Car Price Analysis and Prediction Using Machine Learning

1. Introduction

The automotive industry is one of the most competitive markets globally, with car prices being influenced by a multitude of factors. Understanding these factors and accurately predicting car prices can provide significant insights for manufacturers, dealers, and buyers alike. This project focuses on the detailed analysis of car attributes and the development of a predictive model using machine learning to estimate car prices.

2. Dataset Overview

The dataset used in this project contains various features that are crucial for determining car prices:

- **Brand**: Manufacturer of the car.
- **Model**: Specific model of the car.
- **Year**: Year of manufacture.
- **Transmission**: Type of transmission (e.g., Automatic, Manual).
- **Mileage**: Total distance the car has been driven.
- **Fuel Type**: Type of fuel used by the car (e.g., Diesel, Petrol).
- **Tax**: Annual road tax.
- **MPG**: Miles per gallon, indicating fuel efficiency.
- **Engine Size**: Volume of the car's engine.
- **Price**: Selling price of the car.

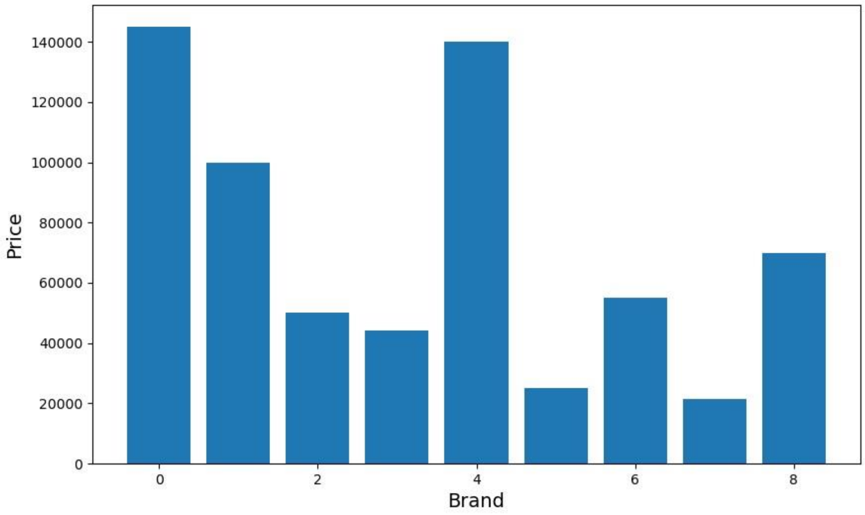
3. Data Preprocessing

- **Handling Missing Values**:
  - **Categorical Features**: Missing values in categorical variables such as Brand and Fuel Type were imputed using the mode (most frequent value).
  - **Numerical Features**: Missing values in numerical variables such as Mileage, MPG, and Engine Size were filled using the mean value of each respective column.
- **Removal of Duplicates**: Duplicate records were identified and removed to ensure data quality and accuracy.
- **Outlier Detection and Treatment**: Outliers in key numerical features like Year, Mileage, Tax, MPG, Engine Size, and Price were identified using the Interquartile Range (IQR) method. These outliers were analyzed to understand their impact on the model.

4. Data Visualization

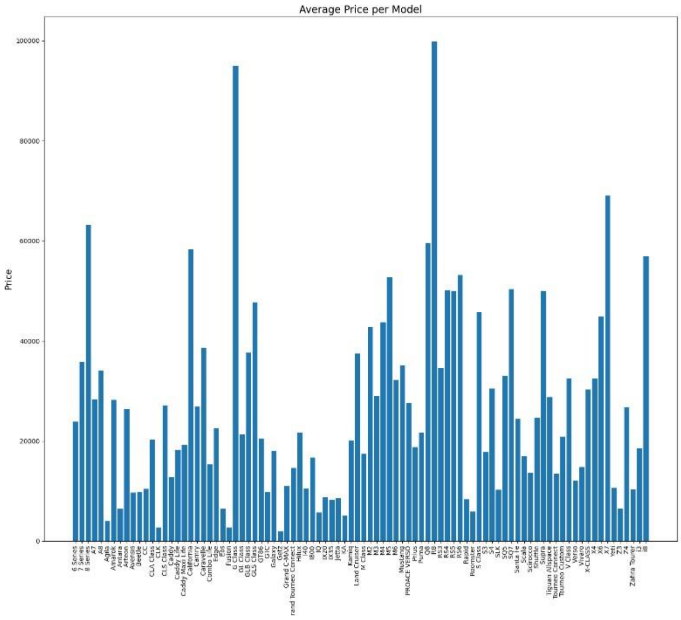
Visualizations were created to explore the relationship between different features:

- **Average Price per Brand**: A bar chart showing the average price of cars for each brand.



