



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**).

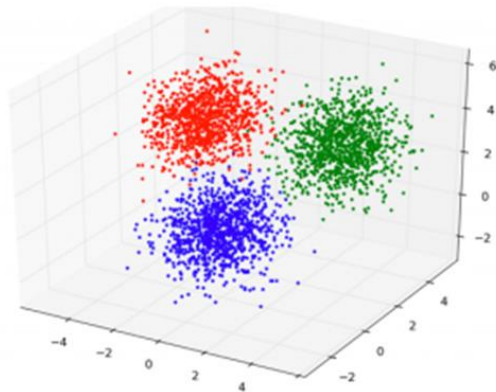


II.1.- Introducción: *Ejemplos*

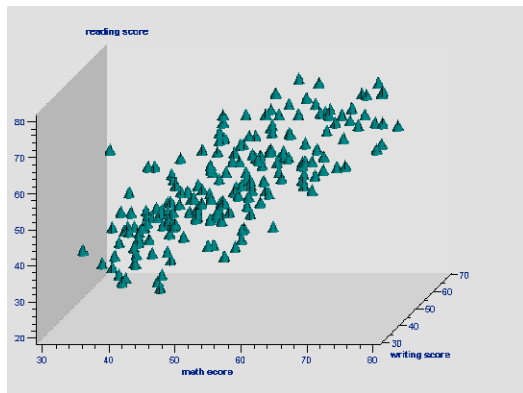
- El departamento de **marketing** de una empresa va a lanzar una campaña publicitaria sobre un nuevo producto. Para ello, desea tener a sus potenciales clientes agrupados según sus necesidades en los distintos aspectos de dicho producto.
- Se desea agrupar a los **clientes de un banco** para determinar diferentes perfiles a los que se les puede conceder un préstamo.
- **En estudios genéticos** es habitual encontrar grupos de individuos con carga genética similar.

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:



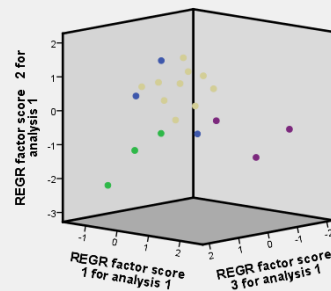
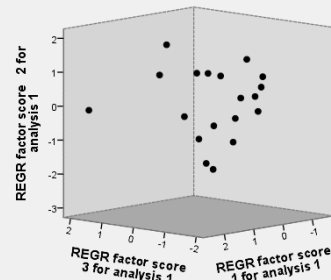
Situaciones con clústeres perfectamente definidos:
Una buena metodología los encontrará



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 de cada variables, por lo que **es conveniente siempre estandarizar las variables** (restar la media y dividir por la desviación estándar). Solo en algunas ocasiones, al homogeneizar la varianza de todas las variables podemos mermar la “capacidad clasificatoria” de alguna variable con gran variabilidad por grupos y en ese caso sería mejor no estandarizar.
- 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("heatmaply")  
install.packages("factoextra")  
install.packages("factoMineR")  
install.packages("NbClust")
```



```
library(Cluster)  
library(ggplot2)  
library("heatmaply")  
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

Importar La base de datos de Esperanza

```
EsperanzaVida <- read_excel("C:/Users/reven/OneDrive/Desktop/Master Big data/Clases/  
Cluster/Cluster/EsperanzaVida.xlsx")  
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?

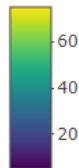


```
heatmaply(dat_EV, seriate = "mean", row_dend_left = TRUE, plot_method = "plotly")
```

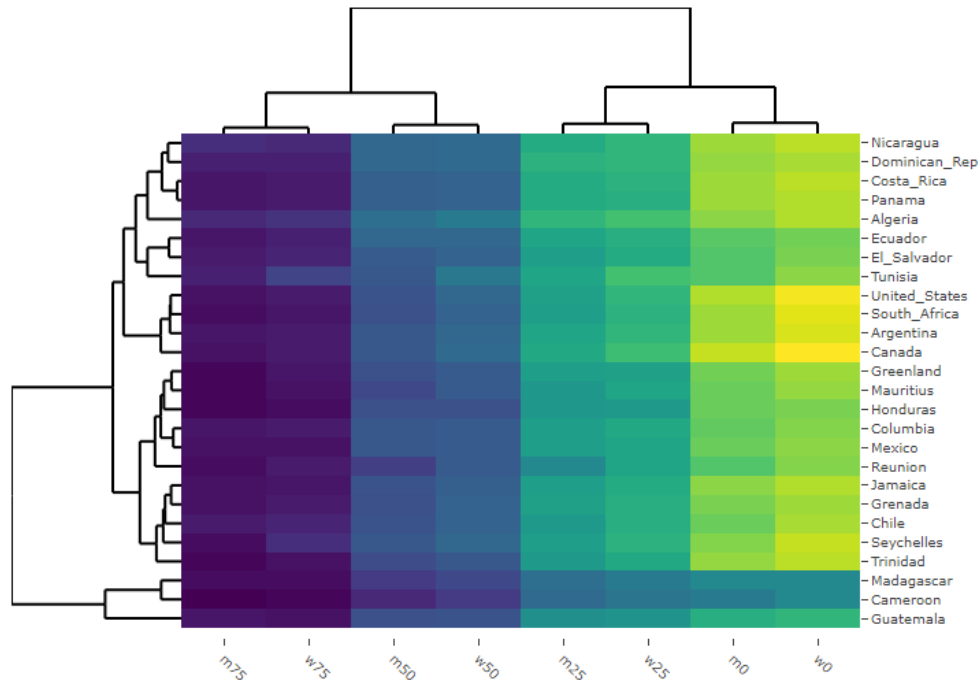
Exploración inicial del fichero de datos:

Creamos un **mapa de calor interactivo** con las filas y columnas ordenadas de forma que estén juntas las mas parecidas.

Cluster de individuos



Cluster de variables



Puesto que **el color es el valor de cada variable**, el mismo color en una variable indica grupos. Tenemos una primera aproximación de los países que tienen valores de las variables más parecidos

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")
```

#Mostramos las primeras seis filas dela matriz de distancias

```
d6<-as.matrix(d)[1:6, 1:6]
```

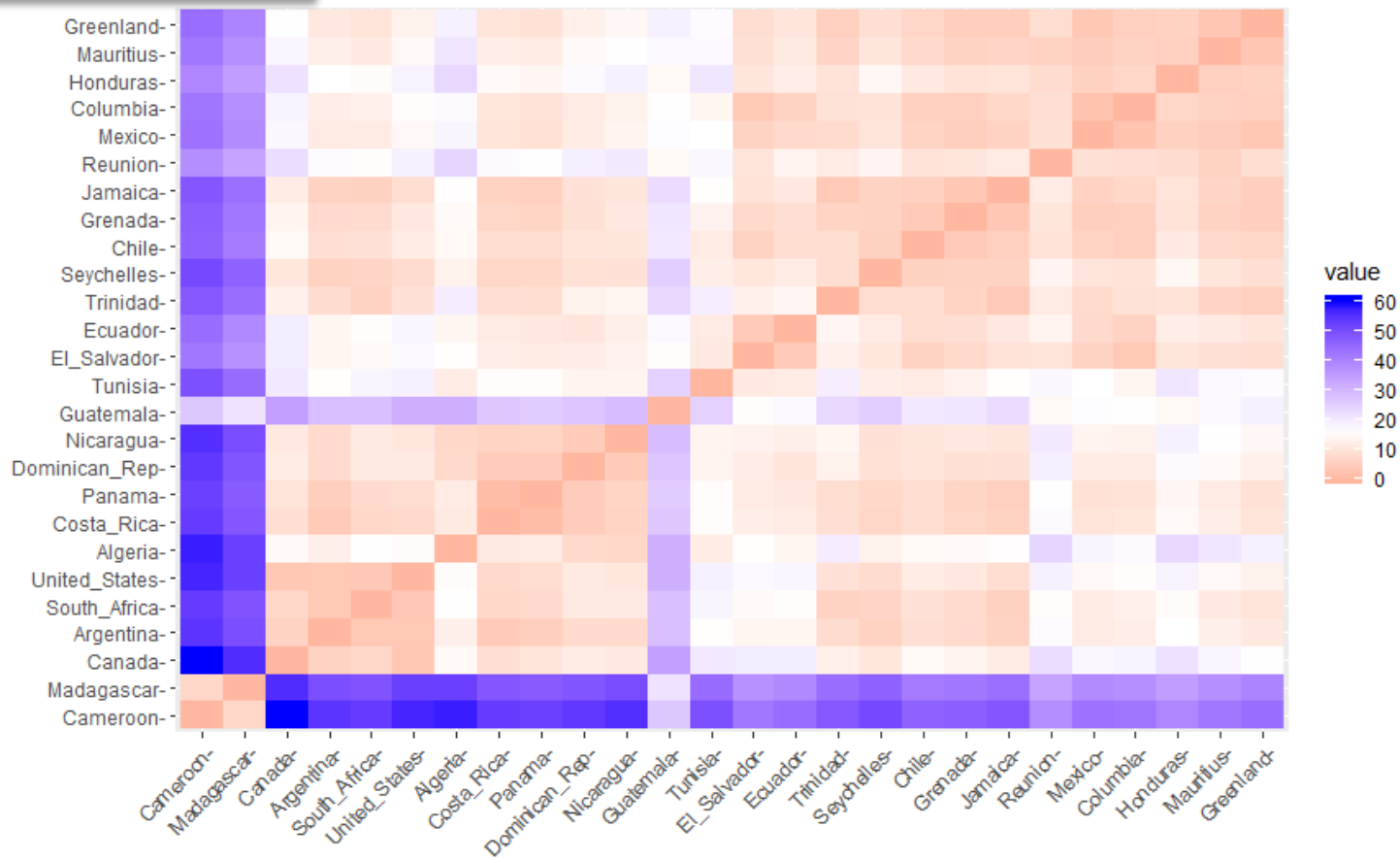
```
knitr::kable(d6, digits =2,caption = "Distancias")
```

