TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

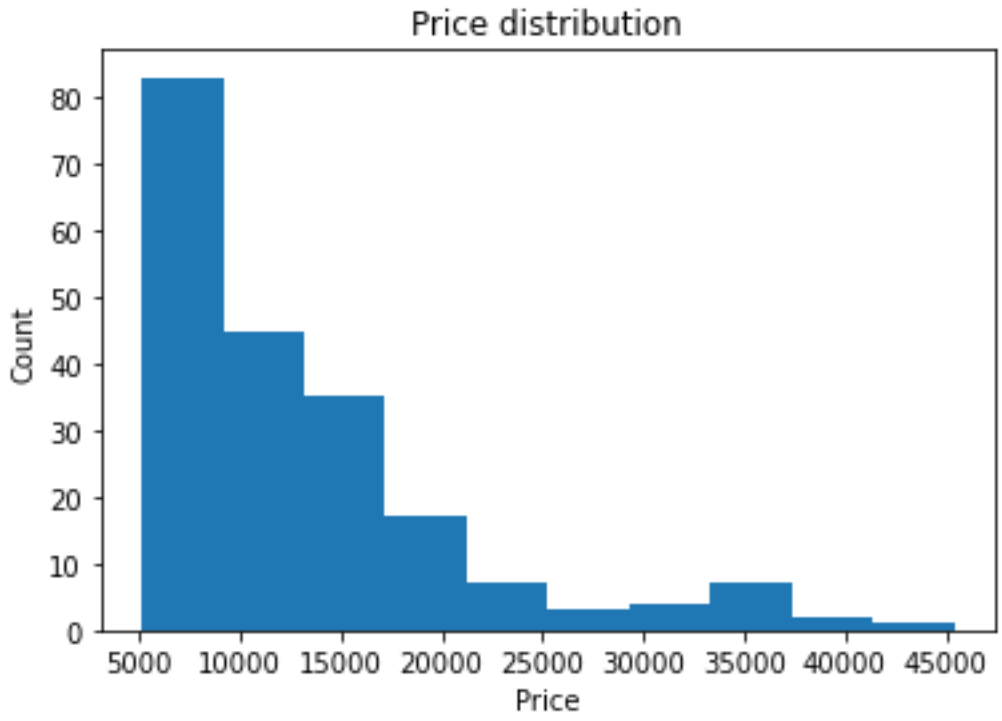
The data set is a summary of different car makes and description of the car in terms of body type, engine type, engine size, horse power, price etc. This report will be heavily focused on what determines a price of a car.

DATA CLEANING

Looking at the data set, it is first discovered that not all columns are relevant for the purpose of this report so a couple of unnecessary columns were dropped using DataFrame.drop() function in pandas.

Furthermore, DataFrame.duplicated().sum() function was used to check if there any duplicated rows of which none were found.

To check if there are any outliers in the price distribution, the following histogram was plotted;

