componente (alrededor del 76%) y por otro lado el componente 2 está principalmente influenciado por la edad.

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

Despues de análisar nuestros resultados podemos concluir que las combinaciones lineales de variables que se pueden recomendar para este análisis de los componentes principales son aquellas que involucran principalmente peso, altura, y edad, por las razones previamente explicadas y justificadas ya que son las variables que más contribuyen a los dos primeros componentes.

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

El análisis de componentes principales que se acaba de realizar nos reveló que una agrupación clara entre variables que reflejan características físicas como peso, altura, y bíceps, las cuales se agrupan en el componete 1, lo cual nos sugiere que estas variables capturan características comunes o relacionadas con el tamaño corporal o el desarrollo físico. Sin embargo tambien cabe destacar que la variable de edad aparece más relacionada con el componente 2, asi qee si lo vemos en términos de agrupación, podríamos decir que las variables físicas forman un grupo de características relacionadas con el "tamaño corporal", y por otro lado la edad podriamos decir que representan un "índice de desarrollo" que capta una dimensión diferente de variabilidad entre las observaciones.

Como conclusión el análisis de componentes principales (PCA) que se realizó en esta actividad me permitió lograr reducir la dimensionalidad del conjunto de datos con el que se esta trabajando, se logra destacar que peso y altura son las variables que más contribuyen al primer componente, ya que explica el 76% de la varianza, y por otro lado edad influye en el segundo componente, ya que explica el 17%. Y se deduce que la matriz de correlación fue la más adecuada debido a las diferentes escalas de las variables como se explicó anteriormente. El enfoque que se tomó a lo largo de la actividad me permitió identificar los patrones clave y simplificar el análisis, para poder resaltar las variables más importantes sin perder información importante.