A4-Componentes Principales

Oskar Arturo Gamboa Reyes

2024-10-10

Análisis Descriptivo

```
M = read.csv("corporal.csv")
M = M[, -which(names(M) == "sexo")]




n=5 #número de variables

d=matrix(NA,ncol=7,nrow=n)
for(i in 1:n){
  d[i,]<-c(as.numeric(summary(M[,i])),sd(M[ ,i]))
}
m=as.data.frame(d)

row.names(m)=c('edad', 'altura','peso', 'muneca','biceps')
names(m)=c("Minimo","Q1","Mediana","Media","Q3","Máximo","Desv Est")

print(m)

##         Minimo      Q1 Mediana      Media     Q3 Máximo  Desv Est
## edad      19.0  24.750   28.00  31.44444  37.00   65.0 10.554469
## altura    42.0  54.950   71.50  68.95278  82.40   98.2 14.868999
## peso     147.2 164.800  172.70 171.55556 179.40  190.5 10.520170
## muneca     8.3   9.475   10.65  10.46667  11.50   12.4  1.175463
## biceps    23.5  25.975   32.15  31.16667  35.05   40.4  5.234392
```

Matrices de Varianza y Correlación

```
S = cov(M)
R = cor(M)

eigen_r = eigen(R)
eigen_s = eigen(S)

print(eigen_s)

## 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

print(eigen_r)

## 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

var_total_s = sum(diag(S))
var_total_r = sum(diag(R))
proporcion_var_s = eigen_s$values/var_total_s
proporcion_var_r = eigen_r$values/var_total_r
print("Valor de variables S")

## [1] "Valor de variables S"

print(diag(S))

##       edad       peso     altura     muneca     biceps
## 111.396825 221.087135 110.673968   1.381714  27.398857

print("Valor de componentes R")

## [1] "Valor de componentes R"

print(diag(R))

##   edad   peso altura muneca biceps
##      1      1      1      1      1

print("Proporción de variables S")

## [1] "Proporción de variables S"

cumsum(proporcion_var_s)

## [1] 0.7615357 0.9318456 0.9903763 0.9995034 1.0000000

print("Proporción de variables R")

## [1] "Proporción de variables R"
```

```r
cumsum(proporcion_var_r)

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

Los componentes más importantes son los primeros dos ya que tienen los valores de eigen más altos, por lo que explican mayor proporción de la varianza.

```r
etiquetas = colnames(M)
# Combinación lineal de CP1
CP1_coeficientes_s = eigen_s$vectors[,1]
CP1_combinacion_s = data.frame(Variable = etiquetas, Coeficiente_CP1_s =
CP1_coeficientes_s)

# Combinación lineal de CP2
CP2_coeficientes_s = eigen_s$vectors[,2]
CP2_combinacion_s = data.frame(Variable = etiquetas, Coeficiente_CP2_s =
CP2_coeficientes_s)

# Imprimir los resultados
print(CP1_combinacion_s)

##   Variable Coeficiente_CP1_s
## 1     edad       -0.34871002
## 2     peso       -0.76617586
## 3   altura       -0.47632405
## 4   muneca       -0.05386189
## 5   biceps       -0.24817367

print(CP2_combinacion_s)

##   Variable Coeficiente_CP2_s
## 1     edad         0.9075501
## 2     peso        -0.1616581
## 3   altura        -0.3851755
## 4   muneca         0.0155423
## 5   biceps        -0.0402221

# Combinación lineal de CP1
CP1_coeficientes_r = eigen_r$vectors[,1]
CP1_combinacion_r = data.frame(Variable = etiquetas, Coeficiente_CP1_r =
CP1_coeficientes_r)

# Combinación lineal de CP2
CP2_coeficientes_r = eigen_r$vectors[,2]
CP2_combinacion_r = data.frame(Variable = etiquetas, Coeficiente_CP2_r =
CP2_coeficientes_r)

# Imprimir los resultados
print(CP1_combinacion_r)
```

```
##   Variable Coeficiente_CP1_r
## 1     edad        -0.3359310
## 2     peso        -0.4927066
## 3   altura        -0.4222426
## 4   muneca        -0.4821923
## 5   biceps        -0.4833139
```

