MINERÍA DE DATOS Y MODELIZACIÓN PREDICTIVA I

ANÁLISIS CLUSTER

Autor:Juana María Alonso Revenga

OBJETIVOS Y COMPETENCIAS A ALCANZAR.

- Saber obtener matrices de distancias de cualquier tipo de datos originales.
- Conocer los fundamentos de los algoritmos de Análisis Clúster Jerárquicos: Método de Ward, del vecino más cercano, del vecino más alejado, del centroide y de la media.
- Conocer la diferencia de eficacia de los métodos Jerárquicos de clasificación en función de las distribuciones originales de los grupos.
- Conocer los fundamentos de los algoritmos de Análisis Cluster No Jerárquico.
- Conocimiento de técnicas que permitan justificar el número de agrupaciones finalmente obtenidas así como la no presencia de cluster en los datos, si así fuera.
- Detectar observaciones atípicas, tanto tras la realización de Análisis Cluster Jerárquico como No Jerárquico.
- Obtención e interpretación de las agrupaciones generadas por cualquiera de los dos métodos.

INTRODUCCIÓN.

Veamos las situaciones que se describen a continuación:

- El departamento de marketing de una empresa va a lanzar una campaña publicitaria sobre un nuevo producto. Para ello, desea tener a sus potenciales clientes agrupados según sus necesidades en los distintos aspectos de dicho producto.
- Para mejorar los métodos terapéuticos a aplicar, el Servicio de Neurología de un Hospital, va a agrupar a sus pacientes, según determinadas variables indicativas del tipo de enfermedad que padecen.
- En un colegio desean crear grupos de apoyo para alumnos con dificultades. Con este fin, se clasifica a los alumnos según las necesidades que puedan presentar.

En todos los casos anteriores, nos encontramos con situaciones similares, en el sentido, de que **nuestro objetivo es formar grupos de individuos con características similares con respecto a determinadas variables**. El Análisis Cluster es la técnica Multivariante cuyo propósito es, a partir de un conjunto de individuos, crear grupos tales que:

- ✓ **Los individuos de cada grupo, deben ser lo más parecidos que sea posible (homogeneidad interna).**
- ✓ **Los grupos deben ser lo más diferentes que sea posible (heterogeneidad entre grupos).**
- ✓ Generalmente **los grupos obtenidos son mutuamente excluyentes** (cada observación pertenece a un solo grupo).

La forma más simple de obtener clusters, es por simple inspección gráfica de cruce de variables, agrupando las observaciones que se encuentren más próximas entre sí, o bien con la representación de las observaciones en los planos factoriales. No obstante, en la práctica serán muchas las observaciones a clasificar por lo que será difícil su representación. Además debemos tener en cuenta las muchas formas a la hora de medir la proximidad entre dos observaciones y por otro los diferentes procedimientos de agrupación. Por lo tanto, se puede obtener una gama amplia de posibles resultados, lo que eleva el riesgo de cometer errores.

Como en la mayoría de las técnicas multivariantes la representatividad de la muestra es tremendamente importante, como también la existencia o no de multicolinealidad.

La diferencia fundamental con respecto a otras técnicas de clasificación es que no se conoce de antemano el número de grupos en los que se divide a la población, ni el valor de la variable que identifica cada grupo.

A la hora de preparar los datos, es frecuente que **las variables vengan en diferentes unidades de medida,** en ese caso conviene normalizarlas. Otra circunstancia no deseable es que las **variables se encuentren correladas.** En ese caso será preferible recurrir previamente a alguna técnica como el análisis factorial o el análisis de componentes principales que sintetice la información, proporcionándonos variables incorreladas. Además es conveniente **corregir el problema de los atípicos** ya que distorsionarían la generación de cluster.

2. ¿CÓMO SE HACE UN ANÁLISIS CLUSTER?.

Como se indicó anteriormente, nuestro objetivo es formar grupos de observaciones homogéneas. Para lograr este objetivo necesitamos:

- Evaluar el parecido entre dos observaciones. Utilizaremos las Medidas de Distancia.
- Una vez que conocemos la manera de determinar la proximidad o el parecido de observaciones, debemos determinar la manera de agruparlas. Existen dos procedimientos de agrupación: **Jerárquicos y No Jerárquicos**. En los primeros no se conoce de antemano el número de clusters o grupos a formar. En los procedimientos No Jerárquicos, el número de grupos está fijado de antemano.
- En tercer lugar, debemos elegir un método de agrupación dentro del procedimiento elegido.
- A continuación, debemos tomar una decisión acerca del número óptimo de clusters.
- Por último, la solución elegida en el paso anterior debe ser interpretada.

En los apartados siguientes, estudiaremos como realizar los pasos propuestos en los puntos anteriores.

Análisis Cluster con R:

Ejemplo guía: *Clúster de Paises según su esperanza de vida.*

Estamos interesados en una **clasificación en grupos de países** según su esperanza de vida a diferentes edades, según sexo.

Jerarquico.R | res.diana | R | EsperanzaVida

Filter

	X__1	m0	m25	m50	m75	w0	w25	w50	w75
1	Algeria	63	51	30	13	67	54	34	15
2	Cameroon	34	29	13	5	38	32	17	6
3	Madagascar	38	30	17	7	38	34	20	7
4	Mauritius	59	42	20	6	64	46	25	8
5	Reunion	56	38	18	7	62	46	25	10
6	Seychelles	62	44	24	7	69	50	28	14
7	South_Africa	65	44	22	7	72	50	27	9
8	Tunisia	56	46	24	11	63	54	33	19
9	Canada	69	47	24	8	75	53	29	10
10	Costa_Rica	65	48	26	9	68	50	27	10
11	Dominican_Rep	64	50	28	11	66	51	29	11
12	El_Salvador	56	44	25	10	61	48	27	12

Showing 1 to 12 of 26 entries

Importamos los datos de Excel y los guardamos como un dataframe con los nombres de los países en la columna rownames.

```
datosEV<- as.data.frame(EsperanzaVida)

rownames(datosEV)<-datosEV[,1]

dat_EV<-datos[,-1]
```

Instalamos las librerías que vamos a necesitar para hacer el análisis Cluster

```
install.packages("cluster")

library(Cluster)

install.packages("ggplot2")
```

```
library(ggplot2)

install.packages( "factoextra")

library(factoextra)

install.packages("factoMineR")

library(FactoMineR)
```

3.1. MEDIDAS DE DISTANCIA ENTRE OBSERVACIONES.

Para medir la semejanza entre dos observaciones se utilizan medidas de distancia (dos observaciones serán más parecidas cuanto menor distancia haya entre ellos) o de similitud (dos observaciones serán más parecidas cuanto mayor sea su similitud).

Suponiendo que cada observación recoge el valor de p variables en un individuo, podemos denotar la observación i por $x_i = (x_{i,1}, x_{i,2},\ldots,x_{ip})$. Se pueden definir las siguientes **medidas de distancia** para variables numéricas, las cuales, si bien son de las más utilizadas, no son las únicas:

La más utilizada es la Euclidea.

Distancia Euclídea: $d(\mathbf{x_i}, \mathbf{x_{i'}}) = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2}$

Distancia de Minkoswski: $d(\mathbf{x_i}, \mathbf{x_{i'}}) = \left(\sum_{j=1}^{p} \left|x_{ij} - x_{i'j}\right|^r\right)^{1/r}$ para r=1 es la distancia de **Manhattan**, para r=2 es la distancia Euclídea.

3. 2.MEDIDAS DE DISTANCIA ENTRE VARIABLES.

Además de poder hacer grupos de observaciones, puede ser de interés hacer grupos de variables cuyo comportamiento sobtre el conjunto de individuos sea similar o estén muy relacionadas. Para ello se definen medias de distancia en tre variables basadas en los diferentes coeficientes de correlación.

