Herramientas computacionales: el arte de la analítica

TC1002S.222

Fecha de entrega: **11 de mayo del 2022**

Equipo 4

Actividad : **“Patrones con K-means”**

Profesor: **Sergio Ruíz Loza**

Nombres	Apellidos	Matrícula
Ana Lucía	Ahedo Reyes	A01661890
Andrea Samantha	Aguilar Ramírez	A01656200
Luis Alfonso	Pérez Mijangos	A01653848
Guillermo	González Jiménez	A01653930
Alfonso	Pineda Cedillo	A01660394

1. **¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?**

   Si, ya que nos ayudan a visualizar qué tan similares son los grupos de datos entre sí a partir del valor de K, así como nos pueden ayudar a especificar entre que variables se busca el agrupamiento.

2. **¿Cómo obtuviste el valor de k a usar?**

   El valor que usamos de K es un valor aproximado, obtenido de la gráfica "Elbow curve", la cual, en este caso, compara el score con el número de clusters. Para hallar dicho valor, tenemos que encontrar el "punto de codo", que es el punto donde la gráfica comienza a "aplanarse", o a tornarse horizontal. Para este caso particular, el valor de K que asignamos fue de 3.

3. **¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo?**

   Si bien el valor que se permite usar para la K puede ser aproximado el valor que se obtiene mediante la regla de 'Elbow curve', es el valor que nos permite observar las relaciones entre los grupos, con menor error de aproximación por parte del usuario.

   Hasta cierto punto sí serían más representativos, porque serían más específicos, sin embargo, esto tiene un límite, el cual si se excede, se trata de hacer tan específico, que se pierde el objetivo inicial de clasificación, debido a la cantidad de datos.

4. **¿Qué distancia tienen los centros entre sí? ¿Hay alguno que esté muy cercano a otros?**

Se puede observar que mediante los centroides a una distancia menor que significa mayor relación entre las variables de 'Average Price' y 'Total Volume' a diferencia de 'Total Bags'.

5. **¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?**

   Significa mayor distancia entre los centros por lo que se tendría menor relación dentro del agrupamiento de las variables, provocando que se grafiquen las variables de una manera más independiente y obteniendo este resultado el método de K-means no nos sería de uso para el análisis del dataset y el agrupamiento.

6. **¿Qué puedes decir de los datos basándose en los centros?**

   Al analizar el gráfico de K-means, se puede observar que hay dos centros/grupos (representados con el color azul y rojo), cuyos elementos guardan una relación estrecha en términos de volumen total y en bajo precio promedio, mientras que los elementos del tercer grupo (representado con el color verde) presentan una relación mínima con el total de bolsas y el precio promedio de los elementos del grupo azul.