ActividadA4

Adrian Pineda Sanchez

2024-10-08

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.

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:

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

head(M)
```

```
##   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'
M <- M[, !colnames(M) %in% c("sexo")]

# Ver las primeras filas para asegurarse que todo está en orden
head(M)
```

```
##   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 varianza-covarianza
S <- cov(M)

# Mostrar las matrices
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 correlación
R <- cor(M)
```

```
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
```

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

```
# Valores y vectores propios de la matriz de covarianza
eigen_S <- eigen(S)

# Mostrar resultados
eigen_S$values  # Valores propios de S
```

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

```
eigen_S$vectors # Vectores propios de S
```

```
##             [,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
```

```
# Valores y vectores propios de la matriz de correlación
eigen_R <- eigen(R)

eigen_R$values  # Valores propios de R
```

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

```
eigen_R$vectors # Vectores propios de R
```

```
##            [,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
```

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.

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

# Mostrar proporciones
proporcion_varianza_S
```

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

```
# Proporción de varianza explicada en la matriz de correlación
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
```

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

```
# Varianza acumulada para la matriz de covarianza
varianza_acumulada_S <- cumsum(proporcion_varianza_S)

# Varianza acumulada para la matriz de correlación
varianza_acumulada_R <- cumsum(proporcion_varianza_R)

# Mostrar resultados de la varianza acumulada
cat("Acumulada S:",varianza_acumulada_S,"\n","\n")
```

```
## Acumulada S: 0.7615357 0.9318456 0.9903763 0.9995034 1
##
```

```
cat("Acumulada R:", varianza_acumulada_R)
```

```
## Acumulada R: 0.7514995 0.8966708 0.9607368 0.9856605 1
```

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

los componentes más importantes son aquellos que explican la mayor proporción de la varianza. Observando los valores de la matriz de covarianza, el primer componente explica aproximadamente el 76.15% de la varianza total, y el segundo componente añade un 17.03%, lo que da un total de 93.18% de la varianza explicada con solo dos componentes.

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.S

Matriz de Covarianzas

```
# Obtener los coeficientes de CP1 y CP2
CP1 <- eigen_S$vectors[,1]
CP2 <- eigen_S$vectors[,2]

# Crear los nombres de las variables
variables <- c("edad", "peso", "altura", "muñeca", "bíceps")

# Crear la ecuación de CP1
cat("Combinación lineal para CP1:\n")

## Combinación lineal para CP1:

cat("CP1 = ")

## CP1 =

for (i in 1:length(CP1)) {
  cat(CP1[i], "*", variables[i])
  if (i != length(CP1)) cat(" + ")
}

## -0.34871 * edad + -0.7661759 * peso + -0.4763241 * altura + -0.05386189 *
muñeca + -0.2481737 * bíceps

cat("\n\n")

# Crear la ecuación de CP2
cat("Combinación lineal para CP2:\n")

## Combinación lineal para CP2:

cat("CP2 = ")

## CP2 =

for (i in 1:length(CP2)) {
  cat(CP2[i], "*", variables[i])
  if (i != length(CP2)) cat(" + ")
}

## 0.9075501 * edad + -0.1616581 * peso + -0.3851755 * altura + 0.0155423 *
muñeca + -0.0402221 * bíceps

cat("\n")
```

CP1: En la combinación lineal de CP1 los coeficientes más altos en valor absoluto son:

Peso (-0.7661759): Este es el coeficiente más grande en magnitud, lo que indica que la variable peso tiene la mayor influencia en la primera componente principal. Altura (-0.4763241): La segunda variable más importante es altura, dado que su coeficiente también es relativamente grande en comparación con las otras variables. *Edad (-0.34871): La edad también tiene una contribución notable, pero menor que las anteriores.* Justificación de CP1: La primera componente principal está principalmente influenciada por el peso y la altura, seguidos por la edad. Estas variables capturan la mayor parte de la variabilidad en los datos, lo que sugiere que las diferencias en estas variables son las que más contribuyen a la varianza total explicada.

CP2: En la combinación lineal de CP2, los coeficientes más altos en valor absoluto son:

Edad (0.9075501): Este es, con mucha diferencia, el coeficiente más grande, lo que indica que la variable edad es la más importante en la segunda componente principal. Altura (-0.3851755): La altura también contribuye de manera significativa, aunque mucho menos que la edad. Justificación de CP2: La segunda componente principal está dominada por la edad, con un coeficiente muy alto en comparación con las otras variables. Esto sugiere que en la segunda dimensión, las diferencias de edad entre los individuos capturan la mayor parte de la variabilidad adicional que no fue explicada por el primer componente principal.

Matriz de Correlaciones

```
# Obtener los coeficientes de CP1 y CP2 para la matriz de correlación
CP1_R <- eigen_R$vectors[,1]
CP2_R <- eigen_R$vectors[,2]

# Crear la ecuación de CP1 para la matriz de correlación
cat("Combinación lineal para CP1 en la matriz de correlaciones:\n")

## Combinación lineal para CP1 en la matriz de correlaciones:

cat("CP1 = ")

## CP1 =

for (i in 1:length(CP1_R)) {
  cat(CP1_R[i], "*", variables[i])
  if (i != length(CP1_R)) cat(" + ")
}

## -0.335931 * edad + -0.4927066 * peso + -0.4222426 * altura + -0.4821923 *
muñeca + -0.4833139 * bíceps

cat("\n\n")

# Crear la ecuación de CP2 para la matriz de correlaciones
cat("Combinación lineal para CP2 en la matriz de correlaciones:\n")

## Combinación lineal para CP2 en la matriz de correlaciones:
```

```r
cat("CP2 = ")

## CP2 =

for (i in 1:length(CP2_R)) {
  cat(CP2_R[i], "*", variables[i])
  if (i != length(CP2_R)) cat(" + ")
}

## 0.8575601 * edad + -0.1647821 * peso + -0.4542223 * altura + 0.1082775 *
muñeca + -0.1392684 * bíceps

cat("\n")
```

CP1: En la combinación lineal de CP1, los coeficientes más altos en valor absoluto son:

Peso (-0.4927066): Este es el coeficiente más grande en magnitud, lo que indica que la variable peso tiene la mayor influencia en la primera componente principal. Bíceps (-0.4833139): La segunda variable más importante es el bíceps, dado que su coeficiente es muy cercano al del peso, lo que indica que también tiene una influencia significativa en esta componente. Muñeca (-0.4821923): La muñeca contribuye de manera significativa, muy cercana en magnitud a las dos variables anteriores, lo que sugiere que estas tres variables capturan gran parte de la variabilidad en los datos. Justificación de CP1: La primera componente principal está principalmente influenciada por el peso, el bíceps, y la muñeca. Estas tres variables tienen coeficientes muy similares en valor absoluto, lo que sugiere que las diferencias en estas medidas corporales son las que más contribuyen a la varianza total explicada por esta componente principal. Esto sugiere que las variaciones en características corporales (peso y dimensiones físicas) son clave para entender la variabilidad capturada por CP1.

CP2: En la combinación lineal de CP2, los coeficientes más altos en valor absoluto son:

Edad (0.8575601): Este es, con diferencia, el coeficiente más grande, lo que indica que la variable edad es la más importante en la segunda componente principal. Altura (-0.4542223): La altura también contribuye de manera significativa, aunque su influencia es mucho menor que la de la edad. Justificación de CP2: La segunda componente principal está dominada por la variable edad, con un coeficiente mucho más grande que los de las otras variables. Esto sugiere que, en esta segunda dimensión, las diferencias en la edad de los individuos explican la mayor parte de la variabilidad adicional que no fue capturada por el primer componente principal. La altura tiene una influencia moderada, pero es notablemente menos importante que la edad en este contexto.