Distancia de correlación de Pearson

$$d(\mathbf{x}, \mathbf{y}) = 1 - \left|r_{xy}\right|$$

Distancia de coseno de Eisen:

Es un caso particular de la Pearson cuando las variables tienen media cero

$$d(\mathbf{x},\mathbf{y}) = 1 - \frac{\left|\sum_{i=1}^{n} x_i y_i\right|}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

Distancia correlación de Spearman:

Es la de Pearson calculada sobre los rangos de las variables. Los rangos de las variables son los valores ordinales, 1,2..,n asignados según el valor de la variable. Se suele utilizar cuando hay mucha diferencia entre los valores de las variables o la distribución es muy asimétrica

$$d(\mathbf{x},\mathbf{y}) = 1 - \left|r_{RxRy}\right|$$

En el caso de que las variables sean binarias u ordinales se utilizan una distancias específicas para este tipo de variables, como:

Distancia correlación de Kendall:

Utiliza las comparaciones entre rangos de las variables

$p_c =$ Número de pares concordantes
$p_d =$ Número de pares discordantes

$$d(\mathbf{x},\mathbf{y}) = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Ejemplo:

```
#Calculamos las distancias con  los valores sin estandarizar
d <- dist(dat_EV, method = "euclidean") # distance matrix
#Mostramos las primeras seis filas dela matriz de distancias
as.matrix(d)[1:6, 1:6]
```

```
            Algeria  Cameroon Madagascar Mauritius   Reunion Seychelles
Algeria     0.00000 58.077534  52.649786 21.189620 24.351591   13.37909
Cameroon   58.07753  0.000000   7.141428 42.237424 38.026307   51.02940
Madagascar 52.64979  7.141428   0.000000 37.960506 33.808283   46.37887
Mauritius  21.18962 42.237424  37.960506  0.000000  6.164414   10.77033
Reunion    24.35159 38.026307  33.808283  6.164414  0.000000   14.07125
Seychelles 13.37909 51.029403  46.378875 10.770330 14.071247    0.00000
```

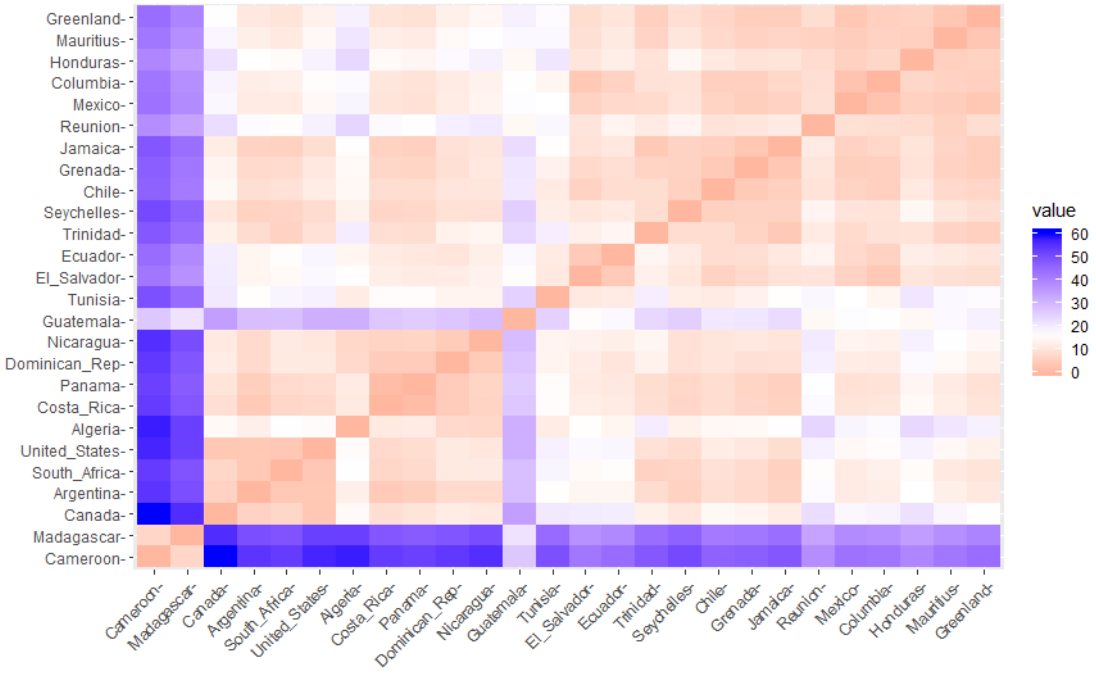
$$d(\mathbf{x_i}, \mathbf{x_{i'}}) = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{i'j} \right)^2}$$

X__1	m0	m25	m50	m75	w0	w25	w50	w75
Algeria	63	51	30	13	67	54	34	15
Cameroon	34	29	13	5	38	32	17	6
Madagascar	38	30	17	7	38	34	20	7

Observemos que la distancia que se ha utilizado es la Euclídea. También podemos observar que hay unas diferencias muy grandes entre los valores de las distancias y estos valores además dependen de las unidades de medida de las variables.

Representamos mediante escalas de color la distancia entre todas las observaciones para ver si detectamos grupos de observaciones.

```
fviz_dist(d, show_labels = TRUE)
```



Como podemos observar hay dos países, Camerún y Madagascar, que tienen una distancia muy superior con el resto de los países.

Para que la distancia entre individuos no dependa de las unidades en las que están medidas las variables y refleje mejor las relaciones entre individuos es conveniente estandarizar las variables.

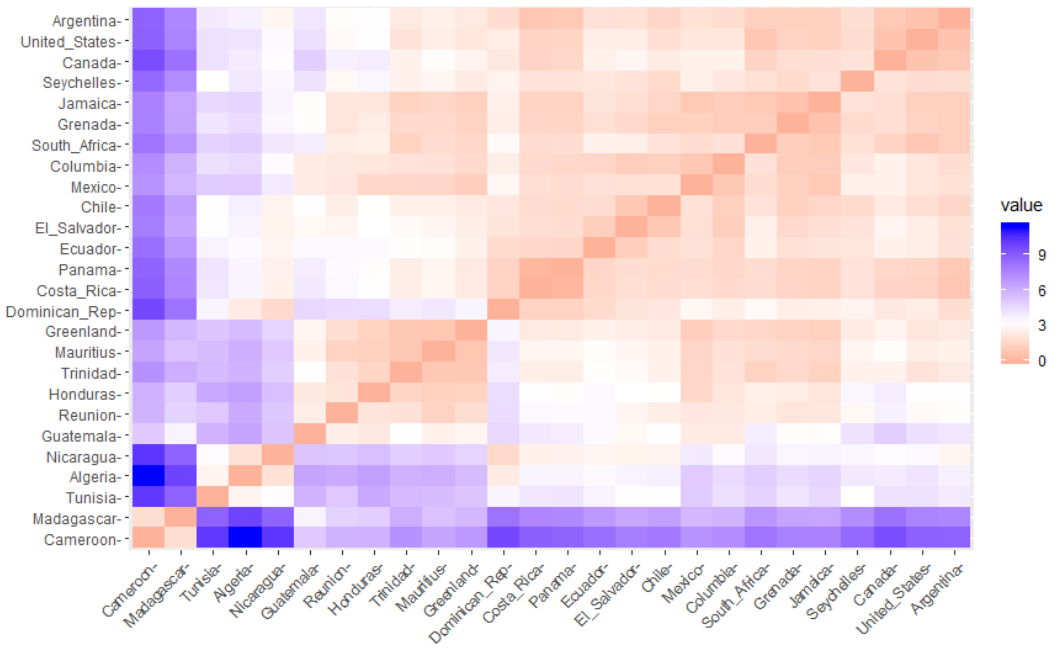
Recordamos que estandarizar es restar su media y dividir por su desviación estándar para conseguir que todas las variables tengan media cero y desviación típica 1:

$$\frac{X_{ij} - \bar{X}_j}{S_j}$$

```
# Standardize the data
datos_ST <- scale(dat_EV)
# Show the first 6 rows
head(datos_ST, nrow = 6)
#Calculamos las distancias con  los valores estandarizados

d_st <- dist(datos_ST, method = "euclidean") # distance matrix

#Mostramos las primeras seis filas de la matriz de distancias

as.matrix(d_st)[1:6, 1:6]
```

	Algeria	Cameroon	Madagascar	Mauritius	Reunion	Seychelles
Algeria	0.000000	11.373268	9.824032	6.070121	6.199760	4.007481
Cameroon	11.373268	0.000000	1.831649	6.428984	5.903757	8.522775
Madagascar	9.824032	1.831649	0.000000	5.393620	4.809238	7.280227
Mauritius	6.070121	6.428984	5.393620	0.000000	1.361332	2.830433
Reunion	6.199760	5.903757	4.809238	1.361332	0.000000	2.947164
Seychelles	4.007481	8.522775	7.280227	2.830433	2.947164	0.000000

```
#Visualizamos

fviz_dist(d_st)
```



Como se puede ver las distancias son diferentes y al representar gráficamente se observa que hay mayores diferencias y menos extremas.

