INFORME LOS COCHES DEL JEFE

Abstract

Se parte de una base de datos compuesta por 125 vehículos de los que se explican quince variables (marca, modelo, precio, número de cilindros, cilindrada, potencia, revoluciones por minuto, peso, número de plazas, consumo a 90 km/h, a 120 km/h, urbano y la aceleración de 0 a 100 así como el tiempo de aceleración). En este informe se recogen los resultados de un exhaustivo análisis exploratorio de datos y se presentan las variables que se consideran decisivas a la hora de separar en grupos los vehículos de manera consistente. La selección de variables para la clasificación se basa tanto en criterios matemáticos (resultados obtenidos del análisis de datos) así como técnicos (informaciones encontradas sobre automoción).

En la base de datos se consideran dos tipos de características (dejando a un lado marca, modelo y precio), aquellas relacionadas con la estructura del coche per sé (peso y números de plazas) y aquellas relacionadas con el motor (número de cilindros, cilindrada, potencia, revoluciones por minuto, consumo (a 90 km/h, 120 km/h y urbano), velocidad y aceleración).

Con una primera visión panorámica de la base de datos se observa que existen muchos huecos en la variable aceleración. En concreto, existen 46 registros vacíos (un 37%) por lo que parece sensato descartar esta variable. Si llevamos a cabo un test ANOVA para estudiar la independencia entre la aceleración y el tiempo de aceleración se obtiene un p valor de 0.0011 luego se descarta la hipótesis nula (aceleración y tiempo de aceleración son independientes). Si se realiza un gráfico de cajas se observa que los tres modelos con tiempo menor de diez segundos son los que poseen una menor aceleración. Así, tiene sentido descartar la aceleración porque aunque se produce una pérdida de información el tiempo de aceleración (que no presenta NA's) la cubre en parte.

En una aproximación naïve se puede suponer que el número de plazas y el peso están correlacionadas lo cual puede permitir prescindir de una de las dos a la hora de proceder a la segmentación. Mediante un test de análisis de la varianza se plantea un contraste de hipótesis que devuelve un valor extremadamente próximo a cero ($2 \cdot 10^{-16}$) lo que nos permite rechazar la hipótesis nula de independencia de variable y deducir que existe una correlación. Sin embargo al visualizar los datos en un boxplot (**Figura 1**) se aprecia la importancia de considerar ambas variables pues la correlación no es completa. Los vehículos biplaza presentan mucho más peso que aquellos de cuatro plazas.

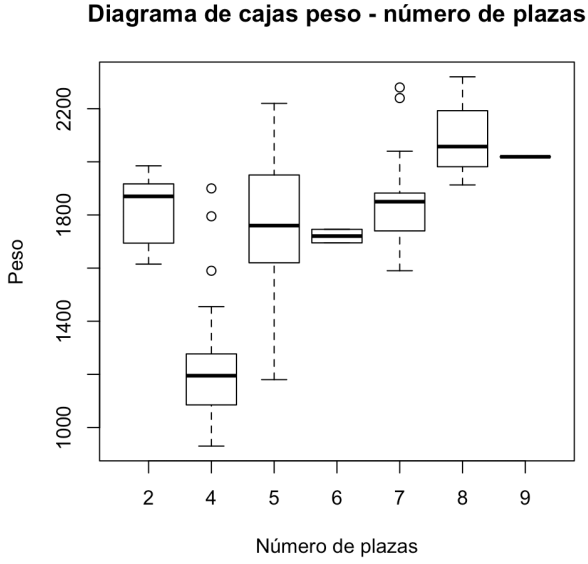

Figura 1. Diagrama de cajas número de plazas vs pesos.

Las marcas no parecen una variable clasificadora importante en este caso. Realmente cuando se recurre a la clasificación a través de marcas, es porque a ciertas marcas se les asocia un mayor prestigio (por ejemplo, Mercedes). Las marcas sabedoras de esto aumentan sus precios por lo que realmente existe una relación directa entre marca prestigiosa y mayor precio. En el código adjunto se presenta un boxplot[1] que muestra que a excepción de Land Rover, Jeep y Mitshubishi todas las marcas concentran los precios de sus modelos en intervalos relativamente pequeños y que las marcas que más distinguen (en el sentido de que presentan vehículos más diferentes coinciden con los outliers de precio (tanto coches más caros como más baratos).

Además al estudiar la distribución del precio en los vehículos se observa que a partir del primer cuartil los precios superan las 2.750.000 pesetas (16.500€ en la actualidad) por lo que se aprecia que se está trabajando con vehículos de gama alta (se necesitaría disponer del año de adquisición o tasación para saber hasta que punto son caros en comparación a los precios de la época).

Entrando en características más técnicas se puede proceder a inspeccionar la importancia de las variables numéricas relacionadas con el motor (se deja a una lado el tiempo de aceleración).

Estudiando la matriz de correlación se observa que la potencia está enormemente asociada a la aceleración, la velocidad, la cilindrada y el consumo urbanos (todos coeficientes con un valor absoluto superior a 0.75). Así, parece evidente que la potencia será de gran importancia a la hora de clasificar los grupos. Además desde un punto de vista técnico la potencia del motor es una de las características esenciales de un vehículo.
De esta misma matriz se observa que las variables de consumo están muy relacionadas entre ellas y que las variables consumo 90 y 120 son extremadamente similares (fuentes externas confirman esta idea). Así se decide trabajar solo con el consumo en 120 pues presenta menores coeficientes de correlación (es decir, distingue más).
Además, se puede observar que las variables velocidad, consumo urbano y potencia están muy correlacionadas con la aceleración (variable previamente descartada) lo que implica que habiéndola descartado no se ha perdido tanta información mientras se conserven estas variables.
Todas estas asociaciones se pueden visualizar de una manera más clara mediante el siguiente gráfico:

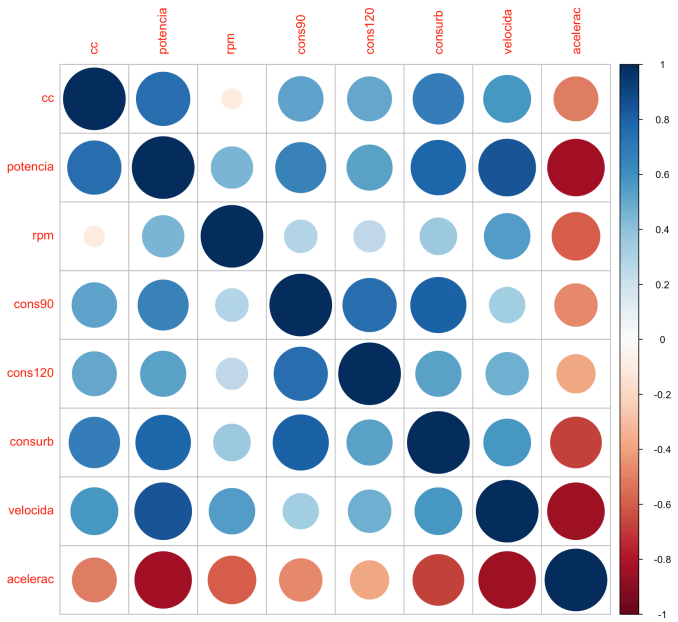

Figura 2. Gráfico de correlación de variables numéricas asociadas al motor.

[1] A lo largo del informe se omiten algunos gráficos que se encuentran presentes en el código para evitar sobrecargar el documento.

Así pues a la hora de clasificar los vehículos parecen prescindibles las variables marca, consumo a 90 km/hora. El resto de variables (descartando evidentemente los modelos) se pueden emplear para clasificar en distintos grupos (análisis cluster).

(**Nota:** En un primer momento interpreté que era necesario realizar dicho análisis y aunque es posible realizarlo empleando todas las variables, es cierto que esta consideración produce resultados poco consistentes en el sentido de que los grupos aparecen poco separados y muchas veces entremezclados.)

Una vez tomada dichas decisiones se procede a continuar profundizando en el análisis explotarlo de datos con el objeto de caracterizar los vehículos.

Una de las características más llamativas es el peso de los vehículos. En 1987 la media de peso de los vehículos en EEUU era de 3221 libras (unos 1400 kilos). En la base de datos solo existen 29 vehículos inferiores a dicho peso. Además, el rango de peso es bastante amplio pensado el vehículo más ligero 930 kg y el más pesado 2320 kg.

