Análisis Estadístico Multivariante

Problemas propuestos de Análisis Cluster

Francisco Javier Mercader Martínez

Problema 1

El conjunto de datos **iris** de `R` contiene las variables Sepal.Length(X1, longitud de los sépalos), Sepal.Width(X2, anchura de los sépalos), Petal.Length(X3, longitud de los pétalos),Petal.Width (X4, anchura de los pétalos) medidas en centímetros, de tres especies diferentes de flor de iris (Species: *setosa*, *versicolor* y *virginica*). Se desea realizar un análisis cluster para determinar las diferentes agrupaciones que pueden obtenerse de las flores de iris a partir de las magnitudes de sus pétalos y sépalos. Se pide:

1. Realizar un agrupamiento de los datos en 3 grupos considerando los datos estandarizados aplicando algoritmo `k-means` con semilla `set.seed(123456)` y 10 inicios aleatorios con todas las variables.

```
set.seed(123456)

d <- iris[, 1:4]
ds <- as.data.frame(scale(d))
```

```
CA <- kmeans(ds, centers = 3, nstart = 10)
```

   a. Indicar cuál es el tamaño de cada grupo.

```
CA$size
```

```
## [1] 53 50 47
```

   b. ¿A qué grupos pertenecen las flores 15, 75 y 123?

```
CA$cluster[c(15, 75, 123)]
```

```
## [1] 2 1 3
```

   c. Obtener los centroides de cada grupo.

```
CA$centers
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1  -0.05005221 -0.88042696    0.3465767   0.2805873
## 2  -1.01119138  0.85041372   -1.3006301  -1.2507035
## 3   1.13217737  0.08812645    0.9928284   1.0141287
```

   d. ¿Qué porcentaje de variabilidad se reduce con estos 3 grupos con respecto de la variabilidad total correspondiente a 1 solo grupo?

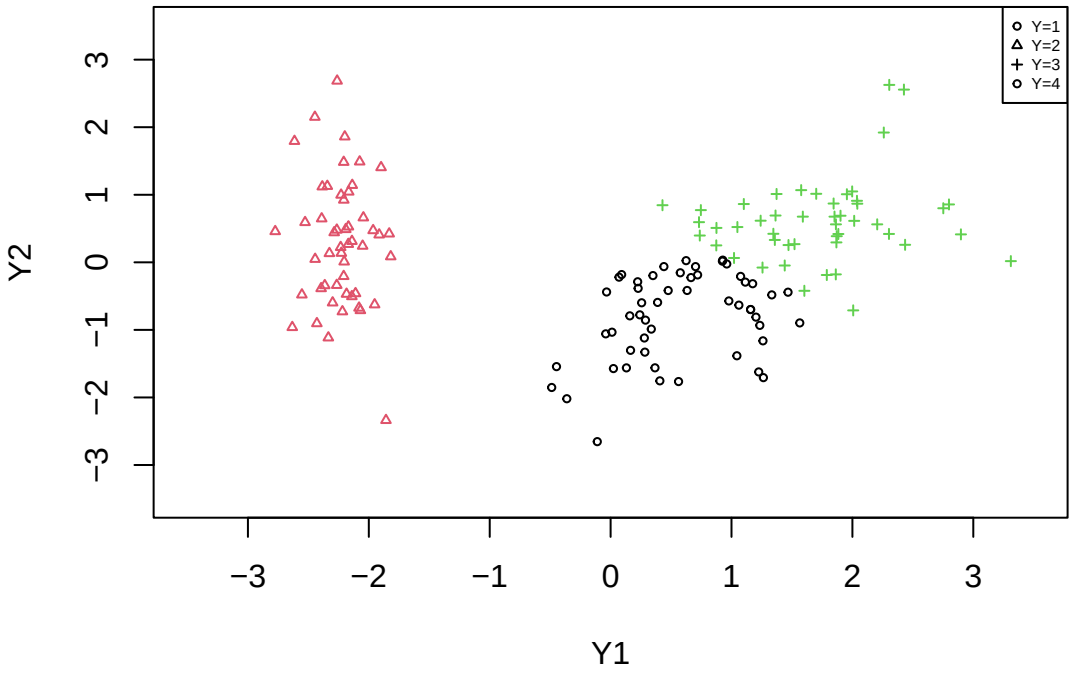
2. Hacer un ACP con las 4 variables.

```
PCA <- princomp(d, cor = TRUE)
L <- PCA$loadings
S <- PCA$scores
Y1 <- S[,1]
Y2 <- S[,2]
```

