A4-Componentes Principales

Catherine Rojas

2024-10-08

En la base de datos Corporal contiene 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.

Análisis descriptivo

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

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Leer datos
data1 <- read_csv("C:/Users/PC/OneDrive - Instituto Tecnologico y de
Estudios Superiores de Monterrey/Documents/Concentracion
Estadistica/corporal.csv")
```

```
## Rows: 36 Columns: 6
```

```
## ── Column specification
────────────────────────────────────────────────────────
```

```
## Delimiter: ","
## chr (1): sexo
## dbl (5): edad, peso, altura, muneca, biceps
##
## ℹ Use `spec()` to retrieve the full column specification for this
data.
```

```
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

```
head(data1)
```

```
## # A tibble: 6 × 6
##    edad  peso altura sexo   muneca biceps
##   <dbl> <dbl>  <dbl> <chr>   <dbl>  <dbl>
## 1    43  87.3   188  Hombre   12.2   35.8
## 2    65  80     174  Hombre   12     35
## 3    45  82.3   176. Hombre   11.2   38.5
## 4    37  73.6   180. Hombre   11.2   32.2
## 5    55  74.1   168. Hombre   11.8   32.9
## 6    33  85.9   188  Hombre   12.4   38.5
```

```
# Verificar los nombres de las columnas
colnames(data1)
```

```
## [1] "edad"   "peso"   "altura" "sexo"   "muneca" "biceps"
```

```
# Verificar los nombres de las columnas
colnames(data1)
```

```
## [1] "edad"   "peso"   "altura" "sexo"   "muneca" "biceps"
```

```
# Eliminar la columna 'sexo' usando el nombre exacto
data <- data1 %>% select(-sexo)

# Mostrar las primeras filas del nuevo dataframe
head(data)
```

```
## # A tibble: 6 × 5
##    edad  peso altura muneca biceps
##   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
## 1    43  87.3   188    12.2   35.8
## 2    65  80     174    12     35
## 3    45  82.3   176.   11.2   38.5
## 4    37  73.6   180.   11.2   32.2
## 5    55  74.1   168.   11.8   32.9
## 6    33  85.9   188    12.4   38.5
```

```
# Medidas descriptivas básicas
summary(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
```

Observaciones * Al observar la distribución general de las variables, podemos ver variaciones considerables sobretodo en los rangos de edad, peso y altura.

```
# Cálculo de la desviación estándar para cada variable
sapply(data, sd, na.rm = TRUE)
```

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

Observaciones * Analizar la desviación estandar nos permite tener una idea de la disperción o variabilidad de los datos en torno a la media para cada variable. En este caso, edad, peso y altura tienen un alta variabilidad, mientras que muñeca y bíceps tienen menor variabilidad. Al observar esto, y con el propósito de realizar PCA, es importante estandarizar las variables para evitar que las que tienen mayor dispersión dominen el análisis.

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.

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:

```
# Matriz de varianzas-covarianzas (S)
S <- cov(data)
cat("Varianzas-Covarianzas (S):", "\n")
```

```
## Varianzas-Covarianzas (S):
```

```
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
```

Observaciones * La matriz de varianzas-covarianzas muestra la relación entre cada par de variables en términos de su varianza conjunta. * La **diagonal** contiene las varianzas de cada variable y estas son las mismas que las desviaciones estándar al

cuadrado. * Las mayores varianzas se observan en edad, peso y altura. * Los valores **fuera d ela diagonal** representan las covarianzas entre pares de variables. * En este caso, todas las covarianzas muestran relaciones positivas entre las variables, lo que indica que a medida que una aumente, la otra también lo hará. * Las relaciones más fuertes son entre **peso, altura y edad.**

```
# Matriz de correlaciones (R)
R <- cor(data)
cat("Correlaciones (R):", "\n")

## Correlaciones (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
```

Observaciones

- La matriz de correlaciones muestra las relaciones lineales entre las variables, estandarizadas entre -1 y 1. Los valores cercanos a 1 indican una correlación positiva fuerte, mientras que los cercanos a -1 indican una correlación negativa fuerte.

Describe las correlaciones que se establecen entre las variables.

- **Edad:** tiene correlaciones moderadas con peso y muñeca. Esto suguiere que a medida que aumenta la edad, también lo hacen el peso y la circunferencia de la muñeca.
- **Peso:** tiene una correlación alta con altura, muñeca y bícep. Esto indica que las personas con mayor peso tienden a tener mayor altura, circunferencia de muñeca y tamaño de bíceps.
- **Altura:** esta moderadamemte correlacionada con muñeca y bíceps. Esto indica que las personas más altas también tienden a tener mayor circunferencia de muñeca y bíceps.
- **Muñeca:** Tiene correlaciones muy altas con pero, bíceps y edad. Esto indica que la circunferencia de la muñeca está fuertemente relacionada con el peso y el tamaño del bíceps, y también tiene una relación moderada con la edad.
- **Bíceps:** Muestra correlaciones fuertes con peso, altura y muñeca, lo que indica que las personas con bíceps más grandes suelen tener mayor peso, altura y circunferencia de muñeca.

- Las variables estan interrelacionadas en el contexto del tamaño corporal general. La edad tiene correlaciones más moderadas, lo que indica que no es tan determinante en las variaciones de las otras medidas.

1.1 Calcule los valores y vectores propios de cada matriz. La función en R es: eigen().

Notas

- eigen_values contiene los valores propios (la varianza explicada por cada componente).
- eigen_vectors contiene los vectores propios (las combinaciones lineales de las variables originales que forman los componentes principales).

```
# Calcular los valores y vectores propios de la matriz de varianzas-covarianzas
eigen_S <- eigen(S)

cat("\n Valores propios de la matriz de varianzas-covarianzas (S):", "\n")

##
##  Valores propios de la matriz de varianzas-covarianzas (S):

eigen_S$values

## [1] 359.3980243  80.3757858  27.6229011   4.3074318   0.2343571

cat("\n Vectores propios de la matriz de varianzas-covarianzas (S):", "\n")

##
##  Vectores propios de la matriz de varianzas-covarianzas (S):

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
```

Observaciones

- **Valores propios:** El primer componente explica la mayor parte de la varianza con un valor propio de 359.40. Esto indica que este componente captura la mayor variabilidad en los datos.
- **Vectores propios:** Estos muestran las combinaciones lineales de las variables originales que forman cada componente principal. El primer vector propio

tiene una fuerte contribución de peso y altura, lo que sugiere que este componente está principalmente relacionado con estas variables.

```
# Calcular los valores y vectores propios de la matriz de correlaciones
eigen_R <- eigen(R)


cat("\n Valores propios de la matriz de correlaciones (R):", "\n")

##
##  Valores propios de la matriz de correlaciones (R):

eigen_R$values

## [1] 3.75749733 0.72585665 0.32032981 0.12461873 0.07169749

cat("\n Vectores propios de la matriz de correlaciones (R):", "\n")

##
##  Vectores propios de la matriz de correlaciones (R):

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
```

Observaciones

- **Valores propios:** El primer valor propio es 3.76, lo que indica que el primer componente principal captura la mayor parte de la varianza estandarizada.
- **Vectores propios:** El primer vector propio tiene una contribución significativa de peso, muñeca y bíceps, lo que significa que estas variables son las que más contribuyen al primer componente principal en la matriz estandarizada.
- **La matriz de varianzas-covarianzas** captura mejor la varianza de los datos originales no estandarizados, por lo que es útil cuando queremos trabajar con las variables en sus escalas originales.
- **La matriz de correlaciones** es más adecuada para comparar variables en diferentes escalas, como en este caso, donde peso, altura, edad, etc., están en diferentes unidades.

1.2 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.

```
# Varianza total (suma de las varianzas en la diagonal de S)
varianza_total_S <- sum(diag(S))

# Proporción de varianza explicada por cada componente (S)
prop_varianza_S <- eigen_S$values / varianza_total_S

cat("Proporción de varianza explicada por cada componente (S):", "\n")

## Proporción de varianza explicada por cada componente (S):

prop_varianza_S

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

Observaciones * El primer componente principal explica 76.15% de la varianza total. Esto indica que la mayor parte de la variabilidad en los datos está capturada por este componente.

```
# Varianza total en la matriz de correlaciones (suma de los valores
propios = número de variables, pues están estandarizadas)
varianza_total_R <- sum(eigen_R$values)

# Proporción de varianza explicada por cada componente (R)
prop_varianza_R <- eigen_R$values / varianza_total_R

cat("Proporción de varianza explicada por cada componente (R):", "\n")

## Proporción de varianza explicada por cada componente (R):

prop_varianza_R

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

Observaciones * El primer componente principal explica 75.15% de la varianza estandarizada, lo cual es muy similar al análisis de la matriz de varianzas-covarianzas.

```
# Verificación: la suma de las varianzas en la diagonal de S es igual a
la suma de los valores propios

cat("Varianza total en la diagonal de S:", varianza_total_S, "\n")

## Varianza total en la diagonal de S: 471.9385
```

```r
cat("Suma de los valores propios de S:", sum(eigen_S$values), "\n")
```

```
## Suma de los valores propios de S: 471.9385
```

```r
# Verificación: la suma de las varianzas en la diagonal deRS es igual a la suma de los valores propios

cat("Varianza total en la diagonal de S:", varianza_total_R, "\n")
```

```
## Varianza total en la diagonal de S: 5
```

```r
cat("Suma de los valores propios de S:", sum(eigen_R$values), "\n")
```

```
## Suma de los valores propios de S: 5
```

Interpretación

- En ambos casos, el primer componente principal es el más importante.
- Los primeros dos componentes juntos explican más del 90% de la varianza en ambos casos, lo que indica que estos dos componentes principales capturan la mayor parte de la información de los datos originales.
- La matriz de correlaciones muestra una distribución de la varianza más equitativa entre los primeros componentes, mientras que la matriz de varianzas-covarianzas da más peso al primer componente.

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

Notas * Se acumulan las proporciones para obtener la varianza acumulada, es decir, la suma progresiva de la varianza explicada por cada componente.

