A4-Componentes Principales

Ozner Leyva

2024-10-10

PARTE I

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

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

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

Varianza-covarianza S con cov(X)

```
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
```

```r
# 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 dominante, explicando el 76.15% de la varianza total. El segundo componente añade un 17.03%. Juntos, los dos primeros componentes abarcan aproximadamente el 93.18% de la varianza, indicando que la mayor parte de la información se concentra en ellos.

```r
# 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
```

```r
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

$CP1(S) = -0.3487 \cdot X_1 - 0.7662 \cdot X_2 - 0.4763 \cdot X_3 - 0.0539 \cdot X_4 - 0.2482 \cdot X_5$

Las variables que más contribuyen son X_2 (peso) y X_3 (altura), ya que sus coeficientes en valor absoluto son los más altos: $|-0.7662|$ y $|-0.4763|$.

$CP2(S) = 0.9076 \cdot X_1 - 0.1616 \cdot X_2 - 0.3852 \cdot X_3 + 0.0155 \cdot X_4 - 0.0402 \cdot X_5$

La variable que más contribuye es X_1 (edad), dado que su coeficiente en valor absoluto es el mayor: $|0.9076|$.

Correlaciones R con cor(X)

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

# Calcular los valores y vectores propios
```

```r
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
```

```r
# 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
```

```r
# 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: El primer componente sigue siendo el más significativo, explicando el 75.15% de la varianza total. El segundo componente añade un 14.51%, sumando aproximadamente un 89.67% de la varianza entre los dos primeros componentes.

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

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

$CP1(R) = -0.3359 \cdot X_1 - 0.4927 \cdot X_2 - 0.4222 \cdot X_3 - 0.4822 \cdot X_4 - 0.4833 \cdot X_5$

Las variables que más contribuyen son X_2 (peso), X_4 (muñeca) y X_5 (bíceps), dado que sus coeficientes en valor absoluto son los más elevados: |−0.4927|, |−0.4822| y |−0.4833|.

$CP2(R) = 0.8576 \cdot X_1 - 0.1648 \cdot X_2 - 0.4542 \cdot X_3 + 0.1083 \cdot X_4 - 0.1393 \cdot X_5$

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 mínima, lo que sugiere que las variables están en escalas similares. Sin embargo, al utilizar la matriz R, se estandarizan las variables, lo cual es más apropiado si existen 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 estandarización en la matriz R explica las diferencias en los coeficientes y en la varianza explicada entre S y R.

Implicaciones prácticas: - Reducir el análisis a dos componentes principales simplifica el estudio sin perder más del 89% de la información en ambos casos.

Resumen final: - En ambos análisis, los dos primeros componentes capturan la mayoría de la varianza. El primer componente representa una medida general del tamaño corporal, mientras que el segundo está relacionado con la edad. Las variables más influyentes son “peso”, “altura” y “edad”.

PARTE II

Obtener gráficas con S y R

```
# Para S (matriz de varianzas-covarianzas)
cpS <- princomp(datos_numericos, cor=FALSE)
```

```r
cpaS <- as.matrix(datos_numericos) %*% cpS$loadings
# Imprimir las puntuaciones (scores)
cat("Puntuaciones (scores) para S:\n")
```

```
## Puntuaciones (scores) para S:
```

```r
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
```

```r
# 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")
```

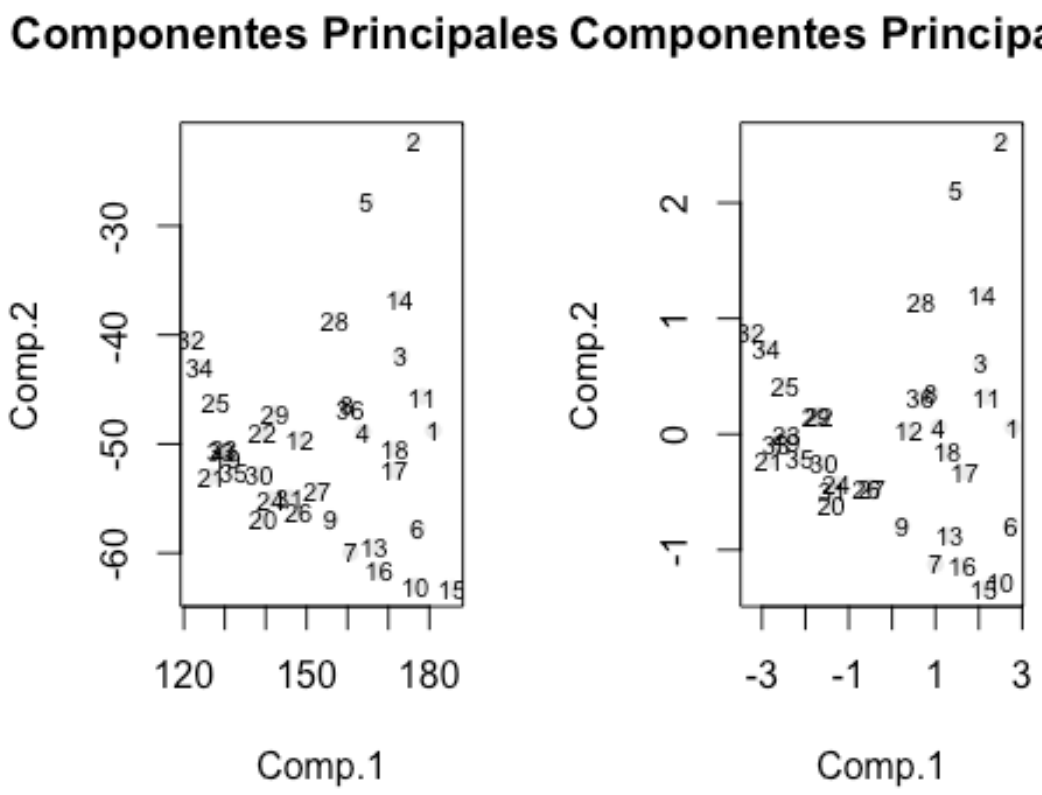
```
##
## Puntuaciones (scores) para R:
```

```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
```

