Examen

Juan Cantero Jimenez

2/8/2022

Primero se cargarán los datos, y se descarta la variable que contiene los nombres de los distintos paises. Estos se usan para nombrar a las distintas filas del data.frame

```
load("datosfinal2022.RData")
head(datosfinal)
```

```
##    country smoking_men alcohol2008 blood_pres_men2008 bmi_men fat_blood_men
## 96  Russia        70.1        16.2                126    22.9          4.70
## 11 Belarus        63.7        18.9                137    26.2          5.02
## 47 Hungary        45.7        16.1                128    25.1          4.31
## 5  Armenia        55.1        13.7                135    25.4          4.71
## 36 Estonia        49.9        17.2                129    20.9          4.11
## 19  Canada        24.3        10.2                124    27.4          5.09
##    mort_c_men TM_Lung_men
## 96       43.1        62.1
## 11       45.2        65.3
## 47       33.0        94.3
## 5        35.1        70.1
## 36       28.6        61.6
## 19       13.2        48.3
```

```
paises <- datosfinal$country
datosfinal.num <- datosfinal[,-1]
rownames(datosfinal.num) <- paises
head(datosfinal.num)
```

```
##         smoking_men alcohol2008 blood_pres_men2008 bmi_men fat_blood_men
## Russia         70.1        16.2                126    22.9          4.70
## Belarus        63.7        18.9                137    26.2          5.02
## Hungary        45.7        16.1                128    25.1          4.31
## Armenia        55.1        13.7                135    25.4          4.71
## Estonia        49.9        17.2                129    20.9          4.11
## Canada         24.3        10.2                124    27.4          5.09
##         mort_c_men TM_Lung_men
## Russia        43.1        62.1
## Belarus       45.2        65.3
## Hungary       33.0        94.3
## Armenia       35.1        70.1
## Estonia       28.6        61.6
## Canada        13.2        48.3
```

