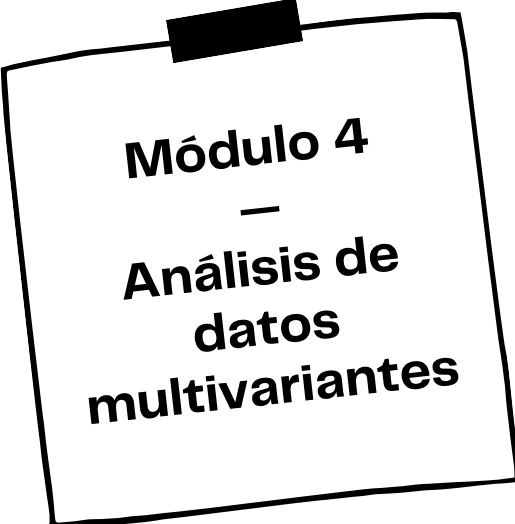


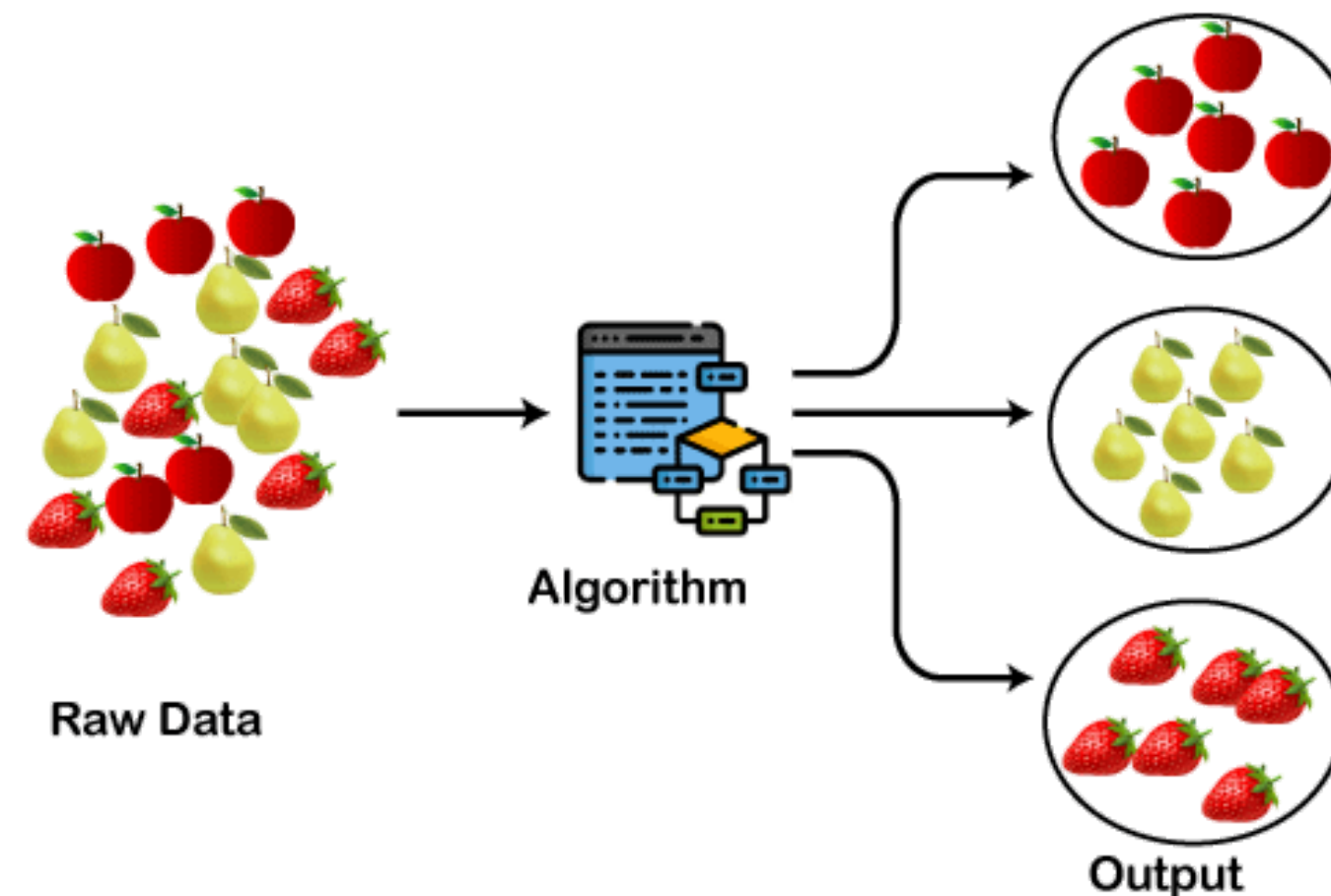
Análisis de conglomerados



Módulo 4
—
**Análisis de
datos
multivariantes**

Análisis de conglomerados

El análisis de conglomerados o clusters tiene como objetivo principal agrupar elementos en grupos homogéneos en función de las similitudes entre ellos



Análisis de conglomerados

Tenemos datos que sospechamos que son heterogéneos y deseamos dividirlos en un número de grupos, de forma que:

1. Cada elemento pertenece a un solo grupo
2. Todos los elementos quedan clasificados
3. Cada grupo es internamente homogéneo

Algoritmo de k-medias

Supongamos que tenemos una muestra de n elementos y p variables. El objetivo es dividir esta muestra en un número de grupos prefijado k . El algoritmo de k-medias consta de 4 etapas:

Algoritmo de k-medias

1. Seleccionar k puntos que serán los centros de los grupos iniciales. Esto puede hacerse de tres formas:
 - Seleccionando aleatoriamente k puntos que serán los centros de los grupos.
 - Tomando como centros los k puntos más alejados entre sí.
 - Seleccionando los centros de cada grupo con información apriori.

Algoritmo de k-medias

2. Calcular las distancias euclidianas de cada elemento al centro de los k grupos y asignar cada elemento al grupo más cercano.


