Los coches del Jefe 3

Alvaro Ferro Perez

20 de diciembre de 2018

Introducción

Este informe es la tercera y última parte de *Los coches del Jefe*. En la primera parte nuestro objetivo fue determinar una serie de grupos a partir de una información proporcinada, en este caso, sobre 125 coches. Cada uno de estos grupos tiene una serie de características en común para hacer posible su pertenecencia a cada grupo. Para hallar las variables representativas partimos de un *Análisis de componentes principales* que nos determinó que las variables que más influencia ejercían en el conjunto de datos eran:

- Potencia
- RPM
- Peso
- Consumo urbano
- Velocidad

En el segundo informe planteabamos con más detalle estas agrupaciones, procediendo en primer lugar a la limpieza y tratamiento del conjunto de datos. En este caso no podíamos deshacernos de las observaciones que contuvieran valores perdidos (*NA*) por lo que procedimos a sustituirlos por lo valores medios por marca de coche puesto que consideramos que era el valor más próximo y real que podíamos darle.

Seguidamente se procedió a realizar el análisis Cluster. Para ello obtuvimos la matriz de distancias y determinamos el número de clústeres desde el punto de vista estadístico y el punto de vista de negocio. En nuestro caso nos quedamos con 6 agrupaciones o clústeres para *aparcar* los coches en las distintas zonas geográficas propuestas.

El objetivo de este último informe es la realización con aún más detalle del Análisis Clúster a través del *K-Means* y el *K-Medoids* ambos similares pero con sus matices y diferencias.

Análisis Exploratorio de datos

Recordando los expuesto en anteriores informes de manera resumida, el análisis exploratorio tenía por objeto limpiar y ordenar las 125 observaciones de 15 variables que se nos daban.

Los valores faltantes, como ha sido expuesto anteriormente, se han reemplazado por la media por marca de coche y posteriormente han procedido a escalarse ya que inicialmente no presentaban la misma escala y no podíamos realizar el análisis. En este caso concreto, se ha creado una nueva variable para los datos sin tipificar que son con los que trabajaremos para realizar el *PAM* (*Partitioning around medoids*).

La columna marca se ha establecido como índice de fila para que posteriormente podemos observar qué coches han sido clasificados en cada clúster así como el medioide en el *PAM* (este será el coche cuyas características representen mejor a cada grupo en particular)

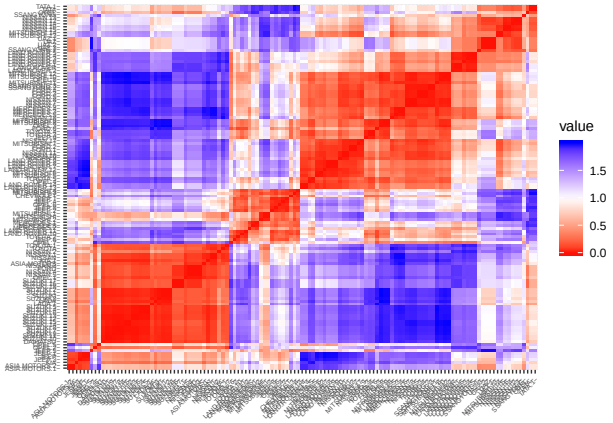
La decisión del uso de las columnas que se han mencionado en el primer apartado viene dada desde el punto de vista de negocio puesto que nuestro objetivo es el traslado de coches desde España hasta distintos puntos de Francia y Suiza y queremos por tanto minimizar el coste lo máximo posible.

Medidas de distancia

Para poder llevar a cabo los métodos de clustering necesitamos definir las similitudes que tienen las observaciones. Cuanto más se asemejen dos observaciones, más próximas estarán en cuanto a distancia y por tanto podrán pertenecer a un mismo grupo.

El primer paso es el cálculo de la matriz de distancias por el método Pearson para establecer estas diferencias. Se realiza igualmente por el método Manhattan y Minkowsky:

```
#Realizamos la representación gráfica.
set.seed(123)
fviz_dist(qdist, lab_size = 5)
```