   a. ¿Qué porcentaje de la variabilidad es explicada por las 2 primeras componentes?

   b. Representar gráficamente las puntuaciones no estandarizadas en el sistema de referencia definido por las 2 primeras componentes principales y etiquetar con símbolos los grupos obtenidos con el algoritmo de `k-means`.

```
plot(Y1, Y2, pch = as.integer(CA$cluster), col = CA$cluster,
     cex = 0.5, xlim = c(-3.5, 3.5), ylim = c(-3.5, 3.5))
legend("topright", legend = c("Y=1", "Y=2", "Y=3", "Y=4"), pch = 1:3, cex = 0.5)
```



3. Realizar un análisis cluster jerarquizado de los datos usando la distancia euclídea con las variables estandarizadas y el método *complete*.

```
D <- dist(ds, method = "euclidean")
dim(ds)
```

```
## [1] 150   4
```

```
n = dim(ds)[1]
M <- as.matrix(D)[1:150, 1:150]
CA2 <- hclust(D, method = "complete")
```

   a. ¿Cuál es la distancia entre las flores 1 y 55?

```
M[1, 55]
```

```
## [1] 3.410625
```

   b. ¿Cuál es la distancia mínima? ¿A qué flores corresponden?

```
CA2$merge[1, ]
```

```
## [1] -102 -143
```

```
M[102, 143]
```

```
## [1] 0
```

   c. ¿Cuál es el primer grupo que se forma? ¿Y el último?

El primero que se forma es uniendo las flores 102 y 143

```r
CA2$merge[149, ]
```

```
## [1] 147 148
```

El último está formado por los grupos que se forman en las etapas 147 y 148

d. Si se formaran 2 grupos, ¿qué tamaño tendría cada grupo? Incluir los 2 grupos en el gráfico de las 2 primeras componentes principales y comentar cómo serían los grupos según este gráfico. ¿Dónde se incluirán las flores 1 y 55?