```
print(CP2_combinacion_r)
```

```
##   Variable Coeficiente_CP2_r
## 1     edad         0.8575601
## 2     peso        -0.1647821
## 3   altura        -0.4542223
## 4   muneca         0.1082775
## 5   biceps        -0.1392684
```

Podemos ver que en el primer componente las variables edad, peso y altura son más dominantes, mientras que en el segundo modelo es la basicamente la edad y un poco de la altura.

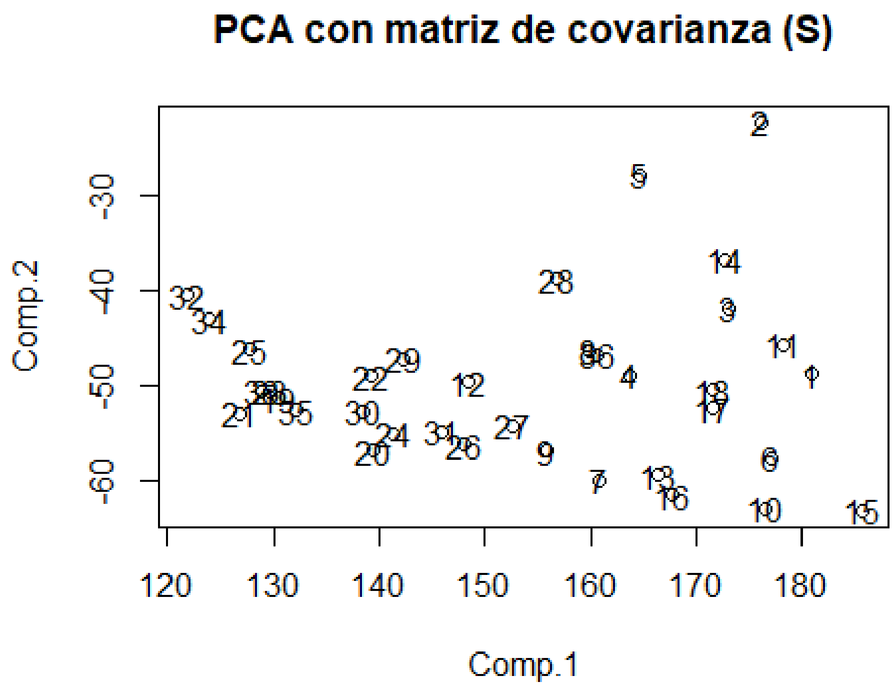
Gráficos de variables

Scores con matriz de varianza-covarianza

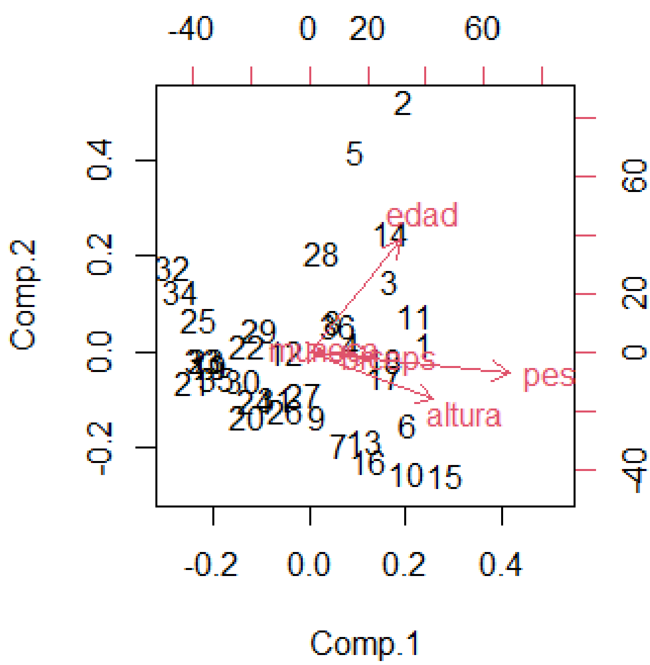
```
cpS = princomp(M, cor = FALSE)


cpaS = as.matrix(M) %*% cpS$loadings

plot(cpaS[,1:2], type = "p", main = "PCA con matriz de covarianza (S)")
text(cpaS[,1], cpaS[,2], labels = 1:nrow(cpaS))
```




```r
biplot(cpS)
```



En la matriz de varianza podemos ver que las variables más determinanates son la de edad, peso y altura. Edad se encuentra en una diagonal casi perfecta por lo que significa que tiene determinancia en los dos componentes, mientras que peso es casi exclusivamente determinante en el primer componente, mientras que altura aunque tiene una gran influencia en la varianza del primer componente también afecta el segundo.

Ahora los datos podemos ver que están dentro del radio de los vectores, pero del lado opuesto a la dirección, esto se debe a que los coeficientes son negativos.

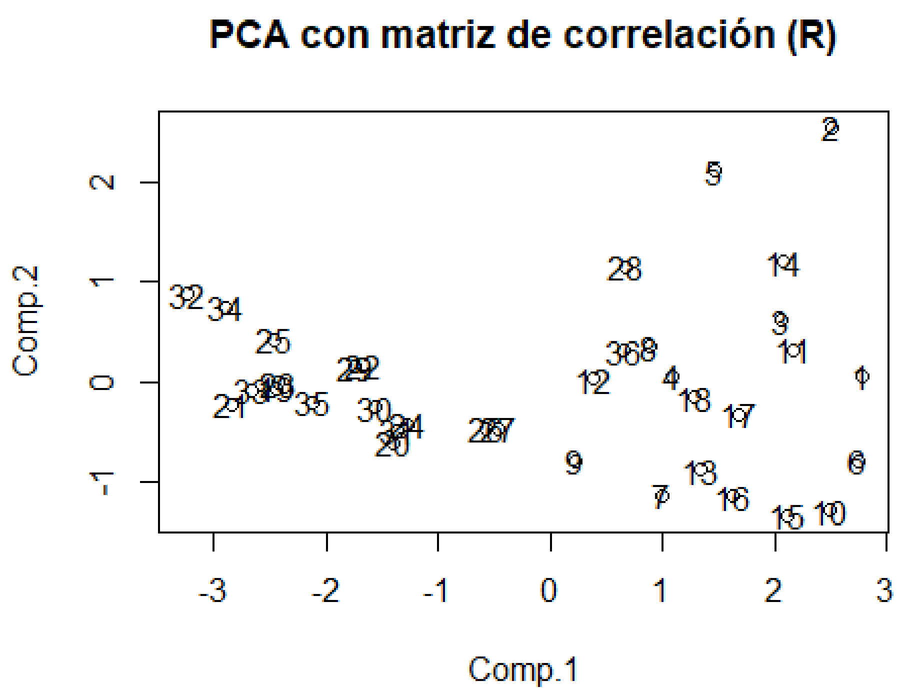
Además podemos ver que existen un par de datos atipicos que su varianza no puede ser explicada por estas variables, lo más probables es que sean casos extremos que no siguen cierto comportamiento.

Scores con matriz de correlación

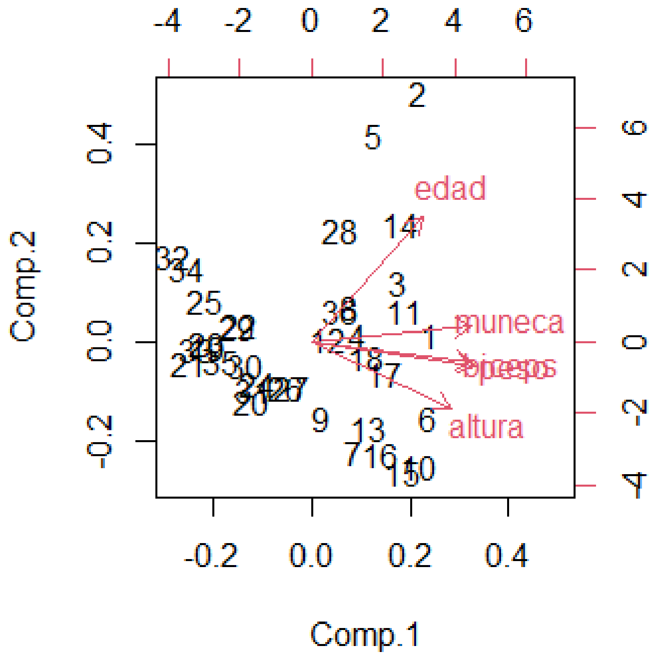
```
cpR <- princomp(M, cor = TRUE)


