A4-Componentes Principales

Luis Maximiliano López Ramírez

2024-10-08

En la base de datos Corporal Download Corporalcontiene 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.

```
datos <- read.csv("corporal.csv")
```

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

```
# Supongamos que tu dataset se llama data y la columna es 'sexo'
datos$sexo <- NULL

# Mostrar las primeras filas del dataset para tener una vista preliminar
head(datos)
```

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

```
# Obtener las estadísticas descriptivas generales
summary(datos)
```

```
##       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
```

```r
# Calcular la desviación estándar de cada columna numérica
# Utilizando lapply para aplicar sd() a cada columna de tipo numérico
sapply(datos, function(x) if(is.numeric(x)) sd(x, na.rm = TRUE))
```

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

```r
# Calcular la matriz de correlación
cor_matrix <- cor(datos, use = "complete.obs")

# Mostrar la matriz de correlación
cor_matrix
```

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

 # 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

```r
X <- datos

# Calcular la matriz de varianza-covarianza S
S <- cov(X)

# Mostrar la matriz de varianza-covarianza
print("Matriz de varianza-covarianza (S):")
```

```
## [1] "Matriz de varianza-covarianza (S):"
```

```r
print(S)
```

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

```r
# Calcular la matriz de correlación R
R <- cor(X)
```

```r
# Mostrar la matriz de correlación
print("Matriz de correlación (R):")
```

```
## [1] "Matriz de correlación (R):"
```

```r
print(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
```

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

```r
# Matriz de varianza-covarianza
S <- cov(X)

# Matriz de correlación
R <- cor(X)

# Calcular autovalores y autovectores para la matriz de varianza-
covarianza S
eigen_S <- eigen(S)

# Calcular autovalores y autovectores para la matriz de correlación R
eigen_R <- eigen(R)

# Imprimir los autovalores y autovectores para S
print("Autovalores de la matriz de varianza-covarianza (S):")
```

```
## [1] "Autovalores de la matriz de varianza-covarianza (S):"
```

```r
print(eigen_S$values)
```

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

```r
print("Autovectores de la matriz de varianza-covarianza (S):")
```

```
## [1] "Autovectores de la matriz de varianza-covarianza (S):"
```

```r
print(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
```

```r
print("")
```

```
## [1] ""
```

```r
# Imprimir los autovalores y autovectores para R
print("Autovalores de la matriz de correlación (R):")
```

```
## [1] "Autovalores de la matriz de correlación (R):"
```

```r
print(eigen_R$values)
```

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

```r
print("Autovectores de la matriz de correlación (R):")
```

```
## [1] "Autovectores de la matriz de correlación (R):"
```

```r
print(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
```

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.

```r
# Calcular la varianza total como la suma de las varianzas de la diagonal de S
varianza_total <- sum(diag(S))

# Mostrar la varianza total
print(paste("Varianza total (suma de las varianzas de la diagonal de S):", varianza_total))
```

```
## [1] "Varianza total (suma de las varianzas de la diagonal de S): 471.9385"
```

```r
# Calcular la suma de los autovalores de S (que debería ser igual a la varianza total)
suma_autovalores_S <- sum(eigen_S$values)

# Verificar que la suma de los autovalores sea igual a la varianza total
```

```r
print(paste("Suma de los autovalores de la matriz S:",
suma_autovalores_S))
```

```
## [1] "Suma de los autovalores de la matriz S: 471.938499999999"
```

```r
# Proporción de varianza explicada por cada componente usando la matriz S
proporcion_varianza_S <- eigen_S$values / varianza_total

# Mostrar la proporción de varianza explicada por cada componente
print("Proporción de varianza explicada por cada componente (S):")
```

```
## [1] "Proporción de varianza explicada por cada componente (S):"
```

```r
print(proporcion_varianza_S)
```

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

```r
# Calcular la varianza total de la matriz de correlación (número de
variables)
varianza_total_R <- ncol(X)

# Proporción de varianza explicada por cada componente usando la matriz R
proporcion_varianza_R <- eigen_R$values / varianza_total_R

# Mostrar la proporción de varianza explicada por cada componente
print("Proporción de varianza explicada por cada componente (R):")
```

```
## [1] "Proporción de varianza explicada por cada componente (R):"
```

```r
print(proporcion_varianza_R)
```

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

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

```r
# Calcular la varianza acumulada
varianza_acumulada_S <- cumsum(proporcion_varianza_S)

# Mostrar la varianza acumulada
print("Varianza acumulada (S):")
```

```
## [1] "Varianza acumulada (S):"
```

```r
print(varianza_acumulada_S)
```

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

```r
# Calcular la varianza acumulada
varianza_acumulada_R <- cumsum(proporcion_varianza_R)

# Mostrar la varianza acumulada
print("Varianza acumulada (R):")
```

```
## [1] "Varianza acumulada (R):"

print(varianza_acumulada_R)

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

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

Los resultados de la varianza acumulada muestran que en ambas matrices, de varianza-covarianza (S) y de correlación (R), los primeros dos componentes son los más importantes. En la matriz de varianza-covarianza, los dos primeros componentes explican aproximadamente el 93% de la varianza total, mientras que en la matriz de correlación explican alrededor del 89% de la varianza. Los componentes adicionales solo añaden una pequeña cantidad de varianza, por lo que no son tan necesarios para el análisis. Por lo tanto, para capturar la mayor parte de la variabilidad de los datos y reducir la dimensionalidad, se recomienda enfocarse en los primeros dos componentes, ya que proporcionan una representación adecuada de la estructura del dataset.

5. Escriba la ecuación de la combinación lineal de los Componentes principales CP1 y CP2 (eiX, donde ei está en eigen(S)$vectors[1], e2X para obtener CP2, donde X = c(X1, X2, ...)) ¿qué variables son las que más contribuyen a la primera y segunda componentes principales? (observe los coeficientes en valor absoluto de las combinaciones lineales). Justifique su respuesta.

```
# Extraer los vectores propios (autovectores) de S
autovectores <- eigen_S$vectors

# Mostrar los vectores propios para ver los coeficientes
print("Vectores propios (autovectores) de la matriz S:")

## [1] "Vectores propios (autovectores) de la matriz S:"

print(autovectores)

##             [,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

# Combinación lineal de las variables para CP1 y CP2
CP1 <- paste0("CP1 = ", round(autovectores[1, 1], 4), " * X1 + ",
              round(autovectores[2, 1], 4), " * X2 + ",
              round(autovectores[3, 1], 4), " * X3 + ",
              round(autovectores[4, 1], 4), " * X4 + ",
              round(autovectores[5, 1], 4), " * X5")

CP2 <- paste0("CP2 = ", round(autovectores[1, 2], 4), " * X1 + ",
```

```
              round(autovectores[2, 2], 4), " * X2 + ",
              round(autovectores[3, 2], 4), " * X3 + ",
              round(autovectores[4, 2], 4), " * X4 + ",
              round(autovectores[5, 2], 4), " * X5")

# Imprimir las ecuaciones de CP1 y CP2
print(CP1)

## [1] "CP1 = -0.3487 * X1 + -0.7662 * X2 + -0.4763 * X3 + -0.0539 * X4 +
-0.2482 * X5"

print(CP2)

## [1] "CP2 = 0.9076 * X1 + -0.1617 * X2 + -0.3852 * X3 + 0.0155 * X4 + -
0.0402 * X5"

contribucion_CP1 <- abs(autovectores[, 1])
contribucion_CP2 <- abs(autovectores[, 2])

# Mostrar las contribuciones de las variables a CP1 y CP2
print("Contribuciones a CP1 (en valor absoluto):")

## [1] "Contribuciones a CP1 (en valor absoluto):"

print(contribucion_CP1)

## [1] 0.34871002 0.76617586 0.47632405 0.05386189 0.24817367

print("Contribuciones a CP2 (en valor absoluto):")

## [1] "Contribuciones a CP2 (en valor absoluto):"

print(contribucion_CP2)

## [1] 0.9075501 0.1616581 0.3851755 0.0155423 0.0402221
```

La variable que más influye en el primer componente PC1 es X2 (peso) con un coeficiente de -0.7662 mientras que la variable que más influye en el segundo componente PC2 es X1 (edad) dado que su coeficiente es el mayor de todo PC2 con 0.9076.

2. No te olvides de seguir los mismos pasos con la matriz de correlaciones (se obtiene con cor(x) si x está compuesto por variables numéricas)

```
# Extraer los vectores propios (autovectores) de S
autovectores <- eigen_R$vectors

# Mostrar los vectores propios para ver los coeficientes
print("Vectores propios (autovectores) de la matriz S:")

## [1] "Vectores propios (autovectores) de la matriz S:"

print(autovectores)
```

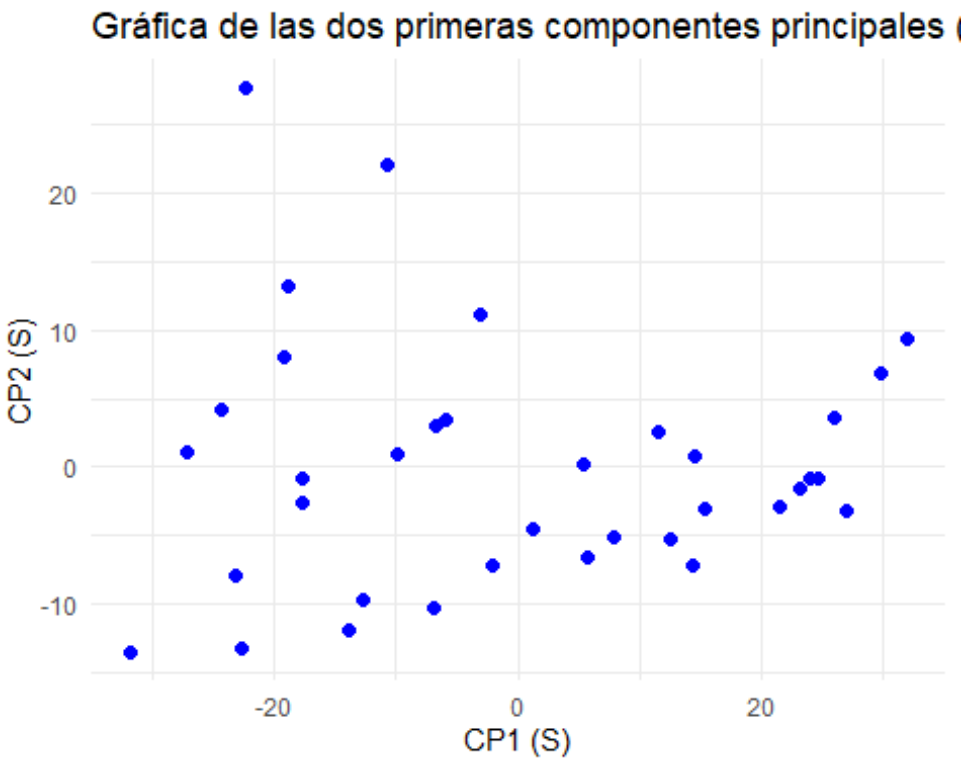
```
##            [,1]       [,2]        [,3]       [,4]       [,5]
## [1,] -0.3359310  0.8575601 -0.34913780 -0.1360111  0.1065123
## [2,] -0.4927066 -0.1647821  0.06924561 -0.5249533 -0.6706087
## [3,] -0.4222426 -0.4542223 -0.73394453  0.2070673  0.1839617
## [4,] -0.4821923  0.1082775  0.36690716  0.7551547 -0.2255818
## [5,] -0.4833139 -0.1392684  0.44722747 -0.3046138  0.6739511

