Análisis Cluster, Los coches del jefe parte 3

Luis Llera García

20 de diciembre de 2018

1.EXECUTIVE SUMMARY

El objetivo de este informe será concluir con la distribución de todos los coches de nuestro jefe en las diferentes localizaciones que nos ha indicado y en función de las características semejantes de los coches, los cuales quedaban distribuidos en la práctica anterior, esta es la tercera y última parte de los coches del jefe. Primeramente nuestro objetivo era establecer un conjunto de grupos a partir de una base de datos que nos aportaba la información, eran 125 coches. Por tanto siguiendo los requisitos del análisis cluster considerar que cada grupo debe ser homogéneo respecto de alguna característica entre las observaciones,pretende generar grupos similares entre si y considerar que cada grupo debe ser diferente de los demás respecto de las mismas característica, grupos heterogéneos entre si.

Primeramente nos basamos del análisis de componentes principales para realizar la selección de variables,el cual busca determinar de algún modo las relaciones íntimas existentes entre todas las variables. Su objetivo primario es construir nuevas variables, artificiales, a partir de combinación lineal de las originales, con la característica de ser independientes entre sí. Con ello el análisis nos arrojó los siguientes resultados.

• Potencia • RPM • Peso • Consumo urbano • Velocidad

En la segunda parte, tuvimos que hacer un análisis más profundo en el que se debía de plantear estas agrupaciones con más detalle, primeramente con la limpieza de datos y su análisis exploratorio, en este caso, a diferencia de un problema de negocio como pueden ser quiebras o creación de perfiles para una pagina web, no se podían eliminar los valores faltantes NAs, ya que todos los coches deben de ser situados en un garaje. Por tanto, en este análisis hemos optado por rellenar dichos valores faltantes con los valores medios de dicha marca de coche, ya que considerabamos que era el valor más cercano a la realidad que podíamos darle para que no influya en nuestro análisis de manera negativa distorsionando la información.

Más tarde se realizó el análisis Cluster, una de las cuestiones básicas del análisis cluster estriba en la selección de la medida de la similitud entre las observaciones, nosotros hemos elegido un critério de distancias que lo que indica es la pertenencia de cada observación a un grupo en función de la menor distancia existente. Para ello obtuvimos la matriz de distancias y estableceremos el número de clusteres óptimo tanto desde el punto de vista de negocio como desde el punto de vista estadístico. Desde el punto de vista de necesidades de negocio nos quedaremos con 6 agrupaciones o clústeres para distribuir los coches en las distintas zonas geográficas establecidas por nuestro jefe.

Por tanto, este informe tiene como objetivo fundamental el análisis Cluster através de dos métodos, del K-Means y K-Medoids, es decir, un análisis cluster en mayor profundidad.

2.ANÁLISIS EXPLORATORIO DE DATOS

Como en todo análisis es necesario realizar un tratamiento a los datos, para que los mismos, se encuentren limpios, en la misma medida, es decir, estandarizados, limpiarlos de valores faltantes... En los informes previos este proceso tenía como objetivo limpiar y ordenar los 125 coches u observaciones distribuidos en las 15 variables.

Como hemos indicado anteriormente los valores faltantes o NAs, los hemos reemplazado por la media por marca de coche y posteriormente las hemos escalado, por que no se encontraban en la misma magnitud, por tanto debemos de tipificarlas o normalizarlas. En este caso hemos creado una nueva variable para los datos sin tipificar que son con los que trabajaremos para realizar el PAM.(Partitioning around medoids).

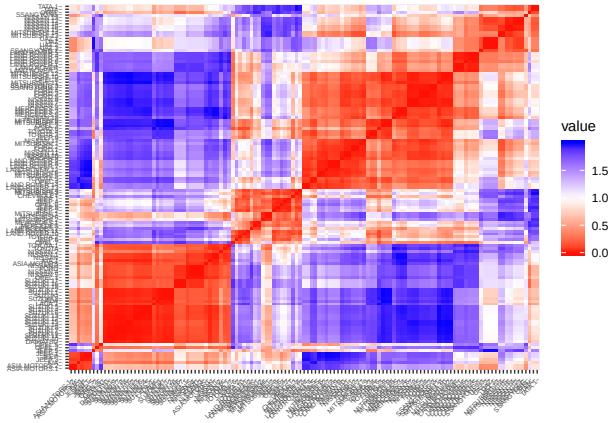
El índice de la columna se ha renombrado como la marca de dicho coche, para posteriormente poder observar en los clusteres o grupos resultantes las marcas que se encuentran en los diferentes clusteres asi como el medioide en el PAM(que será el coche cuyas características definan a cada grupo en concreto)

Por tanto, la selección de variables o el uso de columnas, que hemos mencionado anteriormente se ha realizado así debido a que se ha llevado a cabo dicho análisis desde un punto de vista del negocio, ya que nuestro objetivo es repartir los coches teniendo en cuenta un critério de distancia geográfica(desde España a los diferentes garajes, ya sea en Suiza, Francia...) por tanto, queremos minimizar el coste lo máximo posible, y el ACP tambíen nos arrojó los resultados de la importancia de dichas variables.

