



Minería de datos y Modelización predictiva I

CAPÍTULO II. Análisis Clúster



CAPÍTULO II. Análisis Clúster

II.1.- Introducción

II.2.- Medidas de distancia y similitud

II.3.- Algoritmos de clasificación jerárquica. Distancia entre clústeres.

II.4.- Algoritmos de clasificación no jerárquica

II.5.- Procedimientos para determinar el número de grupos

II.6.- Caracterización de los clústeres

II.7.- Bibliografía.



II.1.- Introducción: Clasificación de la técnica Clúster

El problema de clasificación/agrupación/asignación:

Se trata de clasificar en dos o más grupos a individuos sobre los que se han observado varias variables

Clasificación no supervisada

Se identifican grupos de individuos con características comunes a partir de la observación de varias variables en cada uno de ellos

ANÁLISIS CLÚSTER

Clasificación supervisada

Un individuo se clasifica en un grupo a partir de la información de un conjunto de variables observadas previamente en un conjunto de individuos de **los que se conoce el grupo de clasificación correcto**
(Los grupos están predefinidos)

ANÁLISIS DISCRIMINANTE

II.1.- Introducción: Objetivos

- El Análisis Clúster tiene como objetivo **formar grupos** de individuos con características similares.
- Se cuenta con una matriz de datos X de dimensión $(n \times m)$ cuyas filas y columnas representan las observaciones y las variables, respectivamente.
- La diferencia con el análisis discriminante es que **no se conocen de antemano los** grupos de clasificación de los individuos, ni la caracterización de cada grupo.
- La idea básica es crear grupos **excluyentes y exhaustivos** tales que:
 - Los **individuos** de un mismo grupo deben ser lo más **“parecidos”** posible (homogeneidad interna).
 - Los **grupos** deben ser lo más **“diferentes”** posible (heterogeneidad entre grupos).



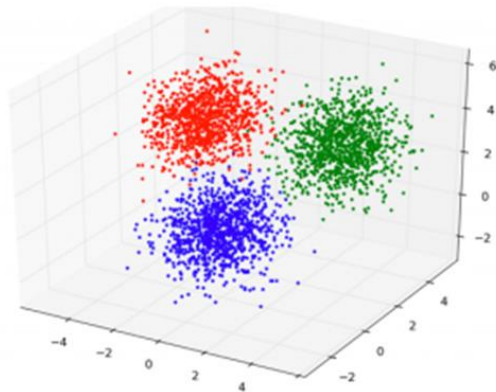
II.1.- Introducción: *Ejemplos*

- 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 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.

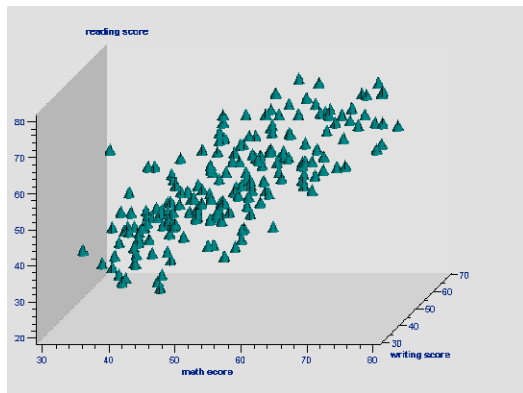
En todos los casos anteriores, nos encontramos con objetivos análogos, pero es evidente que **el mayor o menor grado de consecución, no solo depende de la metodología que se utilice, además tendrá un papel determinante la situación real existente de separación entre elementos.**



Situaciones extremas:



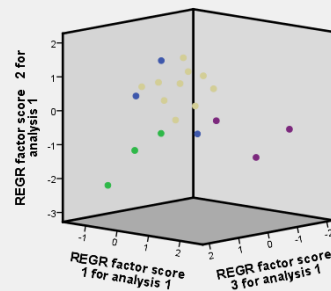
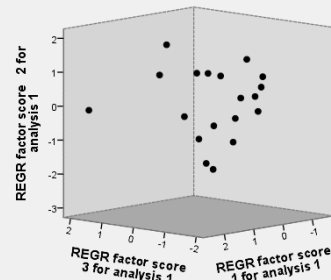
Situaciones con clústeres perfectamente definidos:
Una buena metodología los encontrará



Situaciones con clústeres inexistentes:
Una buena metodología llevará a los menos malos.

Situaciones más comunes:

Algo de separación que deseamos localizar



II.1.- Introducción: Cuestiones previas

- Es frecuente que la medida de parecido dependa de la **escala** de medida, por lo que es frecuente aplicar **normalización**. No se debe abusar de esta técnica puesto que al homogeneizar la varianza de todas las variables podemos mermar la “capacidad clasificatoria” de alguna variable con gran variabilidad por grupos.
- **No deben ser utilizadas** variables que **muy correladas** que pueden estar duplicando discrepancia/parecido respecto de una misma cualidad. En ese caso será preferible recurrir a las puntuaciones factoriales que sintetizan la información mediante variables incorreladas.
- Los términos “**parecidos**/diferentes”, “similares”, “homogéneos”, etc, aparecen tanto relacionados con **individuos** (I_1 I_2) como con **grupos de individuos** (A y B). **Deberán ser bien definidos** en ambos casos mediante indicadores numéricos que resuelvan una y otra cuestión.



II.1.- Ejemplo guía: *Clúster de Países según su esperanza de vida*

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

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

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

```
install.packages("cluster")  
install.packages("ggplot2")  
install.packages("factoextra")  
install.packages("factoMineR")  
install.packages("NbClust")
```



```
library(Cluster)  
library(ggplot2)  
library(factoextra)  
library(FactoMineR)  
library(NbClust)
```

Creamos el conjunto de datos como un dataframe y asignamos la columna de los países como nombres de las filas para usarla como identificador, posteriormente la eliminamos para que todas las columnas sean numéricas

```
datos<- as.data.frame(EsperanzaVida)  
rownames(datos)<-datos[,1]  
dat_EV<-datos[,-1]
```

Opciones a concretar en un análisis clúster

Para alcanzar nuestro objetivo de **formar grupos de observaciones homogéneas**, debemos concretar:

- Decidir la medida de discrepancia entre dos observaciones. Utilizaremos las **medidas de distancia y disimilaridad** entre pares de observaciones (entre cada dos países cualesquiera).
- Decidir la medida de discrepancia entre grupos de observaciones, es decir, elegir una **medida de distancia entre clústeres** (entre dos subconjuntos de países).
- Determinar la metodología con que serán utilizadas las dos distancias elegidas: **Métodos Jerárquicos o No Jerárquicos**.
- En el caso del método jerárquico y, de ser necesario, debemos tomar una decisión acerca del **número óptimo de clústeres** : Será necesario definir indicadores numéricos que nos ayuden a tomar una decisión al respecto. Y quizás a posteriori validarlo con un método de clasificación supervisada.
- Por último, la estructura de clústeres que se proponga como **solución** debe ser **interpretada**.



II.1.- Introducción: Metodologías en un análisis clúster

Fundamentalmente se clasifican en dos grandes grupos:

- **Métodos jerárquicos:** se construye una especie de jerarquía de uniones de observaciones en función de la distancia que haya entre ellas o grupos de ellas. Se obtiene una posible clasificación para **cualquier número de grupos G** ($1 \leq G \leq n$).
- **Ej:** Queremos **conocer la estructura de parecidos** entre todas los países.
- **Métodos no jerárquicos:** se desea **construir un número G** , predefinido, de grupos con los datos. (Indicado por coste computacional cuando hay demasiados casos).
- **Ej:** Queremos formar **tres** grupos de países según su Esperanza de Vida como indicador de su desarrollo. ¿Cómo deberíamos agruparlos?



II.2.- Medidas de distancia entre observaciones

Cuando cada observación está definida por el valor de p variables todas cuantitativas las medidas de discrepancia se denominan medidas de distancia. Si notamos por $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ la i -ésima observación, algunas de las más utilizadas son:

Distancia Euclídea :

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{i'j})^2}$$

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

Distancia de Minkowski (POWER(r,r):

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \left(\sum_{j=1}^p |x_{ij} - x_{i'j}|^r \right)^{1/r}$$

r=1 distancia de Manhattan
r=2 distancia Euclídea.

II.2.- Medidas de distancia entre variables

Distancia de correlación de Pearson

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

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

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

Distancia correlación de Kendall:

