Report

1. Data Preparation

Data type each variable was checked using .dtypes. Data type of the variable 'symboling' was corrected to object data type since it is an ordinal variable. First few rows and count of tuples were verified with the given csv file. The data set Automobile contains 238 entries on the following 26 variables.

1.1 Nominal and Ordinal Data

Frequency tables were generated for all the object type variables using value_counts() method to identify typos and extra whitespaces. Typos were corrected using .replace method and extra whitespace was removed using str.strip() method. All the missing values were replaced with the mode of the variable.

1.2 Numerical Data

Summary of statistics table for all the numerical variables was generated using .describe() method to identify missing values and outliers. Missing values were replaced with the mean value of the particular variables. In addition to the summary statistics table bar graphs were generated to understand the distribution of the variable. Outliers were not replaced or removed from the data set, but identified the variables having high percentage of outliers using the bar graphs.

2. Data Exploration

2.1 Subsection 1 – Individual variables

2.1.1 Nominal: Body styles of different cars

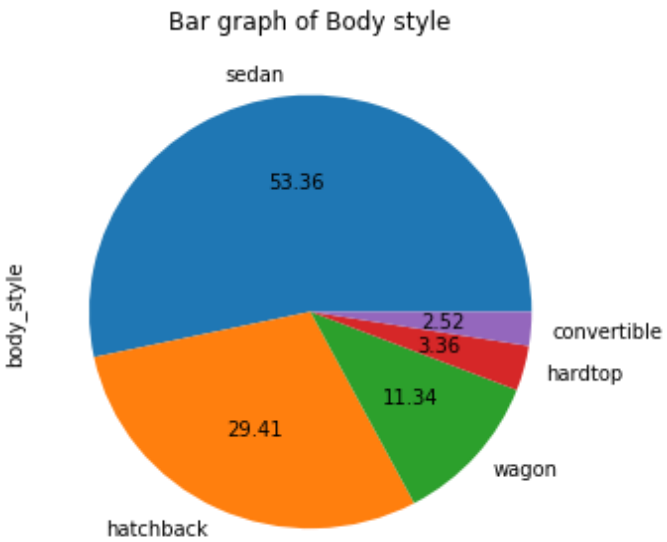

Figure 2.1.1: Pie chart representing body styles of cars

Reason for choosing a pie chart:

Number of categories was less than eight and the categories were not similar sized, hence a pie chart was chosen to analyse the proportion of categories.

Observations:

- More than 50% of the cars are designed in ‘sedan’ style.
- Approximately 40% of cars are ‘hatchback’ or ‘wagon’.
- Very few ‘hardtop’ and ‘convertible’ cars have been recorded in the data sample.