4. ALGORITMOS DE CLASIFICACIÓN JERÁRQUICA.

Los métodos de clasificación jerárquica no producen una clasificación en un número determinado de clusters en un único paso, sino que configuran grupos con estructura arborescente de forma que clusters de niveles más bajos van siendo englobados en otros de niveles superiores.

Los pasos a seguir para realizar un Análisis Cluster Jerárquico son los siguientes:

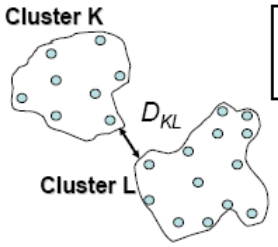
1. Se parte de tantos grupos como observaciones,
2. Se genera una matriz de dimensión $n \times n$ (simétrica) que indique las distancias entre todos los pares de observaciones (esta distancia debe haber sido definida con anterioridad).
3. A continuación, se agrupan las dos observaciones más próximas. Con esto, el número de clusters existentes es uno menos que en el paso anterior.
4. Se vuelve a obtener una matriz de distancias con los clusters formados en el paso 3. Obsérvese que para obtener esta nueva matriz, necesito elegir un método cálculo de distancia entre clusters.
5. A continuación, se vuelven a realizar las agrupaciones.
6. El proceso continúa hasta que todas las observaciones están agrupadas en un solo cluster.

Observemos que es necesario definir con claridad, tanto la distancia entre observaciones, como la distancia entre clusters o grupos de observaciones. Por lo tanto, bajo este esquema y con un mismo conjunto de datos, variando esas dos definiciones se podrán obtener múltiples clasificaciones diferenciadas.
Llegando a este punto nos preguntamos:

4.1. ¿Cómo calcular la distancia entre dos clusters o entre dos grupos de observaciones?

Existen varias maneras, de las que destacamos las siguientes:
Supongamos que tenemos una agrupación denominada C_k y otra agrupación $C_{k'}$, los algoritmos más comunes son:

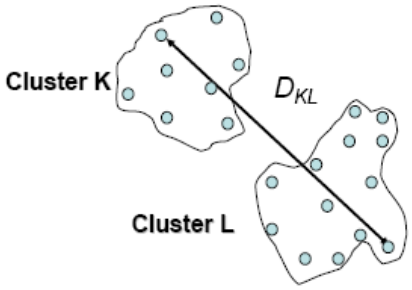
a) **Enlace Simple o del vecino más cercano**: La distancia elegida es:



$$D_{KL} = \min_{\substack{i \in C_K \\ j \in C_L}} d(x_i, x_j)$$

Es decir, la distancia entre dos grupos es la distancia entre las dos observaciones más cercanas, pertenecientes cada una de ellas a un grupo distinto. Tienden a crear grupos con muchas observaciones.

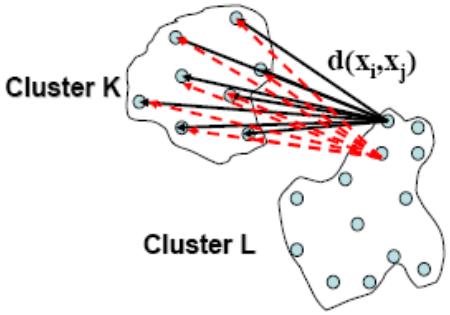
b) **Enlace Completo o del vecino más alejado**. La distancia elegida entre dos cluster será:



$$D_{KL} = \max_{\substack{i \in C_K \\ j \in C_L}} d(x_i, x_j)$$

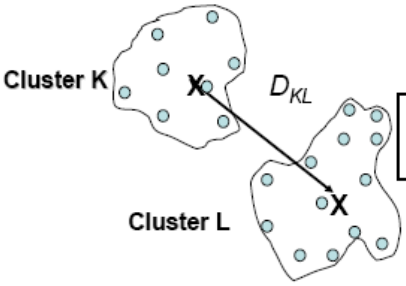
Luego, la distancia entre dos grupos, es la distancia entre las dos observaciones más alejadas de cada uno de los dos grupos. Con este método, los grupos formados son más compactos que los obtenidos con el método del vecino más próximo.

c) **Enlace medio**: La distancia entre dos agrupaciones es la distancia media entre una observación de la agrupación C_k y otra observación de la agrupación $C_{k'}$. Los grupos así formados tienen varianza similar y además pequeña.



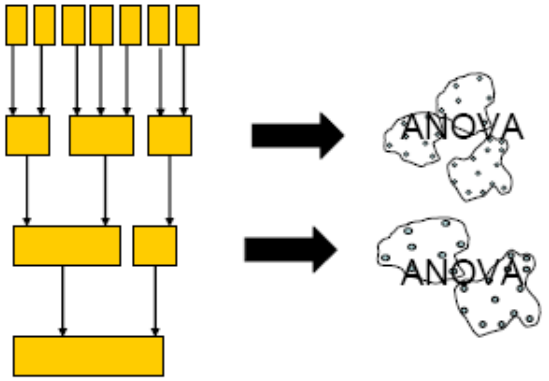
$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j)$$

d) **Distancia entre centroides**: La distancia entre grupos será la distancia entre los centroides, que representarán el vector medio obtenido en las p variables para todos los individuos que formen parte del grupo.



$$D_{KL} = \left\| \bar{x}_K - \bar{x}_L \right\|^2$$

e) **Distancia de Ward o de la mínima varianza**. Este método, entre todas las uniones de cluster posibles en cada nivel, selecciona aquella unión que minimiza la variabilidad interna de los cluster resultantes $\left(\sum W_k\right)$.



$$D_{KL} = \frac{\left\| \bar{x}_K - \bar{x}_L \right\|^2}{\left(\frac{1}{n_K} + \frac{1}{n_L} \right)}$$

Ahora, parece lógico preguntarse: **¿Cómo determinar en cada caso cuál es el método de agrupación más adecuado?** No existe una respuesta exacta a esta pregunta. Siempre es conveniente estudiar varios métodos y tomar una decisión en función de los resultados que se obtengan. Si varios métodos nos dan agrupaciones similares, se puede pensar que existe una forma natural de formarse grupos de observaciones.

Ejemplo.

Se tienen 6 observaciones, llevamos a cabo un ejemplo de agrupación, utilizando el método del vecino más cercano para definir la distancia entre grupos de observaciones. El proceso a seguir, es el indicado en la Figura 1

