ComponentesPrincipales

Facundo Colasurdo Caldironi

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.

```
M=read.csv("file:///Users/facundocolasurdocaldironi/Downloads/corporal.csv") #leer la base de datos
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
```

##Haz un análisis descriptivo de los datos: medidas principales y gráficos caja bijotes de cada varIable y la desviacion estandar

```
summary(M)
```

```
##       edad            peso           altura          sexo
##  Min.   :19.00   Min.   :42.00   Min.   :147.2   Length:36
##  1st Qu.:24.75   1st Qu.:54.95   1st Qu.:164.8   Class :character
##  Median :28.00   Median :71.50   Median :172.7   Mode  :character
##  Mean   :31.44   Mean   :68.95   Mean   :171.6
##  3rd Qu.:37.00   3rd Qu.:82.40   3rd Qu.:179.4
##  Max.   :65.00   Max.   :98.20   Max.   :190.5
##      muneca          biceps
##  Min.   : 8.300   Min.   :23.50
##  1st Qu.: 9.475   1st Qu.:25.98
##  Median :10.650   Median :32.15
##  Mean   :10.467   Mean   :31.17
##  3rd Qu.:11.500   3rd Qu.:35.05
##  Max.   :12.400   Max.   :40.40
```

```
datos_sin_sexo <- M[ , -4]
head(datos_sin_sexo)
```

```
##   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
```

```
summary(datos_sin_sexo)
```

```
##       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
```

```
desviaciones <- apply(datos_sin_sexo, 2, sd)
desviaciones
```

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

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:

```
S <- cov(datos_sin_sexo)
R <- cor(datos_sin_sexo)

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
```

```
##             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().

```
eigen_S <- eigen(S)
valores_propios_S <- eigen_S$values
vectores_propios_S <- eigen_S$vectors

eigen_R <- eigen(R)
valores_propios_R <- eigen_R$values
vectores_propios_R <- eigen_R$vectors

valores_propios_S
```

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

```
vectores_propios_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_propios_R
```

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

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

```
varianza_total_S <- sum(valores_propios_S)
proporcion_varianza_S <- valores_propios_S / varianza_total_S
varianza_acumulada_S <- cumsum(proporcion_varianza_S)

varianza_total_R <- sum(valores_propios_R)
proporcion_varianza_R <- valores_propios_R / varianza_total_R
varianza_acumulada_R <- cumsum(proporcion_varianza_R)

proporcion_varianza_S
```

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

```
varianza_acumulada_S
```

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

```
proporcion_varianza_R
```

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

```
varianza_acumulada_R
```

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

Según los resultados anteriores, ¿qué componentes son los más importantes? Componente uno y componente dos, ya que como se puede ver en la varianza acumulada de S, estas logran explicar el 93% de los datos

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.

```
CP1_S <- vectores_propios_S[, 1]
CP2_S <- vectores_propios_S[, 2]
CP1_R <- vectores_propios_R[, 1]
CP2_R <- vectores_propios_R[, 2]

CP1_S
```

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

```
CP2_S
```

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

```
CP1_R
```

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

```
CP2_R
```

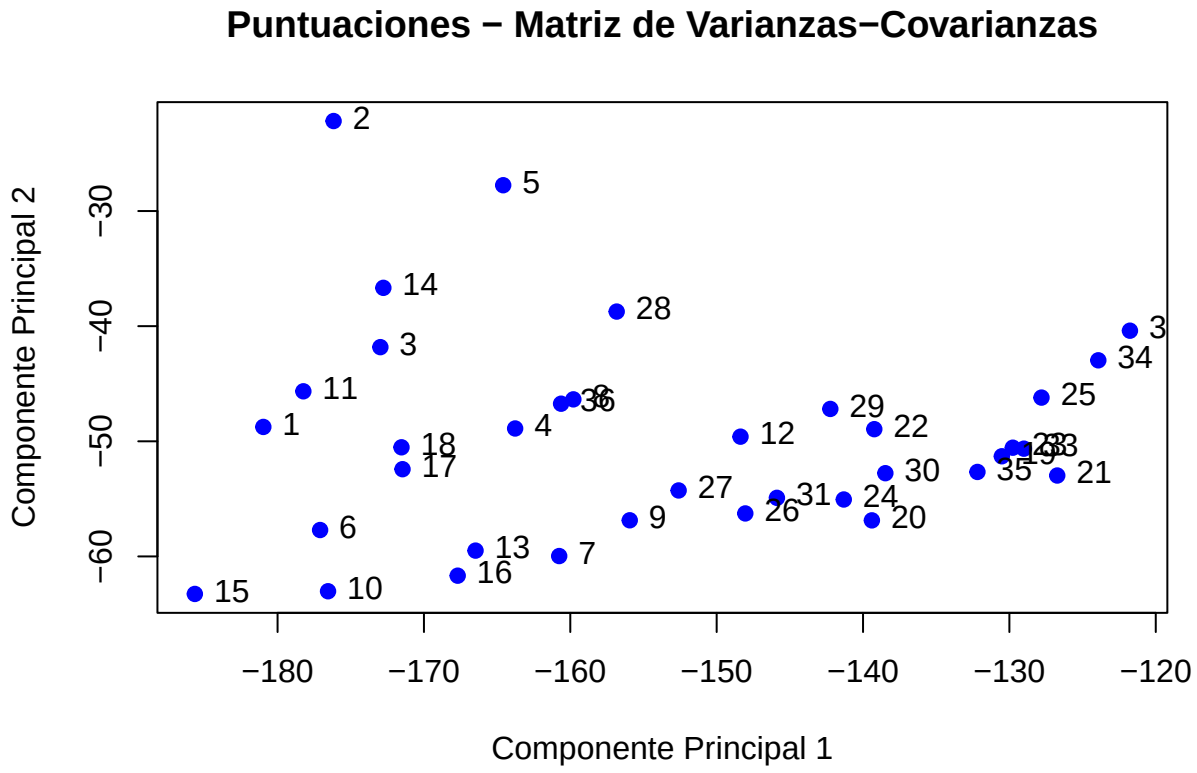
```
## [1]  0.8575601 -0.1647821 -0.4542223  0.1082775 -0.1392684
```

##PARTE II Obtenga las gráficas respectivas con S (matriz de varianzas-covarianzas) y con R (matriz de correlaciones) de las dos primeras componentes. Calcule las puntuaciones (scores) de las observaciones para los componentes obtenidos con la matriz de varianzas-covarianzas 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.

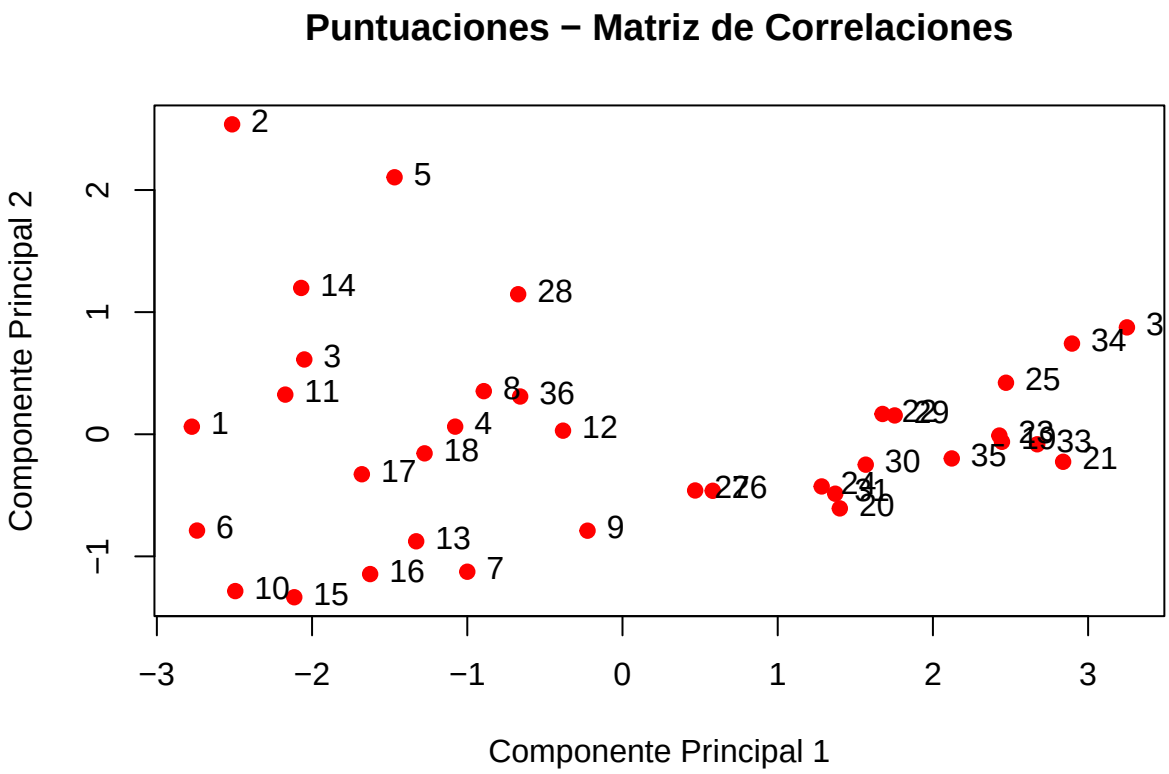
```r
scores_S <- as.matrix(datos_sin_sexo) %*% vectores_propios_S[, 1:2]

