Víctor Escudero Ramos

TheBridge

1. Introducción

Hemos decidido realizar este EDA (Análisis Exploratorio de Datos), tomando un Dataset de Kaggle llamado “Car Price DataSet” facilitado por el usuario Mazaharul Hasnine Mirza, ingeniero y científico de datos con más de 9 años de experiencia. Encontramos datos acerca del precio y características de una muestra de vehículos, tales como el tipo de carrocería, tracción, tipo de carburante, anchura del eje de las ruedas, peso en vacío, etc.

El objetivo principal de este análisis es determinar si realmente la potencia en caballos de un vehículo repercute directamente en precio final del mismo.

En segunda instancia, también hemos querido destacar la importancia de tener una buena fuente de datos y cómo analizar nuestros datasets para saber si éstos nos sirven para realizar afirmaciones concluyentes.

Parte de la intención de elegir este tema, es demostrar lo obvio; muchas veces damos por sentado cosas que realmente no hemos comprobado y damos como ciertas, sin saber realmente si son verdad o no. Por esto, hemos decido comprobar por nosotros mismos si esto realmente se cumple.

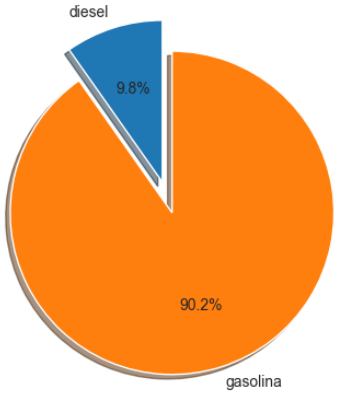
2. Análisis previo

Para plantear nuestra hipótesis, primero de todo hemos analizado los datos de los que disponíamos para poder determinar la fiabilidad de nuestro análisis, ya que unos datos pobres y poco concisos pueden llevar a conclusiones erróneas.

Antes de comparar nuestra variable target (Precio) con el resto de variables para ver su incidencia y correlación, hemos realizado un análisis previo de la distribución de nuestros datos.

Lo primero con lo que nos hemos topado es con la falta de registros, ya que sólo disponíamos de 210 registro, de los cuales, al final nos hemos quedado con 205 debido a la presencia de valores duplicados.

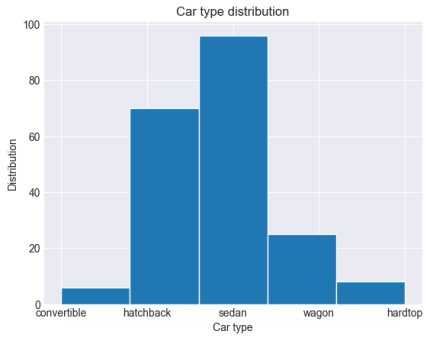
Otro problema que presentaban nuestros datos es la falta de balance; tanto solo el 9% de los vehículos registrados eran de tipo “Fuel” mientras que el otro 91% eran “Gas”.



El rango de precios que abarcaban nuestros datos también estaba sesgado, ya que en su mayoría se encontraban en un rango comprendido entre los 10.000€ y 20.000€.

Aunque esto último contraste con la realidad, ya que los coches de gama media están más generalizados que los de alta gama, la falta de registros y el desbalanceo de nuestro dataset no nos permiten tomar como fiables estos datos.

Esto mismo pasa con el tipo de coche; ya que predominan los de tipo “Sedán” en nuestro dataset.



3. Limpieza de datos

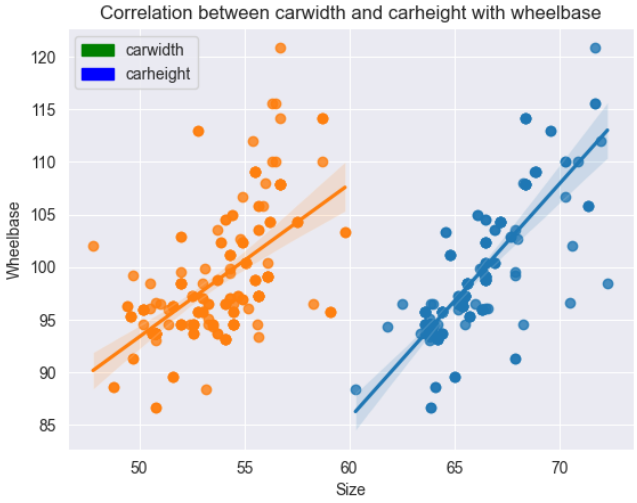
Después de ver la distribución de nuestros datos, hemos realizado un exhaustivo análisis de la correlación entre las variables. Para empezar este proceso, hemos buscado y eliminado valores nulos y repetidos.

Primero hemos convertido las variables de tipo texto con carácter binario en variables de tipo numérico. Se ha realizado la misma transformación para columnas que contenían valores numéricos escritos en como cadena de texto.

