TEMA 4

MÓDULO:
TÉCNICAS AVANZADAS DE DATA MINING

# ANÁLISIS CLUSTER: ALGORITMOS DE CLASIFICACIÓN JERÁRQUICA Y NO JERÁRQUICA

**FERRÁN ARROYO**

Licenciado en Empresariales y en Ciencias Actuariales y Financiaras por la UB. Máster Executive en Data Science por la MBIT School. Data Scientist.


Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

# ÍNDICE

# OBJETIVOS ESPECÍFICOS

- Conocer los principales tipos de clustering.
- Familiarizarse con las técnicas que se utilizan para elegir el número óptimo de clusters.

# 1. INTRODUCCIÓN AL ANÁLISIS CLUSTER

El Análisis Cluster (también conocido como **Análisis de Conglomerados**) consiste en agrupar objetos por similitud, en grupos o conjuntos, de manera que los miembros del mismo grupo tengan características lo más similares posibles y sean lo más distintas posibles a las de los miembros de los otros grupos. Esta es una técnica muy utilizada en el análisis de datos estadísticos debido a su gran versatilidad ante muchos problemas.

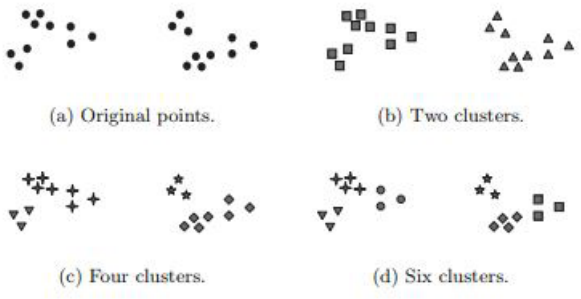
Existen infinidad de algoritmos capaces de resolver el Análisis Cluster, cada uno con sus propias características. Las principales diferencias entre ellos radican en cómo definen lo que es un grupo y la manera de encontrarlo del modo más eficiente posible.

A diferencia de otros problemas, en el Análisis Cluster, no obtendremos una solución directa, ya que es un un proceso iterativo e interactivo que implica ensayo y error. Este proceso de prueba y error es iterativo en la medida que sea automático, e interactivo en la medida que requiera intervención humana. Es una práctica usual ejecutar un algoritmo de clustering (un proceso iterativo), y a partir de los resultados, ajustar determinadas variables o parámetros del algoritmo y repetir la operación (resultando en un proceso interactivo).

El tipo de output que esperamos de este proceso se puede dividir en dos tipos:

- Un conjunto de grupos que constituyen el resultado buscado: por ejemplo, una segmentación de clientes.
- Un conjunto de grupos que se utilizarán como variable explicativa para un modelo de clasificación o regresión posterior: por ejemplo, añadir el cluster de segmentación de clientes dentro de un modelo de propensión a la compra.

En la siguiente imagen, puedes ver el input en (a) y los outputs esperados en (b), (c) y (d):




*Fuente: https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf*

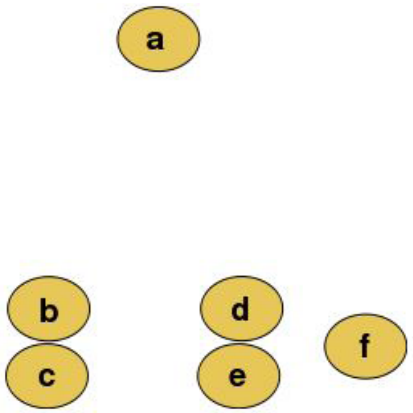
# 2. CONCEPTOS BÁSICOS

## 2.1 TIPOS DE ALGORITMOS

### 2.1.1 CLUSTERING JERÁRQUICO

El Clustering Jerárquico funciona según el principio más simple posible. El punto de datos más cercano al punto base se comportará de manera similar en comparación con un punto de datos que esté más lejos. Vamos a imaginar que tenemos 6 estudiantes que queremos agrupar en clusters y estan categorizados como a, b, c, d, e y f.

