Car Dataset Analysis and Prediction Project

**1. Project Overview**

The goal of this project is to analyze a dataset of cars, clean and preprocess the data, and develop a machine learning model to predict the **total cost** of a car based on its features. This analysis includes understanding the structure of the dataset, handling missing and duplicated values, visualizing data trends, and applying a linear regression model to predict car prices. Below are the steps taken during this project.

**2. Data Overview**

The dataset contains multiple features that describe various attributes of cars such as:

* **brand**: The manufacturer of the car.
* **model**: The specific model of the car.
* **year**: The manufacturing year.
* **transmission**: Type of transmission (e.g., manual, automatic).
* **mileage**: The total miles driven.
* **fuelType**: The type of fuel used (e.g., petrol, diesel).
* **engineSize**: Engine size in liters.
* **mpg**: Miles per gallon (fuel efficiency).
* **price**: The cost of the car.
* **tax**: Tax applicable on the car.

We derived the target variable, **total cost**, as the sum of the car's price and tax.

**3. Data Cleaning and Preprocessing**

**3.1 Handling Missing Values**

* **Mileage**, **mpg**, and **engineSize** had missing values. We filled them using the **mean** of each column.
* Missing categorical values such as **brand** and **fuelType** were filled with their **mode**.

**3.2 Feature Engineering**

* We created a new feature, **total cost**, which is the sum of **price** and **tax**. The original **price** and **tax** columns were dropped.

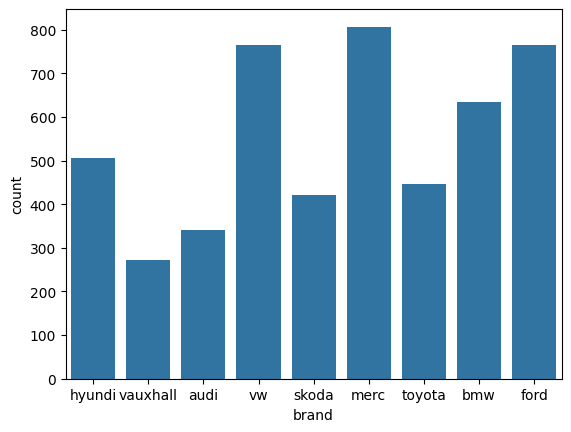
**3.3 Data Transformation**

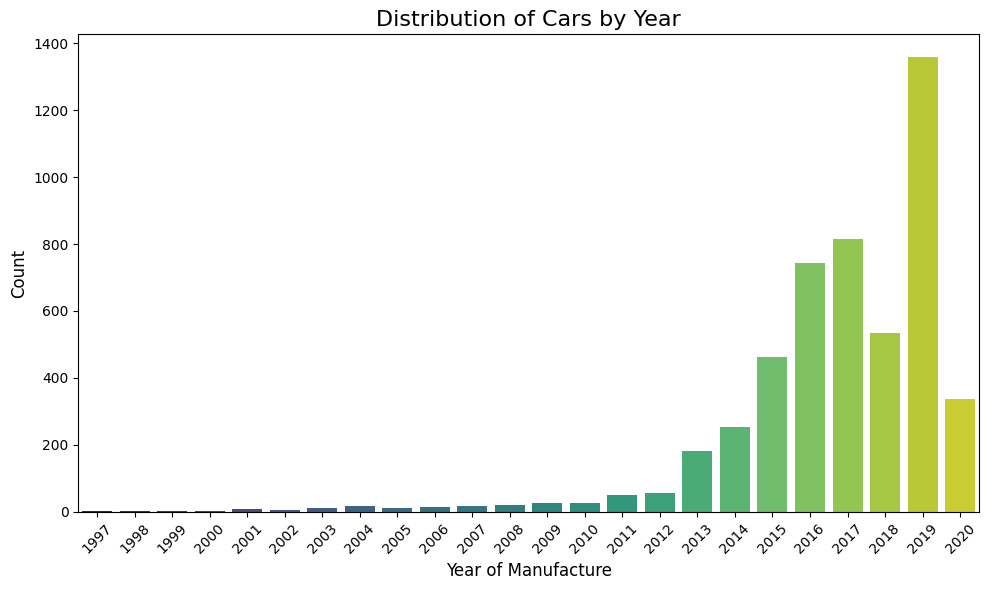
* **Categorical Encoding**: Categorical features like **brand**, **model**, **transmission**, and **fuelType** were label-encoded for machine learning models.
* **Scaling**: Continuous variables such as **mileage**, **engineSize**, **mpg**, and **year** were standardized using StandardScaler to improve model performance.