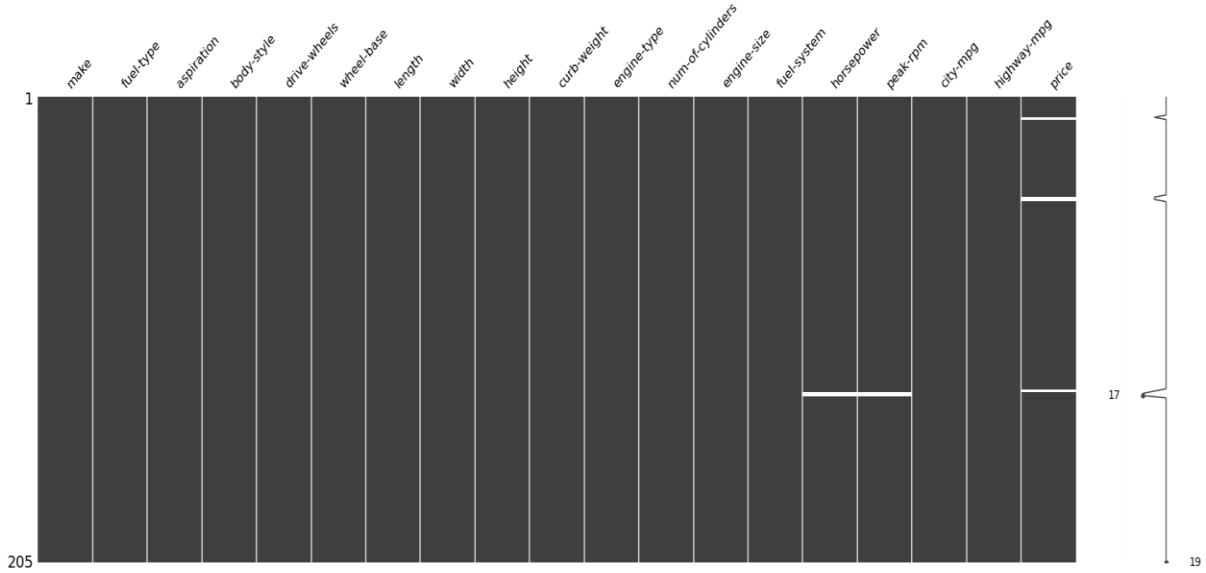


MISSING DATA

The function df.isnull().sum() is used to check if there are any null values and even though the function returned zero, the first inspection after loading the data set resulted in noticing some records presented as question marks(?). A check was made by selecting unique values columns using

DataFrame.Column.unique() in order to make an assumption that ? are all missing values and need to be dealt with. Therefore all ? were replaced with NaN using DataFrame.replace('?', np.nan, inplace=True).

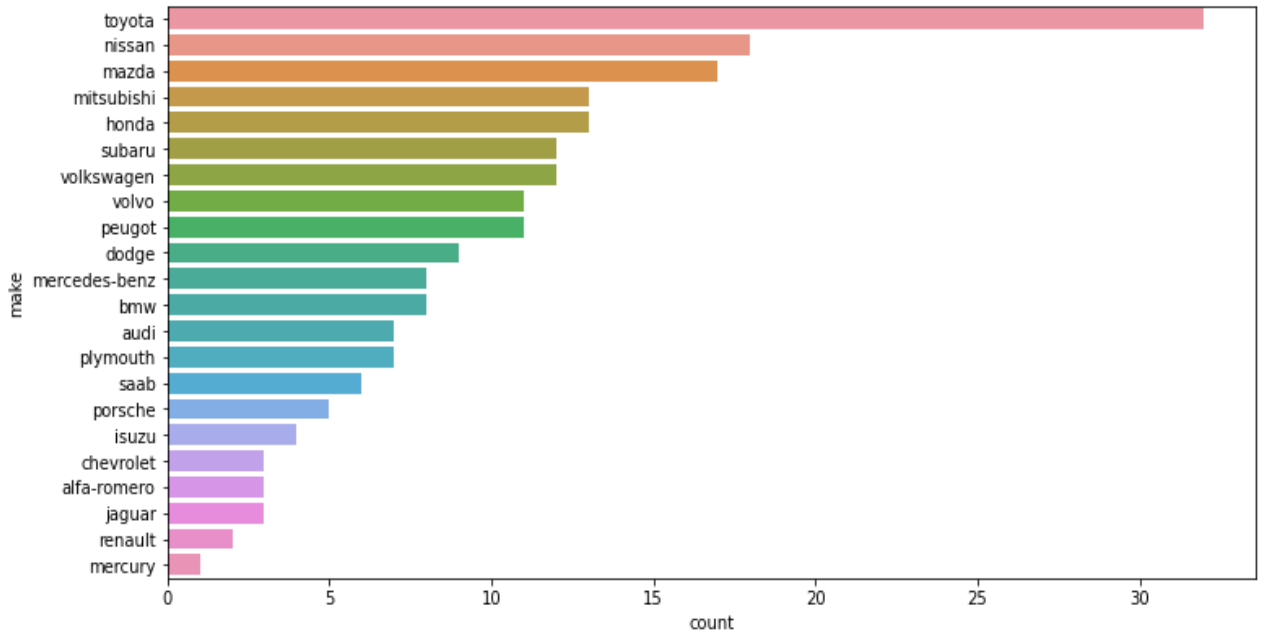
Upon knowledge of the existence of missing values, they were printed out using print(df.isnull().sum()) to see number in each column and potted using missingno.matrix(DataFrame) as shown below;



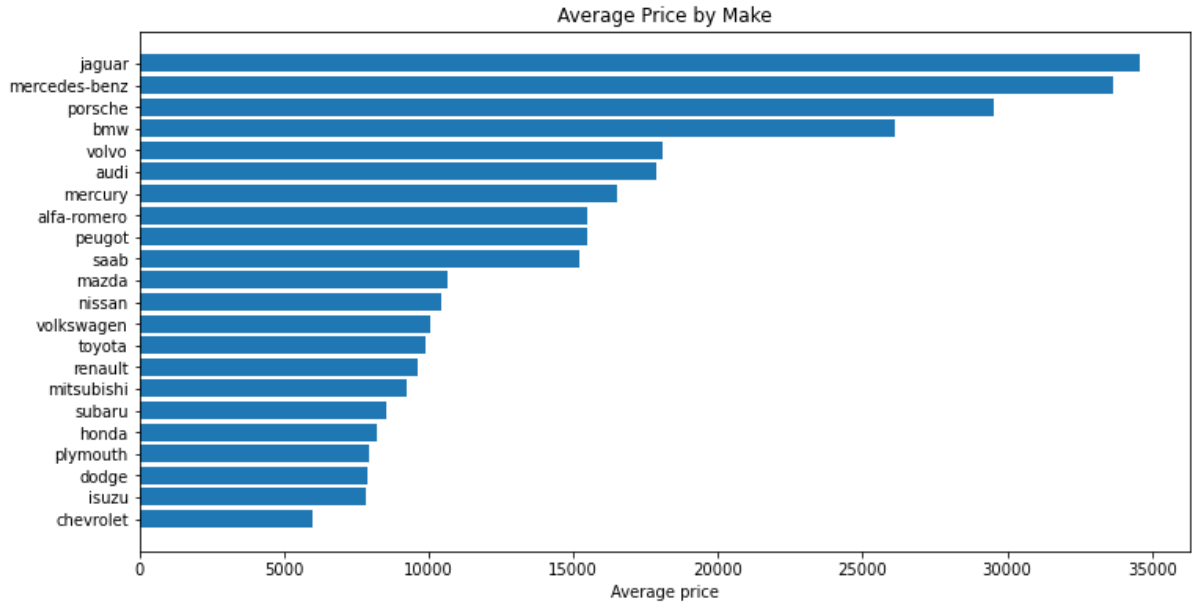
The missing values were thereafter replaced with the median of the variable grouped by car make and body type.

DATA STORIES AND VISUALIZATIONS

First we look at the number of cars by make in the dealership and see that Toyota is leading in terms of count as shown in the figure below;



Is it because Toyota is the cheapest car? To answer this question, we group makes and check average price as below;



We now clearly see that the average price for Toyota cars is probably just below average but there are still way cheaper cars so we conclude the number of cars is probably not correlated to price.

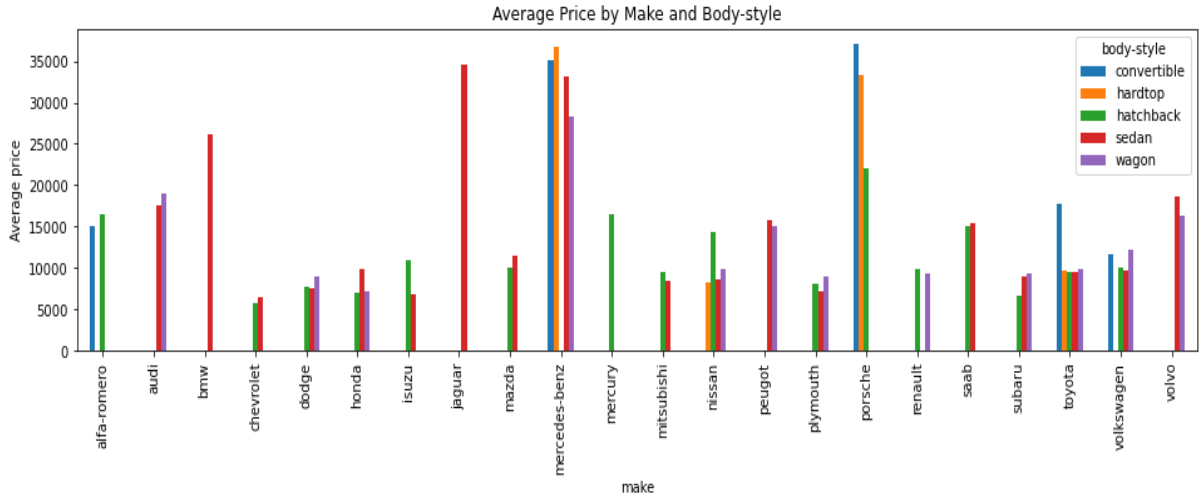
We can further look at how Top 10 most expensive cars compare with Top 10 cheapest cars. We also group by body type to see if it plays a role.

