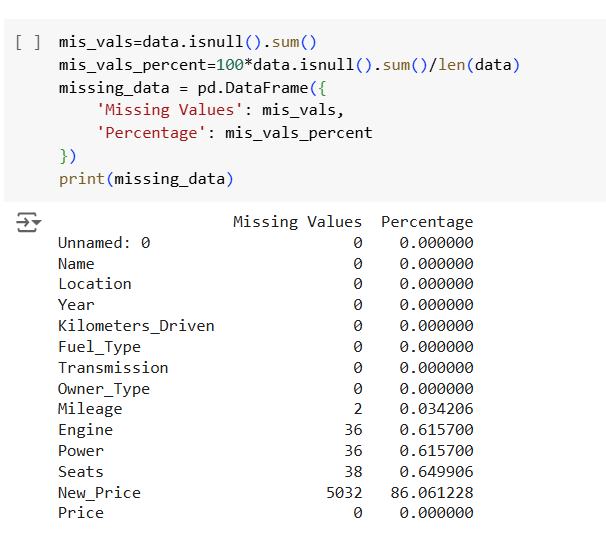
**10-31-2024**

**Principles of Data Science**

**Assignment -2**

**Achyuth Pothuganti-16355349**

**Question A: - Look for the missing values and fill it with mean, median and mode and justify your answer.**



Most columns in the dataset have very few missing values, but *Engine*, *Power*, *Seats*, and particularly *New\_Price* have notable gaps. *New\_Price* is missing in over 86% of entries, which may need special handling in the analysis.

**A screenshot of a computer program

Description automatically generated**

**Justification:**

1. **Filling Missing Values:** For numerical columns with missing values, the mean value was used as a reasonable estimate to retain overall data distribution and minimize data loss.
2. **Handling Categorical Data:** Missing values in categorical columns were filled with the mode, as it represents the most common category and maintains consistency in categorical data.
3. **Dropping New\_Prices Column:** Since over 86% of the New\_Price column values are missing; it was dropped to avoid introducing bias or inaccuracies from excessive imputation.

**Question B: Remove units from attributes and keep only numerical values**

**A screenshot of a computer program

Description automatically generated**

Here I have removed units from mileage, engine and power column and kept only numerical values.

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

**A screenshot of a computer

Description automatically generated**

The Fuel\_Type column was one-hot encoded into separate binary columns: Fuel\_Type\_Diesel, Fuel\_Type\_Electric, and Fuel\_Type\_Petrol. Similarly, the Transmission column was split into Transmission\_Automatic and Transmission\_Manual. Which prepares data for machine learning models.

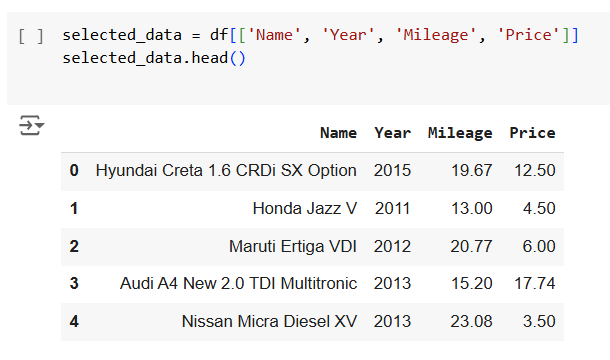
**Question D: Create one or more feature and add the column into the dataset.**

I created a new column called “car\_age” to represent the age of the car, calculated by subtracting the car's manufacturing year from the current year.

A screenshot of a computer

Description automatically generated

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

* **Selecting specific columns: **
* **Filtered the data to include only cars with a manufacturing year after 2015:**

**A screenshot of a computer

Description automatically generated**

* **Renamed the columns kilometers\_driven as odometer and owner\_type as previous owner:**

**A screen shot of a computer code

Description automatically generated**

* **Arranged the data based on their prices in ascending order:**

**A screenshot of a computer

Description automatically generated**

* **Summarize:** here I have summarized the descriptive analysis of the car price,mileage and age

**A screenshot of a computer

Description automatically generated**

**1. Price:** The average price of cars is approximately 9.65 lakhs, with a high standard deviation of 11.28 lakhs, indicating wide price variation. Prices range from 0.44 lakhs to 160 lakhs, suggesting a mix of both budget and luxury cars.

**2. Mileage:** The average mileage is about 18.16 km/l, with most cars achieving between 15.26 km/l (25th percentile) and 21.10 km/l (75th percentile). The minimum recorded mileage is zero, possibly indicating missing or erroneous values.

**3. Car Age:** The average age of cars is around 10.5 years, with a range from 5 to 26 years, indicating the dataset contains both relatively new and older cars.

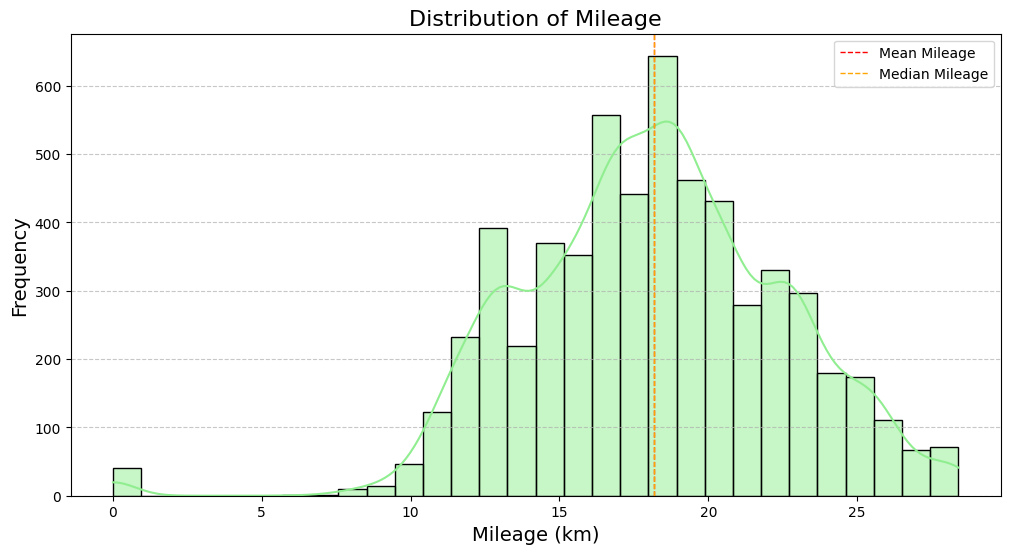
**Here are the some of the visualizations:**

1. **Distribution of car prices:**

A graph of a distribution of car prices

Description automatically generatedThis histogram shows the distribution of car prices, with most cars clustered at lower prices. The red dashed line marks the mean price, while the orange dashed line shows the median. The slight right-skew suggests a few high-priced cars are pulling the mean higher than the median.

1. **Distribution of mileage:**



This histogram illustrates the distribution of car mileage. Most cars are concentrated around moderate mileage values, shown by the green bars and the KDE curve overlay. The red dashed line marks the mean mileage, while the orange dashed line shows the median. The close proximity of the mean and median suggests a relatively balanced distribution with minimal skew.