# Combinación lineal de las variables para CP1 y CP2
CP1 <- paste0("CP1 = ", round(autovectores[1, 1], 4), " * X1 + ",
              round(autovectores[2, 1], 4), " * X2 + ",
              round(autovectores[3, 1], 4), " * X3 + ",
              round(autovectores[4, 1], 4), " * X4 + ",
              round(autovectores[5, 1], 4), " * X5")

CP2 <- paste0("CP2 = ", round(autovectores[1, 2], 4), " * X1 + ",
              round(autovectores[2, 2], 4), " * X2 + ",
              round(autovectores[3, 2], 4), " * X3 + ",
              round(autovectores[4, 2], 4), " * X4 + ",
              round(autovectores[5, 2], 4), " * X5")

# Imprimir las ecuaciones de CP1 y CP2
print(CP1)

## [1] "CP1 = -0.3359 * X1 + -0.4927 * X2 + -0.4222 * X3 + -0.4822 * X4 +
-0.4833 * X5"

print(CP2)

## [1] "CP2 = 0.8576 * X1 + -0.1648 * X2 + -0.4542 * X3 + 0.1083 * X4 + -
0.1393 * X5"

contribucion_CP1 <- abs(autovectores[, 1])
contribucion_CP2 <- abs(autovectores[, 2])

# Mostrar las contribuciones de las variables a CP1 y CP2
print("Contribuciones a CP1 (en valor absoluto):")

## [1] "Contribuciones a CP1 (en valor absoluto):"

print(contribucion_CP1)

## [1] 0.3359310 0.4927066 0.4222426 0.4821923 0.4833139

print("Contribuciones a CP2 (en valor absoluto):")

## [1] "Contribuciones a CP2 (en valor absoluto):"

print(contribucion_CP2)

## [1] 0.8575601 0.1647821 0.4542223 0.1082775 0.1392684
```

La variable que más influye en el primer componente PC1 es X2 (peso) con un coeficiente de -0.4927 siguiendole de cerca X4 (muneca) y X5 (biceps) con -0.4822 y -0.4833 respectivamente. Mientras que la variable que más influye en el segundo componente PC2 es X1 (edad) dado que su coeficiente es el mayor de todo PC2 con 0.8576.

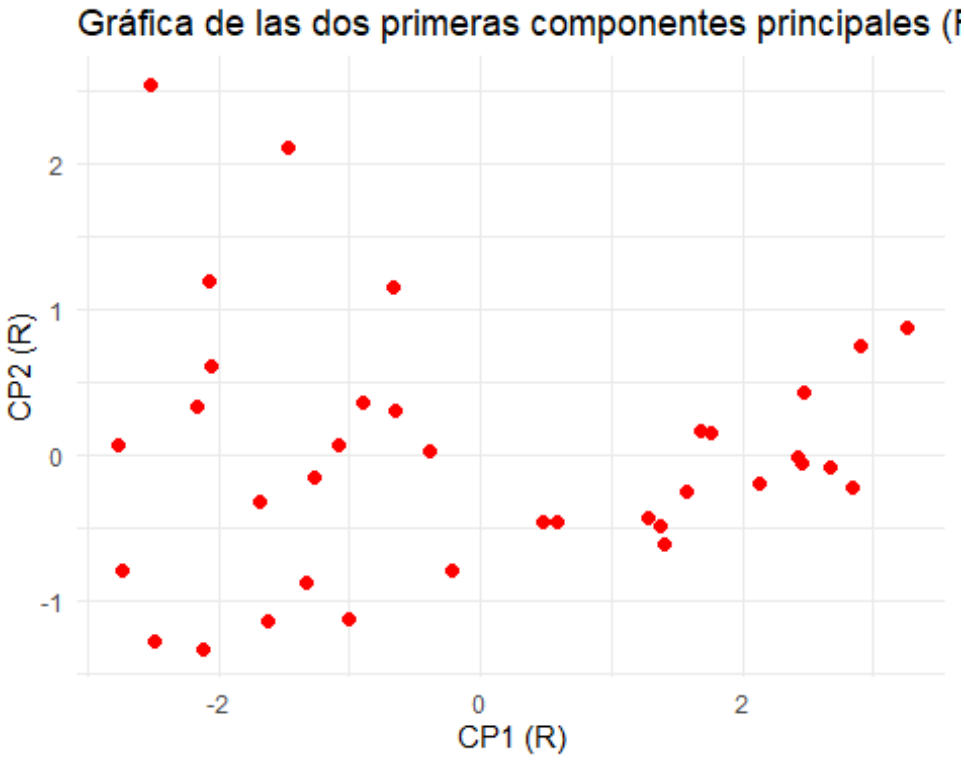
 # Parte 2

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

```
# Paso 2: Realizar el PCA con la matriz de varianza-covarianza (S)
pca_S <- prcomp(X, scale. = FALSE)  # scale. = FALSE para usar la matriz
S

# Paso 3: Realizar el PCA con la matriz de correlación (R)
pca_R <- prcomp(X, scale. = TRUE)   # scale. = TRUE para usar la matriz R

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.2

# Extraer los resultados del PCA con S
pca_S_data <- data.frame(pca_S$x, Group = 1:nrow(X))

# Extraer los resultados del PCA con R
pca_R_data <- data.frame(pca_R$x, Group = 1:nrow(X))

# Graficar la primera y segunda componente principal (PC1 y PC2) para S
plot_S <- ggplot(pca_S_data, aes(x = PC1, y = PC2)) +
  geom_point(color = 'blue', size = 2) +
  ggtitle("Gráfica de las dos primeras componentes principales (S)") +
  theme_minimal() +
  xlab("CP1 (S)") +
  ylab("CP2 (S)")

# Graficar la primera y segunda componente principal (PC1 y PC2) para R
plot_R <- ggplot(pca_R_data, aes(x = PC1, y = PC2)) +
  geom_point(color = 'red', size = 2) +
  ggtitle("Gráfica de las dos primeras componentes principales (R)") +
  theme_minimal() +
  xlab("CP1 (R)") +
  ylab("CP2 (R)")

# Mostrar las gráficas en pantalla
print(plot_S)
```


Gráfica de las dos primeras componentes principales (
20
10
0
-10
CP2 (S)
-20
0
20
CP1 (S)

```
print(plot_R)
```


Gráfica de las dos primeras componentes principales (R
2
1
0
-1
CP2 (R)
-2
0
2
CP1 (R)

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

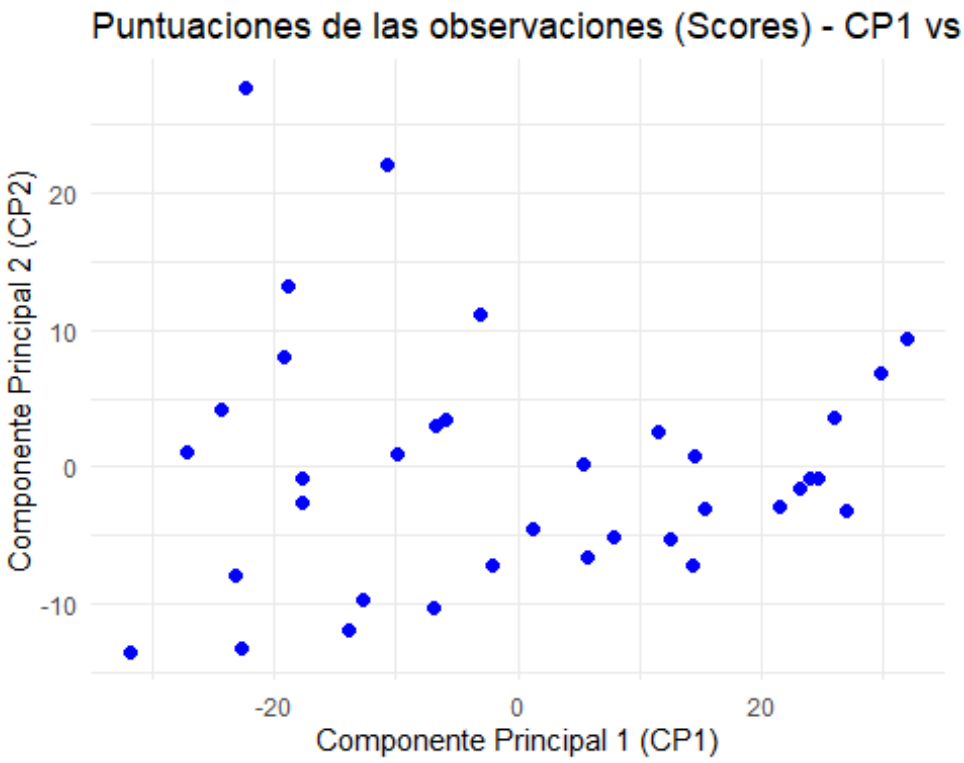
```r
# Paso 2: Calcular la matriz de varianza-covarianza S
S <- cov(X)

# Paso 3: Realizar el PCA con la matriz de varianza-covarianza (sin escalar)
pca_S <- prcomp(X, scale. = FALSE)  # scale. = FALSE asegura que use la matriz de varianza-covarianza S

# Paso 4: Extraer las puntuaciones (scores) de las observaciones
puntuaciones <- pca_S$x

# Mostrar las primeras filas de las puntuaciones
print("Puntuaciones (scores) de las observaciones para los componentes principales:")
```

```
## [1] "Puntuaciones (scores) de las observaciones para los componentes principales:"
```

```r
head(puntuaciones)
```

```
##             PC1        PC2        PC3         PC4         PC5
## [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
```

```r
library(ggplot2)

