A4-Componentes Principales

Eryk Elizondo González A01284899

2024-10-08

```
D <- read.csv("corporal.csv")
head(D)
```

```
##   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
```

Parte 1

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:

```
COV <- cov(D)
COR <- cor(D)

cat("Matriz de Covarianza\n")
```

```
## Matriz de Covarianza
```

```
COV
```

```
##              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
```

```
cat("\n")

cat("Matriz de Correlación\n")
```

```
## Matriz de Correlación
```

```
COR
```

```
##             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().

```
COV_eigen <- eigen(COV)
COR_eigen <- eigen(COR)

cat("Valores Propios de Matriz de Covarianza\n")
```

```
## Valores Propios de Matriz de Covarianza
```

```
COV_eigen
```

```
## eigen() decomposition
## $values
## [1] 359.3980243  80.3757858  27.6229011   4.3074318   0.2343571
##
## $vectors
##             [,1]       [,2]        [,3]         [,4]         [,5]
## [1,] -0.34871002  0.9075501 -0.23248825 -0.001589466  0.026473941
## [2,] -0.76617586 -0.1616581  0.52166894 -0.338508602  0.010707863
## [3,] -0.47632405 -0.3851755 -0.78905759  0.046160807  0.003543154
## [4,] -0.05386189  0.0155423  0.02785902  0.126103480 -0.990039959
## [5,] -0.24817367 -0.0402221  0.22455005  0.931330496  0.137814357
```

```
cat("\n")

cat("Valores Propios de Matriz de Correlación\n")
```

```
## Valores Propios de Matriz de Correlación
```

```
COR_eigen
```

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

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.

```
COV_eigen_prop <- COV_eigen$values/sum(diag(COV))
COR_eigen_prop <- COR_eigen$values/sum(diag(COR))

cat("Proporcion de la Varianza por Componente de Matriz de Covarianza\n")

## Proporcion de la Varianza por Componente de Matriz de Covarianza

COV_eigen_prop

## [1] 0.7615357176 0.1703098726 0.0585307219 0.0091271040 0.0004965839

cat("\n")

cat("Proporcion de la Varianza por Componente de Matriz de Correlación\n")

## Proporcion de la Varianza por Componente de Matriz de Correlación

COR_eigen_prop

## [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.

```
COV_eigen_prop_acum <- cumsum(COV_eigen_prop)
COR_eigen_prop_acum <- cumsum(COR_eigen_prop)

cat("Proporcion Acumulada de la Varianza por Componente de Matriz de
Covarianza\n")

## Proporcion Acumulada de la Varianza por Componente de Matriz de Covarianza

COV_eigen_prop_acum

## [1] 0.7615357 0.9318456 0.9903763 0.9995034 1.0000000

cat("\n")

cat("Proporcion Acumulada de la Varianza por Componente de Matriz de
Correlación\n")
```

```
## Proporcion Acumulada de la Varianza por Componente de Matriz de
Correlación

COR_eigen_prop_acum

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

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

Los componentes más importantes en la matriz de Covarianza son los primeros 2 que explican el 93.2% de la varianza.

Los componentes más importantes en la matriz de Correlación son los primeros 2 que explican el 89.7% de la varianza.

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.

```
cat("Ecuación de Componentes Principales para Matriz de Covarianza\n")

## Ecuación de Componentes Principales para Matriz de Covarianza

cat("CP1 = -", abs(COV_eigen$vectors[1,1]), "* edad +",
COV_eigen$vectors[1,2], "* peso -", abs(COV_eigen$vectors[1,3]), "* altura -
", abs(COV_eigen$vectors[1,4]), "* muneca +", COV_eigen$vectors[1,5], "*
biceps\n")

## CP1 = - 0.34871 * edad + 0.9075501 * peso - 0.2324883 * altura -
0.001589466 * muneca + 0.02647394 * biceps

cat("CP2 = -", abs(COV_eigen$vectors[2,1]), "* edad -",
abs(COV_eigen$vectors[2,2]), "* peso +", COV_eigen$vectors[2,3], "* altura -
", abs(COV_eigen$vectors[2,4]), "* muneca +", COV_eigen$vectors[2,5], "*
biceps\n\n")

## CP2 = - 0.7661759 * edad - 0.1616581 * peso + 0.5216689 * altura -
0.3385086 * muneca + 0.01070786 * biceps
```

Las variables que más influyen en el primer componente son el peso, la edad y la altura respectivamente.

Las variables que más influyen en el segundo componente son la edad, la altura y la muneca respectivamente.

```
cat("Ecuación de Componentes Principales para Matriz de Correlación\n")
```

```
## Ecuación de Componentes Principales para Matriz de Correlación

cat("CP1 = -", abs(COR_eigen$vectors[1,1]), "* edad +",
COR_eigen$vectors[1,2], "* peso -", abs(COR_eigen$vectors[1,3]), "* altura -
", abs(COR_eigen$vectors[1,4]), "* muneca +", COR_eigen$vectors[1,5], "*
biceps\n")

## CP1 = - 0.335931 * edad + 0.8575601 * peso - 0.3491378 * altura -
0.1360111 * muneca + 0.1065123 * biceps

cat("CP2 = -", abs(COR_eigen$vectors[2,1]), "* edad -",
abs(COR_eigen$vectors[2,2]), "* peso +", COR_eigen$vectors[2,3], "* altura -
", abs(COR_eigen$vectors[2,4]), "* muneca -", abs(COR_eigen$vectors[2,5]), "*
biceps")

## CP2 = - 0.4927066 * edad - 0.1647821 * peso + 0.06924561 * altura -
0.5249533 * muneca - 0.6706087 * biceps
```

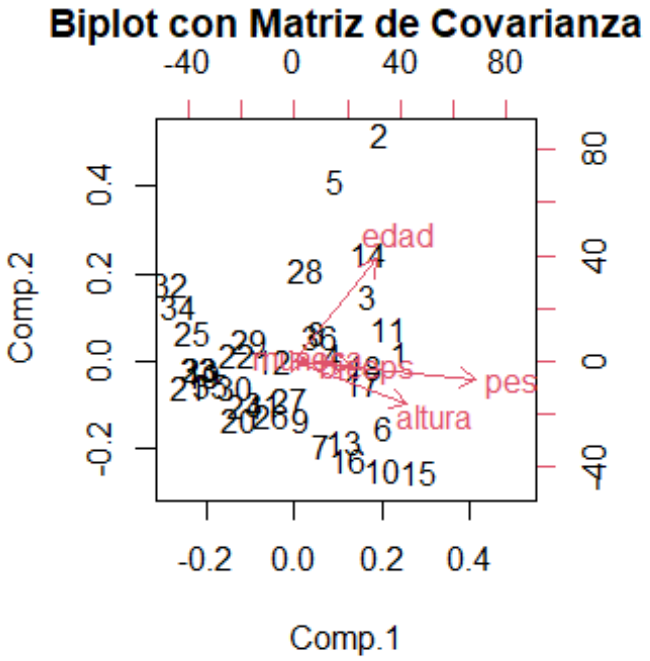
Las variables que más influyen en el primer componente son el peso, la altura y la edad respectivamente.

Las variables que más influyen en el segundo componente son los biceps, la muneca y la edad respectivamente.

Parte 2

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

