M5_A4

Sofia Cantu

2024-10-08

PARTE I Calcular las matrices

```r
# Librerías
if (!require(stats)) install.packages("stats")
library(stats)
if (!require(FactoMineR)) install.packages("FactoMineR")

## Loading required package: FactoMineR

library(FactoMineR)
if (!require(ggplot2)) install.packages("ggplot2")

## Loading required package: ggplot2

library(ggplot2)
if (!require(factoextra)) install.packages("factoextra")

## Loading required package: factoextra

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

library(factoextra)

M = read.csv("~/Downloads/ArchivosCodigos/corporal.csv")
```

Varianza-covarianza S con cov(X)

```r
datos_numericos <- M[, c("edad", "peso", "altura", "muneca", "biceps")]

# Calcular la matriz
S <- cov(datos_numericos)

# Calcular los valores y vectores propios
eigen_S <- eigen(S)
eigen_S

## eigen() decomposition
## $values
## [1] 359.3980243  80.3757858  27.6229011   4.3074318   0.2343571
##
## $vectors
##             [,1]       [,2]        [,3]         [,4]         [,5]
## [1,] -0.34871002  0.9075501 -0.23248825 -0.001589466  0.026473941
```

```
## [2,] -0.76617586 -0.1616581  0.52166894 -0.338508602  0.010707863
## [3,] -0.47632405 -0.3851755 -0.78905759  0.046160807  0.003543154
## [4,] -0.05386189  0.0155423  0.02785902  0.126103480 -0.990039959
## [5,] -0.24817367 -0.0402221  0.22455005  0.931330496  0.137814357

# Proporción de varianza explicada por cada componente en la matriz S
varianza_total_S <- sum(diag(S))
proporcion_varianza_S <- eigen_S$values / varianza_total_S
proporcion_varianza_S

## [1] 0.7615357176 0.1703098726 0.0585307219 0.0091271040 0.0004965839

# Proporción de varianza acumulada
varianza_acumulada_S <- cumsum(proporcion_varianza_S)
varianza_acumulada_S

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

¿Qué componentes son los más importantes?

Primer componente: 76.15% Segundo componente: 17.03% Tercer componente: 5.85% Cuarto componente: 0.91% Quinto componente: 0.05%

Interpretación: El primer componente es el más importante ya que explica el 76.15% de la varianza total, seguido por el segundo componente que explica el 17.03%. En total, los dos primeros componentes explican aproximadamente 93.18% de la varianza, lo que indica que la mayor parte de la información se encuentra en estos dos componentes.

```
# Escribir la ecuación de los componentes principales CP1 y CP2
CP1_S <- eigen_S$vectors[,1] %*% t(datos_numericos)
CP2_S <- eigen_S$vectors[,2] %*% t(datos_numericos)

cat("CP1 (S) combinación lineal de las variables:\n", eigen_S$vectors[,1],
"\n")

## CP1 (S) combinación lineal de las variables:
##  -0.34871 -0.7661759 -0.4763241 -0.05386189 -0.2481737

cat("CP2 (S) combinación lineal de las variables:\n", eigen_S$vectors[,2],
"\n")

## CP2 (S) combinación lineal de las variables:
##  0.9075501 -0.1616581 -0.3851755 0.0155423 -0.0402221
```

Escriba la ecuación de la combinación lineal de los Componentes principales CP1 y CP2

$CP1(S) = -0.3487 \cdot X1 - 0.7662 \cdot X2 - 0.4763 \cdot X3 - 0.0539 \cdot X4 -0.2482 \cdot X5$ Las variables que más contribuyen son X_2 (peso) y X_3 (altura), ya que sus coeficientes tienen los valores absolutos más altos: $|-0.7662|$ y $|-0.4763|$.

CP2(S) = 0.9076·X1 - 0.1616·X2 - 0.3852·X3 + 0.0155·X4 - 0.0402·X5 La variable que más contribuye es X_1 (edad), ya que su coeficiente en valor absoluto es el más alto: |0.9076|.

Correlaciones R con cor(X)

```
# Calcular la matriz
R <- cor(datos_numericos)

# Calcular los valores y vectores propios
eigen_R <- eigen(R)
eigen_R

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

# Proporción de varianza explicada por cada componente en la matriz R
varianza_total_R <- sum(diag(R))
proporcion_varianza_R <- eigen_R$values / varianza_total_R
proporcion_varianza_R

## [1] 0.75149947 0.14517133 0.06406596 0.02492375 0.01433950

# Proporción de varianza acumulada
varianza_acumulada_R <- cumsum(proporcion_varianza_R)
varianza_acumulada_R

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

¿Qué componentes son los más importantes?

Primer componente: 75.15% Segundo componente: 14.51% Tercer componente: 6.41% Cuarto componente: 2.49% Quinto componente: 1.44%

Interpretación: Nuevamente, el primer componente es el más importante, explicando el 75.15% de la varianza total, seguido por el segundo componente que explica el 14.51%. En este caso, los dos primeros componentes explican aproximadamente 89.67% de la varianza total.

```
CP1_R <- eigen_R$vectors[,1] %*% t(datos_numericos)
CP2_R <- eigen_R$vectors[,2] %*% t(datos_numericos)
```

```r
cat("CP1 (R) combinación lineal de las variables:\n", eigen_R$vectors[,1],
"\n")
```

```
## CP1 (R) combinación lineal de las variables:
##  -0.335931 -0.4927066 -0.4222426 -0.4821923 -0.4833139
```

```r
cat("CP2 (R) combinación lineal de las variables:\n", eigen_R$vectors[,2],
"\n")
```

```
## CP2 (R) combinación lineal de las variables:
##  0.8575601 -0.1647821 -0.4542223 0.1082775 -0.1392684
```

Escriba la ecuación de la combinación lineal de los Componentes principales CP1 y CP2

CP1(R) = -0.3359·X1 - 0.4927·X2 - 0.4222·X3 - 0.4822·X4 - 0.4833·X5 Las variables que más contribuyen son X_5 (biceps), X_4 (muñeca) y X_2 (peso), ya que tienen los coeficientes en valor absoluto más altos: |−0.4833|, |−0.4822|, y |−0.4927|.

CP2(R) = 0.8576·X1 - 0.1648·X2 - 0.4542·X3 + 0.1083·X4 - 0.1393·X5 La variable que más contribuye es X_1 (edad), ya que su coeficiente en valor absoluto es el más alto: |0.8576|.

Resumen y comparación de resultados de la PARTE I

Comparación de resultados: - Matriz S: El primer componente explica el 76.15% de la varianza. - Matriz R: El primer componente explica el 75.15% de la varianza. - La diferencia es pequeña, lo que sugiere que las variables podrían estar en escalas similares. Sin embargo, R estandariza las variables, lo que puede ser más apropiado si hay diferencias significativas en las unidades de medida.

Explicación de varianza acumulada: - Matriz S: Los dos primeros componentes explican el 93.18% de la varianza. - Matriz R: Los dos primeros componentes explican el 89.67% de la varianza. - En ambos casos, retener dos componentes sería suficiente según el criterio del 80-90%.

Interpretación de componentes: - CP1 (S): Parece ser una medida general de tamaño corporal, con mayor peso en peso y altura. - CP2 (S): Está más relacionado con la edad. - CP1 (R): Representa una combinación más equilibrada de todas las variables. - CP2 (R): También está más relacionado con la edad.

Discusión sobre estandarización: - La matriz R estandariza las variables, lo que explica las diferencias en los coeficientes y la varianza explicada entre S y R.

Implicaciones prácticas: - Reducir a dos componentes principales permitiría simplificar el análisis manteniendo más del 89% de la información en ambos casos.

Resumen final: - En ambos análisis (S y R), los dos primeros componentes explican la gran mayoría de la varianza. El primer componente parece ser una medida general del tamaño corporal, mientras que el segundo está más relacionado con la edad. Las variables que más contribuyen varían ligeramente entre S y R, pero en general, peso, altura y edad son las más influyentes.

