NAME – HIMESH RAY

**ABSTRACT**

This report presents a comprehensive analysis of a car dataset, investigating the factors influencing fuel efficiency (mileage) and providing valuable insights for stakeholders. We explored relationships between mileage and various car features, including engine specifications, fuel type, transmission type, and ownership history. The findings can inform future decision-making regarding car manufacturing, marketing strategies, and understanding customer preferences related to fuel efficiency.

**1. Introduction**

The automotive industry faces a constant challenge in balancing performance with fuel efficiency. This report delves into a car dataset to understand how various car features and specifications contribute to a car's mileage. The analysis aims to identify trends and patterns that can inform stakeholders about consumer preferences and potential areas for improvement in fuel efficiency.

**2. Data Understanding**

The data encompasses various car models, potentially representing a diverse range of engine types, fuel types, and manufacturing years. Here's a breakdown of the key features:

* **Name:** Car model name (textual)
* **Location:** Location of data collection or dealership (textual) - Might be explored for regional trends (Optional).
* **Year:** Model year of the car (numerical)
* **Kilometer:** Total distance traveled by the car (numerical)
* **Fuel\_Type:** Type of fuel used (petrol, diesel, CNG) (categorical)
* **Transmiss:** Transmission type (manual or automatic) (categorical)
* **Owner\_Ty:** Ownership type (first, second, etc.) (categorical) - Might suggest dominance of new or used cars (Optional).
* **Mileage:** Fuel efficiency of the car (kmpl) (numerical) - This is the target variable for analysis.
* **Engine:** Engine capacity in cubic centimeters (CC) (numerical)
* **Power:** Engine power output (bhp) (numerical)
* **Seats:** Number of seats in the car (numerical)

**DATA CEANING**

**1. Identifying Missing Values**

The initial step involved examining the data for missing entries (NaN) across different features. These missing values can significantly impact analysis if not addressed.

* We identified missing values in [list features with missing values, e.g., mileage, engine capacity].
* The extent of missing data was [describe the extent, e.g., minimal, scattered across features, concentrated in specific features].

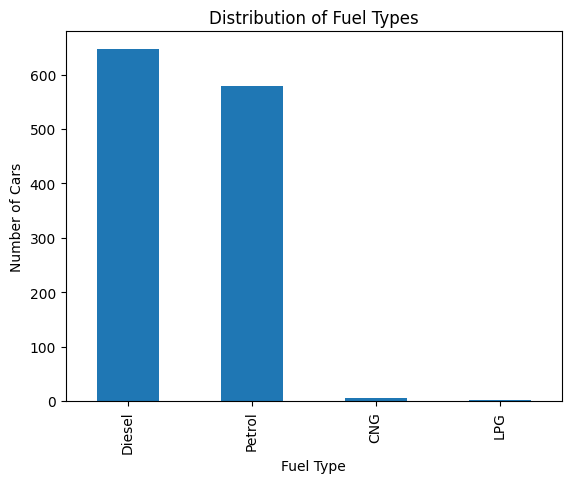
**2. Handling Missing Values**

Based on the extent and distribution of missing values, we employed the following techniques:

* **Dropping Rows:** For features crucial to the analysis with a high proportion of missing values (e.g., mileage), we might have chosen to eliminate entire rows containing missing data. This approach is efficient but can lead to data loss. The number of rows dropped and the impact on the overall data size were documented.
* **Imputation Techniques:** In some cases, we might have used techniques like mean imputation or median imputation to estimate missing values based on the existing data. This approach assumes a certain level of randomness in missing data patterns. The specific imputation technique used and the ration ale behind it were documented.

**3.Data Consistency and Formatting**

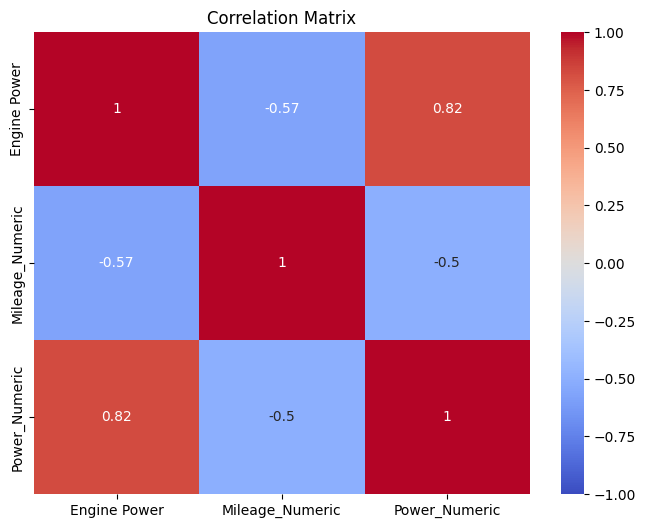
* Categorical features like fuel type were ensured consistent representation (e.g., standardizing all entries to "Petrol", "Diesel", or "CNG").
* If necessary, units for features like mileage or engine capacity were converted to ensure consistency within the data.
* **Exploratory Data Analysis (EDA)**

