A4 - Componentes Principales

Juan Bernal

2024-10-08

En la base de datos Corporal contiene las medidas corporarles de 36 estudiantes de la universidad. Haz una análisis de Componentes principales con la matriz de varianzas-covarianzas y la matriz de correlaciones. Compara los resultados y argumenta cuál es mejor según los resultados obtenidos.

```
data = read.csv('corporal.csv')
head(data)
```

	edad <int>	peso <dbl>	altura <dbl>	sexo <chr>	muneca <dbl>	biceps <dbl>
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

6 rows

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

```
data = data[ , !(names(data) %in% 'sexo')]
summary(data)
```

```
##       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
```

```
sapply(data, sd)
```

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

Eliminamos la variable “sexo” dado que solo podemos aplicar componentes principales a variables numéricas. Y además obtenemos el análisis descriptivo de las variables restantes.

Parte I

Realiza el análisis de los valores y vectores propios con la matriz de covarianzas y con la de correlación. Analiza la varianza explicada por cada componente en cada caso e interpreta dentro del contexto del problema.

1. Calcule las matrices de varianza-covarianza S con cov(X) y la matriz de correlaciones R con cor(X) y realice los siguientes pasos con cada una:

```
cov = cov(data) # 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
```

```
cor = cor(data) # 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
```

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

```
eigcov = eigen(cov) # Eigenvalores y eigenvectores de la matriz de covarianza
eigcov
```

```
## 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
```

```
eigcor = eigen(cor) # Eigenvalores y eigenvectores de la matriz de correlación
eigcor
```

```
eigcor
```

```
## 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
```

b. 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.

```
varcor = eigcor$values / sum(eigcor$values)
cat('Cada componente con la matriz de correlación explica el siguiente porcentaje de
varianza: ',varcor)
```

```
## Cada componente con la matriz de correlación explica el siguiente porcentaje de va
rianza:  0.7514995 0.1451713 0.06406596 0.02492375 0.0143395
```

```
varcor = eigcor$values / sum(diag(cor))
cat('Comprobando sum(diag()) para correlación: ',varcor)
```

```
## Comprobando sum(diag()) para correlación:  0.7514995 0.1451713 0.06406596 0.024923
75 0.0143395
```

```
varcov = eigcov$values / sum(eigcov$values)
cat('Cada componente con la matriz de covarianza explica el siguiente porcentaje de v
arianza: ',varcov)
```

```
## Cada componente con la matriz de covarianza explica el siguiente porcentaje de var
ianza:  0.7615357 0.1703099 0.05853072 0.009127104 0.0004965839
```

```
varcov = eigcov$values / sum(diag(cov))
cat('Comprobando sum(diag()) para covarianza: ',varcov)
```

```
## Comprobando sum(diag()) para covarianza:  0.7615357 0.1703099 0.05853072 0.0091271
```

```
04 0.0004965839
```

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

```
cat('La varianza explicada según el número de componentes principales de la matriz de correlación es:',cumsum(varcor))
```

```
## La varianza explicada según el número de componentes principales de la matriz de correlación es: 0.7514995 0.8966708 0.9607368 0.9856605 1
```

```
cat('La varianza explicada según el número de componentes principales de la matriz de covarianza es:',cumsum(varcov))
```

```
## La varianza explicada según el número de componentes principales de la matriz de covarianza es: 0.7615357 0.9318456 0.9903763 0.9995034 1
```

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

Los componentes 1 y 2 de ambas matrices son los más importantes, pues llegamos a explicar aproximadamente el 90% de la variación de los datos en ambos casos.

e. 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.

Para la matriz de correlación:

```
CP1 = -Edad*0.336 - Peso*0.4927 - Altura*0.4222 - Muneca*0.4822 - Biceps*0.4833

CP2 = Edad*0.8575 - Peso*0.1647 - Altura*0.4542 + Muneca*0.1082 - Biceps*0.1392
```

Donde las variables que más contribuyen al CP1 son Peso, Muñeca y Biceps, pues aproximadamente el 50% de cada variable aporta a la nueva variable, y las que más contribuyen al CP2 son Edad y Altura, pues su el valor de Edad es casi completamente tomado en cuenta y el de altura se cuenta al casi un 50%.

```
X = c(data$edad, data$peso, data$altura, data$muneca, data$biceps)
CP1_cor = X*eigcor$vectors[1:5]
CP2_cor = X*eigcor$vectors[1:5, 2]
CP_cor = data.frame(CP1_cor,CP2_cor)
CP_cor # Componentes principales de la matriz de correlación
```

CP1_cor <dbl>	CP2_cor <dbl>
-14.445034	36.8750827
-32.025929	-10.7108388
-19.000915	-20.4400028
-17.841116	4.0062664
-26.582264	-7.6597635

-11.085724	28.2994820
-12.317665	-4.1195534
-14.778490	-15.8977799
-13.501385	3.0317691
-12.566161	-3.6209791

1-10 of 180 rows Previous **1** 2 3 4 5 6 … 18 Next

Para la matriz de covarianza:

```
CP1 = -Edad*0.3487 - Peso*0.7661 - Altura*0.4763 - Muneca*0.0538 - Biceps*0.2481

CP2 = Edad*0.9075 - Peso*0.1616 - Altura*0.3851 + Muneca*0.0155 - Biceps*0.0402
```

Donde las variables que más contribuyen al CP1 son Peso y Altura, pues se toma un 75% del Peso en cuenta y aproximadamente un 50% del valor de la altura, y las que más contribuyen al CP2 son Edad y Altura, pues Edad aporta un 90% de su valor y Altura aproximadamente un 40% de su valor.

```
CP1_cov = X*eigcov$vectors[1:5]
CP2_cov = X*eigcov$vectors[1:5, 2]
CP_cov = data.frame(CP1_cov,CP2_cov)
CP_cov # Componentes principales de la matriz de covarianza
```

CP1_cov <dbl>	**CP2_cov** <dbl>
-14.9945309	39.0246532
-49.8014308	-10.5077770
-21.4345823	-17.3328959
-1.9928898	0.5750650
-13.6495519	-2.2122155
-11.5074307	29.9491525
-19.1543965	-4.0414527
-16.6713418	-13.4811412
-1.5081328	0.4351843
-6.4525154	-1.0457746

