Coches del Jefe - Parte 2

Hugo César Octavio del Sueldo

11/29/2020

Problema a solucionar

Su jefe tiene una semana complicada y le ha pedido que le haga una propuesta de cómo repartir la colección en las distintas residencias. Como ud. bien sabe, podría repartirlos como máximo en las diez que posee en la actualidad (precisamente está, durante esta semana, cerrando la venta de alguna de ellas, que quizá sustituya por alguna otra), pero, siendo una opción conservadora, quizá no sea la más adecuada, atendiendo a las caracerísticas que ud. ya conoce de los vehículos.

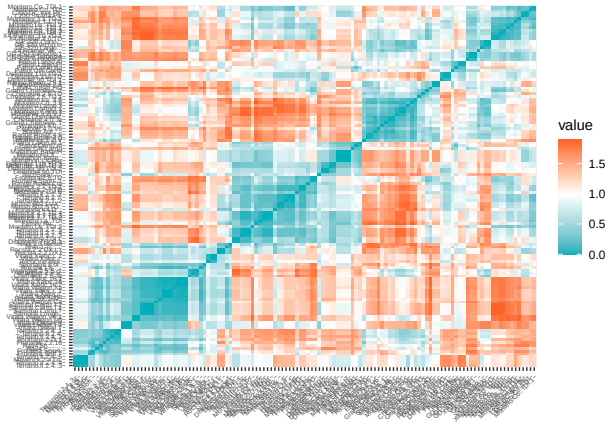
Introduccion

Esta es la segunda parte de la practica relacionada a los coches del jefe. En la primera parte de esta serie de entregas realizamos un analisis explotario de los datos con los que contabamos, analizando aquellas variables que nos resultaban interesantes para la agrupacion y posterior reparto de los coches del jefe en las localidades o residencias con las que el contaba. En esta segunda parte, como habran visto arriba, deberemos hacerle una propuesta de reparto de la coleccion de autos al jefe. Para esto, estudiaremos el numero adecuado de grupos en los que dividir la coleccion. Teniendo en cuenta de que el maximo numero de coches por residencia es de 15 y sabiendo de que en el caso de que propongamos grupos con mas coches, tendremos que escoger las residencias en las que guardarlos, atendiendo al criterio de distancia. En el siguiente enlace tiene un mapa de las residencias actuales. El criterio de reparto debe ser consistente, y debe justificar su decisión en un máximo de 4 páginas.

Análisis Cluster

Creamos un nuevo dataframe, con las columnas numericas que son necesarias en cualquier análisis cluster ya que los mismos no admiten otro tipo de variables. Con estas columnas vamos a realizar un análisis cluster para distribuir los grupos de coches en las diferentes localidades. En esta oportunidad utilizare todas las variables numericas ya que teniendo en cuenta las dimensiones del fichero considero que todas las columnas o variables puede aportarme informacion valiosa de cara al armado de grupos para su distribucion.

Realizamos la representación gráfica.

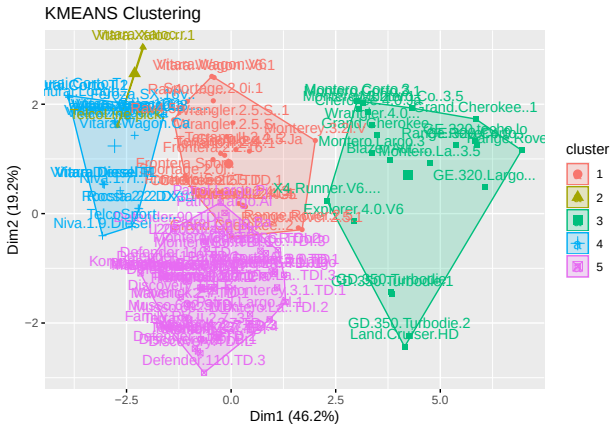


Analizando las variables en base a sus distancias, podemos observar que hay grupos de observaciones bien identificadas en base al tamaño de los cuadrados y de sus colores. Mientras mas azul, podemos observar que los coches son mas homogeneos y los rojos significa que son mas heterogeneos. Aqui para saber lo similares que son los vehiculos, lo analizamos en base a la posicion que tienen los coches en virtud de sus caracteristicas. Este grafico ya nos da la pauta de que seria una muy buena idea separar en diferentes clusters.

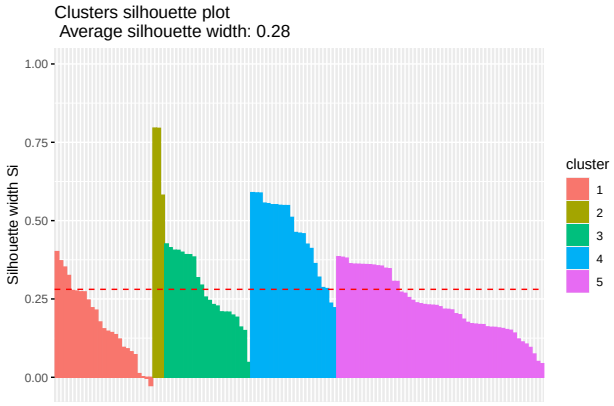
Cluster no jerarquico

Utilizare un metodo de cluster `no jerarquico` como el **kmeans** y el **pam** porque, para la tarea que se nos ha encomendado, no sabemos el numero de grupos en los que debemos agrupar sino que establecemos, de acuerdo a nuestra experiencia, cuál es el número de grupos con el que proceder a la partición de la población; exige un conocimiento inicial de la misma, o una decisión de tipo administrativo (los costes de segmentar la población en muchos grupos pueden ser elevados).