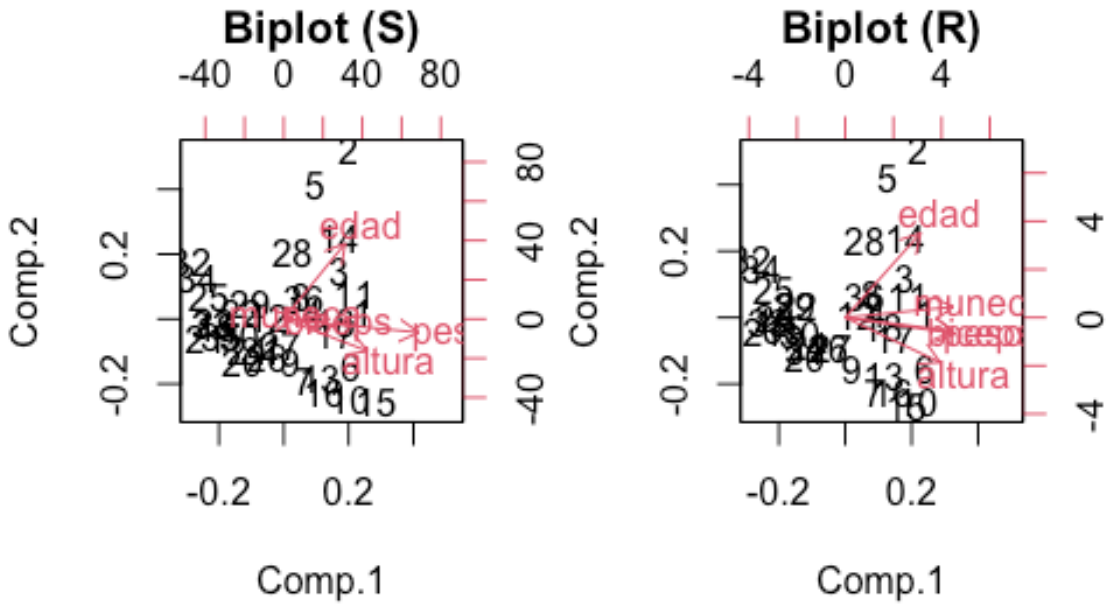
Gráficas

```r
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): Matriz S: Las puntuaciones reflejan las magnitudes originales, mostrando rangos más amplios debido a la falta de estandarización. Matriz R: Las puntuaciones están estandarizadas, siendo comparables entre variables y concentradas alrededor del origen.

Proporción de Varianza Explicada: En ambos casos, los dos primeros componentes explican la mayor parte de la varianza, permitiendo una reducción dimensional efectiva sin pérdida significativa de información. Matriz S: El primer componente explica el 76% y el segundo un 17%, acumulando un 93%. Matriz R: El primer componente explica el 75% y el segundo un 14%, acumulando un 89%.

Gráficos de Componentes Principales: Los gráficos muestran configuraciones similares, aunque con escalas diferentes debido a la estandarización en R. Las observaciones en R están más agrupadas, reflejando puntuaciones más homogéneas tras la estandarización.

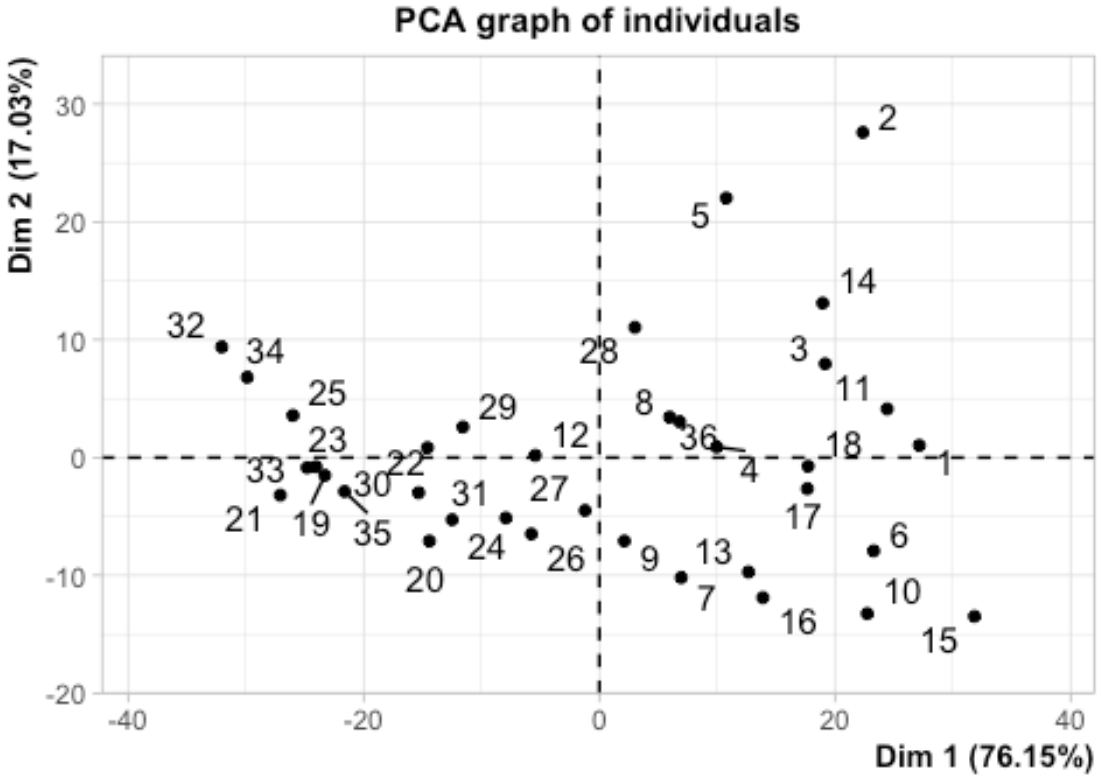