datos_estandarizados <- scale(datos_sin_sexo)
scores_R <- as.matrix(datos_estandarizados) %*% vectores_propios_R[, 1:2]

plot(scores_S, type = "p", col = "blue", pch = 19,
     xlab = "Componente Principal 1",
     ylab = "Componente Principal 2",
     main = "Puntuaciones - Matriz de Varianzas-Covarianzas")
text(scores_S[, 1], scores_S[, 2], labels = 1:nrow(scores_S), pos = 4)
```




```r
# Gráfica de las puntuaciones para la matriz de correlaciones
plot(scores_R, type = "p", col = "red", pch = 19,
     xlab = "Componente Principal 1",
     ylab = "Componente Principal 2",
     main = "Puntuaciones - Matriz de Correlaciones")
text(scores_R[, 1], scores_R[, 2], labels = 1:nrow(scores_R), pos = 4)
```



Interprete los gráficos en términos de: Las relaciones que se establecen entre las variables y los componentes principales La relación entre las puntuaciones de las observaciones y los valores de las variables

La gráfica de matriz correlación muestra cómo las observaciones se agrupan o se dispersan en función de las dos primeras componentes principales. Observaciones como la 2, que están muy alejadas del resto, deben examinarse para ver si representan datos atípicos o un grupo único dentro del conjunto de datos.

La gráfica de matriz de varianza covarianza indica que los primeros dos componentes principales capturan una buena cantidad de la varianza en los datos, permitiendo una diferenciación clara entre algunas observaciones (por ejemplo, observaciones 2 y 15 parecen ser bastante diferentes).

Detecte posibles datos atípicos 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. Sugerencias en R

library(stats) datos=matriz de datos cpS=princomp(datos,cor=FALSE) #Para la matriz de correlación usa cor=TRUE cpaS=as.matrix(datos)%*%cpS$loadings #Calcula las puntuaciones plot(cpaS[,1:2],type="p", main = "Título") text(cpaS[,1],cpaS[,2],1:nrow(cpaS)) biplot(cpS)

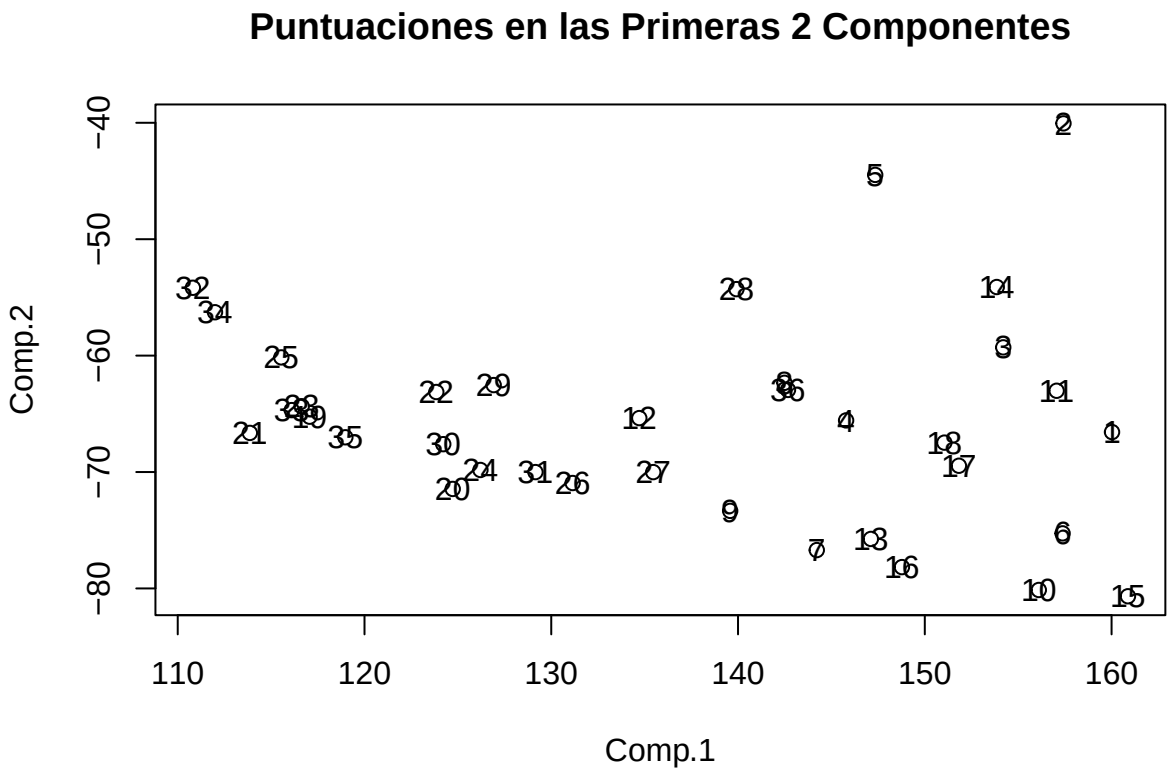
```
cpS <- princomp(datos_sin_sexo, cor = TRUE)
summary(cpS)
```

```
## 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
```

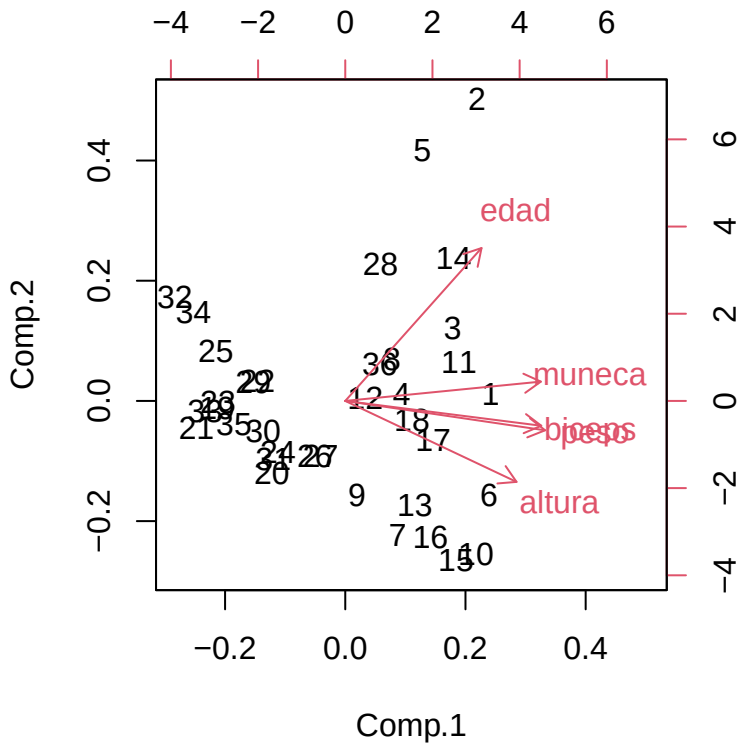
```r
cpS$loadings
```