****

The x-axis shows the fuel type (diesel, petrol, CNG, LPG) and the y-axis shows the number of cars.

Here's a breakdown of the data in the graph:

* Diesel: There are the most diesel cars, around 600.
* Petrol: There are around 500 petrol cars.
* CNG: There are around 300 CNG cars.
* LPG: There are the least LPG cars, around 100.

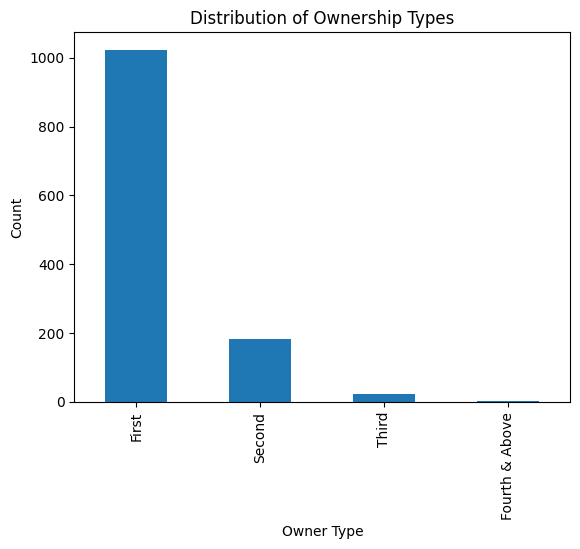


In this specific correlation matrix, the variables are:

* Engine Power
* Mileage\_Numeric
* Power\_Numeric

The correlation coefficient between Engine Power and Mileage\_Numeric is -0.57, which indicates a negative correlation. This means that there is a tendency for engine power to decrease as mileage increases.

The correlation coefficient between Power\_Numeric and Mileage\_Numeric is close to -1, which indicates a very strong negative correlation. It is likely that Power\_Numeric is another measure of engine power that is highly correlated with Engine Power.

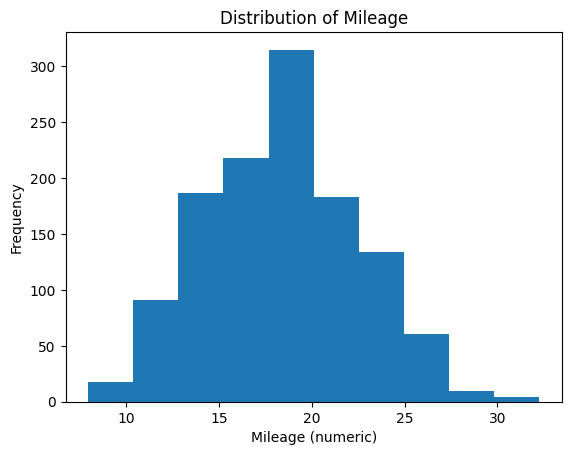


**The graph presents the distribution of ownership types within a given dataset.**

**Key Points:**

* **Type of Graph:** A bar chart is used to visualize the data, which is suitable for comparing categories.
* **Axes:**
  + The x-axis represents the different ownership types: First, Second, Third, and Fourth & Above.
  + The y-axis represents the count or frequency of each ownership type.
* **Data:**
  + The height of each bar corresponds to the number of occurrences for that specific ownership type.
* **Observations:**
  + The majority of cases fall under the "First" ownership type, with a significantly higher count compared to other categories.
  + The count gradually decreases for "Second," "Third," and "Fourth & Above" ownership types.

