A4_A01571214_Lautaro_Coteja

A01571214 - Lautaro Coteja

2024-10-10

R Markdown

A4 - Componentes Principales

Parte I

```
# Cargar Datos
corporal = read.csv("C:/Users/lauta/Downloads/corporal.csv")

head(corporal)
```

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

```
# Eliminar la columna sexo
corporal_num = corporal[ , !(names(corporal) %in% "sexo")]

# Verificamos las nuevas variables
head(corporal_num)
```

```
##   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
```

```
# Matriz de covarianzas
S = cov(corporal_num)
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
```

```
# Matriz de correlaciones
R = cor(corporal_num)
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
```

```
# Eigen descomposicion para la matriz de covarianzas
eigen_cov = eigen(S)
eigen_cov$values
```

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

```
eigen_cov$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
```

```
# Eigen descomposicion para la matriz de correlaciones
eigen_corr = eigen(R)
eigen_corr$values
```

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

```
eigen_corr$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 - Matriz de covarianzas
var_explained_cov = eigen_cov$values / sum(eigen_cov$values)
var_explained_cov
```

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

```r
# Varianza acumulada - Matriz de covarianzas
cum_var_explained_cov = cumsum(var_explained_cov)
cum_var_explained_cov
```

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

```r
# Proporción de varianza explicada - Matriz de correlaciones
var_explained_corr = eigen_corr$values / sum(eigen_corr$values)
var_explained_corr
```

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

```r
# Varianza acumulada - Matriz de correlaciones
cum_var_explained_corr = cumsum(var_explained_corr)
cum_var_explained_corr
```

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

```r
# Componentes Principales (CP1 y CP2) usando la matriz de covarianzas
CP1_cov = eigen_cov$vectors[, 1]  # Primer vector propio
CP2_cov = eigen_cov$vectors[, 2]  # Segundo vector propio

# Componentes Principales (CP1 y CP2) usando la matriz de correlaciones
CP1_corr = eigen_corr$vectors[, 1]  # Primer vector propio
CP2_corr = eigen_corr$vectors[, 2]  # Segundo vector propio

# Mostrar las combinaciones lineales de CP1 y CP2 (Covarianzas)
CP1_cov
```

```
## [1] -0.34871002 -0.76617586 -0.47632405 -0.05386189 -0.24817367
```

```r
CP2_cov
```

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

```r
# Mostrar las combinaciones lineales de CP1 y CP2 (Correlaciones)
CP1_corr
```

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

```r
CP2_corr
```

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

¿Qué componentes son los más importantes?

Los componentes más importantes son aquellos que explican la mayor proporción de la varianza. En ambos casos (matriz de covarianzas y matriz de correlaciones):

El primer componente principal (CP1) es el más importante, ya que explica aproximadamente el 75% de la varianza en ambos casos. El segundo componente principal

(CP2) añade alrededor de 15% a la explicación de la varianza. Estos dos componentes juntos explican aproximadamente el 90% de la variabilidad total de los datos, lo que indica que son suficientes para describir gran parte de la estructura de las variables originales.

Ecuación de la combinación lineal para CP1 y CP2

La combinación lineal para los componentes principales se obtiene multiplicando los valores de los vectores propios por las variables originales. Para la matriz de covarianzas y correlaciones, las fórmulas serían: Para la matriz de covarianzas: CP1=0.35·edad+0.77·peso+0.48·altura+0.05·muneca+0.25·biceps CP2=0.91·edad−0.16·peso−0.39·altura+0.01·muneca−0.04·biceps Para la matriz de correlaciones: CP1=−0.34·edad−0.49·peso−0.42·altura−0.48·muneca−0.48·biceps CP2=−0.86·edad+0.16·peso+0.45·altura−0.11·muneca+0.14·biceps

En ambos casos, las variables que más contribuyen a la primera componente principal (CP1) son peso y altura, mientras que en CP2 la variable más significativa es edad.

Estas combinaciones lineales ayudan a interpretar cómo las variables originales se combinan para formar los componentes principales. El objetivo es maximizar la varianza explicada con el menor número de componentes.

Conclusion

Componentes principales más importantes: El primer componente (CP1) explica aproximadamente el 75% de la varianza, tanto con la matriz de covarianzas como con la de correlaciones. Combinaciones lineales: La variable peso contribuye significativamente a CP1 en la matriz de covarianzas, mientras que en la matriz de correlaciones, todas las variables tienen contribuciones similares.

Parte II

```
library(ggplot2)
library(dplyr)
```

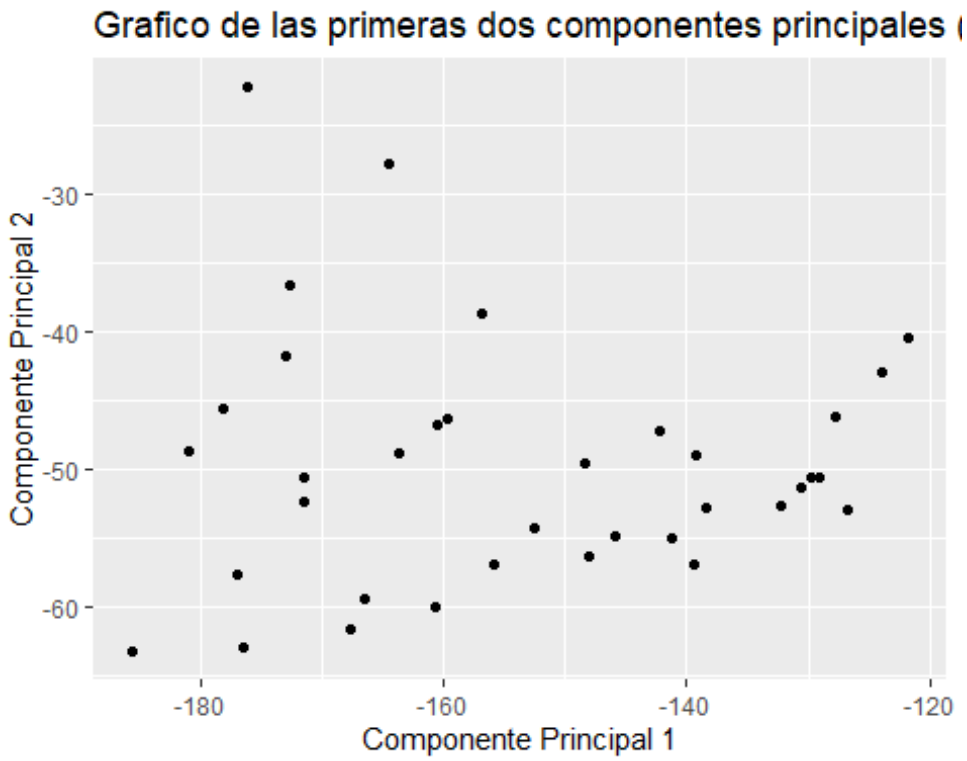
```
##
## Adjuntando el paquete: 'dplyr'
```

