Los coches del Jefe 3

Jorge Casan VÃ¡zquez

22 de Diciembre de 2018

EXECUTIVE SUMMARY

El objetivo de este informe es realizar una compilación de todo el análsis cluster que habiamos efectuado en la parte 1 y 2,en donde en la primera tuvimos que realizar un análisis de componentes principales y determinar la pertenencia de cada coche en un grupo determinado, realizando para ello un análisis de componentes principales en donde vimos que las variables estadísticamente significativas eran la potencia,el RPM, el peso, el consumo urbano y la velocidad.Por otro lado, en el segundo informe realizamos un tratamiento de los NA, los cuales debemos tratarlos, y para ello los reemplazamos por los valores medios de la marca de coche.

Con este informe realizaremos una matriz de distancias y determinanaremos el número de clusteres óptimo desde un punto de vista estadístico, los cuales y en virtud de nuestro análisis son 6 y desde un punto de vista de negocio,continuando con la realización aún más en detalle del K-means y el k-mediods, ambos similares pero con sus diferencias,las cuales las explicaremos a continuación.

Finalmente asignaremos el conjunto total de coches a 5 grupos diferentes, por razón de abaratamiento de los costes logísticos y de transporte.

INTRODUCCIÓN

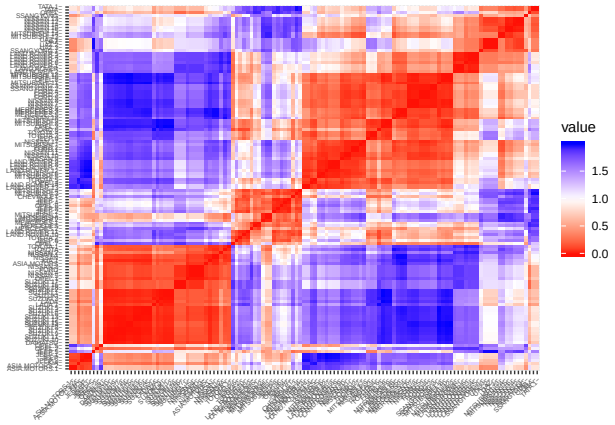
Con el análisis exploratorio de datos que realizamos en la parte 1 limpiamos y ordenamos las 125 observaciones de las 15 variables, las cuales a través de la matriz de correlaciones pudimos determinar cuáles eran las variables estadísticamente significativas. Una vez realizado este paso previo fundamental hemos procedido a escalar las variables puesto que estaban en diferentes unidades métricas, lo cual resulta imprescindible si queremos compararlas aunque perdamos varianza explicada.

La decisión del uso de las variables anteriormente mencionadas viene dada desde un punto de vista de negocio puesto que nuestro objetivo principal es asignar el reparto de coches lo más eficientemente posible hacia los diferentes garajes distribuidos geográficamente por el territorio europeo y por tanto minimizar el coste de transporte y en general logístico lo máximo posible.

MEDIDAS DE DISTANCIA, DENDOGRAMA

Para poder llevar a cabo el clustering tenemos que definir las similitudes que tienen los coches. Cuánto más se asemejen las observaciones más próximas estarán en cuanto a distancia y por tanto podrán peternecer a un mismo grupo, para ello obtenemos la matriz de distancias calculadas a través del método de Pearson, Manhattan y Minkowski.

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



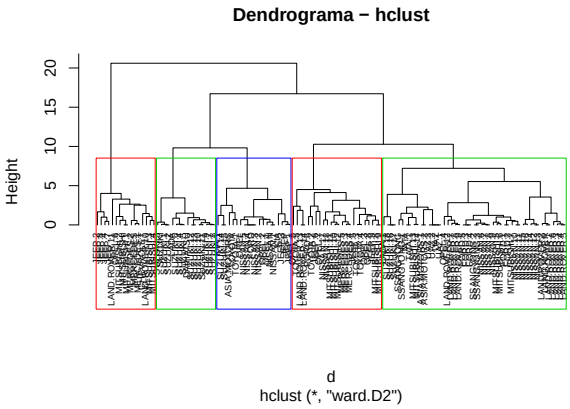
A simple vista podemos ver que existen 4 posibles agregaciones, 3 bien definidas, arriba y debajo a la izquierda, con color azul y rojo, respectivamente y abajo a la derecha. También podríamos pensar que la parte de arriba a la derecha forma una posible agrupación de coches, pero genera dudas.

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.

El gráfico nos impide ver con claridad el conjunto total de coches representados en el eje de abcisas y en el corrdenadas, por lo que seguiremos con nuestro análisis.