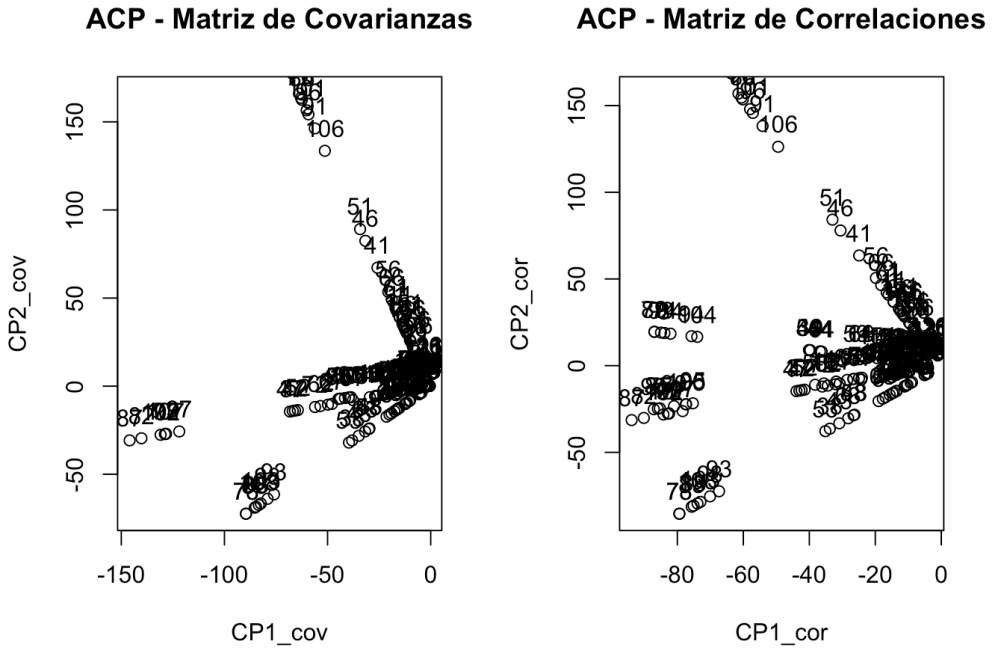
1-10 of 180 rows Previous **1** 2 3 4 5 6 … 18 Next

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

Parte II

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

```
# Graficar las puntuaciones de las dos primeras componentes para ambas matrices
par(mfrow = c(1, 2))
# Covarianzas
plot(CP_cov[,1:2], type = "p", main = "ACP - Matriz de Covarianzas")
text(CP_cov[,1], CP_cov[,2], labels = 1:nrow(CP_cov), pos = 3)
# Correlaciones
plot(CP_cor[,1:2], type = "p", main = "ACP - Matriz de Correlaciones")
text(CP_cor[,1], CP_cor[,2], labels = 1:nrow(CP_cor), pos = 3)
```



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

```
head(CP_cov)
```

	CP1_cov <dbl>	CP2_cov <dbl>
1	-14.99453	39.024653
2	-49.80143	-10.507777
3	-21.43458	-17.332896
4	-1.99289	0.575065
5	-13.64955	-2.212215
6	-11.50743	29.949152

6 rows

b. 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.

```
head(CP_cor)
```

	CP1_cor <dbl>	CP2_cor <dbl>
1	-14.44503	36.875083
2	-32.02593	-10.710839
3	-19.00092	-20.440003
4	-17.84112	4.006266
5	-26.58226	-7.659763
6	-11.08572	28.299482

6 rows

2. Interprete los gráficos en términos de:

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

La gráfica de la izquierda muestra un análisis de componentes principales (ACP) basado en la matriz de covarianzas. La gráfica de la derecha utiliza la matriz de correlaciones. En ambos casos, los ejes CP1 y CP2 representan los dos primeros componentes principales que capturan la mayor parte de la varianza de los datos.

En la matriz de covarianzas (izquierda), las variables se asocian más fuertemente con los componentes principales cuando están más alejadas del origen. Las variables que se encuentran en direcciones similares pueden estar correlacionadas, mientras que las variables en direcciones opuestas pueden estar negativamente correlacionadas con esos componentes.

En la matriz de correlaciones (derecha), las variables con mayor dispersión a lo largo de CP1 y CP2 son las que más contribuyen a los componentes principales. Aquí, las correlaciones estandarizan los datos, permitiendo que las variables con diferentes escalas sean comparadas en términos relativos.

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

Las puntuaciones de las observaciones representan las coordenadas de las variables en el nuevo espacio generado por los componentes principales (CP1 y CP2). En ambos gráficas los números alrededor de los puntos indican observaciones específicas. Si observaciones similares se agrupan en el mismo cuadrante, entonces las variables asociadas en ese cuadrante tienen relaciones similares con esas observaciones.

c. Detecte posibles datos atípicos

Se pueden detectar posibles datos atípicos (outliers) al observar las observaciones que están alejadas del resto en cada gráfico. En ambos gráficos, se ven observaciones como la 106 y la 41 que parecen estar bastante alejadas del resto del grupo, lo que sugiere que podrían ser atípicas. Esto indica que estas observaciones no siguen las mismas tendencias que el resto de los datos y podrían requerir un análisis más detallado.

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?

```
library(stats)
# Matriz de covarianza
cpS=princomp(data,cor=FALSE) #Para la matriz de correlación usa cor=TRUE
cpaS=as.matrix(data)%*%cpS$loadings  #Calcula las puntuaciones
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
```

```
cpS$loading
```

```
##
## 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
```

```
head(cpS$scores)
```

```
##         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
```

```
# Matriz de correlación
cpR=princomp(data,cor=TRUE) #Para la matriz de correlación usa cor=TRUE
cpaR=as.matrix(data)%*%cpR$loadings  #Calcula las puntuaciones
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
```

```
cpR$loading
```

```
##
## 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
```

```
head(cpR$scores)
```

```
##        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
```

El comando princomp() realiza el PCA utilizando la descomposición en valores singulares (SVD) de la matriz de datos, aplicando preferentemente la matriz de covarianzas o la matriz de correlaciones.

- summary(cpaS): Ayuda a elegir cuántos componentes retener, mirando la varianza explicada.
- cpaS$loadings: Muestra qué variables están más asociadas con cada componente.
- cpaS$scores: Permite visualizar las observaciones en el nuevo espacio de los componentes, detectando patrones o outliers.