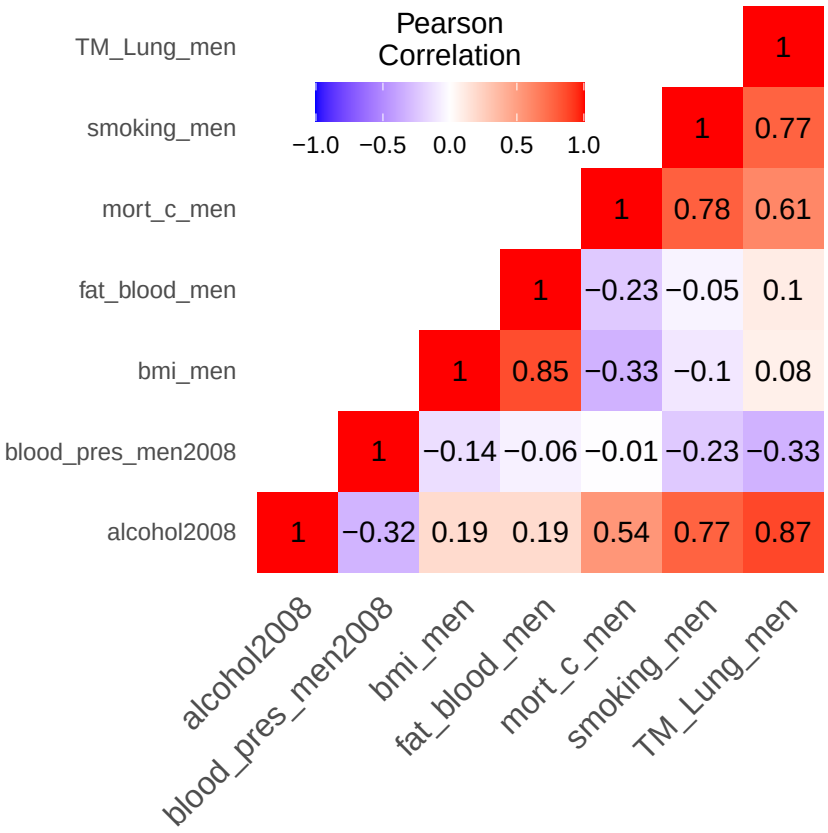
Ejercicio 1

1. Muestra UN RESULTADO (numérico o gráfico) que permita conocer la relación entre las variables del banco de datos y coméntalo brevemente.

```
correlation_ggplot <- function(data){ # Se crea una función que genera gráficos
  require(ggplot2)
  #de correlación para la variables numéricas en el argumento data, usando
  #el motor gráfico ggplot2
  tipos <- sapply(data, function(x){
    is.numeric(x)
  })#Se obtiene la posición de las columnas de tipo numerico en el data.frame

  numeric_names <- sort(names(data)[tipos])#Se seleccionan los nombres de las
  #variables de tipo factor. Notese que se han ordenado los nombres, esto es
  #necesario debido a que ggplot2 ordenará posteriormente las variables a
  #representar.
  data_new <- data[,numeric_names] #Se crea un data.frame adicional que
  #facilitará la representación con ggplot2
  correlation_mat <- round(cor(data_new),2)#Se crea la matriz de correlación
  #en este caso con la función cor

  correlation_mat[upper.tri(correlation_mat)] <- NA #Se elimina la información
  # de la diagonal superior, pues esta repetida
  correlation_mat <- t(correlation_mat) # Por conveniencia se transpone la
  #matriz



  melt_corr <- data.frame(list(Var1=rep(numeric_names, length(numeric_names)),
                               Var2=rep(numeric_names,
                                        rep(length(numeric_names),
                                            length(numeric_names))),
                               value=rep(NA, length(numeric_names)^2)))
  #Se crea un data.frame auxiliar con toda la informacióna representar. Este
  #se encuentra vacio y se rellenará con el bucle for que se encuentra más abajo.
  for (x in 1:nrow(melt_corr)){
    if(!is.na(correlation_mat[melt_corr[x, 2], melt_corr[x,1]])){
      melt_corr[x, 3] <- correlation_mat[melt_corr[x, 2], melt_corr[x,1]]
    }
  }#Se rellena el campo value del dataframe creado con la información presente
  #en correlation_mat gracias a la información aportada por las variables
  #Var1 y Var2

  melt_corr$Var1 <- as.factor(melt_corr$Var1)
  melt_corr$Var2 <- as.factor(melt_corr$Var2)
  melt_corr$value <- as.numeric(melt_corr$value)#Se convierten las variables
  #al tipo neceario
  melted_cormat <- melt_corr[!is.na(melt_corr$value),]#Se eliminan los NA

  h <- ggplot(data = melted_cormat, aes(Var1, Var2, fill = value))+ #Se definen
    #el data.set de donde saldrán los datos, así como que hacer con cada
    #variable
    geom_raster() + #Se genera un geom de tipo raster, una imagen.
```

```
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1,1), space = "Lab",
                         name="Pearson\nCorrelation") + #Se crea una escala
    #de color para usar en la imagen creada con geom_raster.
    theme_minimal()+
    theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                     size = 12, hjust = 1))+#Se define como se
    #representará la etiqueta del eje x.
    coord_fixed()#Se fijan las coordenadas del plot creado.
  j <-  h +
    geom_text(aes(Var1, Var2, label = value), color = "black", size = 4) +
    theme(
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      panel.grid.major = element_blank(),
      panel.border = element_blank(),
      panel.background = element_blank(),
      axis.ticks = element_blank(),
      legend.justification = c(1, 0),
      legend.position = c(0.6, 0.7),
      legend.direction = "horizontal")+
    guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                                 title.position = "top", title.hjust = 0.5))
  #Se añaden los valores del coeficiente de correlación en la posición adecuada.
  print(j)#Se imprime el gráfico creado.

}

correlation_ggplot(datosfinal.num)
```

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



En la imagen anterior se muestra un plot de las correlaciones que poseen las distintas variables del banco de datos entre si. En este se puede aprecira como la variable TM_Lung_men se encuentra relacionada positivamente con smoking_men y alcohol2018, esto es lógico puesto que el consumo de alcohol y tabaco aumenta el riesgo de padecer cancer de pulmon. Tambien es lógico la correlación positiva entre smoking_men y mot_c_men puesto que el tabaco no solo aumenta el riesgo de padecer cancer de pulmon, sino que contribuye al desarrollo de otras enfermedades como cancer de boca o garganta entre otros.

2. ¿Te parece adecuado realizar un ACP sobre este banco de datos? Es decir, ¿la relación entre las variables de este banco de datos recomienda la aplicación de esta técnica o no? Justifica brevemente tu respuesta.

Si considero adecuado la realización de un ACP sobre el banco de datos puesto que las variables poseen correlación lineal entre ellas, aunque en distinto grado.

Ejercicio 2

1. Muestra UN RESULTADO (numérico o gráfico) que permita conocer la relación entre los individuos y coméntalo brevemente.

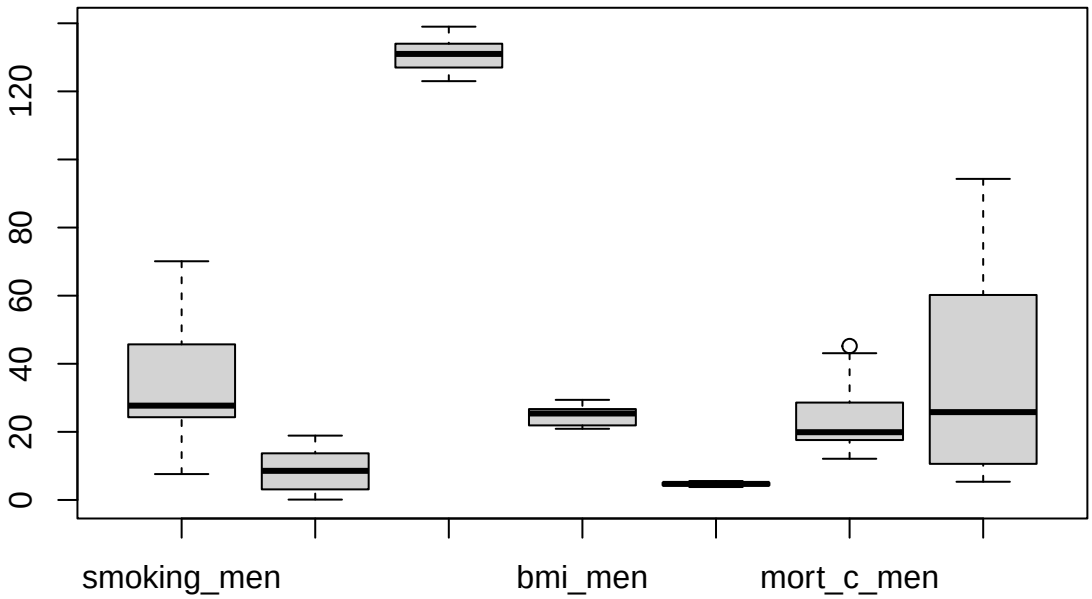
```
describe_custom <- function(data){
  require(e1071)
result <- apply(data, 2, function(x){


  c(media=mean(x),
    mediana=median(x),
    varianza = var(x),
    des_tipic = sd(x),
```

```
    skew = e1071::skewness(x),
    kurto = e1071::kurtosis(x),
    maximo = max(x),
    minimo = min(x),
    rango = max(x)- min(x),
    quantile(x , 0.25 ),
    quantile(x, 0.50),
    quantile(x, 0.75),
    shapiro_pvalor = shapiro.test(x)$p.value)

})
return(result)
}



describe_custom(datosfinal.num)### Debido a la heterogeneidad de las escalas, así como la diferencia en
```

```
## Loading required package: e1071
```

```
##                 smoking_men alcohol2008 blood_pres_men2008     bmi_men
## media            33.9880000  8.71240000        130.8400000 24.74400000
## mediana          27.7000000  8.55000000        131.0000000 25.40000000
## varianza        248.7477667 38.27894400         20.8900000  7.22006667
## des_tipic        15.7717395  6.18699798          4.5705580  2.68701817
## skew              0.5729497 -0.00770998          0.1394046 -0.10219206
## kurto            -0.4857348 -1.50048886         -1.1662420 -1.51195197
## maximo           70.1000000 18.90000000        139.0000000 29.40000000
## minimo            7.6000000  0.11000000        123.0000000 20.90000000
## rango            62.5000000 18.79000000         16.0000000  8.50000000
## 25%              24.3000000  3.11000000        127.0000000 21.90000000
## 50%              27.7000000  8.55000000        131.0000000 25.40000000
## 75%              45.7000000 13.70000000        134.0000000 26.70000000
## shapiro_pvalor    0.1684845  0.05886849          0.3733001  0.02814777
##                 fat_blood_men   mort_c_men   TM_Lung_men
## media               4.6732000 23.184000000  34.795600000
## mediana             4.7100000 19.900000000  25.800000000
## varianza            0.2252477 88.285566667 681.603659000
## des_tipic           0.4746026  9.396039946  26.107540271
## skew               -0.1713004  0.908977567   0.485648110
## kurto              -1.0942494 -0.330223279  -1.130678141
## maximo              5.5600000 45.200000000  94.300000000
## minimo              3.8000000 12.100000000   5.350000000
## rango               1.7600000 33.100000000  88.950000000
## 25%                 4.3100000 17.600000000  10.600000000
## 50%                 4.7100000 19.900000000  25.800000000
## 75%                 5.0700000 28.600000000  60.200000000
## shapiro_pvalor      0.6475618  0.009242466   0.005654532
```

