**Car Price Prediction Project Report**

# Executive Summary

The primary objective of this project was to develop a predictive model to estimate car prices based on various features. By utilizing machine learning techniques, we aimed to understand the relationships between features such as mileage, engine size, fuel type, and transmission type. This information is crucial for stakeholders in the automotive market, allowing them to make informed decisions.

# Project Overview

The dataset used in this project comprises various attributes of cars, which were analyzed to create a model capable of predicting their market prices. The following sections detail the methodology, data preprocessing steps, model training, evaluation, and results achieved.

# Dataset Description

The dataset includes the following columns:

* **Brand**: The car manufacturer.
* **Model**: The specific model of the car.
* **Year**: The year of manufacture.
* **Transmission**: Type of transmission (e.g., Manual, Automatic).
* **Mileage**: Total distance traveled by car.
* **Fuel Type**: The fuel type used (e.g., Petrol, Diesel).
* **Tax**: The applicable tax for the car.
* **MPG**: Fuel efficiency (miles per gallon).
* **Engine Size**: Size of the engine in liters.
* **Price**: The market price of the car.
* **Car Age**: Age of the car (calculated as the difference between the current year and the year of manufacture).
* **Mileage per Age**: Average mileage per year.

# Methodology

## Data Preprocessing

### Data Cleaning:

* + **Checking for Missing Values**: The dataset was checked for missing values using the isnull() function to ensure data quality.
  + **Handling Missing Values**: Missing values were dealt with either by deleting rows or filling them with mean or mode values as appropriate.
  + **Detecting Outliers**: Techniques such as box plots were utilized to identify outliers, which were removed if they affected model accuracy.

### Data Visualization

#### Mileage vs. Car Age by Fuel Type

A graph of mileage and car age

Description automatically generated

#### Price Distribution by Transmission and Year

A graph showing the price of a product

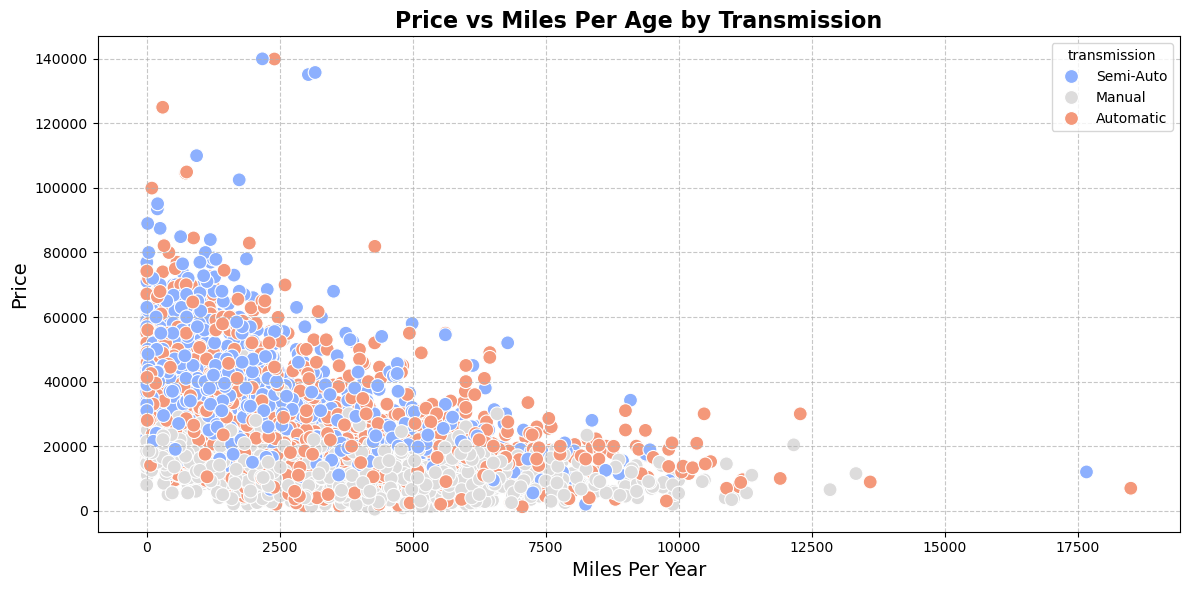
Description automatically generated with medium confidence

