CHINESE AUTOMOBILE

COMPANY GEELY AUTO PREDICTION OVERVIEW

NAME: ESTHER PAM

CONTRACTED DATA SCIENTIST

OBJECTIVES

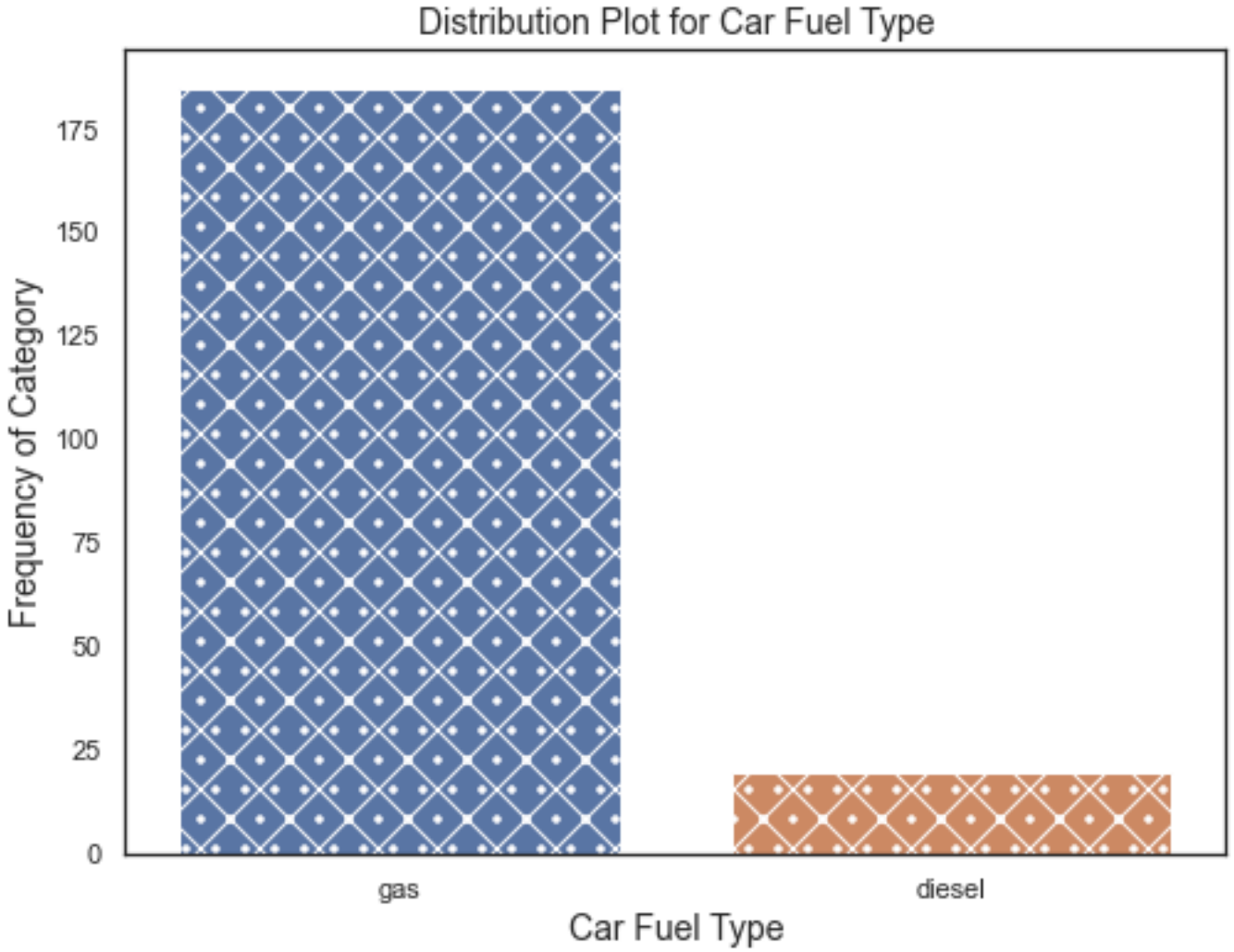
The project is aimed at building a machine learning model to predict the most influential factors to the prices of different car brands in the Nigeria Automobile Industry and to help management accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels.

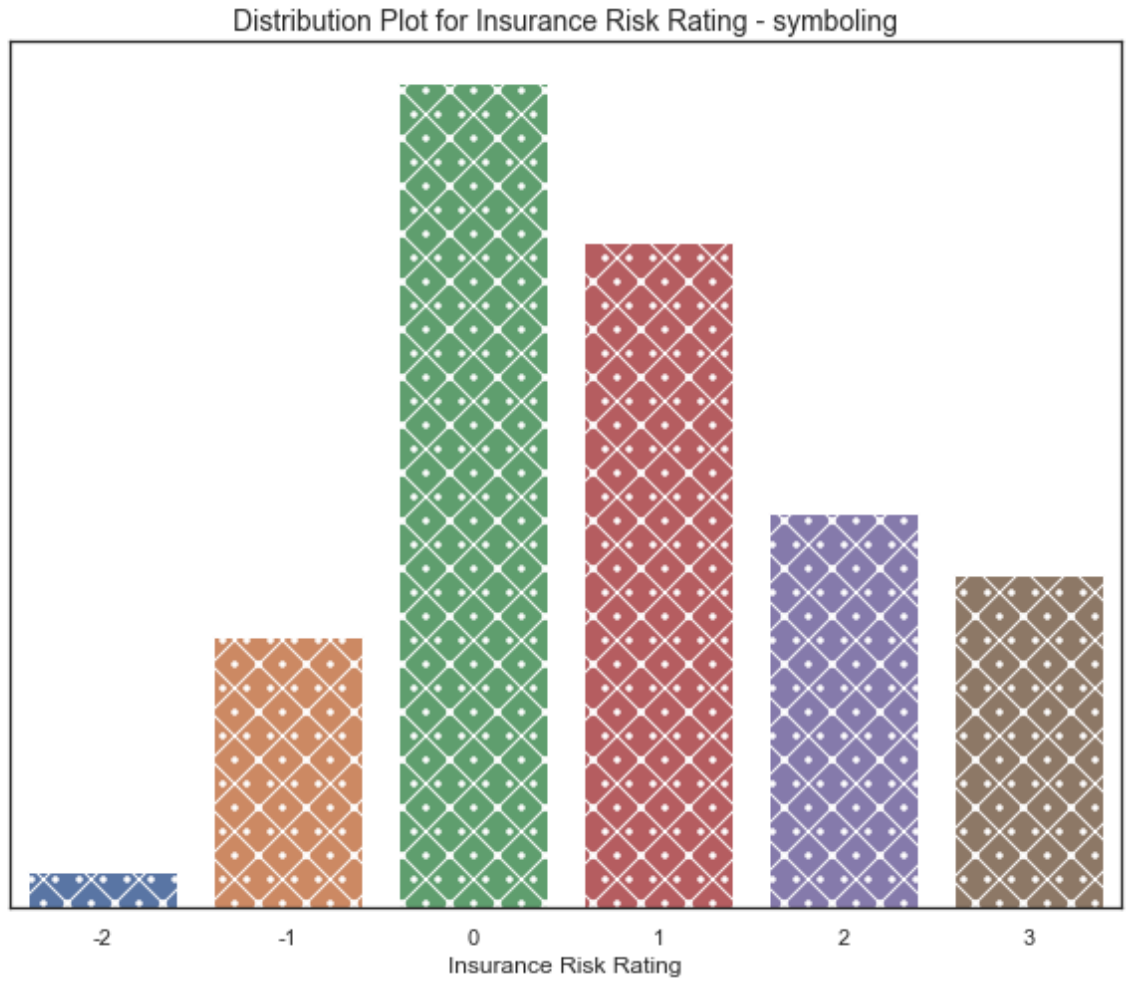
DATA OVERVIEW

- Car_ID - Unique id of each observation (Interger)
- Symboling - Its assigned insurance risk rating,A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
- carCompany - Name of car company (Categorical)
- fueltype - Car fuel type i.e gas or diesel (Categorical)
- aspiration - Aspiration used in a car (Categorical)
- doornumber - Number of doors in a car (Categorical)
- carbody - body of car (Categorical)
- drivewheel - type of drive wheel (Categorical)
- enginelocation - Location of car engine (Categorical)
- wheelbase - Weelbase of car (Numeric)
- carlength - Length of car (Numeric)
- carwidth - Width of car (Numeric)
- carheight - height of car (Numeric)
- curbweight - The weight of a car without occupants or baggage. (Numeric)
- enginetype - Type of engine. (Categorical
- cylindernumber - cylinder placed in the car (Categorical)
- enginesize - Size of car (Numeric)
- fuelsystem - Fuel system of car (Categorical)
- boreratio - Boreratio of car (Numeric)
- stroke - Stroke or volume inside the engine (Numeric)
- compressionratio - compression ratio of car (Numeric)
- horsepower - Horsepower (Numeric)
- peakrpm - car peak rpm (Numeric)
- citympg - Mileage in city (Numeric)
- highwaympg - Mileage on highway (Numeric)
- price(Dependent variable) - Price of car (Numeric)