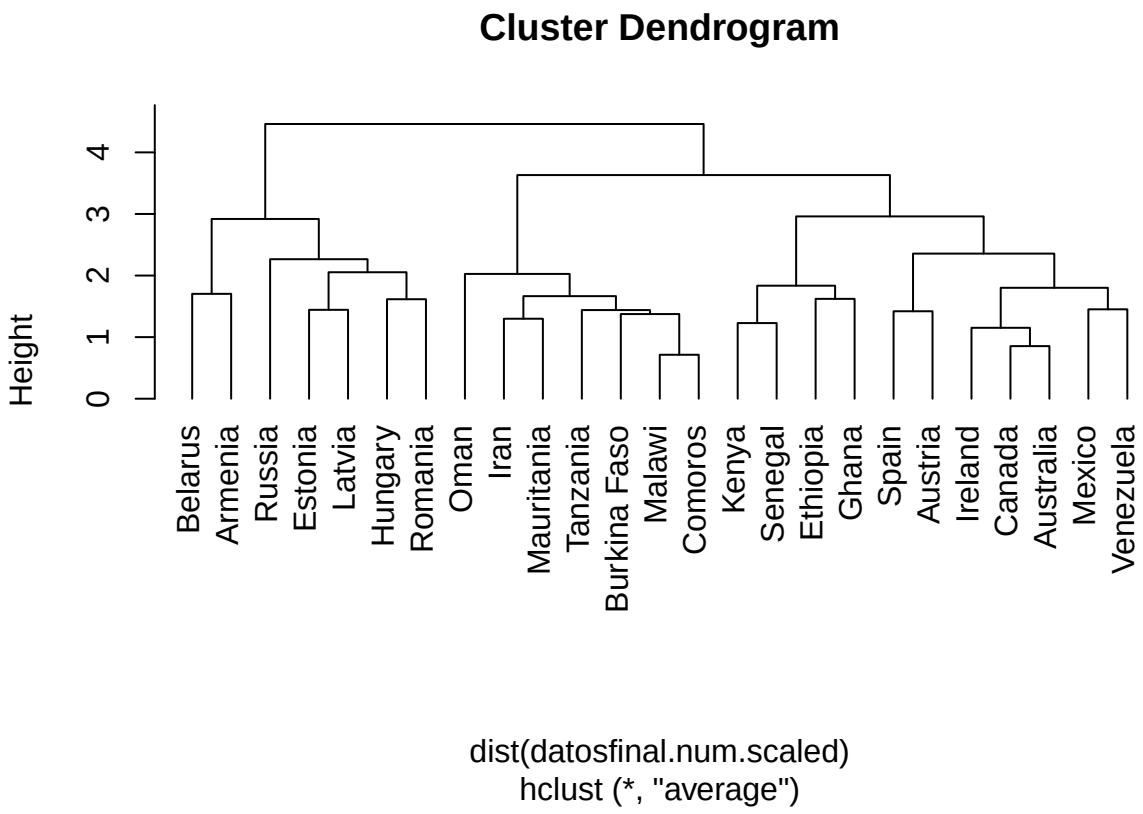
```
boxplot(datosfinal.num)
```




```
datosfinal.num.scaled <- scale(datosfinal.num)
hcl1 <- hclust(dist(datosfinal.num.scaled), method = "average")
cor(dist(datosfinal.num.scaled), cophenetic(hcl1))
```

```
## [1] 0.8255034
```

```
plot(hcl1,hang=-1)
```



Se ha decidido realizar un análisis de cluster jerarquico con un linkage de tipo average. Este posee una correlación cofenética de 0.82, por lo que se puede considerar una análisis cluster optimo. En el gráfico

se puede observar como paises con cercania geográfica, vease Burquina Faso o Mauritania, se encuentran proximos entre sí. Además es interesante como paises que han poseido una fuerte relación a lo largo de la historia reciente se encuentrán próximos entre si, vease Rusia, Estonia y Lituania.

2. ¿Te parece adecuado realizar un análisis de agrupamiento sobre este banco de datos? Justifica brevemente tu respuesta.

Si me parece adecuado realizar un análisis cluster del banco de datos puesto que permitirá obtener las relaciones entre las variables. El banco de datos posee una gran diferencia entres las escalas, pero esto es subsanable con un escalado. Además el banco de datos carece de gran cantidad de outlaiers que pudierán dificultar el aglomeramiento.