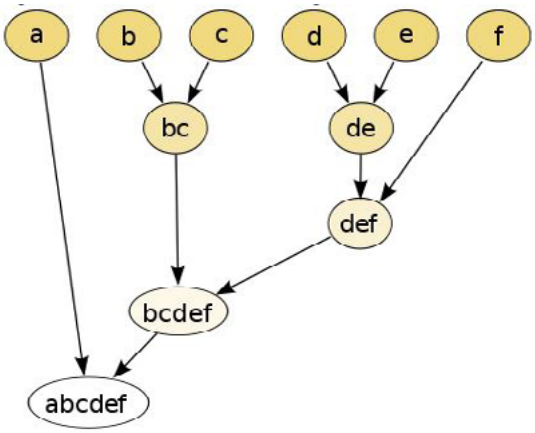
A continuación, puedes visualizar cómo estarían distribuidos en un plano de dos dimensiones:




*Fuente: https://www.analyticsvidhya.com/blog/2013/11/getting-clustering-right/*

Mediante la utilización de este algoritmo, se agruparán de forma secuencial cada uno de los estudiantes en distintos grupos. Nosotros debemos ser los que elijamos dónde fijar el punto de corte (es decir, el número de clusters con el que nos quedamos).

Gráficamente:




*Fuente: https://www.analyticsvidhya.com/blog/2013/11/getting-clustering-right/*

Por lo tanto, en este ejemplo concreto, si quisiésemos 4 clusters, elegiríamos:

- Cluster 1: estudiante a.
- Cluster 2: estudiantes b y c.
- Cluster 3: estudiantes d y e.
- Cluster 4: estudiante f.

Tal y como puedes observar en la imagen, el "estudiante a" está situado mucho más lejos que el resto y podría ser un **outlier.**

El principal problema con esta técnica es que puede manejar un número relativamente pequeño de puntos de datos y es costoso, computacionalmente. Esto se debe a que trata de calcular la distancia entre todas las combinaciones posibles y, luego, toma una decisión para combinar dos grupos/puntos de datos individuales.

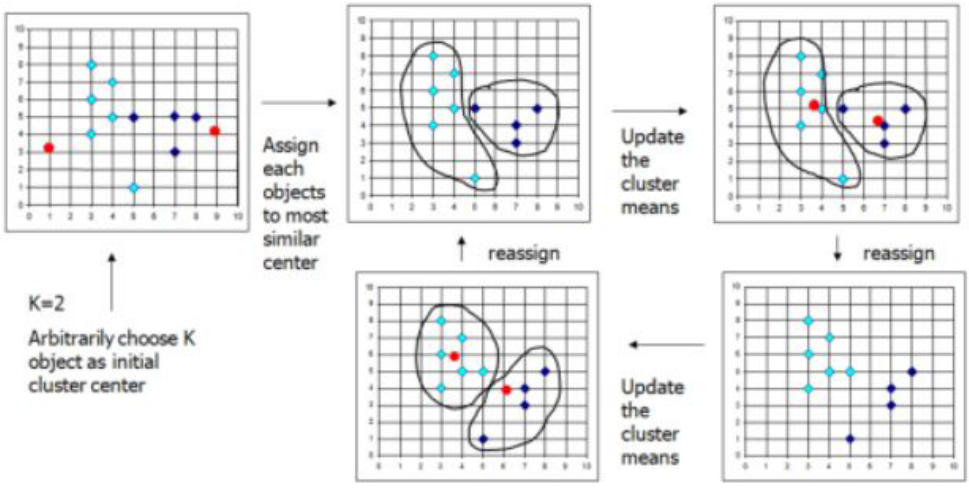
## 2.1.2 CLUSTERING K-MEANS

Este algoritmo es el más común y utilizado y permite trabajar con grandes volúmenes de datos. Básicamente, el algoritmo sigue los siguientes pasos:

- Elección aleatoria del centroide de cada grupo y posicionamiento en el espacio.
- Asignación de los puntos más cercanos a cada centroide. Así se crea el cluster.
- Actualización de los centroides en función de los puntos. El centroide siempre ha de encontrarse en el punto más equidistante a todos los puntos pertenecientes a su cluster.
- Reasignación de los puntos una vez recalculados los centroides.
- Actualización de los clusters.

Tal y como puedes ver, este es un proceso iterativo. El número de iteraciones hay que determinarlas.