	Algeria	Cameroon	Madagascar	Mauritius	Reunion	Seychelles
Algeria	0.00	58.08	52.65	21.19	24.35	13.38
Cameroon	58.08	0.00	7.14	42.24	38.03	51.03
Madagascar	52.65	7.14	0.00	37.96	33.81	46.38
Mauritius	21.19	42.24	37.96	0.00	6.16	10.77
Reunion	24.35	38.03	33.81	6.16	0.00	14.07
Seychelles	13.38	51.03	46.38	10.77	14.07	0.00

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

Representamos mediante escalas de color la distancia entre todas las observaciones

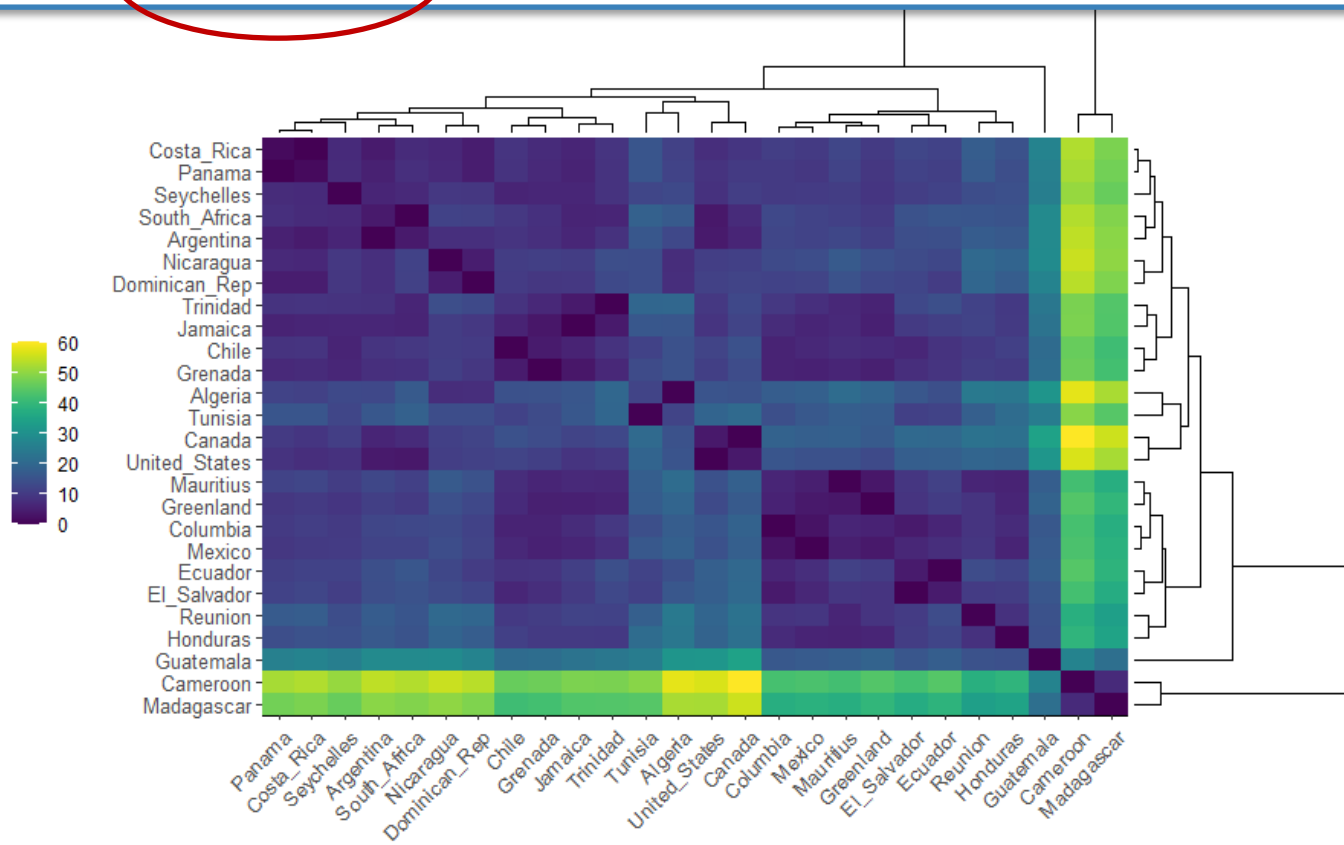
```
fviz_dist(d, show_labels = TRUE)
```



#Reordenamos para agrupar las observaciones que están más próximas y visualizar los posibles clusters

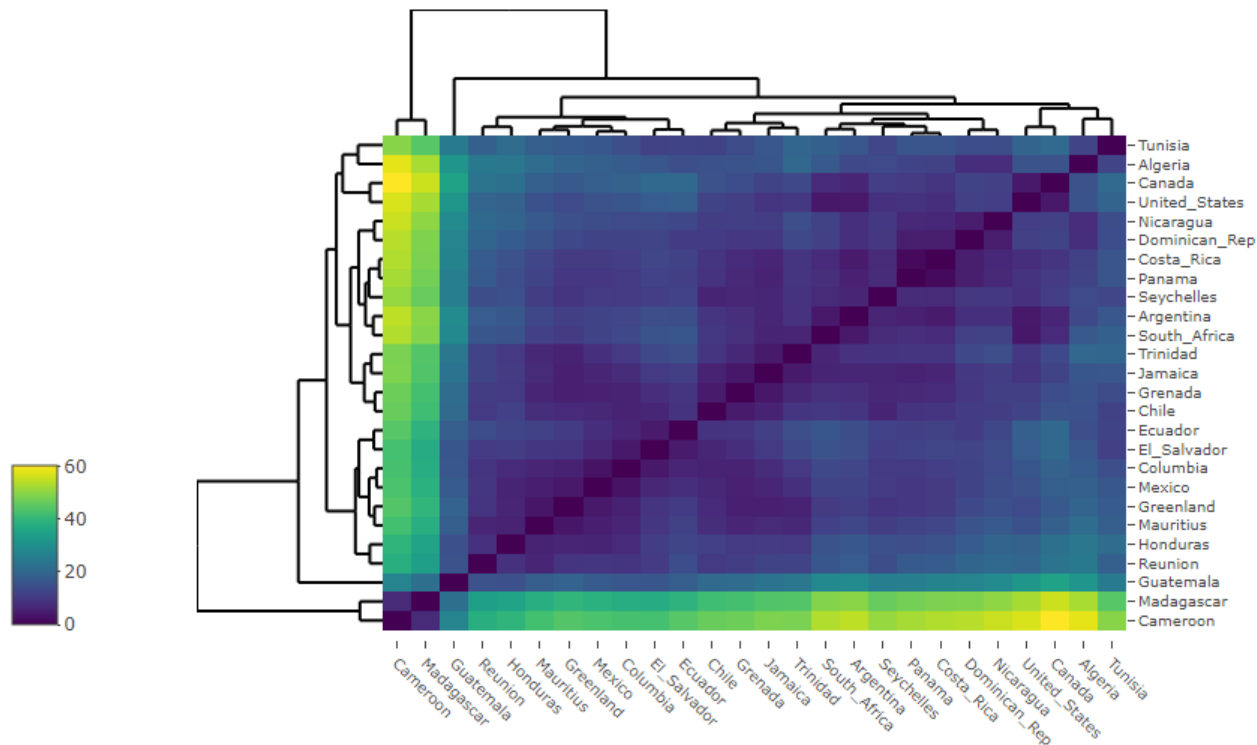
```
ggheatmap(as.matrix(d), seriate="mean")
```

La entrada es una matriz normal, no de distancias



```
heatmaply(as.matrix(d), seriate = "OLO", row_dend_left = TRUE, plot_method = "plotly")
```

Esta función nos permite
hacer **un mapa interactivo**
y con mayor flexibilidad en
los argumentos, además
hemos cambiado el
algoritmo de ordenación



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

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

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

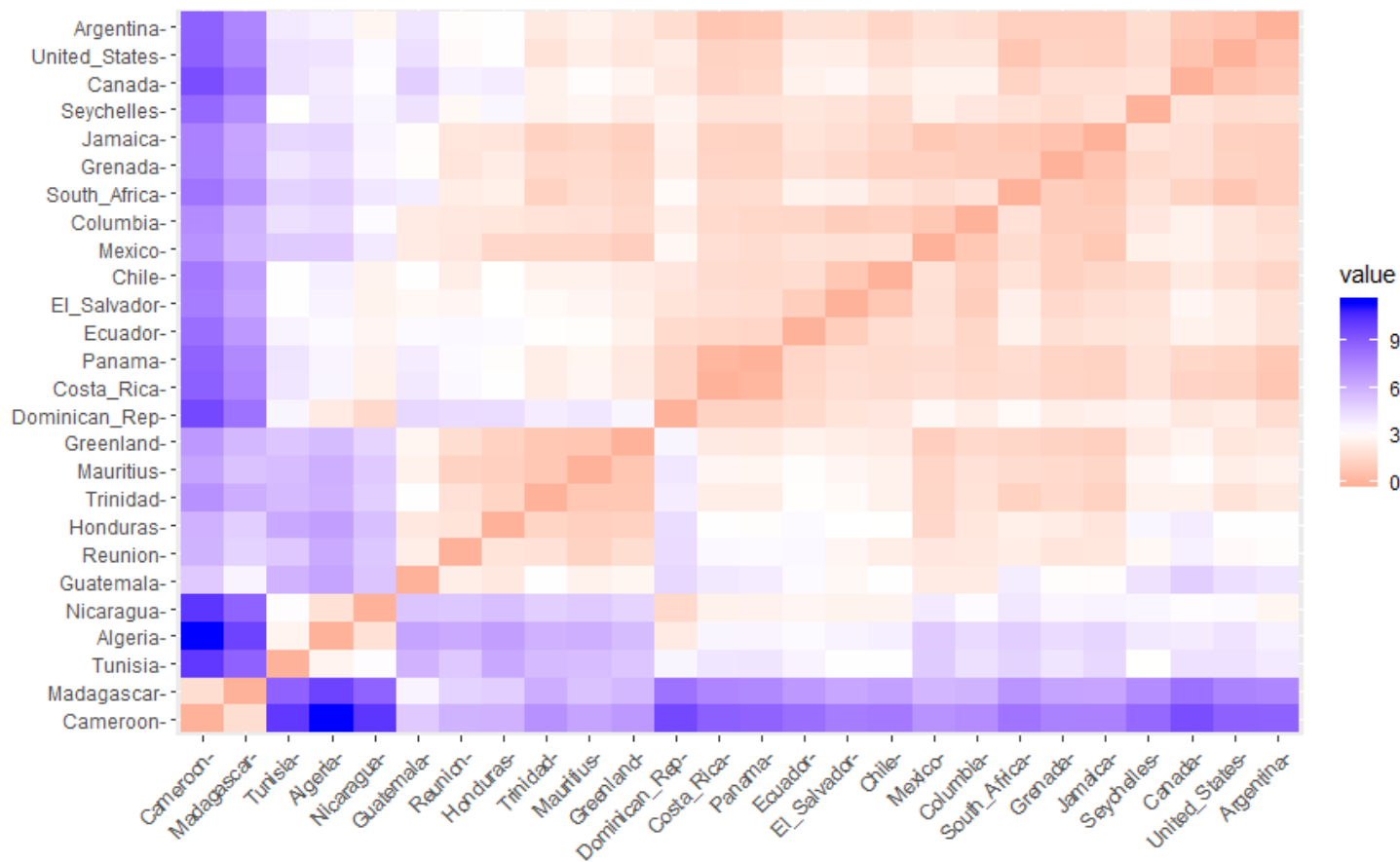
```
d_st <- dist(datos_ST, method = "euclidean")
```

```
d_st6 <- as.matrix(d_st)[1:6, 1:6]
```

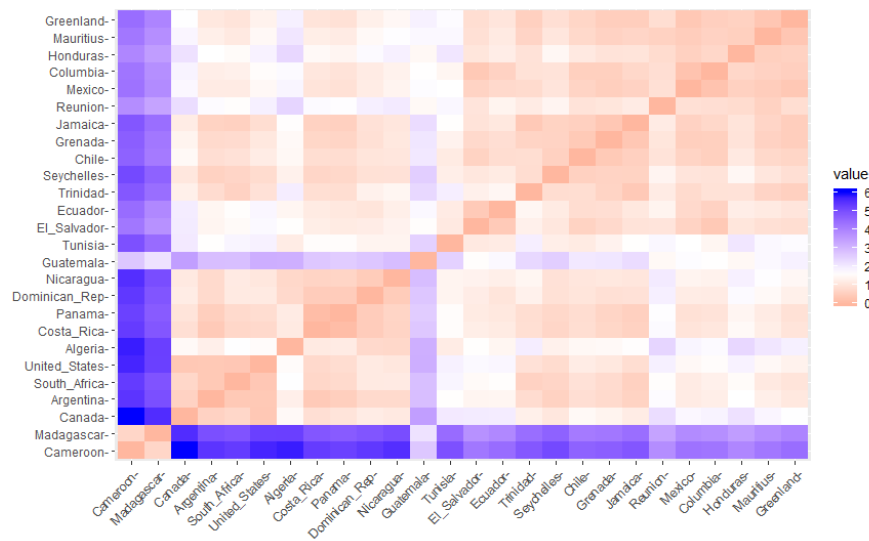
```
knitr::kable(d_st6, digits = 2, caption = "Distancias datos estandarizados")
```

	Algeria	Cameroon	Madagascar	Mauritius	Reunion	Seychelles
Algeria	0.00	11.37	9.82	6.07	6.20	4.01
Cameroon	11.37	0.00	1.83	6.43	5.90	8.52
Madagascar	9.82	1.83	0.00	5.39	4.81	7.28
Mauritius	6.07	6.43	5.39	0.00	1.36	2.83
Reunion	6.20	5.90	4.81	1.36	0.00	2.95
Seychelles	4.01	8.52	7.28	2.83	2.95	0.00