Utiliza las comparaciones entre rangos de las variables

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

p_c = Número de pares concordantes

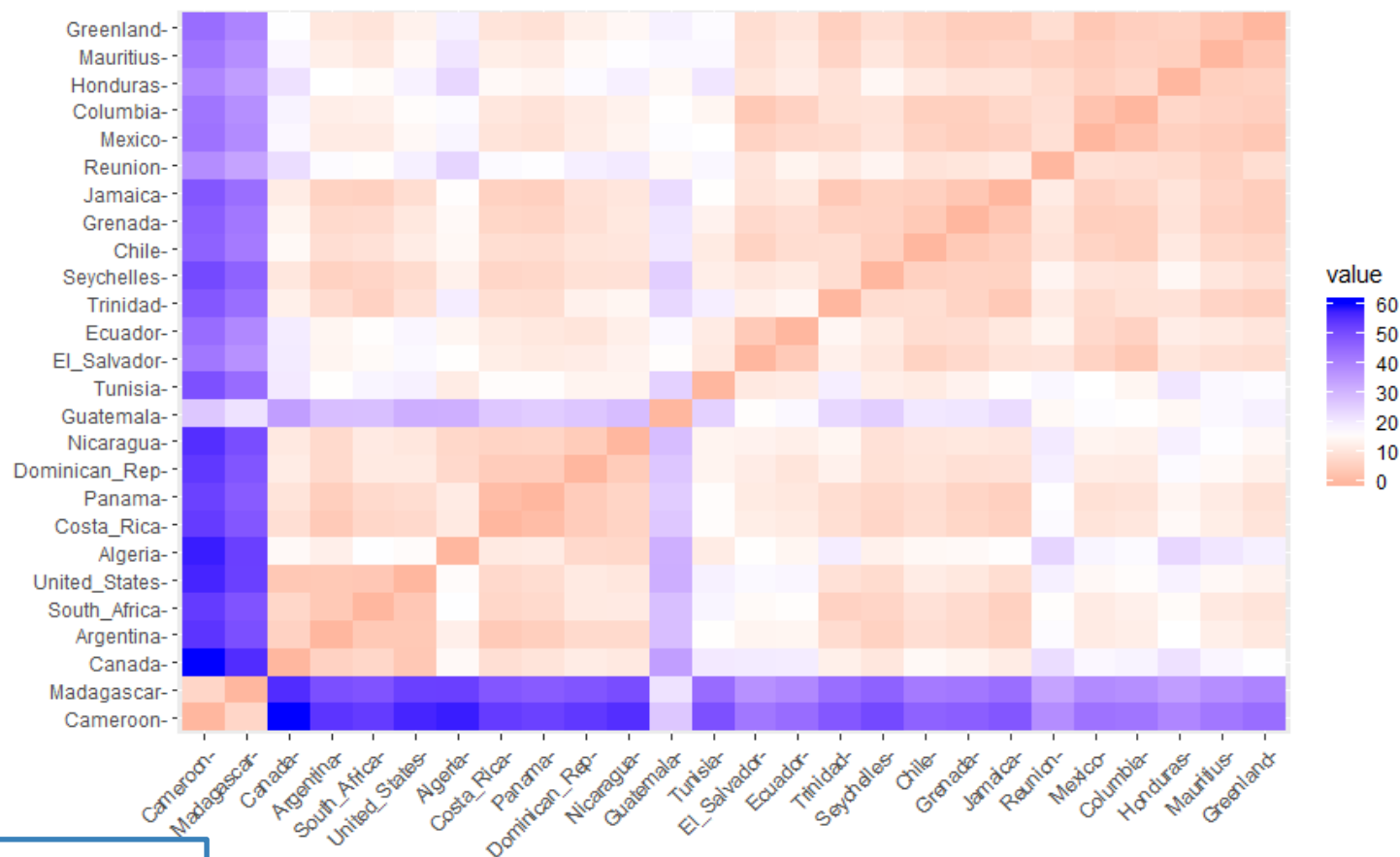
p_d = Número de pares discordantes

```
#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

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{i',j})^2}$$

Representamos mediante escalas de color la distancia entre todas las observaciones



`fviz_dist(d, show_labels = TRUE)`

```
# Standardize the data  
datos_ST <- scale(dat_EV)  
# Show the first 6 rows  
head(datos_ST, nrow = 6)
```

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