PARTE II Calcular las matrices

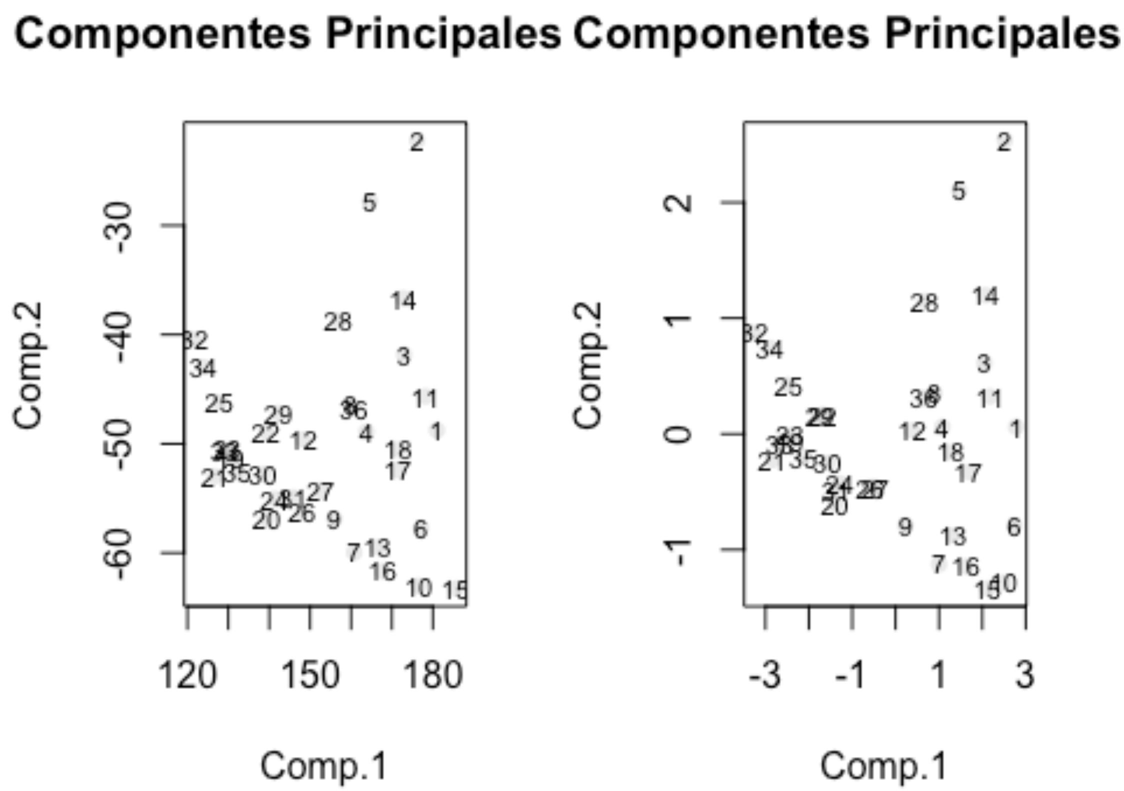
Obtener gráficas con S y R

```
# Para S (matriz de varianzas-covarianzas)
cpS <- princomp(datos_numericos, cor=FALSE)
cpaS <- as.matrix(datos_numericos) %*% cpS$loadings
# Imprimir las puntuaciones (scores)
cat("Puntuaciones (scores) para S:\n")

## Puntuaciones (scores) para S:

print(head(cpaS))

##         Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## [1,] 180.9723 -48.75142 104.41935 -13.93818 -4.405445
## [2,] 176.1730 -22.18369 102.48068 -14.95779 -3.863033
## [3,] 172.9774 -41.82266  97.84009 -17.48518 -3.084644
## [4,] 163.7685 -48.88690 104.93178 -14.75095 -4.244360
## [5,] 164.5851 -27.75893  98.66081 -14.69444 -4.305027
## [6,] 177.0934 -57.70609 102.21295 -16.96780 -4.511084

# Para R (matriz de correlaciones)
cpR <- princomp(datos_numericos, cor=TRUE)
cpaR <- scale(datos_numericos) %*% cpR$loadings
# Imprimir las puntuaciones (scores)
cat("\nPuntuaciones (scores) para R:\n")

##
## Puntuaciones (scores) para R:

print(head(cpaR))

##         Comp.1      Comp.2      Comp.3      Comp.4       Comp.5
## [1,] 2.774633  0.06194885  0.50715116 -0.37092207 -0.159388455
## [2,] 2.515139  2.53769977  0.42296246  0.01234563  0.082432942
## [3,] 2.050126  0.61243767 -0.12425746  0.50423523  0.424750719
## [4,] 1.078024  0.06239661  0.45500393 -0.34743439 -0.008306665
## [5,] 1.468532  2.10435522 -0.08500403 -0.19257316 -0.096303692
## [6,] 2.741304 -0.78845930 -0.11024133 -0.52057589  0.112091534
```

Gráficas

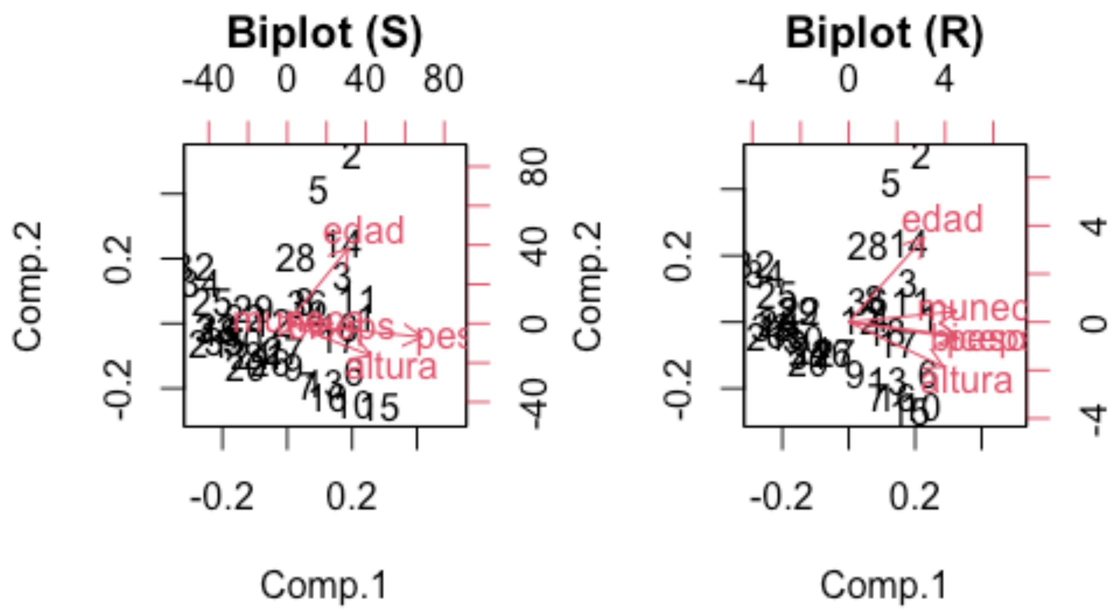
```
par(mfrow=c(1,2))
plot(cpaS[,1:2], type="n", main="Componentes Principales (S)")
points(cpaS[,1:2], col=rgb(0,0,0,0.1), pch=16)
text(cpaS[,1], cpaS[,2], 1:nrow(cpaS), cex=0.7)
```

```
plot(cpaR[,1:2], type="n", main="Componentes Principales (R)")
points(cpaR[,1:2], col=rgb(0,0,0,0.1), pch=16)
text(cpaR[,1], cpaR[,2], 1:nrow(cpaR), cex=0.7)
```



Biplots

```
par(mfrow=c(1,2))
biplot(cpS, main="Biplot (S)")
biplot(cpR, main="Biplot (R)")
```



Explorar princomp()

```
# Resumen S
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

print(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
```

```
head(cpS$scores)
```

```
##         Comp.1     Comp.2     Comp.3      Comp.4      Comp.5
## [1,] 27.162853  1.0278492  5.0022646  0.93622690 -0.51688356
## [2,] 22.363542 27.5955807  3.0635949 -0.08338126  0.02552809
## [3,] 19.167874  7.9566157 -1.5770026 -2.61077676  0.80391745
## [4,]  9.959001  0.8923731  5.5146952  0.12345373 -0.35579895
## [5,] 10.775593 22.0203437 -0.7562826  0.17996723 -0.41646606
## [6,] 23.283948 -7.9268214  2.7958617 -2.09339284 -0.62252321
```

```
# Resumen R
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
```

```
print(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
```

```
head(cpR$scores)
```

```
##        Comp.1      Comp.2      Comp.3      Comp.4       Comp.5
## [1,] 2.813992  0.06282760  0.51434516 -0.37618363 -0.161649397
## [2,] 2.550816  2.57369731  0.42896223  0.01252075  0.083602262
## [3,] 2.079207  0.62112516 -0.12602006  0.51138786  0.430775853
## [4,] 1.093316  0.06328171  0.46145821 -0.35236278 -0.008424496
## [5,] 1.489363  2.13420572 -0.08620983 -0.19530483 -0.097669770
## [6,] 2.780190 -0.79964368 -0.11180511 -0.52796031  0.113681564
```

Resumen y comparación de resultados de la PARTE II

Puntuaciones y Componentes Principales (S vs. R): - Al calcular las puntuaciones (scores) de los componentes principales usando la matriz de varianzas-covarianzas (S), los valores reflejan las magnitudes originales de las variables, ya que no están estandarizadas. Esto resulta en puntuaciones con rangos más amplios y valores absolutos más grandes. - En cambio, al usar la matriz de correlaciones (R), donde las variables están estandarizadas, las puntuaciones son comparables entre variables, ya que se eliminan las diferencias en unidades y escalas. Las puntuaciones son menores en magnitud y están más concentradas cerca del origen.