PARTE II

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

```r
# Realizar el análisis de componentes principales (PCA) utilizando la matriz
de varianza-covarianza
```

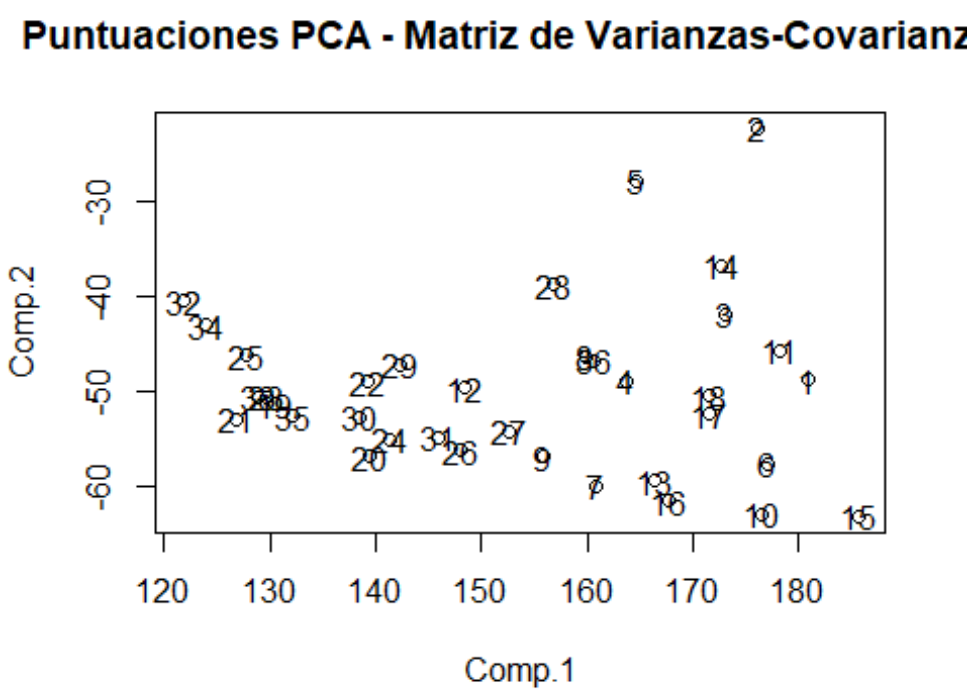
```r
cpS <- princomp(M, cor = FALSE)

# Obtener las puntuaciones (scores) de las observaciones para la matriz de covarianza
scores_cov <- cpS$scores
head(scores_cov)  # Ver las primeras puntuaciones
```

```
##         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
```

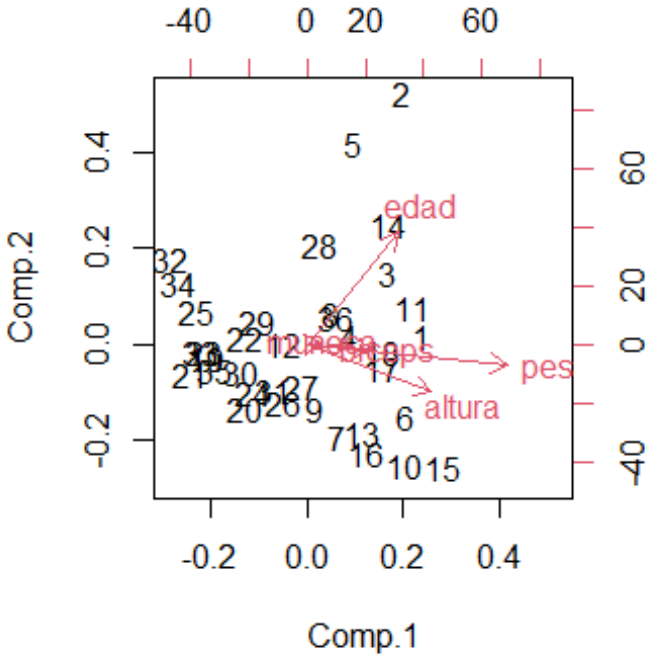
```r
# Realizar el análisis de componentes principales (PCA) utilizando la matriz de correlación
cpR <- princomp(M, cor = TRUE)

# Obtener las puntuaciones (scores) de las observaciones para la matriz de correlación
scores_cor <- cpR$scores
head(scores_cor)  # Ver las primeras puntuaciones
```

```
##        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
```

```r
cpS <- princomp(M, cor = FALSE)
cpaS <- as.matrix(M) %*% cpS$loadings
plot(cpaS[,1:2], type = "p", main = "Puntuaciones PCA - Matriz de Varianzas-Covarianzas")
text(cpaS[,1], cpaS[,2], labels = 1:nrow(cpaS))
```


Puntuaciones PCA - Matriz de Varianzas-Covarianz
Comp.2
-30
-40
-50
-60
120
130
140
150
160
170
180
Comp.1

```
biplot(cpS)
```


-40
0
20
60
Comp.2
0.4
0.2
0.0
-0.2
edad
pes
altura
60
20
0
-40
-0.2
0.0
0.2
0.4
Comp.1

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

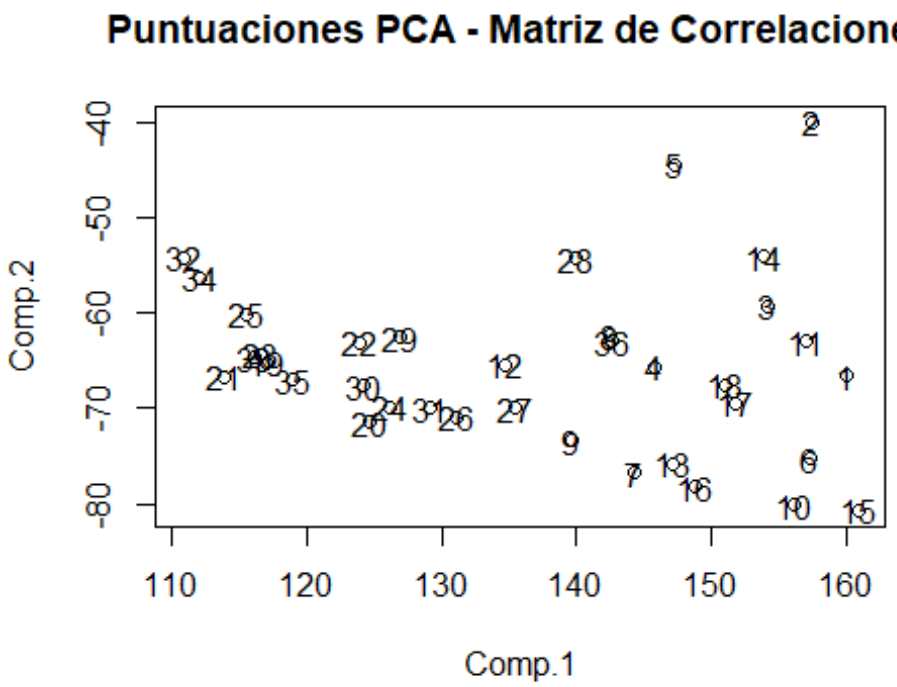
En el gráfico de varianza-covarianza, CP1 captura un 76.15% de la variabilidad total, y CP2 captura un 17% adicional. Esto sugiere que CP1 está fuertemente influenciada por variables con alta varianza como peso y altura, ya que estas tienen los coeficientes de carga más altos. Por ejemplo, en el gráfico de cargas de las variables, peso tiene una carga alta en CP1, lo que indica que contribuye significativamente a la variabilidad en CP1. La altura también contribuye, aunque en menor medida. En cambio, CP2 captura la variabilidad residual no explicada por CP1, como la relacionada con edad.

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

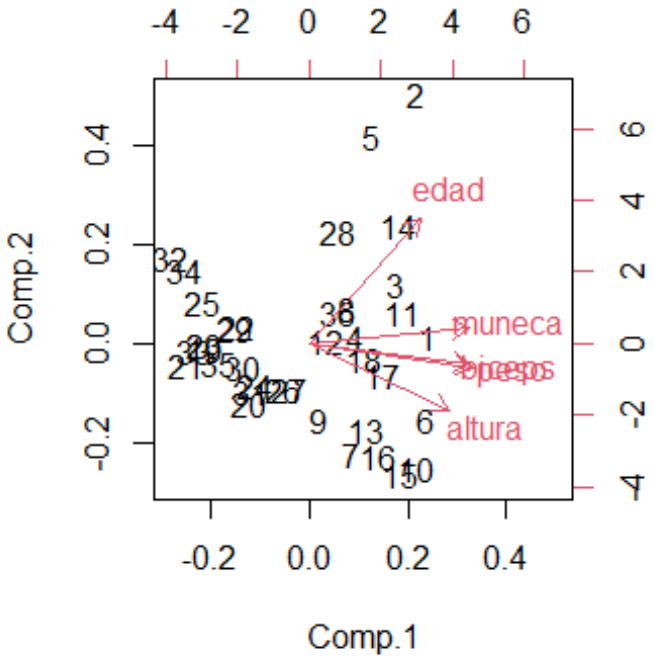
En las puntuaciones de las observaciones, podemos ver que los puntos más altos y a la derecha (por ejemplo, observaciones 2 y 5) tienen altos valores en variables como edad la cual es predominante en comp2. Por el contrario, las observaciones más a la izquierda (como la 32) tienen valores más bajos en esas variables. Las observaciones más altas en el eje CP1 son las relacionadas a las carcateristicas fisicas propias del individuo, peso altura, biceps etc, ye stas predominan valores a la derecha como 10,15, 6 etc. ### Detecte posibles datos atípicos

