Tarea 04 - Repartición de Coches

Mayra Goicochea Neyra
27/11/2019

Introducción

Se tiene la información de 125 vehículos Todo Terreno clásicos adquiridos por el dueño de Family Office. La información que se nos ha entregado está clasificada en 15 características: *Marca, Modelo, PVP, Cilindro, CC, Potencia, RPM, Peso, Plazas, Cons90, Cons120, ConsUrb, Velocida, Acelerac y Acel2.*
En el anterior trabajo, se realizó un análisis de las características que son relevantes para hacer la distribución, considerando que se cuenta con 10 lugares de estacionamiento. El análisis comprendió una revisión exploratoria de los datos, matriz de correlación de las variables y un estudio de las distancias en base a sus índices de correlación. Se concluyó con que las características relevantes son *Marca, Plazas, Aceleración, Velocidad Máxima, Consumo Urbano y el Precio*.
La finalidad de éste informe es realizar el agrupamiento de los vehículos de forma eficiente, considerando las 10 propiedades del dueño y que cada residencia puede albergar 15 coches. Para lograr éste objetivo primero se completará la información ausente (en la revisión exploratoria se encontraron 83 casos con valores NA, y dado que la muestra es de los 125 vehículos a distribuir no se puede omitir ninguno), luego se escalarán las características numéricas, y finalmente se realizará el análisis clúster de las observaciones.

Análisis Exploratorio de Datos

Existen 83 casos con observaciones ausentes que luego se deben imputar antes del análisis clúster.

```
##         marca       modelo               pvp              cilindro
##  NISSAN    :19   Length:125         Min.   : 1367000   Min.   :4.000
##  SUZUKI    :19   Class :character   1st Qu.: 2721000   1st Qu.:4.000
##  LAND ROVER:15   Mode  :character   Median : 3730000   Median :4.000
##  MITSUBISHI:15                      Mean   : 4004459   Mean   :4.592
##  JEEP      :10                      3rd Qu.: 4675406   3rd Qu.:6.000
##  OPEL      : 9                      Max.   :10419200   Max.   :8.000
##  (Other)   :38
##        cc          potencia          rpm            peso
##  Min.   :1298   Min.   : 64.0   Min.   :3600   Min.   : 930
##  1st Qu.:2184   1st Qu.: 95.0   1st Qu.:4000   1st Qu.:1462
##  Median :2497   Median :112.0   Median :4500   Median :1750
##  Mean   :2570   Mean   :117.1   Mean   :4671   Mean   :1675
##  3rd Qu.:2835   3rd Qu.:125.0   3rd Qu.:5200   3rd Qu.:1909
##  Max.   :5216   Max.   :225.0   Max.   :6500   Max.   :2320
##                                                NA's   :2
##      plazas          cons90          cons120         consurb
##  Min.   :2.000   Min.   : 6.600   Min.   : 8.40   Min.   : 8.10
##  1st Qu.:4.000   1st Qu.: 7.800   1st Qu.:10.53   1st Qu.:10.43
##  Median :5.000   Median : 8.600   Median :12.20   Median :12.00
##  Mean   :5.184   Mean   : 8.897   Mean   :12.25   Mean   :12.59
##  3rd Qu.:5.000   3rd Qu.: 9.700   3rd Qu.:13.90   3rd Qu.:13.57
##  Max.   :9.000   Max.   :13.700   Max.   :18.50   Max.   :22.10
##                  NA's   :10       NA's   :15      NA's   :7
##     velocida        acelerac                     acel2
##  Min.   :120.0   Min.   : 9.40   Menor a 10 segundos:  3
##  1st Qu.:140.0   1st Qu.:13.20   Mayor a 10 segundos:122
##  Median :146.5   Median :15.60
##  Mean   :150.6   Mean   :15.43
##  3rd Qu.:160.8   3rd Qu.:18.50
##  Max.   :196.0   Max.   :22.00
##  NA's   :3       NA's   :46
```

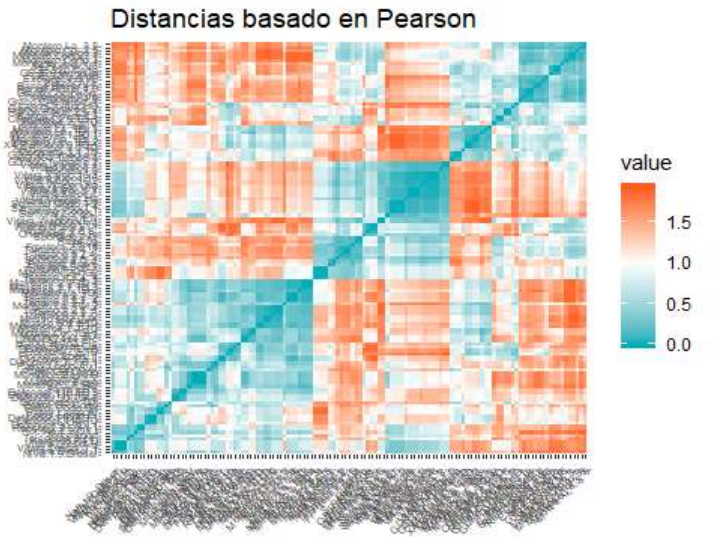
Sobre la imputación de las observaciones ausentes, se buscó en internet los pesos de los modelos “Maverick 2.7 TD GL 3” y “Maverick 2.7 TD GLS” y se añadió al dataframe. Los modelos NISSAN Patrol, ASIA MOTORS Rocsta, JEEP Cherokee 2.5 TD Jamb, Korando K4 D, UAZ Marathon, Rav4 y Niva 1.9 Diesel no cuentan con información en los campos de consumo, es así que se reemplazan con el valor medio del grupo que se asemeja en las características de peso, potencia y rpm de cada uno. En cuanto a los modelos Vitara Xaloc y TelcoLine pick-up no tienen valor en velocidad máxima. Se reemplaza con la información de internet.

Análisis Clúster

Antes de utilizar las funciones clúster, se debe verificar si la información cuenta con rasgos suficientes para la segmentación, en éste caso se utiliza el estadístico de Hopkings, si el valor del estadístico de Hopkins es cercano a 0 (menor a 0.5) se rechaza la H0, que significa que la data no está distribuida uniforme y se concluye que la data puede ser segmentada.

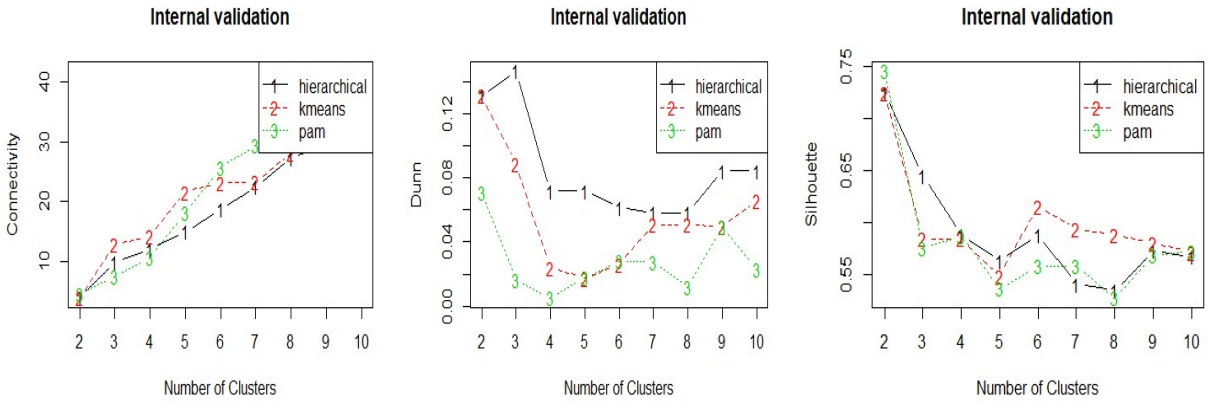
El estadístico de Hopkins resulta 0.276129, nos indica que se tiene información que puede segmentarse y por tanto podemos continuar con el análisis Clúster.

Otra forma de justificar la idoneidad de aplicar la técnica de agrupamiento en los datos es mediante el algoritmo de Evaluación visual de la tendencia de agrupación (o VAT). El gráfico de disimilitud utiliza medidas de distancia para dar una visión general de los atributos de agrupación de los datos. Por ejemplo, utilizando el método de pearson, el gráfico muestra que existen posibles clústeres.



Con el paquete clValid, se compara los algoritmos de agrupamiento considerando las medidas internas (incluyen la conectividad, el coeficiente de silueta y el índice Dunn) para para encontrar el más adecuado para los datos.

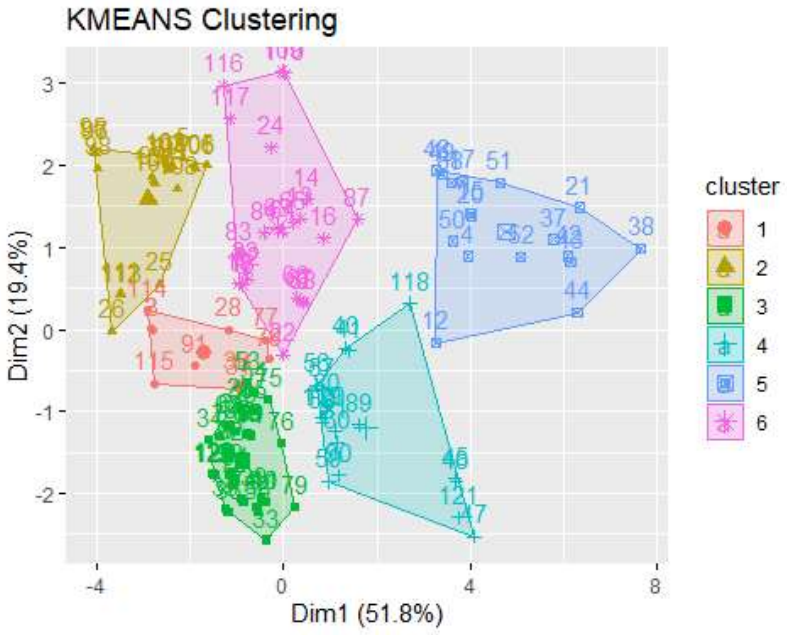
```
## Optimal Scores:
##
##              Score  Method       Clusters
## Connectivity 3.8714 hierarchical 2
## Dunn         0.1464 hierarchical 3
## Silhouette   0.7459 pam          2
```



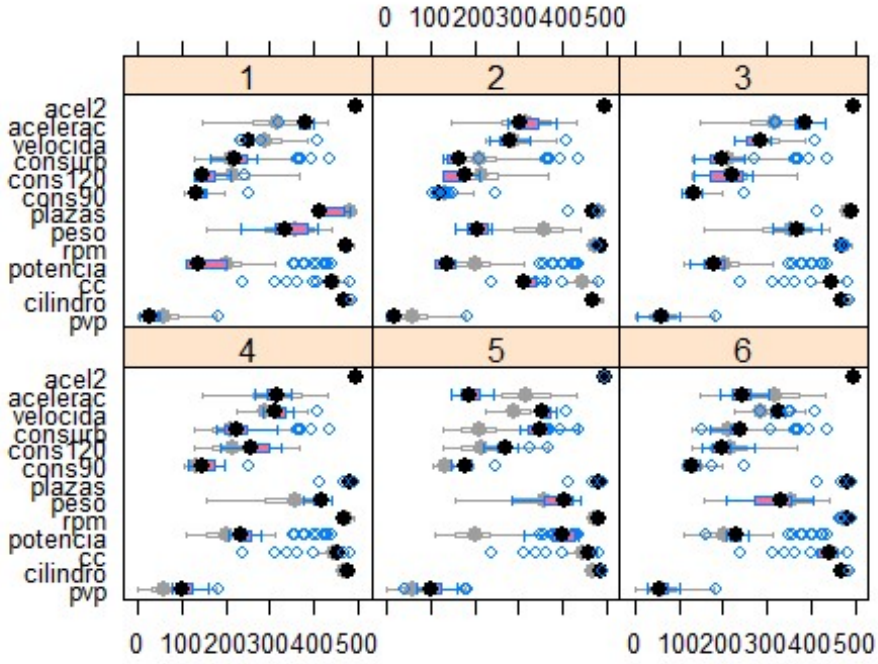
El resultado muestra que el método Jerárquico es eficiente en las métricas Connectividad e índice Dunn y Pam es más efectivo en el ancho de silueta o perfil. Sin embargo, muestran 2 o 3 clústeres como número óptimo de grupos, lo que es ineficiente en este caso porque implicaría tener grupos de más de 15 vehículos.

Algoritmos de Clustering

Se realizaron clústeres con los métodos Jerárquico y Pam, pero presentaron solapamientos a partir de 3 grupos lo cual como explicamos antes es ineficiente porque excederíamos con los coches asignados por garaje. Es así que se utilizó el método K Means y pudo agrupar si solapamiento 6 clústeres.



Finalmente, se realizó un gráfico de cajas para representar las características por clústeres.



Conclusiones

La distribución de los coches no ha sido sencilla, dado las distintas características que tiene, pero se puede concluir en lo siguiente:

- El Método KMeans, a pesar de no ser el óptimo en conservar medidas internas de cada clúster, permitió clasificar la muestra en 6 grupos.
- Si bien el jefe solicitó 10 grupos de coches, desde el punto estadístico hubiera sido correcto elegir 2 o 3 clústeres, pero finalmente desde el punto de vista del negocio y ahorro de coste es preferible 6 grupos.
- Dado que no se tiene otro criterio que la distancia geográfica, se considera el costo de transporte para distribuir los coches:
  - Punto 3 y Casa 5, se asignan los clústeres 1 y 2 debido a que su peso es menor y permitirá ahorrar en costes de transporte dado que puede realizarse por barco.
  - Punto 9, los coches del clúster 5 debido a que tienen mayor consumo y es más a propiado para ahorrar costes asignarlos más cerca de España.
  - Punto 8, de manera similar asignaremos a los coches del clúster 6 que son los se gundos de mayor consumo.
  - Punto 2, continuando con el ahorro de transporte (costo de gasolina), se deben colocar los coches del clúster 4.
  - Punto 1, se colocarán los coches del clúster 3 que son los restantes.

Bibliografía

- Clustering Pokemon, en RPubs: https://rpubs.com/Buczman/ClusteringPokemon
- Electric Vehicles Analytics - part (3/3) - Machine Learning on 80 EVs driving behaviour, en RPubs: https://rpubs.com/jianlee/ev03_clustering

Anexo

El archive Rmd con el analisis detallado se encuentra en el repositorio: https://github.com/Mayragn387/MasterDS2019/tree/master/T%C3%A9cnicas%20de%20Agrupaci%C3%B3n%20y%20de%20Reducci%C3%B3n%20de%20la%20Dimensi%C3%B3n/Research