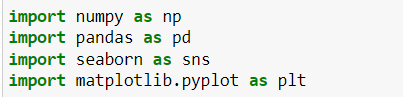
**Car Price Analysis : Using Synthetic Data**

Step 1: Import Python Libraries

Import all libraries which are required for our analysis, such as Data Loading, Statistical analysis, Visualizations, Data Transformations, Merge and Joins, etc.

**Pandas and Numpy have been used for Data Manipulation and numerical Calculations**

**Matplotlib and Seaborn have been used for Data visualizations**



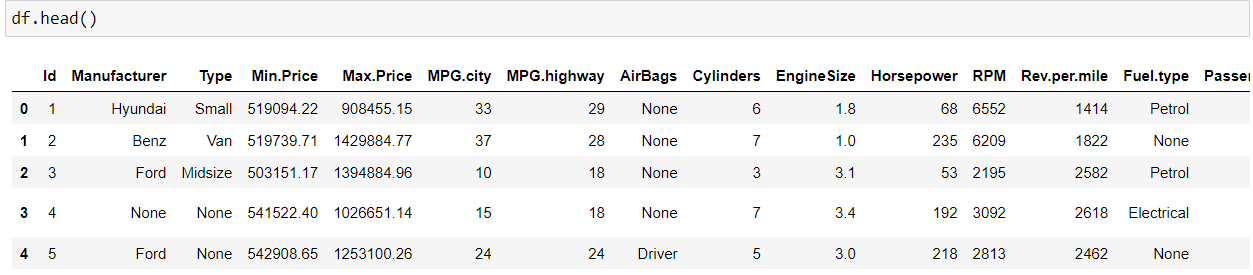
Step 2: Dataset

In this blog, the Synthetic data to predict **car price** is being used as an example. In this dataset, we are trying to analyse the car’s price and how EDA focuses on identifying the factors influencing the car price. We have stored the data in the Data Frame **df.**

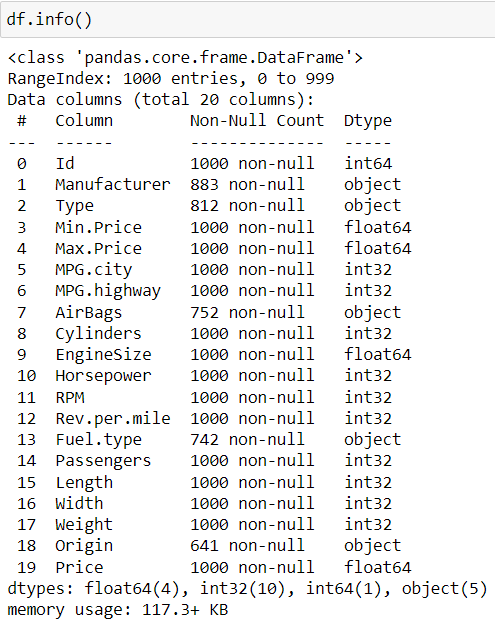


#### **Analysing the Data**

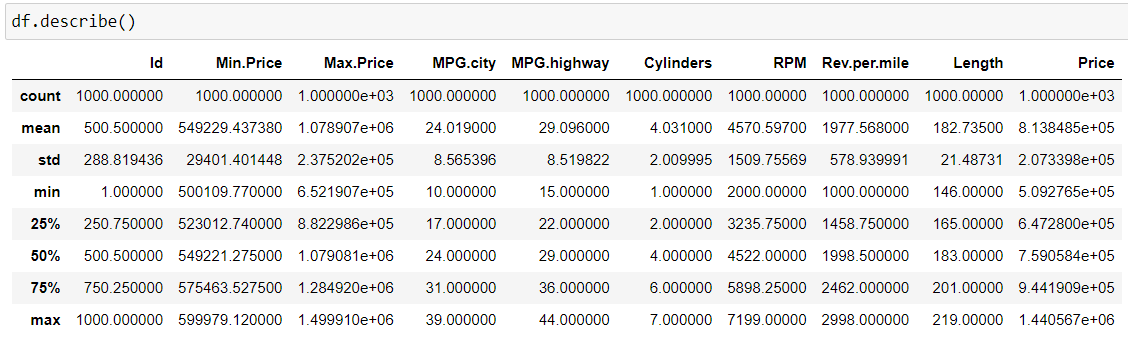
**head()** will display the top 5 observations of the dataset

****

**info()**helps to understand the data type and information about data, including the number of records in each column, data having null or not null, Data type, the memory usage of the dataset

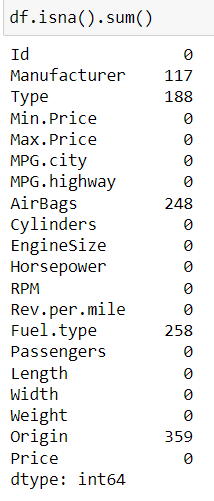
****

**Describe()** It provides a summary of key statistical measures for numerical columns, including count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum values .

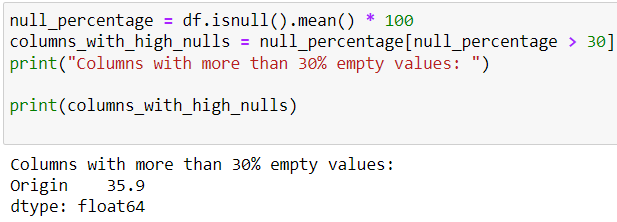
****

#### **Missing Values Calculation**

**isnull()**is widely been in all pre-processing steps to identify null values in the data. In our example,**data.isnull().sum()** is used to get the number of missing records in each column.

****

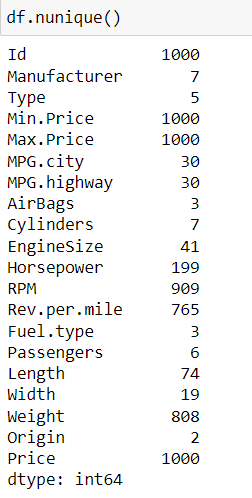
Finding out Columns which have null values greater than 30%.



Here the Origin column has 35.9 % of missing values.

#### **Check for Duplication**

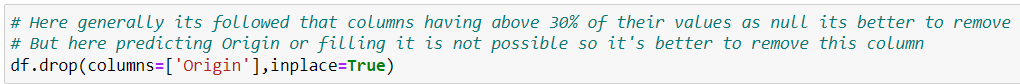
**nunique() based on** several unique values in each column and the data description, we can identify the continuous and categorical columns in the data. Duplicated data can be handled or removed based on further analysis



### Step 3: Data Reduction

Some columns or variables can be dropped if they do not add value to our analysis.

In our dataset, the column S.No have only ID values, assuming they don’t have any predictive power to predict the dependent variable.

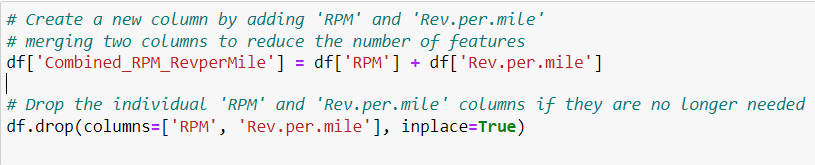


Dropping id column

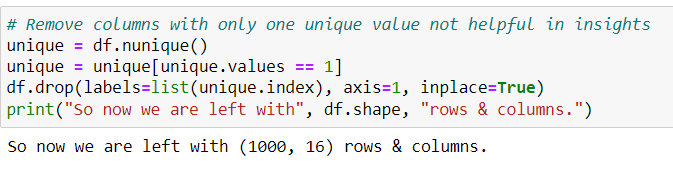


**Combining Related Columns together**

Added both of the columns and not multiply them as values were increasing a large extent due to this more importance will be given to this column during knn imputation so adding was the best option



**Removing Columns with Unique Value as 1 (Same values through out the column)**

****

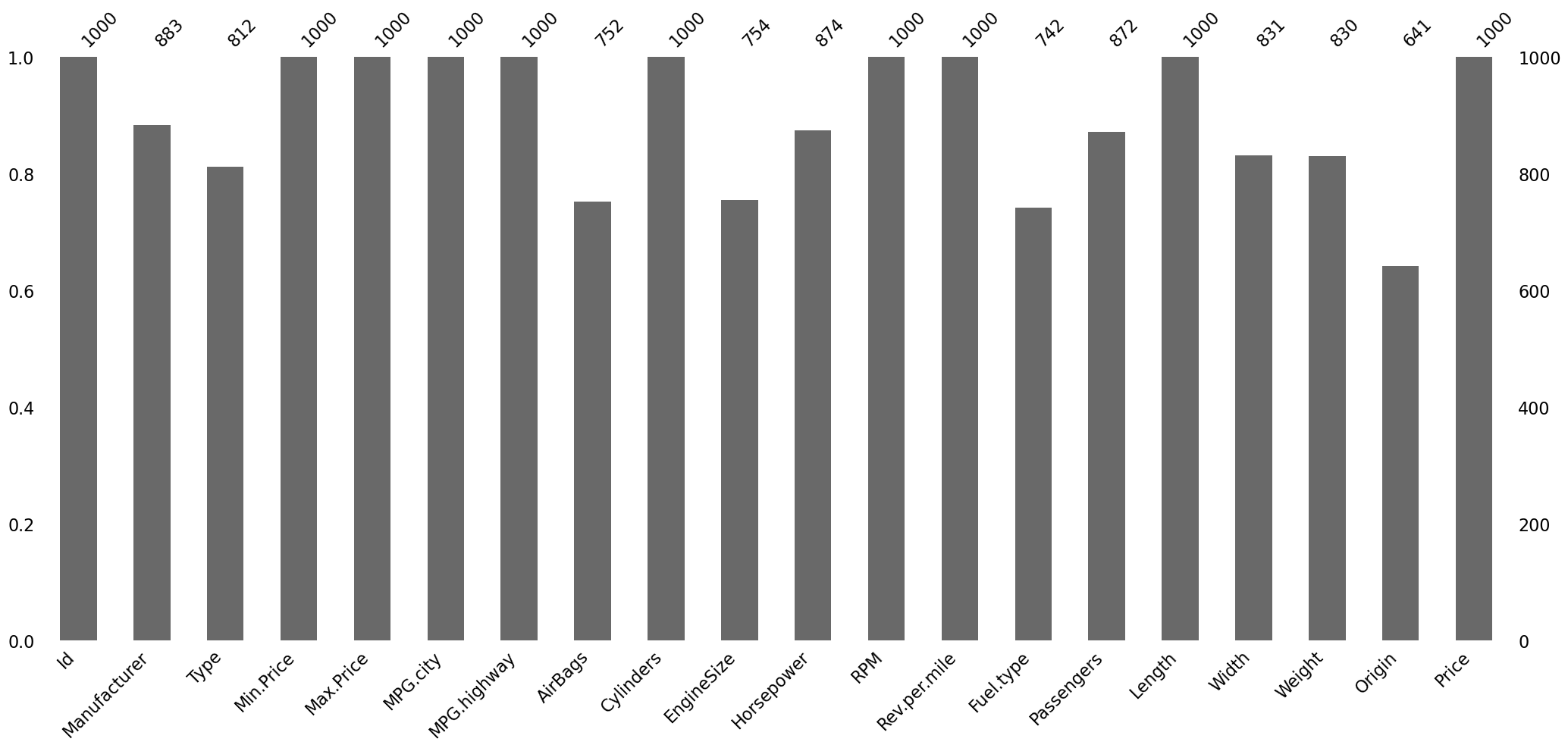
****

****

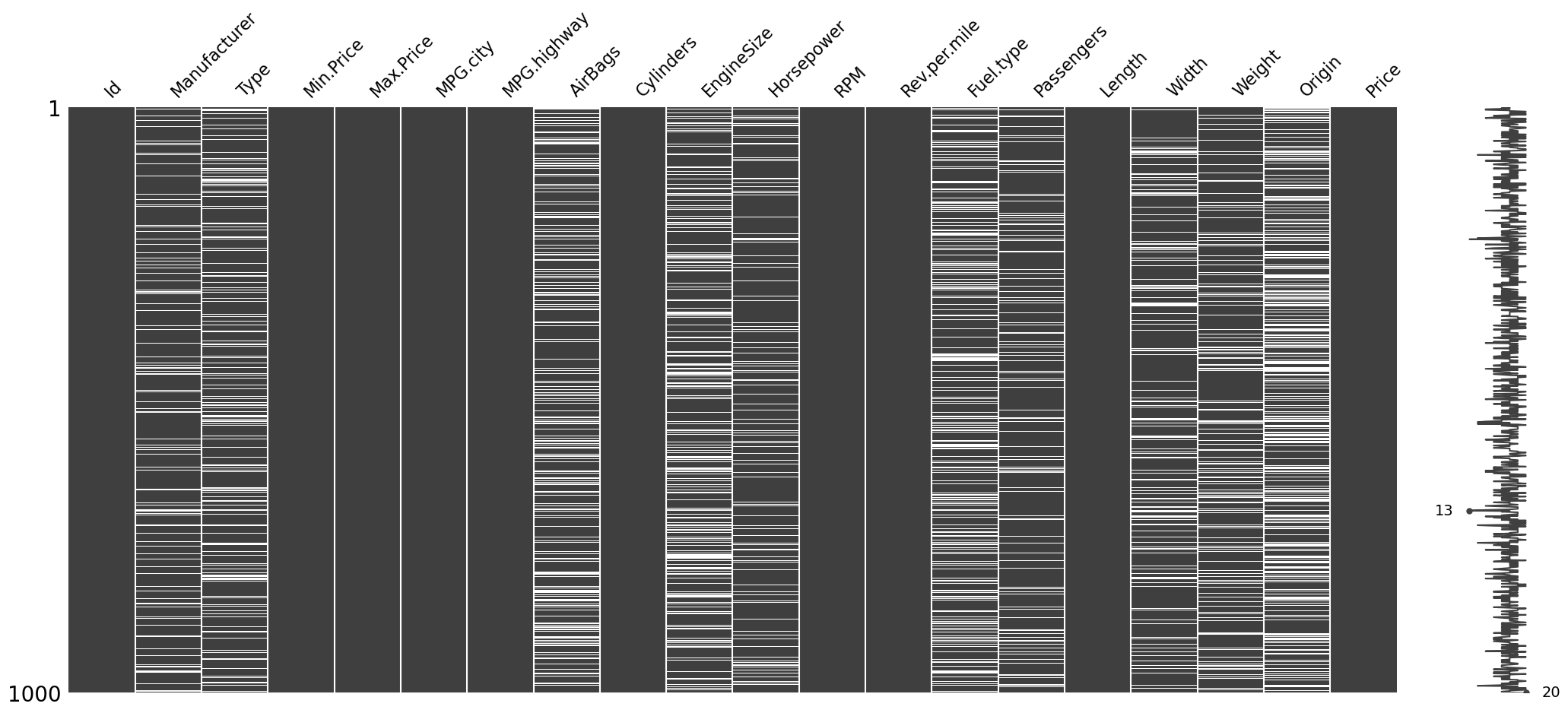
**Step 4 : Imputation Filling**

**These is a visualisation of missing values in my Dataset**

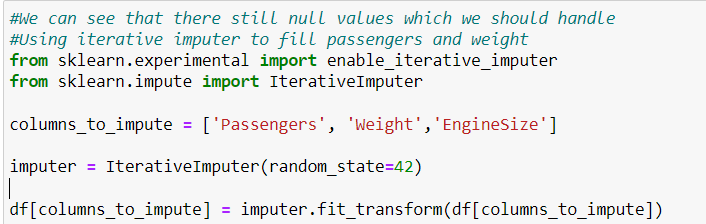
In Bar format



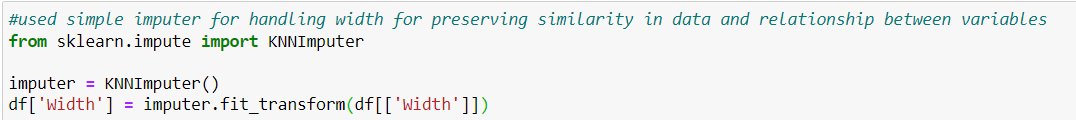
In Matrix Format



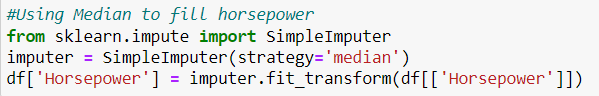
Used Iterative Imputer to fill Passenger, Weight ,Engine Size

****

Used KNN imputer to fill Width



Used Median to fill Horse Power



### Step 5: Feature Engineering

### The column horsepower is a numerical value but to get a better insight used binning to classify them into low, medium, high

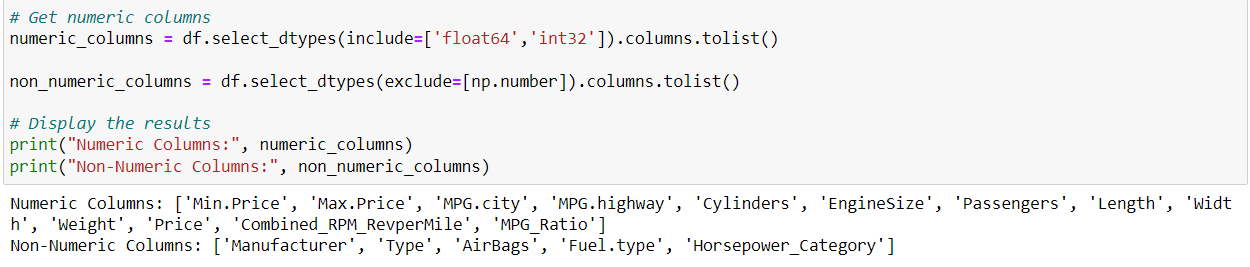
### 

Generate new features by computing ratios between two relevant columns (e.g., 'MPG.city' / 'MPG.highway' to derive a ratio between city and highway mileage efficiency).

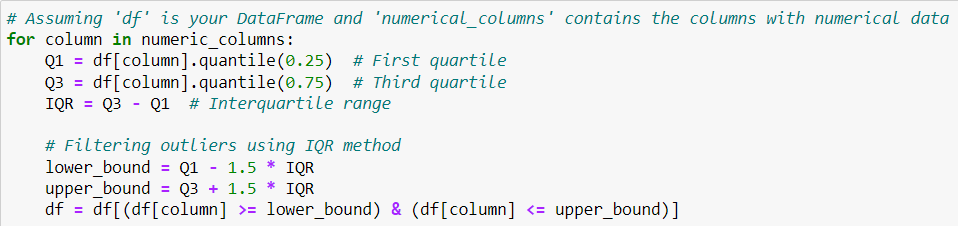


**Step 6: Removing Outliers**

**Dividing into numerical and catrgorical**

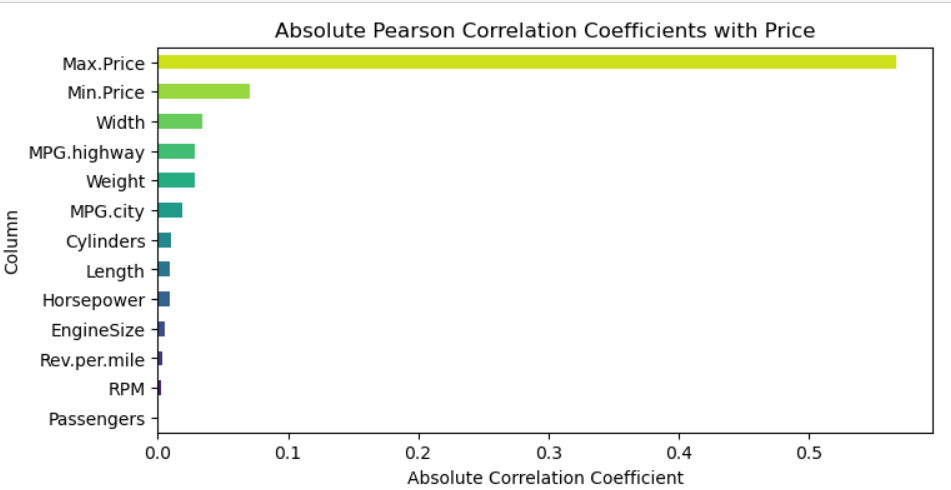
****

Removed the outliers using IQR method for numerical columns



**Pearson Coefficient (linear Relation Ship)**

This is used so as to find the out the linear connectivity between all the independent columns with dependent columns in our data frame so as to see which columns which we can further remove which are unnecessary .

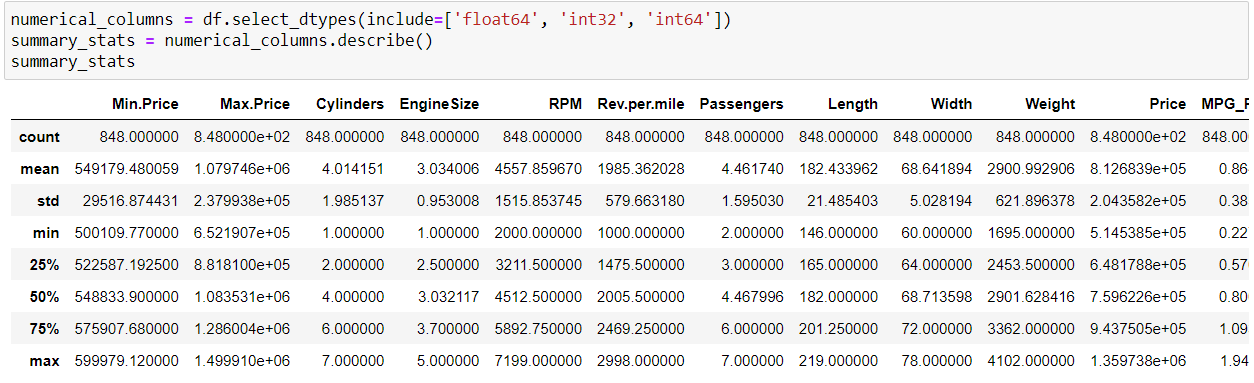


From this we can tell use highly correlated columns for predicting our model further. From this we can see passengers al most has zero correlation with the price column so we can neglect its effects while predicting the prices .

**Step 7 : EDA**

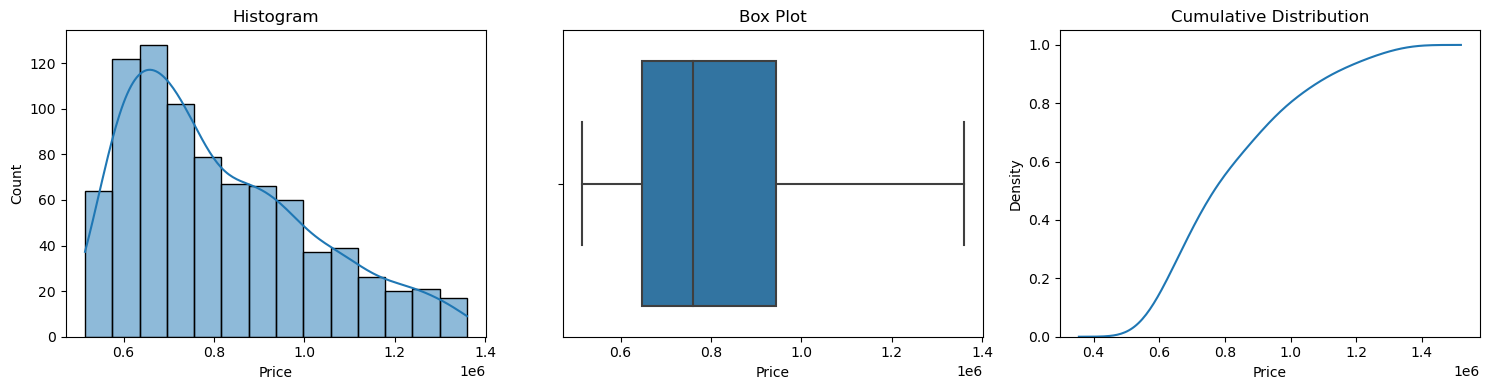
**Univariate Analysis**

Finding out numerical and categorial columns in the dataset

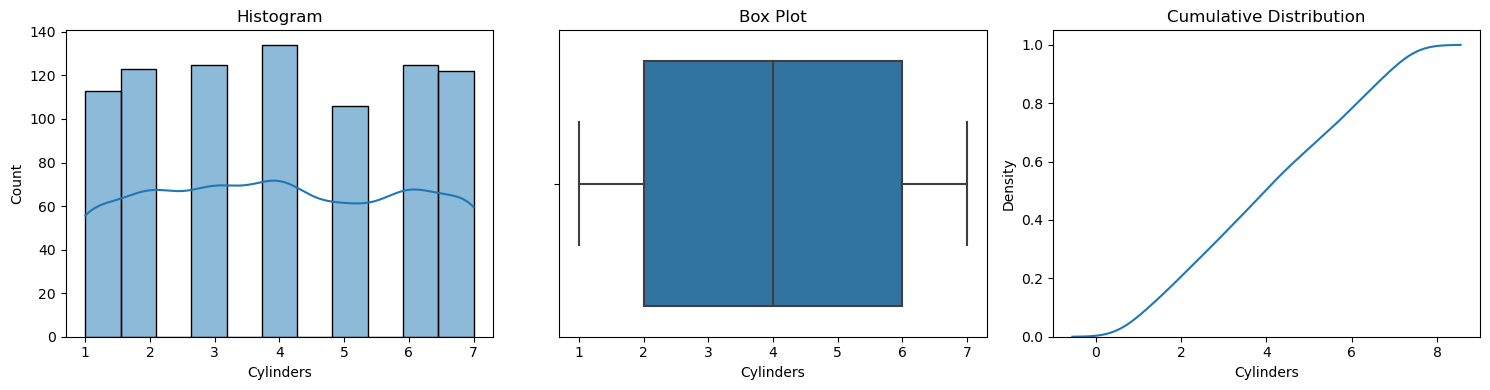


Used Box plot , histogram , Cumulative distribution charts to find the skewness and its distribution

Below is the price column as we can see its left skewed in nature (6-7.5 lakhs was most)

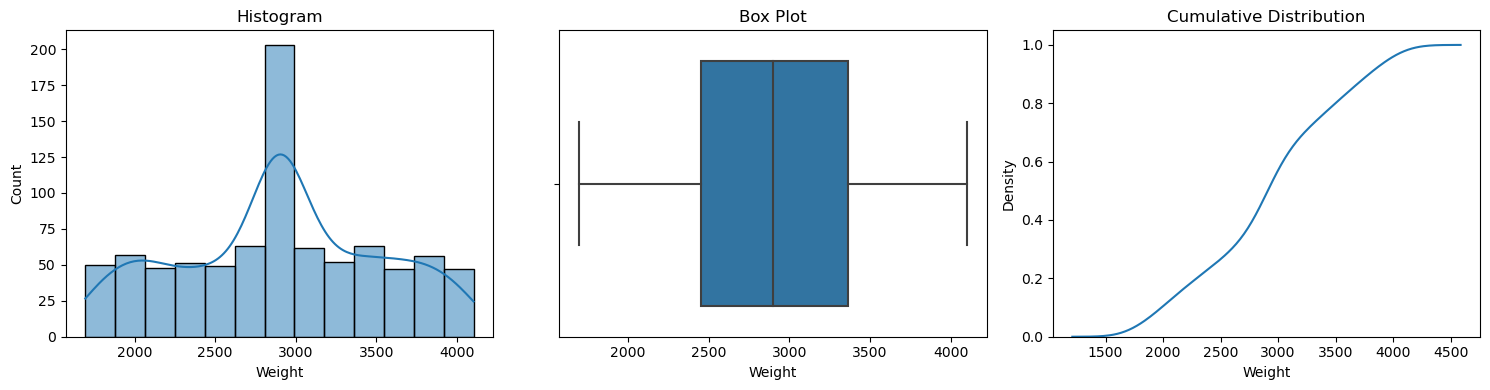


Below is for Cylinders column its not skewed and uniformly distributed

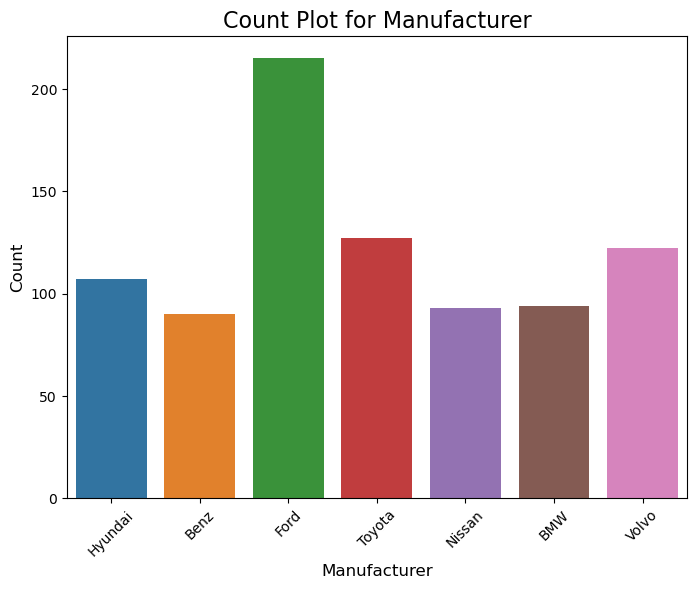


From my observation engine size , weight and width had a value which was most repeating

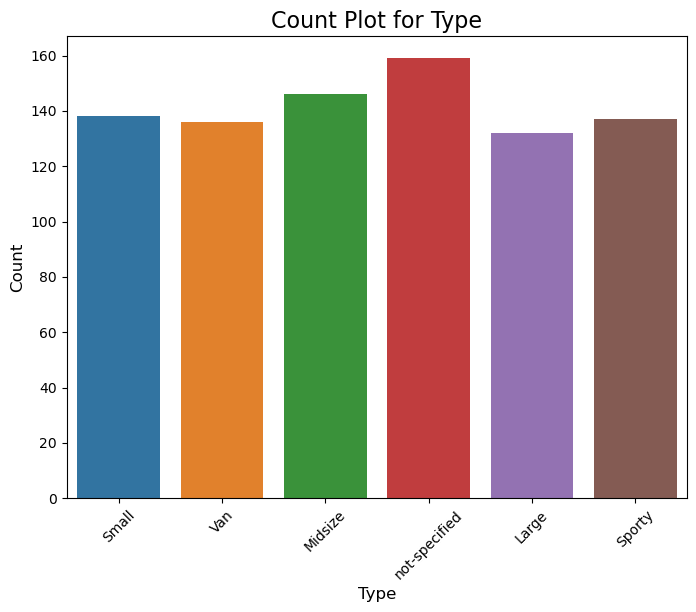
Below is for Weight (2833.3,3000) was most repeating



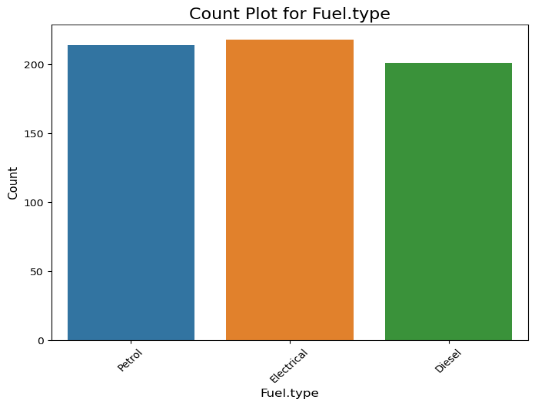
Used count plot for visualising Categorical Columns



Most of the cars were of type Ford followed by Toyota

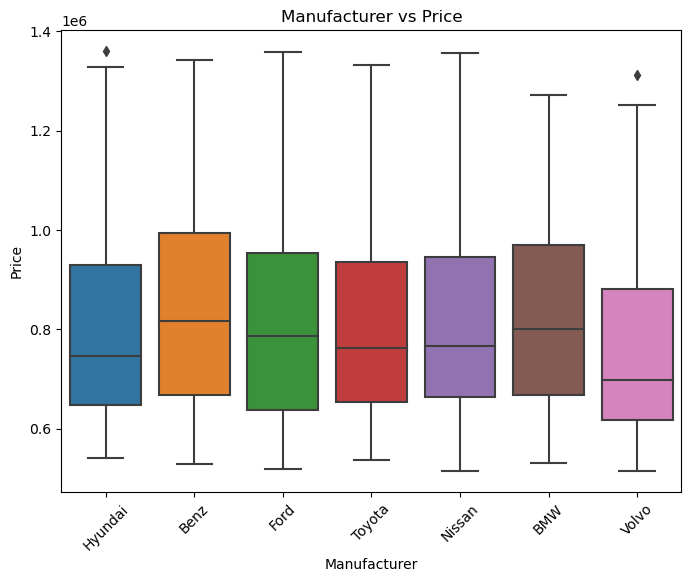


In our dataset most of the cars were not specified by the company followed by midsize then sporty.

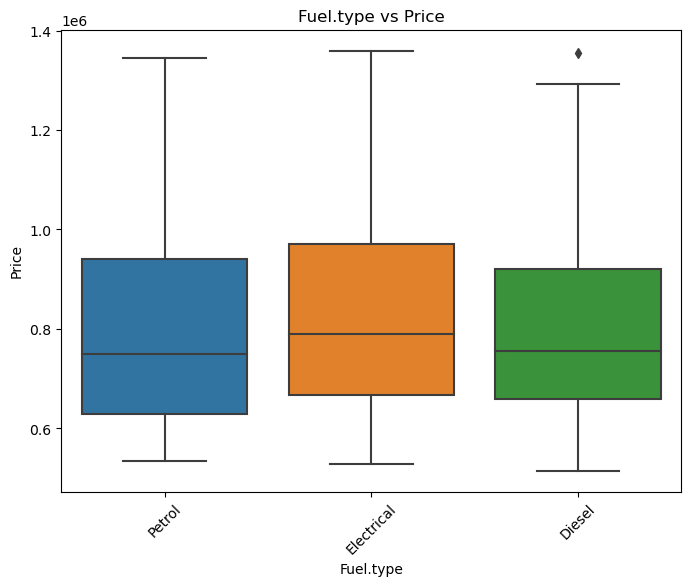


Most of the cars were of Electrical then petrol the lest was Diesel.

**Comparison of Categorical Columns with Price and Visualising through Box-plot**



Toyota and Hyundai have the lowest median car prices Benz and BMW have the highest median car prices

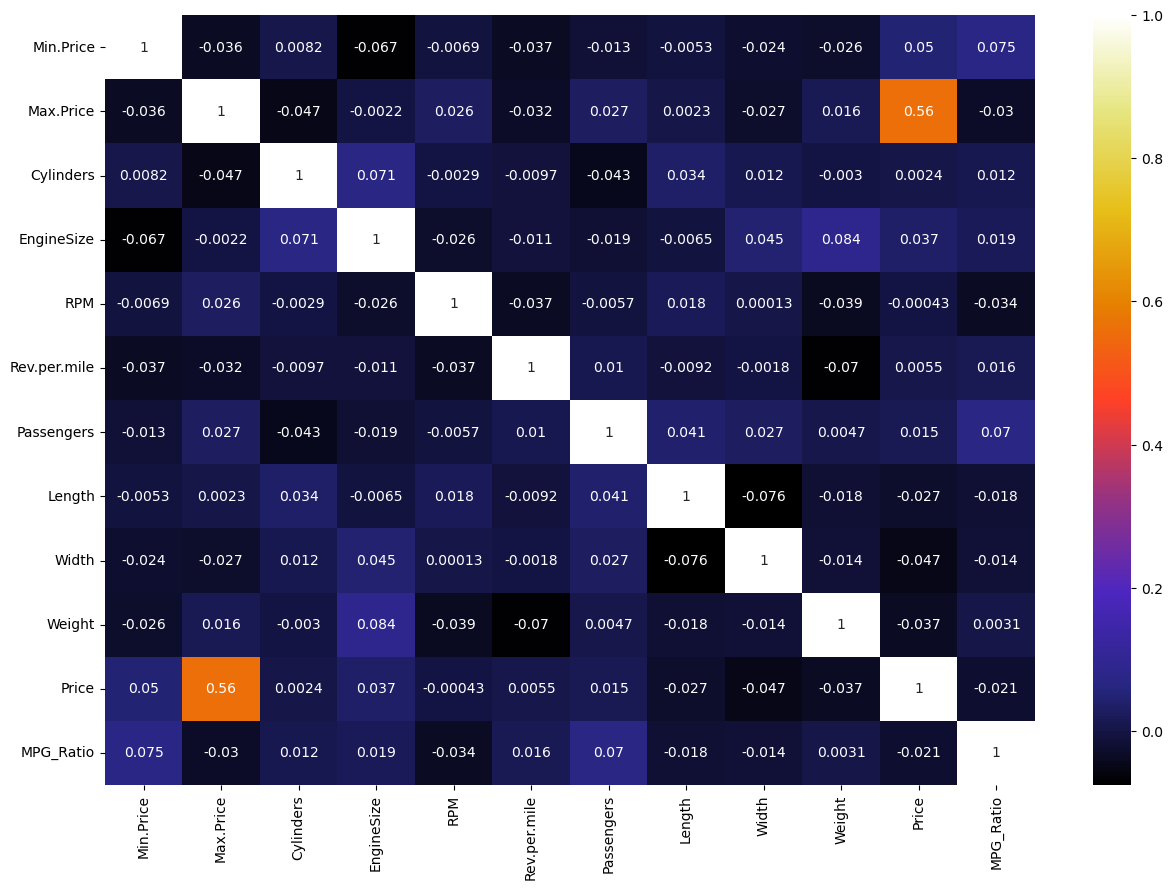


Electrical fuel type has the most consistent pricing. The median price for Petrol is the lowest the median for Petrol prices is lower than the medians for Electrical and Diesel prices.

**Multivariate Data Analysis**

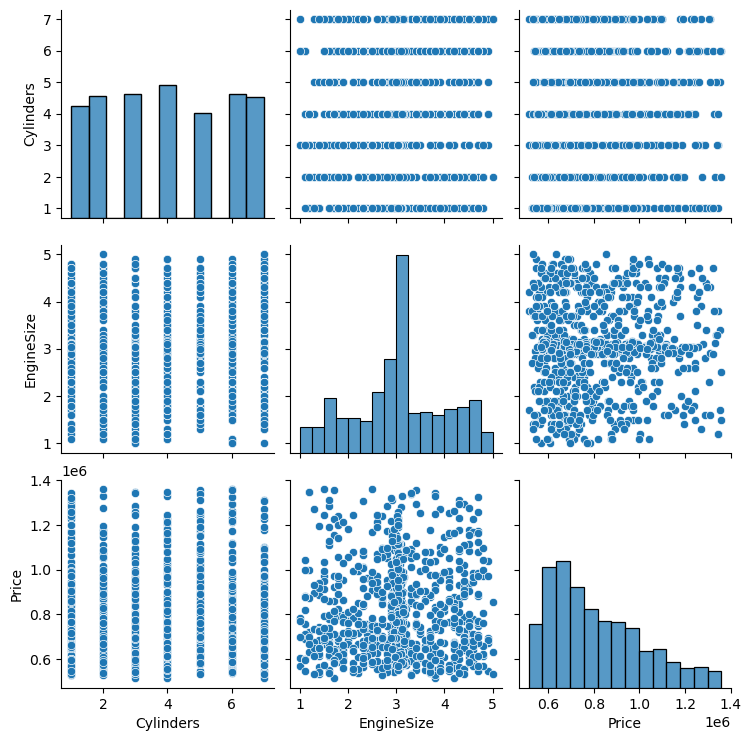
Heatmaps transform numerical data into vibrant visual landscapes, revealing patterns and relationships that might otherwise remain hidden. These colourful matrices paint a compelling picture of data density and variation, with each cell's hue representing a corresponding value. Heatmaps excel at untangling the complexities of two-dimensional data, offering an intuitive visual summary that empowers exploration and discovery.

Below is the correlation matrix between different Columns of the numerical Dataset



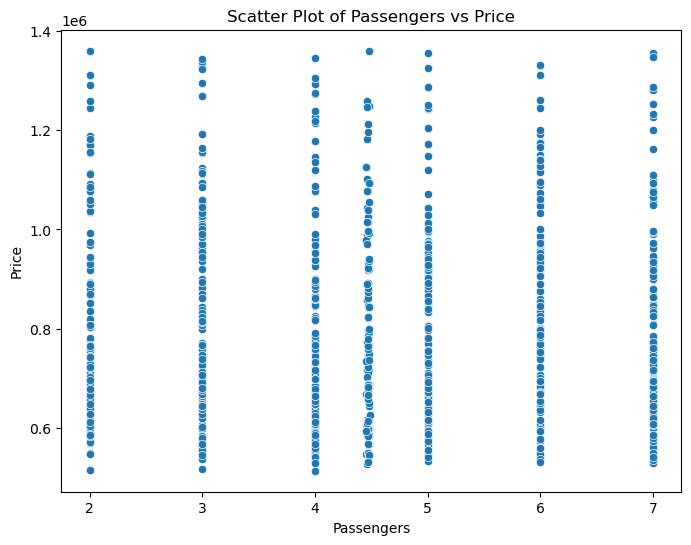
As we can see Max\_Price and Price column are closely related and have a correlation factor of 0.56

**Visualising Engine Size , Cylinder relationship with Price**

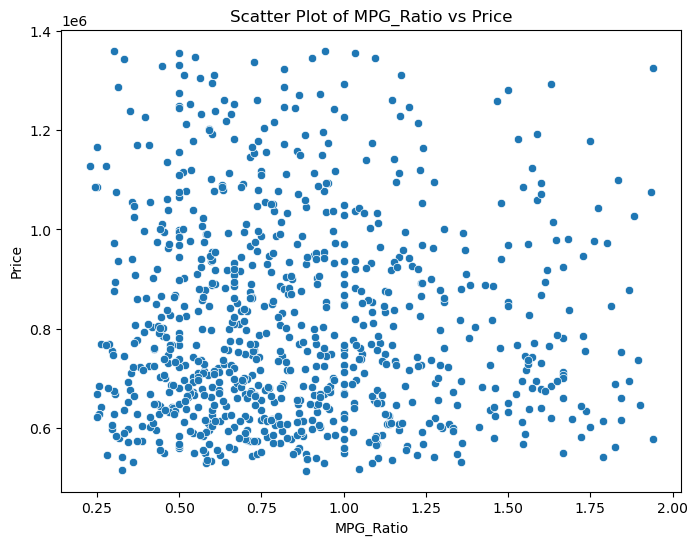


Bigger engines (more cylinders) generally mean bigger price tags, but not always .Price clusters reveal popular engine sizes within certain budget ranges. Outliers like luxury cars or fuel-efficient hybrids defy the trend. These are the observations from this plot .

**Scatter Plot**

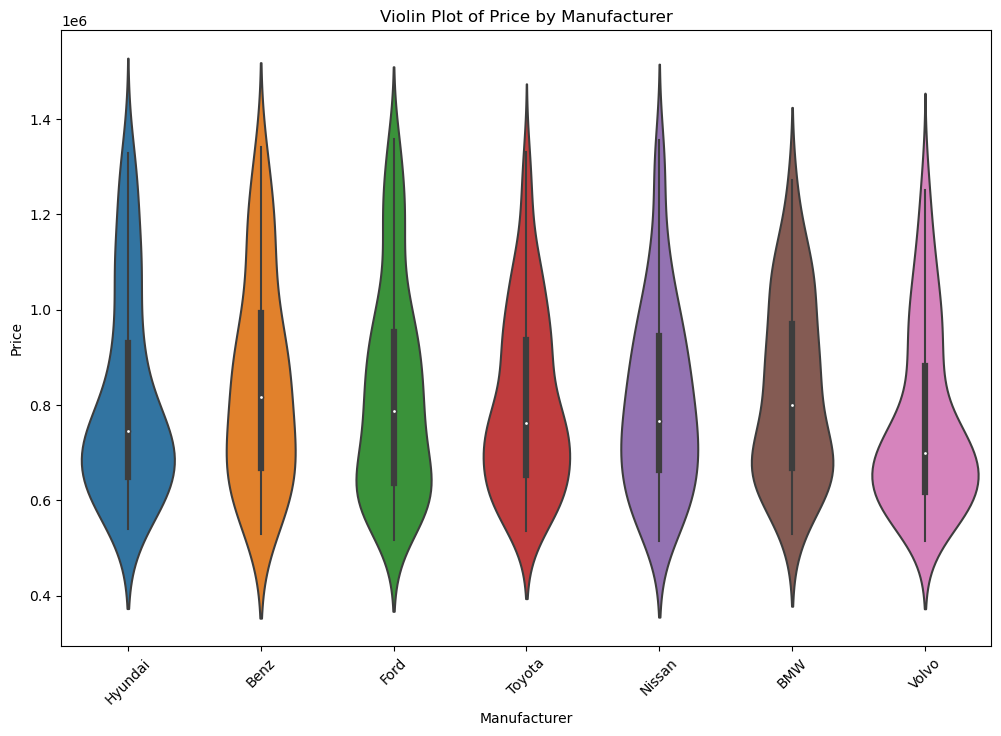


From this plot we can see the relationship between passenger and price they are somewhat distributed because price also depends on brand , engine , fuel type etc



MPG\_Raito defines the mileage ratio of the car from the above plot shows the vehicle mileage ratios relation with the price they are somewhat directly proportional as high mileage means efficiency of the car is high hence price is also higher.

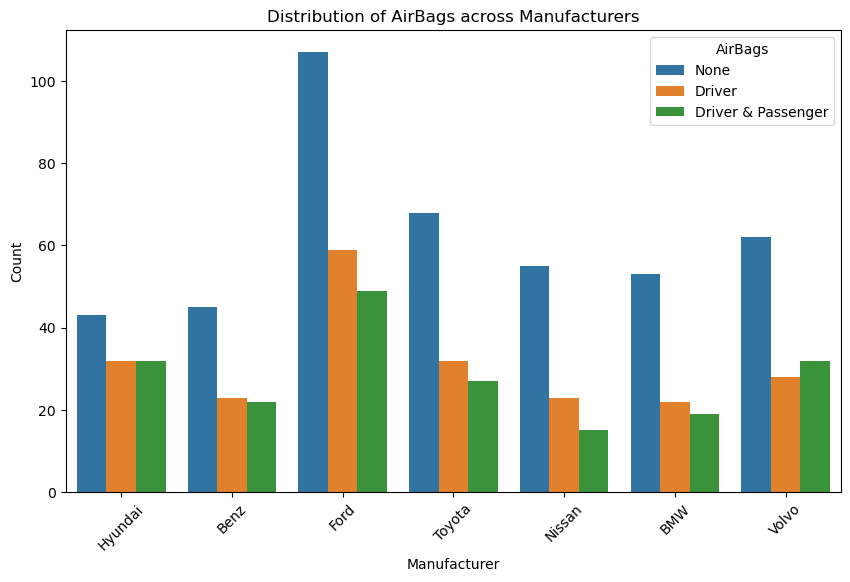
**Violin Plot for Manufacturer and Price**



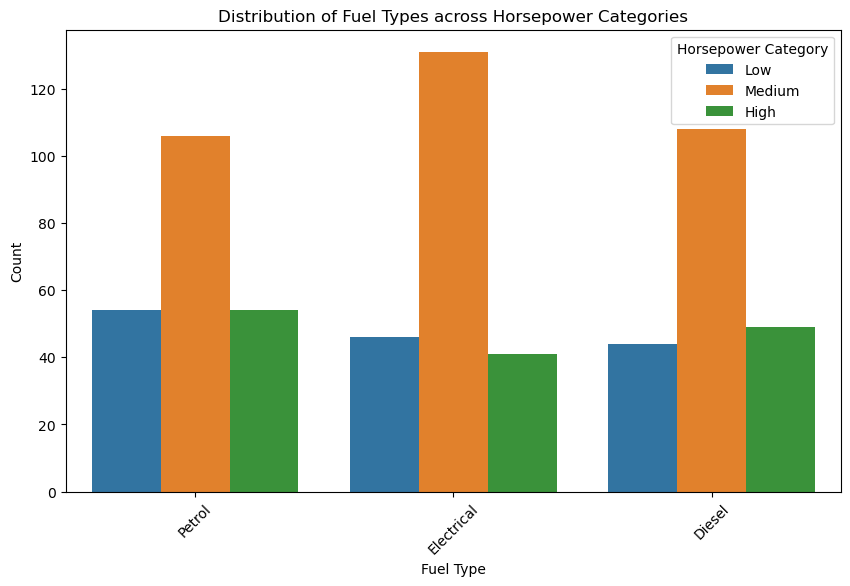
For the above violin plot Benz has a median value of 9,00,000 which is the highest of all the manufacturers medians followed by bmw . Hyundai has most cars with highest value 6,80,000 the car with highest price in our dataset is also Hyundai while most cars of volvo have less price . Benz also has the car with the least price followed by Nissan .

Hyundai’s violin plot is bigger indication greater spread of the prices where as Nissan’s violin plot in thinner indicating most of the cars have similar prices. In length Benz is the bigger one of all indicating highest difference between lowest and highest cars.

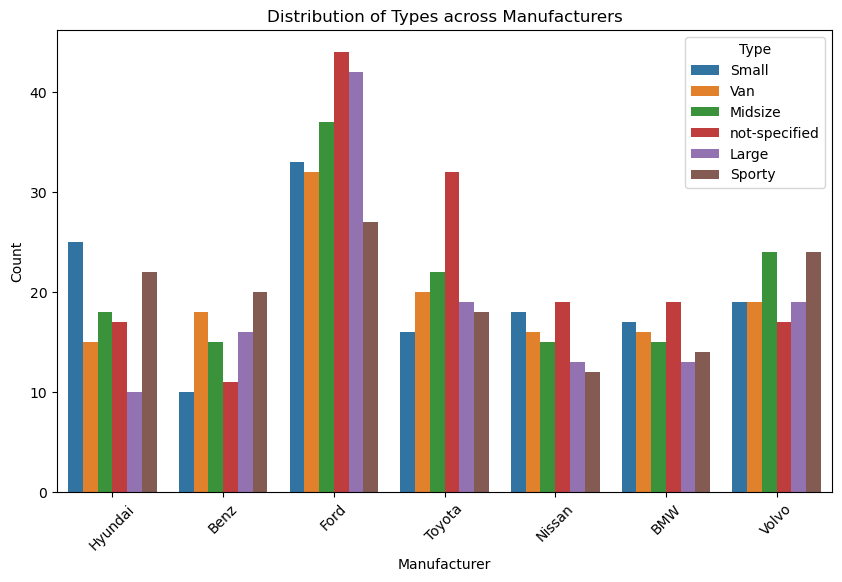
**Visualising Categorical Values using Counter plot**



Here in ford most of the cars do not have airbags so in safety measures its lowly rated but in compared to all manufacturers proportionately Hyundai is advisable as equal number of cars have airbags for both drivers & drivers and passengers similarly with volvo most of the cars have airbags for drivers and passengers . While Nissan has cars with least number of airbags for both drivers and passengers while Benz has cars with least no of airbags for drivers .

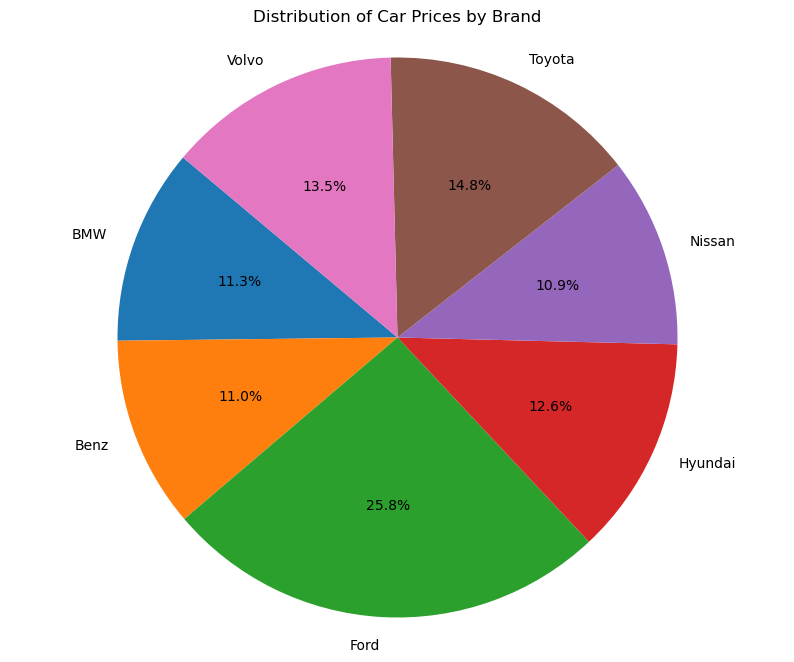


If you are planning to buy a car having medium horsepower its better to chose an electrical cars where as if you want to buy a car which has high horse power its better to choose a petrol car



Here we can see Hyundai has most cars of small types so its advisable to look into Hyundai cars first if you want to take a small car as it has more options . But if you want to buy car of large type choose ford and for sporty you can choose volvo for variety of options to choose Ford also caters to van type cars the most .

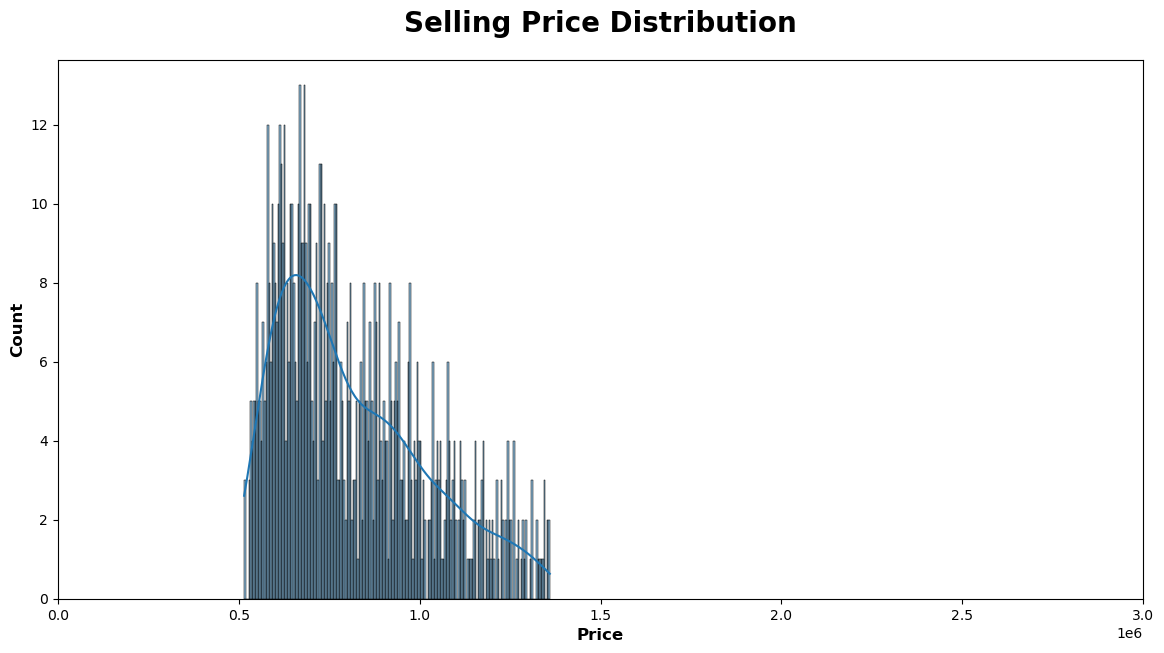
**Pie Chart :**



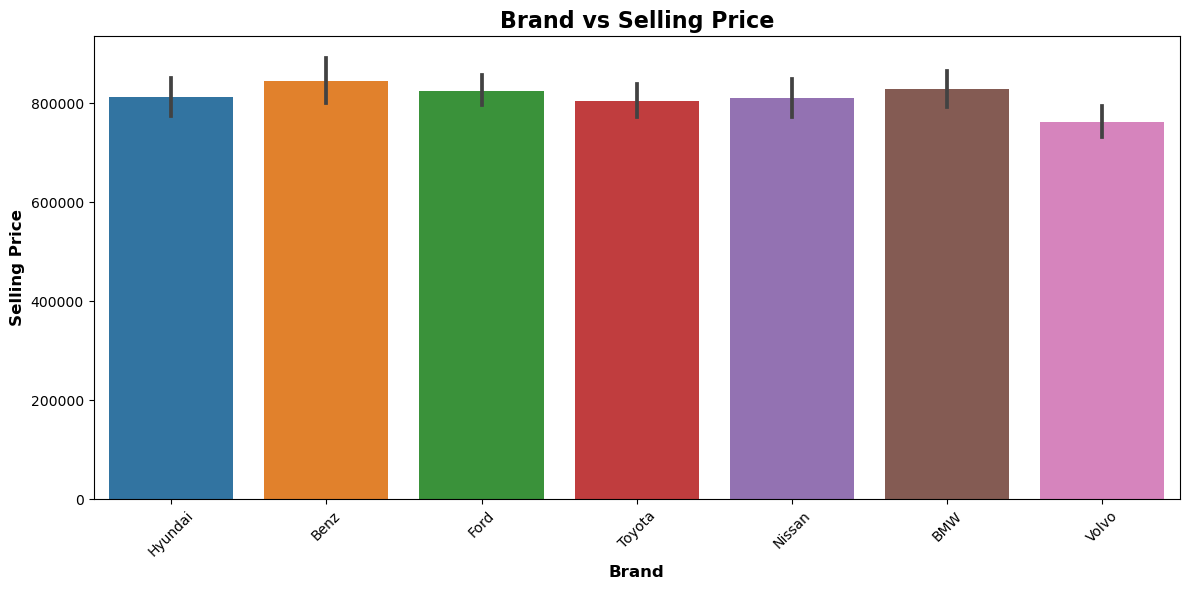
In our data set ford cars are most followed by toyota then Hyundai and bmw and so on form this we can know the distribution of the cars and their percentages in our dataset meaning data on ford cars is more in general .

Some More Insights :-

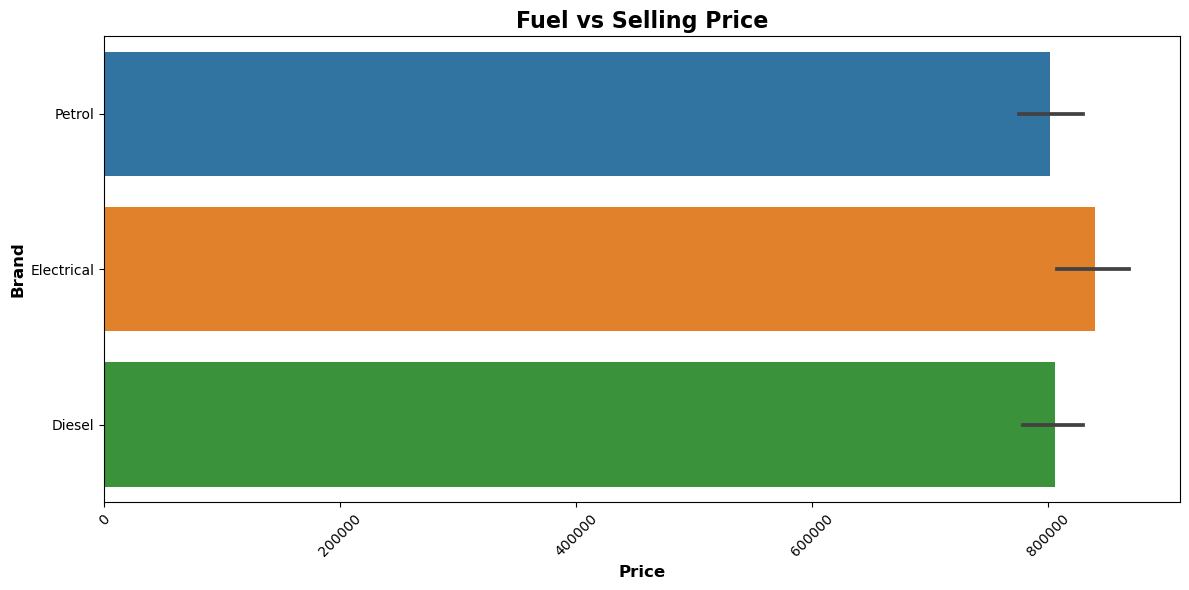
**Histogram plot for price and Count**



Here most of the cars are between (0.5-1.0)\*10^6 range from the above the image which shows the average price of most of the cars in our dataset



From this selling price of Benz is most followed by bmw then Hyundai then ford so that we can get a clear visualization of the cars price according to the company they belong to.



From here we can tell that Electrical Cars have higher selling price compared to Petrol and diesel so they are more expensive in general if you want to buy it followed by Diesel least is petrol.