La asignación de los elementos se hace de forma secuencial y cada vez que se asigna un elemento a un grupo se recalculan las coordenadas de la nueva media del grupo

Algoritmo de k-medias

3. Definir un criterio de homogeneidad y comprobar si asignando de nuevo un elemento a otro grupo mejora este criterio
4. Continuar con el proceso hasta que no se pueda mejorar el criterio de homogeneidad

Criterio de homogeneidad

El criterio de homogeneidad más utilizado consiste en minimizar las distancias al cuadrado entre los centros de los grupos y los puntos que pertenecen a ese grupo

$$\min \sum_{g=1}^G \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)' (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g) = \sum_{g=1}^G \sum_{i=1}^{n_g} d^2(i, g)$$


Cuadrado de la distancia euclidiana entre el elemento **i** del grupo **k** y su media de grupo

El algoritmo de k-medias es un proceso iterativo en el que en cada iteración solo se permite mover un elemento de un grupo a otro:

1. Se parte de una asignación inicial
2. Comprobar si moviendo algún elemento se reduce **W**
3. Si se reduce **W**, mover el elemento, recalcular las medias de los dos grupos afectados y volver a (2). Cuando ya no sea posible reducir **W** terminar el proceso o se alcance el máximo número de iteraciones, termina el proceso

<https://www.tidymodels.org/learn/statistics/k-means/>

Practiquemos...