Biplots e Interpretación de las Variables: Los biplots revelan las relaciones entre las variables y los componentes. Matriz S: Variables como “peso” y “edad” tienen una fuerte asociación con el primer componente. Matriz R: Las variables presentan contribuciones más uniformes; “peso” y “altura” son especialmente relevantes.

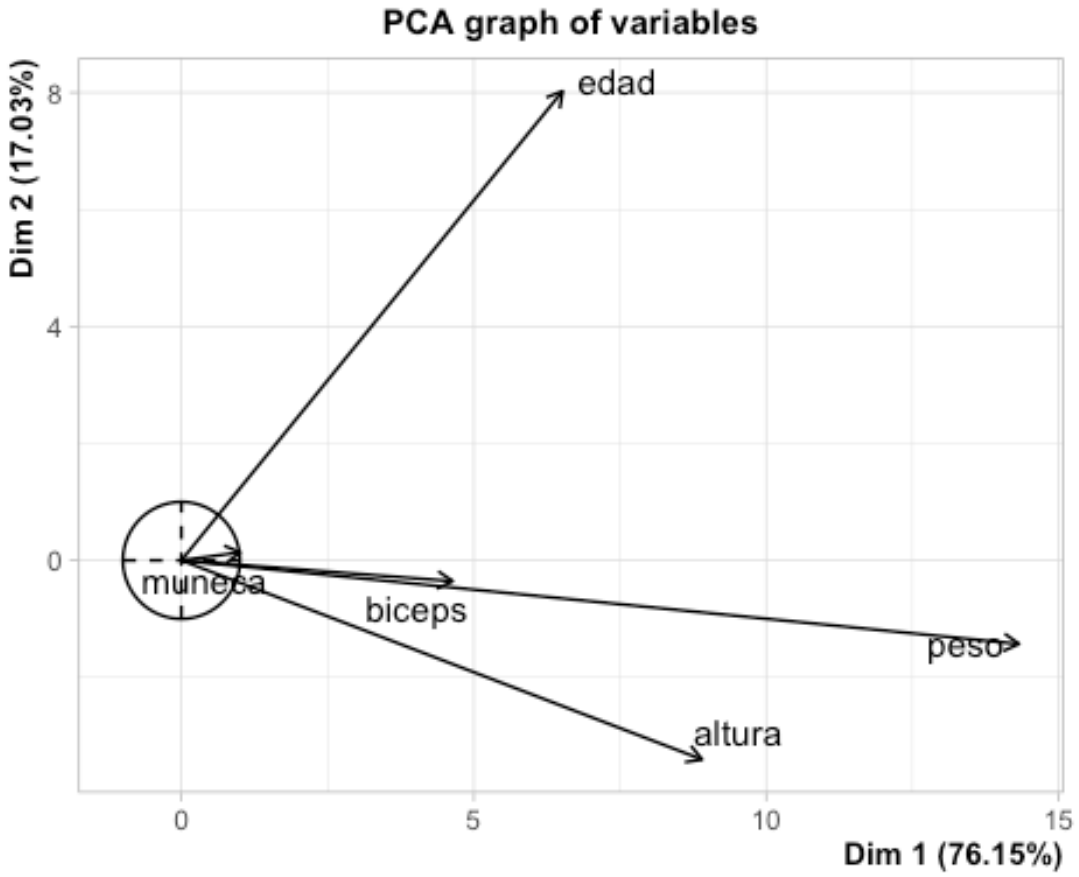
Detección de Datos Atípicos: Algunos puntos alejados sugieren la presencia de posibles valores atípicos que requieren análisis adicional.

Conclusiones Reducción dimensional efectiva: Los primeros dos componentes capturan la mayoría de la variabilidad. Importancia de la estandarización: La matriz R proporciona una visión equilibrada, evitando que variables con mayor varianza dominen el análisis. Identificación de variables clave: “Peso”, “altura” y “edad” son las variables más influyentes. Necesidad de análisis adicional: Los posibles datos atípicos identificados deben ser examinados con más detalle.

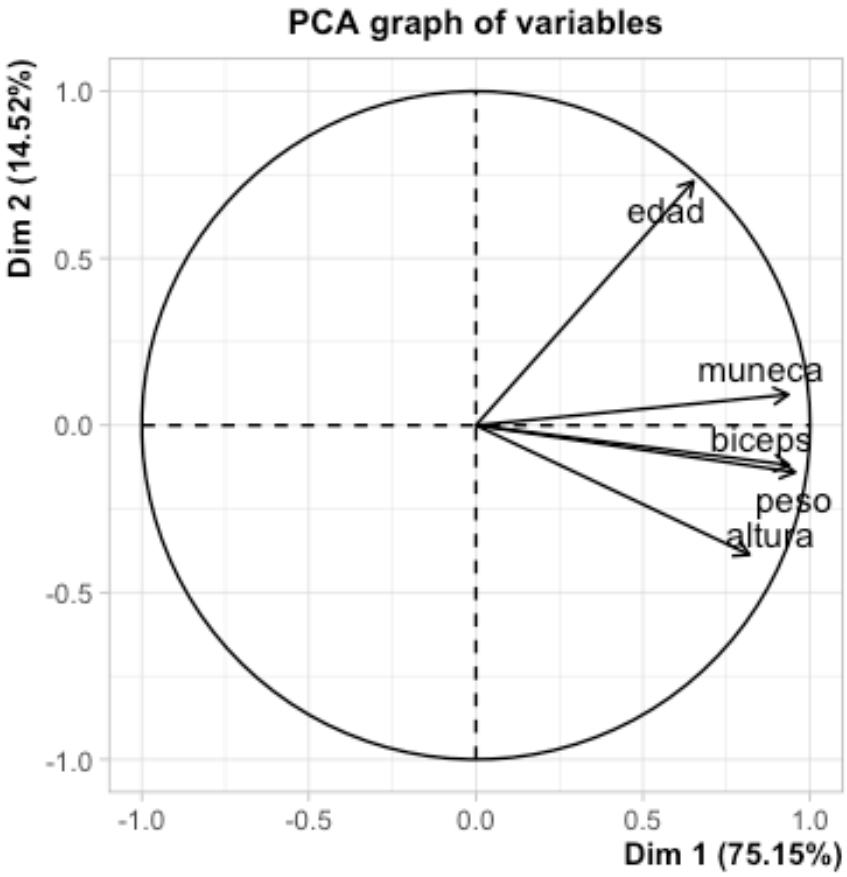
PARTE III

