Los coches del jefe, reparto

Isabel Afán de Ribera

02 diciembre 2020

Introducción

Tras haber realizado un análisis exploratorio sobre las características de los coches y haber seleccionado las variables más relevantes procedemos a realizar un análisis cluster.

Objetivo

El objetivo de nuestro análisis es estudiar el número adecuado de grupos en los que dividir la colección de 125 todo terreno.

Descripción del dataset

El dataset final cuenta con 125 observaciones y 9 variables que se especifican a continuación:

- cc: Cilindrada (cm cúbicos)
- potencia: Potencia (CV)
- rpm: Revoluciones por minuto
- peso: Peso en kg *plazas: Número de plazas* cons120: Consumo 120 km/h *consurb: Consumo urbano* velocida: Velocidad máxima
- acel2: Tiempo de aceleración, 2 grupos 1(Menor a 10 seg) 2(Mayor a 10 seg)

Desarrollo: análisis cluster

Para proceder con el análisis cluster, y tras la exploración de los datos con los que vamos a trabajar, es fundamental contar con datos homogéneos en cuanto a su unidad de medida por ello el primer paso es escalar las variables.

Evaluación de la bondad del análisis cluster

Estadístico Hopkins

Antes de aplicar un método de clustering a los datos es conveniente evaluar si hay indicios de que realmente existe algún tipo de agrupación en ellos. Un método es el estadístico de Hopkins que permite evaluar la tendencia de clustering de un conjunto de datos mediante el cálculo de la probabilidad de que dichos datos procedan de una distribución uniforme, es decir, estudia la distribución espacial aleatoria de las observaciones.

Valores de estadístico en torno a 0.5 indican que los datos estudiados se distribuyen uniformemente (hipotesis nula) y que por lo tanto no tiene sentido aplicar clustering. Cuanto más se aproxime a 0 más evidencias se tienen a favor de que existen agrupaciones en los datos y si debe aplicarse clustering (hipótesis alternativa).

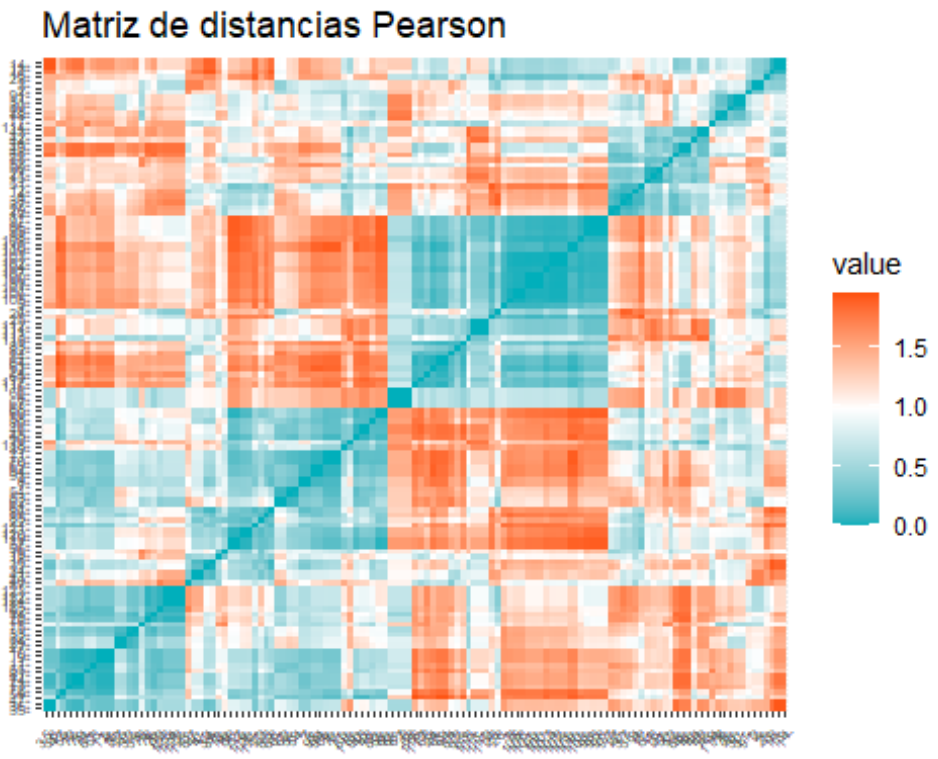
En nuestro caso se ha obtenido un estadístico de Hopkins de 0.1919, por tanto, existen agrupaciones en los datos y podemos realizar la técnica clustering.

Distancia Pearson

Otra forma de inspeccionar la posibilidad de agrupamiento de los datos analizados es mediante la visualización de similitudes entre los elementos.

Para ello pasamos a calcular las distancias mediante la *matriz de distancias* calculada a través de las correlaciones y de las distancias euclídeas de las observaciones.

Con respecto a las distancias calculadas mediante las correlaciones utilizamos el método Pearson para medir la similitud. Y lo representamos gráficamente con un *heatmap*, el cual nos permite visualizar las distintas correlaciones entre las variables en una escala de color del 0 al 1.5 donde el 0 significa que no hay distancia entre las variables y, por tanto, se da la máxima correlación.



Como puede observarse en el gráfico de distancias de Pearson existe posibilidad de formación de grupos, los cuales pueden identificarse mediante el color azul y rojo, siendo los azules los grupos con observaciones que menor distancia presentan y, por tanto, los más similares.