cpaR <- as.matrix(scale(M)) %*% cpR$loadings

plot(cpaR[,1:2], type = "p", main = "PCA con matriz de correlación (R)")
text(cpaR[,1], cpaR[,2], labels = 1:nrow(cpaR))
```




```
biplot(cpR)
```



En la matriz de correlación podeoms ver que la edad tiene un comportamiento similar, al ser incluido en los dos componentes, la altura también tiene una determinación en la correlación de los dos componentes, sin embargo la muñeca, los biceps y el peso son variables que solo tienen influencia en el primer componente, lo que explicaría la gran importancia que tiene este componente.

Los datos estan agrupados de manera similar, principalmente opuestos a las variables. También podemos ver un par de datos atipicos que también están señalados en la matriz de varianza.

Exploración de función princomp()

```
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$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
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
```

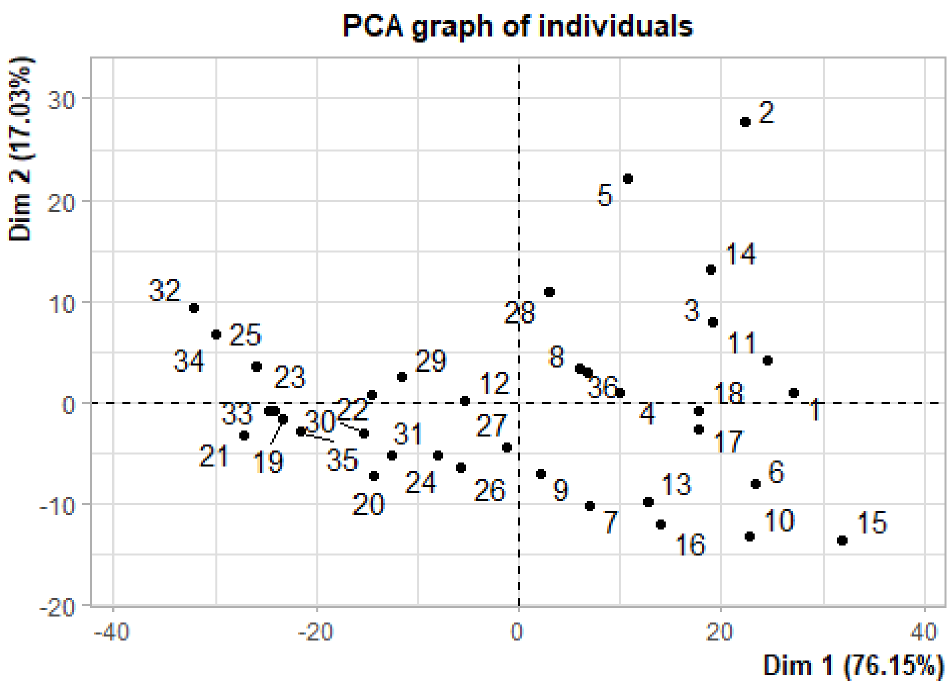
La primera función ayuda a ver la varianza explicada por cada componente, que fue lo mismo que encontramos en la primera parte del ejercicio, podemos ver que con dos es suficiente para explicar la varianza de las variables.

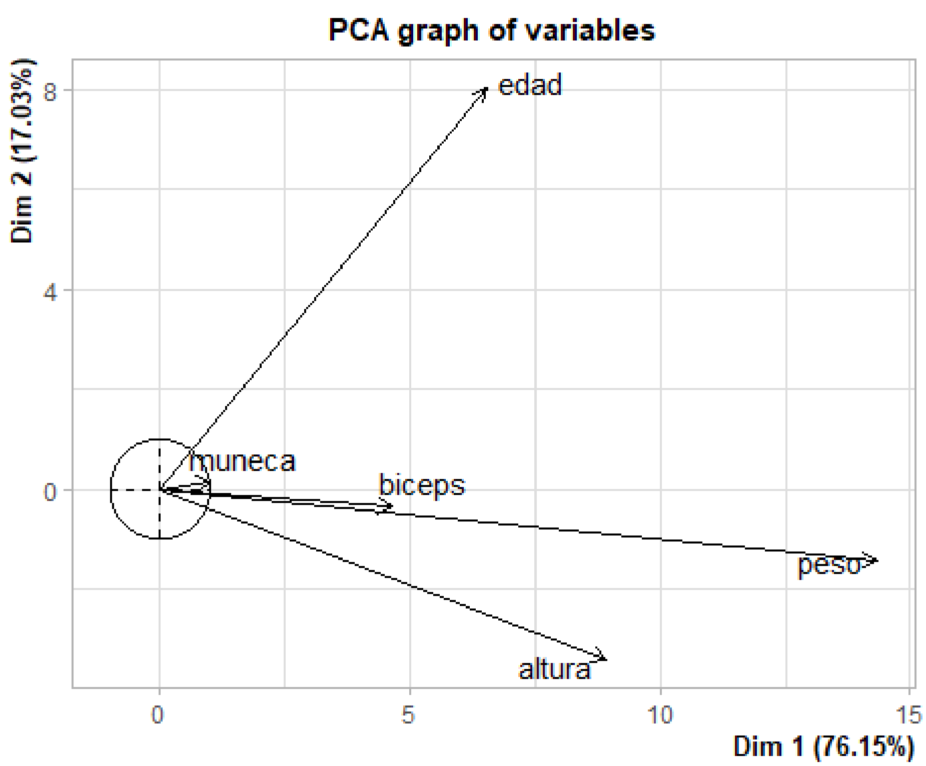
La segunda muestra que variables contribuyen a cada componente, obtenemos que el peso, la altura y la edad contribuyen al primer componente mientras que la edad explica casi todo el segundo componente.

La última función muestra las nuevas coordenadas de las observaciones proyectadas en el espacio de los componentes principales.

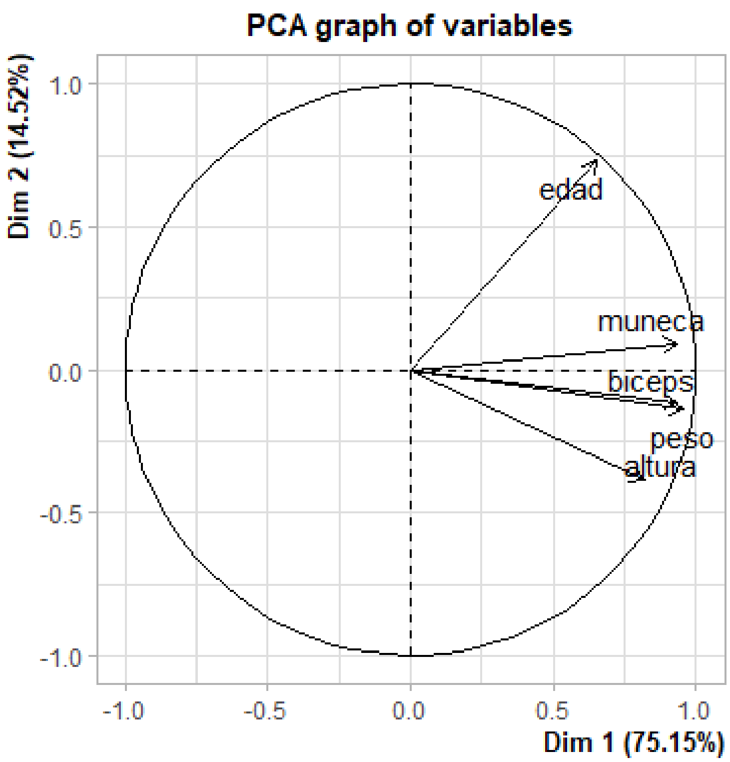
Gráficas de Componentes principales