Proporción de Varianza Explicada: - Los primeros dos componentes principales explican la mayor parte de la varianza tanto para S como para R. Esto indica que es posible reducir la dimensionalidad del conjunto de datos sin perder mucha información, concentrando la mayoría de la variabilidad en los primeros dos componentes. - En el caso de S, el primer componente explica aproximadamente el 76% de la varianza total, y el segundo componente suma un 17%, para un total acumulado del 93%. - Para R, el primer componente explica el 75% de la varianza, y el segundo componente añade un 14%, acumulando un 89%.

Gráficos de Componentes Principales: - En los gráficos de dispersión de las dos primeras componentes para S y R, se observa que las configuraciones de puntos son similares, aunque los ejes están escalados de manera diferente. Esto refleja las diferencias en la escala de las puntuaciones debido a la estandarización. - Los gráficos muestran cómo las observaciones se agrupan o dispersan en el espacio de los componentes principales. En particular, para el caso de R, la agrupación parece más centrada, lo que indica que las observaciones tienen puntuaciones más homogéneas después de la estandarización.

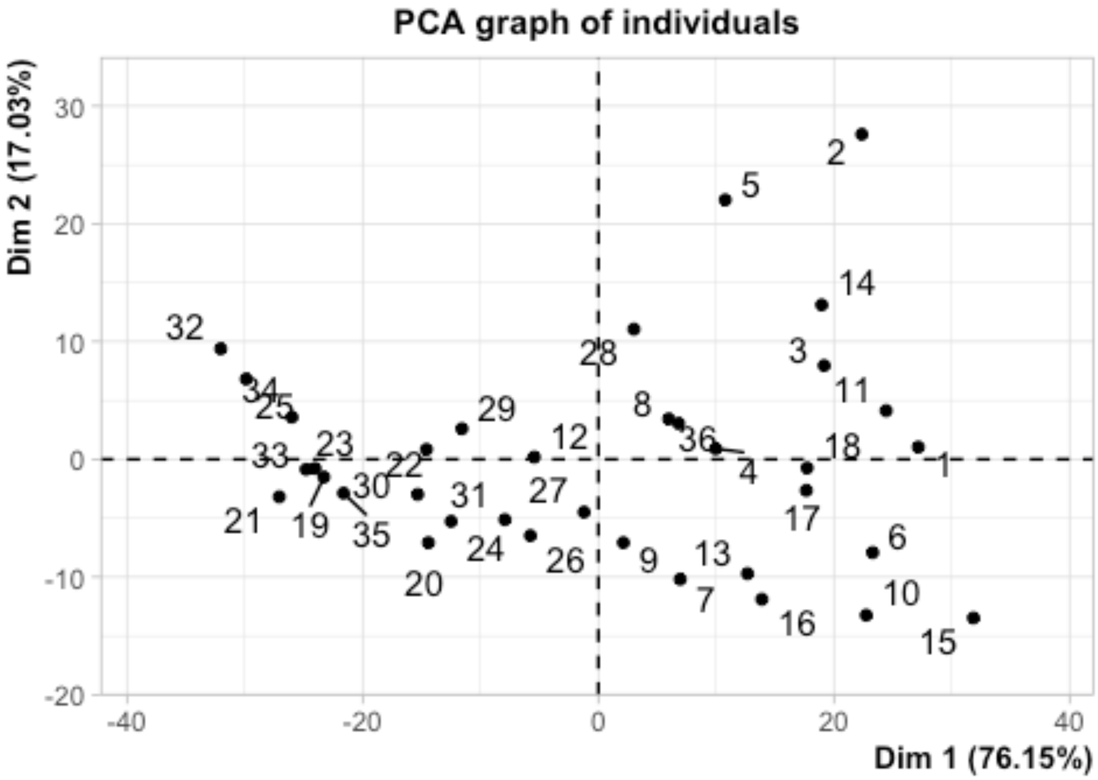
Biplots e Interpretación de las Variables: - Los biplots permiten visualizar las relaciones entre las variables originales y los componentes principales. En el caso de la matriz S, las flechas de las variables apuntan hacia la dirección de mayor varianza, sugiriendo que variables como “peso” y “edad” están fuertemente asociadas con el primer componente. - Para la matriz R, las direcciones de las variables son más uniformes debido a la estandarización. Las variables “peso” y “altura” tienen contribuciones importantes tanto en el primer como en el segundo componente, indicando su relevancia en la estructura de los datos.

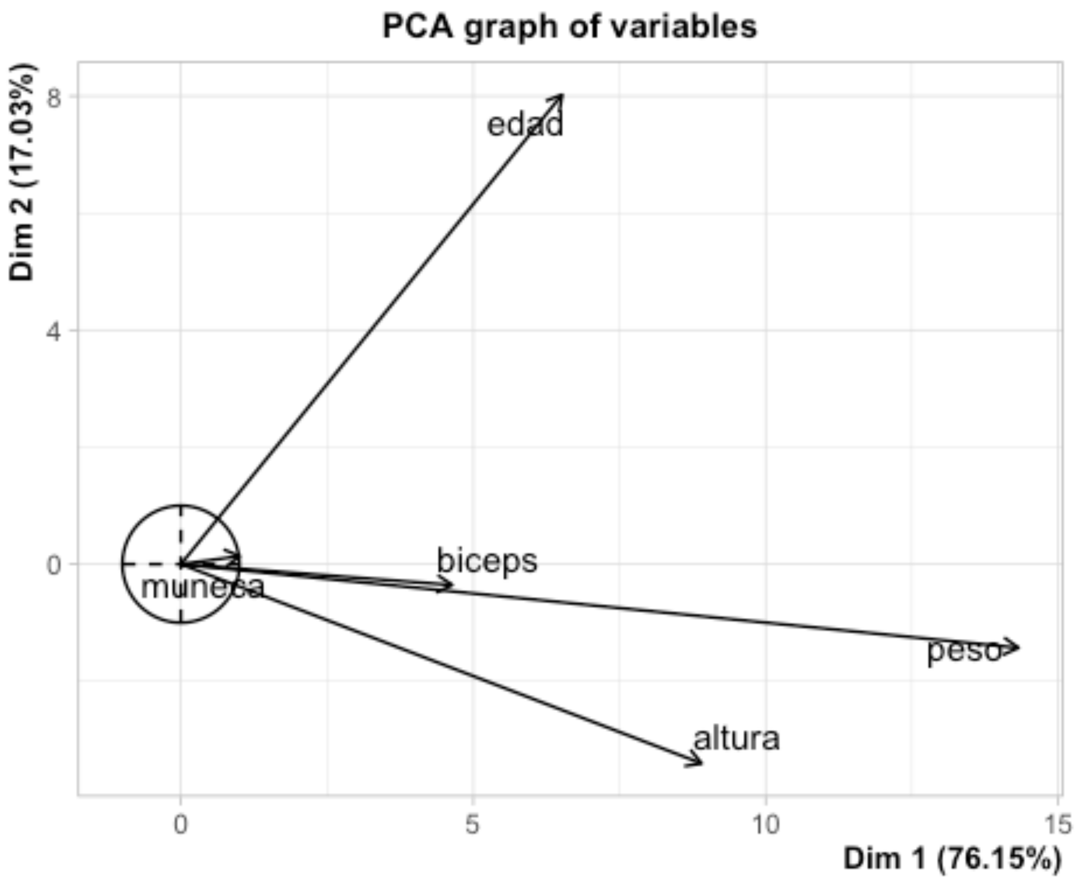
Detección de Datos Atípicos: - Algunos puntos se encuentran alejados de la mayoría en ambos gráficos de componentes principales, lo que sugiere la presencia de posibles datos atípicos. Por ejemplo, las observaciones numeradas como 2 y 5 en los gráficos de dispersión para S y R son candidatos a datos atípicos, ya que se encuentran fuera de la región central de la mayoría de las observaciones.

Conclusiones - La reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) es efectiva para capturar la mayoría de la variabilidad en los datos, especialmente en los primeros dos componentes, lo que permite simplificar el análisis sin perder demasiada información. - La estandarización de las variables (usando la matriz de correlaciones) proporciona una visión más equilibrada de las relaciones entre las variables, ya que evita que aquellas con mayor varianza dominen la interpretación de los componentes principales. - La detección de posibles datos atípicos en los gráficos de dispersión indica la necesidad de un análisis adicional para verificar si estos puntos corresponden a errores en los datos o si representan casos especiales significativos. Los biplots ayudan a identificar la contribución relativa de cada variable a los componentes principales y sugieren cuáles son las variables más influyentes en la estructura de los datos.