```
cpS = princomp(D, cor=FALSE)
biplot(cpS, main="Biplot con Matriz de Covarianza")
```



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

La variable peso y altura parecen estar fuertemente correlacionada con el Componente 1, ya que las flechas están alineadas con este eje.

La variable edad tiene una mayor correlación con el Componente 2, ya que la flecha apunta hacia ese eje.

Las variables biceps y muneca parecen estar correlacionadas tanto con el Componente 1 como con el 2, pero con una influencia menor.

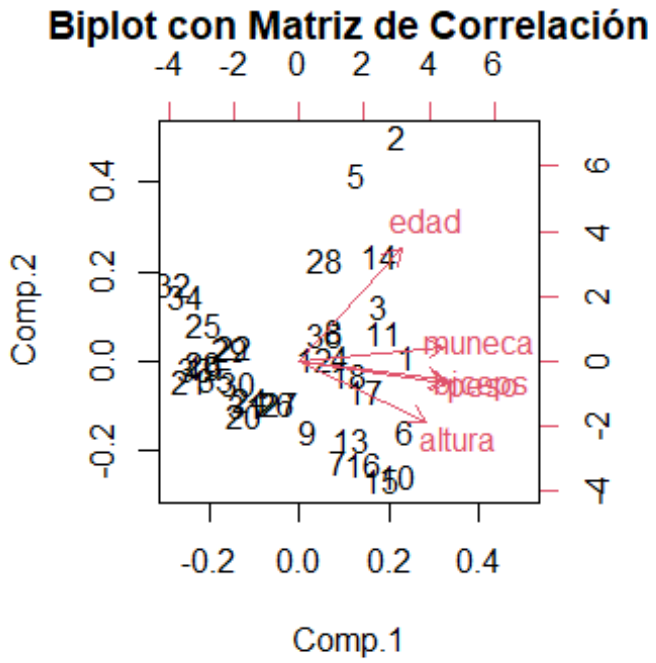
2. La relación entre las puntuaciones de las observaciones y los valores de las variables

Se puede apreciar que las observaciones tienden a ser opuestas a las variables de las que provienen, respaldado por los signos negativos en las equaciones anteriores.

3. Detecte posibles datos atípicos

Datos como el 2, 5 y 14 parecen estar alejados del agrupamiento general de las observaciones y podrían representar datos atípicos.

```
cpR = princomp(D, cor=TRUE)
biplot(cpR, main="Biplot con Matriz de Correlación")
```



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

Las variables muneca, biceps y altura parecen estar fuertemente correlacionada con el Componente 1, ya que las flechas están alineadas con este eje.

La variable edad tiene una mayor correlación con el Componente 2, ya que la flecha apunta hacia ese eje.

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

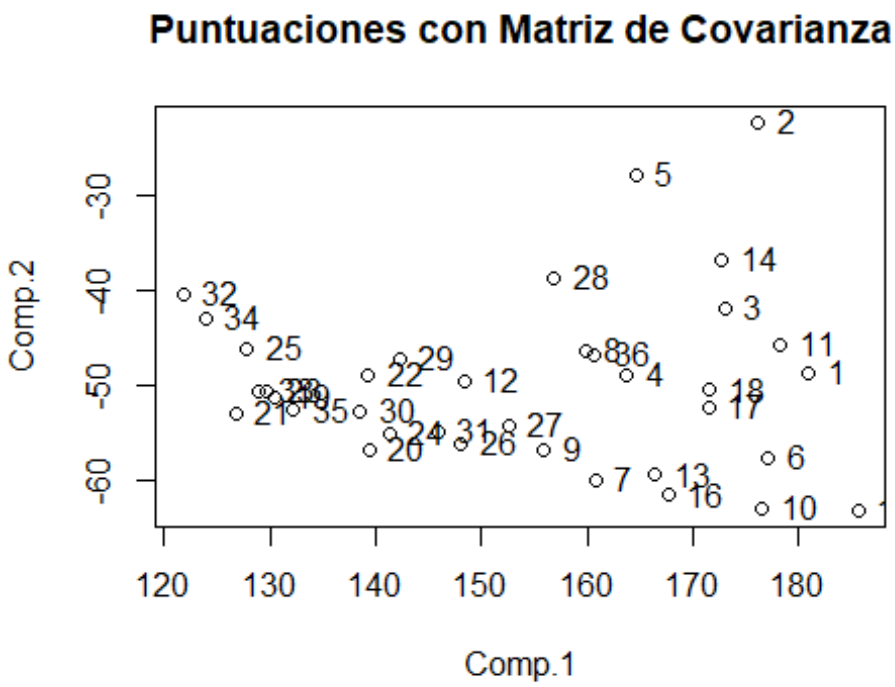
Se puede apreciar que las observaciones tienden a ser opuestas a las variables de las que provienen, respaldado por los signos negativos en las equaciones anteriores.

3. Detecte posibles datos atípicos

Datos como el 2, 5 y 14 parecen estar alejados del agrupamiento general de las observaciones y podrían representar datos atípicos.

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

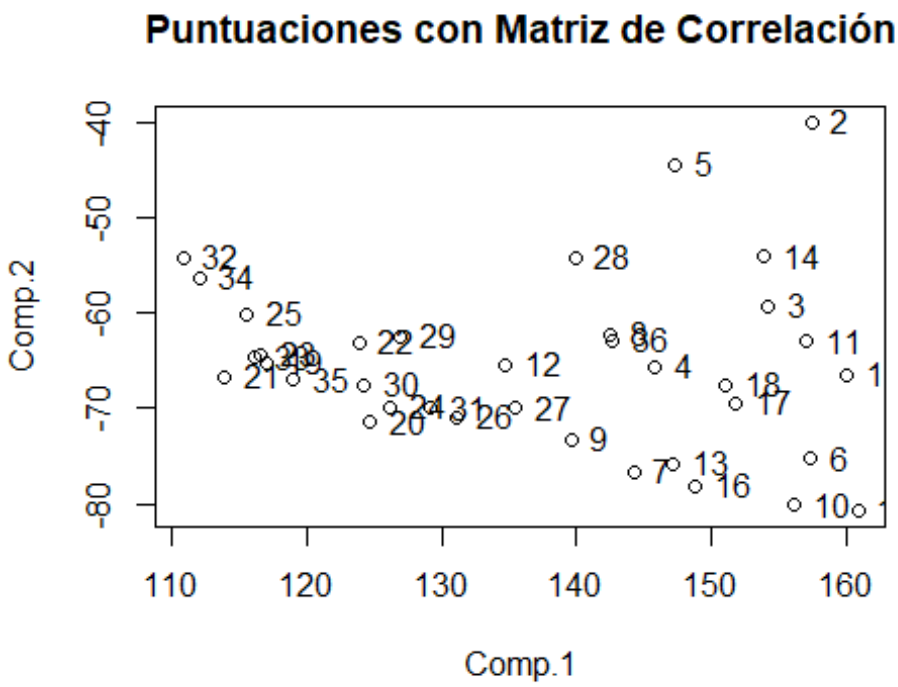
```
cpaS = as.matrix(D) %*% cpS$loadings
plot(cpaS[,1:2], main="Puntuaciones con Matriz de Covarianza")
text(cpaS[,1], cpaS[,2], labels=1:nrow(cpaS), pos=4)
```



De igual forma que en el anterior, los datos tienen a estar mpas aluneados con el eje del componente 1 en un rango positivo mientras que en el componente 2 en un rango negativo con datos atípicos como el 2, 5, 14 y 15.

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.