Top 10 most expensive **Top 10 cheapest**

make	body-style	price
mercedes-benz	hardtop	45400.0
bmw	sedan	41315.0
mercedes-benz	sedan	40960.0
porsche	convertible	37028.0
bmw	sedan	36880.0
jaguar	sedan	36000.0
jaguar	sedan	35550.0
mercedes-benz	convertible	35056.0
mercedes-benz	sedan	34184.0
porsche	hardtop	34028.0

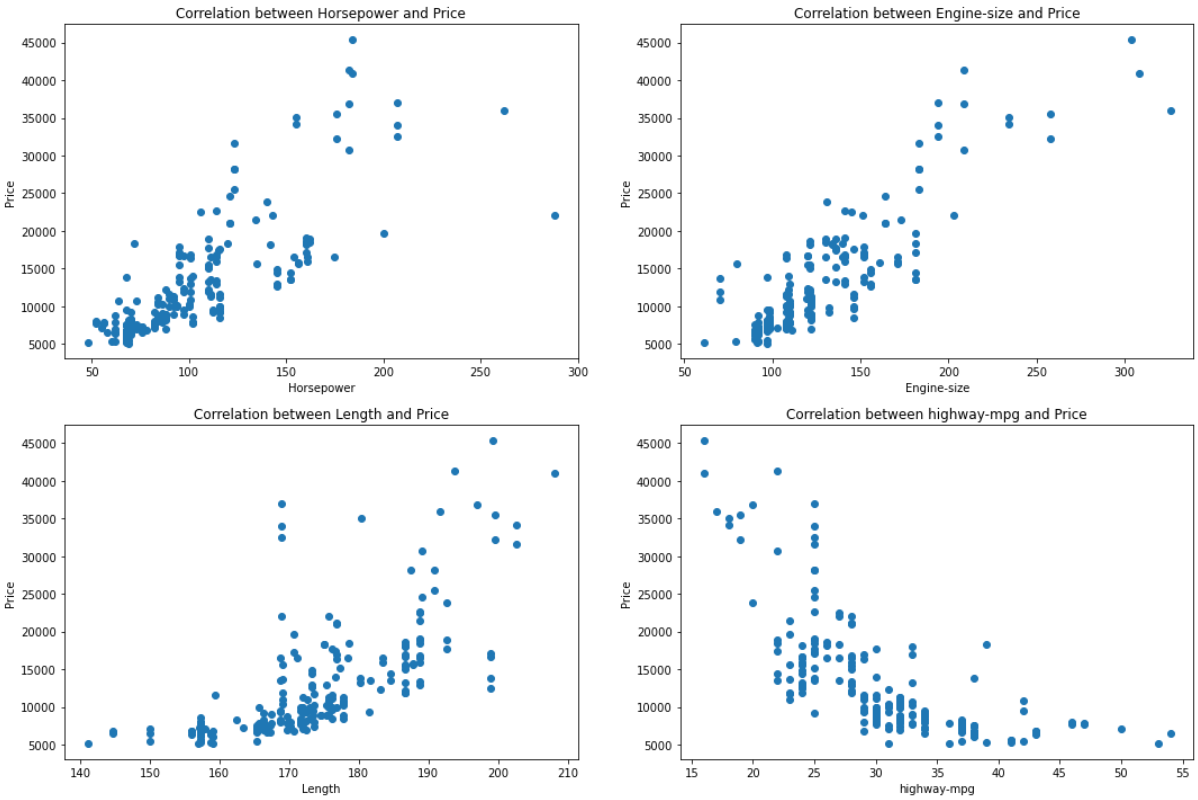
make	body-style	price
subaru	hatchback	5118.0
chevrolet	hatchback	5151.0
mazda	hatchback	5195.0
toyota	hatchback	5348.0
mitsubishi	hatchback	5389.0
honda	hatchback	5399.0
nissan	sedan	5499.0
dodge	hatchback	5572.0
plymouth	hatchback	5572.0
mazda	hatchback	6095.0

From the 2 tables, we see Mercedes-benz is the most is expensive car make and Subaru is the cheapest in terms of price averages. We also see that most cheap cars are hatchbacks and most expensive cars are sedans. Below is a graph of car makes and body types and average prices;



We can see that Toyota has the widest range of body types with average prices highest for converstibles and the rest being about equal.

To conclude, below are some interesting price correlations;



The price increases with horsepower, engine-size and length but decreases with highway-mpg.

THIS REPORT WAS WRITTEN BY : Dalitso Chomey