```r
library(FactoMineR)
library(ggplot2)
datos=M
cpS = PCA(datos,scale.unit=FALSE)
```





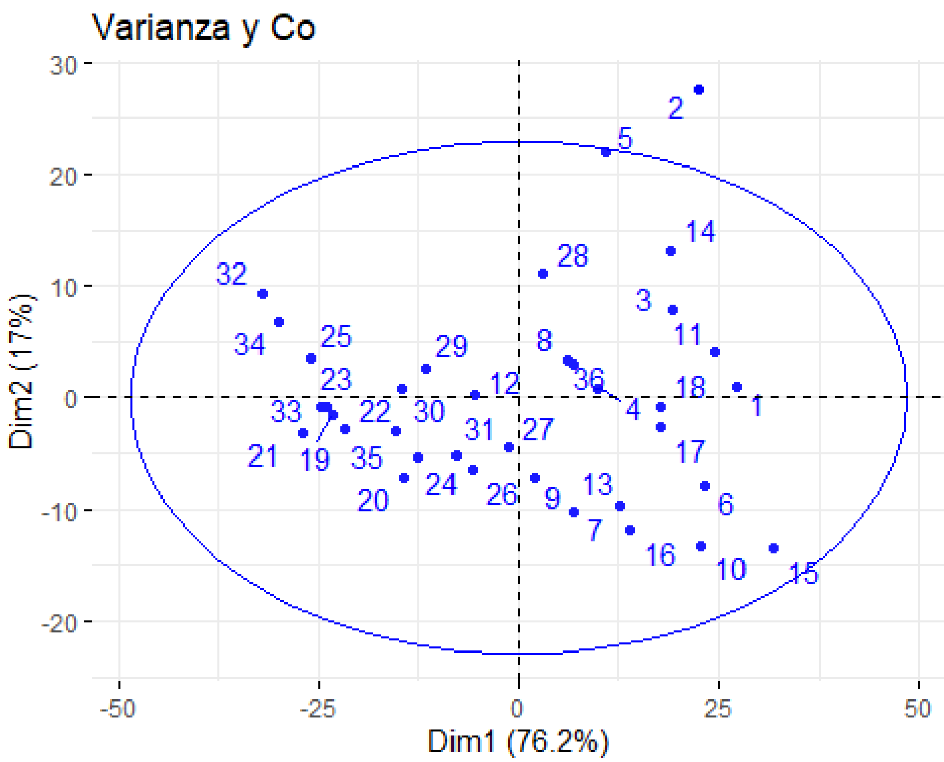