Parte III

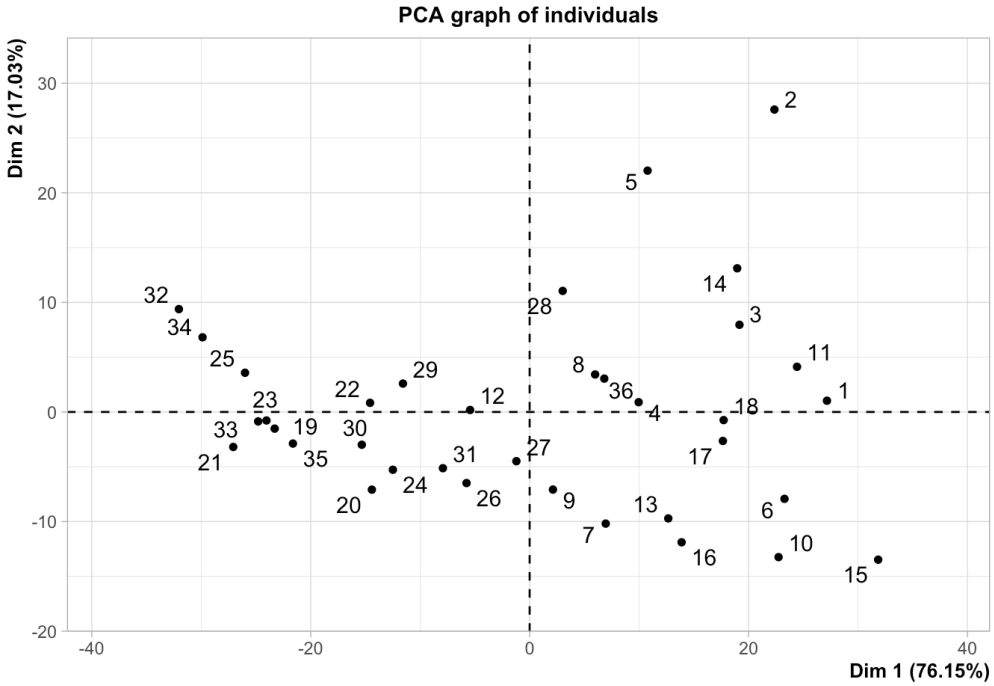
1. Explore los siguientes gráficos relativos a Componentes Principales.

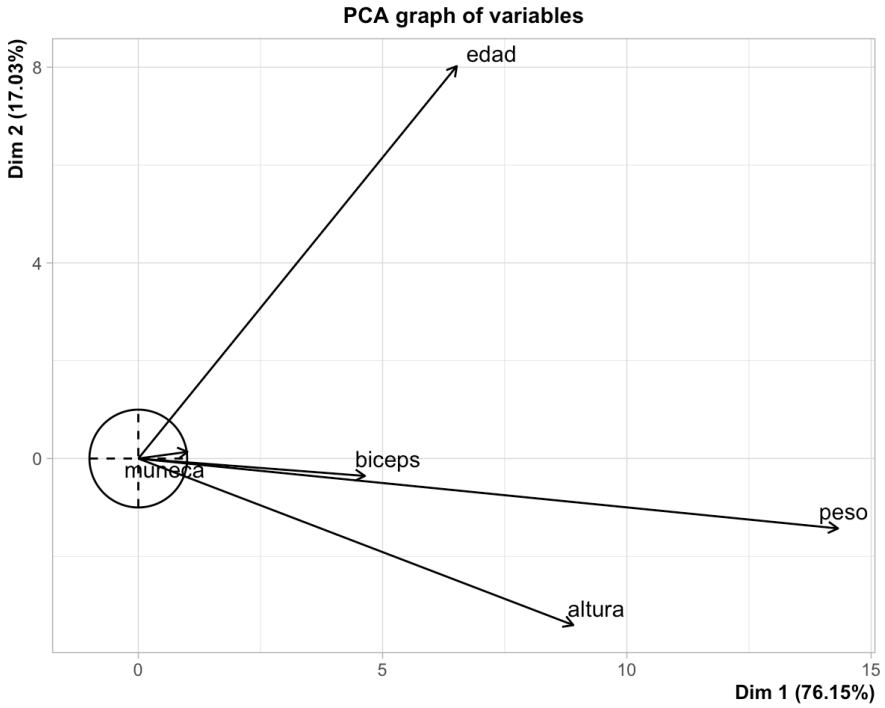
2. Interprete cada gráfico e identifica qué es lo que se está graficando en cada uno. Realiza el análisis con la matriz de varianzas y covarianzas y correlación.

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

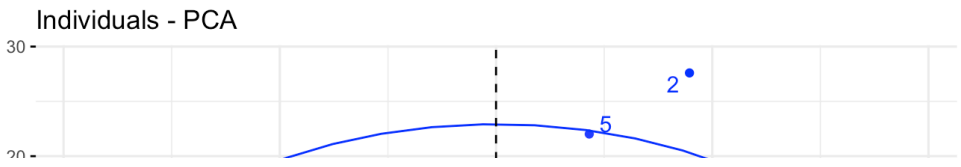
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve
3WBa
```

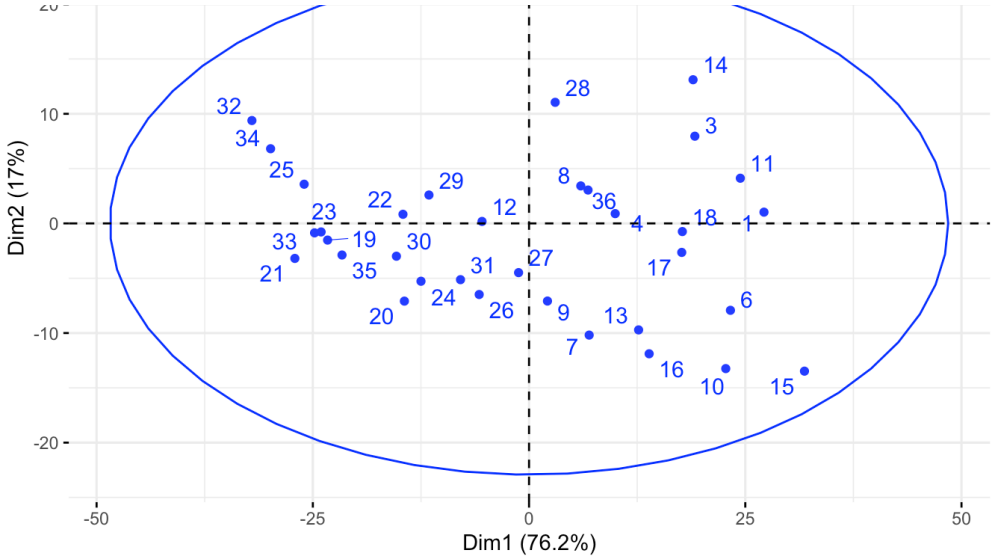
```
cpS = PCA(data,scale.unit=FALSE) #Para matriz de correlaciones usa scale.unit=TRUE
```





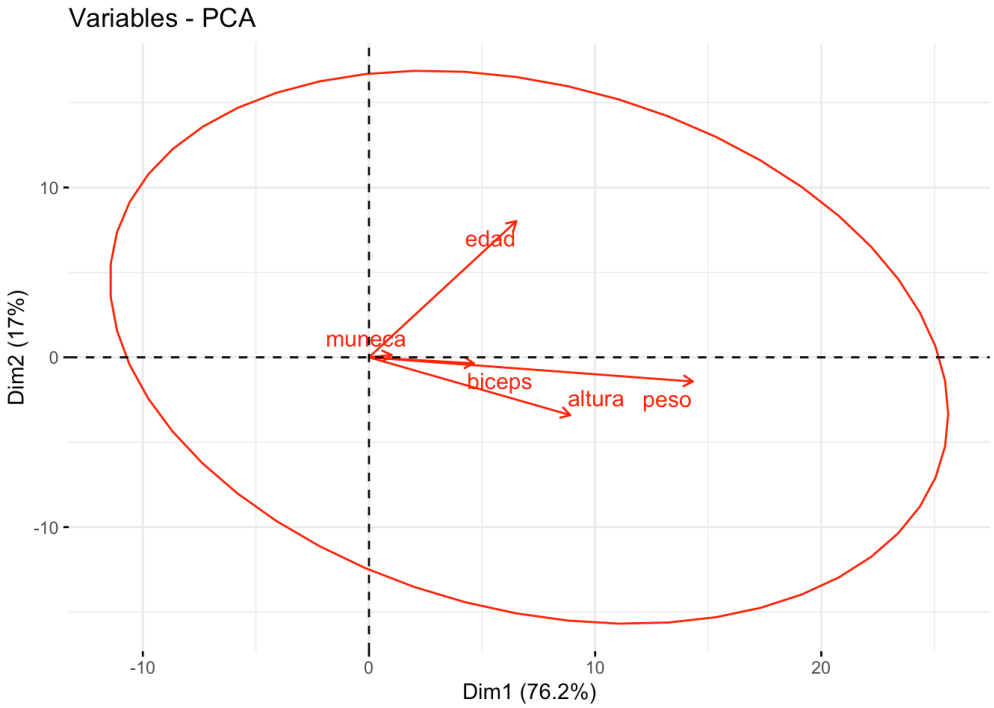

