Los coches del jefe

Isabel Afán de Ribera

24 noviembre 2020

Resumen Ejecutivo

El dueño de un family office ha comprado 125 vehículos de todo terreno clásicos con distintas características, las cuales ha recopilado en un documento que se nos ha presentado. Como trabajadores de este family office se nos ha pedido realizar un análisis sobre cada una de las características de estos vehículos para que nuestro cliente pueda distribuirlos de forma eficiente en 10 propiedades distintas. En el presente informe se recogen los resultados obtenidos del análisis exploratorio sobre las características de los vehículos y la selección de variables más relevantes para una futura propuesta de agrupación y distribución de los vehículos.

1. Objetivo

El objetivo del presente trabajo es realizar un análisis sobre las características de vehículos todo terreno para facilitarle a nuestro cliente una futura propuesta de distribución en grupos con características similares entre sí y distintas con el resto.

2. Descripción del dataset

El dataset objeto de estudio cuenta con 125 observaciones relativas a los distintos vehículos de los que el cliente es propietario y 15 variables.

- marca: Marca del todo terreno
- modelo: Modelo de todo terreno
- pvp: Precio (pesetas)
- cilindro: Número de cilindros
- cc: Cilindrada (cm cúbicos)
- potencia: Potencia (CV)
- rpm: Revoluciones por minuto
- peso: Peso en kg plazas: Número de plazas cons90: Consumo 90 km/h cons120: Consumo 120 km/h consurb: Consumo urbano velocida: Velocidad máxima
- acelerac: Aceleración de 0 a 100
- acel2: Tiempo de aceleración, 2 grupos 1(Menor a 10 seg) 2(Mayor a 10 seg)

3. Desarrollo

3.1. Selección de variables

En una primera aproximación, y partiendo de la lógica de que nuestro cliente quiere distribuir sus vehículos según características técnicas se ha decidido prescindir de las variables modelo, marca y precio. De las dos primeras cabe decir, tras analizar los datos, que existen 109 modelos diferentes y 17 marcas distintas con lo cual se entiende innecesaria la agrupación de vehículos según tales características, pues resulta muy complicado dividir en solo 10 grupos. Igualmente, teniendo en cuenta que nuestro cliente tiene como objetivo la repartición de sus coches a modo de colección y no de venta se entiende que puede prescindirse de la variable precio.

En el caso de las características técnicas, a priori solo resulta discutible la utilidad de la variable "acelerac" pues no se conocen los datos para esta característica de 46 de los 125 vehículos, representando más de la mitad de los valores nulos que contiene el dataset. Por tanto, también vamos a prescindir de ella.

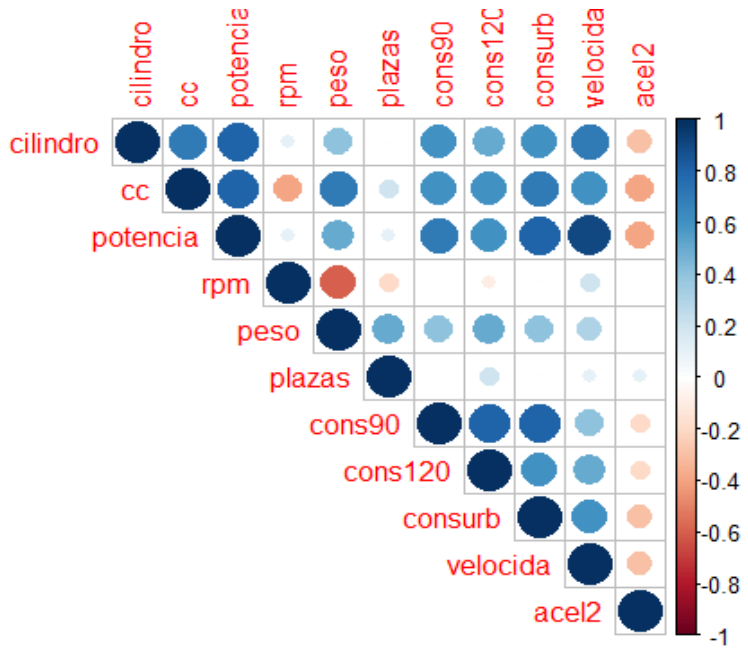
Además, podría discutirse la necesidad de tener en cuenta tanto la variable "cilindro" como "cc" pues ambas son características relativas al motor siendo la segunda (cilindrada en cm) dependiente de la primera (número de cilindros) pues la cilindrada de un motor es el volumen unitario -de cada uno de sus cilindros- multiplicado por el número de cilindros. Por lo tanto, si conocemos la cilindrada de los vehículos implícitamente conocemos también el número de cilindros y podemos prescindir de ella.

Con respecto, a las variables "cons90", "cons120" y "consurb" puede decirse que todas se refieren a características del consumo según velocidad siendo innecesario utilizar las tres, por ejemplo, podría eliminarse la variable cons90 o cons120 ya que ambas van referidas al consumo en carretera.

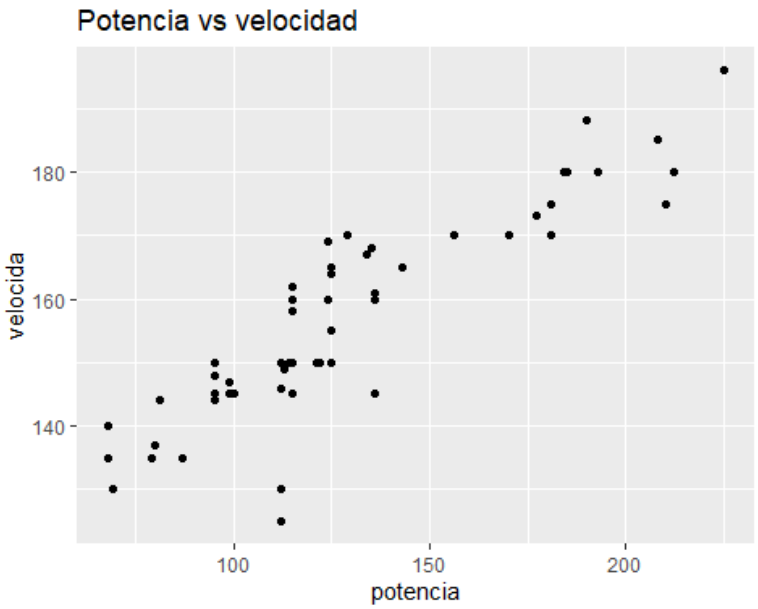
3.2. Análisis exploratorio

Pasando a analizar relaciones entre variables, primero se puede visualizar un gráfico general de correlaciones entre todas las variables de la que puede extraerse que la característica "acel2" es la que mayor correlación negativa o nula presenta, lo cual puede ayudar a diferenciar grupos en el análisis cluster. Igualmente, las variables "rpm" y "plazas" tienen poca relación con las demás.

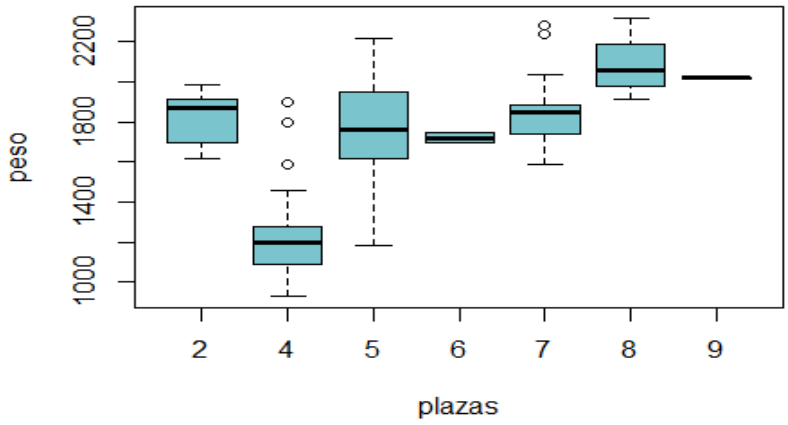
Por el contrario, las variables "potencia", "cilindro" y "cc" tienen una alta asociación con las 3 variables relativas a consumo y con velocidad, siendo la que mayor correlación positiva tiene la variable velocidad con potencia, a medida que aumenta una lo hace también la otra.



Como nos ha mostrado la *matriz de correlaciones*, si analizamos por separado mediante un *diagrama de dispersión* la relación entre potencia y velocidad se ve de forma muy clara como a medida que aumenta la potencia, y como es lógico, la velocidad también lo hace. Se conoce pues que a mayor potencia tenga el vehículo más velocidad será capaz de alcanzar, por tanto, se podría llegar a prescindir de alguna de ellas.

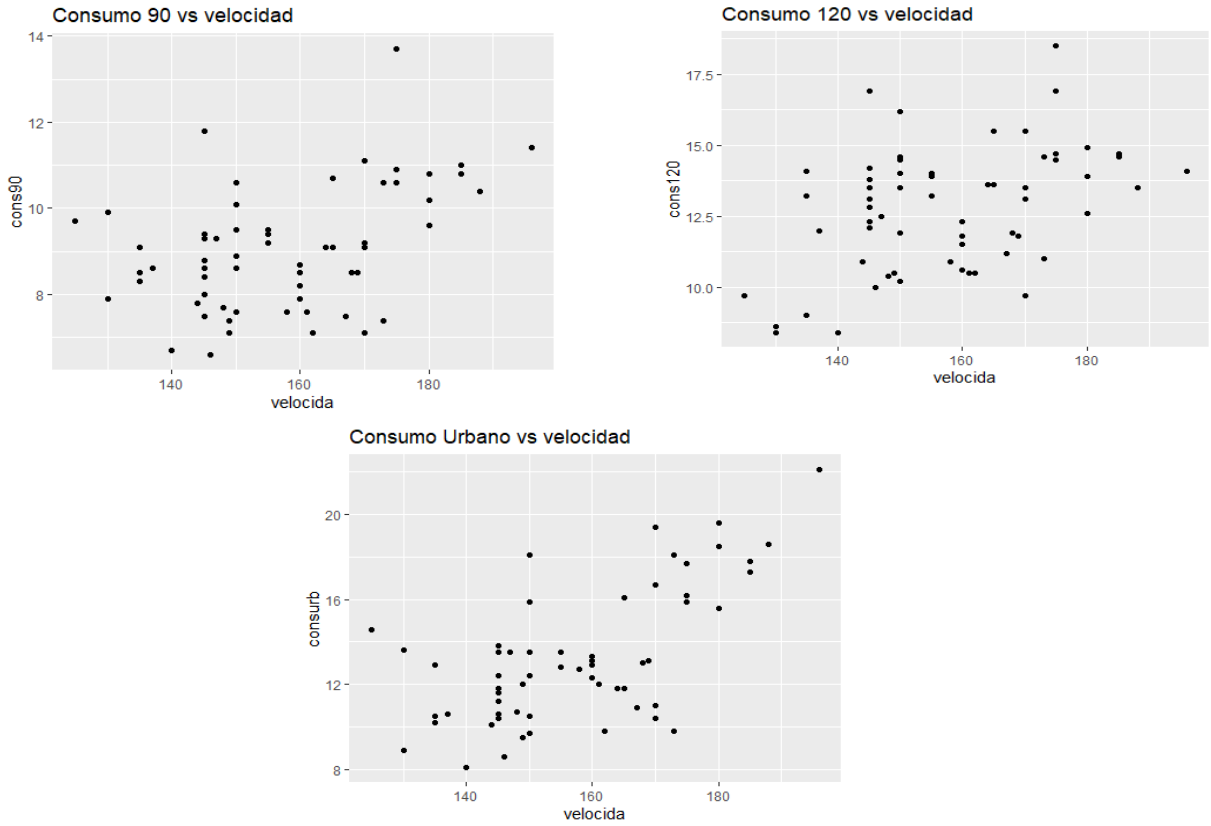


En el caso de las variables número de plazas y peso, puede llegar a pensarse, en un primer momento, que tienen una alta correlación positiva pues a mayor número de plazas más grande será el vehículo y con ello más pesado. Sin embargo, en este *diagrama de cajas* esta asunción no parece ser tan clara ya que se aprecia que los vehículos de 2 plazas pesan más que los de 4 e incluso que vehículos de 5 plazas pesan más que de 6 o 7 y de 8 más que los de 9. Por tanto, es adecuado mantener ambas variables.



Otras variables que por lógica se entienden muy relacionadas son el consumo con la velocidad. En los siguientes *diagramas de dispersión* se comparan las variables consumo a 90 km/h, consumo a 120 y consumo urbano con la variable velocidad. Como puede observarse, de manera general a medida que aumenta la velocidad aumenta el consumo, pero no siempre es así.

Por ejemplo, puede verse que en una velocidad de entre 140 y 170 hay vehículos que consumen bastante más que otros, lo cual dependerá de otras características. Si bien, puede concluirse que no es necesaria la inclusión de los tres tipos de consumo pues tienen relaciones muy parecidas con las demás variables como ya se vio en el *gráfico de correlaciones*, podría eliminarse la variable "cons90", por ejemplo, ya que es muy similar a "cons120" pero presenta más relaciones con el resto de variable con lo cual puede ser de menos ayuda que "cons120" a la hora de diferenciar en grupos.



4. Conclusiones

Las variables modelo y marca pueden ser descartadas en un posterior análisis cluster pues con ellas no es fácil agrupar los vehículos en grupos homogéneos, especialmente la característica modelo ya que hay hasta 109 modelos distintos.

La variable "cons90" también se ha eliminado al igual que "cilindro" pues ambas pueden ser explicadas por otras variables resultando redundante su información. También, la variable "precio" ya que el cliente quiere realizar la asignación por grupos a modo de colección y no de venta. Igualmente, puede descartarse la variable "aceleración" al no disponerse de una gran parte de sus datos.

Es importante destacar que a pesar de que hay variables que en un primer momento pueden parecer muy relacionadas, como el "peso" y la "velocidad", no siempre lo están y pueden ayudar en la distinción de grupos por ello nos quedamos con ambas.

Referencias

- El cilindro y la cilindrada. Disponible en: https://www.motor.es/que-es/cilindro

- Velocidad y Aceleración Vs. Consumo de Gasolina. Disponible en: https://blog.genesis.es/velocidad-y-aceleracion-vs-consumo-de-gasolina/