**Assignment-2**

**Yogitha Mekala**

**16354649**

**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.**

A screenshot of a computer

Description automatically generated

* First, we must read the dataset given.

A screenshot of a computer code

Description automatically generated

Columns like Engine, Power, Seats, and New Price have missing values. We can utilize 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 proceeding with the preceding step.

The pre-processing methods listed below can help improve the accuracy of the data.

* A screenshot of a computer

  Description automatically generated

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

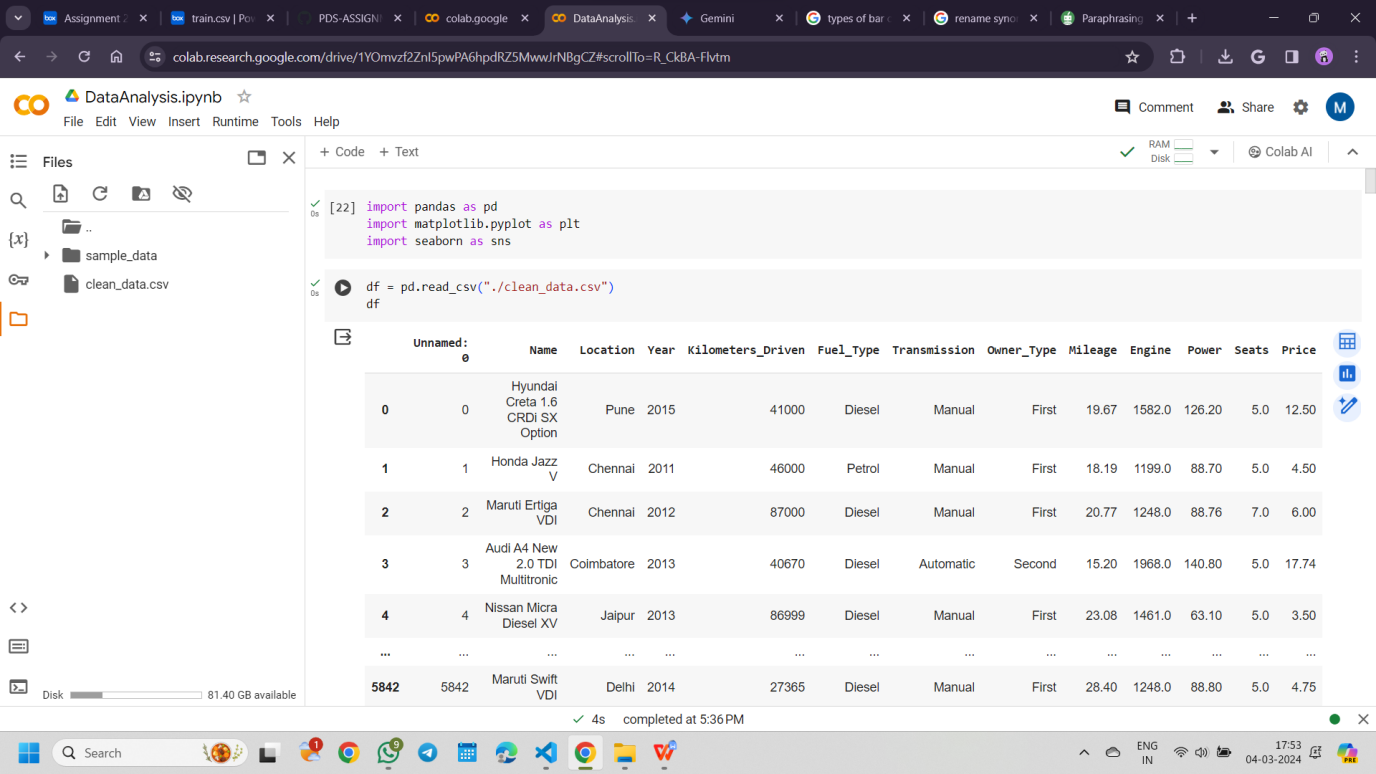
* At last the pre-processed dataset converted into the **clean\_data.csv**.