Matriz de varianza-covarianza: En el primer gráfico, observaciones que están muy alejadas de la nube de puntos principal, como los puntos en el extremo derecho o el extremo superior del gráfico como la 2 y 5, que están muy alejadas del centro del gráfico, podrían considerarse datos atípicos, . Estas observaciones tienen características que las diferencian significativamente del resto de los datos en las variables que influyen en CP1 y CP2. asi como tambien podriamos darles un vistaso a los datos a la izquierda y ver si no tienen correlaciones negativas cayendo en datos invalidos con datos que deben tener un valor positivo coherente

```
cpR <- princomp(M, cor = TRUE)
# Calcula las puntuaciones (scores)
cpaR <- as.matrix(M) %*% cpR$loadings
# Graficar las puntuaciones de las dos primeras componentes principales
plot(cpaR[,1:2], type = "p", main = "Puntuaciones PCA - Matriz de
Correlaciones")
text(cpaR[,1], cpaR[,2], labels = 1:nrow(cpaR))
```




```r
# Graficar el biplot
biplot(cpR)
```



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

En el gráfico con la matriz de correlación, el primer componente principal (CP1) está principalmente influenciado por variables como peso, biceps, muñeca y altura, ya que estas variables se alinean más con CP1, como lo podemos ver en el gráfico de "loadings". En cambio, el segundo componente (CP2) está principalmente influenciado por la variable edad, la cual tiene una dirección notablemente alineada con el eje de CP2. Esto sugiere que CP1 captura la mayor variabilidad de las demás variables (peso, altura, etc.), mientras que CP2 está centrado en la variabilidad de la edad.

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

En el gráfico de la matriz de correlación, las puntuaciones de las observaciones reflejan cómo las distintas observaciones se proyectan en el espacio estandarizado de los componentes principales (CP1 y CP2). Las observaciones cercanas en el gráfico comparten patrones similares en cuanto a sus valores de las variables. Observaciones en el lado derecho del gráfico (como la observación 2) tienen valores más altos en variables como peso y altura asimismos influenciados por la edad aunque lo vemos algo retidao del grupo, mientras que las observaciones cercanas a la parte superior reflejan mayor peso en la variable edad, como las observaciones 14, 5, y 3. Sin embargo todos los datos de la izquieda 32,34.25 muchos numeros de rangos del 20-30 parecen tener una correlacion inversa tantos en la parte inferior izquiera con edad asi como en la izquieda media con las variables fisicas.

Detecte posibles datos atípicos

En el gráfico de correlación, las observaciones alejadas del centro y de la mayoría de los puntos (como la observación 2 en la parte superior derecha y la observación 15 en la parte inferior derecha) son posibles datos atípicos. Estas observaciones tienen características que difieren significativamente del patrón general en las variables que influyen en CP1 y CP2. Asimismo podiamos notar los valores extremos taales como 32,34 o los valores mas extremos, aunque si lo vemos por la cuestion de la combinacion lineal podemos ver que estas caracteristicas fisicas tienen pesos coherentes,

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

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?

```
cpS = princomp(M, cor = FALSE)  # Mantener cor = FALSE para usar la matriz de varianza-covarianza

cpR = princomp(M, cor = TRUE)  # Usar cor = TRUE para estandarizar las variables y usar la matriz de correlación
```

```r
# Resumen de la varianza explicada por cada componente
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
# Cargas de las variables en los componentes principales
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
# Puntajes de las observaciones
cpS$scores
```

```
##           Comp.1      Comp.2      Comp.3       Comp.4      Comp.5
##  [1,]  27.162853   1.0278492   5.0022646  0.936226898 -0.51688356
##  [2,]  22.363542  27.5955807   3.0635949 -0.083381259  0.02552809
##  [3,]  19.167874   7.9566157  -1.5770026 -2.610776762  0.80391745
##  [4,]   9.959001   0.8923731   5.5146952  0.123453725 -0.35579895
##  [5,]  10.775593  22.0203437  -0.7562826  0.179967226 -0.41646606
##  [6,]  23.283948  -7.9268214   2.7958617 -2.093392841 -0.62252321
##  [7,]   6.949553 -10.1882447   1.5804639 -5.636477243  0.75692216
##  [8,]   5.981213   3.4214568  -7.0113449 -0.999845471 -0.13795746
##  [9,]   2.128453  -7.0823040   9.6199213 -2.402765355  0.30931008
## [10,]  22.742222 -13.2447241  -5.8006902 -1.900258608 -0.11415400
## [11,]  24.427931   4.1227827  -3.0914640  1.417935347  0.45836253
## [12,]  -5.438123   0.1807499   1.3551969 -5.147087631 -0.71928452
## [13,]  12.665261  -9.7148314  -4.4445147  0.469977365 -0.44199755
## [14,]  18.962350  13.1080907   4.5325770  0.310839551 -0.27648044
## [15,]  31.842783 -13.4784052  -1.4672915  5.610391303  0.61177438
## [16,]  13.884278 -11.8930081  -6.4032979 -2.225813208 -0.01138562
## [17,]  17.653813  -2.6451319  -0.8986274 -0.529020358  0.37187295
```

```
## [18,]  17.723299  -0.7428241   0.1219847  1.785013852  0.68809035
## [19,] -23.293603  -1.5208783   0.2627514  1.143811767 -0.16480880
## [20,] -14.414169  -7.0887516   0.1030611  0.006854239 -0.32687435
## [21,] -27.078917  -3.1933468  -0.4483831  0.722326288 -0.02028518
## [22,] -14.579228   0.8324474  -9.1400445  1.717699742  0.23470254
## [23,] -24.042246  -0.7779288  -5.8550300 -0.340341079  0.26832127
## [24,] -12.494468  -5.2751971   3.0622990  1.094339917 -0.51675730
## [25,] -26.002609   3.5759758   1.6616974  0.054118319 -0.33475598
## [26,]  -5.766003  -6.4856729  -6.5862305  2.330421808 -0.76268815
## [27,]  -1.211876  -4.4901315   4.4920764  1.153351801  0.26364518
## [28,]   3.020501  11.0467489 -10.8052957  0.255974364 -0.43453383
## [29,] -11.574038   2.5907341   9.5304169  1.466717121  0.84144772
## [30,] -15.335150  -2.9912143   6.9968010  0.493427421 -0.36660212
## [31,]  -7.926087  -5.1312097   4.1467185  2.808113699  0.29328661
## [32,] -32.046176   9.3863372   0.8359798 -1.341797979  0.73976836
## [33,] -24.800765  -0.8616289  -0.1246471 -0.477476584  0.58698947
## [34,] -29.884003   6.8137270  -9.5237493 -0.372525171  0.27802711
## [35,] -21.626441  -2.8831824   7.4391447  0.704477945 -0.64549912
## [36,]   6.819433   3.0436244   1.8163894  1.375519851 -0.34623005
```

*Varianza explicada: Utilizando summary(cpS), en la matriz de varianza-covarianza (S), CP1 explica el 76.15% de la varianza total, y CP2 el 17.03%, lo que en conjunto suma 93.18%. Esto significa que las dos primeras componentes capturan la mayor parte de la variabilidad presente en los datos originales (no estandarizados), especialmente en variables como peso y altura.

*Cargas (loadings): En la matriz de varianza-covarianza, los coeficientes más grandes en valor absoluto para CP1 son:

–Peso (-0.766) –Altura (-0.476) –Bíceps (-0.248)

Esto indica que el peso y la altura son las variables que más contribuyen a la variabilidad capturada por CP1. En CP2, la variable que más contribuye es la edad (0.908), lo que sugiere que CP2 captura variaciones en la edad que no fueron explicadas por CP1.

*Puntuaciones (scores): En la matriz de varianza-covarianza, las puntuaciones muestran cómo cada observación se proyecta en los componentes principales. Por ejemplo, la primera observación tiene una puntuación de 27.16 en CP1, lo que indica que tiene un valor relativamente alto en las variables que dominan CP1, como el peso y la altura. Observaciones con puntuaciones altas o bajas en CP1 o CP2 indican diferencias significativas en las variables que más influyen en esos componentes.

Dado que se utiliza la matriz de varianza-covarianza, las variables con mayor varianza (como el peso) tendrán un mayor impacto en los componentes principales, ya que no están estandarizadas.