A continuación se estudia la potencia y la cilindrada. Para el estudio de ambos encuadrado en una época se realizarán comparativas respecto a la media en 1990. La Asociación Manufacturera Europea de Automóviles presenta en su página web los datos de cilindrada y potencia media (ojo, esta se presenta en kw y no en cavallos de vapor) en Europa en general y en cada país de Europa en particular.

Así se puede caracterizar también los vehículos por su altísima cilindrada (exceptuando los Suzuky Samurai Corto todos superan la cilindrada media en Europa (1587) y aun más la media española (1532)).

De la misma manera, transformando los kilowatios a cavallos se observa que los vehículos recogidos en la base de datos superan la potencia media europea y española en la década de los 90 (en torno a 80 cv) pues a partir del primer cuartil los automóvil poseen una potencia superior a 95 cavallos.

La **Figura 3** muestra un histograma con la cilindrada de los vehículos en la base de datos, en rojo aparece la media española en 1990 y en azul la media europea. La **Figura 4** por su parte, muestra la potencia de los vehículos en cavallos de vapor siendo la recta la media (europea y española pues eran prácticamente iguales) en 1990:

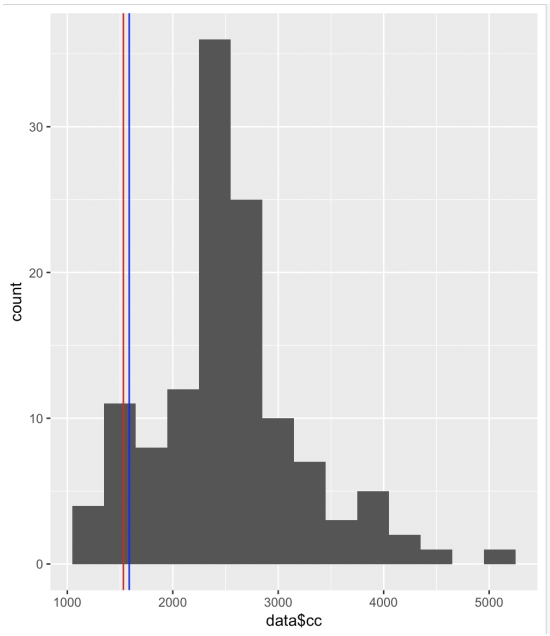

Figura 3. Histograma de la cilindrada.

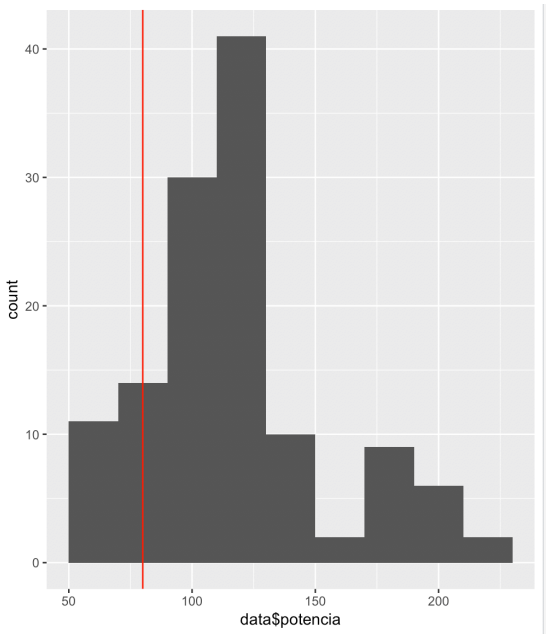

Figura 4. Histograma de la potencia.