```
#Calculamos las distancias con los valores estandarizados
```

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

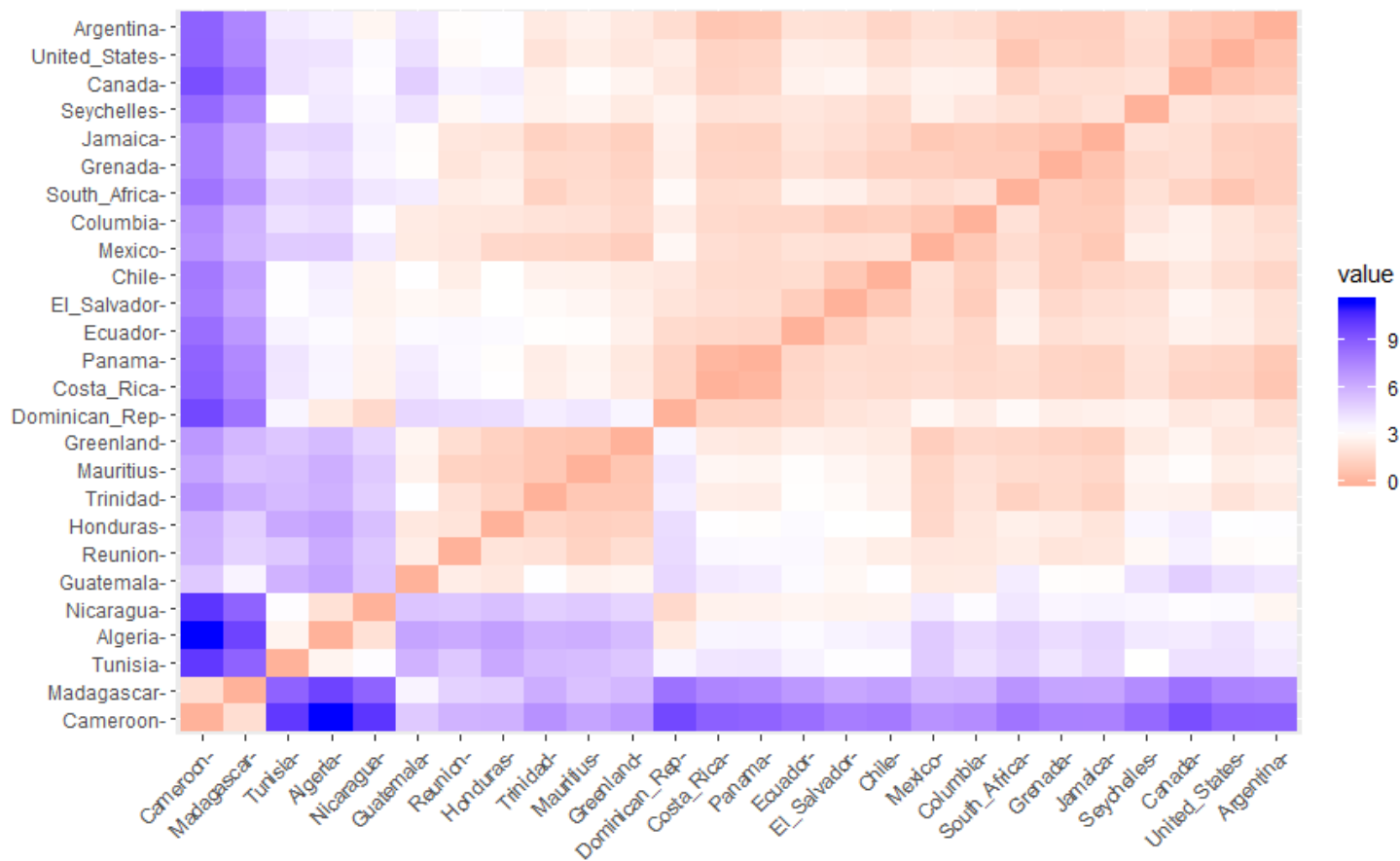
```
#Mostramos las primeras seis filas de la matriz de distancias
```

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

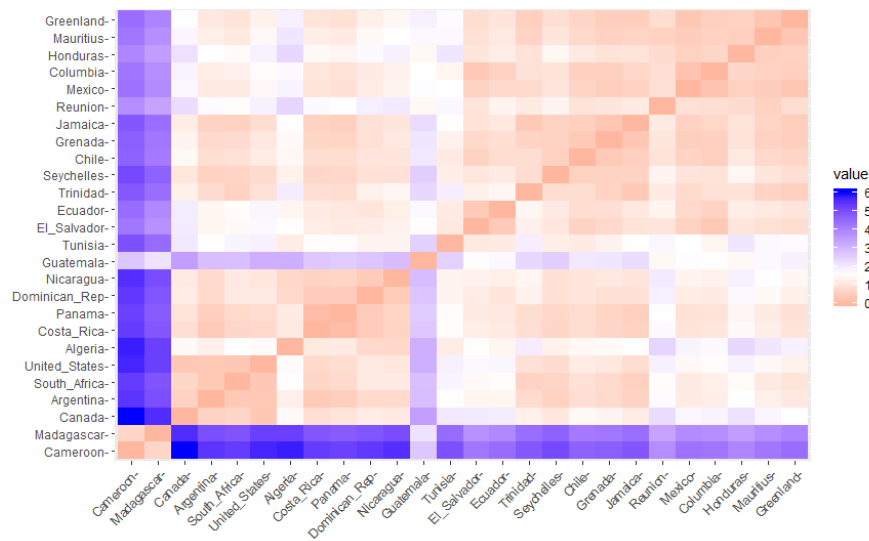
	Algeria	Cameroon	Madagascar	Mauritius	Reunion	Seychelles
Algeria	0.000000	11.373268	9.824032	6.070121	6.199760	4.007481
Cameroon	11.373268	0.000000	1.831649	6.428984	5.903757	8.522775
Madagascar	9.824032	1.831649	0.000000	5.393620	4.809238	7.280227
Mauritius	6.070121	6.428984	5.393620	0.000000	1.361332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

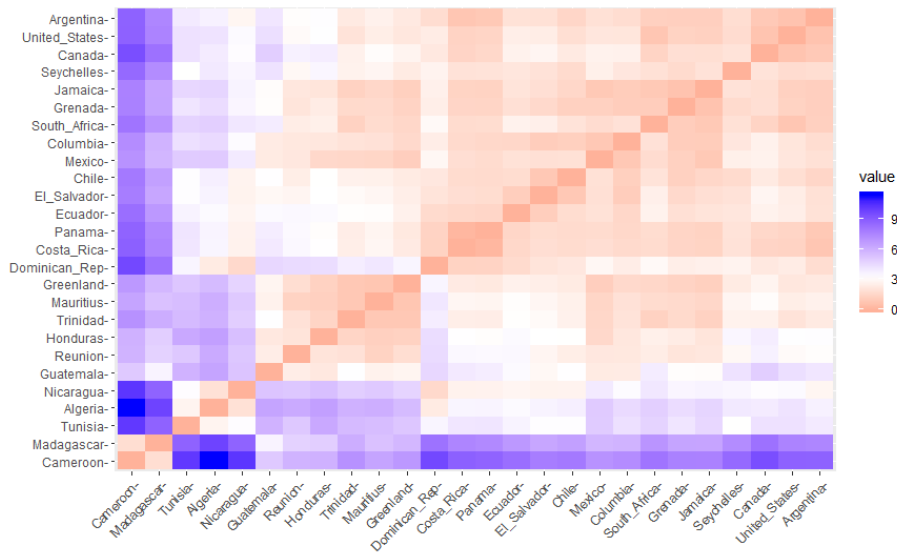
fviz_dist(d_st)



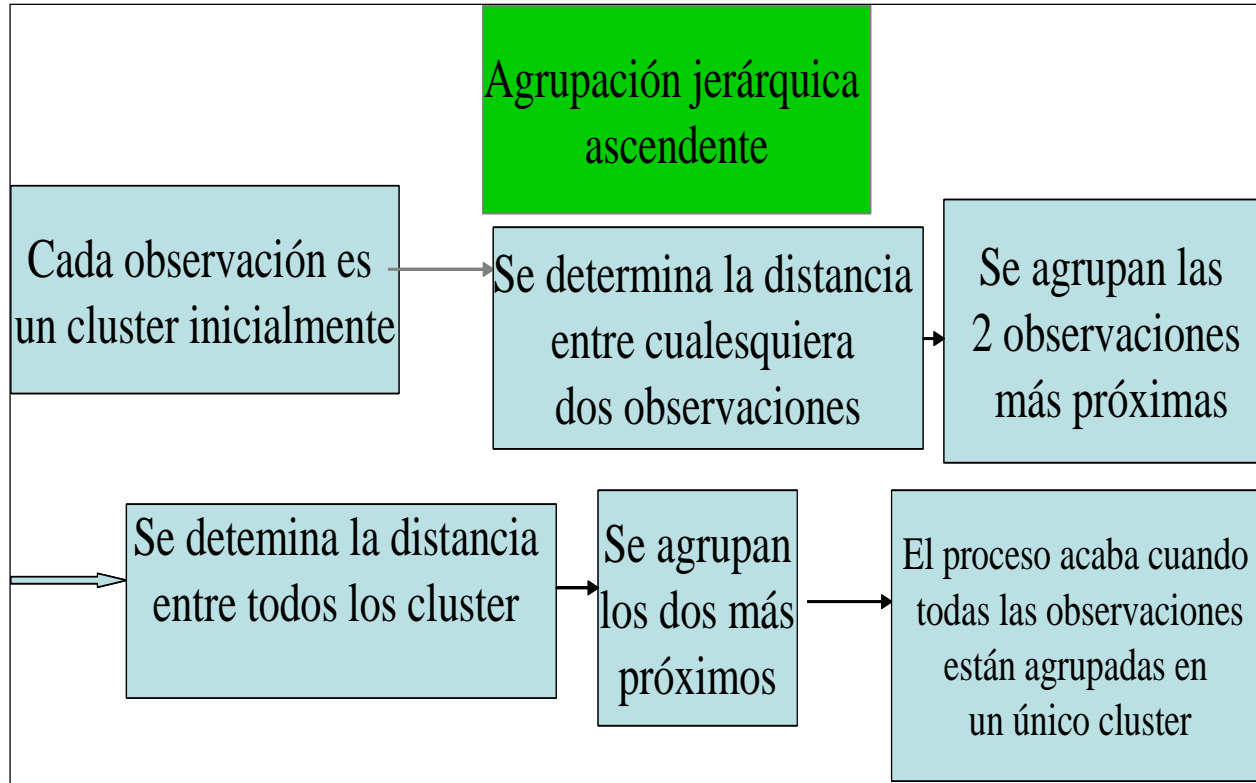
Datos originales



Datos estandarizados



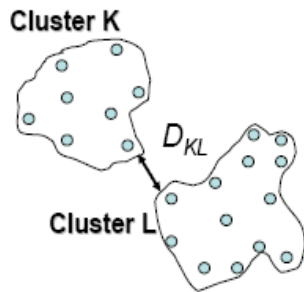
Modelos Jerárquicos



Enlace Simple o del vecino más cercano (single):

La distancia entre dos clústeres viene dada por la **distancia mínima** entre pares de observaciones cada una perteneciente a uno de los dos clústeres.

$$d(C_k, C_{k'}) = \min_{\substack{i=1, \dots, n_k \\ i'=1, \dots, n_{k'}}} d(x_{ki}, x_{k'i'})$$



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

Enlace simple: Tiende a crear grupos con **muchas observaciones y alargados**, que pueden incluir elementos muy distintos en los extremos.

Método del vecino más cercano

Distancia (euclídea) entre 6 observaciones

	1	2	3	4	5	6
1		0.31	0.23	0.32	0.26	0.25
2			0.34	0.21	0.36	0.28
3				0.31	0.04	0.07
4					0.31	0.28
5						0.09
6						

$$C_1 = \{[1],[2],[3,5],[4],[6]\}$$

	1	2	[3,5]	4	6
1		0.31	0.23	0.32	0.25
2			0.34	0.21	0.28
[3,5]				0.31	0.07
4					0.28
6					

$$C_2 = \{[1],[2],[3,5,6],[4]\}$$

	1	2	[3,5,6]	4
1		0.31	0.23	0.32
2			0.28	0.21
[3,5,6]				0.28
4				

$$C_3 = \{[1],[2,4],[3,5,6]\}$$

	1	[2,4]	[3,5,6]
1		0.31	0.23
[2,4]			0.28
[3,5,6]			

$$C_4 = \{[1,3,5,6],[2,4]\}$$

	[2,4]	[3,5,6]
[2,4]		0.28
[1,3,5,6]		

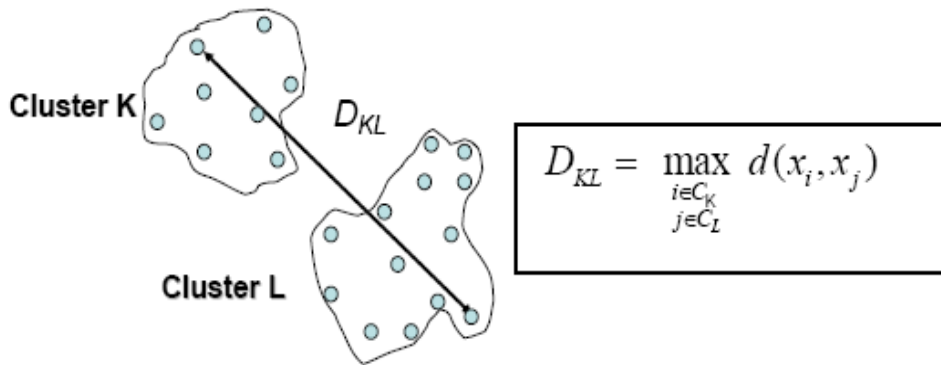
$$C_5 = \{[1,2,3,4,5,6]\}$$

Enlace Completo o del vecino más alejado (complete):

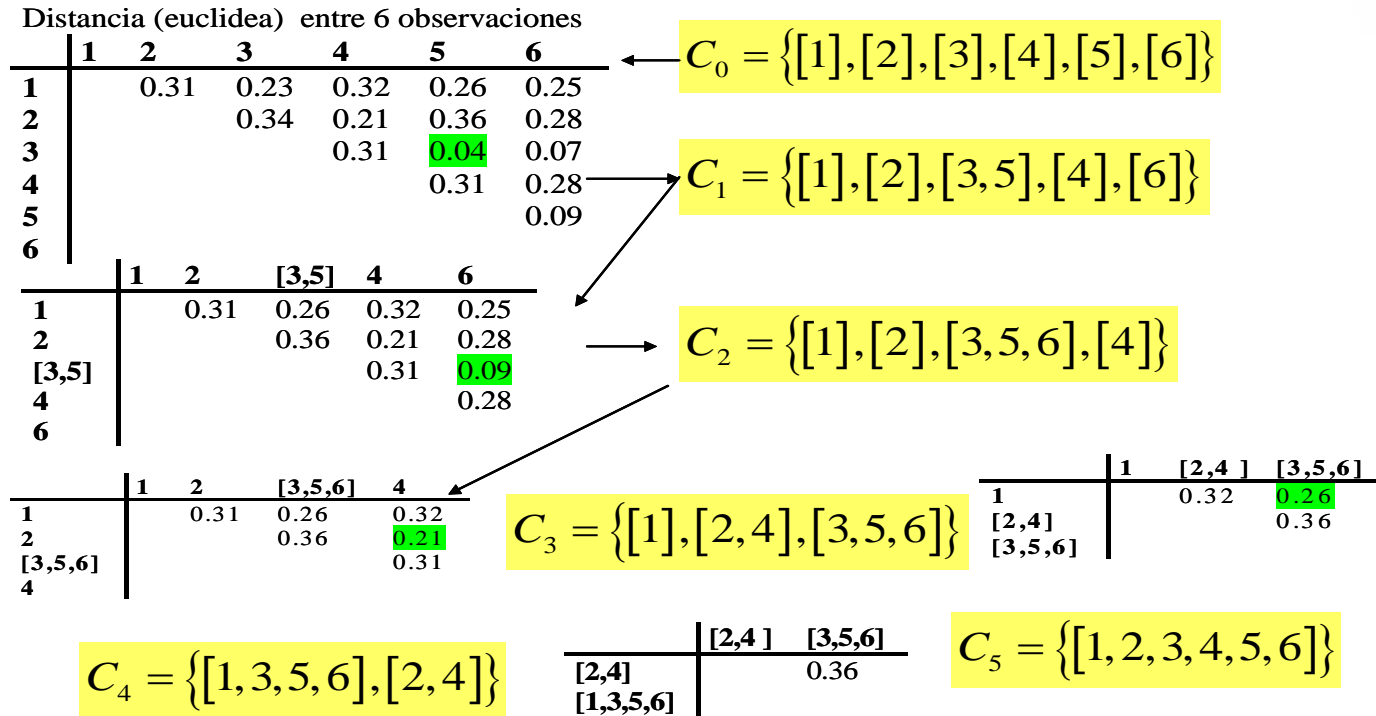
La distancia entre dos clústeres viene dada por la **distancia máxima** entre pares de observaciones cada una perteneciente a uno de los dos clústeres.

$$d(C_k, C_{k'}) = \max_{\substack{i=1, \dots, n_k \\ i'=1, \dots, n_{k'}}} d(x_{ki}, x_{k'i'})$$

Enlace más lejano: Los grupos obtenidos con este método son **más compactos** que los obtenidos con el método del vecino más próximo.



Método del vecino más alejado



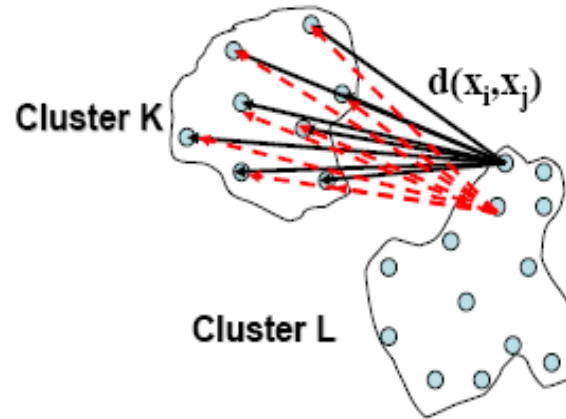
Distancias entre clústeres

Enlace medio (average):

La distancia entre dos clústeres viene dada por la **distancia media** entre observaciones de distintos grupos.

$$d(C_k, C_{k'}) = \frac{\sum_{i=1}^{c_1} \sum_{i'=1}^{c_2} d(x_{ki}, x_{k'i'})}{n_k n_{k'}}$$

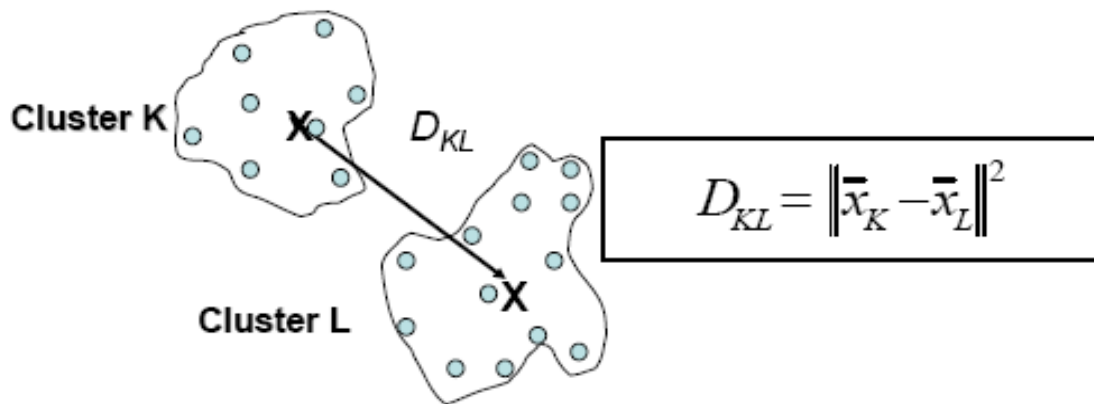
Enlace Medio: Los grupos así formados tienen **varianza similar y pequeña**.



Distancia entre centroides (centroid): La distancia entre dos clústeres viene dada por la distancia entre los centroides de cada grupo (**vector de medias** obtenido para las m variables desde los datos correspondientes a los individuos que formen parte del grupo).

$$d(C_k, C_{k'}) = d(\bar{x}_k, \bar{x}_{k'})$$

Enlace Centroide: Más sensible a datos extraños.



Método de Ward o de la mínima varianza (**Ward**):

Más que definir distancia entre cada dos clústeres, este método selecciona, entre todas las uniones posibles de dos clústeres, aquella unión que **minimiza la variabilidad interna** de los clústeres resultantes.

Este método tiende a generar **conglomerados pequeños** y equilibrados en tamaño

$$d(C_k, C_{k'}) = \frac{\sum_{j=1}^p (\bar{x}_{k,j} - \bar{x}_{k',j})^2}{\frac{1}{n_k} + \frac{1}{n_{k'}}}$$

¿Cuál es el método de agrupación más adecuado para definir la estructura de parecidos presente en los datos?

No existe una respuesta exacta a esta pregunta, aunque los tres últimos son los más utilizados.

Como técnica exploratoria es conveniente **estudiar varios métodos** y comparar resultados antes de tomar una decisión.

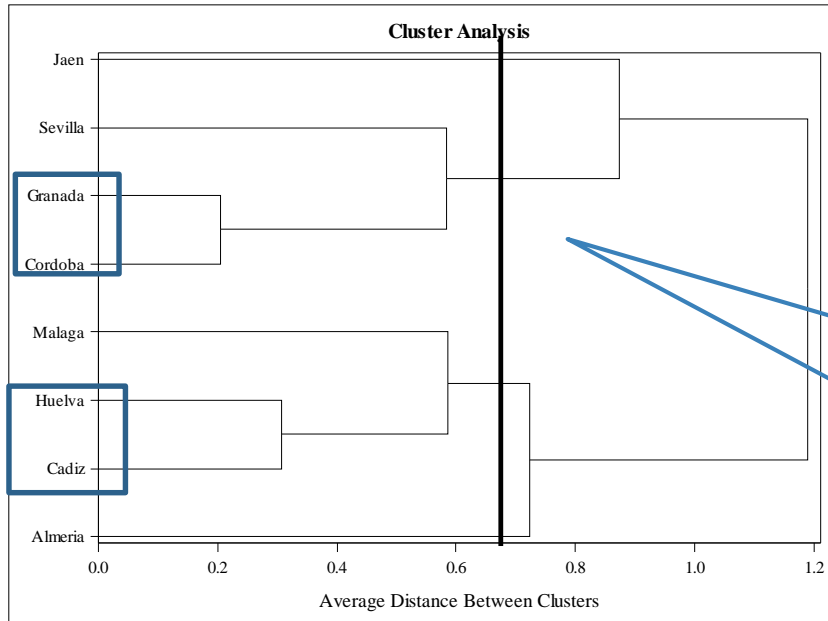
```
#Hacemos el cluster jerárquico con las distancias entre los datos sin estandarizar  
res.hc <- hclust(d, method="ward.D2")
```

Matriz de distancias

Método para medir las distancias entre clusters

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 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 $k = 4$.

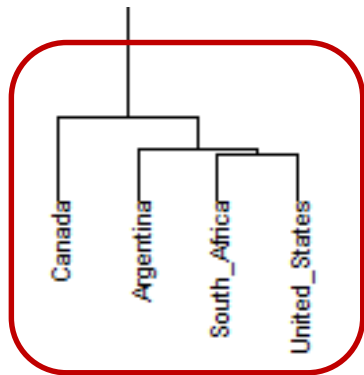
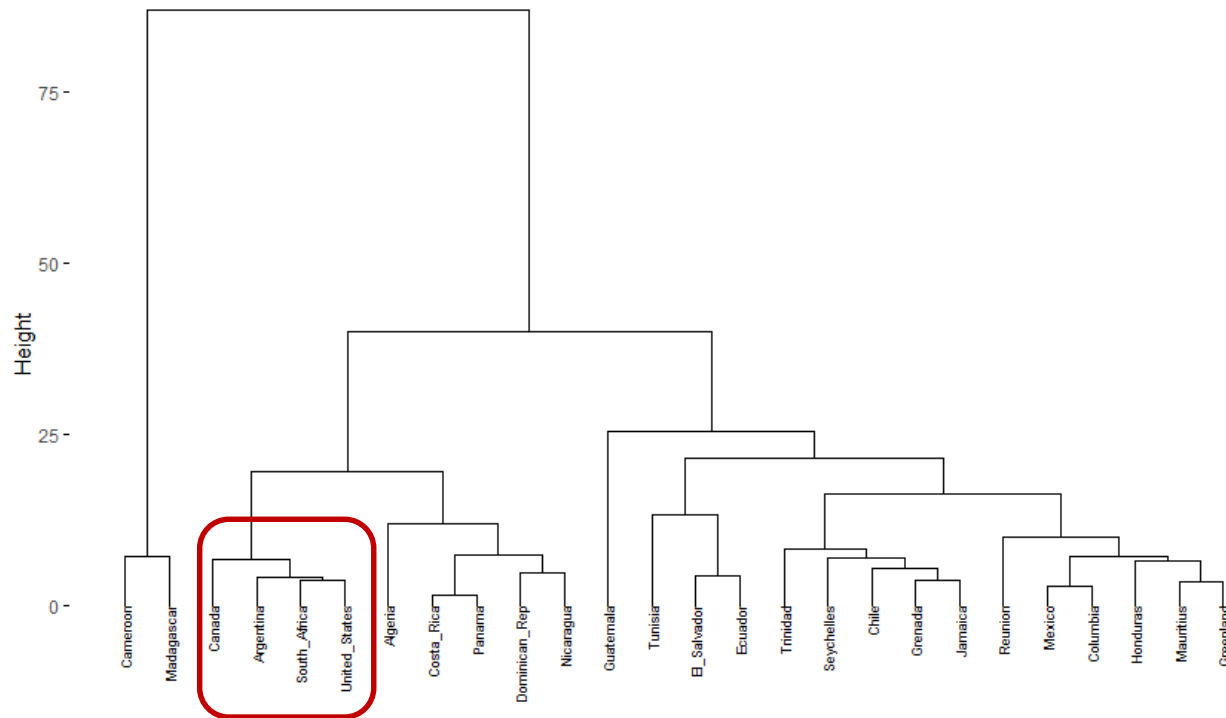
```
#Dibujamos el dendrograma correspondiente
```