```
fviz_pca_ind(cpS, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```





En esta gráfica con la matriz de varianzas y covarianzas, se están destacando observaciones con mayor varianza en las variables originales. Las observaciones alejadas del centro podrían tener valores extremos o ser casos atípicos.

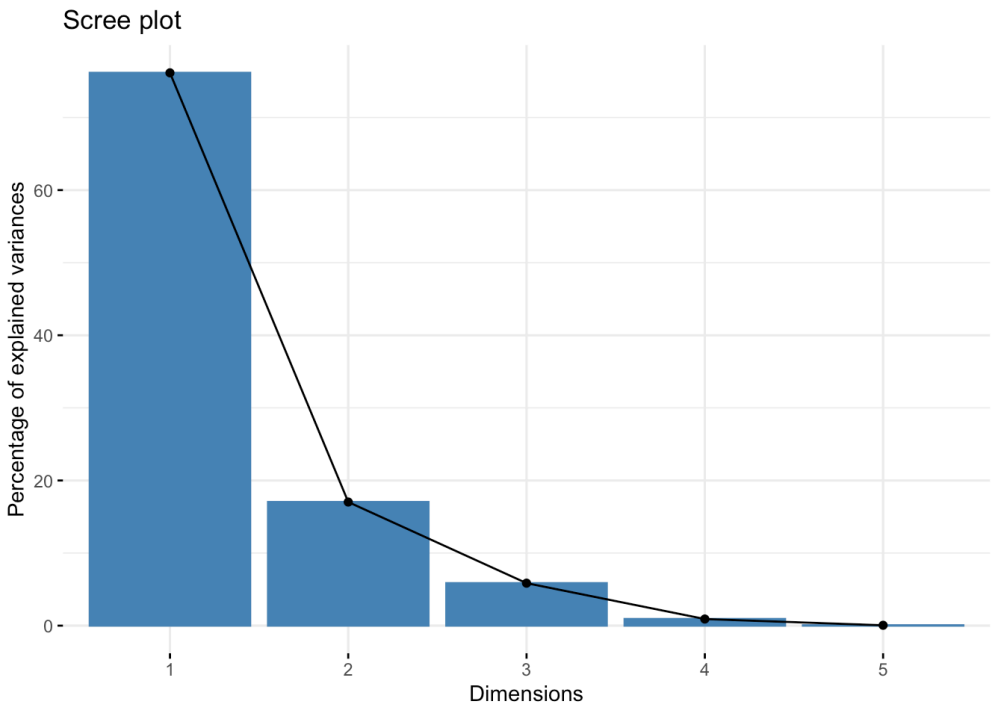
```
fviz_pca_var(cpS, col.var = "red", addEllipses = TRUE, repel = TRUE)
```



En esta gráfica de variables (PCA), se proyectan las variables originales en el espacio de los dos primeros componentes principales, Dim1 y Dim2. El análisis refleja las relaciones entre las variables y cómo contribuyen a los componentes principales. Dim1 parece capturar variabilidad relacionada con medidas corporales como peso y altura, que están altamente correlacionadas entre sí. Dim2 parece captar información más relacionada con edad. El ángulo entre las flechas de las variables da información sobre su grado de correlación: flechas cercanas en dirección indican una correlación positiva, mientras que flechas opuestas sugieren una correlación negativa.

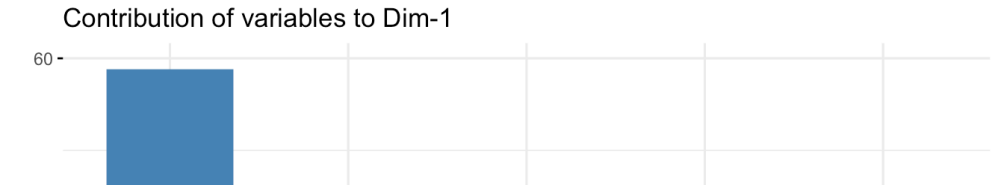
sugieren una correlación negativa.