```
# 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
2
5
14
32
34
28
3
25
29
8
11
23
12
36
18
1
33
22
4
7
30
31
27
17
21
19
35
24
26
9
13
6
20
7
10
16
15

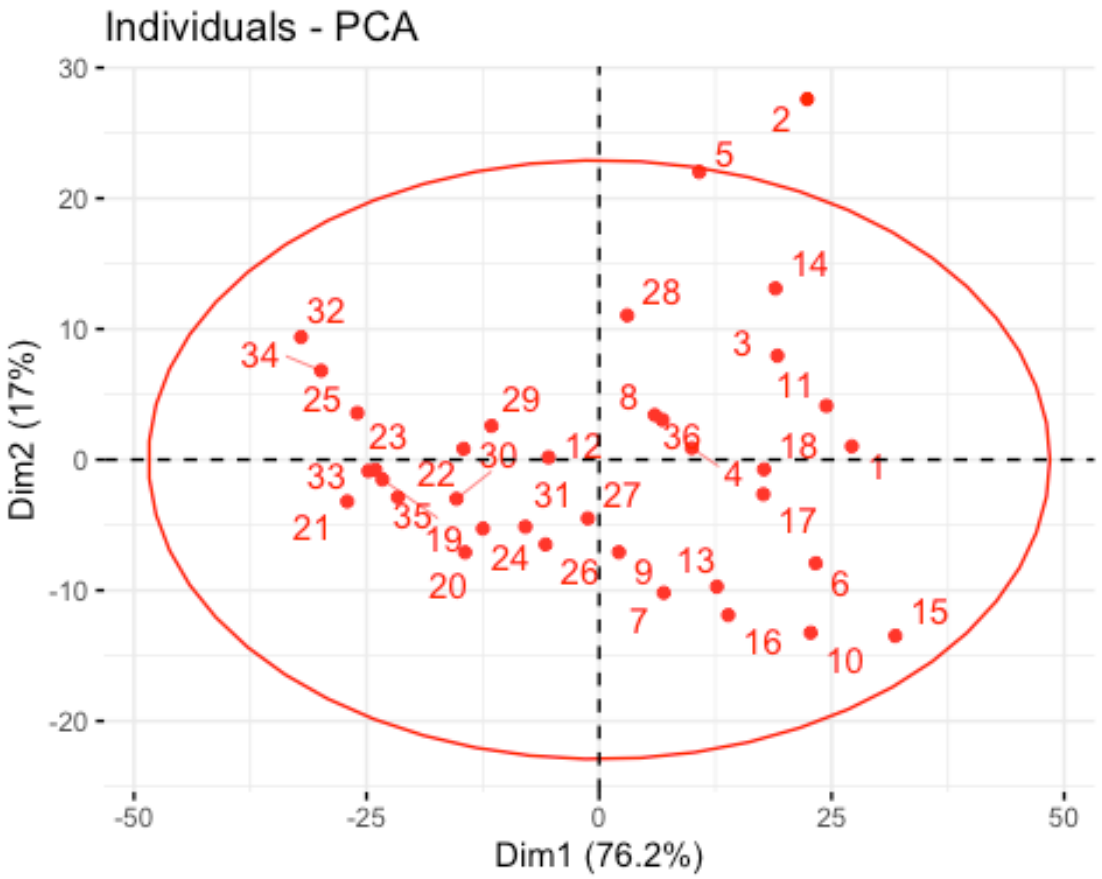


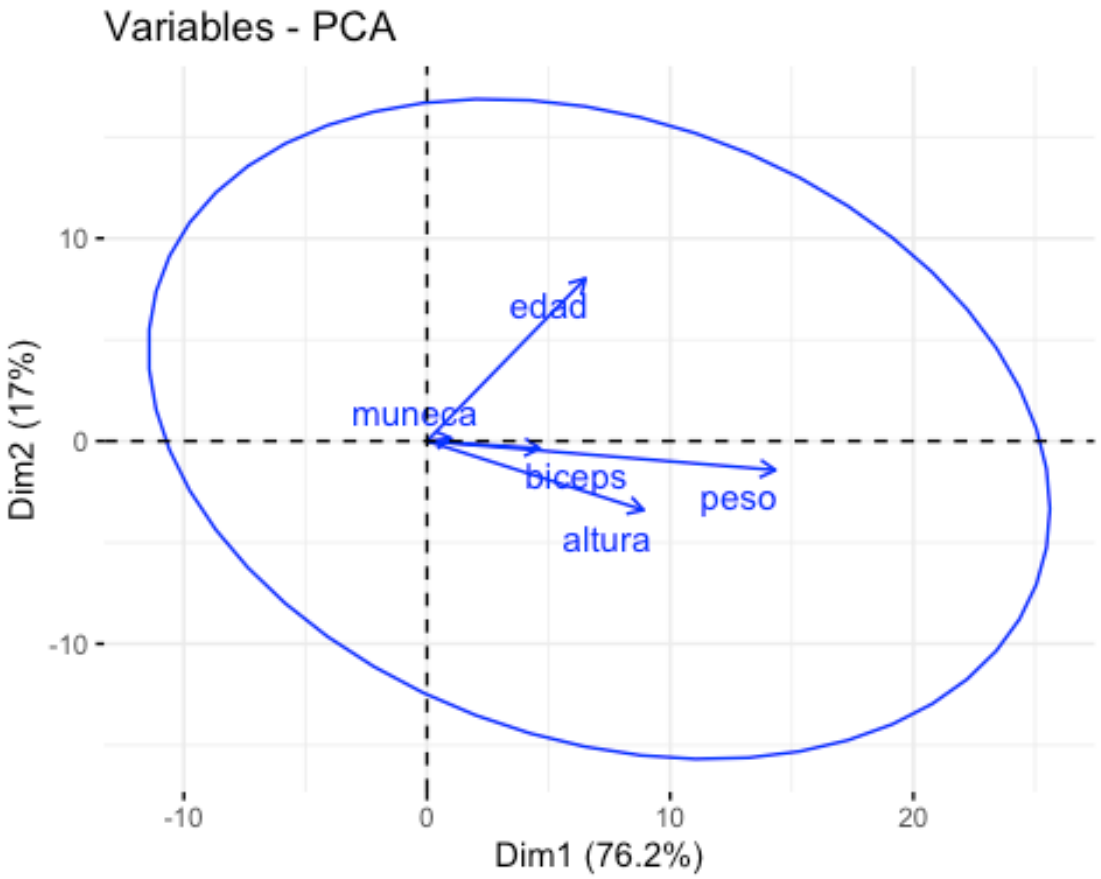

```r
# Para matriz de correlaciones
cpR <- PCA(datos_numericos, scale.unit = TRUE)
```

