Problem Statement:

Objective:

Imagine you are working as a data scientist in Car Dheko, your aim is to enhance the customer experience and streamline the pricing process by leveraging machine learning. You need to create an accurate and user-friendly streamlit tool that predicts the prices of used cars based on various features. This tool should be deployed as an interactive web application for both customers and sales representatives to use seamlessly.

Project Scope:

We have historical data on used car prices from CarDekho, including various features such as make, model, year, fuel type, transmission type, and other relevant attributes from different cities. Your task as a data scientist is to develop a machine learning model that can accurately predict the prices of used cars based on these features. The model should be integrated into a Streamlit-based web application to allow users to input car details and receive an estimated price instantly.

Approach:

1. **Data Processing**
    a. **Import and concatenate:**
        i. Import all city's dataset which is in unstructured format.
        ii. Convert it into a structured format.
        iii. Add a new column named 'City' and assign values for all rows with the name of the respective city.
        iv. Concatenate all datasets and make it as a single dataset.
    b. **Handling Missing Values:** Identify and fill or remove missing values in the dataset.
        i. For numerical columns, use techniques like mean, median, or mode imputation.
        ii. For categorical columns, use mode imputation or create a new category for missing values.
    c. **Standardising Data Formats:**

        i. Check for all data types and do the necessary steps to keep the data in the correct format.
            1. Eg. If a data point has string formats like 70 kms, then remove the unit ‘kms’ and change the data type from string to integers.
    d. **Encoding Categorical Variables:** Convert categorical features into numerical values using encoding techniques.
        i. Use one-hot encoding for nominal categorical variables.
        ii. Use label encoding or ordinal encoding for ordinal categorical variables.
    e. **Normalizing Numerical Features:** Scale numerical features to a standard range, usually between 0 and 1.(For necessary algorithms)
        i. Apply techniques like Min-Max Scaling or Standard Scaling.
    f. **Removing Outliers:** Identify and remove or cap outliers in the dataset to avoid skewing the model.
        i. Use IQR (Interquartile Range) method or Z-score analysis.

2. Exploratory Data Analysis (EDA)

a. **Descriptive Statistics:** Calculate summary statistics to understand the distribution of data.
    i. Mean, median, mode, standard deviation, etc.
b. **Data Visualization:** Create visualizations to identify patterns and correlations.
    i. Use scatter plots, histograms, box plots, and correlation heatmaps.
c. **Feature Selection:** Identify important features that significantly impact the car prices.
    i. Use techniques like correlation analysis, feature importance from models, and domain knowledge.

3. Model Development

a. **Train-Test Split:** Split the dataset into training and testing sets to evaluate model performance.
    i. Common split ratios are 70-30 or 80-20.

b. **Model Selection:** Choose appropriate machine learning algorithms for price prediction.
    i. Linear Regression, Decision Trees, Random Forests, Gradient Boosting Machines, etc.
c. **Model Training:** Train the selected models on the training dataset.
    i. Use cross-validation techniques to ensure robust performance.
d. **Hyperparameter Tuning:** Optimize model parameters to improve performance.
    i. Use techniques like Grid Search or Random Search.

4. Model Evaluation

a. **Performance Metrics:** Evaluate model performance using relevant metrics.
    i. Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared.
b. **Model Comparison:** Compare different models based on evaluation metrics to select the best performing model.

5. Optimization

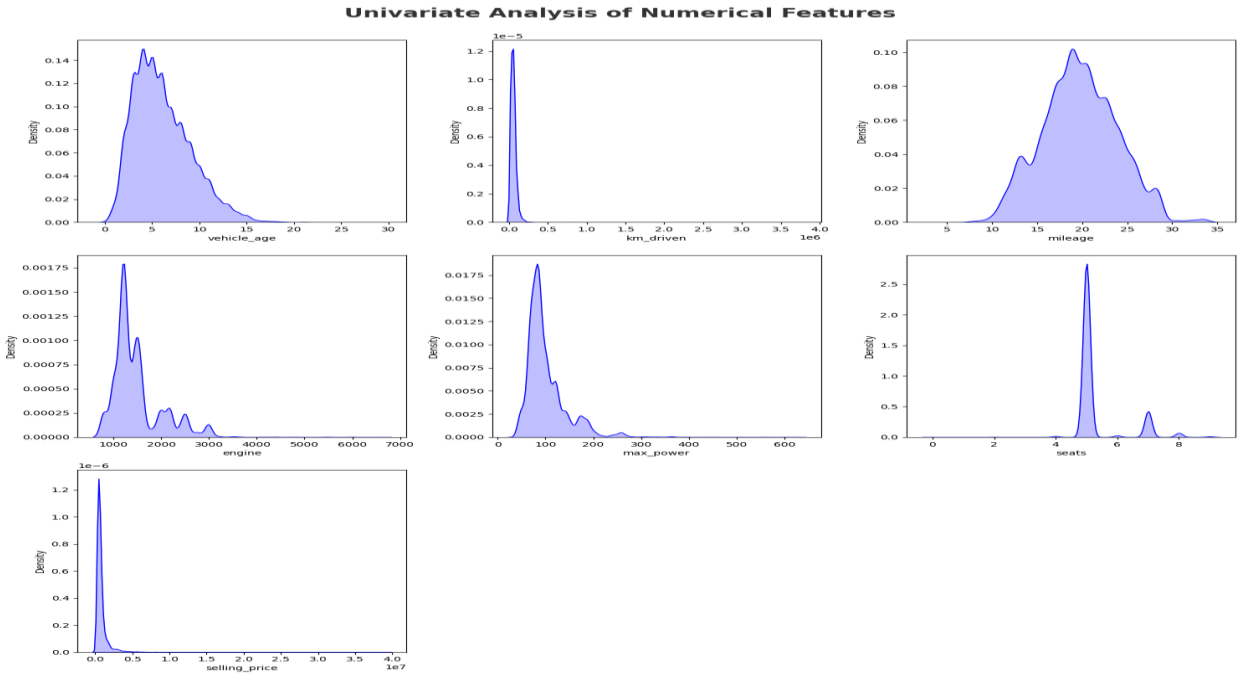
a. **Feature Engineering**: Create new features or modify existing ones to improve model performance.
    i. Use domain knowledge and exploratory data analysis insights.
b. **Regularization:** Apply regularization techniques to prevent overfitting.
    i. Lasso (L1) and Ridge (L2) regularization.

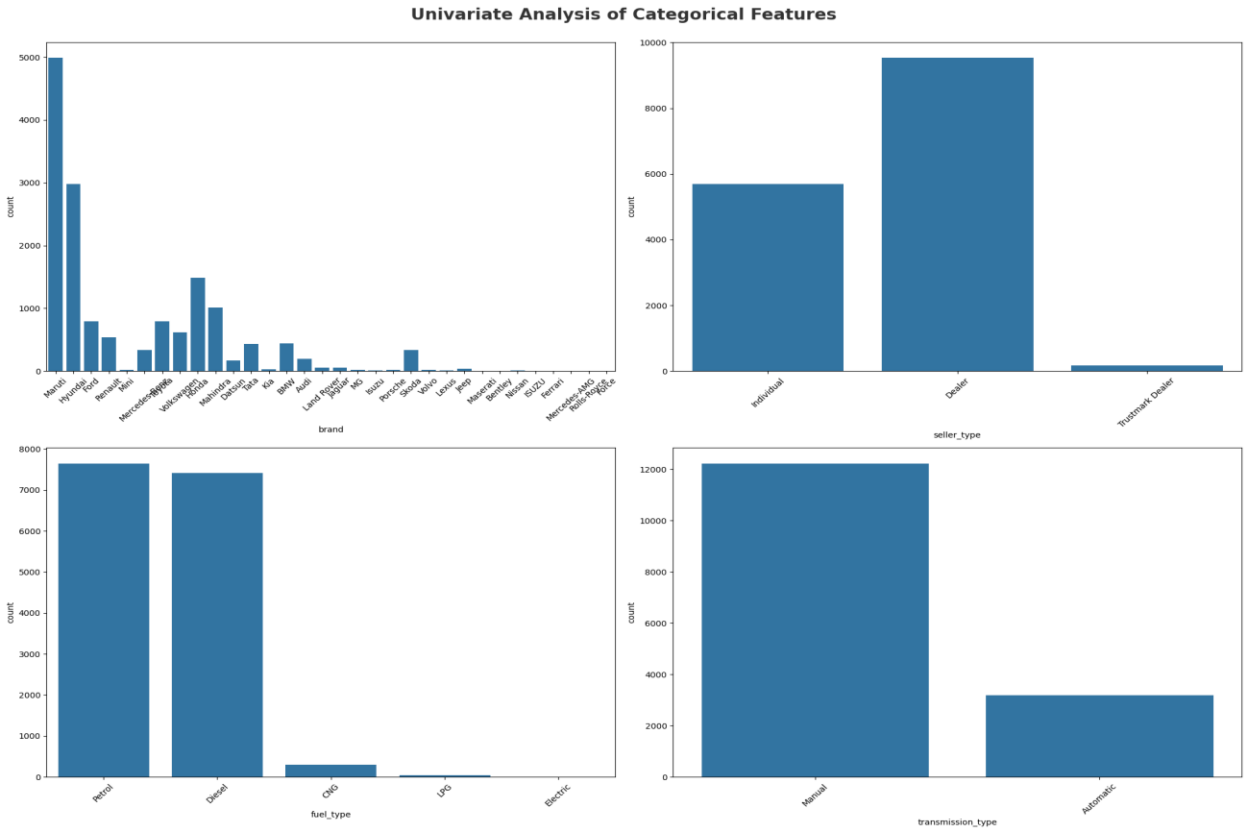
6. Deployment

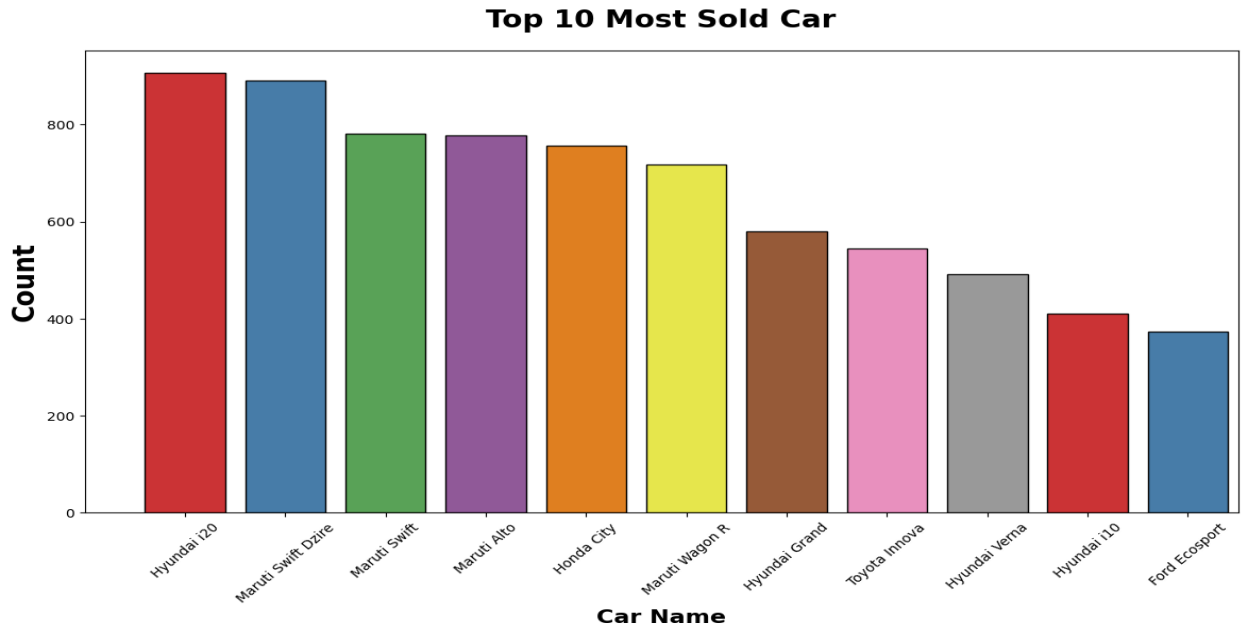
a. **Streamlit Application:** Deploy the final model using Streamlit to create an interactive web application.
    i. Allow users to input car features and get real-time price predictions.

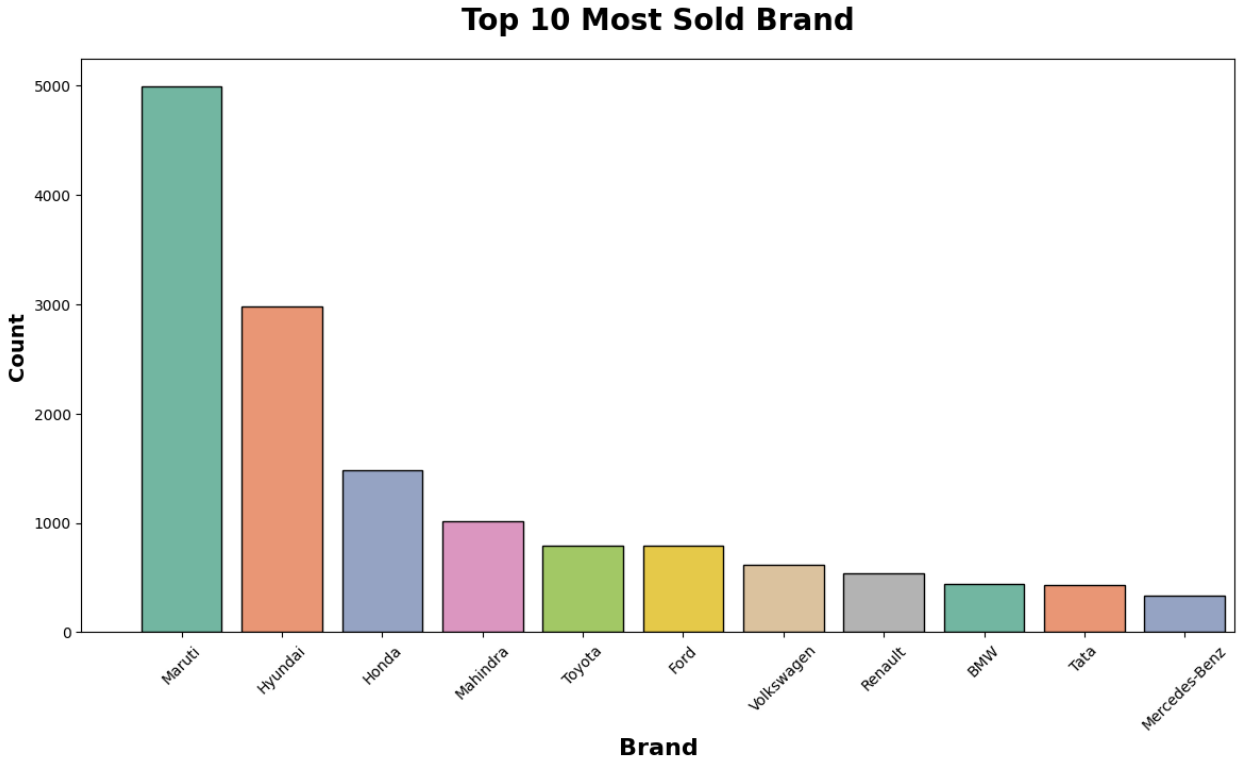
b. **User Interface Design**: Ensure the application is user-friendly and intuitive.
    i. Provide clear instructions and error handling.

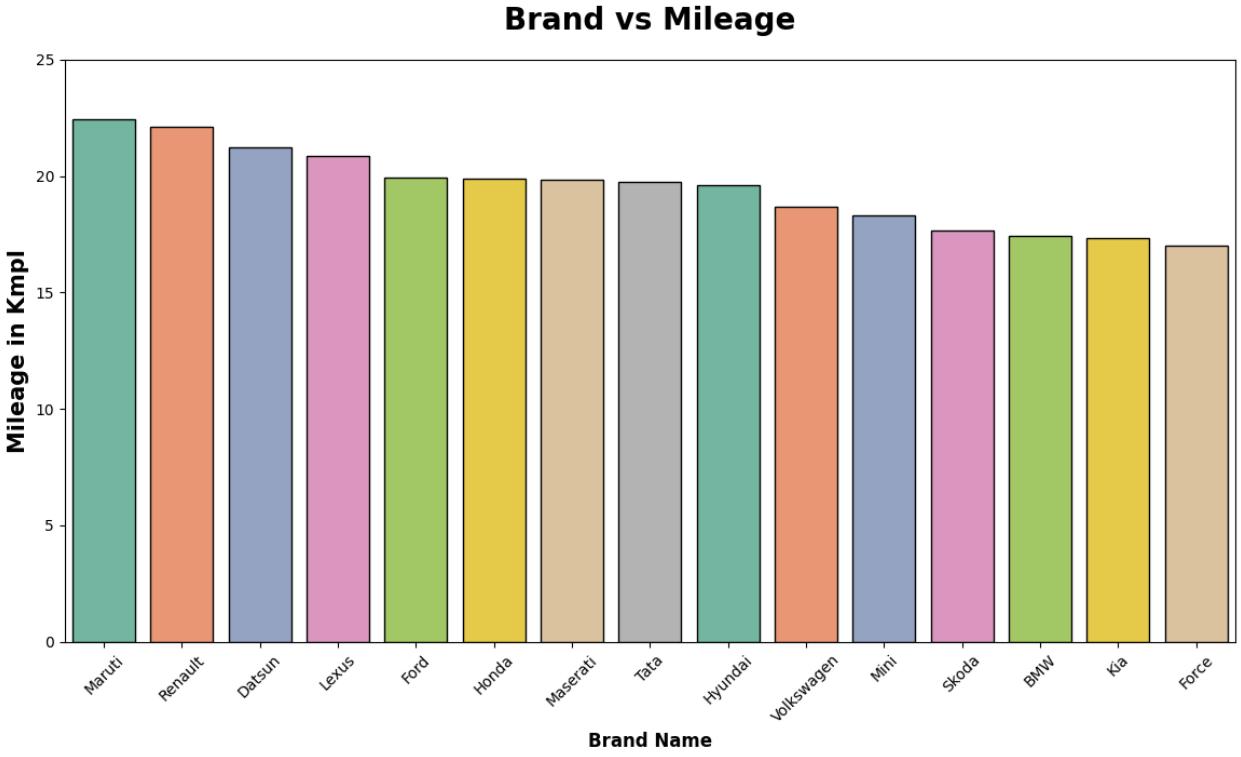
Data Analysis:

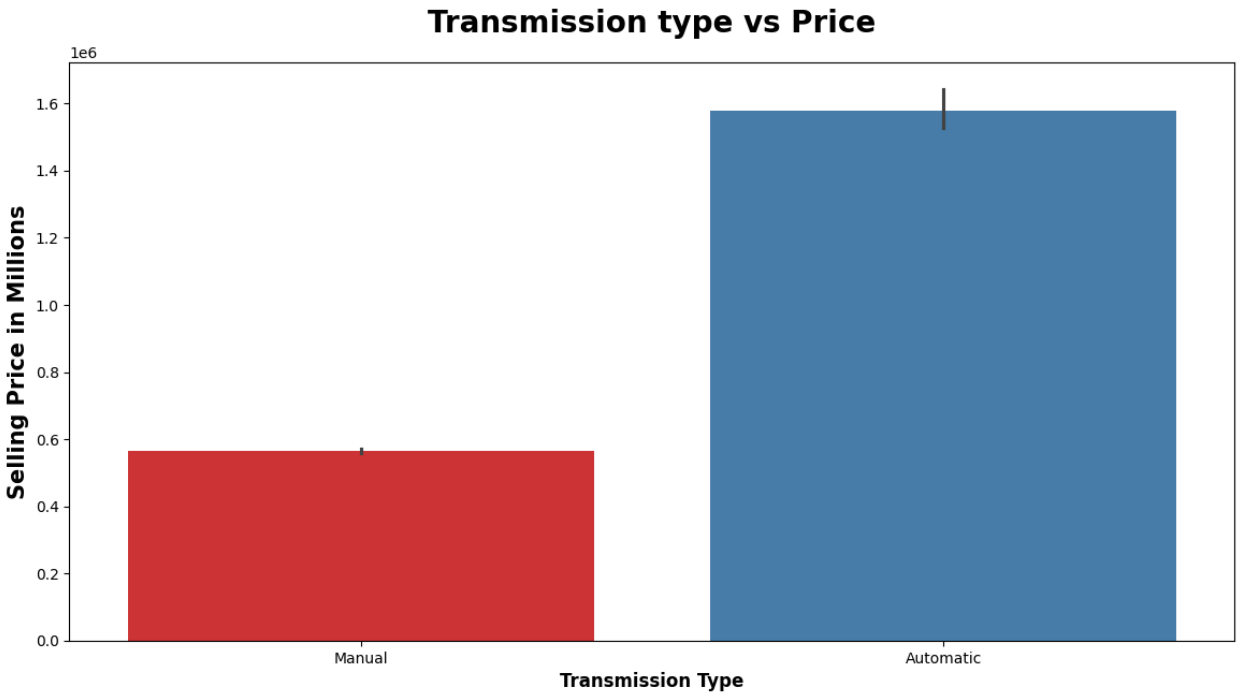


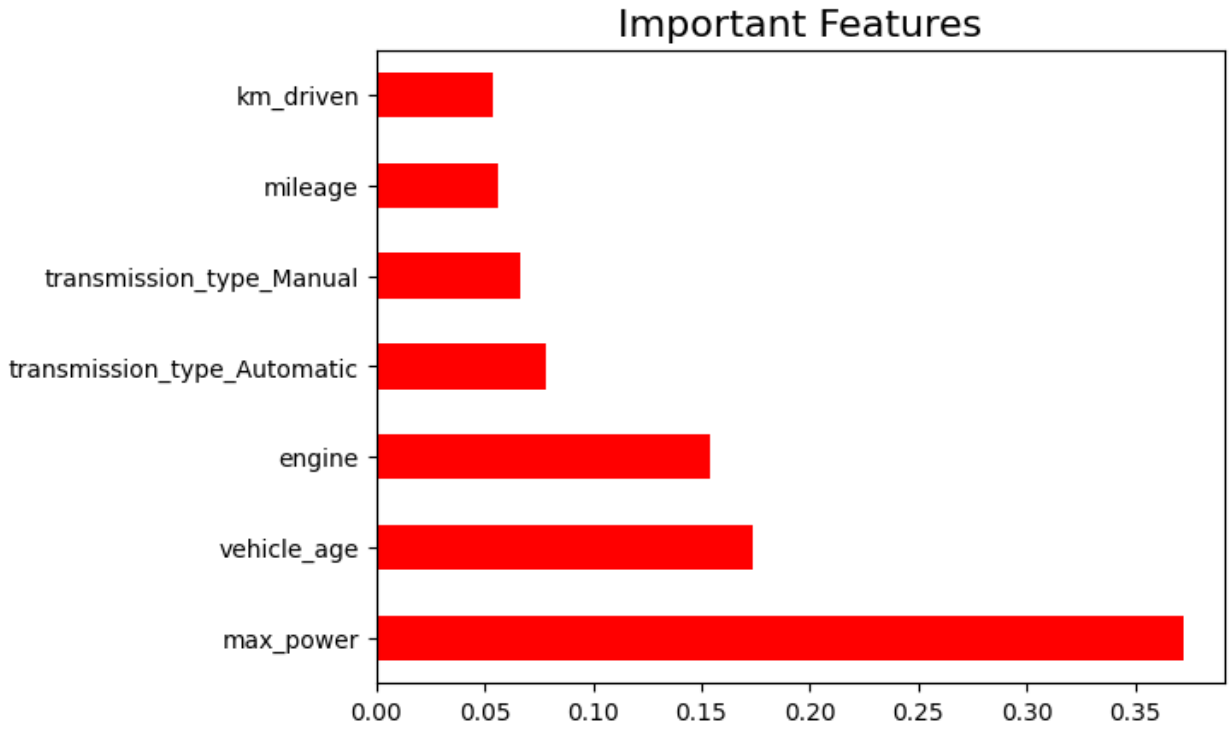

Univariate Analysis of Categorical Features
count
brand
Maruti
Hyundai
Ford
Renault
Mini
Mercedes-Benz
Skoda
Volkswagen
Honda
Mahindra
Datsun
Tata
Kia
BMW
Audi
Land Rover
Jaguar
MG
Isuzu
Porsche
Skoda
Volvo
Lexus
Jeep
Maserati
Bentley
Nissan
ISUZU
Ferrari
Mercedes-AMG
Rolls-Royce
Force
seller_type
Individual
Dealer
Trustmark Dealer
fuel_type
Petrol
Diesel
CNG
LPG
Electric
transmission_type
Manual
Automatic


Top 10 Most Sold Car
Count
Car Name
Hyundai i20
Maruti Swift Dzire
Maruti Swift
Maruti Alto
Honda City
Maruti Wagon R
Hyundai Grand
Toyota Innova
Hyundai Verna
Hyundai i10
Ford Ecosport


Top 10 Most Sold Brand
Count
5000
4000
3000
2000
1000
0
Maruti
Hyundai
Honda
Mahindra
Toyota
Ford
Volkswagen
Renault
BMW
Tata
Mercedes-Benz
Brand


Brand vs Mileage
Mileage in Kmpl
25
20
15
10
5
0
Maruti
Renault
Datsun
Lexus
Ford
Honda
Maserati
Tata
Hyundai
Volkswagen
Mini
Skoda
BMW
Kia
Force
Brand Name





Comparision:

Model	Mean Squarred Error	Root Mean Squarred Error	Explained Variance Score	R-Sqaure Score(Accuracy)
Linear Regression	2.009861e+11	448314.693363	0.669843	0.669251
Support Vector Rregression	6.474655e+11	804652.385273	0.000083	-0.065490
Decision Tree Regressor	8.755613e+10	295898.842803	0.855932	0.855915
Random Forest Regressor	7.058197e+10	265672.673833	0.883900	0.883848
Ridge	2.009844e+11	448312.849115	0.669846	0.669253
Lasso	2.009860e+11	448314.633565	0.669843	0.669251