```
library("factoextra")
```

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

Dendrograma con los datos sin estandarizar

Cluster Dendrogram



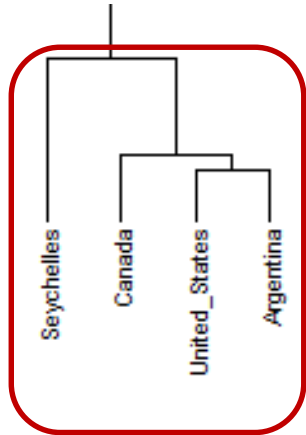
#Hacemos el cluster jerárquico con las distancias entre los
datos estandarizados

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

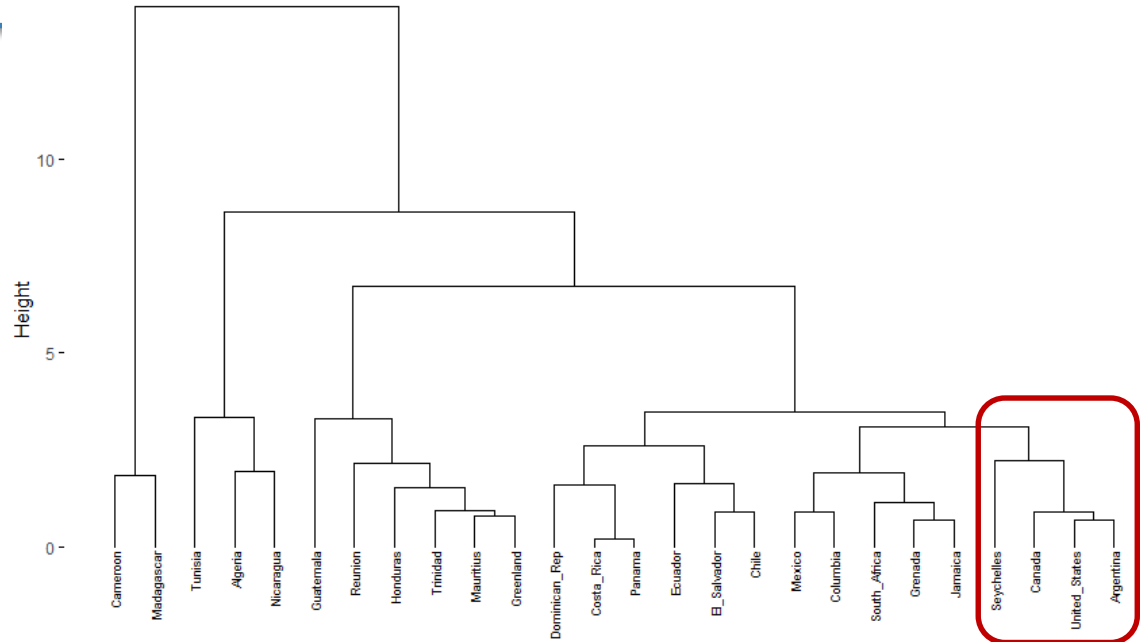
#Dibujamos el dendrograma correspondiente

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

Dendrograma con los datos estandarizados



Cluster Dendrogram



Seleccionamos el número de clusters que nos parece “lógico”