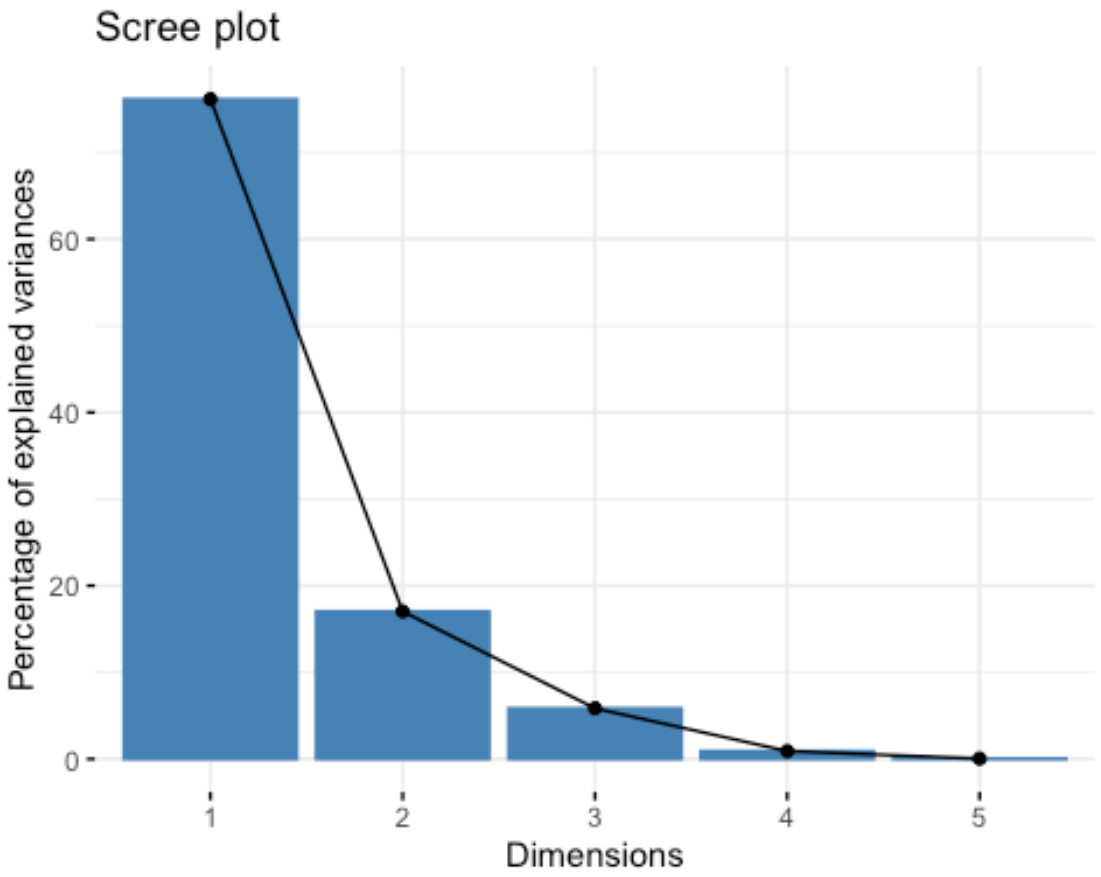


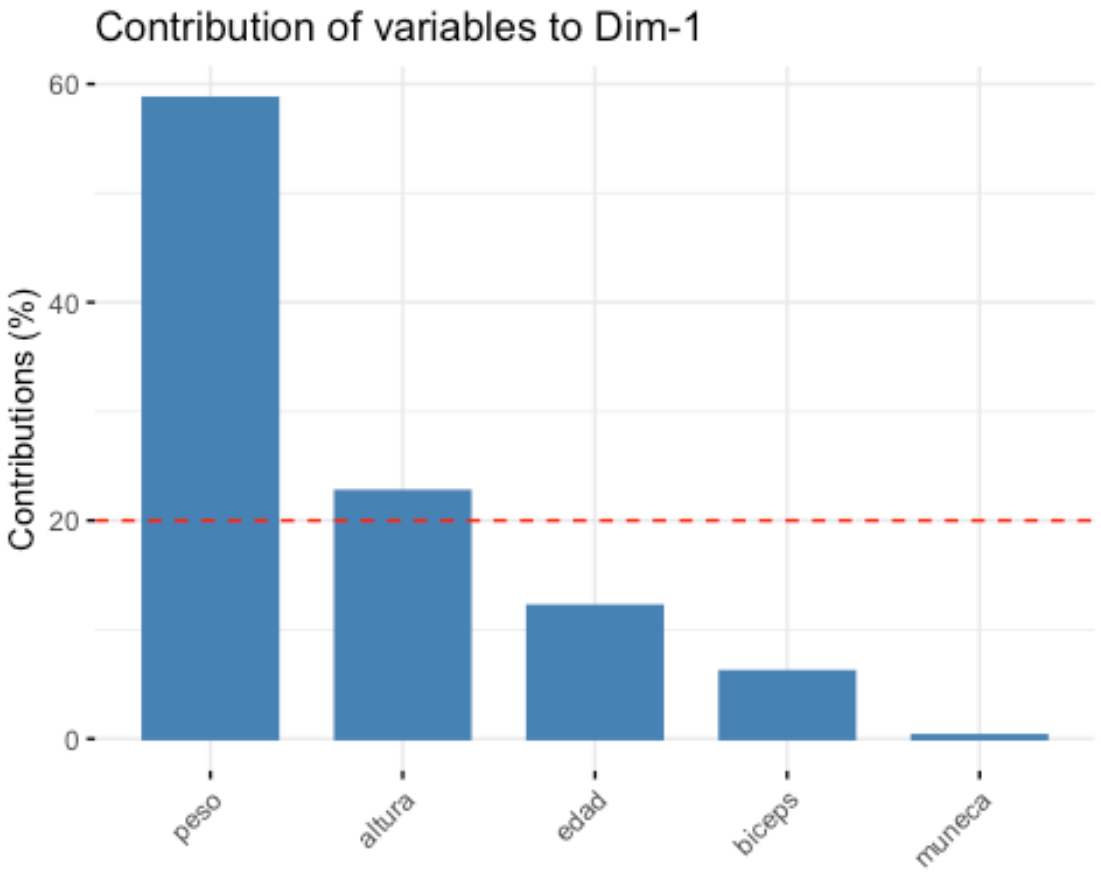

```r
# Función para generar e imprimir gráficos
generar_graficos <- function(pca_result) {
  # Gráfico de individuos
  p1 <- fviz_pca_ind(pca_result, col.ind = "red", addEllipses = TRUE, repel = TRUE)
  print(p1)

  # Gráfico de variables
  p2 <- fviz_pca_var(pca_result, col.var = "blue", 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 = "blue", col.ind =
```

