Los coches del jefe

Sara Bengoechea Rodríguez

11/23/2020

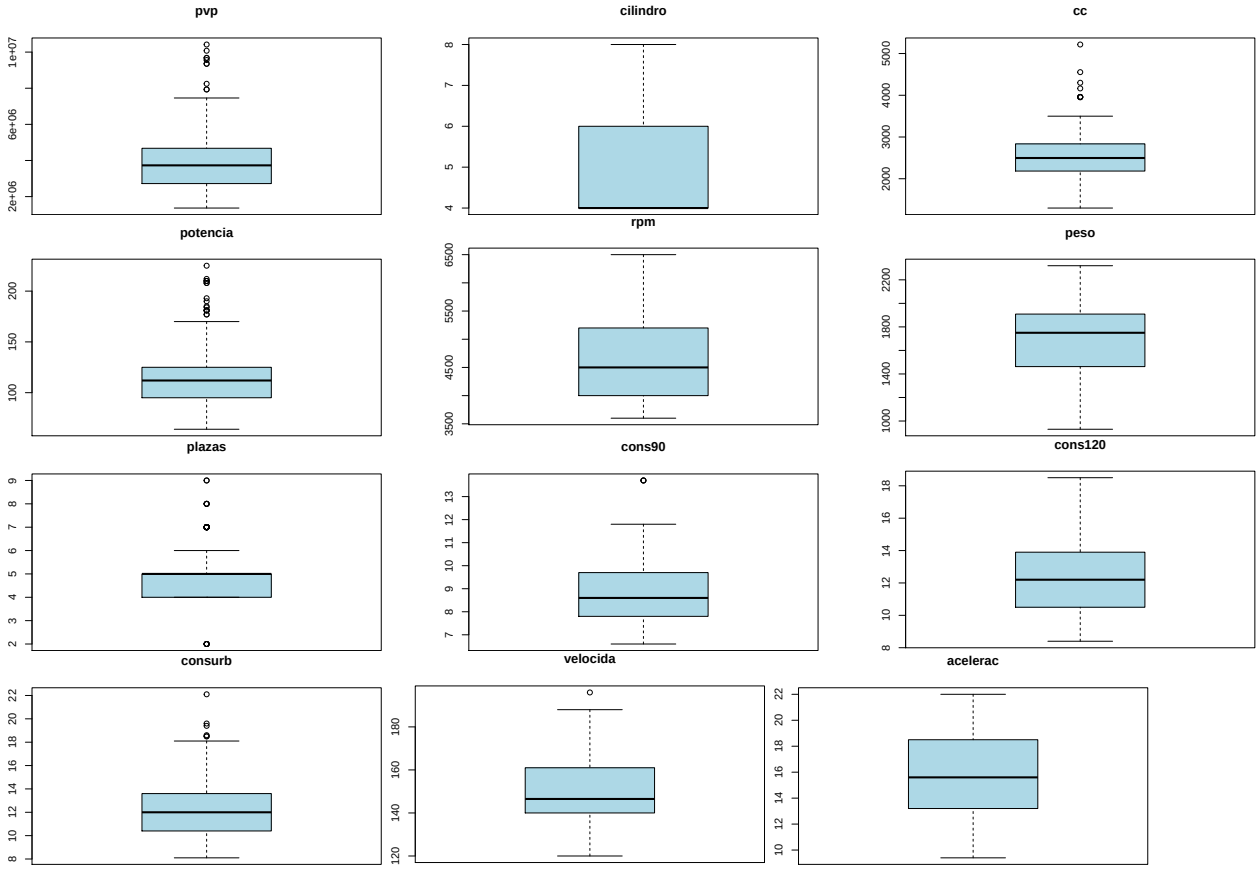
Introducción

El presente informe tiene como objetivo estudiar las características más relevantes de un dataset formado por de 125 vehículos clásicos de un coleccionista, para posteriormente asignar dichos vehículos a 10 lugares distintos, agrupados de manera homogénea.

Este dataset contine 125 observaciones y 15 variables que aportan información sobre el tipo de vehículo.Entre las 15 variables mencionadas, 3 de ellas son de tipo factor (categóricas) y el resto numéricas.

Tratamiento de valores nulos

En total hay 83 valores nulos, y dado que el número de valores nulos es alto en comparación con el tamaño de nuestro dataset, debemos tratar dichos valores. Para ello, sustituiremos los valores nulos por la media o por la mediana, en función de si existen outliers o no en dicha variable. Para ello, primero realizamos un boxplot de todas las variables numéricas y vemos el número de NAs que hay en cada columna.



Mediante los boxplots, estudiamos la concentración de los valores alrededor de la media y podemos observar los outliers. Ya que la mayoría de variables tienen outliers, utilizaremos la mediana en los valores nulos excepto en las variables que no tienen valores atípicos ("peso" y "cons120").

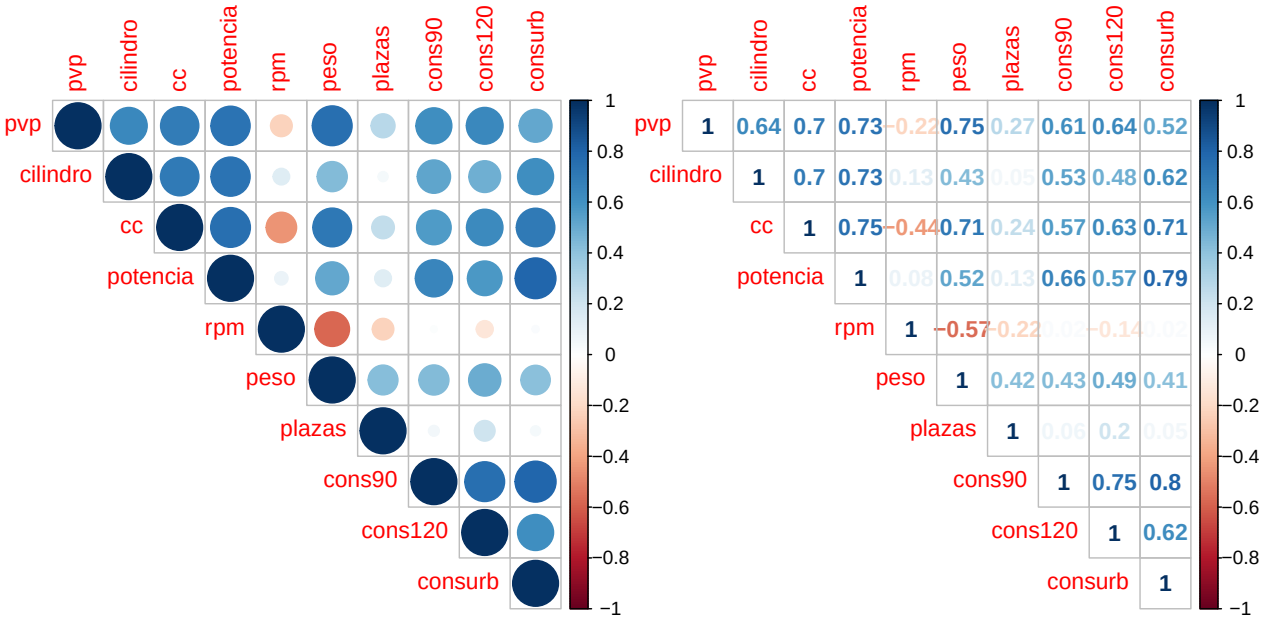
Es destacable el alto número de valores nulos(46) que hay en la columna "acelerac", que es una variable que informa sobre la aceleración en una escala del 0 a 100 segundos. Por otro lado, la variable "acel2" es una variable binaria que clasifica en dos grupos cada coche en función de si su aceleración está por encima o por debajo de los 10 segundos. Ya que esta última no tiene valores nulos y ambas aportan información similar, podemos eliminar la variable "acelerac".

Correlación entre variables

A continuación, estudiamos las correlaciones de las variables para poder decidir cuáles de ellas no son relevantes de cara a la futura segmentación.