```
cpR = PCA(datos,scale.unit=TRUE)
```



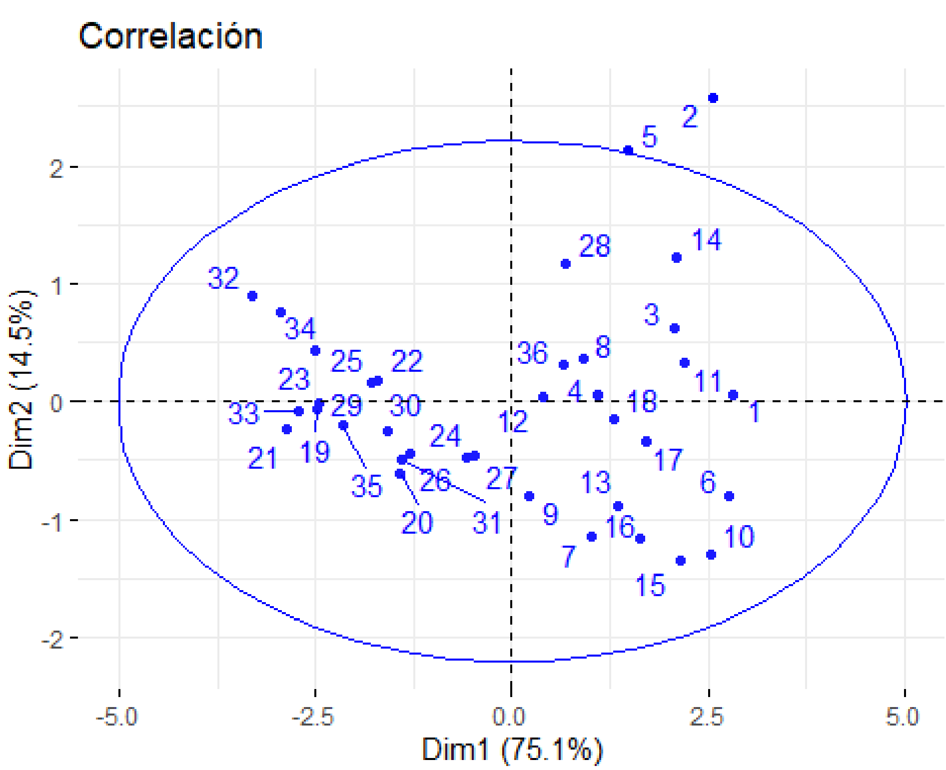

```
library(factoextra)

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

fviz_pca_ind(cpS, col.ind = "blue", addEllipses = TRUE, repel = TRUE,
title="Varianza y Co")
```



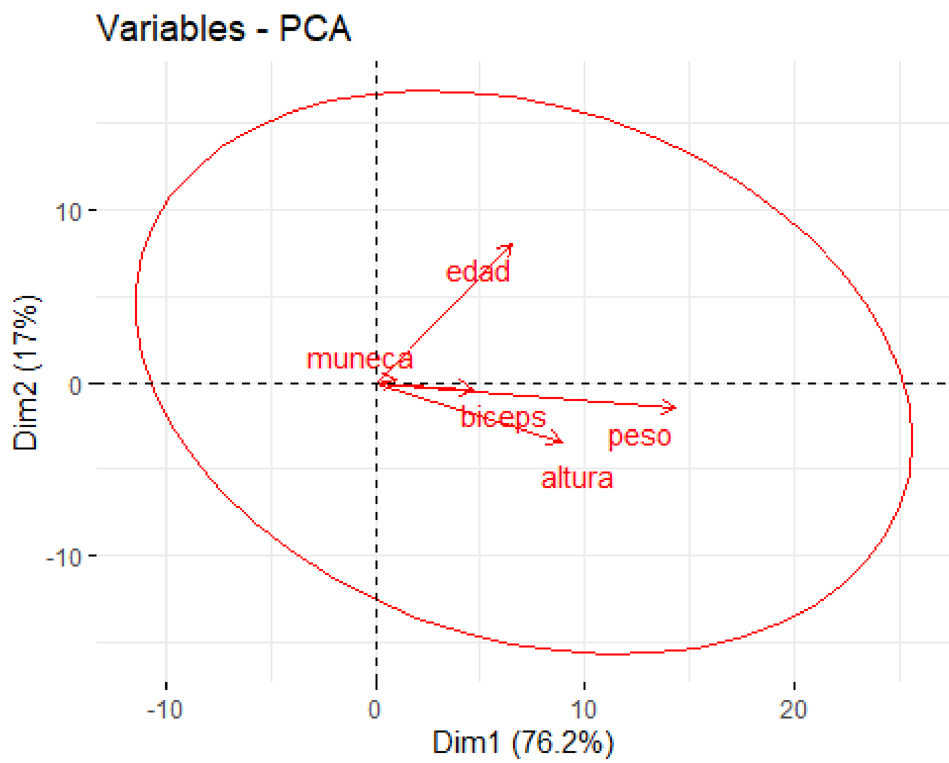

```r
fviz_pca_ind(cpR, col.ind = "blue", addEllipses = TRUE, repel = TRUE,
title="Correlación")
```



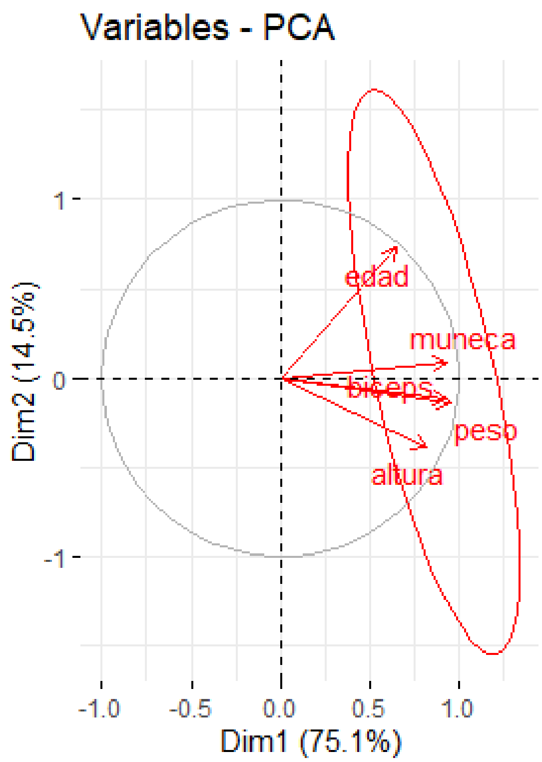
1. La primera gráfica muestra las observaciones, las que están cerca comparten caracteristicas similares

en las variables. El elipse nos ayuda a ver agrupaciones posibles. Aqui podemos ver que todos los datos están compartiendo un espacio ya que tienen las mismas caracteristicas, el unico dato atípico que podemos encontrar con esta gráfica es el numero 2.

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

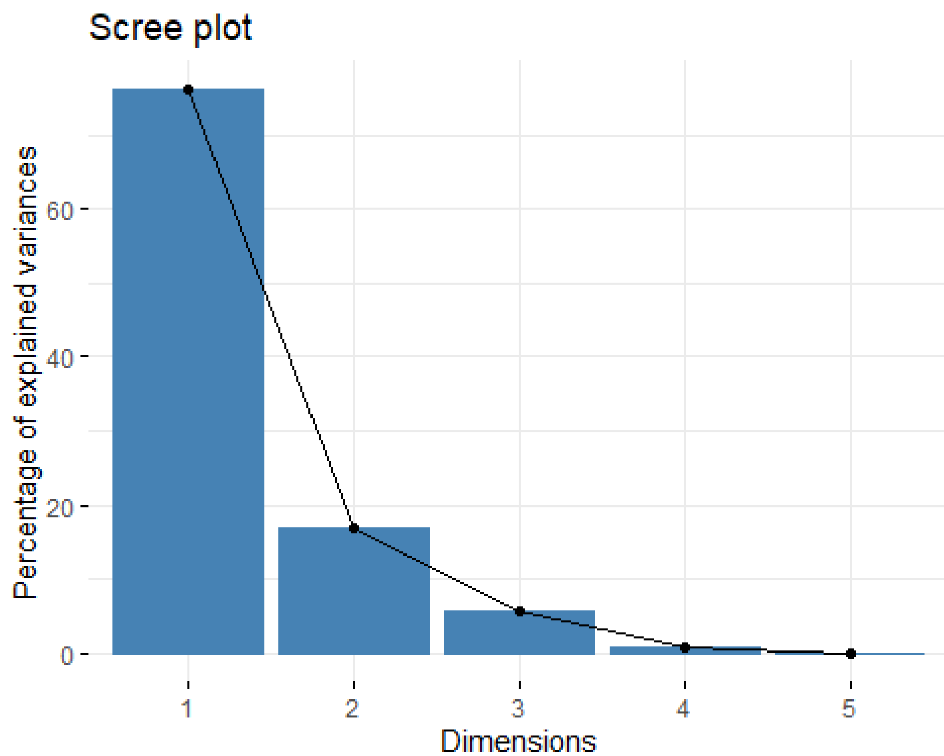



```
fviz_pca_var(cpR, col.var = "red", addEllipses = TRUE, repel = TRUE)
```

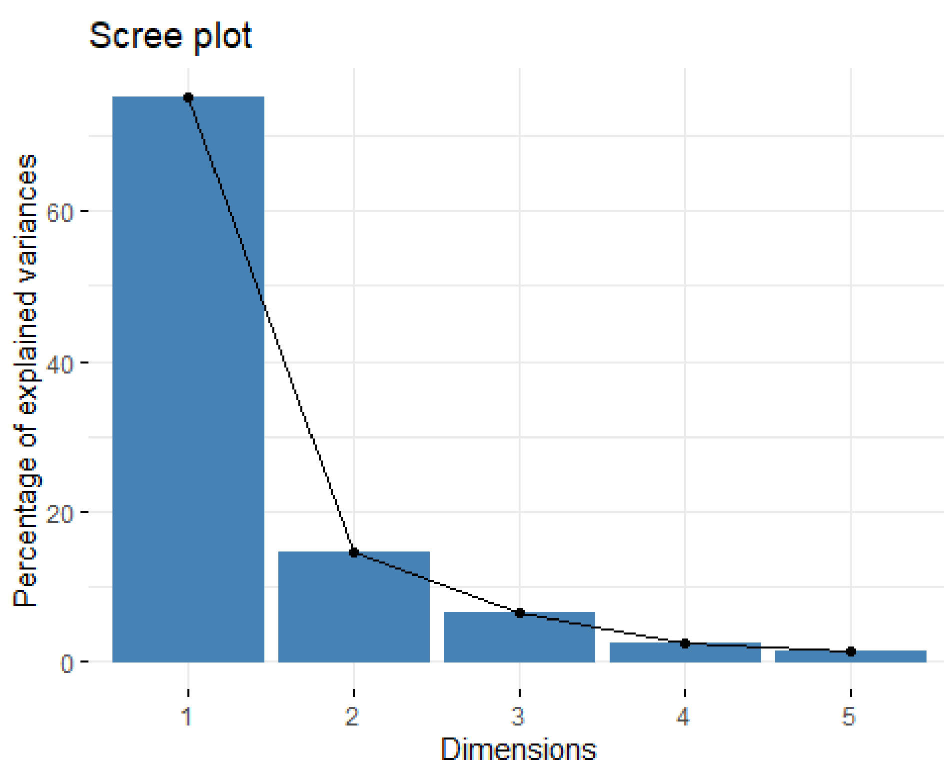


2. Esta gráfica muestra la correlacionalidad entre variables, mientras más cercanas tienen una mayor correlación y mientras más grandes sean mayor afectan a el componente.