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

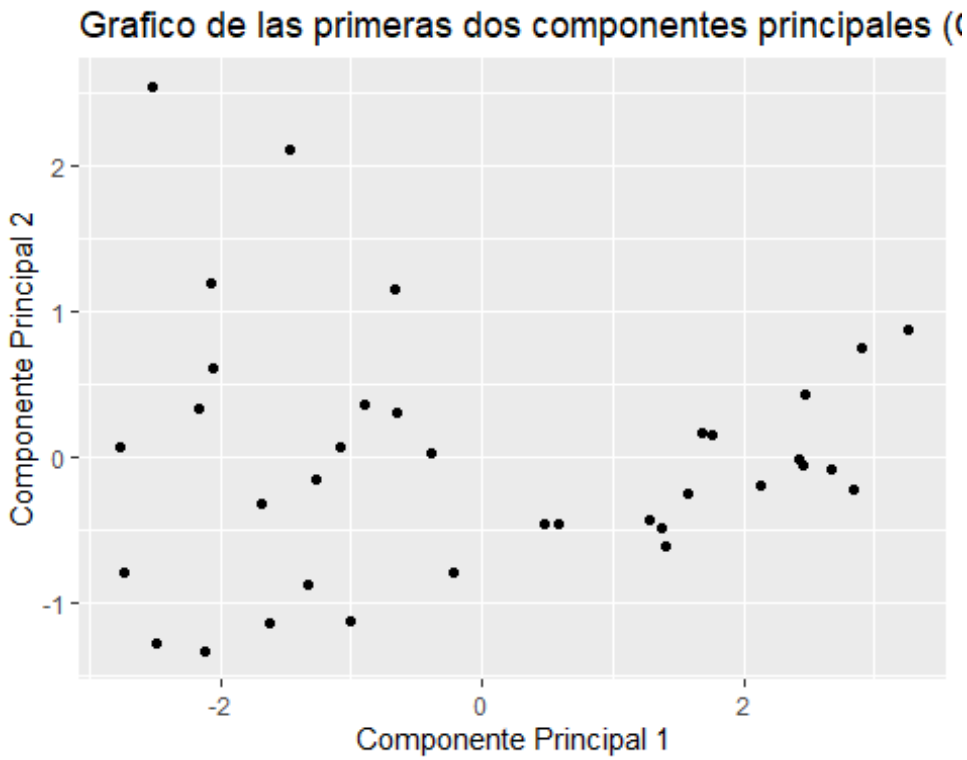
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Matriz de varianzas-covarianzas
S = cov(corporal_num)
```

```r
# Matriz de correlaciones (con variables estandarizadas)
R = cor(scale(corporal_num))
# Eigen descomposicion para la matriz de covarianzas
eigen_cov = eigen(S)

# Eigen descomposicion para la matriz de correlaciones
eigen_corr = eigen(R)
# Scores con la matriz de covarianzas
scores_cov = as.matrix(corporal_num) %*% eigen_cov$vectors

# Scores con la matriz de correlaciones (con variables estandarizadas)
corporal_scaled = scale(corporal_num)
scores_corr = as.matrix(corporal_scaled) %*% eigen_corr$vectors

# Convertir los scores a un dataframe
scores_cov_df = as.data.frame(scores_cov)
colnames(scores_cov_df) = paste0("PC", 1:ncol(scores_cov_df))

# Gráfico PC1 vs PC2
ggplot(scores_cov_df, aes(x = PC1, y = PC2)) +
  geom_point() +
  ggtitle("Grafico de las primeras dos componentes principales
(Covarianzas)") +
  xlab("Componente Principal 1") +
  ylab("Componente Principal 2")
```



```r
# Convertir los scores a un dataframe
scores_corr_df = as.data.frame(scores_corr)
colnames(scores_corr_df) = paste0("PC", 1:ncol(scores_corr_df))