```
##
## 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
```

```r
cpaS <- as.matrix(datos_sin_sexo) %*% cpS$loadings
plot(cpaS[, 1:2], type = "p", main = "Puntuaciones en las Primeras 2 Componentes")
text(cpaS[, 1], cpaS[, 2], labels = 1:nrow(cpaS))  # Etiquetar las observaciones
```




```r
biplot(cpS)
```



¿Cómo se interpreta el resultado?

Los resultados nos indican que las dos primeras componentes capturan la mayor parte de la variabilidad, con la primera componiendo una mezcla de todas las variables y la segunda dominada principalmente por la edad.

El gráfico de puntuaciones en las primeras dos componentes nos muestra cada individuo y la posición, su relación con los componentes, se puede observar que la mayoría se encuentran mas cercanos al primero de los dos componentes.

Por otro lado en el biplot es posible ver como edad, muñeca, bíceps y altura influyen en las dos primeras componentes principales, en donde se puede ver que las variables como muñeca y bíceps están correlacionadas y afectan principalmente al primer componente, mientras que la edad se asocia más con al segundo

#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) datos=matriz de datos cpS = PCA(datos,scale.unit=FALSE) #Para matriz de correlaciones usa scale.unit=TRUE library(factoextra) fviz_pca_ind(cpS, col.ind = “blue”, addEllipses = TRUE, repel = TRUE) fviz_pca_var(cpS, col.var = “red”, addEllipses = TRUE, repel = TRUE) fviz_screeplot(cpS) fviz_contrib(cpS, choice = c(“var”) fviz_pca_biplot(cpS, repel=TRUE, col.var=“red”, col.ind=“blue”)

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?

Otras opciones adicionales de FactoMineR afectan la visualizacion de los gráficos y agregar más detalles a los mismos, tales como etiquetas, elipses de confianza y colores.

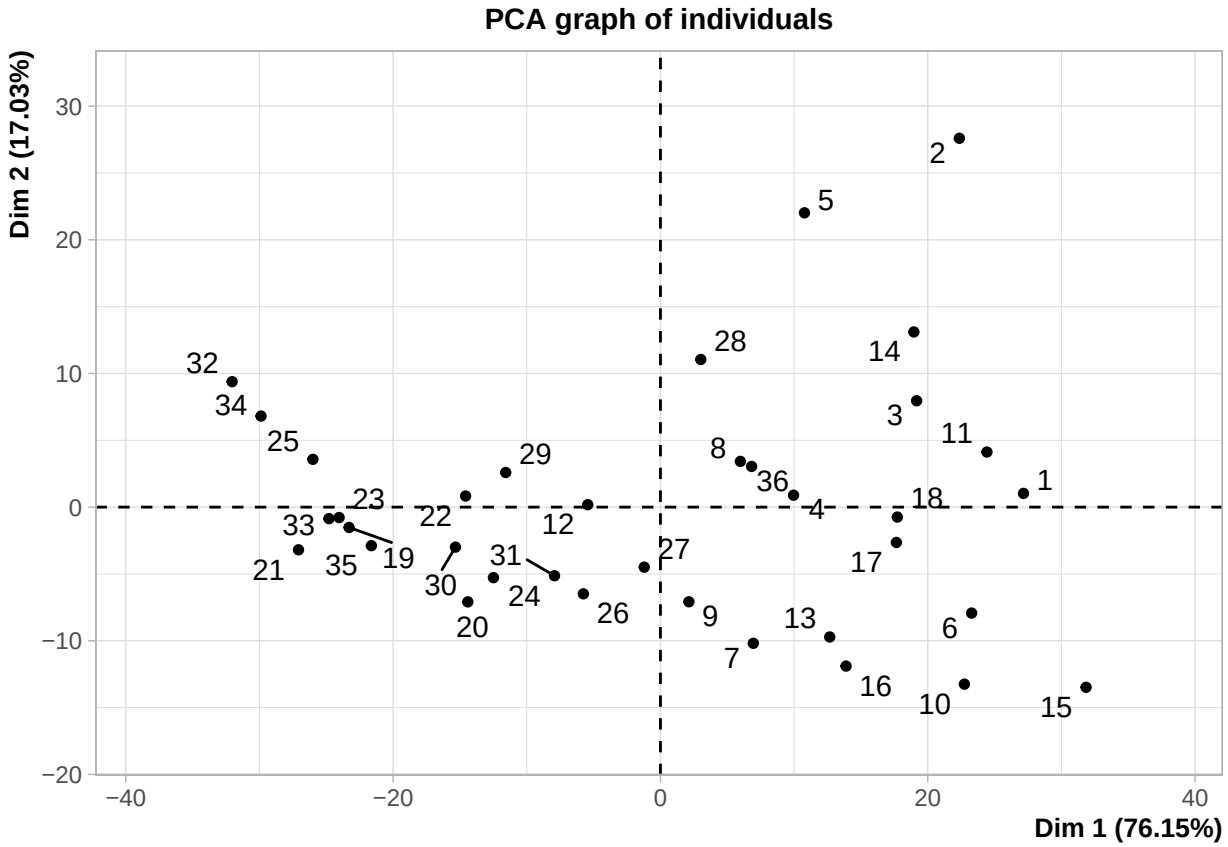
```
library(FactoMineR)
library(factoextra)
```

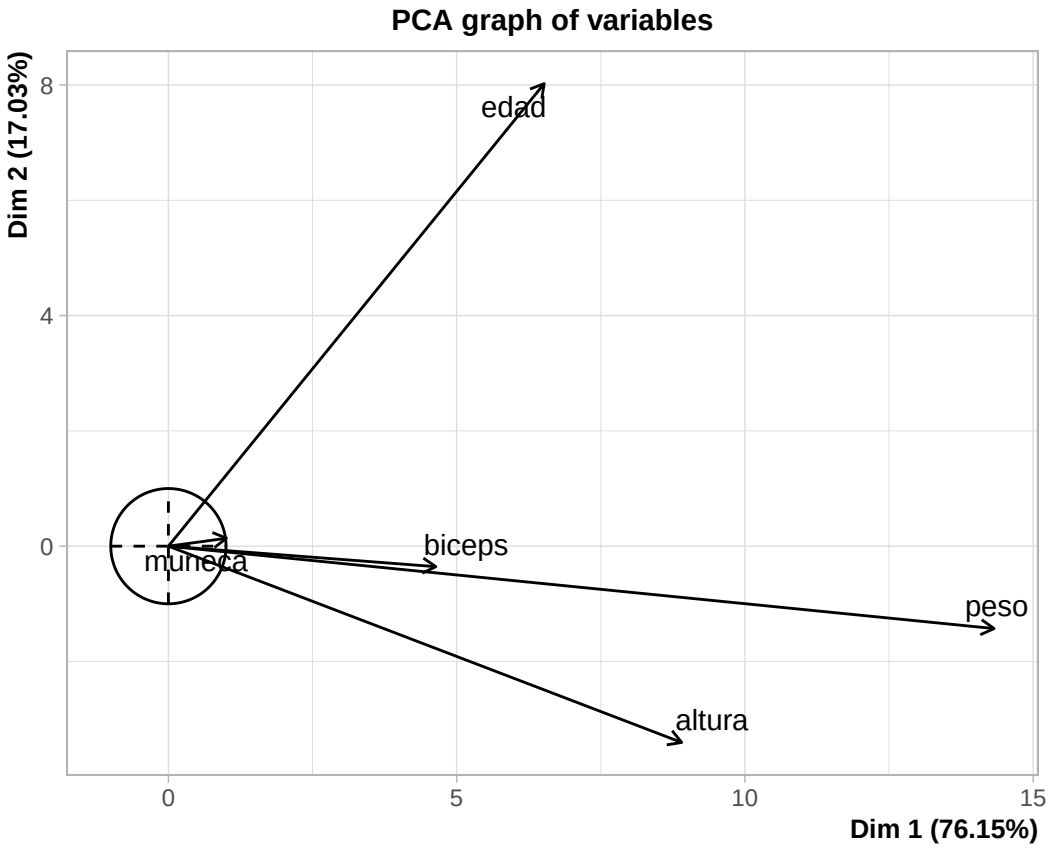
```
## Loading required package: ggplot2
```

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

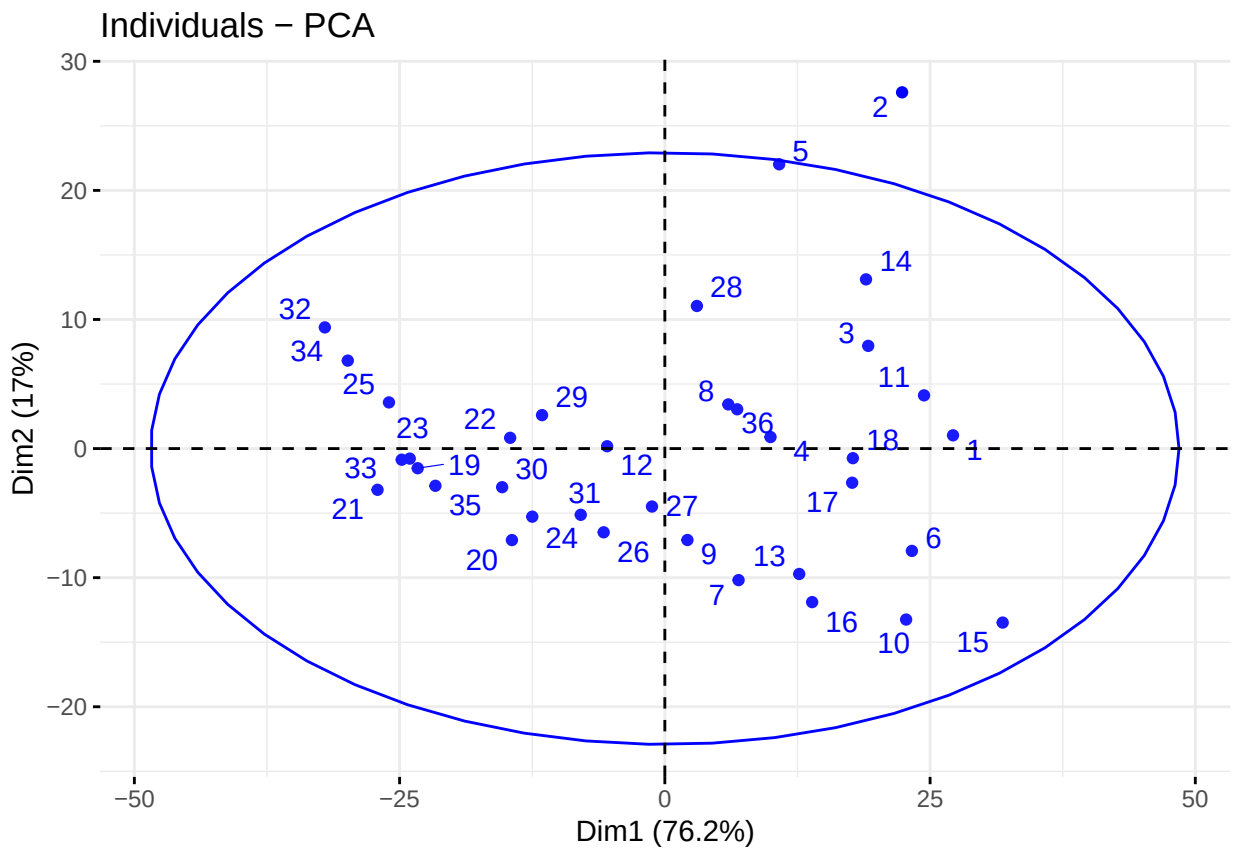
```
library(ggplot2)