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.

```
# Resumen de la varianza explicada por cada componente
summary(cpR)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4    Comp.5
## Standard deviation     1.9384265 0.8519722 0.56597686 0.35301378 0.2677639
## Proportion of Variance 0.7514995 0.1451713 0.06406596 0.02492375 0.0143395
## Cumulative Proportion  0.7514995 0.8966708 0.96073676 0.98566050 1.0000000
```

```
# Cargas de las variables en los componentes principales
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
```

```
# Puntajes de las observaciones
cpR$scores
```

```
##           Comp.1      Comp.2      Comp.3      Comp.4       Comp.5
##  [1,]  2.8139915  0.06282760  0.51434516 -0.37618363 -0.161649397
##  [2,]  2.5508161  2.57369731  0.42896223  0.01252075  0.083602262
##  [3,]  2.0792069  0.62112516 -0.12602006  0.51138786  0.430775853
##  [4,]  1.0933160  0.06328171  0.46145821 -0.35236278 -0.008424496
##  [5,]  1.4893629  2.13420572 -0.08620983 -0.19530483 -0.097669770
##  [6,]  2.7801900 -0.79964368 -0.11180511 -0.52796031  0.113681564
##  [7,]  1.0141243 -1.14171806 -0.27787746  0.22743193  0.800375496
##  [8,]  0.9063369  0.35803327 -0.79126430  0.07179533 -0.031461084
##  [9,]  0.2285350 -0.80075813  0.71215644 -0.15394896  0.481123407
## [10,]  2.5302453 -1.30235901 -0.76205083  0.03215070  0.050616130
## [11,]  2.2033222  0.32934887  0.10037610  0.49363388 -0.135246631
## [12,]  0.3885728  0.02978904 -0.70291329 -0.72426251  0.460456523
## [13,]  1.3480354 -0.88888844 -0.48237353 -0.13878866 -0.248233214
## [14,]  2.0994018  1.21514134  0.47434543 -0.23319402 -0.019726560
## [15,]  2.1447355 -1.35354752  0.76511713  0.71259130 -0.587575667
## [16,]  1.6489148 -1.16117562 -0.85070099  0.08586963  0.111234627
## [17,]  1.7030809 -0.33209829  0.01673614  0.27827557  0.099895723
## [18,]  1.2932746 -0.15858301  0.48173868  0.55369253 -0.076249945
```

```
## [19,] -2.4795617 -0.06280633  0.02839564 -0.11803106 -0.136704692
## [20,] -1.4200084 -0.61570309 -0.15277478 -0.25447677 -0.063137788
## [21,] -2.8791600 -0.22853227 -0.06023367 -0.03148088 -0.068564803
## [22,] -1.6992789  0.16837324 -0.63755548  0.43611800 -0.277172176
## [23,] -2.4625686 -0.01072936 -0.59031600  0.26691381  0.024784946
## [24,] -1.3015384 -0.43354360  0.20575074 -0.40705451 -0.177314913
## [25,] -2.5058729  0.42780280 -0.01308499 -0.30917018 -0.015086855
## [26,] -0.5896282 -0.46963951 -0.61738513 -0.25029697 -0.536163469
## [27,] -0.4747287 -0.46682854  0.62201914  0.09167385 -0.007586913
## [28,]  0.6816507  1.16291258 -1.08391248  0.03253793 -0.282947483
## [29,] -1.7786024  0.15640801  1.29302710  0.33642964  0.183446578
## [30,] -1.5894735 -0.25254138  0.54948615 -0.44020946 -0.006577363
## [31,] -1.3903223 -0.49360911  0.76675148  0.17233872 -0.188151664
## [32,] -3.2962547  0.88748511  0.06759476  0.35410490  0.371715392
## [33,] -2.7100620 -0.08340844  0.02833828  0.31628667  0.201732879
## [34,] -2.9371073  0.75312128 -0.93702305  0.36683866 -0.011037680
## [35,] -2.1514986 -0.20099407  0.51126095 -0.63846467 -0.074866432
## [36,]  0.6685529  0.31355440  0.25564126 -0.20140147 -0.201892385
```

*Varianza explicada: Utilizando summary(cpR), en la matriz de correlación, CP1 explica el 75.15% de la varianza total y CP2 el 14.52%, sumando 89.67% en conjunto. Esto significa que CP1 y CP2 capturan la mayor parte de la variabilidad presente en los datos estandarizados, donde todas las variables tienen la misma importancia inicial.

*Cargas (loadings): Las cargas indican cómo contribuyen las variables a cada componente:

Para CP1, los valores más altos en valor absoluto son:

-Peso: -0.493 -Bíceps: -0.483 -Muñeca: -0.482 Esto indica que estas tres variables (peso, bíceps, muñeca) capturan la mayor parte de la variabilidad en CP1.

Para CP2, la variable más influyente es:

-Edad: 0.858 Esto sugiere que CP2 captura principalmente la variabilidad relacionada con la edad, que no fue explicada por CP1.

*Puntuaciones (scores): Las puntuaciones proyectan cada observación en los componentes principales. Por ejemplo, en la matriz de correlación, la primera observación tiene una puntuación de 2.81 en CP1 y 0.06 en CP2, lo que indica que esta observación tiene valores altos en las variables que dominan CP1 (peso, bíceps, muñeca). Las observaciones alejadas del centro del gráfico pueden considerarse datos atípicos.

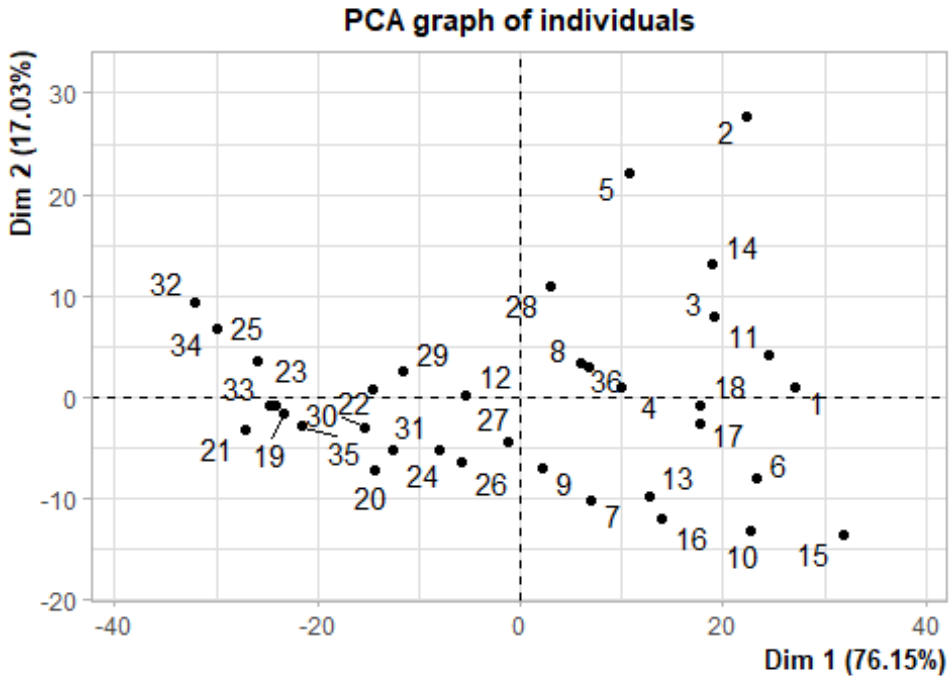
En resumen, en la matriz de correlación, CP1 está dominada por peso, bíceps y muñeca, mientras que CP2 está influenciada principalmente por la edad. Esto sugiere que las diferencias en las medidas corporales capturan la mayor parte de la variabilidad, mientras que la edad añade una dimensión secundaria significativa.

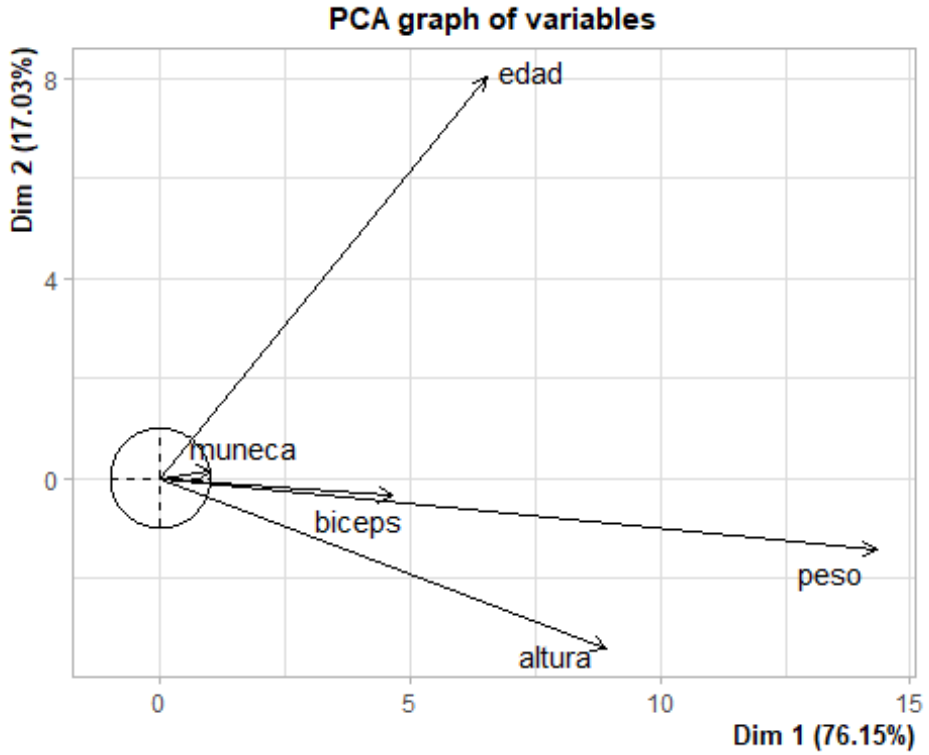
PARTE III

Explore los siguientes gráficos relativos a Componentes Principales.