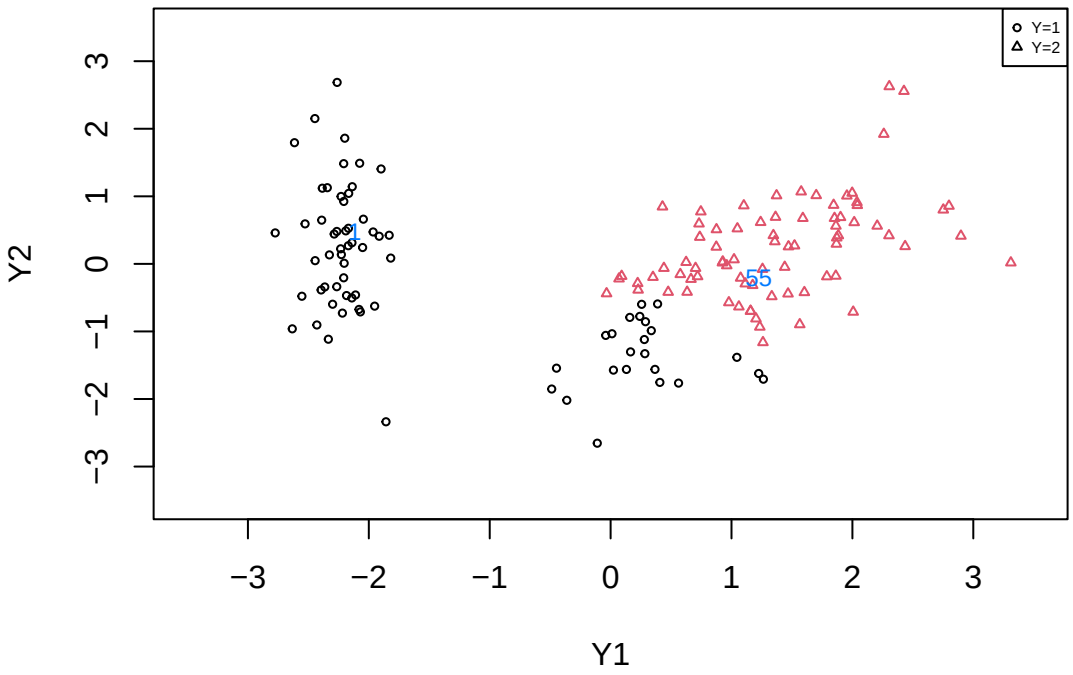
```r
grupos <- cutree(CA2, k = 2)
table(grupos)
```

```
## grupos
##  1  2
## 73 77
```

```r
grupos[c(1, 55)]
```

```
## [1] 1 2
```

```r
plot(Y1, Y2, pch = as.integer(grupos), col = grupos,
     cex = 0.5, xlim = c(-3.5, 3.5), ylim = c(-3.5, 3.5))
legend("topright", legend = c("Y=1", "Y=2"), pch = 1:2, cex = 0.5, col = grupos)
text(Y1[1]+0.15, Y2[1], "1", col = "#007AFF", cex = 0.75)
text(Y1[55]+0.15, Y2[55], "55", col = "#007AFF", cex = 0.75)
```



e. Si se formaran 3 grupos, ¿qué tamaño tendría cada grupo? Incluir los 3 grupos en el gráfico de las 2 primeras componentes principales y comentar cómo serían los grupos según este gráfico. ¿Dónde se incluirán las flores 1 y 55?

```r
grupos <- cutree(CA2, k = 3)
table(grupos)
```

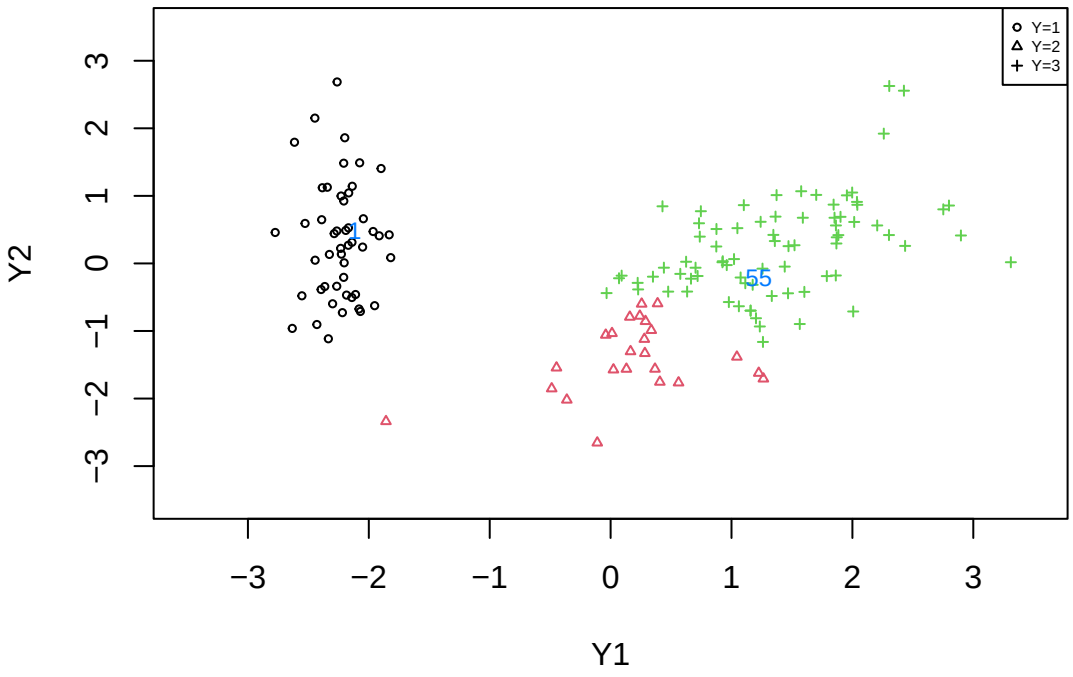
```
## grupos
```

```
##  1  2  3
## 49 24 77
```

```
grupos[c(1, 55)]
```