# Grafico PC1 vs PC2
ggplot(scores_corr_df, aes(x = PC1, y = PC2)) +
  geom_point() +
  ggtitle("Grafico de las primeras dos componentes principales
(Correlaciones)") +
  xlab("Componente Principal 1") +
  ylab("Componente Principal 2")
```



Interpretación de los resultados Relaciones entre las variables y los componentes principales: En ambos casos, el primer componente principal (CP1) explica la mayor parte de la varianza. En la matriz de varianzas-covarianzas, las variables que más contribuyen a CP1 son peso y altura. En la matriz de correlaciones, las contribuciones de las variables son más equilibradas, pero peso sigue siendo una de las variables más influyentes.

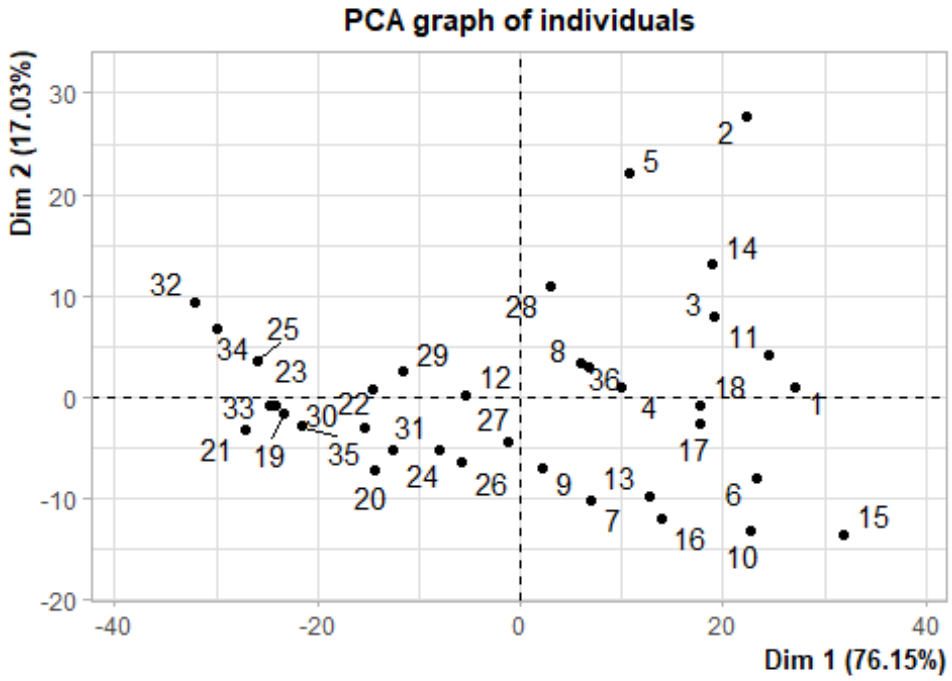
Relación entre las puntuaciones de las observaciones y los valores de las variables: Los gráficos de PC1 vs PC2 nos permiten ver la distribución de las observaciones en relación con los dos primeros componentes. Las observaciones que están más alejadas del centro en estas gráficas suelen tener valores extremos en las variables que más contribuyen a estos componentes.

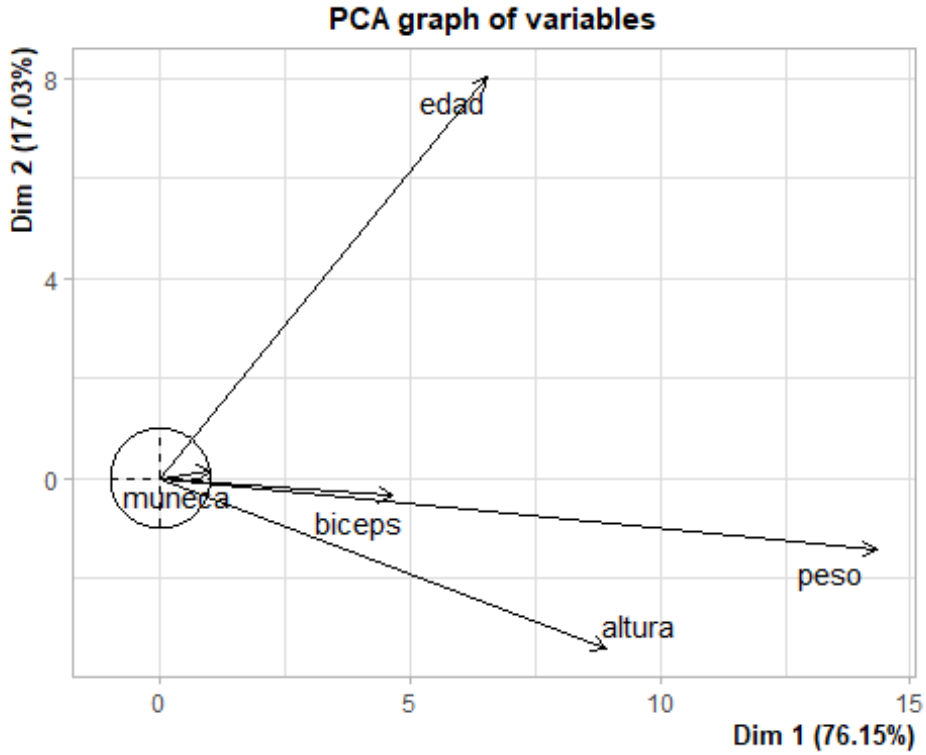
Detección de datos atípicos: Observaciones fuera del grupo principal en los gráficos de PC1 vs PC2 podrían ser consideradas datos atípicos. Estos puntos merecen una investigación

adicional para determinar si representan errores o casos especiales en el conjunto de datos.

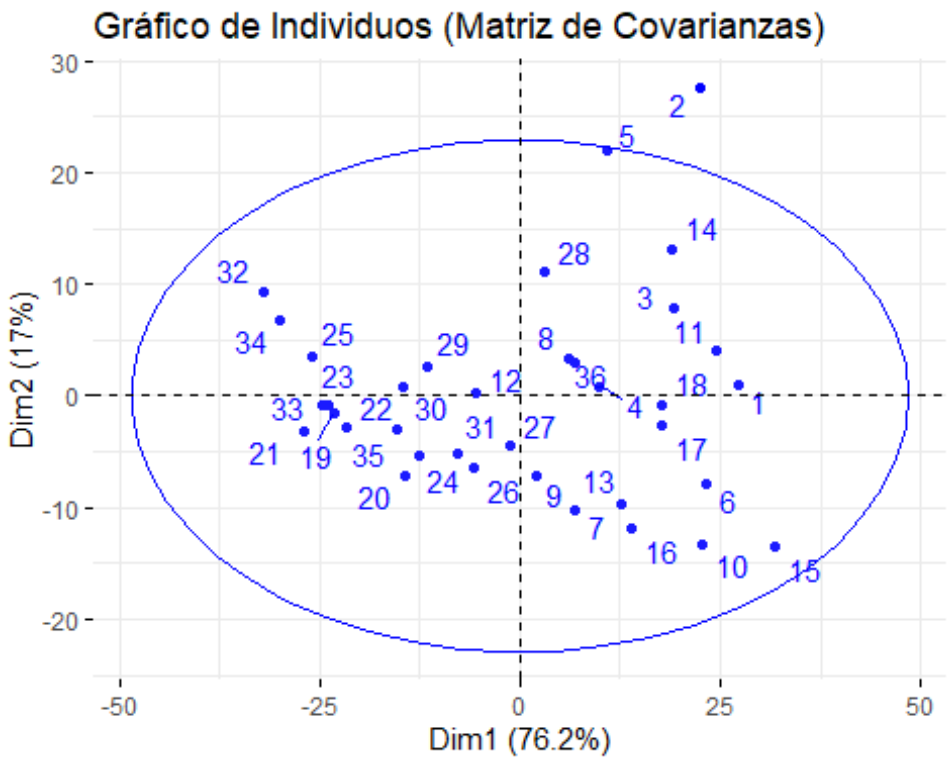
Parte III

```
library(FactoMineR)
library(factoextra)

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

library(ggplot2)