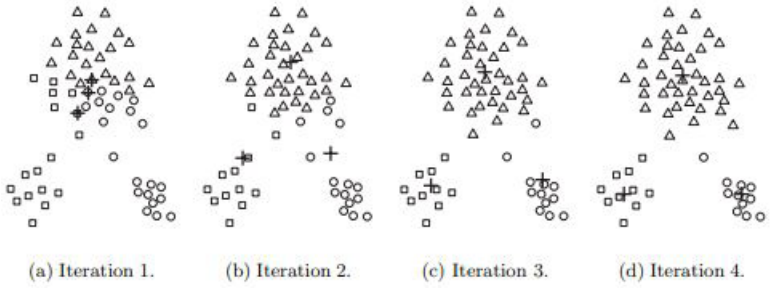
Gráficamente:




*Fuente: https://www.analyticsvidhya.com/blog/2013/11/getting-clustering-right/*

Tal y como puedes ver en la imagen anterior, comenzamos con un número definido de clusters (en este caso k=2). El algoritmo toma dos puntos al azar y mapea todos los demás puntos de datos en base a los dos puntos elegidos. El algoritmo se repite hasta que se minimiza el término de penalización global.

En la siguiente imagen, puedes ver cómo se van desplazando los centroides en cada iteración:




*Fuente: https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf*

Si comparamos las dos técnicas, la principal diferencia es que el Clustering Jerárquico no requiere predefinir un número de clusters mientras que el Clustering K-Means sí.

## 2.2 EVALUACIÓN DEL NÚMERO DE CLUSTERS

La evaluación del número de clusters consiste, básicamente, en determinar cuál es el número óptimo de clusters en función de una determinada función o criterio.

Existen distintos métodos para evaluar el número óptimo de clusters, en este curso veremos:

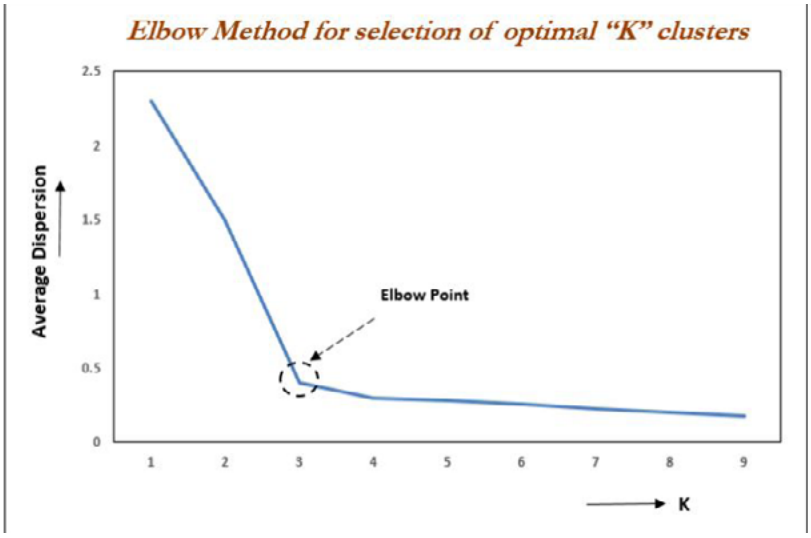
- Método del codo (Elbow Method).
- Dendrograma.

### 2.2.1 ELBOW METHOD

En Análisis de Clusters, el método del codo (también conocido como Elbow Method), consiste en graficar la variación explicada como una función del número de grupos, y elegir el codo de la curva como el número de grupos a utilizar, es decir, aquél punto donde se suavize la pendiente.