3.MEDIDAS DE DISTANCIA

La más importante entre ellas es la distancia euclídea, la distancia "ordinaria" entre dos puntos de un espacio euclídeo. Al realizar el análisis es necesario enunciar que criterio vamos a definir para la medida de la similitud entre las observaciones, de ella dependerán los resultados, y con ello la dependencia a cada grupo. La elección de los métodos dependerán fundamentalmente de nuestros datos, primeramente calcularemos la matriz de distancias por el método de Pearson, de la misma manera que lo realizaremos por el método Manhattan, que la cual funciona mejor con vectores de alta dimensión que la euclidiana, y Minkowsky:

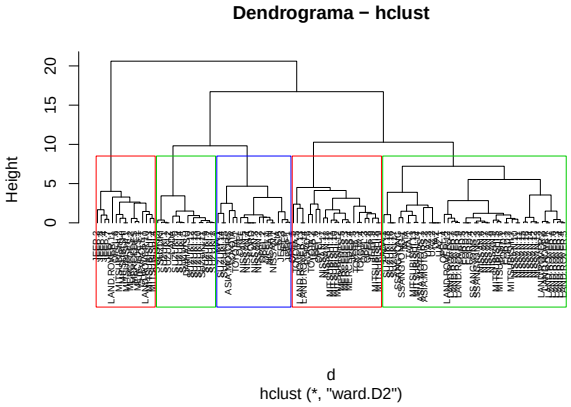
```
##Realizamos la representación gráfica.
fviz_dist(qdist, lab_size = 5)
```



Como podemos observar en la siguiente representacion, buscamos que los colores tengan una distribución lo más similar posible, pero desde los ejes X e Y no se puede observar con claridad las observaciones. Podemos deducir que hay agrupaciones en nuestos datos, debido a la diferencia de colores parece que generan varios grupos, dos azules en la parte superior izquierda y otro en la parte inferior derecha y los tres rojos a través de la diagonal con esto podemos deducir que las marcas están biendefinidas dentro de los grupos, por tanto podemos continuar con el análisis.

Ahora fijándonos en el dendograma se puede apreciar como el algoritmo ha realizado una serie de agrupaciones, definiremos el parámetro K=5 para que realice cinco divisiones ya que queremos un numero más alto de dos grupos, debido a las exigencias del jefe.

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



Gracias al dendograma que es un gráfico perfecto para una visualización rápida se puede deducir que los coches están agrupados mas o menos en el mismo conjunto(en la misma caja digamos), esto significa que las observaciones de cada grupo son homogéneas, que es uno de los requisitos básicos del análisis cluster.

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.

4.K-MEANS CLUSTERING

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática.

Dentro del algoritmo depende de nosotros el número de clusteres a utilizar, como vimos en el informe anterior este es uno de los problemas más grandes ya que existe una disyuntiva entre la visión de negocio y el análisis estadístico, que tendremos que solventarlo nosotros como científicos de datos. Utilizaremos un rango de gráficos que nos muestran las diferentes combinaciones, estableceremos el límite en seis grupos ya que nos interesan desde el punto de vista de distancia geográfica con los coches nos interesan 6 grupos:

```
library(gridExtra)
```

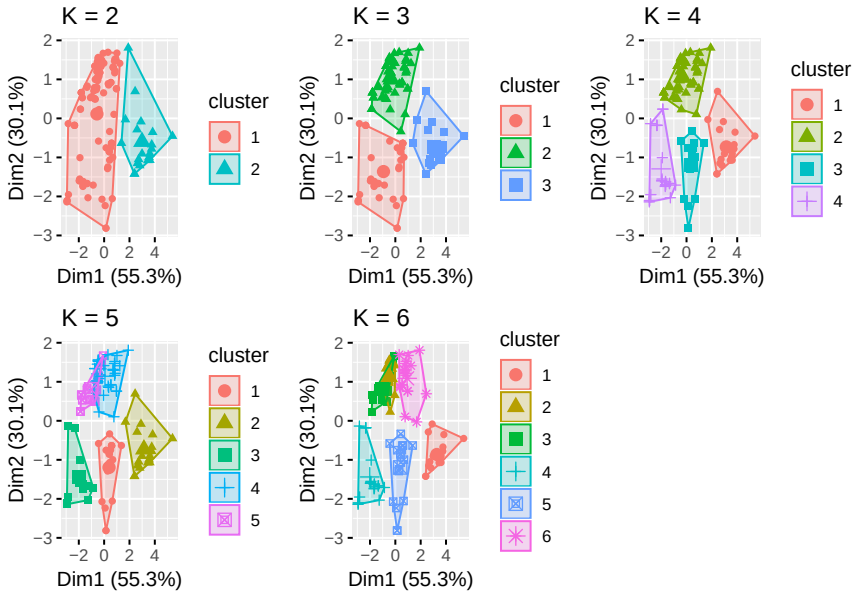
```
##
## Attaching package: 'gridExtra'
```