# PCA usando la matriz de covarianzas (scale.unit=FALSE)
cpS_cov = PCA(corporal_num, scale.unit = FALSE)
```

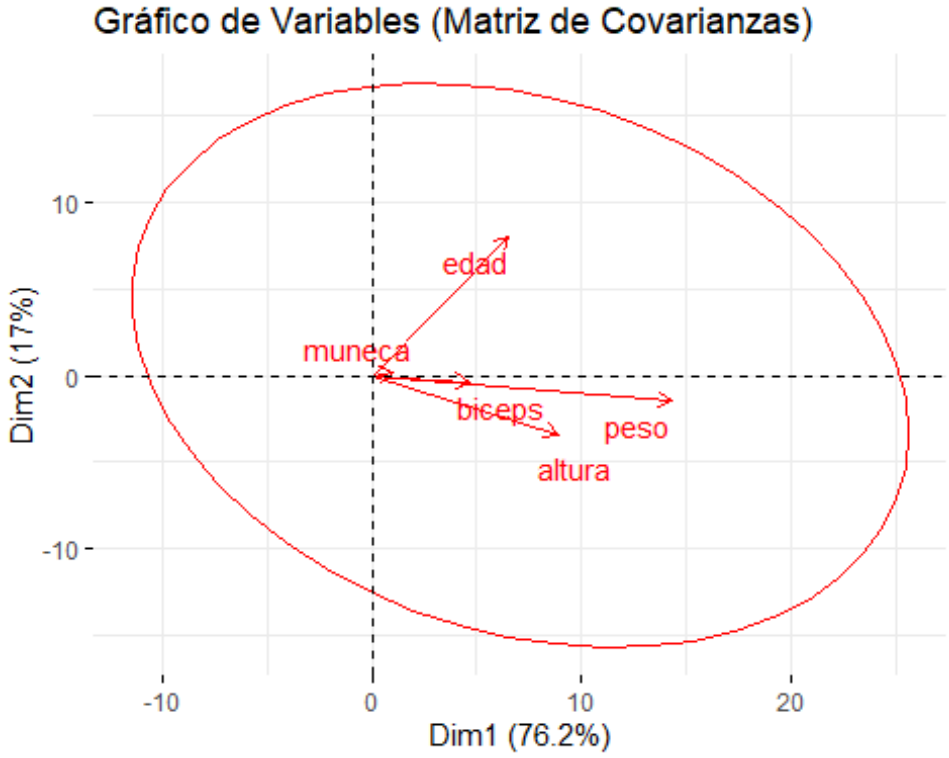




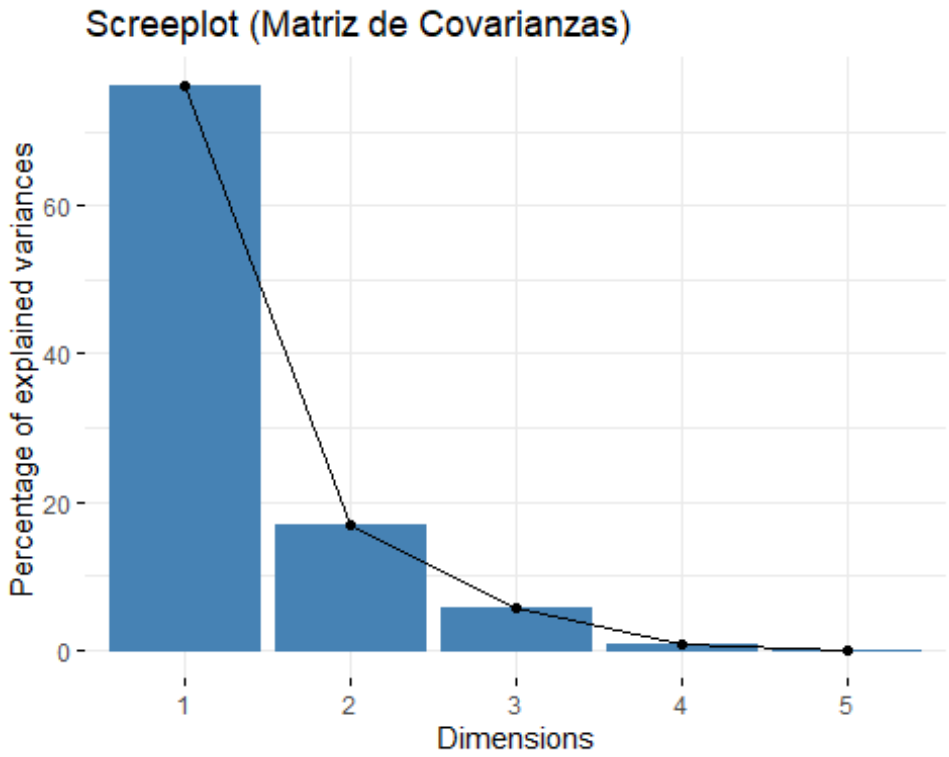

```r
# Gráfico de individuos
fviz_pca_ind(cpS_cov, col.ind = "blue", addEllipses = TRUE, repel = TRUE) +
  ggtitle("Gráfico de Individuos (Matriz de Covarianzas)")
```



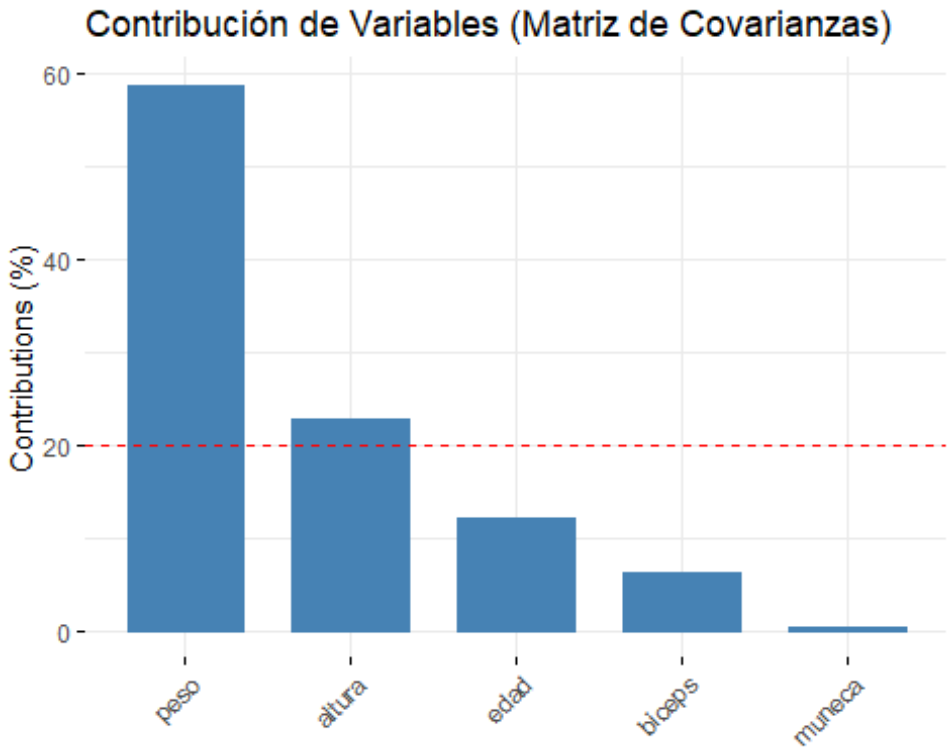