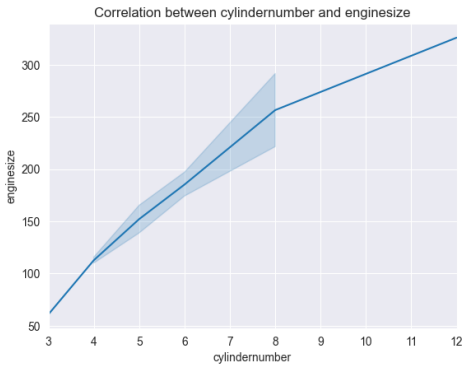
También hemos transformado las columnas que indican la eficiencia del combustible en ciudad y la eficiencia del combustible autopista de un vehículo en una sola columna sacando la media de ambas.

Una vez tratados estos datos, hemos podido advertir que muchas de las variables de nuestro dataset son simplemente la suma o consecuencia directa de otras variables.

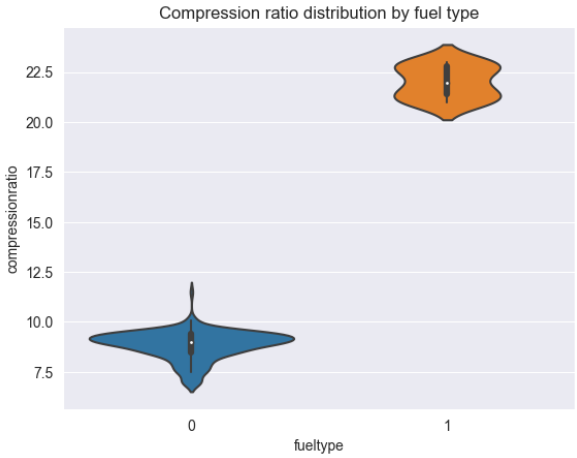
Por ejemplo, las variables del largo y ancho determinan la distancia entre los ejes, variable también presente en nuestros datos, por lo que hemos eliminado el ancho y largo de nuestro dataset.



Otras variables directamente relacionadas entre sí son el tamaño del motor y el número de cilindros; a mayor número de cilindros mayor tamaño, por lo que eliminamos la variable del tamaño del motor.



Esto se cumple también para las variables "CompressionRatio" (Ratio de compresión del motor) y "FuelType" (Tipo de combustible), ya que los coches de tipo diésel tienen un ratio de compresión del motor de 18 a 24 y los de gasolina uno de 6 a 12. Eliminamos la columna de "CompressionRatio" ya que repite información que nos da el tipo de combustible.



4. Hipótesis

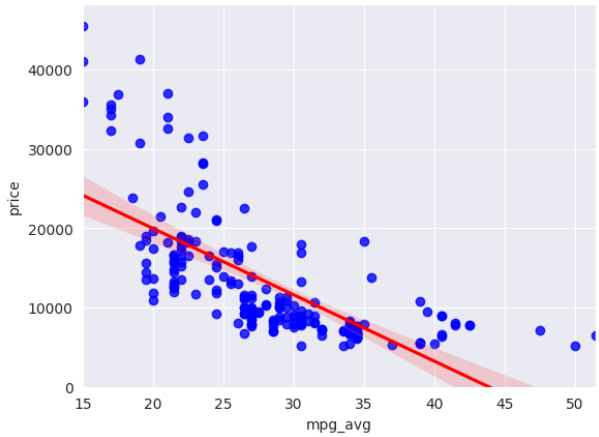
Una vez analizados y contextualizados los datos, podemos tener ya una pequeña idea del grado de fiabilidad de nuestras conclusiones.

Nuestra hipótesis pretende demostrar que la relación entre los caballos de potencia de un vehículo tiene una relación directamente proporcional al precio.

Para desarrollar la hipótesis nos centraremos en las variables más correlacionadas con el precio: “mpg_avg” (millas por galón), “cylindernumber” (número de cilindros), “horsepower” (caballos de potencia), “drivewheel” (Tipo de tracción), “wheelbase” (distancia entre los ejes).

- MPG

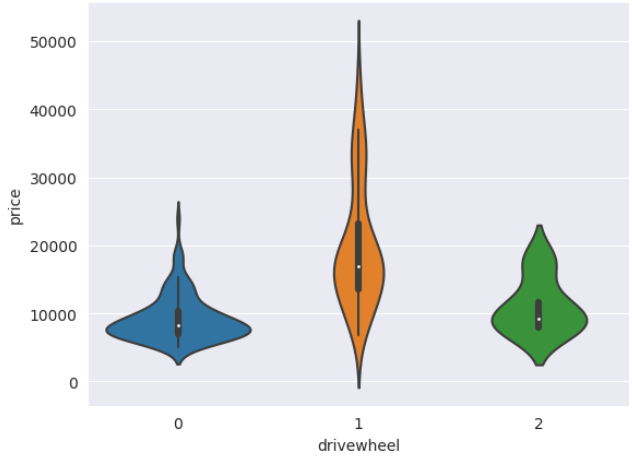
Podemos observar que, a menor capacidad de recorrer millas por galón, el precio aumenta.



Como el número de millas por galón es inversamente proporcional a la potencia en caballos, podemos deducir que la relación que tienen las millas por galón con el precio se debe a su relación directa con la potencia en caballos.