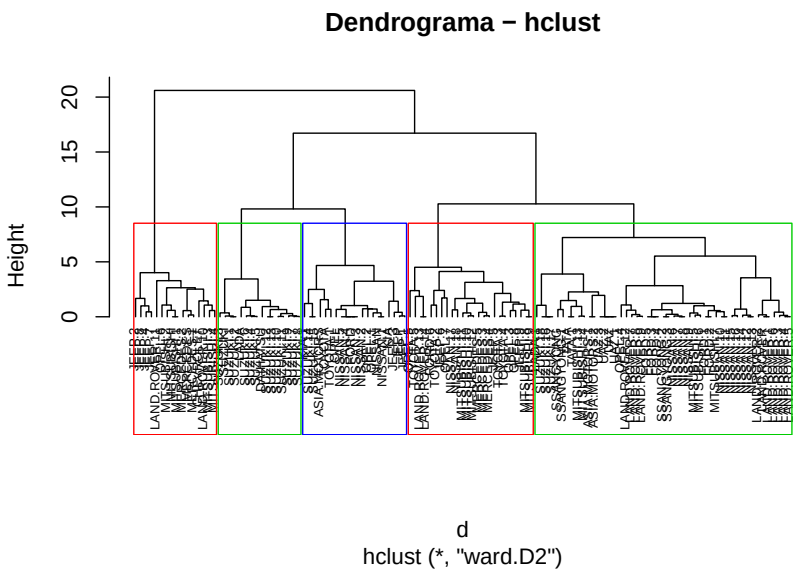


Como se puede observar en el gráfico, para que la distribución fuera perfecta, la distribución por colores debería serlo igualmente. El gráfico nos impide ver con claridad las observaciones en los ejes X e Y pero esto no importa ya que lo que queríamos era la visión conjunta.

Podemos apreciar que hay cierta diferencia en la gráfica por lo que podemos intuir agrupaciones en el *dataset*, por ello vamos a continuar con nuestro análisis.

Usando el dendrograma podemos ver como el algoritmo ha agrupado de manera más visible e intuitiva. Para ello vamos a pasarle el parámetro $k = 5$ para que nos haga cinco divisiones.

```
plot(fit, cex = 0.6, hang = -1, main="Dendrograma - hclust")
rect.hclust(fit, k=5, border = 2:4)
```

