Report for Practical Data Science Assignment1 Task3

Data Preparation

During the data presentation step, firstly, I input the Automobile.csv file from data folder, use '#' to separate each column and name each attribute as symboling, normalized-losses, make, fuel-type, aspiration, num-of-doors, body-style, drive-wheels, engine-location, wheel-base, length, width, height, curb-weight, engine-type, num-of-cylinders, engine-size, fuel-system, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg and price. Secondly, I deal with the potential errors of data, and the subsection includes Typing error, Extra Whitespaces, Impossible values, and Missing values. Then, I will describe these errors found in each attribute and explain how to deal with it.

Typing error

For the typing error, I found it in make, fuel-type, aspiration, num-of-doors, body-style, drive-wheels, engine-location, engine-type, num-of-cylinders, and fuel-system attributes, where they all include capital letters in the data. For this typing error, I transform them to the lowercase by using 'str.lower()'.

There is another kind of typing error that includes the repeated letter or unrelated number, such as, 'turrrrbo', 'vol00112ov'. I found this kind of error in make, aspiration and num-of-doors attributes. For this error, I refer to the data description document and replace them with their right form.

Extra whitespaces

For the extra whitespaces, I found them included in 'make' attribute et al. To clean these error quickly, I use 'str.strip()' method implemented in all the data, and clean all the extra whitespaces.

Impossible values

There are 3 impossible values include in these attributes:

symboling, which includes the value out of the range, 4, so I replace it with maximum value, 3.

normalized-losses, which include the value out of the range, 25. I guess maybe it is typing error so replace it with 65.

price, which includes 0.0 and it is not possible as price. I observe this feature data and found 0.0 is a little far away from the minimum value which is 5118. And the distribution of this data is range from 5118 to 45400 evenly, although most of them are bias to 5118. So, I replace this value with the median of this data.

Missing values

For the missing value, I found it in 7 attributes: normalized-losses, num-of-doors, bore, stroke, horsepower, peak-rpm, and price attributes. Since each of these attributes is numeric and distributed evenly, I fill these missing values with their median. The reason why I use the median rather than mean is median can always better than mean, since in some case when the distribution of the data bias to one side that median is better than mean.

After these errors cleaned, I store all the processed data to an 'automobile_data.csv' file and separate by ',' symbol. I also pick up the numeric data to store in an 'automobile_numeric.csv' file prepared for the next task.

Data Exploration

Part1: Data visualization

In this section, I select drive-wheels, symboling, and horsepower values to created visualization for each of them.

For the nominal value, drive-wheels, it represents the drive-type of the car, which includes fwd, rwd, and 4wd. In order to give a feeling about what percentage of the different type of drive-type accounts for the total cars, I select the Pip chart to present these data (**Figure1.1**).

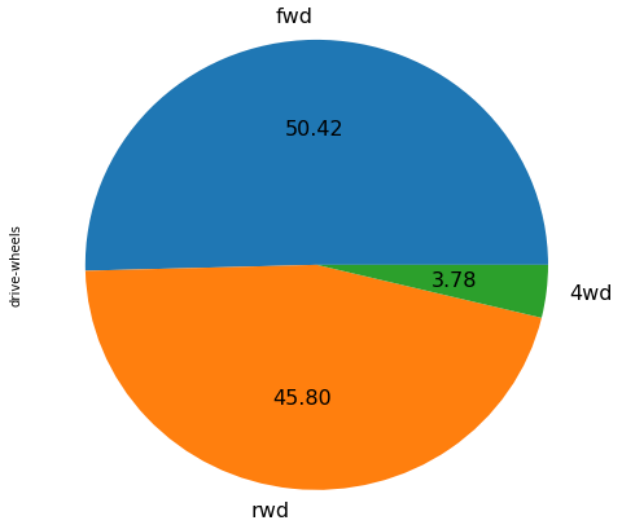

Figure 1.1: Pip chart of drive-wheels value.