# Crear un dataframe con las puntuaciones para ggplot
puntuaciones_df <- data.frame(puntuaciones)

# Graficar las puntuaciones de las dos primeras componentes principales
ggplot(puntuaciones_df, aes(x = PC1, y = PC2)) +
  geom_point(color = 'blue', size = 2) +
  ggtitle("Puntuaciones de las observaciones (Scores) - CP1 vs CP2") +
  theme_minimal() +
  xlab("Componente Principal 1 (CP1)") +
  ylab("Componente Principal 2 (CP2)")
```



El resultado es una matriz de puntuaciones donde cada fila representa una observación y cada columna corresponde a un componente principal (PC1, PC2, PC3, etc.).

Las puntuaciones indican la posición de cada observación en el nuevo espacio de componentes principales.

2. Calcule las puntuaciones (scores) de las observaciones para los componentes obtenidos con la matriz de correlaciones. Recuerde que en la matriz de correlaciones las variables tienen que estar estandarizadas.

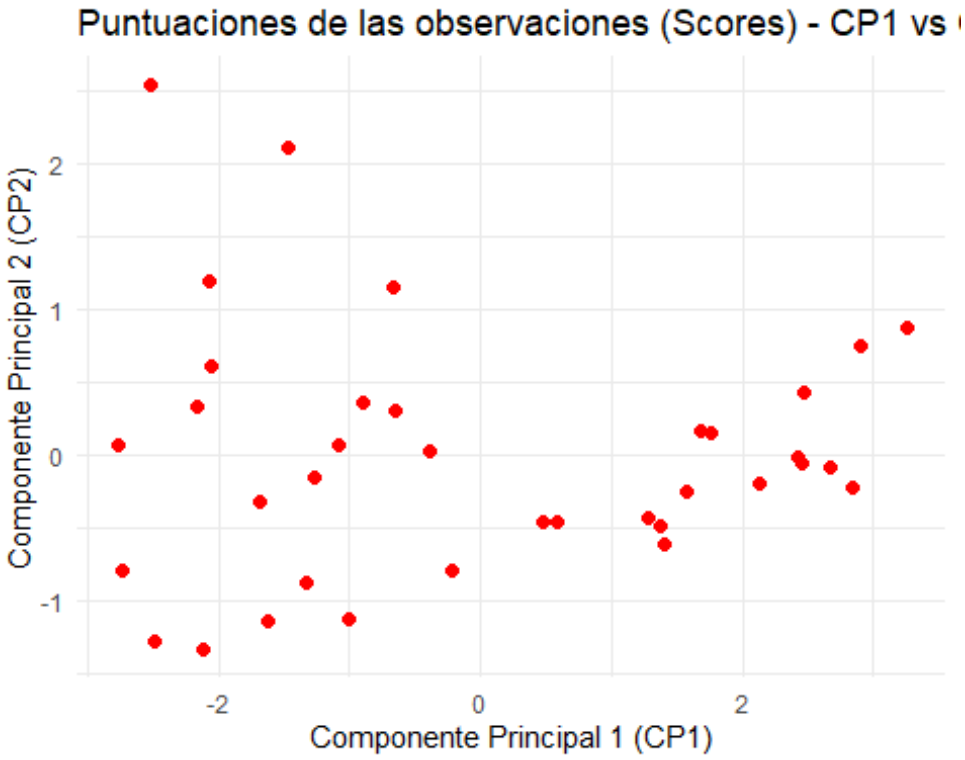
```
# Paso 2: Realizar el PCA usando la matriz de correlación (variables estandarizadas)
pca_R <- prcomp(X, scale. = TRUE)  # scale. = TRUE estandariza las variables antes de realizar el PCA

# Paso 3: Extraer las puntuaciones (scores) de las observaciones
puntuaciones_R <- pca_R$x

# Mostrar las primeras filas de las puntuaciones
print("Puntuaciones (scores) de las observaciones para los componentes principales (Matriz de correlaciones):")
```

```
## [1] "Puntuaciones (scores) de las observaciones para los componentes principales (Matriz de correlaciones):"
```

```
head(puntuaciones_R)
```

```
##             PC1         PC2         PC3         PC4          PC5
## [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
```

```r
library(ggplot2)

# Crear un dataframe con las puntuaciones para ggplot
puntuaciones_R_df <- data.frame(puntuaciones_R)

# Graficar las puntuaciones de las dos primeras componentes principales
ggplot(puntuaciones_R_df, aes(x = PC1, y = PC2)) +
  geom_point(color = 'red', size = 2) +
  ggtitle("Puntuaciones de las observaciones (Scores) - CP1 vs CP2
(Matriz de correlaciones)") +
  theme_minimal() +
  xlab("Componente Principal 1 (CP1)") +
  ylab("Componente Principal 2 (CP2)")
```



La matriz de puntuaciones (puntuaciones_R) tendrá tantas filas como observaciones y tantas columnas como componentes principales (PC1, PC2, PC3, etc.).

Las puntuaciones indican la posición de cada observación en el espacio de los componentes principales.

2. Interprete los gráficos en términos de:

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

Gráfico de las componentes principales utilizando la matriz de varianza-covarianza (S):

- En este gráfico, las dos primeras componentes principales (CP1 y CP2) explican la mayor parte de la varianza de los datos originales.
- La dispersión de los puntos a lo largo de CP1 indica que esta componente capta la mayor variabilidad en los datos, lo que sugiere que las observaciones con valores extremos en CP1 son aquellas con diferencias significativas en las variables originales.
- La relación entre CP1 y CP2 parece lineal en algunas partes, lo que indica que hay una correlación entre ciertas variables en los componentes.
- Como las variables no están estandarizadas, las variables con varianzas más altas tienen un mayor impacto en la construcción de CP1 y CP2.

Gráfico de las componentes principales utilizando la matriz de correlación (R):

- La utilización de la matriz de correlación implica que las variables han sido estandarizadas, por lo que cada variable contribuye de igual manera al análisis.
- En este gráfico, CP1 y CP2 se distribuyen más simétricamente alrededor del origen (0,0), lo que refleja que las variables tienen el mismo peso en la construcción de estos componentes.
- La dispersión de puntos a lo largo de CP1 y CP2 indica que algunas variables están más relacionadas con CP1, mientras que otras están más relacionadas con CP2. Por ejemplo, puntos que están lejos del centro indican observaciones con características que difieren de la media en las variables correspondientes.
- La diferencia en la dispersión entre el gráfico con S y el gráfico con R muestra cómo la estandarización afecta la construcción de los componentes principales, dándole a cada variable la misma importancia.

Comparación entre ambos gráficos:

- Matriz de varianza-covarianza (S): Las variables con varianzas mayores dominan la construcción de los componentes, lo que puede sesgar la interpretación hacia aquellas variables.
- Matriz de correlación (R): Las variables tienen la misma importancia, permitiendo que los componentes principales reflejen mejor las relaciones estructurales sin sesgo de escala.

2. La relación entre las puntuaciones de las observaciones y los valores de las variables

Las puntuaciones obtenidas de la matriz de varianza-covarianza (S) muestran que la mayor variabilidad se captura en los primeros dos componentes principales (PC1 y PC2), con valores extremos para algunas observaciones (e.g., observaciones 1, 2 y 6). Esto indica que estas observaciones se desvían significativamente de la media en las variables originales que dominan estas componentes. Los componentes PC3, PC4, y PC5 tienen valores más cercanos a cero, lo que sugiere que explican una porción menor de la variabilidad en el dataset. En general, las observaciones con puntuaciones muy negativas o positivas en PC1 y PC2 reflejan diferencias notables en las características que influyen en estos componentes.

Por otro lado, las puntuaciones calculadas con la matriz de correlación (R) muestran valores más equilibrados entre los componentes debido a la estandarización de las variables. Las observaciones con valores extremos en PC1 y PC2 (e.g., observaciones 2 y 5) tienen características distintivas en las combinaciones de las variables estandarizadas. A diferencia de la matriz de varianza-covarianza, en la matriz de correlación cada componente principal captura una parte más homogénea de la variabilidad relativa de las observaciones. Esto implica que la matriz de correlación proporciona una perspectiva más equilibrada de las relaciones entre las observaciones y las variables cuando estas tienen diferentes escalas o varianzas.

3. Detecte posibles datos atípicos

```
# Paso 1: Crear un vector de puntuaciones para la primera componente
(PC1)
puntuaciones_PC1_S <- pca_S$x[, 1]  # Puntuaciones de PC1 para matriz de
varianza-covarianza
puntuaciones_PC2_S <- pca_S$x[, 2]  # Puntuaciones de PC2 para matriz de
varianza-covarianza

puntuaciones_PC1_R <- pca_R$x[, 1]  # Puntuaciones de PC1 para matriz de
correlación
puntuaciones_PC2_R <- pca_R$x[, 2]  # Puntuaciones de PC2 para matriz de
correlación

# Paso 2: Calcular el umbral para detectar outliers (e.g., 3 desviaciones
estándar)
umbral_S_PC1 <- mean(puntuaciones_PC1_S) + 3 * sd(puntuaciones_PC1_S)
umbral_S_PC2 <- mean(puntuaciones_PC2_S) + 3 * sd(puntuaciones_PC2_S)

umbral_R_PC1 <- mean(puntuaciones_PC1_R) + 3 * sd(puntuaciones_PC1_R)
umbral_R_PC2 <- mean(puntuaciones_PC2_R) + 3 * sd(puntuaciones_PC2_R)

# Paso 3: Detectar outliers en las puntuaciones de PC1 y PC2
outliers_S <- which(abs(puntuaciones_PC1_S) > umbral_S_PC1 |
abs(puntuaciones_PC2_S) > umbral_S_PC2)
outliers_R <- which(abs(puntuaciones_PC1_R) > umbral_R_PC1 |
```

```r
abs(puntuaciones_PC2_R) > umbral_R_PC2)

# Mostrar los outliers detectados
print("Outliers detectados en matriz de varianza-covarianza (S):")
```

```
## [1] "Outliers detectados en matriz de varianza-covarianza (S):"
```

```r
print(outliers_S)
```

```
## [1] 2
```

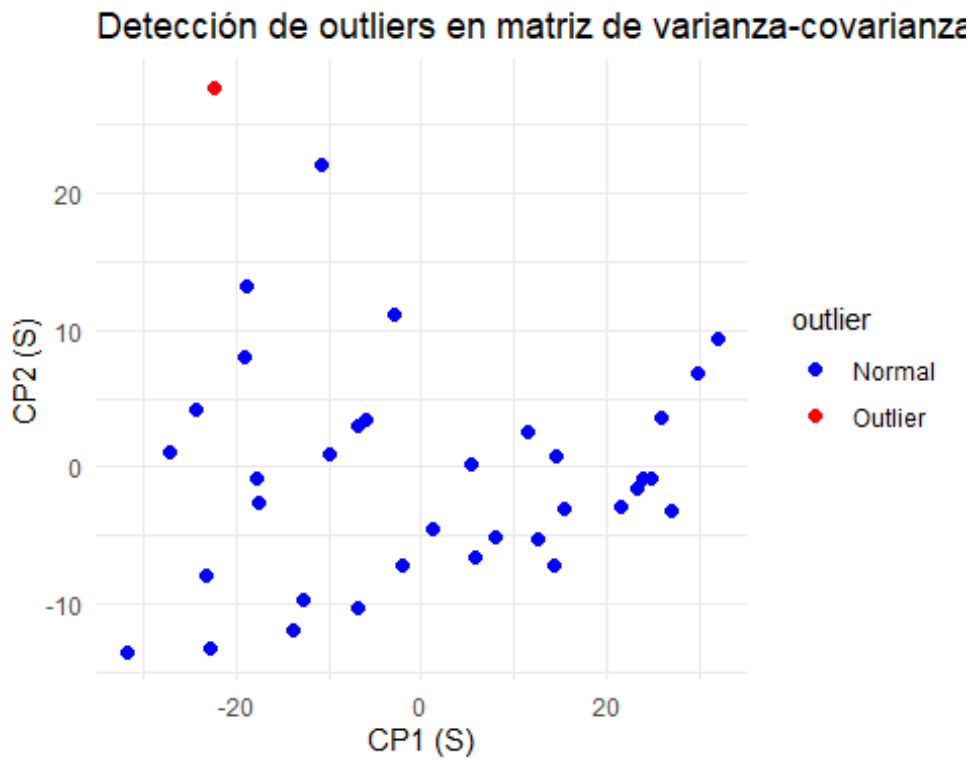
```r
print("Outliers detectados en matriz de correlación (R):")
```