- varianza_acumulada son los vectores que contienen la proporción acumulada de la varianza explicada por los componentes principales.

```r
# Cálculo de la proporción de varianza acumulada para la matriz de varianzas-covarianzas
varianza_acumulada_S <- cumsum(prop_varianza_S)

cat("Varianza acumulada por componente (S):", "\n")
```

```
## Varianza acumulada por componente (S):
```

```r
varianza_acumulada_S
```

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

```r
# Cálculo de la proporción de varianza acumulada para la matriz de correlaciones
varianza_acumulada_R <- cumsum(prop_varianza_R)

cat("Varianza acumulada por componente (R):", "\n")
```

```
## Varianza acumulada por componente (R):

varianza_acumulada_R

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

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

Tanto en la matriz de varianzas-covarianzas (S) como en la matriz de correlaciones (R), los primeros dos componentes explican la mayor parte de la varianza (más del 90%), lo que sugiere que se podría reducir la dimensionalidad del problema utilizando solo dos componentes sin perder demasiada información.

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, …))

- X1 = Edad
- X2 = Peso
- X3 = Altura
- X4 = Muñeca
- X5 = Bíceps

Con la matriz de varianzas-covarianzas (S)

Ecuación de (CP1) (Componente Principal 1):

$$CP1 = e1_1 \cdot X1 + e1_2 \cdot X2 + e1_3 \cdot X3 + e1_4 \cdot X4 + e1_5 \cdot X5$$

Donde: (e1_1), (e1_2), …, (e1_5) son los valores del primer vector propio (eigen(S)$vectors).

Para (CP1), utilizando los valores del vector propio de la matriz de varianzas-covarianzas (S):

$$CP1 = (-0.3487) \cdot X1 + (-0.7662) \cdot X2 + (-0.4763) \cdot X3 + (-0.0539) \cdot X4 + (-0.2481) \cdot X5$$

Ecuación de (CP2) (Componente Principal 2):

$$CP2 = e2_1 \cdot X1 + e2_2 \cdot X2 + e2_3 \cdot X3 + e2_4 \cdot X4 + e2_5 \cdot X5$$

Donde: (e2_1), (e2_2), …, (e2_5) son los valores del segundo vector propio (eigen(S)$vectors).

$$CP2 = (0.9076) \cdot X1 + (-0.1616) \cdot X2 + (-0.3852) \cdot X3 + (0.0155) \cdot X4 + (-0.0402) \cdot X5$$

Con la matriz de correlaciones (R)

Ecuación de (CP1) (Componente Principal 1):

(e1_1), (e1_2), …, (e1_5) son los valores del primer vector propio (eigen(R)$vectors).

Para (CP1), utilizando los valores del vector propio de la matriz de varianzas-covarianzas (R):

$$CP1 = (-0.3359) \cdot X1 + (-0.4927) \cdot X2 + (-0.4222) \cdot X3 + (-0.4821) \cdot X4 + (-0.4833) \cdot X5$$

Ecuación de (CP2) (Componente Principal 2):

(e2_1), (e2_2), …, (e2_5) son los valores del segundo vector propio (eigen(R)$vectors).

$$CP2 = (0.8575) \cdot X1 + (-0.1647) \cdot X2 + (-0.4542) \cdot X3 + (0.1082) \cdot X4 + (-0.1392) \cdot X5$$

1.5.1 ¿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.

Con la matriz de varianzas-covarianzas (S)

CP1:

Los valores absolutos de los coeficientes son:

- -0.3487 | = 0.3487 (Edad)
- -0.7662 | = 0.7662 (Peso)
- -0.4763 | = 0.4763 (Altura)
- -0.0539 | = 0.0539 (Muñeca)
- -0.2481 | = 0.2481 (Bíceps)

Análisis de CP1: * La variable que más contribuye a CP1 es **Peso (X2)**, con un coeficiente de 0.7662. * La segunda variable que más contribuye es **Altura (X3)**, con un coeficiente de 0.4763. * **Muñeca (X4)** tiene la menor contribución, con un coeficiente muy pequeño (0.0539).

CP2:

Los valores absolutos de los coeficientes son:

- 0.9076 | = 0.9076 (Edad)
- -0.1616 | = 0.1616 (Peso)
- -0.3852 | = 0.3852 (Altura)
- 0.0155 | = 0.0155 (Muñeca)

- -0.0402 | = 0.0402 (Bíceps)

Análisis de CP2: - La variable que más contribuye a CP2 es **Edad (X1)**, con un coeficiente de 0.9076. - La segunda variable que más contribuye es **Altura (X3)**, con un coeficiente de 0.3852. - **Muñeca (X4)** tiene la menor contribución, con un coeficiente muy pequeño (0.0155).

Justificación

La contribución de cada variable a un componente principal está determinada por el valor absoluto de su coeficiente en las combinaciones lineales. En **CP1**, **Peso** es la variable más significativa debido a su coeficiente elevado, mientras que en **CP2**, **Edad** tiene la mayor influencia. Las variables con coeficientes pequeños, como **Muñeca**, tienen poca influencia en ambos componentes.

Con la matriz de correlaciones (R)

CP1:

Los valores absolutos de los coeficientes son:

- -0.3359 | = 0.3359 (Edad)
- -0.4927 | = 0.4927 (Peso)
- -0.4222 | = 0.4222 (Altura)
- -0.4821 | = 0.4821 (Muñeca)
- -0.4833 | = 0.4833 (Bíceps)

Análisis de CP1: * La variable que más contribuye a CP1 es **Peso (X2)**, con un coeficiente de 0.4927. * La segunda variable que más contribuye es **Bíceps (X5)**, con un coeficiente de 0.4833. * **Edad (X1)** tiene la menor contribución, con un coeficiente muy pequeño (0.3359).

CP2:

Los valores absolutos de los coeficientes son:

- 0.8575 | = 0.8575 (Edad)
- -0.1647 | = 0.1647 (Peso)
- -0.4542 | = 0.4542 (Altura)
- 0.1082 | = 0.1088 (Muñeca)
- -0.1392 | = 0.1392 (Bíceps)

Análisis de CP2: - La variable que más contribuye a CP2 es **Edad (X1)**, con un coeficiente de 0.8575 - La segunda variable que más contribuye es **Altura (X3)**, con

un coeficiente de 0.4542. - **Muñeca (X4)** tiene la menor contribución, con un coeficiente muy pequeño (0.1088).

Justificación

La contribución de cada variable a un componente principal está determinada por el valor absoluto de su coeficiente en las combinaciones lineales. En **CP1**, **Peso** es la variable más significativa debido a su coeficiente elevado, mientras que en **CP2**, **Edad** tiene la mayor influencia. Las variables con coeficientes pequeños, como **Muñeca**, tienen poca influencia en ambos componentes.

PARTE II

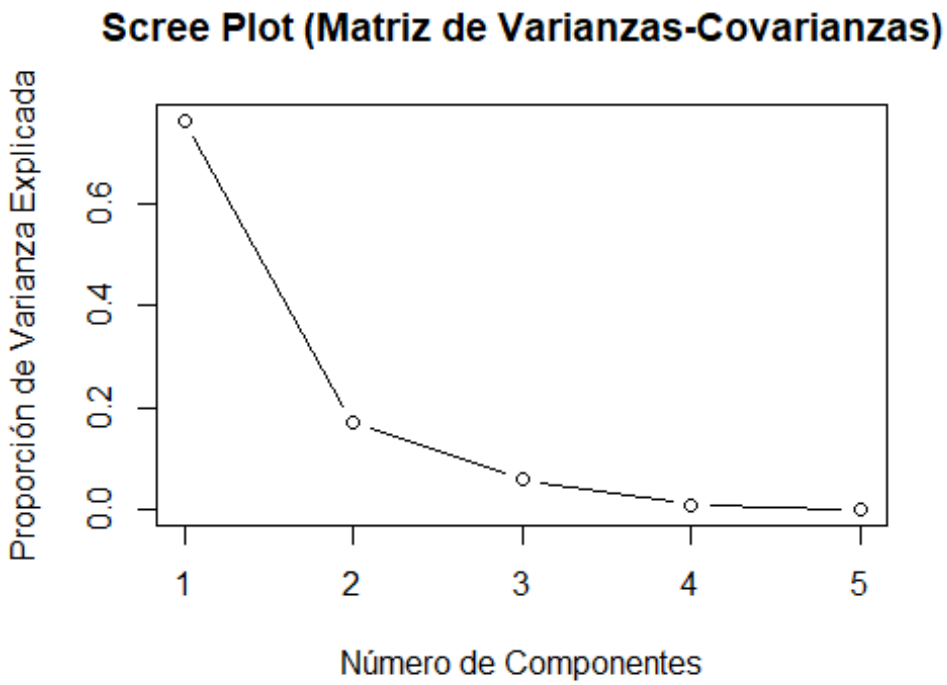
1. Obtenga las gráficas respectivas con S (matriz de varianzas-covarianzas) y con R (matriz de correlaciones) de las dos primeras componentes.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

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

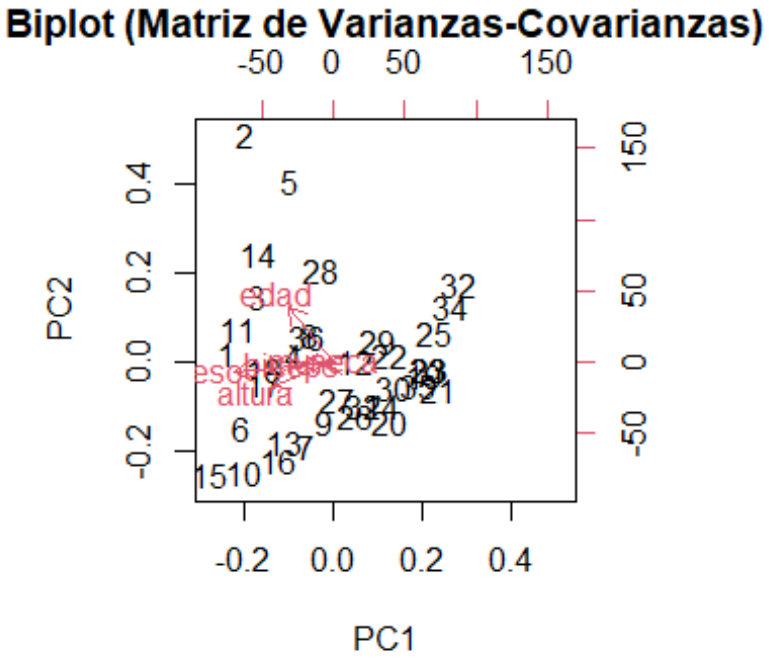
```
# Obtener las puntuaciones (scores) para la matriz de varianzas-
covarianzas

# Las puntuaciones son obtenidas multiplicando los datos centrados por
los vectores propios

# scores_S <- pca_S$x

 scores_S <- as.matrix(scale(data, center = TRUE, scale = FALSE)) %*%
eigen_S$vectors