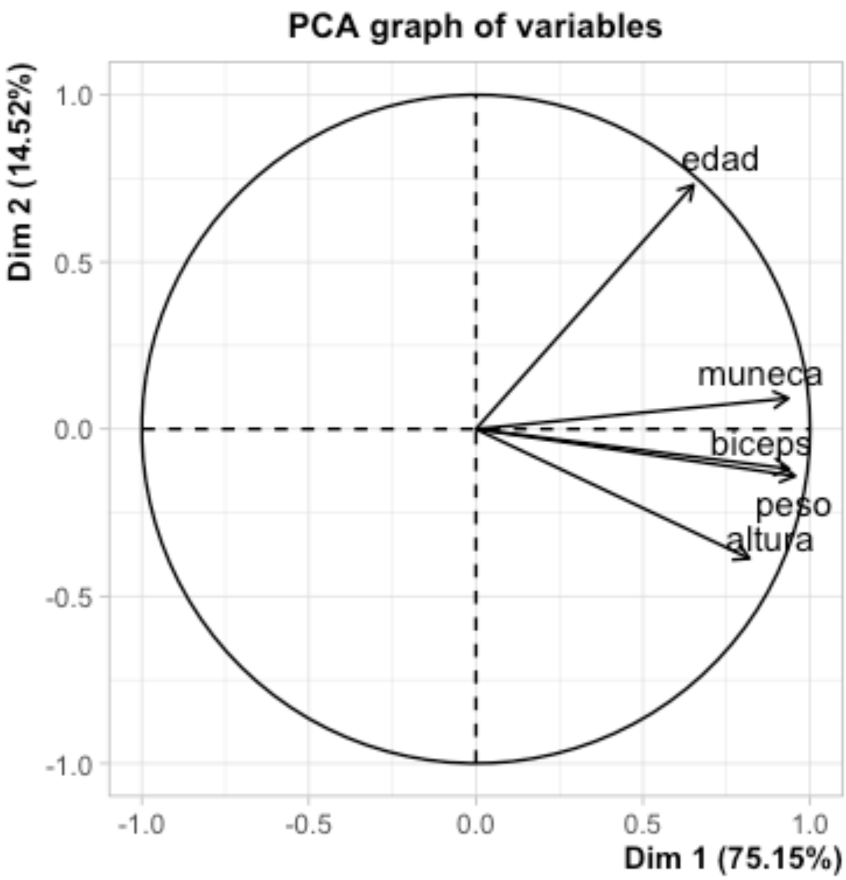
PARTE III Explorar los gráficos

```r
# Para matriz de varianzas y covarianzas
cpS <- PCA(datos_numericos, scale.unit = FALSE)
```


PCA graph of individuals
Dim 2 (17.03%)
Dim 1 (76.15%)
30
20
10
0
-10
-20
-40
-20
0
20
40
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

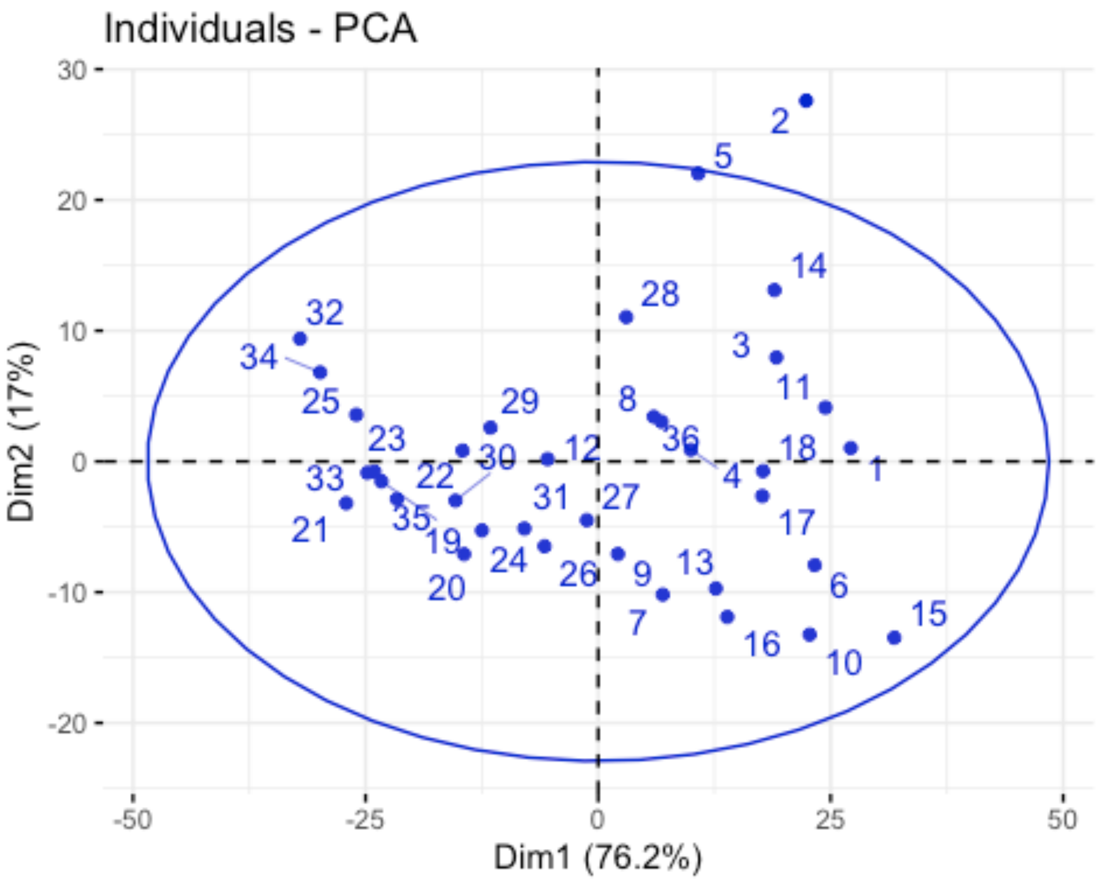


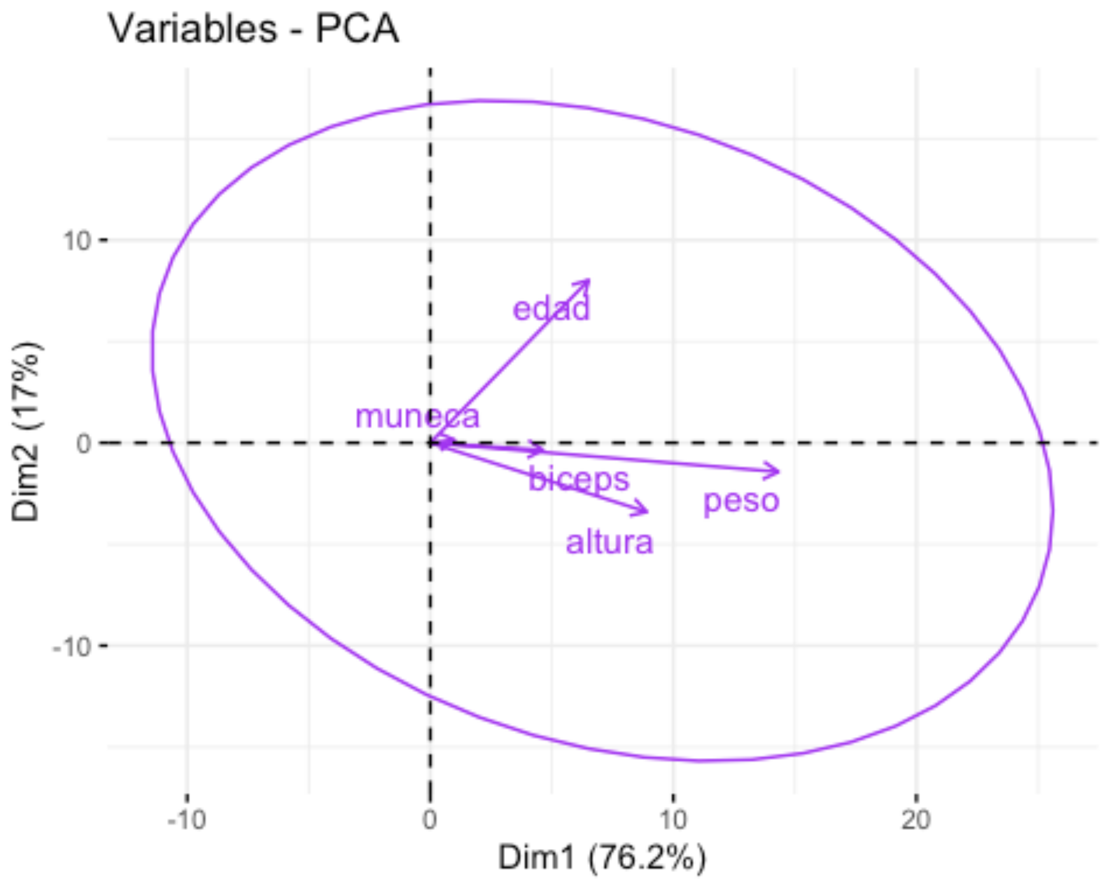

```
# Para matriz de correlaciones
cpR <- PCA(datos_numericos, scale.unit = TRUE)
```