```
# Cut tree into 4 groups  
grp <- cutree(res.hc_st, k = 4)  
head(grp, n = 4)
```

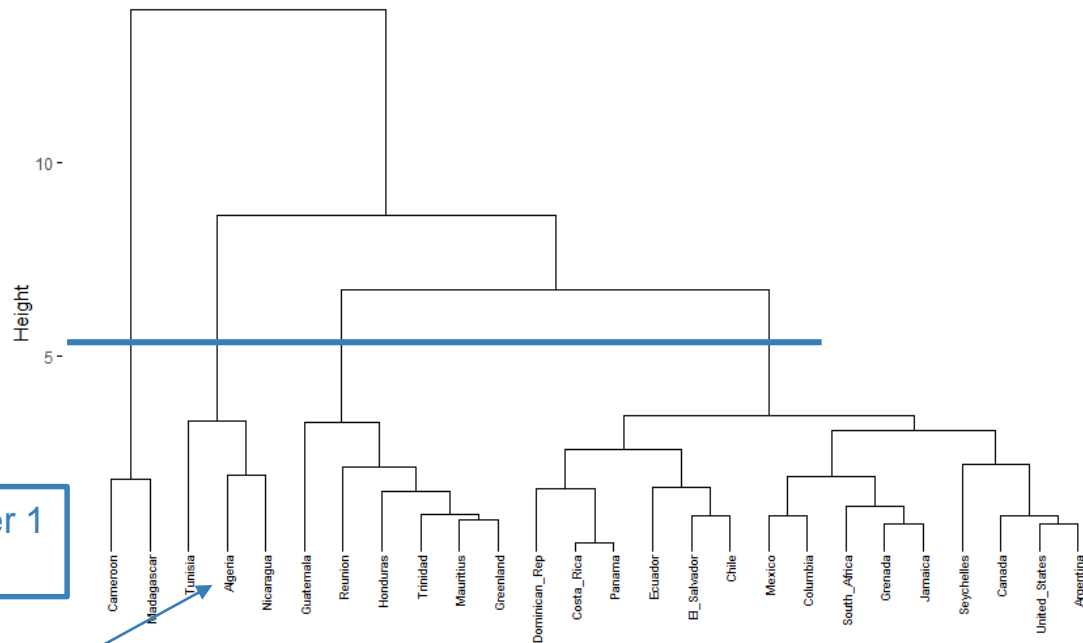
```
# Number of members in each cluster  
table(grp)
```

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

```
# Get the names for the members of cluster 1  
rownames(dat_EV)[grp == 1]
```

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

Cluster Dendrogram




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

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

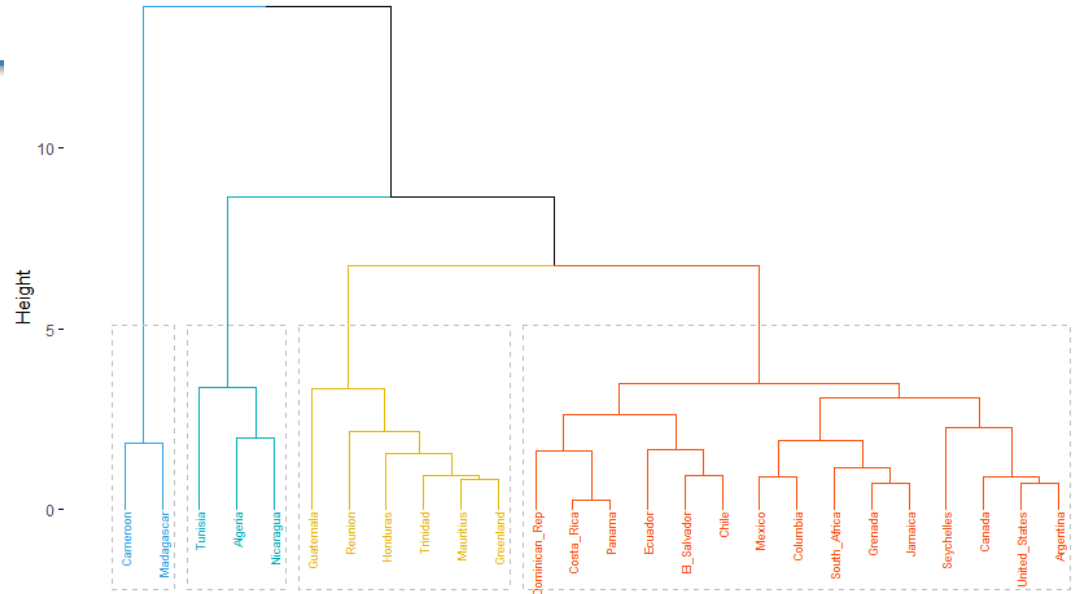
```
cex = 0.5, # label size
```

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