EJERCICIO 3: Análisis de componentes principales

1. ¿Crees que el ACP se debería realizar sobre la matriz de varianzas-covarianzas o sobre la de correlaciones? Justifica brevemente tu respuesta.

Puesto que existe una notable diferencia en las escalas de las distinta variables, así como en sus varianzas, la opción más óptima es realizarlo sobre la matriz de correlaciones.

2. Realiza el PCA y contesta qué porcentaje de varianza original del banco de datos explican las dos primeras componentes principales.

```
pca1<-princomp(datosfinal.num, cor=TRUE)
summary(pca1)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.8085007 1.4346620 0.9641982 0.57536976 0.40414434
## Proportion of Variance 0.4672392 0.2940364 0.1328112 0.04729291 0.02333324
## Cumulative Proportion  0.4672392 0.7612757 0.8940869 0.94137976 0.96471299
##                            Comp.6     Comp.7
## Standard deviation     0.37199801 0.32958538
## Proportion of Variance 0.01976893 0.01551807
## Cumulative Proportion  0.98448193 1.00000000
```

Las dos primeras componentes principales explican el 76 % de la varianza.

3. Interpreta brevemente como se define la primera componente principal.

```
pca1$loadings
```

```
##
## Loadings:
##                    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## smoking_men         0.512         0.102  0.244  0.701  0.259  0.322
## alcohol2008         0.499 -0.162        -0.473  0.145        -0.691
## blood_pres_men2008 -0.197  0.177  0.916 -0.288
## bmi_men                    -0.665  0.120         0.208 -0.679  0.188
## fat_blood_men              -0.642  0.249  0.315 -0.241  0.593 -0.130
## mort_c_men          0.434  0.272  0.270  0.602 -0.372 -0.343 -0.213
## TM_Lung_men         0.511                -0.409 -0.495         0.564
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
```

```
## Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
## Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```

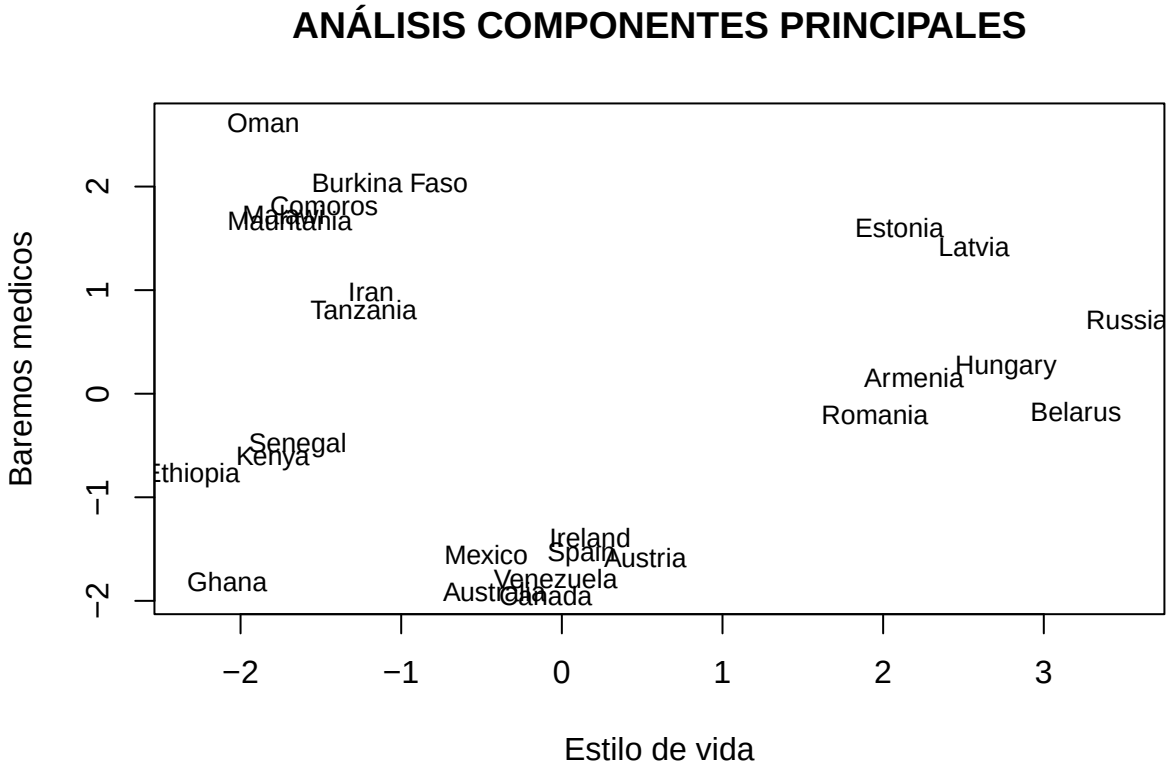
En la primera componente principal podemos observar como aquellas observaciones que posean una baja presion sanguínea, una alta tasa de consumo de alcohol, tabaco, así como una alta tasa de mortalidad en hombres ya sea general o provocada por el cancer de pulmon, obtendran valores positivos grandes. Se podría decir que esta columna resume la mortalidad debido a efectos de estilo de vida.

4. Interpreta brevemente como se define la segunda componente principal.

En esta segunda componente princpal podemos ver como paises que posean un escaso consumo de alcohol, bajo BMI así como un nivel bajo de colesterol en sangre y una mortalidad alta de hombres y una media de la presión sanguínea alta poseeran valores positivos grandes. Se podria decir que esta columna resumen ciertos baremos que describen el estado de salud general.

5. Realiza un gráfico en el que sitúes los paises sobre las dos primeras componentes principales y comenta brevemente el resultado en función de lo que representa cada componente principal y la situación de los paises en el gráfico.

