Tarea Clúster - Los coches del jefe 2

Diego Senso González

2/12/2020

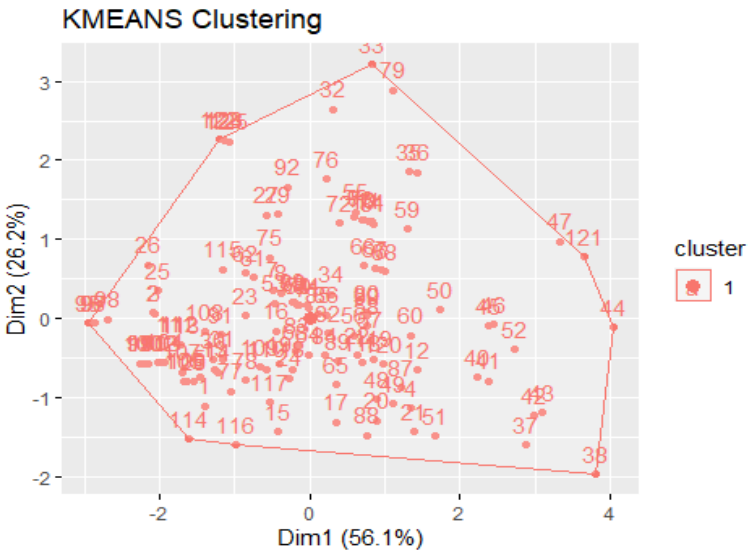
Table of Contents

Objetivo

El objetivo del presente informe es estudiar el número adecuado de grupos en los que dividir la colección de coches existente.

Primera observación

De cara a realizar una primera aproximación, es posible observar en un plano dónde se sitúan cada una de las observaciones del dataset. Esto puede ofrecer una primera visión de las distancias entre los puntos, viendo cuáles están cercanos entre sí (observaciones que se asemejan) y en qué zonas las distancias se incrementan (observaciones con poca relación o poco similares).

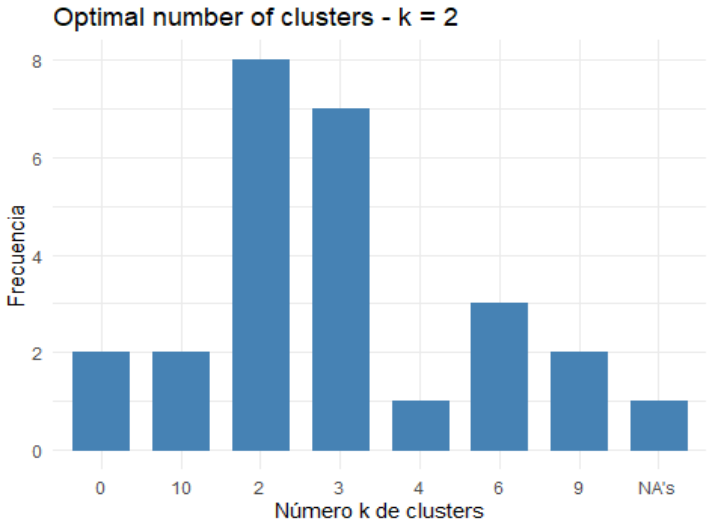


A simple vista, parece haber una alta concentración de datos en la parte centro-izquierda del plano, siendo menos los que se agrupan a la derecha. Esto podría significar una clara distinción entre unos y otros vehículos, algo que podría ser importante para la configuración posterior de grupos.

Adicionalmente, este plano recoge los vehículos colocados en dos dimensiones. Como se puede observar en los ejes, la primera dimensión permite explicar más de la mitad de los casos (56.1% concretamente). Añadiendo la segunda dimensión, es posible estar explicando algo más de un 82%. Es una cantidad a tener en cuenta, ya que contar con dos dimensiones reduce mucho la dimensión y a cambio se logra explicar una cantidad considerablemente buena.

Número óptimo de clusters (30 índices)

Para resolver el problema de determinar el número óptimo de clústers en los que agrupar todas las observaciones disponibles, la librería "NbClust" ofrece en un mismo contraste 30 índices diferentes que estudian esta cuestión. Se trabajará con un mínimo de clusters igual a 2 (por crear una pluralidad de grupos), y un máximo igual a 10 (que son los garajes que el jefe posee).



A la vista de los resultados, el número óptimo de clusters es 2, seguido de cerca de 3. Bastante más atrás está la opción de crear 6 clusters. Es un resultado comprensible, parece que al crear el número de grupos mínimo la agrupación es mejor y obtiene menos fallos. Al hacer dos grupos, gráficamente se observa que no hay solapamientos y las observaciones que entran en uno y otro grupo quedan bien definidas. Sin embargo, el jefe posee 10 garajes, por lo que no parecería lógico intentar almacenar los coches en solo dos. Además, los garajes cuentan con una capacidad limitada (15 plazas). Dadas estas circunstancias, se va a tratar de crear otra agrupación diferente con mayor número de clusters, pese a que el agrupamiento sea de peor calidad y puedan producirse solapamientos o errores en la inclusión de alguna observación dentro de un grupo.

Cluster pam

Se prueba a realizar el cluster con la función de abreviación "pam". Se visualizan todas las posibles agrupaciones desde los dos hasta los diez grupos. No parece ser el mejor en este caso ya que desde la realización de 2 grupos ya aparecen solapamientos.

Cluster kmeans

Utilizando el cluster bajo la función "kmeans", la agrupación parece ligeramente mejor, y los solapamientos comienzan a aparecer a partir de los 5 grupos.

Cluster jerárquico

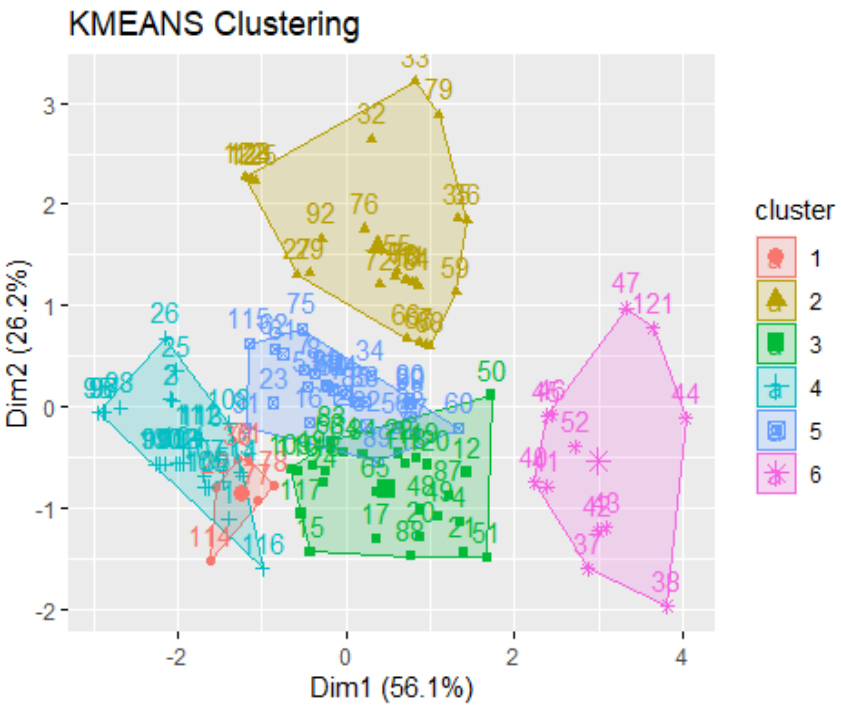
Utilizando el cluster jerárquico, los solapamientos aparecen a partir de los dos grupos. En el momento de aumentar el número de estos, comienza a haber más de dos clusters que se solapan entre sí, por lo que no parece el mejor método de agrupación.

Otros métodos

Pese a no representarlos, se han probado el resto de funciones de agrupación como la "clara", "hclust" y "fanny" entre otras. Todas ofrecen solapamientos con menos grupos que el kmeans, por lo que se decide realizar la agrupación con esta función. Adicionalmente, otras distancias como la "manhattan" tampoco han ofrecido mejores resultados.

Número de grupos escogidos

Como se ha comentado anteriormente, lo recomendado por los 30 índices era agrupar en dos o tres clusters. Sin embargo, dadas las características concretas de la situación no parecía lo más adecuado. Con 5 grupos comienzan a aparecer los primeros solapamientos. Observando el test de los 30 índices realizado al comienzo, 6 clusters era la tercera opción más recomendada. Se procede a visualizar la creación de 6 grupos:



Con esta configuración, aparecen solapamientos entre los grupos 1-4 y 3-5. Sin embargo, los otros dos grupos aparecen muy bien definidos. Pese a existir estos solapamientos, estos son en observaciones puntuales. Se procede a visualizar la silueta del cluster creado:

Se puede observar que esta configuración, pese a no obtener un buen average, no comete muchos errores a la hora de agrupar. Probando con otro número de clusters el "average" no mejora significativamente, y en varios casos se cometen más errores. Adicionalmente, se ha tenido en cuenta la distribución geográfica de los garajes para elegir el número de clusters a realizar. Por último, podemos observar el número de coches incluidos en cada uno de los 6 grupos creados, a fin de localizarlos geográficamente en el apartado de conclusiones:

```
##
##   1   2   3   4   5   6
##   6  26  27  26  28  12
```

Conclusiones

Como se puede observar, existen solapamientos claros con la creación de 6 grupos. Sin embargo, se ha decidido este proceder en este caso ya que pese a colocar mal algunas observaciones, podremos distribuir correctamente los coches por diferentes localizaciones, no dejando muy saturadas unas y muy vacías otras, lo que habría ocurrido en caso de haber seleccionado un bajo número de grupos. Además, al tener grupos con muchas observaciones habría que haber separado geográficamente coches con características similares o del mismo grupo.

Observando el mapa de localizaciones de los garajes, las distribución sería la siguiente:

- Los grupos 3 y 5 (que se solapan en alguna observación y en total son los más numerosos) se localizarían en las propiedades en el sureste de Francia (zona de Mónaco y Cannes) y en la isla de Córcega. Estarían muy próximos geográficamente y llenarían 55 plazas del total de 60 que tendrían los cuatro garajes juntos.

- El grupo 2, que a simple vista parece más disperso en las observaciones (26 coches) se repartiría entre los dos garajes del este, ya en territorio suizo.

- El grupo 4 (26 coches) se situaría en los dos garajes del norte del mapa (cercanos a París). Cubriría 26 plazas de un total de 30 entre ambos garajes.

- El grupo 6 (12 coches) que está más distanciado en la representación de clusteres anterior, se situará en el garaje del oeste de Francia, cerca de La Rochelle.

- El grupo 1 (el menos numeroso con 6 vehículos) se lleva a Andorra.

De esta forma, todos los garajes del jefe quedan ocupados. A nivel estádistico, lo más recomendable parecía inclinarse por realizar dos o tres grupos. Sin embargo, ante el caso presentado parecía major alternativa tratar de aumentar el número de grupos pese a perder algo de calidad en la agrupación de vehículos. En caso contrario, habría supuesto desperdiciar varios de los garajes que el jefe posee y dejarlos vacíos. Trabajar con un número tan reducido de grupos no habría permitido diferenciar entre los coches, agrupando en clusters formados por coches más diferentes entre sí de lo que se puede obtener aumentando a 6 grupos.