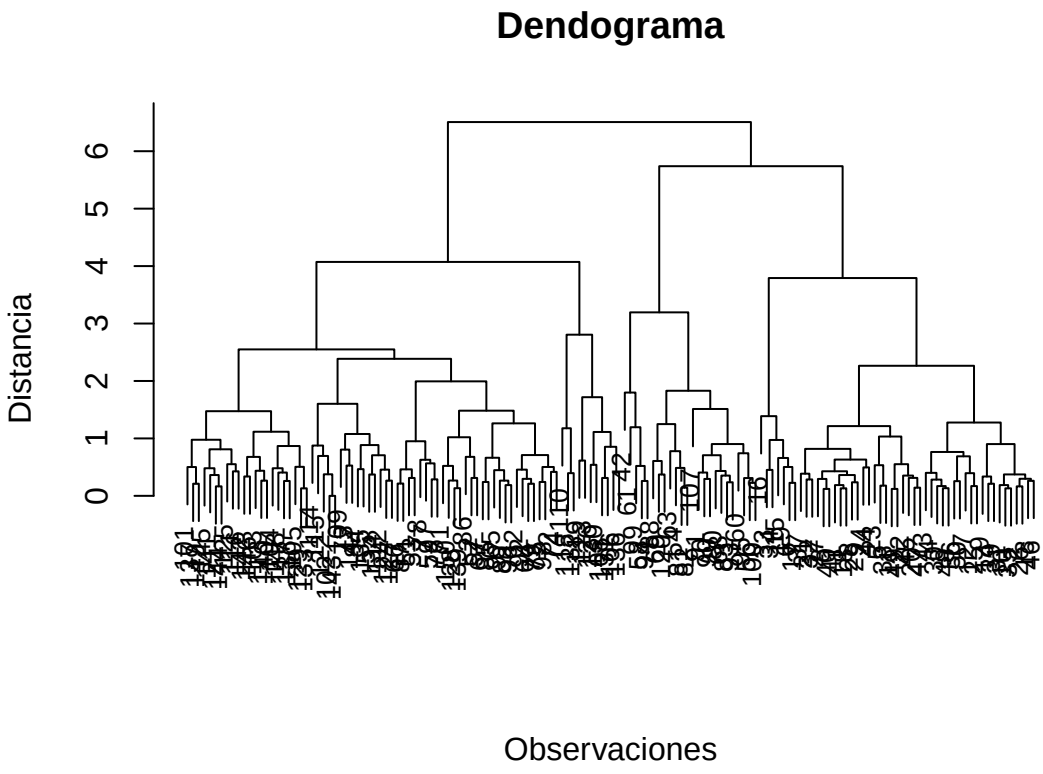
```
## [1] 1 3
```

```
plot(Y1, Y2, pch = as.integer(grupos), col = grupos,
     cex = 0.5, xlim = c(-3.5, 3.5), ylim = c(-3.5, 3.5))
legend("topright", legend = c("Y=1", "Y=2", "Y=3"), pch = 1:3, cex = 0.5, col = grupos)
text(Y1[1]+0.15, Y2[1], "1", col = "#007AFF", cex = 0.75)
text(Y1[55]+0.15, Y2[55], "55", col = "#007AFF", cex = 0.75)
```