cpS_varianza <- PCA(datos_sin_sexo, scale.unit = FALSE)
```

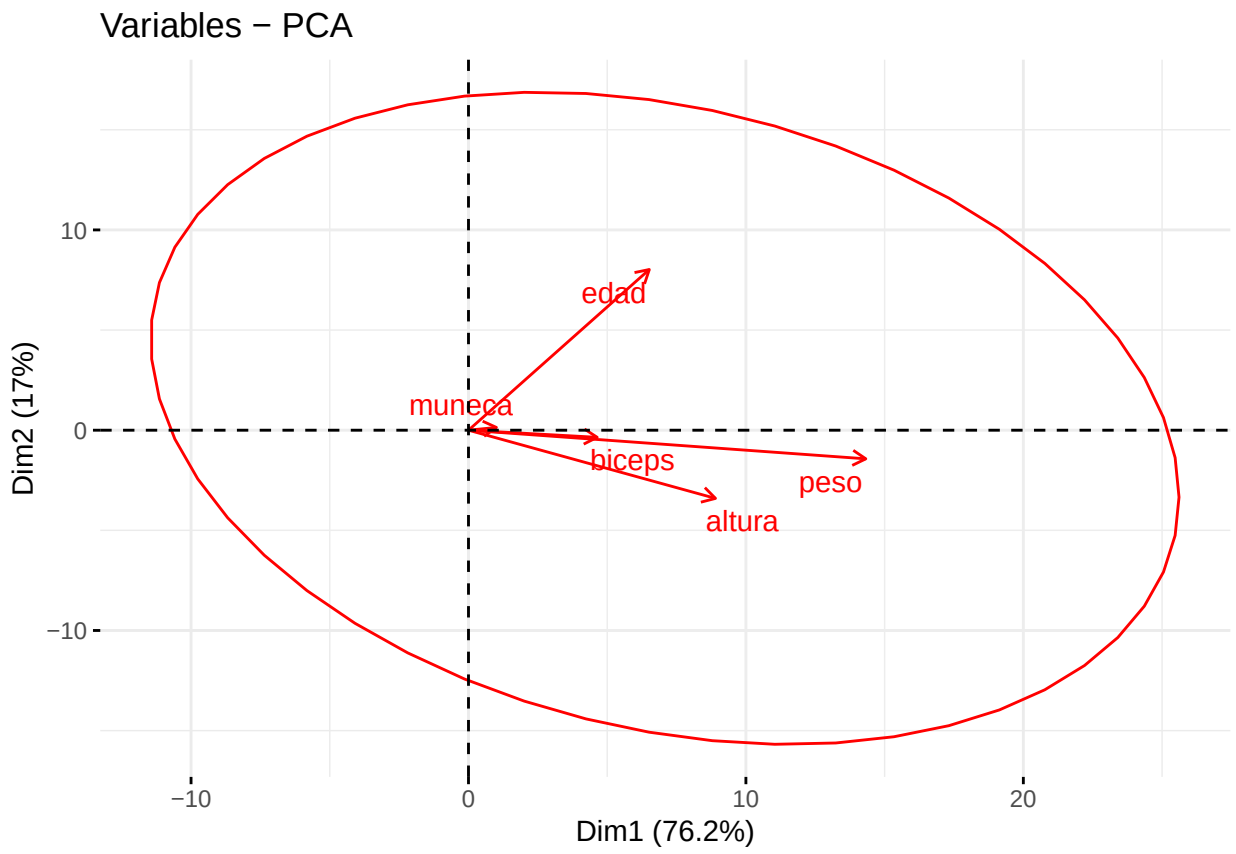




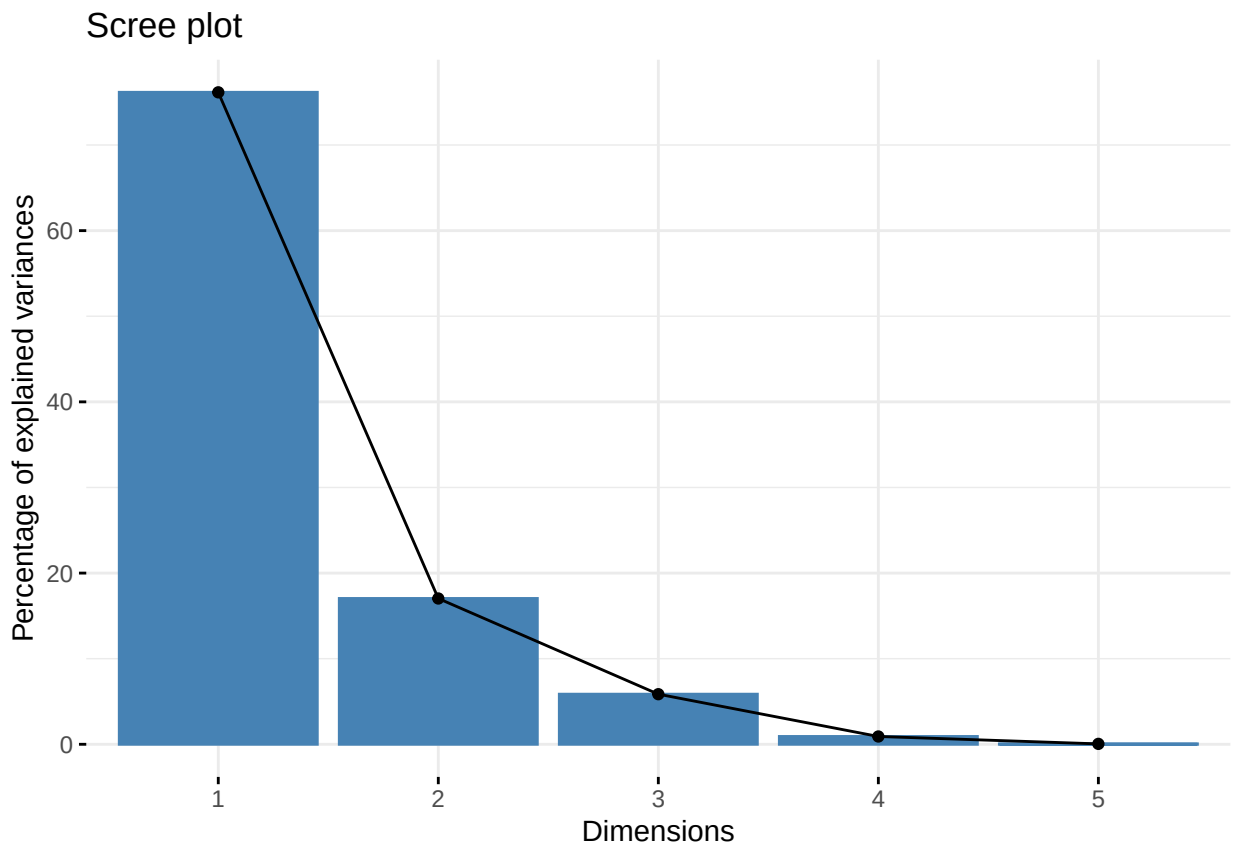