# Cálculo de PCA para la matriz de varianzas-covarianzas
pca_S <- prcomp(data, scale = FALSE)

# Scree plot para la matriz de varianzas-covarianzas
var_explicada_S <- pca_S$sdev^2 / sum(pca_S$sdev^2)
plot(var_explicada_S, type = "b", main = "Scree Plot (Matriz de
Varianzas-Covarianzas)",
     xlab = "Número de Componentes", ylab = "Proporción de Varianza
Explicada")
```



Interpretación

- Este gráfico muestra la proporción de la varianza explicada por cada uno de los componentes principales.
- Se observa que el primer componente principal explica la mayor parte de la varianza (más del 60%), seguido del segundo componente con una proporción menor.
- A partir del tercer componente, la proporción de varianza explicada es mucho menor, lo que indica que los primeros dos componentes capturan la mayor parte de la información en los datos.
- Los primeros componentes son los más significativos y podría ser suficiente trabajar solo con los primeros dos o tres componentes.

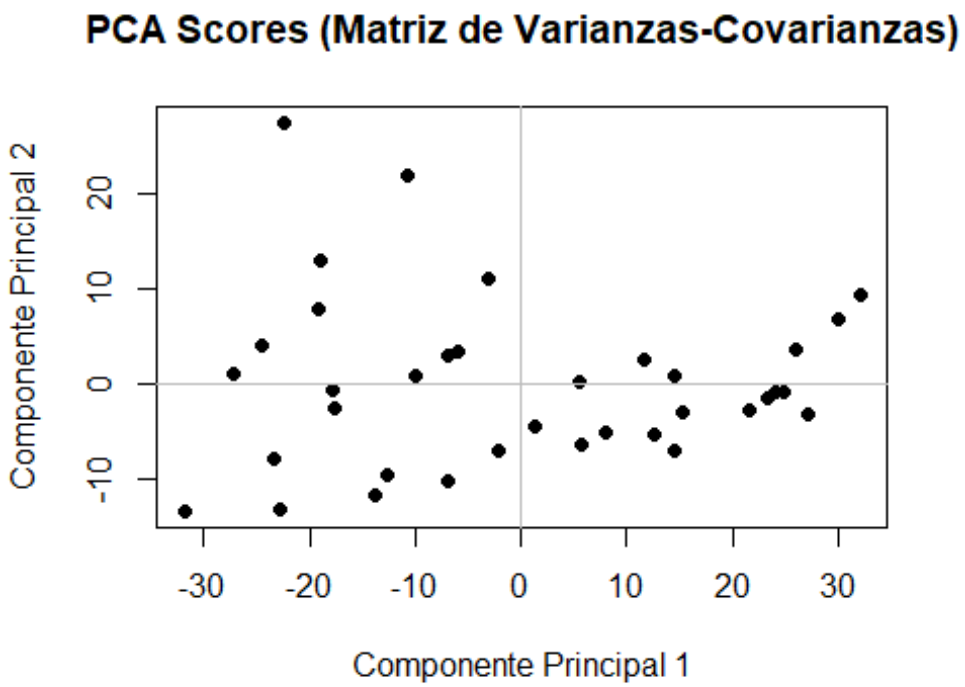
```
# Biplot para la matriz de varianzas-covarianzas
biplot(prcomp(data, scale = FALSE), main = "Biplot (Matriz de Varianzas-
Covarianzas)")
```



Interpretación

- Las variables como peso y altura tienen una mayor contribución al primer componente (PC1).
- Edad está más orientada hacia el segundo componente (PC2).

```
# Graficar las dos primeras componentes principales para la matriz de varianzas-covarianzas
plot(scores_S[, 1], scores_S[, 2], main = "PCA Scores (Matriz de Varianzas-Covarianzas)",
     xlab = "Componente Principal 1", ylab = "Componente Principal 2", pch = 19)
abline(h = 0, v = 0, col = "gray")
```



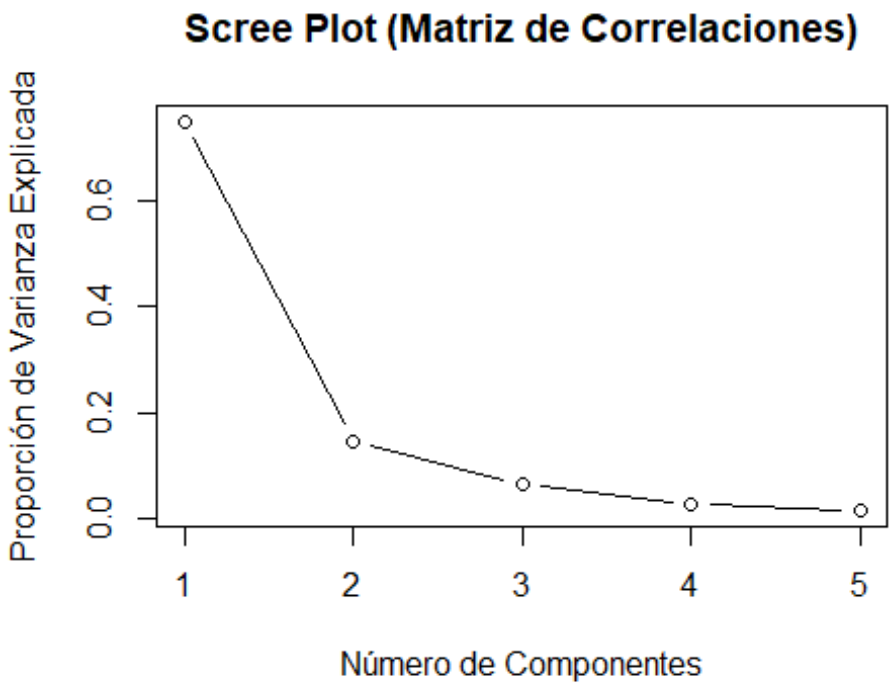
Interpretación * Este gráfico muestra cómo las observaciones (individuos) están distribuidas en función de las dos primeras componentes principales.

- Se observan algunos puntos que están más alejados, lo que sugiere que estas observaciones tienen características que difieren más del promedio de la muestra (posibles outliers).
- Las observaciones distribuidas en la parte positiva del eje del Componente Principal 1 parecen estar asociadas con valores altos en variables como peso y altura, mientras que aquellas en la parte negativa del eje PC1 podrían estar asociadas con valores más bajos en esas variables.

1.3 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.

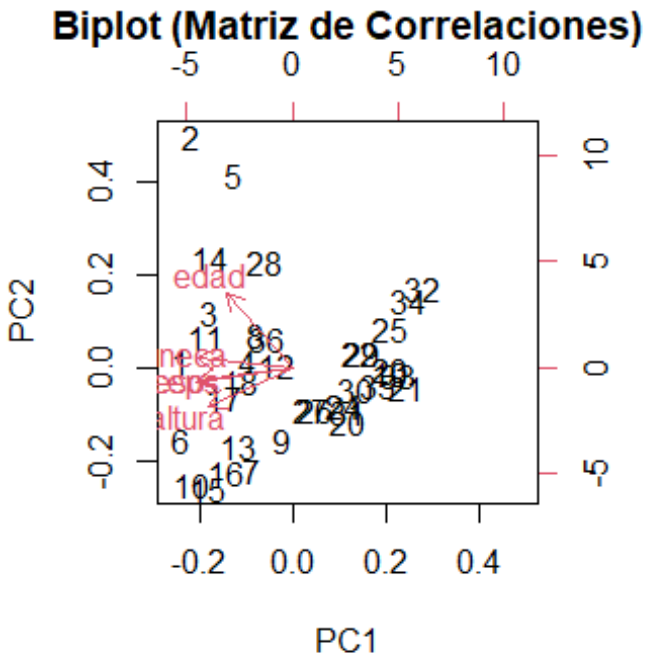
```
# Obtener las puntuaciones (scores) para la matriz de correlaciones
(variables estandarizadas)
# scores_R <- pca_R$x

scores_R <- as.matrix(scale(data)) %*% eigen_R$vectors

# Cálculo de PCA para la matriz de correlaciones
pca_R <- prcomp(data, scale = TRUE)

# Scree plot para la matriz de correlaciones
var_explicada_R <- pca_R$sdev^2 / sum(pca_R$sdev^2)
```

```r
plot(var_explicada_R, type = "b", main = "Scree Plot (Matriz de
Correlaciones)",
     xlab = "Número de Componentes", ylab = "Proporción de Varianza
Explicada")
```



Interpretación * Se observa que el primer componente explica una proporción significativa de la varianza, superior al 60%. El segundo componente también tiene una contribución importante, aunque menor. * A partir del tercer componente, la proporción de varianza explicada disminuye considerablemente, lo que sugiere que los primeros dos componentes principales son suficientes para explicar la mayoría de la variabilidad en los datos.

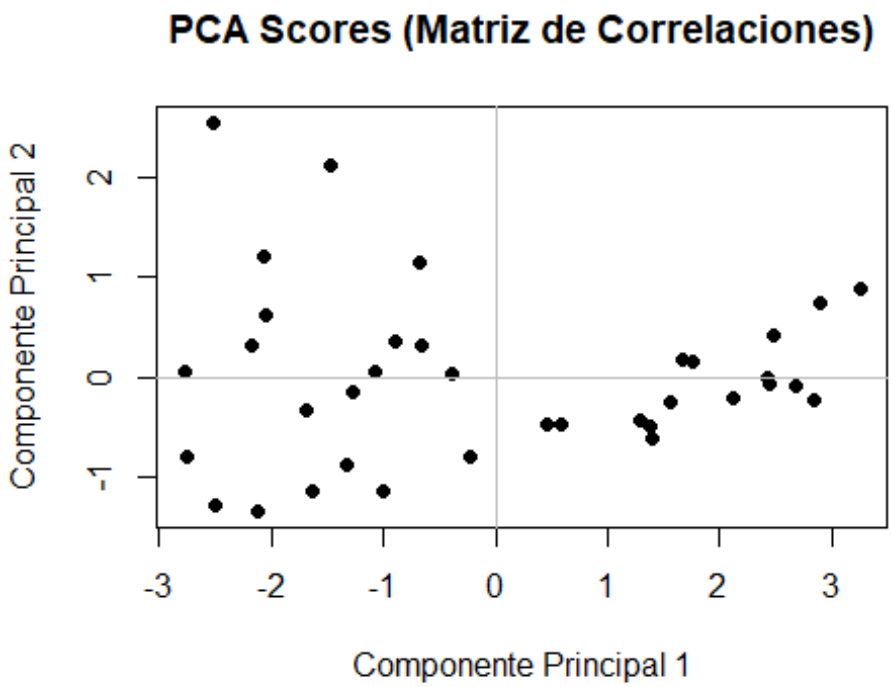
```r
# Biplot para la matriz de correlaciones (variables estandarizadas)
biplot(prcomp(data, scale = TRUE), main = "Biplot (Matriz de
Correlaciones)")
```



Interpretación

- Las variables como peso y altura parecen tener una mayor contribución al primer componente principal (PC1), mientras que edad tiene más influencia sobre el segundo componente (PC2).