En un principio surgió la idea de transformar el precio de pesetas a euros sin embargo este perdería mucha representación por no saber en qué época fueron adquiridos los vehículos (no es lo mismo un millón de pesetas en 1950, 1970 o 1999). En esta misma línea parece una buena idea la retasación de los vehículos en la actualidad pues el valor de los vehículos ha variado sustancialmente. En primer lugar, todos los vehículos deben tener una antigüedad superior a veinte años (el euro se implanta en el 2000) por lo que se deben haber revalorizado en gran medida. En segundo lugar, la peseta no es una unidad representativa en la actualidad por lo anteriormente expuesto. Además existe otro factor de gran relevancia, el consumo de todoterreno tan extendido en Europa y en España hoy en día no lo era en el siglo XX lo cual le otorga un valor añadido a los vehículos de esta base. Por hacer una idea se muestra en la siguiente gráfica (**Figura 5**) la evolución del uso de todoterrenos en España durante los últimos treinta años:

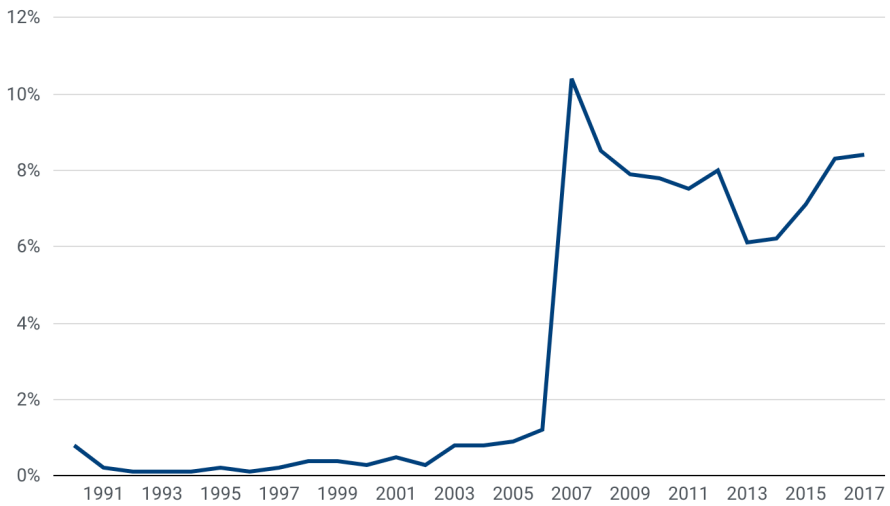

Figura 5. Adquisición de todoterrenos en España.
Fuente: AAA

Así se observa que las piezas no son solo únicas por su antigüedad si no que además ya entonces eran artículos de una gran exclusividad. Por todo ello es aconsejable recalcular el valor económica de los coches.

Conclusiones

Tras un exhaustivo análisis de las variables es claro que la colección de autos aquí expuesta habla de un tipo muy particular de vehículo. Son vehículos de lujo (sobre todo encuadrados en su época) tanto por su exclusividad entonces como por sus elevadísimos precios (existen vehículos con valores que superan los diez millones de pesetas (más de 60.000€), de gran potencia y cilindrada (valores muy altos tanto en su época como actualmente) y con un consumo muy elevado (tal y como viene siendo habitual en los todoterrenos). Son vehículos de gran tamaño y en su gran mayoría con amplia capacidad para pasajeros (exceptuando los seis modelos pick up).

A la hora de realizar una clasificación para su posterior reparto en los distintos garajes se descartan algunas variables por considerarse incompletas (aceleración de 0 a 100) o superfluas (marcas, consumo a 90 km/h) por considerar que su información se puede obtener a partir de otras variables (precio y potencia o consumo a 120 km/h y consumo urbano). Estas hipótesis se verificarán y revisarán en la siguiente práctica mediante el análisis cluster.

CÓDIGO

En el siguiente enlace se puede acceder al repositorio de GitHub en el que se encuentra el código empleado para el estudio de esta base de datos y una primera aproximación a la clasificación de los vehículos. El código incluye a su vez más gráficos y los detalles del análisis de cada variable (media, cuartiles, máximo, mínimo…)

BIBLIOGRAFÍA

https://www.r-bloggers.com/clustering-mixed-data-types-in-r/

https://www.tecnocoche.com/mecanica/mecanica_basica/cilindrada_relevante_comp.html

https://slate.com/business/2011/06/american-cars-are-getting-heavier-and-heavier-is-that-dangerous.html?via=gdpr-consent

https://www.acea.be/statistics/tag/category/cubic-capacity-average-power