**4. Data Visualization and Insights**

**4.1 Brand Distribution**

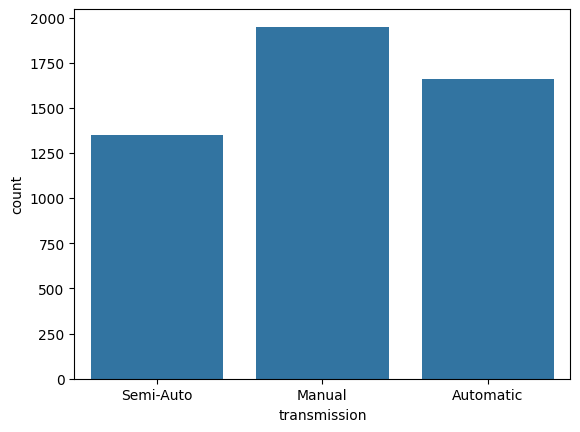
* The most common car brands in the dataset were visualized using a **countplot**, showing the distribution of cars across different brands.

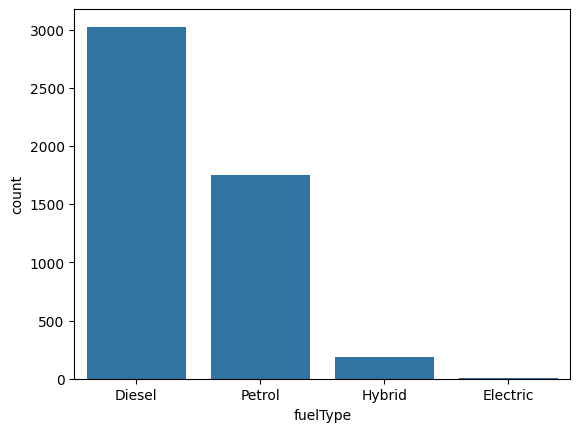
** 4.2 Year of Manufacture**

* A **bar chart** was created to show the distribution of cars by their manufacturing year. Most cars in the dataset were manufactured in recent
* years, with the number of cars increasing steadily over time. ****

**4.3 Transmission and Fuel Type**

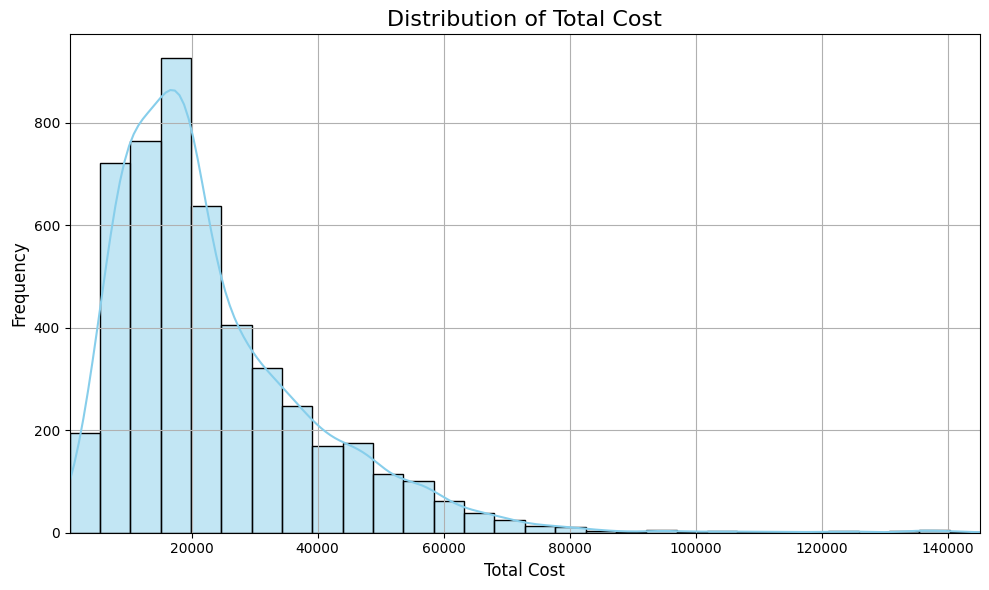
* The dataset contains various types of transmission and fuel types. We used **count plots** to show the distribution of cars based on these categories.





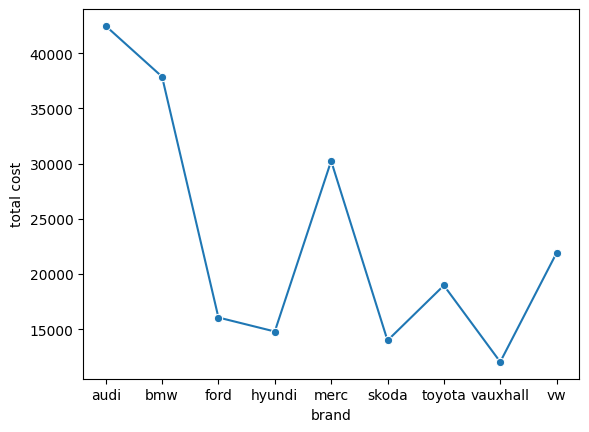
**4.4 Total Cost Distribution**

* A **histogram** of the total cost of cars was plotted, showing the range and frequency of car prices. The majority of cars had total costs within a specific price range.



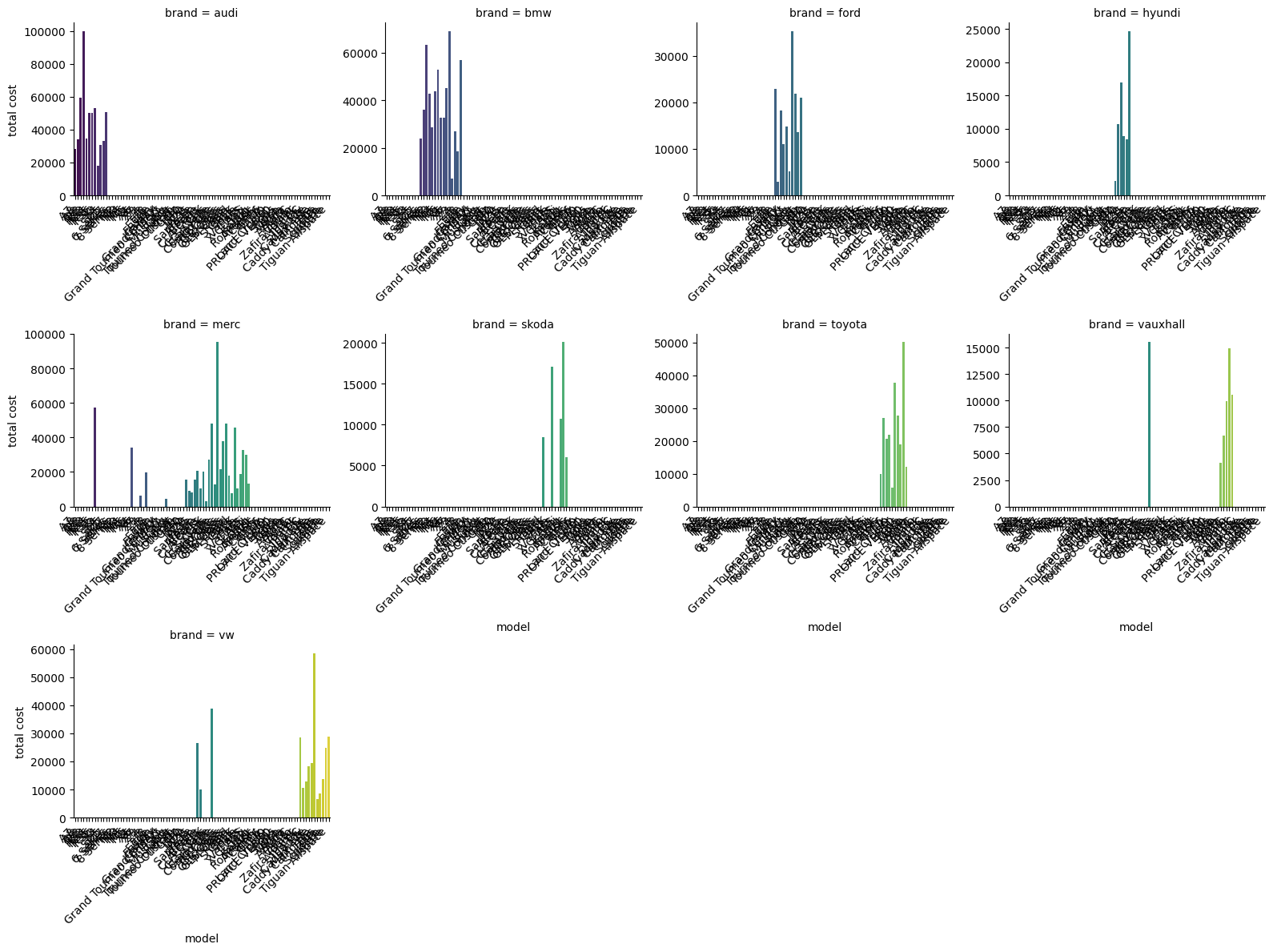
**4.5 Brand and Total Cost**

* A **line plot** was used to show the average total cost per car brand. This helps us understand how different brands vary in terms of pricing.



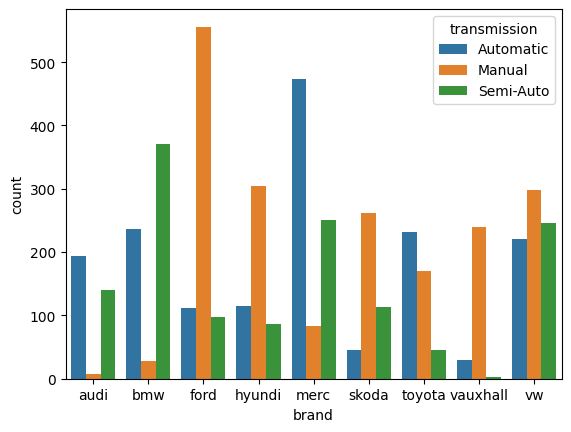
**4.6 Brand, Model, and Total Cost**

* A **facet grid** of bar plots was created, breaking down the average total cost by car model for each brand. This provides more granular insight into how pricing varies across models within the same brand.



**4.7 Brand and Transmission**

* We also analyzed the relationship between **brand** and **transmission** type by plotting a **bar chart** showing the count of each transmission type for each brand.



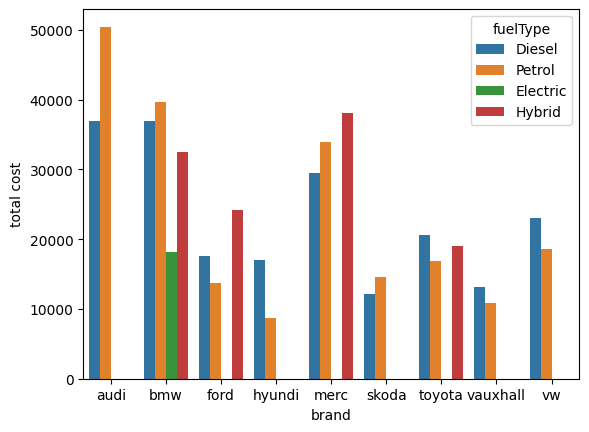
**4.8 Fuel Type, Brand, and Total Cost**

* A key analysis in this project was to examine how different **fuel types** affect the **total cost** across different car brands.

Using a **bar plot**, we visualized the relationship between **fuelType**, **brand**, and **total cost**. The following insights were observed:

* + **Diesel** and **Petrol** were the most common fuel types across brands.
  + Cars running on **Diesel** generally had higher average total costs compared to **Petrol** cars for most brands.
  + Some brands showed more variation in total cost across different fuel types, indicating that the choice of fuel type could have a significant impact on the total cost depending on the brand.

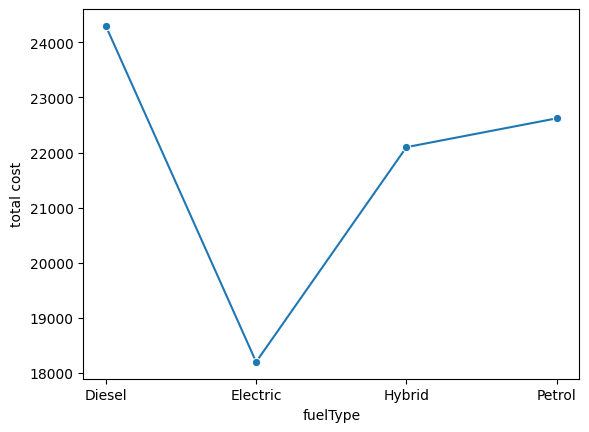
This analysis helps in understanding how fuel type influences car pricing, providing buyers with insights on the expected cost based on their preferred fuel type.



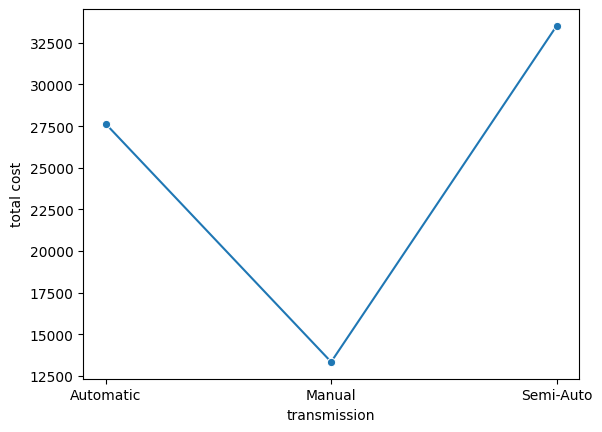
**4.9 Total Cost and Fuel Type**

* We explored the direct relationship between **fuel type** and **total cost** using a **line plot**. Key insights include:
  + Cars running on **Diesel** had a higher average total cost compared to **Petrol**, **Hybrid**, and **Electric** cars.
  + **Hybrid** and **Electric** cars had more moderate total costs, with **Electric** cars tending to have lower costs on average.

This analysis shows that fuel type plays a significant role in determining the total cost of a car.



**4.10 Total Cost and Transmission**

* The relationship between **transmission** type and **total cost** was analyzed using a **line plot**. Insights from this analysis include:
  + Cars with **Automatic** transmission generally had higher total costs compared to **Manual** cars.
  + **Semi-Automatic** transmission cars also tended to have higher total costs, reflecting the technological complexity and consumer demand for these types of cars.
* 

**5. Machine Learning Model: Linear Regression**

**5.1 Feature Selection**

* We applied SelectKBest using **f\_regression** to select the top 8 features most relevant to predicting the target variable (**total cost**).

**5.2 Train-Test Split**

* The dataset was split into training and testing sets with an 80/20 split, ensuring that the model could be evaluated on unseen data.

**5.3 Model Training**

* A **Linear Regression** model was trained on the selected features. Linear regression was chosen due to the continuous nature of the target variable.

**5.4 Model Performance**

* The model was evaluated using the following metrics:
  + **Mean Absolute Error (MAE)**: This measures the average difference between the predicted and actual values. The MAE for this model was computed as:

MAE = mean\_absolute\_error(pred, y\_test)

* + **R-squared (R²)**: This explains how well the model fits the data. An R² score closer to 1 indicates a better fit. The R² for this model was:

r2 = r2\_score(y\_test, pred)

**6. Conclusion**

* **Fuel type** plays a significant role in determining the total cost of a car. Cars that use **diesel** tend to have the highest prices despite the higher cost of diesel fuel itself. On the other hand, **electric** cars, while cheaper for consumers and environmentally friendly, are still not as widely adopted, leading to lower sales. Therefore, it is crucial to **increase awareness of the benefits and advantages of electric cars** to promote their usage and highlight their long-term cost-saving potential.
* **The car model** has a greater impact on the price than the brand itself. The most sold models are from **2019**, indicating a high demand for cars of that year, likely due to their balance of modern features and affordability.
* **Transmission type** also affects car sales through pricing. Although **automatic** cars are easier to drive, **manual** cars tend to sell more because they are generally **cheaper**. This reflects the demand for more affordable vehicles, particularly among cost-conscious consumers.
* Interestingly, **the most sold brands** are not necessarily the most expensive ones. **Mid-range** and **lower-priced brands**, especially those using **manual transmission** and **gas**, tend to be the most popular. This further emphasizes the demand for affordable vehicles in the market.

In summary, the analysis reveals that fuel type, model year, and transmission significantly affect car prices and sales, and **raising awareness of electric vehicles** is essential for promoting more sustainable transportation options.