2.1.2 Ordinal : Insurance risk rating

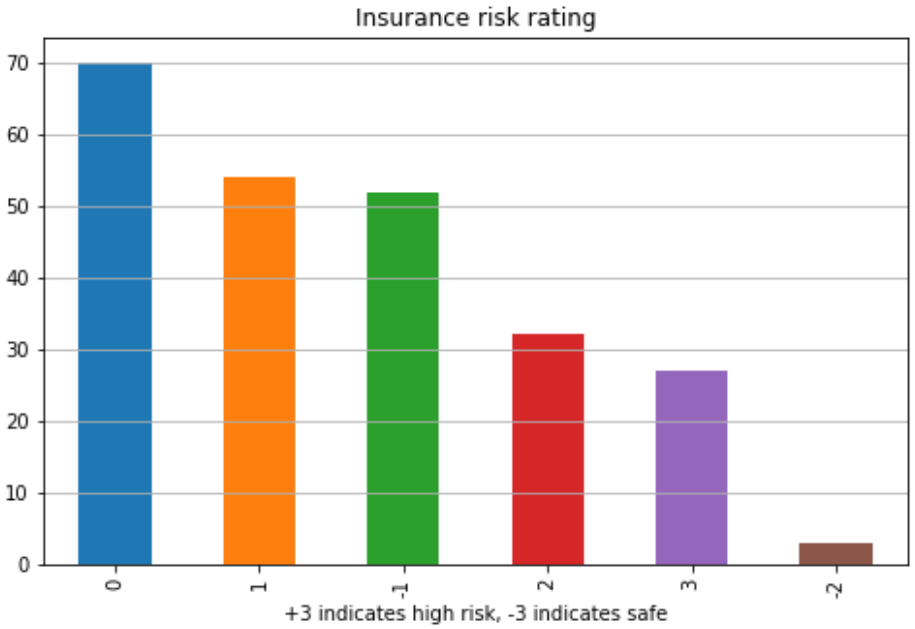

Figure 2.1.2: Bar graph representing insurance risk rating

Reason for choosing the variable:

The only ordinal variable present in the data set is ‘symboling’. When using a bar chart it is easy to compare the frequencies of different categories (rankings).

Observations:

(Here +3 indicates high risk cars and -3 indicates safest cars)

- Majority of the cars risk level is recorded as 0.
- Though there are no safest cars, there are some high risk cars.
- More than 50% of the cars in the range of moderate risk level and moderate safe level.

2.1.3 Numerical: Price of the car

Reason for choosing the variable:

Price is a continuous numerical variable depending on most of the other variables. Hence distribution of this variable is very important. A histogram was used to plot the frequency distribution to observe and identify the distributional properties.

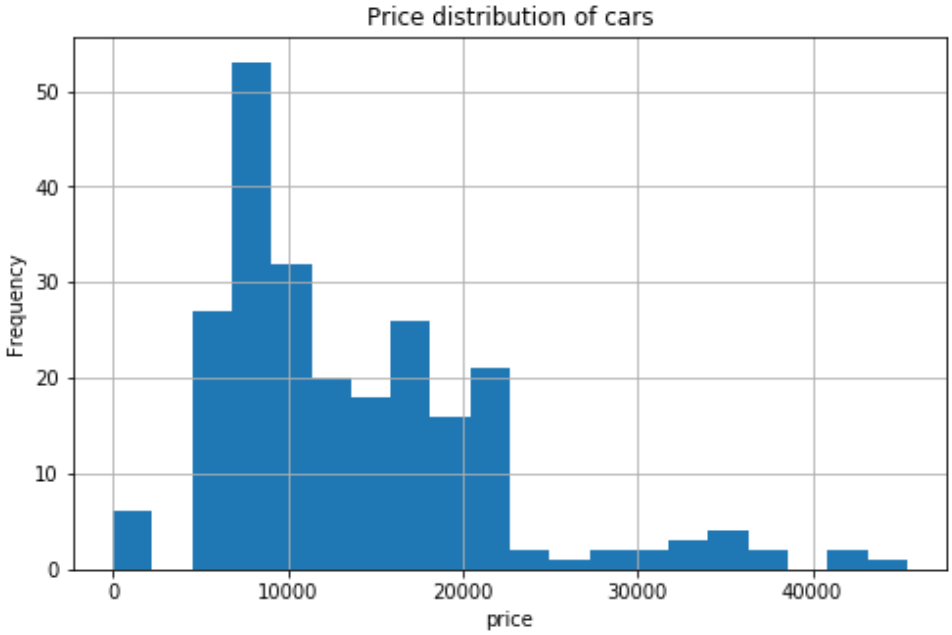

Figure 2.1.3: Price distribution of Cars

Observations:

- The distribution curve is positively skewed and hence it is possible to conclude that the price of cars does not follow a normal distribution.
- Most cars are in price range of 5000 – 20 000.