A screenshot of a computer

Description automatically generated

**Data\_Analysis**

* Read the pre-processed dataset **clean\_data.csv**.



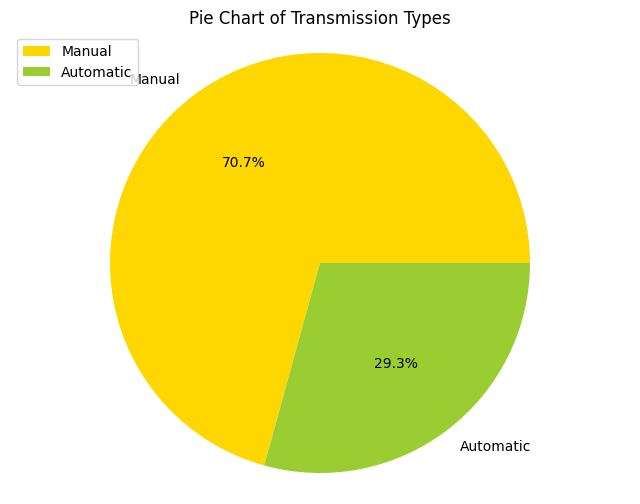
**Histogram**

A blue graph with white text

Description automatically generated

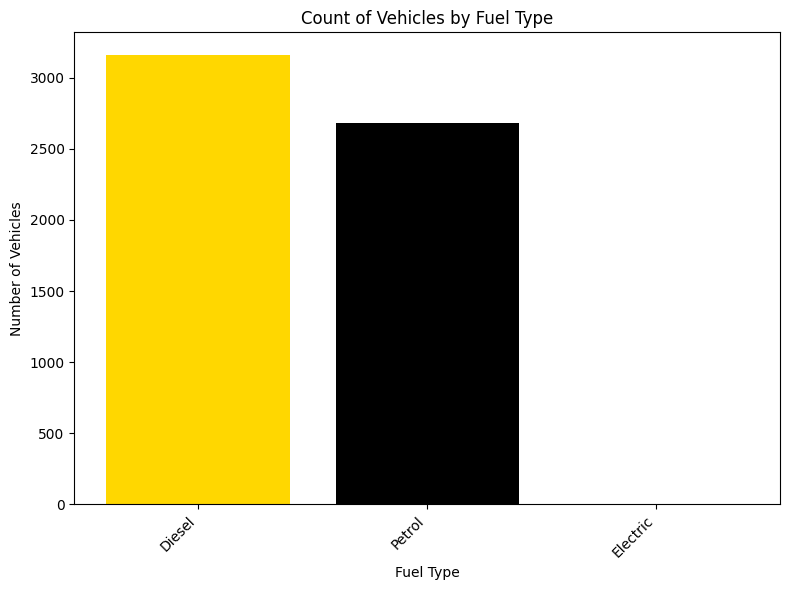
* The histogram illustrates the frequency of cars manufactured per year, offering valuable insights into the automotive industry's production trends over time.

**Pie Chart**

****

* The pie chart showing the number of cars with automatic and manual transmissions. The y-axis shows the number of cars, and the x-axis shows the two transmission types (automatic and manual).
* It appears that there are more cars with automatic transmissions than manual transmissions. The number of cars with automatic transmissions is around 1600 while the number of cars with manual transmissions is around 4100. As the data is from India, most of the cars are manual.

**Bar Chart**

****

* The percentages of the image labeled "Bar Chart of Fuel Types" are displayed in the bar chart. It displays the proportion of each of the three fuel types utilized in India: electric, diesel, and gasoline. The above chart shows us the percentages of it.

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

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

* The categorical variables Fuel\_Type and Transmission are transformed into the one hot encoded values. It is evident that every distinct category in "Fuel\_Type" and "Transmission" has been transformed into a distinct binary column, either zero or one.