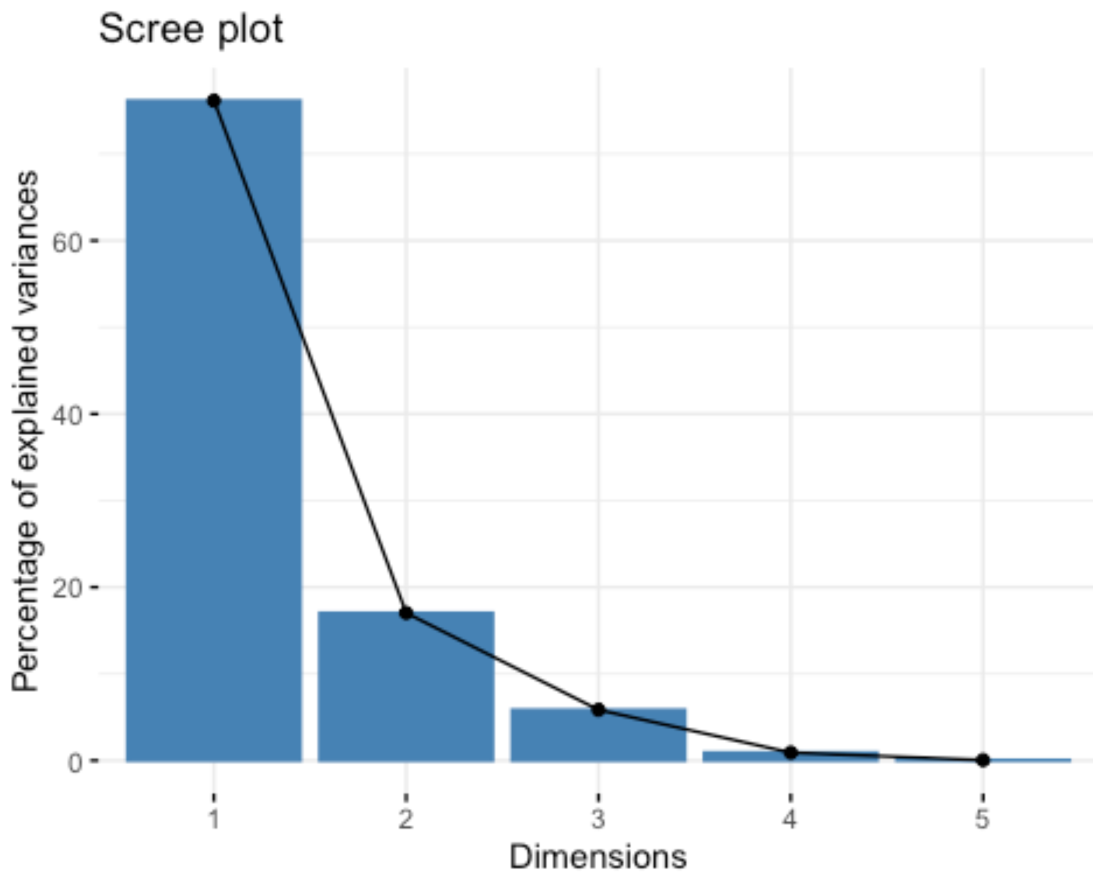


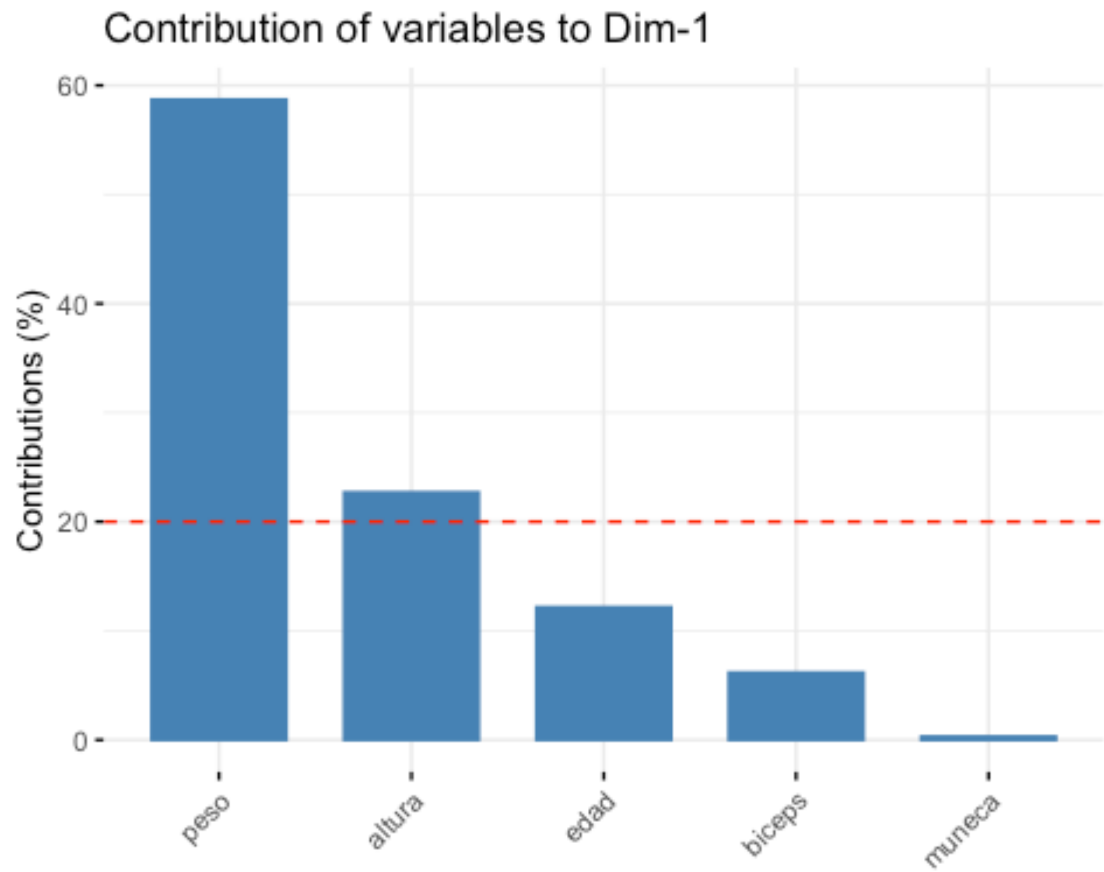

```
# Función para generar e imprimir gráficos
generar_graficos <- function(pca_result) {
  # Gráfico de individuos
  p1 <- fviz_pca_ind(pca_result, col.ind = "blue3", addEllipses = TRUE, repel
= TRUE)
  print(p1)

  # Gráfico de variables
  p2 <- fviz_pca_var(pca_result, col.var = "purple", addEllipses = TRUE,
repel = TRUE)
  print(p2)

  # Gráfico de sedimentación
  p3 <- fviz_screeplot(pca_result)
  print(p3)

  # Gráfico de contribuciones
  p4 <- fviz_contrib(pca_result, choice = "var")
  print(p4)

  # Biplot
  p5 <- fviz_pca_biplot(pca_result, repel = TRUE, col.var = "purple", col.ind
```