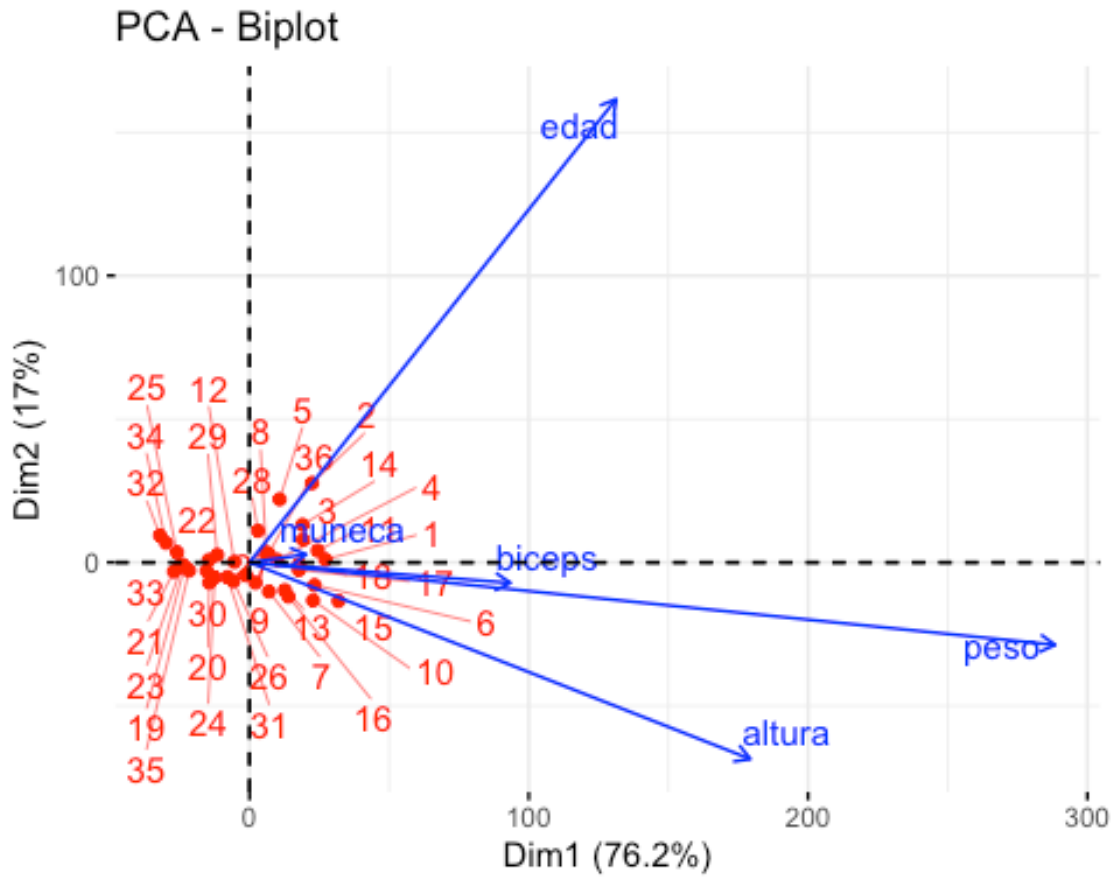
```
"red")
  print(p5)
}

# 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
2
5
14
28
32
34
3
25
29
8
11
23
36
18
30
12
4
1
33
22
31
27
35
17
21
19
24
26
9
13
6
20
7
15
16
10

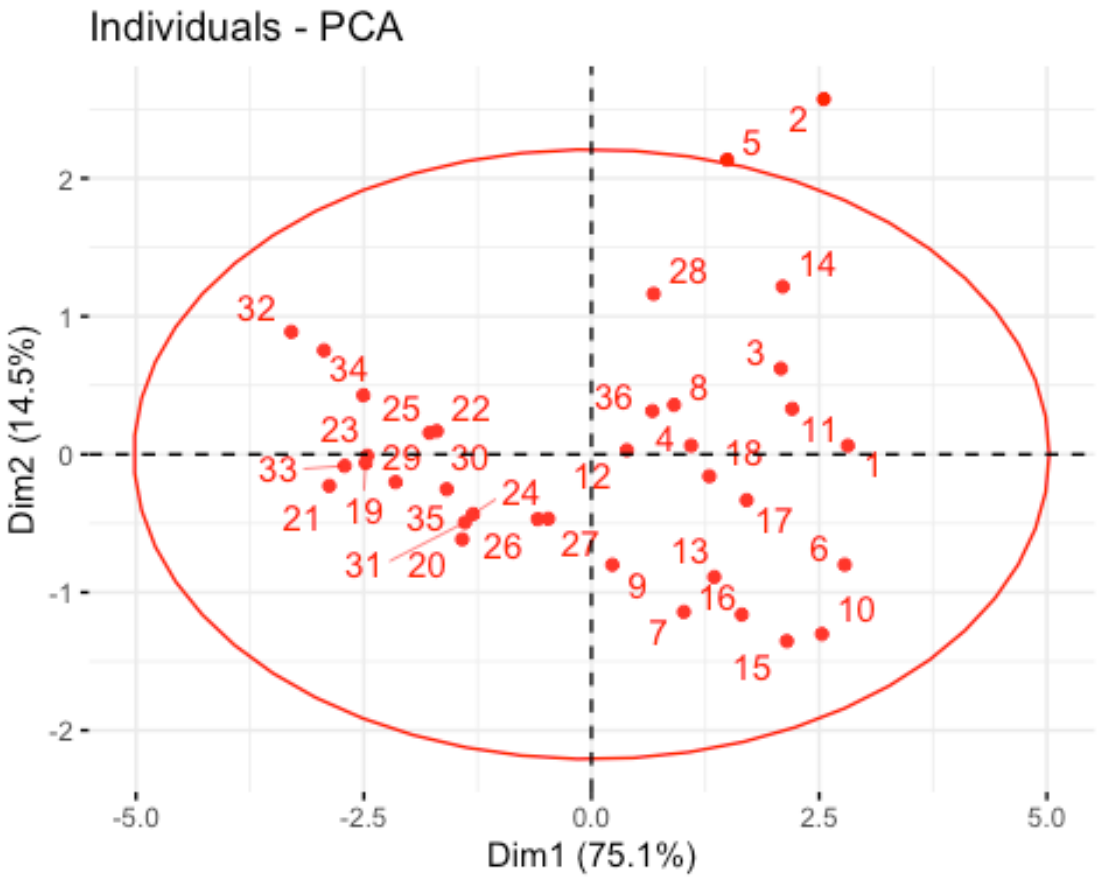

Variables - PCA
Dim2 (17%)
Dim1 (76.2%)
10
0
-10
-10
0
10
20
edad
muneca
biceps
peso