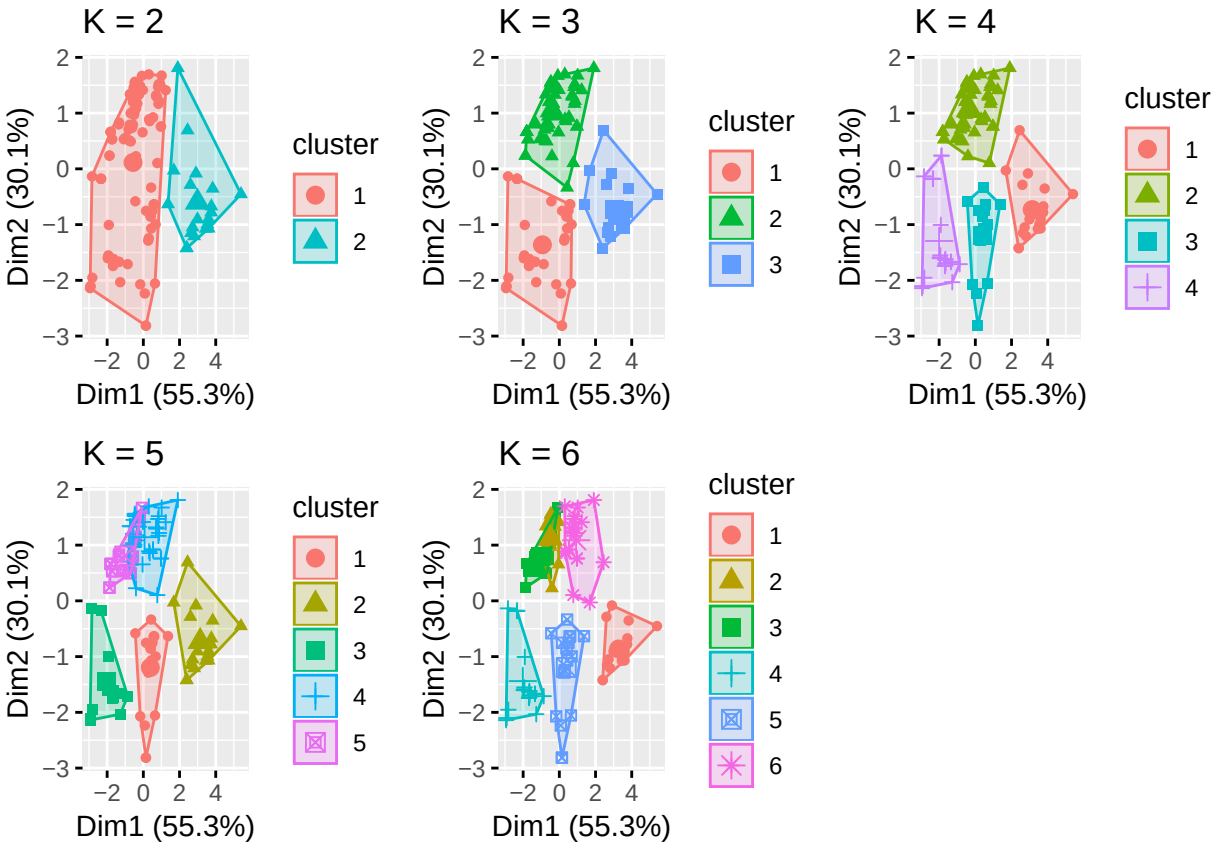


De manera rápida e intuitiva podemos ver que las marcas de coches están todas más o menos agrupadas en el mismo conjunto, lo que nos lleva a pensar que las características son parecidas dentro de cada grupo.

K-Means Clustering

El método K-Means agrupa las observaciones en un número de clústeres distintos, donde el número lo tiene que determinar el analista. Ya vimos en el informe anterior que esto supone uno de los mayores problemas. Desde el punto de vista estadístico una agrupación perfecta serían 4 clústeres, pero en nuestro caso y desde el punto de vista de negocio nos interesan 6. Veámoslo gráficamente:

```
grid.arrange(p1, p2, p3, p4, p5, nrow = 2)
```



Como vemos 4 suponen una agrupación perfecta pues no hay superposición de los clúster pero a la hora de tomar la decisión de manera geográfica veremos que con solo 4 tenemos un problema, y es que hay demasiados coches por grupo y la división no podría hacerse de manera correcta.

Para ver las características de cada grupo podemos ver las características de los centroides para así hacernos una idea del grupo completo:

```
set.seed(123)
caracteristicas <- kmeans(cochesescalados, 6)
caracteristicas$centers
```

```
##      potencia        rpm        peso     consurb    velocida
## 1 -0.08457694 -0.9607736  0.69317121 -0.34440906 -0.02020228
## 2 -1.19323727  0.9018852 -1.71572242 -1.22755771 -0.84434559
## 3  2.03798055  0.4152770  0.64604432  1.87075218  1.70276046
## 4 -0.30105803  1.5231790 -1.43514160 -0.57770395  0.14502543
## 5 -0.71635049 -0.4306737  0.11354843 -0.05158258 -1.24063422
## 6  0.23594436  0.6691238 -0.09170227  0.39362736  0.55018907
```

Podemos observar lo siguiente: - Potencia:Los clusteres que más potencia tienen serán el 3 y 6, y los de menores serán el 2 y 5. - RPM: El de mayor revoluciones por minuto es el 4 y el 2 y los menores el 1 y 4. - Peso: Los de mayor peso son el 1 y el 3 y los menores el 2 y el 4. - Consumo urbano: Los que mayor consumo tienen son los del 3 y el 6 y los menores los del 2 y el 4. - Velocidad: Y los más rápidos son los del grupo 3 y 6 y los más lentos son los del 5 y 2.

Por tanto según este criterio podremos deducir los garajes donde debemos de introducir a cada uno de los coches, en función de las variables seleccionadas.

Los grupos 3 y 6 son los coches más potenes, los que mas consumen y los que más velocidad pueden conseguir,

por tanto podemos deducir que son coches deportivos por que no tienen mucho peso pero tienen mucha velocidad y mucha potencia por tanto los mandaremos en ferry a Corcega y a los alrededores

K-Medoids clustering (PAM)

Este es muy similar a K-Means en cuanto que ambos agrupan las observaciones en K-clusters. La diferencia fundamental radica en que cada clúster está representado por una observación presente en el clúster (medoid), mientras que en K-Means cada clúster está representado por su centroide, que es el promedio de todas las observaciones pero ninguna en particular como hemos vista anteriormente.