**El uso del "codo" o "rodilla de una curva" como punto de corte** es una heurística común en la optimización matemática para elegir un punto en el que los rendimientos decrecientes ya no valen el coste adicional. En el Análisis Clúster, esto significa que **se debe elegir un número de agrupaciones para que, al agregar otra agrupación, no se obtenga un mejor modelado de los datos.**

Gráficamente:

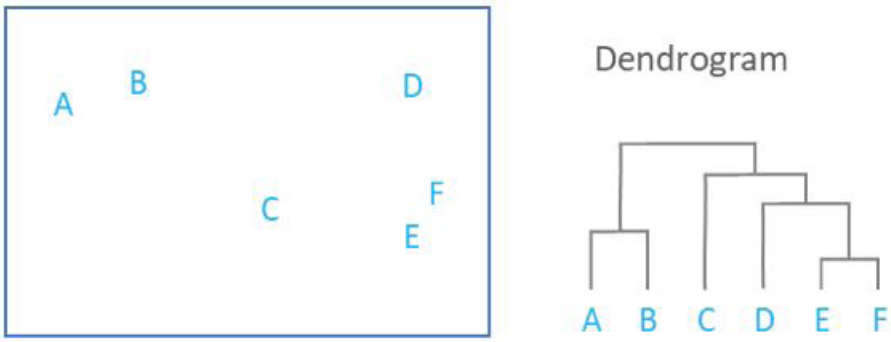



*Fuente: https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/c71ea970-0f3c-4973-8d3a-b09a7a6553c1.xhtml*

## 2.2.2 DENDROGRAMA

Un dendrograma es un diagrama que muestra la relación jerárquica entre los objetos.

Se crea más comúnmente como una salida del Clustering Jerárquico. **El uso principal de un dendrograma es encontrar la mejor manera de asignar objetos a los clusters.** El dendrograma de abajo muestra la agrupación jerárquica de seis observaciones que se muestran en el diagrama de dispersión de la izquierda:



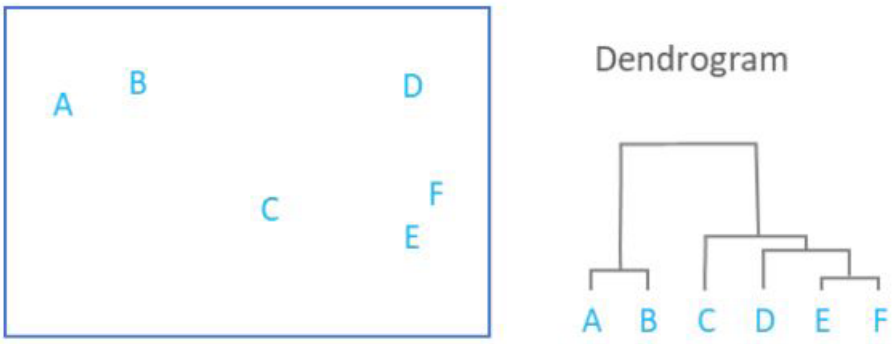

*Fuente: https://www.displayr.com/what-is-dendrogram/*

La clave para interpretar un dendrograma está en centrarse en la altura a la que dos objetos cualesquiera se unen. En el ejemplo anterior, podemos ver que E y F son más similares, ya que la altura del enlace que los une es la más pequeña. Los dos siguientes objetos más similares son A y B.

En el dendrograma de arriba, la altura indica el orden en el que se unieron los clusters.

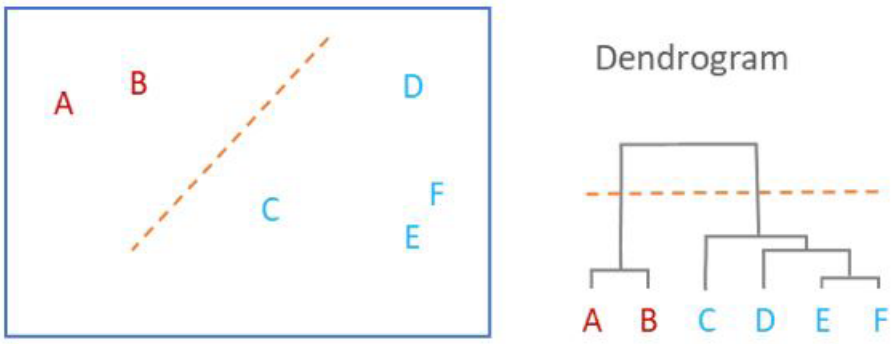
Se puede crear un dendrograma más informativo donde **las alturas reflejan la distancia entre los clusters** como se muestra a continuación.