```
plot(pca1$scores[,1], pca1$scores[,2], main = "ANÁLISIS COMPONENTES PRINCIPALES",xlab = "Estilo de vida
text(pca1$scores[,1], pca1$scores[,2], labels = rownames(datosfinal.num),cex=0.8)
```



Como se puede observar en el plot anterior, países del mundo islámico que no consumen alcohol, se encuentran a la izquierda del gráfico. Algo logico en función de la componente principal 1, en el gráfico Estilo de vida. Tambien es interesante observar el comportamiento en la segunda componente principal, en el gráfico Baremos Médicos, Irlanda, España, Asutria o Canada poseen un valor negativo muy alto. Esto puede indicar una baja mortalidad por cáncer o enfermedades cardiovasculares, y un alto BMI o colesterol total en sangre, esto es lógico pues ambas se pueden ver como consecuencia del grado de desarrollo de estos países.

Ejercicio 4: Analisis cluster.

1. ¿Crees que en este caso debes estandarizar las variables antes de realizar un análisis de agrupamiento o no? Justifica brevemente tu respuesta.

```
describe_custom(datosfinal.num)
```

```
##                smoking_men alcohol2008 blood_pres_men2008     bmi_men
## media           33.9880000  8.71240000        130.8400000 24.74400000
## mediana         27.7000000  8.55000000        131.0000000 25.40000000
## varianza       248.7477667 38.27894400         20.8900000  7.22006667
## des_tipic       15.7717395  6.18699798          4.5705580  2.68701817
## skew             0.5729497 -0.00770998          0.1394046 -0.10219206
## kurto           -0.4857348 -1.50048886         -1.1662420 -1.51195197
## maximo          70.1000000 18.90000000        139.0000000 29.40000000
## minimo           7.6000000  0.11000000        123.0000000 20.90000000
## rango           62.5000000 18.79000000         16.0000000  8.50000000
## 25%             24.3000000  3.11000000        127.0000000 21.90000000
## 50%             27.7000000  8.55000000        131.0000000 25.40000000
## 75%             45.7000000 13.70000000        134.0000000 26.70000000
## shapiro_pvalor   0.1684845  0.05886849          0.3733001  0.02814777
##                fat_blood_men   mort_c_men   TM_Lung_men
## media              4.6732000 23.184000000  34.795600000
## mediana            4.7100000 19.900000000  25.800000000
## varianza           0.2252477 88.285566667 681.603659000
## des_tipic          0.4746026  9.396039946  26.107540271
## skew              -0.1713004  0.908977567   0.485648110
## kurto             -1.0942494 -0.330223279  -1.130678141
## maximo             5.5600000 45.200000000  94.300000000
## minimo             3.8000000 12.100000000   5.350000000
## rango              1.7600000 33.100000000  88.950000000
## 25%                4.3100000 17.600000000  10.600000000
## 50%                4.7100000 19.900000000  25.800000000
## 75%                5.0700000 28.600000000  60.200000000
## shapiro_pvalor     0.6475618  0.009242466   0.005654532
```

Si puesto que las variables no poseen unidades similares, y la varianza varia mucho a lo largo de las distintas observaciones, vease las varianzas de las variables fat_blood_men, 0.22, y TM_lung_men, 681.60.

2. Realiza un análisis de agrupamiento jerárquico probando los algoritmos "ward.D2","single" y "average" y comprueba qué algoritmo obtiene una mayor correlación cofenética.

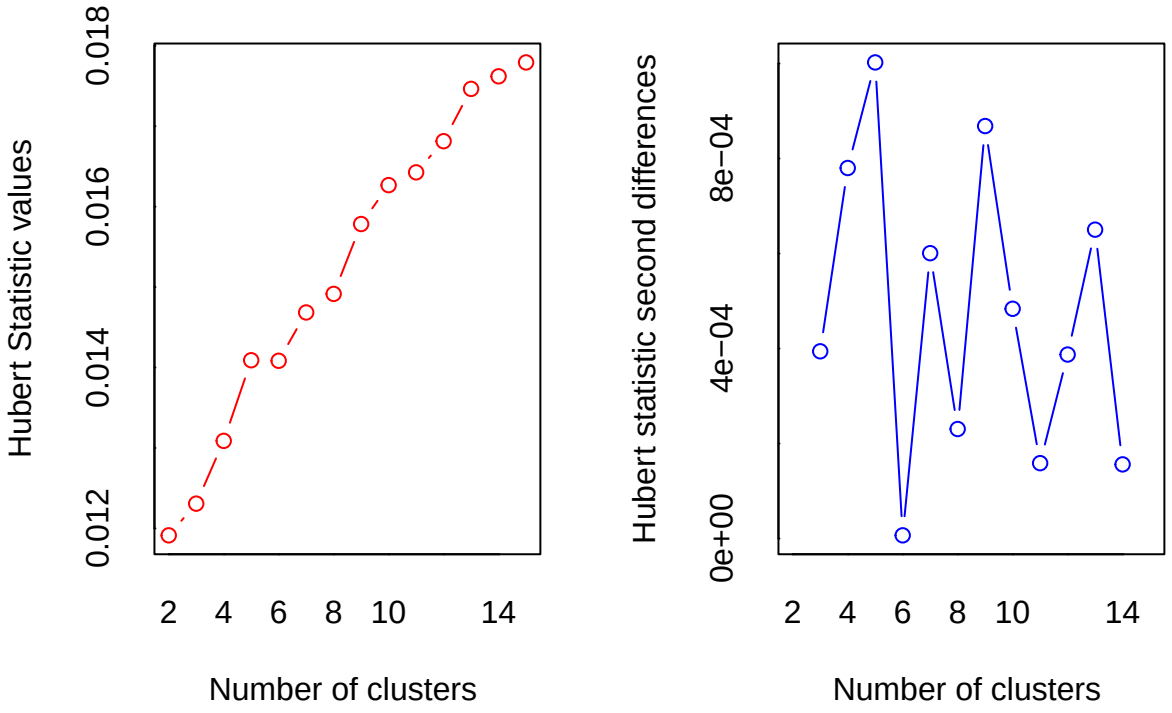
```
result <- sapply(c("ward.D2", "single","average"), function(x){
  clust <- hclust(dist(datosfinal.num.scaled), method = x)
  return(c( cor(dist(datosfinal.num.scaled), cophenetic(clust))))
})
result
```

