Tarea Clúster - Los coches del jefe

Diego Senso González

24/11/2020

Contents

Objetivo

El objetivo de la presente práctica es realizar una primera aproximación hacia el dataset propuesto, escogiendo las variables idóneas para realizar el análisis que se completará más adelante.

Explicación del dataset

Para la presente práctica, se cuenta con un dataset en el que cada observación (de un total de 125) representa una combinación de un coche adquirido junto con una serie de características referentes a ese vehículo.

Carga de librerías y el dataset

Se cargan las librerías necesarias, además del dataset. En este caso viene en formato sav, por lo que la función a utilizar para leer los datos será diferente. Tras leer el archivo, se pasa a convertirlo a formato csv para no encontrar problemas en su posterior tratamiento.

Observamos el dataframe gracias a la función "skim" (no se ha incluido en el informe), de cara a obtener una idea de cómo se distribuyen y se comportan las variables. Esto facilitará la elección de variables posterior, ya que observar el comportamiento y valores de cada columna ofrecerá la lectura de que algunas de ellas no son interesantes para incluirlas.

Selección de variables a utilizar

De cara a agrupar los vehículos más adelante, se deben elegir una serie de variables consideradas como importantes. Estas se han seleccionado observando cuáles podrían ser diferenciadoras dentro del dataset. Tras haber realizado una aproximación a los datos y ver cómo se distribuyen, las variables que se ha decidido seleccionar son las siguientes:

- **marca**: variable categórica que expresa la marca del fabricante del vehículo en formato numérico. Toma valores entre 1 y 17. Interesante por poder organizar coches en función de marcas.
- **pvp**: precio de venta al público (en euros). Interesante para saber cuáles son más caros o baratos para almacenarlos en un lugar u otro.
- **peso**: peso del vehículo (en kilogramos). Bueno para temas logísticos, transporte o de almacenaje de los vehículos.
- **plazas**: número de plazas del vehículo. Buena para conocer el tamaño de los coches.
- **velocida**: velocidad máxima que el vehículo alcanza (km por hora). Para observar si tienen un estilo más deportivo o familiar, entre otras cosas.

Además de por los valores de estas variables, se han escogido en base a que parecen características lógicas en las que se podría pensar si la finalidad es clasificar grupos de vehículos que se parezcan entre sí y se diferencien con respecto a otros grupos.

Variables descartadas

- **modelo**: se trata de una variable de tipo texto que tiene 111 valores únicos, por lo que no parece adecuada para realizar una agrupación. Además, en el caso de desear agrupar coches por fabricante ya es útil la variable "marca", que sí se ha seleccionado.
- **cilindro**: este caso es similar pero opuesto al anterior, ya que con tan sólo valores de cilindro entre 4, 6 y 8 no es clara la diferenciación.
- **potencia**: la potencia del vehículo medida en caballos. Se ha descartado al finalizar el análisis dada su estrecha asociación la variable "velocidad".
- **cc**: centímetros cúbicos. En esta sí que existe mayor dispersión de los datos, pero existen otras variables más claras que explican algo similar como la potencia. Por ello, en un ejercicio de reducción se prefiere dejar fuera esta variable.
- **rpm**: revoluciones por minuto. Mismo caso que con la anterior varible, además de que los vehículos con una mayor velocidad tendrá ya mayor número de rpm, con lo que ya estaría explicada en parte.
- **cons90**, **cons120**, **consurb**: consumo del vehículo en 90km/h, 120km/h y en contexto urbano. Pese a ser variables que ofrezcan una información que podría ser interesante, no se ha considerado como del todo relevantes dado que hay otras variables según las cuales parece más lógico agrupar una serie de coches si el objetivo va a ser distribuirlos en diferentes garajes.
- **acelerac**: tiempo en segundos que el vehículo tarda en ir de 0 a 100 km/h. No se ha seleccionado porque cuenta con una gran cantidad de NAs.
- **acel2**: se trata de una categórica que explica la aceleración. Al tener dos valores no parece tener demasiada calidad explicativa como para incluirla.

Procedemos a selecionar las variables para el análisis.

Tratamiento de los datos

Pasamos a sustituir los valores NA por la media de cada columna, de cara a no perder ninguna observación, ya que de eliminarlas estaríamos perdiendo registros de coches que se deben agrupar posteriormente. Tras esto, observamos las primeras filas del dataframe resultante.

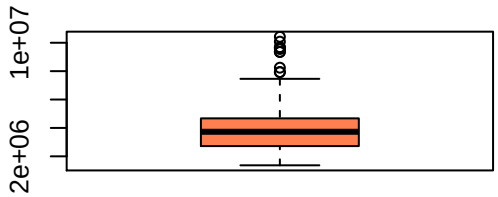
```
## # A tibble: 6 x 6
##   marca           pvp potencia  peso    plazas velocida
##   <fct>         <dbl>    <dbl> <dbl> <dbl+lbl>    <dbl>
## 1 ASIA MOTORS 2274590       85  1220     4 [4]      160
## 2 ASIA MOTORS 2161913       72  1270     4 [4]      130
## 3 ASIA MOTORS 2274590       72  1270     4 [4]      130
## 4 CHEVROLET   4745000      193  1915     5 [5]      180
## 5 DAIHATSU    2693460       95  1250     4 [4]      150
## 6 FORD        4461000      124  1750     7 [7]      160
```

Descripción de los vehículos adquiridos

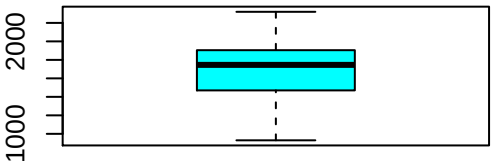
Gracias a la función "summary", podemos observar la distribución de las diferentes variables seleccionadas, para coger una idea de cómo se comportan.

```
##       pvp              potencia          peso          plazas
##  Min.   : 1367000   Min.   : 64.0   Min.   : 930   Min.   :2.000
##  1st Qu.: 2721000   1st Qu.: 95.0   1st Qu.:1470   1st Qu.:4.000
##  Median : 3730000   Median :112.0   Median :1746   Median :5.000
##  Mean   : 4004459   Mean   :117.1   Mean   :1675   Mean   :5.184
##  3rd Qu.: 4675406   3rd Qu.:125.0   3rd Qu.:1905   3rd Qu.:5.000
##  Max.   :10419200   Max.   :225.0   Max.   :2320   Max.   :9.000
##     velocida
##  Min.   :120.0
##  1st Qu.:140.0
##  Median :148.0
##  Mean   :150.6
##  3rd Qu.:160.0
##  Max.   :196.0
```

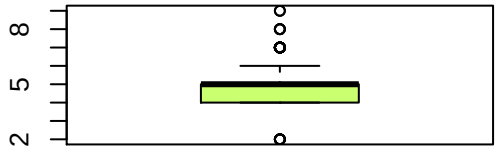
Para obtener una idea de los coches adquiridos es fundamental, pues ofrece entre otros los valores medios, mínimos y máximos de todas las observaciones, esto es, de los coches con los que contamos y pretendemos ordenar próximamente de forma eficiente y consistente. A continuación se realizarán gráficos boxplot. Así se pueden observar cómo se comportan la mediana, los cuartiles y los outliers existentes en cada una de las variables numéricas. En general, se puede obtener una rápida visión de los valores que suelen tomar los coches comprados en cada uno de estos aspectos.



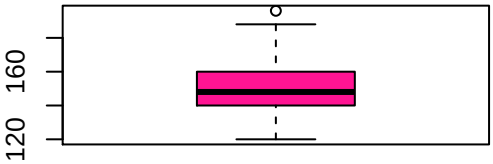
Precio de venta al público



Peso en kilogramos



Número de plazas



Velocidad en Km/h

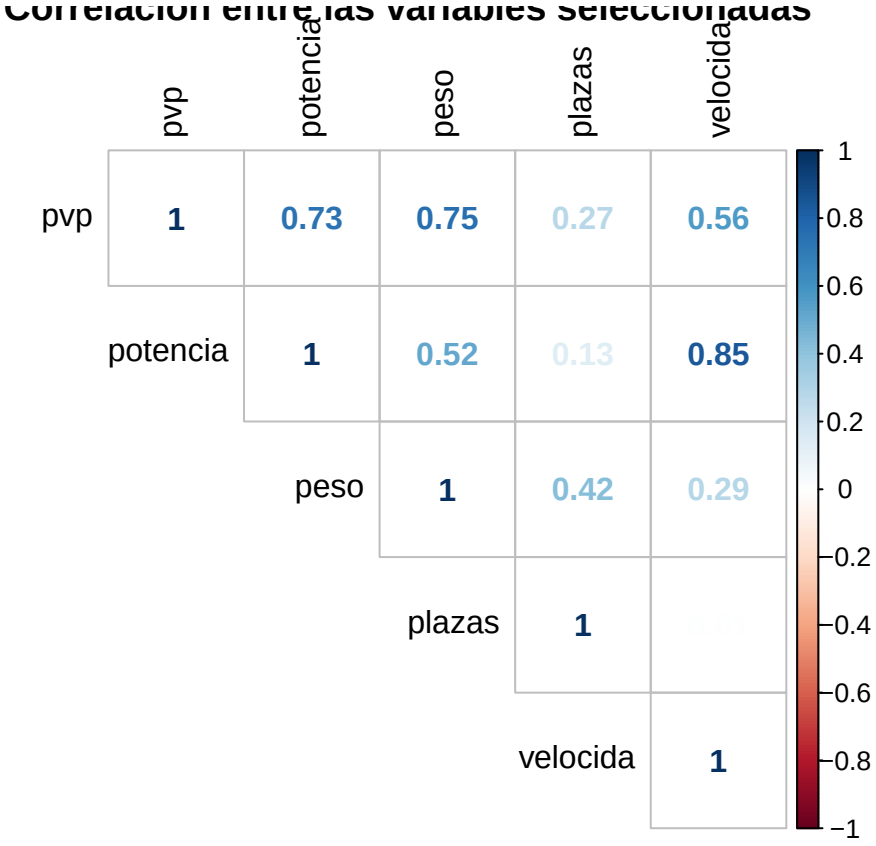
Posteriormente, se pueden observar las distancias entre las variables escogidas. Cuanto más cercano es el valor a cero, más cercanas estarán dos variables entre sí. Mientras, cuanto más elevado es el dato más distanciadas se encontrarán.

```
##           pvp potencia peso plazas velocida
## pvp      0.00     0.27 0.25   0.73     0.44
## potencia 0.27     0.00 0.48   0.87     0.15
## peso     0.25     0.48 0.00   0.58     0.71
## plazas   0.73     0.87 0.58   0.00     0.99
## velocida 0.44     0.15 0.71   0.99     0.00
```

Cercanas parecen entre sí variables como “velocidad” y “pvp”. Parece una asociación lógica en principio, ya que los coches más rápidos suelen pertenecer a una categoría de automóviles de lujo y por ende su precio es más elevado. También parecen estar cercanas “peso” y “pvp”, una asociación interesante, pues no parece tan obvia como en el caso anterior.

En caso de variables lejanas entre sí, destaca el número de plazas con respecto a la velocidad y la potencia, entendiendo que normalmente los coches que sobresalen en estos dos últimos aspectos no suelen albergar a muchos ocupantes.

Continuamos con la matriz de correlaciones:



Se puede observar una muy alta correlación entre las variables "potencia" y "velocidad", la cual destaca por encima del resto. El resto de correlaciones son positivas pero no presentan valores muy elevados.

Tras esta visión y dada la estrecha relación entre "potencia" y "velocidad", se procede a eliminar también una de estas dos variables, debido a que están explicando cosas muy parecidas de los vehículos. Se elimina la "potencia" dado que resulta más explicativo y claro a priori agrupar por velocidad, pues los caballos de potencia son algo menos visuales.

Conclusiones

No todas las variables con las que cuenta el dataset son fundamentales a la hora de agrupar. Por ello, se ha decidido eliminar una parte de ellas del análisis. Las decisiones han sido tomadas siguiendo los valores de las variables y configuración del dataset, las conclusiones extraídas de las diferentes medidas aplicadas, y por último la lógica sobre cuáles son los atributos que tiene sentido tener en cuenta a la hora de ordenar un grupo de coches. Finalmente, las variables escogidas y ya tratadas son: "marca", "precio", "peso", "plazas", "velocidad".