IT 462
Exploratory Data Analysis

VEHICLE SALES

Presented by Group - 4

START

Topic

Price Prediction of Used cars

Introduction

Buying and selling used cars can be challenging due to determining fair prices and identifying fuel-efficient vehicles based on factors like age and mileage.

As a part of the EDA project we have decided to perform the analysis to solve the above problem of finding the best price for a used car to based on several factors like KM-Driven, age of the car, type of fuel etc.,

This project analyzes used car sales data to uncover hidden patterns and trends that influence pricing and other key aspects. By examining factors like mileage and pricing, it aims to provide insights that help understand the dynamics of the used car market and assist in decision-making

Problem Statement

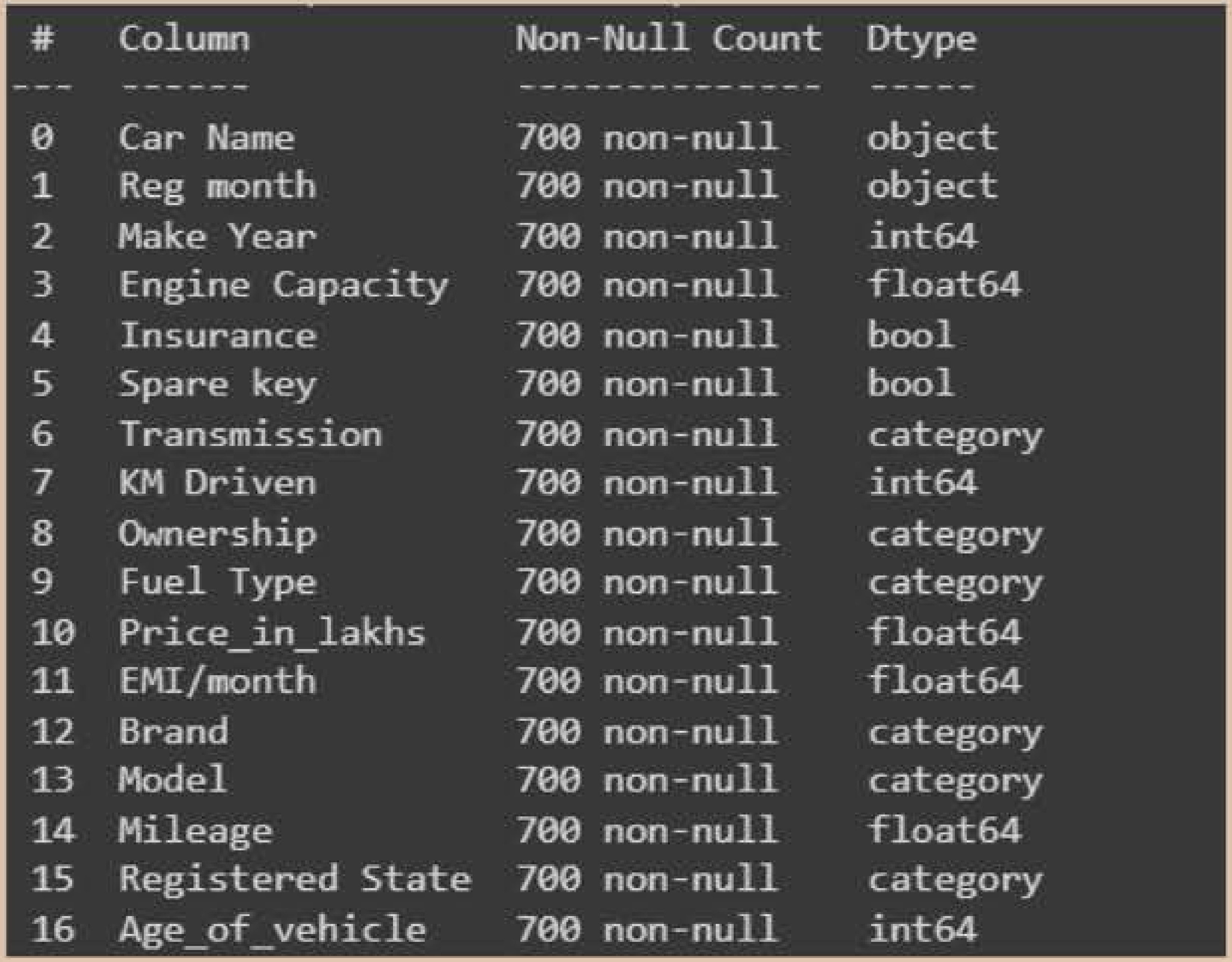
Objective : The main objective is to develop a predictive model in orderto predict the selling price of used cars for customers using different car attributes, such as engine capacity, mileage, brand of the car, model of the car, and make year. This model should allow both the dealerships and independent sellers to place competitive prices that meet the conditions of the market and the characteristics of the vehicles

We have performed regression analysis to analyze the data and predict accurate prices based on many dependable variables.

Data Collection

- Data Collection Methodology: Utilized the Beautiful-Soup library toscrape data from Cars24.com website, a prominent website for buying and selling used cars across India
- Data Source and Tool: Data was obtained through web scrapingtechniques using Beautiful-Soup, effectively handling HTML and XMLdocuments to gather comprehensive data on used cars from 2010 to2023.
- Data Attributes: Extracted detailed information on used cars,including name, model, manufacture year, price, mileage, enginecapacity, and kilometers driven, resulting in a dataset of 700 entriesacross 16 different columns.

Data Description



```
#   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Car Name           700 non-null    object
 1   Reg month          700 non-null    object
 2   Make Year          700 non-null    int64
 3   Engine Capacity    700 non-null    float64
 4   Insurance          700 non-null    bool
 5   Spare key          700 non-null    bool
 6   Transmission       700 non-null    category
 7   KM Driven          700 non-null    int64
 8   Ownership          700 non-null    category
 9   Fuel Type          700 non-null    category
 10  Price_in_lakhs     700 non-null    float64
 11  EMI/month          700 non-null    float64
 12  Brand              700 non-null    category
 13  Model              700 non-null    category
 14  Mileage            700 non-null    float64
 15  Registered State   700 non-null    category
 16  Age_of_vehicle     700 non-null    int64
```

Data Cleaning

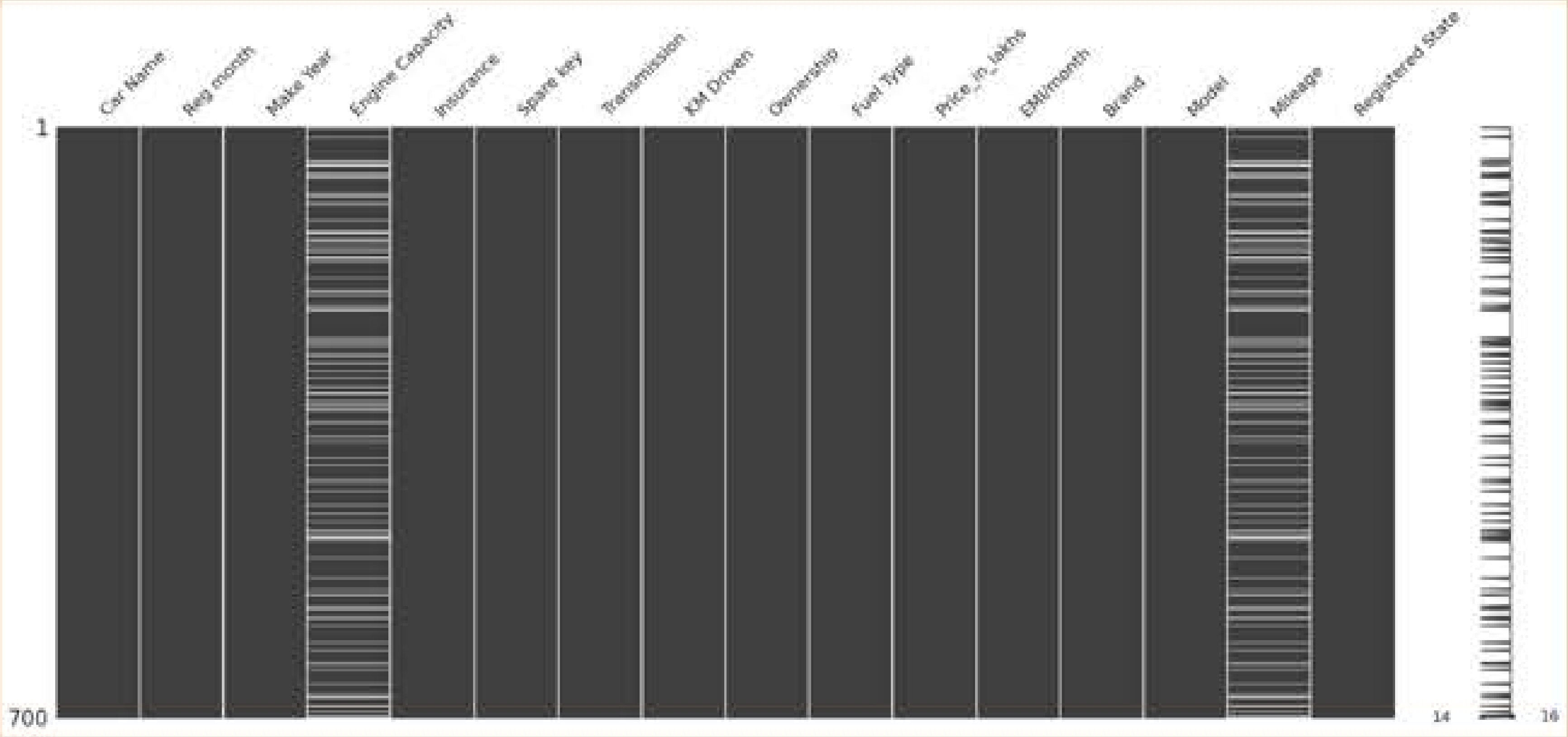
1. Data Formatting and Cleanup:

- Standardized categorical data by converting features like Insurance and Spare Key from 'Yes/No' to booleanvalues (True/False).
- Transformed numerical data by removing comma separators from the 'EMI/month' field, converting it into a float format for accurate numerical analysis

2. Outcome of Data Cleaning:

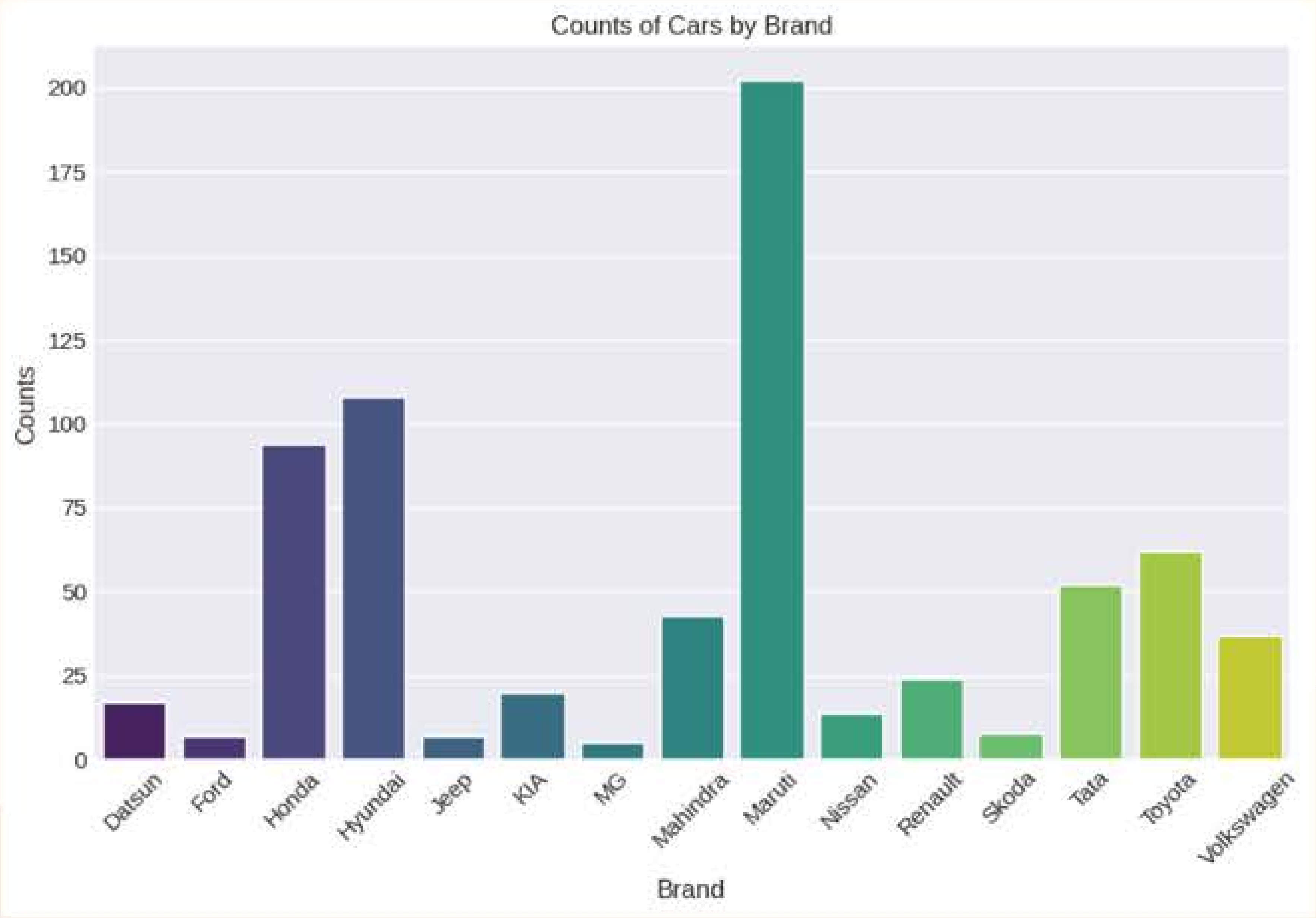
- These steps ensured the data was well-prepared for accurate and reliable exploratory analysis, visualization,and modeling, enhancing the overall quality and usability of the dataset.