```
# Graficar las dos primeras componentes principales para la matriz de correlaciones
plot(scores_R[, 1], scores_R[, 2], main = "PCA Scores (Matriz de Correlaciones)",
     xlab = "Componente Principal 1", ylab = "Componente Principal 2", pch = 19)
abline(h = 0, v = 0, col = "gray")
```



Interpretación

- A diferencia de la matriz de varianzas-covarianzas, las puntuaciones ahora están más equilibradas alrededor del origen debido a la estandarización de las variables.
- Las observaciones que se encuentran más alejadas del origen indican valores más extremos en las combinaciones lineales de las variables originales, mientras que aquellas más cercanas al origen tienen valores más promedio.
- Al igual que en los otros gráficos, parece que los primeros dos componentes capturan gran parte de la variabilidad en los datos, y los patrones en la dispersión pueden indicar posibles agrupamientos o diferencias entre las observaciones.

2. Interprete los gráficos en términos de:

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

Matriz de varianzas y covarianzas

- Peso y Altura tienen flechas largas y están más alineadas con el eje del primer componente principal (PC1), lo que sugiere que estas dos variables tienen una fuerte influencia en este componente.

- Edad tiene una flecha orientada más hacia el segundo componente (PC2), lo que sugiere que el segundo componente captura más variabilidad relacionada con la edad de los individuos.

Matriz de correlaciones

- Peso y Altura parecen contribuir fuertemente al primer componente principal (PC1), ya que las flechas de estas variables están alineadas con este eje. Esto sugiere que el primer componente refleja variaciones en el tamaño corporal de los individuos.
- Edad tiene más influencia en el segundo componente principal (PC2), lo que indica que este componente está más relacionado con la variabilidad en la edad de los individuos.

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

Matriz de varianzas y covarianzas

- Las observaciones ubicadas hacia los extremos positivos del eje de PC1 corresponden a individuos con valores más altos en peso y altura, dado que estas variables contribuyen fuertemente a este componente.
- Las observaciones más cercanas a los valores negativos de PC1 tienden a representar individuos con menor peso y altura.
- En cuanto a PC2, las observaciones con valores más altos pueden tener edades mayores, ya que la variable edad tiene una mayor influencia en este componente.

Matriz de correlaciones

- Las observaciones que se encuentran en los extremos del eje PC1 probablemente corresponden a individuos con valores extremos en variables como peso y altura, ya que estas variables contribuyen significativamente al primer componente.
- Las observaciones en los extremos del eje PC2 pueden representar individuos con valores más extremos en edad, ya que esta variable está más relacionada con el segundo componente.
- Las observaciones cercanas al origen del gráfico (0,0) tienen puntuaciones cercanas a la media en las variables originales, lo que indica que estos individuos no tienen valores extremos en ninguna de las variables.

2.3 Detecte posibles datos atípicos

Matriz de varianzas y covarianzas

- Observaciones que están alejadas del centro del gráfico, especialmente en los extremos de PC1 o PC2, podrían ser consideradas como datos atípicos o individuos con características que difieren considerablemente del resto del grupo.

Matriz de correlaciones

- Las observaciones que se encuentran muy alejadas del origen (0,0) en cualquiera de los ejes (PC1 o PC2) podrían ser consideradas como datos atípicos. Estas observaciones tienen valores muy diferentes de la media en las variables que más influyen en esos componentes.

3. Explora el: princomp() en library(stats).

Puedes poner help(princomp) en la consola o buscarlo en la ventana de ayuda.

```
library(stats)

# Aplicar princomp() para la matriz de varianzas-covarianzas
cpS <- princomp(data, cor = FALSE)

# Aplicar princomp() para la matriz de correlaciones
cpR <- princomp(data, cor = TRUE)
```

3.1 Indaga: ¿qué otras opciones tiene para facilitarte el análisis? En particular, explora los comandos y subcomandos: summary(cpS), cpaSloading, cpaSscores.

```
# Resumen del análisis de componentes principales para la matriz de varianzas-covarianzas
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
```

```
# Obtener las cargas (loadings) de los componentes principales (para la matriz de varianzas-covarianzas)
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
```

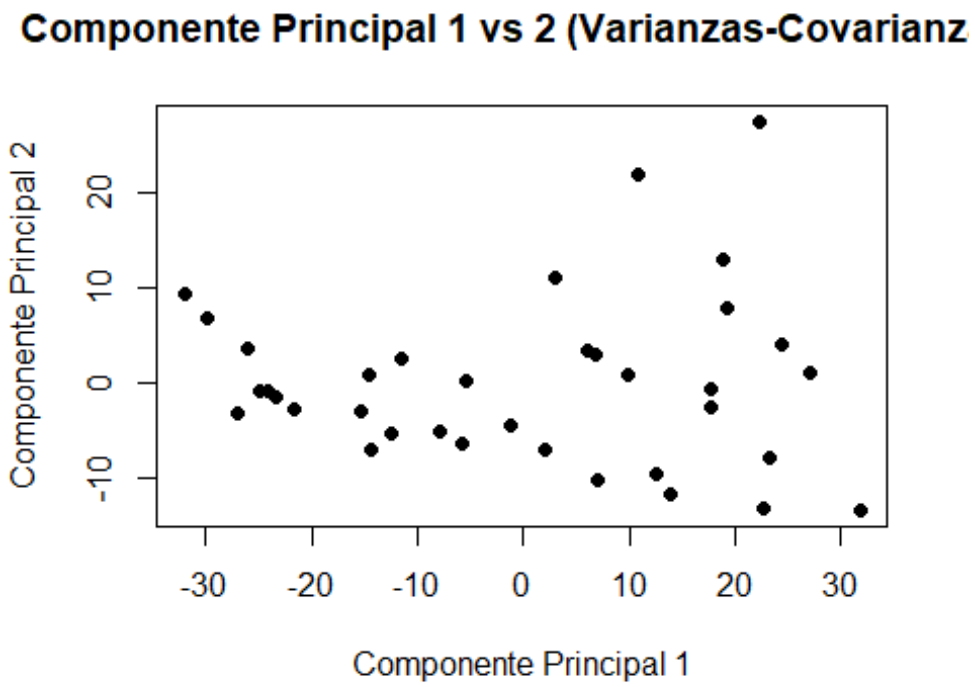
```
## biceps  0.248        -0.225 -0.931  0.138
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

```
# Obtener las puntuaciones (scores) de las observaciones para la matriz de varianzas-covarianzas
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
```

```
# Gráfico de los componentes principales (usando las puntuaciones)
plot(cpS$scores[,1], cpS$scores[,2], main = "Componente Principal 1 vs 2
```

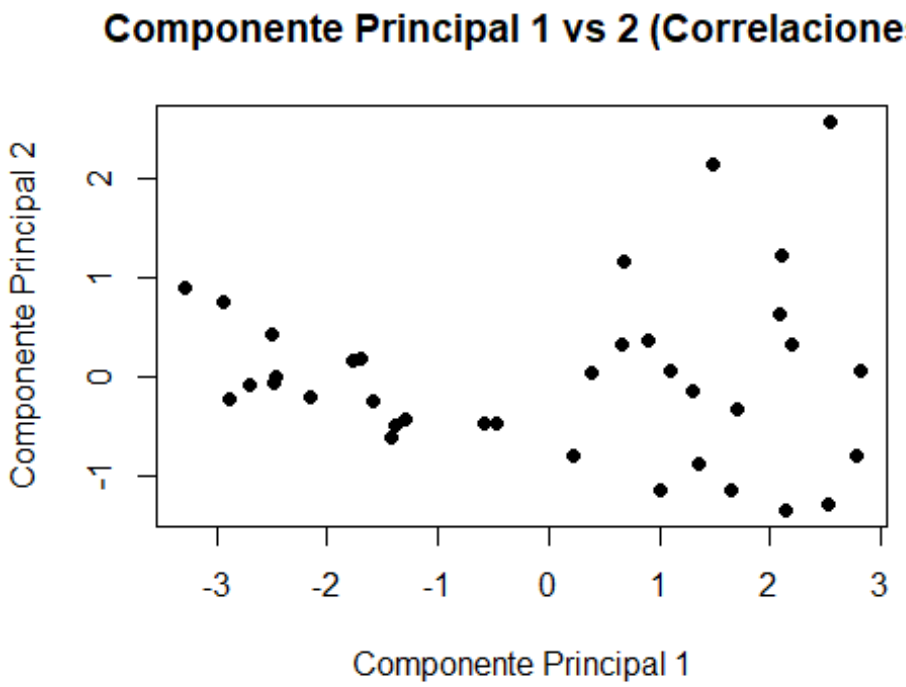
```
(Varianzas-Covarianzas)",
     xlab = "Componente Principal 1", ylab = "Componente Principal 2",
pch = 19)
```




```
# Resumen del análisis de componentes principales para la matriz de
correlaciones
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

# Cargas (loadings) para la matriz de correlaciones
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

# Puntuaciones (scores) de las observaciones para la matriz de
correlaciones
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

# Gráfico de los componentes principales (usando las puntuaciones)
plot(cpR$scores[,1], cpR$scores[,2], main = "Componente Principal 1 vs 2
```

```
(Correlaciones)",
     xlab = "Componente Principal 1", ylab = "Componente Principal 2",
pch = 19)
```



3.2 ¿Cómo se interpreta el resultado?

- cpS <- princomp(data, cor = FALSE): Ejecuta el análisis de componentes principales utilizando la matriz de varianzas-covarianzas. Si se desea usar la matriz de correlaciones, se establece cor = TRUE.
- summary(cpS): Proporciona un resumen del análisis, incluyendo la proporción de la varianza explicada por cada componente principal.
- cpS$loadings: Muestra las cargas (loadings) de los componentes principales. Esto indica cómo contribuyen las variables originales a los componentes principales.
- cpS$scores: Proporciona las puntuaciones (scores) de las observaciones en el espacio de los componentes principales. Estas puntuaciones indican la posición de cada observación en el nuevo espacio reducido.
- Gráficos: Los gráficos de las puntuaciones permiten visualizar cómo se distribuyen las observaciones en el espacio de los componentes principales.

Interpretación general de los gráficos

- Tanto en la matriz de varianzas-covarianzas como en la matriz de correlaciones, el primer componente principal explica una gran parte de la

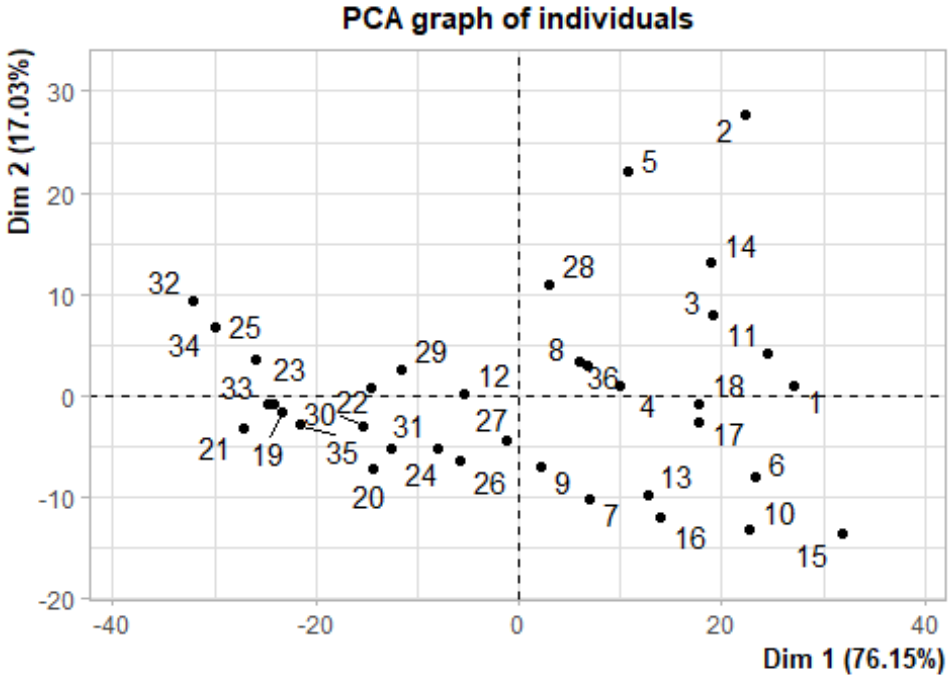
varianza, con una fuerte influencia de las variables relacionadas con el tamaño corporal, como peso y altura.