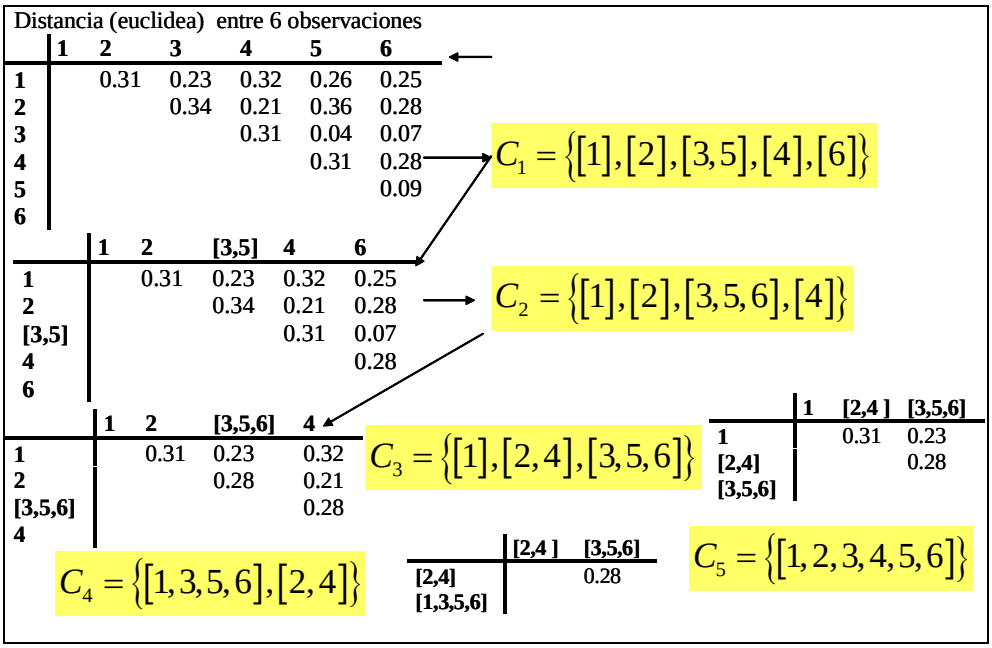


Figura 1. Desarrollo del ejemplo. Método del vecino más cercano

Repetimos el proceso utilizando el procedimiento del vecino más alejado:

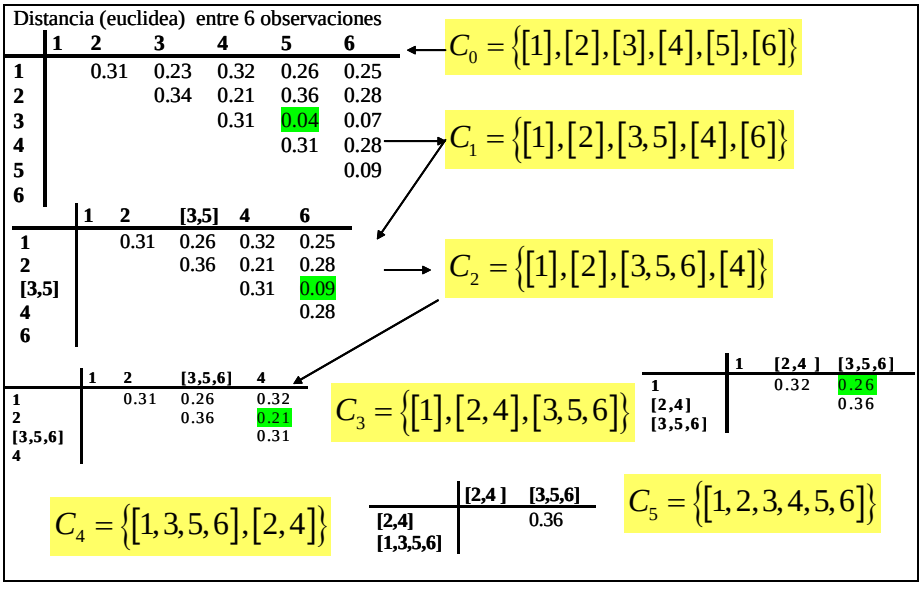
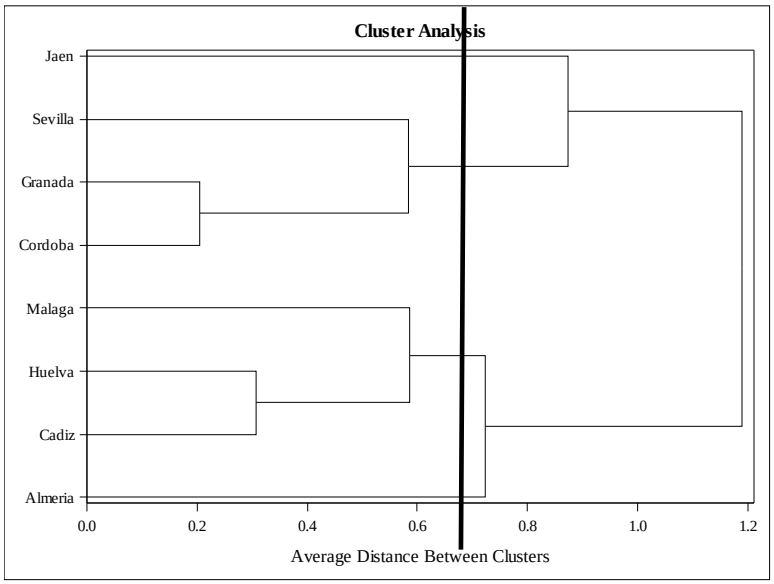


Figura 2. Desarrollo del ejemplo. Método del vecino más alejado

4.2. Resultados del clúster jerárquico: El Dendrograma

Es frecuente presentar los resultados del análisis clúster jerárquico con este gráfico. Tiene la estructura de un árbol que permite plasmar el **proceso de aglomeración** y composición de grupos (para cualquier número de ellos) junto con la distancia entre cada dos grupos unidos en una gráfica.. Este tipo de diagrama contiene ramas que unen puntos y muestran el orden en que se asignan las observaciones a los agrupamientos. **Las longitudes de las ramas son proporcionales a las distancias entre los puntos y agrupamientos**, cuando los puntos y los agrupamientos se combinan.

Este diagrama depende de la distancia entre elementos y entre clústeres utilizada, y nos **puede ayudar** a determinar en qué momento del proceso de agrupación nos deberemos detener

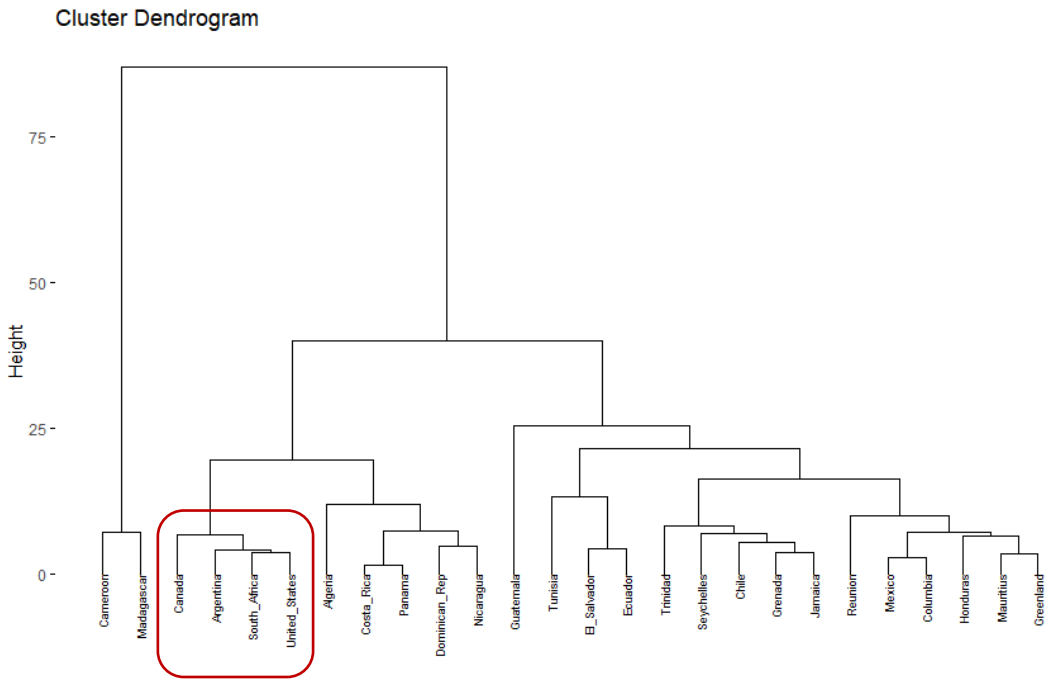


Dependiendo por dónde cortemos vemos la estructura de k-ramas cada una correspondiente a un clúster. En nuestro ejemplo vemos la composición para k = 4.

