Exploratory Data Analysis on the automobiles Dataset

Report

Introduction

The goal of this analysis is to perform **exploratory data analysis (EDA)** on the automobile's dataset. EDA allows us to understand the structure, patterns, and relationships in the dataset. Key questions addressed include:

- What are the distributions of car prices?
- Which cars are the most fuel-efficient or have the largest engines?
- Which manufacturers produce the most models?
- What are the trends in car characteristics such as num-of-doors, engine-size, and fuel-type?

The analysis includes **data cleaning, transformation, and visualisation** to provide meaningful insights.

Data cleaning

Summary of Methods

1. **Removed unnecessary columns**:
    - normalized losses and symboling were dropped because they were not relevant for theanalysis.
2. **Handled duplicates**:
    - Duplicate rows were identified and removed.
3. **Handled missing values**:
    - Cells with "?" were replaced with NaN.
    - Rows containing NaN in critical numeric columns like price were dropped.
4. **Categorical column handling**:
    - Categorical columns were reviewed for unique values.
    - Columns like num-of-doors and num-of-cylinders were mapped to integers.
5. **Data type conversion**:
    - price was converted to int64.
    - Other numeric columns were converted from string/text to numeric types.
6. **Feature engineering**:

- car_price_groupings was created to categorize cars into **Low-End**, **Medium**, and **Luxury** groups.

Missing data

- The dataset initially contained **missing values represented as "?"**.
- After replacement and removal:
  - **All critical numeric columns** (like price, num-of-doors, num-of-cylinders) contain no missing values.
  - Remaining categorical variables were cleaned by mapping or encoding.

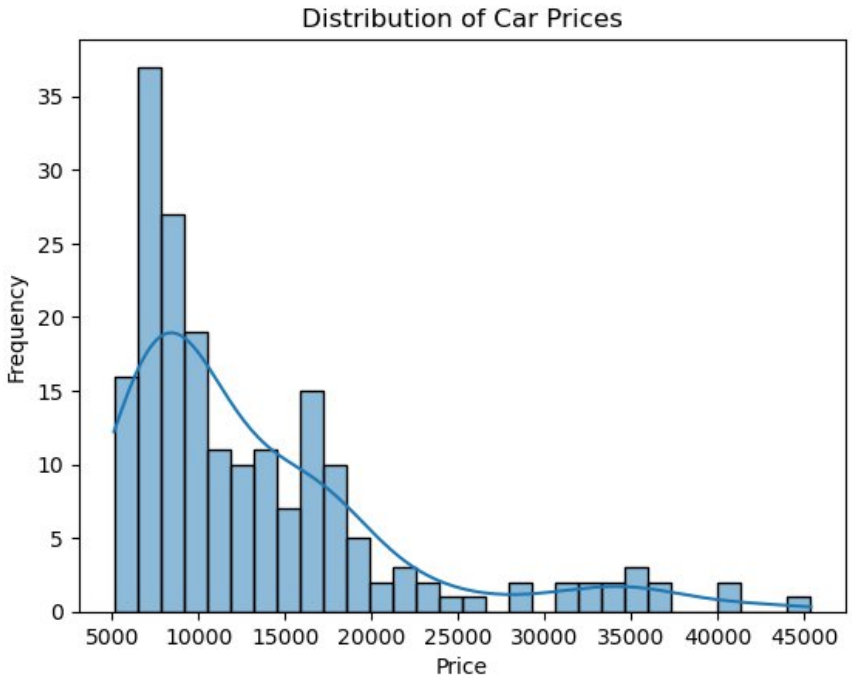
Handling Method:

- Replaced "?" with np.nan.
- Dropped rows with missing values in important columns to ensure analysis accuracy.

Data stories and visualisations

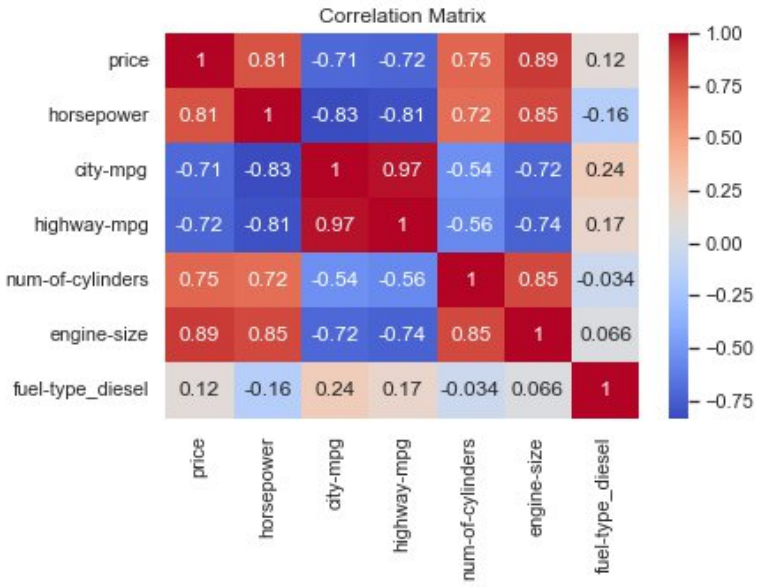
Price Distribution

Based on the charts provided, car prices are not evenly distributed. The "Distribution of Car Prices" chart shows a right-skewed distribution, meaning most cars are clustered at the lower end of the price range. The highest frequency of cars falls between $5,000 and $10,000, with a sharp drop-off in frequency as the price increases. This indicates that there are many more affordable cars on the market than expensive ones. The "Top 5 Cheapest Cars" table supports this, listing five cars all priced under $5,400, while the "Top 5 Expensive Cars" table shows a price range starting from $36,880.



Relationship between price and other KPIs

Based on the correlation matrix, the three most important KPIs to measure for car pricing, besides city-mpg, are **horsepower**, **engine-size**, and **highway-mpg**.



- **Horsepower**: Horsepower has a very strong positive correlation with price (0.81). This means that as horsepower increases, the price of the car tends to increase significantly.
- **Engine-Size**: Engine-size also has a strong positive correlation with price (0.89). This is the strongest correlation with price shown in the matrix, indicating that a larger engine size is a powerful predictor of a higher car price.
- **Highway-MPG**: Highway-mpg has a strong negative correlation with price (-0.72). This suggests that as highway fuel efficiency (miles per gallon) increases, the price of the car tends to decrease. This makes sense as more expensive cars often have larger, less fuel-efficient engines.

More expensive cars are generally less fuel-efficient in the city, while cheaper cars are more fuel-efficient. The chart demonstrates a clear inverse relationship, showing that as a car's **City MPG** increases, its average price decreases. The correlation matrix supports this, indicating a strong negative correlation of -0.71 between **price** and **city-mpg**.

Top 5 Most expensive cars compared to Top 5 Cheapest

High-priced luxury cars like Mercedes-Benz, BMW, and Porsche have a low fuel efficiency, with City MPG ratings of 14 to 17. The provided chart further illustrates this inverse relationship, showing that as a car's City MPG increases, its mean price drops significantly.

Top 5 Expensive Cars:

	make	body-style	car_price_groupings	price	city-mpg	highway-mpg
74	mercedes-benz	hardtop	Luxury	45400	14	16
16	bmw	sedan	Luxury	41315	16	22
73	mercedes-benz	sedan	Luxury	40960	14	16
128	porsche	convertible	Luxury	37028	17	25
17	bmw	sedan	Luxury	36880	15	20

The **Top 5 Cheapest Cars** are all low-end vehicles from makes like Subaru, Chevrolet, Mazda, Toyota, and Mitsubishi. These cars are highly fuel-efficient, with City MPG ratings ranging from 30 to 47. In contrast, the **Top 5 Expensive Cars** are luxury brands such as Mercedes-Benz, BMW, and Porsche, and they are significantly less fuel-efficient, with City MPG ratings ranging from 14 to 17.

Top 5 Cheapest Cars:

	make	body-style	car_price_groupings	price	city-mpg	highway-mpg
138	subaru	hatchback	Low-End	5118	31	36
18	chevrolet	hatchback	Low-End	5151	47	53
50	mazda	hatchback	Low-End	5195	30	31
150	toyota	hatchback	Low-End	5348	35	39
76	mitsubishi	hatchback	Low-End	5389	37	41

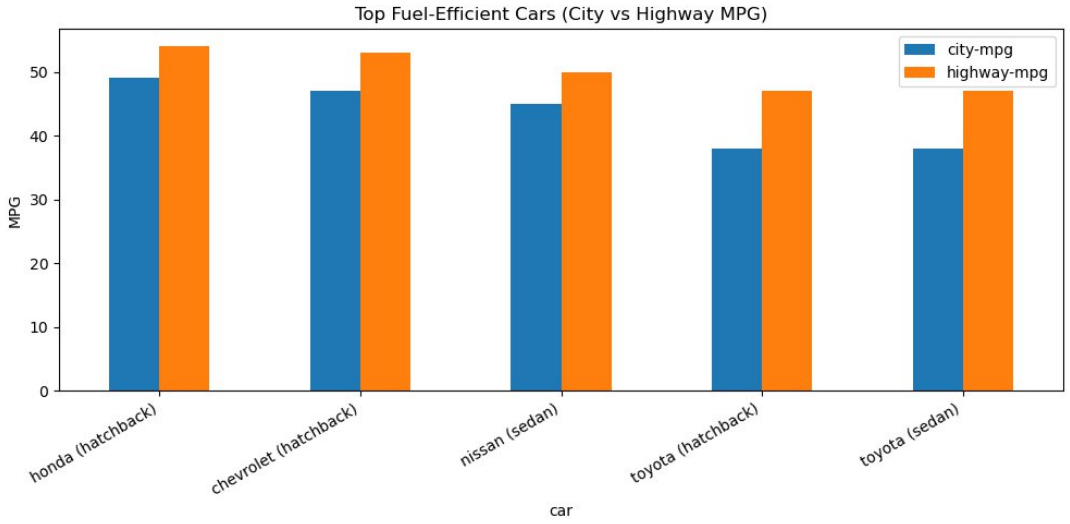
The cheapest cars are predominantly **hatchbacks**, while the most expensive cars are a mix of **sedans**, **hardtops**, and **convertibles**.

Manufacturers with most fuel-efficient cars

The bar chart shows the top 5 most fuel-efficient cars, comparing their City and Highway MPG.

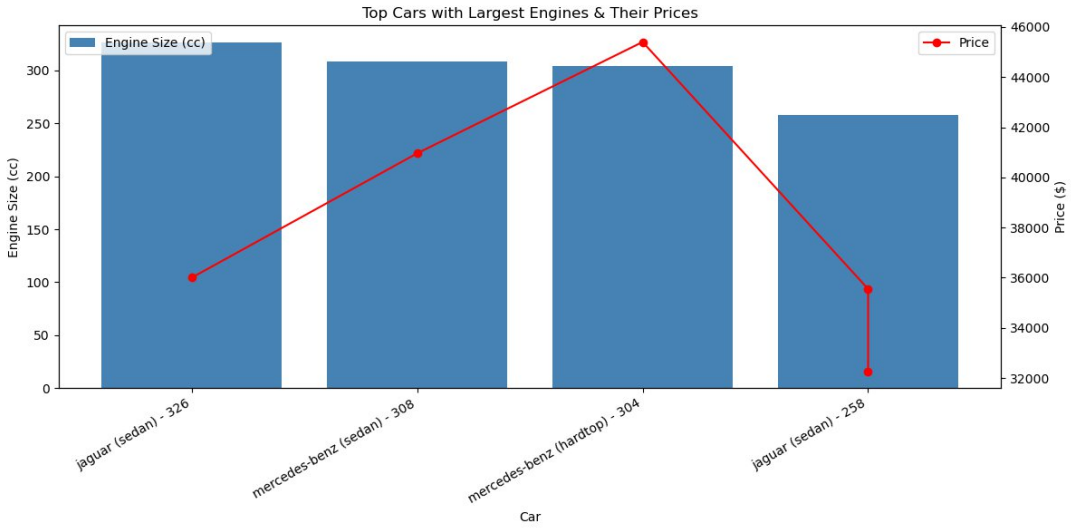
- The most fuel-efficient cars are predominantly hatchbacks from brands like Honda, Chevrolet, and Toyota.
- Highway MPG is consistently higher than City MPG for all cars shown, which is typical as steady speeds on highways require less engine work than you would experience in the city, with constant stop and go.
- The Honda (hatchback) and Chevrolet (hatchback) are the top two most fuel-efficient cars, with both achieving a City MPG of around 45-50 and Highway MPG exceeding 50.

- The Nissan (sedan) and Toyota (sedan), despite being sedans, are also among the top fuel-efficient vehicles.



Vehicles with largest Engine Capacity

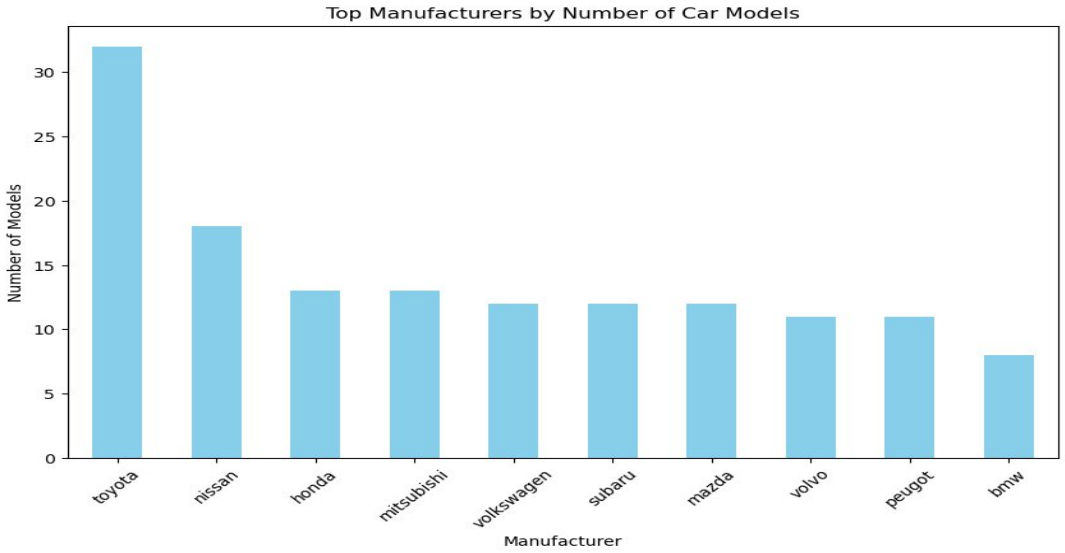
The "Top Cars with Largest Engines & Their Prices" chart shows a clear relationship between engine size and a car's price. The cars with the largest engines (Jaguar sedan-326 and Mercedes-Benz sedan-308) are also the most expensive. The Jaguar sedan (engine size 326) and the Mercedes-Benz (engine size 304) have the highest prices, while the Jaguar (engine size 258) and Mercedes-Benz (engine size 308) have the lowest.



Manufactures with most car models

Based on the graph below, Toyota is the leading car manufacturer in terms of the number of models offered, with over 30 different models. The chart "Top Manufacturers by Number of Car Models" shows Toyota well ahead of its

competitors, such as Nissan, Honda, and Mitsubishi. This suggests Toyota has a wide variety of vehicles to appeal to different customers, including both fuel-efficient and expensive models. The "Mean Price by City MPG and Price Group" chart reinforces this by showing that as a car's City MPG increases, its mean price decreases, indicating that manufacturers like Toyota, which offer many models, can cater to both ends of the price spectrum.



Also, interesting to note that Luxury car manufacturers, such as **Mercedes-Benz, BMW, and Porsche**, tend to produce fewer models compared to high-volume, mainstream manufacturers like **Toyota**, Nissan, and Honda. For example, the chart "Top Manufacturers by Number of Car Models" clearly shows Toyota with the highest number of models, while BMW has a much smaller number.

Summary

Expensive luxury cars have larger engines, higher horsepower, and lower fuel efficiency, while cheaper cars are the opposite. The market is dominated by affordable, fuel-efficient vehicles from manufacturers like Toyota, which produce a wide variety of models. In contrast, luxury brands focus on fewer, high-priced models with less emphasis on fuel economy.

This report was written by: Juleiga Regal