Los coches del jefe

Hugo César Octavio del Sueldo

11/21/2020

Introduccion

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 del 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.

A la hora de realizar la división de los vehículos, voy a prestarle atencion a las características técnicas de los mismos. Características relacionadas con los atributos del motor (cilindrada, revoluciones por minuto, potencia y numero de cilindros), especificaciones generales y capacidad (velocidad máxima, aceleración y número de plazas) y eficiencia (consumos).

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 10 lugares agrupándolos en virtud de las características que considere oportunas.

Desarrollo

Tomando como referencia los objetivos de la practica planteada, algunos atributos como el precio en pesetas, la marca y el modelo, no resultan útiles de cara a realizar la división y, por ello, no serán tenidos en cuenta. Si bien podrían ser tenidas en cuenta de cara a la clasificación de los vehículos para su venta, no es nuestro principal objetivo en este momento.

A la hora de realizar una clasificación para su posterior reparto se descartan algunas variables por considerarse incompletas (aceleración de 0 a 100).

Las variables asociadas al consumo (consumo a 90 km/h, a 120 km/h y urbano), seran desechadas ya que utilizaremos el consumo urbano que no tiene fuerte correlacion con ninguna de las otras variables, por lo tanto, a la hora de un analisis cluster creo que seria interesante mantenerla y no a consumo a 90 y 120 km/hr.

Tratamiento de los datos

El dataset se compone de tres tipos de variables, numéricas continuas (precio en pesetas, cc, potencia, 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). El dataframe tiene muchos nulos para sus dimensiones por lo que deberemos trabajar estos Na.

El criterio que voy a utilizar sera completar los datos faltantes de cada una de las variables con el promedio de aquellos vehiculos de la misma marca o modelo. Asi voy a ir revisando el fichero fila a fila observando

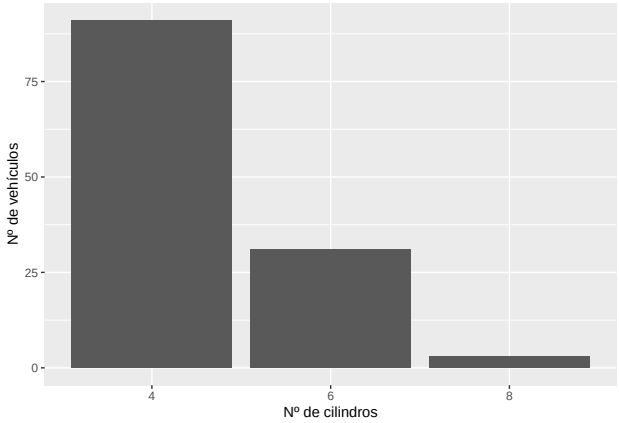
a que dato corresponde el Na faltante asignandole un valor en base a esta regla de decision. He tomado esta decision ya que si descartaria los valores Na directamente estaria perdiendo muchisima informacion. Por otro lado, aquellos variables que no tengan valores de referencia de otros vehiculos para usarlos en las imputaciones faltantes, lo que se hizo fue reemplazarlos por la media global de las observaciones.

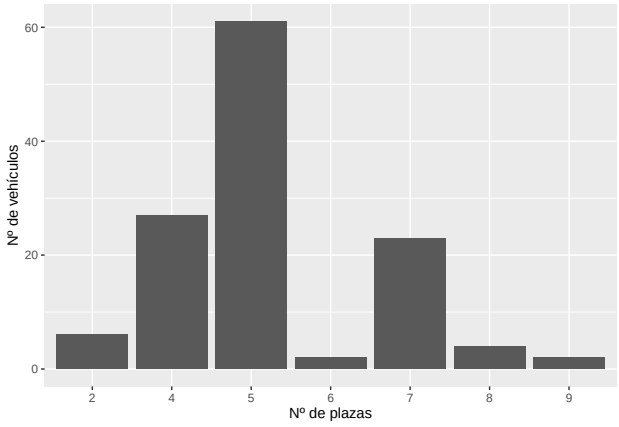
Ya completados los datos del dataframe, procedemos a analizar en detalle cada una de las variables con el fin de filtrar aquellas variables que consideramos relevantes para el analisis dejando afuera algunas que no consideramos oportundas.

Procederemos primero a escalar las variables numericas para poder compararlas. A traves de la funcion scale, escalamos los valores numericos para que esten en la misma escala metrica para que podamos realizar el analisis

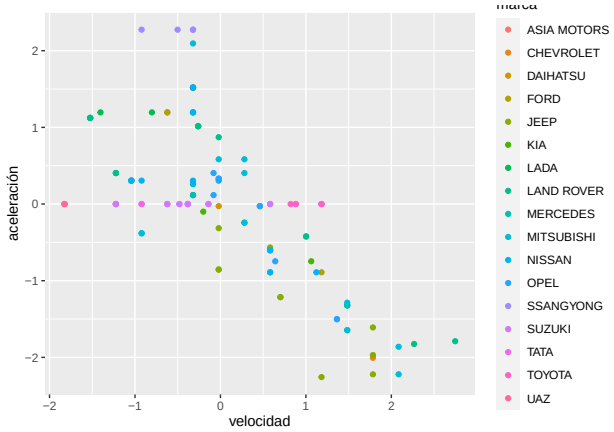
Una vez hecho esto, procederemos a graficar las variables que caracterizan a los vehiculos para poder conocer mejor los coches del jefe.