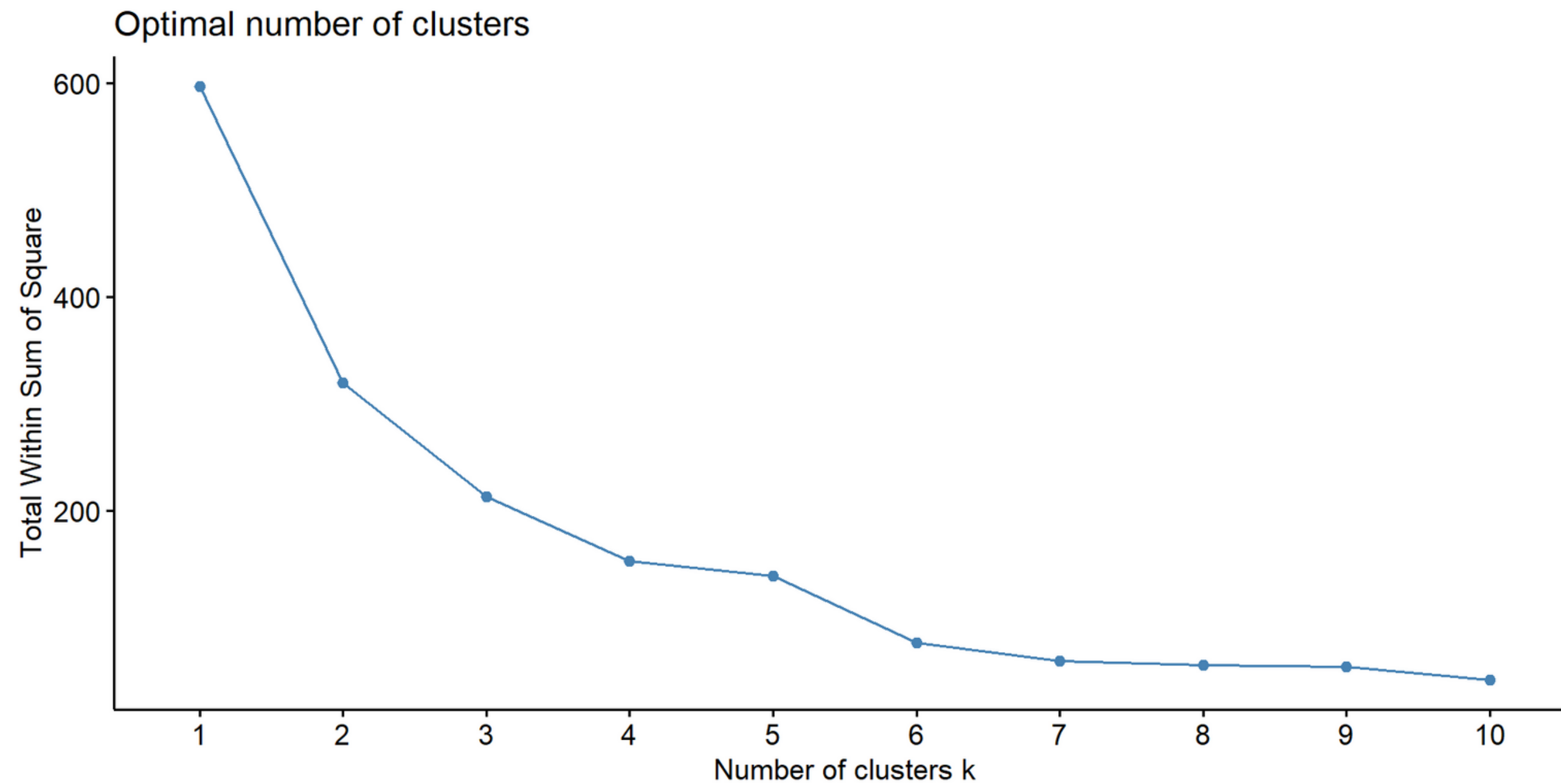
Apliquemos el algoritmo de k-medias a una base de datos que contiene información de clientes de un Mall referente a su salario anual (miles de dólares) y a un puntaje de gastos que indica cuánto ha gastado el cliente en el Mall (entre mayor sea el puntaje, mayor ha sido el gasto del cliente)

Número óptimo de clusters

Método del codo

1. Implementar el algoritmo de k-medias usando diferentes valores de **k** (variar el número de clusters de 1 a 10).
2. Calcular y graficar W para cada valor de **k**.
3. La ubicación de una curva o codo se considera un indicador el número apropiado de clusters o conglomerados.

Número óptimo de clusters



Número óptimo de clusters