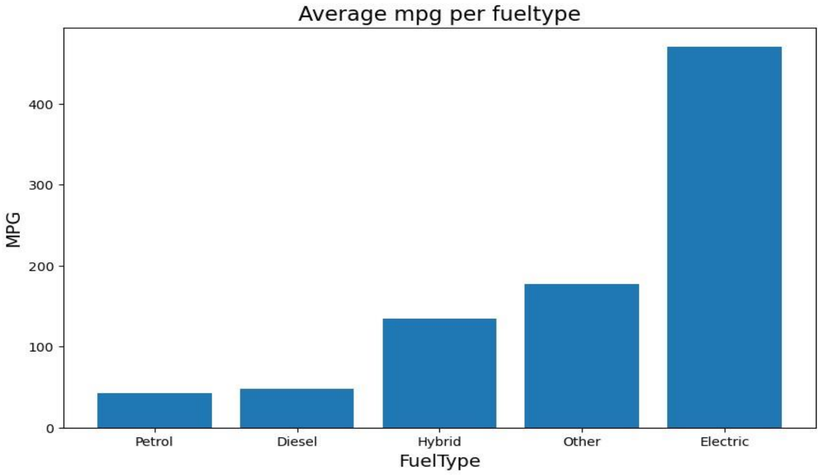
-

- **Average Price per Model**: A bar chart showing the average price for each car model.

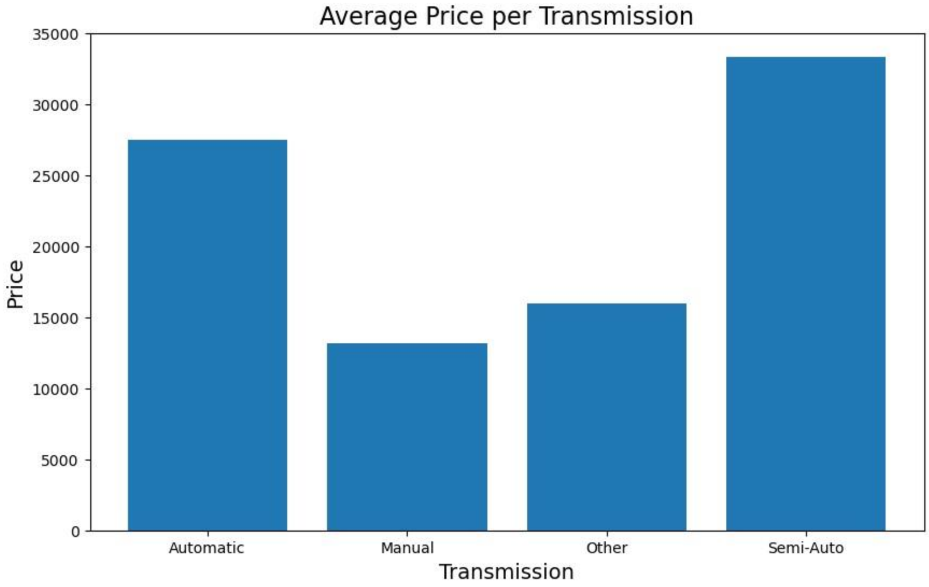


-

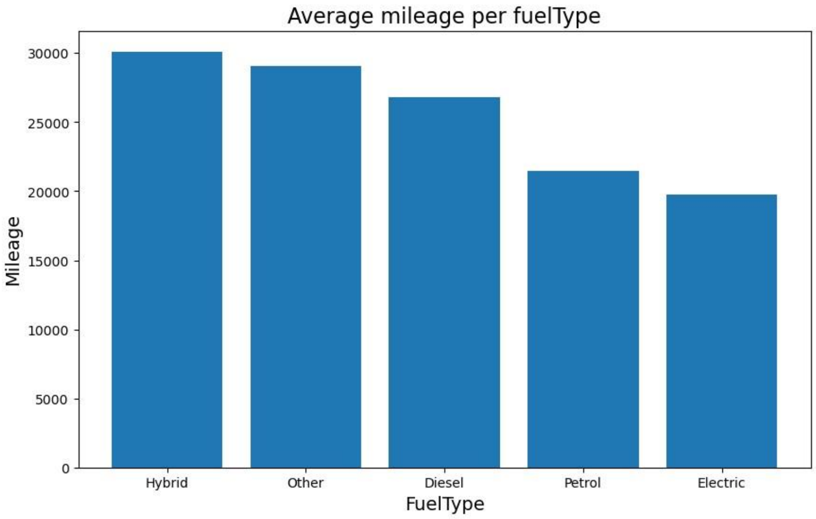
- **Average MPG per Fuel Type**: A bar chart illustrating the fuel efficiency across different fuel types.