For the ordinal value, symbolling, it represents the insurance risk rating, which ranges from +3 (high risk auto) to -3 (safe), although there is not absolutely safe (-3) car in this data. To compare

the number of different insurance risk car in this dataset, I select the Histogram graph to present it (**Figure 1.2**).

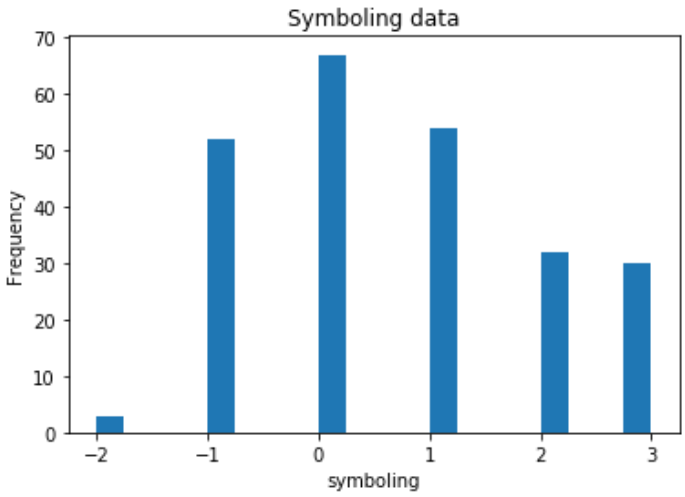

Figure 1.2: Histogram of symboling value.

For the numeric value, horsepower, it represents the engine power of the car. In order to have a good understanding of the distribution of this value in all the data, I display this data by using BoxPlot diagram (**Figure 1.3**). It can draft the key figures in the distribution and help you spot outliners.

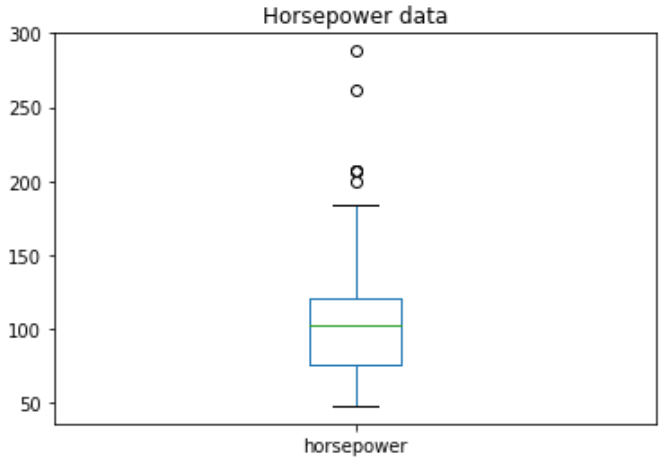

Figure 1.3: BoxPlot of horsepower value.

Part2: Data relationships

In this section, I assume that if there are some interesting relationship between some values, so I make assumptions, and draw graphs to display their relationship and explain if the assumption can be proved.

Firstly, I find the horsepower and drive-wheels. I guess if different drive-wheels have the different horsepower, and I am also interested in which kind of drive type can have the highest horsepower, so I draw a BloxPlot graph to display the distribution of each type of drive-wheels to the horsepower and compare them (**Figure 2.1**). Then I found drive-wheels indeed has the relationship to the horsepower, and rwd type of drive-wheels have the highest horsepower on the average.

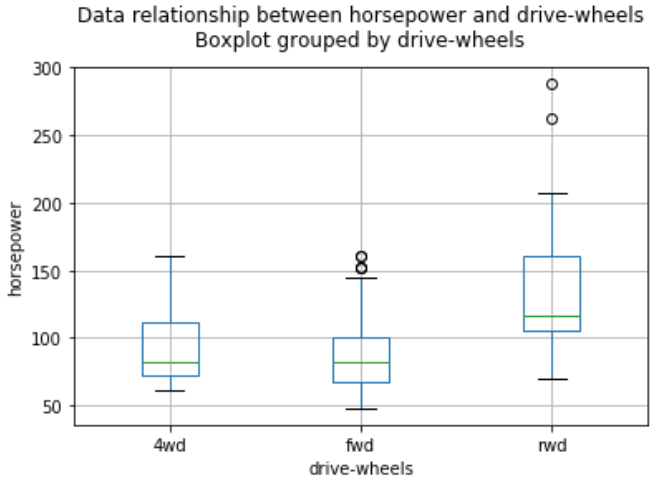

Figure 2.1: BoxPlot of the relationship between horsepower and drive-wheels.