En R podemos utilizar la función hclust de la librería stats. Esta función necesita la matriz de distancias d

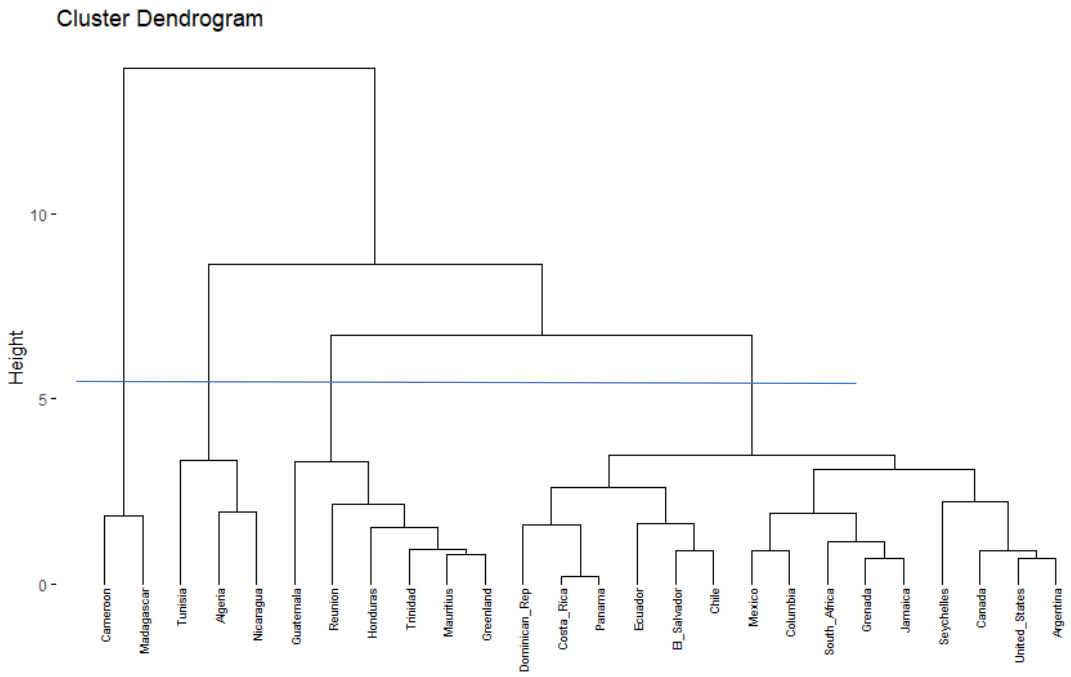
```
#Hacemos el cluster jerárquico con las distancias entre los datos sin estandarizar y

# la distancia de ward

res.hc <- hclust(d, method="ward.D2")

#Dibujamos el dendograma correspondiente
```

```
library("factoextra")

fviz_dend(res.hc, cex = 0.5)
```



Vamos a comparar como serían las agrupaciones calculando Dendograma con los datos estandarizados

```
#Hacemos el cluster jerárquico con las distancias entre los datos estandarizados

res.hc_st <- hclust(d_st, method="ward.D2")

#Dibujamos el dendograma correspondiente

fviz_dend(res.hc_st, cex = 0.5)
```



Como podemos observar los cluster cambian, por esta razón es siempre conveniente realizar la agrupación con las distancias estandarizadas.

Observando la estructura del dendrograma podemos hacernos una idea de cual podría ser el número más adecuado de cluster. Por ejemplo, en nuestro caso podría ser cuatro, como aparece en la imagen en donde hemos dibujado una línea. En ese caso podemos guardar los 4 clusters que hemos elegido para más tarde poder estudiar la caracterización de cada uno.

Seleccionamos el número de clusters que nos parece "lógico", mediante la función cutree que actúa sobre el resultado del algoritmo jerárquico (res.hc). El objeto grp contiene la información de los 4 clusters creados

```
# Cut tree into 4 groups

grp <- cutree(res.hc_st, k = 4)

head(grp, n = 4)

# Number of members in each cluster

table(grp)
```

```
grp 1 2 3 4
    3 2 6 15
```

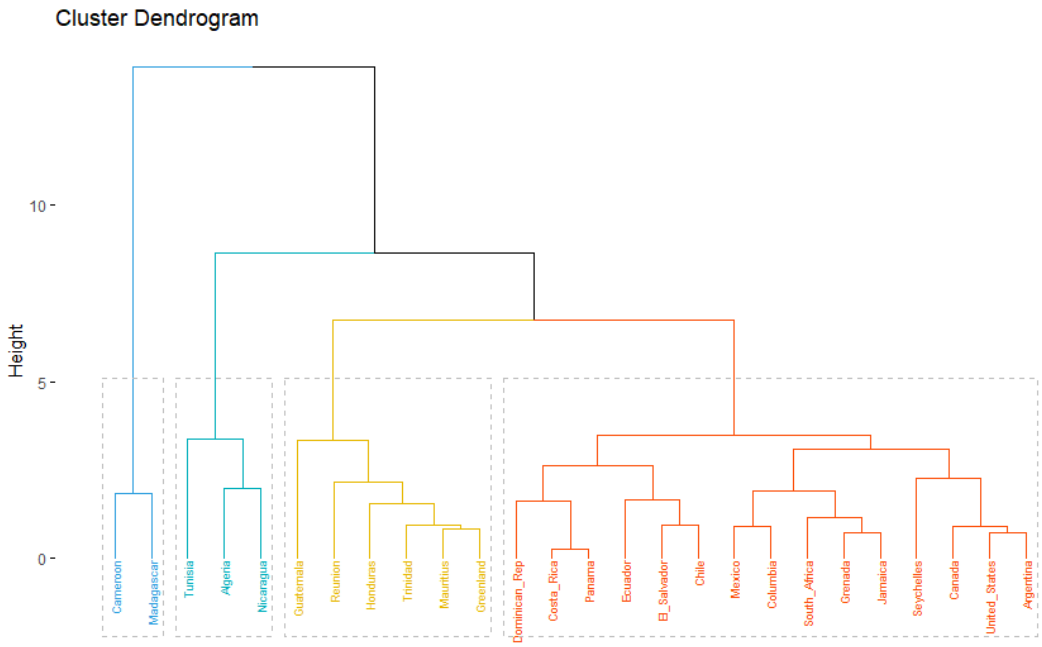
Si queremos mostrar una lista con los individuos de cada cluster, podemos utilizar la función rownames:

```
# Get the names for the members of cluster 1

rownames(dat_EV)[grp == 1]
```

```
[1] "Algeria" "Tunisia" "Nicaragua"
```

Esta información, con el número de clusters decidido, puede reflejarse en el dendrograma, mostrando las observaciones de cada cluster en diferente color mediante la función fviz_dend.

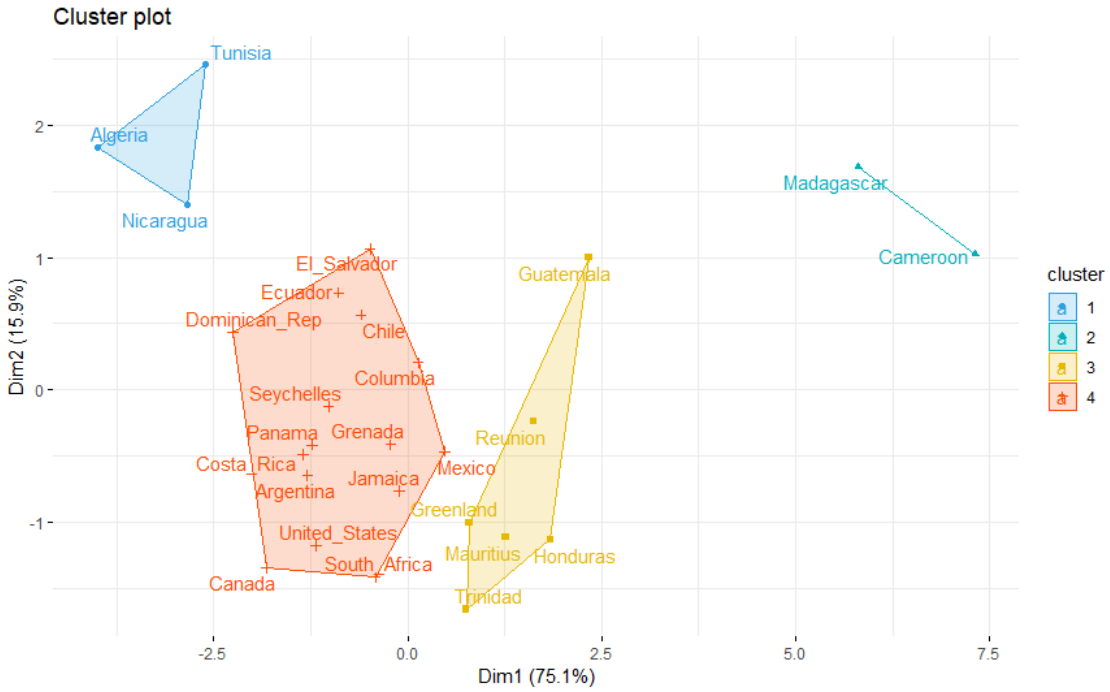
```
# Cut in 4 groups and color by groups

fviz_dend(res.hc_st, k = 4, # Cut in four groups

      cex = 0.5, # label size

      k_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),

      color_labels_by_k = TRUE, # color labels by groups

      rect = TRUE # Add rectangle around groups

      )
```