- Drivewheel

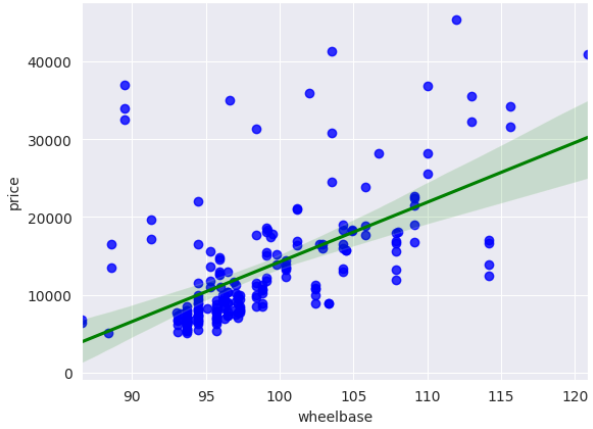
La distribución con mayor rango de precios se encuentra en los coches de tracción trasera (tipo 1) siendo de rango más reducido los de tipo 0 y tipo 2 (tracción delantera y tracción total respectivamente)



Como la distribución de precios de los vehículos con tracción trasera (tipo 1) abarca un rango comprendido entre los más bajos y los más altos, y los otros tipos de tracción están comprendidos entre los 10.000€ y 20.000€, no podemos concluir con que tenga relación directa con el precio del vehículo.

- WheelBase

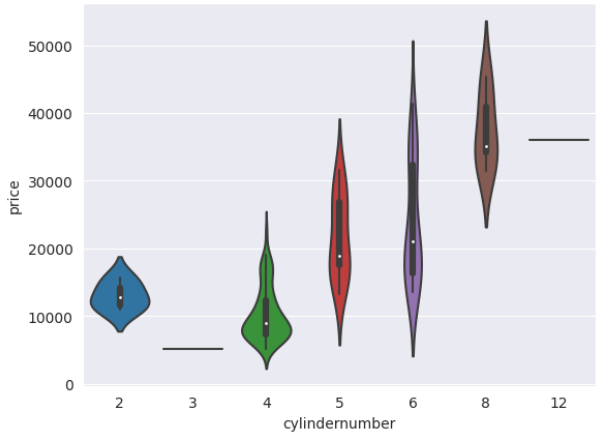
La distancia entre los ejes de los vehículos presenta una distribución dispersa.



Aunque tiende a ascender el precio con el tamaño del vehículo, vemos que la distribución de los precios más altos también se encuentra en vehículos de menor tamaño y al contrario, la distribución de los precios bajos se encuentran en vehículos de mayor tamaño.

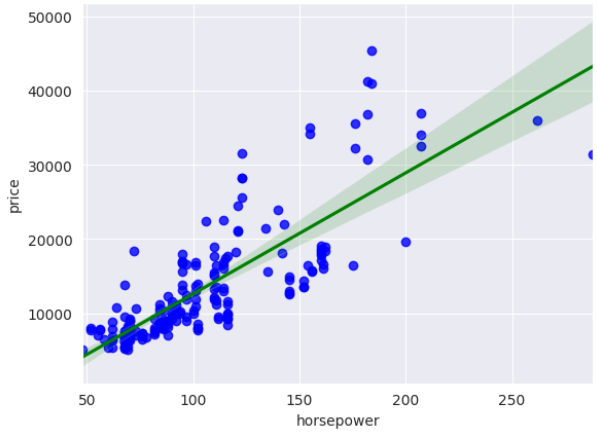
- CylinderNumber

Se observa que, a mayor número de cilindros, mayor es el precio, pero hay coches con precios muy elevados con un número de cilindros menor a 12.



No podemos concluir que el número de cilindros repercute en el precio del vehículo, ya que observamos una distribución de precios muy dispar en cada uno de ellos. Ya que, por ejemplo, un vehículo con 6 cilindros puede costar menos que uno de 4 e igual que uno de 8.

- HorsePower



En esta gráfica es en la única que podemos observar una correlación directa sin ningún tipo de anomalía ni distribución anormal, por lo que, con lo previamente analizado, podemos concluir que la relación más determinante es la potencia en caballos.

5 - Conclusión

Pese a la baja calidad y cantidad de datos de los que disponemos, vemos que nuestra hipótesis es cierta; los caballos de potencia son directamente proporcionales al precio de un coche.

Aunque veamos una clara relación, no podemos afirmar que esto se cumpla en todos los casos, ya queda demostrado que los datos de los que hemos partido presentan irregularidades y están desbalanceados, por lo que en esta conclusión tenemos que tener en cuenta estos factores.

La intención de este análisis es demostrar que, aunque los resultados que estamos observando después de un análisis parezcan lógicos y demuestren algo obvio, no podemos quitar importancia a la calidad de los datos con los que hemos llegado a esa conclusión.

En este caso la conclusión podría darse por válida dada su obviedad, pero, debido a la fuente de datos de la que se ha extraído, no podemos considerarla correcta.