altura

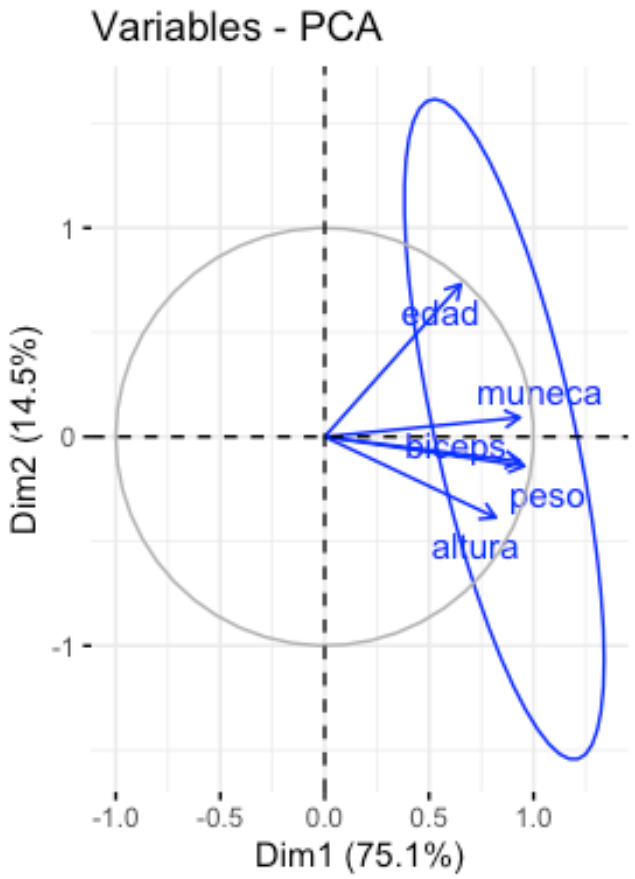

Scree plot
Percentage of explained variances
60
40
20
0
1
2
3
4
5
Dimensions

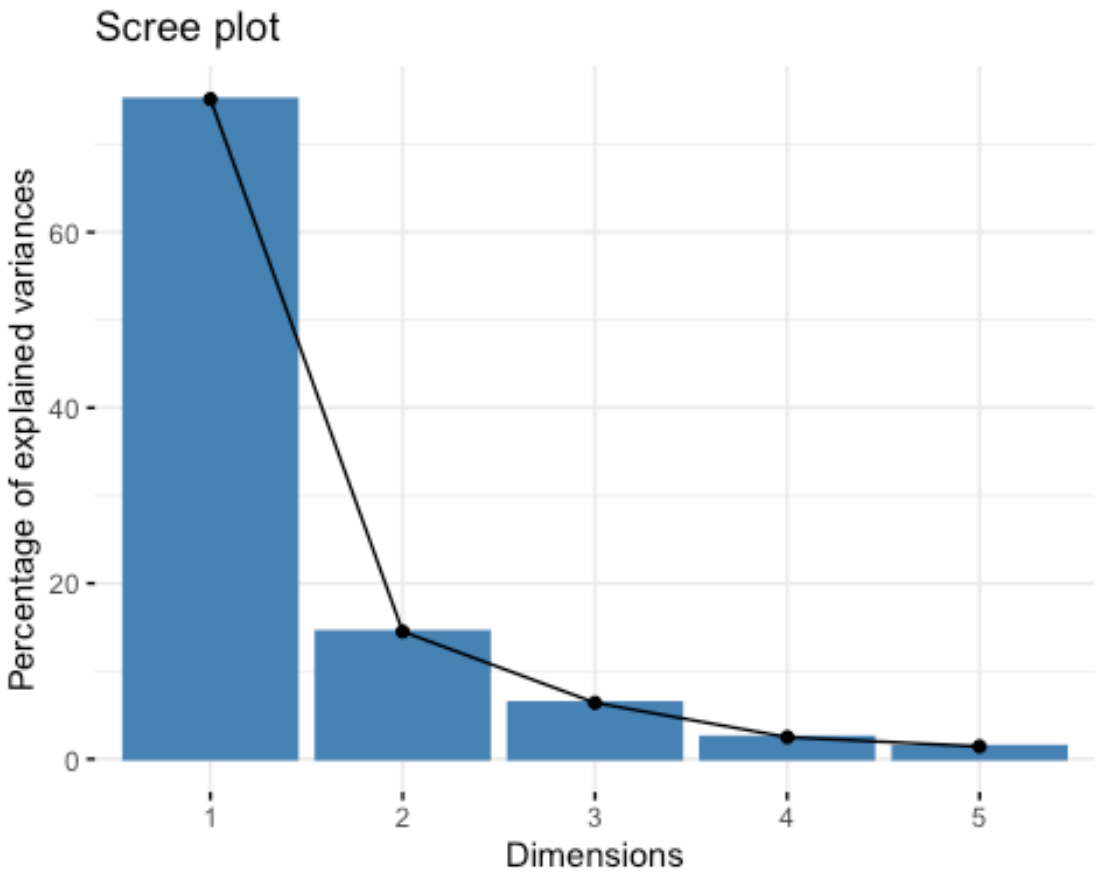

Contribution of variables to Dim-1
Contributions (%)
60
40
20
0
peso
altura
edad
biceps
muneca

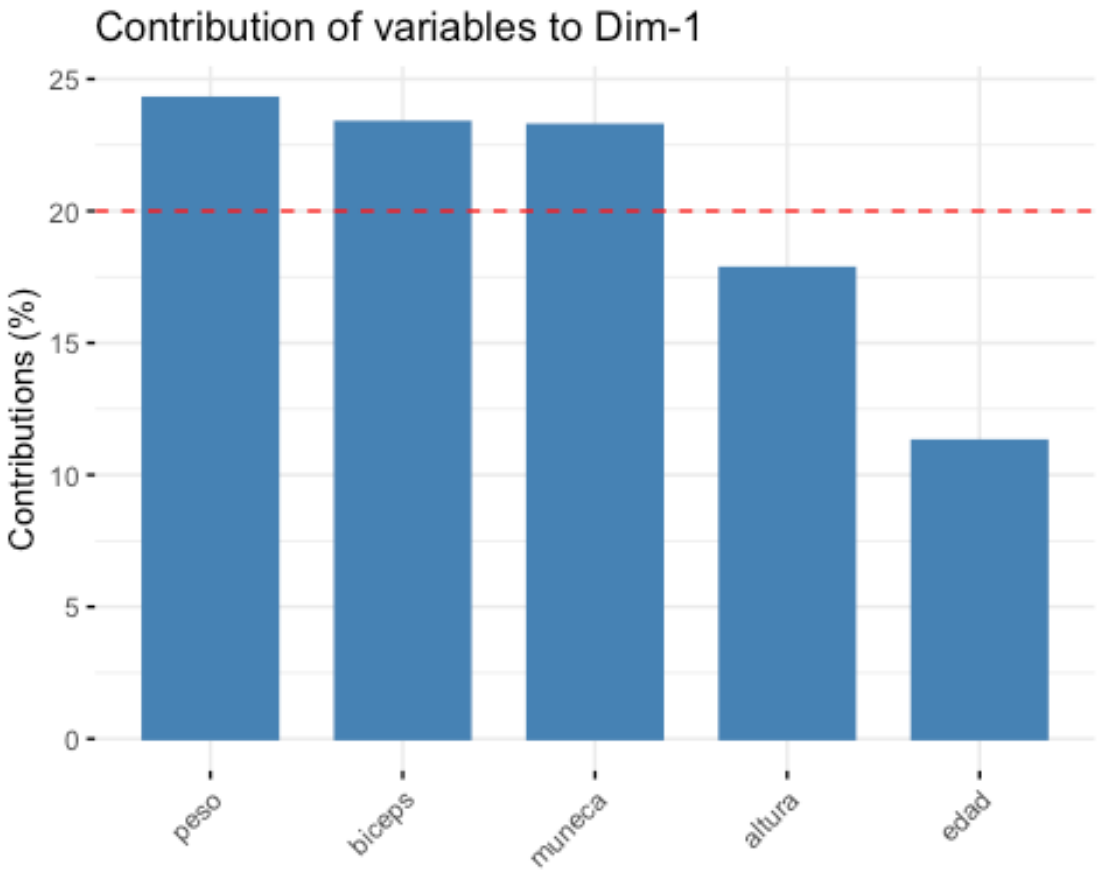


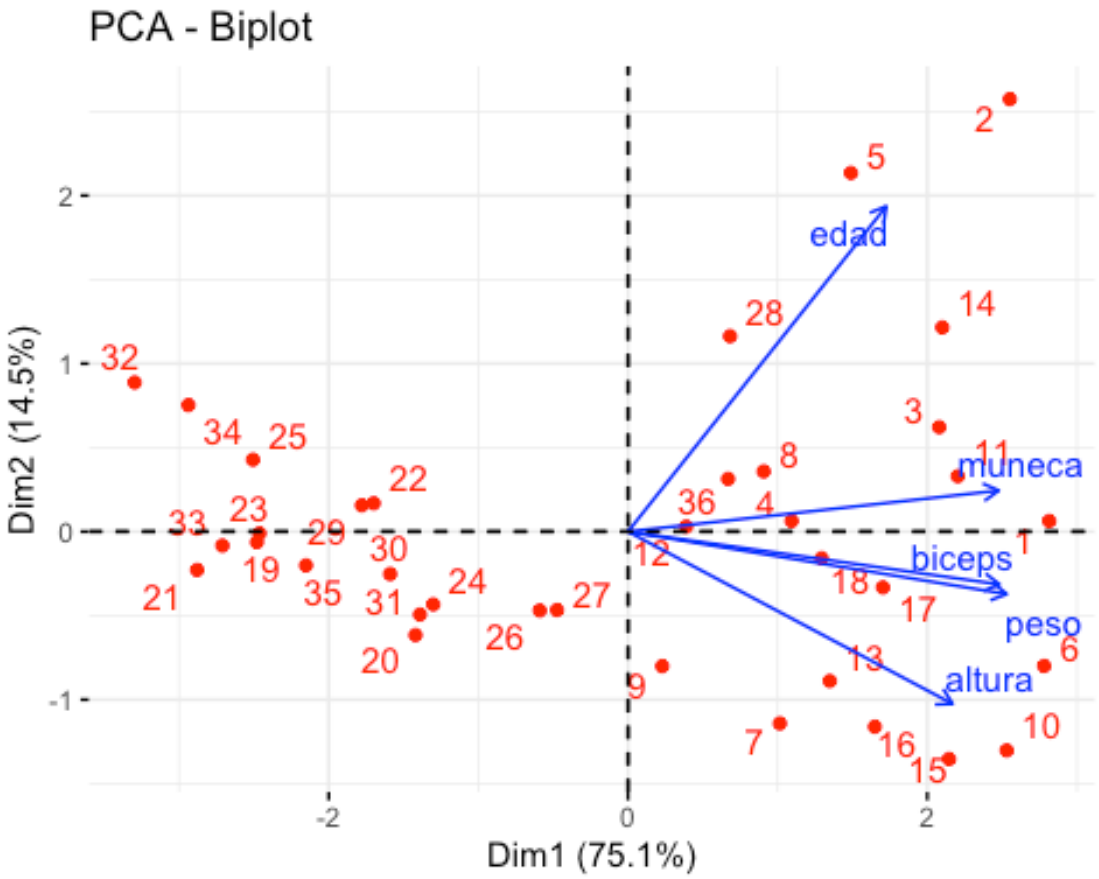