```
## [1] "Outliers detectados en matriz de correlación (R):"
```

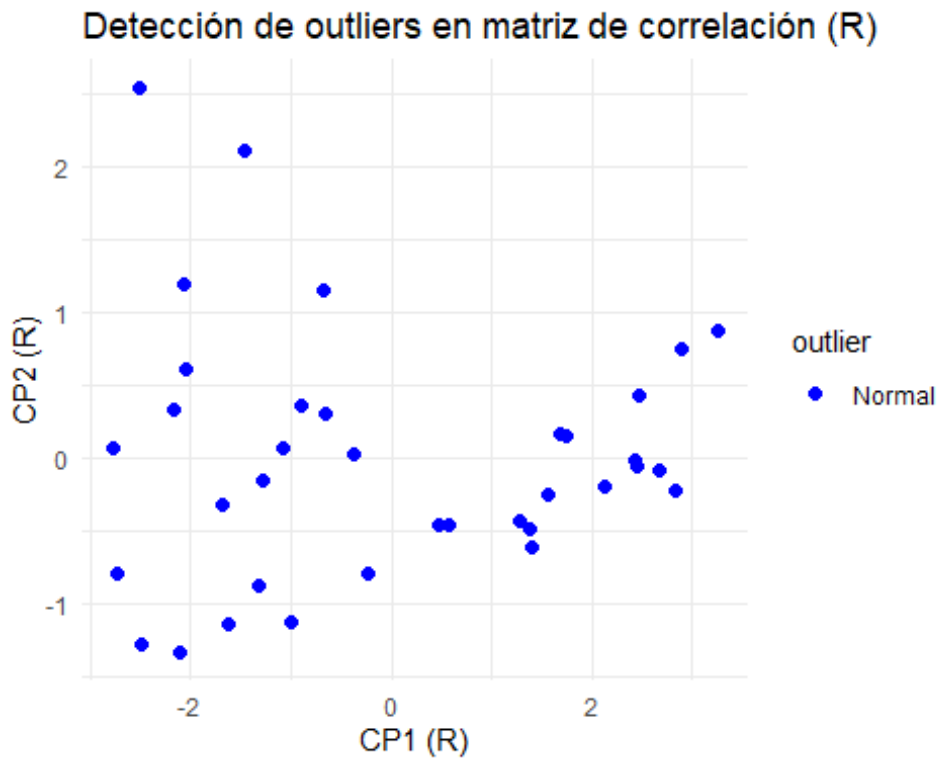
```r
print(outliers_R)
```

```
## integer(0)
```

```r
# Crear un dataframe con las puntuaciones y etiquetas de outliers para S
pca_S_data$outlier <- ifelse(1:nrow(pca_S_data) %in% outliers_S,
"Outlier", "Normal")

# Crear un dataframe con las puntuaciones y etiquetas de outliers para R
pca_R_data$outlier <- ifelse(1:nrow(pca_R_data) %in% outliers_R,
"Outlier", "Normal")

# Graficar los outliers en el gráfico de CP1 y CP2 para S
ggplot(pca_S_data, aes(x = PC1, y = PC2, color = outlier)) +
  geom_point(size = 2) +
  ggtitle("Detección de outliers en matriz de varianza-covarianza (S)") +
  theme_minimal() +
  xlab("CP1 (S)") +
  ylab("CP2 (S)") +
  scale_color_manual(values = c("blue", "red"))
```




```r
# Graficar los outliers en el gráfico de CP1 y CP2 para R
ggplot(pca_R_data, aes(x = PC1, y = PC2, color = outlier)) +
  geom_point(size = 2) +
  ggtitle("Detección de outliers en matriz de correlación (R)") +
  theme_minimal() +
  xlab("CP1 (R)") +
  ylab("CP2 (R)") +
  scale_color_manual(values = c("blue", "red"))
```



Análisis gráfico de los scores:

- Observar las puntuaciones en gráficos de dispersión de las dos primeras componentes principales (CP1 y CP2).
- Las observaciones que están alejadas del centro o que se diferencian notablemente del grupo principal podrían ser outliers.

Identificación numérica usando puntuaciones:

- Considerar un umbral para las puntuaciones: por ejemplo, observaciones cuya puntuación está a más de 3 desviaciones estándar de la media.
- También se puede aplicar el criterio de los percentiles: puntos en el 1% más bajo o alto en las puntuaciones podrían ser candidatos a outliers.

En este caso, los "outliers detectados en matriz de varianza-covarianza (S):" fue solamente el dato 2

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

```
# Paso 2: Calcular el PCA usando la matriz de covarianza
cpS <- princomp(datos, cor = FALSE)  # cor = FALSE usa la matriz de
varianza-covarianza
```

```r
# Paso 3: Mostrar el resumen del PCA
print("Resumen del PCA:")
```

```
## [1] "Resumen del PCA:"
```

```r
summary(cpS)
```

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

```r
# Paso 4: Mostrar las cargas (loadings) de los componentes principales
print("Cargas de los componentes principales (Loadings):")
```

```
## [1] "Cargas de los componentes principales (Loadings):"
```

```r
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
```

```r
# Paso 5: Calcular manualmente las puntuaciones usando las cargas
cpaS <- as.matrix(datos) %*% cpS$loadings

# Mostrar las primeras filas de las puntuaciones calculadas manualmente
print("Puntuaciones calculadas manualmente:")
```

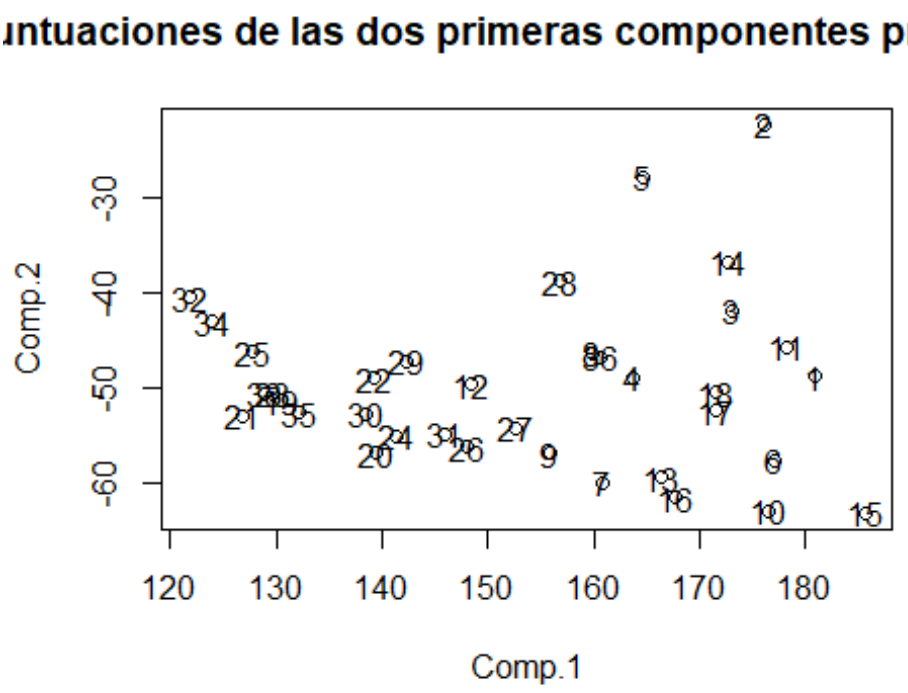
```
## [1] "Puntuaciones calculadas manualmente:"
```

```r
head(cpaS)
```

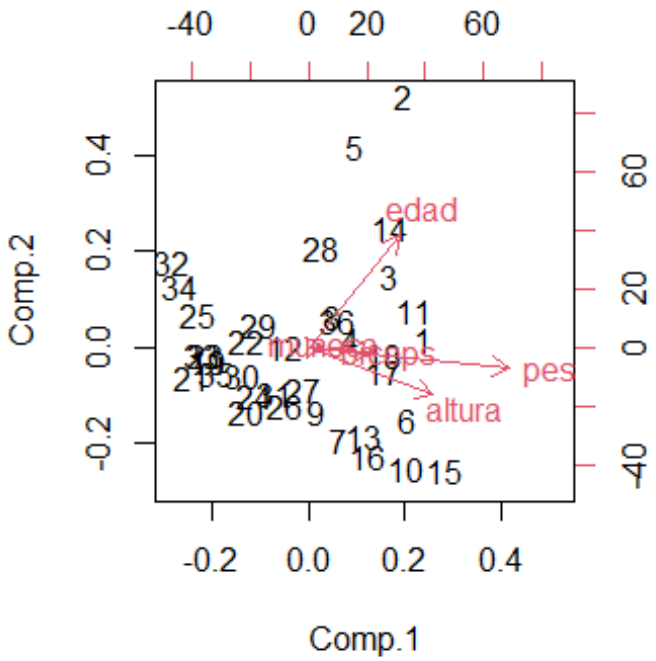
```
##        Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## [1,] 180.9723 -48.75142 104.41935 -13.93818 -4.405445
## [2,] 176.1730 -22.18369 102.48068 -14.95779 -3.863033
```

```
## [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
# Paso 6: Graficar las puntuaciones de las dos primeras componentes principales
plot(cpaS[, 1:2], type = "p", main = "Puntuaciones de las dos primeras componentes principales")
text(cpaS[, 1], cpaS[, 2], 1:nrow(cpaS))  # Etiquetar los puntos con el número de observación
```