En este caso, el dendrograma nos muestra que la gran diferencia entre los clusters es entre el cluster de A y B frente al de C, D, E y F:




*Fuente: https://www.displayr.com/what-is-dendrogram/*

Las observaciones son situadas trazando líneas en horizontal a través del dendograma. Por ejemplo, en la imagen de abajo puedes ver cómo la recta separa, perfectamente, A y B del resto de observaciones, creando dos clusters:

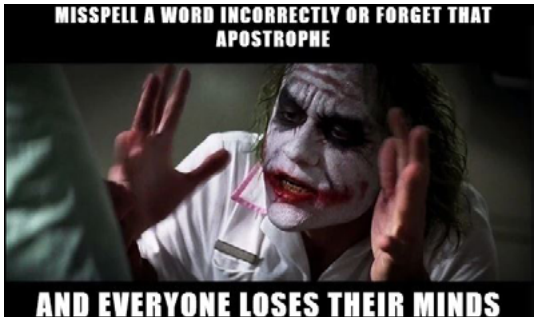



*Fuente: https://www.displayr.com/what-is-dendrogram/*

En general, es un error utilizar los dendrogramas como herramienta para determinar el número de clusters en los datos. Cuando hay un número obviamente "correcto" de clusters, esto será a menudo evidente en un dendrograma. Sin embargo, los dendrogramas, a menudo, sugieren un número correcto de clusters cuando no existen pruebas reales que apoyen la conclusión. **No deja de ser un soporte gráfico y es necesario combinarlo con otros métodos** (como el Elbow o el GAP, por ejemplo).

## SABÍAS QUE...

El dendrograma, a menudo, se escribe mal, llamándose dendograma. Tómalo en consideración si en una charla informal con tus compañeros/as alguno de ellos/as lo menciona sin la primera r:




*Fuente: https://imgur.com/gallery/VUbJKhe*

# 3. ACTIVIDAD GUIADA

El ejercicio propuesto consiste en una tarea de clustering sobre un conjunto de datos públicos, el famoso dataset de USArrests.

Este conjunto de datos contiene estadísticas, en detenciones por cada 100.000 residentes por asalto, asesinato y violación en cada uno de los 50 estados de Estados Unidos en el año 1973. También se da el porcentaje de la población que vive en zonas urbanas.

Podrás encontrar más detalles acerca del conjunto de datos en el siguiente enlace: https://www.kaggle.com/deepakg/usarrests

## 3.1 CLUSTERING K-MEANS

Lo primero que haremos será visualizar nuestro conjunto de datos:

```
# Carga del conjunto de datos
data("USArrests")

# Visualizamos algunas observaciones
head(USArrests,5)
```

| | Murder <dbl> | Assault <int> | UrbanPop <int> | Rape <dbl> |
|---|---|---|---|---|
| Alabama | 13.2 | 236 | 58 | 21.2 |
| Alaska | 10.0 | 263 | 48 | 44.5 |
| Arizona | 8.1 | 294 | 80 | 31.0 |
| Arkansas | 8.8 | 190 | 50 | 19.5 |
| California | 9.0 | 276 | 91 | 40.6 |

Tal y como puedes comprobar, la variable UrbanPop hace referencia al porcentaje de población urbana. Así, California tiene, casi su totalidad, población categorizada como población urbana mientras que Arkansas, sólo la mitad.

Por otra parte, las variables Murder, Assault y Rape, hacen referencia al número de asesinatos, asaltos y secuestros por cada 100.000 residentes.

Si visualizamos la distribución de las variables, podrás comprobar que los rangos son muy distintos entre ellas, por lo que es necesario escalarlas.

```
# Sumarización del Dataset
summary(USArrests)
```