2.2 Subsection 2 – Pairs of variables

2.2.1 Price Compared with City mpg

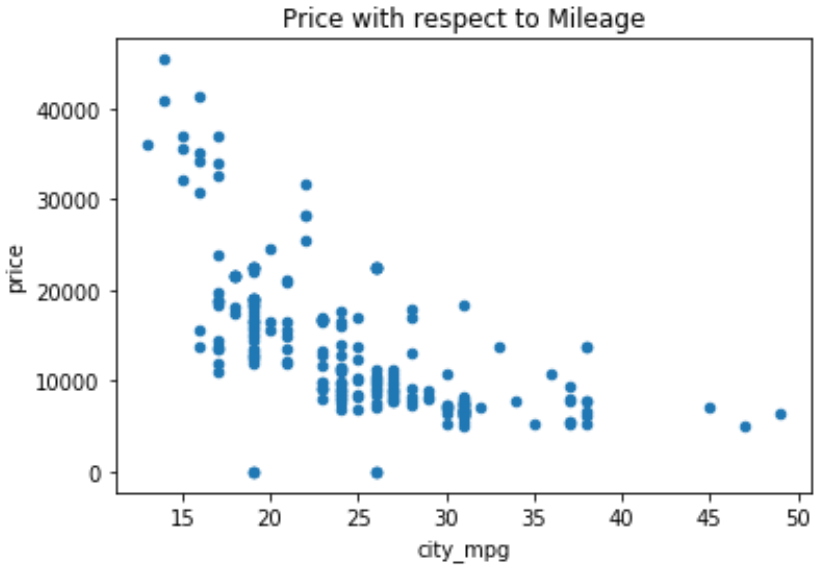

Figure 2.2.1: Price by city mpg

Hypothesis: Whether there is a relationship between the price and the fuel economy of an auto

Observations:

- Mileage and price are negatively correlated.
- Majority of the expensive cars are not fuel economical.

2.2.2 City Mpg compared with fuel type

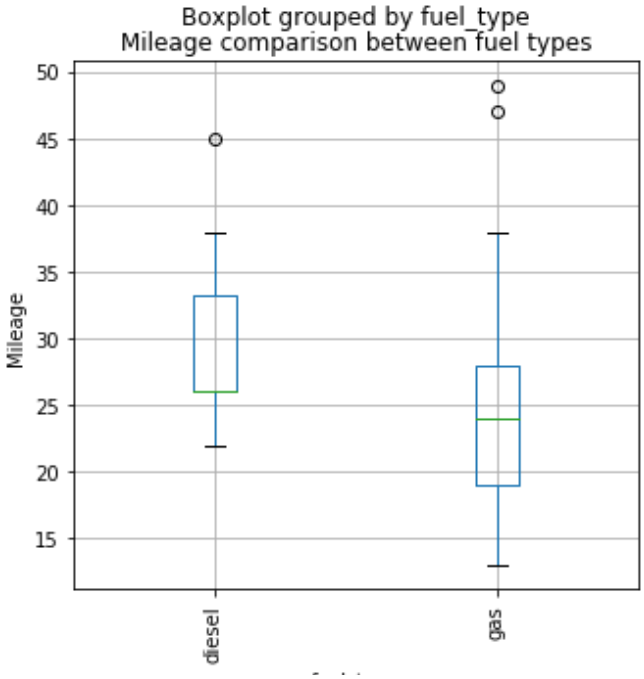

Figure 2.2.2: City mileage by fuel type

Hypothesis: Whether the fuel type has an effect on miles per gallon

Observations:

- Cars with gas engines have a wide range of mpg distribution.
- Diesel vehicles are more fuel efficient in average when compared to gas vehicles.