```r
# Paso 7: Graficar el biplot para visualizar la relación entre variables y observaciones
biplot(cpS)
```



El comando princomp(datos, cor = FALSE) realiza el análisis de componentes principales usando la matriz de varianza-covarianza; si se establece cor = TRUE, usa la matriz de correlación. El resumen del PCA (summary(cpS)) muestra la proporción de varianza explicada por cada componente, mientras que las cargas (cpS$loadings) indican cómo se construye cada componente a partir de las variables originales. Las puntuaciones (cpS$scores) proyectan cada observación en el espacio de componentes principales, lo que permite analizar patrones y detectar outliers. La primera gráfica muestra las puntuaciones de las observaciones en las dos primeras componentes principales, y el biplot combina las puntuaciones y las cargas, facilitando la interpretación conjunta de observaciones y variables. Además, princomp() permite manejar datos con NA (na.action), realizar preprocesamiento de datos y obtener información detallada de la estructura de los componentes con print.summary.

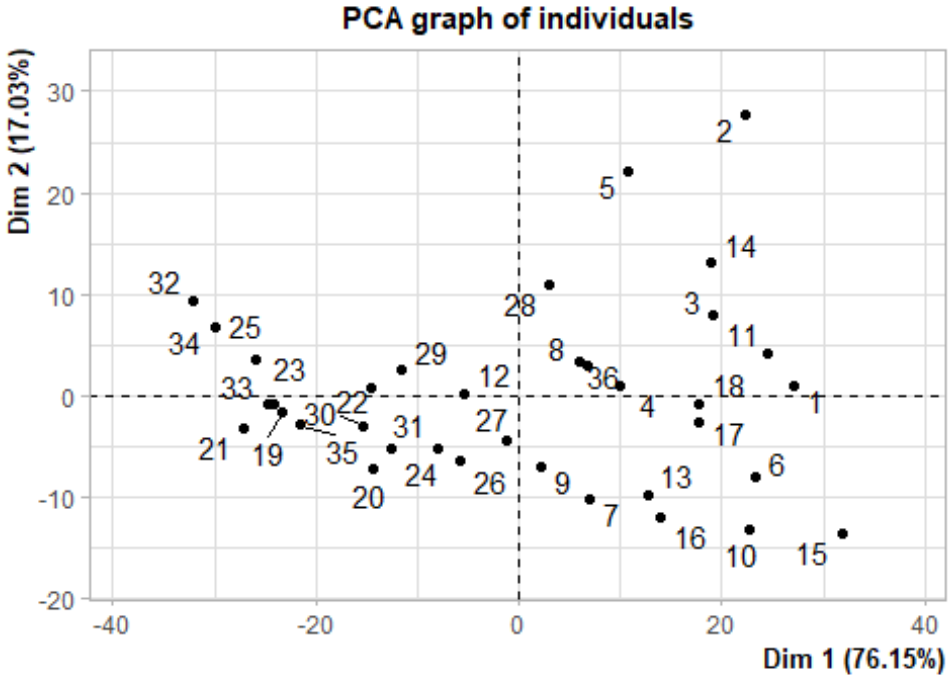
 # Parte 3

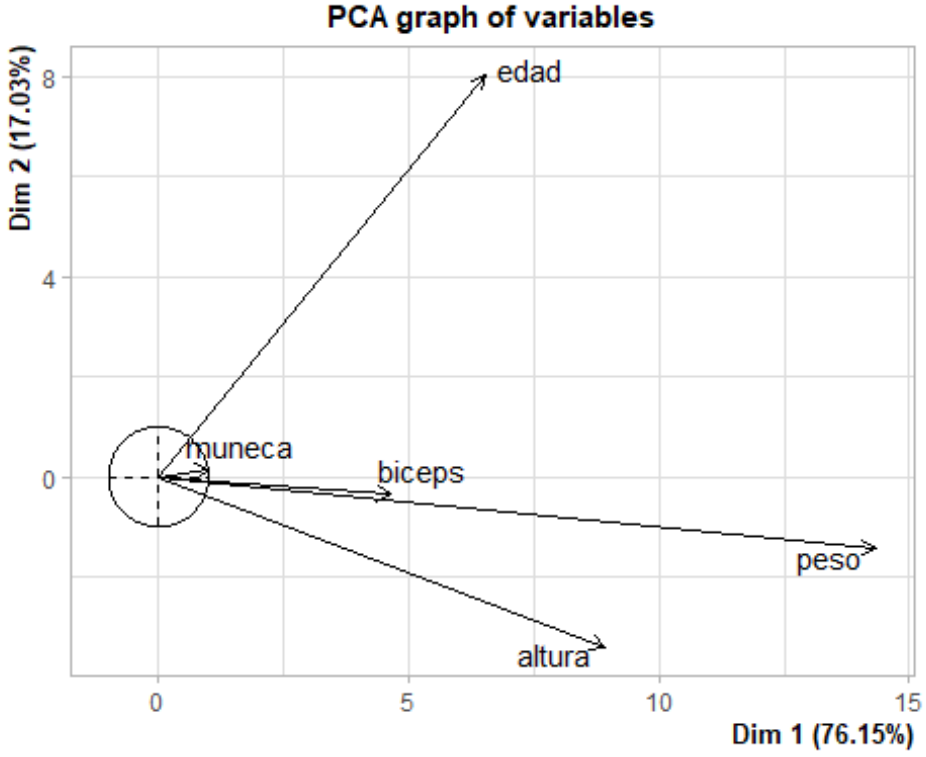
1. Explore los siguientes gráficos relativos a Componentes Principales.

Matriz de Varianza-covarianza

```
library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 4.3.2

library(ggplot2)
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.2
```

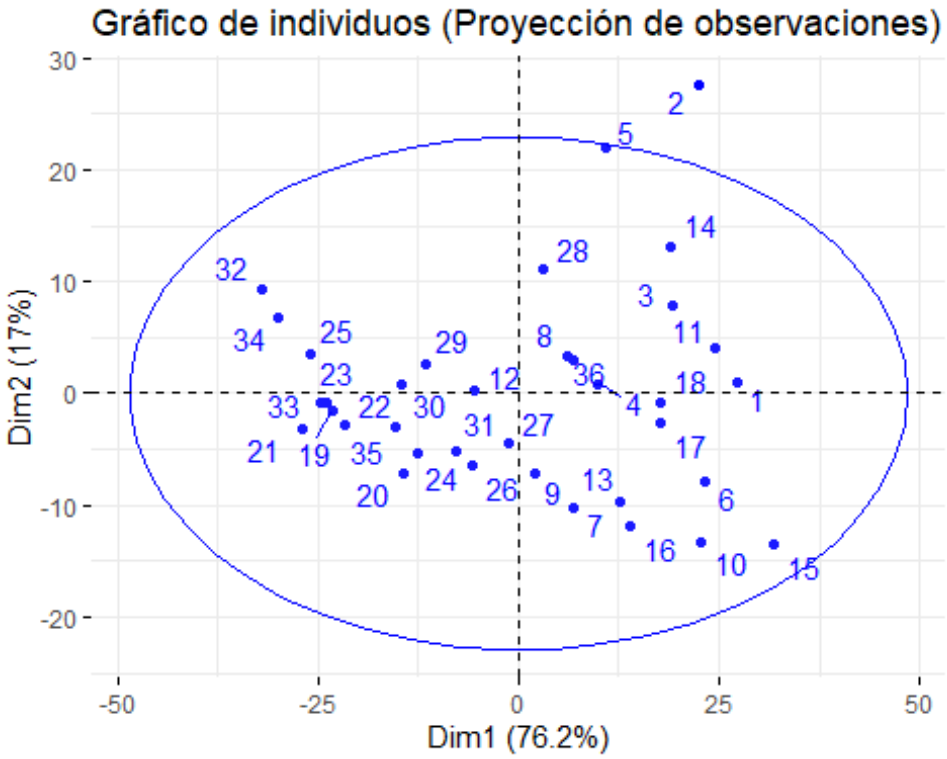
```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

# Paso 3: Realizar el PCA usando la matriz de varianza-covarianza
(scale.unit = FALSE)
cpS <- PCA(datos, scale.unit = FALSE)  # Para matriz de correlación, usar
scale.unit = TRUE
```



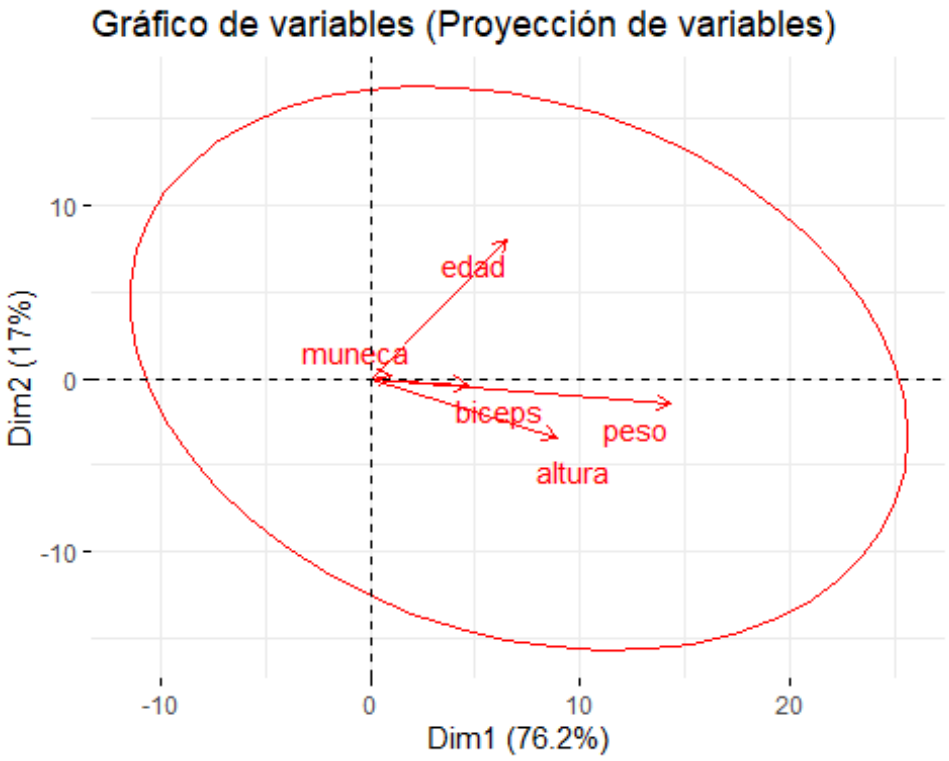