```
cpaR = as.matrix(D) %*% cpR$loadings
plot(cpaR[,1:2], main="Puntuaciones con Matriz de Correlación")
text(cpaR[,1], cpaR[,2], labels=1:nrow(cpaR), pos=4)
```



De forma similar a la matriz de covarianza, los datos atipicos son el 2, 5, 15 y ahora el 1 con los mismos límites de los ejes.

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?

summary(cpS): Muestra la varianza explicada por cada componente principal. Los primeros componentes explican la mayor parte de la variabilidad;

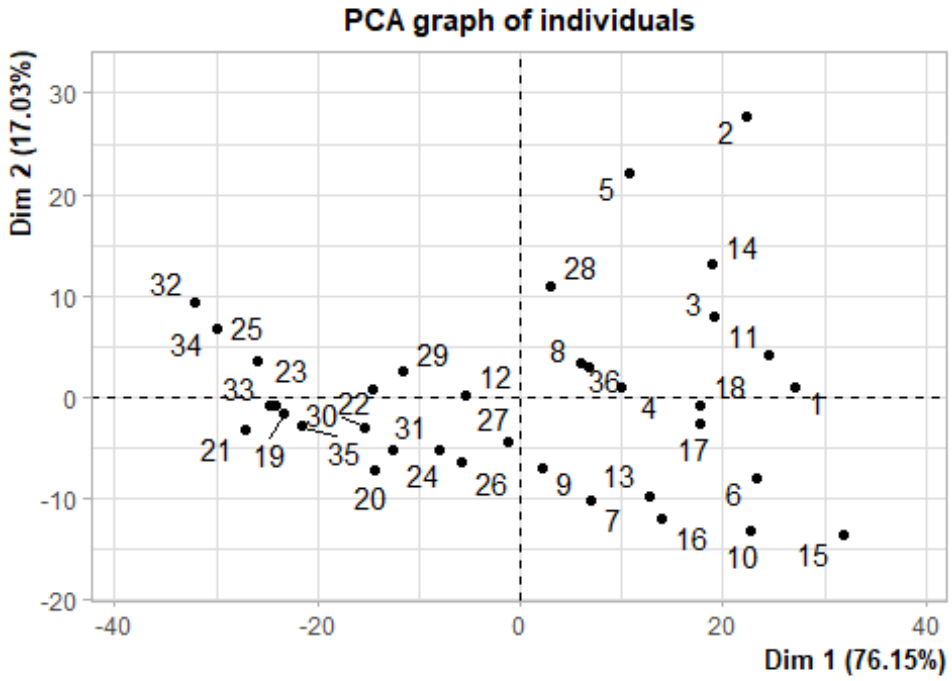
cpS$loadings: Muestra las cargas de las variables originales en cada componente principal. Las variables con mayores cargas (en valor absoluto) son las que más contribuyen a ese componente principal.

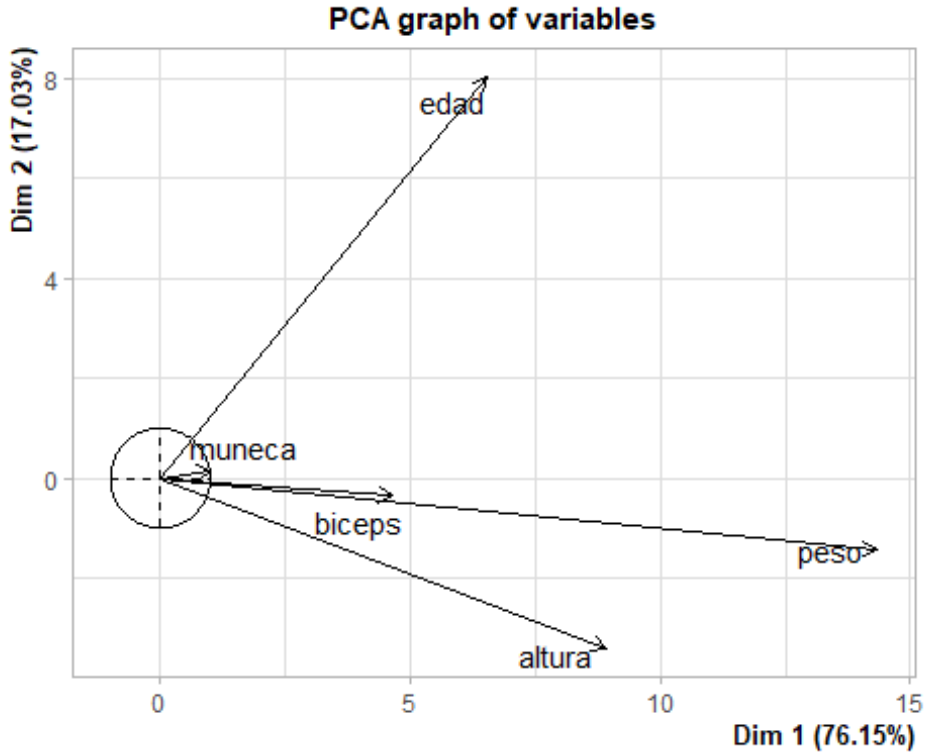
cpS$scores: Proporciona las puntuaciones de las observaciones en el espacio de los componentes principales. Representa las coordenadas de cada observación en los nuevos ejes de los componentes principales.

Parte 3

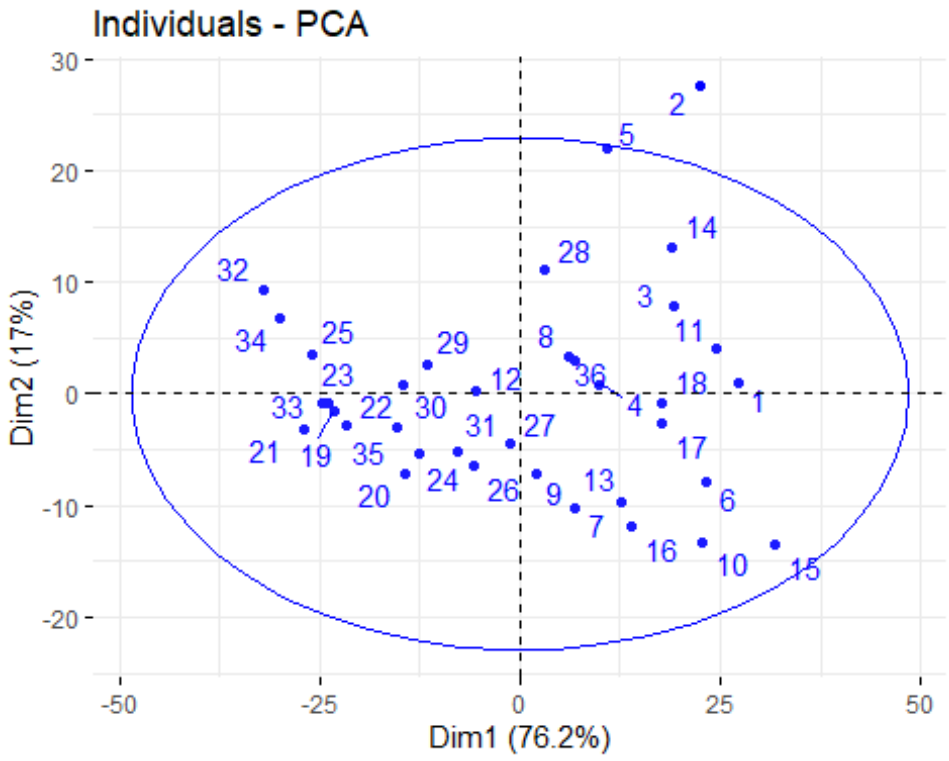
Explore los siguientes gráficos relativos a Componentes Principales. Interprete cada gráfico e identifica qué es lo que se está graficando en cada uno. Realiza el análisis con la matriz de varianzas y covarianzas y correlación.

```
cpS = PCA(D,scale.unit=FALSE) #Para matriz de correlaciones usa
scale.unit=TRUE
```

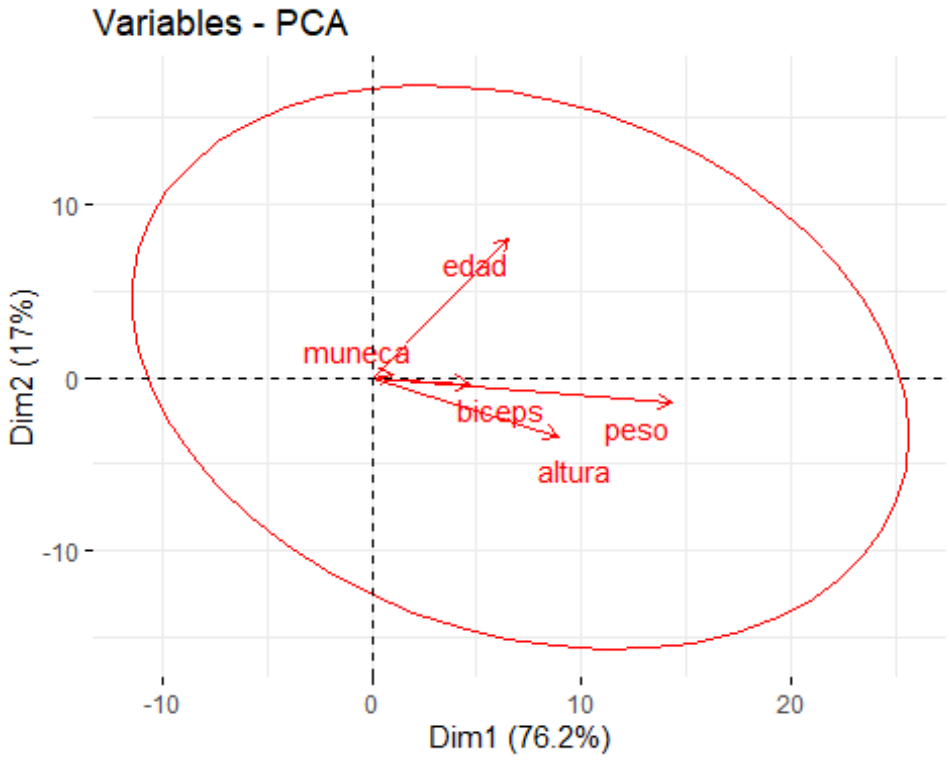





```r
fviz_pca_ind(cpS, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```




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



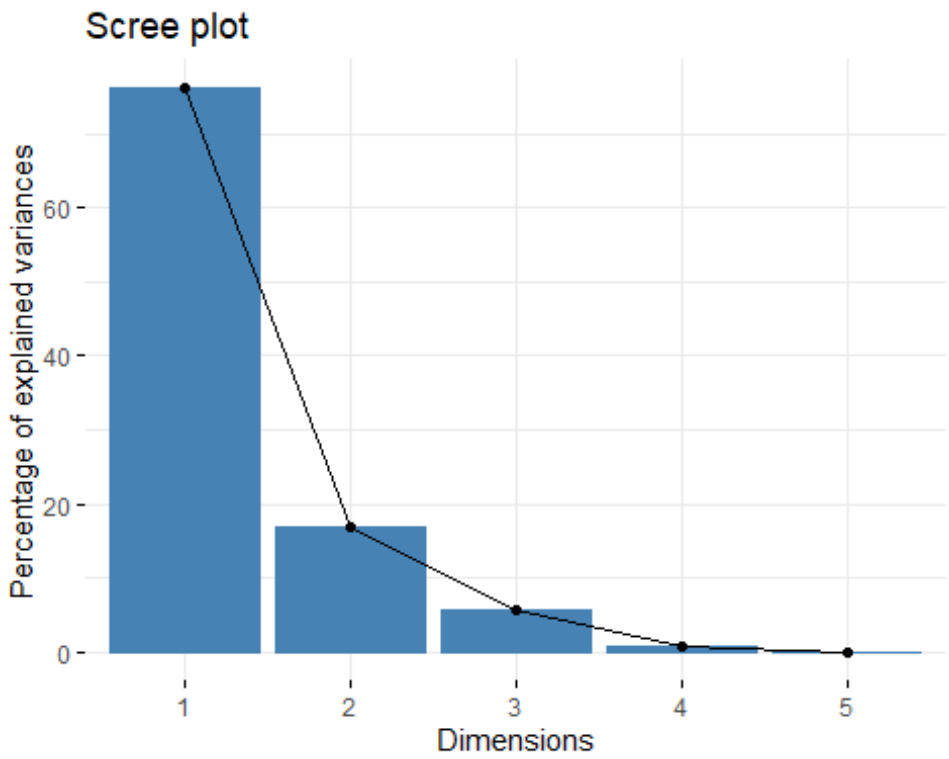
El gráfico de individuos muestra la distribución de las observaciones en el espacio de los componentes principales.

Se puede identificar que la observación 2 está fuera del rango de los componentes y muestra tendencias a favor de mayor variedad en el componente 1.

El gráfico de variables muestra las cargas de las variables en los componentes principales.

Muestra qye todas las variables tienen un peso positivo sobre el componente 1 y positivo y negativo más leve para el componente 2 que con la elipse muestra una tendencia al componente 1.

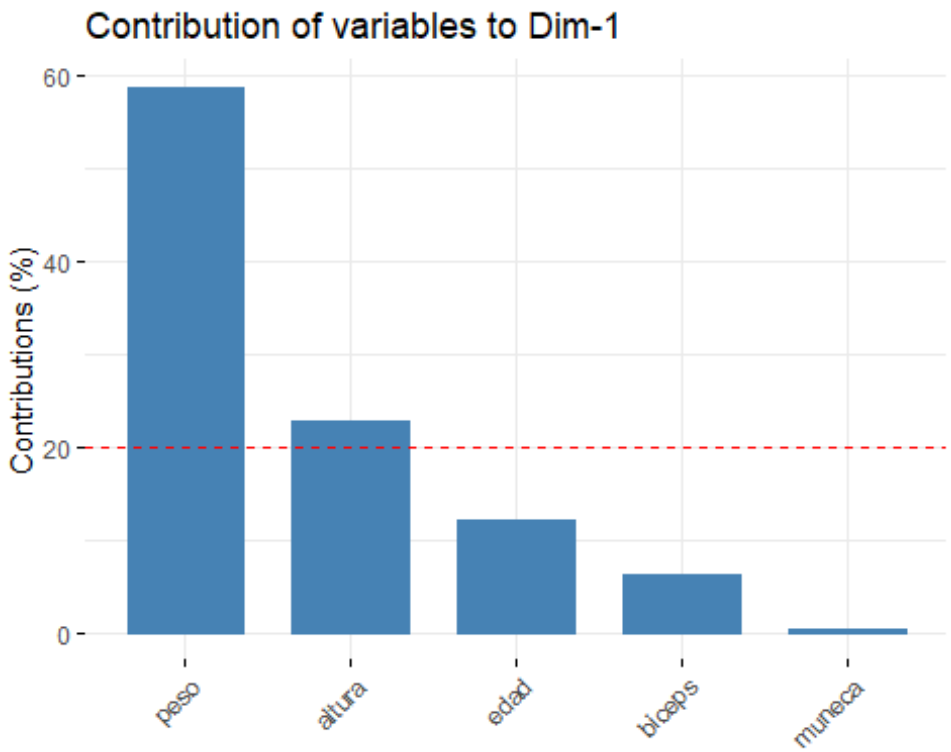
```
fviz_screeplot(cpS)
```



Muestra la varianza explicada por cada componente.

Donde la mayoría de la variación es explicada por el componente 1, seguido del componente 2 con una mucho menor variación explicada.

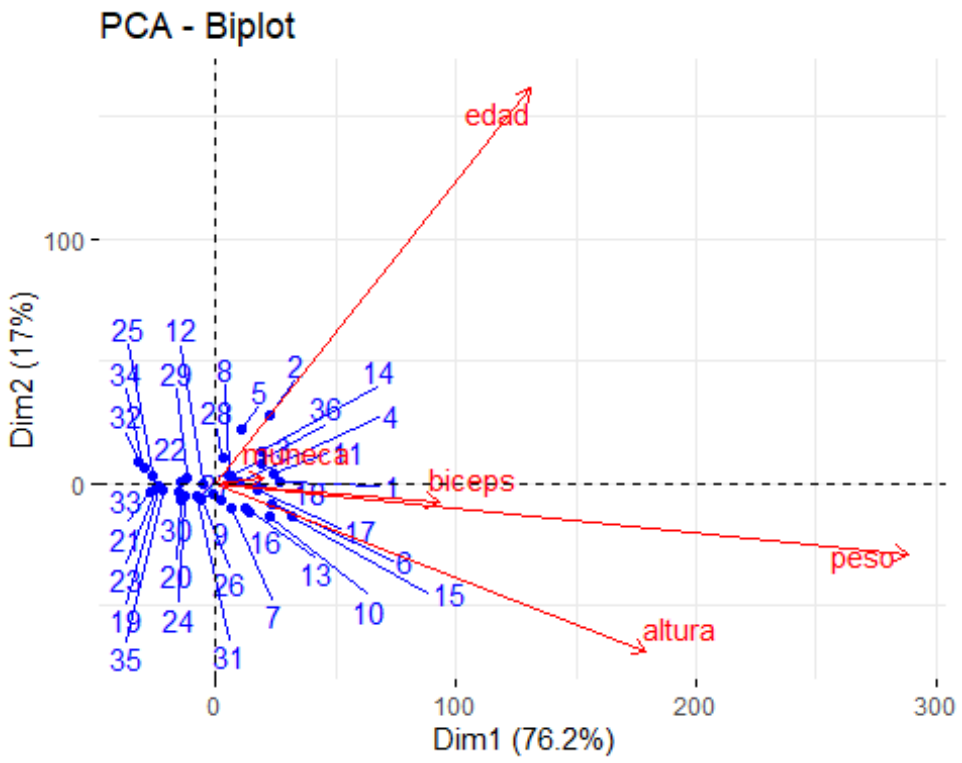
```
fviz_contrib(cpS, choice = c("var"))
```



Muestra la contribución de cada variable en los componentes seleccionados.

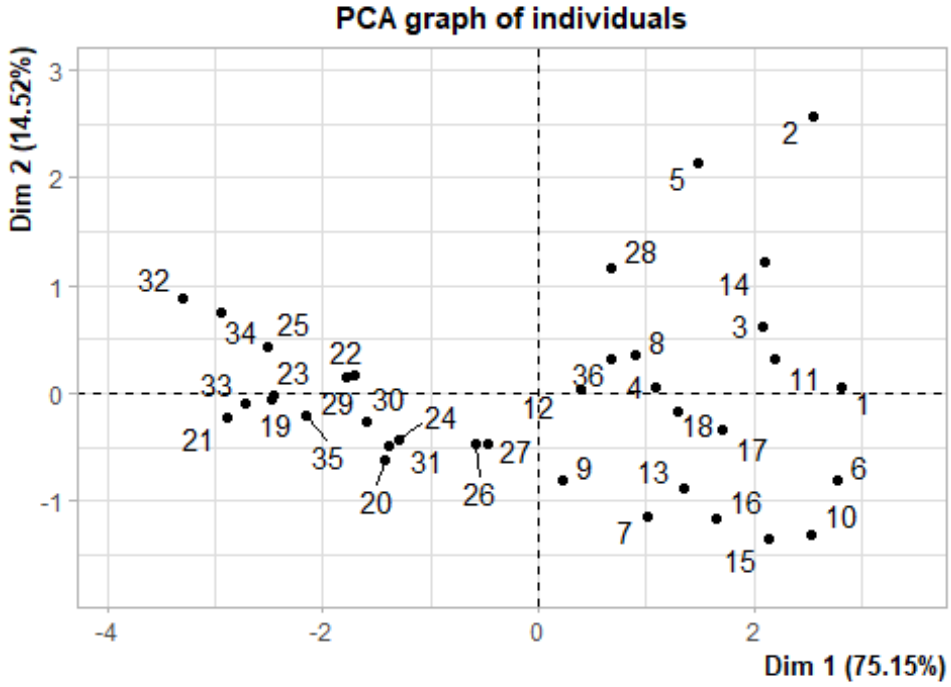
Se observa que la variable con mayor contribución al componente 1 es el peso seguido de la altura.

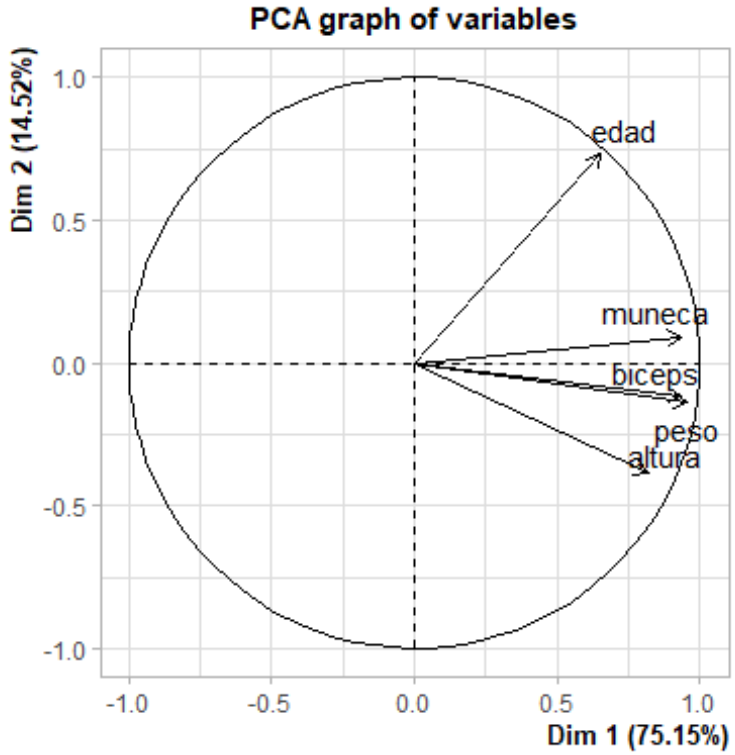
```
fviz_pca_biplot(cpS, repel=TRUE, col.var="red", col.ind="blue")
```