Representamos el conjunto de marcas de coches a través de un Dendograma a través del método de Ward. Podemos pensar que existen 5 agrupaciones, la resaltada en color verde es la que menos duda genera. Sin embargo, la agrupacion vista de izquierda a derecha es la que presenta mayor complejidad.

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



K-MEANS CLUSTERING

El método del K-means agrupa las observaciones en un número de clusteres distintos, donde el número se tiene que determinar a priori. Desde el punto de vista estadístico, segun el criterio de optimalidad analizado en el parte 2, el número perfecto de clusteres era de 4. Sin embargo, según el punto de vista de asignación de coches por garaje pero sobre todo por abaratamiento de costes logísticos hemos decidido realizar 6 clusteres.

```
library(gridExtra)
```

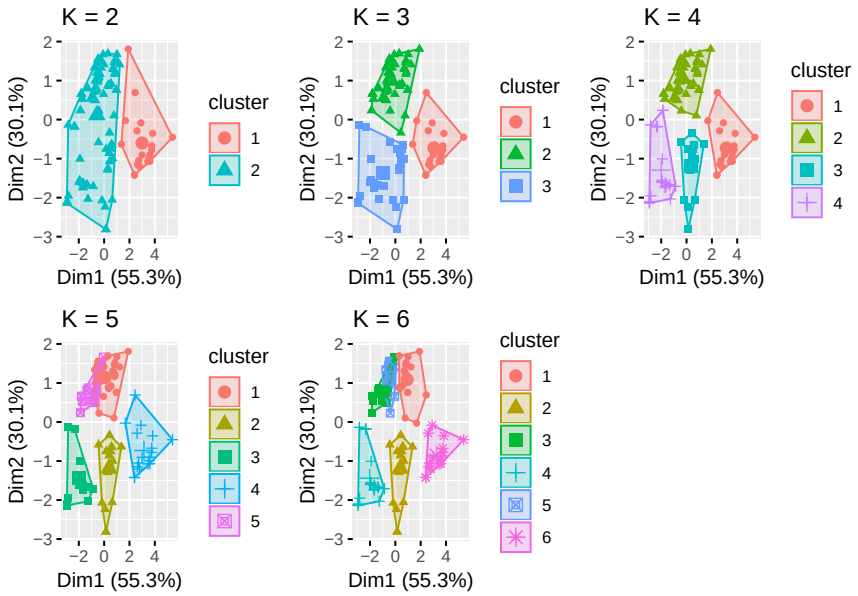
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
require(ggrepel)
```

```
## Loading required package: ggrepel
```

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



Para ver las características de cada grupo podemos ver las características de sus centroides para así hacernos una idea del grupo completo.0

```
caracteristicas_kmeans <- kmeans(cochesescalados, 6)
caracteristicas_kmeans$centers
```

```
##     potencia        rpm       peso     consurb   velocida
## 1 -0.3536947 -0.9252917  0.4506508 -0.65989656 -0.2649616
## 2  2.1006086  0.4823328  0.6067122  1.89605693  1.7352927
## 3 -0.9463969  1.0914195 -1.6250101 -1.02777815 -0.5605657
## 4 -0.7163505 -0.4306737  0.1135484 -0.05158258 -1.2406342
## 5  0.3345749 -0.9439126  0.9396305  0.28620537  0.4147487
## 6  0.2627092  1.0392144 -0.3341895  0.19431522  0.5923261
```

El grupo que tiene los coches más potentes lo conforma el cluster 3. Por RPM, la mayor representatividad la tendrá el cluster 3 y 6. Por variable peso la mayor será la del cluster 1 y la 3 la menor. Finalmente, consumo urbano y velocidad, la mayor representatividad la tendrán el 2 para ambos y la menos el cluster 3 y 4, respectivamente.

K-MEDIODS CLUSTERING

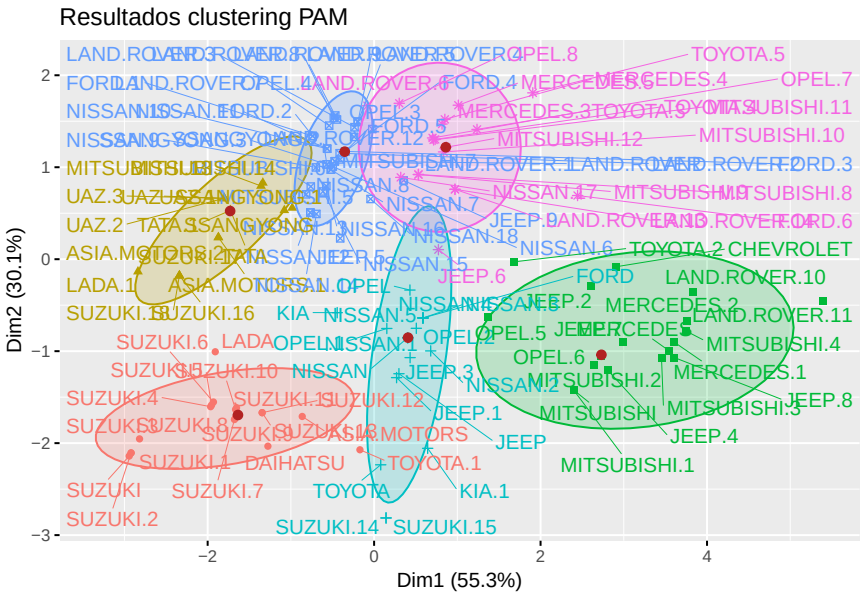
A través del algoritmo PAM, muy similar al K-means, en cuanto a que ambos agrupan las observaciones en función del número de clusters, pero con el k-mediods, cada cluster está representado por una observación (el mediode), mientras que con el K-means cada cluster está representado por su centroide.