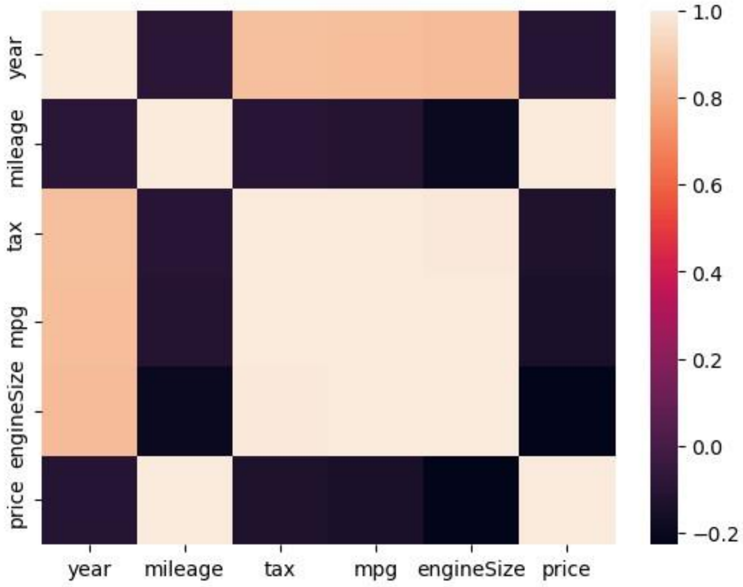
-
- **Average Price per Transmission**: A bar chart comparing car prices based on transmission type.



-
- **Average Mileage per Fuel Type**: A bar chart showing the average mileage for each fuel type.

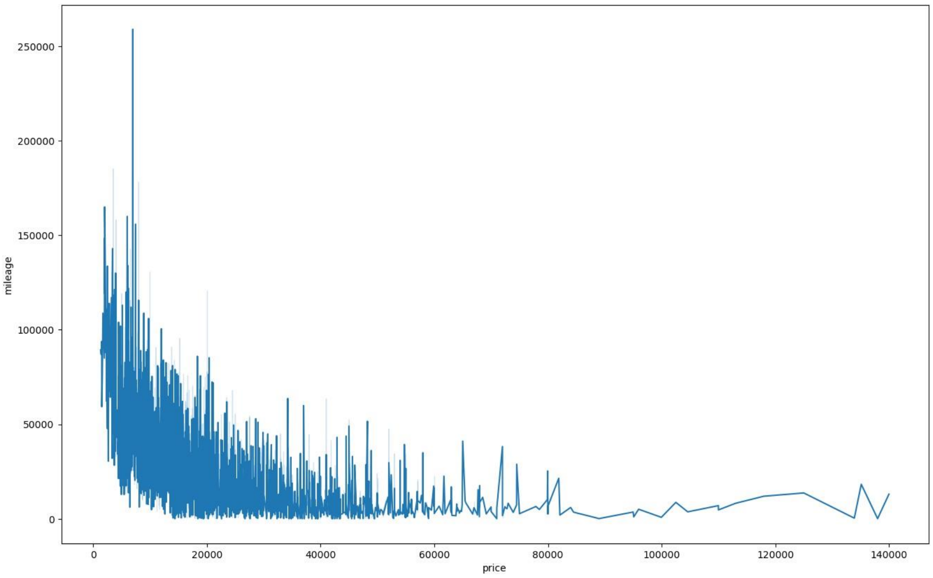


-
- **Correlation Matrix**: A heatmap showing the correlation between numerical features.