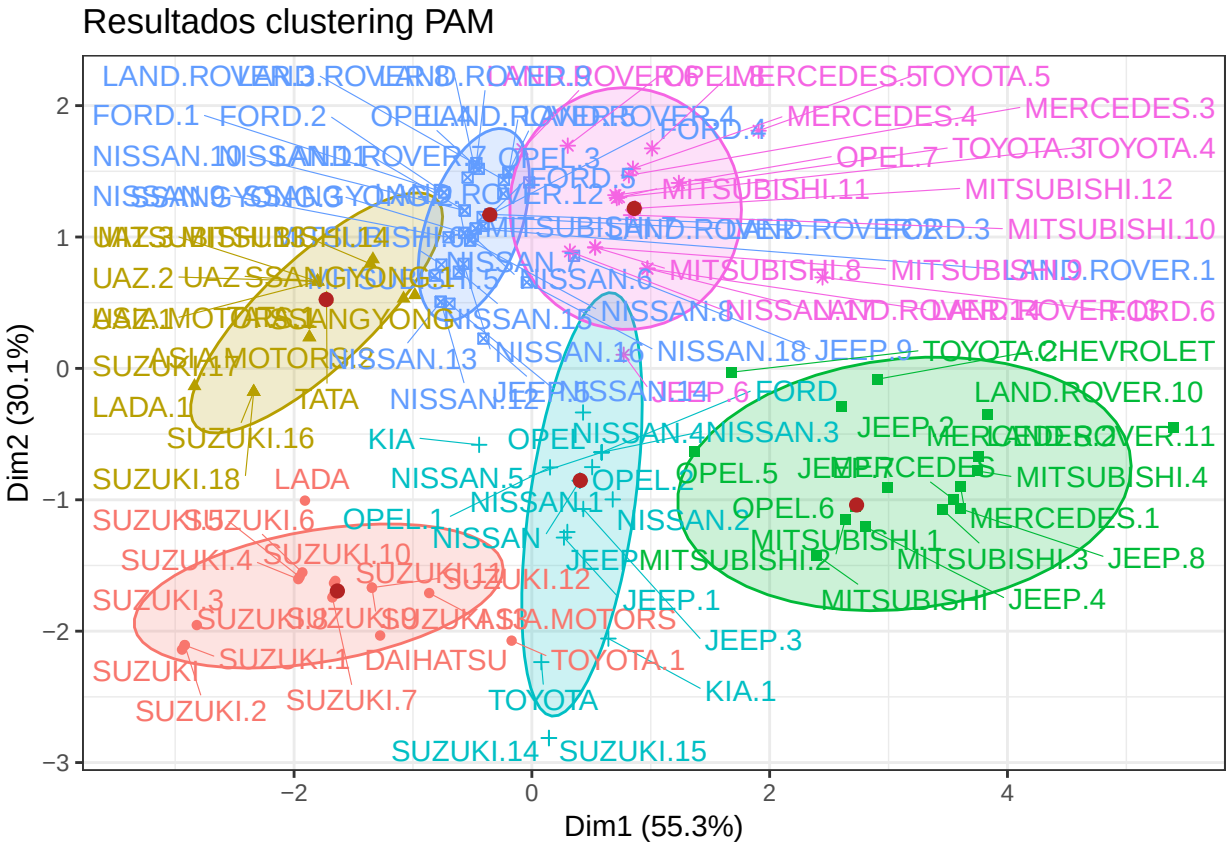
```
pam_clusters$medoids
```

```
##                 potencia        rpm       peso    consurb   velocida
## SUZUKI.9      -0.5877020  1.2975795 -1.6173020 -0.8757261 -0.3828449
## TATA.1        -1.3063592 -0.2386456 -0.1784654 -0.8402472 -0.9246066
## OPEL.6         1.5948868  0.7389522  0.3610983  1.9625797  1.3628316
## NISSAN.1       0.1841892  0.7389522 -0.1634776  0.1886386  0.5802869
## FORD.3        -0.4546173 -0.9369297  0.5259649 -0.2725861 -0.3226492
## MITSUBISHI.12  0.2108061 -0.9369297  1.0955044  0.3305539  0.2793082
```

Como puede verse, ahora los medoids están representados por una observación de cada grupo (Suzuki.9, Tata.1, Opel.6...) donde sus características en potencia, rpm, peso, consurb y velocidad están representadas en la tabla. Cabría esperar pues que cada grupo en el que estuviera cada uno de esos coches, presentara características parecidas.

De manera gráfica se pueden obtener las agrupaciones y sus medoids

```
# Creación del gráfico
set.seed(123)
fviz_cluster(object = pam_clusters, data = cochesescalados, ellipse.type = "t",
             repel = TRUE) +
  theme_bw() +
  # Se resaltan las observaciones que actúan como medoids
  geom_point(data = medoids, color = "firebrick", size = 2) +
  labs(title = "Resultados clustering PAM") +
  theme(legend.position = "none")
```