Para además, representar gráficamente la situación relativa de los clusters calculados podemos realizar un análisis de componentes principales sobre la matriz de correlaciones de los datos y utilizar las dos primeras componentes para representar los individuos en estos planos. Esto lo podemos hacer con la función de la librería factominer: fviz_cluster.

```
#Visualizamos los clusters

fviz_cluster(list(data = datos_ST, cluster = grp),

        palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),

        ellipse.type = "convex", # Concentration ellipse

        repel = TRUE, # Avoid label overplotting (slow)

        show.clust.cent = FALSE, ggtheme = theme_minimal())
```

Resultado del cluster Jerárquico:



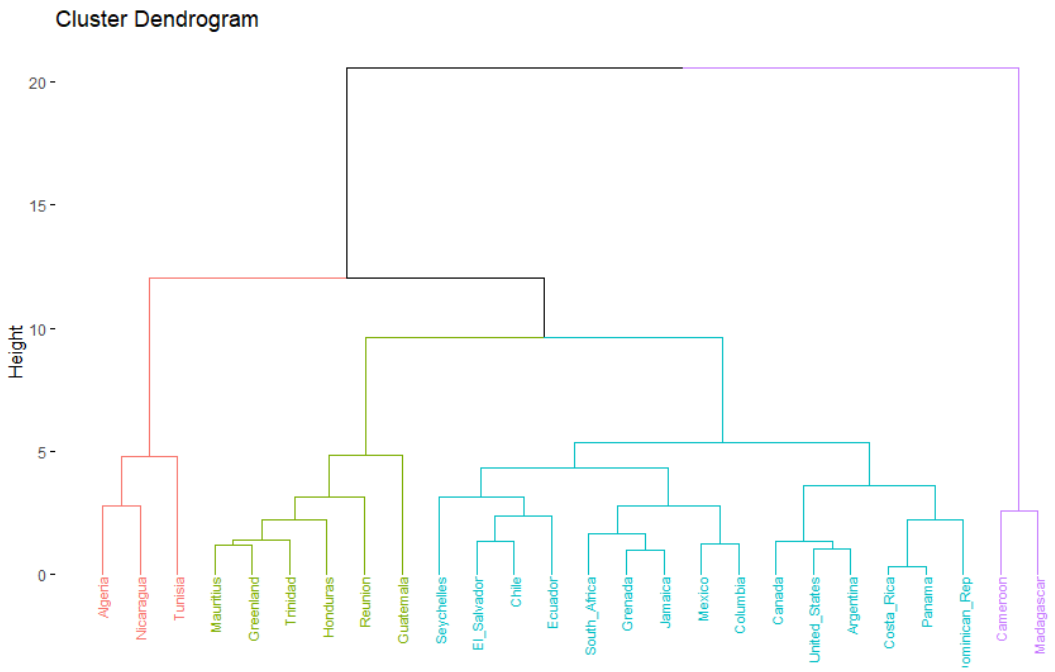
Otras librerias de R para cluster jerárquico.

Podemos realizar los pasos anteriores a las representaciones con **la función agnes** de R:

```
library("cluster")

# Agglomerative Nesting (Hierarchical Clustering)

res.agnes <- agnes(x =dat_EV, # data matrix   stand = TRUE, # Standardize the data

                   metric = "euclidean", # metric for distance matrix

                   method = "ward" # Linkage method)
```

Observemos que la función Agnes actúa directamente sobre los datos, no sobre la matriz de distancias y además, la estandarización se le puede indicar en la función

```
fviz_dend(res.agnes, cex = 0.6, k = 4)
```



5. ALGORITMOS DE CLASIFICACIÓN NO JERÁRQUICOS.

Para utilizar los algoritmos de agrupación no Jerárquicos, es necesario fijar de antemano el número de grupos en que se pretende dividir las observaciones. Por esta razón, normalmente se utiliza un cluster jerárquico previamente para tener una idea aproximada del número de clusters adecuado. Las técnicas de agrupación para realizar análisis Cluster No Jerárquico, siguen básicamente los siguientes pasos:

1. **Seleccionar K observaciones como centroides iniciales** de los Clusters a construir, siendo K el número deseado de Clusters.
2. **Asignar cada una de las observaciones restantes al Cluster más próximo**.
3. **Reasignar cada observación a uno de los K Clusters de acuerdo con una regla de parada** determinada previamente.
4. **Parar si no se reasignan observaciones a un grupo distinto** del de partida, o si la reasignación satisface la regla de parada. En caso contrario, volver a 2.

Los métodos para realizar agrupaciones o Clusters No jerárquicos, se diferencian entre sí por el modo de escoger los centroides iniciales y por el criterio empleado para reasignar observaciones a los distintos grupos.

Algunos de los métodos utilizados para obtener los centroides iniciales son:

a) **Seleccionar las K primeras observaciones con datos no-missing**.
b) **Seleccionar la primera observación con datos conocidos como primer centroide**. **El segundo** centroide, es aquella observación **cuya distancia al primer centroide sea tan grande como una distancia seleccionada previamente.** El tercer centroide, es la observación cuya distancia a los dos primeros centroides sea mayor que la distancia seleccionada. Y así sucesivamente.
c) **Seleccionar aleatoriamente K observaciones con datos conocidos.**
d) **Elegir centroides que estén entre sí lo más lejanos posible**.
e) Utilizar centroides que proporcione el investigador.

Una vez que se han identificado los centroides, se pueden formar los Clusters iniciales asignando cada una de las N-K observaciones restantes al Cluster correspondiente al centroide más próximo.

En cuanto **a la forma de reasignar observaciones, algunas de las reglas más utilizadas** son las siguientes:

a) Calcular el centroide de cada Cluster y reasignar sujetos a los Clusters cuyo centroide es el más cercano. Los centroides no varían mientras se reasignan observaciones, sino que se recalculan después de hacer la nueva asignación. Si el cambio en el centroide, es mayor que el valor determinado por un criterio de convergencia, se vuelve a hacer una reasignación de observaciones, y se vuelven a calcular los centroides. El proceso de reasignación continúa hasta que el cambio en los centroides es menor que el valor dado por el criterio de convergencia.

b) Calcular el centroide de cada grupo y reasignar sujetos al Cluster cuyo centroide es el más cercano. Para la asignación de cada observación, se recalcula el centroide del Cluster al que se asigna la observación, y el centroide del Cluster del que proviene la observación. La reasignación continúa hasta que la diferencia entre centroides es menor que el valor dado por el criterio de convergencia.

c) Reasignar observaciones de acuerdo con algún criterio estadístico tal como, minimizar la traza de la matriz SSCP(dentro de los grupos). (VER NOTACIÓN), Sumas de Cuadrados de Productos Cruzados dentro de los grupos.

Por lo tanto, combinando los distintos métodos de selección de centroides iniciales y de reasignación de observaciones, se pueden desarrollar un gran número de algoritmos de Clusters No jerárquicos.