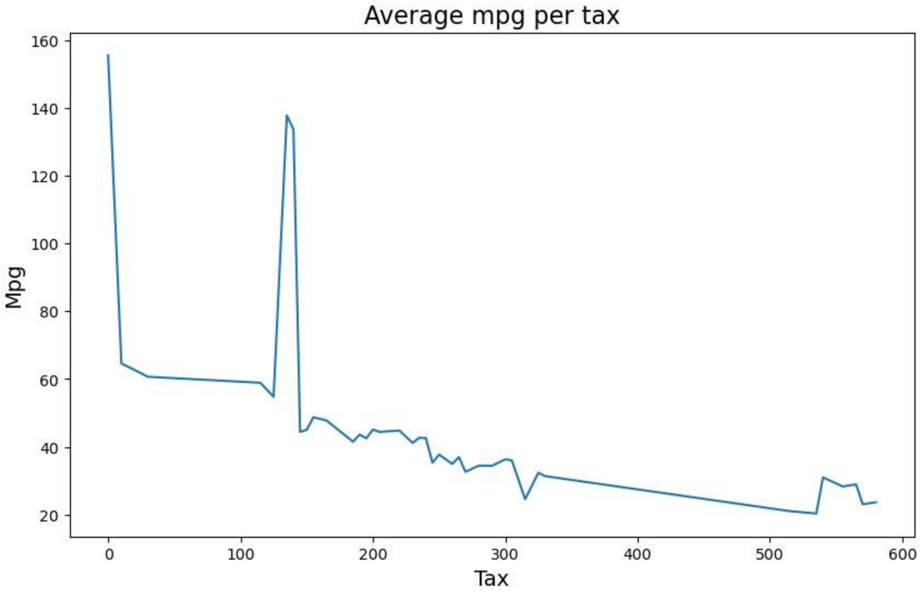
-

- **Price vs. Mileage**: A line plot illustrating the relationship between mileage and price.



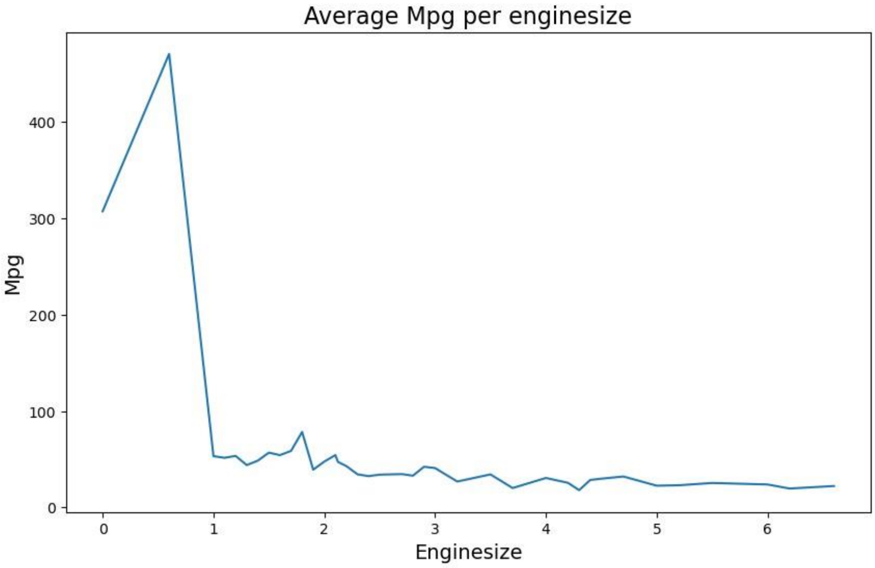
-

- **Tax vs. MPG**: A line plot depicting the relationship between tax and mpg.



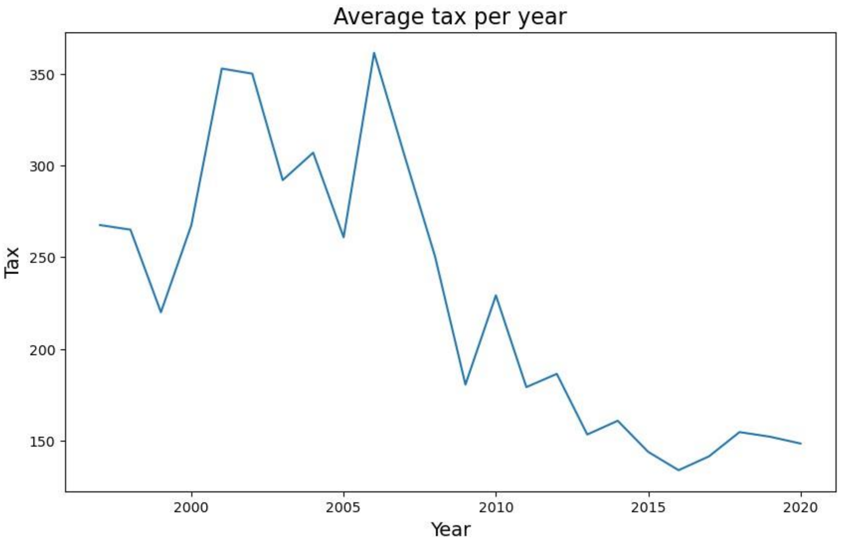
-

- **Engine Size vs. MPG**: A line plot showing the relationship between engine size and mpg.