```
fviz_screeplot(cpS)
```

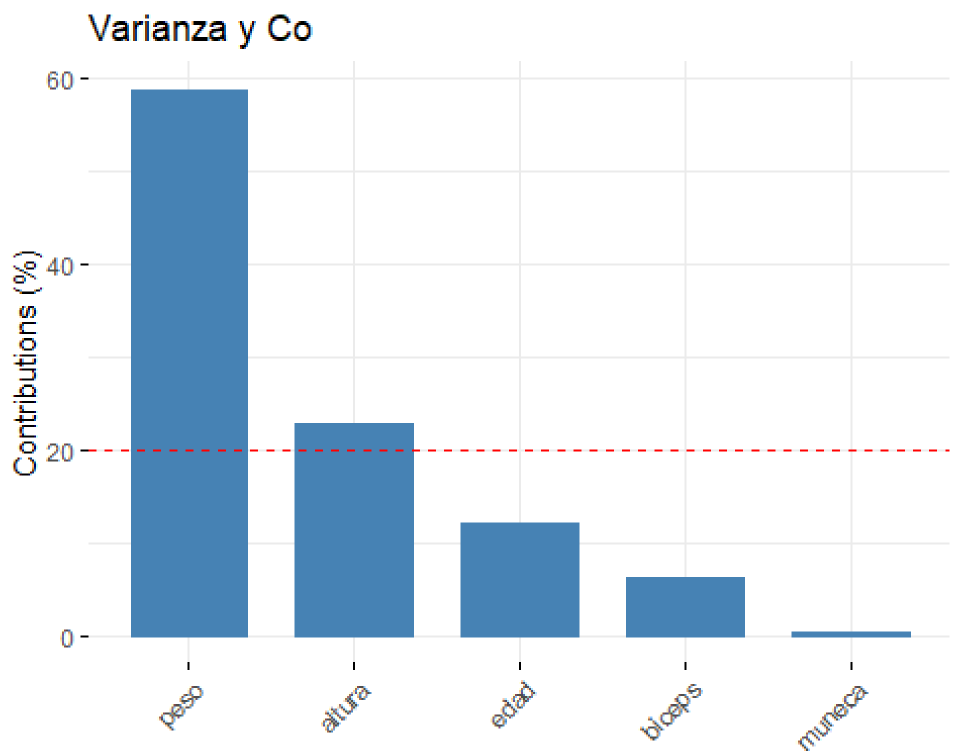



```
fviz_screeplot(cpR)
```

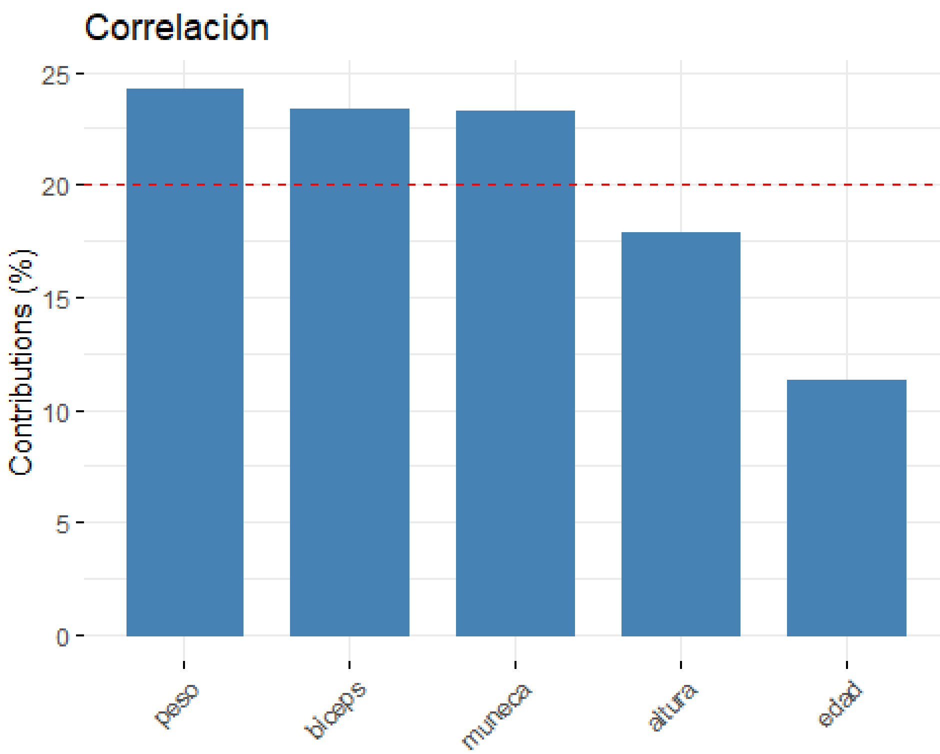


3. Este gráfico muestra la varianza o correlación explicada por cada componente, como podemos ver el primero explica la mayoría de la varianza y si incluimos al segundo obtenemos casi toda la varianza.