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”.**

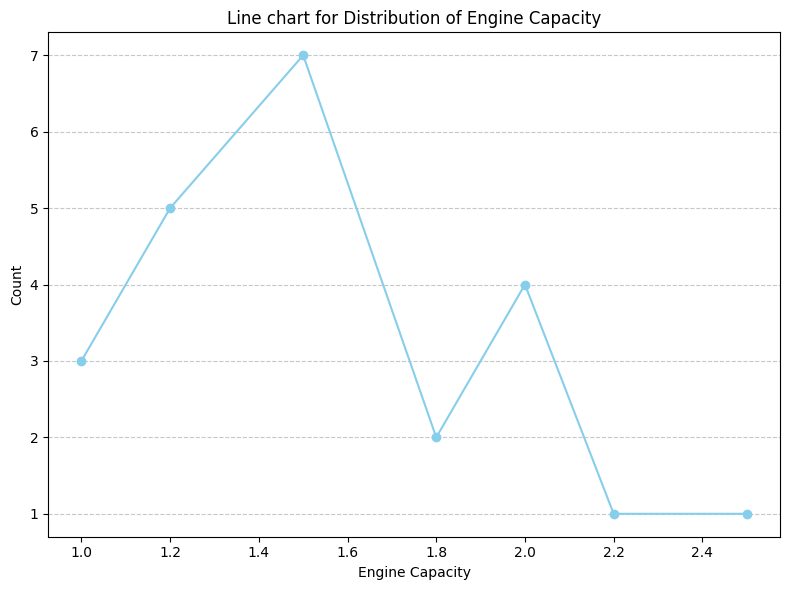
**A screenshot of a computer

Description automatically generated**

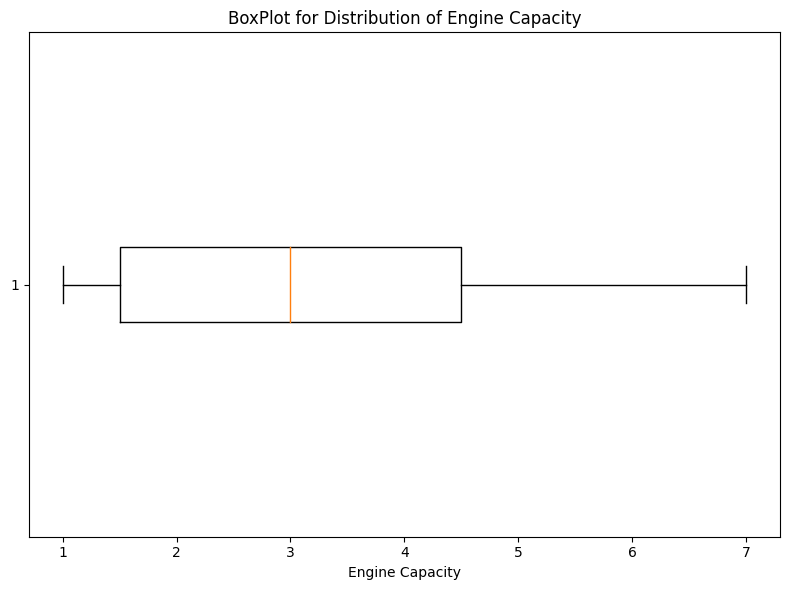
**A screenshot of a computer

Description automatically generated**

* **Line Chart**

****

* The line chart shows the Distribution of the Engine Capacity that we added a new feature to the dataset.
* **BoxPlot chart**



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.**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

