Data Preparation:

- I have loaded the dataset from local storage as “Car_data” and checked the data is properly loaded or not using **head (), shape & dtypes** function to check sample records, number of observations & data types of each columns. And then checked for following types of potential issues/errors :

    - **Checking for Typos:** I have made the list of categorical columns which need to be checked for typos and then passed this list in for loop for calling **value_counts ()** function to check frequency of each values. There were typos in terms of values, case-sensitivity of values and whitespaces.

        - To remove those, I have used “.str.**lower ()**” to convert those values into a lowercase.
        - To replace the false value with correct values, “**replace_string**” function is defined and passed the false values and true values as an arguments to function using lambda to pass column. I have replaced **‘vo100112oo’** with **‘volvo’** in **Make**, **‘turrrrbo’** with **‘turbo’** in **Aspiration** and **‘fourr’** with **‘four’** in Num-of-doors.
        - To remove whitespace, I have used “.str.**strip ()**” to trim those values. And then checked for final value counts of each columns to verify data.

    - **Checking for Sanity Checks:** I have created a masking variables to check impossible values for price & symbolling to check if the price is 0 or negative and if the values of symboling are out of range -3 to +3.

        - For values 0 of price, I have removed those records using **masking** variable.
        - For values outside of the range in symboling, I’ve replaced the values with median **0**.

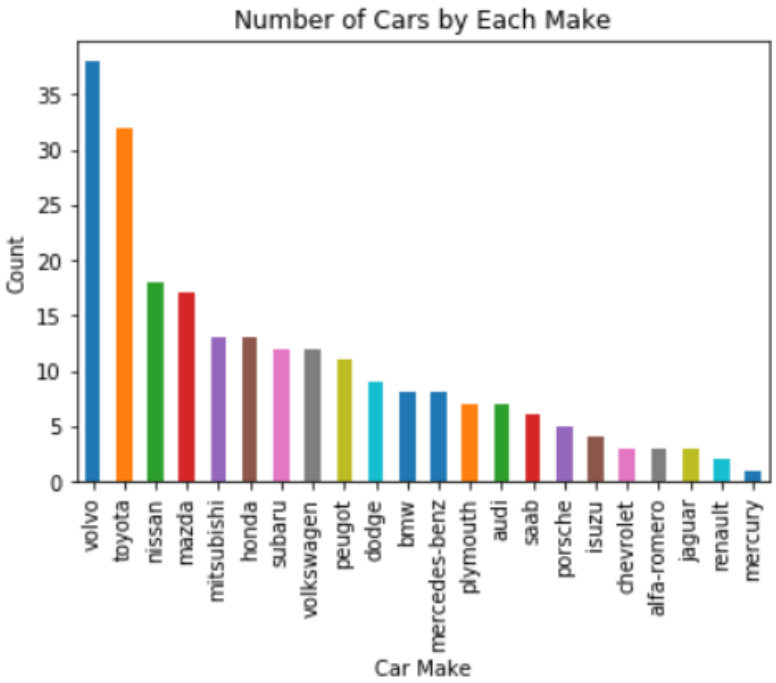
    - **Checking Missing Values & Outliers:** I have checked missing values using “isnull” function and displayed sum of it to check missing values of each column. Also, plotted box plot to check outliers for “normalized-losses” column.

        - I have imputed missing values with vertical **mean** of each column **(Normalized-losses, Bore, Stroke, Horsepower, Price and Peak-rpm)**. For “**num-of-doors**” column I’ve replaced missing value with **‘four’** as four cylinder is more common (**median**) in cars.

Data Exploration

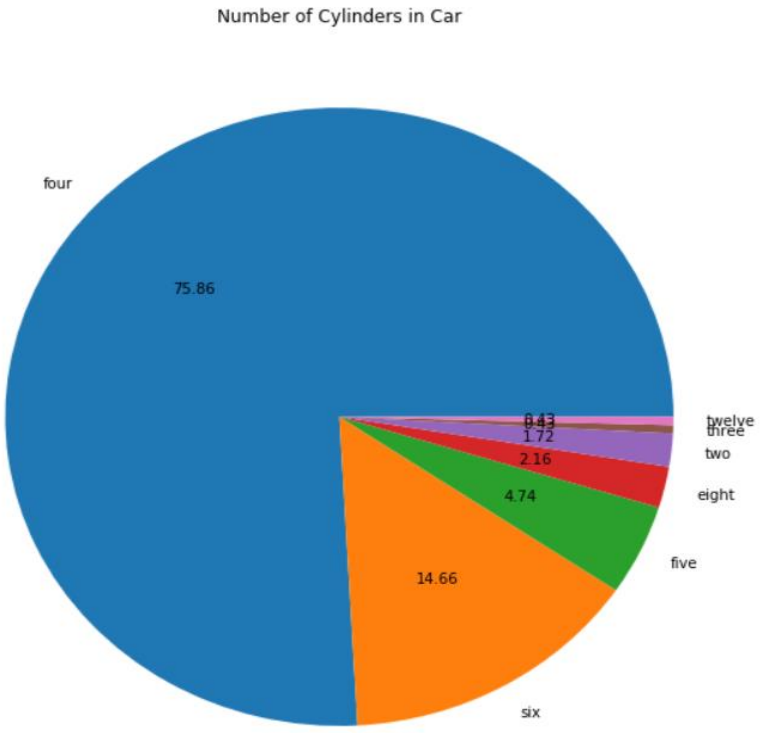
1) **Task 2.1:**

    - I have choose **'make' as Nominal variable** to plot graph to visualize number of cars sold by each make.



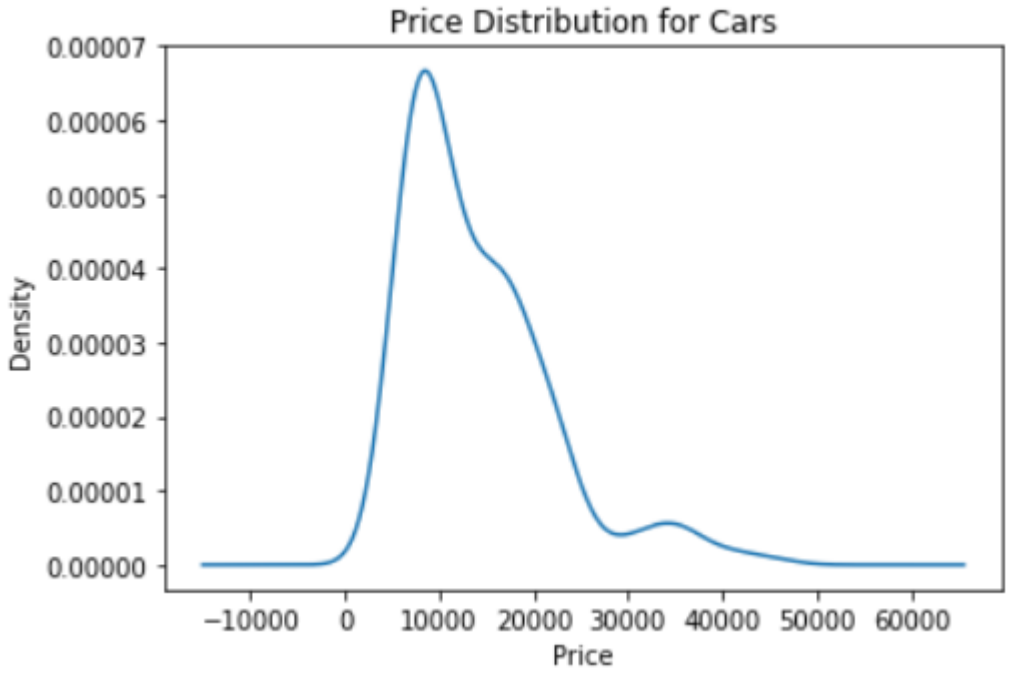
As I wanted to visualize counts for each make, I chose bar chart to represent this categorical variable. It can be observed that volvo & Toyota have significantly more cars than any other cars.

- I have choose **'num-of-cylinders' as Ordinal variable** to plot graph to visualize what is the most common number of cylinders in cars.