```
##   ward.D2    single   average
## 0.8149534 0.7874055 0.8255034
```

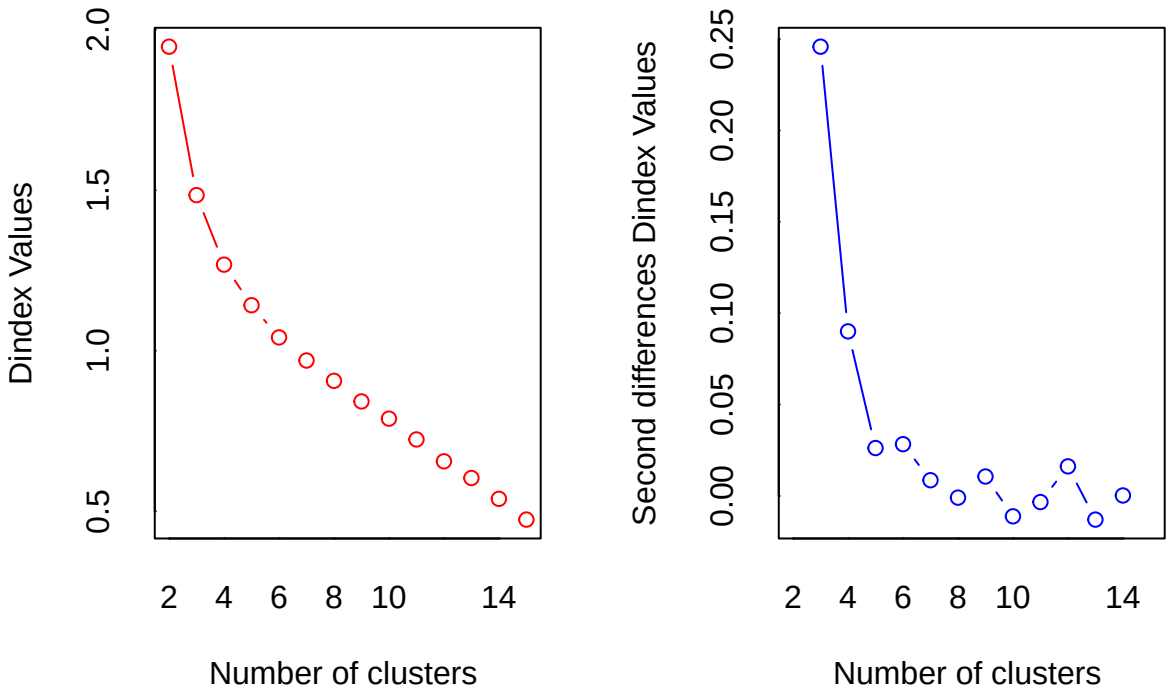
El metodo average obtiene la mayor correlación cofenética.

```
library(NbClust)
nbclust.complete <- NbClust(data=datosfinal.num.scaled,
diss=NULL,
distance="euclidean",
method="average")
```

```
## Warning in pf(beale, pp, df2): NaNs produced

## Warning in pf(beale, pp, df2): NaNs produced

## Warning in pf(beale, pp, df2): NaNs produced

## Warning in pf(beale, pp, df2): NaNs produced
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                In the plot of Hubert index, we seek a significant knee that corresponds to a
##                significant increase of the value of the measure i.e the significant peak in Hubert
##                index second differences plot.
##
```




```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 1 proposed 2 as the best number of clusters
## * 11 proposed 3 as the best number of clusters
## * 2 proposed 4 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 1 proposed 11 as the best number of clusters
## * 6 proposed 15 as the best number of clusters
##
##                   ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************
```

Según los resultados obtenidos el mejor número de particiones es 3.

4. Realiza un análisis de agrupamiento mediante el algoritmo de k-medias considerando como centroides iniciales los que se derivan de los grupos definidos como mejor partición en el apartado anterior. ¿Cambia la composición de los grupos formados o se mantiene igual o muy similar?

```
partition <- nbclust.complete$Best.partition
countri_clusters <-list(clust1=names(partition[partition==1]),
clust2=names(partition[partition==2]),
```

```
clust3=names(partition[partition==3]))
centroides <- t(as.data.frame( lapply(countri_clusters, function(x, y){
clust <- y[which(rownames(y) %in% x),]
apply(clust, 2, mean)
}, y = datosfinal.num.scaled)))

clust_kmeans <-kmeans(datosfinal.num.scaled,centroides,nstart = 100)

nbclust.complete$Best.partition
```

```
##       Russia      Belarus      Hungary      Armenia      Estonia       Canada
##            1            1            1            1            1            2
##    Australia       Mexico    Venezuela      Ireland         Iran     Ethiopia
##            2            2            2            2            3            2
##        Ghana        Spain       Malawi   Mauritania      Comoros Burkina Faso
##            2            2            3            3            3            3
##     Tanzania       Latvia        Kenya      Romania         Oman      Austria
##            3            1            2            1            3            2
##      Senegal
##            2
```

```
clust_kmeans$cluster
```

```
##       Russia      Belarus      Hungary      Armenia      Estonia       Canada
##            1            1            1            1            1            2
##    Australia       Mexico    Venezuela      Ireland         Iran     Ethiopia
##            2            2            2            2            3            2
##        Ghana        Spain       Malawi   Mauritania      Comoros Burkina Faso
##            2            2            3            3            3            3
##     Tanzania       Latvia        Kenya      Romania         Oman      Austria
##            3            1            2            1            3            2
##      Senegal
##            2
```