I assume that there are some relationships between horsepower and price, and I also guess the car with higher horsepower should be more expensive, so I draw a Scatterplots graph to present the values for price to horsepower (**Figure 2.2**). The result shows that it does hold on. Higher horsepower can have a higher price.

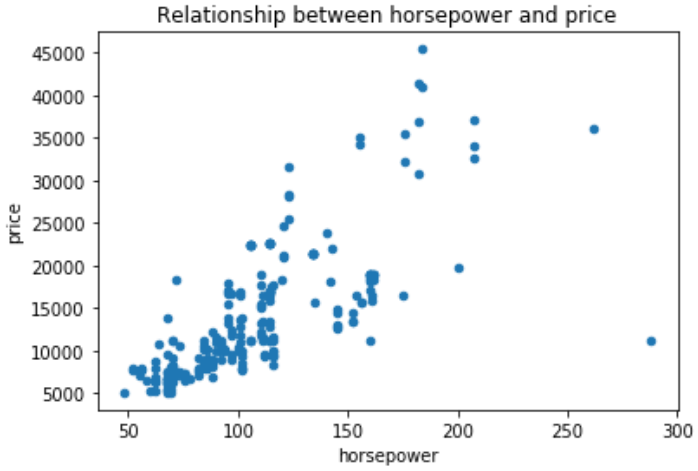

Figure 2.2: Scatterplots of the relationship between horsepower and price.

For the car size, I assume if different drive-wheels might have some restriction to the car size in terms of drive type. Then I build a Scatterplots graph to display the distribution of drive-wheels to length and width (**Figure 2.3**). From the graph, we can see that the car with rwd drive type are usually longer and wider, while fwd-type car can be smaller. 4wd-drive-wheels car is between rwd-type car and fwd-type car and close to the fwd-type car.

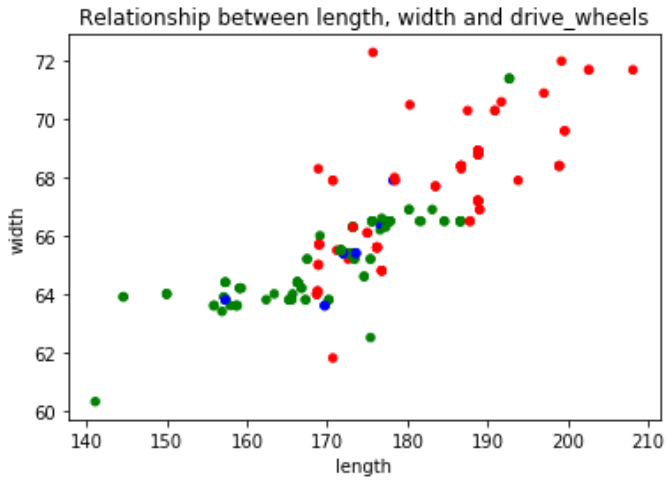

Figure 2.3: Scatterplots of the relationship between length, width and drive-wheels (red: rwd, green: fwd, blue: 4wd).

Part3: Scatter matrix

In this section, I build a Scatter matrix for all numerical columns (**Figure 3.1**). We can observe something interesting from it:

1. wheel-base is positively correlated with length, width, and curb-weight.
2. curb-weight is positively correlated with length and width.
3. engine-size is positively correlated with price and horsepower, while it has a negative correlation with highway-mpg and city-mpg.
4. horsepower has a negative correlation with highway-mpg and city-mpg.
5. city-mpg is positively correlated with highway-mpg.