```
# Paso 4: Generar los gráficos correspondientes

# Gráfico de individuos con elipses
```

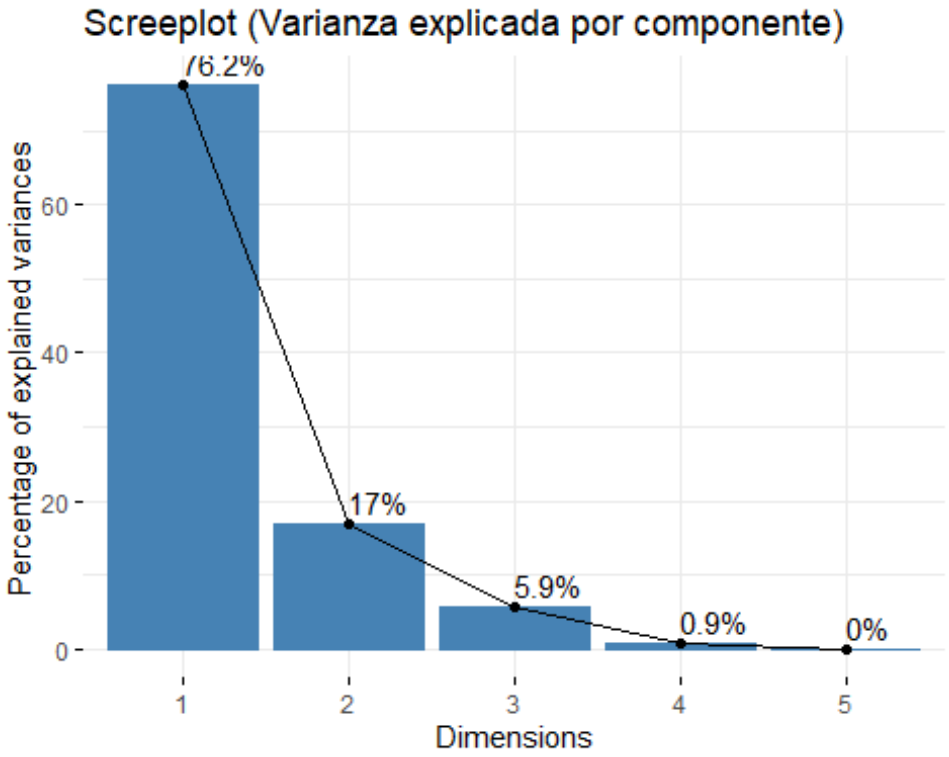
```r
fviz_pca_ind(cpS, col.ind = "blue", addEllipses = TRUE, repel = TRUE) +
  ggtitle("Gráfico de individuos (Proyección de observaciones)")
```



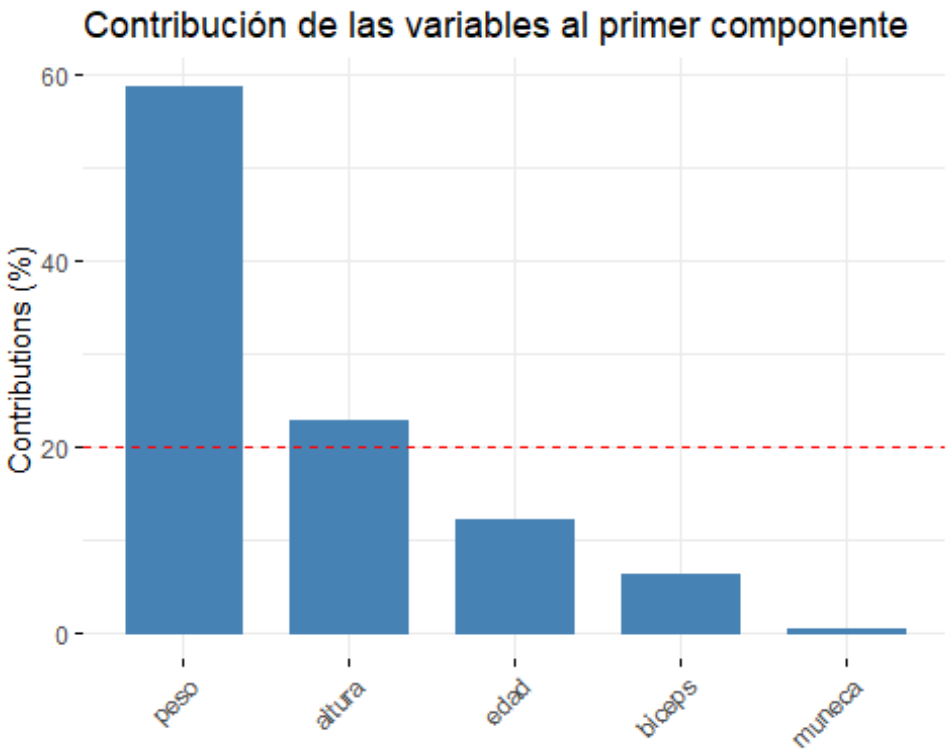

```r
# Gráfico de variables con elipses
fviz_pca_var(cpS, col.var = "red", addEllipses = TRUE, repel = TRUE) +
  ggtitle("Gráfico de variables (Proyección de variables)")
```



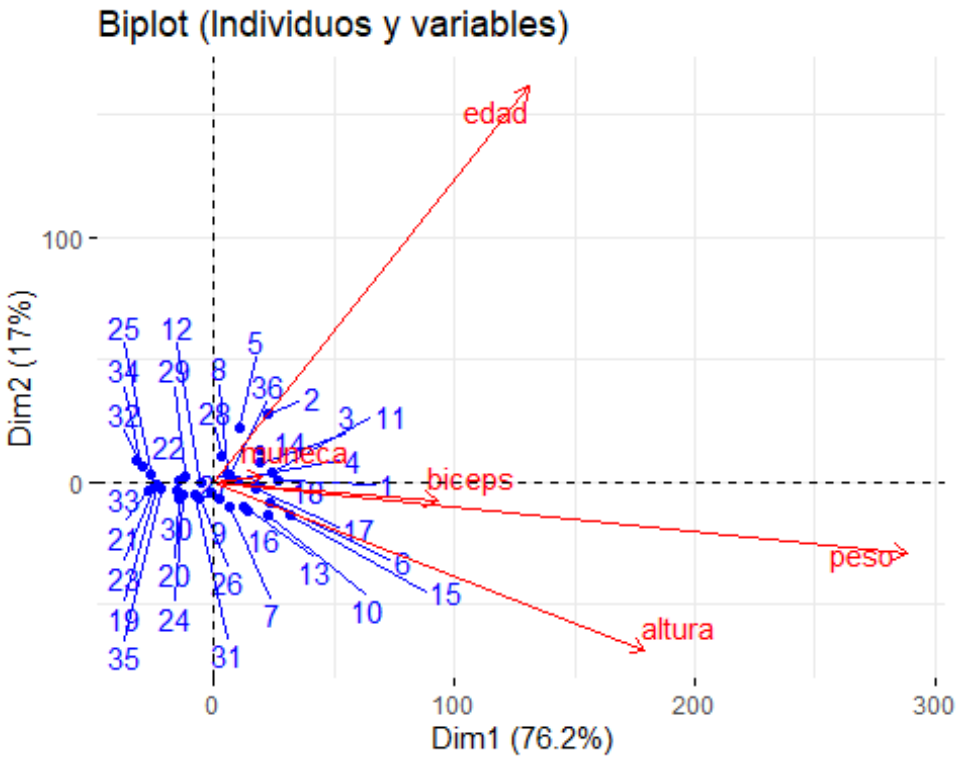

```r
# Gráfico de screeplot
fviz_screeplot(cpS, addlabels = TRUE, barfill = "steelblue") +
  ggtitle("Screeplot (Varianza explicada por componente)")
```



```r
# Gráfico de contribución de variables al primer componente
fviz_contrib(cpS, choice = "var", axes = 1) +
  ggtitle("Contribución de las variables al primer componente")
```

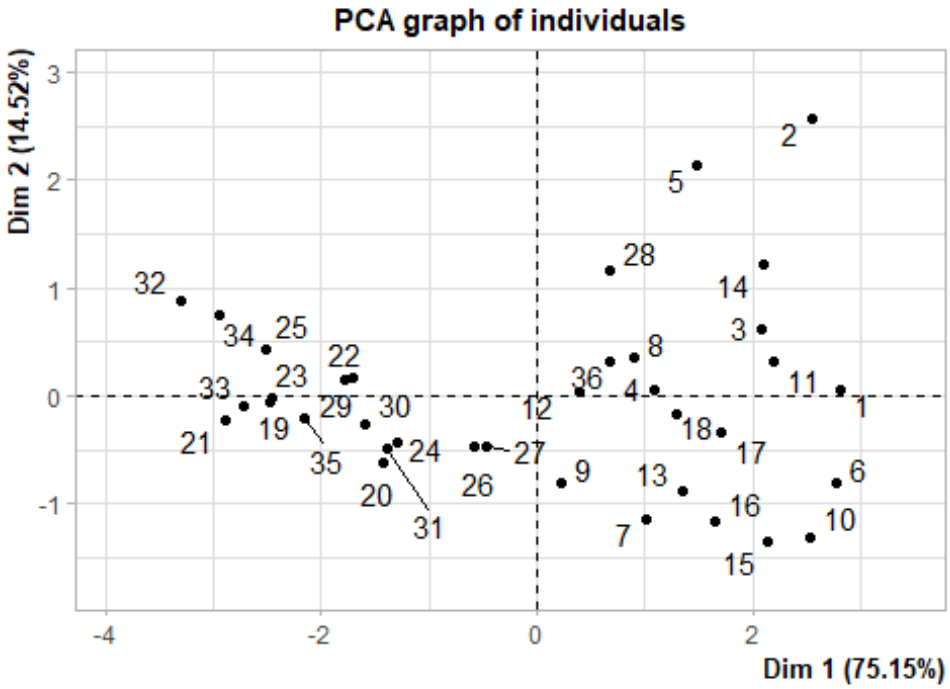


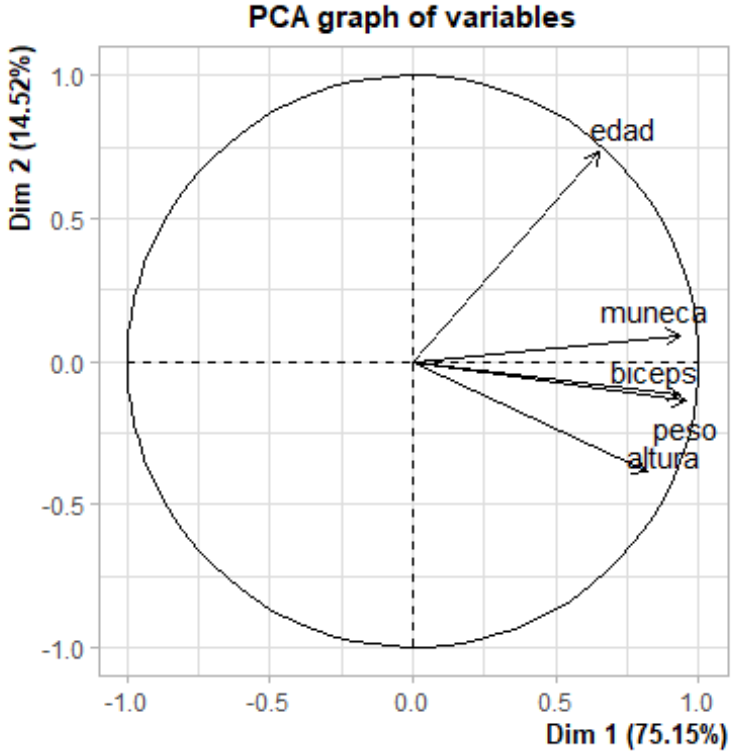