Algunas de las variables que tienen más correlación son "cons90" junto con "consurb" y con "cons120". Es por ello, que puesto que la información que aportan es muy similar, se rechaza la variable "cons90". Sin embargo, se dejan "consurb" porque "cons120" su correlación es mucho menor y la información que aportan no es tan similar.

Por otro lado, existe una alta correlación entre potencia con las variables "cc" y "consurb", es por ello que también podríamos prescindir de la variable "potencia".



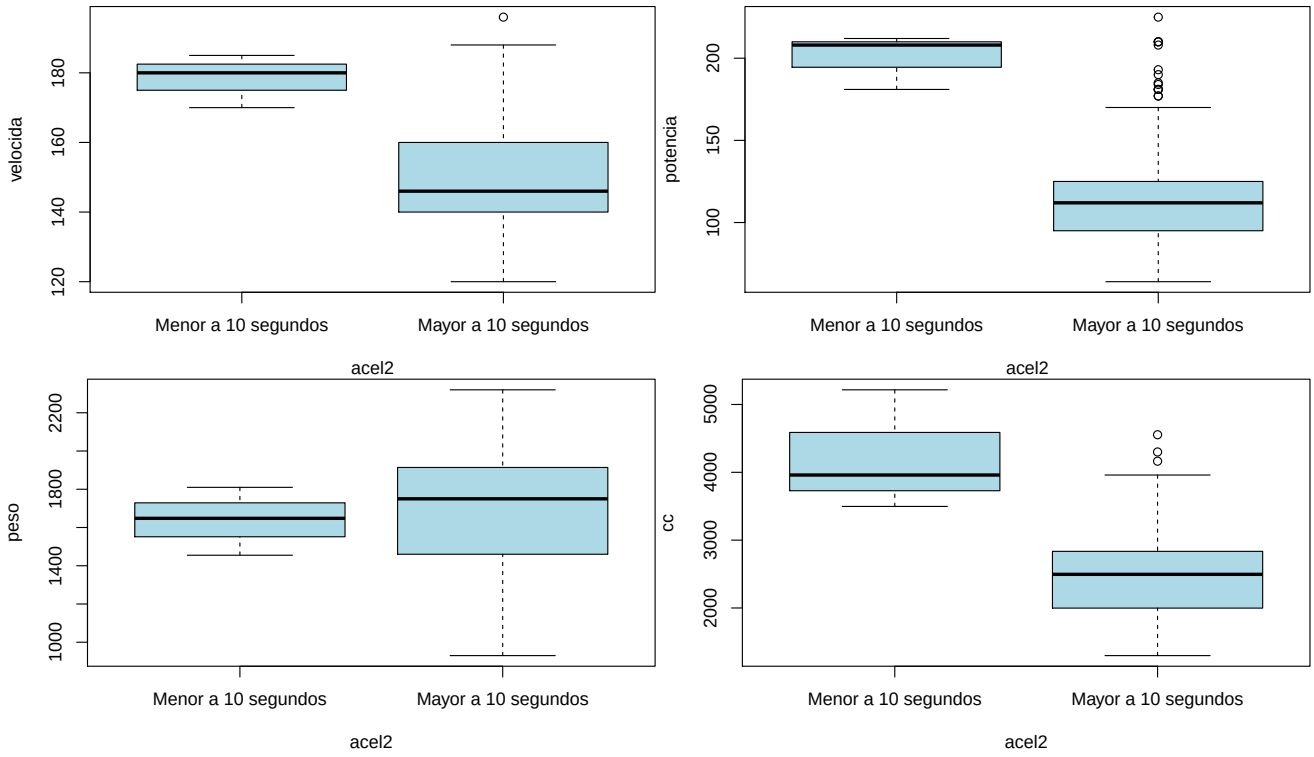
	pvp	cilindro	cc	potencia	rpm	peso	plazas	cons90	cons120	consurb
pvp	1	0.64	0.7	0.73	−0.22	0.75	0.27	0.61	0.64	0.52
cilindro		1	0.7	0.73	0.13	0.43	0.05	0.53	0.48	0.62
cc			1	0.75	−0.44	0.71	0.24	0.57	0.63	0.71
potencia				1	0.08	0.52	0.13	0.66	0.57	0.79
rpm					1	−0.57	−0.22	0.02	−0.14	0.02
peso						1	0.42	0.43	0.49	0.41
plazas							1	0.06	0.2	0.05
cons90								1	0.75	0.8
cons120									1	0.62
consurb										1

