Data Preparation

Data preparation is the process of gathering, combining, structuring and organizing data so it can be analyzed as part of data visualization, analytics, and machine learning applications.

The main purpose of data preparation is that the data is being prepared for analysis in an accurate and consistent manner. Data often consist of missing values, typos, impossible values, and many more errors. The process of correcting these inaccuracies, and verify the information and combining the datasets plays a major role in the data preparation process. (Ren, D. ('2019')).

Errors that considered while cleaning the data:

1. Missing Value
2. Typos
3. Extra White Spaces
4. Sanity check

Cleaning of Dataset Column wise:

Symboling: In symboling, I have done a sanity check and found there were 3 impossible values in the column. I have done a sanity check based on the values given in the UCI Machine Learning repository on the automobile.

I have found value '4' as an outlier(impossible) value so I just removed that value because it is not always mandatory that all cars of the same type can have same symboling criteria so I just dropped the records to remove the outlier.

Normalized Losses

For the normalized losses columns if there are enough number of records with valid normalized loss values (>50%) then I have taken mean or median and replace all the NaN.

If less than 50% of records have NaNs for normalized loss if taken the mask with 'Make' column values then I removed the records.

Fuel Type

In Fuel type column there were white spaces and typos. I removed the spaces by using trim() function and in typos, the values are given in both lower and upper case for similar value so I just converted all the values in lower case.

Aspiration

In Aspiration column, there were white spaces, spelling error (typo), and words are written in both the letter cases lower and upper. I cleaned the column first by lowering the case of a letter, then removing the whitespaces and finally by replacing the character of column values with correct values.

NumOfDoors

In NumOfDoors column, there were white spaces, spelling error (typo), words are written in both the letter cases lower and upper, and also there were missing values. I cleaned the column first by lowering the case of a letter, then removing the whitespaces, then by replacing the character of column values with correct values and finally by filling the null values.

BodyStyle

In Bodystyle column, there were white spaces, and words are written in both the letter cases lower and upper. I cleaned the column first by lowering the case of a letter, then removing the whitespaces.

DriveWheels

In DriveWheels column, there were white spaces, and words are written in both the letter cases lower and upper. I cleaned the column first by lowering the case of a letter, then removing the whitespaces.

EngineLocation

In EngineLocation column, there were white spaces, and words are written in both the letter cases lower and upper. I cleaned the column first by lowering the case of a letter, then removing the whitespaces.

Length, Width, Height, and CurbWeight

All these columns have all the values that are needed for analysis.

There are no null values in these columns and also I have done a sanity check on all the values and all values are coming in the range as mentioned in the UCI Machine Learning Repository(https://archive.ics.uci.edu/ml/datasets/automobile).

EngineType

In the EngineType column, there were white spaces, and words are written in both the letter cases lower and upper. I cleaned the column first by lowering the case of a letter, then removing the whitespaces.

NumOfCylinders

In the NumOfCylinders column, there were white spaces, and words are written in both the letter cases lower and upper. I cleaned the column first by lowering the case of a letter, then removing the whitespaces.

FuelSystem

In the FuelSystem column, there were white spaces, and words are written in both the letter cases lower and upper. I cleaned the column first by lowering the case of a letter, then removing the whitespaces.

Bore

In the Bore column, there were four values are missing which belong to ‘Mazda’ value in make column. So I just take the mean of all the values in the Bore column by applying a filter with ‘make’ columns ‘Mazda’ value and then used fillna() condition to fill the empty values in Bore column.

Stroke

In the stroke column, there were four values are missing which belong to ‘Mazda’ value in make column. So I just take the mean of all the values in the Stroke column by applying a filter with ‘make’ columns ‘Mazda’ value and then used fillna() condition to fill the empty values in Stroke column.

CompressionRatio

Compression ratio has all the value as expected. Done with all the conditional checks on the column and all values are in the range as mentioned in the UCI Repository (https://archive.ics.uci.edu/ml/datasets/automobile).

HorsePower

HorsePower has all the value as expected. Done with all the conditional checks on the column and all values are in the range as mentioned in the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/automobile).

PeakRPM

PeakRPM has all the value as expected. Done with all the conditional checks on the column and all values are in the range as mentioned in the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/automobile).

CiytMpg

CityMPG has all the value as expected. Done with all the conditional checks on the column and all values are in the range as mentioned in the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/automobile).

HighwayMpg

Highway MPG has all the value as expected. Done with all the conditional checks on the column and all values are in the range as mentioned in the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/automobile).

Price

In the price column, there are 2 NaN values. Both the NaN positions are filled by taking mean and median using the filter of make column. There are also some zeroes (0) in the Price column so I have taken the median of the columns by creating the mask and applying a filter with make column.

Data Exploration

Data exploration process involves exploring a large dataset in an unstructured way to uncover patterns, characteristics, and point of interest (Ren, D. ('2019')).

Nominal data:

Nominal data are used to label variables without providing any quantitative values.

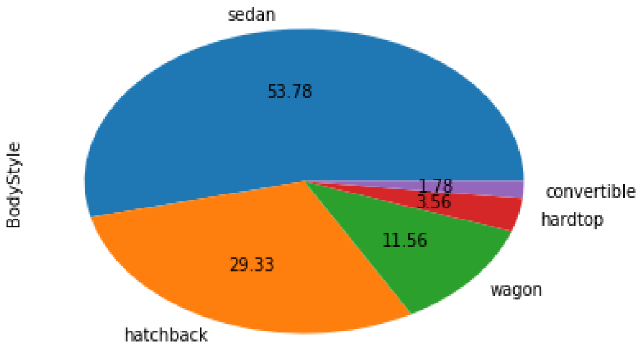

Fig: 1

The pie chart(as shown in fig.1) is used here because it works best on the composition of data. Here in the above pie chart Body Style column is depicted. Body Style column shows the body of the car and consists of categorical values like a sedan, hatchback, wagon, hardtop, and convertible.

From the above chart, I can analyze that the value of sedan is the highest in the body style column with 53.78% percentage in a given dataset. The second highest number of body style data given is of hatchback cars with 29.33% and the least high number of body style data is of convertible body style type with 1.78%.

Ordinal data:

Ordinal data consists of a value that follows a natural order.

In the given dataset I have taken symboling column as an ordinal value. I have created a bar chart (as shown in fig. 2) for the symboling column and analyzed that as the graph is right skewed most of the cars falls under high-risk rating criteria. The highest insurance risk rating given car in dataset is of 0 and the lowest risk rating car given in dataset is of -2.

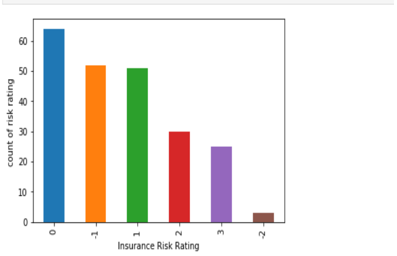

Fig: 2

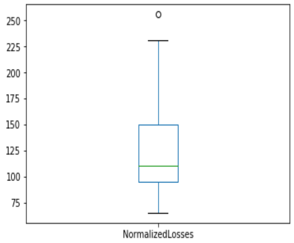

Fig: 3

Numeric value:

As shown in Fig. 3 for the automobile dataset given I have taken Normalized Losses as numeric column and have depicted box plot for that column. This Box plot provides a revealing summary of the Normalized Losses column. The minimum value shown is less than 75, the first quartile is near 100, the median value shown is of between 100 and 125, the third quartile is 150 and the maximum value is shown around 225. Outside the maximum value, all values are considered as an outlier. In the above chart, we are able to see one outlier value.

Task 2:

2.1 The relationship between Price and Number of Cylinders

The graph below (as shown in fig: 4) summarizes Prices of the number of cylinders i.e. two, three, four, five, six, eight, and twelve. The graph shown above is left skewed and I can analyze from it that as the numbers of cylinders are increasing accordingly the prices of cars are also increasing but for three cylinders car the price is less than two cylinder car and again for twelve Cylinder cars the price is less than 6 and 8 cylinder cars.
The reason for this may be people mostly prefer six or eight cylinders cars and not 12 cylinders car. This may be one of the reasons of drop in price of 12 cylinders cars.

2.2 The Relationship between Engine Size and City Mileage

The below graph (as shown in fig: 5) determines the relationship between Engine Size and City Mileage. The Scatter plot depicts that as the size of the engine increases the city mileage decreases. So the Engine Size is inversely proportional to City Mileage.
I think as the engine size increases the weight of the car will increase so this may be the reason for decrease in the mileage of the car.

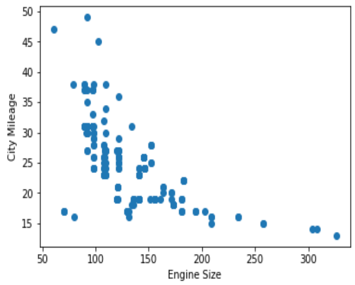

Fig: 4

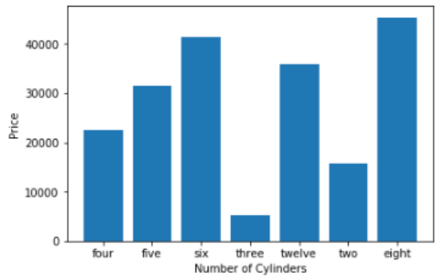

Fig: 5

2.3 The relationship between Price and Highway MPG

The below scatter plot (as shown fig: 6) determines the correlation between Price and Highway MPG. The scatter plot depicts that as the as the prices of vehicle increases the Highway MPG decreases. So the price is inversely proportional to Highway MPG.

Therefore here I conclude that if customer wants the car with high mileage he can purchase the car between range 5000 to 20000 but if customers are not willing to give high importance to mileage then they can go for car price range between 30000 to 45000.

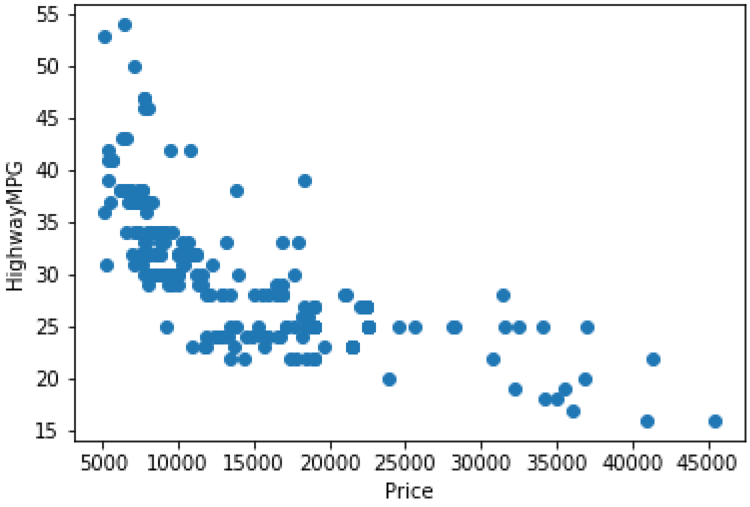

Fig: 6

Task 3:

In the scatter matrix plot given below (as shown in fig:7) there are all three types of correlation like Positive, Negative and Normal between the columns in an Automobile dataset.

1. In a positive relationship between the columns both the columns are proportional to each other. For example:

- As the length of the automobile increases with respect to that price will also increase.
- As the engine size increases the height, width and length of automobile will also increases.

2. In a negative relationship between the columns both the columns are inversely proportional to each other.

For example:

- As the City Mileage increases, Curb Weight and Engine Size decreases.
- As the horse power increases City Mileage decreases.

3. In a normal relationship there is no correlation between the columns.

For example:

- Length has no relationship with compression ratio.
- Highway MPG has no correlation with stroke.

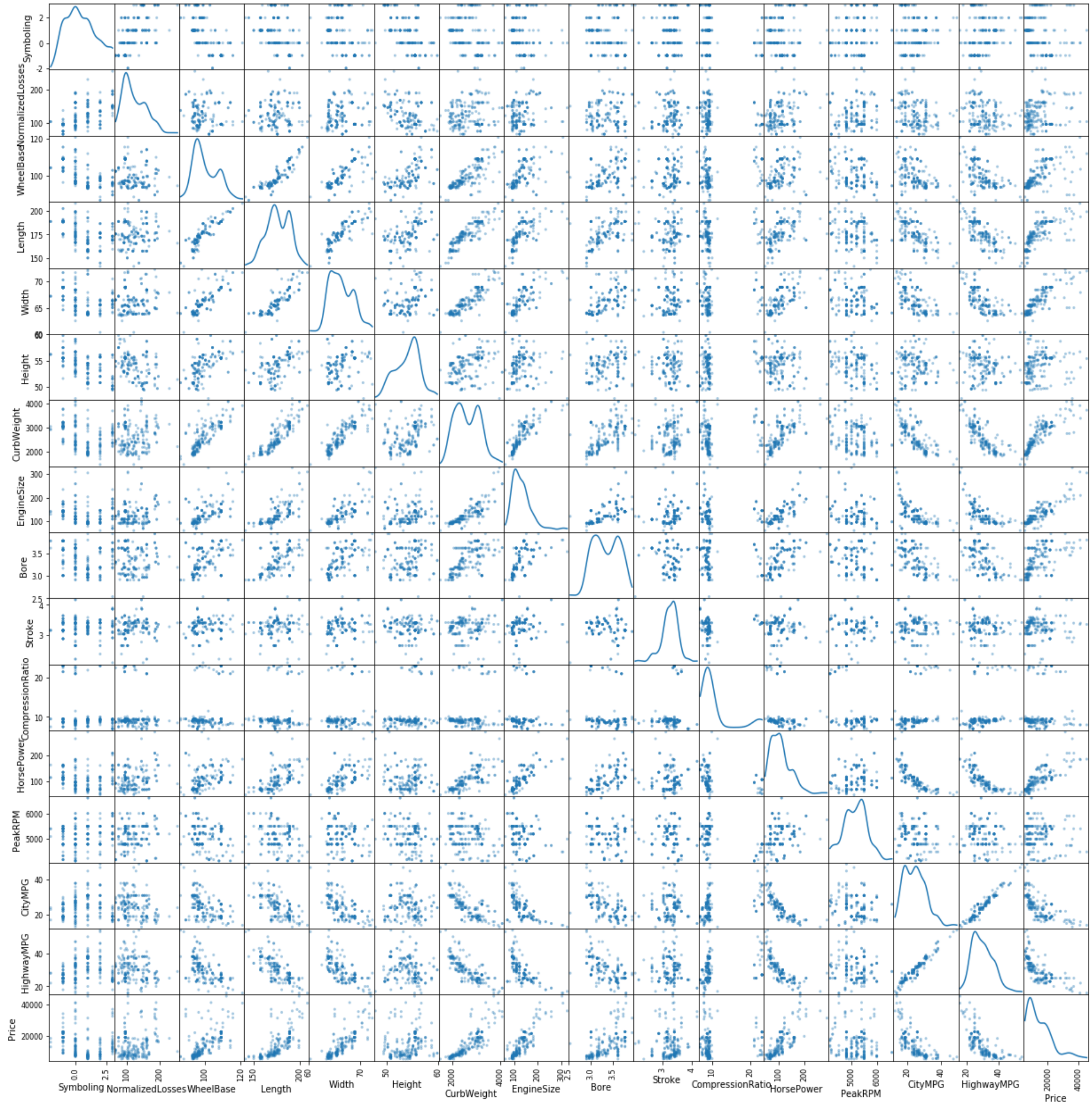

Fig:7

References:

1. Ren, D. (2019). *Data Curation, Data Exploration*. Presentation, RMIT University.

2. UCI Machine Learning Repository: Automobile Data Set. (2019). Retrieved from https://archive.ics.uci.edu/ml/datasets/automobile