```r
# Gráfico de variables
fviz_pca_var(cpS_cov, col.var = "red", addEllipses = TRUE, repel = TRUE) +
  ggtitle("Gráfico de Variables (Matriz de Covarianzas)")
```



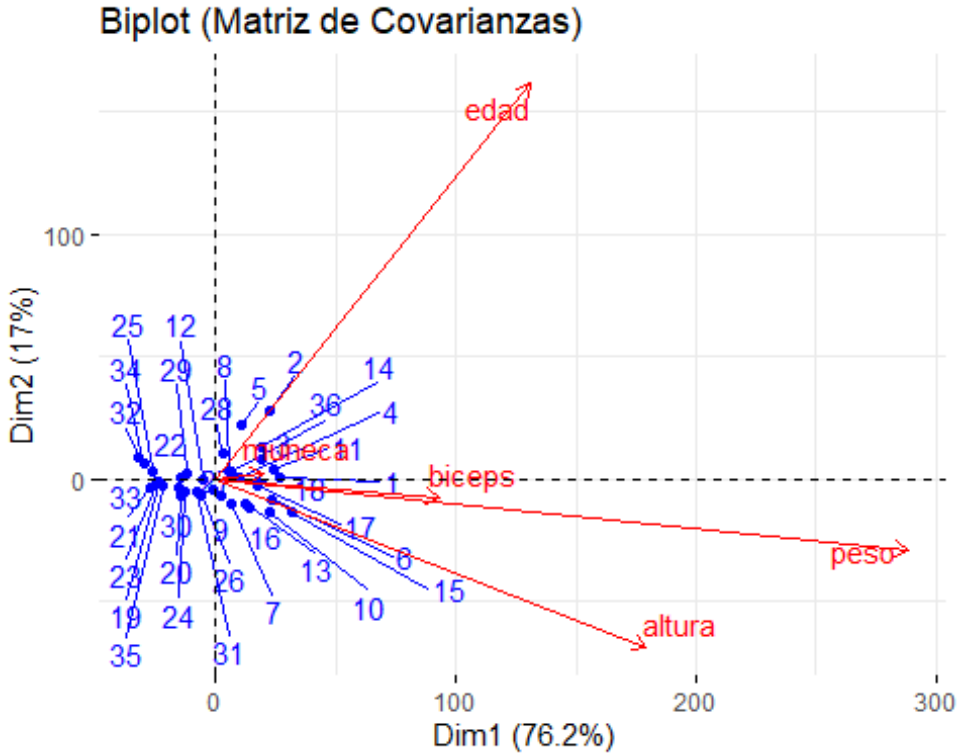
```r
# Screeplot (Varianza explicada por componente)
fviz_screeplot(cpS_cov) +
  ggtitle("Screeplot (Matriz de Covarianzas)")
```



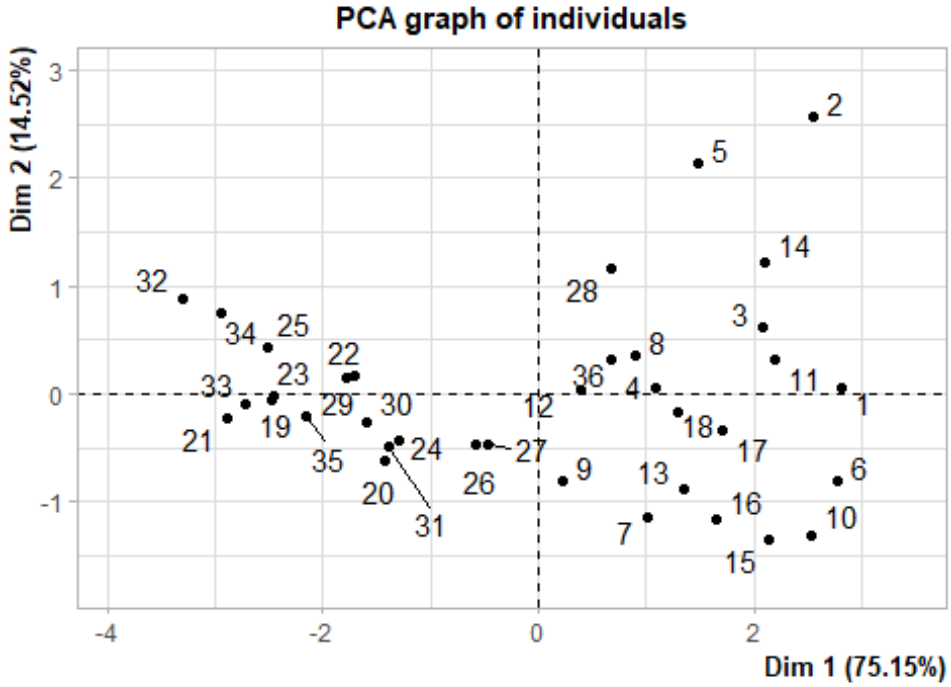