```
fviz_pca_ind(cpS_varianza, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```



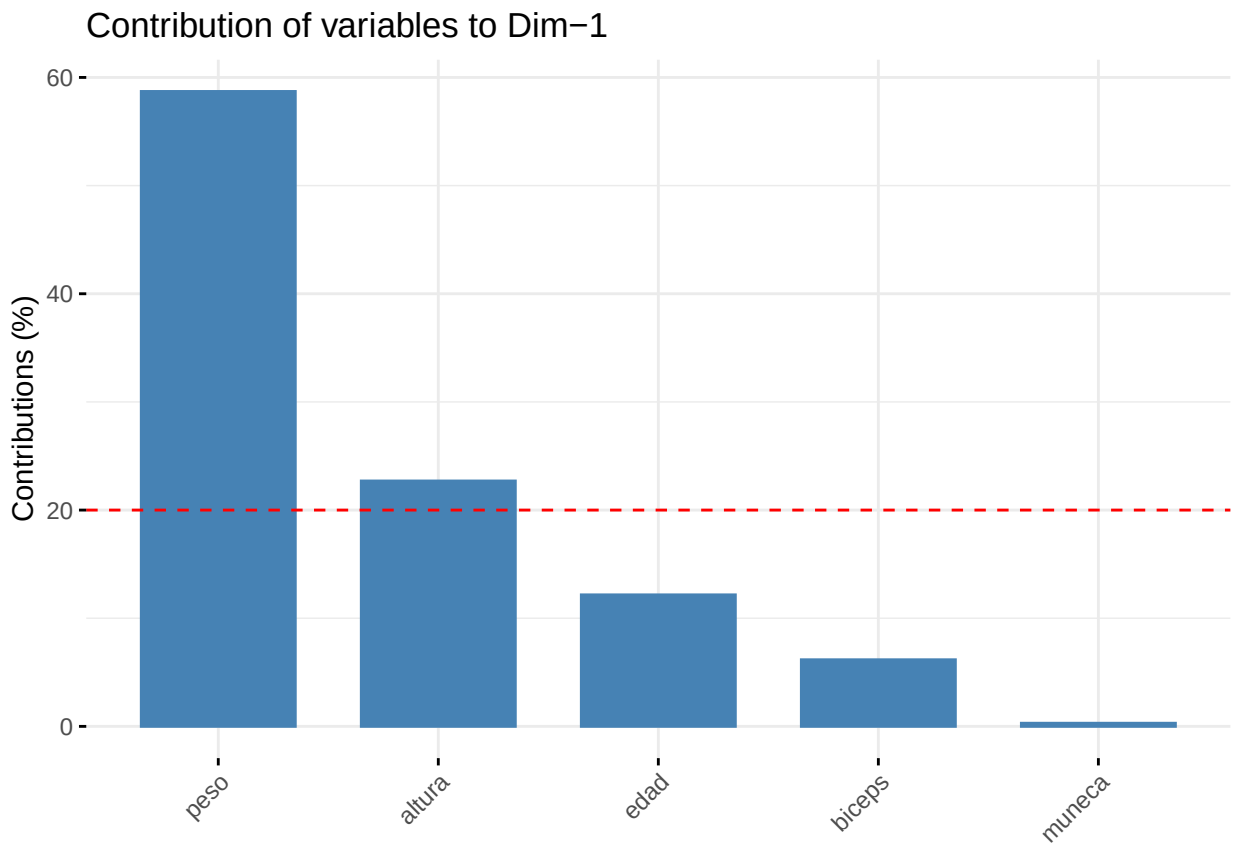

```
fviz_pca_var(cpS_varianza, col.var = "red", addEllipses = TRUE, repel = TRUE)
```