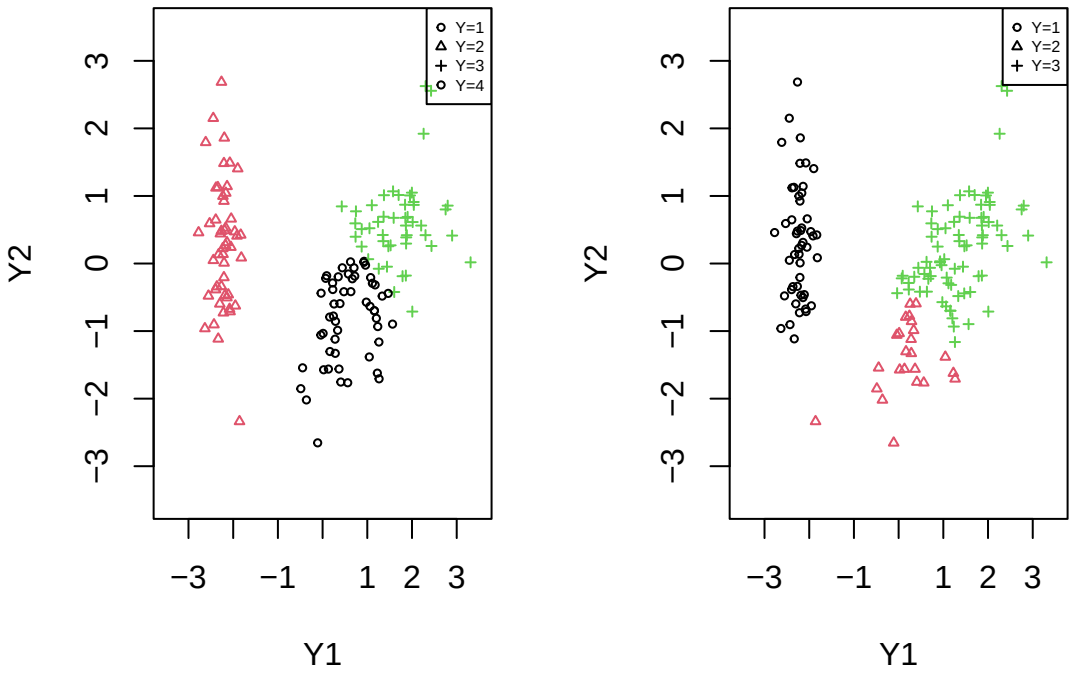
f. Representar el dendograma e indicar dónde se sitúan esas dos agrupaciones.

```
plot(CA2, cex=0.8, main="Dendograma", ylab="Distancia", xlab="Observaciones", sub='')
```



4. Comparar los 3 grupos formados con **kmeans** y con **hclust**.

```
par(mfrow = c(1, 2))
plot(Y1, Y2, pch = as.integer(CA$cluster), col = CA$cluster,
     cex = 0.5, xlim = c(-3.5, 3.5), ylim = c(-3.5, 3.5))
legend("topright", legend = c("Y=1", "Y=2", "Y=3", "Y=4"), pch = 1:3, cex = 0.5)
plot(Y1, Y2, pch = as.integer(grupos), col = grupos,
     cex = 0.5, xlim = c(-3.5, 3.5), ylim = c(-3.5, 3.5))
legend("topright", legend = c("Y=1", "Y=2", "Y=3"), pch = 1:3, cex = 0.5, col = grupos)
```



a. ¿Hay grupos coincidentes?

Los grupos no son coincidentes

b. ¿Qué grupos de `kmeans` se parece más al grupo 1 de `hclust`?