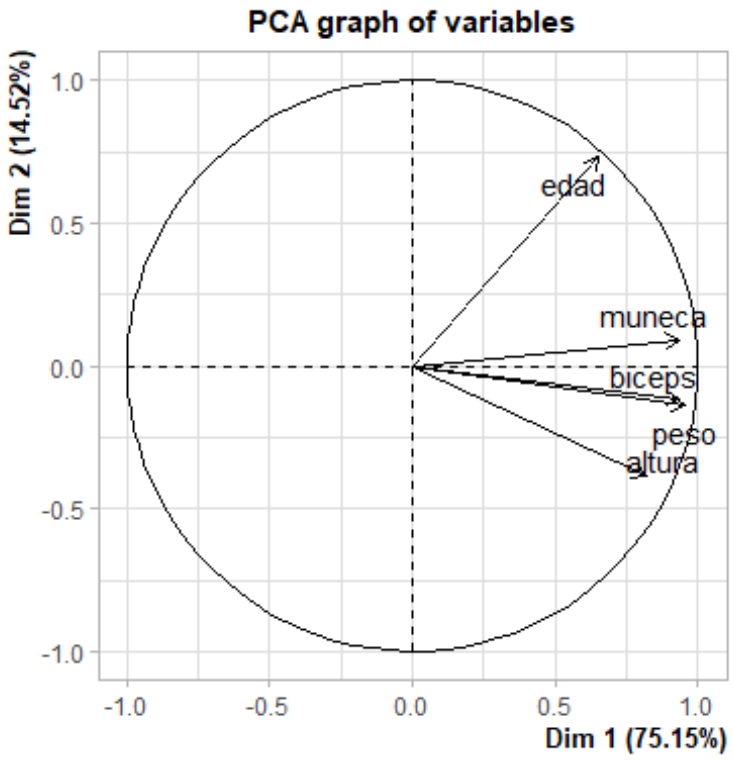
Interprete cada gráfico e identifica qué es lo que se está graficando en cada uno. Realiza el análisis con la matriz de varianzas y covarianzas y correlación.

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

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

cpS = PCA(M,scale.unit=FALSE)
```





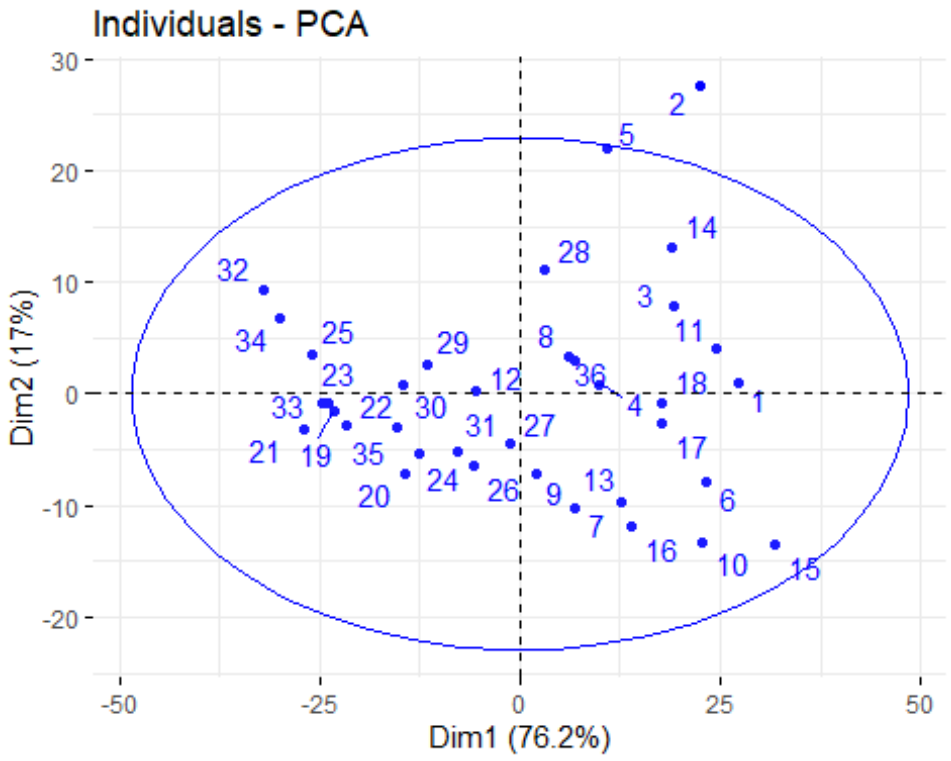

```
cpR = PCA(M,scale.unit=TRUE)
```



Matriz Varianzas y Covarianzas

1.1. Gráfico de las observaciones en el espacio de los componentes principales

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

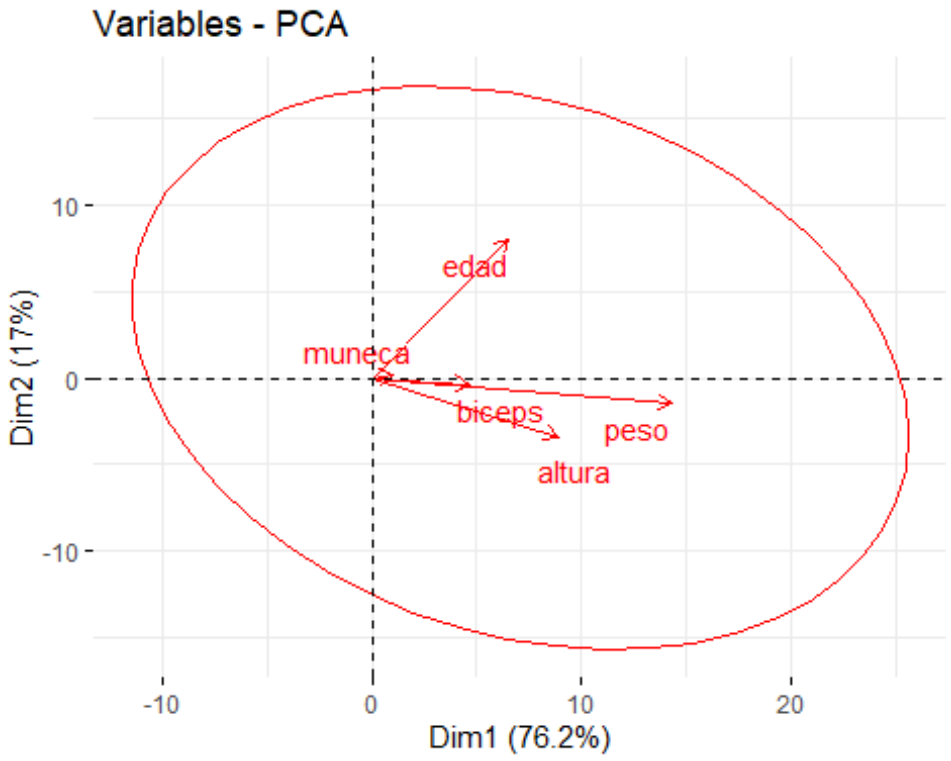


Este gráfico muestra cómo se distribuyen las observaciones (individuos) en el espacio de los dos primeros componentes principales (Dim1 y Dim2). Dim1 (76.2% de la varianza explicada) y Dim2 (17%) capturan la mayor parte de la variabilidad en los datos. Cada punto representa un individuo, y las observaciones más cercanas entre sí son más similares en términos de las variables originales.

*Los puntos más alejados del centro pueden indicar posibles datos atípicos (como el individuo 2), que podrían ser outliers o representar valores extremos en alguna de las variables.

1.2. Gráfico de las variables originales en el espacio de los componentes principales:

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



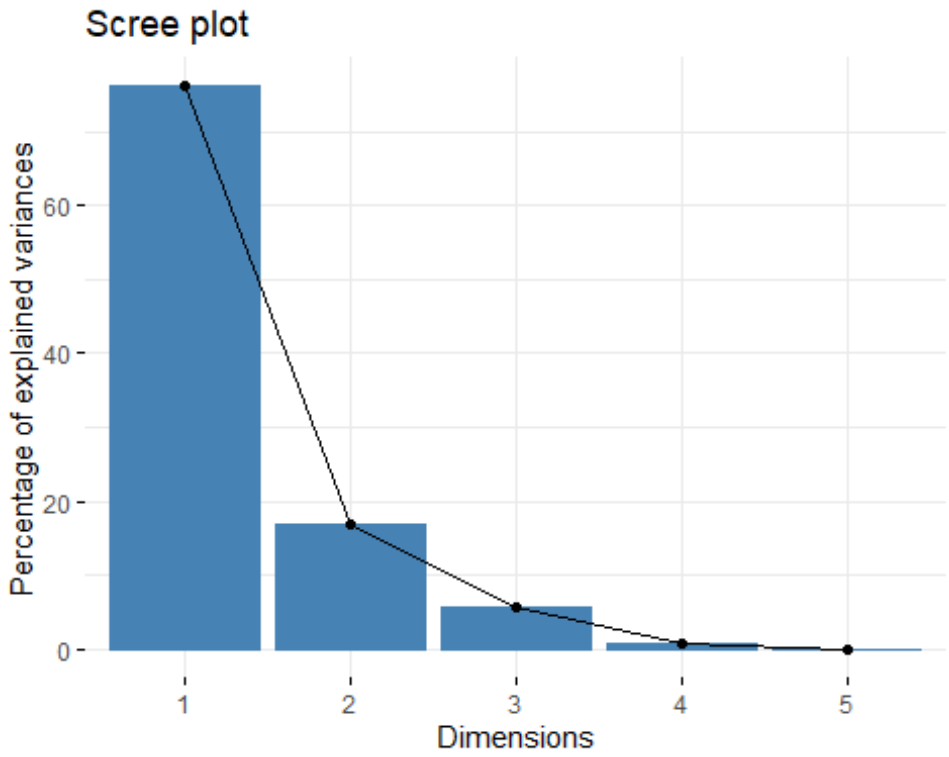
*Este gráfico representa las variables originales proyectadas en el espacio de los componentes principales.

*La dirección y longitud de las flechas indican cuánto contribuye cada variable a los componentes principales. Peso y altura están fuertemente correlacionados con Dim1, mientras que edad y muñeca contribuyen menos.

*Las variables que apuntan en direcciones similares están correlacionadas positivamente, mientras que aquellas que están en direcciones opuestas están correlacionadas negativamente.

1.3. Scree plot (gráfico de sedimentación):