```
color_labels_by_k = TRUE, # color labels by groups
```

```
rect = TRUE) # Add rectangle a
```

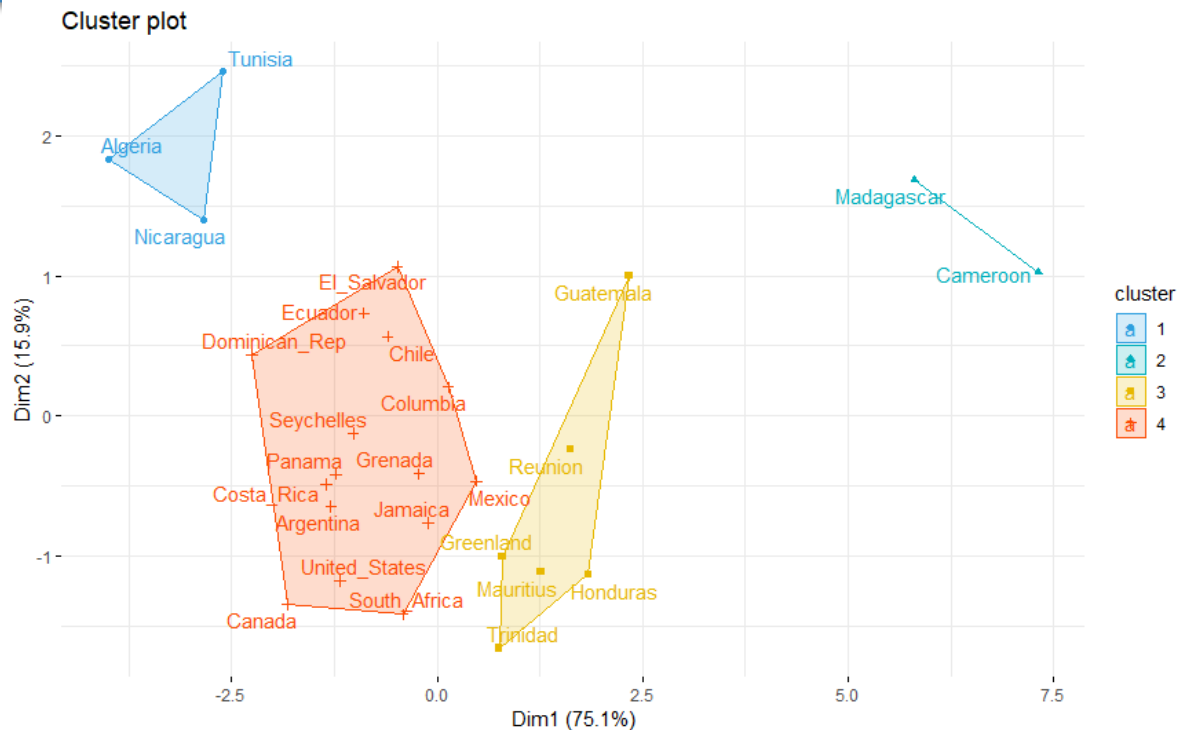
Cluster Dendrogram



#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())
```

Representamos los
países en los planos de
las dos primeras
Componentes
Principales

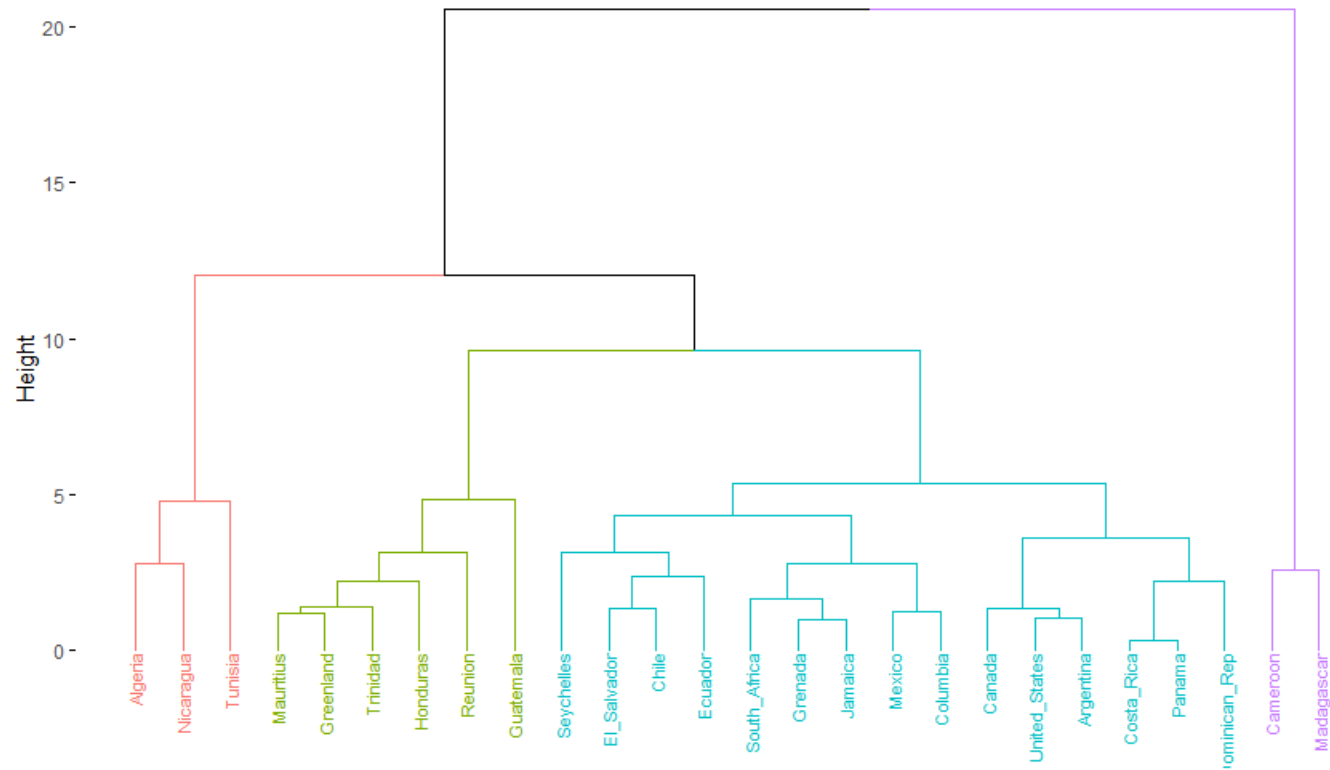


Podemos realizar los pasos anteriores a las representaciones con las siguientes funciones

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

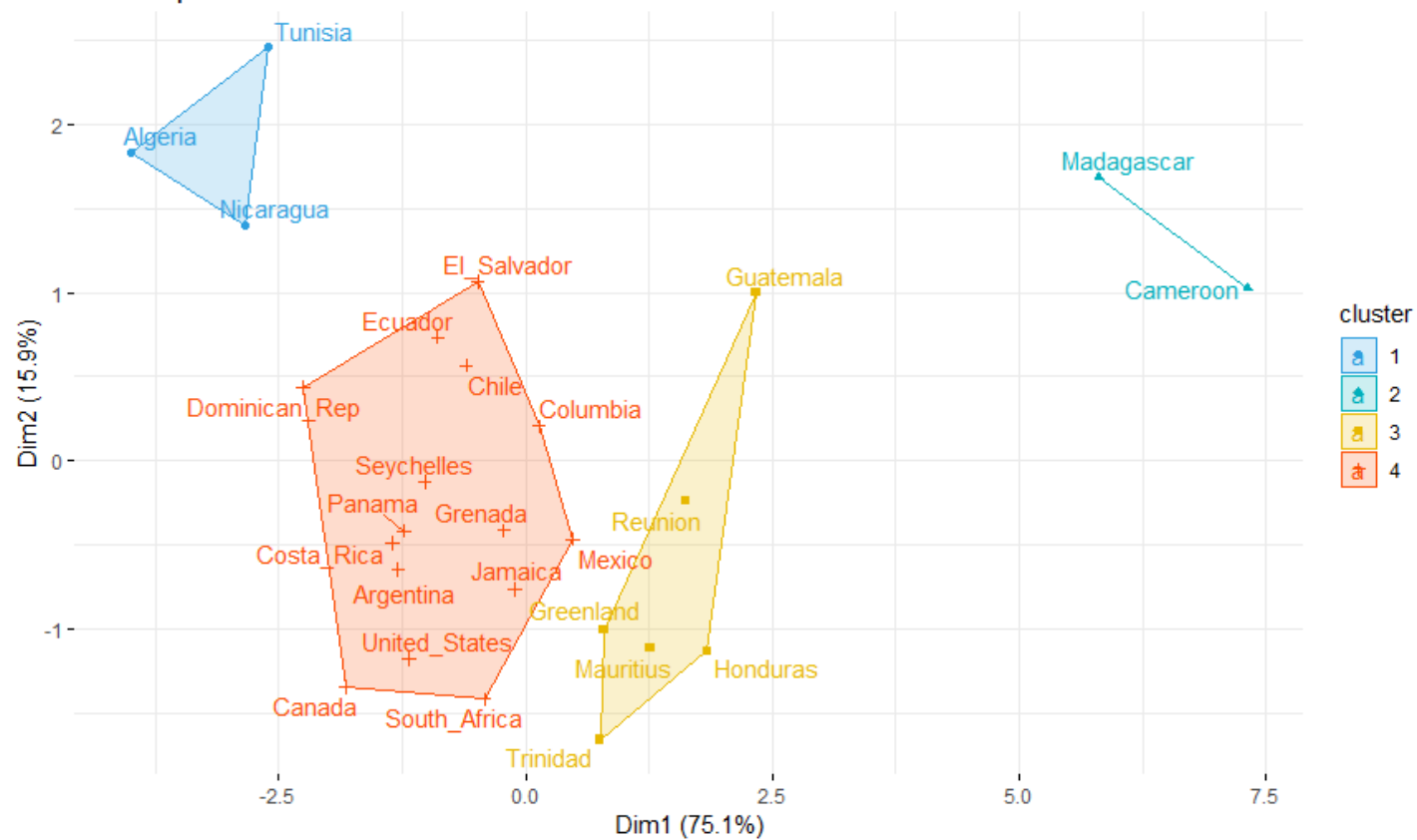


Cluster Dendrogram



```
fviz_cluster(list(data = dat_EV, 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())
```

Cluster plot



II.5.- Algoritmos de clasificación no jerárquica

En el análisis clúster no jerárquico es necesario **fijar de antemano** el número **k** de grupos en que se pretende dividir las observaciones. La clasificación admite variantes dependiendo de:

- El modo de **escoger k semillas iniciales** para generar los k grupos
- El criterio empleado para relacionar cada **observación con cada una de ellas**.

Pasos del algoritmo:

1. **Seleccionar k puntos** como semillas iniciales de los clústeres a construir, siendo k el número deseado de clústeres.
2. **Asignar** cada una de las observaciones restantes al clúster **más próximo**.
3. **Redefinir** las K semillas.
4. **Reasignar** cada observación a uno de los k clústeres de acuerdo con el criterio de proximidad.
5. **Parar** si no se reasignan observaciones de forma distinta a como se hizo en la iteración anterior, o si la reasignación satisface alguna otra regla de parada. En caso contrario, volver a 3.