El grupo 2 de **`kmeans`** se parece más al grupo 1 de **`hclust`**.

c. ¿Cuántas flores tienen en común?

```
sum(which(CA$cluster == 2) %in% which(grupos == 1))
```

```
## [1] 49
```

```
sum(which(CA$cluster == 1) %in% which(grupos == 2))
```

```
## [1] 23
```

Tienen en común 49 flores

c. ¿Cuáles son las flores que no tienen en común?

```
sum(which(CA$cluster == 1) %in% which(grupos == 1))
```

```
## [1] 0
```

```
sum(which(CA$cluster == 2) %in% which(grupos == 2))
```

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

```
which(CA$cluster == 2) %in% which(grupos == 2)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [37] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [49] FALSE FALSE
```

No tienen en común la flor 42

d. ¿Cuáles son las principales diferencias entre todos los grupos de ambos métodos?

e. Comparar los 3 grupos formados por cada método con los definidos por la especie de iris. ¿Cuáles es la especie más en cada grupo formado por cada método?

```
table(CA$cluster, iris$Species)
```

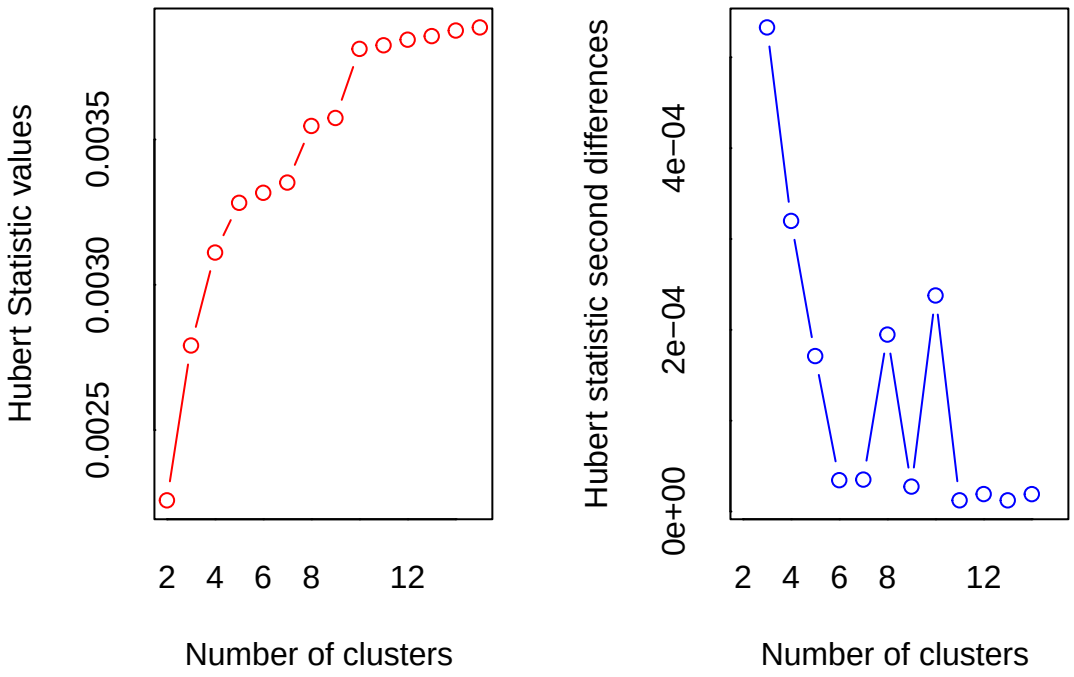
```
##
##     setosa versicolor virginica
##   1      0         39        14
##   2     50          0         0
##   3      0         11        36
```

```
table(grupos, iris$Species)
```