Variables categóricas

Con respecto a las variables categóricas, hay dos que no vamos a tener en cuenta en nuestro estudio: marca y modelo. Esta decisión es tomada en base a dos criterios principalmente:

- El primero y más importante es porque el objetivo final es agrupar los coches para conservarlos en distintos lugares, por lo que el modelo o la marca del coche no son características influyentes.
- El segundo criterio es que hay 17 marcas distintas de coches y 111 modelos. Este número tan alto de clases hace su estudio y posterior agrupación mucho más difícil.

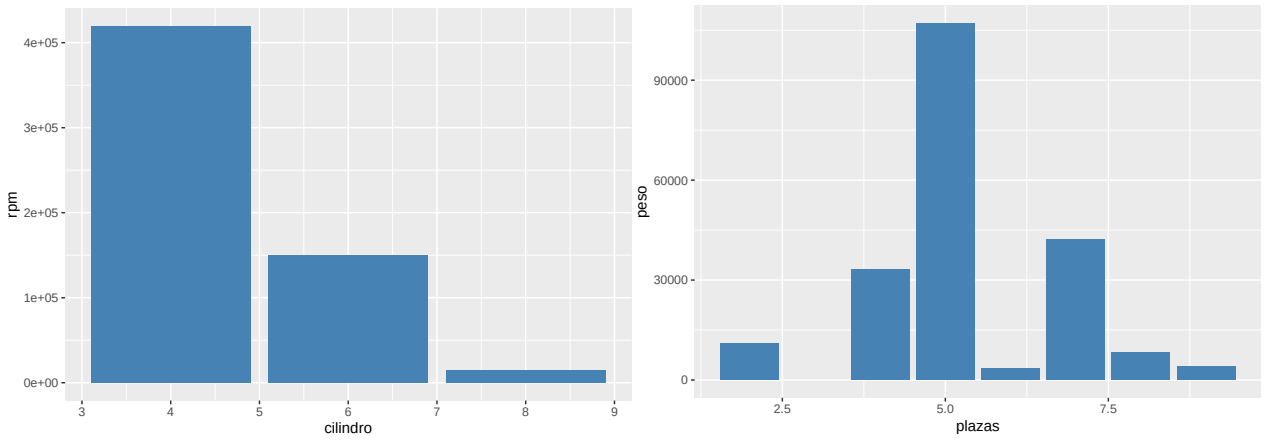
La variable categórica restante es acel2, mencionada anteriormente. Para estudiar su relación con otras variables se han realizado boxplots con aquellas variables que intuitivamente, podría parecer que tendrían más relación. Se puede observar como eso ocurre con todas menos con el peso, donde se puede ver que aunque un coche sea de mayor peso, no significa que vaya a tener un menor tiempo de aceleración.



Por razones de negocio, ya que no se pretende vender dichos coches sino conservarlos en los lugares adecuados, el precio de venta no es relevante. Es por ello que la variable "pvp" se rechace también.

Variables numéricas discretas

Podemos ver cómo a mayor número de cilindros, mayores son las revoluciones por minuto del coche en cuestión. Sin embargo, el número de plazas no tiene relación con el peso, ya que los de mayor peso son los de 5 plazas.



Conclusiones

Tras realizar un análisis exploratorio, hemos visto que para la asignación futura de los vehículos en distintos lugares debemos tener en cuenta 10 variables, de las cuales solo una es categórica y podríamos prescindir de un total de 5 variables.

Las principales razones para rechazar esas 5 variables han sido el alto número de observaciones nulas, la

no relevancia para la consecución de nuestro objetivo (conocimiento del negocio) y la alta correlación entre variables.