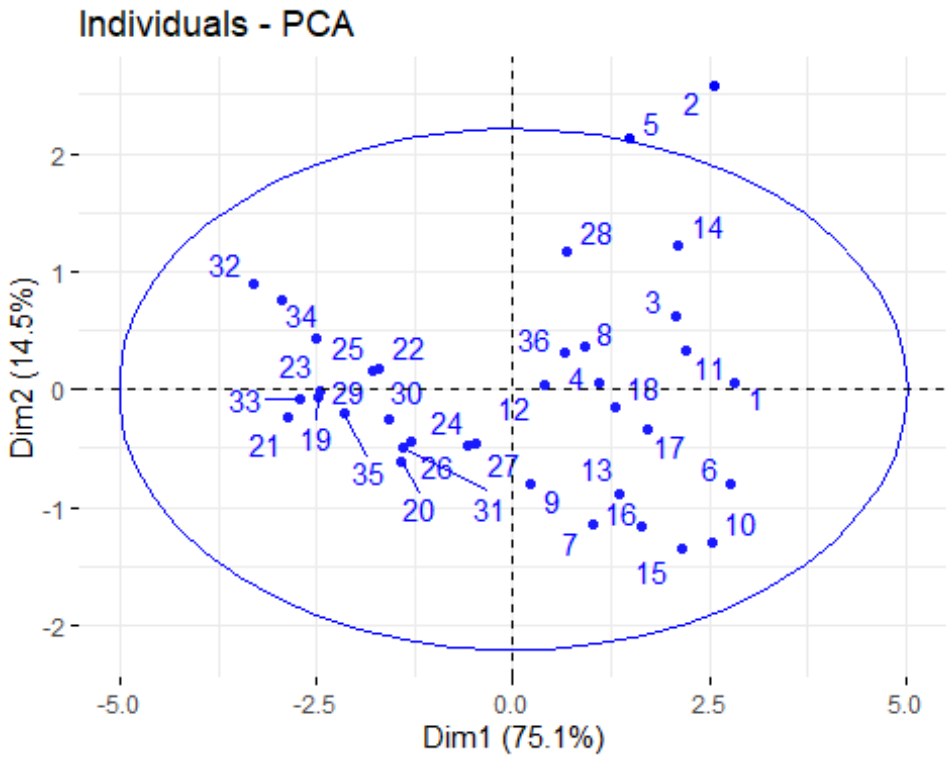
Éste gráfico combina el gráfico de individuos y variables en un solo gráfico y las interpretaciones serían iguales ya que sus tendencias se reflejan en este gráfico.

```
cpR = PCA(D,scale.unit=TRUE) #Para matriz de correlaciones usa
scale.unit=TRUE
```