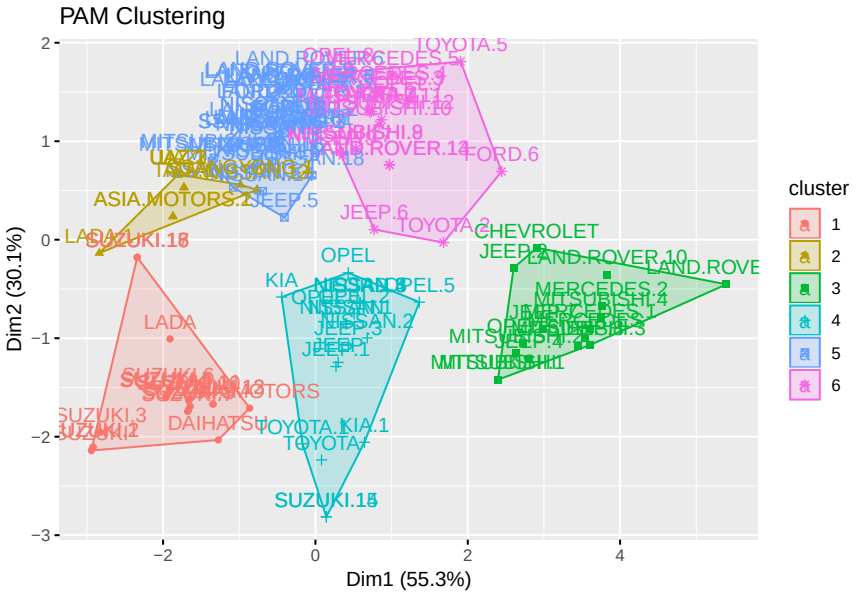
Al igual que el K-Means aquí observamos los 6 grupos aunque también podemos observar los medoids de cada uno de ellos que como es lógico, se situarán más al centro.

Conclusiones

Para concluir y resumir la información presentada, hemos de exponer que la decisión del número de clústeres no ha sido fácil. Por un lado el jefe nos pedía 10 agrupaciones, por el punto de vista estadístico asumíamos 4 y por el punto de vista de negocio, fueron 6.

Este último ha sido con el que finalmente hemos hecho la agrupación quedando el resultado así:

```
set.seed(123)
coches.eclust = eclust(cochesescalados, FUNcluster = "pam", stand = TRUE,
                       hc_metric = "euclidean", k = 6)
```



De manera detallada y basándonos en la información de las variables dentro de cada grupo se ha decidido realizar 3 agrupaciones de coches que mandaremos a las zonas geográficas propuestas. Primero tenemos que establecer las características generales de cada grupo para que la agrupación se haga de la forma más correcta posible.

Grupo 1: Son coches de poca potencia de poco peso y de poco consumo. Son todos practicamente suzuki menos un toyota, un lada y un asia motors. Sin embargo este grupo presenta un RPM muy alto. Este grupo incluye en su mayoría coches de la marca Suzuki.

Grupo 2: Son coches de poca potencia de poco peso y de poco consumo. Es un conjunto un tanto heterogéneo. Estos presentan tambien una velocidad baja. Este grupo incluye coches de diferentes marcas como: Tata, UAZ, Asia motors, Suzuki, Ssangyong y mitsubishi.

Grupo 3: Son los coches que más consumen, más potencia y más RPM por minuto (velocidad) tienen. Se incluyen Mitsubishi, Mercedes, Jeep, Land Rover y Chevrolet entre otros.

Grupo 4: Tienen una potencia baja, un peso bajo pero sin embargo el consumo, las RPM y la velocidad son altas. Este grupo incluye Nissan, Ford, Opel, Jeep, Kia, Opel, Suzuki, Kia y Toyota

Grupo 5: Todas las caracteristicas son bajas excepto el peso que es alto. Este grupo incluye coches de prácticamente todas las marcas. Es el grupo que peor diferenciado está.

Grupo 6: Estos son los más pesados e incluyen marcas como Mercedes, Opel, Ford, Nissan y Jeep.

El reparto por tanto lo haremos de la siguiente manera:

Conjunto del grupo 3, 4 y 6: Debido a que son los que más consumen, el desplazamiento por ferry a la zona de Niza y Córcega será mucho más rentable que enviarlos por carretera a otras zona.

Conjunto del grupo 1 y 2: Debido a que son, en general coches de poco peso y consumo, podrán desplazarse con poco gasto a la zona de Suiza. Los cuatro coches sobrantes (debido a que los garajes solo pueden incluir a 15 coches cada uno), se irán a la zona de París.

Conjunto 5: Son coches en general, heterogéneos y el reparto de hará de la siguiente manera teniendo siempre en cuenta el punto de vista de negocio y las restricciones que tenemos: 16 coches a París (recordando que en esta zona ya hay cuatro), esto nos resultaría en un total de 10 coches en cada garaje en esta ubicación. Otros 10 coches se irán a la zona de La Rochelle que solo dispone de un garaje y los 10 restantes a Andorra. Es importante mencionar que a esta última zona mandaremos los que, dentro de este grupo reúnan las condiciones de más consumo.