* **Renaming the columns**

**A screenshot of a computer

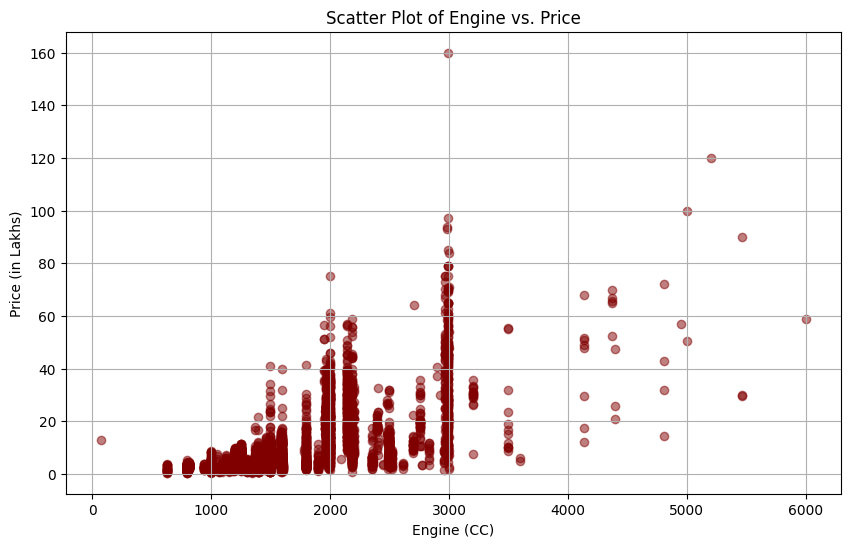
Description automatically generated**

* **Add or modifying the columns**

**A screenshot of a computer

Description automatically generated**

* **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.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* **Group by Operations**

**A screenshot of a computer

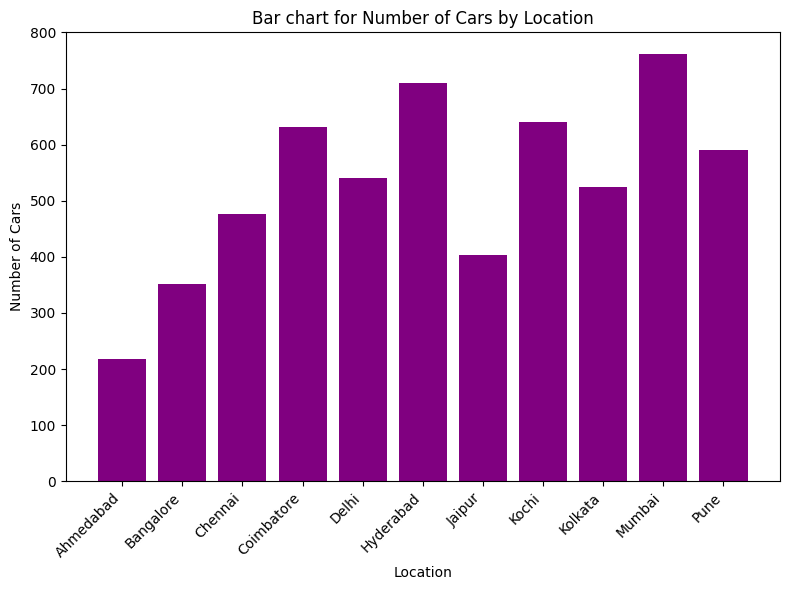
Description automatically generated**

**A screen shot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

* **Summarizing the whole dataset.**
* **Bar Chart**

****

The Bar 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 table that arranges correlation coefficients to display the strength and direction of the linear correlations between two variables is called a correlation matrix. The attributes of the used cars, such as mileage, engine size, price, and year, are the variables in this particular matrix. From -1 to 1, the correlation coefficient is the range.

A perfect positive correlation, or one in which the values of the two variables increase proportionately as one variable's value rises, is shown by a value of 1.   
A perfect negative correlation, or one in which the value of one variable increases proportionately to the value of the other, is represented by a value of -1.   
There is no association between the variables when the value is 0.

A screenshot of a graph

Description automatically generated

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

**Year:** The price of the car and its year show a weak negative link (correlation coefficient of -0.3). In other words, while there are few exceptions, newer cars are often more expensive than older ones.

**Mileage:** The price of the car and its year show a weak negative link (correlation coefficient of -0.3). In other words, while there are few exceptions, newer cars are often more expensive than older ones.

**Engine Size:** Price and engine size have a somewhat positive link (correlation coefficient of 0.33). This indicates that, although there are few exceptions, cars with larger engines are often more expensive than those with smaller engines.

**Transmission:** Given that the correlation coefficients for the two features—price and transmission type—are not displayed in the portion of the matrix that was photographed, it is challenging to ascertain the relationship between these two variables.