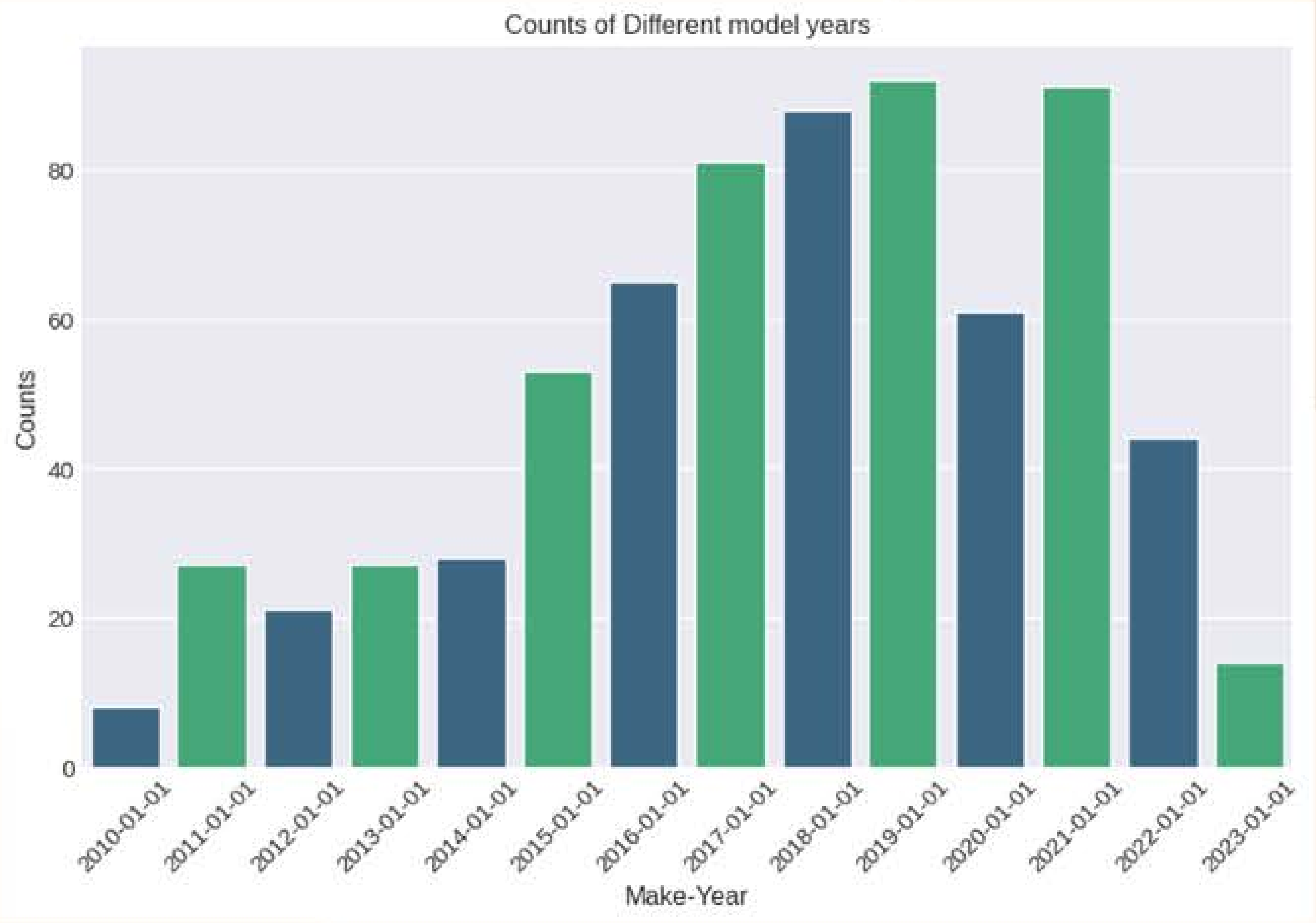
Missing Data



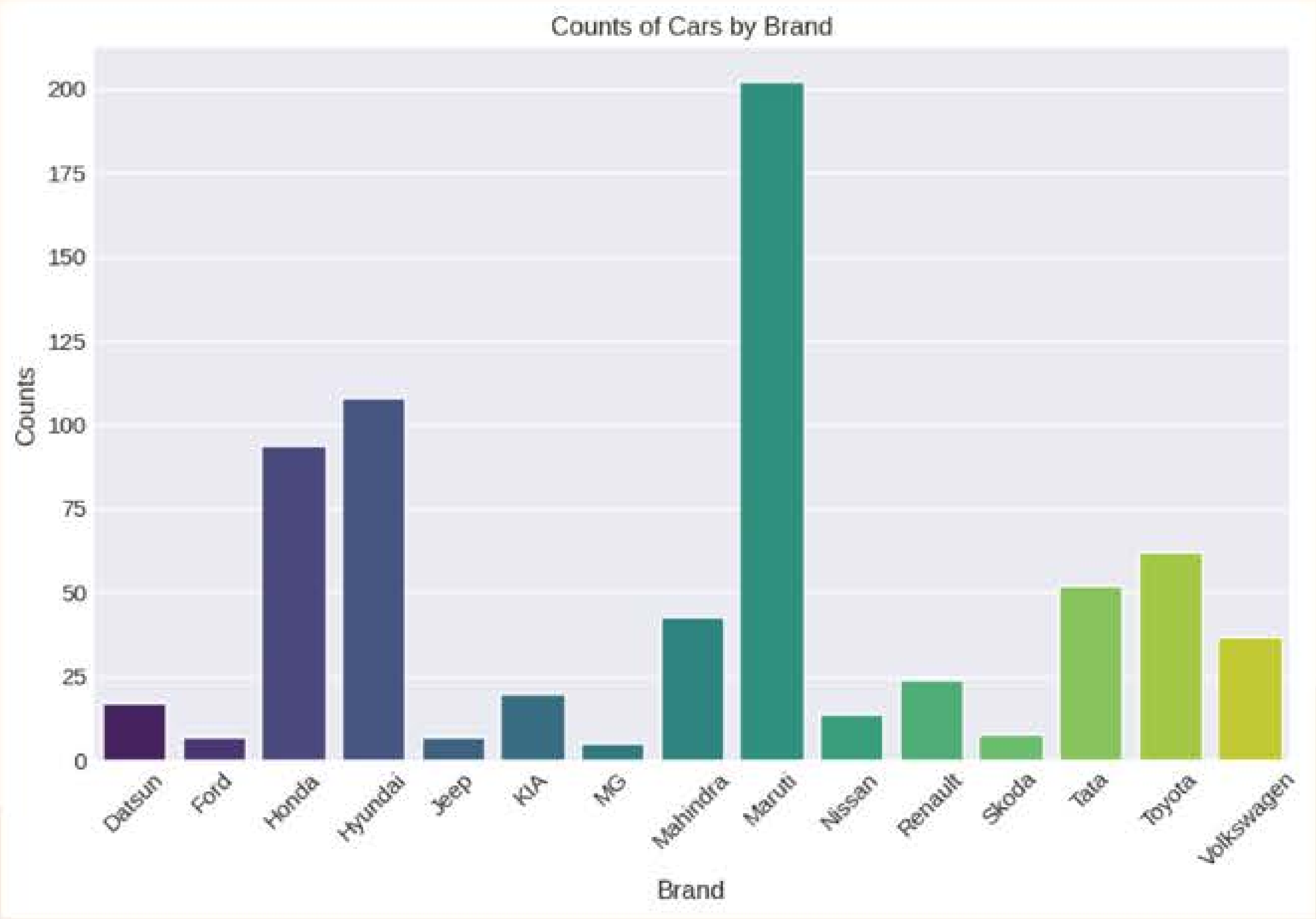
Analyzed missing data crucial for predictions such as engine capacity and mileage. Employed mean imputation for engine capacity and median imputation for mileage to handle missing values ffectively, ensuring minimal bias and maintaining data integrity.

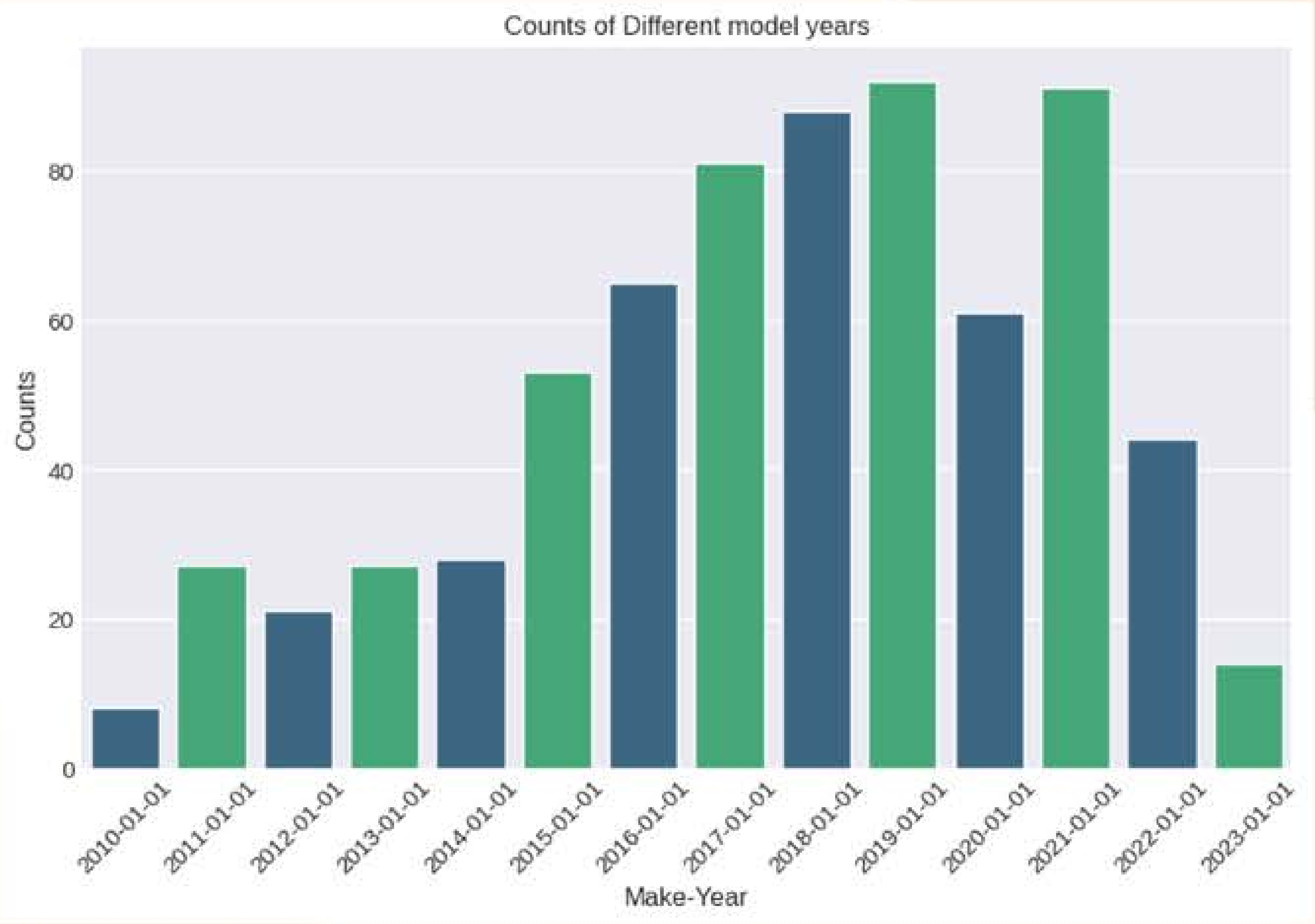
Visualization



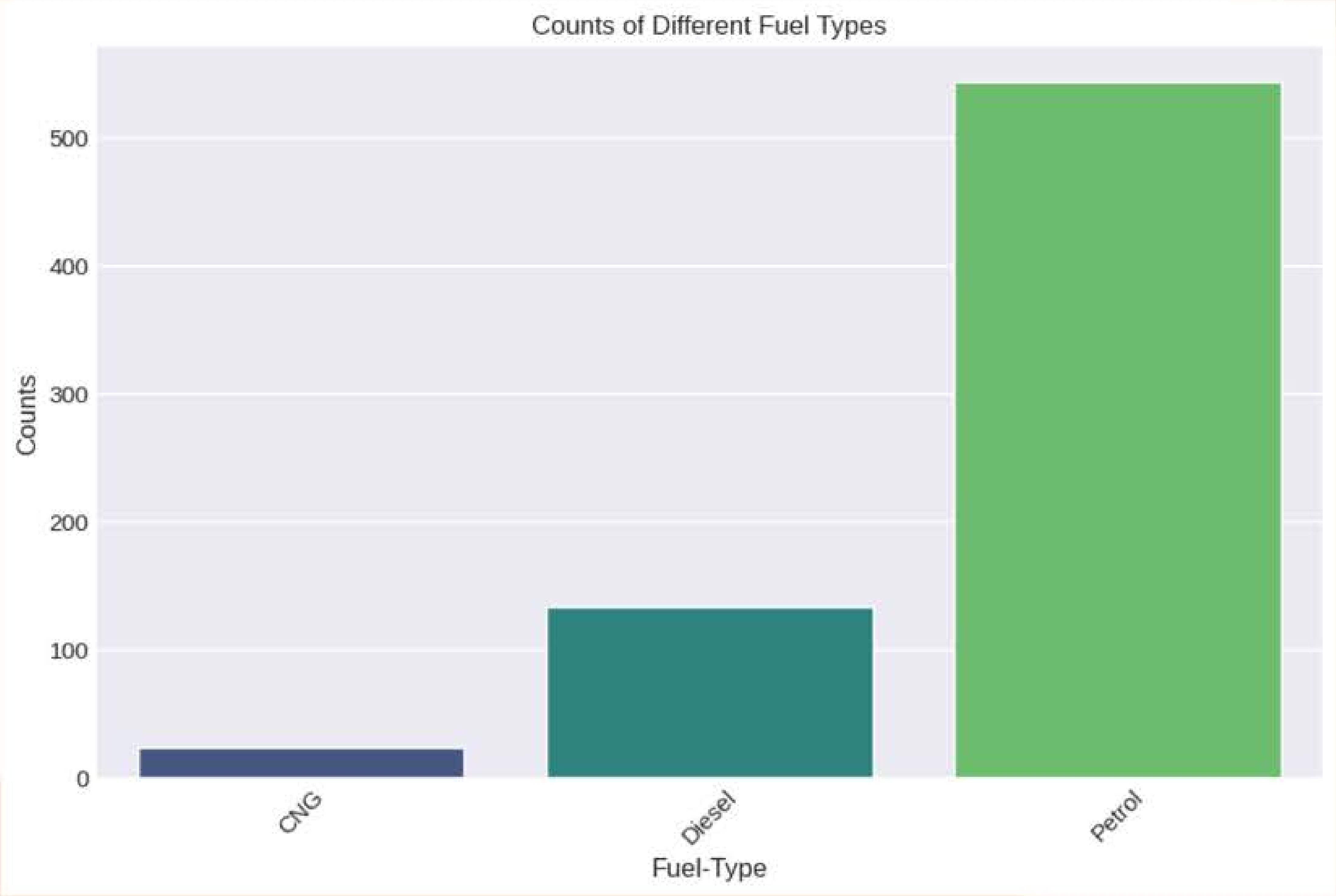


Visualization

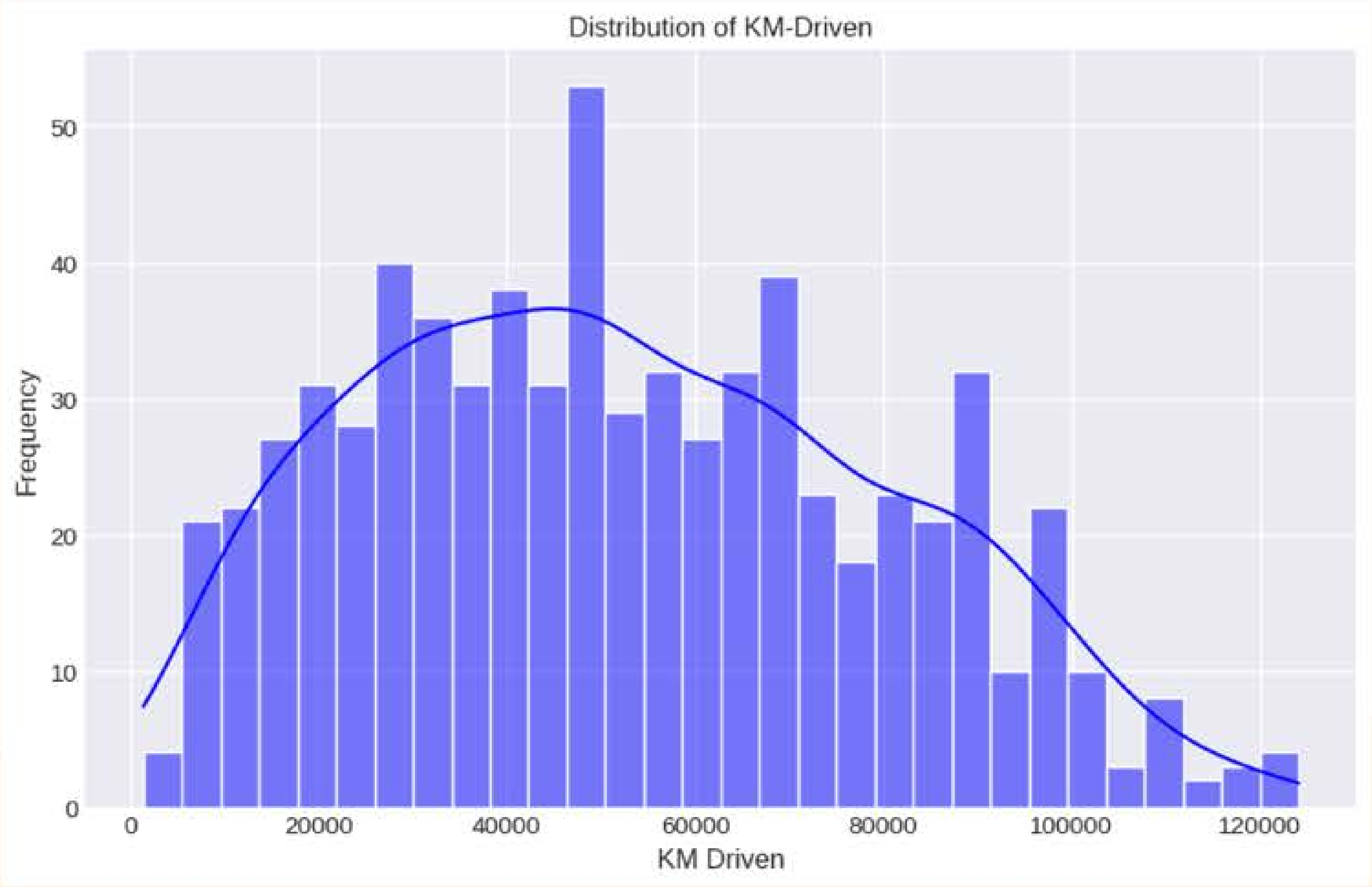




Visualization

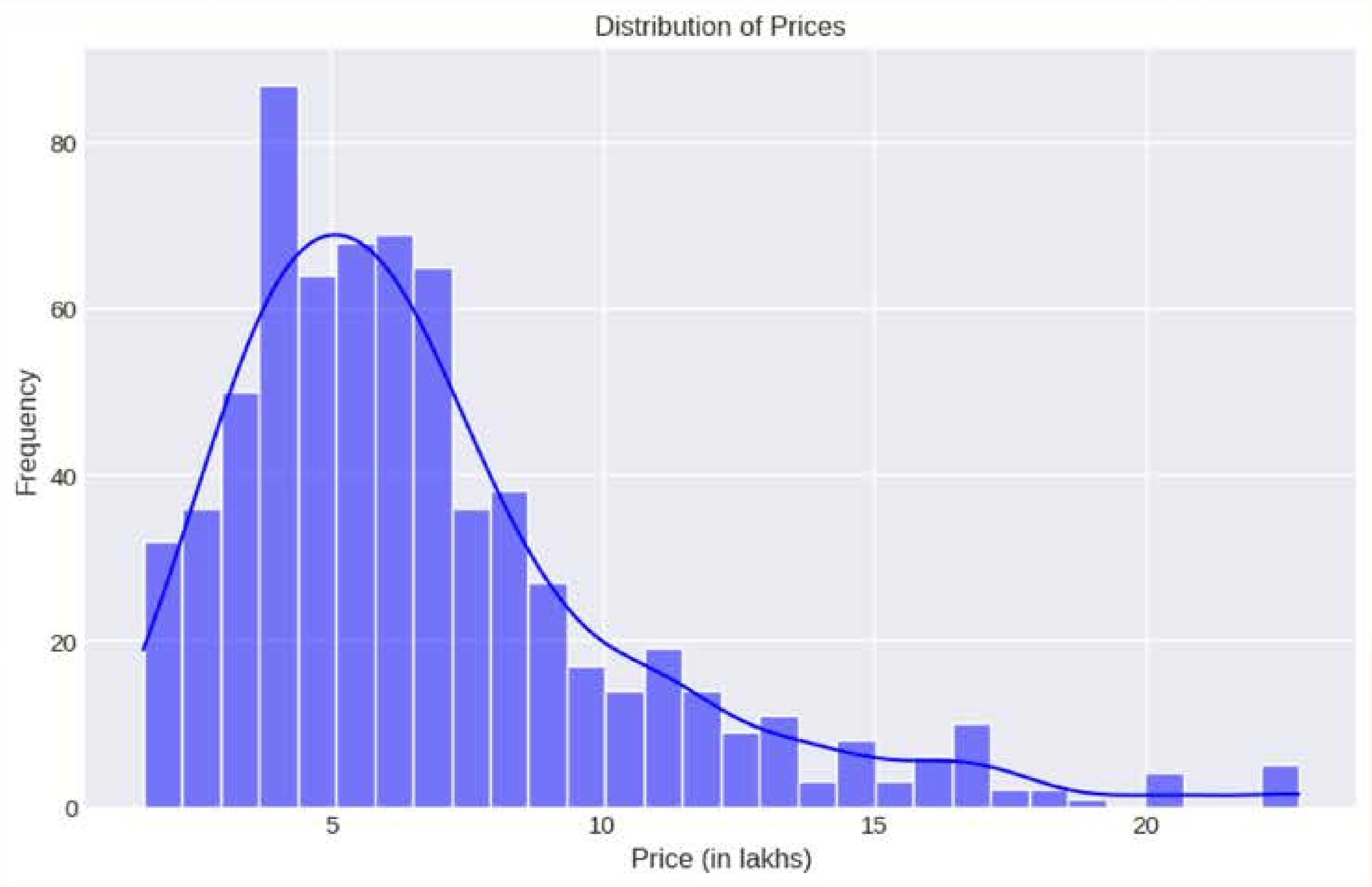


Univariate Analysis



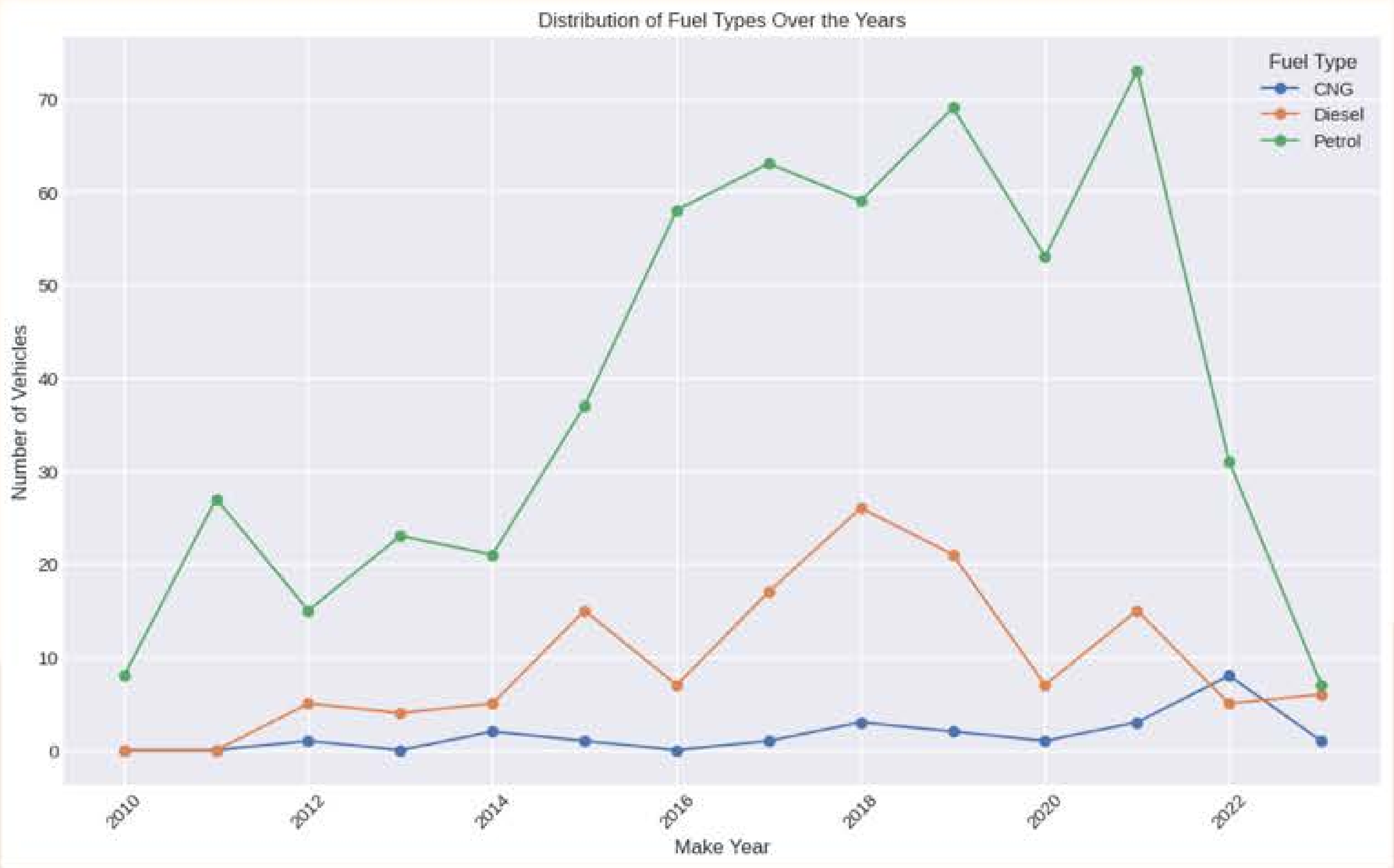
The histogram of kilometers driven shows a right-skewed distribution, indicating most used cars have distance travelled around 20,000km - 60,000km which is relatively less travelled.

Univariate Analysis



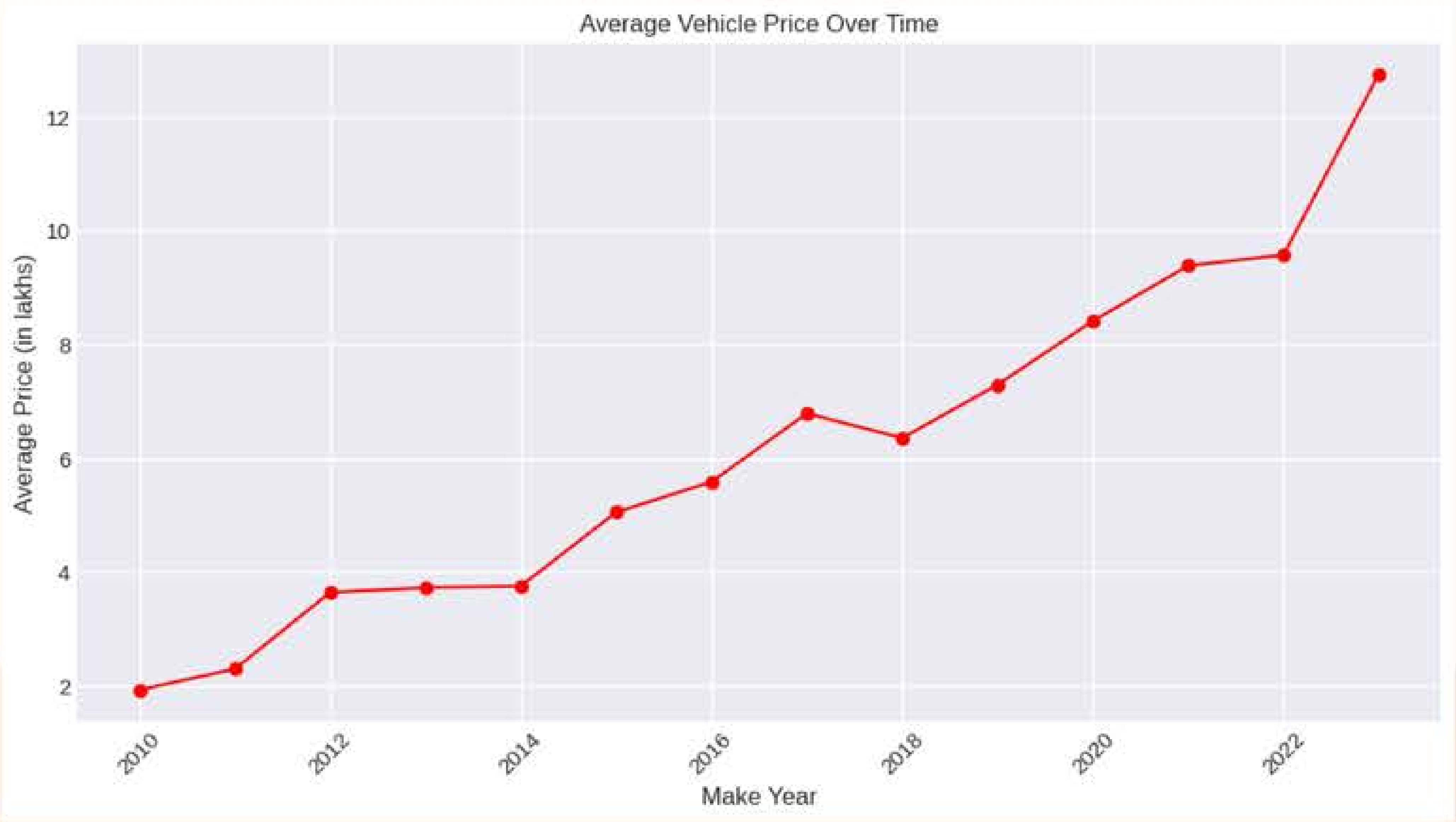
The vehicle price distribution is right-skewed,showing a concentration of cars in the lower price range with fewer high-priced outliers, suggesting a higher demand or availability of more affordable vehicles. The presence of high-cost outliers inflates the average price,indicating a lesser supply or demand for luxury vehicles in the used car market.

Univariate Analysis



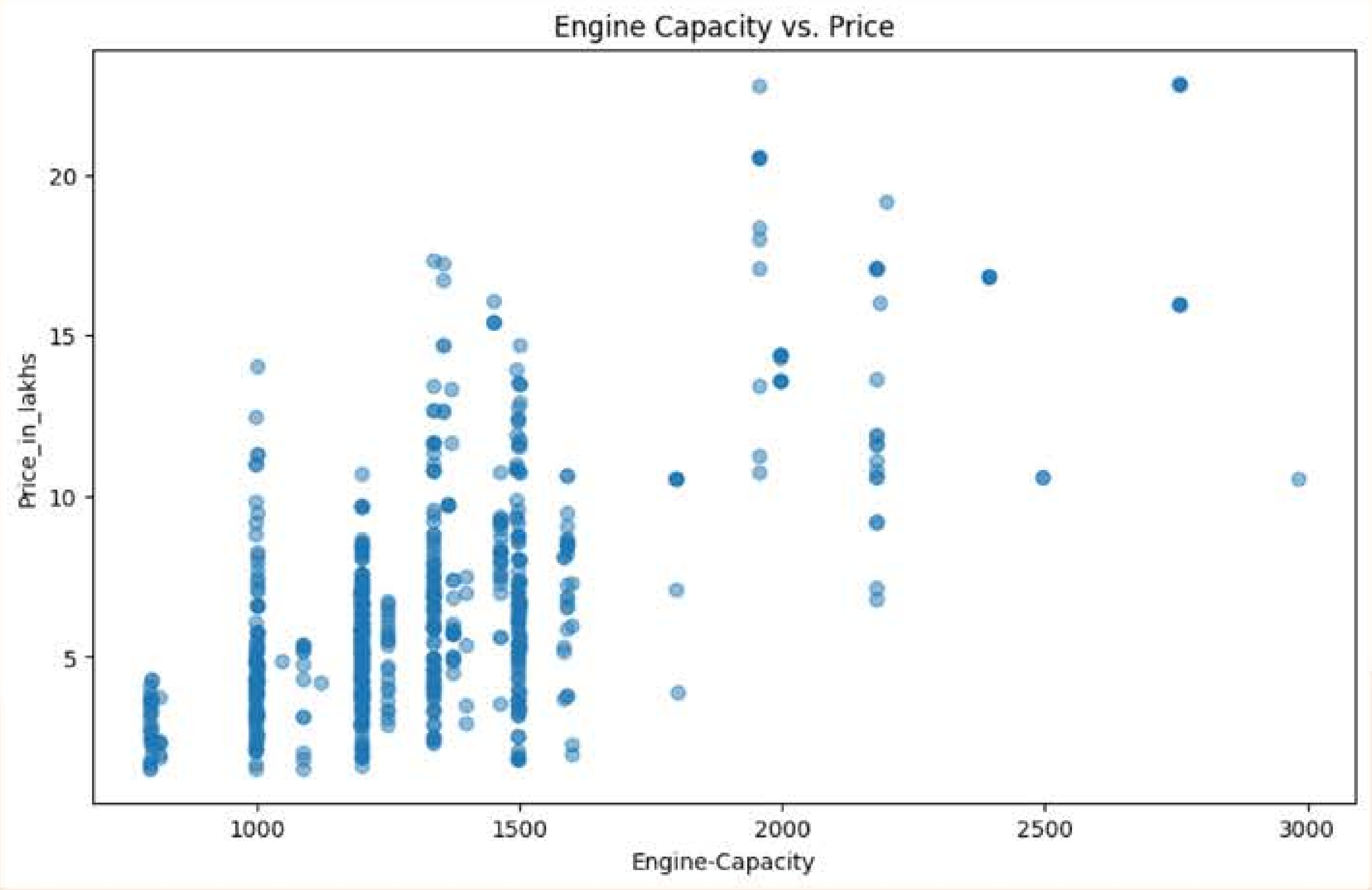
The line plot reveals an overall upward trend in car prices over the years, influenced by changes in consumer preferences, inflation, and technological advancements. Diesel cars have seen a price increase, likely due to enhanced fuel efficiency and technology, while CNG vehicles maintain consistent pricing, with newer models experiencing thehighest price hikes.