Vemos como hacerlo con R:

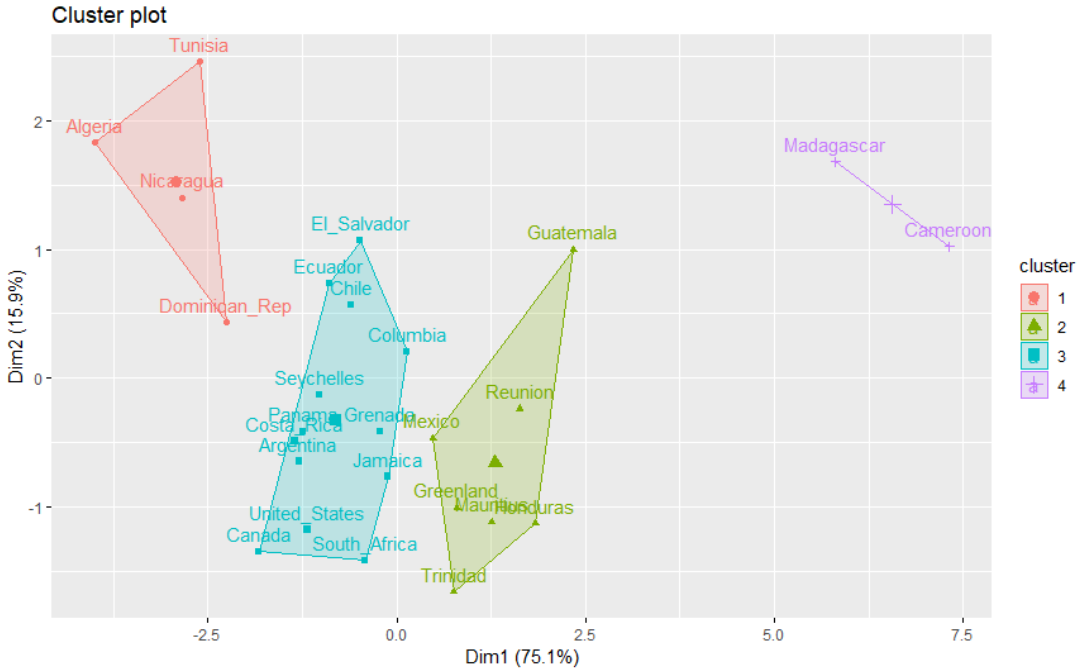
Primero estandarizamos los datos.

```
# Standardize the data

datos_ST <- scale(dat_EV)
```

Utilizamos la función Kmeans con una semilla fijada (123) indicando que el número de clusters va a ser 4

```
# Compute k-means

set.seed(123)

km.res <- kmeans(datos_ST , 4)

head(km.res$cluster, 20)
```

```
# Visualize clusters using factoextra

library("factoextra")

fviz_cluster(km.res, datos)
```



Resultado del cluster no jerárquico.

6. PROCEDIMIENTOS PARA DETERMINAR EL NÚMERO ADECUADO DE GRUPOS.

Después de obtener una solución en análisis Cluster, el paso siguiente es evaluar esta solución y determinar el número real de grupos existentes en nuestros datos. Nuestro objetivo es formar grupos lo más homogéneos "dentro de sí" y lo más diferentes "entre sí". Para medir esto, la variabilidad total de los datos la dividiremos en dos partes la que mide la variabilidad dentro de los grupos y la que mide la variabilidad entre los grupos.

Variabilidad Total $$T = \sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij} - \overline{x}_j)^2$$

Variabilidad dentro del cluster K $$W_k = \sum_{i \in C_k}\left(\sum_{j=1}^{p}\left(x_{ij} - \overline{x}_j^k\right)^2\right)$$

Variabilidad dentro de los G cluster

$$W = \sum_{k=1}^{G} W_k$$

Variabilidad entre los cluster

$$E = \sum_{k} \sum_{j=1}^{p} \left(\bar{x}_{jk} - \bar{x}_j \right)^2$$

Se demuestra que la variabilidad Total es igual a la suma de la variabilidad dentro de los clusters más la variabilidad entre clusters.

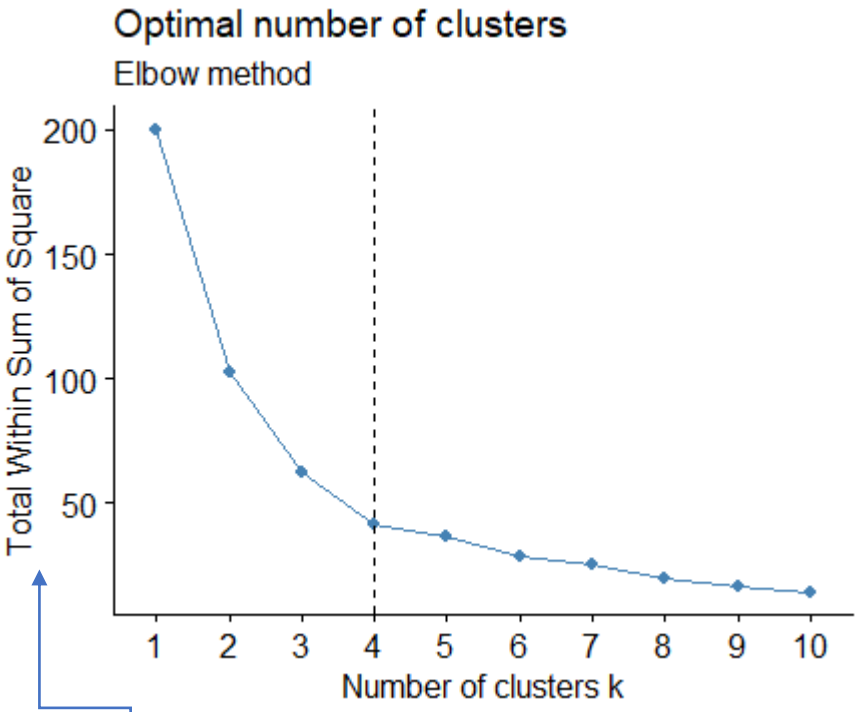
$$T = W + E$$

Tanto dentro del análisis Jerárquico, como en el No Jerárquico, existen procedimientos (además de las gráficas) que nos ayudan a decidir el número más adecuado de cluster.

En R utilizaremos la siguiente librería para determinar el número óptimo de clusters:

```
#Determinación del número óptimo de clusters

install.packages("NbClust")

library(NbClust)


# Elbow method

fviz_nbclust(datos_st, kmeans, method = "wss") +

  geom_vline(xintercept = 4, linetype = 2)+

  labs(subtitle = "Elbow method")
```

El criterio de Elbow elige como número óptimo de clusters aquel número de clusters en el que la Variabilidad total intra-clústeres ya no se reduce de forma significativa al aumentar uno más



$$W_k = \sum_{j=1}^{p} \sum_{i \in C_k} \left(x_{ij} - \bar{x}_{jK} \right)^2$$

$$W = \sum_{k} W_k$$

Gráfica de la Silueta.

Es una medida de como de compactos son los clusters y cuanto de separados están unos de otros Dada una observación i, se denota como:

a_i =la distancia media todos los otros puntos de su propio cluster.