UNIVARIATE ANALYSIS

UNIVARIATE ANALYSIS - DISTRIBUTION OF CAR FUEL TYPE AND SYMBOLLING



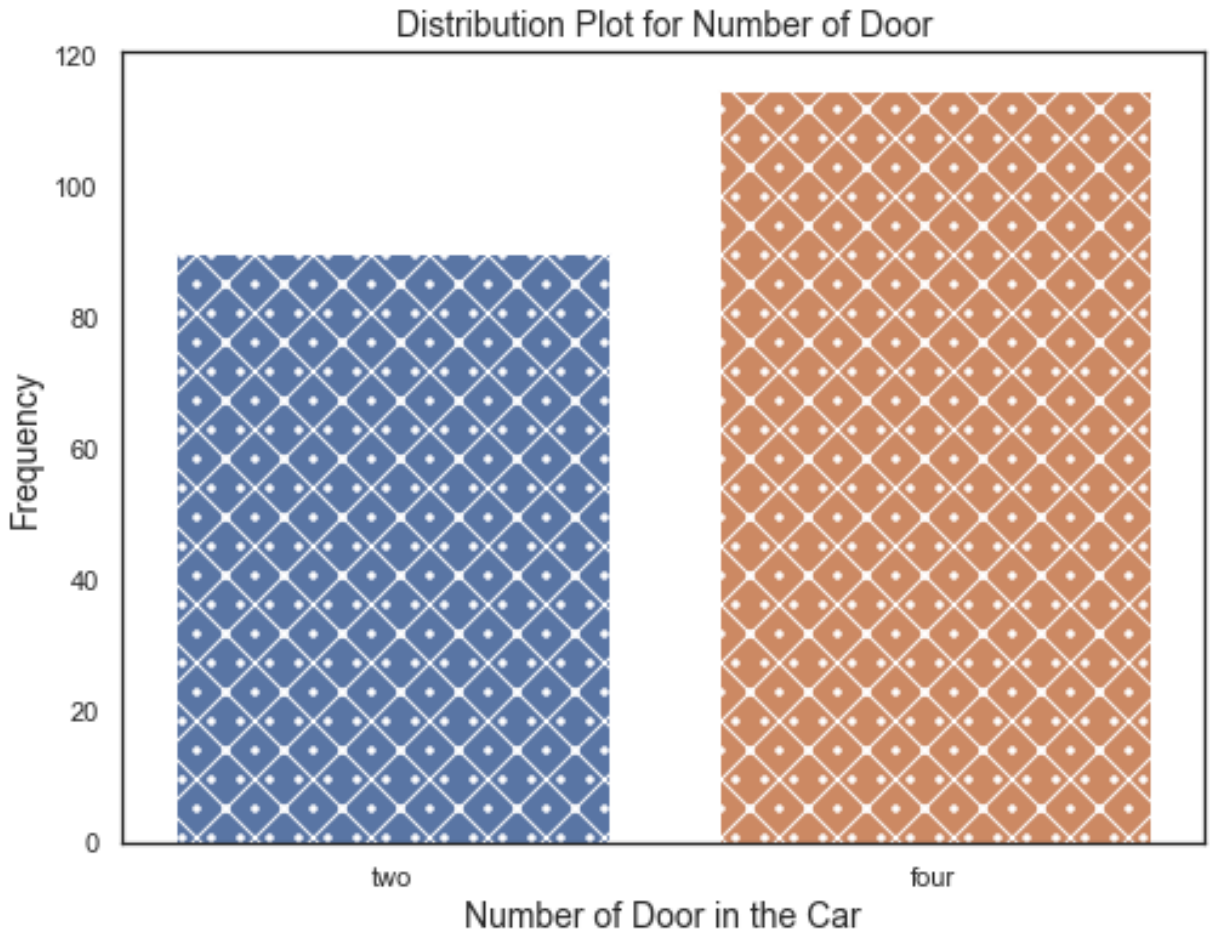


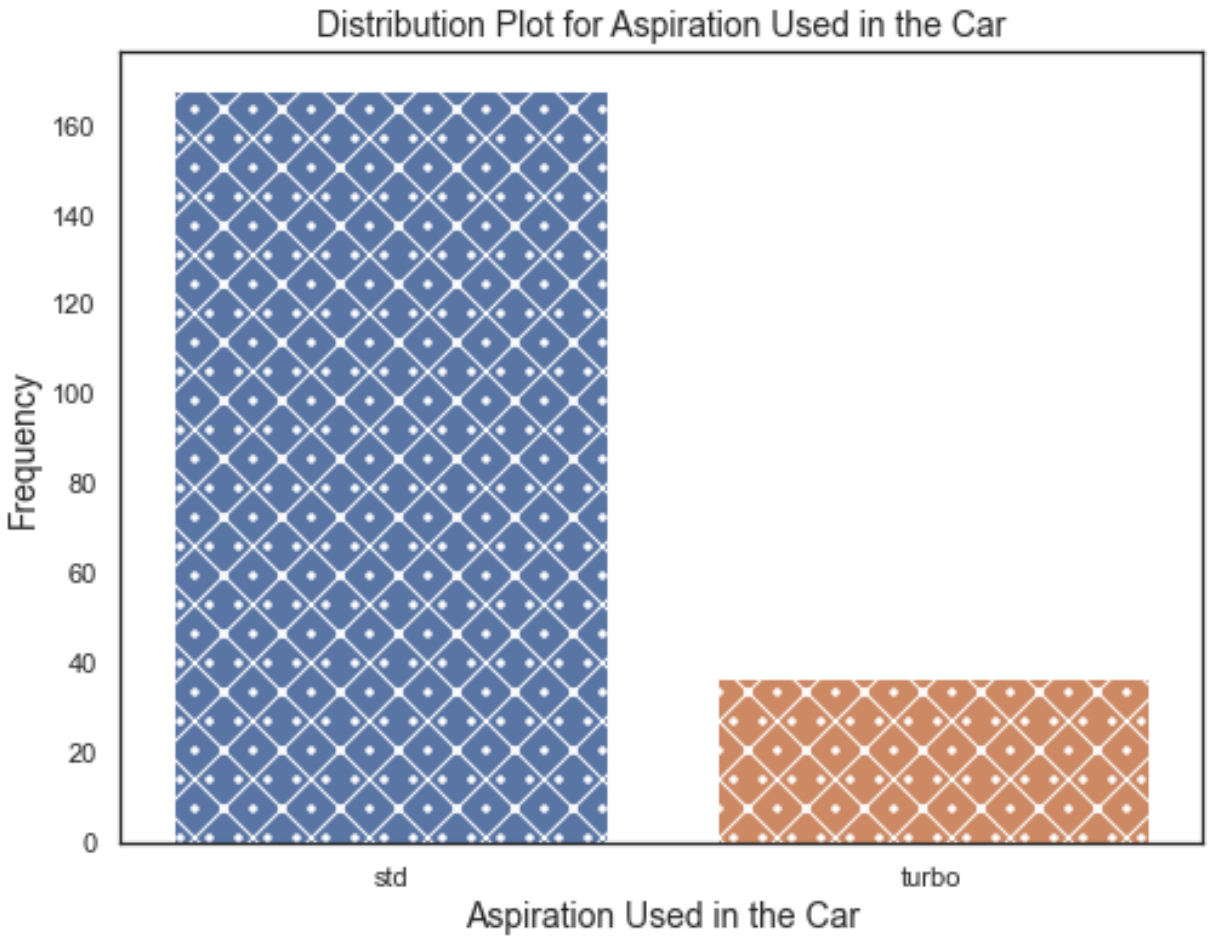
OBSERVATION

- From the distribution plot of car fuel type, it was seen that more car type purchased in the car are fuel gas powered cars rather than diesel as the plot shows a significant difference.

- To the insurance risk rating plot, it is seen that major of the car produced has a IRR of greater +1 which is risky as compared to the number of car -1 that are safe as the data shows a margin level greater than 70% as risky.

UNIVARIATE ANALYSIS - DISTRIBUTION OF NUMBER OF DOORS AND ASPIRATION FOR THE PURCHASE



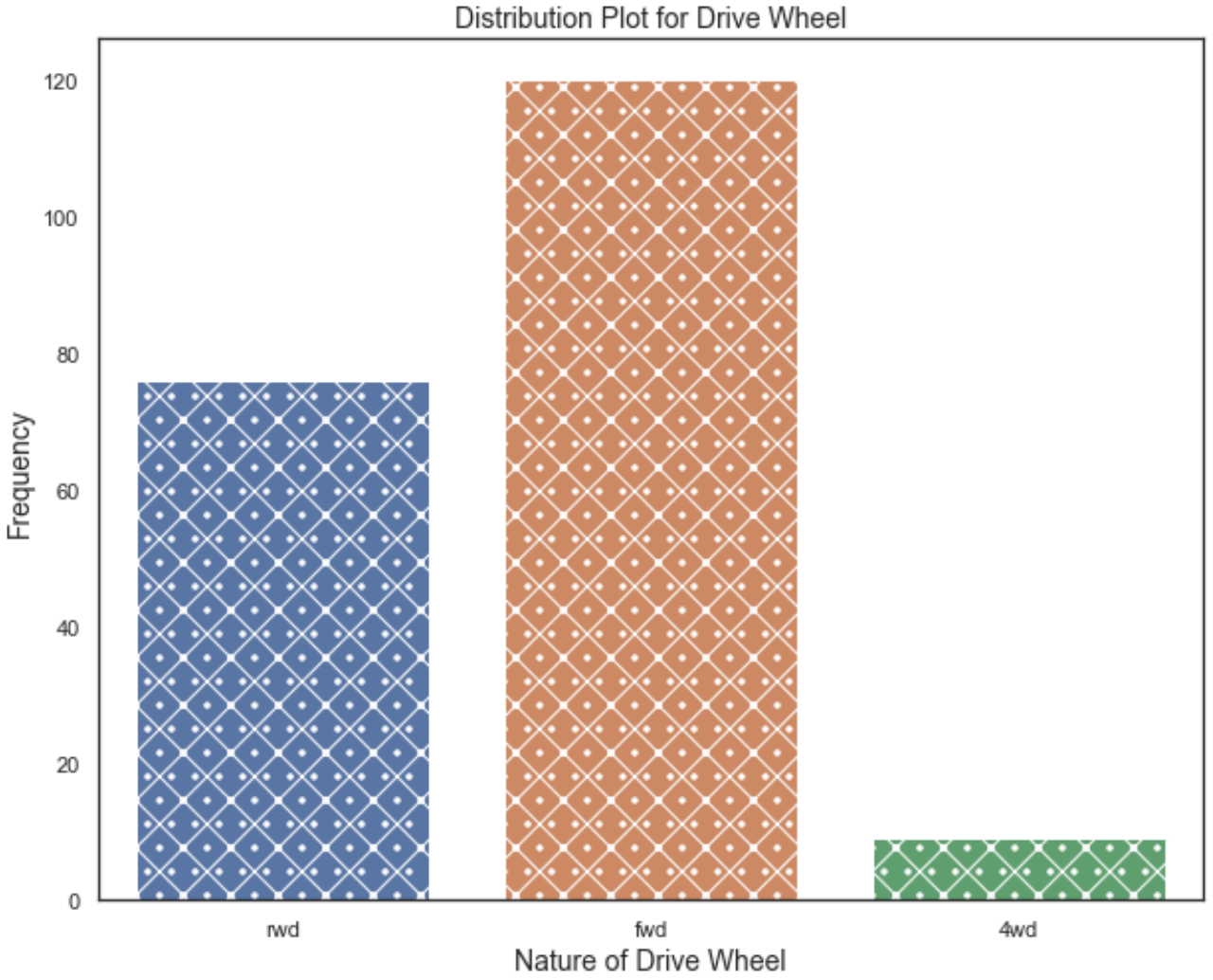


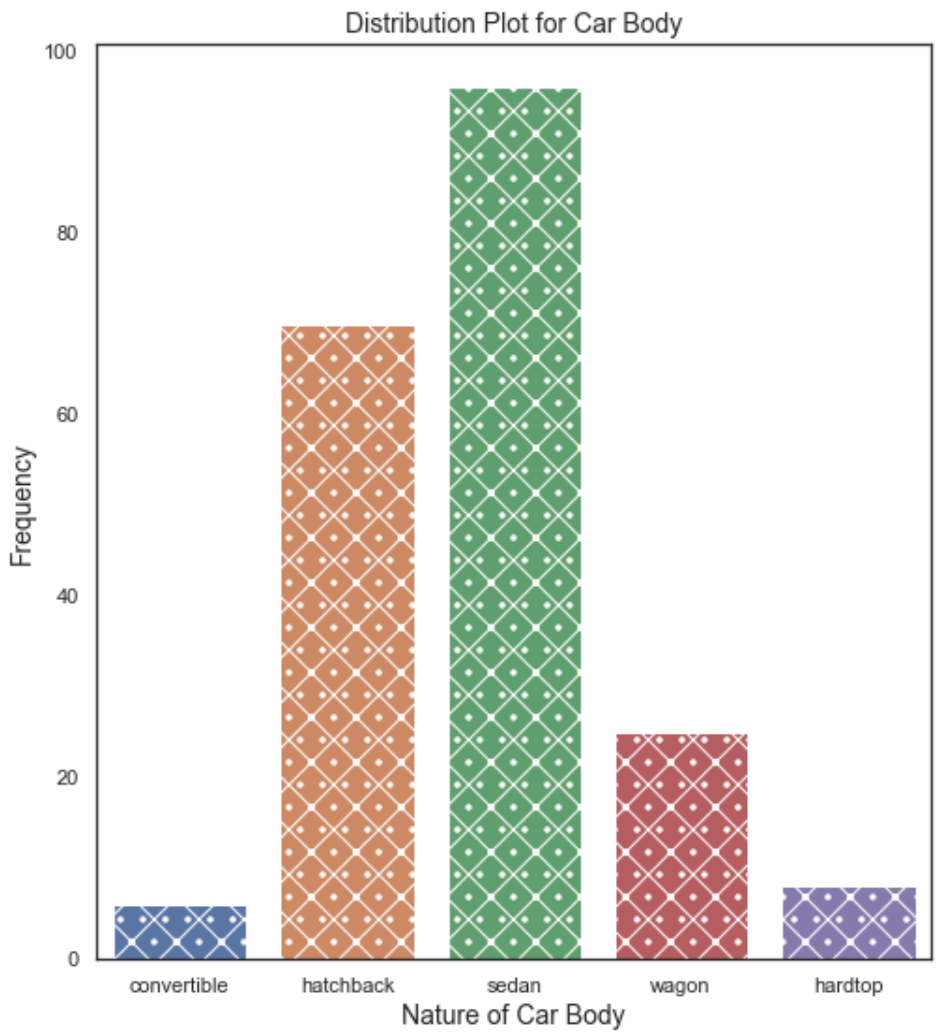
OBSERVATION

- On the number of car produced with doors, the data shows that majority of the cars are usually 4 doors automobile as there is little difference between both four doors and two doors.

- Aspiration of most cars produced has a std value rather than a turbo value which is has a large significant difference in both categories.