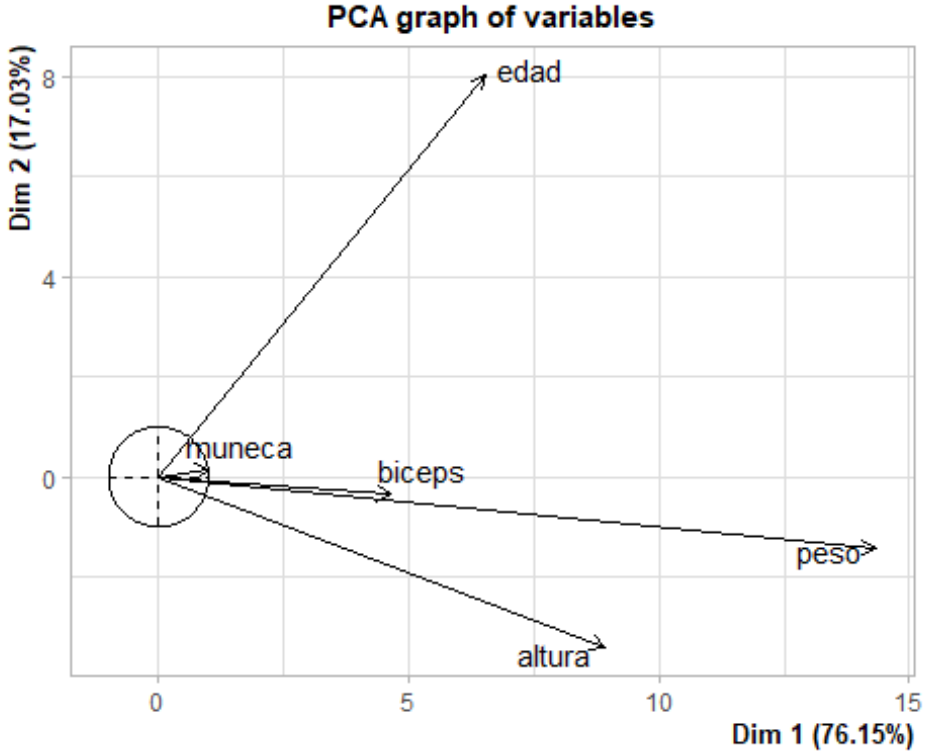
- El segundo componente principal en ambos casos está principalmente relacionado con la variable edad, lo que refleja que hay una fuente importante de variabilidad en la edad de los individuos.
- En ambos casos, los dos primeros componentes explican más del 90% de la varianza, lo que sugiere que una gran parte de la información puede ser capturada usando solo estos dos componentes, permitiendo una reducción significativa de la dimensionalidad del problema.

PARTE 3

1. Explore los siguientes gráficos relativos a Componentes Principales.

```
# Cargar las librerías necesarias
library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 4.3.3

library(ggplot2)
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.3

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

# Para la matriz de varianzas-covarianzas
cpS = PCA(data, scale.unit = FALSE)  # PCA usando la matriz de varianzas-
covarianzas
```



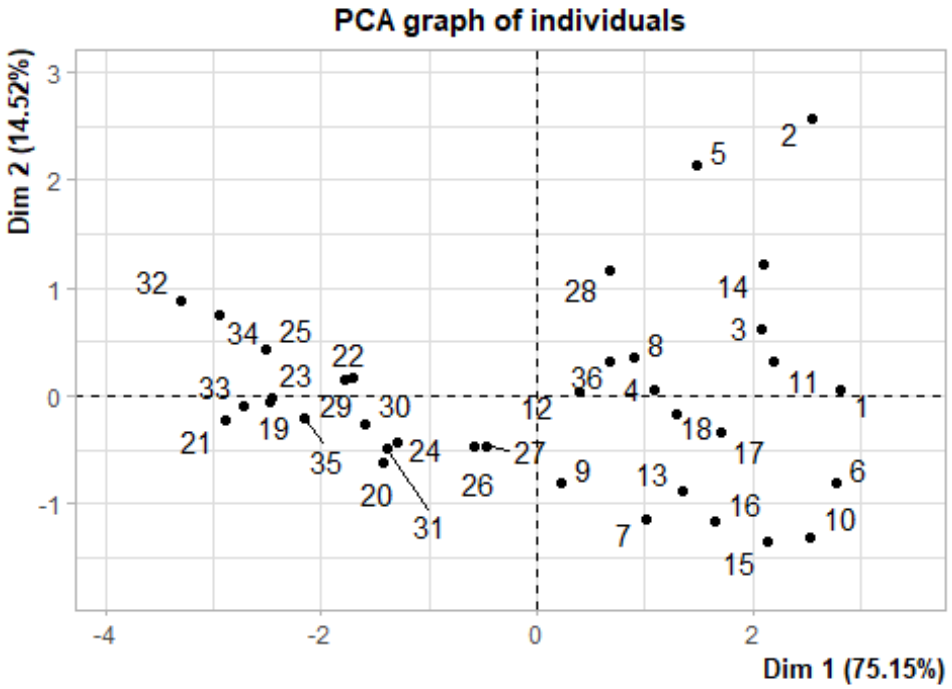


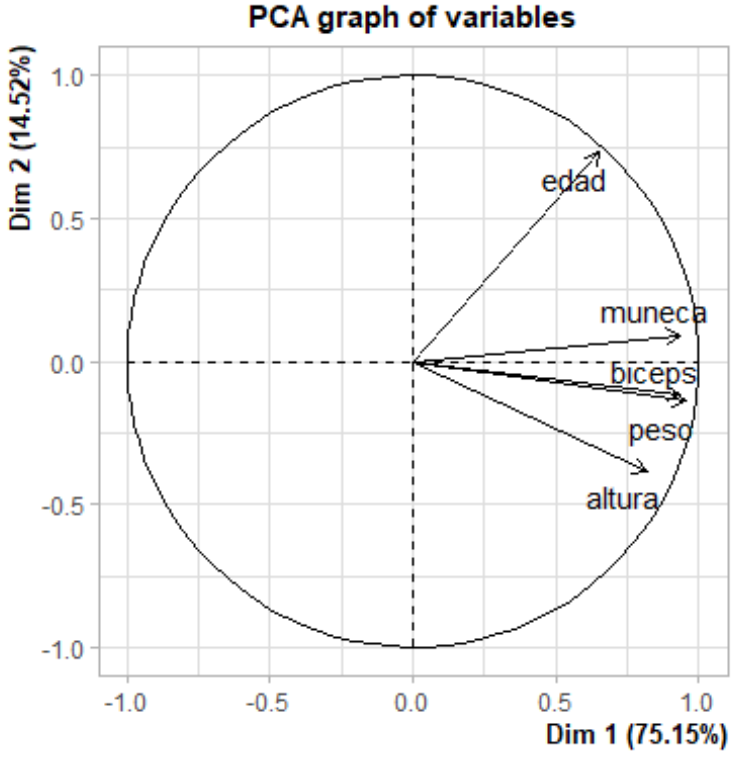
Interpretación de los graficos

- El primer componente principal (Dim 1) parece capturar las variaciones en Peso y Altura, que son las variables que más contribuyen a la varianza total.

- El segundo componente principal (Dim 2) está más influido por la Edad.
- Las observaciones que están lejos del origen en el gráfico de individuos probablemente tienen valores extremos en una combinación de estas variables (por ejemplo, personas más pesadas y altas).
- Las flechas en el gráfico de variables sugieren que Peso y Altura están altamente correlacionadas y definen las diferencias clave entre los individuos en el primer componente.

```
# Para la matriz de correlaciones
cpR = PCA(data, scale.unit = TRUE)   # PCA usando la matriz de
correlaciones
```



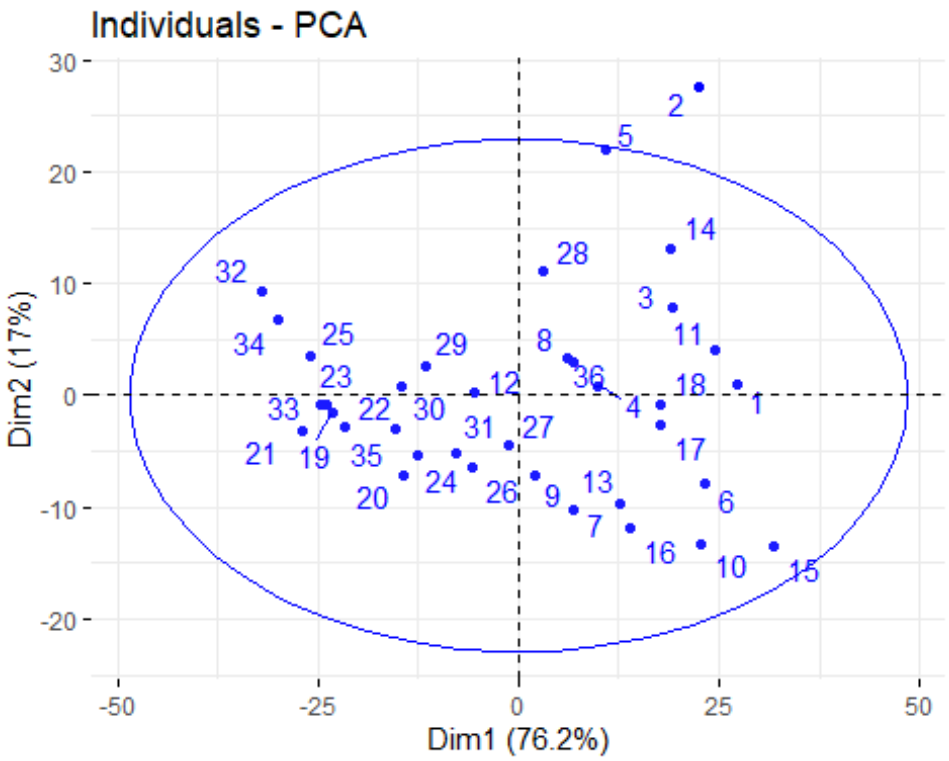


Interpretación

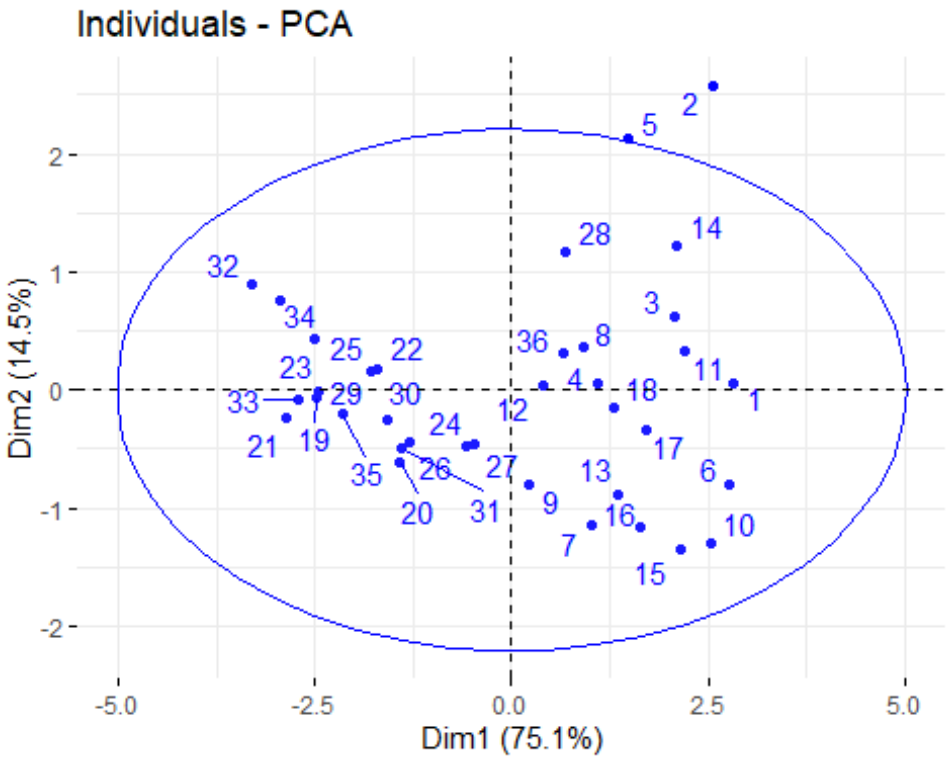
- Dim 1 refleja principalmente las diferencias en tamaño corporal (Peso, Altura, Bíceps), mientras que Dim 2 captura información relacionada con Edad.