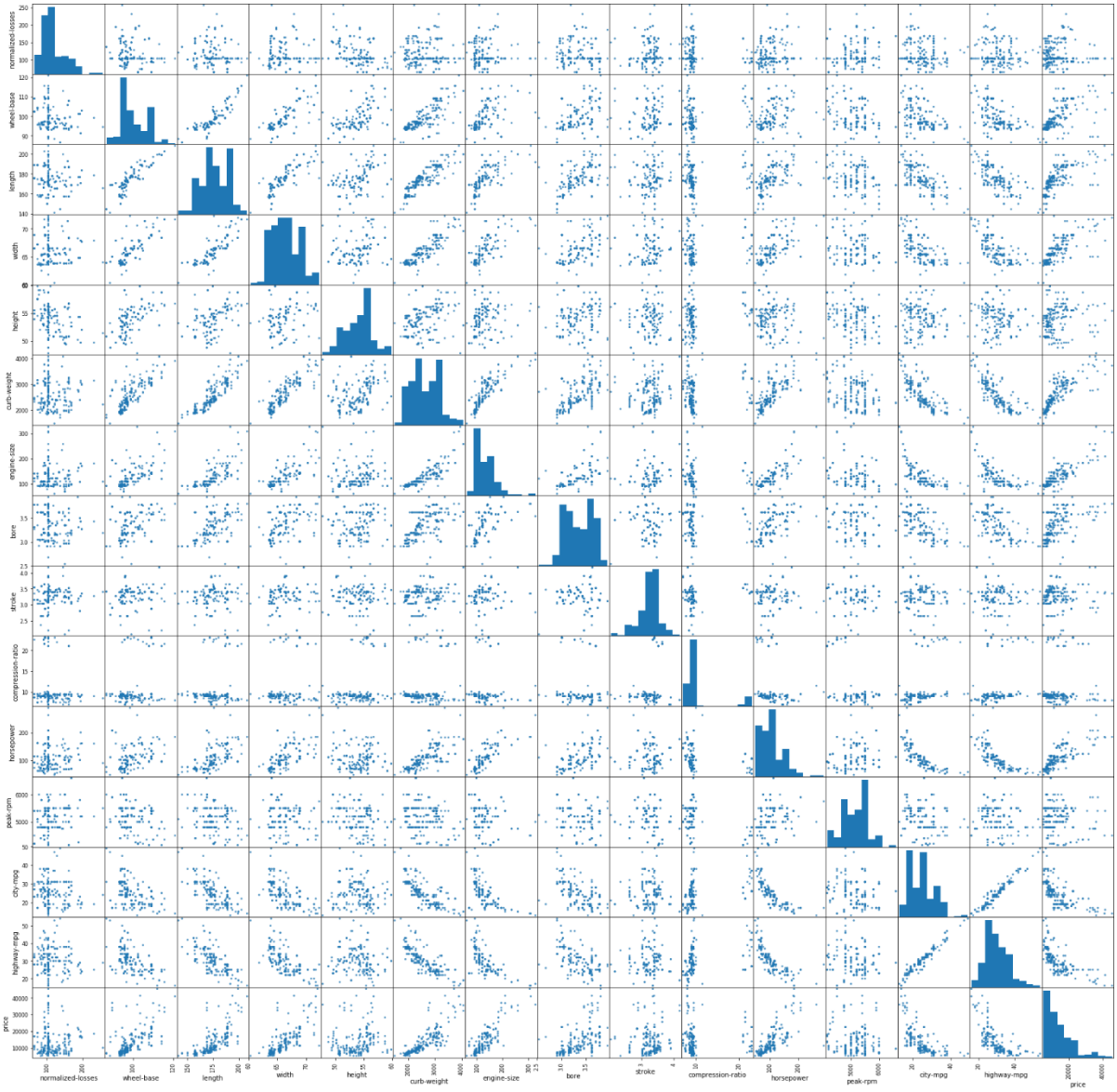

Figure 3.1: Scatter matrix of all numerical columns.