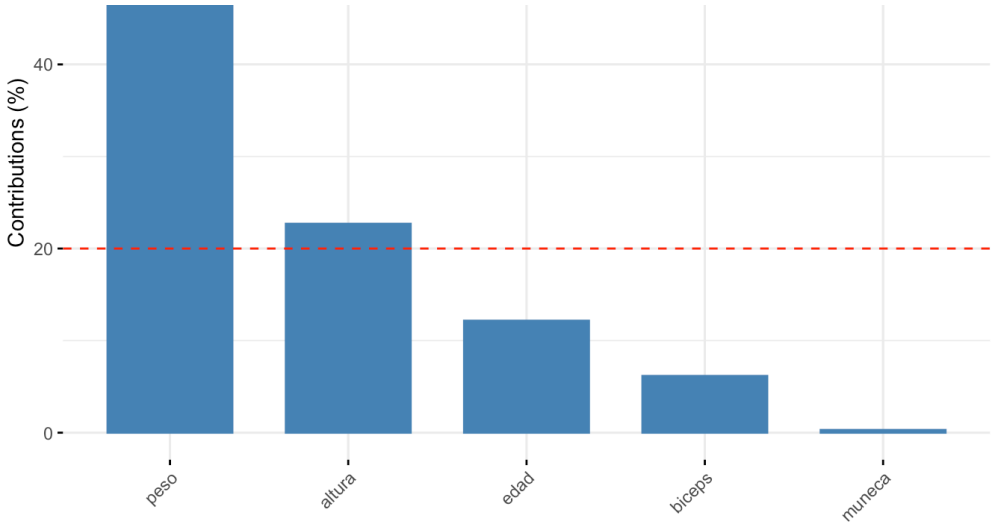
```
fviz_screeplot(cpS)
```



El screeplot ayuda a determinar cuántos componentes principales se deben conservar. Generalmente, se eligen los componentes que explican un porcentaje significativo de la varianza total. Notemos que con los primeros dos componentes se puede explicar más del 90% de la variabilidad total de los datos, lo que suele ser suficiente para interpretar los patrones principales. Los componentes adicionales no aportan mucha información adicional relevante.

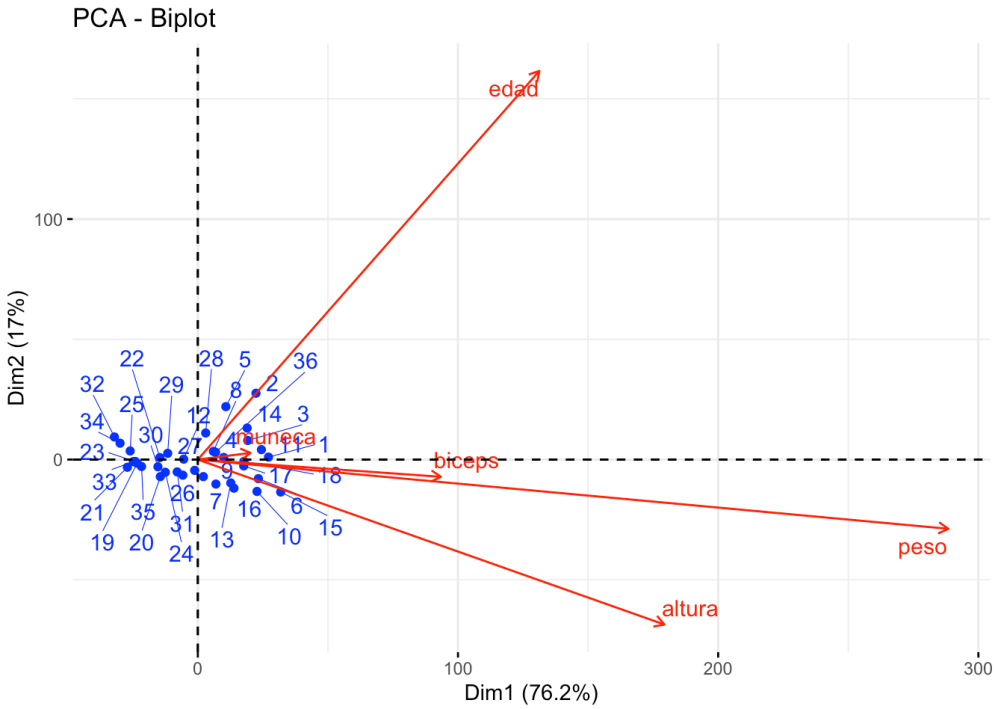
```
fviz_contrib(cpS, choice = c("var"))
```





Este gráfico muestra qué variables tienen una mayor influencia en la construcción de los componentes principales. Las variables con mayor contribución son las que más influyen en el PCA. En el caso de la gráfica con la matriz de covarianza, las variables que más contribuyen son Peso y Altura, mientras que la influencia de Edad, Biceps y Muñeca es tan pequeña que no es indispensable guardar dichas variables.

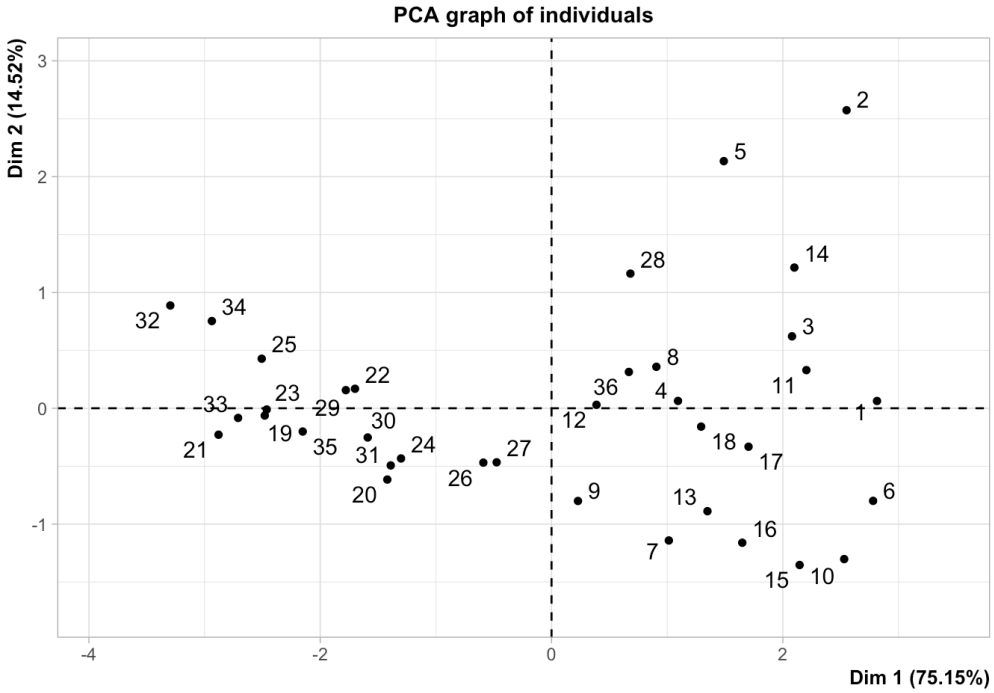
```
fviz_pca_biplot(cpS, repel=TRUE, col.var="red", col.ind="blue")
```