Algunos métodos para definir las semillas iniciales

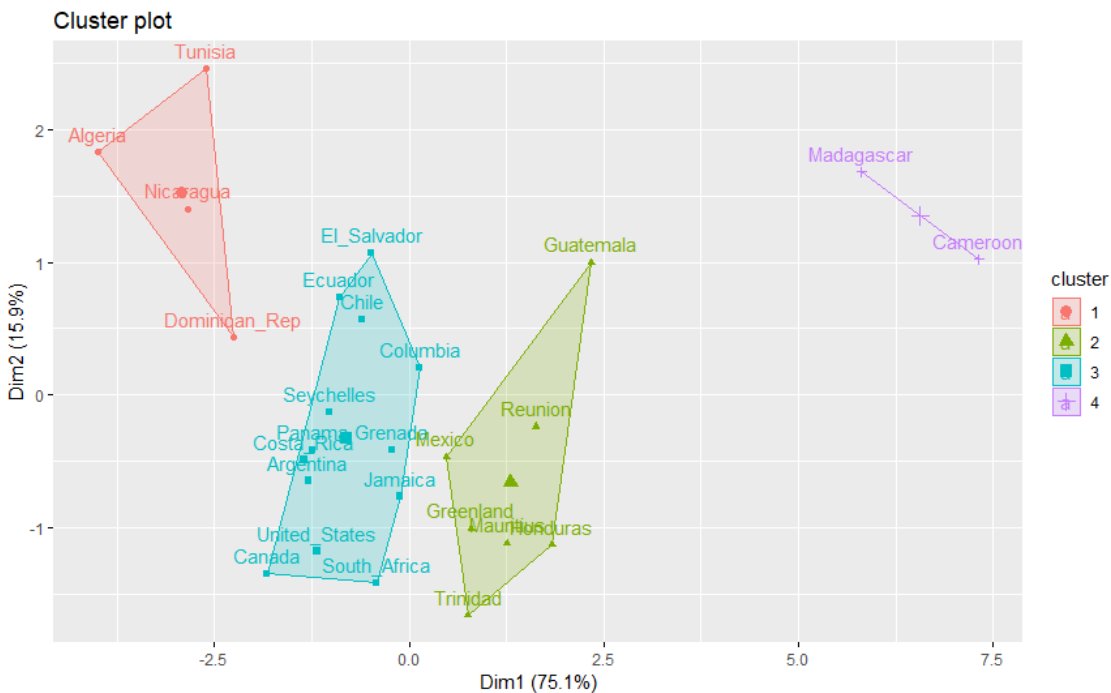
- Seleccionar las **k primeras observaciones** con datos no-missing.
- Seleccionar la **primera observación** como primera semilla.
 1. La segunda semilla será aquella observación cuya distancia a la primera sea tan **grande** como una distancia predefinida.
 2. La tercera semilla será la observación cuya **distancia** a las dos primeras sea tan grande como la distancia prefijada.
 3. Y así sucesivamente.
- Seleccionar **aleatoriamente k** observaciones con datos conocidos.
- Elegir semillas que estén entre sí lo más **lejanas** posible.
- Utilizar k semillas que propone el **investigador**.



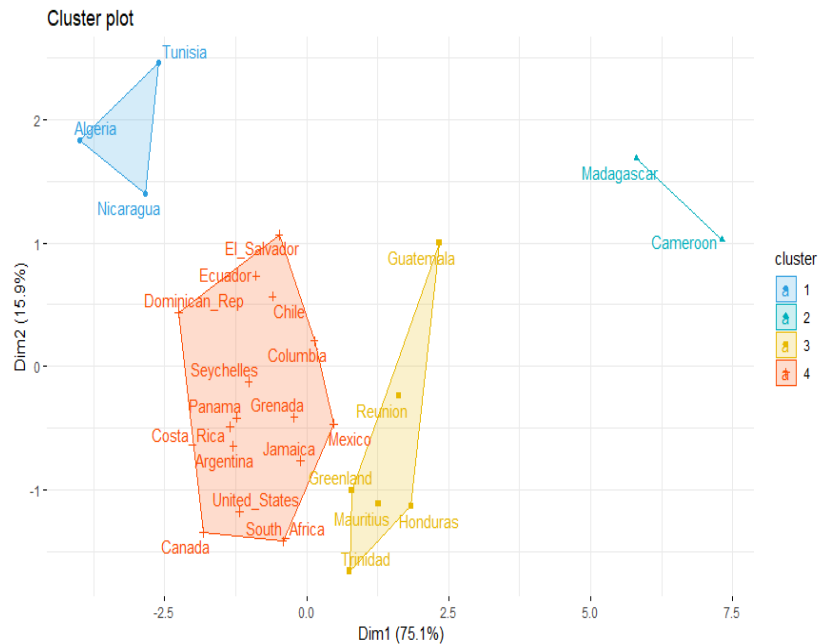

```
# Standardize the data
datos_ST <- scale(dat_EV)
```

```
# Compute k-means
set.seed(123)
km.res <- kmeans(datos_ST, 4)
head(km.res$cluster, 20)
```

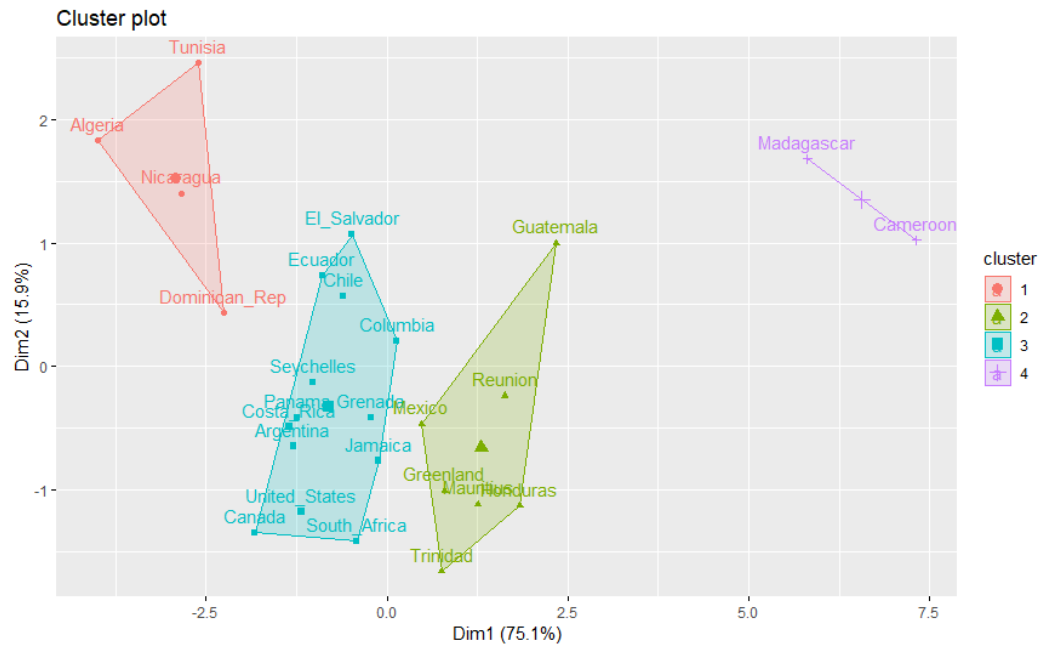
```
# Visualize clusters using factoextra
library("factoextra")
fviz_cluster(km.res, datos)
```



Jerárquico



No Jerárquico



¿Cuál es el número óptimo de clusters?

II.4.- Procedimientos para determinar el número de clusters

Para determinar el **número de clusters** existentes en nuestros datos, serán de utilidad las siguientes medidas donde i representa observación, j variable, k clúster:

Variabilidad total:

$$T = \sum_{j=1}^p \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

Variabilidad dentro del clúster k :

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

Variabilidad total intra-clústeres:

$$W = \sum_k W_k$$

Variabilidad total entre-clústeres:

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

Se demuestra que: $T = W + E$

$$T = \sum_{j=1}^p \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

totss	double [1]	200
withinss	double [4]	10.95 11.15 17.82 1.68
tot.withinss	double [1]	41.59675
betweenss	double [1]	158.4032
size	integer [4]	4 7 13 2

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

$$W = \sum_k W_k$$

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

```
#Determinación del número óptimo de clusters  
install.packages("NbClust")  
library(NbClust)
```

Elbow method

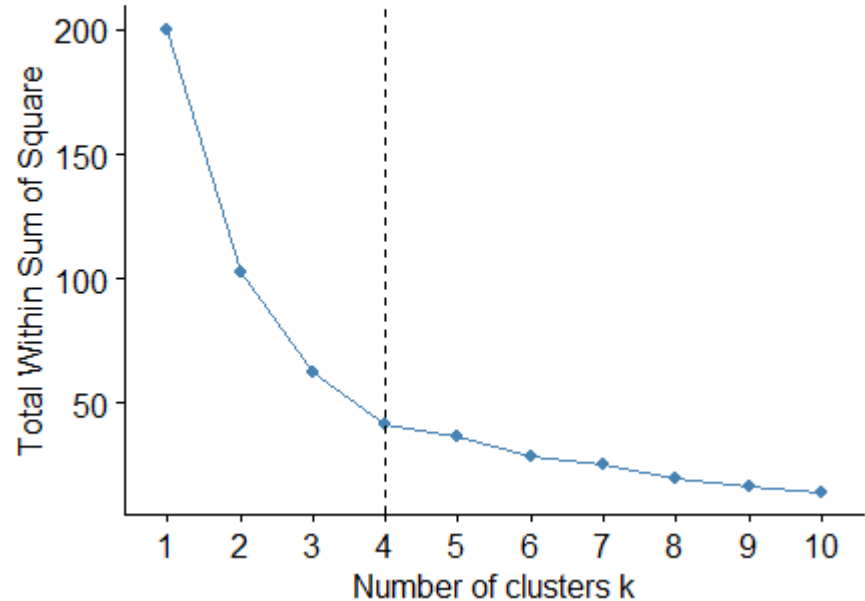
```
fviz_nbclust(datos_st, kmeans, method = "wss") +  
  geom_vline(xintercept = 4, linetype = 2)+  
  labs(subtitle = "Elbow method")
```