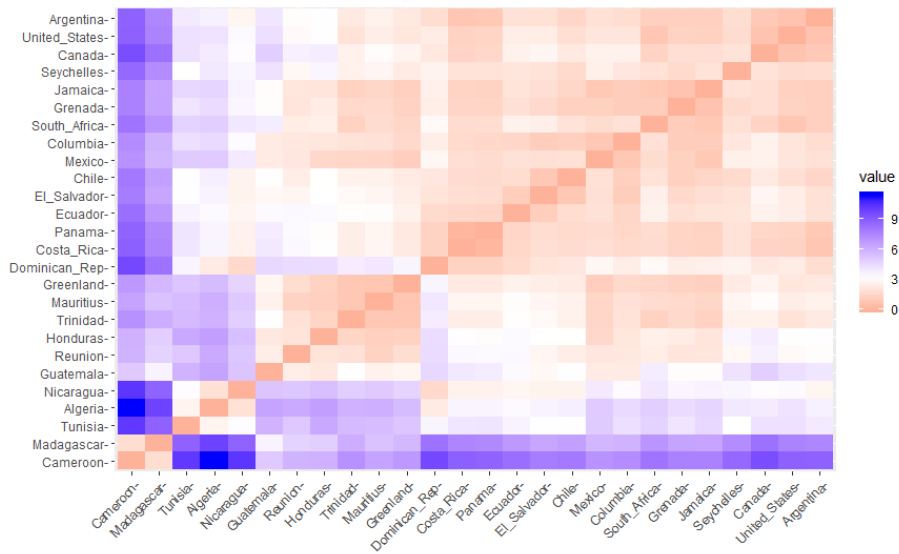
```
#Visualizamos  
fviz_dist(d_st)
```



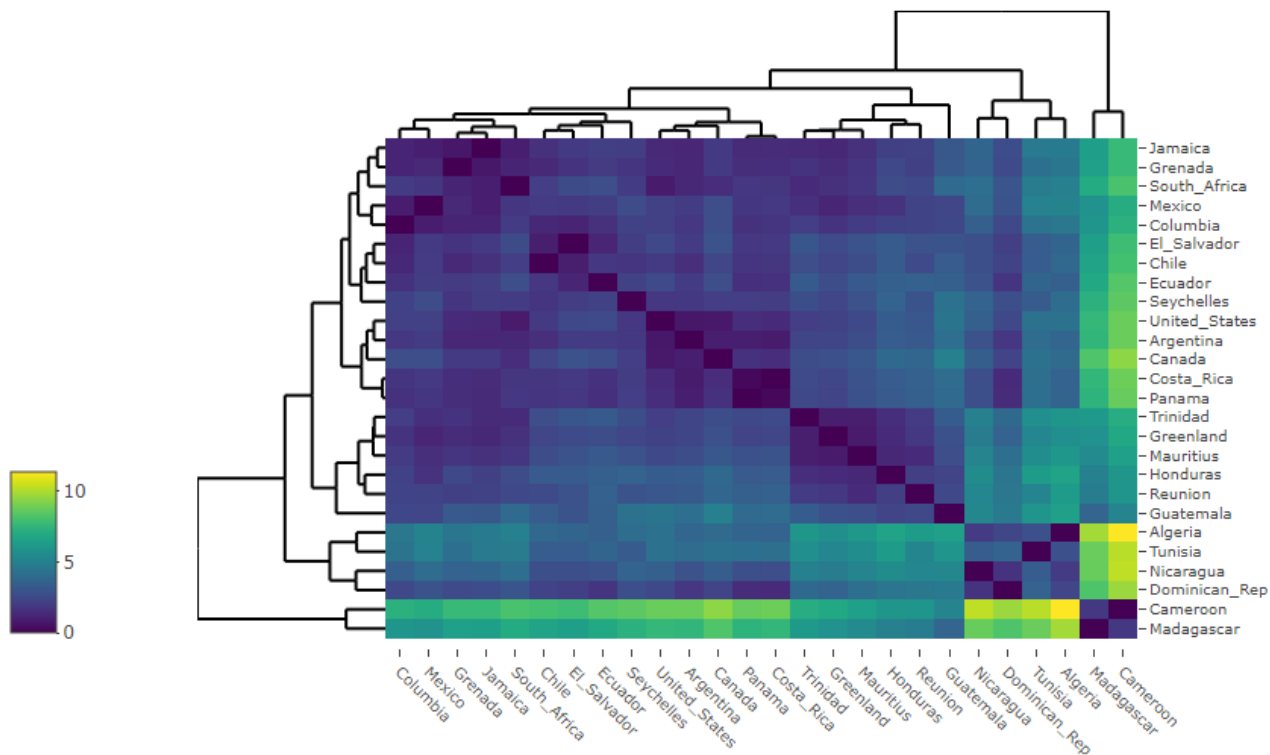
Datos originales



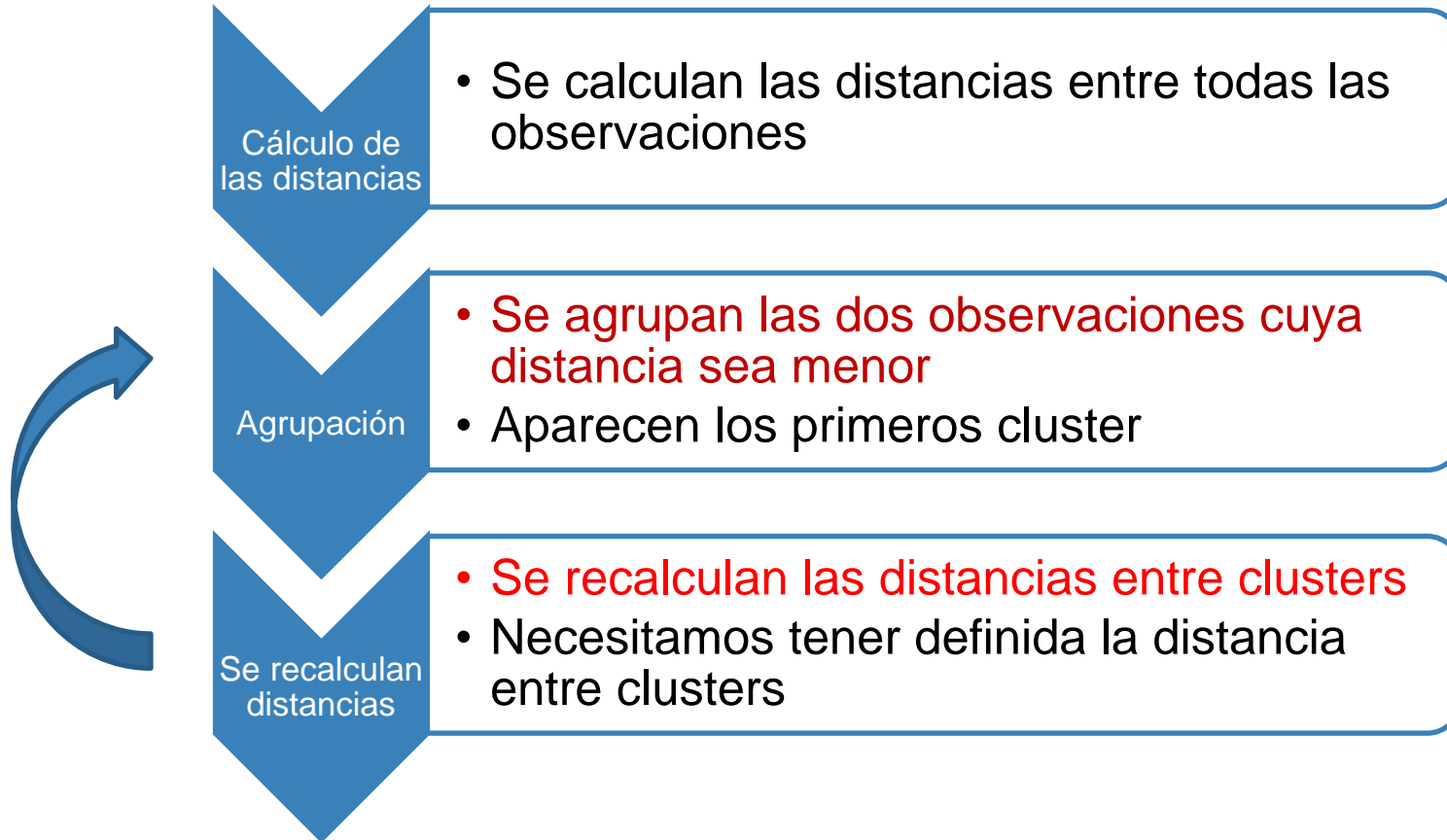
Datos estandarizados



```
#Podemos también calcular el mapa interactivo además incluimos la opción seriate="mean"  
heatmaply(as.matrix(d_st), seriate = "mean", row_dend_left = TRUE, plot_method = "plotly")
```



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

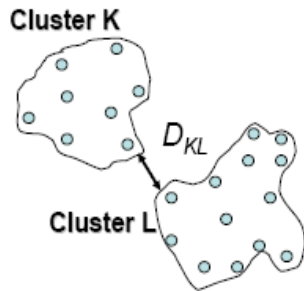


Distancias entre clústeres

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]		

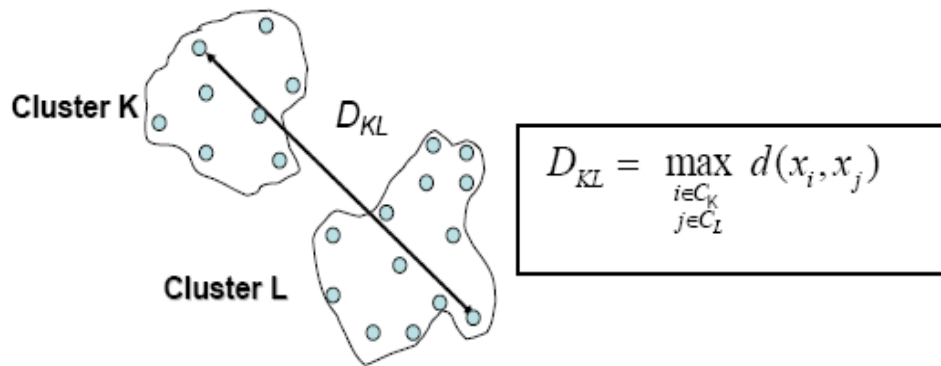
$$C_5 = \{[1,2,3,4,5,6]\}$$

Enlace Completo o del vecino más alejado (complete):

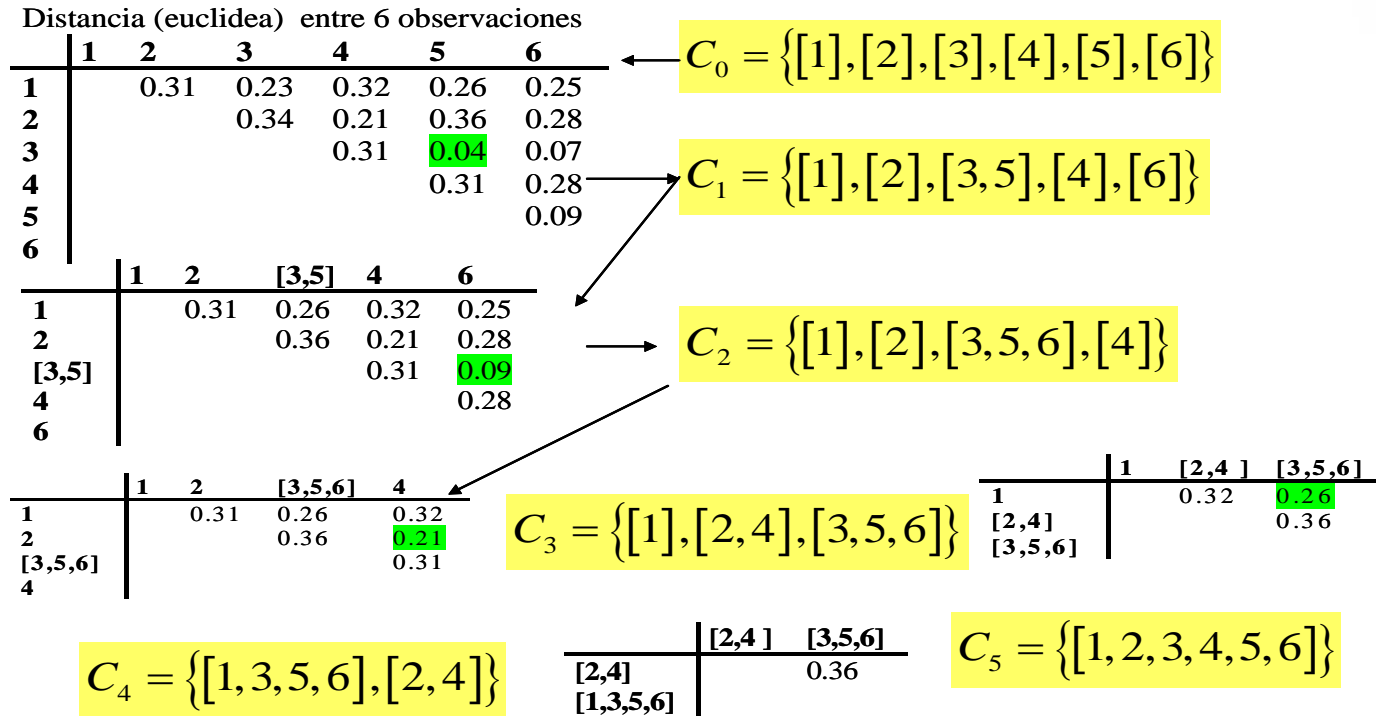
La distancia entre dos clústeres viene dada por la **distancia máxima** entre pares de observaciones cada una perteneciente a uno de los dos clústeres.

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

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

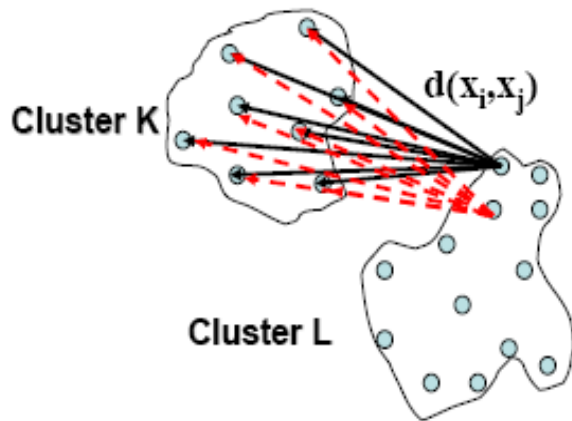


Enlace medio (average):

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

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

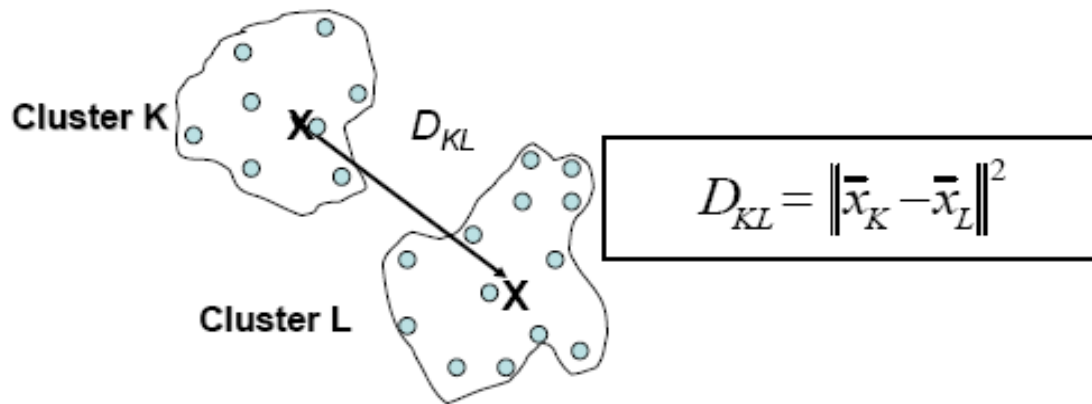
Enlace Medio: Los grupos así formados tienen **varianza similar y pequeña**.