```
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
```

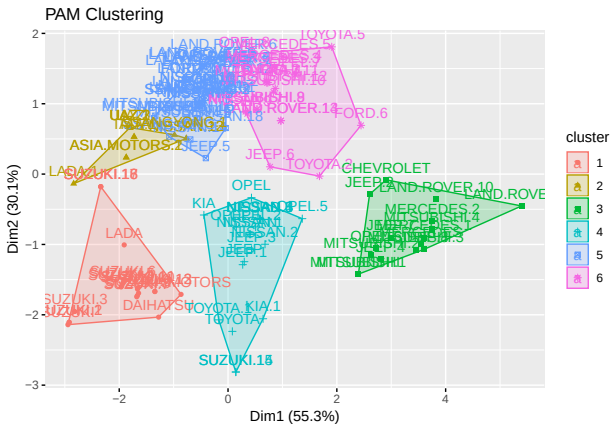
Los mediodes están representados a través de un modelo de coche, en donde el conjunto de características están representadas en la tabla. Lo podemos ver gráficamente, en donde se espera que dentro de cada cluster estuviera cada uno de los mediodes, presentando características similares cada grupo.

```
#Resalamos las observaciones que actúan como mediodes
fviz_cluster(object = pam_clusters, data = cochesescalados, ellipse.type = "t",
             repel = TRUE) +
  geom_point(data = medoids, color = "firebrick", size = 2) +
  labs(title = "Resultados clustering PAM") +
  theme(legend.position = "none")
```



Representamos gráficamente la cantidad total de clusteres y nos quedan 6 conjuntos

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



CONCLUSIONES

La decisión del número de clusters no ha sido tarea fácil. Por un lado, nuestro jefe nos pedía asignar a cada grupo un máximo de 10 coches, es decir, hasta un máximo de 10. Sin embargo, y siguiendo con criterios de orden estadístico vimos que lo mejor era realizar 4 clusteres pero lo más importante para nosotros es realizar la asignación de clusteres por criterio de negocio para abaratar los costes logísticos de cara a la distribución de los cohes en diferentes puntos geográficos, por lo que al final hemos decidio realizar 6 clusteres.

Basándonos en la información media por marca de coches dentro de los 6 clusteres hemos decidido distribuirlos de la siguiente manera:

- Grupo 1_ Estos coches se distribuirán a la zona de Niza, Mónaco y Córcega, ya que por razones de distancia abarataríamos los costes de distribución y trasporte marítimo para entregarlos a la zona de Córcega.
- Grupo 2_ Este conjunto de coches se distribuirán a la zona de Suiza ya que el conjunto de coches agregados en este cluster son idóneos para clientes con un alto poder adquisitivo.
- Grupo 3_ Estos coches irán a parar a la zona de París debido a que son los que menor consumo de gasolina producen y los clientes que tenemos en la zona están interesados por estos coches en particular.
- Grupo 4_: Estos coches son los que mayor consumo presentan y en base a este criterio consideramos que por raones topográficas de La Rochelle con alto nivel de renta per capita, estos coches deberían ir a parar allí.
- Grupo 5_ : Finalmente, estos coches irán a parar a la zona de Andorra por razones de abaratamiento y por criterio de cercanía.

Enlace de GITHUB:

Formato RMD:

https://github.com/JORGECASAN/TecnicasReduccion/blob/master/COCHES%20DEL%20JEFE/COCHES_JEFE_PARTE3_JORGE.Rmd