#### Price Distribution by Brand and Transmission

A graph of different colored bars

Description automatically generated

#### Price vs. Miles Per Age by Transmission

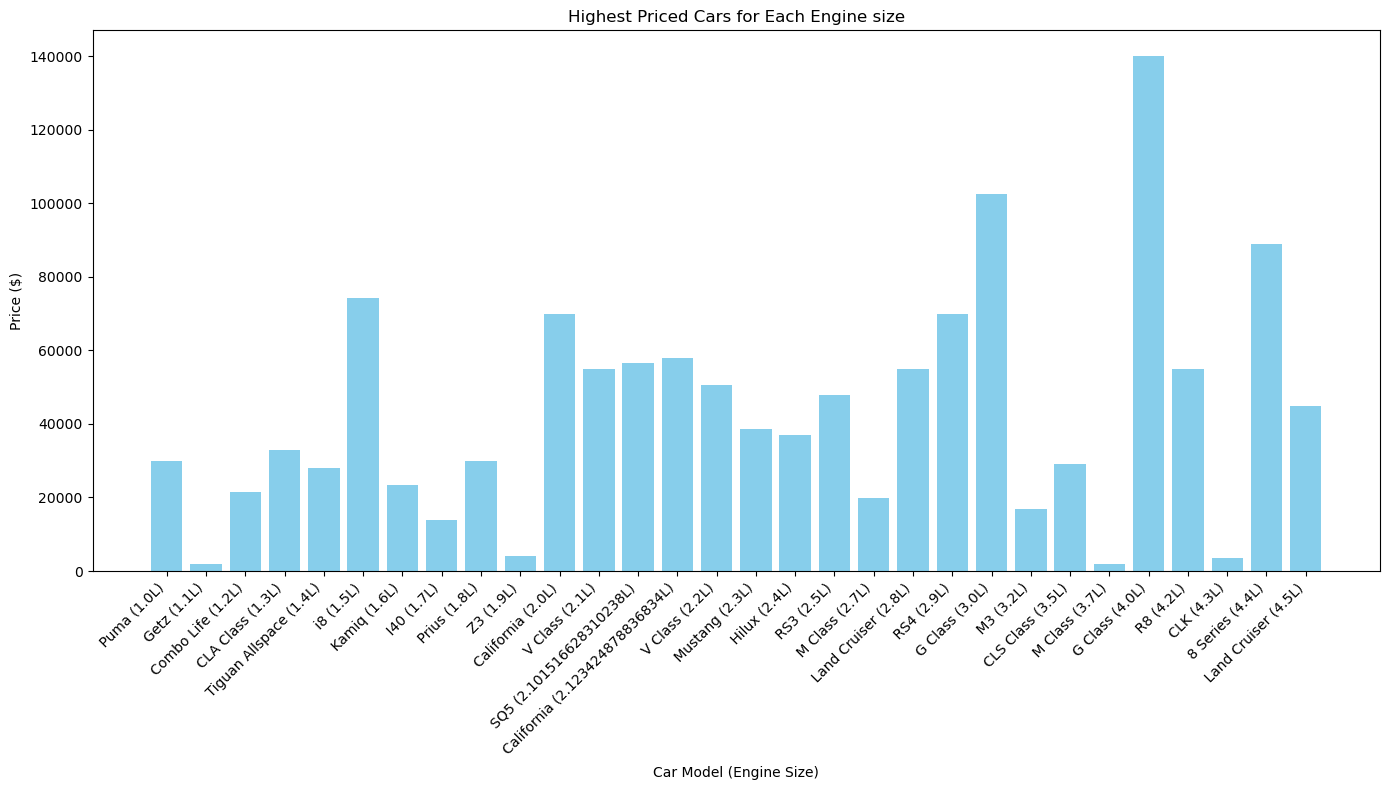


#### Mileage per Age by Brand and Fuel Type

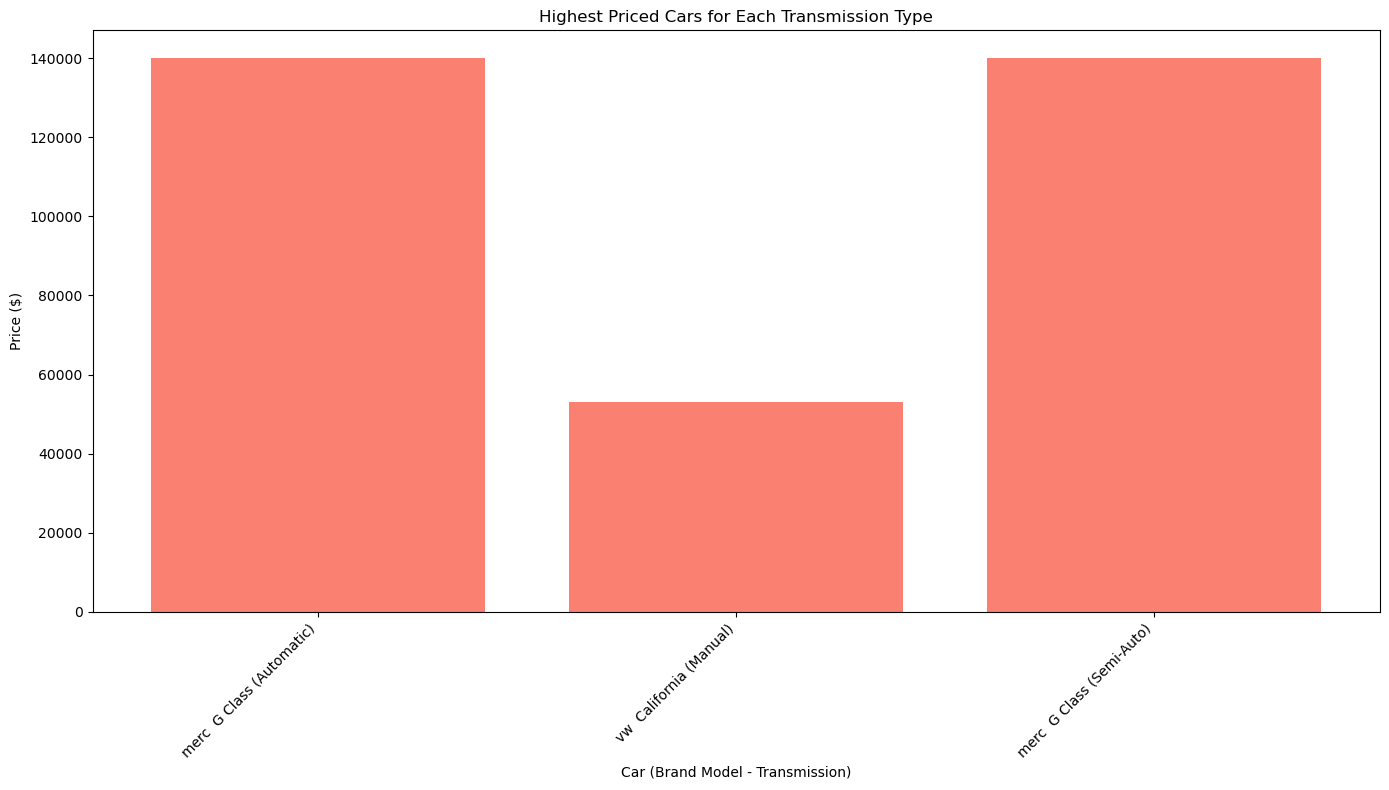
A graph of different colored squares

Description automatically generated

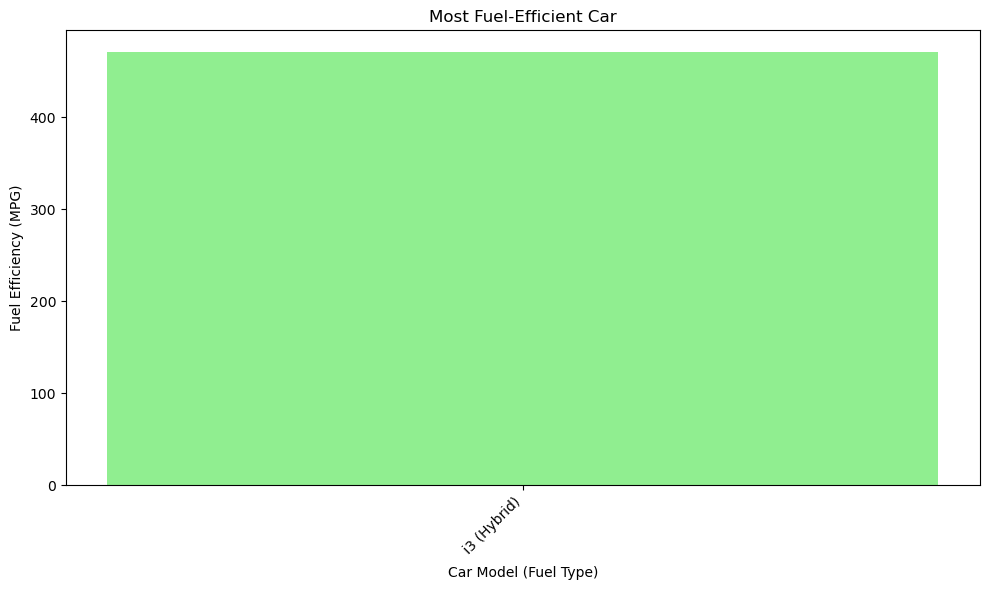
#### Highest priced cars for each engine type



#### Highest priced cars for each transmission type



#### Most fuel-efficient car (highest MPG)



### Encoding Categorical Variables

* + **Converting Categorical Variables**: Categorical columns like brand, model, fuel type, and transmission were converted to numerical format using One-Hot Encoding.
