Task 1: Data Acquisition & Preparation

1.1 Problem

Assume someone want to sell their car, but the problem is the person don't know how much money he/she can get for the car. But there should be a price on car to sell it for a reasonable price and only then a costumer shows interest to purchase it.

As the data analyst we have to go through various questions to solve the problem. As for the problem described above, we have to identify what are the features that affect the price of car. I can be Brand, Milage, Engine Size etc.

1.2 Understanding the data

There are three data set given "data1.csv", "data2.csv", and "data3.csv" which contain the data from the 1985 Ward's Automotive Yearbook. The files "data1.csv" and "data2.csv" contain the similar set of cars but different sets of attributes for describing the features of car, where each car has its unique attribute "ID". The third data file "data3.csv" contains different set of cars with all the attributes. In total there are 27 different attributes.

Merging data:

As the data files "data1.csv" and "data2.csv" both have the same set of cars but different attributes. We use **pandas.merge()** function to merge both the files into a data frame in pandas.

```
df12 = pd.merge(df1, df2) #these two dataframes has different columns
```

For the mergers with third data file i.e., "data3.csv" it contains different set of cars but have same attributes to present features of car, we use concatenation of this file with the previously created data frame. The **pandas.concat()** function is used.

```
dfall = pd.concat([df12,df3], sort = False) #the two dataframes contains same colums so concatination should be done.
```

"dfall" is the data frame where all the data is merged together

It is a data-set with 199 rows and 27 columns. Starts with the very first attribute "ID" means to give a unique identity to individual vehicle. Second attribute is "Symbolling", corresponds to the level of insurance risk of particular vehicle. Cars are initially

assigned with a risk factor symbol which is related to their price. Then, if any car is riskier, the symbol is adjusted accordingly by moving it up the scale. Value +3 indicates the car is risky and -3 that it is probably much safer.

The third attribute “normalised-losses” represents the relative average loss payment on one insured vehicle year. The range of values are from 65 to 256. The other attributes are easy to understand, the attributes like, “fuel-type”, “engine-size”, “city-mpg”, “highway-mpg” etc.

The last and 27th attribute is “price”. This is the target value which means “price” is the value that we have to focus on so we can find the relationship how the other attributes are affecting the focus area which is “price”. The attributes like, “symbolling”, “normalised-losses”, “make” and so on can affect this.

1.3 Data Cleaning

Tidy data is best suitable for data analysis. It is necessary that your data has less or no null values, no duplicate rows and no impossible values.

i) Check the data: To check the data we use info() function to check the data format of all the cells are right or not and it also describes the number of non-null cells(non-empty cells) and about the data type of particular attribute.

```
dfall.info()
```

The dfall represents the merged file of all data files. And info() function gives a detailed overview about the data frame.

ii) First, we detect some null values to handle. To tackle the empty cells, we have two methods either to remove the entire row or fill the null values with most suitable method so that it cannot impact the data to misinterpret our analysis. Here we are using mean a linear regression technique to get the average value to replace the empty cells as it is the most suitable to fill the empty cells in both the attributes “normalised-losses” and “price”. As these two are the only attributes with empty cells. The function isnull().sum() shows the total number of empty cells in particular attribute

```
dfall.isnull().sum()
```

As mentioned above there is only two attributes with empty cells and handled by putting the mean values of column.

iii) Secondly, once the null values being handled now, we look for duplicate values by using the duplicated function of pandas.

```
dfall[dfall.duplicated(keep = False)] #Checking the duplicate rows
```

	id	symboling	normalised-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower	pe r
16	10180	-1	90.0	toyota	gas	std	four	sedan	rwd	front	...	171	mpfi	3.27	3.35	9.2	156	5
80	10180	-1	90.0	toyota	gas	std	four	sedan	rwd	front	...	171	mpfi	3.27	3.35	9.2	156	5

2 rows × 27 columns

To drop the duplicate values there is a function named drop_duplicates in pandas to remove the duplicated row.

```
dfall.drop_duplicates(inplace = True) #drop the duplicates
```

iv) After performing the above steps, the data is almost tidy. But to recheck the info() function is being used again. Once the data is cleaned, we are all set to perform statistics and visualization on the data for analysis purposes.

1.4 Problems During Data Acquisition and Preparation

Although, there is not much complexities with data as the data is already in its good form there are not any major problem occur during the whole process. But still it is little difficult to understand the data and one of the major problem is to apply an accurate method to replace the empty cells. It is the sensitive part where no mistakes are acceptable. One wrong step choosing the wrong method to handling the empty cell may result in a totally incorrect data analysis or incorrect predictions. Also, there are many functions like drop_duplicates it is recommended to use there parameters very carefully as it can result in deleting all the duplicate rows.

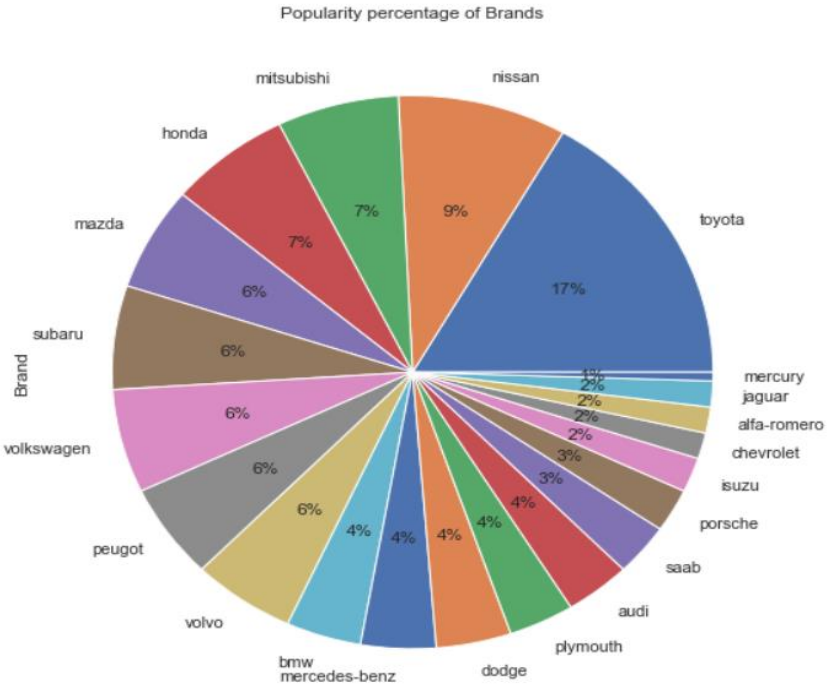
Task 2: Data Exploration

2.1 Visualisation of Categorical and Numerical data

2.1.1 Categorical Data:

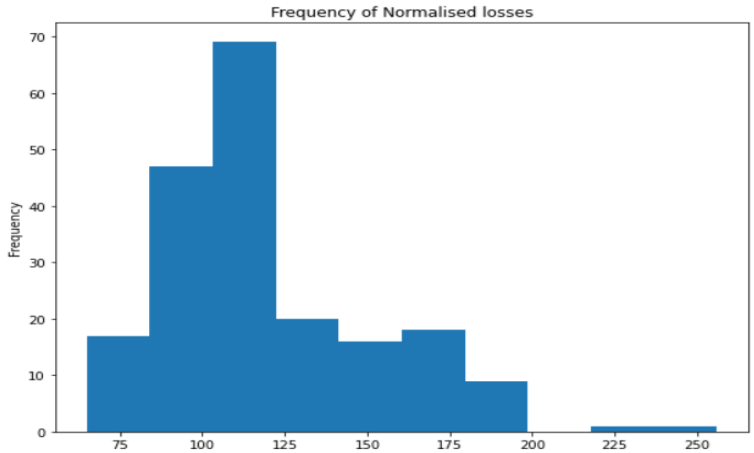
Here we are choosing the attribute “make” to identify which car maker brand is much popular among all brands and preferred by people. Pie Chart is used for this analysis

because it can display the percentage of particular brand very accurately and visualisation with different cut-offs and colours makes it self-explanatory.



2.1.2 Numerical Data:

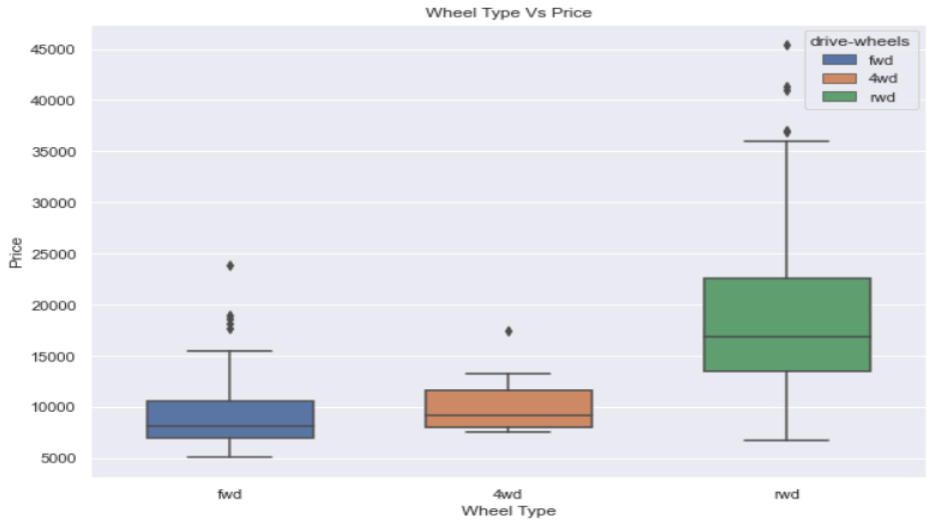
For numerical data we choose normalised-losses attribute it is being selected to check the frequency values of normalised losses where they are occurring more frequently. It represents the average loss by most of the cars are occurring between 100 to 125. The histogram is selected for the frequency because of its gap and it can explain very clearly where the most of the losses occurs.



2.2 Visualising Relational Attributes

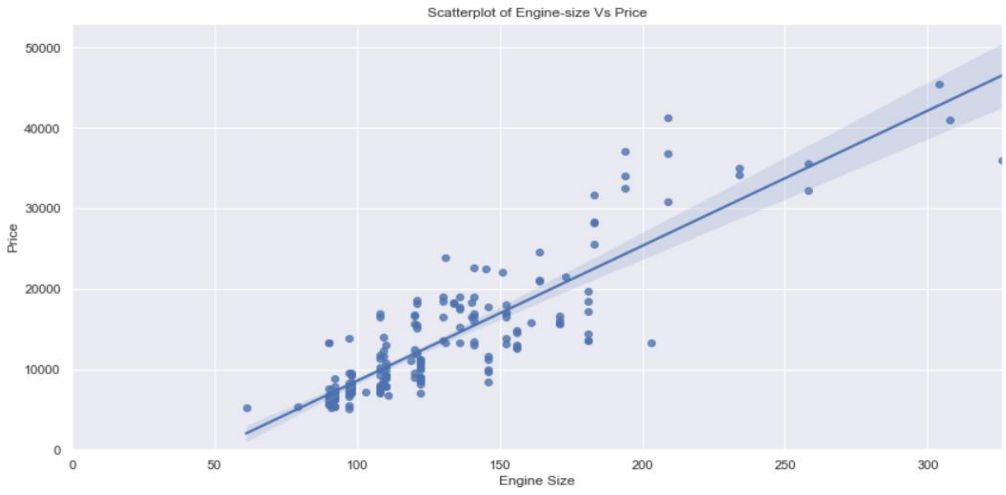
2.2.1 Wheel Type Vs Price

These columns “drive-wheels” i.e. wheel Type is shown with respect to the price of car using the boxplot. The box plot represents the percentile values. The major features of boxplot is that it shows the median of the data. The uppermost quartile shows where the 75th percentile is and lower most shows the 25th percentile. It also visualise outliers as individual dots that occur outside the upper quartile.



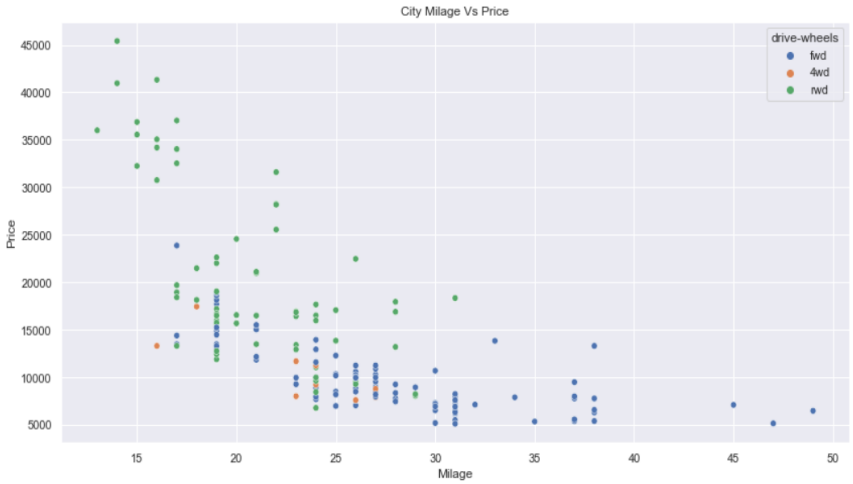
2.2.2 Regression plot of Engine-size Vs Price

The regression plot shows the relationship between the two attributes that how they are dependent on each other the graph visualize a positive regression and shows that Price is directly dependent on the size of engine.



2.2.3 Scatterplot of City Milage Vs Price

Scatterplot visualize individual data by dots it represents how the 2 attributes have their relation with each other as the graph displays the scattered sots are going downwards to the price as the milage increases. Which shows the negative corelation between two attributes.



2.3 Plot of Scatter Matrix

Plot of scatter matrix display the relationship between all the attributes with respect to each other. It represents the scatter plot with respect to the different attributes but when compared to themselves it shows a histogram. The Scatter matrix plots almost all the relationships there are several outcomes some shows positive corelation some shows negative corelation. To sum up is used to identify corelation or covariance with all the attributes.