Los coches del jefe: clustering

Sara Bengoechea Rodríguez

Introducción

El presente informe tiene como objetivo estudiar el número adecuado de grupos en los que dividir 125 vehículos de un coleccionista y asignarlos mediante un criterio de distancia en un máximo de 10 residencias que este posee.

Tras el análisis exploratorio y la eliminación de variables del anterior informe, partimos desde el dataset resultante y generamos los estadísticos principales del mismo.

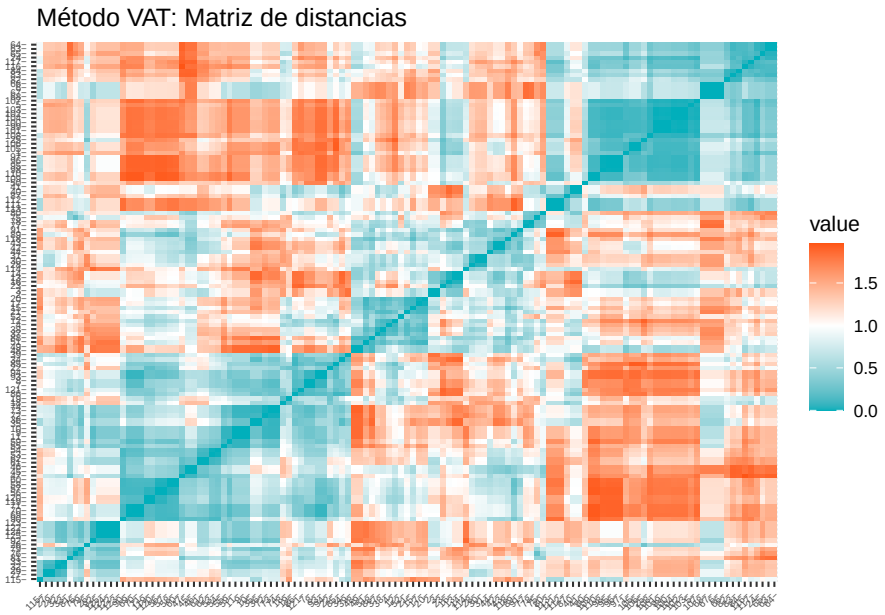
```
##             Min    Med   Mean    SD    Max
## cilindro    4.0    4.0    4.6   1.0    8.0
## cc       1298.0 2497.0 2569.8 691.5 5216.0
## rpm      3600.0 4500.0 4670.9 716.0 6500.0
## peso      930.0 1746.0 1674.5 330.9 2320.0
## plazas      2.0    5.0    5.2   1.4    9.0
## cons90      6.6    8.6    8.9   1.4   13.7
## cons120     8.4   12.2   12.2   2.2   18.5
## consurb     8.1   12.0   12.6   2.8   22.1
## velocida  120.0  146.5  150.5  16.5  196.0
```

Estudio de la distancia de las observaciones

Para proceder al análisis cluster, previamente debemos estudiar la distancia entre las observaciones para saber si es adecuado dicho análisis. Las dos técnicas utilizadas para ello son: el estadístico de Hopkins y el método VAT.

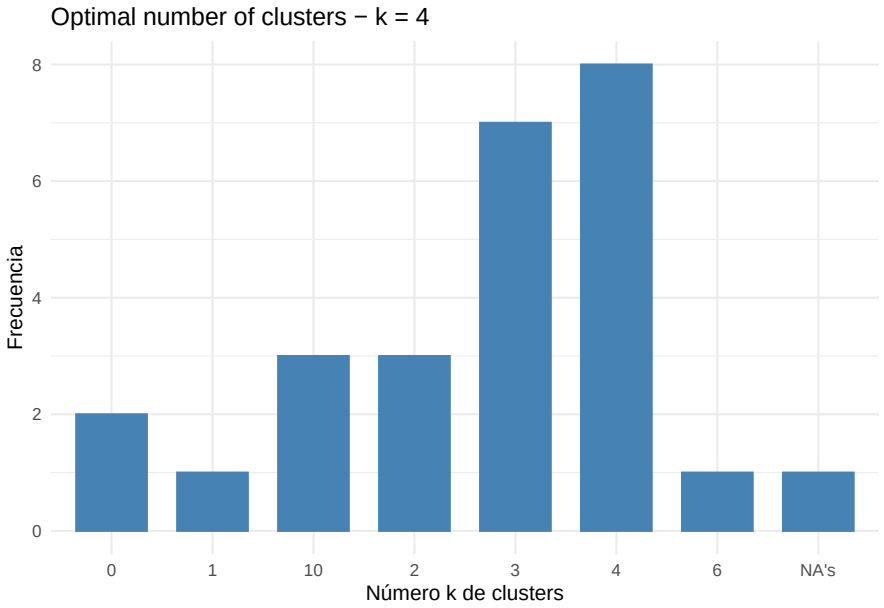
Mediante el estadístico de Hopkins estudiamos la distribución de las observaciones. El valor obtenido es 0.16 lo que significa, que por su proximidad a cero, rechazamos la hipótesis de aletoriedad y avalamos la presencia de dos o más clusters en el conjunto de observaciones.

Esta misma conclusión se puede estudiar de manera visual mediante el método VAT, donde el color azul implica poca distancia entre las observaciones y el color rojo, lejanía.



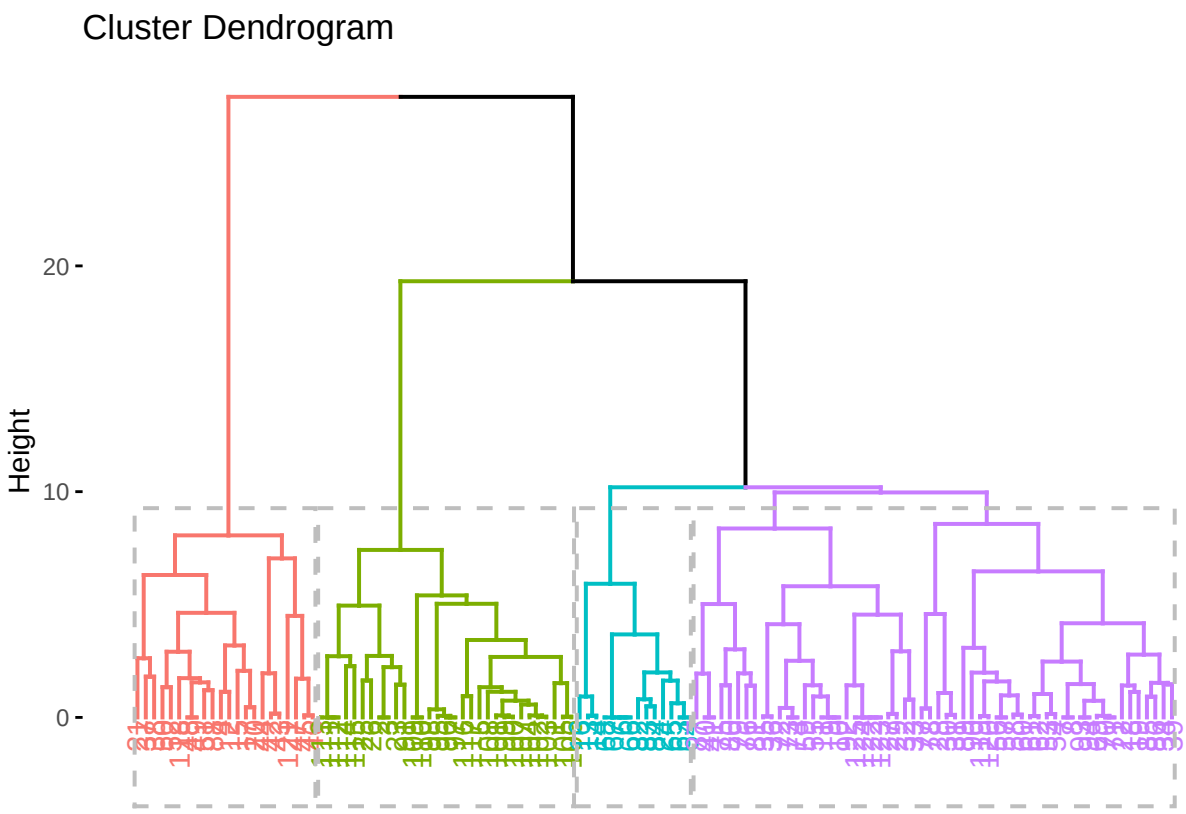
Número óptimo de clusters

Para conocer cuál es el número óptimo de clusters utilizamos el paquete NbClust. Este combina los distintos número de clusters, medidas de distancia y métodos de clustering, para determinar el número óptimo de clusters. Como se ve representado a continuación, el número óptimo es 4.



El método de segmentación seleccionado es un método jerárquico aglomerativo: hclust. Esta decisión ha sido tomada ya que, a pesar de que el objetivo final es agrupar las observaciones en un máximo de 10 grupos, es preferible un método que agrupe las observaciones en virtud de su similaridad, sin grupos preestablecidos y obtener un número óptimo.

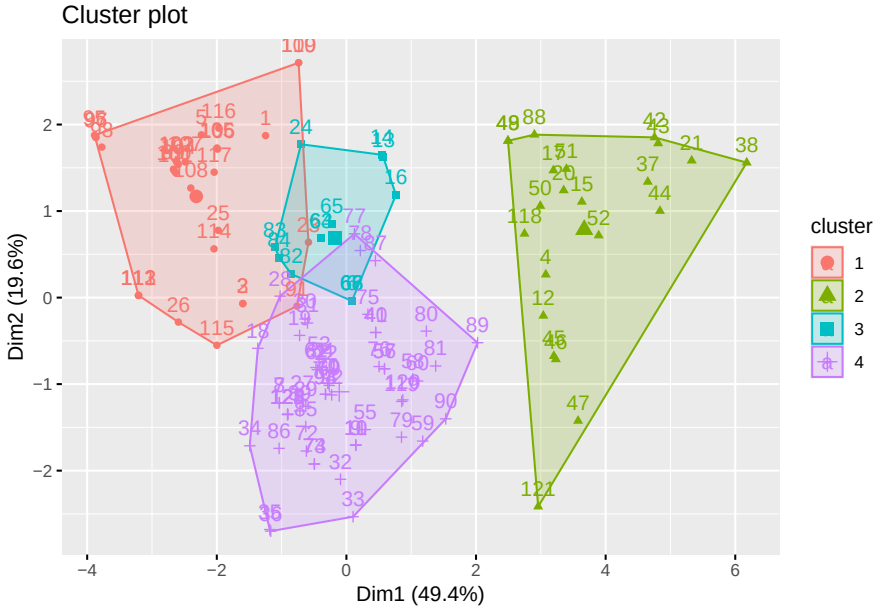
A continuación se muestra un dendograma mediante hclust y método ward con cuatro clusters.



Este sería entonces el tamaño de los grupos y la anchura de la silueta:

```
##   cluster size ave.sil.width
## 1       1   31          0.26
## 2       2   22          0.29
## 3       3   14          0.39
## 4       4   58          0.14
```

En un espacio de 3 dimensiones (al añadir el color), los clusters se representarían de la siguiente manera:



Conclusiones:

Dados los 4 grupos definidos y que la capacidad máxima de las residencias es de 15 vehículos , la distribución será de la siguiente manera:

- **Cluster 1**: Consta de 31 observaciones que se dividirá en dos grupos de 15 en las dos localidades de París. La observación restante es la observación número 23 que será agrupada con el cluster 3, ya que atendiendo a la representación anterior, podemos ver que hay mucha cercanía al centroide en las dos primeras dimensiones.
- **Cluster 2**: Este cuenta con 22 observaciones, se dividirá en 2 grupos a partes iguales y estarán en las residencias de Suiza.
- **Cluster 3**: Los 14 vehículos de este (más la observación 23 del cluster 1) estarán en la residencia de La Rochelle.
- **Cluster 4**: Los 58 vehículos restantes se repartirán en dos grupos de 14 y otros dos grupos de 15 y sus residencias serán las tres residencias cercanas a Mónaco y Niza y por último, la residencia próxima a la frontera de Andorra.

La única residencia de la cuál no se va a hacer uso es la que está situada en la isla de Córcega ya que el transporte de los coches a esta puede ser de gran coste. De esta manera, todos los vehículos quedarían distribuidos de manera eficiente en virtud de su similaridad y distancia en 9 residencias distintas.