UNIVARIATE ANALYSIS - DISTRIBUTION OF DRIVE WHEEL AND CAR BODY



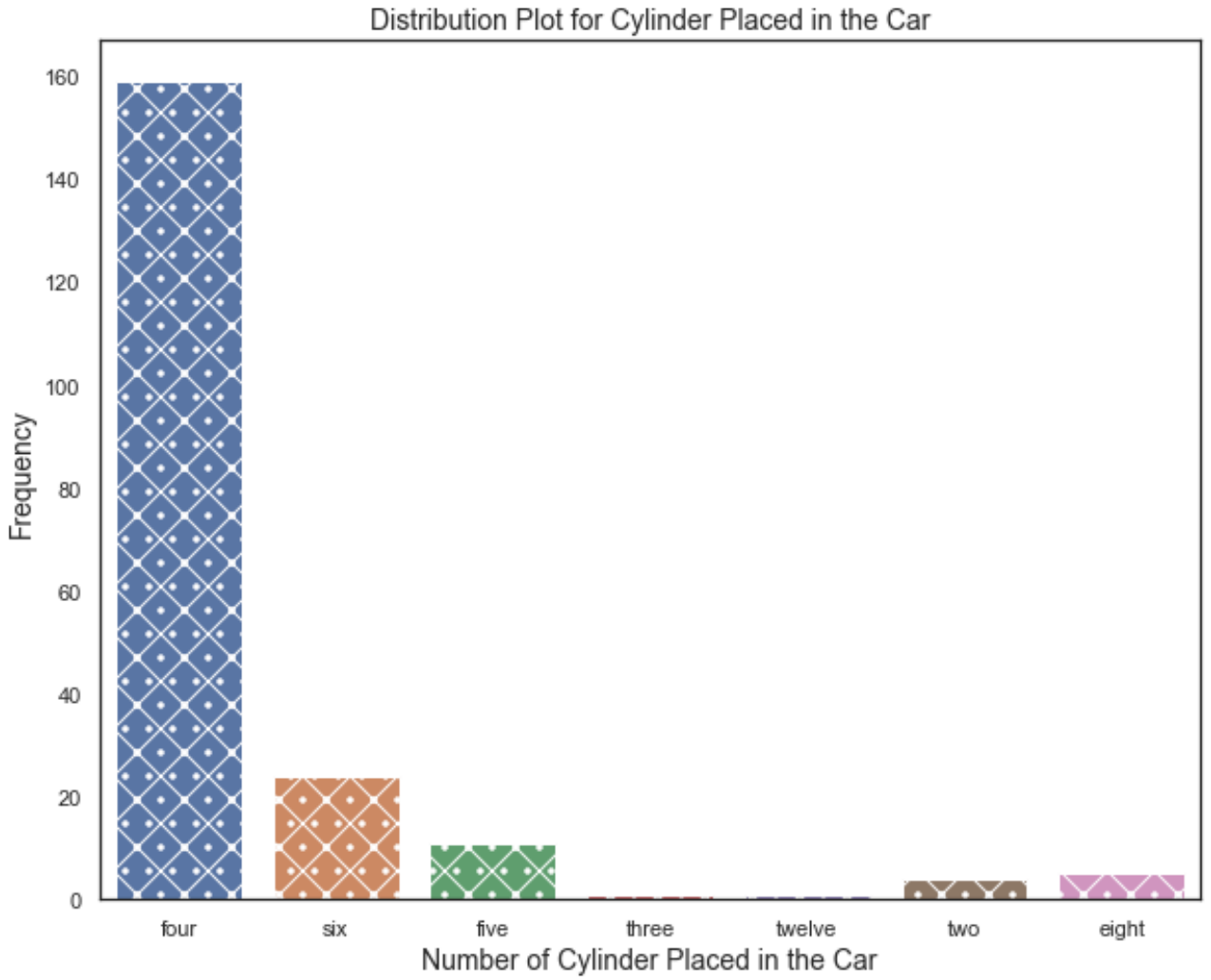


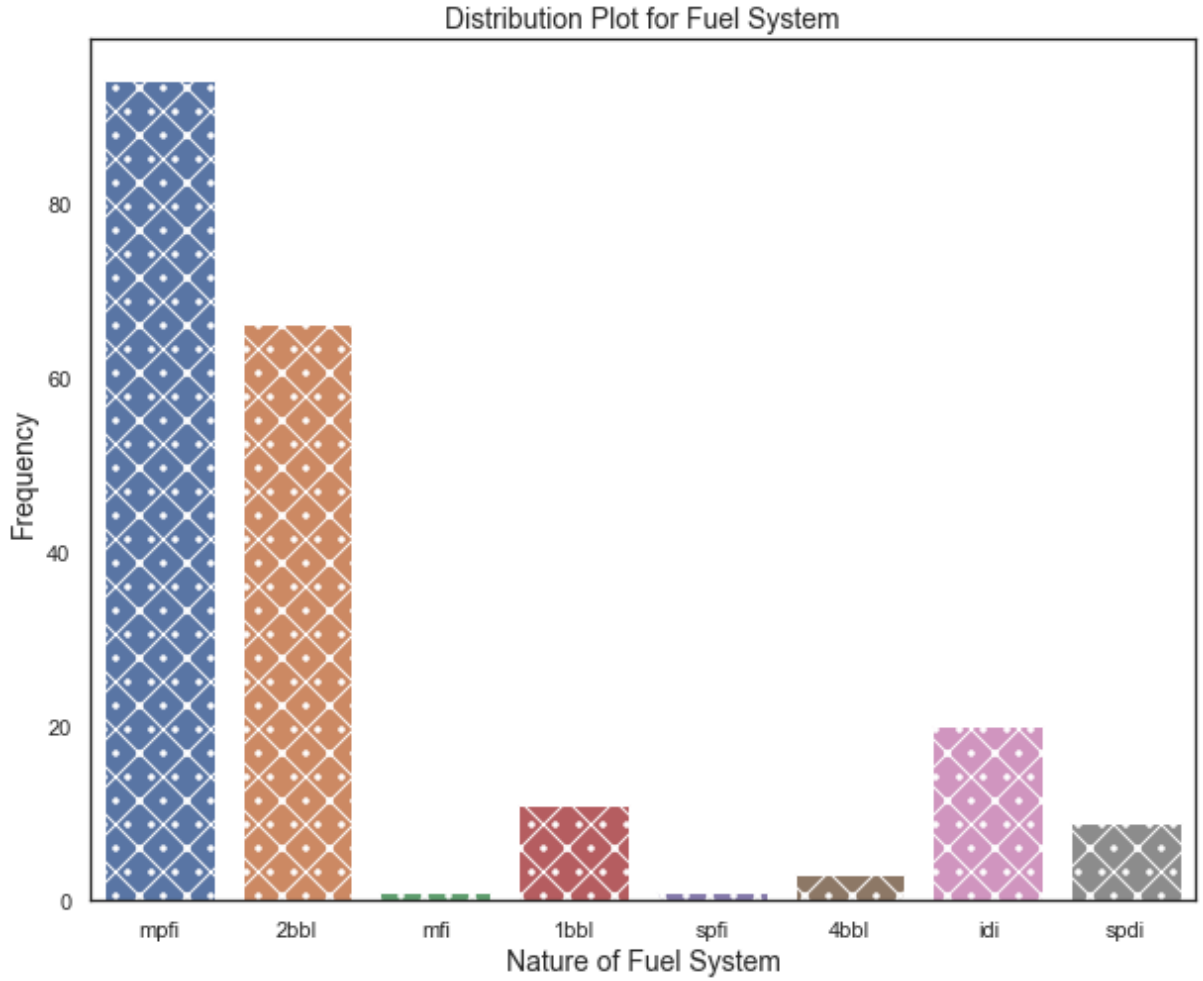
OBSERVATION

- To the drive wheel, it was observed that more than 50% of the cars are fwd wheel, followed by rwd and 4wd as the plot show a difference gap in the bars.

- The nature of the car body shows that sedan motors was predominant in the markets followed by hatchback, wagon, hardtop and convertible which is least. This shows a actual trend of car used and purchase in the markets.

UNIVARIATE ANALYSIS - DISTRIBUTION OF CYLINDER IN CARS AND FUEL SYSTEM





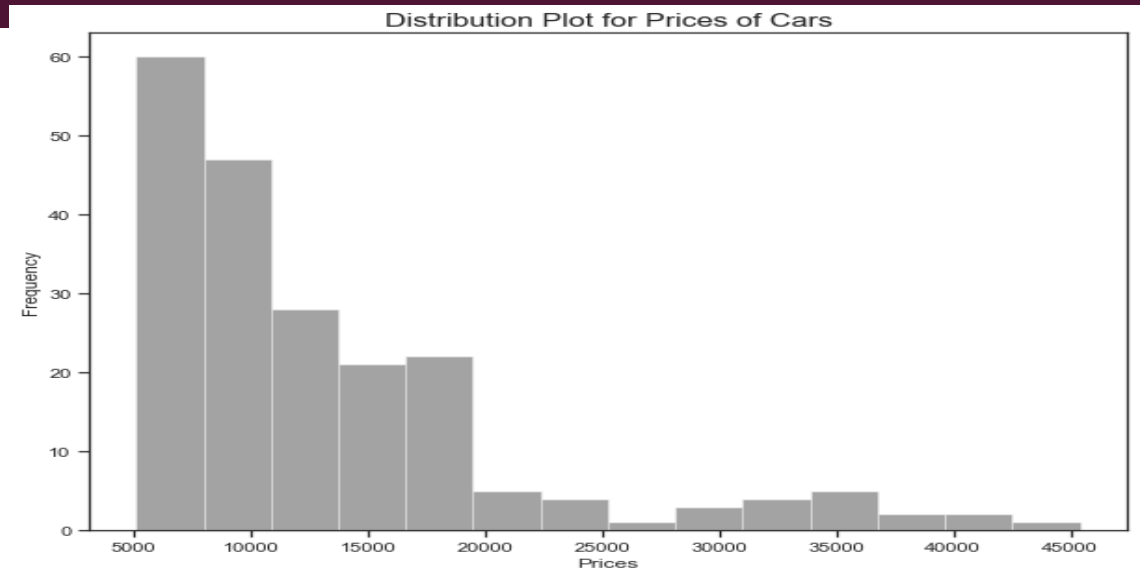
OBSERVATION

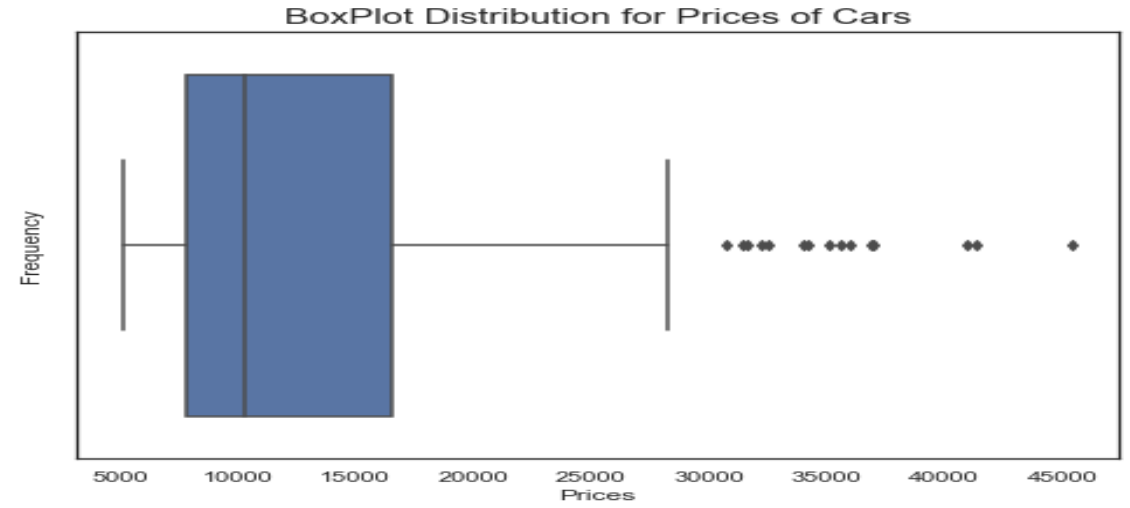
- The graph to number of cylinder placed in the cars shows that majority of the car has four cylinders placed in them as compared to three and twelve which is a rare scenario for most cars.

- To the nature of fuel system, it can be seen that most cars have a mpfi and 2bbi fueling system to the counterpart of nfi and spfi which exist on rare occasion with other fueling system such as 1bbl, 4bbl, idi and spdi.

UNIVARIATE ANALYSIS - DISTRIBUTION OF PRICES OF CARS.