```r
# Generar gráficos para matriz de correlaciones
generar_graficos(cpR)
```


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


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


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 III

Gráfico de individuos (PCA - matriz de varianzas-covarianzas y correlación): Matriz S: Los individuos están más dispersos, reflejando las diferencias de escala entre variables. Matriz R: Los individuos se agrupan más cerca del origen debido a la estandarización, facilitando la comparación.

Gráfico de variables (PCA - matriz de varianzas-covarianzas y correlación): Matriz S: Variables como “peso” y “altura” tienen mayor influencia en el primer componente. Matriz R: Las contribuciones de las variables son más equilibradas; “peso”, “bíceps” y “muñeca” destacan en el primer componente.

Gráfico de sedimentación (Scree plot): Ambos análisis muestran que los dos primeros componentes explican la mayoría de la varianza, justificando la reducción dimensional.

Gráfico de contribuciones (Contribution plot): Identifica las variables que más aportan a cada componente. Matriz R: “Peso”, “bíceps” y “muñeca” son las más influyentes en el primer componente.

Biplot: Combina la información de individuos y variables. Matriz R: “Edad” está más asociada con el segundo componente, mientras que “peso” y “altura” influyen en el primero.

Conclusiones Estandarización beneficiosa: La matriz R ofrece una representación más equilibrada de las variables. Variables clave identificadas: “Peso”, “bíceps” y “muñeca” son fundamentales en la estructura de los datos. Dimensiones interpretables: Los componentes principales pueden interpretarse en términos de agrupaciones significativas de variables. Análisis adicional recomendado: Observaciones alejadas en los gráficos sugieren la necesidad de un estudio más profundo.

PARTE IV

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

Matriz de varianzas-covarianzas: Mantiene las escalas originales, por lo que las variables con mayor variabilidad tienen más influencia. El primer componente explica más del 75% de la varianza, pero puede no ser adecuado si las unidades de las variables difieren significativamente, ya que podría sesgar el análisis.

Matriz de correlación: Estandariza las variables, equilibrando su influencia independientemente de sus unidades originales. Los primeros dos componentes explican conjuntamente cerca del 90% de la varianza, proporcionando una representación más justa cuando las variables son heterogéneas.

Conclusión Comparativa: Procedimiento recomendado: Utilizar la matriz de correlación, especialmente cuando las variables tienen diferentes escalas y unidades, como en datos económicos y sociales.

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

Primera Componente (Dim1): Las variables “peso”, “bíceps” y “muñeca” son las que más contribuyen, indicando su importancia en la variabilidad capturada por esta dimensión.

Segunda Componente (Dim2): La variable “edad” es la más influyente, representando una dimensión distinta no relacionada con las variables físicas.

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 reflejan la contribución relativa de cada variable en los componentes principales.

4.4 Interpretación en términos de agrupación de variables

“Índice de desarrollo físico” La primera componente puede interpretarse como un indicador del desarrollo físico general, dado que variables como “peso”, “bíceps” y “altura” son las más influyentes.

“Índice de edad o madurez”: La segunda componente representa principalmente la “edad”, diferenciando las observaciones en función de este factor.

Conclusión General

El análisis de componentes principales utilizando la matriz de correlación proporciona una comprensión más equilibrada y significativa de los datos. Las variables físicas como “peso” y “bíceps” dominan la primera dimensión, mientras que “edad” es crucial en la segunda. Este enfoque permite identificar las dimensiones clave que explican las diferencias entre las observaciones, facilitando la interpretación y la toma de decisiones basadas en los datos.