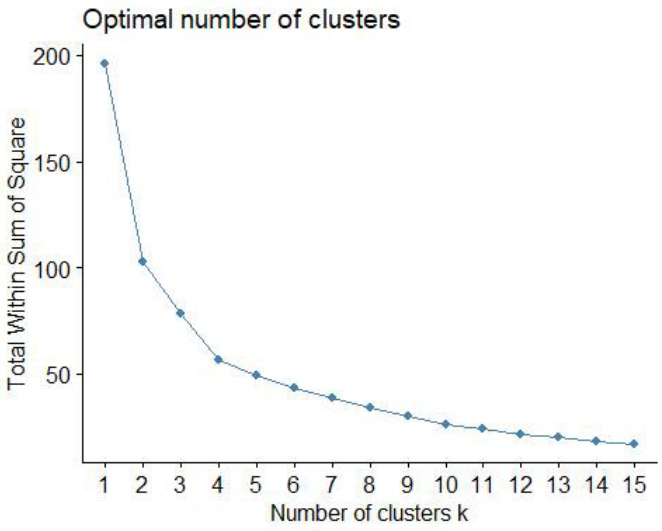
```
     Murder          Assault         UrbanPop          Rape       
 Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30  
 1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07  
 Median : 7.250   Median :159.0   Median :66.00   Median :20.10  
 Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23  
 3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18  
 Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00  
```

```r
# Reescalamoos los datos
datos <- scale(USArrests)
```

Una forma sencilla de estimar el número K óptimo de clusters, cuando no se dispone de información adicional en la que basarse, consiste en aplicar el algoritmo de K-means para un rango de valores de K e identificar aquel valor a partir del cual la reducción en la suma total de varianza intra-cluster deja de ser sustancial. A esta estrategia se la conoce como **método del codo o elbow method.**

La función **fviz_nbclust()** automatiza este proceso y genera una representación de los resultados:

```r
# Visualización del elbow method
fviz_nbclust(x = datos, FUNcluster = kmeans, method = "wss", k.max = 15,
             diss = get_dist(datos, method = "euclidean"), nstart = 50)
```

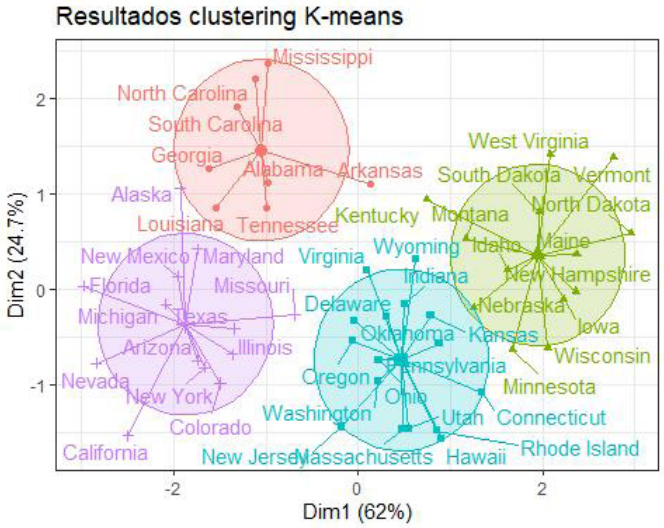



En este análisis, a partir de cuatro clusters, la reducción en la suma total de cuadrados internos parece estabilizarse, indicando que K = 4. Parece una buena opción.

El paquete **factoextra,** también, permite obtener visualizaciones de las agrupaciones resultantes. Si el número de variables (dimensionalidad) es mayor de dos, automáticamente, realiza un PCA y representa los dos primeros componentes principales:

```r
set.seed(123)
km_clusters <- kmeans(x = datos, centers = 4, nstart = 50)

# Las funciones del paquete factoextra emplean el nombre de las filas del
# dataframe que contiene los datos como identificador de las observaciones.
# Esto permite añadir labels a los gráficos.
fviz_cluster(object = km_clusters, data = datos, show.clust.cent = TRUE,
             ellipse.type = "euclid", star.plot = TRUE, repel = TRUE) +
  labs(title = "Resultados clustering K-means") +
  theme_bw() +
  theme(legend.position = "none")
```




Tal y como puedes ver en el gráfico, los dos componentes principales utilizados representan más del 85% de la varianza.

Mediante la información de los clusters, podríamos intuir qué estados de la costa oeste y adyacentes parecen tener datos de criminalidad significativamente distintos con los estados situados más al norte y al centro.

