LOS COCHES DEL JEFE

AGRUPACIÓN Y REDUCCIÓN DE LA DIMENSIÓN

Beltrán Aller López

26/11/2019

CONCLUSIONES

Basándome en el análisis exploratorio de la colección de vehículos y el análisis de conglomerados realizado, se comprueba que:

- Se pueden dividir los vehículos en grupos claramente diferenciados.
- El número óptimo de clústeres a formar es de 6.
- A tenor de las capacidades de los garajes, será necesario llevar a cabo divisiones entre los grupos respecto de los garajes (partir un grupo y enviar una mitad a cada garaje).
- Para lo anterior se atenderá a la distancia entre garajes y las características del vehículo necesarias según el terreno.

Ilustración 1: Localizaciones de los garajes

Por tanto, se concluye que:

- Los vehículos más potentes, es decir, aquellos con una mayor cilindrada, aceleración y revoluciones por minuto, en principio deberán ser asignados a los garajes ubicados en Andorra y Suiza dadas las condiciones orográficas del terreno. Localizaciones escarpadas y que requieren de un motor con empuje para subir las cuestas.
- Los vehículos más ligeros, con mayor velocidad punta y menor consumo se repartirán entre el resto de las localizaciones, si bien se pueden realizar algunas consideraciones:
    - Se ha de atender al lugar de residencia habitual, es probable que allí sean necesarios vehículos con una mayor capacidad para la realización de la vida cotidiana (se intuye que puede ser París).
    - El transporte de los vehículos a Córcega ha de realizarse obligatoriamente por barco/ferry, lo que implica un coste en función del número de vehículos y el peso de estos. Variables a tener en cuenta.
    - Cannes, la Rochelle, Mónaco y San Remo se caracterizan por ser localizaciones glamurosas, con vehículos más de tipo *sport,* que de tipo sedán. Es por ello por lo que pueden ser propicios los vehículos de 2 plazas y gran velocidad punta (esta siempre ha estado asociada a los coches deportivos de lujo).

RESUMEN

Acorde a lo expuesto en el anterior informe, partimos de una base de datos ya limpia compuesta por 119 observaciones y 11 variables. La marca, el precio, el modelo y el tiempo de aceleración han resultado eliminadas del dataset.

El objetivo de esta nueva práctica será llevar a cabo la división de los distintos vehículos a fin de asignarlos más adelante a un garaje específico.

DESARROLLO

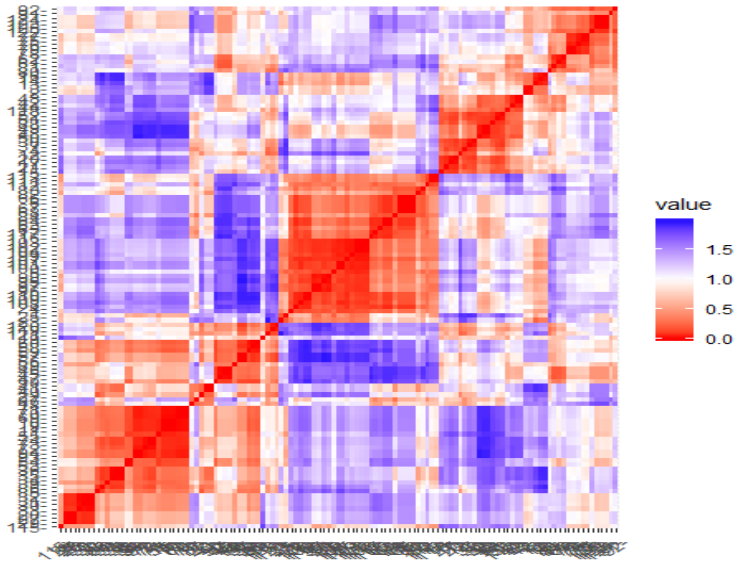
Para la realización del análisis de conglomerados o análisis clúster, una vez hecho previamente el análisis exploratorio correspondiente y escaladas las variables, el primer paso será el cálculo de las distancias. Representaré la matriz de distancias calculada a través de las correlaciones y de las distancias euclídeas de las observaciones. Dichas matrices nos sirven como medida de la similitud entre las observaciones.

El coeficiente de correlación de Pearson es una medida que puede emplearse para identificar la similitud, aunque no se la considera una métrica propiamente dicha.

La distancia euclídea entre dos puntos *p* y *q* se define como la longitud del segmento que une ambos puntos. En coordenadas cartesianas, la distancia euclídea se calcula empleando el teorema de Pitágoras.

La primera matriz representada será la de correlaciones, mostrándose un mapa de calor que nos indica en función del color, las distintas correlaciones entre las variables.

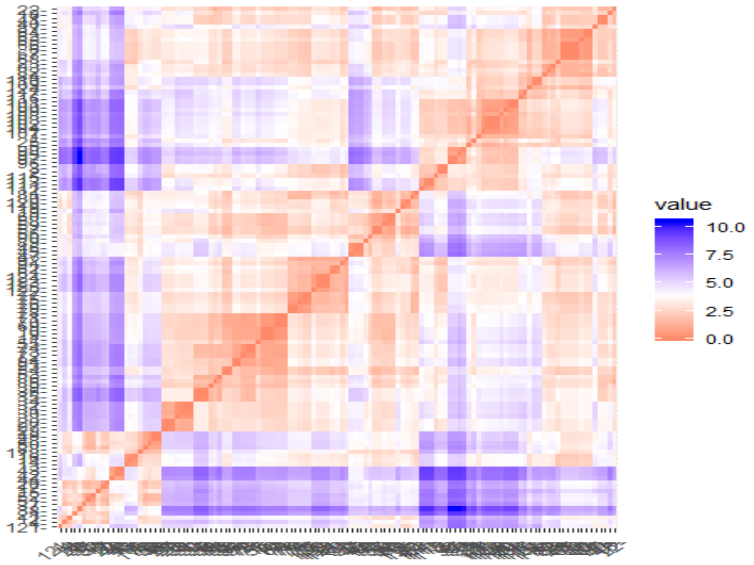
Ilustración 2: Matriz de distancias por correlaciones



Del análisis de la matriz de distancias basadas en la correlación no se puede extraer el número de clúster, en los cuales a priori va a ser dividida la muestra.