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**):

Este método utiliza también la distancia entre centroides pero con ponderación inversa de los tamaños de los clusters. Este método **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.

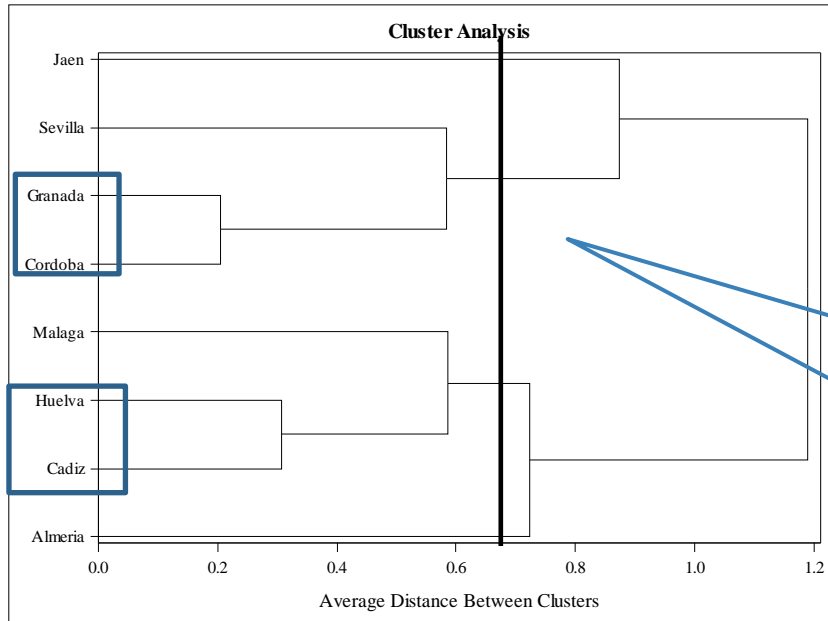
```
#Agrupamos las observaciones según el criterio de ward  
res.hc <- hclust(d, method="ward.D2")
```

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.

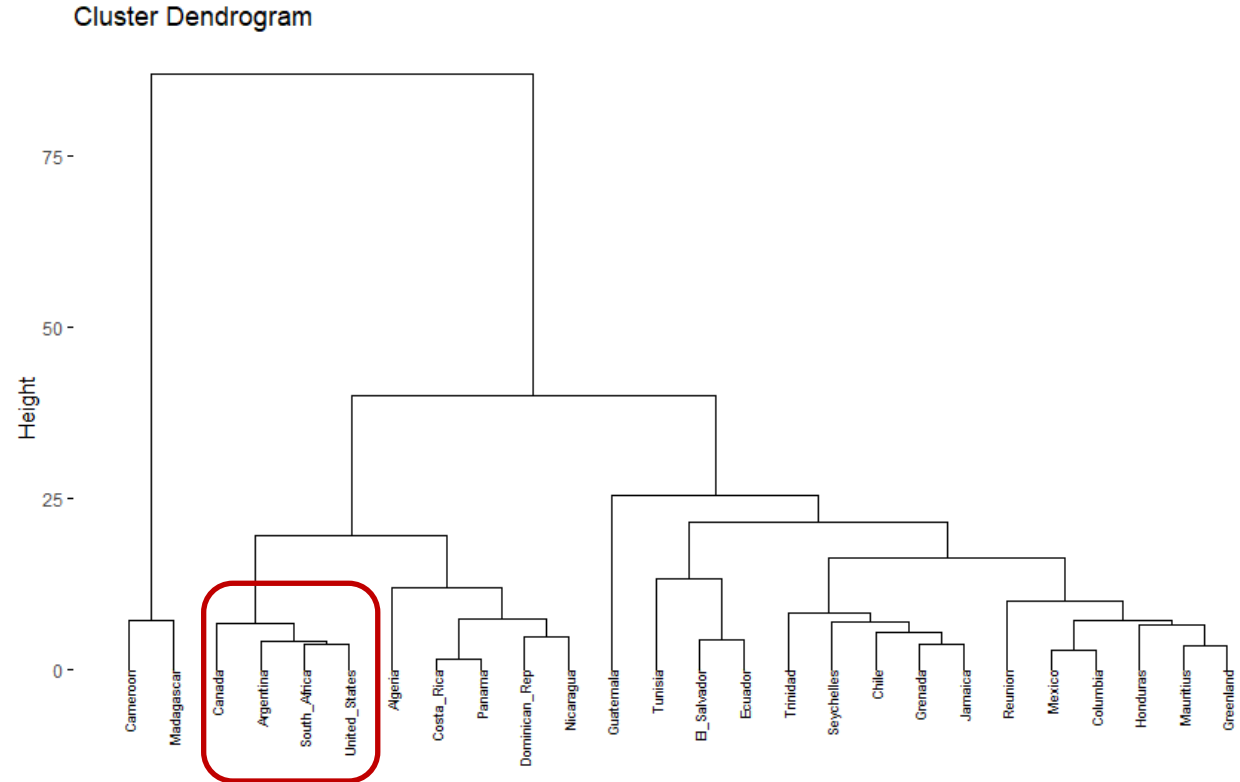
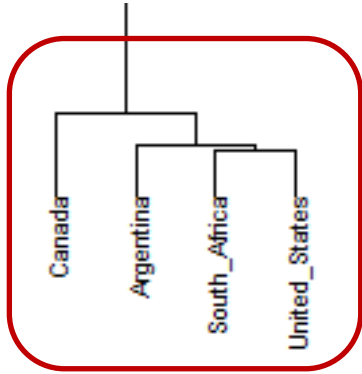


Este diagrama depende de la distancia entre elementos y entre clústeres utilizada, y nos **puede ayudar** a determinar en qué momento del proceso de agrupación nos deberemos detener

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

Dendrograma con los datos sin estandarizar



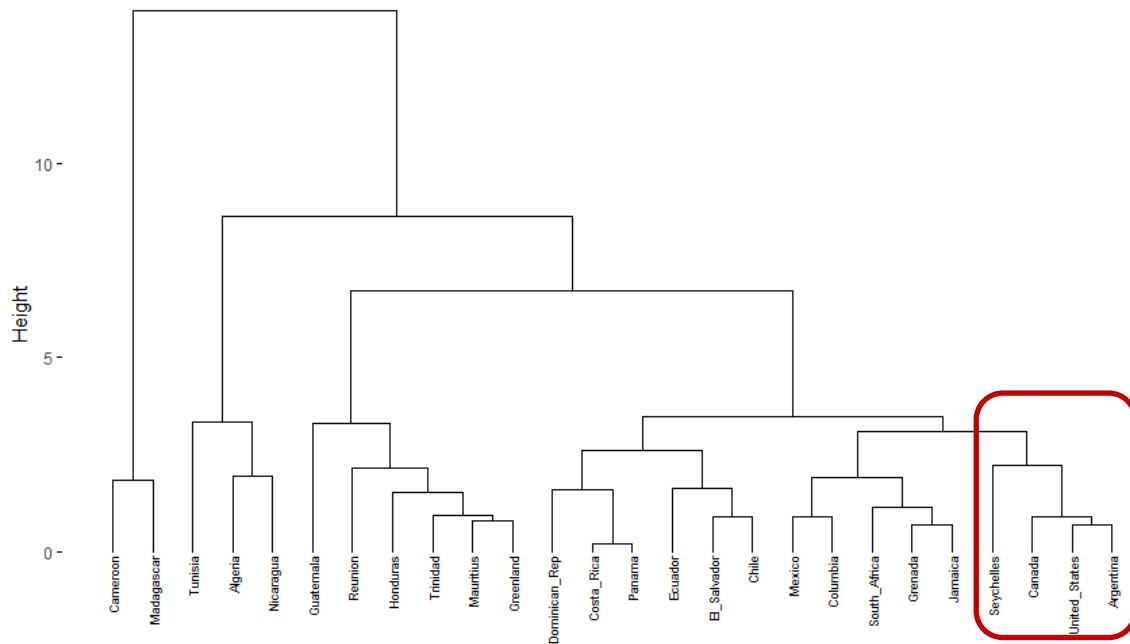
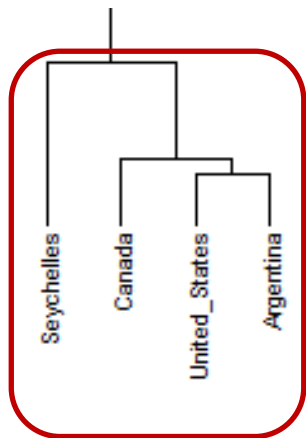
Dendrograma con los datos estandarizados

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

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

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

Cluster Dendrogram



Seleccionamos el número de clusters que nos parece “lógico”

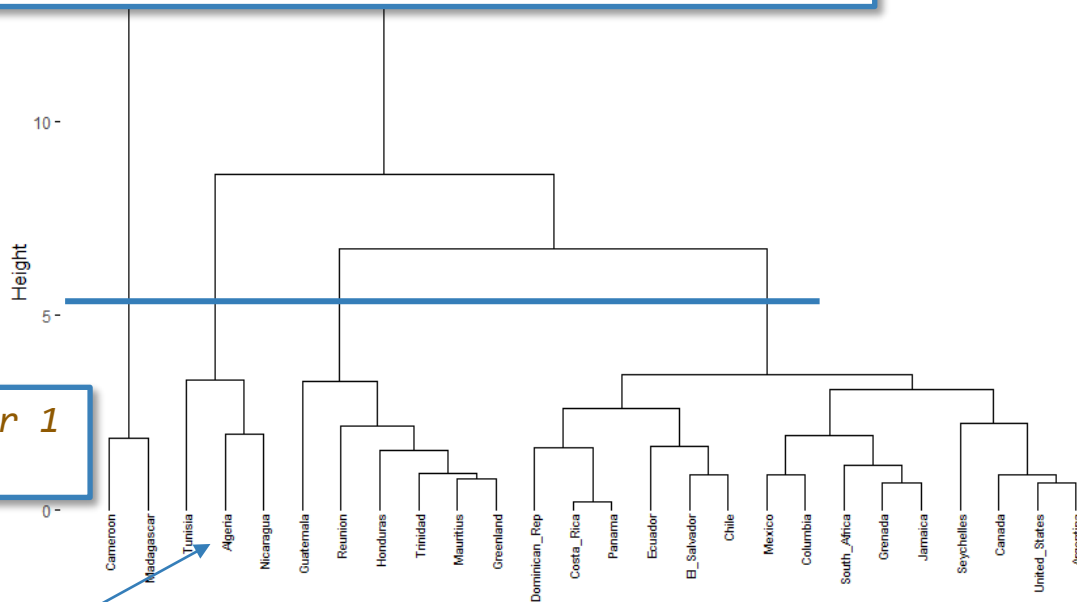
```
# Seleccionamos 4 clusters  
grp <- cutree(res.hc_st, k = 4)
```

```
# Number of members in each cluster  
knitr::kable(table(grp), caption = "Número de individuos por cluster")
```

grp	Freq
1	3
2	2
3	6
4	15

```
# Podemos ver los países del cluster 1  
rownames(dat_EV)[grp == 1]
```

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



Esta función representa el dendrograma con el número de clusters decidido

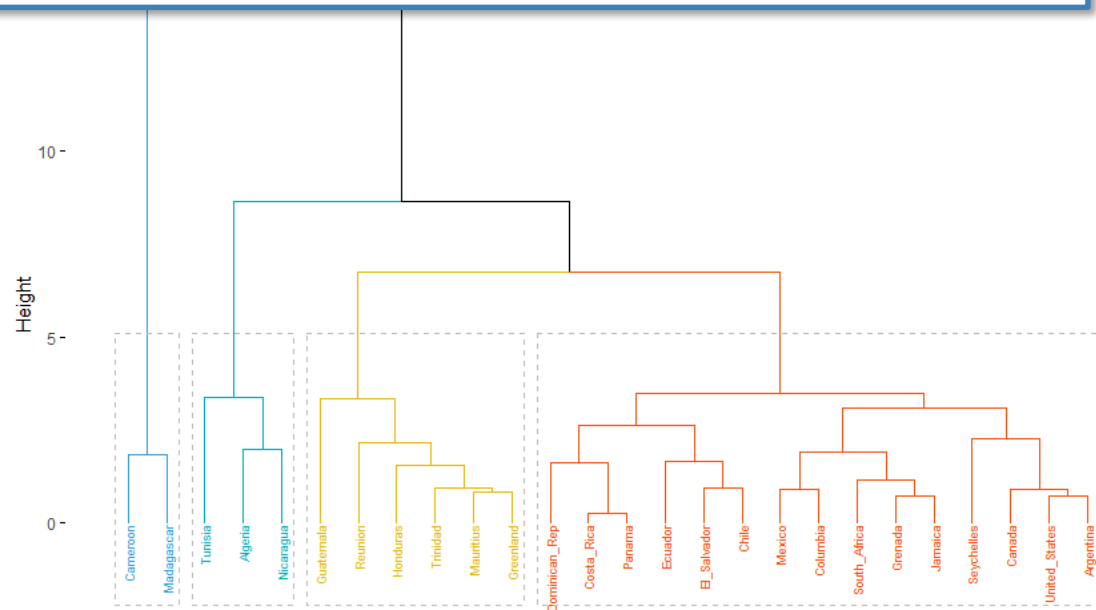
```
fviz_dend(res.hc_st, k = 4, # Cuatro Clusters
```

```
cex = 0.5, #tamaño
```

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

```
color_labels_by_k = TRUE, #Diferentes colores a los clusters
```

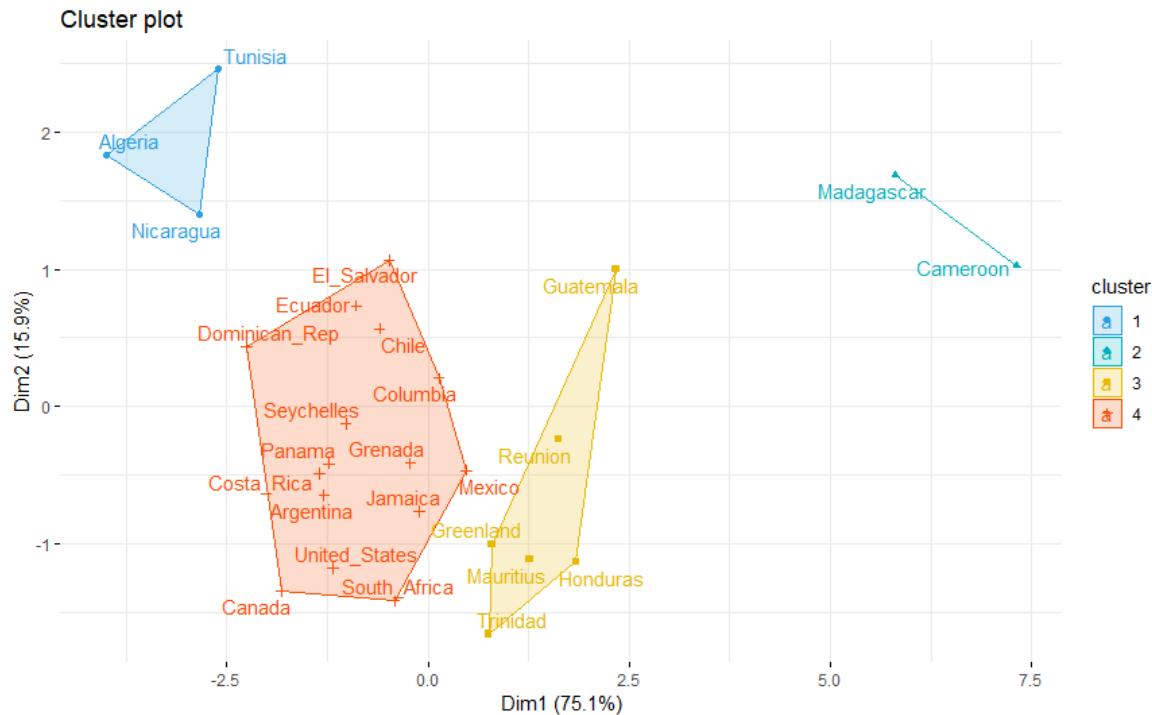
```
rect = TRUE) #añade un rectángulo alrededor
```



```
#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 **la función agnes** que directamente estandariza, calcula las distancias entre individuos y realiza el cluster jerárquico

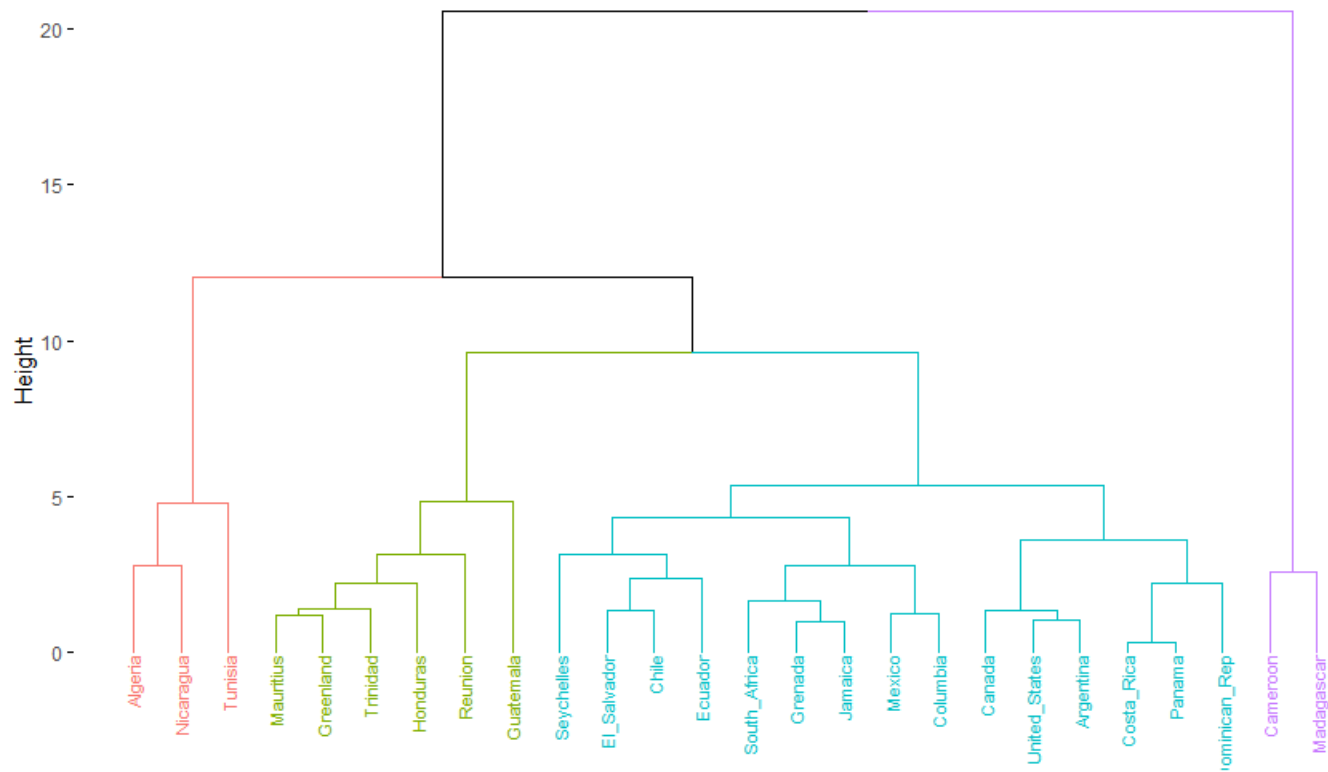
```
# Agglomerative Nesting (Hierarchical Clustering)
res.agnes <- agnes(x = dat_EV, # datos
                  stand = TRUE, # Standardizamos
                  metric = "euclidean", # distancia entre individuos
                  method = "ward") # distancia entre clusters
```



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

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

Cluster Dendrogram



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)
```

Para fijar la generación de valores aleatorios que se utilizarán como semillas y así obtener los mismos agrupamientos

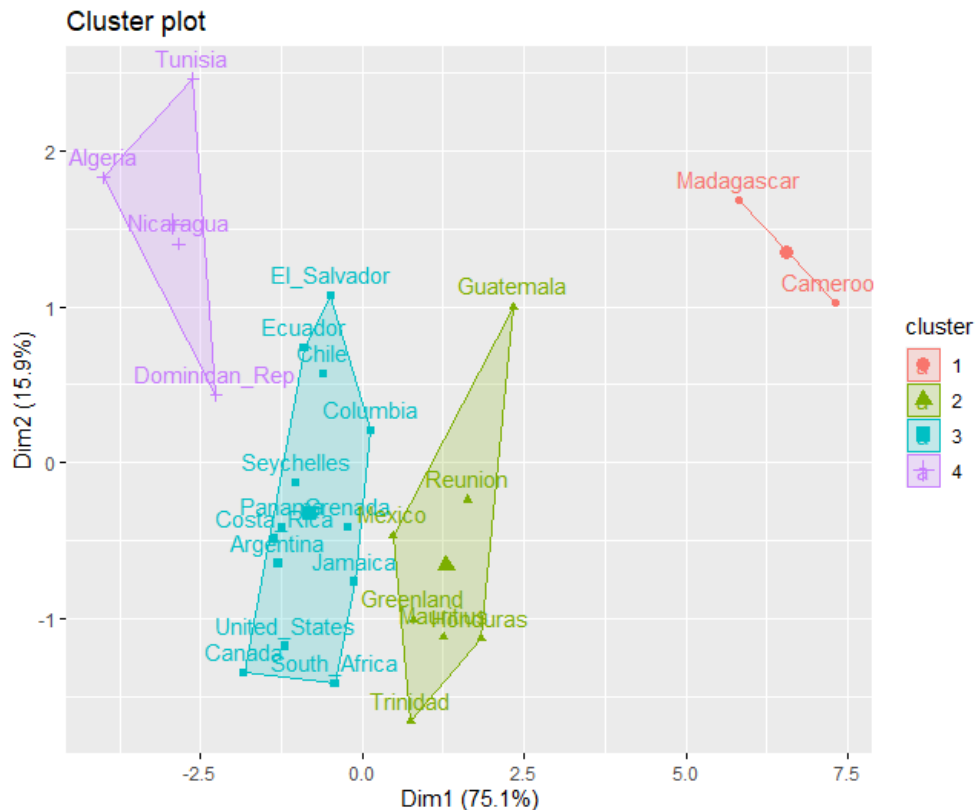
```
#Nos aseguramos que tenemos todos la misma semilla
RNGkind(sample.kind = "Rounding")
```

```
set.seed(1234)
```

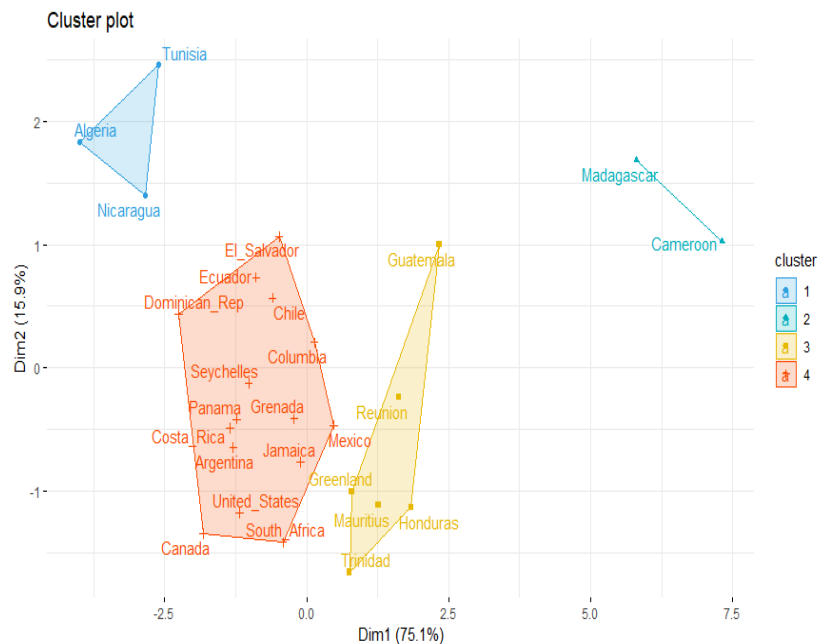
```
RNGkind(sample.kind = "Rejection")
#Para versiones a partir de 3.6
```

```
# Compute k-means
km.res <- kmeans(datos_ST, 4)
```

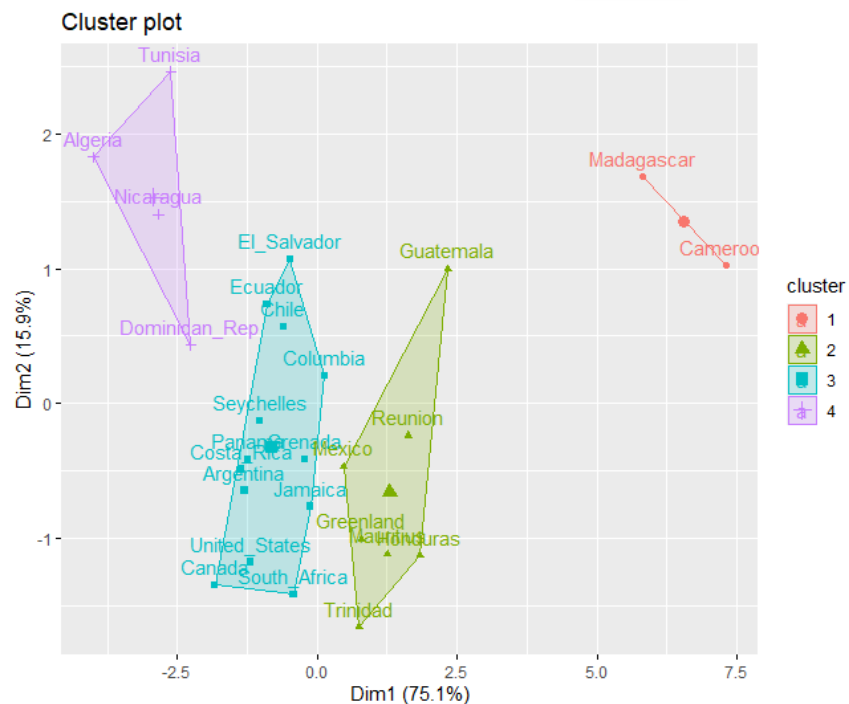
```
fviz_cluster(km.res, datos_ST)
```



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

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

$$W = \sum_k W_k$$

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

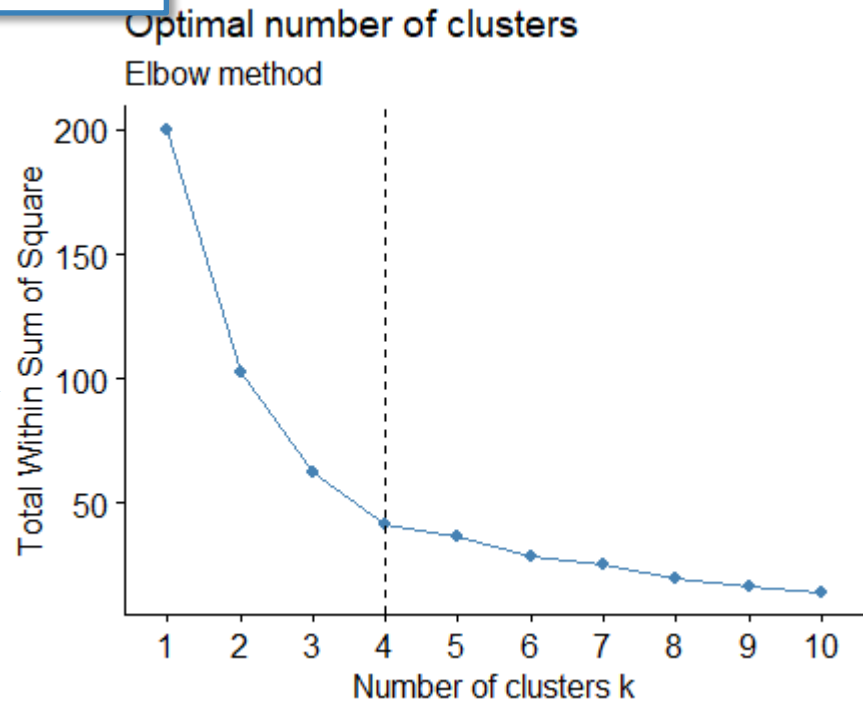
Determinación del número óptimo de clusters
library(NbClust)

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

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



```
# Silhouette method
```

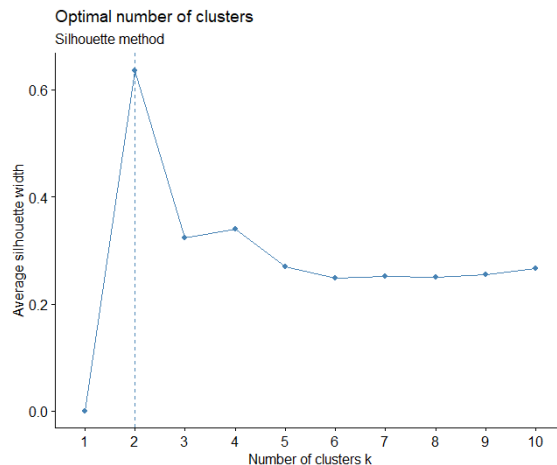
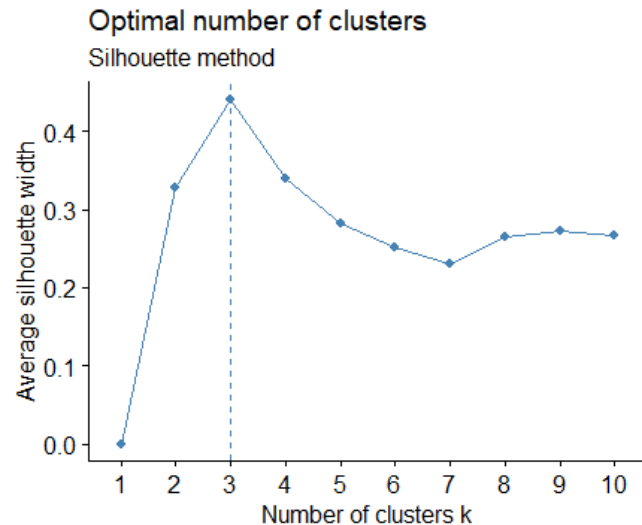
```
fviz_nbclust(datos_ST, kmeans, method = "silhouette")+  
  labs(subtitle = "Silhouette method")
```