```
fviz_screeplot(cpS)
```

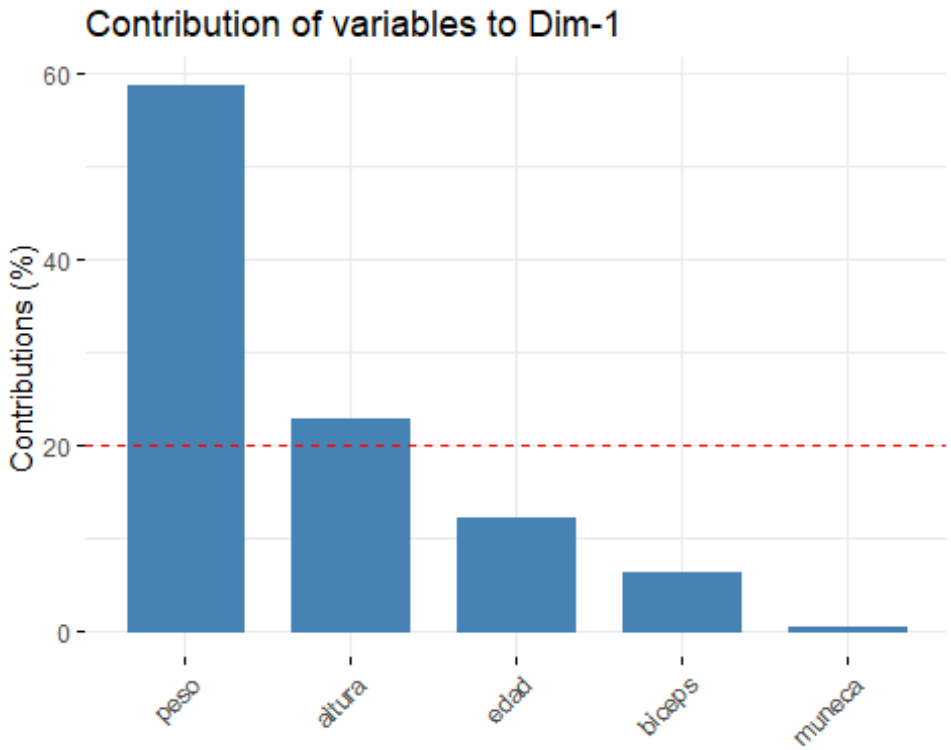


*El Scree plot muestra la varianza explicada por cada componente principal.

*El gráfico sugiere que el primer componente (Dim1) explica una porción significativa de la varianza (76.2%), seguido por el segundo componente (17%). A partir del tercer componente, la varianza explicada disminuye drásticamente, lo que indica que probablemente no se necesiten más de dos componentes para capturar la mayor parte de la información en los datos.

1.4. Contribución de las variables a cada componente principal:

```
fviz_contrib(cpS, choice = c("var"))
```



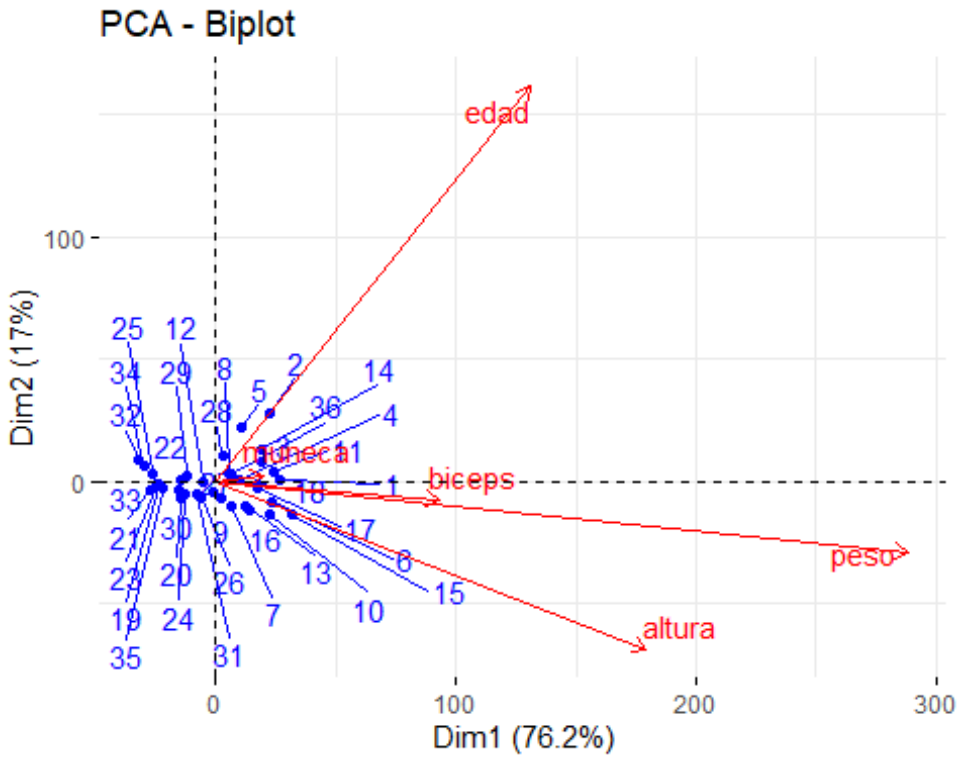
*Este gráfico muestra la contribución de cada variable a Dim1.

*Peso es la variable que más contribuye al primer componente principal (< 60%), seguido de altura y edad. Esto sugiere que Dim1 está principalmente influenciado por el peso de los individuos.

*Las variables con menor contribución son bíceps y muñeca, lo que indica que estas variables no aportan tanta variabilidad en Dim1.

1.5. Biplot de individuos y variables:

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



*El biplot combina las observaciones y las variables en un solo gráfico, mostrando cómo las observaciones se distribuyen en función de las variables.

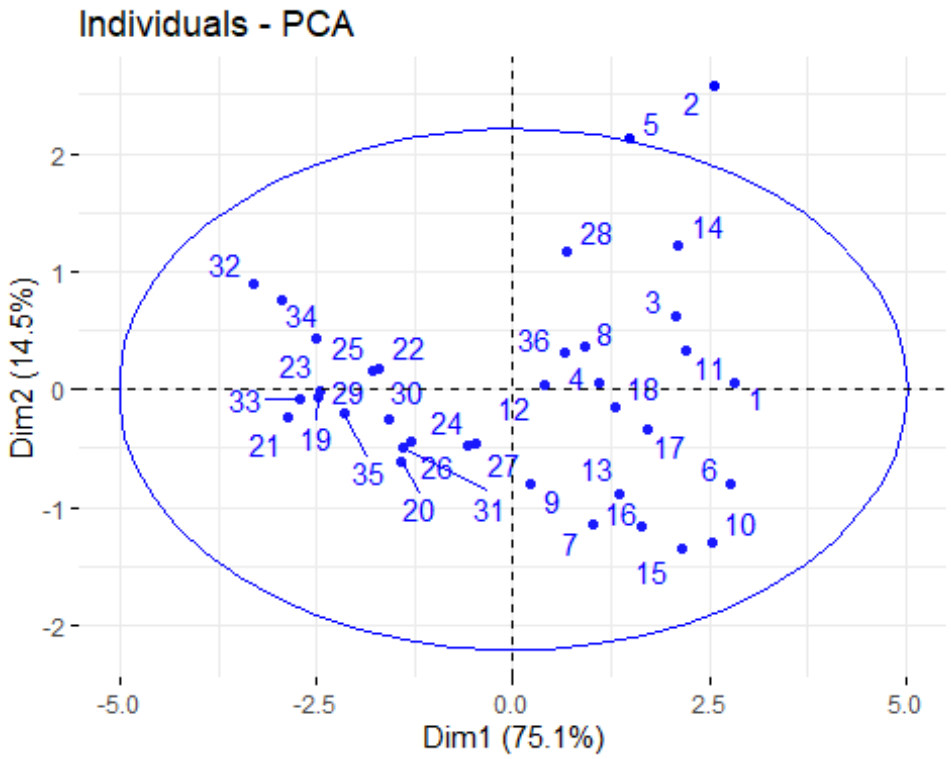
*Las direcciones de las flechas representan las variables originales, y su relación con las observaciones indica cómo cada individuo se alinea con respecto a las variables.

*Por ejemplo, los individuos que están más alineados con la flecha de peso tienen valores más altos de peso, mientras que aquellos alejados tienen valores más bajos.

Matriz Correlaciones

1.1. Gráfico de las observaciones en el espacio de los componentes principales

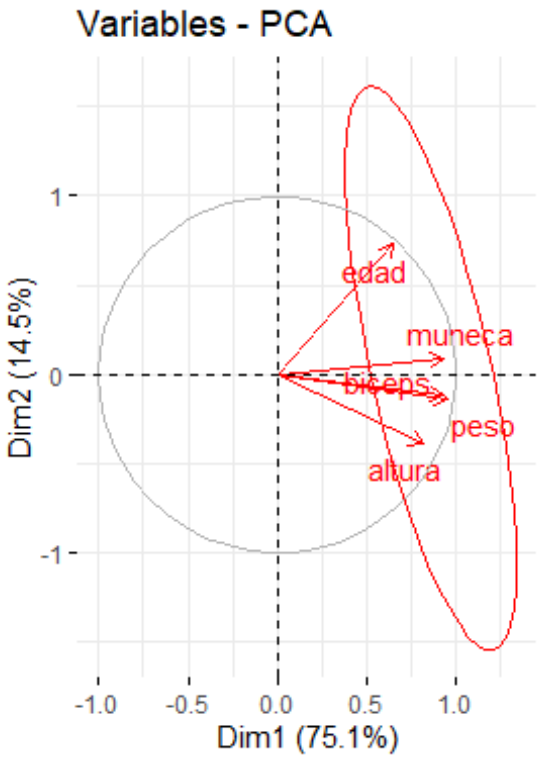
```
fviz_pca_ind(cpR, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```



Este gráfico proyecta a las observaciones (individuos) en el espacio definido por los dos primeros componentes principales, Dim1 (75.1%) y Dim2 (14.5%), capturando en total el 89.6% de la varianza de los datos. Las observaciones cercanas entre sí indican similitud en las variables originales. Los puntos alejados del centro, como los individuos 2, 5, 14 y 28, pueden considerarse observaciones algo atípicas, mientras que otros puntos como el 7 y 16 están más agrupados, lo que indica patrones similares en esas observaciones.

1.2. Gráfico de las variables originales en el espacio de los componentes principales:

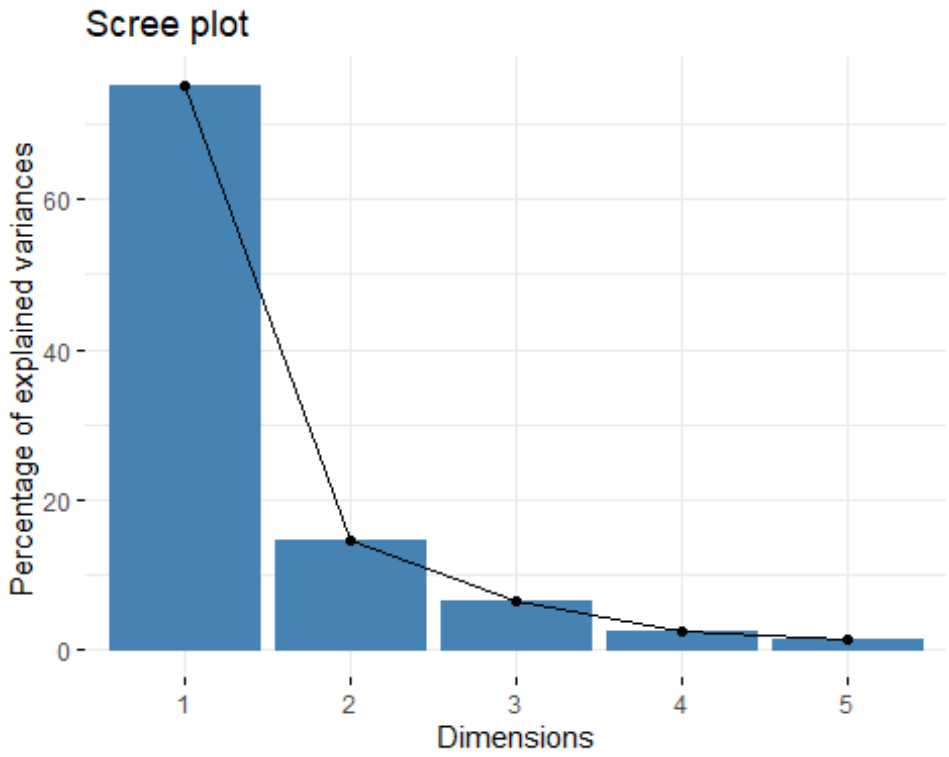
```
fviz_pca_var(cpR, col.var = "red", addEllipses = TRUE, repel = TRUE)
```



Este gráfico muestra cómo las variables contribuyen a los componentes principales. Se observa que "peso", "bíceps", "muñeca" tienen la mayor correlación con el primer componente (Dim1), ya que sus flechas son más largas en esta dimensión. "Edad" contribuye más a Dim2, lo que sugiere que en el segundo componente, las diferencias de edad capturan variabilidad adicional que no fue explicada por Dim1. Las flechas cercanas a la línea horizontal indican que estas variables están relacionadas positivamente con Dim1.

1.3. Scree plot (gráfico de sedimentación):

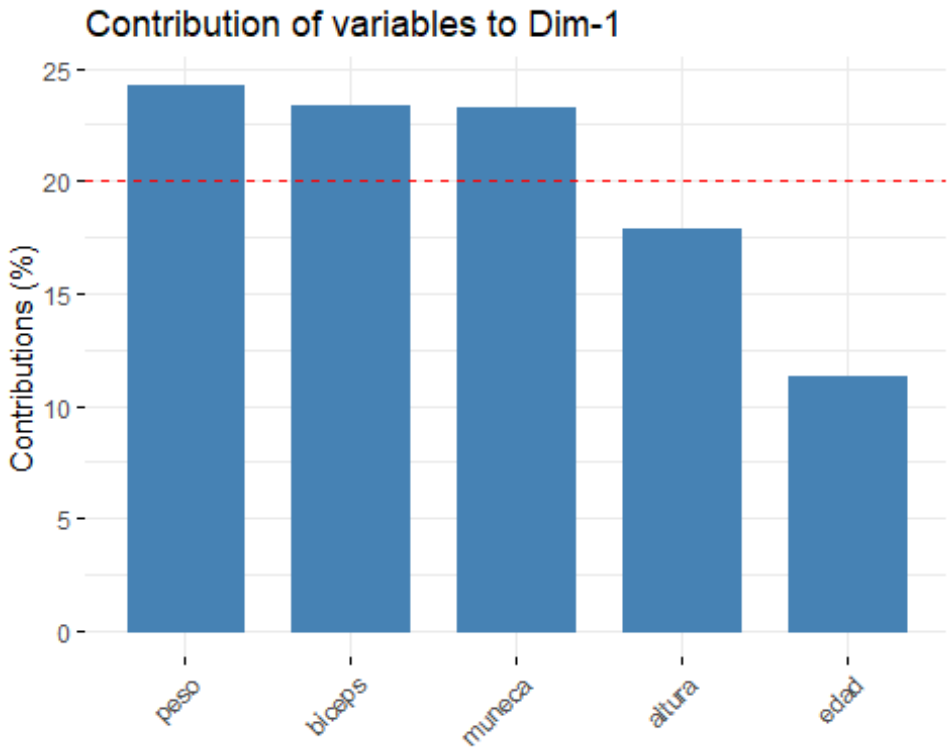
```
fviz_screeplot(cpR)
```



El gráfico muestra que el primer componente (Dim1) explica el 75.1% de la varianza, mientras que el segundo componente (Dim2) captura el 14.5%, lo que sugiere que la mayor parte de la variabilidad se puede reducir a dos dimensiones. Los componentes adicionales (Dim3, Dim4, y Dim5) explican muy poca varianza adicional, lo que justifica enfocarse solo en los primeros dos componentes para la reducción de dimensionalidad.