A continuación, calculo la matriz de distancias euclídeas como segunda medida de similitud.

Ilustración 3: Matriz de distancias euclídeas



De la matriz de distancias euclídeas no se extrae un conocimiento mucho más profundo que de la matriz de correlaciones, en este caso. Sin embargo, es cierto que se aprecia como una buena parte de las observaciones del dataset presenta altas distancias respecto a la otra.

A continuación, debo dirimir entre llevar a cabo un clúster jerárquico o un clúster no jerárquico. Los clústeres jerárquicos son llevados a cabo en situaciones en las que el analista, a priori, desconoce el número de clústeres a formar. En los clústeres no jerárquicos, el analista fija de antemano el número máximo de clústeres a realizar. Esta decisión se basa en el conocimiento del negocio por parte del analista.

En el caso que nos acontece, nuestro jefe dispone de hasta 10 garajes situados en distintas localizaciones, por tanto, ya sabemos de antemano que el número máximo de clústeres será de 10.

Una vez determinado el clúster a realizar, en este caso no jerárquico, la siguiente duda se cierne sobre el método que llevar a cabo, *k-mean o k-medians.*

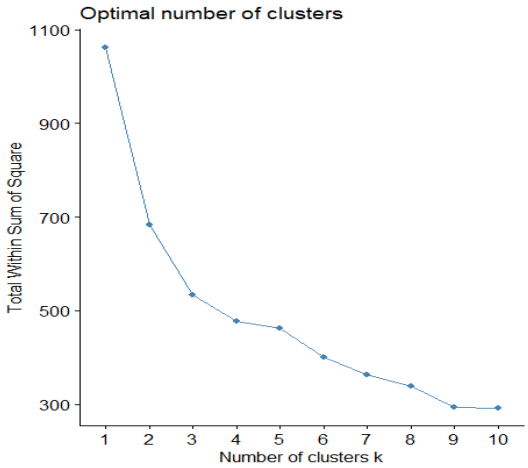
K-means minimiza la varianza intra-clúster y se calcula mediante el cuadrado de las distancias euclídeas entre las observaciones.

K-medoids minimiza las desviaciones absolutas y su cálculo se lleva a cabo a través de la distancia de Manhattan, la cual define la distancia entre dos puntos *p* y *q* como el sumatorio de las diferencias absolutas entre cada dimensión.

En el problema al que nos enfrentamos, calculo las distancias euclídeas, por lo que aplicaré el método *k-means.*

Ahora represento gráficamente cómo evoluciona la suma total de cuadrados internos en función del número de clúster.

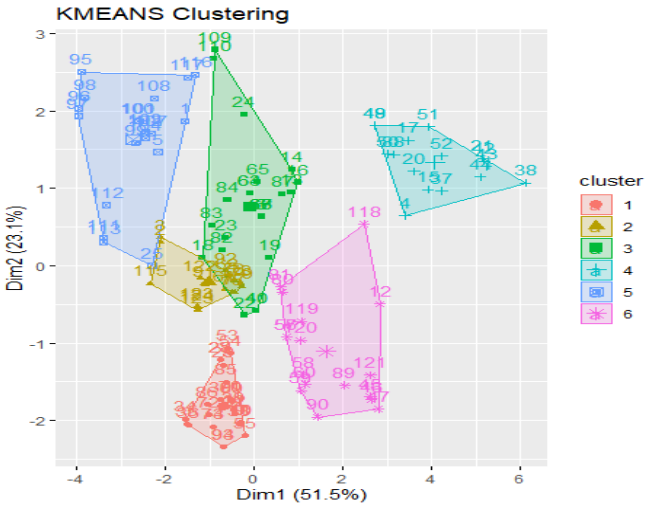
Ilustración 4: Análisis gráfico del número óptimo de clústeres



En el gráfico se puede observar que a partir de 6 grupos la reducción de la suma total de cuadrados internos se estabiliza, por tanto y, atendiendo también a la regla del codo, establezco el número óptimo de clúster en 6. Dicha decisión es momentánea, puesto que se limita a que en cada garaje haya como máximo 15 coches, todavía no se sabe si pueden hacer falta más grupos para cumplir tal condición.

Procedo a graficar los clústeres con sus distintas observaciones.

Ilustración 5: Clústeres por k-means



El tamaño obtenido para cada uno de los clústeres es el siguiente.

Tabla 1: Tamaño de clúster

Clúster	1	2	3	4	5	6
N.º coches	27	16	23	16	20	17

Se puede observar que han de realizarse agrupaciones, así como particiones a la hora de distribuir los vehículos entre los 10 garajes, puesto que únicamente se dispone de 15 plazas por garaje.

Bibliografía

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer.