Introducción

Este informe es la segunda parte de *Los coches del jefe*. En la primera parte nuestro objetivo era establecer una serie de grupos a partir de unas observaciones. Cada uno de estos grupos debía tener una serie de características comunes para poder pertenecer a él.

Prosiguiendo esta dinámica se va a proceder a agrupar el conjunto (esta vez de 125 coches) en diversos grupos con el objetivo de separarlos por garajes situados en distintos puntos geográficos de Europa.

Las variables con las que contamos son Marca, Modelo, Precio, Cilindros, Cilindrada, Potencia, Revoluciones, Peso, Plazas, Consumo 90, Consumo 120, Consumo urbano, Velocidad, Aceleración. Y nos quedaremos con Potencia, Revoluciones por minuto, peso, consumo urbano y velocidad. Ya que son las más relevantes y las únicas que podemos utilizar debido a los valores faltantes.

De estas hemos seleccionado Potencia, Revoluciones por minuto, peso, consumo urbano y velocidad de las cuales hemos completado los valores faltantes para cada una de ellas con las medias por columnas filtradas por marca de coche.

Gracias al análisis cluster se pueden separar estos 125 coches en grupos más o menos homogéneos y mandarlos al mismo garaje. Este tipo de análisis es de carácter no jerárquico ya que el jefe nos ha especificado que tiene que ser en 10 garajes, aunque haremos variaciones si fuera necesario dependiendo de la viabilidad de los grupos.

Vamos a ayudarnos de todas las librerías especializadas para el análisis cluster además de análisis gráfico para entender los resultados.

Análisis exploratorio de datos

Los datos que hemos recibido, que poseen 125 observaciones y 15 variables, tienen una serie de datos ausentes y además tenemos que tener en cuenta que no están en la misma escala.

Por tanto, en el pasado informe eliminamos todos los valores faltantes que encontremos en nuestros datos, pero en este caso no vamos a eliminarlos puesto que es un caso real y se disponen de 125 coches que deben ser colocados, no podemos deshacernos de ningún coche según especificó el jefe.

De la misma manera que tanto los de tipo *integer* los convertiremos en tipo numérico y las escalaremos más tarde. Estableceremos como índice la columna marca, para que nos distribuya las observaciones en función de esta y eliminamos las columnas que no sean de tipo numérico, no utilizaremos las categóricas.

La decisión del uso de solo las columnas mencionadas anteriormente viene determinada por el desplazamiento que harán los coches desde el punto de partida hacia los distintos destinos. Es por ello que no todas las variables tienen la misma importancia en este caso.

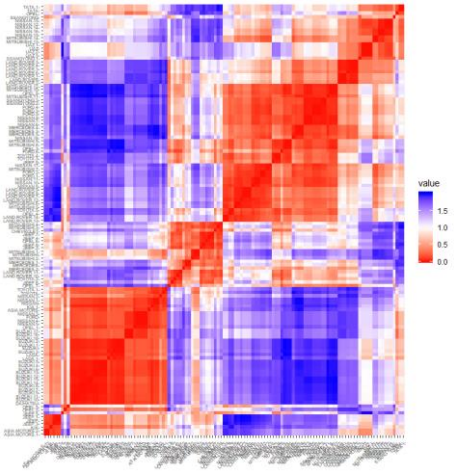
Por tanto, con nuestros datos limpios, ya podemos realizar la segmentación de los coches en cada garaje, teniendo en cuenta la distancia desde España que es el país donde tenemos los coches a los diferentes garajes.

Análisis clúster

Como ya se mencionó en el anteriormente informe, antes del análisis cluster debemos realizar el estadístico de Hopkings que es un contraste frente a la estructura aleatoria a través de una distribución uniforme del espacio de datos. La idea por tanto es contrastar una hipótesis de distribución uniforme de los datos frente a la alternativa.

En nuestro caso un valor de 0.2012 nos indica que los datos pueden segmentarse y por tanto podemos realizar el análisis.

El primer paso ya dentro de nuestro análisis es el contraste de la matriz de distancias por el método de Pearson para establecer las diferentes distancias entre observaciones y con las que más tarde se clasificarán las mismas en los diferentes grupos. Esto se ha realizado igualmente con Manhattan y Minkowsky.



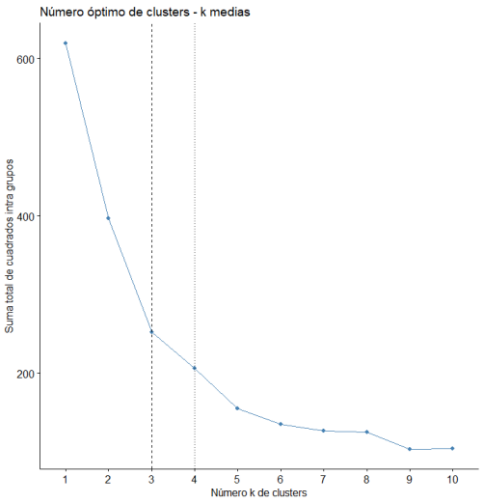
Como se puede observar en el gráfico, para que el agrupamiento fuera perfecto la distribución por colores y zonas debería ser perfecta también.

Vamos a proseguir con el análisis pues, aunque no haya grupos perfectamente distintos si que se aprecia una cierta diferencia en la gráfica

Fuente: Elaboración propia

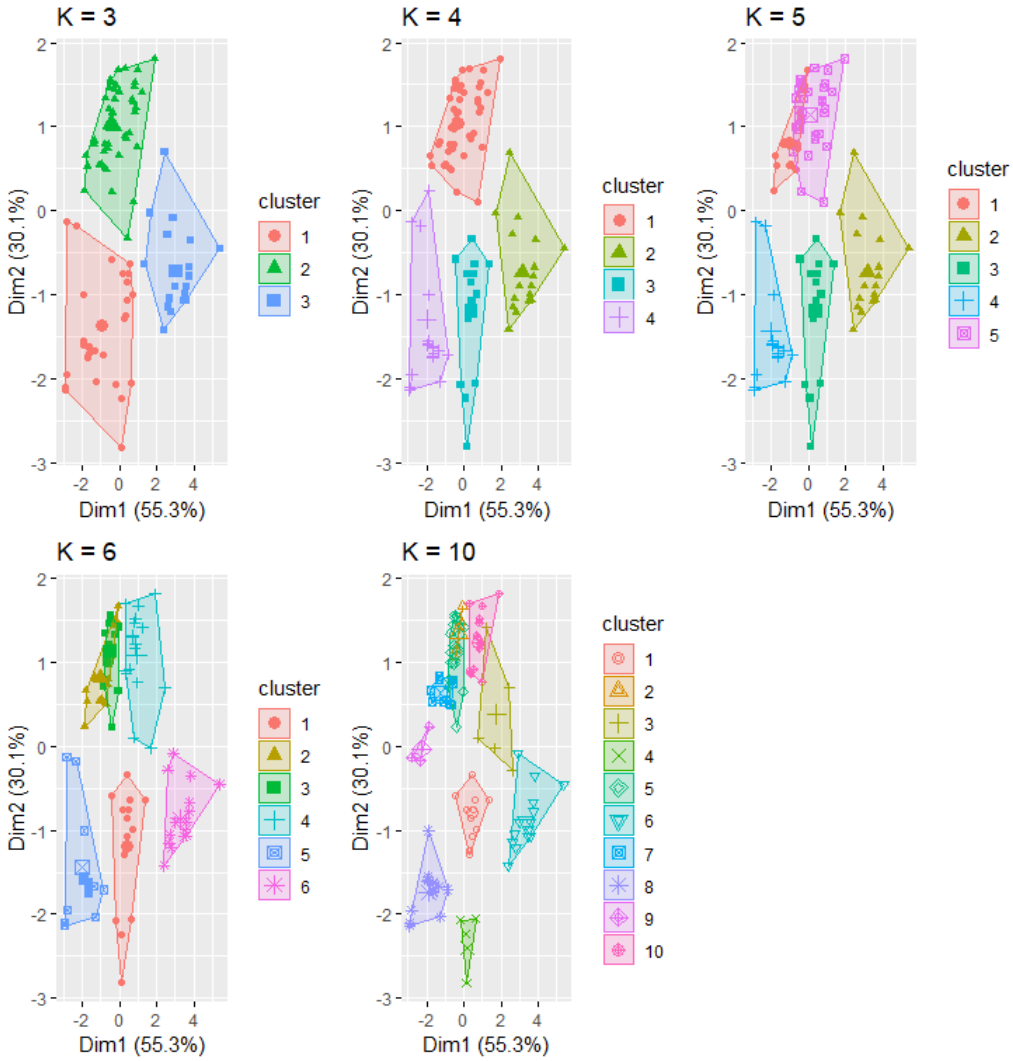
Uno de los problemas más grandes a los que nos enfrentamos cuando realizamos este tipo de análisis es la decisión de cuantos clústeres usar. El jefe ha propuesto 10 agrupaciones, pero desde un punto de vista de negocio esto no es rentable.