Graficos de las variables de los vehiculos





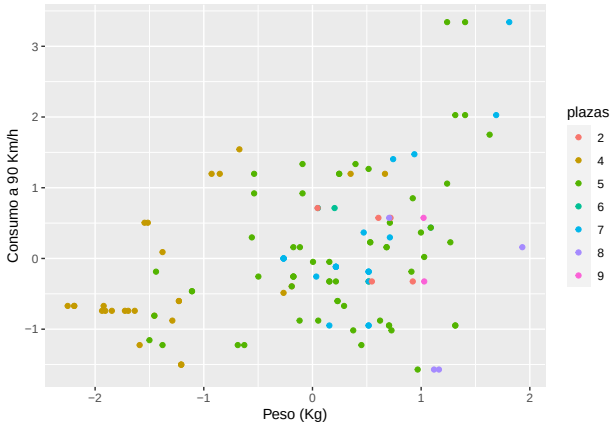
Aqui podemos observar que los coches del jefe son coches grandes ya que la mayoria son vehiculos entre 4 y 7 plazas.

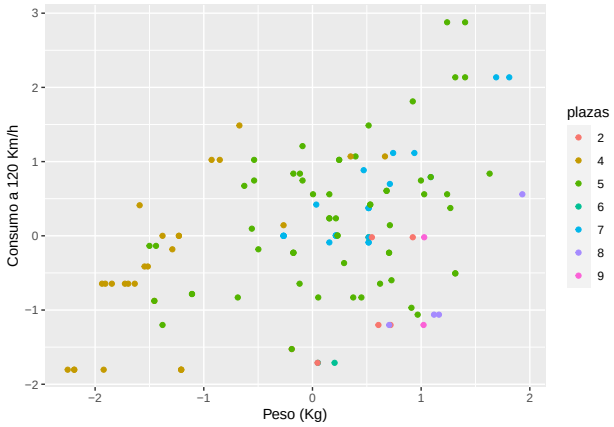


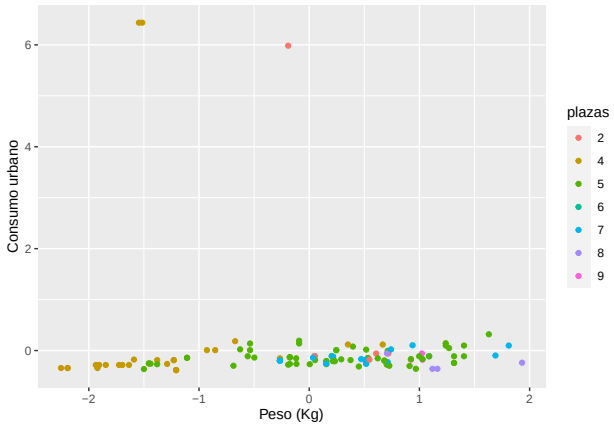
Aqui podemos apreciar como a mayor velocidad, menor es la aceleracion que se genera por lo vehiculos.

Veamos ahora la relacion existente entre peso, consumo y plazas de los vehiculos

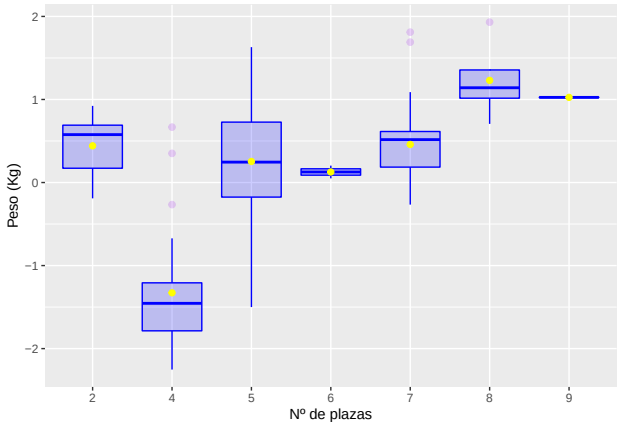
En una aproximación naive, se podría suponer que los consumos han de estar relacionados con el peso y, este a su vez con el número de plazas del vehículo; lo cual podría conducir a prescindir del peso o del número de plazas. A continuación, se exponen en los siguientes gráficos, la relación entre las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





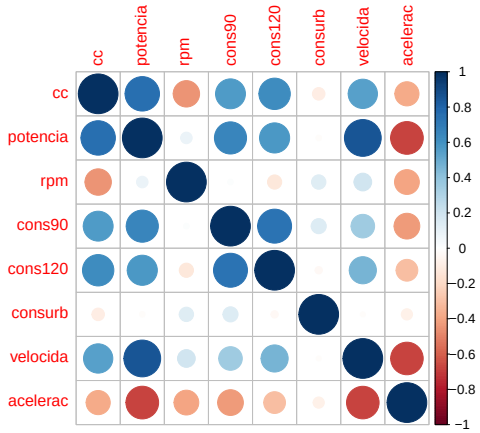


En los diagramas de dispersión se aprecia la relación existente entre las variables, por norma general, a mayor peso, mayor consumo. Sin embargo, la relación existente entre el peso y el número de plazas no queda del todo bien definida. En la siguiente ilustración se observa mejor.



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 en esta primera etapa.

Respecto a las variables numéricas de motor se calcula la matriz de correlaciones:



De acuerdo a la matriz de correlaciones obtenida, se visualiza que la potencia se encuentra muy asociada a los cilindrada del vehiculo, los consumos a 90 y 120 km/hr y la velocidad máxima. Esto es un indicador de que puede ser una variable 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 podria llegar a ser un atributo diferenciador a la hora de realizar grupos en el análisis clúster. Lo mismo para velocidad y rpm ya que no estan muy correlacionadas a ninguna de las otras variables.

Los consumos a 90 y 120 km/hr presentan correlaciones altas entre sí. El consumo urbano 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 a 120 km/hr; 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 en este momento