```r
# Gráfico biplot combinando observaciones y variables
fviz_pca_biplot(cpS, repel = TRUE, col.var = "red", col.ind = "blue") +
  ggtitle("Biplot (Individuos y variables)")
```



Matriz de Correlación

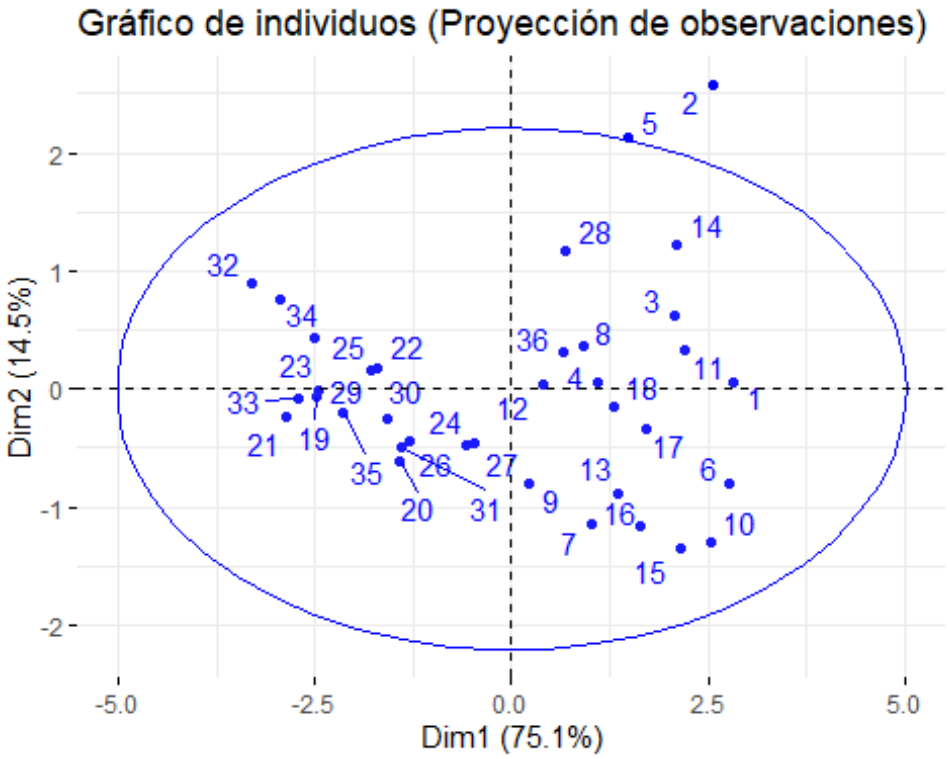
```
# Paso 3: Realizar el PCA usando la matriz de varianza-covarianza (scale.unit = FALSE)
cpS <- PCA(datos, scale.unit = TRUE)  # Para matriz de correlación, usar scale.unit = TRUE
```



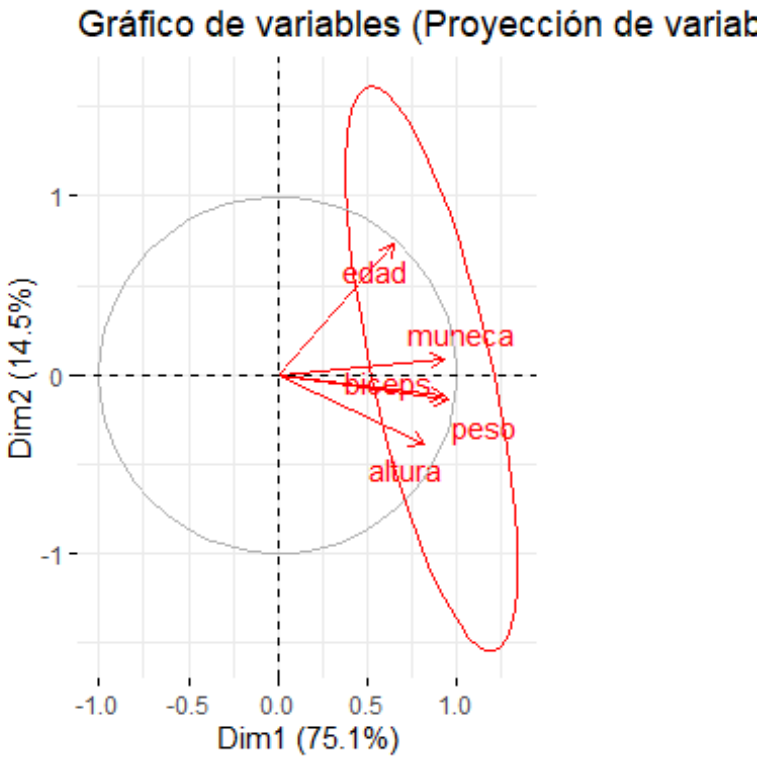



```
# Paso 4: Generar los gráficos correspondientes

# Gráfico de individuos con elipses
```

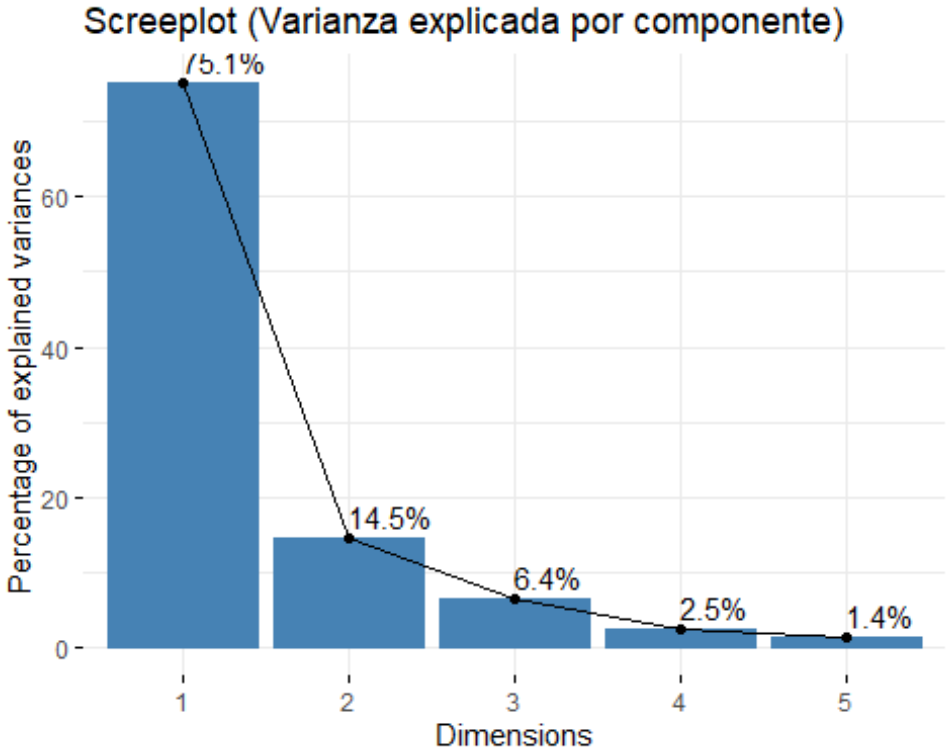
```r
fviz_pca_ind(cpS, col.ind = "blue", addEllipses = TRUE, repel = TRUE) +
  ggtitle("Gráfico de individuos (Proyección de observaciones)")
```



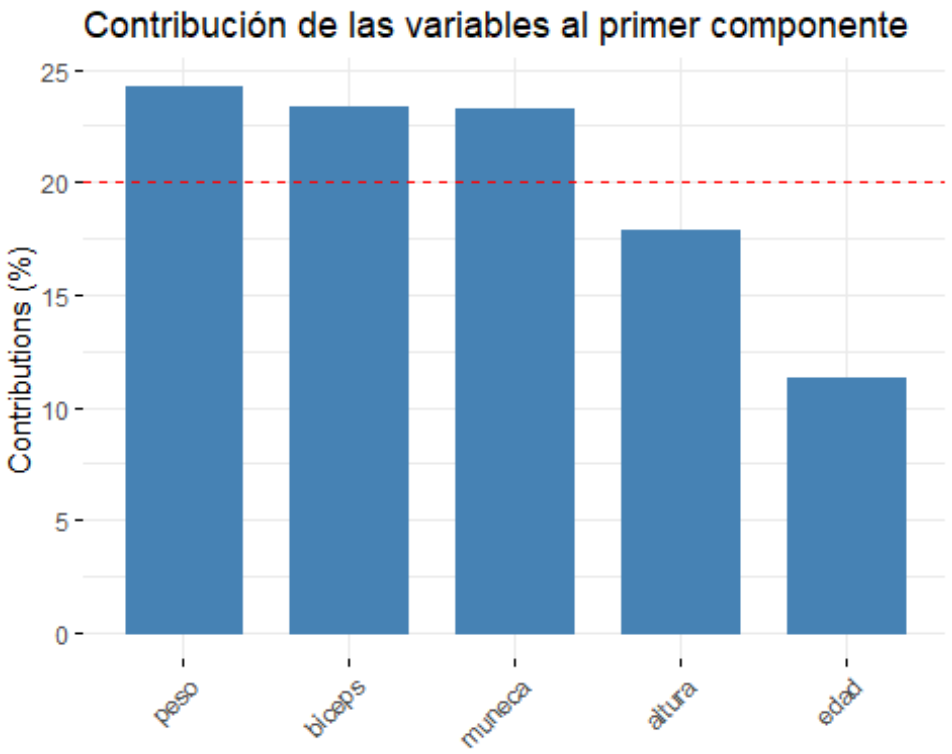