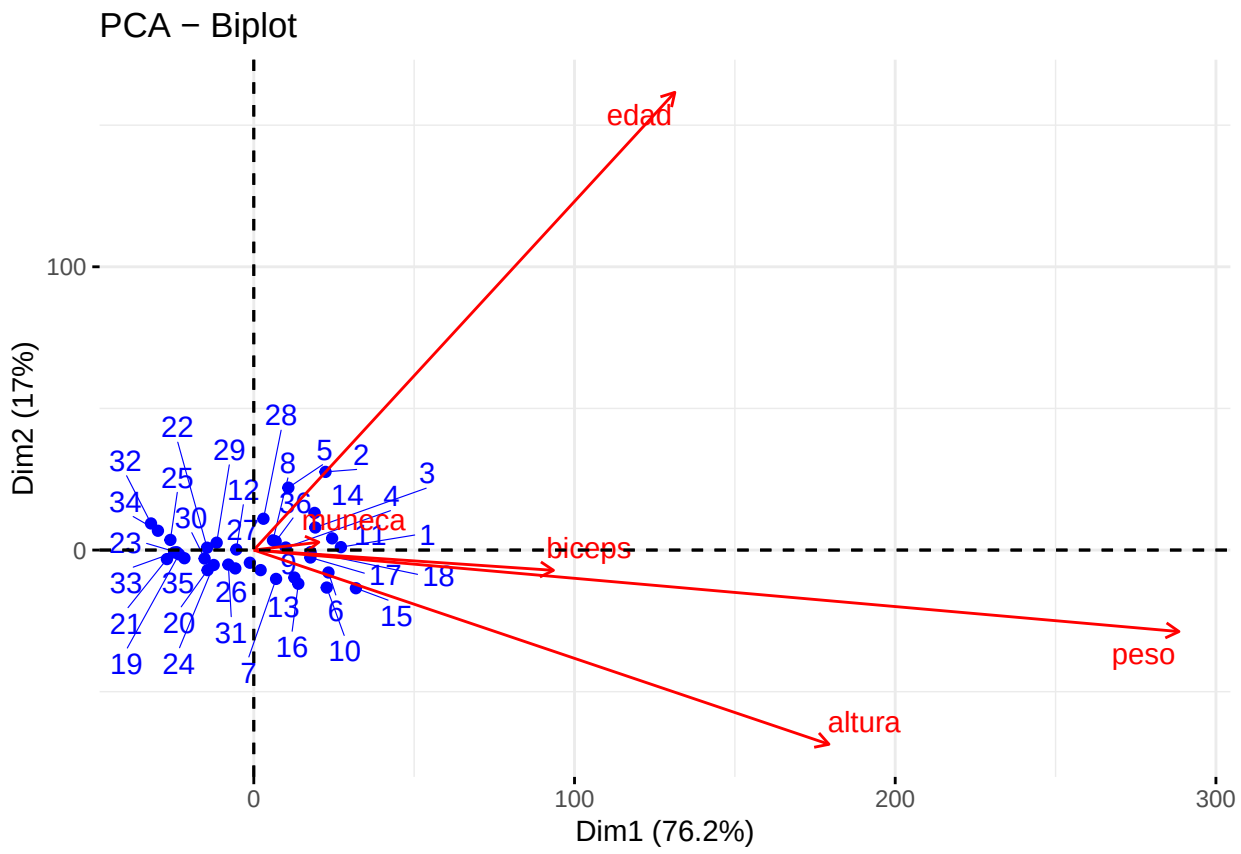

```
fviz_screeplot(cpS_varianza)
```



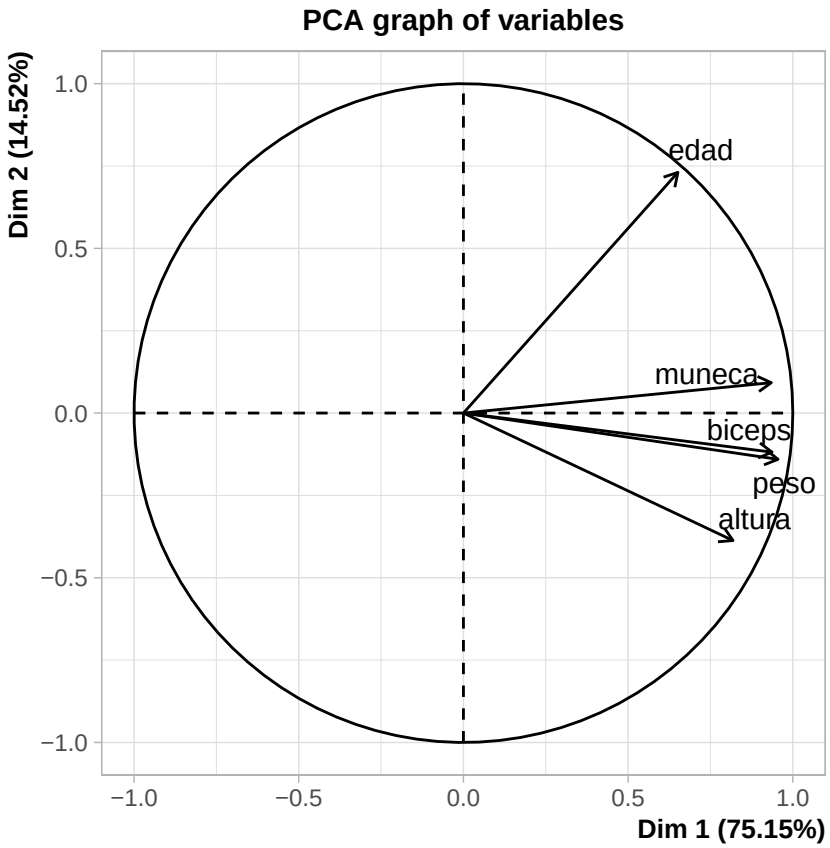

```
fviz_contrib(cpS_varianza, choice = "var")
```



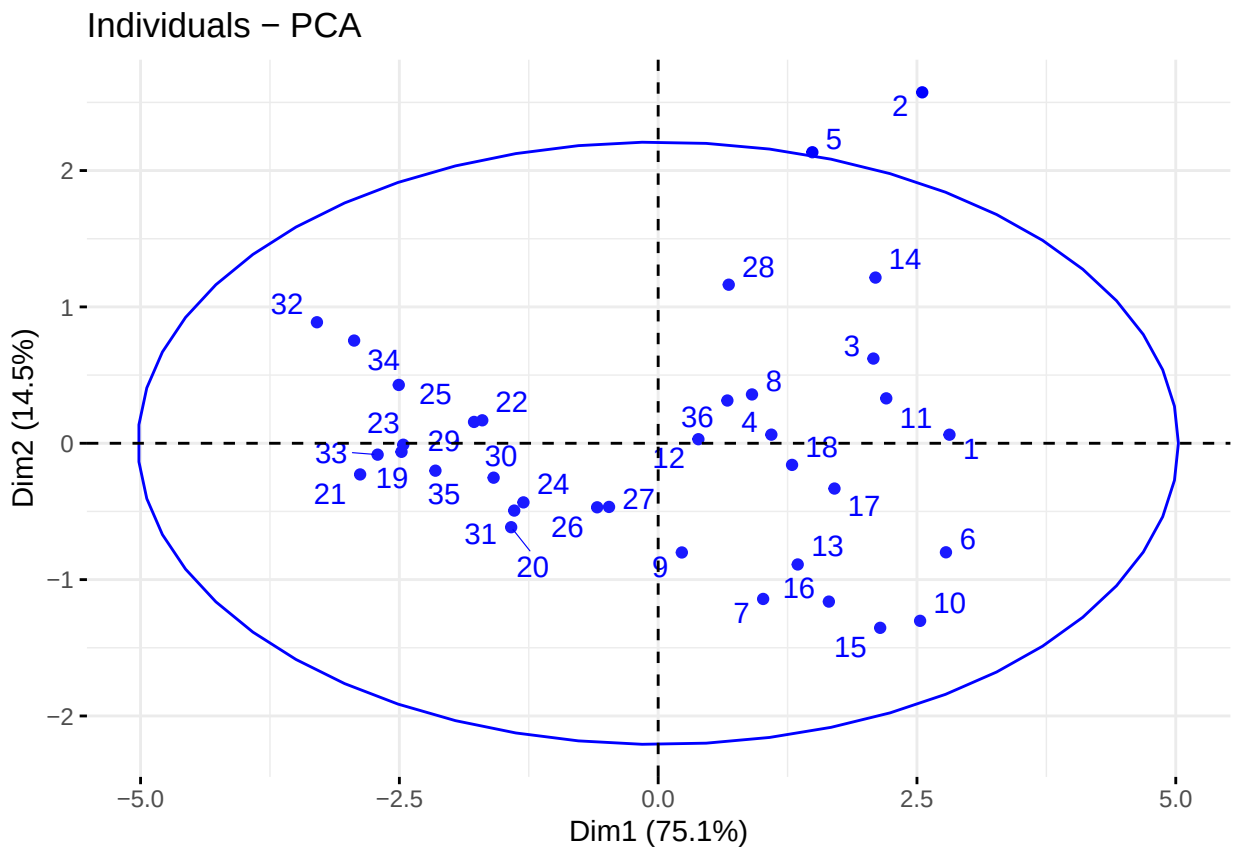

```
fviz_pca_biplot(cpS_varianza, repel = TRUE, col.var = "red", col.ind = "blue")
```