Para el trabajo realizare un cluster de 5 grupos ya que con este tamaño estoy segmentando bastante bien mis coches con algún pequeño solapamiento pero sin que nos dificulte mucho la tarea de la distribucion. El algoritmo de la libreria eclust nos daba como óptimo un número de 10 cluster, pero, para esta cantidad de clusters habia muchos solapamientos entre los vehiculos.




```
##   cluster size ave.sil.width
## 1       1   25          0.18
## 2       2    3          0.72
## 3       3   22          0.28
## 4       4   22          0.46
## 5       5   53          0.23
```



Con el algoritmo `pam` nos sucede algo similar al kmeans, donde para el numero k optimo que propone el mismo de 9, tenemos muchos solapamientos entre los coches, dificultandonos la tarea que nos han planteado el jefe. Entonces procederemos con el numero de cinco cluster que a nuestro entender nos parece razonable.

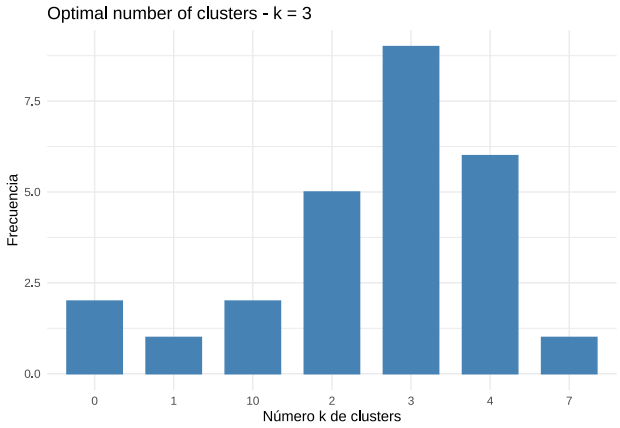
Identificación del número de grupos Pruebas que permiten establecer el número adecuado de grupos antes de proceder a la segmentación, que establecerá las observaciones en grupos de acuerdo con la minimización del indicador establecido.

```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##

## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 5 proposed 2 as the best number of clusters
## * 9 proposed 3 as the best number of clusters
## * 6 proposed 4 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 2 proposed 10 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************

## Among all indices:
## ==================
## * 2 proposed  0 as the best number of clusters
## * 1 proposed  1 as the best number of clusters
```

```
## * 5 proposed  2 as the best number of clusters
## * 9 proposed  3 as the best number of clusters
## * 6 proposed  4 as the best number of clusters
## * 1 proposed  7 as the best number of clusters
## * 2 proposed  10 as the best number of clusters
##
## Conclusion
## =========================
## * According to the majority rule, the best number of clusters is  3 .
```



Entre los distintos metodos utilizados para conocer el numero optimo de clusters como la regla del perfil y el método del hombre, el paquete NBClust nos dice que son 3 el optimo de clusters que deberiamos generar. No obstante esto, utilizaremos un numero de 5 grupos, como comentamos anteriormente, porque con ese numero de cluster estamos segmentando bastante bien y en calidad de la representacion no perdemos mucho.

Conclusiones

La distribución de los coches según el análisis cluster la vamos a realizar del siguiente modo:

- El Método KMeans y PAM permitieron clasificar la muestra en 5 grupos bastante bien diferenciados.
- Si bien el jefe solicitó 10 grupos de coches, desde el punto estadístico hubiera sido correcto elegir 3 clusters, pero finalmente desde el punto de vista del negocio y la distancia es preferible 5 grupos.
- Dado que no se tiene otro criterio que la distancia geográfica, se considera la distancia como unico criterio para distribuir los clusters:
- Casa 3, Casa 5 y Casa 10, se asignan al cluster 5 conseguido con kmeans debido a que es un grupo grande con 56 coches por lo que de estos, 45 iran a esas tres casas y luego, los 11 coches restantes de este cluster seran localizados en la Casa 4 en Corse junto con el cluster 2 que en similitud no son tan distintos.
- Casa 2, se asigna con los coches del cluster 3 debido a que en distancia estan muy alejados respecto del cluster 5 y el cluster 2.
- Casa 1, de manera similar asignaremos a los coches del cluster 3 que no tenian espacio en la Casa 2 junto con 8 coches del cluster 1 que en terminos de distancia estan bastante cerca.

- Casa 7 y Casa 6, se deben colocar los coches del cluster 1 que no tenian espacio en la casa 1 (unos 15 coches para Casa 7 y 2 Coches para Casa 6) y los 13 coches restantes, para completar Casa 6 (13 coches) deben ser del cluster 4.
- Casa 9, se colocaran los coches del cluster 4 que son los restantes.