De manera gráfica podríamos ver cuantas agrupaciones serían las óptimas:



Según el gráfico entre 3 y 4 agrupaciones estadísticamente sería viable pero a continuación será necesario analizar el punto de vista de negocio.

Podríamos hacer un gráfico agrupando un número de clústeres entre 3 y 10 para asegurarnos de que no cometemos ningún error a la hora de tomar la decisión:



Se puede observar de manera precisa como se obtienen 3 y 4 agrupaciones perfectas de los coches y como, a partir de la quinta, los grupos empiezan a colisionar entre ellos.

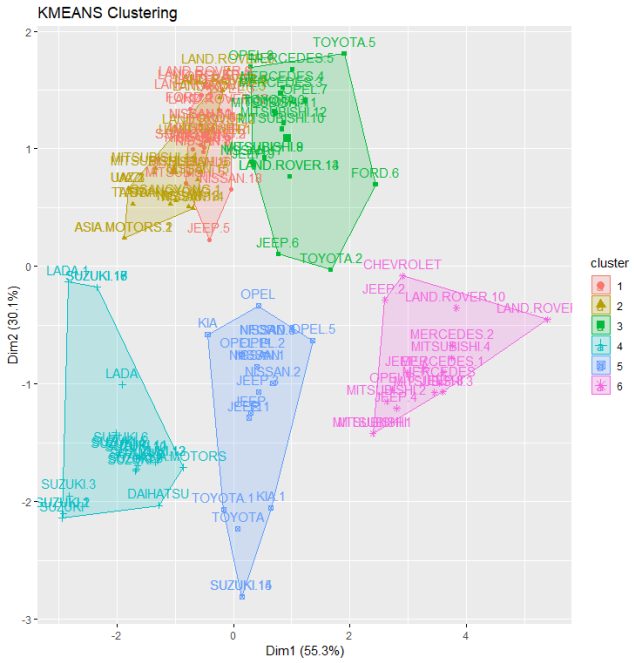
Según nos había encomendado nuestro jefe las agrupaciones ideales para él serían 10, pero como observamos en la última gráfica esto no tendría mucho sentido desde el punto de vista económico ya que hay grupos colapsados a la vez que muy cerca unos de otros.

Utilizando a la vez el mapa que se nos ha proporcionado vemos que se pueden hacer 6 agrupaciones geográficas, ordenadas por distancias entre ellas. Debido a que la diferencia entre 5 y 6 clústeres es mínima vamos a quedarnos con este último para el reparto final.

Conclusiones

Para concluir y resumir la información presentada, la tarea de agrupación de los coches no ha sido fácil. En primer lugar, hemos afrontado la decisión de la elección de clústeres. Por un lado, el jefe nos había pedido 10, por otro, desde el punto de vista estadístico y a través de complejas fórmulas la elección correcta hubieran sido 3 o 4 pero finalmente y desde un punto de vista enfocado más al negocio y ahorro de coste hemos decidido que serán 6 las agrupaciones.

Ahora solo queda ver que coche pertenece a cada agrupación:



Fuente: Elaboración propia

De manera detallado y basándonos en la información media por variables dentro de cada grupo se ha decidido reunir los 6 clústeres en 5 grupos:

- Grupo 1: Es la agrupación de los clústeres uno y dos. Las características de estos grupos son similares en cuanto a peso. Estos irán a la zona de Niza y Córcega ya que el transporte por mar nos permitirá ahorrar en costes ya que este se paga en función de la dimensión del coche y, por ende, su peso.
- Grupo 2: Los coches relativos al clúster seis son los que presentan mayor consumo y por tanto deberán ir a la zona de Andorra para un mayor ahorro en gasolina.
- Grupo 3: Coches pertenecientes al clúster tres, porque siguiendo el criterio de consumo estos son los siguientes y en consecuencia irán a La Rochelle
- Grupo 4: Pertenecientes al clúster cinco por criterios de velocidad y consumo irán a París
- Grupo 5: El clúster restante (clúster cuatro) irán directos a los dos garajes de Suiza