- Las variables altamente correlacionadas entre sí se proyectan en direcciones similares (por ejemplo, Peso y Altura), mientras que las variables que no están correlacionadas se proyectan en ángulos más perpendiculares (como Edad en relación con Peso).

```
# 1. Gráfico de individuos (Proyección de las observaciones)
fviz_pca_ind(cpS, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```




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



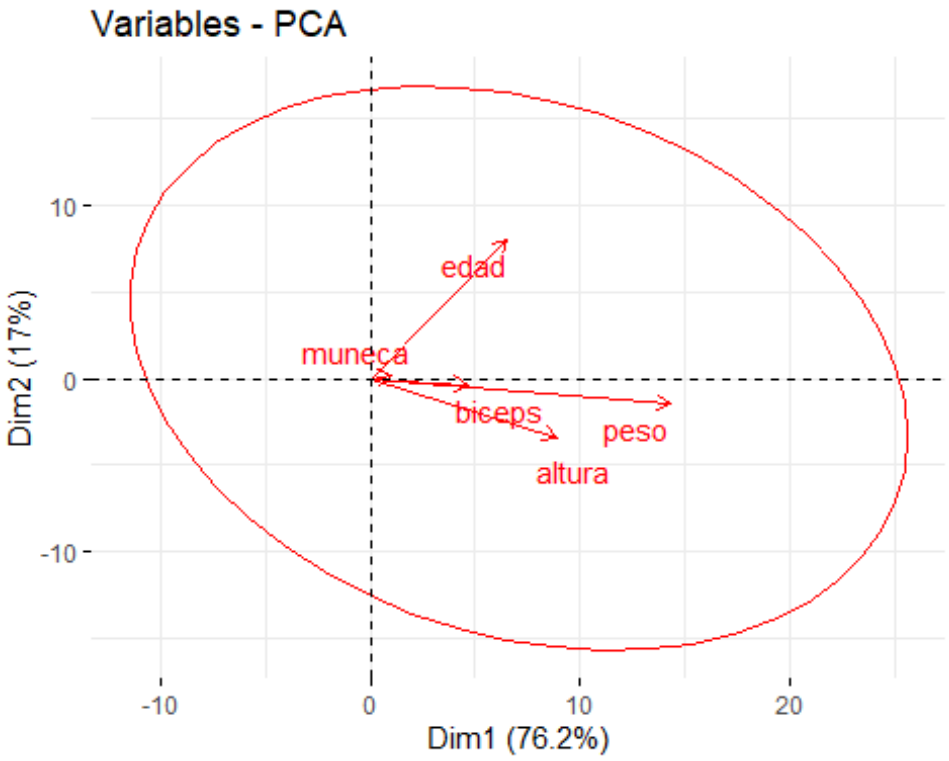
Interpretación

Martiz S * El gráfico sugiere que la Dim 1 es la más importante para diferenciar las observaciones, mientras que Dim 2 ofrece información adicional pero menos relevante. Las observaciones más alejadas del centro son las que tienen valores más extremos en las variables que más contribuyen a la varianza capturada por estos componentes principales.

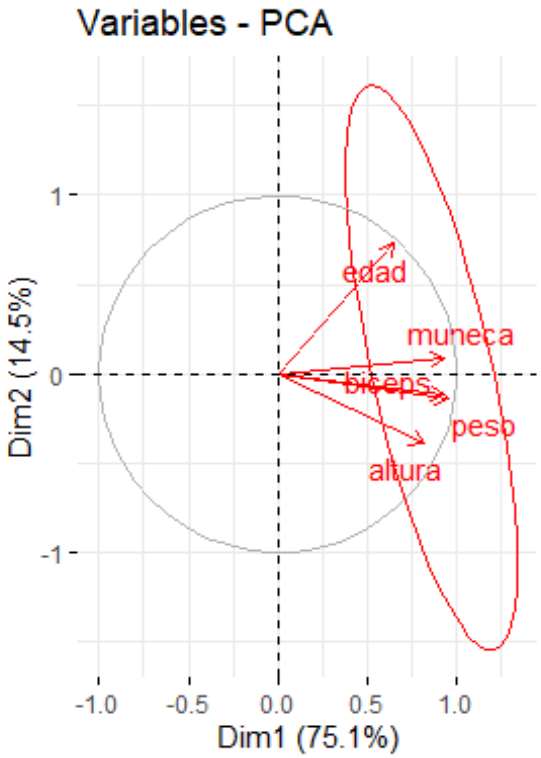
Matriz R * Dim 1 (75.1%): Explica la mayor parte de la variabilidad en los datos, lo que indica que las diferencias entre los individuos se explican principalmente por las variables relacionadas con el tamaño corporal (como peso y altura).

- Dim 2 (14.5%): Añade información adicional que podría estar relacionada con variables como edad, pero explica una cantidad mucho menor de la varianza. Observaciones clave:
- el gráfico sugiere que las diferencias en el tamaño corporal son la principal fuente de variabilidad entre los individuos, mientras que el segundo componente agrega información relacionada con otras características como la edad.

```
# 2. Gráfico de variables (Proyección de las variables originales)
fviz_pca_var(cpS, col.var = "red", addEllipses = TRUE, repel = TRUE)
```




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



Interpretación

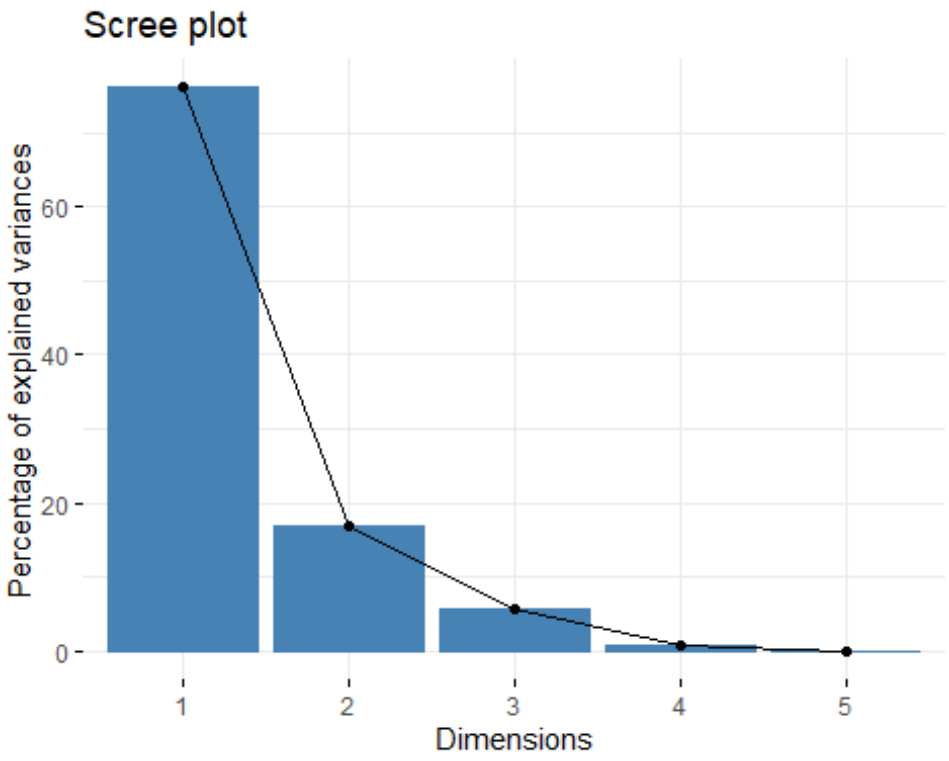
Matriz S

- Dim 1 (76.2%): Las variables peso, altura, y bíceps están altamente correlacionadas y contribuyen significativamente a este primer componente.
- Dim 2 (17%): La variable edad contribuye más al segundo componente, que explica una menor parte de la varianza.
- Las variables muñeca, bíceps, peso, y altura están alineadas, lo que indica que están altamente correlacionadas entre sí.
- Dim 1 captura principalmente las diferencias en el tamaño corporal, mientras que edad tiene una influencia más importante en Dim 2.

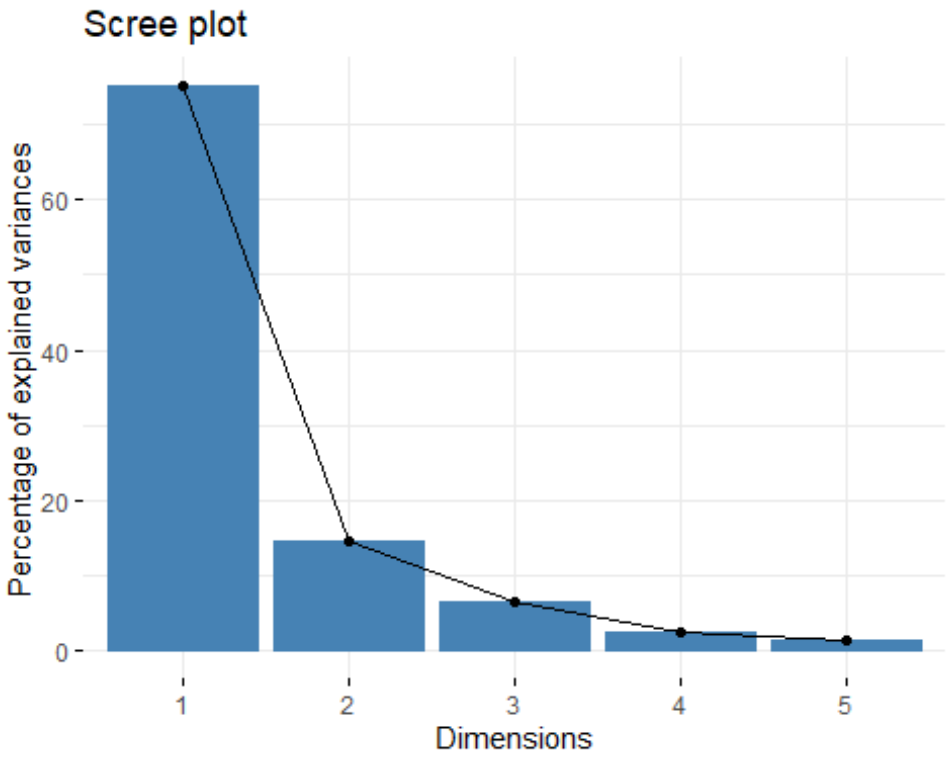
Matriz R

- Dim 1 (75.1%): Explica la mayor parte de la varianza, con variables como peso, bíceps, altura, y muñeca contribuyendo significativamente a este componente.
- Dim 2 (14.5%): La variable edad es la que más contribuye a este segundo componente.
- Las variables relacionadas con el tamaño corporal (peso, altura, bíceps, muñeca) están estrechamente agrupadas, lo que indica que están altamente correlacionadas entre sí.
- Dim 1 refleja principalmente el tamaño corporal, mientras que edad se correlaciona más con Dim 2, capturando otras diferencias entre los individuos.

```
# 3. Scree plot (Gráfico de codo - proporción de varianza explicada por cada componente)
fviz_screeplot(cpS)
```



```
fviz_screeplot(cpR)
```



Interpretación

Matriz S

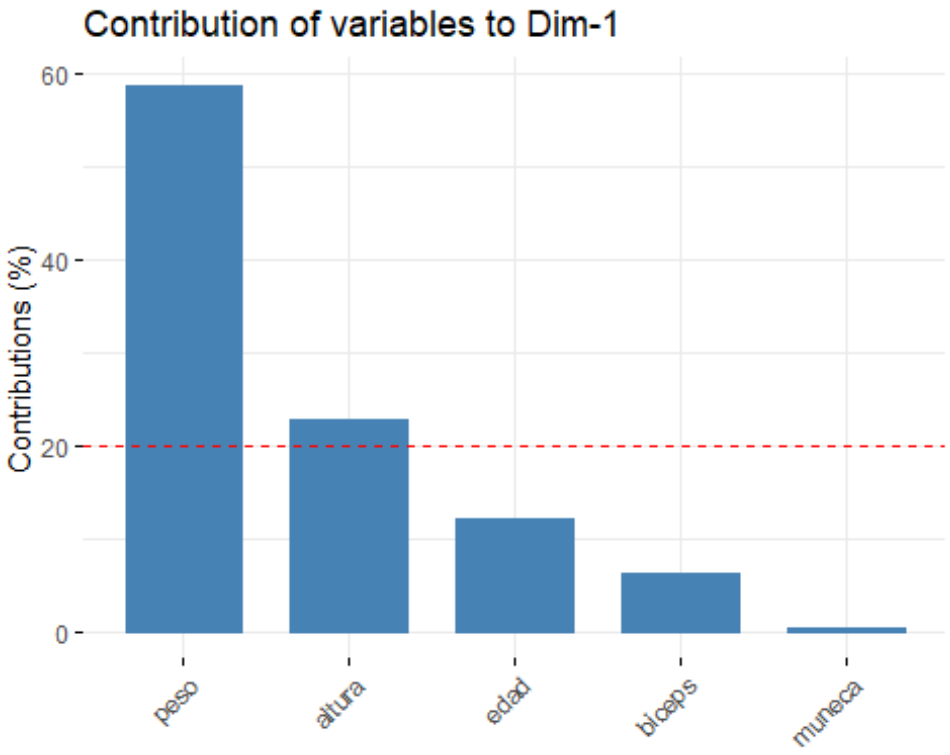
- Dim 1 explica más del 70% de la varianza, lo que sugiere que la mayor parte de la información de los datos está concentrada en este componente.
- Dim 2 explica alrededor del 20% de la varianza, lo que también es significativo, pero mucho menor que el primer componente.
- A partir del tercer componente, la varianza explicada es mucho menor, lo que indica que los primeros dos componentes son los más importantes para capturar la estructura de los datos.
- El gráfico sugiere que se puede reducir dimensionalidad conservando solo los primeros dos componentes, ya que capturan la mayor parte de la información.

Matriz R

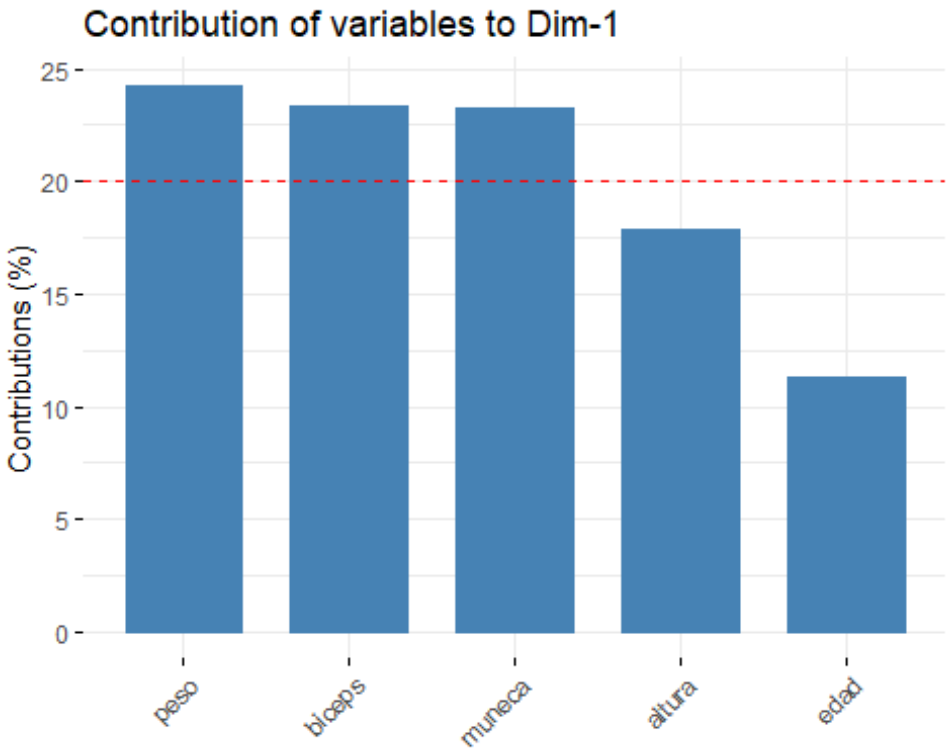
- El primer componente explica más del 70% de la varianza total, lo que sugiere que captura la mayor parte de la información presente en los datos.
- El segundo componente explica alrededor del 20% de la varianza.
- A partir del tercer componente, la cantidad de varianza explicada disminuye drásticamente, indicando que los primeros dos componentes principales son los más relevantes y que los demás componentes contribuyen muy poco a la explicación de la variabilidad de los datos.

- Se puede decir que los primeros dos componentes principales son suficientes para capturar la mayoría de la variabilidad en los datos, lo que permite reducir la dimensionalidad sin perder demasiada información.

```r
# 4. Gráfico de contribución de las variables a los componentes
fviz_contrib(cpS, choice = "var")
```




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



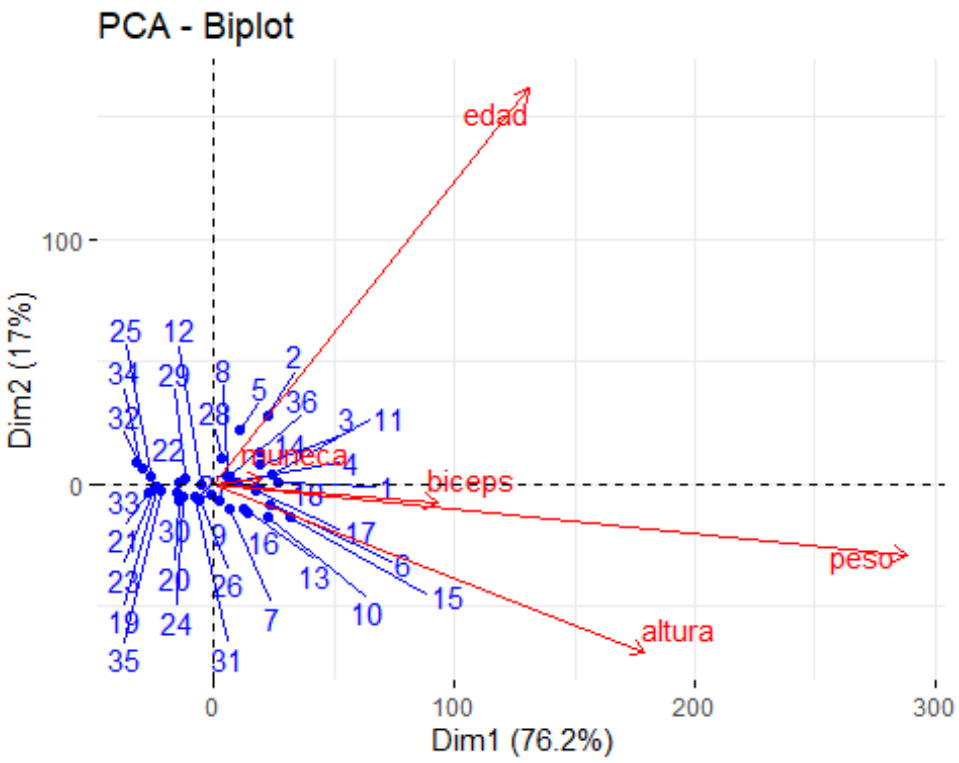
Interpretación

Matriz S * Peso tiene la mayor contribución, explicando cerca del 60% de la variabilidad en el primer componente. Esto indica que esta variable tiene una influencia significativa en el primer componente principal. * Altura y Edad también contribuyen de manera considerable, aunque en menor medida, con un poco más del 20% y menos del 20%, respectivamente. * Bíceps tiene una contribución pequeña y Muñeca apenas contribuye al primer componente principal. * El peso es la variable dominante en el primer componente, seguido de la altura y la edad.

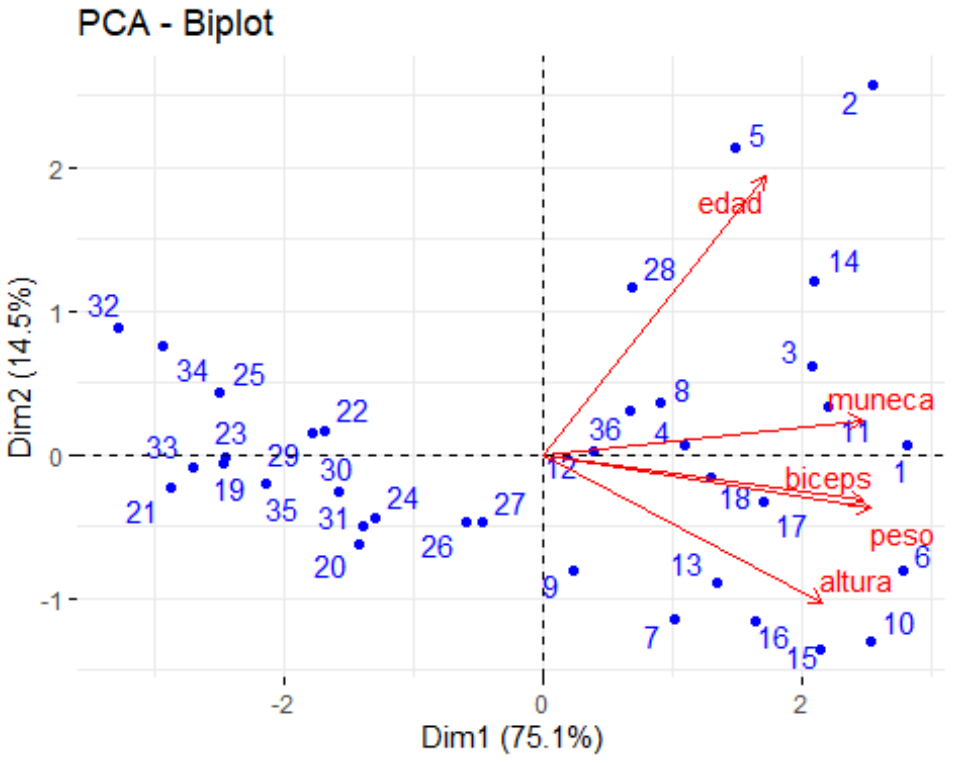
Matriz R

- Peso, bíceps y muñeca son las variables que más contribuyen al primer componente, cada una superando el 20% de contribución.
- Altura tiene una contribución menor pero aún significativa.
- Edad es la variable que menos contribuye al primer componente.
- Las variables que más influyen en el primer componente son peso, bíceps y muñeca.

