LOS COCHES DEL JEFE

AGRUPACIÓN Y REDUCCIÓN DE LA DIMENSIÓN

Beltrán Aller López

19/11/2019

CONCLUSIONES

Tras llevar a cabo un análisis exploratorio de la colección de vehículos, se comprueba que estos presentan por término general las siguientes características:

- Son vehículos pesados.
- Presentan una gran capacidad, oscilando el número de plazas entre cuatro y siete.
- Vehículos con un alto consumo, tanto por carretera como por zona urbana.
- Vehículos potentes.

El objetivo final será llevar a cabo un análisis clúster de los vehículos, dividiéndolos así en grupos homogéneos internamente y heterogéneos entre sí. Dichos vehículos serán guardados en garajes ubicados en distintas localizaciones.

Se puede asumir que las localizaciones presentan condiciones climatológicas, así como orográficas distintas. Por lo que, a la hora de realizar la división de los vehículos, esta ha de atender a las características técnicas. Características relacionadas con los atributos del motor (cilindrada, revoluciones por minuto, potencia y centímetros cúbicos), especificaciones generales y capacidad (velocidad máxima, aceleración y número de plazas) y eficiencia (consumos).

Es por esto, por lo que atributos como el precio en pesetas, la marca y el modelo, no resultan útiles de cara a realizar la división y, por ello, serán extraídos del dataset y no tenidos en cuenta. Si bien podrían ser tenidas en cuenta de cara a la clasificación de los vehículos para su venta, no es nuestro objetivo (a priori).

Las variables asociadas al consumo (consumo a 90 km/h, a 120 km/h y urbano), aunque pueden resultar redundantes, no serán desechadas en un primer momento. En prácticas posteriores se procederá a su tratamiento.

RESUMEN

Parto de una base de datos compuesta por un total de 125 vehículos y 15 variables asociadas a estos (marca, modelo, precio en pesetas, número de cilindros, centímetros cúbicos, potencia en caballos de vapor, revoluciones por minuto, peso en kilogramos, número de plazas, consumo en carretera a 90 y 120 km/h, consumo urbano, velocidad máxima en km/h, aceleración de 0 a 100 km/h y tiempo de esta aceleración). En el siguiente informe se recogen los resultados surgidos del análisis exploratorio de datos y la selección de las variables a tener en cuenta de cara a un futuro análisis de conglomerados de los vehículos.

DESARROLLO

El dataset se compone de tres tipos de variables, numéricas continuas (precio en pesetas, centímetros cúbicos, potencia en caballos de vapor, revoluciones por minuto, peso en kilogramos, consumo en carretera a 90 y 120 km/h, consumo urbano y aceleración de 0 a 100 km/h), numéricas discretas (número de cilindros y número de plazas) y categóricas (marca, modelo y tiempo de aceleración).

A la hora de realizar una primera exploración sobre los datos de los vehículos registrados en el dataset, se observa un elevado porcentaje de omisiones en la variable correspondiente a la aceleración, un 36,8 %. La omisión de todas estas observaciones reduciría el número de vehículos de 125 a 79; esto generaría una disrupción grave en el análisis. Sin embargo, el tiempo de aceleración resulta ser una variable dicotómica; la cual adopta sus valores en función de si la aceleración de 0 a 100 km/h se produce en un intervalo de tiempo superior, o inferior, a los 10 segundos. Puesto que la aceleración resulta una variable imprescindible a la hora de clasificar un vehículo a mi parecer, procedo a realizar una imputación de los valores ausentes a partir de la variable tiempo de aceleración. El método será el siguiente:

- Se partirá de un valor base igual a 10.
- En función de los valores que adopte la variable tiempo de aceleración (1 = inferior a 10 segundos, 2 = superior a 10 segundos), se sumará (2) o restará (1) la desviación típica de la aceleración de los vehículos para los cuales si existe el dato en cuestión.
- A fin de no generar por defecto valores homogéneos que distorsionen la muestra, la desviación típica a sumar (o restar) será multiplicada por un valor entre 0 y 1 generado de forma aleatoria para cada vehículo.
- Se lleva a cabo la suma (o resta) y la imputación del valor.

Una vez imputados los valores para la aceleración procedo a estudiar el resto de los valores omitidos en el dataset. El consumo en todos sus niveles (90 km/h, 120 km/h y urbano) es la siguiente variable que mayor número de omisiones representa (8 %, 12 % y 5,6 % respectivamente), por tanto, se procede a la imputación de sus valores de la siguiente forma:

- Para aquellos vehículos de los cuales existan otros de la misma marca y modelo, o sólo de la misma marca y, esos a su vez presenten información sobre los consumos; se les imputará el consumo de los segundos.
- Para aquellos vehículos sobre los que no exista ningún otro vehículo de la misma marca y modelo, ni siquiera de la misma marca; se les imputará el valor de los consumos de los vehículos más parecidos. Las variables tenidas en cuenta de

cara a seleccionar los vehículos más parecidos son el número de cilindros, las revoluciones por minuto, los centímetros cúbicos, la potencia en caballos de vapor, el número de plazas y el peso en kilogramos. Las cuatro primeras variables se corresponden con atributos del motor, que influyen directamente sobre la aceleración de un vehículo; y las dos últimas, con características generales del vehículo, que a mi parecer condicionan de una forma u otra la aceleración de este.

Una vez realizada la imputación de los valores omitidos en las anteriores variables, se comprueba que el número de vehículos que presentan omisiones se reduce de un total de 36,8 % de la muestra, a un 4,8 %. Subjetivamente, el sesgo generado por la pérdida de este porcentaje de vehículos no es significativo, por lo que elimino tales vehículos del dataset. El número final de vehículos a tener en cuenta es de un total de 119.

En una primera aproximación, se podría suponer que los consumos han de estar relacionados con el peso y, este a su vez con el número de plazas del vehículo; lo cual podría conducir a prescindir del peso o del número de plazas. A continuación, se exponen en los siguientes gráficos, la relación entre las tres variables mencionadas.

Ilustración 1: Peso - Consumo a 90

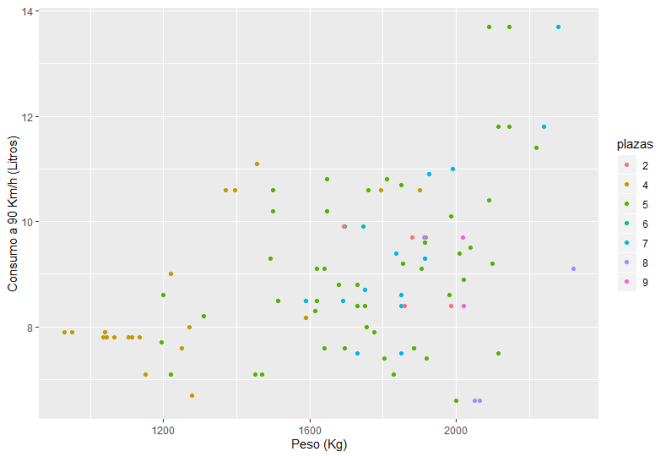


Ilustración 2: Peso - Consumo a 120

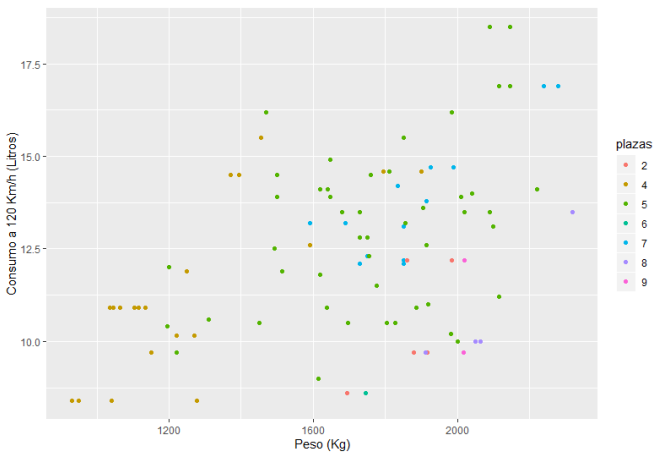
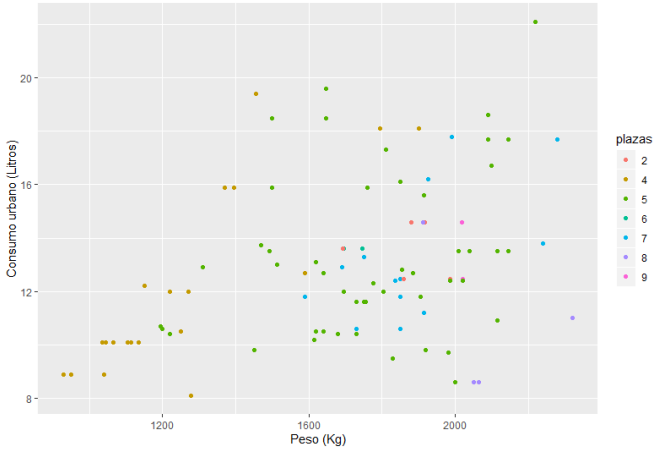
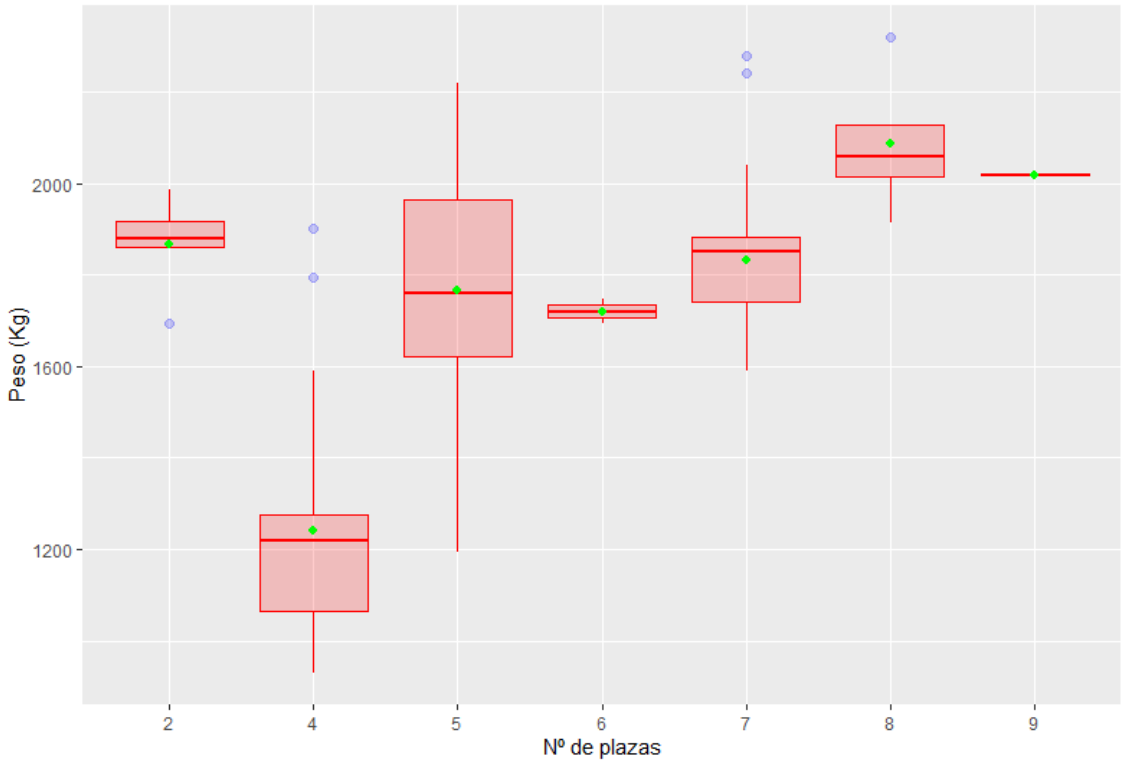