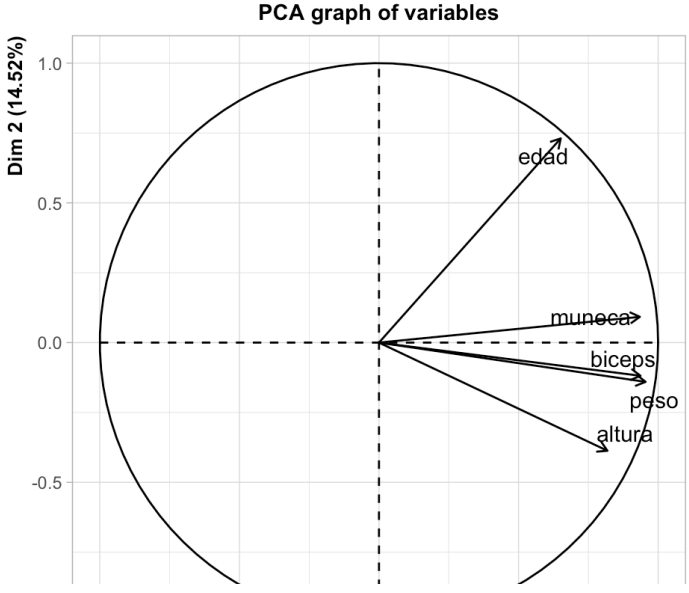


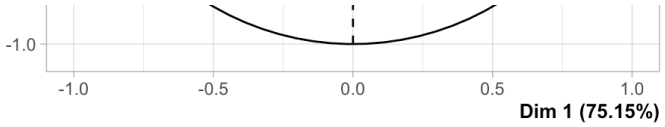
El biplot es útil para ver simultáneamente cómo las variables influyen en las observaciones. Permite identificar

El biplot es útil para ver simultáneamente cómo las variables influyen en las observaciones. Permite identificar relaciones entre variables y grupos de observaciones, así como patrones de agrupamiento.

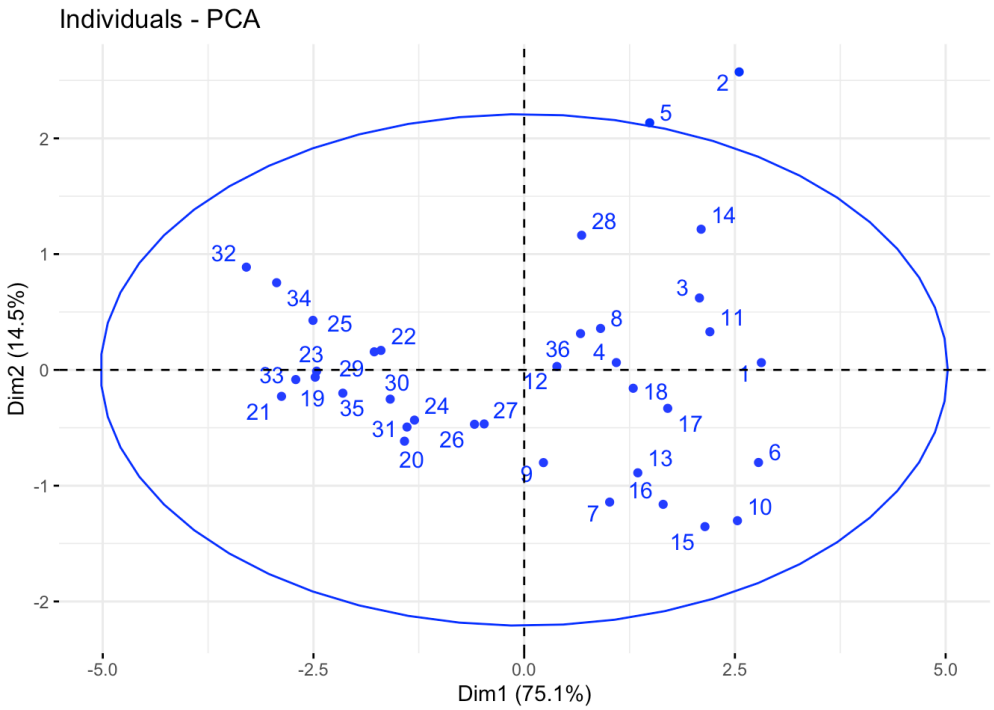
```
cpR = PCA(data,scale.unit=TRUE) #Para matriz de correlaciones usa scale.unit=TRUE
```





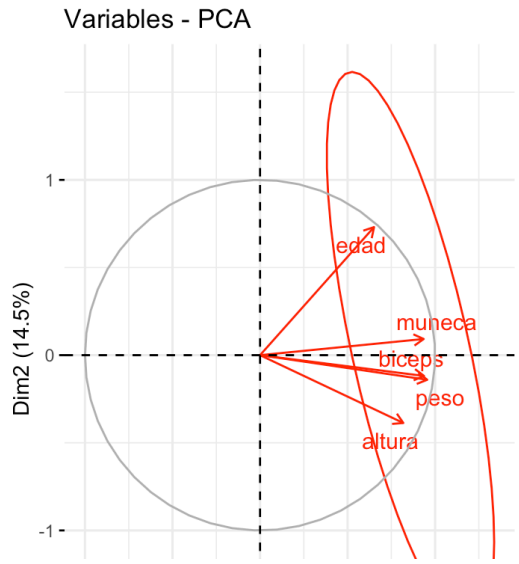


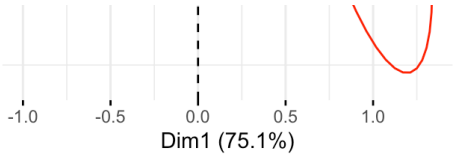

```
fviz_pca_ind(cpR, col.ind = "blue", addEllipses = TRUE, repel = TRUE)
```



En esta gráfica con la matriz de correlación se están destacando observaciones con mayor varianza en las variables originales. Las observaciones alejadas del centro podrían tener valores extremos o ser casos atípicos.