```r
# Gráfico de contribución de las variables
fviz_contrib(cpS_cov, choice = "var") +
  ggtitle("Contribución de Variables (Matriz de Covarianzas)")
```

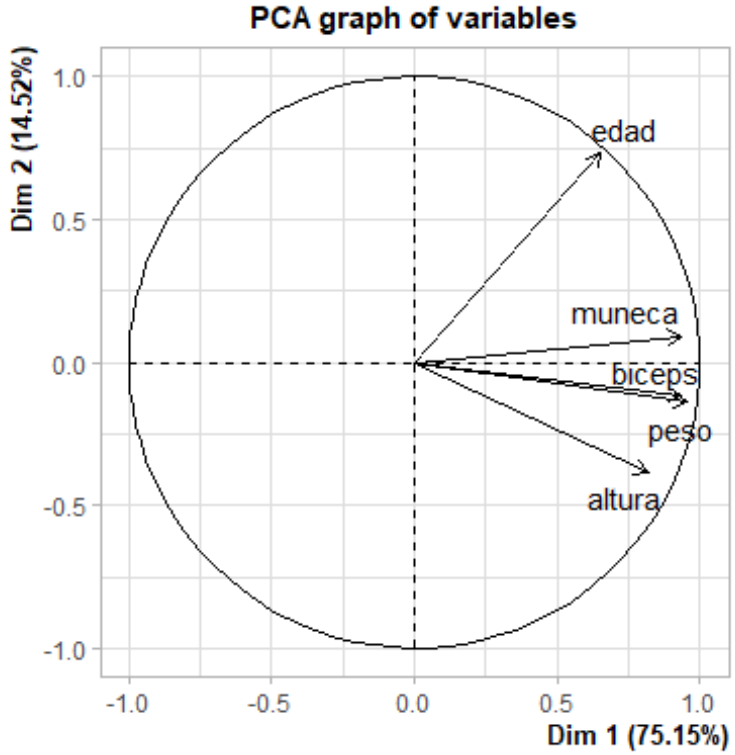



```r
# Biplot
fviz_pca_biplot(cpS_cov, repel=TRUE, col.var="red", col.ind="blue") +
  ggtitle("Biplot (Matriz de Covarianzas)")
```

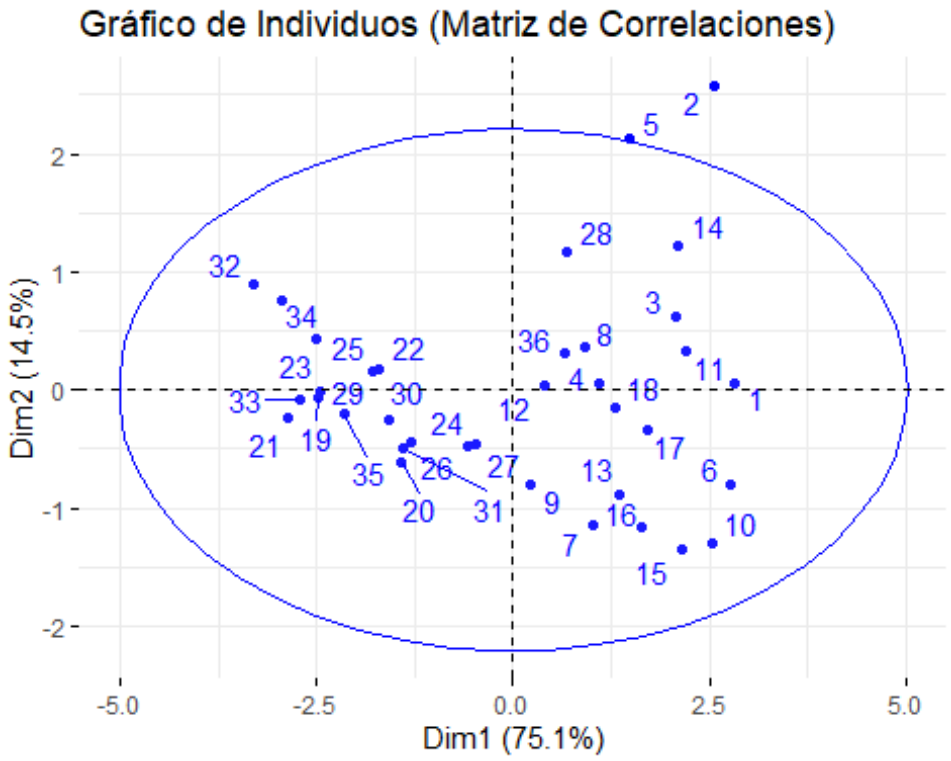