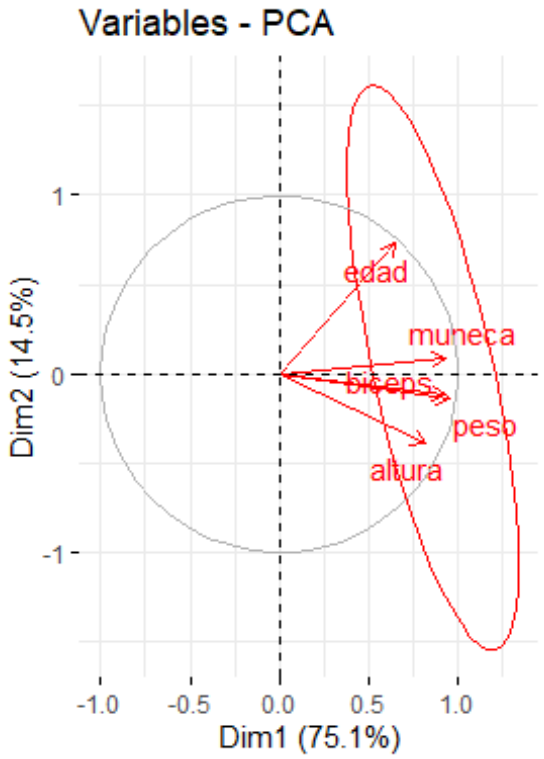





```
fviz_pca_ind(cpR, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```




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



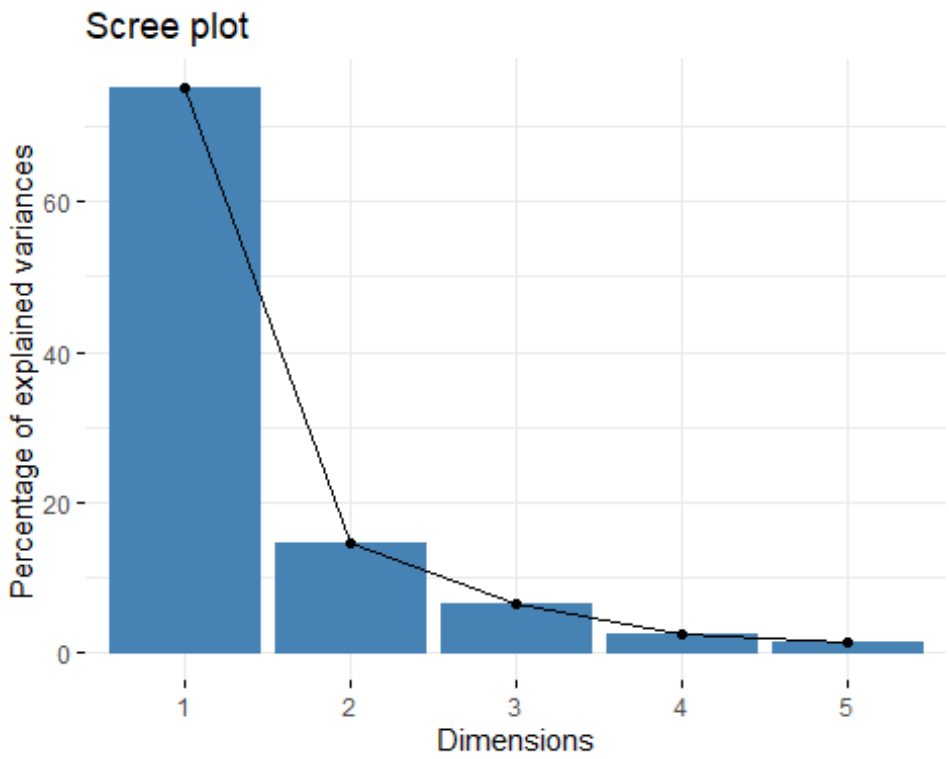
El gráfico de individuos muestra la distribución de las observaciones en el espacio de los componentes principales.

Se puede identificar que la observación 2 y 5 están afuera del rango de los componentes y muestra tendencias a favor de mayor variedad en el componente 1, sin embargo, con mayor variedad en el componente 2 que en el gráfico anterior.

El gráfico de variables muestra las cargas de las variables en los componentes principales.

Muestra que todas las variables tienen un peso positivo sobre el componente 1 y positivo y negativo más leve para el componente 2.

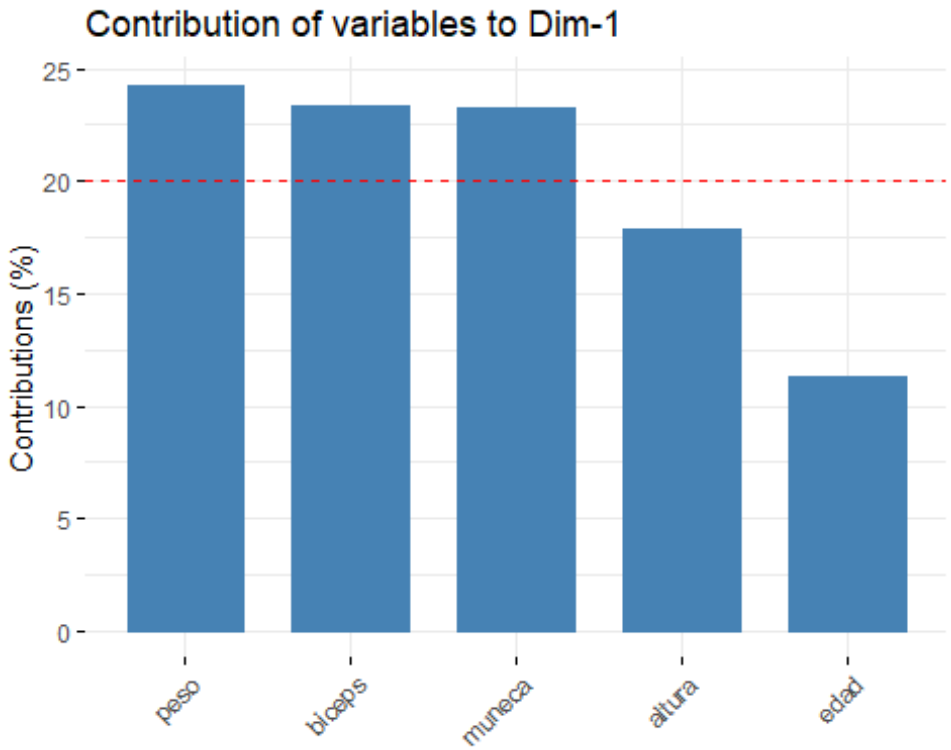
```
fviz_screeplot(cpR)
```



Muestra la varianza explicada por cada componente.

Donde la mayoría de la variación es explicada por el componente 1, seguido del componente 2 con una mucho menor variación explicada.

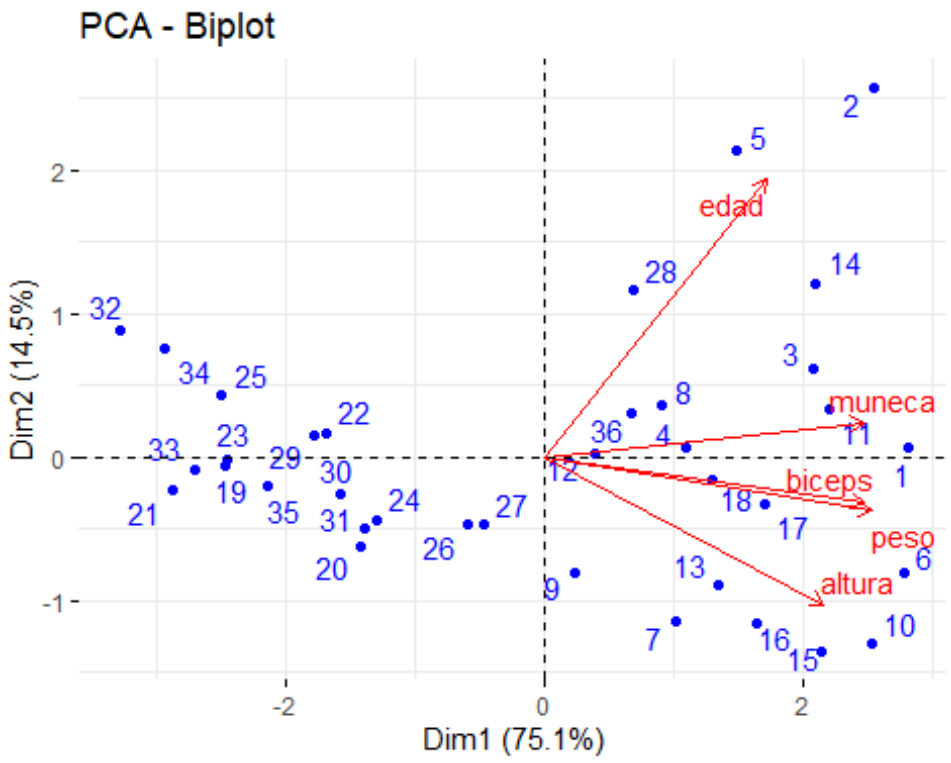
```
fviz_contrib(cpR, choice = c("var"))
```



Muestra la contribución de cada variable en los componentes seleccionados.

Se observa que as variables con mayor contribución al componente 1 son el peso seguido de biceps y muneca de forma similar.

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



Éste gráfico combina el gráfico de individuos y variables en un solo gráfico y las interpretaciones serían iguales ya que sus tendencias se reflejan en este gráfico.

3. Explora el comando PCA, (puedes poner help(PCA) en la consola o buscarlo en la ventana de ayuda) ¿qué otras opciones tiene para facilitarte el análisis?

summary(cpS): Muestra un resumen detallado de los componentes principales obtenidos.

cpSvarcoord: Devuelve las coordenadas de las variables en el espacio de los componentes principales.

cpSindcoord: Devuelve las coordenadas de los individuos (observaciones) en el espacio de los componentes principales.

Parte 4

Matriz de varianza-covarianza: Este método es adecuado cuando las unidades de medida de las variables son comparables.

Matriz de correlación: Este método es más adecuado cuando las variables tienen diferentes escalas, ya que las estandariza antes de realizar el análisis.

El análisis con la matriz de correlación aporta componentes más equilibrados y, por lo tanto, de mayor interés, ya que ajusta las variables a una escala común, permitiendo una comparación equitativa.

Las variables, como peso, altura, y edad, poseen diferentes unidades de medida. Usar la matriz de correlación elimina el problema de escala, permitiendo una interpretación más justa de la influencia de cada variable en el análisis.

Las variables que más influyen en el primer componente son el peso, la altura y la edad respectivamente.

Las variables que más influyen en el segundo componente son los biceps, la muneca y la edad respectivamente.

Las combinaciones finales que se recomiendan para hacer el análisis de componentes principales CP1 = - 0.335931 * edad + 0.8575601 * peso - 0.3491378 * altura - 0.1360111 * muneca + 0.1065123 * biceps CP2 = - 0.4927066 * edad - 0.1647821 * peso + 0.06924561 * altura - 0.5249533 * muneca - 0.6706087 * biceps

Podemos agrupar peso, biceps y muñeca ya que para la medida del BMI se necesitan peso que esta conectado con biceos y muñeca, altura y edad que cada uno está en su propia unidad y resulta en un análisis más preciso.