Aquel número de clusters en el que la
Variabilidad total intra-clústeres ya no se reduce
de forma significativa al aumentar uno más

$$W_k = \sum_{j=1}^p \sum_{i \in C_k} (x_{ij} - \bar{x}_{jk})^2$$
$$W = \sum_k W_k$$

Optimal number of clusters

Elbow method



```
# Silhouette method
```

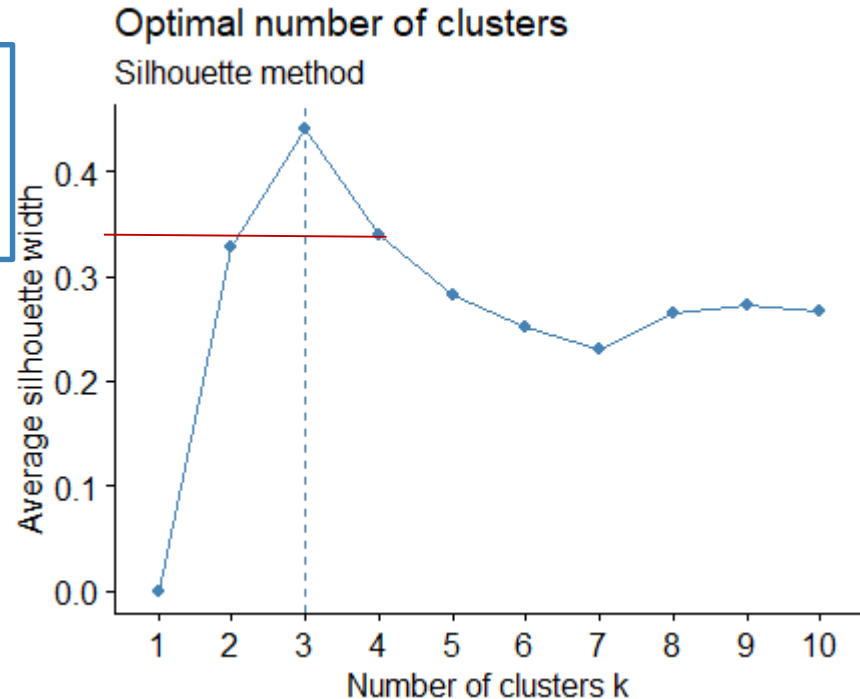
```
fviz_nbclust(datos_st, kmeans, method = "silhouette")+  
  labs(subtitle = "Silhouette method")
```

Es una medida de como de compactos son los clusters y cuanto de separados están unos de otros

Cuanto mayor sea su valor mejor

$$s(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}}$$

$$\bar{s} = \frac{\sum_i s(i)}{n}$$



#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)
```

$$s(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}}$$

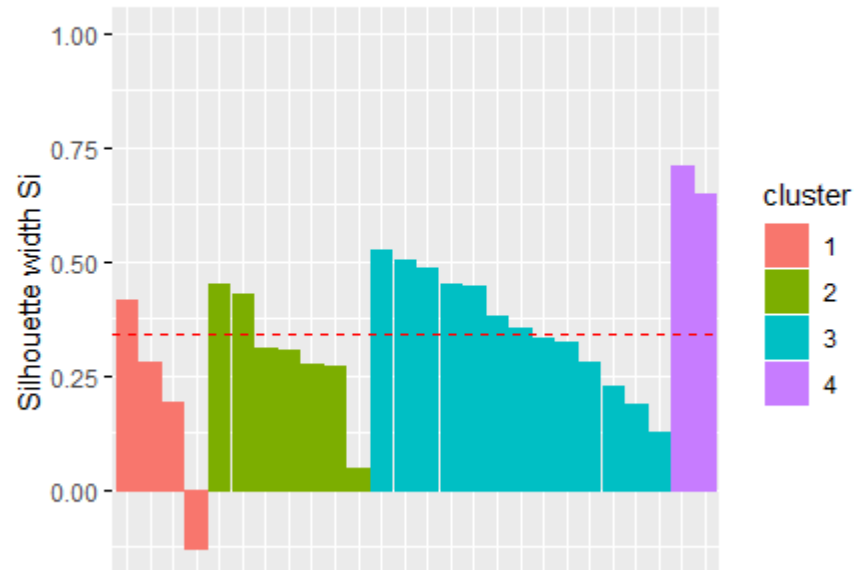
Algeria	1	3	0.4147402
Cameroon	4	2	0.7116156
Madagascar	4	2	0.6480743
Mauritius	2	3	0.4295660
Reunion	2	3	0.3098000
Seychelles	3	2	0.3536013

Evaluación de la calidad de los clusters

$a(i)$ = distancia media de la observación i -ésima a las observaciones de su cluster

$b(i)$ = distancia media de la observación i -ésima a las observaciones de otros clusters

Clusters silhouette plot
Average silhouette width: 0.34



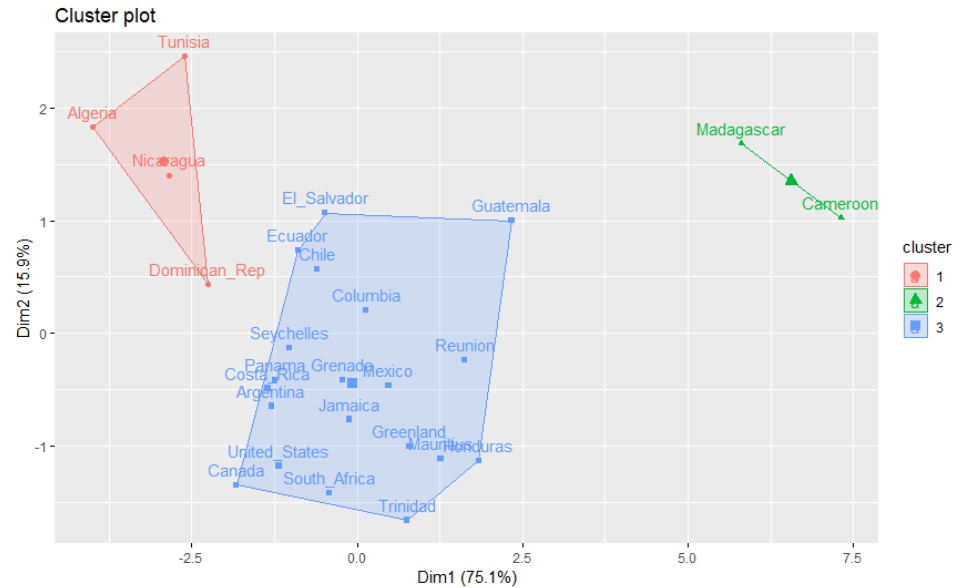
```
# Probamos con 3 Clusters que es lo que nos  
recomienda el criterio Silouette
```

```
set.seed(123)
```

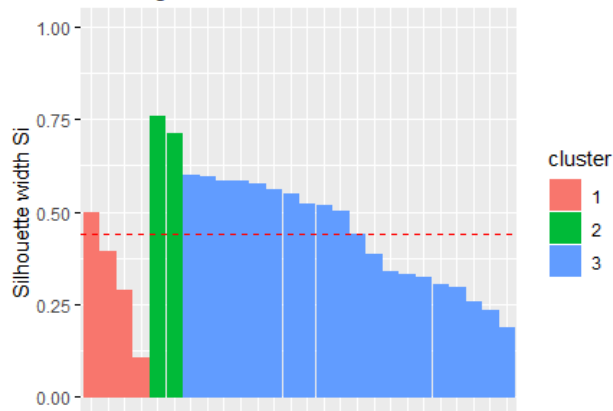
```
km.res <- kmeans(datos_st, 3)
```

```
# Visualize clusters using factoextra
```

```
fviz_cluster(km.res, datos_st)
```



Clusters silhouette plot
Average silhouette width: 0.44



```
#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)
```


II.6.- Caracterización de los clústeres

Una vez que se ha decidido la partición de los clústeres, se desea **caracterizarlos**:

- Por un lado, se realiza un análisis descriptivo sobre las **variables activas** utilizadas en el análisis, con lo que se determinarán las medias y varianzas de todas las variables.
- Un gráfico box-plot de cada una de ellas según clúster puede ser útil.
- Los diagramas de dispersión con marcas de clúster pueden ser útiles.



```
print(km.res)
```

Mostramos los estadísticos que identifican a los clusters

K-means clustering with 4 clusters of sizes 4, 7, 13, 2

Cluster means:

	m0	m25	m50	m75	w0	w25	w50	w75
1	0.3516843	1.0104699	1.1855842	1.75526068	0.2793340	0.9587342	1.3928580	1.5380631
2	-0.1425747	-0.3562519	-0.5247420	-0.74334463	-0.1676004	-0.4955725	-0.6258107	-0.7007392
3	0.4087142	0.3126737	0.2646393	0.03563981	0.4426369	0.3952034	0.2512607	0.1108514
4	-2.8609996	-2.8064371	-2.2547271	-1.14047395	-2.8492065	-2.7517866	-2.2285728	-1.3440731

Clustering vector: [1] 1 4 4 2 2 3 3 1 3 3 1 3 2 3 2 2 3 2 1 3 2 3 3 3 3 3

#Se puede 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

- 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 Analysis in R. A. Kassambara. Ed. STHDA. 2017
- <http://www.sthda.com/english/>

UNIVERSIDAD
COMPLUTENSE
DE MADRID