```r
# PCA usando la matriz de correlaciones (scale.unit=TRUE)
cpS_corr = PCA(corporal_num, scale.unit = TRUE)
```

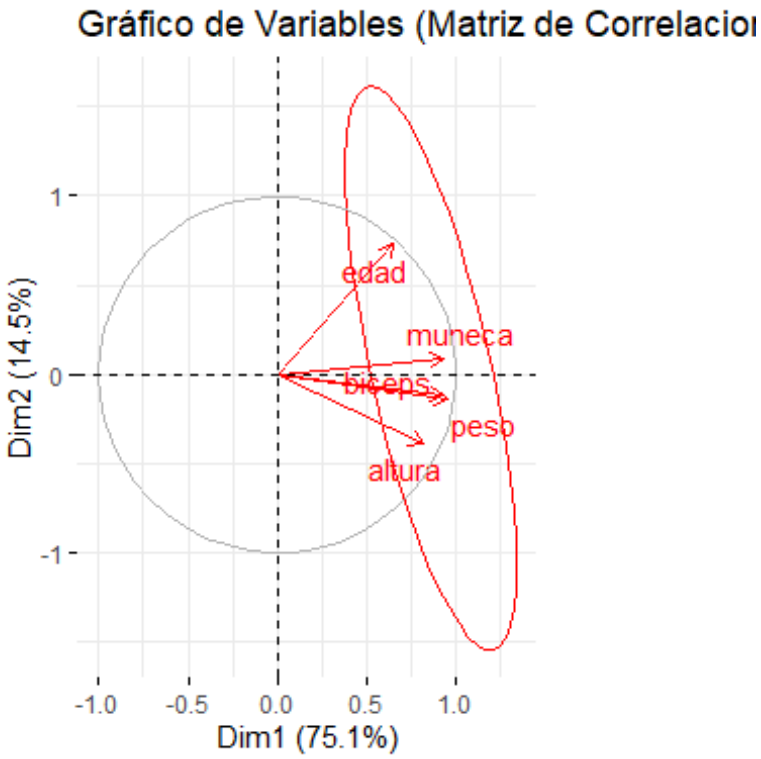




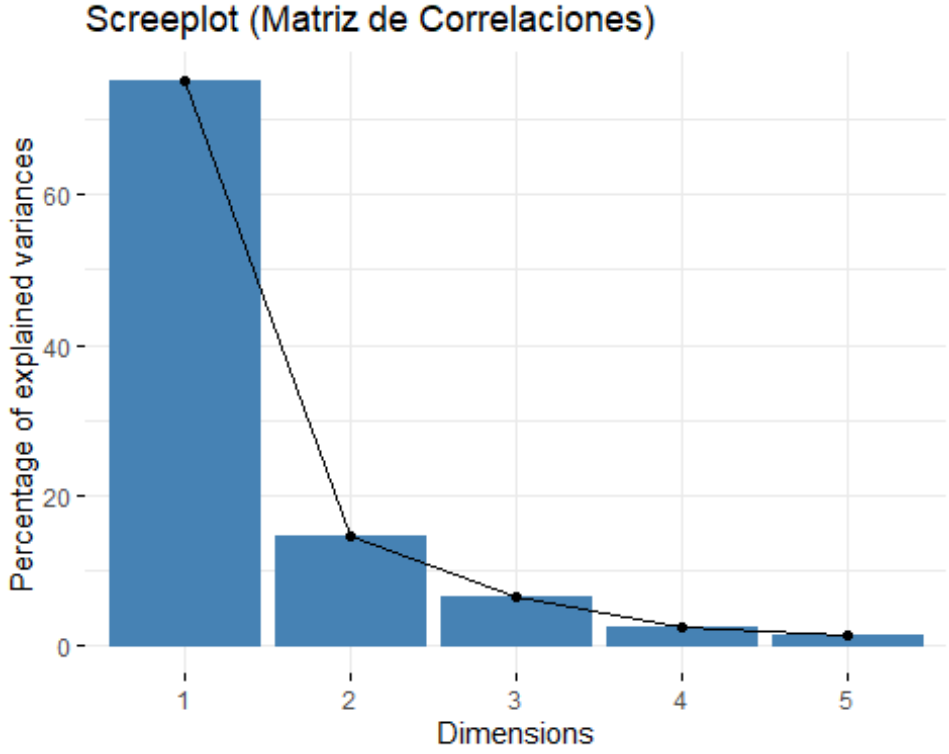
```r
# Gráfico de individuos
fviz_pca_ind(cpS_corr, col.ind = "blue", addEllipses = TRUE, repel = TRUE) +
  ggtitle("Gráfico de Individuos (Matriz de Correlaciones)")
```



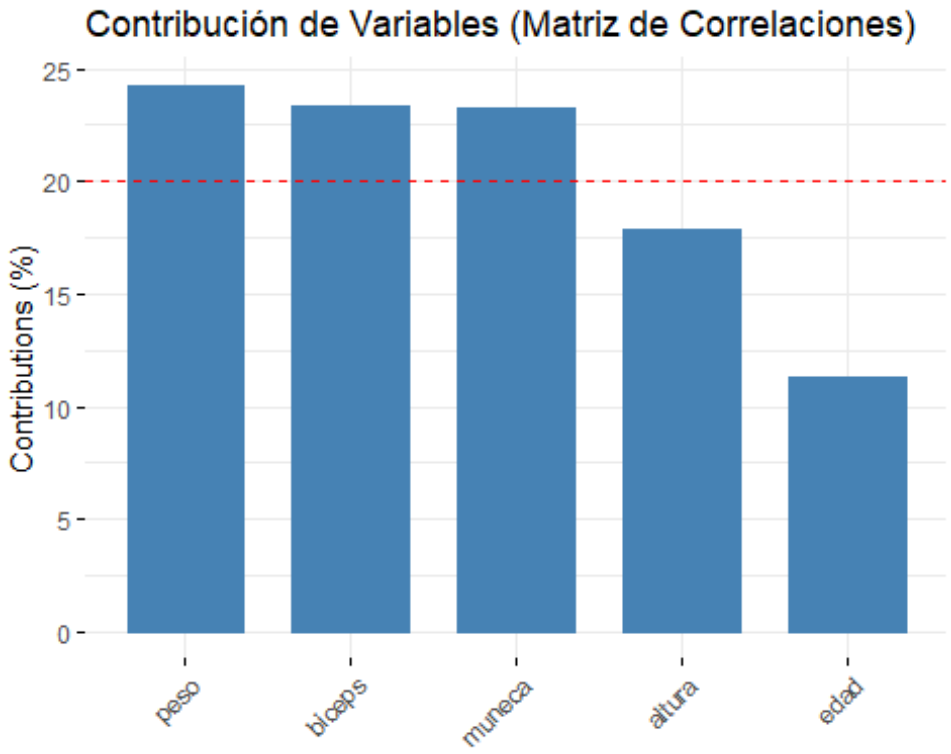

```r
# Gráfico de variables
fviz_pca_var(cpS_corr, col.var = "red", addEllipses = TRUE, repel = TRUE) +
  ggtitle("Gráfico de Variables (Matriz de Correlaciones)")
```



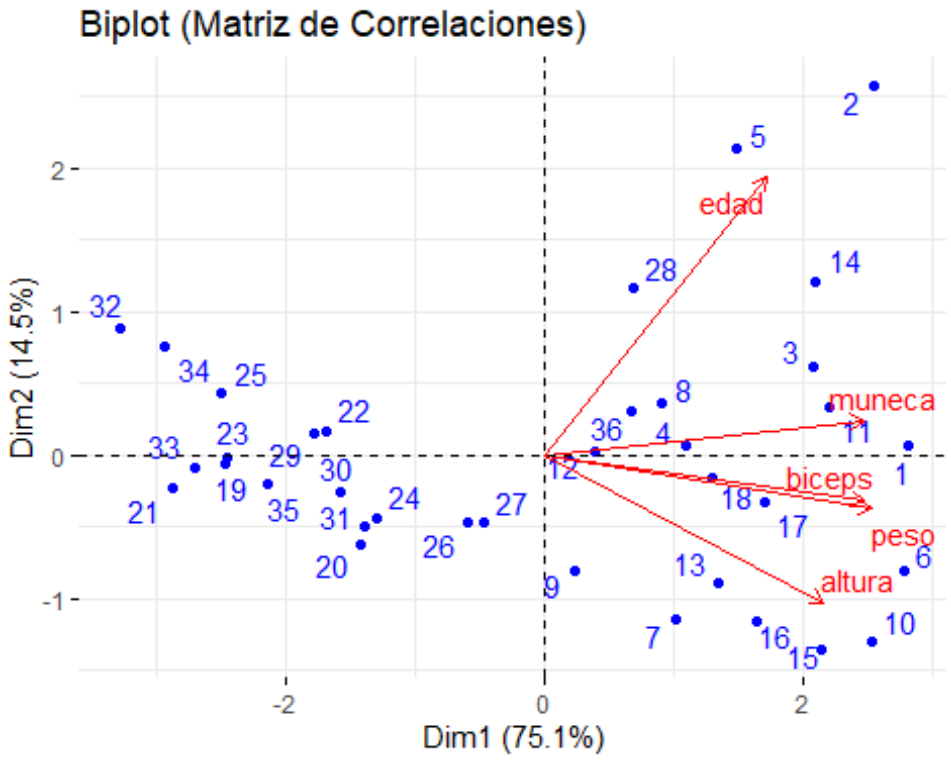