```r
# Gráfico de variables con elipses
fviz_pca_var(cpS, col.var = "red", addEllipses = TRUE, repel = TRUE) +
  ggtitle("Gráfico de variables (Proyección de variables)")
```



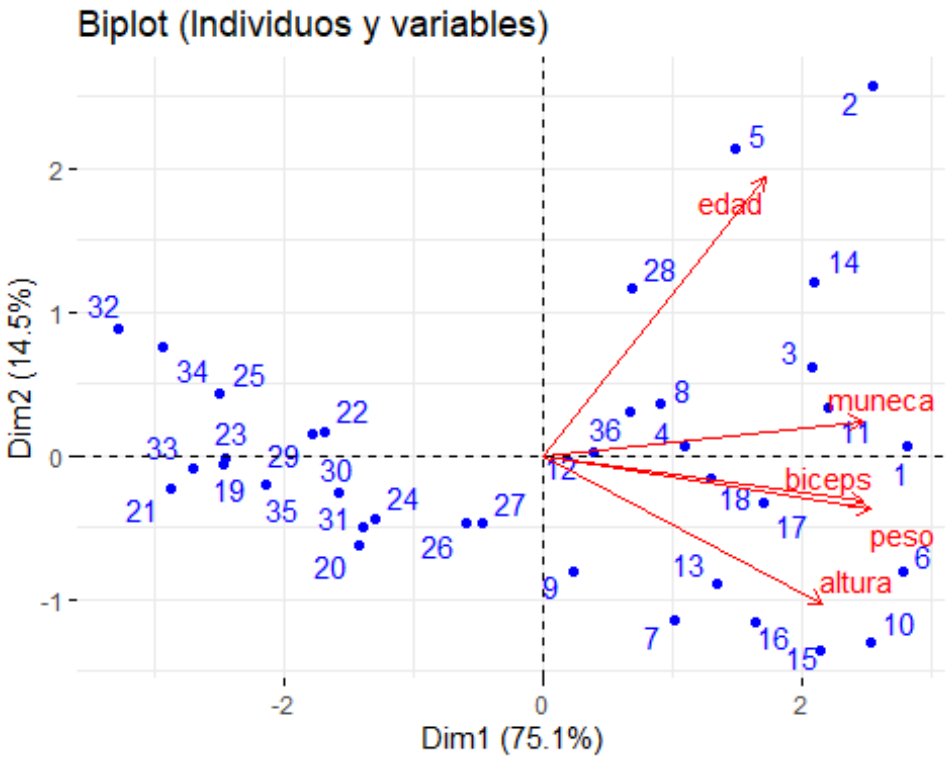

```r
# Gráfico de screeplot
fviz_screeplot(cpS, addlabels = TRUE, barfill = "steelblue") +
  ggtitle("Screeplot (Varianza explicada por componente)")
```



```r
# Gráfico de contribución de variables al primer componente
fviz_contrib(cpS, choice = "var", axes = 1) +
  ggtitle("Contribución de las variables al primer componente")
```




```r
# Gráfico biplot combinando observaciones y variables
fviz_pca_biplot(cpS, repel = TRUE, col.var = "red", col.ind = "blue") +
  ggtitle("Biplot (Individuos y variables)")
```



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.

- El gráfico de individuos muestra cómo se agrupan las observaciones y permite identificar posibles outliers.
- El gráfico de variables ayuda a entender qué variables influyen más en cada componente y cómo se relacionan entre sí.
- El screeplot facilita la selección del número de componentes a retener, indicando la proporción de varianza explicada.
- El gráfico de contribución resalta qué variables son las más importantes en cada componente.
- El biplot ofrece una visión integrada de observaciones y variables, permitiendo ver cómo se relacionan entre sí y cuál es la influencia de cada variable en la estructura de los datos.

El análisis de componentes principales con la matriz de varianzas-covarianzas es apropiado cuando las variables tienen la misma escala y la magnitud de sus varianzas es relevante para la interpretación. En este contexto, el screeplot indica la proporción de varianza explicada por cada componente; si el primer componente principal (CP1) explica una proporción alta (por ejemplo, >70%), entonces la mayor parte de la variabilidad se concentra en ese componente. En el biplot, las variables con flechas más largas tienen mayor varianza, lo que significa que dominan la estructura del

componente. Las observaciones alejadas del centro o alineadas con estas flechas reflejan características dominadas por esas variables. En el gráfico de contribución, las variables con varianza elevada se destacan, lo cual puede sesgar la interpretación hacia estas variables específicas.

Por otro lado, el análisis con la matriz de correlación estandariza las variables a media 0 y varianza 1, garantizando que todas tengan la misma importancia en el análisis, sin importar sus unidades o magnitudes originales. En este caso, el screeplot suele mostrar una distribución más equilibrada de la varianza explicada entre los componentes. En el biplot, las flechas de las variables tienen longitudes más uniformes, lo que indica que contribuyen de manera similar a la estructura de los componentes. Las observaciones que se alinean con ciertas flechas pueden interpretarse como outliers en términos de combinaciones lineales de las variables estandarizadas. El gráfico de contribución permite identificar claramente qué variables son más importantes para cada componente, proporcionando una interpretación más balanceada de la relación entre las variables y los componentes principales.

 # 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? ¿Cuál de los dos procedimientos aporta componentes con de mayor interés?

El análisis de componentes principales utilizando la matriz de varianza-covarianza tiende a estar dominado por las variables con varianzas más grandes, ya que estas influyen más en la construcción de los componentes. Esto puede llevar a interpretaciones sesgadas si las variables tienen magnitudes muy distintas, haciendo que los componentes principales capturen la mayor parte de la variabilidad de solo algunas variables. Por otro lado, con la matriz de correlación, las variables se estandarizan a media 0 y varianza 1, lo que asegura que todas contribuyan de manera equilibrada al análisis. Como resultado, el impacto de cada variable es uniforme, y los componentes reflejan mejor las relaciones estructurales entre las variables sin sesgos de magnitud.

En la mayoría de los casos, la matriz de correlación suele ser preferida, ya que facilita la comparación entre variables y componentes, proporcionando resultados más interpretables y relevantes para identificar patrones generales.

2. Indique cuál de los dos análisis (a partir de la matriz de varianza y covarianza o de correlación) resulta mejor para los datos indicadores económicos y sociales del 96 países en el mundo. Comparar los resultados y argumentar cuál es mejor según los resultados obtenidos.

Para los datos económicos y sociales de 96 países, el uso de la matriz de correlación es generalmente la opción más adecuada. Esto se debe a que los indicadores económicos

y sociales (como PIB, tasa de alfabetización, esperanza de vida, etc.) suelen tener diferentes escalas y unidades. Algunos indicadores pueden tener varianzas mucho mayores que otros, lo que podría sesgar el análisis si se utiliza la matriz de varianza-covarianza, haciendo que solo algunos indicadores dominen los componentes principales.

Al comparar los resultados, la matriz de varianza-covarianza probablemente muestra que las variables con mayor varianza (como el PIB o la población) explican la mayor parte de la variabilidad total, mientras que otras variables de menor varianza (como tasas de desempleo o mortalidad) tienen un impacto reducido. Por el contrario, la matriz de correlación estandariza todas las variables, permitiendo que cada una contribuya de manera equitativa en la construcción de los componentes. Esto lleva a una interpretación más equilibrada y significativa de las relaciones entre los indicadores.

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)

```
# Extraer los vectores propios (autovectores) de S
autovectores <- eigen_R$vectors

# Combinación lineal de las variables para CP1 y CP2
CP1 <- paste0("CP1 = ", round(autovectores[1, 1], 4), " * X1 + ",
              round(autovectores[2, 1], 4), " * X2 + ",
              round(autovectores[3, 1], 4), " * X3 + ",
              round(autovectores[4, 1], 4), " * X4 + ",
              round(autovectores[5, 1], 4), " * X5")

CP2 <- paste0("CP2 = ", round(autovectores[1, 2], 4), " * X1 + ",
              round(autovectores[2, 2], 4), " * X2 + ",
              round(autovectores[3, 2], 4), " * X3 + ",
              round(autovectores[4, 2], 4), " * X4 + ",
              round(autovectores[5, 2], 4), " * X5")

# Imprimir las ecuaciones de CP1 y CP2
print(CP1)
```

```
## [1] "CP1 = -0.3359 * X1 + -0.4927 * X2 + -0.4222 * X3 + -0.4822 * X4 +
-0.4833 * X5"
```

```
print(CP2)
```

```
## [1] "CP2 = 0.8576 * X1 + -0.1648 * X2 + -0.4542 * X3 + 0.1083 * X4 + -
0.1393 * X5"
```

```
contribucion_CP1 <- abs(autovectores[, 1])
contribucion_CP2 <- abs(autovectores[, 2])
```

```r
# Mostrar las contribuciones de las variables a CP1 y CP2
print("Contribuciones a CP1 (en valor absoluto):")
```

```
## [1] "Contribuciones a CP1 (en valor absoluto):"
```

```r
print(contribucion_CP1)
```

```
## [1] 0.3359310 0.4927066 0.4222426 0.4821923 0.4833139
```

```r
print("Contribuciones a CP2 (en valor absoluto):")
```

```
## [1] "Contribuciones a CP2 (en valor absoluto):"
```

```r
print(contribucion_CP2)
```

```
## [1] 0.8575601 0.1647821 0.4542223 0.1082775 0.1392684
```

El método seleccionado fue la matriz de correlación

La variable que más influye en el primer componente PC1 es X2 (peso) con un coeficiente de -0.4927 siguiendole de cerca X4 (muneca) y X5 (biceps) con -0.4822 y -0.4833 respectivamente. Mientras que la variable que más influye en el segundo componente PC2 es X1 (edad) dado que su coeficiente es el mayor de todo PC2 con 0.8576.

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

[1] "CP1 = -0.3359 * X1 + -0.4927 * X2 + -0.4222 * X3 + -0.4822 * X4 + -0.4833 * X5"
[1] "CP2 = 0.8576 * X1 + -0.1648 * X2 + -0.4542 * X3 + 0.1083 * X4 + -0.1393 * X5"

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

Ejemplo de agrupación:

Primera componente principal (CP1): Si las variables que más contribuyen a CP1 incluyen indicadores como PIB per cápita, esperanza de vida, y nivel de alfabetización, podemos interpretar esta componente como un "índice de riqueza o desarrollo socioeconómico". Un valor alto en esta componente indicaría que los países tienen un nivel alto de desarrollo económico y social, mientras que valores bajos reflejan países con menores niveles de desarrollo.

Segunda componente principal (CP2): Si CP2 está dominado por variables como porcentaje de población rural, tasa de mortalidad infantil, o tasa de desempleo, esta componente podría ser interpretada como un "índice de ruralidad o pobreza estructural". Un valor alto en esta componente reflejaría un entorno rural con altos niveles de pobreza y menos acceso a recursos, mientras que valores bajos indicarían países con características más urbanas y una estructura socioeconómica más estable.

Agrupaciones adicionales:

Dependiendo de las variables incluidas en el análisis, podríamos identificar otras dimensiones como un “índice de calidad de vida”, que agrupe variables como el acceso a servicios básicos, educación y salud, o un “índice de estabilidad económica” que esté relacionado con la inflación, desempleo y producción.

Interpretación de agrupaciones:

Cada componente principal se interpreta como una dimensión que resume el comportamiento de varias variables relacionadas. Estas agrupaciones reflejan diferentes características de los países y ayudan a comprender cómo se relacionan los indicadores económicos y sociales. Por ejemplo, un país que tiene un valor alto en el índice de desarrollo (CP1) pero bajo en el índice de ruralidad (CP2) puede ser interpretado como un país en desarrollo con buenas condiciones de vida, pero con una estructura social aún dependiente de la ruralidad.