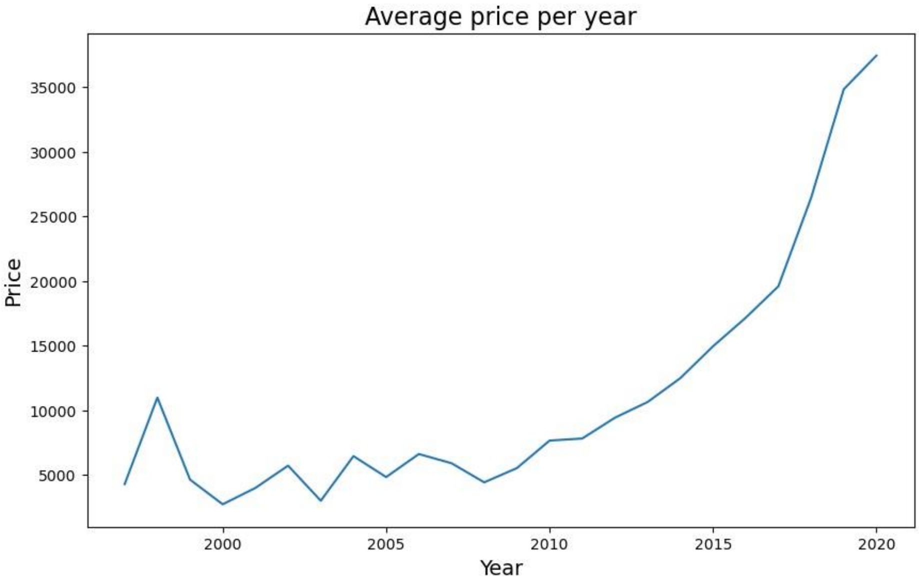
-

- **Tax vs. Year**: A line plot showing the trend of average tax over the years.



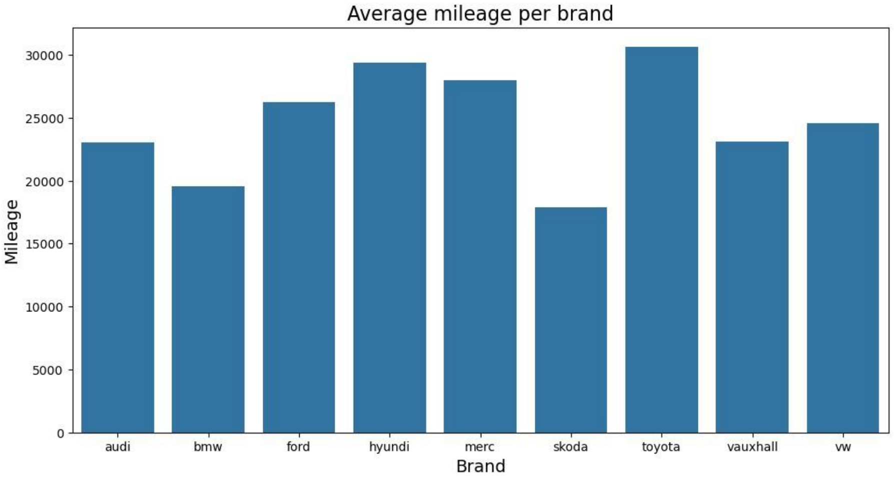
-

- **Price vs. Year**: A line plot depicting the relationship between car price and production year.