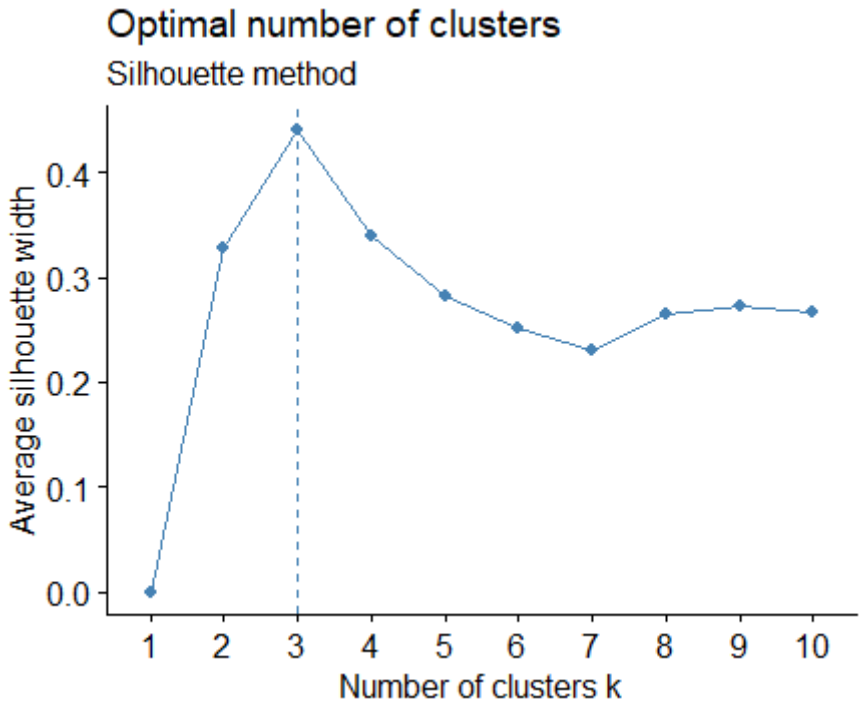
b_i= distancia media de la observación i-ésima a las observaciones de otros clusters

La anchura de la silueta para la i-esima observación es: $sw_i = \dfrac{(b_i - a_i)}{\max(a_i - b_i)}$

Se encuentra la anchura media de las siluetas como: $\overline{sw} = \dfrac{1}{n} \sum_{i=1}^{n} sw_i$

Este valor se calcula para todos los números de cluster. Si la silueta media es superior a 0.5 se produce una partición de los datos razonable mientras que un valor por debajo de 0.2 indica que los datos no exhiben ninguna estructura de cluster.

```
# Silhouette method

fviz_nbclust(datos_st, kmeans, method = "silhouette")+

  labs(subtitle = "Silhouette method")
```



Si por ejemplo, decidiéramos hacer 4 clusters estudiaríamos la calidad de estos calculando el valor del índice para cada uno de los elementos en cada uno de los clusters

```
#Evaluación de la calidad de los clusters

sil <- silhouette(km.res$cluster, dist(datos_st))

rownames(sil) <- rownames(dat_EV)

head(sil[, 1:3])

fviz_silhouette(sil)
```

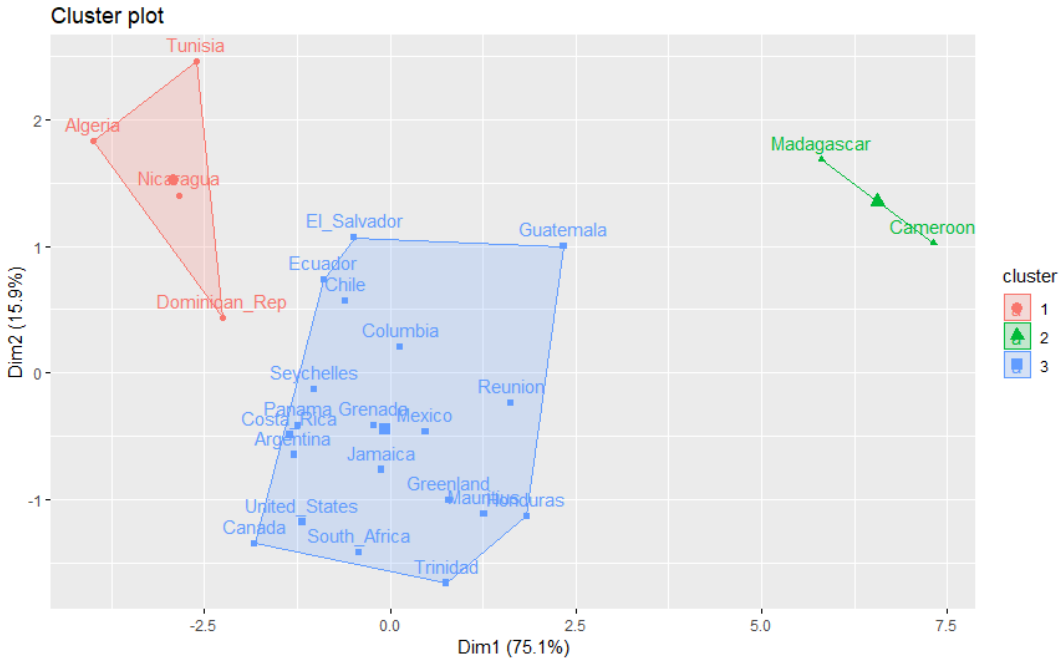
Algeria	1	3	0.4147402
Cameroon	4	2	0.7116156
Madagascar	4	2	0.6480743
Mauritius	2	3	0.4295660
Reunion	2	3	0.3098000
Seychelles	3	2	0.3536013

Las siluetas se encuentran entre -1 y 1. Si la silueta esta próxima a 1 eso querría indicar que la observación se encuentra bien agrupada, mientras que si vale 0 indica que la observación podría pertenecer a su cluster actual o a otro cercano a él. Si la silueta es negativa indicaría una mala agrupación para la observación.

Existe un gráfico denominado grafico de las siluetas que representa los valores de las siluetas para cada observación, agrupadas por cluster, esto permite analizar rápidamente de forma lineal la estructura de los cluster.

Por ejemplo la observación 4 del cluster 1 está mal clasificada porque su silueta es negativa.

```
# Probamos con 3 Clusters que es lo que nos recomienda el criterio Silouette

set.seed(123)

km.res <- kmeans(datos_st, 3)

# Visualize clusters using factoextra

fviz_cluster(km.res, datos_st)
```




```
#Evaluación de la calidad de los clusters

library("factoextra")

library("cluster")

sil <- silhouette(km.res$cluster, dist(datos_st))

rownames(sil) <- rownames(datos)

head(sil[, 1:3])

fviz_silhouette(sil)
```

Observemos que ahora no existen observaciones con valor de la silueta negativo

7. CARACTERIZACIÓN DE LOS CLUSTER.

Una vez que se ha decidido la partición de los cluster, el siguiente paso consistirá en caracterizarlos. Se debe realizar análisis descriptivo sobre las variables activas utilizadas en el análisis, con ello se determinará las medias y varianzas todas las observaciones. Ello nos permitirá una primera caracterización permitiendo interpretar las características de cada cluster con respecto a las variables originales.

```
#Se pueden calcular las medias de las variables originales

aggregate(dat_EV, by=list(km.res$cluster),mean)
```

Group.1		m0	m25	m50	m75	w0	w25	w50	w75
1	1	62.00000	48.75000	27.50000	12.250000	66.00000	52.50000	31.25000	14.500000
2	2	58.00000	41.85714	21.28571	6.857143	62.00000	44.85714	24.14286	8.285714
3	3	62.46154	45.23077	24.15385	8.538462	67.46154	49.53846	27.23077	10.538462
4	4	36.00000	29.50000	15.00000	6.000000	38.00000	33.00000	18.50000	6.500000

Por ejemplo, los cluster 1 y tres son los que tiene mayor esperanza de vida tanto en hombres como en mujeres al nacer. Mientras que el cluster 4 es el que tiene menor.
Todo lo anterior nos dará una explicación verosímil del trabajo de clasificación realizado. Sin esta última parte el análisis quedaría muy incompleto.

8. SISTEMÁTICA DEL ANÁLISIS CLUSTER.

1) Elección de la medida de distancia entre observaciones a utilizar.
2) Realizar un Análisis **Jerárquico.** Elección del método de agrupación de observaciones: Centroide, Ward, etc.
3) Decisión del número aproximado de grupos o Clusters a formar
4) Calcular los coeficientes que ayudan a decidir el número de grupos a formar.
5) Realizar un Análisis **No Jerárquico con el número decidido**
6) Estudiar la calidad de los cluster formados
7) Caracterización de los distintos grupos formados.

Bibliografía:

Análisis de Datos Multivariantes. Peña D. 2002.

Nuevos Métodos de Análisis Multivariante. Cuadras C.M. 2014

Análisis Multivariante de Datos. Pérez, C. Ed. Garceta. 2013

Practical guide to Cluster Analisys in R. A. Kassambara.Ed. STHDA. 2017

http://www.sthda.com/english/