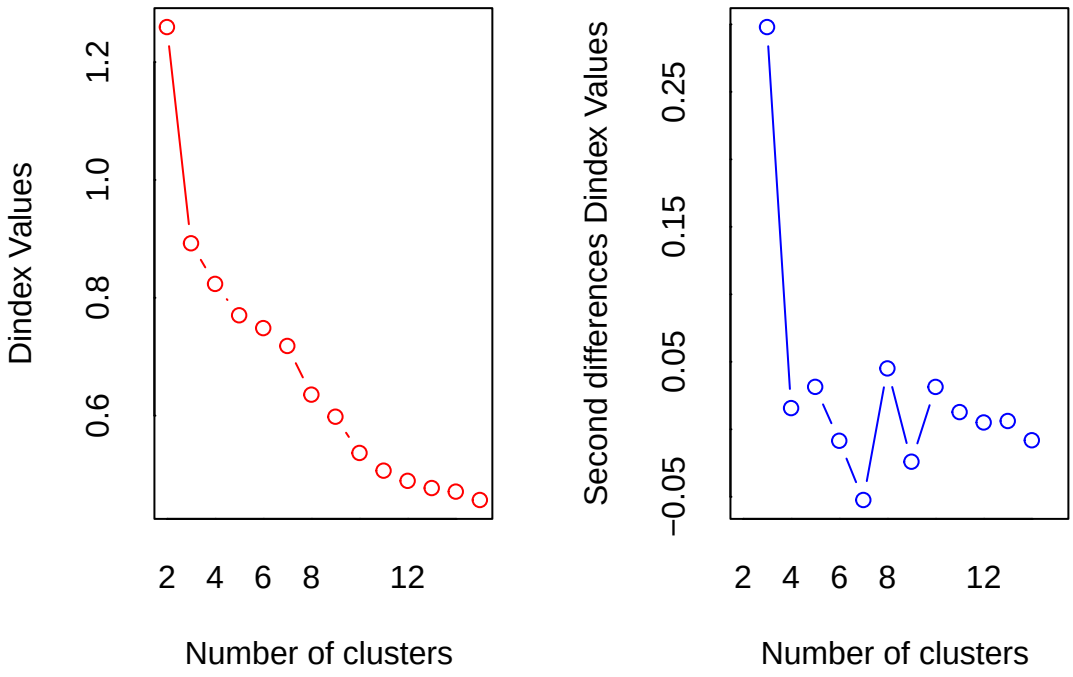
```
##
## grupos setosa versicolor virginica
##      1     49          0         0
##      2      1         21         2
##      3      0         29        48
```

5. ¿Se podría obtener un número óptimo de grupos?

```
library(NbClust)
NbClust(ds, method = 'complete', index = "all")$Best.nc
```




```
## *** : 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:
## * 2 proposed 2 as the best number of clusters
## * 17 proposed 3 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 1 proposed 12 as the best number of clusters
## * 2 proposed 15 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************

##                       KL       CH Hartigan     CCC    Scott Marriot   TrCovW
## Number_clusters 3.0000   3.0000   3.0000 12.0000   3.0000       3     3.00
## Value_Index     9.8471 213.0817 103.8142  4.8266 238.6703 2042179 10381.44
##                   TraceW Friedman   Rubin Cindex     DB Silhouette   Duda
## Number_clusters   3.0000    3.000  3.0000 2.0000 3.0000     3.0000 3.0000
## Value_Index     113.3546   15.107 -0.9934 0.2978 0.8581     0.4496 0.7005
##                  PseudoT2  Beale Ratkowsky   Ball PtBiserial Frey McClain
## Number_clusters   3.0000 3.0000    3.0000  3.000     3.0000    1  2.0000
## Value_Index      32.0629 1.0185    0.4958 96.241     0.7169   NA  0.5265
##                     Dunn Hubert SDindex Dindex    SDbw
## Number_clusters 15.0000      0  5.0000      0 15.0000
## Value_Index      0.1563      0  1.7359      0  0.0655
```

6. ¿Cómo cambiará la distribución de los grupos si se realizara un cluster jerárquico utilizando el método del enlace simple o del vecino más próximo (*single*)?

```
CA3 = hclust(D, method = "single")
CA3
```

```
##
## Call:
## hclust(d = D, method = "single")
##
## Cluster method   : single
## Distance         : euclidean
## Number of objects: 150
```