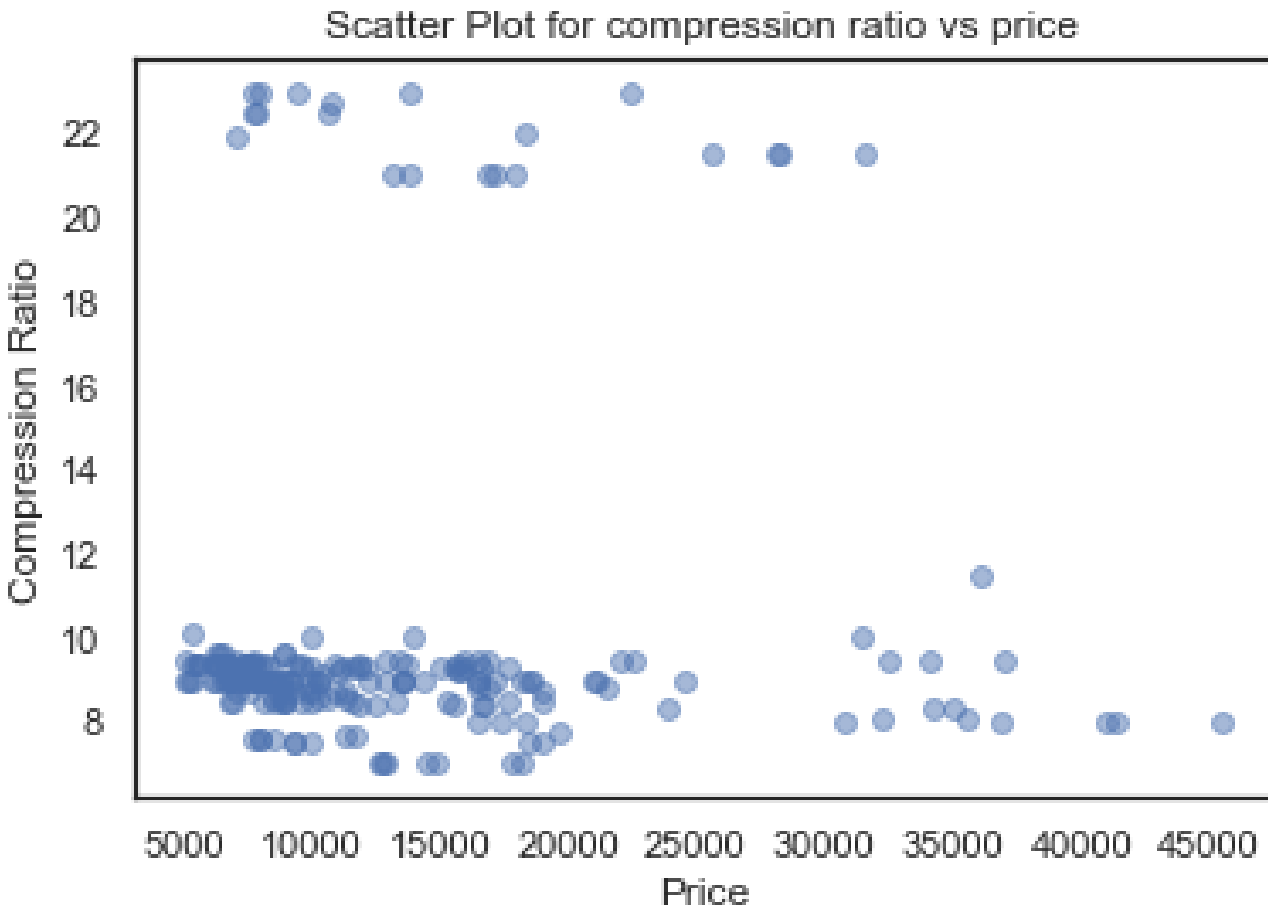
The price plot shows a trend in most purchase of cars as most cars purchased were between the prices of 5000 – 10, 000. This is important feature to be noted as cheap cars sells more as compared to expensive cars in the range of 40000 – 45000.