```
nbclust.complete$Best.partition == clust_kmeans$cluster
```

```
##       Russia      Belarus      Hungary      Armenia      Estonia       Canada
##         TRUE         TRUE         TRUE         TRUE         TRUE         TRUE
##    Australia       Mexico    Venezuela      Ireland         Iran     Ethiopia
##         TRUE         TRUE         TRUE         TRUE         TRUE         TRUE
##        Ghana        Spain       Malawi   Mauritania      Comoros Burkina Faso
##         TRUE         TRUE         TRUE         TRUE         TRUE         TRUE
##     Tanzania       Latvia        Kenya      Romania         Oman      Austria
##         TRUE         TRUE         TRUE         TRUE         TRUE         TRUE
##      Senegal
##         TRUE
```

Como se puede observar, ambas particiones son idénticas.

5 A partir de los clusters obtenidos mediante el algoritmo de k-medias del apartado 4, realiza un análisis exploratorio a tu elección que te permita caracterizar cada grupo en función de las variables disponibles

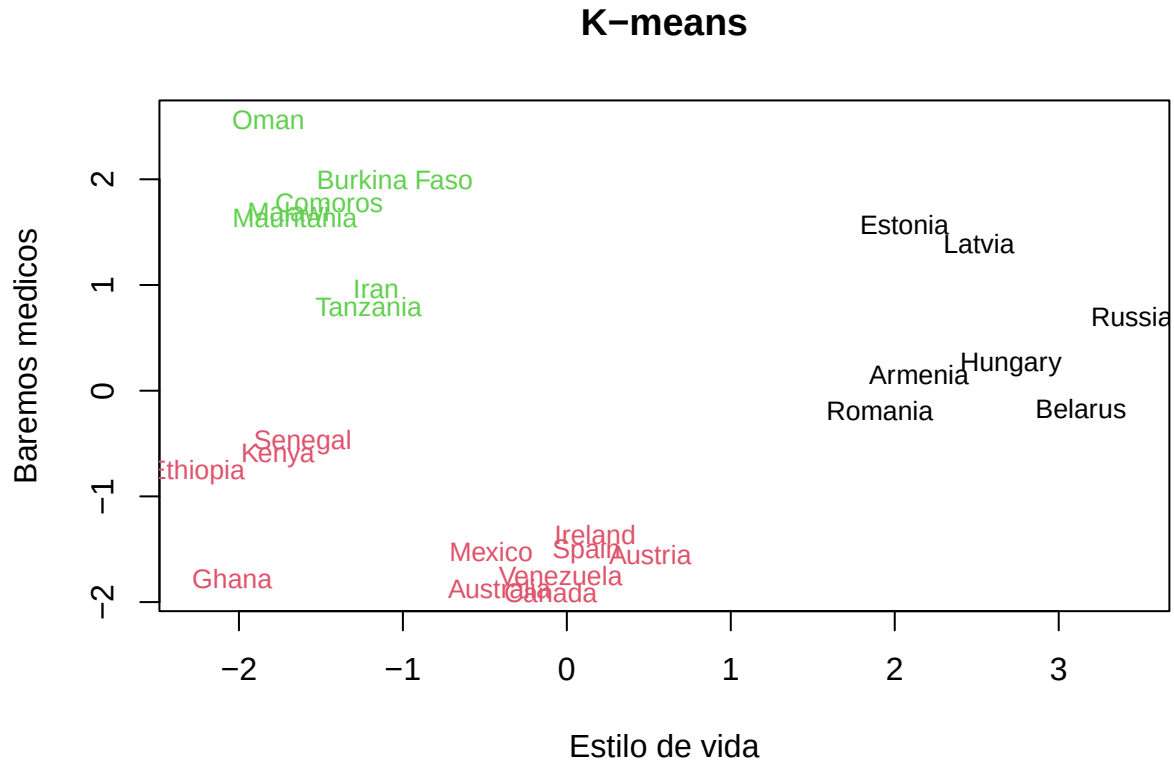
```
pca2 <- princomp(datosfinal.num.scaled)
summary(pca2)
```

```
## Importance of components:
##                            Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.7719615 1.4056760 0.9447175 0.56374493 0.39597897
## Proportion of Variance 0.4672392 0.2940364 0.1328112 0.04729291 0.02333324
## Cumulative Proportion  0.4672392 0.7612757 0.8940869 0.94137976 0.96471299
##                            Comp.6     Comp.7
## Standard deviation     0.36448213 0.32292640
## Proportion of Variance 0.01976893 0.01551807
## Cumulative Proportion  0.98448193 1.00000000
```

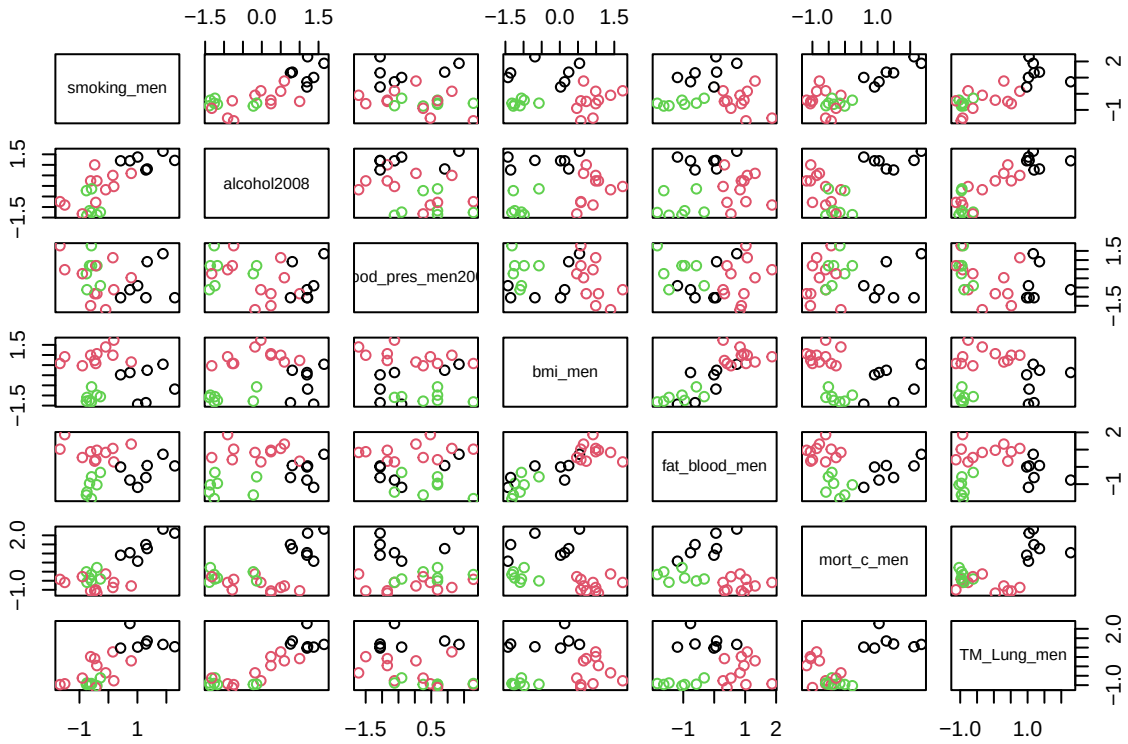
```
pca2$loadings ### como se puede observar, los loadings son idénticos a
```