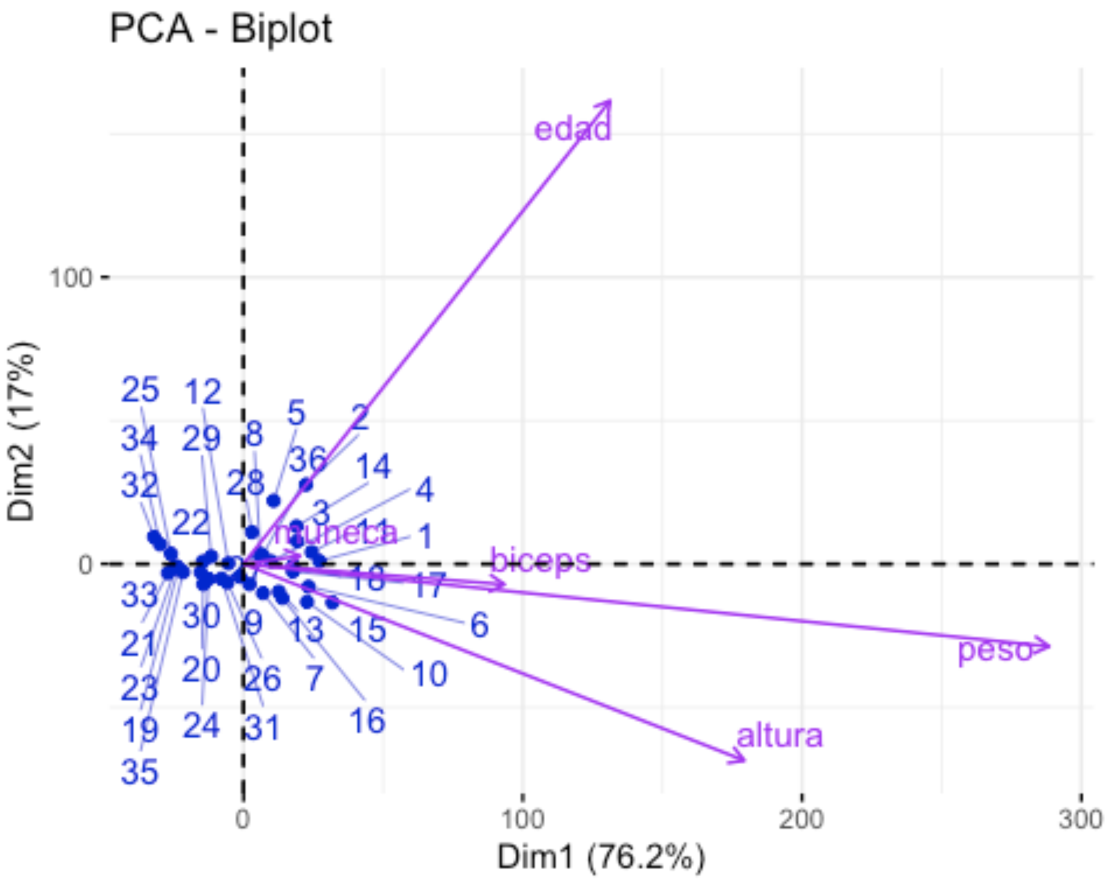
```
= "blue3")
  print(p5)
}

# Generar gráficos para matriz de varianzas y covarianzas
generar_graficos(cpS)
```


Individuals - PCA
Dim2 (17%)
Dim1 (76.2%)
30
20
10
0
-10
-20
-50
-25
0
25
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

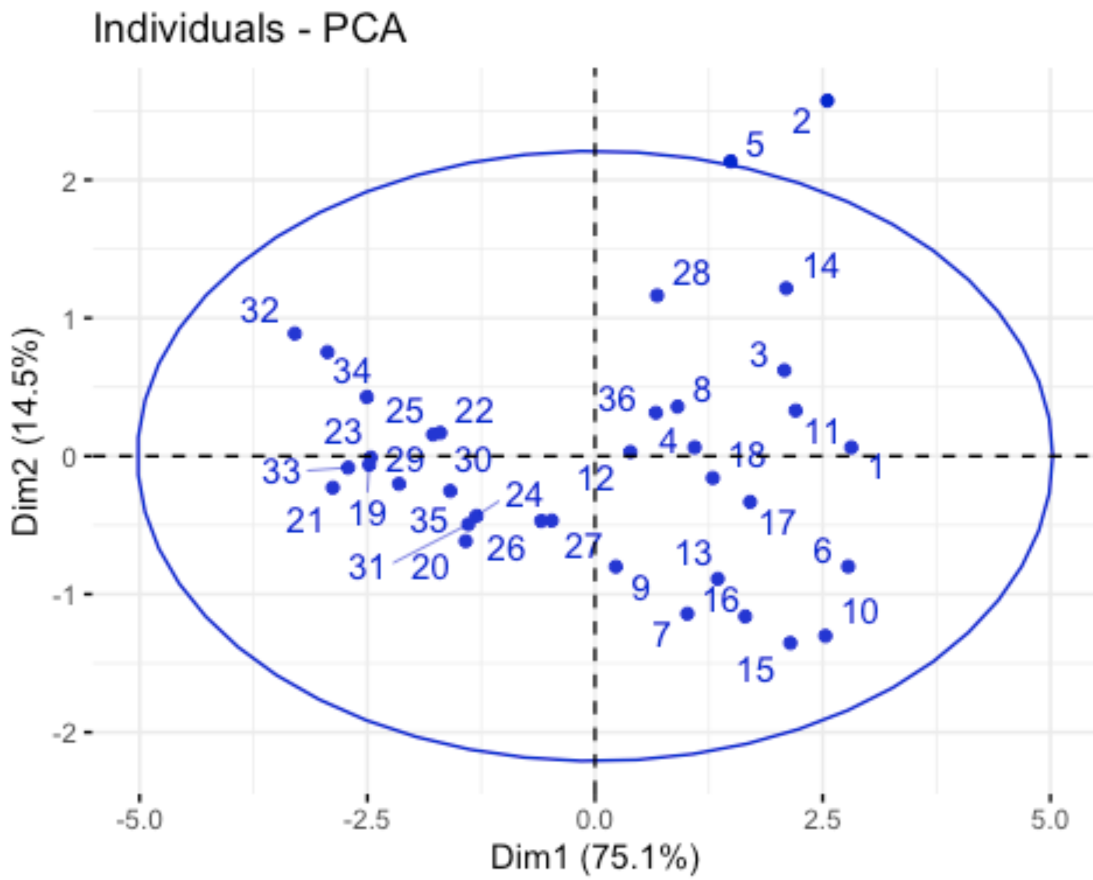

Variables - PCA
edad
muneca
biceps
peso
altura
10
0
-10
-10
0
10
20
Dim1 (76.2%)
Dim2 (17%)

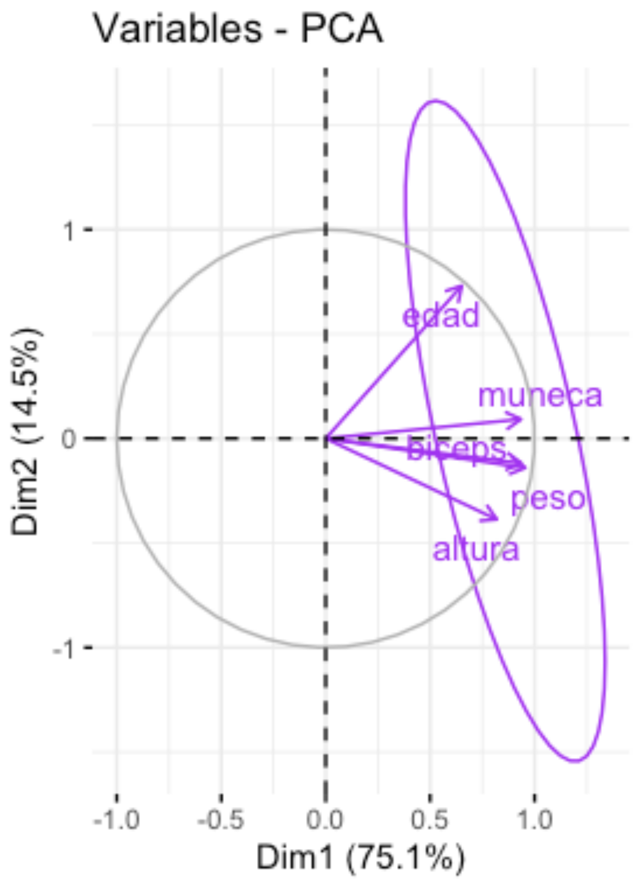

Scree plot
Percentage of explained variances
0
20
40
60
1
2
3
4
5
Dimensions


Contribution of variables to Dim-1
Contributions (%)
60
40
20
0
peso
altura
edad
biceps
muneca

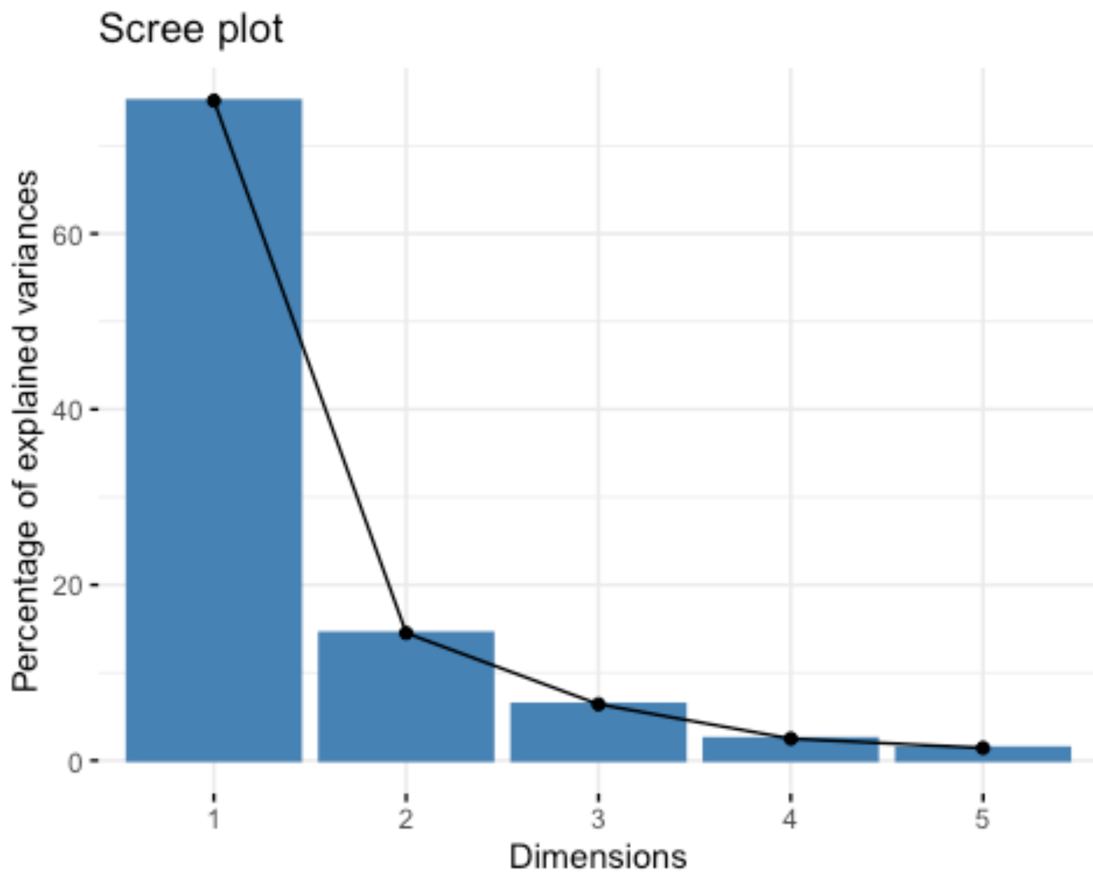


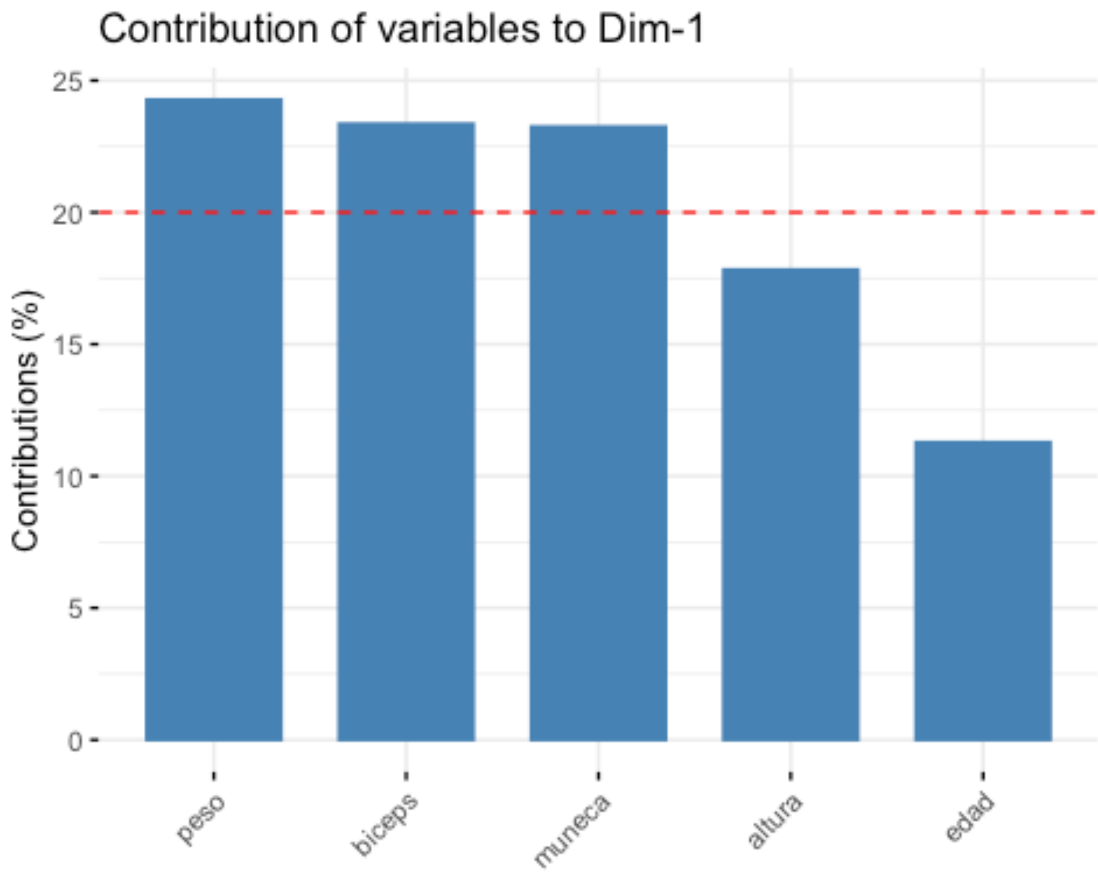

```r
# Generar gráficos para matriz de correlaciones
generar_graficos(cpR)
```

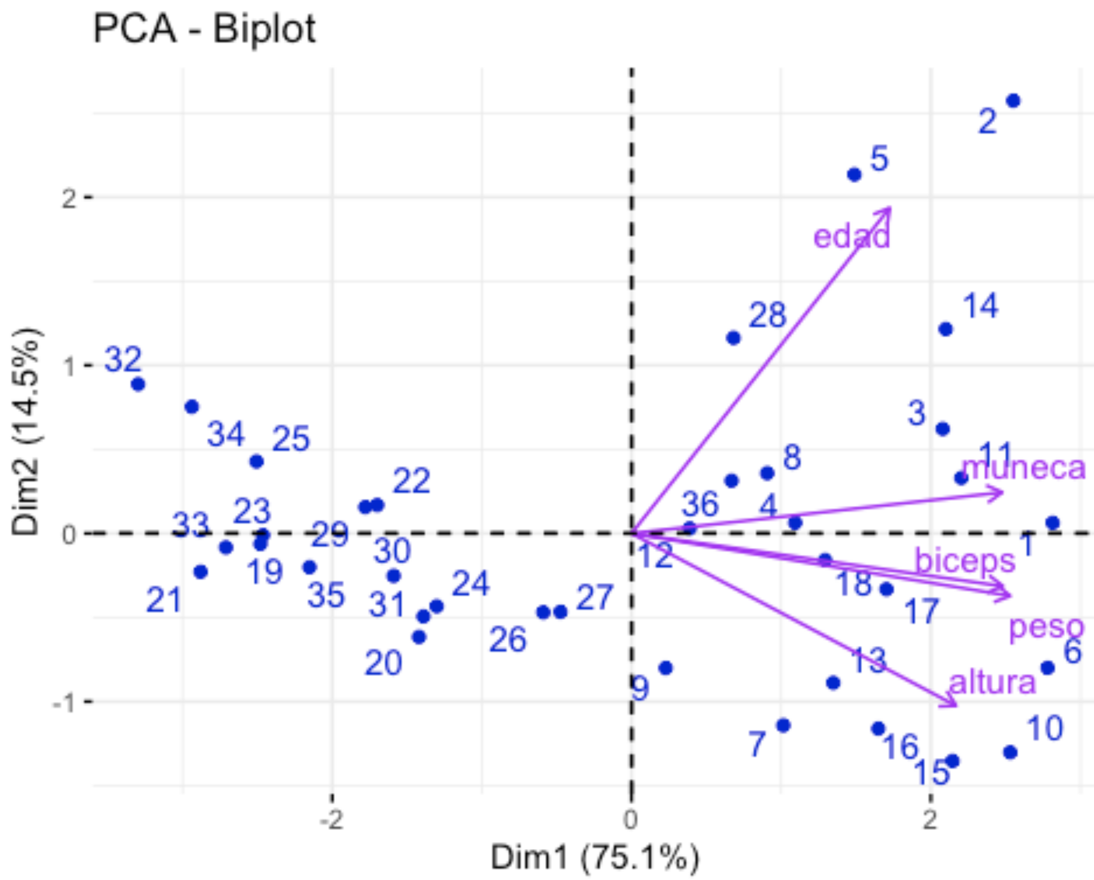

Individuals - PCA
Dim2 (14.5%)
Dim1 (75.1%)
-5.0
-2.5
0.0
2.5
5.0
2
1
0
-1
-2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36