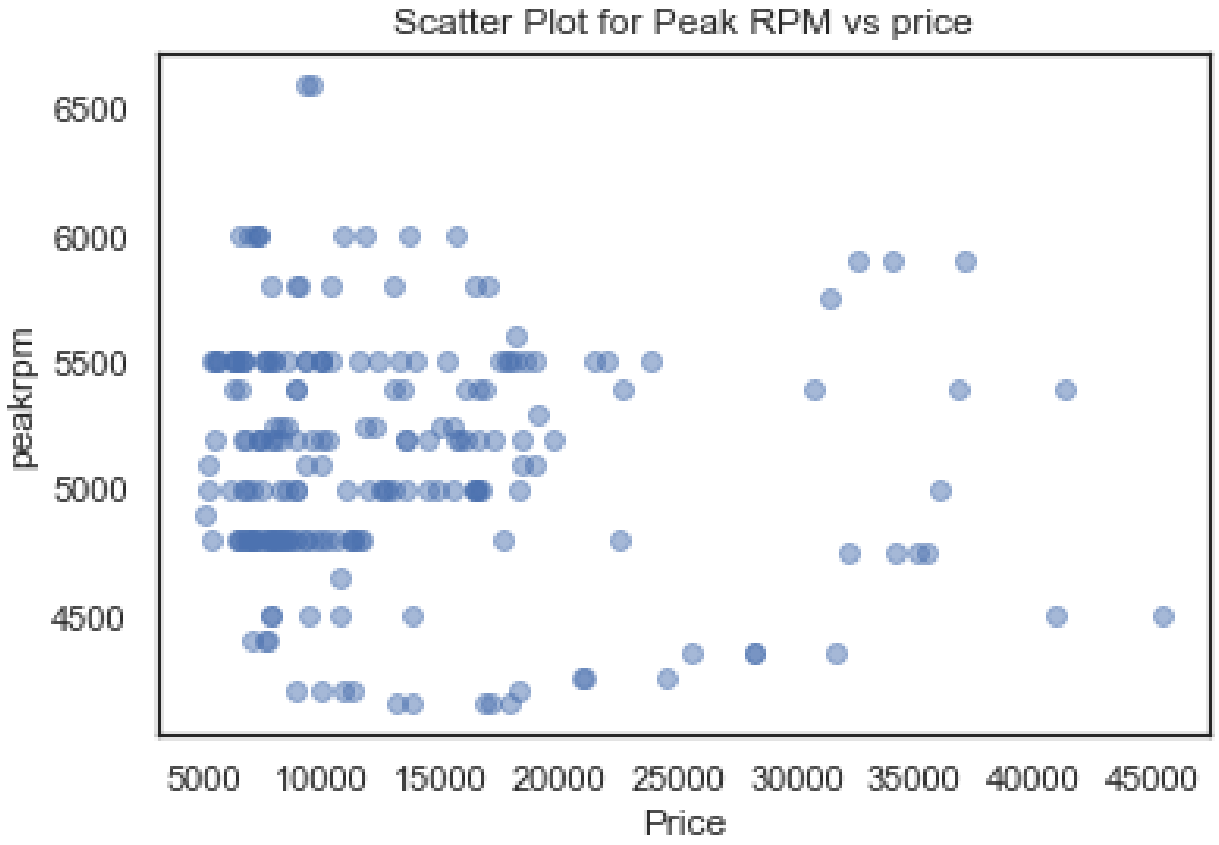




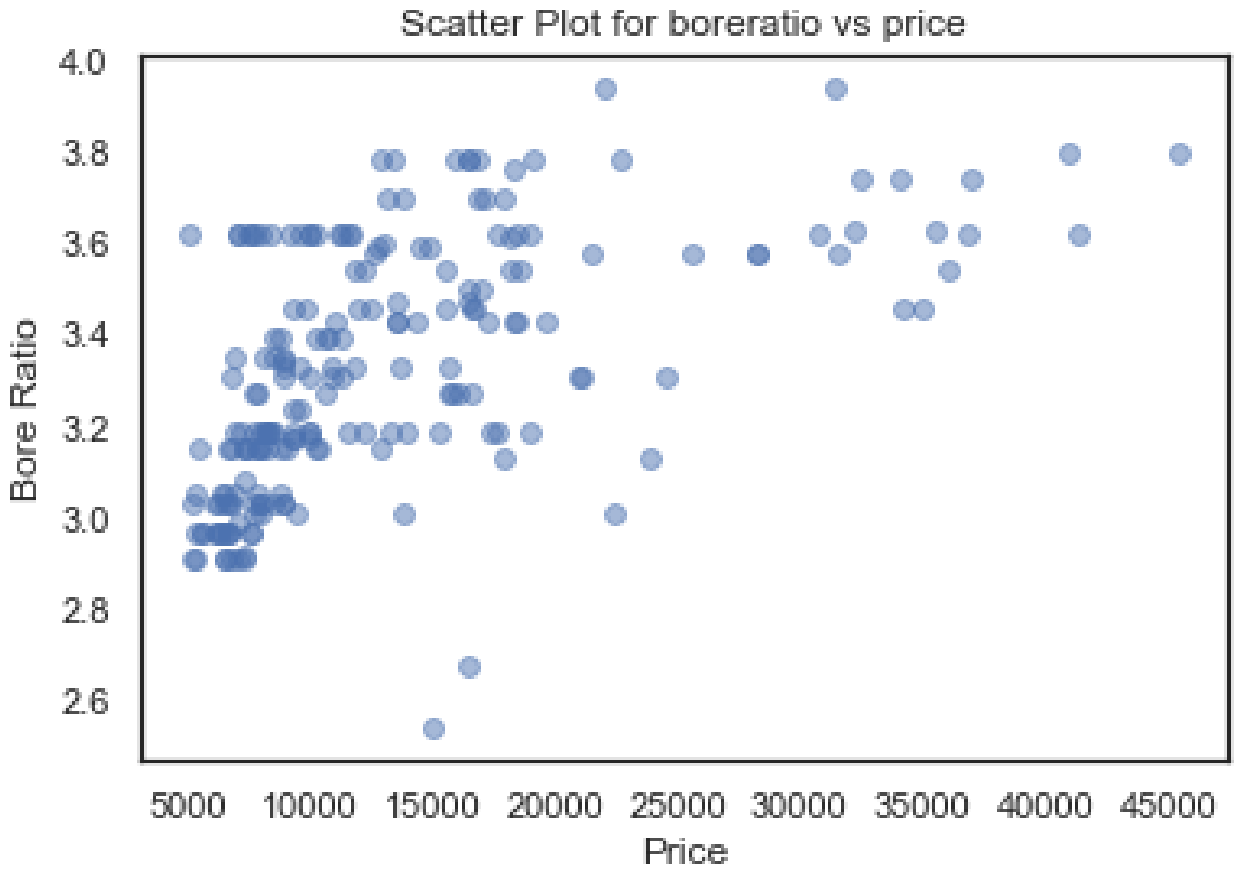
BIVARIATE ANALYSIS

BIVARIATE ANALYSIS

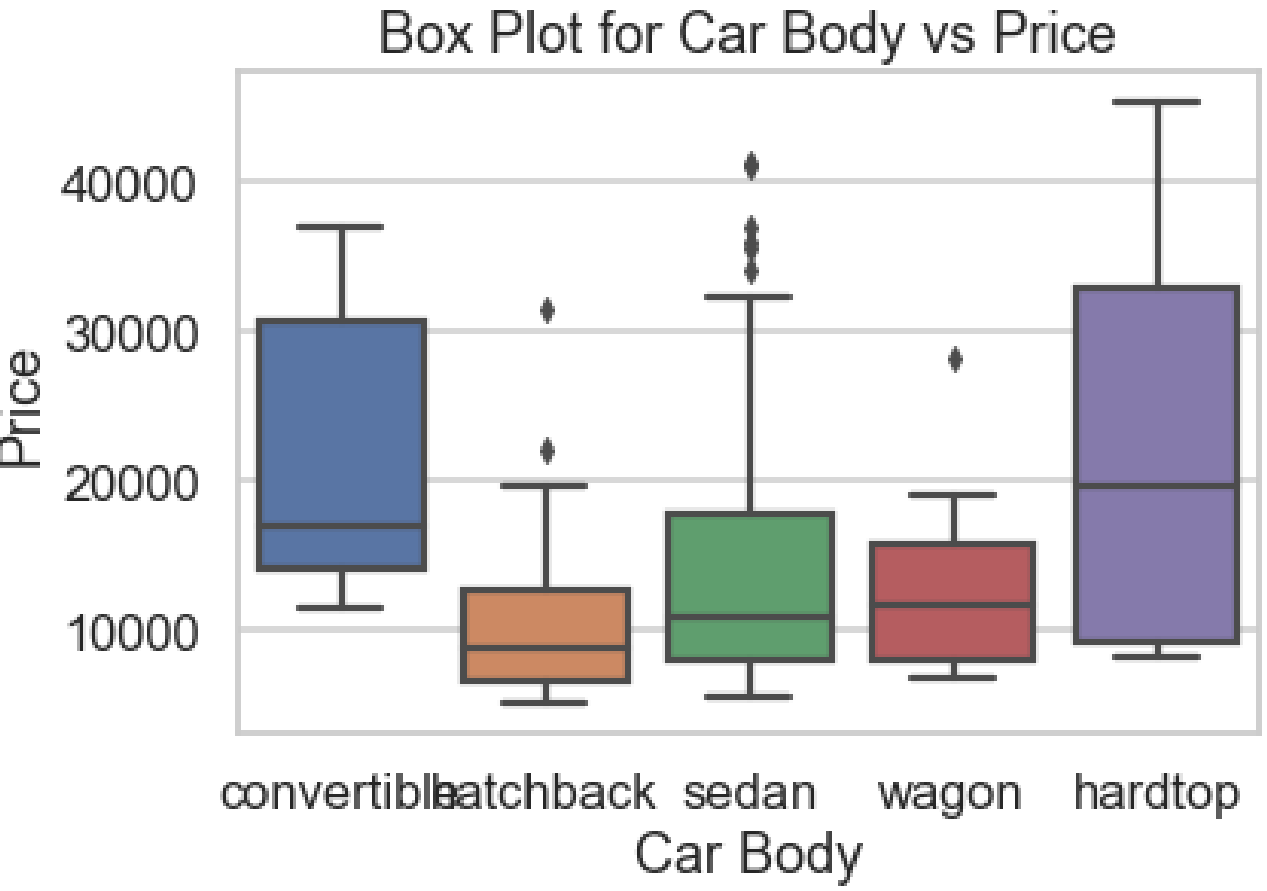


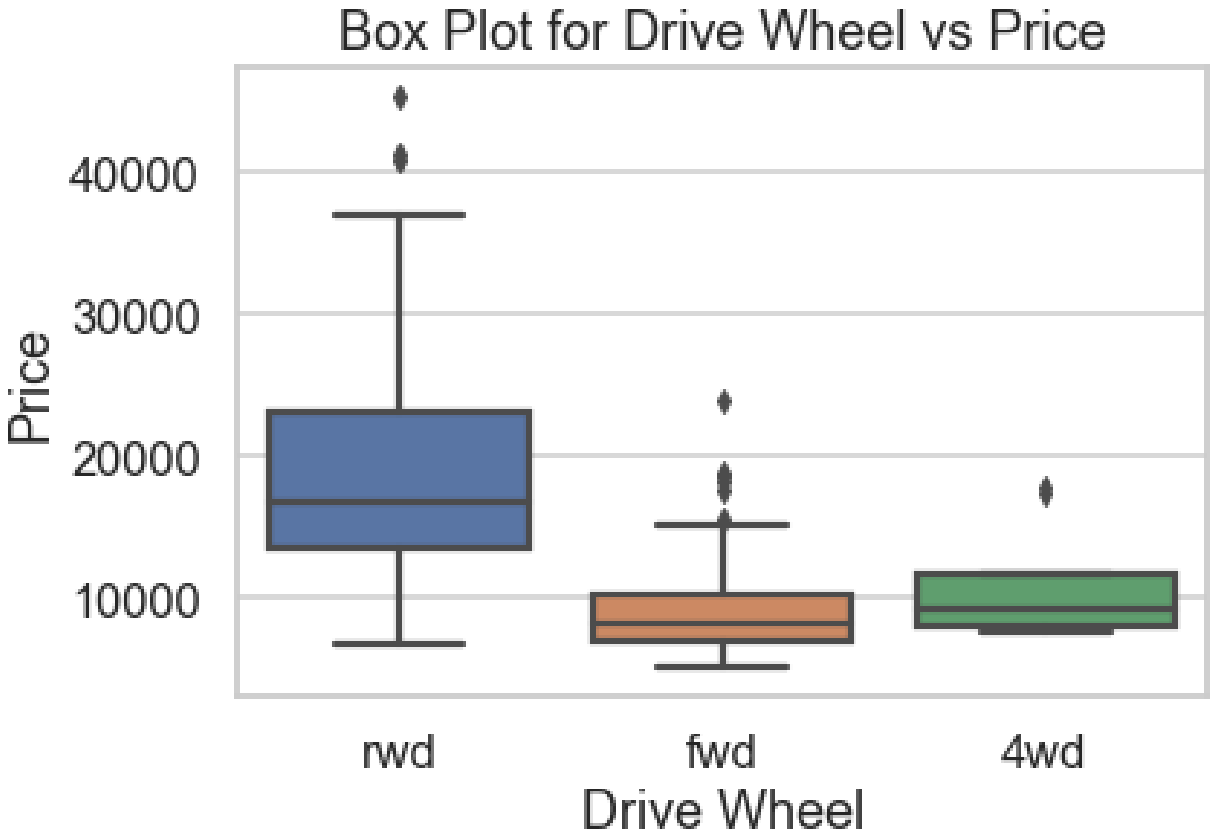


BIVARIATE ANALYSIS



BIVARIATE ANALYSIS





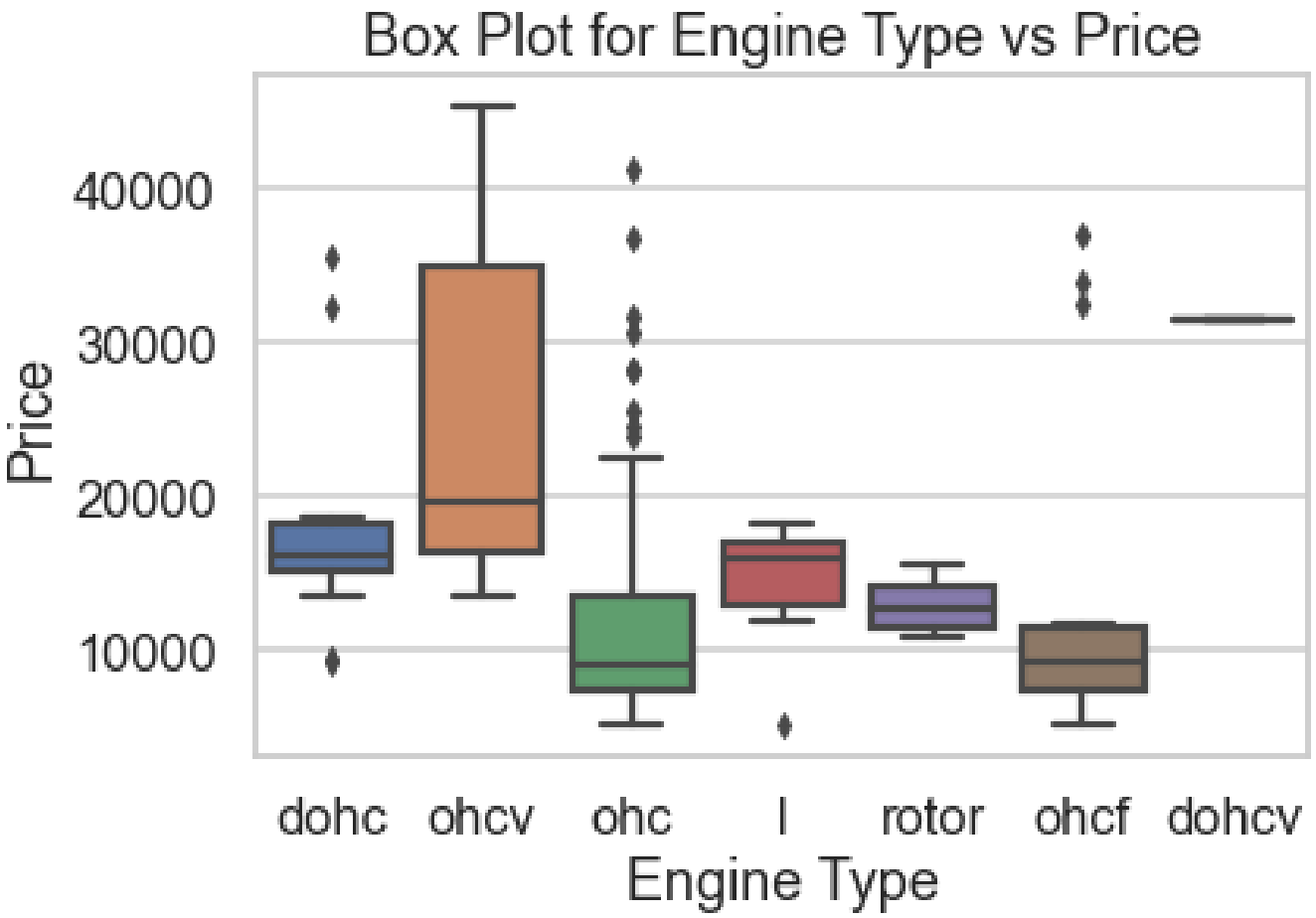
OBSERVATION

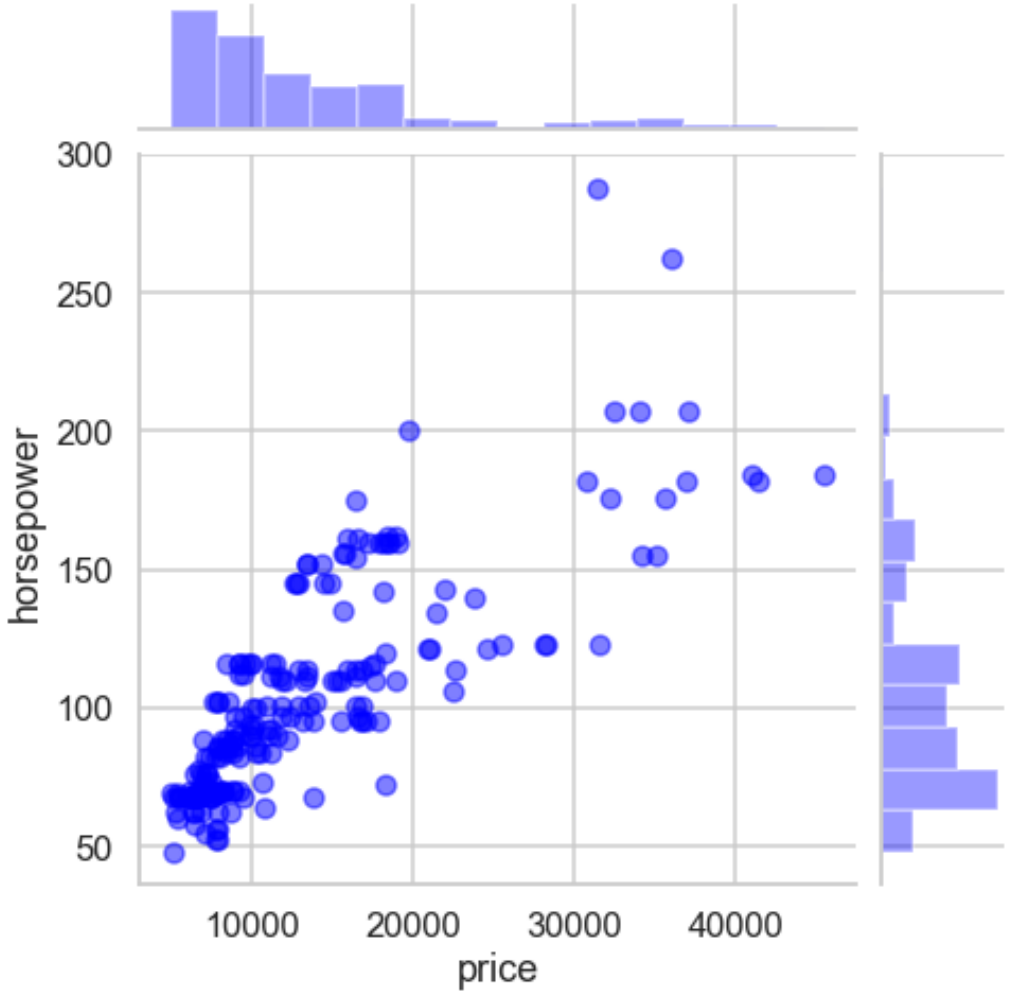
- From the comparison of compression ratio to price, it was observed that there is a concentration in the lower region around 8 - 10 of the compression ratio to 5000 - 20000 of the price with few observation in upper region of compression ratio of 22.

- For the comparison of peakrpm and price, the exist a high density between 4800 and 5500 in peakrpm and 5000 to 15000.

- Other variables shows a significant behaviour in comparison to price such as engine size, car body, bore ratio and drive wheel.