2.2.3 Price vs Make of the car

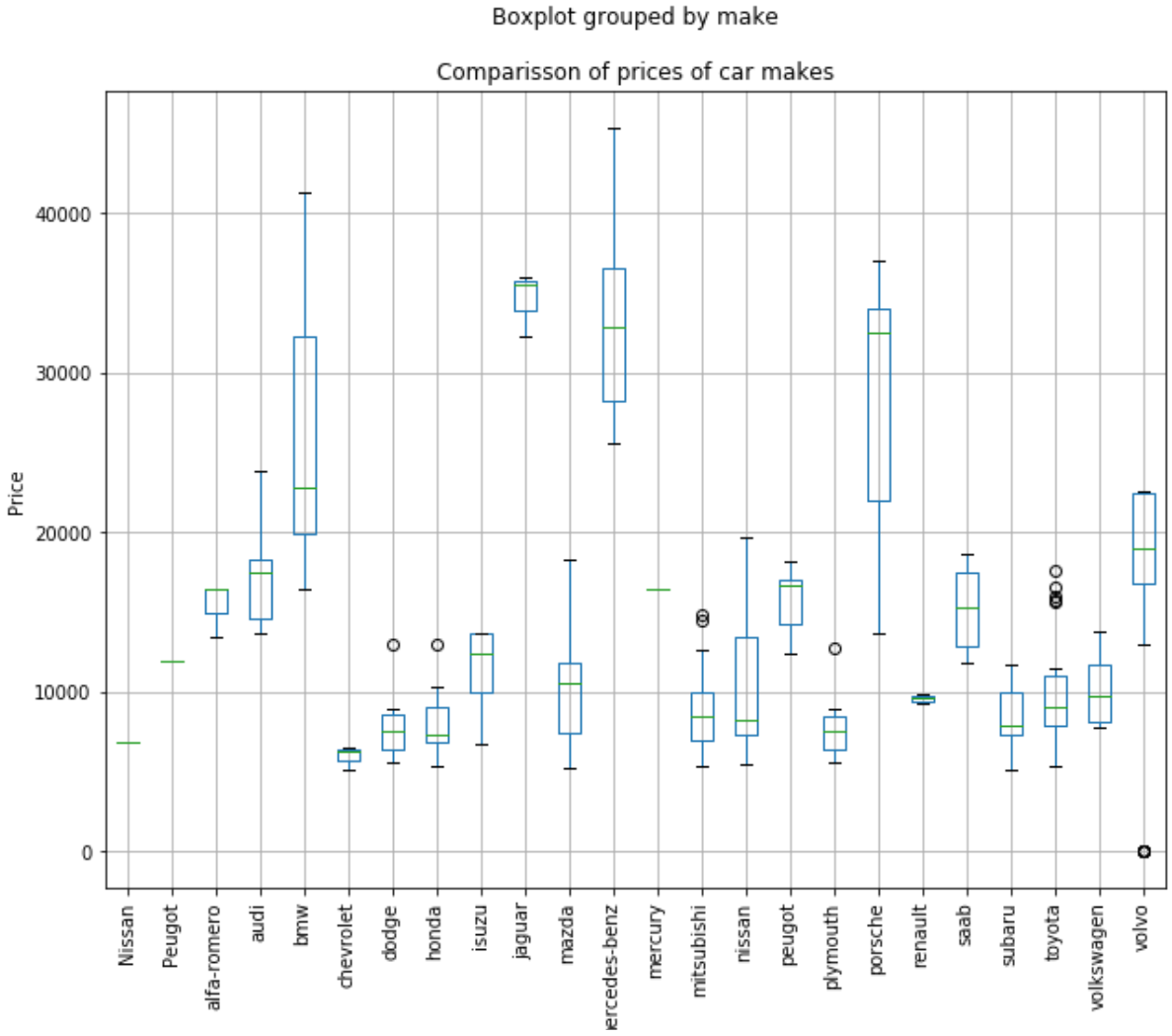

Figure 2.2.3: Price by Car make

Hypothesis: Whether it can be identified the manufactures of the most expensive cars and the most economical cars.

Observations:

- Volvo produces the least expensive car, however when comparing the total price ranges, Chevrolet can be identified as the manufacturer of the most economical cars.
- Other budget car manufacturers who produces cars valued less than 10 000 are Chevrolet, Dodge, Honda, Mitsubishi, Plymouth, Renault, Subaru and Toyota.
- Apparently Mercedes –Benz is the manufacturer of the most expensive cars.
- BMW, Jaguar and Porsche can also be identified as expensive car manufactures.
- While majority of the manufacturers produce cars within a narrow price range, BMW and Porsche manufactures cars in a wide price range.
- Majority of the cars are valued less than 20 000.

2.3 Subsection 3 – Scatter Matrix For all Numerical Variables

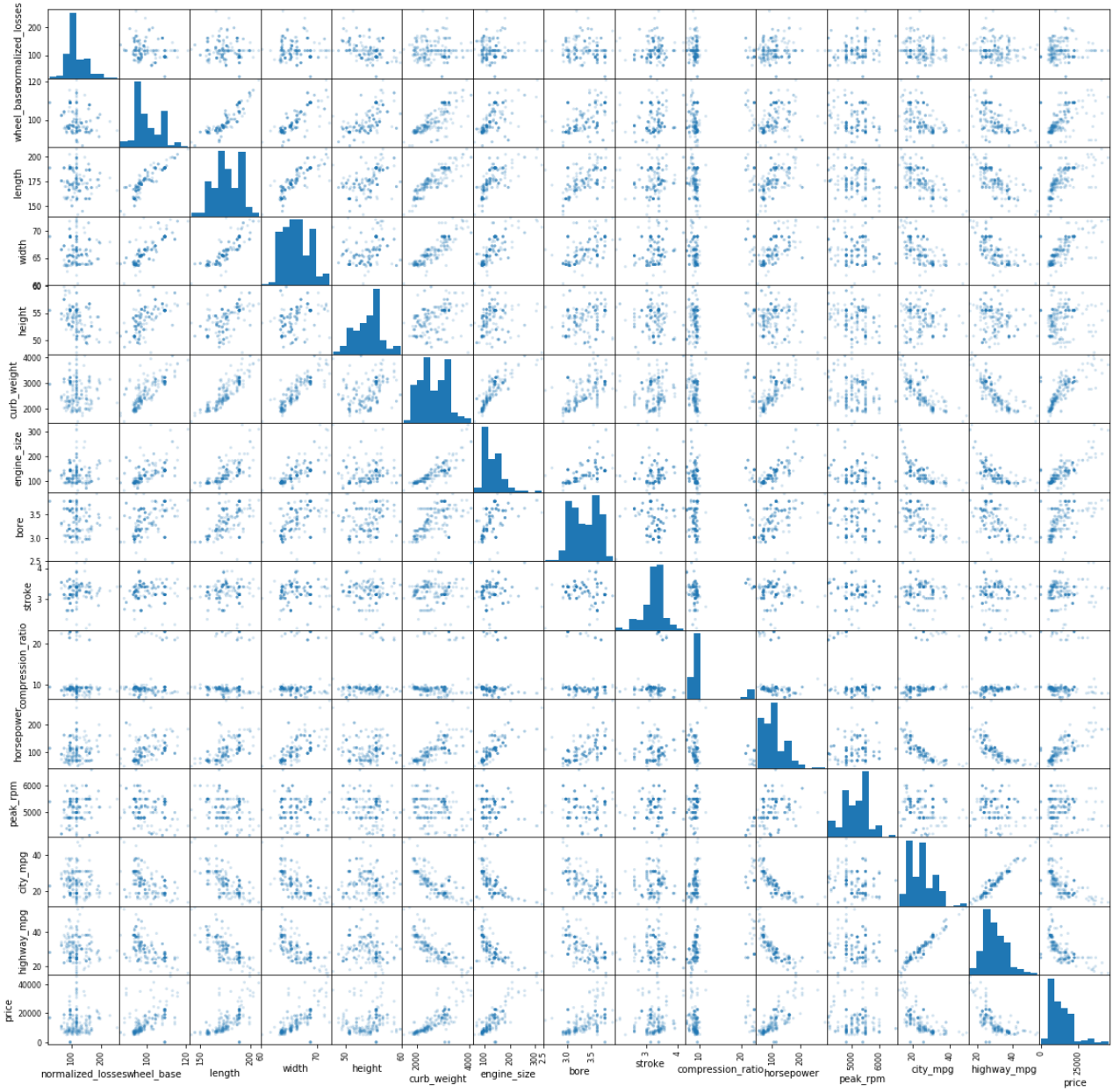

Figure 2.3.1: Scatter Matrix for all the numerical variables

Observations:

- Mileage rate decreases when curb weight, engine size and horse increases.
- Horsepower increases with curb weight and the engine size.
- Price increases with size of the vehicle, curb weight and engine size.