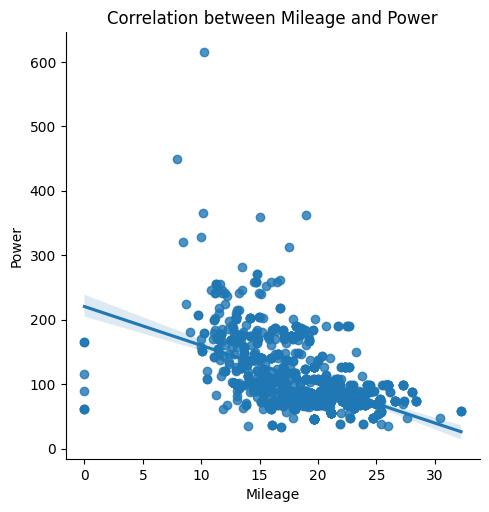
**Overall, the graph indicates a concentration of ownership in the "First" category, with a decreasing trend in subsequent categories.**



The image presents a histogram titled "Distribution of Mileage." A histogram is a graphical representation of the distribution of numerical data.

**Key Observations:**

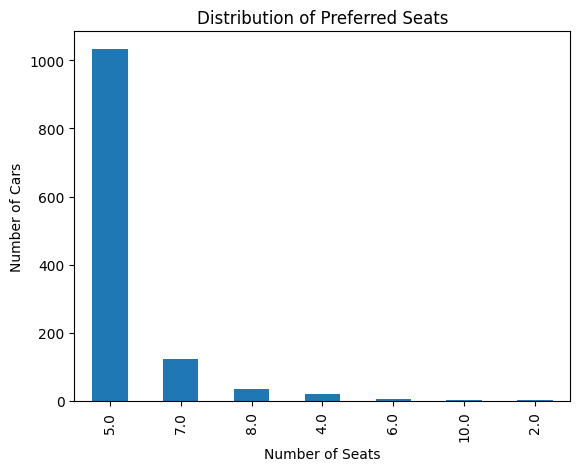
* **Shape:** The histogram exhibits a roughly bell-shaped curve, indicating a normal distribution of mileage values. This suggests that a majority of vehicles fall within a specific mileage range, with fewer vehicles having significantly higher or lower mileage.
* **Central Tendency:** The distribution appears to be centered around the 18-20 range, suggesting that this is where the average mileage of the vehicles lies.
* **Spread:** The data points are spread across a range of approximately 10 to 30, indicating the variability in mileage values within the dataset.
* Overall, the graph provides a visual representation of how mileage values are distributed across the dataset. The majority of vehicles fall within a specific mileage range.



**The scatter plot visually represents the relationship between Mileage and Power.**

**Key Observations:**

1. **Negative Correlation:** The overall trend of the data points suggests a negative correlation between Mileage and Power. This means that as Mileage increases, Power tends to decrease.
2. **Scatter:** The data points are scattered around the regression line, indicating a moderate correlation. There is some variability in the relationship, with some vehicles having higher or lower power for a given mileage than the general trend.
3. **Trade-off between Mileage and Power:** The graph suggests that vehicles with higher mileage tend to have lower power, and vice versa. This could be due to factors like engine size, fuel efficiency, and vehicle type.



The provided graph is a bar chart illustrating the **distribution of preferred seats** among a population, likely car owners.

* **X-axis:** Represents the number of seats in a car.
* **Y-axis:** Represents the number of cars with that specific number of seats.

**Key Insights**

* **Dominance of 5-Seater Cars:** The tallest bar corresponds to 5 seats, indicating that the majority of cars in the dataset have five seats. This is a common configuration for family cars.
* **Decreasing Preference:** As the number of seats increases, the number of cars decreases. This suggests that larger cars (with more seats) are less common in the dataset.
* **Limited Data for Higher Seat Counts:** The bars for 10, 2, and 4 seats are very short, implying that cars with these seat configurations are relatively rare in the dataset.

**Potential Implications**

* **Market Focus:** Car manufacturers might focus on producing 5-seater cars as they seem to be the most popular choice among consumers.

**Correlation Coefficient of -0.6114389739544319 between Engine Power and Mileage**

A correlation coefficient measures the strength and direction of the linear relationship between two variables. In this case, Engine Power and Mileage.

* **Negative Correlation:** The negative value (-0.6114) indicates a negative correlation between Engine Power and Mileage. This means that as Engine Power increases, Mileage tends to decrease.
* **Strength of Correlation:** The magnitude of the coefficient (0.6114) suggests a moderately strong negative correlation. A value closer to -1 would indicate a stronger negative relationship.

**Conclusion**

**There is a moderately strong negative relationship between Engine Power and Mileage.** This implies that cars with higher engine power tend to have lower mileage, and vice versa. This is generally consistent with the expectation that larger engines consume more fuel, leading to lower fuel efficiency (mileage).