```
# 5. Biplot (Combinación de individuos y variables)
fviz_pca_biplot(cpS, repel = TRUE, col.var = "red", col.ind = "blue")
```




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



Interpretación

Matriz S * Peso tiene la mayor longitud de flecha, lo que indica que es la variable que más contribuye a la variabilidad explicada por el primer componente (Dim1).

- Altura y bíceps también tienen una influencia importante, aunque menor que el peso.

*Edad contribuye más al segundo componente (Dim2), lo que se refleja en la orientación de su flecha hacia esa dimensión.

- Este gráfico muestra que peso, altura y bíceps dominan en la variabilidad explicada por el primer componente, mientras que edad tiene más influencia en el segundo componente.

Matriz R

- Dim1 (75.1%) y Dim2 (14.5%) explican conjuntamente el 89.6% de la varianza total.
- Peso, altura, y bíceps son las variables que más contribuyen a la variabilidad en el primer componente (Dim1), como lo indican las flechas largas apuntando hacia esa dirección.

*Edad tiene una mayor influencia en el segundo componente (Dim2), con una flecha que se orienta hacia arriba.

- Peso y altura dominan en Dim1, mientras que edad es más relevante en Dim2.

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.

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` en la librería `FactoMineR` de R permite realizar un análisis de componentes principales de forma sencilla y visual.

scale.unit * Esta opción permite especificar si las variables deben ser estandarizadas antes de realizar el PCA. * Uso: scale.unit = TRUE o scale.unit = FALSE. * Función: Si 'TRUE' (valor por defecto), las variables se estandarizan para tener varianza 1. Esto es útil cuando las variables están en diferentes escalas.

ncp * Define el número de componentes principales que quieres calcular. * Uso: ncp = 5 (por ejemplo, si quieres extraer los primeros cinco componentes). * Función: Permite controlar el número de componentes a retener, lo cual puede ser útil si ya tienes una idea clara de cuántos componentes son relevantes para tu análisis.

ind.sup * Permite especificar las observaciones que quieres considerar como suplementarias (no influirán en el análisis, pero se proyectarán en el gráfico de componentes principales). * Uso: ind.sup = c(5, 10) (para hacer que las observaciones

5 y 10 sean suplementarias). * Función: Útil cuando se desea evaluar la influencia de ciertas observaciones sin que afecten el cálculo de los componentes.

quali.sup * Especifica las variables cualitativas que deben considerarse suplementarias, por lo que no afectarán al análisis principal pero se proyectarán en los gráficos. * Uso: quali.sup = c(3) (indicar la columna que contiene variables cualitativas suplementarias). * Función: Útil cuando tienes variables categóricas en el dataset que no deben ser incluidas en el PCA, pero te interesa ver cómo se proyectan.

PARTE 4

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

1. Compare los resultados obtenidos con la matriz de varianza-covarianza y con la correlación. ¿Qué concluye?

Matriz de varianzas y covarianzas

- El primer componente principal (CP1) explica una porción muy grande de la varianza (alrededor del 76%). Esto ocurre porque en esta matriz no se estandarizan las variables, por lo que aquellas con mayor variabilidad, como peso y altura, tienen un mayor impacto en el análisis.
- Usar la matriz de varianzas-covarianzas es útil cuando se quiere preservar la magnitud original de las variables y se desea identificar qué variables, en su escala original, tienen mayor impacto en la variabilidad total. En este caso, variables con grandes variaciones (peso y altura) dominan la varianza explicada.

Matriz de correlaciones

- El primer componente principal explica un porcentaje similar de la varianza (alrededor del 75%), pero en este caso, las variables se han estandarizado. Esto significa que todas las variables tienen una varianza unitaria y ninguna domina el análisis solo por tener una escala mayor.
- La matriz de correlaciones es más apropiada cuando las variables están en diferentes escalas o unidades y se busca un análisis más equilibrado. Al estandarizar todas las variables, evitamos que aquellas con mayores varianzas (como el peso y la altura) dominen el análisis, permitiendo que otras variables también contribuyan de manera significativa.
- Si el objetivo es dar igual importancia a todas las variables, la matriz de correlaciones es la mejor opción. Por otro lado, si es importante que las variables se consideren en su escala original, la matriz de varianzas-covarianzas es más apropiada.

1.2 ¿Cuál de los dos procedimientos aporta componentes de mayor interés?

- Siel interes es en un análisis donde las variables con mayor varianza (como peso y altura) dominen y se refleje la escala original de las variables, el análisis con la matriz de varianzas-covarianzas es de mayor interés.

- Si se prefiere un análisis más equilibrado, donde cada variable tiene una influencia similar en los componentes, entonces los componentes obtenidos con la matriz de correlaciones son más interesantes.

- En general, el análisis basado en la matriz de correlaciones, como en este caso, suele aportar componentes de mayor interés en muchos casos, ya que permite comparar variables en diferentes escalas y no deja que una variable con mayor varianza domine el análisis. Esto resulta más útil cuando se quieren identificar patrones entre todas las variables, sin que algunas de ellas tengan un peso desproporcionado en los componentes principales.

2. 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.

- Las variables en la base de datos tienen escalas diferentes. Por ejemplo, la edad se mide en años, el peso en kilogramos, la altura en centímetros, y las medidas de muñeca y bíceps en circunferencia. Esto significa que las varianzas de las variables no son comparables entre sí, ya que las variables con escalas más grandes tienden a dominar el análisis en la matriz de varianzas-covarianzas.

- En este caso, variables como peso y altura, que tienen varianzas más grandes debido a su escala, dominarían el primer componente principal si se utiliza la matriz de varianzas-covarianzas. En cambio, la matriz de correlaciones estandariza las variables (haciéndolas comparables), dándoles igual peso, lo cual es crucial cuando las variables están en diferentes escalas.

- En los resultados de la matriz de correlaciones, observamos que todas las variables (como edad, peso, altura, muñeca y bíceps) tienen contribuciones más equitativas a los primeros componentes.

- En este caso, el análisis basado en correlaciones ofrece una mejor interpretación de las relaciones entre las medidas corporales de los estudiantes, capturando de manera más justa las correlaciones entre las variables y permitiendo una comparación directa entre ellas.

3. ¿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)

- La primera componente captura la mayor parte de la varianza y está dominada por las variables que representan el tamaño corporal.

- Peso: Es la variable que más contribuye a la primera componente. Esto se debe a que tiene el mayor coeficiente en el vector propio asociado con CP1, lo que

indica que las diferencias en peso entre los individuos son una de las principales fuentes de variabilidad en los datos.

- Podemos decir que, Peso, Bíceps, Muñeca, y Altura son las variables que más contribuyen al primer componente principal (CP1).
- La segunda componente explica una parte menor de la varianza y está más relacionada con variaciones en una variable distinta.
- Edad: Es la variable que más contribuye a la segunda componente. Esto significa que las diferencias de edad entre los individuos explican una parte importante de la variabilidad en esta dimensión.
- Altura: Aunque en menor medida, también tiene una contribución relevante al segundo componente.
- CP1 está dominado por variables relacionadas con el tamaño corporal: Peso, seguido de Bíceps, Muñeca, y Altura.
- CP2 está dominado por Edad, con una menor contribución de Altura.

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

- X1 = Edad
- X2 = Peso
- X3 = Altura
- X4 = Muñeca
- X5 = Bíceps

Con la matriz de correlaciones (R)

Ecuación de (CP1) (Componente Principal 1):

(e1_1), (e1_2), …, (e1_5) son los valores del primer vector propio (eigen(R)$vectors).

Para (CP1), utilizando los valores del vector propio de la matriz de varianzas-covarianzas (R):

$$CP1 = (-0.3359) \cdot X1 + (-0.4927) \cdot X2 + (-0.4222) \cdot X3 + (-0.4821) \cdot X4 + (-0.4833) \cdot X5$$

Ecuación de (CP2) (Componente Principal 2):

(e2_1), (e2_2), …, (e2_5) son los valores del segundo vector propio (eigen(R)$vectors).

$$CP2 = (0.8575) \cdot X1 + (-0.1647) \cdot X2 + (-0.4542) \cdot X3 + (0.1082) \cdot X4 + (-0.1392) \cdot X5$$

- Estas combinaciones son las ecuaciones finales recomendadas para interpretar los resultados del análisis de componentes principales.
- Utilizando estas combinaciones lineales, se puede proyectar cada observación en el espacio de los componentes principales, reduciendo la dimensionalidad de los datos sin perder la mayor parte de la información relevante.

5. Interpreta los resultados en término de agrupación de variables

- El primer componente principal agrupa variables que están relacionadas con el tamaño corporal. Las variables que más contribuyen a CP1 son:
- Peso
- Bíceps
- Muñeca
- Altura
- Estas variables están altamente correlacionadas entre sí, lo que sugiere que CP1 captura una dimensión de tamaño físico. Individuos con valores altos en CP1 tienden a ser más pesados, más altos, y con mayores circunferencias de muñeca y bíceps. En términos de agrupación, estas cuatro variables forman un grupo que explica gran parte de la variabilidad en los datos y están alineadas en una dimensión de tamaño general.
- El segundo componente principal agrupa principalmente la variable:
- La edad no está tan correlacionada con el tamaño corporal como las otras variables. Su mayor contribución a CP2 sugiere que esta dimensión está más relacionada con las diferencias de edad, lo que significa que se comporta de manera más independiente respecto a las variables físicas como peso, altura, bíceps y muñeca, y está asociada con una dimensión de variabilidad independiente del tamaño físico.

Conclusión

- El primer grupo de variables (peso, altura, bíceps, muñeca) describe el tamaño corporal general de los individuos, reflejando una relación clara entre estas características físicas. Este grupo domina el análisis en el primer componente principal (CP1).
- La edad se agrupa en el segundo componente principal (CP2) como una variable independiente del tamaño corporal, capturando la variabilidad relacionada con las diferencias de edad entre los individuos.