```r
# Screeplot (Varianza explicada por componente)
fviz_screeplot(cpS_corr) +
  ggtitle("Screeplot (Matriz de Correlaciones)")
```



```r
# Gráfico de contribución de las variables
fviz_contrib(cpS_corr, choice = "var") +
  ggtitle("Contribución de Variables (Matriz de Correlaciones)")
```




```r
# Biplot
fviz_pca_biplot(cpS_corr, repel=TRUE, col.var="red", col.ind="blue") +
  ggtitle("Biplot (Matriz de Correlaciones)")
```



Interpretación de las Gráficas Gráfico de Individuos (PC1 vs PC2): Este gráfico muestra cómo se distribuyen las observaciones (los individuos) en el espacio de las dos primeras componentes principales. Los puntos que están más alejados del centro representan observaciones con valores extremos en las variables que más contribuyen a CP1 y CP2. Por ejemplo, individuos que están lejos en el eje de CP1 probablemente tienen valores extremos de peso o altura, mientras que los que se encuentran lejos en CP2 podrían tener edades fuera del promedio.

Gráfico de Variables: El gráfico de variables muestra cómo las variables originales (peso, altura, edad, etc.) están relacionadas entre sí y con las componentes principales. Las variables más alejadas del centro tienen la mayor contribución a los componentes principales. En este caso, peso y altura están muy cercanas al eje de CP1, indicando que son las variables más influyentes en esa componente. Esto significa que CP1 puede ser interpretado como una medida general de constitución física.

Screeplot: El screeplot visualiza la cantidad de varianza explicada por cada componente principal. Aquí podemos ver que los dos primeros componentes explican una gran proporción de la varianza, mientras que los siguientes componentes aportan muy poca información adicional. Esto confirma que es suficiente trabajar con CP1 y CP2 para capturar la mayoría de la estructura de los datos.

Biplot: El biplot combina las dos gráficas anteriores y nos da una imagen más completa de cómo se relacionan las variables y los individuos. Aquí podemos ver, por ejemplo, que los individuos con valores altos en CP1 tienden a tener también valores altos en peso y altura. Este gráfico es útil para interpretar las relaciones entre observaciones y variables al mismo tiempo.

Parte IV

#Interpretaciones y Conclusiones Finales del Análisis de Componentes Principales

En este análisis de componentes principales (PCA), se trabajó con dos enfoques: uno basado en la matriz de varianzas-covarianzas y otro basado en la matriz de correlaciones. Cada uno de estos enfoques tiene implicaciones importantes según las características de los datos.

#Comparación entre Covarianzas y Correlaciones Para este conjunto de datos, que incluye variables como edad, peso, altura, muñeca y bíceps, las escalas son bastante similares. Por ejemplo, tanto el peso como la altura están en unidades comparables, lo que hace que la matriz de varianzas-covarianzas sea adecuada porque no necesitamos normalizar los datos. Este enfoque preserva la magnitud de las variables y su influencia en los componentes principales.

Sin embargo, si los datos tuvieran escalas muy diferentes o si trabajáramos con indicadores sociales y económicos de países, la matriz de correlaciones sería más apropiada. Esto es porque estandariza las variables, dándoles a todas un peso equivalente en el análisis, lo que evita que una variable domine simplemente por tener una escala mayor (como PIB en millones frente a tasa de alfabetización en porcentaje).

#Componentes Principales: ¿Qué Variables Contribuyen Más? Al analizar los resultados, encontramos que el primer componente principal (CP1) explica aproximadamente el 75% de la varianza en ambos enfoques (covarianzas y correlaciones), mientras que el segundo componente principal (CP2) agrega otro 15%. Juntos, estos dos componentes explican más del 90% de la variabilidad total, lo que significa que capturan la mayor parte de la estructura subyacente en los datos.

Lo que esto nos dice es que en el CP1, las variables peso y altura son las más importantes, ya que tienen la mayor influencia en este componente. Esto tiene sentido, porque estas dos variables están altamente correlacionadas con la constitución física de una persona. El CP2, por otro lado, está más influido por la edad, lo cual también es lógico, ya que la edad puede tener un impacto diferente en las características físicas como el tamaño de bíceps.

#Conclusión: ¿Cuál Enfoque Es Mejor? En resumen, para este conjunto de datos, el análisis basado en la matriz de varianzas-covarianzas es más adecuado debido a las escalas comparables de las variables. Sin embargo, si estuviéramos trabajando con datos más diversos, como indicadores sociales o económicos, la matriz de correlaciones sería preferible, ya que estandariza las variables y permite que cada una tenga un peso equitativo en el análisis.

El análisis muestra que el primer componente principal está fuertemente influenciado por las variables peso y altura, lo que sugiere que este componente refleja principalmente la constitución física de los individuos. El segundo componente está más relacionado con la edad, lo que podría indicar cambios físicos asociados con el envejecimiento.

En general, este análisis nos permite identificar rápidamente las variables más importantes y agrupar a los individuos según sus características físicas, lo que puede ser muy útil para estudios que analicen la relación entre estas medidas y otros factores, como la salud o el rendimiento físico.