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

```
require(ggrepel)
```

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

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



Estadísticamente hablando se puede observar que de dos a cuatro grupos la división es perfecta pero desde el punto de vista de las exigencias del jefe y pensando en la distribución de los garajes realizaremos seis grupos. Para observar la pertenencia a cada grupo podemos observar los centroides, para hacernos una idea general de cada grupo:

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

Esta salida nos muestra los siguientes resultados: - 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 .

5.K-MEDIODS (PAM)

El algoritmo k medoids es un algoritmo de agrupamiento relacionado con el algoritmo k significa y el algoritmo de cambio medoid. Los algoritmos k significa y k medoids son particionales (dividiendo el conjunto de datos en grupos). K significa intentar minimizar el error cuadrado total, mientras que k medoids minimiza la suma de las diferencias entre los puntos etiquetados para estar en un grupo y un punto designado como el centro de

ese grupo. En contraste con el algoritmo k significa, k medoids elige puntos de datos como centros (medoides o ejemplares). Un medoide de un conjunto de datos finitos es un punto de datos de este conjunto, cuya disimilitud promedio con todos los puntos de datos es mínima, es decir, es el punto más centralmente ubicado en el conjunto.

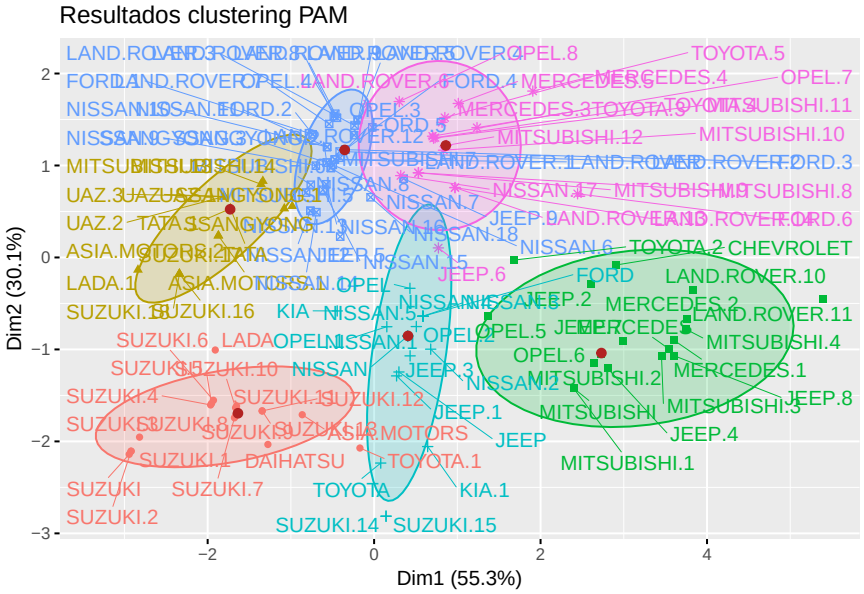
```
pam_clusters$medoids
```

```
##                  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
```

La siguiente tabla nos muestra en el índice los diferentes modelos de coche(SUZUKI.9, TATA.1, OPEL.6, NISSAN.1, FORD.3. . .)para cada observación por cada variable(potencia, revoluciones por minuto. . .) donde cabe esperar que cada grupo tendrá unas características similares.

Mostraremos los diferentes clusteres en forma de elipses, y mostrando sus medioides.

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



Como hemos afirmado anteriormente el algoritmo de K-medioids, PAM, es una alternativa robusta a k-means(ya que tiene una mejora respecto al k-means, que es el problema con su alta sensibilidad respecto a los outliers) para dividir un conjunto de datos en grupos de observación, como en el método k-medoids, cada grupo está representado por un objeto seleccionado dentro del grupo. Los objetos seleccionados se denominan medoides y corresponden a los puntos ubicados más centralmente dentro del grupo, se presupone que las observaciones entre los grupos son muy parecidas entre si.