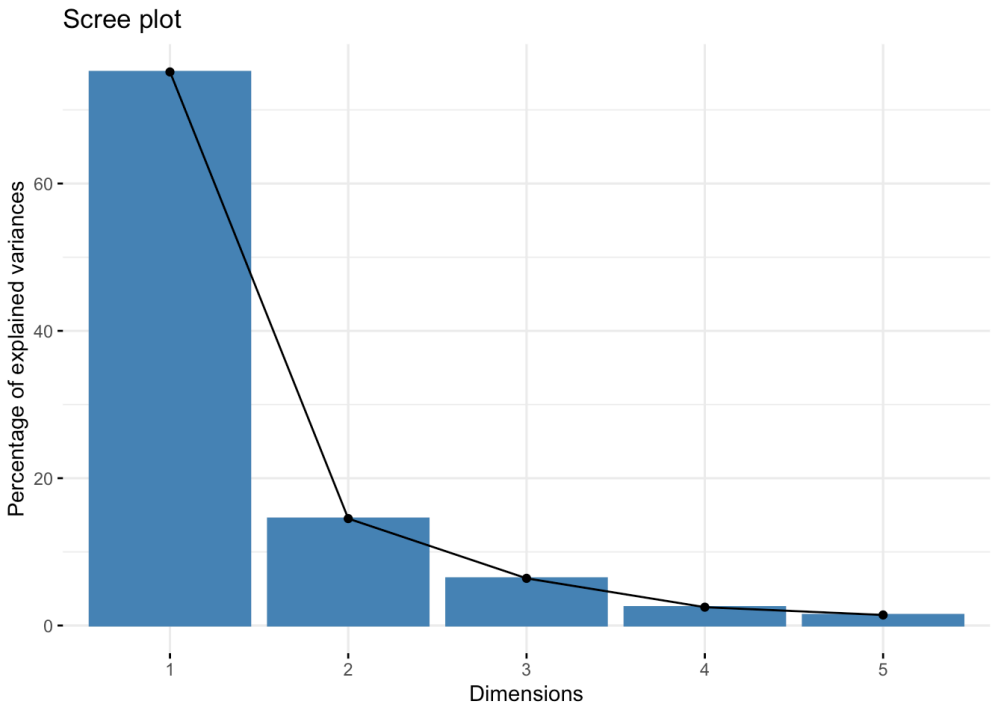
```
fviz_pca_var(cpR, col.var = "red", addEllipses = TRUE, repel = TRUE)
```





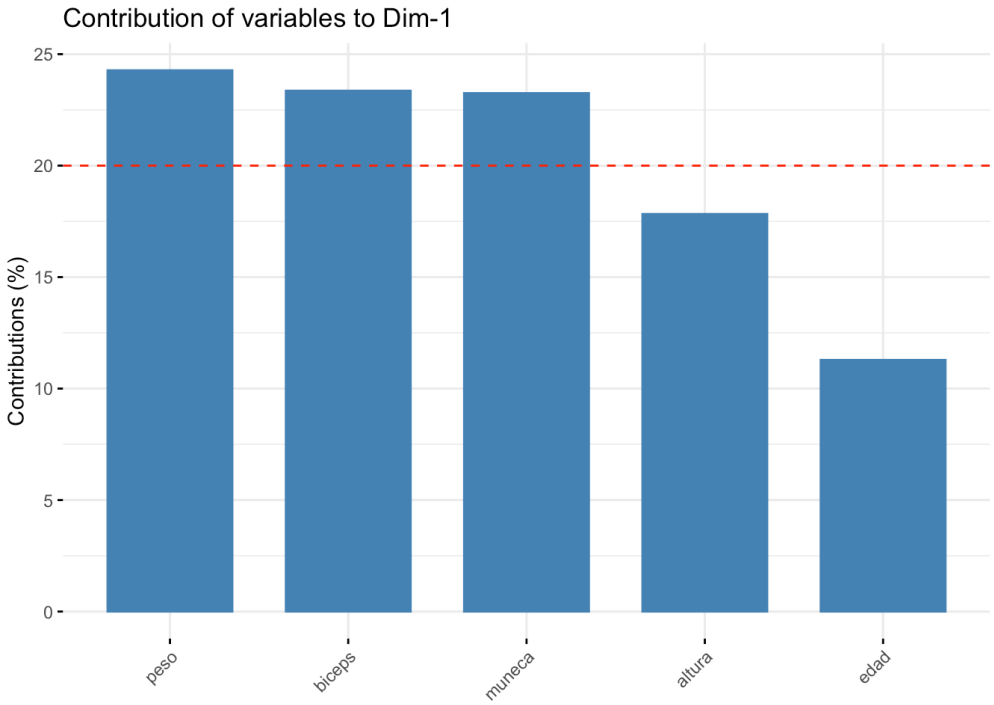
En esta gráfica de variables (PCA), se proyectan las variables originales en el espacio de los dos primeros componentes principales, Dim1 y Dim2. El análisis refleja las relaciones entre las variables y cómo contribuyen a los componentes principales. Dim1 parece capturar variabilidad relacionada con medidas corporales como peso, altura, biceps y muñeca que están altamente correlacionadas entre sí. Dim2 parece captar información más relacionada con edad. El ángulo entre las flechas de las variables da información sobre su grado de correlación: flechas cercanas en dirección indican una correlación positiva, mientras que flechas opuestas sugieren una correlación negativa.

```
fviz_screeplot(cpR)
```



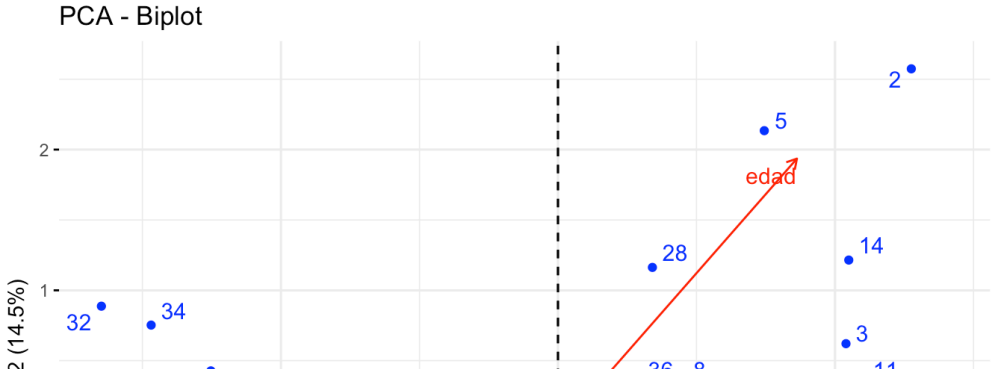
El screeplot ayuda a determinar cuántos componentes principales se deben conservar. Generalmente, se eligen los componentes que explican un porcentaje significativo de la varianza total. Notemos que con los primeros dos componentes se puede explicar más del 90% de la variabilidad total de los datos, lo que suele ser suficiente para interpretar los patrones principales. Los componentes adicionales no aportan mucha información adicional relevante.