Univariate Analysis



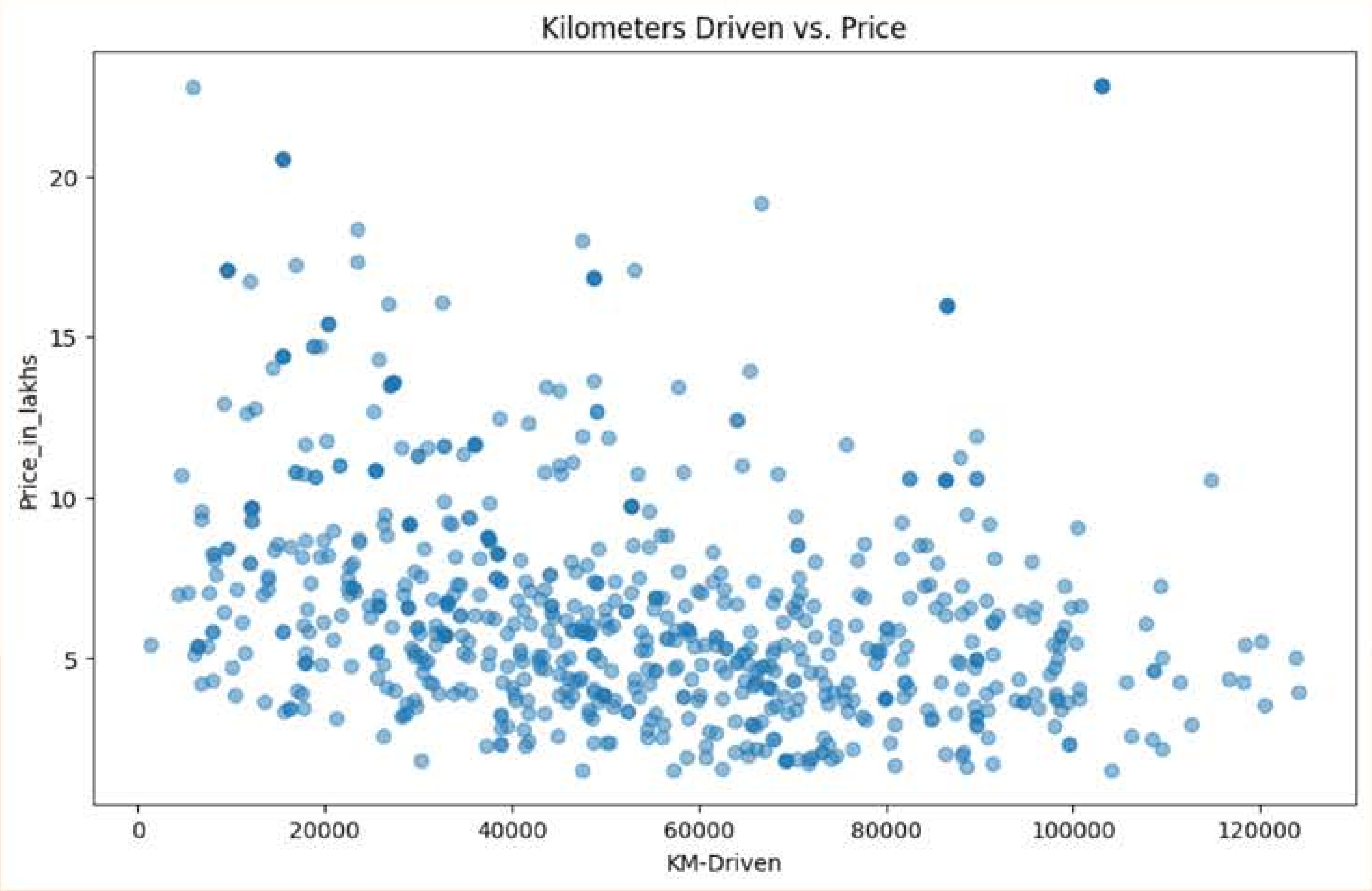
The line plot highlights a notable increase in car prices in the final year, attributed to market trends and consumer preferences, with significant variances in price distribution across different states, showing both linear increases and fluctuations over time.

Multivariate Analysis



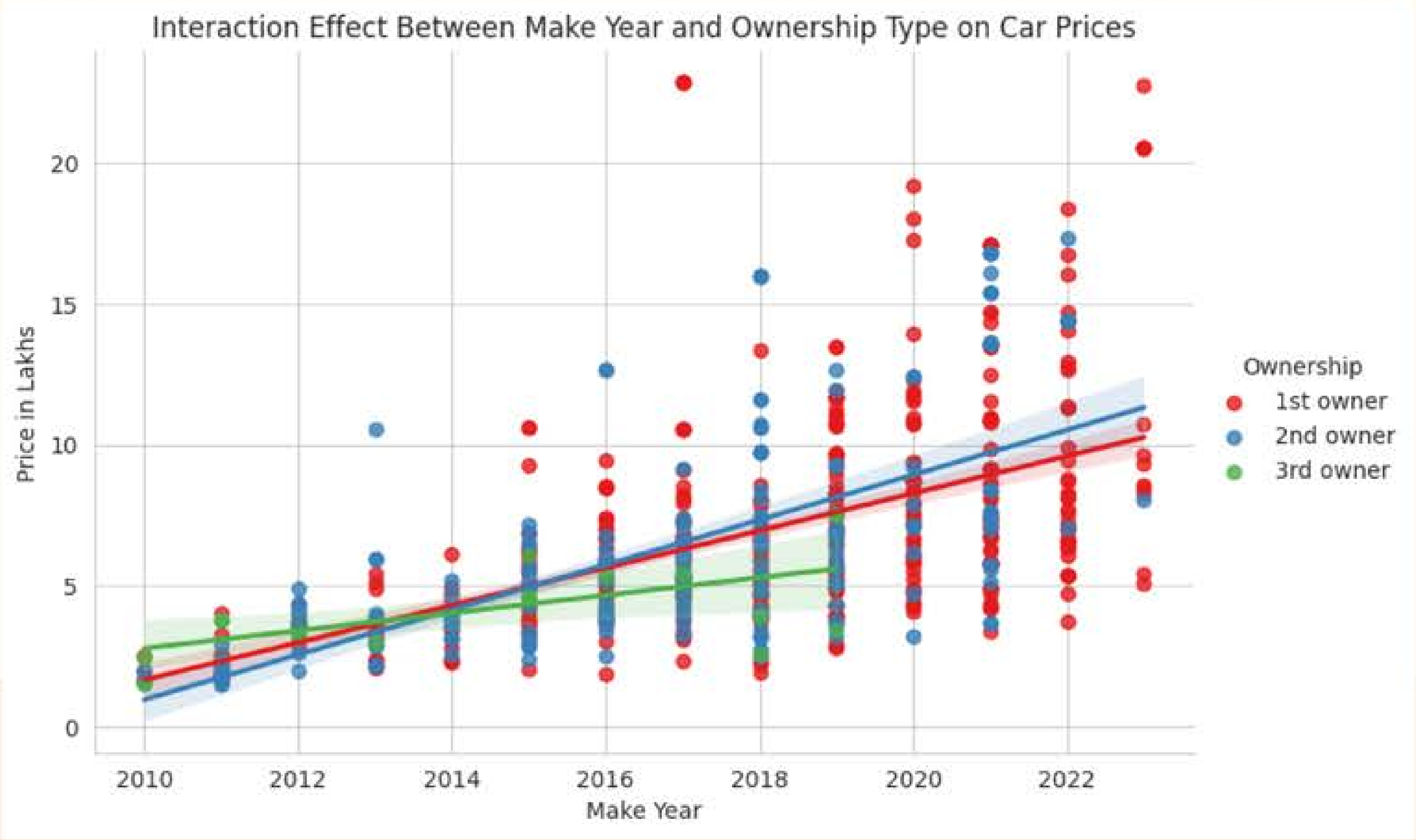
The scatter diagram reveals a general trend where car prices increase with larger engine capacities, although significant price variability across all engine sizes suggests that other factors also influence automobile costs. Brands that prefer larger engines may impact pricing due to their specific market positioning and consumer perception.

Multivariate Analysis



The scatter plot analysis shows a negativecorrelation between car prices and kilometers driven,with most data points indicating that lower km-driven corresponds to higher prices. The plot shows denser distributions for lowerkilometers and wider spreads for the larger ones

Multivariate Analysis


Interaction Effect Between Make Year and Ownership Type on Car Prices

Scatter plots demonstrate that cars with a single owner typically retain higher values and show less depreciation compared to those with multiple owners, indicating a positive correlation between fewer owners and higher resale prices. The overall average car price has remained stable from 2010 to 2014, with a significant increase observed in subsequent years.

Feature Engneering

Feature Extraction :

Enhance predictive model accuracy by introducing relevant new variables from existing data.

Key Features Extracted :

- Reg Month:
  - Extracted from the vehicle registration year. Useful in assessing seasonal effects on vehicle pricing and demand.
- Boolean Conversion :
  - Convert "Insurance" and "Spare Key" from text ('Yes'/'No') to boolean (True/False).
  - Ensures data uniformity, aiding in simpler and more efficient processing by machine learning models.

Feature Engneering

Key Features Extracted :

- Missing Value Imputation:
  - Essential for "Engine Capacity" and "Mileage" where data gaps can skew analysis.
  - Employ mean or median imputation techniques to provide complete datasets for model training.
- Age of Vehicle:
  - Calculated by subtracting the 'Make Year' from the current year.
  - Critical for modeling as older vehicles typically decrease in value.
- Km-Driven Numeric Extraction:
  - Extract pure numerical values from entries formatted as "number KM", enhancing data cleanliness for analysis.