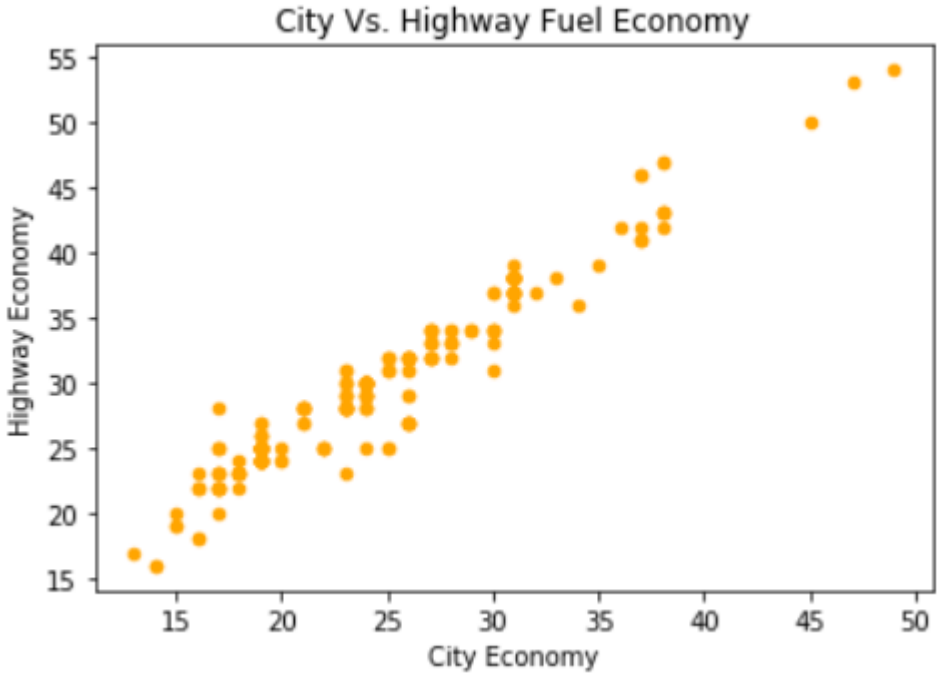
As I wanted to visualize proportions of number of cylinders in car, I have used bar chart for this variable. The plot suggest that **four cylinder** cars have larger proportions – **75%** followed by six and five cylinders. While there are only **5%** of cars with cylinders higher than six or lower than four.

- I have choose **'price' as Numeric variable** to plot graph to visualize - in what range the mean price of all cars falls. As I wanted to visualize distribution of price, I've used density graph to analyze it. It can be seen that **more than 75%** of the cars' price **fall in the range of 5000-25000**.



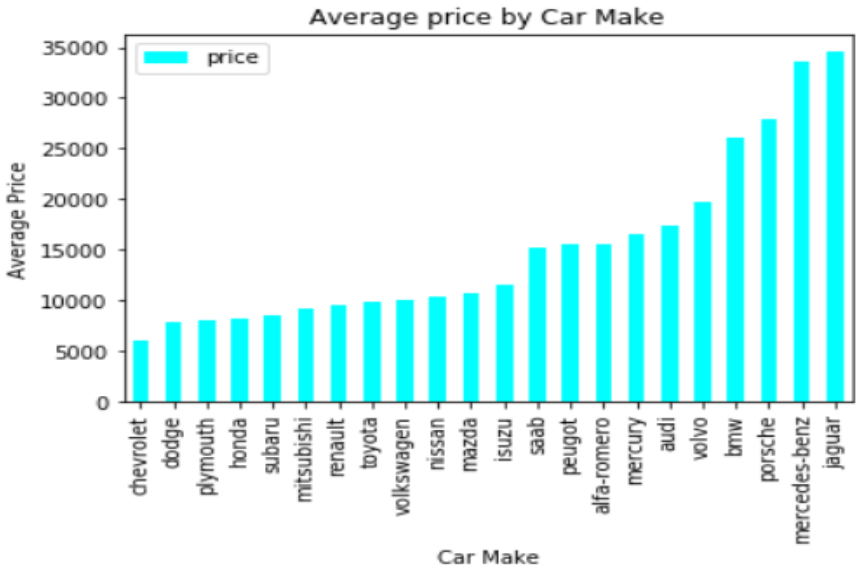
2) Task 2.2:

- **Hypothesis - 1: Do city fuels economy increases as highway fuel economy increases in the car?**



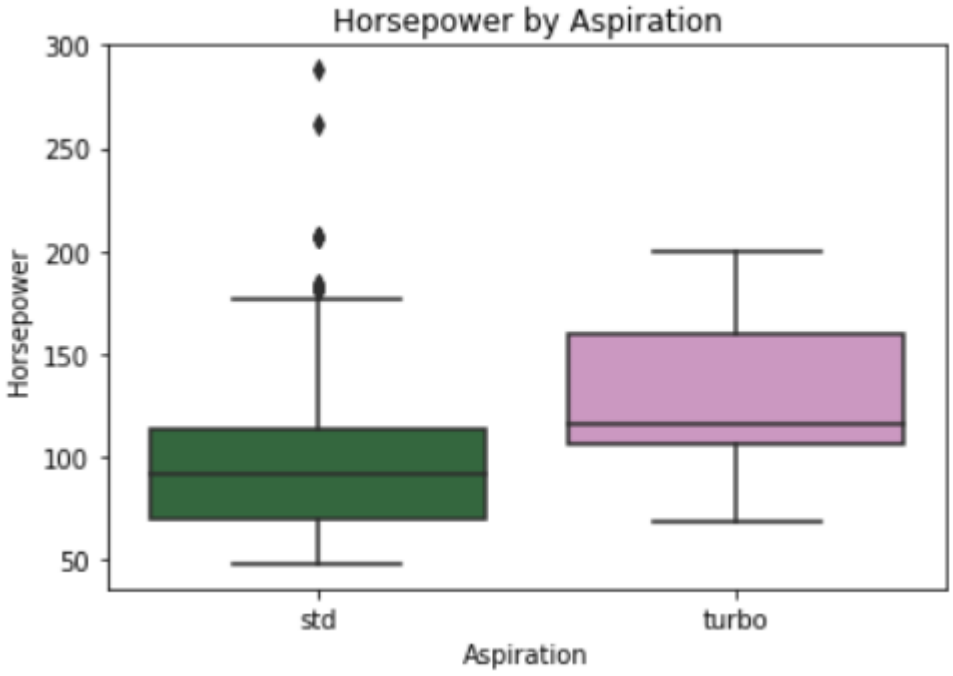
As can be seen from the graph, we can observe that they are in positive linear correlation which suggest that our hypothesis is plausible and city fuel economy increases as highway fuel economy increases.

- **Hypothesis - 2: Do the average price of car tends to be higher based on make?**



As can be seen from the graph, certain car makes tends to have significantly higher average price for the cars than others. So we can say that our hypothesis holds true for our dataset.

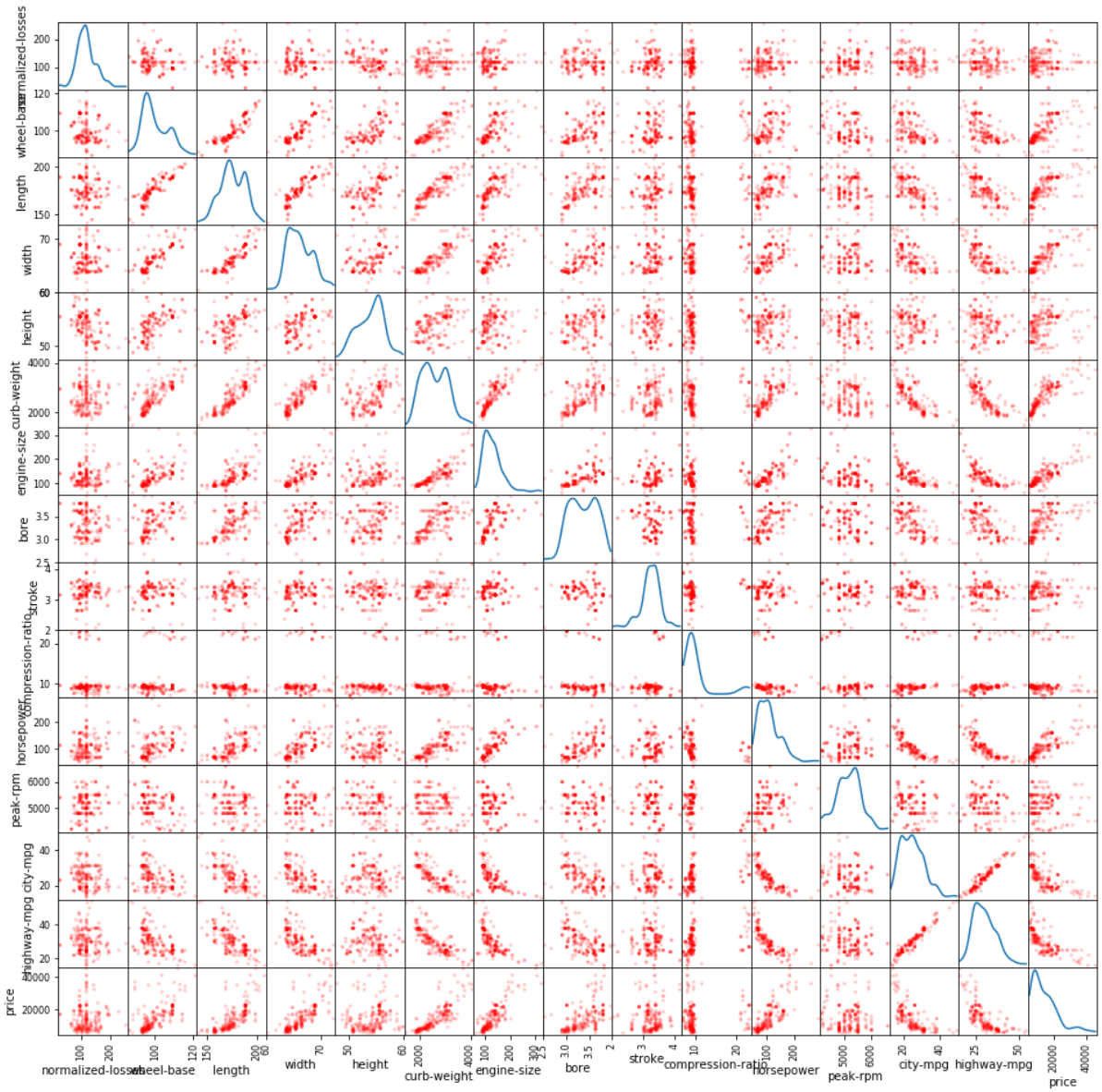
- **Hypothesis - 3: Can Aspiration influence the horsepower of the Car?**



As can be seen from the graph, Turbo aspiration tends to have higher horsepower than the standard one on an average and it concludes that it increases horsepower and are more powerful.

3) Task 2.3:

- Scatter matrix are very useful for visualizing the pairwise relationship & correlation between all pairs of variables.
- As can be seen from the graph, “Engine size”, “Compression Ratio” and “horsepower” are left skewed.
- Variables “bore”, “stroke” and “peak rpm” doesn’t seem to have a linear correlation with other variables. Scatter plot of these variables are sparse without.
- Density plot in the diagonal suggest how values of variables are distributed in the dataset (i.e. Left-skewed, Normal, multi-modal etc.).
- We can establish the positive (Wheel-base to length, Wheel-base to width, Curb-weight to Engine-Size, city-mpg to highway-mpg) and negative linear correlation ship (Curb-weight to City-mpg and highway-mpg, mpg to bore, mpg to horsepower, engine-size -> mpg) between the variables and can infer the result of some hypothesis.


normalized-losses
wheel-base
length
width
height
curb-weight
engine-size
bore
stroke
compression-ratio
horsepower
peak-rpm
city-mpg
highway-mpg
price