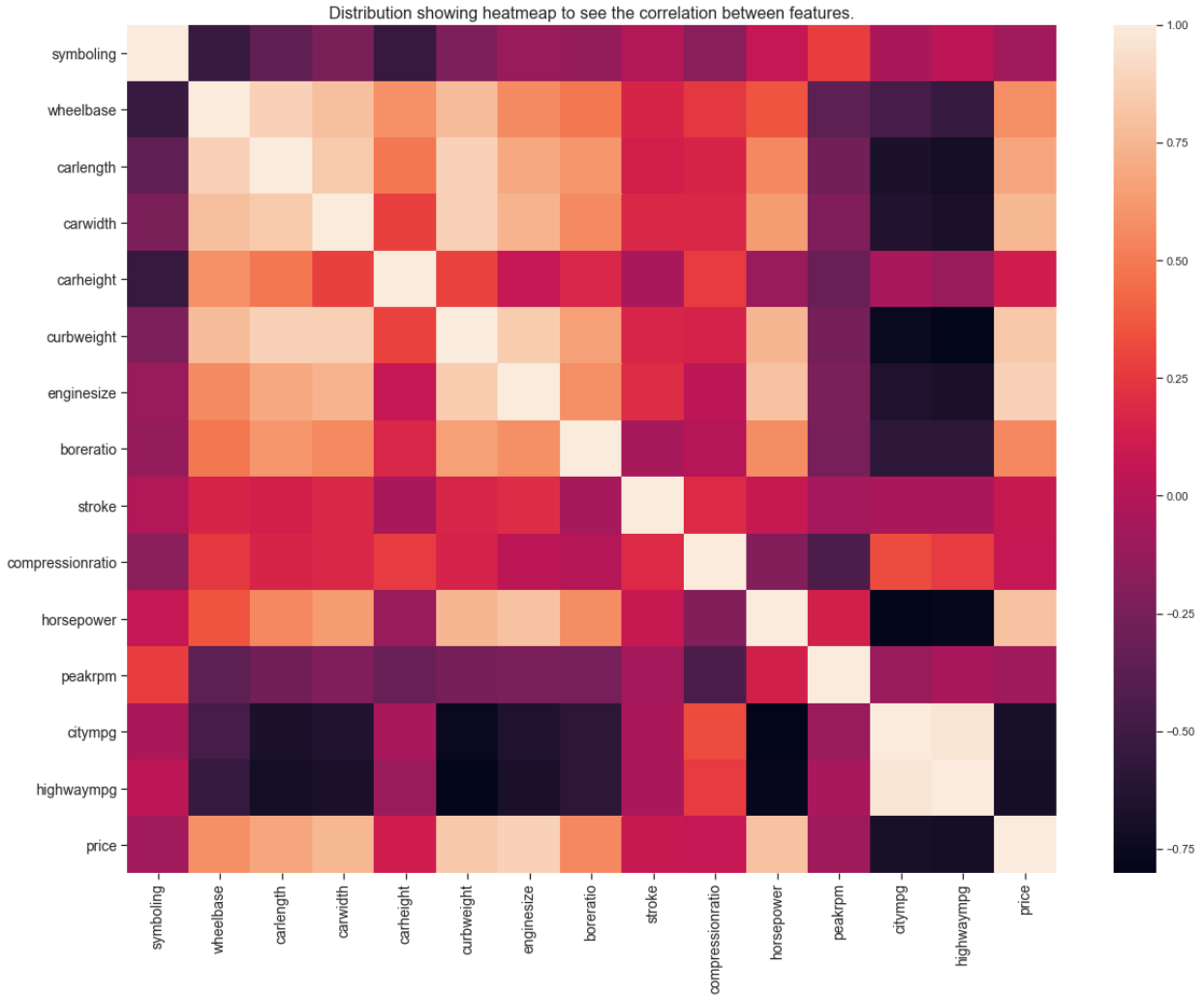
BIVARIATE ANALYSIS

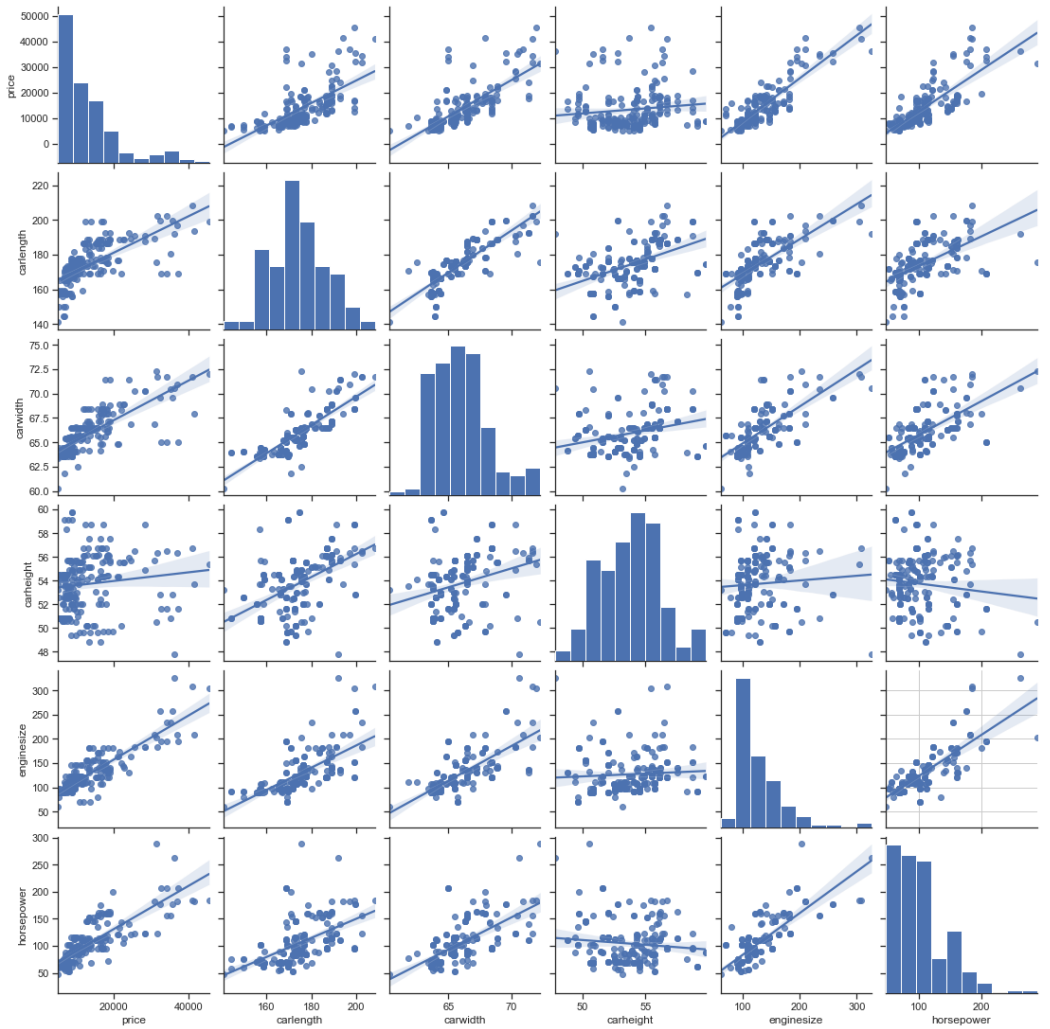




MULTIVARIATE ANALYSIS

MULTIVARIATE ANALYSIS





OBSERVATION

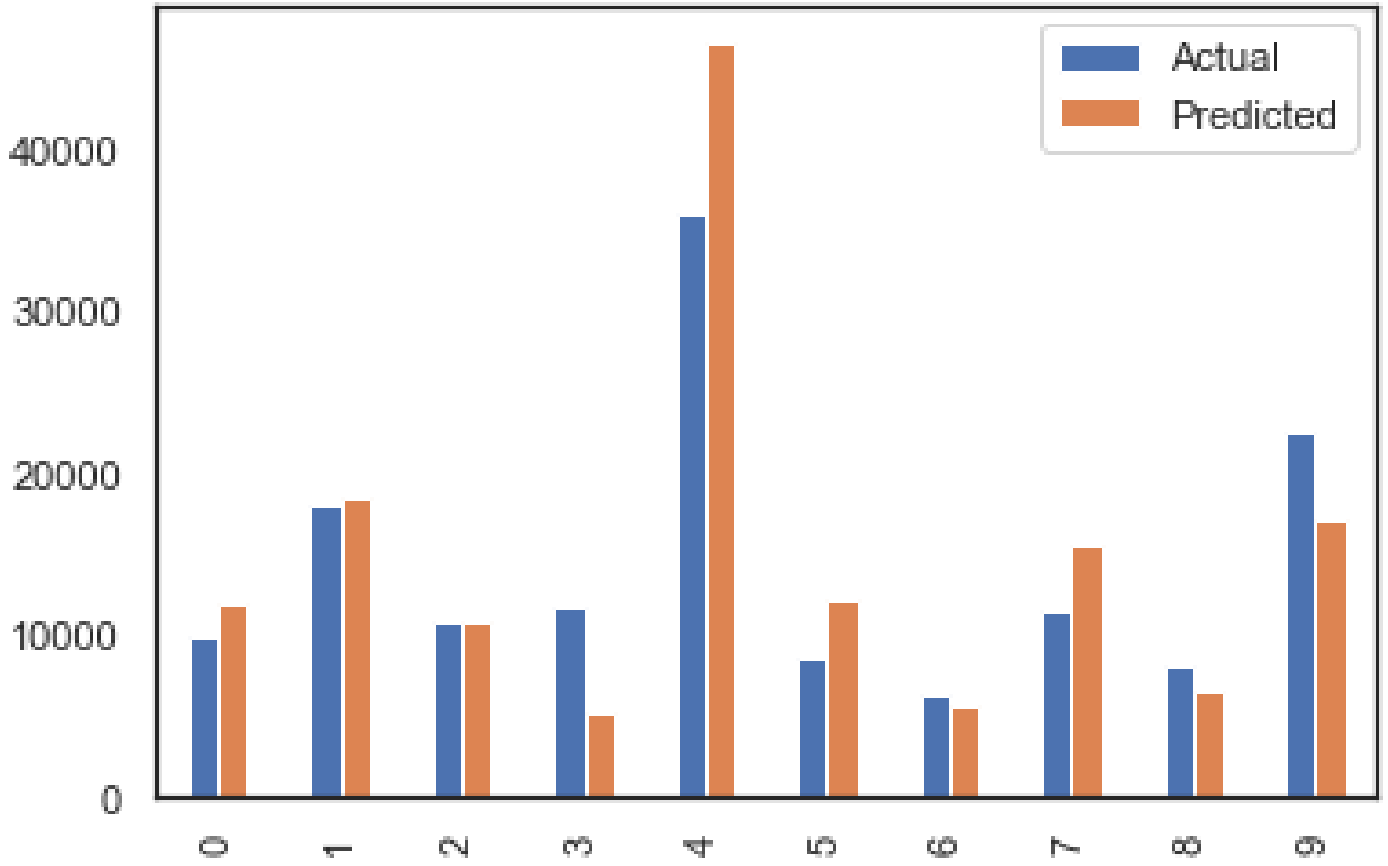
- From observation, the heatmap shows a strong correlation between enginesize to horsepower, and price to enginesize and also price to carwidth and weak correlation between price to highwaympg and citympg.

NB: A coefficient close to 1 means that there's a very strong positive correlation between the two variables. The diagonal line is the correlation of the variables to themselves – so they'll, of course, be 1.

MULTIPLE LINEAR REGRESSION ANALYSIS

OBSERVATION

- Note that most of the predicted values are higher than the actual values



ACTION AND RECOMMENDATION

- Multiple linear regression model was employed for the prediction of the variable dependent on influencing the price of car. The model gave a regression score of 63% with queried with the entire feature against prices as compared to a selection of dependent variable which give a regression score of 71%. That is to say, with proper investigation of individual variable, a more accurate regression score will be obtained and compared among each other to understand the dependent with more influence to the price of car in the Nigeria market.

ACTION AND RECOMMENDATION CONT'D

- OLS model was also implemented to obtain the p-value and coef which from the results, it was observed that certain features plays a significant role in sales in the Nigeria markets such as aspiration, carheight, carwidth, boreratio, highwaympg, cylindernumber, enginesize and others with low p-value such as horsepower, highwaympg, wheelbase, enginetype, enginelocation and boreratio.
- Concentration needs to be given to the doornumber, carbody, stroke and symboling as these feature if not properly handle could cost the company money in relation to competition among other car production company.

THANK YOU

VERY MUCH!!!.