Feature Engineering

Feature Selection:

- Objective:
  - Streamline the predictive model by reducing the number of input variables, which helps in reducing complexity and improving interpretability.

Selection Techniques :

- Variance Threshold:
  - Eliminates features that do not meet a threshold of variance, under the premise that low variance features contribute minimal information.

Feature Engneering

Selection Techniques :

- **Mutual Information Scores:**
  - Assess the mutual dependence between independent variables and the target variable ('Price in Lakhs').
  - Helps in retaining highly predictive features, improving model reliability.
- **Pearson Correlation:**
  - Identifies and removes highly correlated features to combat multicollinearity.
  - Enhances model performance by ensuring independent feature inputs.

Application of Feature selection

- By use of Feature Selcection method like Variance Threshold, Mutual -Information,Pearson Coe-relation through which we choosen top K features,in our case we have taken top 7 features and can be applied in Feature transformation technique inorder to transform feature distribution to Gaussian Distribution for better model fitting.
- Can be used remove irrelevant feature to decrease model complexity.

Data Transformation

- Goal:
  - ○ Modify data distribution to approximate normality, which is often a requirement in many statistical modeling techniques.

- <u>Transformation Techniques:</u>

1. Logarithmic Transformation

2. Box-Cox Transformation

Transformation Techniques

- Logarithmic Transformation:
    - Ideal for data with exponential growth or heavy right-skewed distributions.
    - Applied to features like 'KM Driven' and 'Mileage' to stabilize variance and achieve symmetry in data distribution.

- Box-Cox Transformation:
    - A parametric method that efficiently finds a transformation lambda to normalize data.
    - Applied specifically to 'Engine Capacity', demonstrating flexibility in handling varying data distributions.

Model Fitting

Linear Regression

- Linear regression is a statistical method used to model the relationship between one or more independent variables (predictors) and a dependent variable (target). It assumes a linear relationship between the predictors and the target variable.
- The simple linear regression model can be represented by the following equation:
  - $Y=\beta 0+\beta 1X+\epsilon$Y=β0+β1X+ϵ

Model Fitting

Linear Regression

Using Feature Selection:

- Now Using Previously used Data Transformation and feature selection, now we have taken top features and tried to apply model Fitting.
- Utilized mutual information scores and correlation for selecting impactful features like 'BoxCox Engine Capacity', 'Log KM Driven','Log_Mileage', 'Insurance', 'Spare key', 'Ownership','MakeYear','Transmission_Automatic' 'Transmission_Manual' and 'Price_in_lakhs' as target variable.

Model Fitting

Linear Regression

Using Without Feature Selection:

- Based on practical intuition, included features beyond basic model requirements for comprehensive analysis.
- Includes features like ‘Engine Capacity’,’Insurance’,’Spare-key’, ‘Transmission’,’KM Driven’, ‘Ownership’,’Fuel’, ‘EMI/month’,’Mileage’ etc and ‘price_in_lakh’ as Target Variable.

Model Fitting

RandomForest Regressor

- Model Description:
  - Ensemble learning method combining multiple decision trees to output the mean prediction, enhancing accuracy and robustness.
- Encoding Techniques:
  - Categorical variables (e.g., Transmission, Fuel Type) converted to binary formats using one-hot encoding to facilitate model input and improve training efficiency.

Model Fitting

RandomForest Regressor

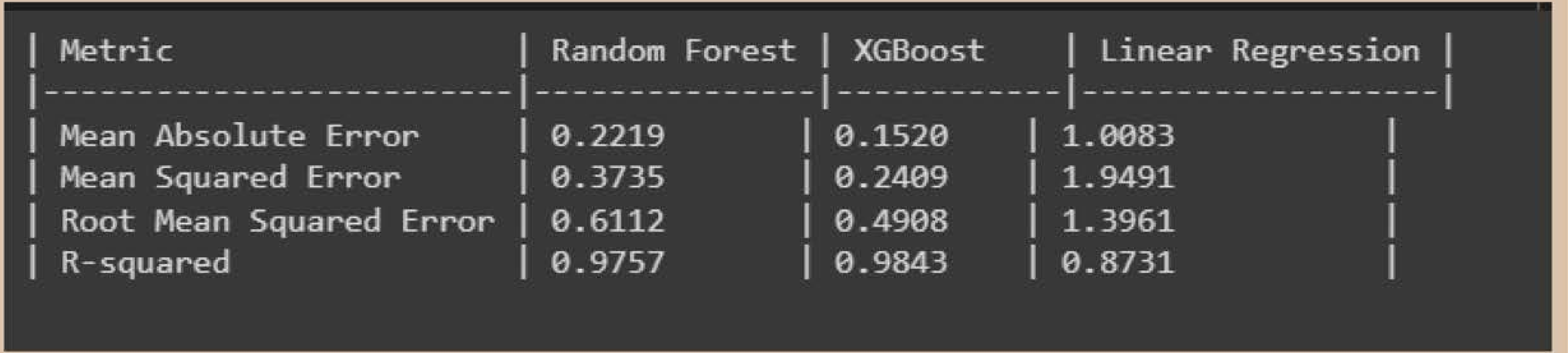
- Key Parameters :
  - n_estimators=100: Enhances model performance through 100 decision trees.
  - max_depth: Controls complexity by limiting tree depth, crucial for avoiding overfitting.
  - random_state: Guarantees that model results are consistent and replicable across different runs.

Model Fitting

XGBoost Regressor

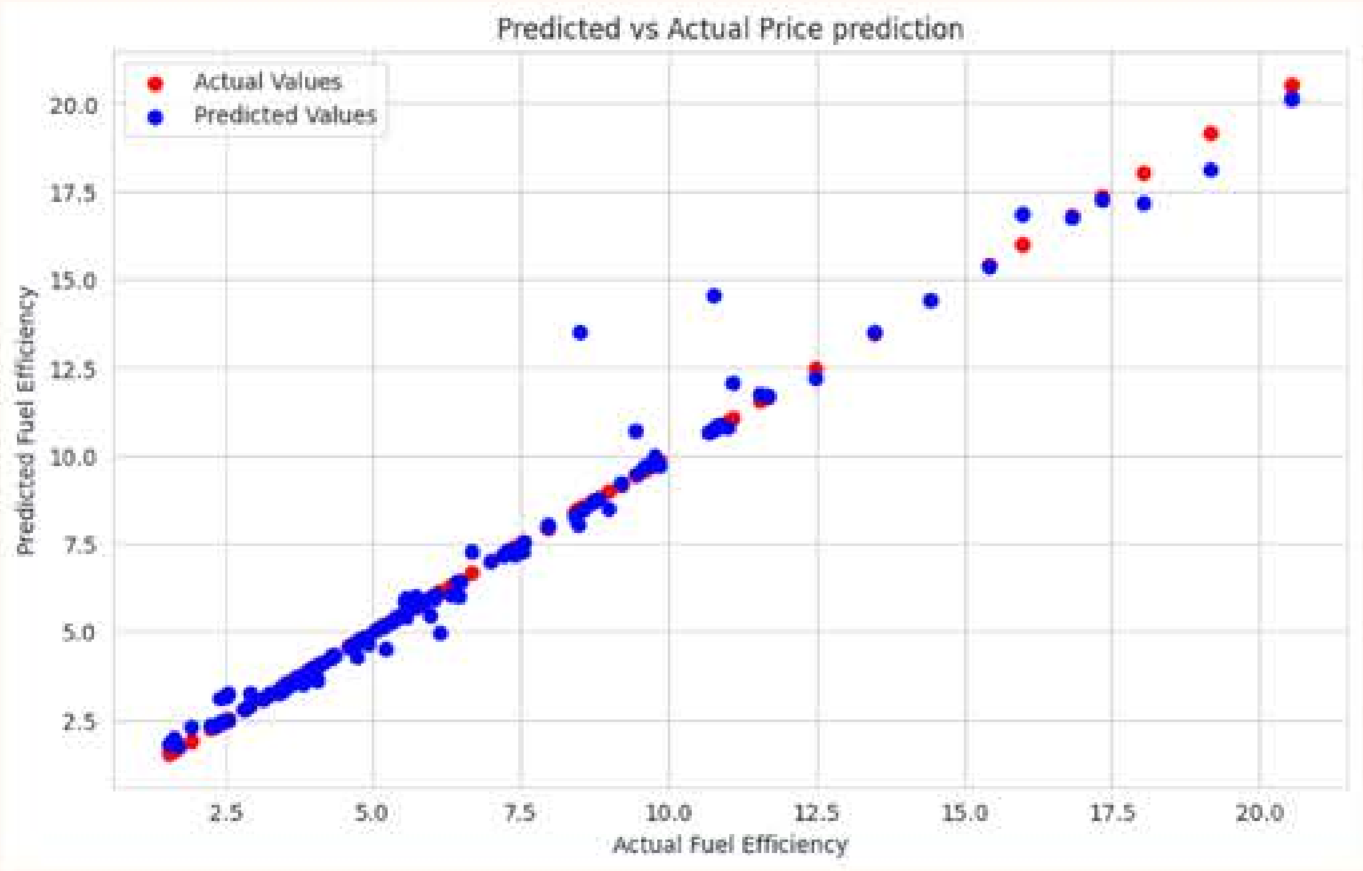
- XGBoost (Extreme Gradient Boosting) is a powerful and effic algorithm that uses family of gradient boosting methods. It builds an ensemble of weak learners (typically decision trees) sequentially, where each new learner corrects the errors made by the previous ones.
- Objective='reg:squarederror': This parameter specifies the objective function to optimize during training. In this case, it's set to 'reg:squarederror', indicating that the model should minimize the mean squared error.
- Includes features like ‘Engine Capacity’,’Insurance’,’Spare-key’, ‘Transmission’,’KM Driven’, ‘Ownership’,’Fuel’, ‘EMI/month’,’Mileage’ etc and ‘price_in_lakh’ as Target Variable.

