**Assignment-2**

**Abhinav Reddy Ayyadapu**

**16354733**

**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 need to read the dataset.

A screenshot of a computer code

Description automatically generated

* There are missing values in the columns like Engine, power, Seats, New\_Price. To fill those NA values, we can use mean, mode and median.

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”).**

* For doing the above step, we must ensure to remove the units of those columns to get better result.
* The below are some pre-processing techniques to follow to get better 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**.

A screenshot of a computer

Description automatically generated

**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.

**Bar Chart**

A bar chart with green squares

Description automatically generated

* The bar 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.

**Pie Chart**

A chart of fuel types

Description automatically generated

* The pie chart shows the percentages of the image titled "Pie Chart of Fuel Types". It shows the percentage of three types of fuel used in India: Petrol, Diesel and Electric. As we can see the percentages of it from the above chart.

**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 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”.**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

* **Bar Chart**

A graph of a distribution of engine capacity

Description automatically generated

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

A pie chart with different colored circles

Description automatically generated

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**

A graph with blue dots

Description automatically generated

* 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.**
* **Pie Chart**

A colorful pie chart with text

Description automatically generated

The pie 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.

A screenshot of a graph

Description automatically generated

**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.