```r
fviz_contrib(cpS, choice = c("var"), title="Varianza y Co")
```

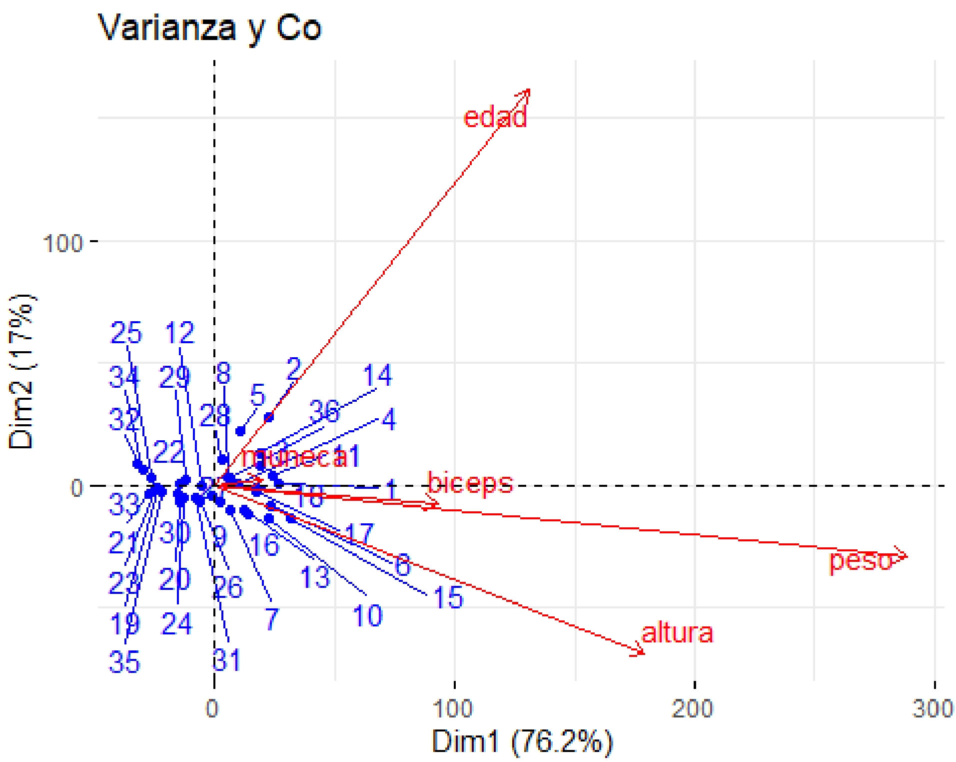



```r
fviz_contrib(cpR, choice = c("var"), title="Correlación")
```

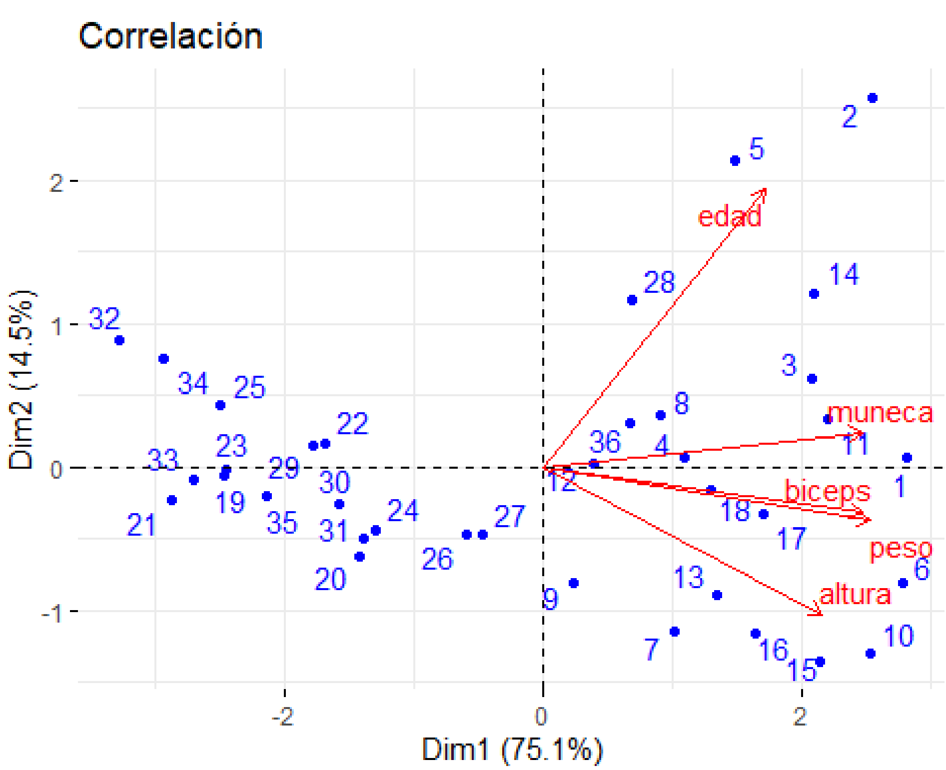


4. Muestra la aportación de cada variable al primer componente, el peso y la altura determinan casi toda la varianza, mientras que en la correlación la altura y la edad son los menos determinantes.

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




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



5. Esta gráfica es basicamente la misma que encontramos en la parte 2 de esta actividad. Los puntos azules

muestran los datos y su relación entre ellos y con las variables, mientras que las lineas rojas muestran la influencia de las variables en los componentes principales, la edad tiene efecto en los dos componentes, mientras que los demás influyen principalmente en el primero.

Conclusiones

Creo que para esta actividad es más apropiado usar el método de correlación ya que todas las variables tienen escalas diferentes, esto nos permite que cada variable tenga el mismo peso en el análisis, llegando a un resultado más certero y con variables más significativas.

Haría un análisis con todas las variables, agrupando bíceps, muñeca y peso, mientras que altura y edad estarían en su propia categoría, esto a partir de que tienen un comportamiento diferente a las demás variables.