Ilustración 3: Peso - Consumo urbano



En los diagramas de dispersión se aprecia la relación existente entre las variables, por norma general, a mayor peso, mayor consumo. Sin embargo, la relación existente entre el peso y el número de plazas no queda del todo bien definida. En la siguiente ilustración se observa mejor.

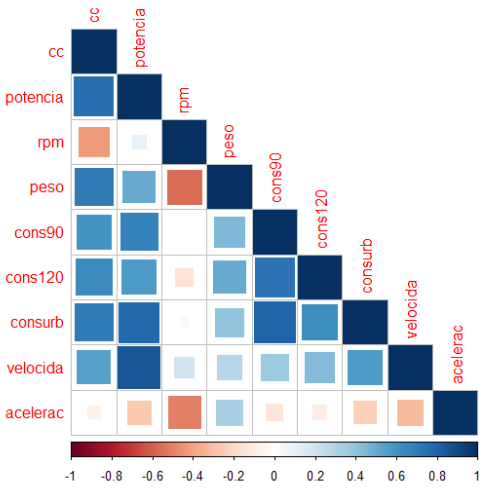
Ilustración 4: Diagrama de cajas N.º de plazas - Peso



En el diagrama de cajas se observa que aquellos vehículos de 2 plazas, pese a que el sentido común nos haga pensar que tienen que pesar menos que aquellos con un mayor número de plazas, pesan más en muchas ocasiones que el resto de los vehículos. Por tanto, se conservan ambas variables.

A continuación, procedo al análisis de las correlaciones de las variables asociadas a los aspectos mecánicos del vehículo, en concreto con el motor.

Ilustración 5: Matriz de correlaciones de las variables del motor



Acorde a la matriz de correlaciones obtenida, se observa que la potencia se encuentra muy asociada a los centímetros cúbicos, el peso, los distintos consumos y la velocidad máxima. Es un indicador de que puede ser una variable muy a tener en cuenta de cara a la realización de un análisis clúster. La aceleración resulta ser una de las variables más incorreladas de todas las relacionadas con el motor, esto indica que poder llegar a ser un atributo diferenciador a la hora de realizar grupos en el análisis clúster. Los tres tipos de consumo presentan correlaciones altas entre sí, si bien es cierto que el consumo a 120 km/h es el que menores correlaciones presenta con el resto de las variables, lo cual puede indicar que posee una mayor capacidad discriminante que el consumo a 90 km/h y el consumo urbano; se podría prescindir de estos dos últimos.

Una vez llevado a cabo todo el análisis exploratorio, la marca y el modelo de los vehículos presentan poca importancia a la hora clasificar estos, al igual que el precio en pesetas. Si bien estas variables podrían ser tenidas en cuenta a la hora de realizar grupos de cara a proceder a la venta de los vehículos, no es nuestro objetivo. Por lo que prescindo de ellas.

Bibliografía

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer.