**Assignment-2**

**Vedanth goud Nagapola**

**16354630**

**Data Clean:**

**Given Dataset:**

A screenshot of a computer

Description automatically generated

There are different attributes present in the dataset.

1. **Look for the missing values in all the columns and either impute them (replace with mean, median, or mode) or drop them. Justify your action for this task.**

* First, we need to read the dataset and get to know the null values in each attributes for data preprocessing .

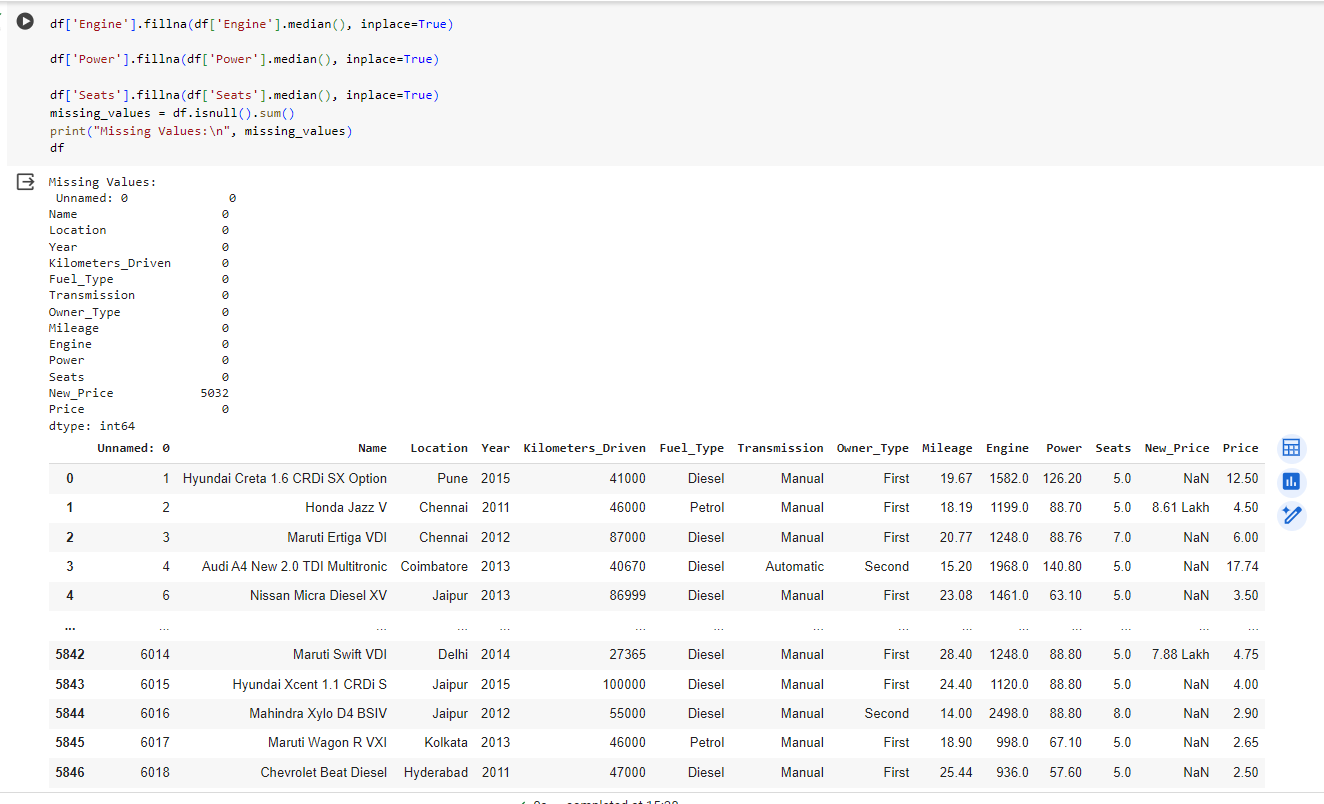
A screenshot of a computer code

Description automatically generated

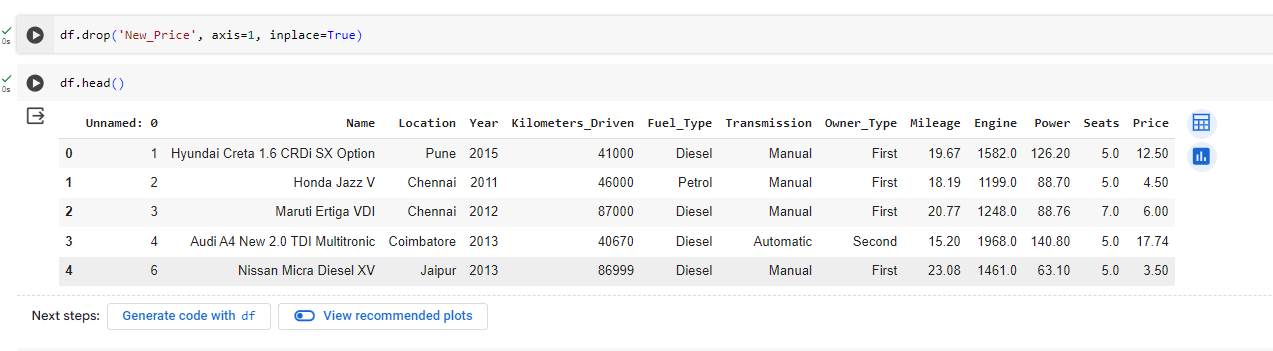
* Columns such as Engine, Power, Seats, and New Price have missing values. We can utilise the mean, mode, and median to fill in those NA values.

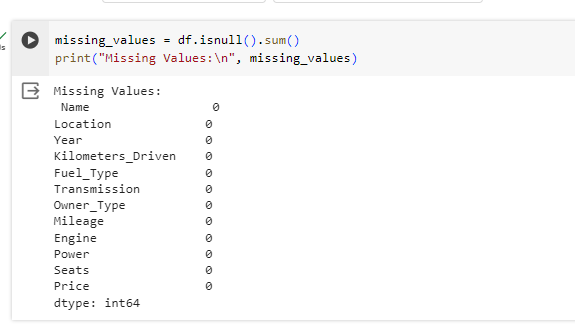
1. **Remove the units from some of the attributes and only keep the numerical values (for example remove kmpl from “Mileage”, CC from “Engine”, bhp from “Power”, and lakh from “New\_price”).**

* To achieve a better outcome, we must be sure to remove the units from those columns before doing the preceding step.
* The pre-processing methods listed below can help improve the accuracy of the data.

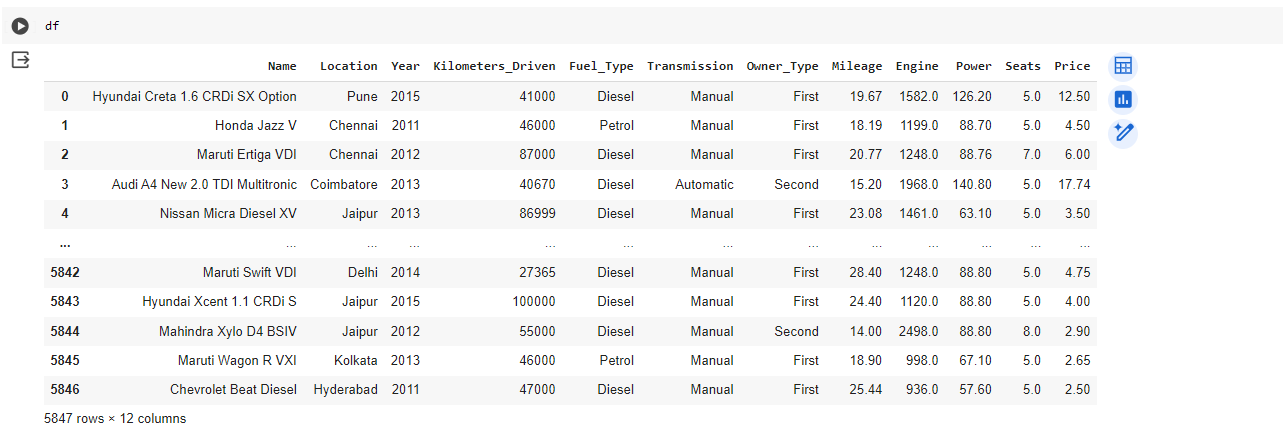


* As we see from the above fig. we can analyse the missing fields are sorted with median values.



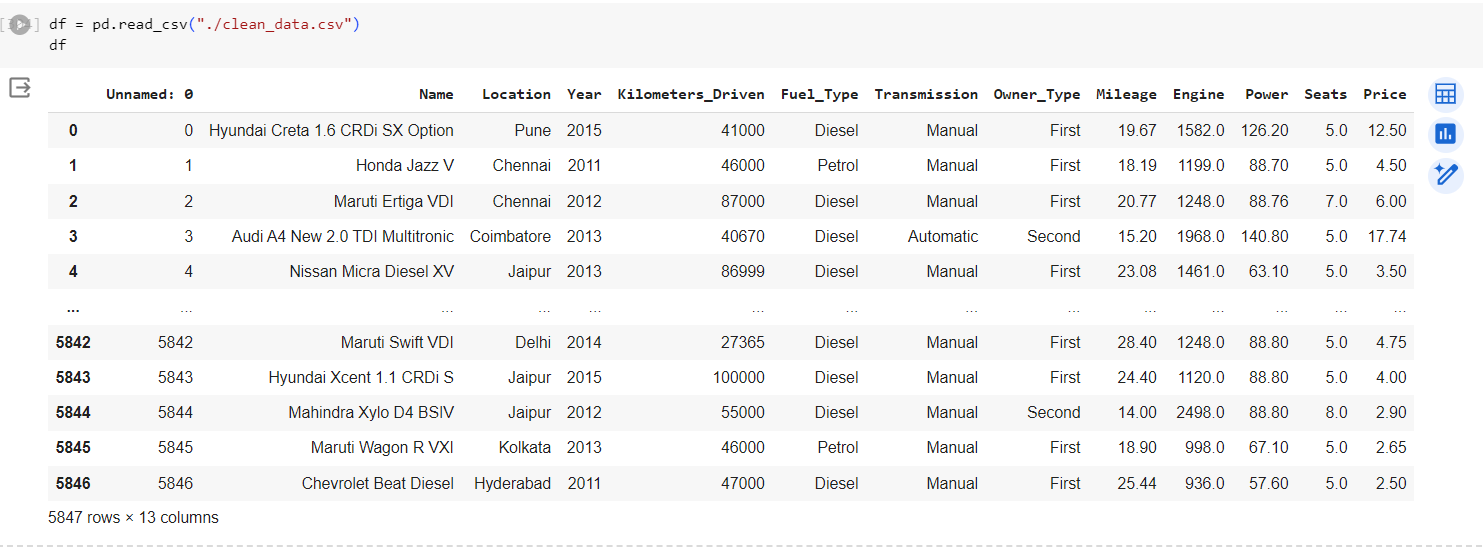


* At last the pre-processed dataset converted into the **clean\_data.csv**.

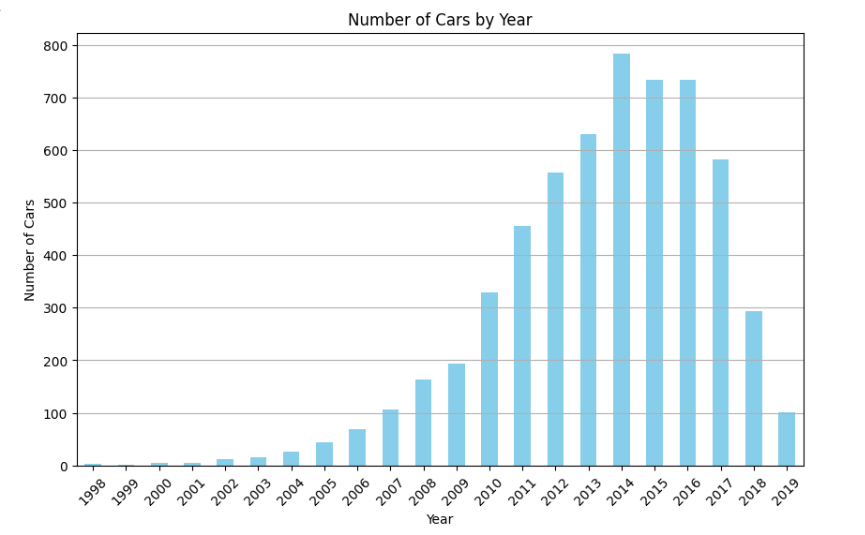


**Data Analysis:**

* Read the pre-processed dataset **clean\_data.csv** to perform analysis on the data .

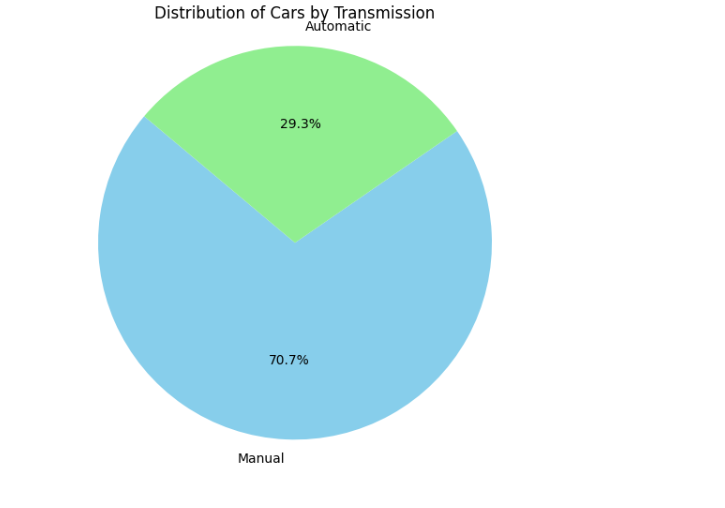


**Line graph:**

****

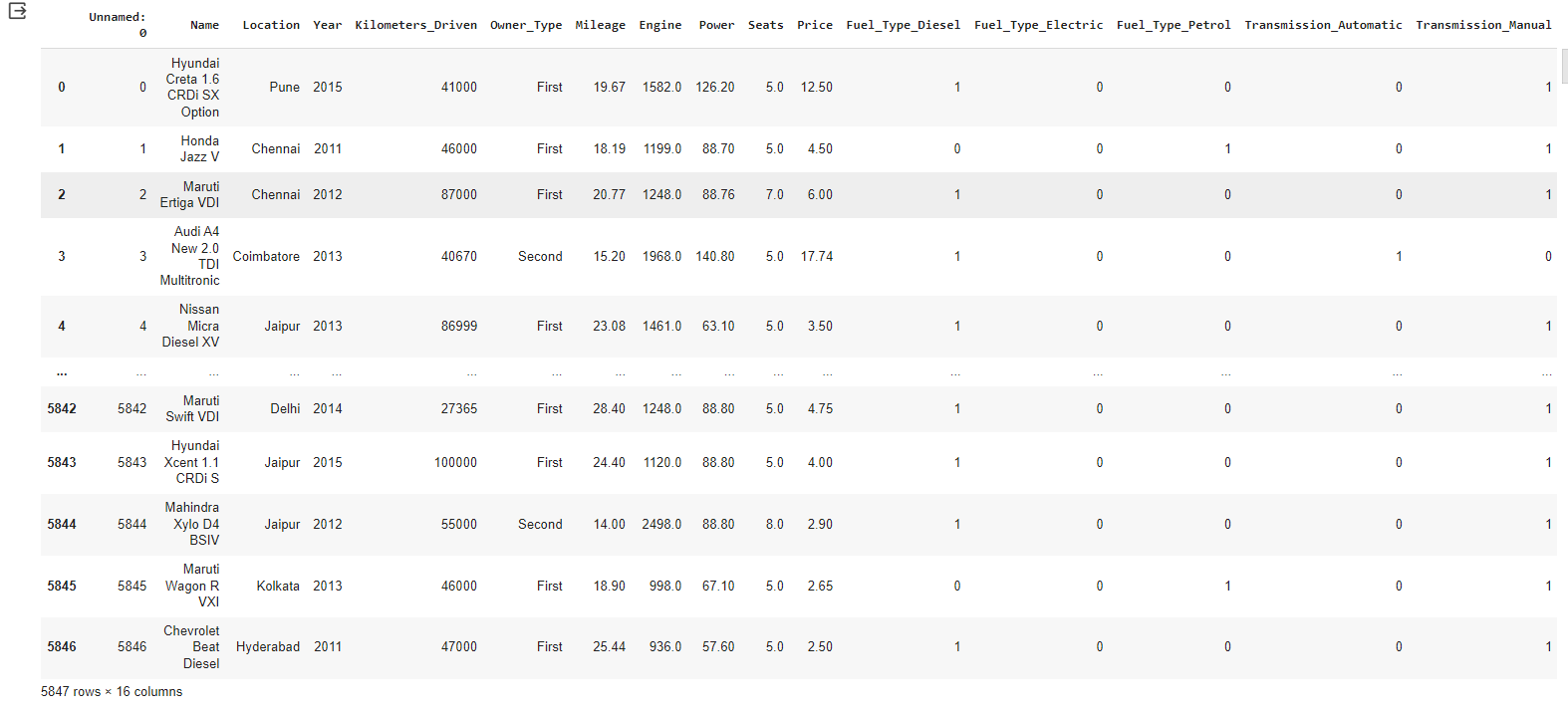
* According to the line graph, the number of cars sold appears to have increased steadily between 1998 and 2008. There is a sharp decrease in car sales between 2008 and 2010. Sales appear to have increased somewhat since then, but they level off around 2014 and don't seem to surpass pre-2008 levels.
* It appears that there are more cars in the year 2013 compared to that of other years .

**Pie Chart:**

****

* This pie chart shows the distribution of cars by two categories: automatic and manual transmission. The green circle slice represents the percentage of cars with automatic transmission, while the blue slice represents the percentage of cars with manual transmission.
* The percentage for automatic transmissions is labeled as 29.3%. The percentage for manual transmissions is labeled as 70.7%.

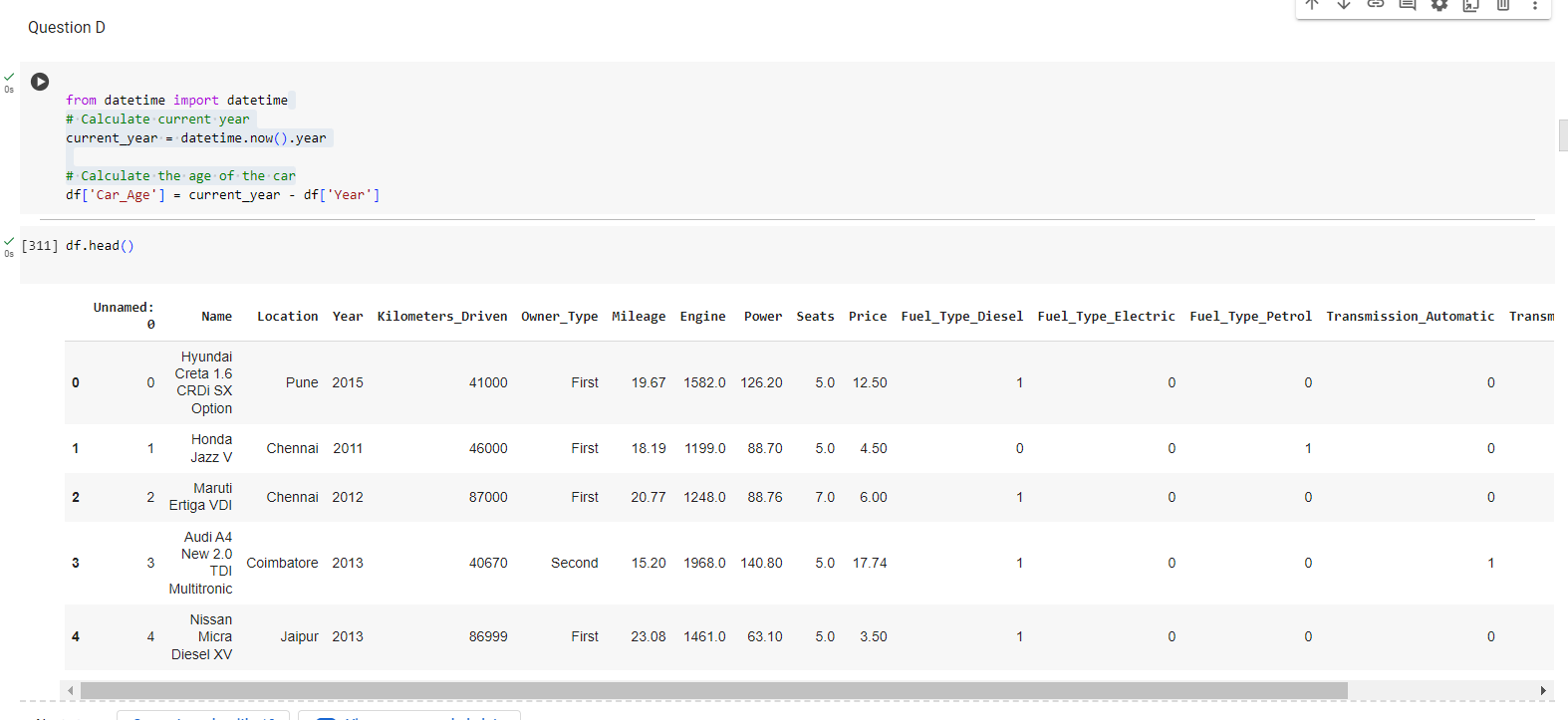
1. **Change the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot encoded value.**

****

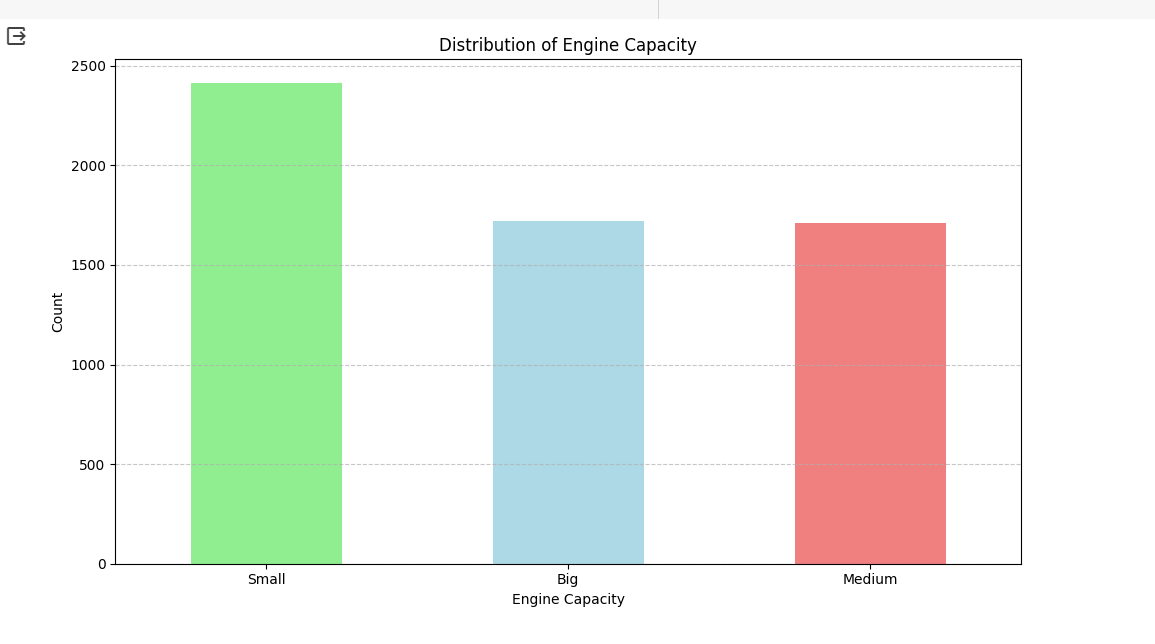
* The Fuel\_Type and Transmission categorical variables are converted into the one hot encoded values. we can see that each unique category within "Fuel\_Type" and "Transmission" has been converted into a separate binary column (0 or 1).

1. **Create one more feature and add this column to the dataset (you can use mutate function in R for this). For example, you can calculate the current age of the car by subtracting “Year” value from the current year.**

* **Added new column “Car\_Age” and “Engine\_Capacity”.**

****

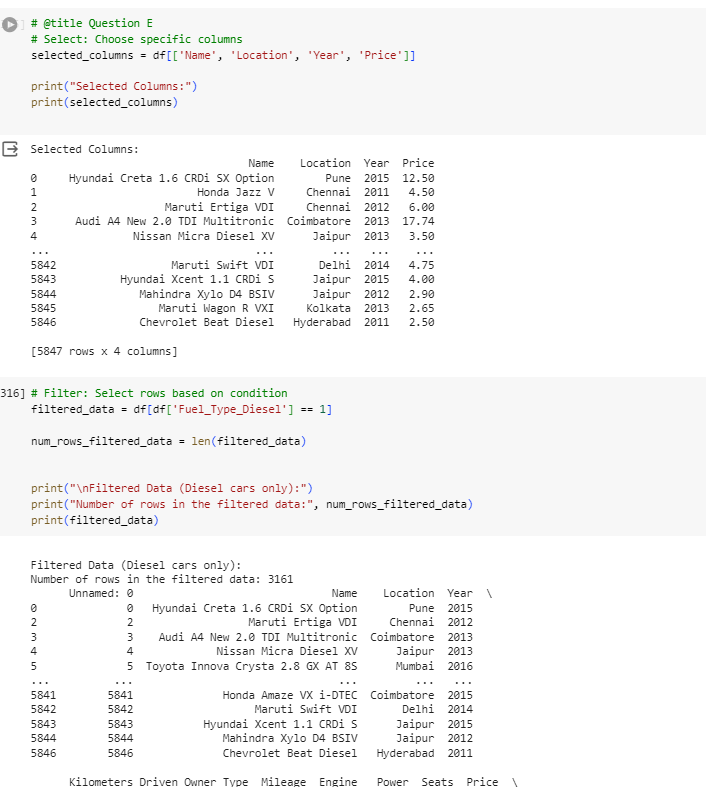
**Bar Chart:**

****

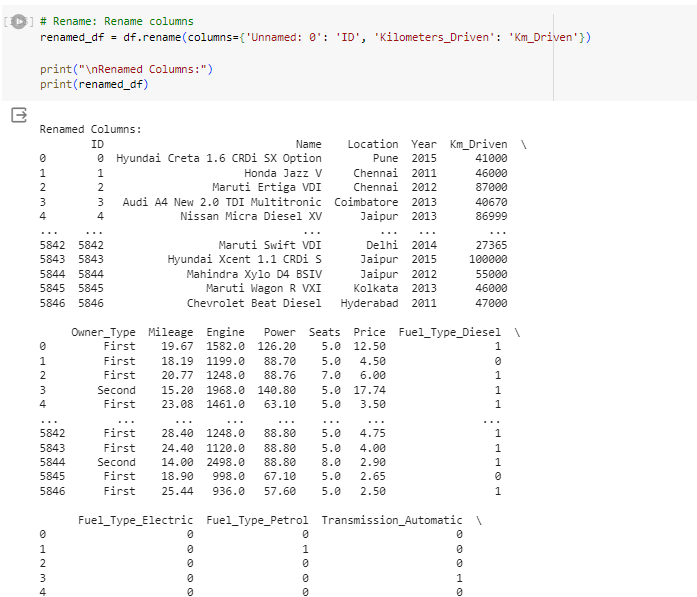
The distribution of cars into three engine capacity categories (small, medium, and large) is depicted in the graph. The graph's y-axis displays the number of cars, but the scale is left unlabelled. The "Small" engines category is the largest. Compared to cars with medium and big engines, there are more cars with small engines. Because the scale on the y-axis is not labelled, it is challenging to estimate the precise number of cars in each group. Conclusions regarding the causes of the engine size distribution are hard to come by in the absence of additional data.

1. **Perform select, filter, rename, mutate, arrange and summarize with group by operations (or their equivalent operations in python) on this dataset**.

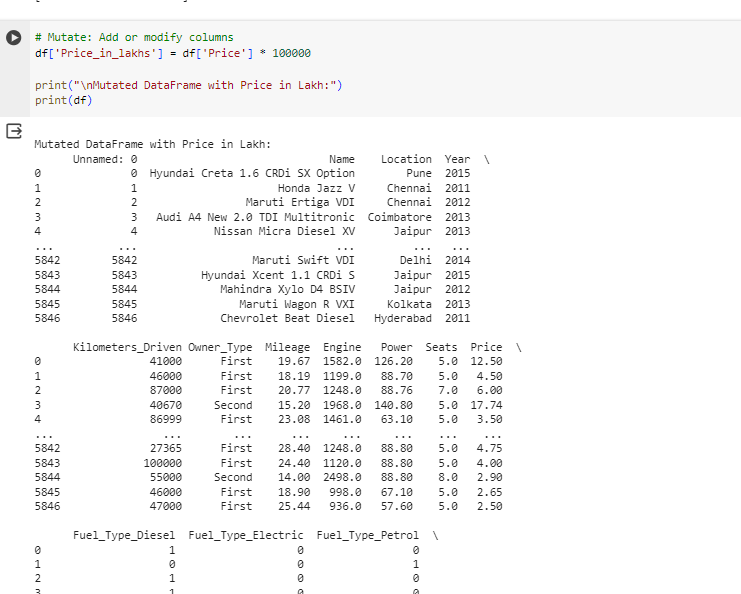
* **Returning the selecting data:**

****

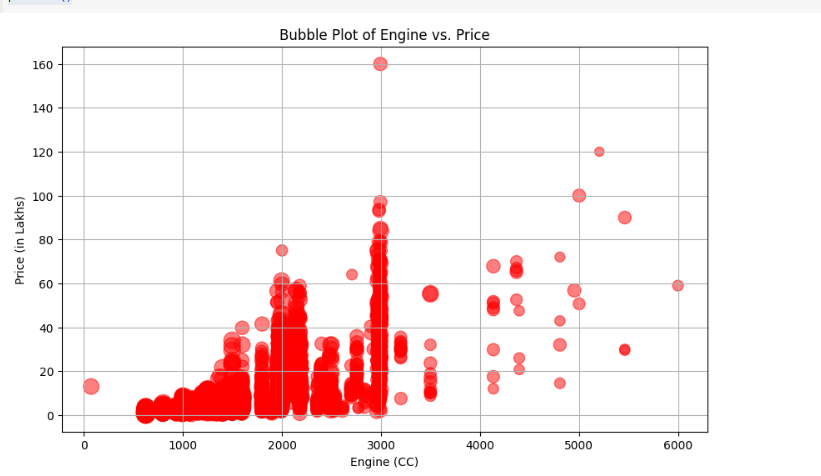
* **Renaming the columns:**

****

* **Add or modifying the columns:**

****

* **Scatter Plot:**

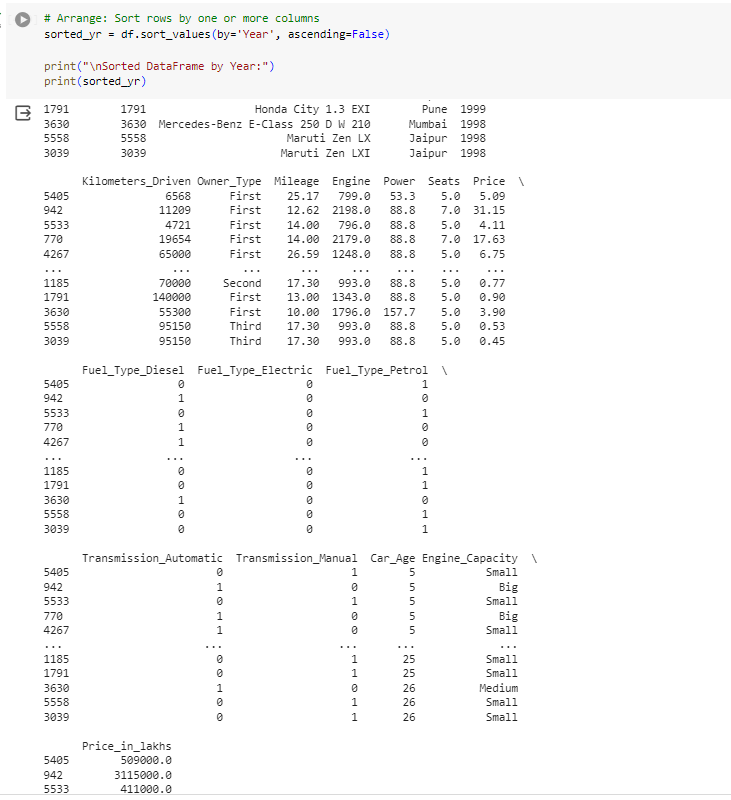
****

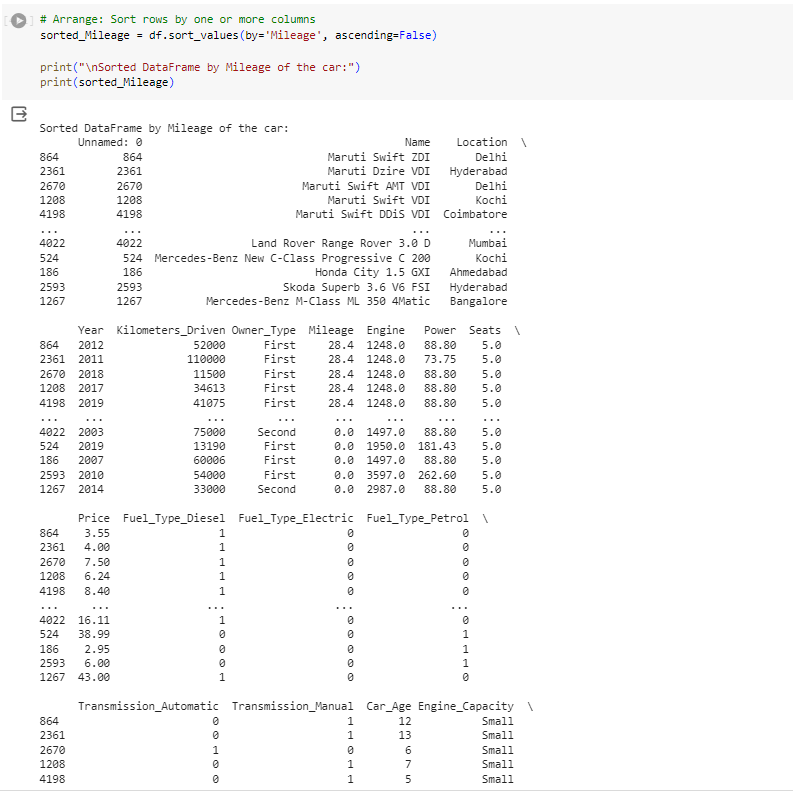
* The image of scatter plot titled "Scatter Plot of Engine vs. Price". It shows the relationship between the price of an engine and the engine size measured in cubic centimeters (cc).

The x-axis represents the engine size (cc).

The y-axis represents the price of the engine in lakhs.

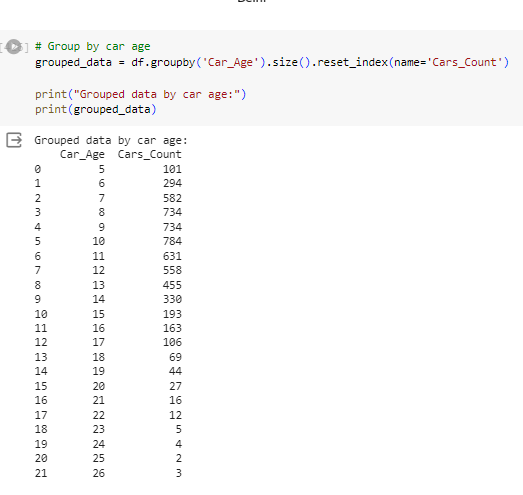
* There is a positive correlation between engine size and price. This means that as the engine size increases, the price of the engine also tends to increase. However, the data points are scattered, which means that there is not a perfect linear relationship between the two variables. Some large engines are relatively inexpensive and some small engines that are relatively expensive which depends on the other attributes too.
* Arranging the columns by ascending or descending order.

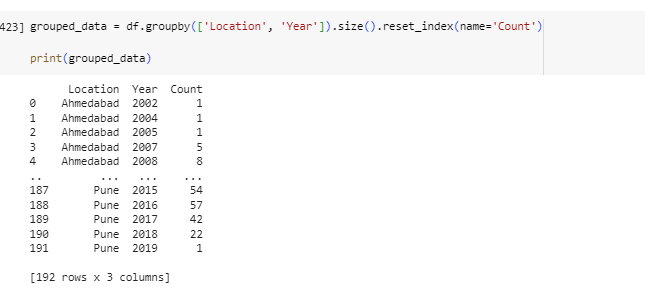




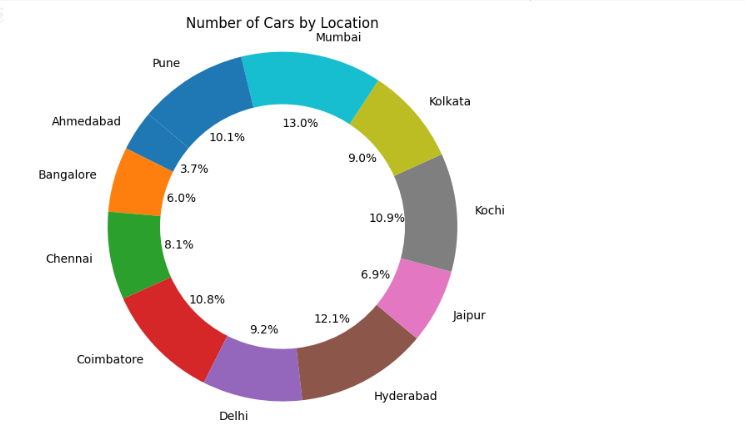
* **Group by Operations:**

****

****

****

* **Summarizing the whole dataset.**
* **Donut chart :**

****

A donut chart is similar to a pie chart in that both types of charts display categorical data as proportions of a whole. The donut chart showing the number of cars by location given in the Dataset. While it shows the number of cars in different locations. Here's a breakdown of the locations included in the chart:

Delhi makes up 9.2%.

Kolkata makes up 9.0%.

Chennai makes up 8.1%.

Hyderabad makes up 12.1%.

Bangalore makes up 6.0%.

Ahmedabad makes up 3.7%.

Pune makes up 10.1%.

Jaipur makes up 6.9%.

Kochi makes up 10.9%.

Coimbatore makes up 10.8%.

Mumbai makes up 13.0%.

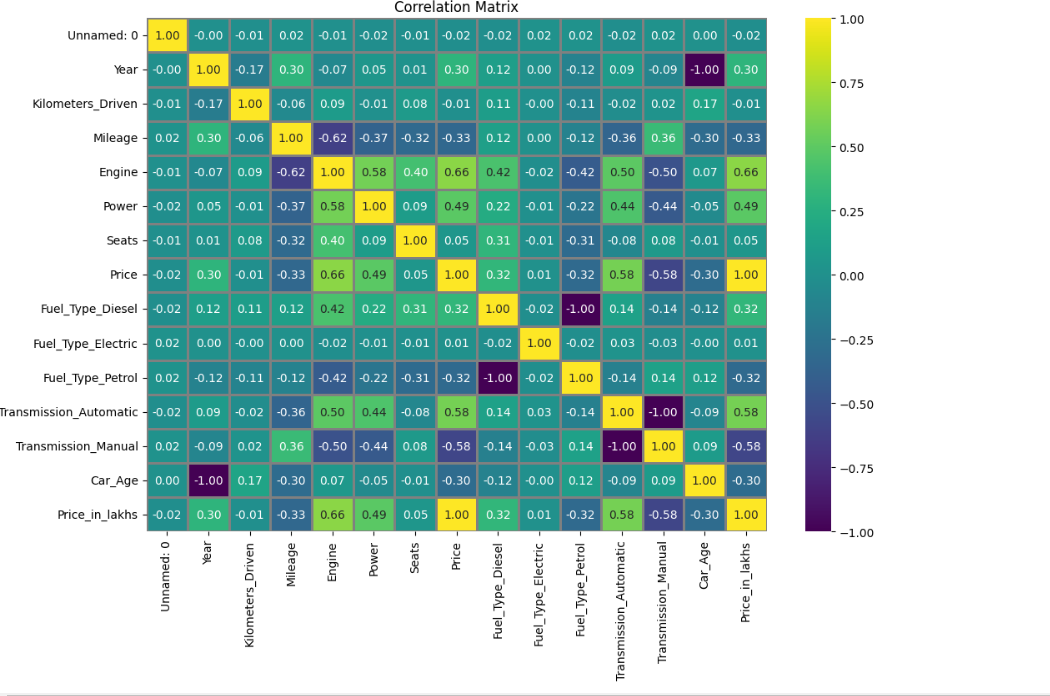
* **Correlation Matrix:**

A correlation matrix is a table that organizes correlation coefficients to show the strength and direction of the linear relationships between pairs of variables. In this specific matrix, the variables are features of the used cars, like mileage, engine size, price, and year. The correlation coefficient ranges from -1 to 1.

A value of 1 indicates a perfect positive correlation, which means as the value of one variable increases, the value of the other variable also increases proportionally.

A value of -1 indicates a perfect negative correlation, which means as the value of one variable increases, the value of the other variable decreases proportionally.

A value of 0 indicates no correlation between the variables.



**Here are some main observations about the correlation matrix in the image:**

**Year:** There is a weak negative correlation between the car's year and its price (correlation coefficient of -0.3). This means that newer cars tend to be more expensive than older cars, but there are exceptions.

**Mileage:** There is a weak negative correlation between mileage and price (correlation coefficient of -0.33). This means that cars with lower mileage tend to be more expensive than cars with higher mileage, but there are exceptions.

**Engine Size:** There is a weak positive correlation between engine size and price (correlation coefficient of 0.33). This means that cars with larger engines tend to be more expensive than cars with smaller engines, but there are exceptions.

**Transmission:** It is difficult to determine the correlation between transmission type (automatic or manual) and price because the correlation coefficients for these two features are not shown in the part of the matrix captured in the image.