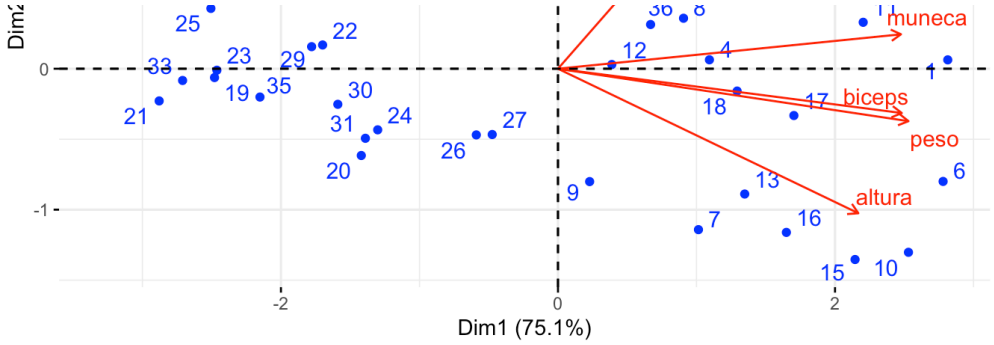
```
fviz_contrib(cpR, choice = c("var"))
```



Este gráfico muestra qué variables tienen una mayor influencia en la construcción de los componentes principales. Las variables con mayor contribución son las que más influyen en el PCA. En el caso de la gráfica con la matriz de correlación, las variables que más contribuyen son Peso, Biceps y Muñeca, mientras que la influencia de Edad y Altura no son indispensable guardar dichas variables.

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





El biplot es útil para ver simultáneamente cómo las variables influyen en las observaciones. Permite identificar relaciones entre variables y grupos de observaciones, así como patrones de agrupamiento.

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(PCA_result): Resumen detallado de los resultados del análisis, incluyendo varianza explicada y correlaciones.
- PCA_result$eig: Muestra los autovalores (varianza explicada por cada componente).
- PCA_resultvarcoord: Coordenadas de las variables en el espacio de componentes principales.
- PCA_resultindcoord: Coordenadas de los individuos en el espacio de componentes principales.
- plot.PCA(): Genera gráficos personalizados de los resultados del PCA.
- dimdesc(): Proporciona una descripción de las dimensiones principales.

Parte IV

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

1. Compare los resultados obtenidos con la matriz de varianza-covarianza y con la correlación . ¿Qué concluye? ¿Cuál de los dos procedimientos aporta componentes con de mayor interés?

La principal diferencia entre el uso de la matriz de varianza-covarianza y la matriz de correlación radica en cómo tratan la escala de las variables:

Matriz de varianza-covarianza: Cuando las variables están en diferentes escalas (como edad, peso, y altura), el análisis basado en la varianza-covarianza podría estar sesgado hacia aquellas variables con una mayor dispersión o rango. En este caso, variables como el peso y la altura tienen varianzas mucho mayores que la edad o el perímetro del bíceps, lo que les otorga un peso excesivo en los componentes principales. Esto puede ocultar la relación real entre las variables. Matriz de correlación: Este método estandariza las variables antes de realizar el análisis, lo que asegura que cada variable tenga la misma influencia en la formación de los componentes principales, independientemente de sus unidades de medida o varianza original. Por lo tanto, la matriz de correlación es preferible cuando las variables tienen diferentes escalas o unidades, como es el caso en este análisis.

2. Indique cuál de los dos análisis (a partir de la matriz de varianza y covarianza o de correlación)

resulta mejor para los datos indicadores económicos y sociales del 96 países en el mundo. Comparar los resultados y argumentar cuál es mejor según los resultados obtenidos.

No se entiende la pregunta. Pero en la respuesta anterior se menciona con qué matriz se obtuvo mejores resultados y por qué.

3. ¿Qué variables son las que más contribuyen a la primera y segunda componentes principales del método seleccionado? (observa los coeficientes en valor absoluto de las combinaciones lineales, auxíliate también de los gráficos)

Según el análisis realizado con la matriz de correlación, las variables que más contribuyen a los primeros dos componentes principales (CP1 y CP2) son:

Componente principal 1 (CP1): Las variables peso, muñeca y bíceps son las que más influyen, con coeficientes cercanos al 50% en valor absoluto. Esto sugiere que CP1 está relacionado principalmente con variables corporales, específicamente con el tamaño y masa corporal de los estudiantes. Componente principal 2 (CP2): Las variables edad y altura son las que más contribuyen a este componente, con la edad dominando significativamente. Esto indica que CP2 captura variación relacionada con la edad de los estudiantes y, en menor medida, con su altura.

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

Para las combinaciones lineales de los primeros dos componentes principales (CP1 y CP2), obtenidos con la matriz de correlación, las ecuaciones son:

```
CP1 = -Edad*0.336 - Peso*0.4927 - Altura*0.4222 - Muneca*0.4822 - Biceps*0.4833

CP2 = Edad*0.8575 - Peso*0.1647 - Altura*0.4542 + Muneca*0.1082 - Biceps*0.1392
```

Estas combinaciones permiten reducir el número de variables originales a dos componentes principales, que explican gran parte de la variabilidad de los datos.

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

En este análisis, los componentes principales pueden interpretarse en términos de agrupación de variables:

Componente principal 1 (CP1): Agrupa las variables relacionadas con el tamaño y la masa corporal de los estudiantes. Es una medida que puede interpretarse como un “índice de masa corporal” o un indicador de la corpulencia general, dado que peso, muñeca y bíceps son las principales variables que lo componen. Componente principal 2 (CP2): Está relacionado principalmente con la edad de los estudiantes y, en menor medida, con la altura. Este componente podría interpretarse como un “índice de edad y crecimiento”, capturando las diferencias en la madurez física entre los estudiantes.

Conclusión

El análisis de componentes principales basado en la matriz de correlación es más adecuado para los datos corporales de los estudiantes debido a las diferentes escalas de las variables. Los primeros dos componentes principales proporcionan un resumen útil de las características físicas de los estudiantes, permitiendo interpretar las relaciones entre variables clave como el peso, la altura, la edad y otras medidas corporales.