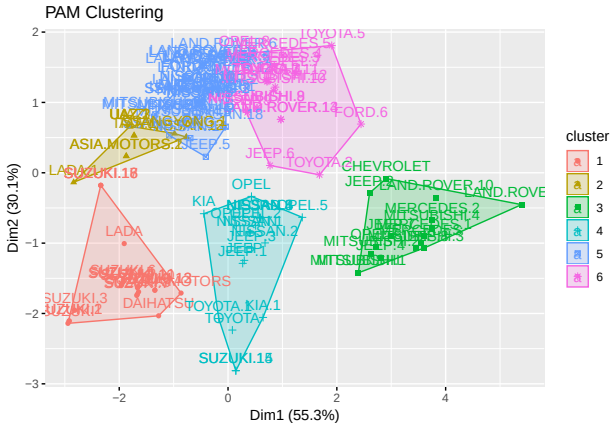
6.CONCLUSIONES

Como hemos podido estudiar tanto el PAM como el K-MEANS forman parte de los métodos no jerárquicos o de partición ya que existe un problema que exigen un conocimiento previo del conjunto de datos a analizar para determinar apropiadamente el número de grupos a priori, ese es uno de los problemas básicos de estos métodos. Teniendo en cuenta los criterios de negocio y estadística, el jefe nos pedía 10 grupos y los análisis estadísticos nos indicaban que el número óptimo era 4, teniendo en cuenta ambos acabamos eligiendo 6. Previamente hemos realizado un análisis de componentes principales que nos ha reducido a dos dimensiones, las cuales explican entre las dos un 85.4% de varianza. Hemos llevado a cabo una reducción de 15 variables a 5.

A partir del ACP la solución originada de dos dimensiones, se ha realizado un análisis cluster para agrupar los coches en función de sus características semejates.

Finalmente nos hemos quedado con los resultdos que nos muestra la siguiente gráfica.

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



Utilizando el PAM y utilizando la información de las variables con las que se ha generado el modelo, y teniendo sobre todo en cuenta la situación geográfica de los diversos garajes hemos decidido unificar seis clusteres en cinco.

Por tanto, con la información que nos aportan las cinco variables que hemos elegido, decidimos realizar trés agrupaciones de los diferentes coches que mandaremos a los diferentes garajes de nuestros jefes, pero primero vamos a definir las características de los diferentes clústeres que hemos realizado.

Grupo 1: Aquí se encuentran los coches de poca potencia y con un consumo urbano bajo. Son todos practicamente de la marca Suzuki menos un Toyota, un lada y un Asia motors. Sin embargo este grupo presenta un RPM muy alto. Grupo 2: Estos tambien son coches con potencia baja, pero a su vez tambien tienen poco peso y consumen poco. Estos presentan tambien una velocidad baja. Este grupo incluye coches de diferentes marcas como: Tata, UAZ, Asia Motors, Suzuki, Ssangyong y Mitsubishi. Grupo 3: Esos son los coches que possen un cnsumo más alto al igual que su potencia y mas RPM. Se incluyen Mitsubishi, Mercedes, Jeep, Land Rover y Chevrolet entre otros. Grupo 4: Sus revoluciones por minuto a la vez que su velocidad son muy altas, pero su potencia y peso son bajos. 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.

Con todo esto podemos concluir la siguiente división de los coches en los diferentes garajes, los clusteres 3, 4 y 6 los colocaremos en la zona de Niza y las zonas costera debido a que son los que más consumen y por temas de ahorro de costes los mandaremos allí en ferry. Los grupos uno y dos debido a que son coches con poco peso y consumo, podran desplazarse pon un coste reducido hasta la zona de Suiza, pese a que sobren 4(ya

que los garajes como máximo son 15 coches) los meteremos en uno de los garajes de París. El grupo 5 que es el más numeroso y grupo más heterogéneo, desde un punto de vista de negocio y nuestras limitaciones 16 coches los meteremos en París(teniendo en cuenta los cuatro colocados ateriormente), podriamos incluir diez coches en cada garaje de dicha ubicación, y otros diez sobrantes a la zona de la Rochelle que solo tiene un garaje disponible y los diez que quedan a Andorra, los coches que más consumen irán a la zona de Andorra y los siguientes a la zona de La Rochelle y asi sucesivamente, para ahorrar costes desde un punto de vista de consumo

7.Bibliografia

https://www.ecured.cu/Distancia_eucl%C3%ADdea https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html#kmeans http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html Lopez Zafra, J. (2018). El Analisis Cluster. Madrid: Máster en Data Science para Finanzas - CUNEF.