```
##
## Loadings:
##                    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## smoking_men         0.512         0.102  0.244  0.701  0.259  0.322
## alcohol2008         0.499 -0.162        -0.473  0.145        -0.691
## blood_pres_men2008 -0.197  0.177  0.916 -0.288
## bmi_men                   -0.665  0.120         0.208 -0.679  0.188
## fat_blood_men             -0.642  0.249  0.315 -0.241  0.593 -0.130
## mort_c_men          0.434  0.272  0.270  0.602 -0.372 -0.343 -0.213
## TM_Lung_men         0.511                -0.409 -0.495         0.564
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
## Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```

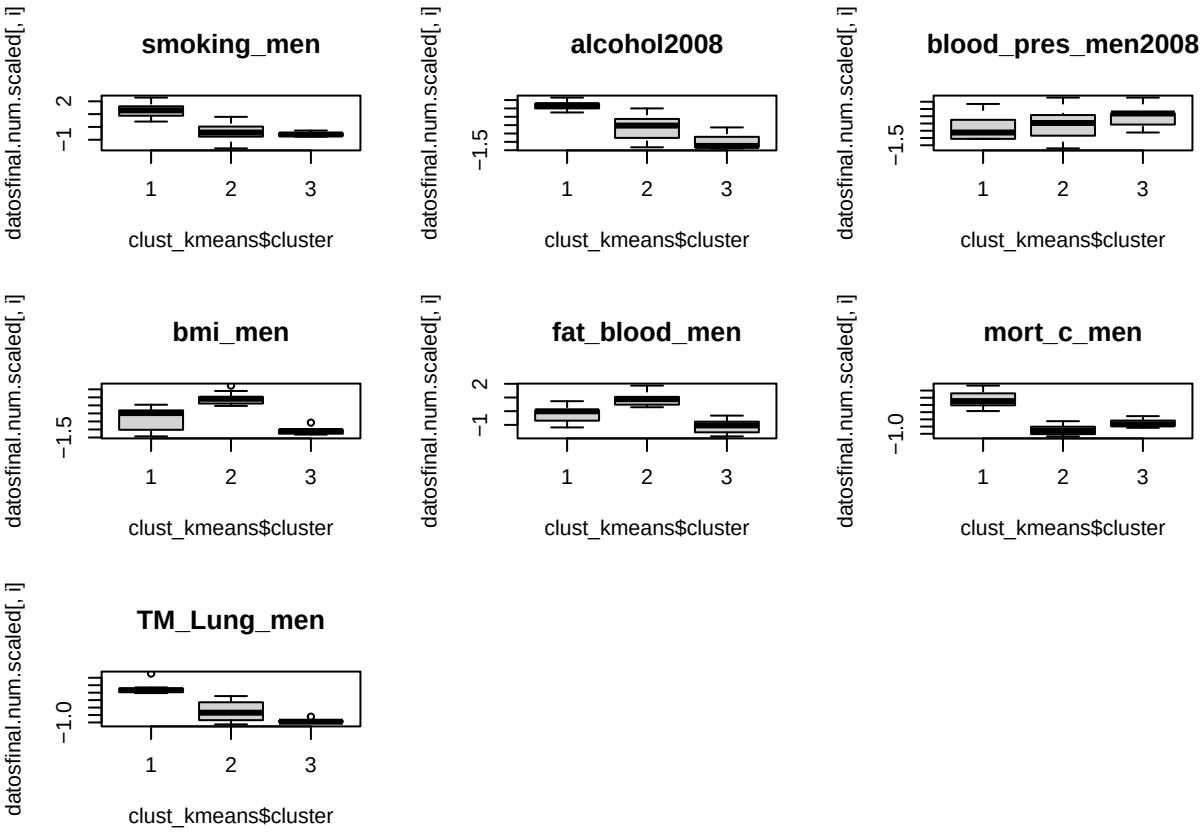
```
### los de pca1

plot(pca2$scores[,1], pca2$scores[,2], main = "K-means",xlab = "Estilo de vida", ylab="Baremos medicos"
text(pca2$scores[,1], pca2$scores[,2], labels = rownames(datosfinal.num),cex=0.8,col = clust_kmeans$clus
```




```
pairs(datosfinal.num.scaled, col=clust_kmeans$cluster)
```




```
par(mfrow=c(3,3))
for (i in 1:7){
  boxplot(datosfinal.num.scaled[,i] ~ clust_kmeans$cluster, main=colnames(datosfinal.num.scaled)[i])
}
```



Como se puede observar existen grandes diferencias entre las medias de las distintas variable cuantitativas agrupadas en función de los distintos cluster, siendo la única que no presenta esta heterogeneidad la presión sanguínea.