1.4. Contribución de las variables a cada componente principal:

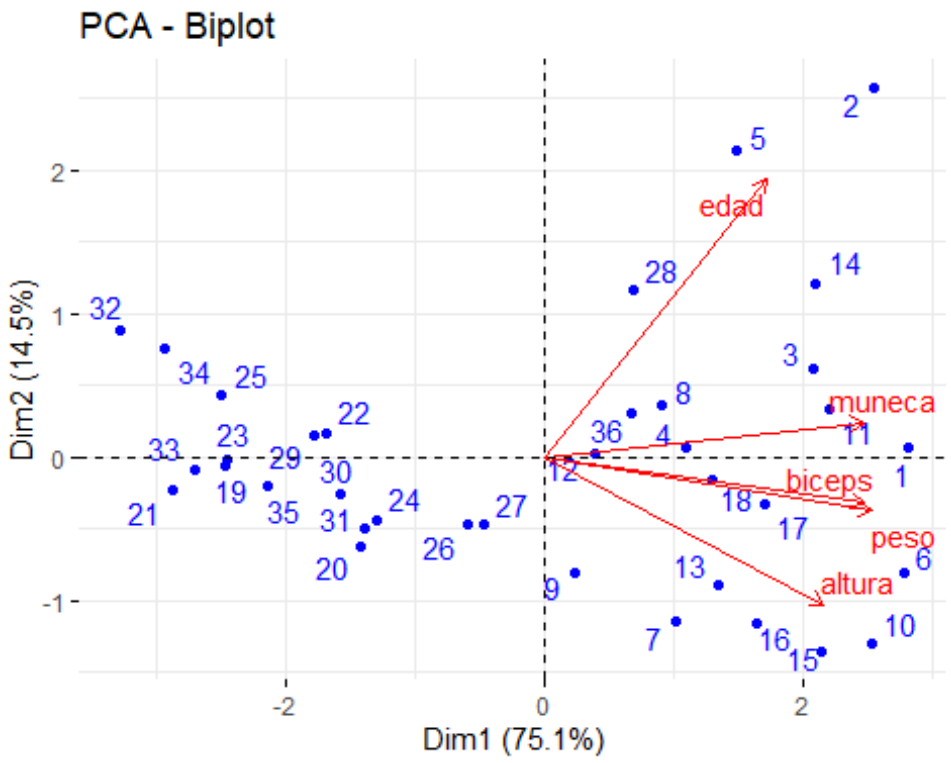
```
fviz_contrib(cpR, choice = "var")
```



El gráfico de contribución muestra que “peso” (>25%), “bíceps” (aprox 24%), y “muñeca” (aprox 24%) son las variables que más contribuyen a la varianza capturada en Dim1. Estas tres variables juntas representan más del 70% de la variabilidad en el primer componente, lo que sugiere que las diferencias en estas medidas corporales son las más importantes en este conjunto de datos. “Altura” (17.6 aprox %) y “edad” (10.7%) también contribuyen, pero en menor medida.

1.5. Biplot de individuos y variables:

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



En el biplot se pueden observar tanto las proyecciones de los individuos como de las variables. Las flechas de las variables indican la dirección y magnitud de su contribución a cada componente. Individuos cercanos a una flecha, como el individuo 2 cercano a la flecha de “edad”, tienen altos valores en esa variable. Los individuos 13,15,10,16…etc, por ejemplo, se encuentran más relacionados con las variables “altura” mientras que 18,17,7 con biceps y peso, lo que sugiere que tienen valores altos en estas características.

PARTE IV

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

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?

*Matriz de varianza-covarianza: Este análisis captura la varianza en términos absolutos, lo que significa que las variables con mayores magnitudes, como el peso y la altura, tienen un impacto más significativo en los componentes. Los resultados muestran que CP1 captura el 76.2% de la variabilidad y CP2 el 17%, siendo el peso la variable dominante en la primera componente.

*Matriz de correlación: Aquí las variables son estandarizadas, eliminando el sesgo causado por las diferentes unidades de medida. Los resultados muestran que CP1 captura el 75.1%

de la variabilidad y CP2 el 14.5%, con peso, bíceps y muñeca como las variables más importantes en CP1. Este enfoque es útil cuando las variables están en diferentes escalas.

** Conclusión sobre qué procedimiento aporta componentes de mayor interés: El análisis de la matriz de correlación parece más robusto en este caso, ya que elimina el sesgo causado por las diferencias en las escalas de las variables. Al estandarizar las variables, el análisis de correlación ofrece una visión más equilibrada y generalizable de la relación entre las variables, particularmente cuando se comparan variables con unidades muy diferentes.

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 datos socioeconómicos de 96 países, el análisis basado en la matriz de correlación sería más adecuado. En este tipo de datos, las variables suelen estar en diferentes escalas (por ejemplo, PIB per cápita, índice de alfabetización, mortalidad infantil), por lo que estandarizar las variables evitaría que las que tienen mayores varianzas dominen el análisis. Esto permite captar patrones más equitativos entre las diferentes dimensiones de los indicadores económicos y sociales.

¿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)

*Matriz de varianza-covarianza (METODO DESCARTADO):

**CP1: Las variables que más contribuyen son peso (76.2%), seguido de altura y bíceps. Esto muestra que las diferencias en el peso y altura explican la mayor parte de la varianza en este conjunto de datos.

**CP2: La edad es la variable que más contribuye en este componente (principalmente en CP2), lo que indica que la variabilidad restante se explica por diferencias en la edad.

*Matriz de correlación (METODO SELECCIONADO):

**CP1: Las variables que más contribuyen son peso (25%), bíceps (24.8%) y muñeca (24.1%). Esto resalta la importancia de las medidas corporales.

**CP2: La edad es nuevamente la variable dominante en CP2.

En cuestion con los componentes al observar la primera dimension que explica aprox el 76% de la variabilidad de los datos. Dado que el metodo seleccionado es el de correlacion podemos ver la distribucion en el biplot y observamos una tendencia donde las agrupaciones con las complexiones fisicas propias del individuo como ” peso, biceps, muneca, asi como la altura guardan un cierto grado de correlacion entre si por la linea observada en la direccion.

Mientras que la edad predomina en el Componente 2, y aunque no cumple un sentido de ortogonalidad con respecto de las otras variables para descartar completamente una correlacion, si vemos que se podria encontra a un grado mayor que con respecto de las otras, podriamos calcular el producto punto para observar el efecto, pero parece ser un angulo >55 o >60 grados sino es que mas, lo que esta lejos incluso de un paralelimos notable que podemos notar entre peso y biceps por ejemplo lo cual indica una akta correlacion significativa

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

METODO SELECCIONADO:

$$\text{CP1} = -0.33593 \times \text{edad} + -0.49271 \times \text{peso} + -0.42224 \times \text{altura} + -0.48219 \times \text{muñeca} + -0.48331 \times \text{bíceps}$$

$$\text{CP2} = 0.85756 \times \text{edad} + -0.16478 \times \text{peso} + -0.45422 \times \text{altura} + 0.10828 \times \text{muñeca} + -0.13927 \times \text{bíceps}$$

Interpreta los resultados en término de agrupación de variables (puede ayudar “índice de riqueza”, “índice de ruralidad”, etc)

A partir de los componentes principales, podemos interpretar que las variables peso, bíceps y muñeca están estrechamente relacionadas y podrían formar un “índice de masa corporal” o un índice de medidas físicas. Estos tres factores explican la mayor parte de la variabilidad en CP1, lo que sugiere que las diferencias en el tamaño corporal son las más relevantes en este conjunto de datos.

En cuanto a CP2, la variable edad se destaca significativamente, lo que podría asociarse con un índice de envejecimiento o de cambio generacional, explicando la variabilidad adicional que no es capturada por las medidas físicas.