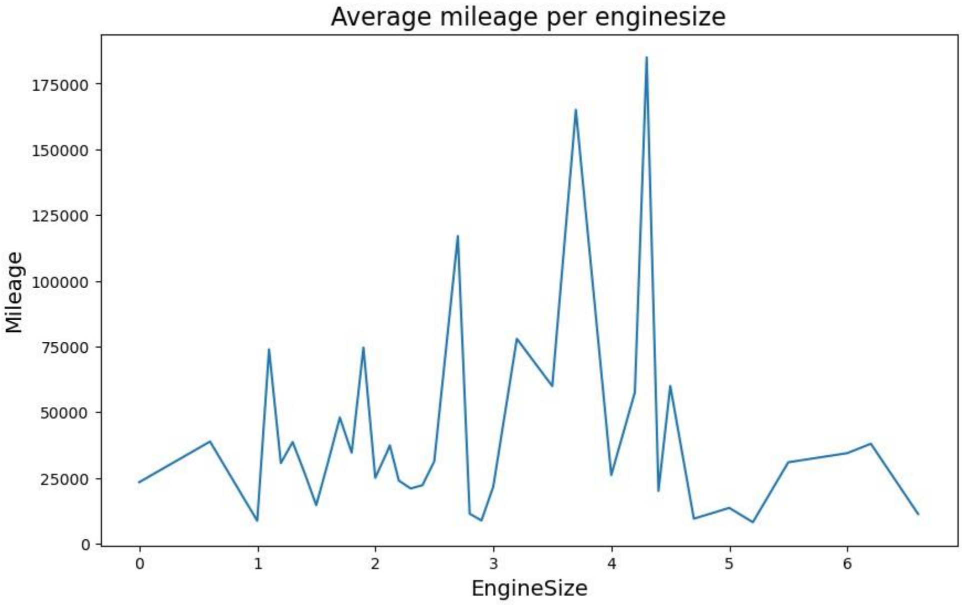
-

- **Average Mileage per Brand**: A bar chart showing the average mileage for each brand.



-

- **Mileage vs. Engine Size**: A line plot showing the relationship between engine size and mileage.



-

Data Encoding and Scaling

- **Label Encoding**: Categorical variables such as Brand, Model, Transmission, and Fuel Type were converted into numerical values using label encoding to make them suitable for model training.
- **Feature Scaling**: Numerical features were scaled using StandardScaler to normalize the data, ensuring all features contributed equally to the predictive model.

6. Feature Selection

Using SelectKBest with f_regression as the scoring function, the most significant features were selected for model training. This step ensured that only the most relevant features were used, improving the model's accuracy and reducing overfitting.

7. Model Development

- **Data Splitting**: The dataset was split into training and test sets in an 80:20 ratio to evaluate the model's performance on unseen data.
- **Model Selection**: A RandomForestClassifier was chosen for its robustness and ability to handle large datasets with high dimensionality. The model was trained on the selected features to predict car prices.

8. Model Evaluation

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

- **Accuracy**: The model's accuracy was computed, reflecting the proportion of correct predictions.

- **Mean Absolute Error (MAE)**: The MAE was calculated to measure the average magnitude of errors between predicted and actual prices.
- **R-squared (R^2) Score**: The R^2 score was used to assess how well the model explains the variance in car prices.

9. Results and Discussion

The Random Forest model demonstrated strong predictive capabilities, with high accuracy, low mean absolute error, and a good R^2 score. The analysis showed that factors such as Brand, Model, Year, Mileage, and Engine Size significantly influence car prices.

10. Key Factors Influencing Car Price:

1. **Brand: Brand has a significant impact on the price. Some brands have a strong reputation and command higher prices.**
2. **Model: Certain well-known or luxury models tend to have higher prices.**
3. **Year of Manufacture: Newer cars tend to be more expensive compared to older models.**
4. **Mileage: The higher the mileage, the lower the price of the car.**
5. **Engine Size: Engine size directly affects the price; cars with larger engines are typically more expensive.**
6. **Fuel Type: The type of fuel used can influence the price, especially with the rising costs of fuel and the shift toward electric or hybrid cars.**
7. **Transmission Type: Cars with automatic transmissions tend to be priced higher than those with manual transmissions.**

Best Cars:

1. **By Brand:**
   - **Luxury brands like BMW and Mercedes command higher prices due to their high quality and superior performance.**
2. **By Fuel Efficiency:**
   - **Cars with high fuel efficiency, such as the Toyota Prius (hybrid cars), offer good long-term value due to their fuel savings.**
3. **By Newer Models:**
   - **Newer models are generally more technologically advanced and attract higher prices in the market.**

Future Expectations:

1. **Increasing Demand for Hybrid and Electric Cars: With the global shift toward sustainable energy, prices for electric and hybrid cars are expected to continue rising.**
2. **Technological Advancements: Cars equipped with advanced technological features, such as self-driving systems, will have higher prices and attract greater interest.**

3. **Fluctuations in the Used Car Market: As sales of used cars increase, there is an opportunity for higher demand for cars with excellent features and high quality but at lower prices than new cars.**

Conclusion:

The report indicates that the most influential factors on car prices include brand, model, year of manufacture, mileage, and engine size. As for the best cars, luxury brands and newer models with high fuel efficiency stand out as top choices.