Identificación del número óptimo de grupos.

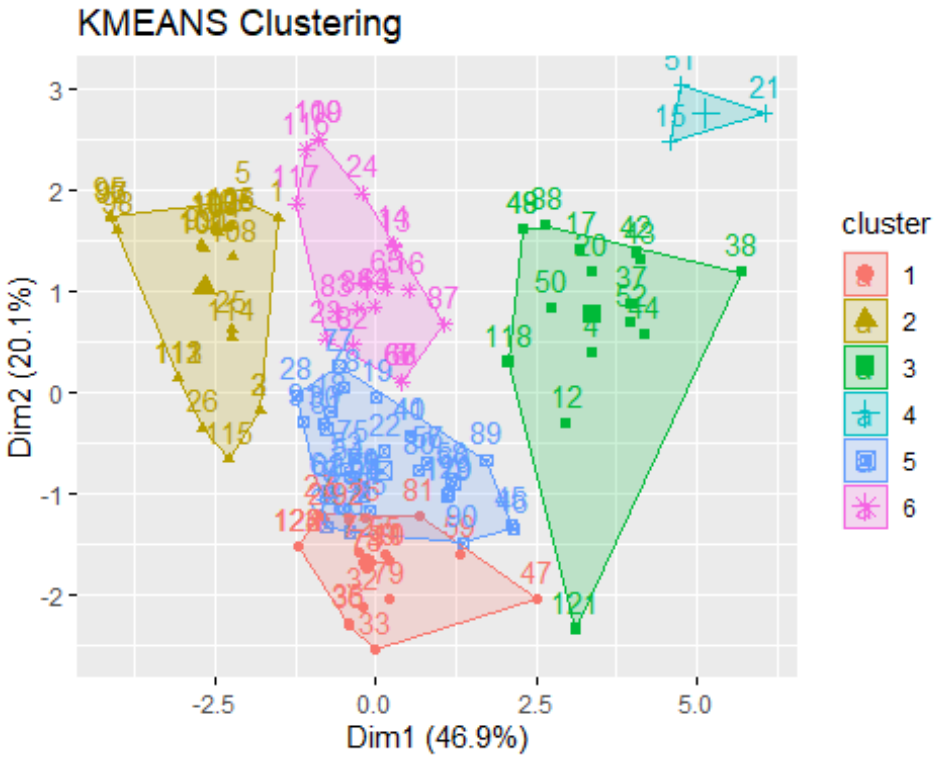
Una vez aceptada la conveniencia de llevar a cabo el análisis cluster, pasamos a determinar el número óptimo de grupos. En este caso, se nos ha especificado de antemano el número máximo de grupos en los que pueden dividirse los datos con los que estamos trabajando.

Trabajaremos con un máximo de 10 grupos correspondientes a los 10 lugares de los que dispone nuestro cliente para conversar sus vehículos todo terreno.

Tras realizar el análisis estadístico correspondiente hemos obtenido que el número óptimo de grupos es de 2, sin embargo, esto no concuerda con los que el cliente nos solicita. Nuestro cliente dispone de 10 lugares en los que conservar sus vehículos y desea repartirlos de tal forma que pueda disponer de ellos en distintas ubicaciones de Francia, Italia y Suiza.

No resulta pues una forma eficiente conservar los 125 vehículos en dos únicos grupos ya que no se cumple con el objetivo del cliente ni los garajes tienen capacidad suficiente para albergar tantos coches. Por ello, se ha procedido a realizar una repartición por *kmeans*.

Para ello se han realizado distintas comprobaciones con distintos números de grupos representándolos gráficamente y se ha llegado al resultado de que a partir de 4 grupos dejan de producirse solapamientos. Sin embargo, teniendo en cuenta que nuestro cliente desea distribuir sus vehículos en 6 ciudades europeas y que se produce poco solapamiento (como puede observarse en la gráfica más abajo), se considera más eficiente realizar la repartición en 6 grupos.



Conclusiones

Tras el análisis se llega a la conclusión de que para cumplir con el objetivo de nuestro cliente resulta más conveniente atender a un a un criterio *k-medias* pues nos da un resultado más acorde a la distribución geográfica de los lugares en los que pueden conservarse los vehículos. Lugares que se corresponden a las ciudades de: Andorra, La Rochelle, París, Suiza, zona costera de Francia-Italia y Córcega.

La distribución será la siguiente:

- Grupo 1: formado por 23 vehículos se distribuirán en Andorra y alguno de ellos en la zona costera de Francia-Italia ya que existen muchas semejanzas con los vehículos del grupo 5.
- Grupo 2: formado por 25 vehículos se distribuirán en los dos garajes de París.
- Grupo 3: formado por 16 vehículos se distribuirán en La Rochelle.
- Grupo 4: formado por 3 vehículos se distribuirán en Córcega.
- Grupo 5: formado por 38 vehículos se distribuirán en los 3 garajes de la zona costera de Francia-Italia.
- Grupo 6: formado por 20 vehículos se distribuirán en los dos garajes de Suiza.

Referencias

- Amat Rodrigo, J (2017) Clustering y heatmaps: aprendizaje no supervisado. Disponible en: https://rpubs.com/Joaquin_AR/310338
- Zafra, JM (2020). Análisis Cluster. Colegio Universitario de Estudios Financieros.