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

Cuanto mayor sea su valor mejor

```
RNGkind(sample.kind = "Rejection")  
#Si hemos utilizado esta opción
```

$$\bar{s} = \frac{\sum_{i=1}^n s(i)}{n}$$



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

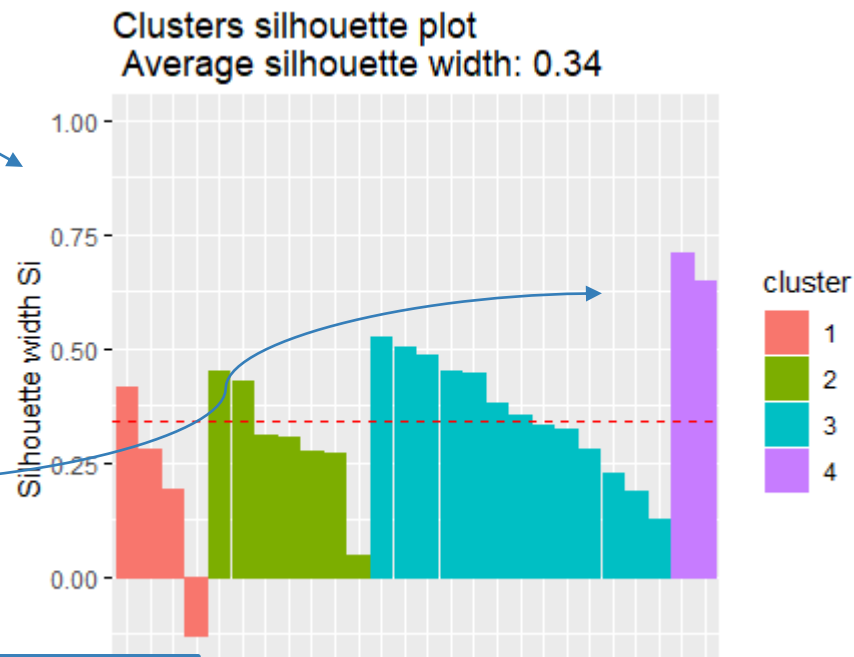
b(i)= distancia media de la observación i-ésima a las observaciones de otros clusters

```
sil <- silhouette(km.res$cluster, dist(datos_ST))  
rownames(sil) <- rownames(dat_EV)  
head(sil[, 1:3])
```

`fviz_silhouette(sil)`

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad -1 \leq s(i) \leq 1$$

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



Valores negativos indican que esa observación no está bien clasificada

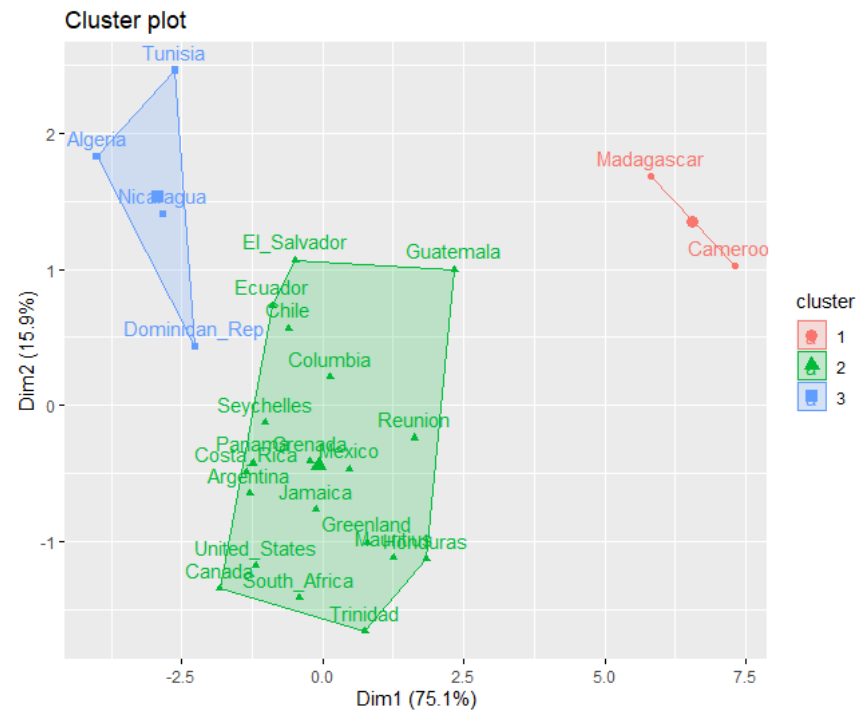
Probamos con 3 Clusters que es lo que nos recomienda el criterio Silhouette

```
RNGkind(sample.kind = "Rounding")
```

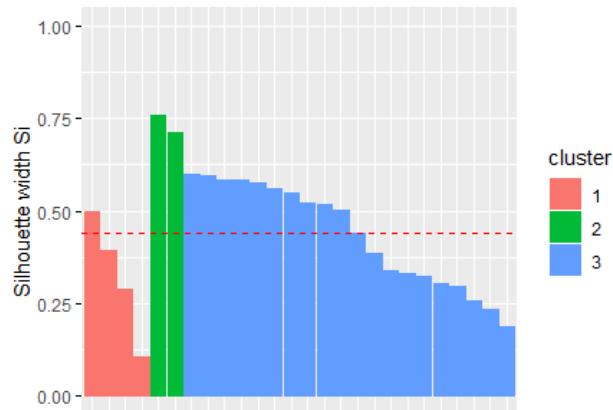
```
set.seed(1234)
```

```
km.res3 <- kmeans(datos_ST, 3)
```

```
fviz_cluster(km.res3, datos_ST)
```



Clusters silhouette plot
Average silhouette width: 0.44



```
sil <- silhouette(km.res3$cluster, dist(datos_ST))  
rownames(sil) <- rownames(datos)  
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.



Mostramos los estadísticos resumen de los individuos de cada cluster

```
knitr::kable(km.res$centers, digits =2,caption = "Estadísticos de los clusters,  
datos STD")
```

Estadísticos de los clusters, datos STD

m0	m25	m50	m75	w0	w25	w50	w75
0.41	0.31	0.26	0.04	0.44	0.40	0.25	0.11
-0.14	-0.36	-0.52	-0.74	-0.17	-0.50	-0.63	-0.70
0.35	1.01	1.19	1.76	0.28	0.96	1.39	1.54
-2.86	-2.81	-2.25	-1.14	-2.85	-2.75	-2.23	-1.34

#Se puede calcular las medias de las variables originales

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

```
knitr::kable(Est_Clus, digits =2,caption = "Estadísticos de los clusters")
```

Estadísticos de los clusters

Group.1	m0	m25	m50	m75	w0	w25	w50	w75
1	62.46	45.23	24.15	8.54	67.46	49.54	27.23	10.54
2	58.00	41.86	21.29	6.86	62.00	44.86	24.14	8.29
3	62.00	48.75	27.50	12.25	66.00	52.50	31.25	14.50
4	36.00	29.50	15.00	6.00	38.00	33.00	18.50	6.50

Pais y cluster

	x
Seychelles	1
South_Africa	1
Canada	1
Costa_Rica	1
El_Salvador	1
Grenada	1
Jamaica	1
Panama	1
United_States	1
Argentina	1
Chile	1
Columbia	1
Ecuador	1
Mauritius	2
Reunion	2
Greenland	2
Guatemala	2
Honduras	2
Mexico	2
Trinidad	2
Algeria	3
Tunisia	3
Dominican_Rep	3
Nicaragua	3
Cameroon	4
Madagascar	4

Group.1	m0	m25	m50	m75	w0	w25	w50	w75
1	62.46	45.23	24.15	8.54	67.46	49.54	27.23	10.54
2	58.00	41.86	21.29	6.86	62.00	44.86	24.14	8.29
3	62.00	48.75	27.50	12.25	66.00	52.50	31.25	14.50
4	36.00	29.50	15.00	6.00	38.00	33.00	18.50	6.50

```
ordenado<-sort(km.res$cluster)
knitr::kable(ordenado, digits =2, caption = "Pais y cluster")
```

El cluster 4 es el que tiene menor esperanza de vida a todas las edades,
El Cluster 3 es el que tiene mayor esperanza de vida con datos muy similares a los del Cluster1

Bibliografía

- ✓ An Introduction of Applied Multivariate Analysis with R. Everitt B, Hothorn T. Ed Wiley. 2011. [Libro completo con explicaciones teóricas y ejemplos resueltos en R aunque con librerías básicas.](#)
- ✓ Nuevos Métodos de Análisis Multivariante. Cuadras C.M. 2014 . [Libro completo con explicaciones teóricas y otras técnicas multivariantes](#)
- ✓ Practical guide to Cluster Analysis in R. A. Kassambara.Ed. STHDA. 2017. [Libro completo que explica las librerías factominer y factoextra mediante ejemplos. Sin explicaciones teóricas](#)
- ✓ Package 'factoextra' . [Explicación del funcionamiento de la librería y la sintaxis detallada](#)
- ✓ Package 'factominer' . [Explicación del funcionamiento de la librería y la sintaxis detallada](#)
- ✓ Getting Things in Order: An Introduction to the R Package seriation. Hahsle. M, Hornik K, Buchta C. Journal of Statistical Software. 2008. [Artículo con explicación detallada de los algoritmos de ordenación utilizados en los heatmap.](#)
- ✓ <http://www.sthda.com/english/>

UNIVERSIDAD
COMPLUTENSE
DE MADRID