Evulation Metric of Models

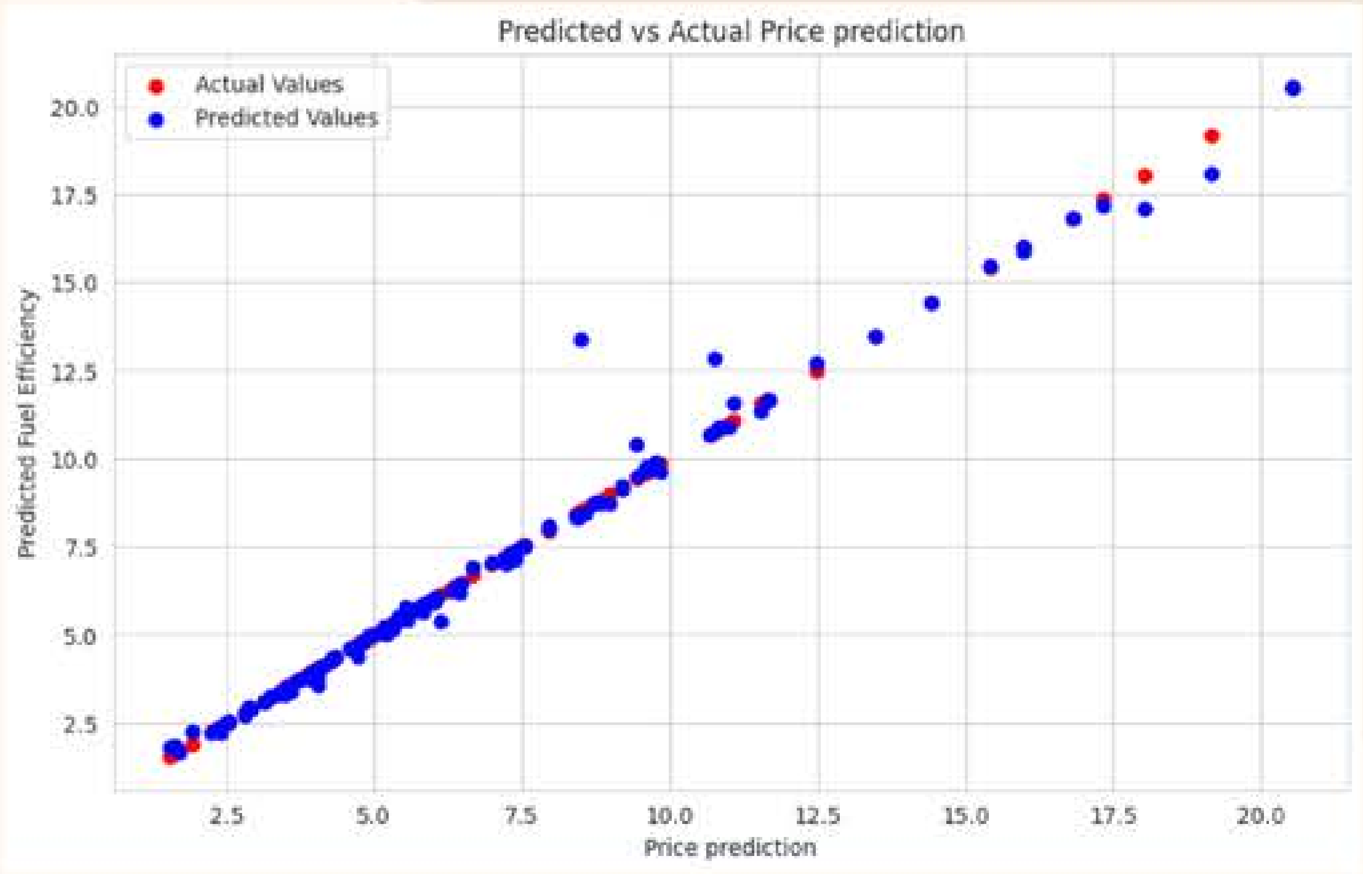


```
| Metric                    | Random Forest | XGBoost     | Linear Regression |
|---------------------------|---------------|-------------|-------------------|
| Mean Absolute Error       | 0.2219       | 0.1520     | 1.0083           |
| Mean Squared Error        | 0.3735       | 0.2409     | 1.9491           |
| Root Mean Squared Error   | 0.6112       | 0.4908     | 1.3961           |
| R-squared                 | 0.9757       | 0.9843     | 0.8731           |
```

Model Prediction Plots

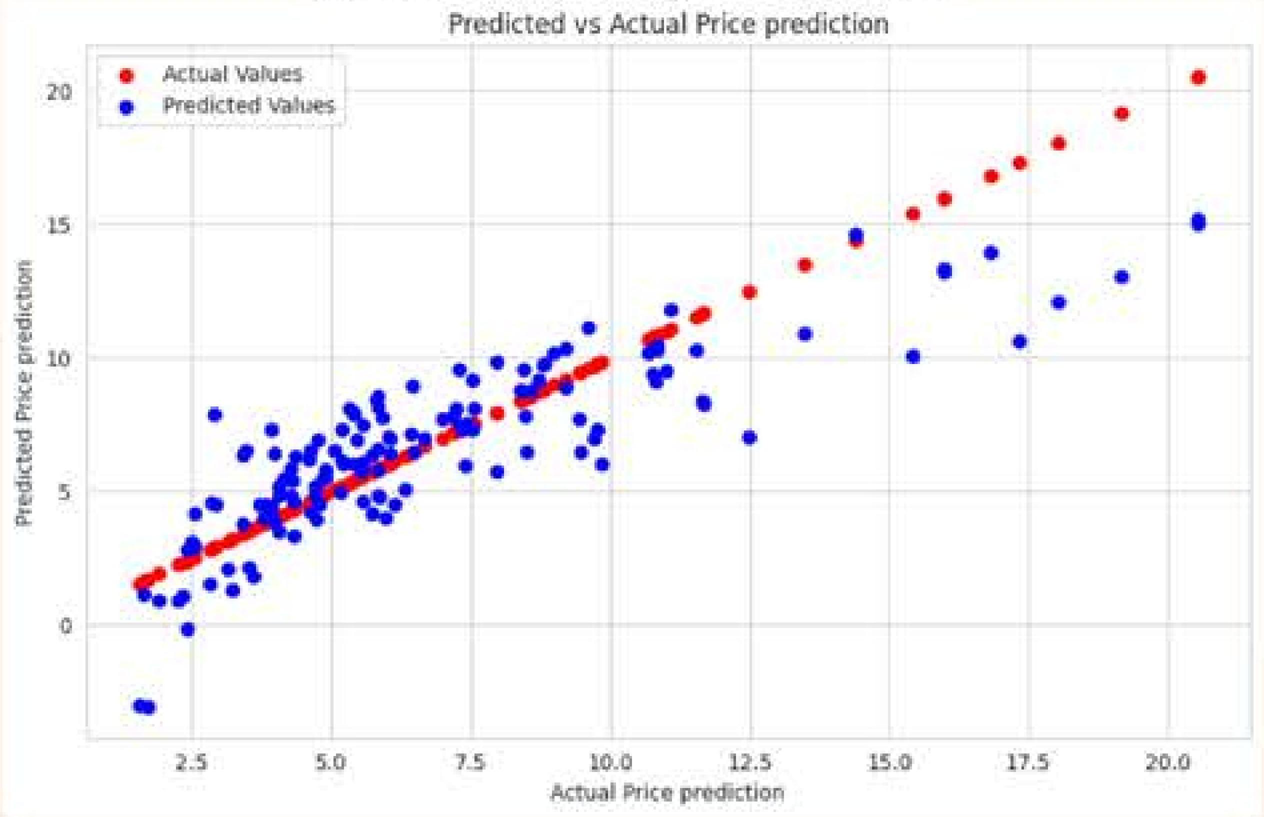


RandomForest Regressor



XGBoost Regressor

Linear Regression



Strategic Implications and Market Impact

Strategic Benefits

- Accurate predictions empower dealers to optimize pricing strategies, aligning prices closely with market dynamics.

Enhance Trust:

- Transparent, data-backed pricing fosters consumer trust, enhancing their purchase confidence and potentially boosting sales through improved customer satisfaction.

Conclusion

1. Our analysis utilized regression models such as XGBoost for mileage estimation and Linear Regression and Random Forest for pricing prediction, effectively mapping the complex relationships between vehicle attributes and outcomes like price .

2. Critical factors identified include brand, make year, and engine capacity for influencing both the pricing.

3. The models developed aid both sellers in refining pricing strategies and buyers in making informed decisions by providing insights on expected fuel economy, serving as a valuable tool for stakeholders in the automobile industry.

Future Plans

1. If we can have a more detailed dataset consisting more detail like the wear and tear of the car, any damage or accidents in the past, color etc., with more data in general would greatly enhance the accuracy of the prediction and analysis of the data.
2. Using more sophisticated learning algorithms and approaches we can find more patterns and interactions between variables that are currently unrecognized by cur rent models, which will in turn help increase theaccuracy of the prediction

Group Contribution

Member1 : Aadesh Minz has contributed in collecting the data by web scraping and fitting and prediction of the data.

Member2 : Asish Joel has contributed in cleaning the data and generated all the visual plots and analysis of the data.

Member3 : Raj Kariya contributed in the section of Feature Engineering and feature selection

Refrences

[1] Cars24 Services Pvt. Ltd., Vehicle Sales Data, obtained from Cars24 website :
https://www.cars24.com

[2] Group-4 Exploratory data Analysis of Vehicle sales :
https://colab. research.google.com/drive/1TQ1zclkL2vH5Z--69-D_GJya4RoPlaa_ ?
usp=sharing

[3] Group- 4 Web- Scrapping from Cars24 :
https://colab.research.google.com/drive/1qE87qLJ9hngMW0Zc0aWpb89-
jyy0hOYX?usp=sharing

Question Time

THANK YOU SO MUCH!