  + **Removing Original Columns**: The original categorical columns were removed after encoding to avoid redundant information.

### Normalization

* + **Applying Normalization**: Min-Max Scaling was used to normalize numerical columns like mileage and engine size to a uniform scale between 0 and 1.
  + **Confirming Normalization**: It was confirmed that all values were in accordance with the new scale

### Feature Selection

* + **Applying SelectKBest**: SelectKBest was used to identify the top features impacting car prices based on ANOVA testing.
  + **Analyzing Results**: The selected features were reviewed based on the resulting values, determining the relative importance of each feature.

## Model Development

### Data Splitting

* + **Splitting the Data**: The dataset was divided into training (80%) and testing (20%) sets using train\_test\_split to ensure an accurate evaluation of model performance.
  + **Ensuring Distribution**: It was verified that the distribution of features was balanced in both the training and testing sets.

### 2- Model Selection

* + **Choosing the Model**: A Random Forest Regressor was selected due to its effectiveness in handling large and complex datasets.
  + **Analyzing Features**: The model’s characteristics, such as the number of trees (n\_estimators) and the maximum depth of trees (max\_depth), were reviewed.

### 3- Model Training

* + **Applying Training**: The model was trained on the training set using the selected features.
  + **Tuning Hyperparameters**: Techniques such as Grid Search were employed to optimize the model’s hyperparameters and enhance performance.