Vamos a echar un vistazo:

```r
# Selecciono los estados que quiero ver
estados<-c("California","Nevada","North Dakota","Minnesota")

# Subset por estado
subset(USArrests, rownames(USArrests) %in% estados)
```

| | Murder <dbl> | Assault <int> | UrbanPop <int> | Rape <dbl> |
|---|---|---|---|---|
| California | 9.0 | 276 | 91 | 40.6 |
| Minnesota | 2.7 | 72 | 66 | 14.9 |
| Nevada | 12.2 | 252 | 81 | 46.0 |
| North Dakota | 0.8 | 45 | 44 | 7.3 |

Efectivamente, California y Nevada tienen más casos de asesinatos, asaltos y secuestros por cada 100K habitantes con un porcentaje de población urbana mucho más alta que Minnesota o Dakota del Norte.

## 3.2 CLUSTERING JERÁRQUICO

Tal y como habrás podido ver en el ejercicio anterior, el algoritmo del K-Means arroja resultados entendibles y segmenta bastante bien, es por eso que es uno de los métodos más utilizados. Sin embargo, tiene la limitación de necesitar que se especifique el número de clusters de antemano y de que sus resultados puedan variar en función de la iniciación aleatoria. Una forma de contrarrestar estos dos problemas consiste en combinar el K-means con el Clustering Jerárquico.

Los pasos a seguir son:

- Aplicar el Clustering Jerárquico a los datos y cortar el árbol en k clusters. El número óptimo puede elegirse, de forma visual, en el dendrograma o con cualquier otro método.
- Calcular el centro (por ejemplo, la media) de cada cluster.
- Aplicar el algoritmo K-Means empleando como centroides iniciales los centros calculados en el paso anterior.

El algoritmo del K-means tratará de mejorar la agrupación hecha por el Clustering Jerárquico del paso uno. Las agrupaciones finales puedan variar respecto a las iniciales:

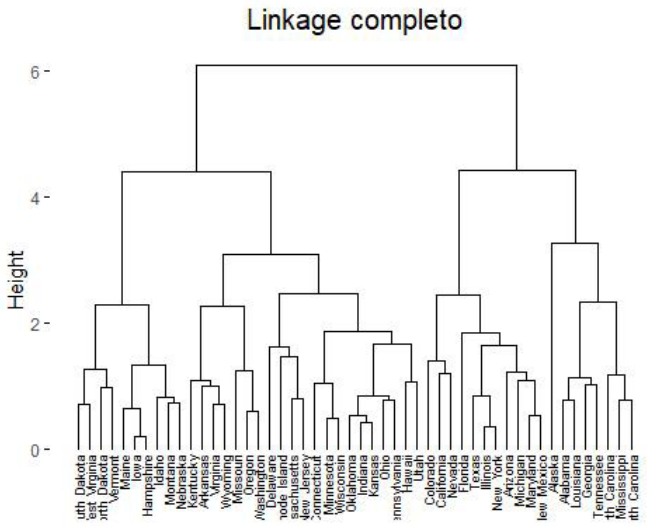
```r
# Carga del conjunto de datos
data("USArrests")

# Reescalamos los datos
datos <- scale(USArrests)
```

La función **hkmeans()** del paquete **factoextra** permite aplicar el método hierarchical K-means clustering de forma muy similar a la función estándar **kmeans():**

```r
library(factoextra)
# Se obtiene el dendrograma de Clustering Jerárquico para elegir el número de clusters.
set.seed(101)
hc_euclidea_completo <- hclust(d = dist(x = datos, method = "euclidean"),
                               method = "complete")

fviz_dend(x = hc_euclidea_completo, cex = 0.5, main = "Linkage completo",
          sub = "Distancia euclídea") +
  theme(plot.title =  element_text(hjust = 0.5, size = 15))
```




Empleando la representación del dendrograma consideraríamos que existen 4 grupos.

# IDEAS CLAVE

- En el Análisis Cluster, no obtendremos una solución directa a nuestro problema.
- El Análisis Cluster consiste, básicamente, en realizar agrupaciones en las que los miembros del mismo grupo deben tener características lo más similares posibles y deben ser lo más distintas posibles a los miembros de los otros grupos.
- No es necesario definir el número de clusters en el Clustering Jerárquico mientras que sí lo es en el K-Means.
- Existen distintos métodos para evaluar el número óptimo de clusters, tales como el Elbow Method o el Dendrograma.
- El punto óptimo para el Elbow Method es aquél donde se suaviza la pendiente.
- Las alturas en el árbol reflejan las distancias entre los clusters en el método del Dendrograma.