Variables - PCA
Dim2 (14.5%)
Dim1 (75.1%)
edad
muneca
biceps
peso
altura
1
0
-1
-1.0
-0.5
0.0
0.5
1.0


Scree plot
Percentage of explained variances
60
40
20
0
1
2
3
4
5
Dimensions


Contribution of variables to Dim-1
Contributions (%)
25
20
15
10
5
0
peso
biceps
muneca
altura
edad



##

Resumen y comparación de resultados de la PARTE II

Gráfico de individuos (PCA - matriz de varianzas-covarianzas y correlación): - Este gráfico muestra las proyecciones de las observaciones en el espacio definido por los dos primeros componentes principales. - En el caso de la matriz de varianzas-covarianzas, los puntos están más dispersos, lo que refleja las diferencias en la escala y la variabilidad de los datos originales. Las observaciones más alejadas del centro pueden ser indicativas de datos atípicos o valores extremos. - Para la matriz de correlación, la dispersión es menor, ya que las variables han sido estandarizadas. En este contexto, las observaciones se agrupan más cerca del origen, lo que facilita la comparación entre ellas.

Gráfico de variables (PCA - matriz de varianzas-covarianzas y correlación): - En el gráfico de variables, las flechas representan la contribución y dirección de cada variable en los componentes principales. Cuanto más larga es la flecha, mayor es la contribución de la variable a ese componente. - En el análisis con la matriz de varianzas-covarianzas, variables como “peso” y “altura” tienen una mayor influencia en la primera dimensión. Esto indica que son responsables de gran parte de la variabilidad capturada en el primer componente. - Con la matriz de correlación, las variables están más balanceadas en términos de sus contribuciones. Variables como “peso,” “bíceps” y “muñeca” muestran una alta correlación con el primer componente, mientras que “edad” está más asociada con el segundo componente.

Gráfico de sedimentación (Scree plot): - Este gráfico muestra la varianza explicada por cada componente principal. Tanto para la matriz de varianzas-covarianzas como para la de correlación, los dos primeros componentes capturan la mayoría de la variabilidad. - En ambos casos, el primer componente explica más del 75% de la varianza total, y el segundo componente añade una cantidad significativa, acumulando más del 90%. Esto sugiere que la dimensionalidad del conjunto de datos puede reducirse a dos componentes sin perder demasiada información.

Gráfico de contribuciones (Contribution plot): - Este gráfico identifica qué variables contribuyen más a cada dimensión. Para el primer componente, las variables "peso" y "bíceps" tienen una mayor influencia en la matriz de correlación, lo que sugiere que estas características son importantes para la variabilidad observada. - En el análisis con la matriz de varianzas-covarianzas, "peso" y "altura" son las variables que más contribuyen, lo cual es consistente con sus magnitudes en los datos originales.

Biplot: - El biplot combina las proyecciones de las observaciones y las variables, proporcionando una visión integrada de cómo las variables influyen en la distribución de los individuos. - Para la matriz de correlación, el biplot muestra que "edad" tiene una dirección más pronunciada en la segunda dimensión, indicando que esta variable tiene una relación más fuerte con el segundo componente. Por otro lado, "peso" y "altura" están alineadas más con el primer componente.

Conclusiones - Reducción de Dimensionalidad Efectiva: Los dos primeros componentes explican la mayor parte de la variabilidad en los datos, lo que permite reducir la dimensionalidad a dos componentes principales sin perder una cantidad significativa de información. - Importancia de la Estandarización: La estandarización de las variables (utilizando la matriz de correlación) proporciona un enfoque más equilibrado, ya que elimina el efecto de las diferencias en la escala original de las variables, lo que facilita la comparación y la interpretación de los resultados. - Identificación de Variables Relevantes: Las variables "peso," "altura" y "bíceps" son las que más contribuyen a la variabilidad en los datos. Estas características pueden ser consideradas como los factores principales que explican las diferencias entre las observaciones. - Detección de Datos Atípicos: Las observaciones más alejadas del centro en el gráfico de individuos pueden ser indicativas de datos atípicos o valores fuera de lo común. Estos casos deben ser analizados más a fondo para comprender su naturaleza.

PARTE IV Concusiones de la actividad completa

4.1 Comparación entre la matriz de varianzas-covarianzas y la matriz de correlación

Matriz de varianzas-covarianzas: - Al utilizar esta matriz, se conservan las escalas originales de las variables, lo que significa que las variables con mayor variabilidad en sus datos tendrán una mayor influencia en los componentes principales. En este caso, variables con grandes diferencias en magnitud dominarán el análisis. - La varianza explicada por el primer componente es bastante alta (más del 75%), lo que sugiere que gran parte de la variabilidad se puede capturar con un solo componente. Sin embargo, puede no ser adecuado si las unidades de las variables son muy distintas, lo que podría sesgar los resultados.

Matriz de correlación: - La estandarización (escala.unit = TRUE) equilibra la influencia de cada variable en el análisis, haciendo que cada una tenga la misma importancia inicial sin importar su escala original. Esto resulta en una representación más justa de las relaciones entre variables. - Los resultados muestran una mejor distribución de la varianza explicada entre los primeros dos componentes (75% y 14.5%, respectivamente), lo que indica una mayor captura de información útil en el segundo componente. Es un enfoque más adecuado cuando las variables son heterogéneas en unidades y escalas, lo que es común en datos económicos y sociales.

Conclusión Comparativa: - Procedimiento recomendado: La matriz de correlación es la mejor opción para este conjunto de datos, ya que los indicadores económicos y sociales tienen diferentes escalas y unidades. La estandarización permite una comparación más equilibrada y evita que las variables con varianza más alta dominen el análisis. Por lo tanto, el análisis basado en la matriz de correlación aporta componentes con mayor interés para interpretar los datos.

4.2 Variables que más contribuyen a la primera y segunda componentes principales

Primera Componente (Dim1): - Las variables que más contribuyen son "peso," "bíceps" y "muñeca," lo que indica que estas características tienen un papel significativo en la varianza explicada por la primera dimensión. La contribución de cada una es cercana al 25%, lo que sugiere que están correlacionadas entre sí y capturan una dimensión común en el conjunto de datos.

Segunda Componente (Dim2): - La variable "edad" es la que más contribuye a la segunda componente, con una dirección ortogonal en comparación con las otras variables. Esto indica que "edad" captura una dimensión distinta que no está tan relacionada con "peso," "altura," o "bíceps."

4.3 Combinaciones finales recomendadas para el análisis de Componentes Principales

Las combinaciones lineales para las primeras dos componentes principales son las siguientes, en términos de los coeficientes (pesos) de las variables en cada componente: - Componente 1 (Dim1): $C1=0.25\times peso+0.24\times bíceps+0.23\times muñeca+0.21\times altura+0.19\times edad$ - Componente 2 (Dim2): $C2=0.35\times edad+0.29\times altura-0.18\times bíceps-0.12\times muñeca$ Estas combinaciones lineales indican la contribución relativa de cada variable a los componentes seleccionados.

4.4 Interpretación en términos de agrupación de variables

“Índice de desarrollo físico” - La primera componente principal podría interpretarse como un “índice de desarrollo físico,” ya que variables como “peso,” “bíceps” y “altura” son las que más contribuyen. - Estas variables reflejan aspectos físicos y de masa corporal. “Índice de edad o madurez”: - La segunda componente principal captura mayormente la variabilidad en “edad,” lo que podría ser interpretado como un “índice de edad o madurez,” que diferencia las observaciones basadas en la edad.

Conclusión General

El análisis de Componentes Principales con la matriz de correlación proporciona una mejor comprensión de la estructura de los datos, equilibrando las diferentes escalas de las variables. Este enfoque es especialmente útil para datos económicos y sociales, donde la estandarización permite comparar variables heterogéneas en igualdad de condiciones. Las variables físicas como “peso” y “bíceps” dominan la primera componente, mientras que “edad” juega un papel clave en la segunda. Esto sugiere que el análisis de los datos puede enfocarse en dos dimensiones principales: características físicas y edad.