## Model Evaluation

The model's performance was evaluated using the following metrics:

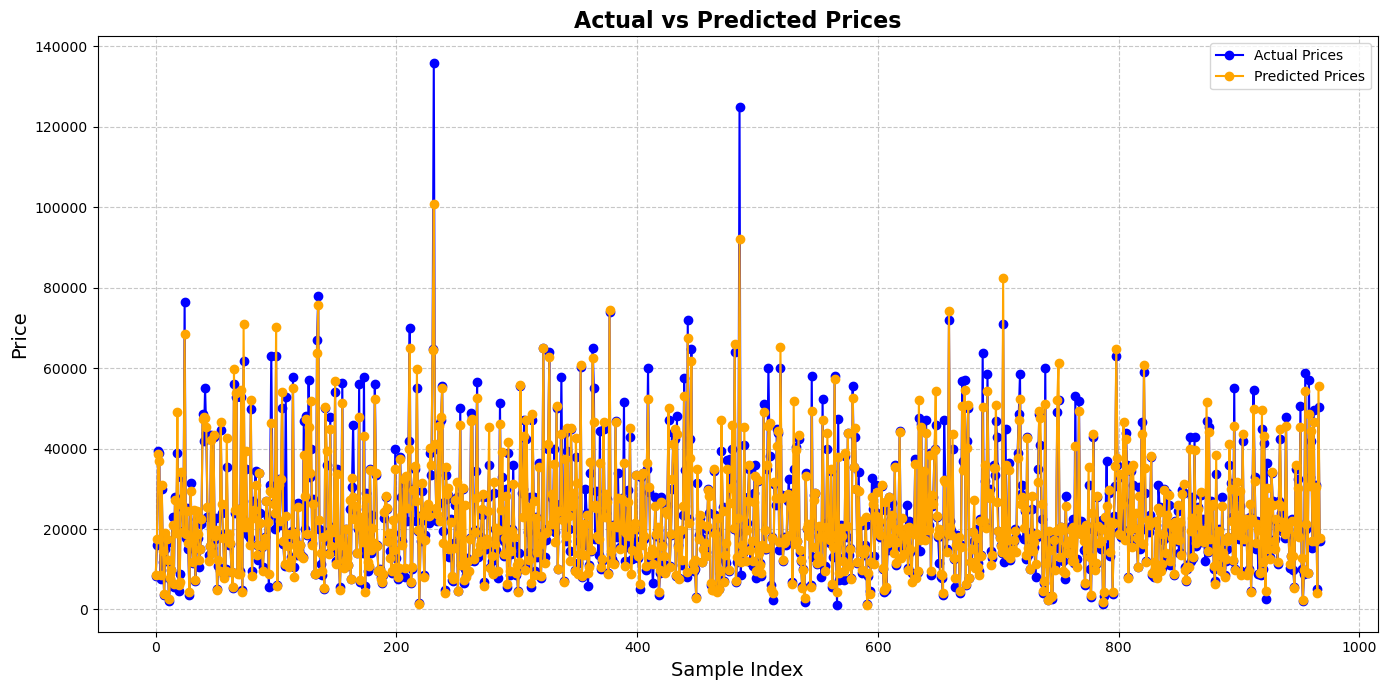
### Mean Squared Error (MSE)

* + **Calculating MSE**: MSE was computed using the mean\_squared\_error function to compare predicted values with actual prices.
  + **Analyzing Results**: Mean Absolute Error: 2134.03 and R^2 score: 0.94

### R² Score

* + **Calculating R² Score:** R² was calculated using the r2\_score function, achieving a score of 0.94, suggesting that the model explains 94% of the variance in car prices.
  + **Interpreting Results**: The R² value was utilized to assess the model's fit to the data. Conclusion

# Results

* Mean Absolute Error: 2134.03
* The R² Score achieved was 0.94, indicating that the model explains 94% of the variance in car prices based on the selected features.
* **Visualization**: To illustrate the model's performance, a comparison of actual versus predicted prices was plotted. This visualization showed a close alignment between the predicted and actual values, reinforcing the model's accuracy.

# Conclusion

The car price prediction project successfully demonstrated the effectiveness of machine learning techniques in estimating car prices. The Random Forest model produced strong results, with a high R² score, suggesting that it can reliably predict car prices based on key features. These insights can provide significant value to stakeholders in the automotive market, enabling them to make informed pricing decisions.