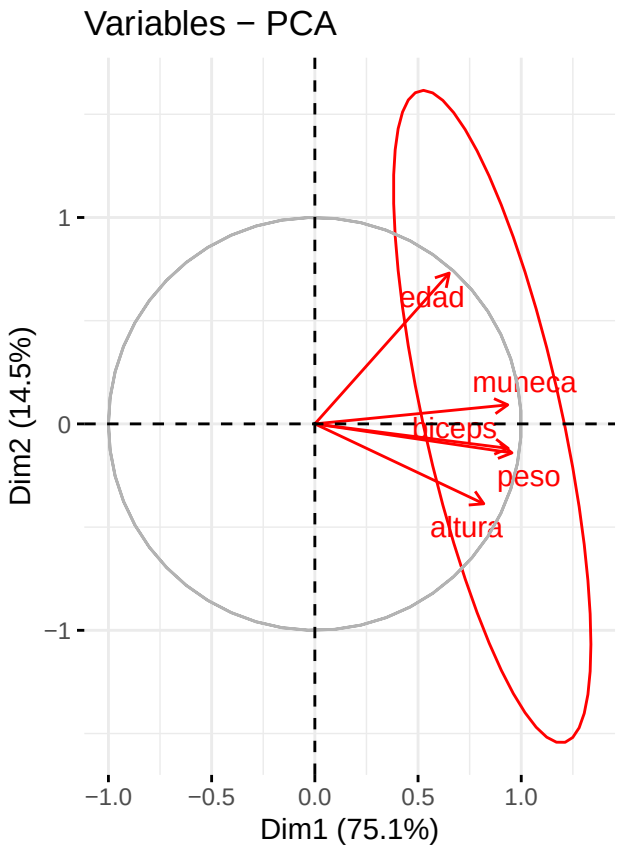

```
cpS_correlacion <- PCA(datos_sin_sexo, scale.unit = TRUE)
```



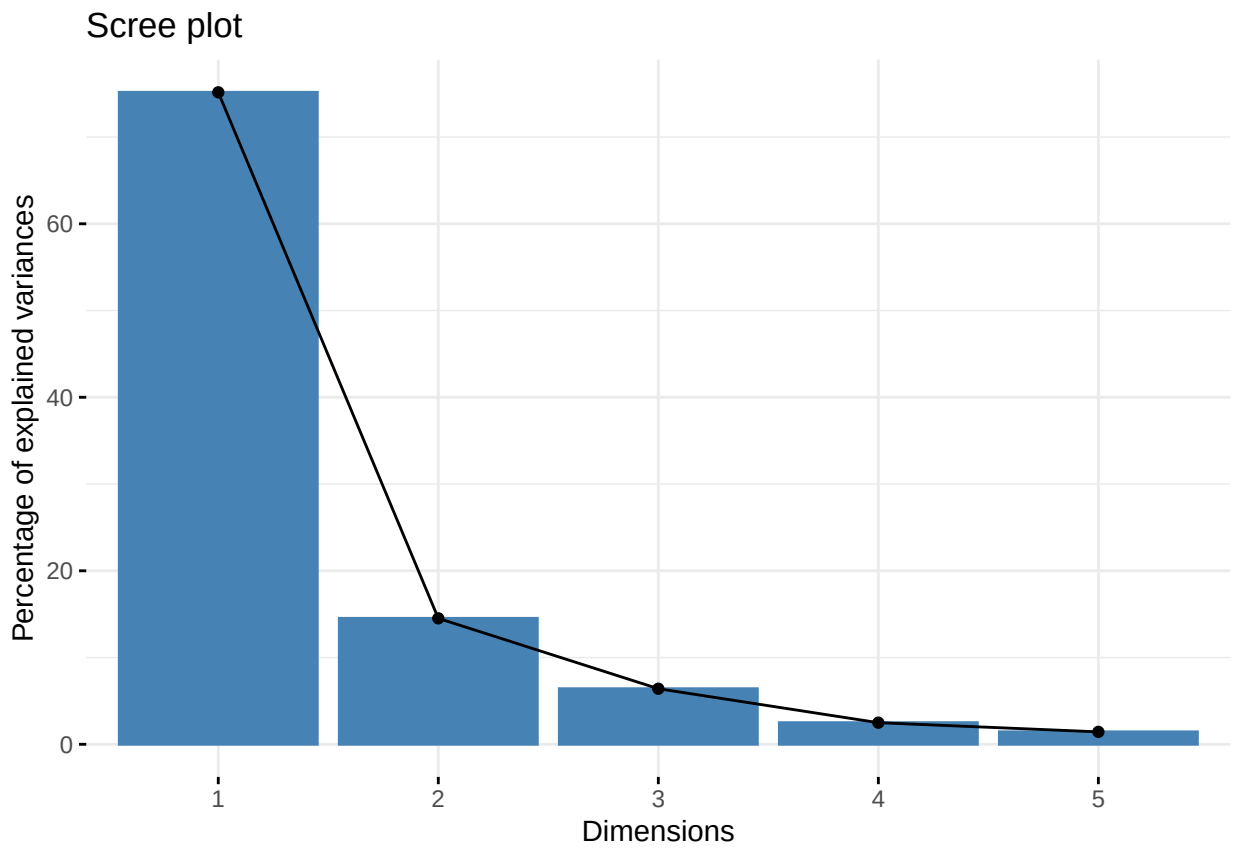

```
fviz_pca_ind(cpS_correlacion, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```



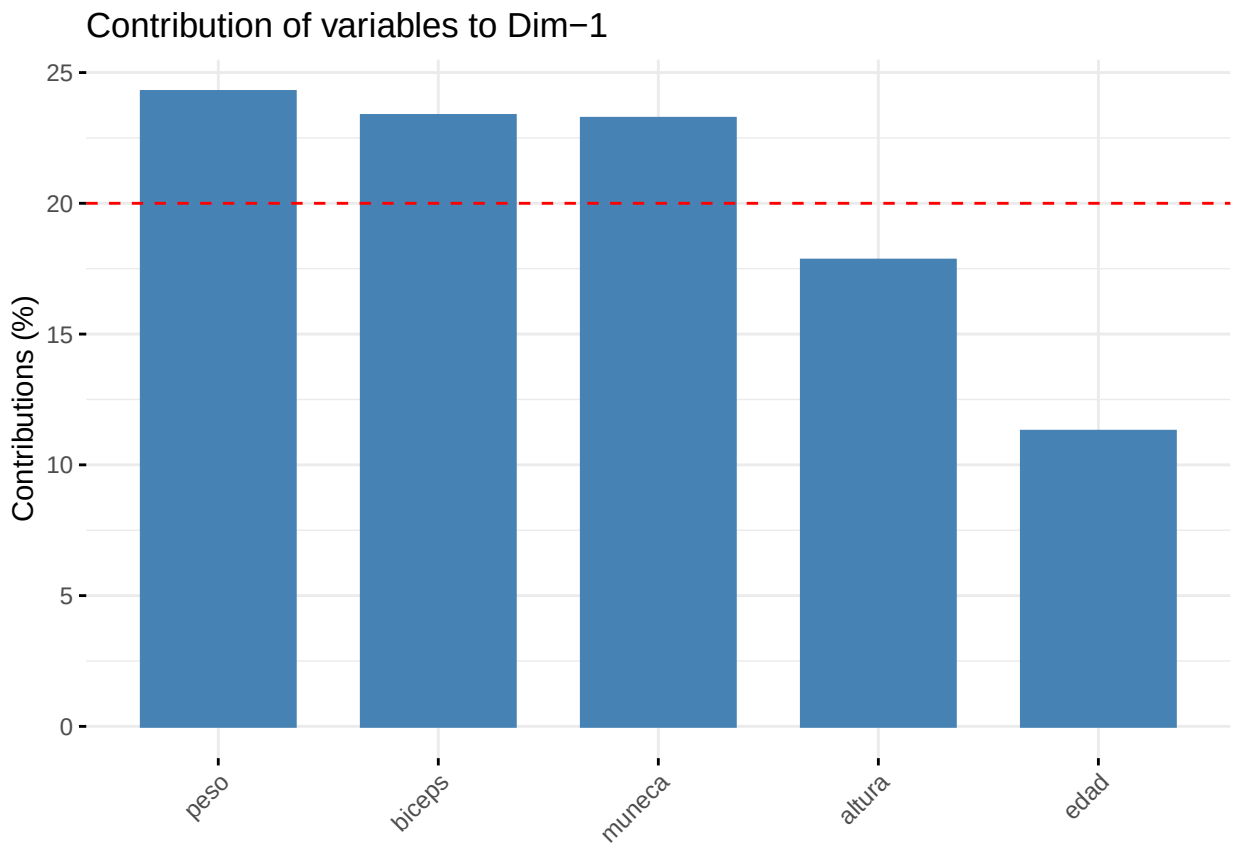

```
fviz_pca_var(cpS_correlacion, col.var = "red", addEllipses = TRUE, repel = TRUE)
```



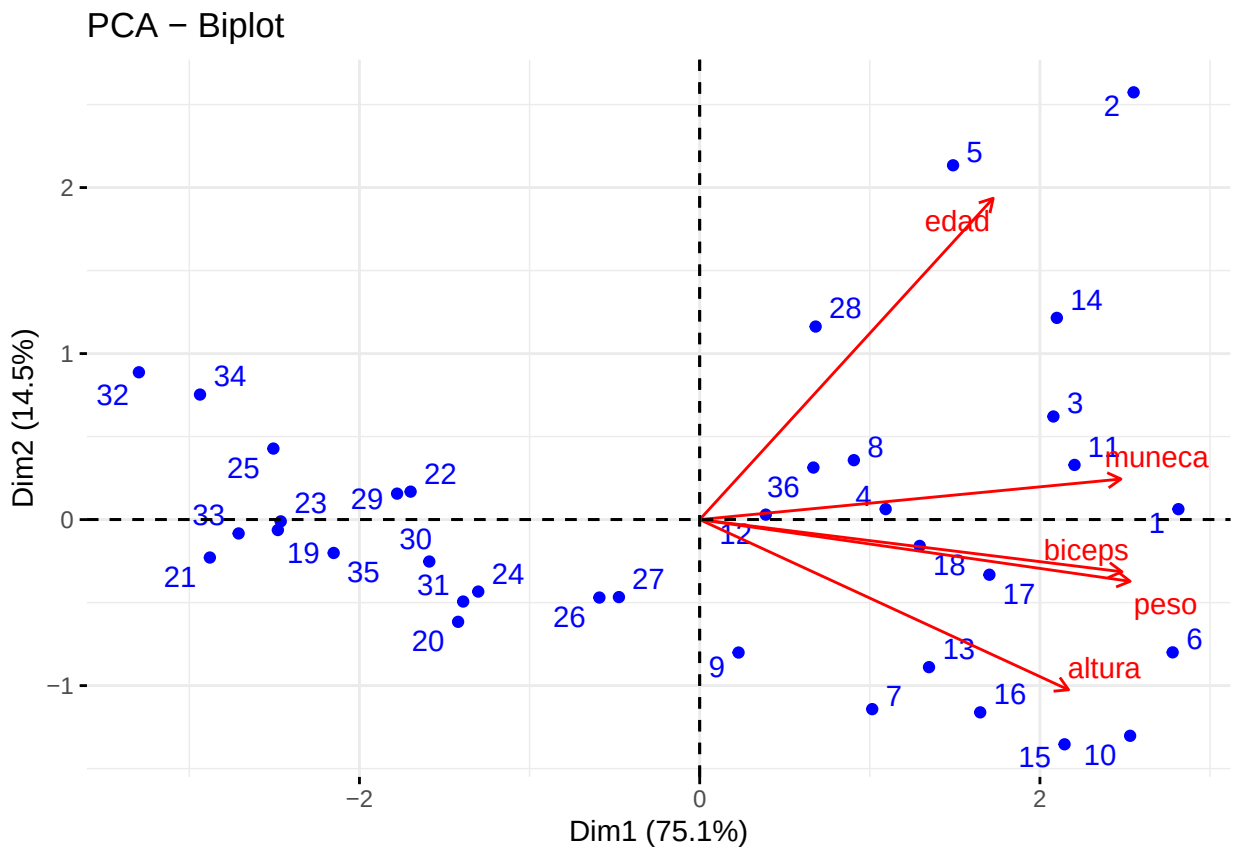

```
fviz_screeplot(cpS_correlacion)
```




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




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



Las gráficas de varianza nos muestran que edad, peso y altura son los que más afecta a los componentes, PCA nos da a entender que existe una fuerte correlación entre los individuos, con solamente un individuo fuera del alcance, el scree plot nos demuestra que el porcentaje se explica en su mayoría en la segunda dimensión, y que de los valores peso y altura son los que más contribuyen

Las gráficas de correlación nos muestra que edad, muñecas, biceps, peso y altura son los que afectan a los componentes, gracias a la gráfica PCA de las variables, al mismo tiempo que existe una fuerte correlación entre los individuos, el scree plot nos demuestra que el porcentaje se explica en su mayoría en la segunda dimensión, y que peso, biceps e muneca son los que más contribuyen.

#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? Se concluye que los dos componentes capturan una cantidad importante de los datos, por otro lado, se puede determinar que la matriz de correlación aporta un mayor interés, debido a que es más representativo al estar estandarizado, lo que facilita la identificación de patrones y relaciones más claras entre las variables.

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 de 96 países en el mundo. Comparar los resultados y argumentar cuál es mejor según los resultados obtenidos. La matriz de correlación resulta ser el mejor para los datos económicos y sociales de 96 países, ya que ayuda a revelar las relaciones entre las variables y los componentes principales

¿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, auxiliate también de los gráficos) En la primera, se descubrió que las variables peso, altura y bíceps tienen los coeficientes más altos, mientras que en la segunda es la edad.

Escriba las combinaciones finales que se recomiendan para hacer el análisis de componentes principales. Serie a1+peso * a2+altura * a3+biceps para el componente 1, y para el componente dos seria a1+edad * a2+muñeca

Interpreta los resultados en término de agrupación de variables (puede ayudar “índice de riqueza”, “índice de ruralidad”, etc) Al analizar los resultados, es posible que que las variables tanto de peso como de altura, se encuentran correlacionadas, esto puede intuir que estas variables tienen que ver con la calidad socioeconómica de lasa personas, por otra parte, la edad, también influye, ya que determina como esas caracteristicas afecta a la persona, a su vez, las variables muñeca y bicep nos ayudan a comprender que están relacionadas con el desarrollo físico de la persona, lo que ayuda a reforzar la idea que permite analizar la calidad socio económica.