Exploratory Data Analysis on the Automobile Dataset

Report

Introduction

The Automobile dataset contains various attributes of vehicles, including engine specifications, fuel efficiency, price, body-style, and manufacturer details. The purpose of this exploration is to clean the data, identify insights, understand patterns, and determine how different features influence car price and performance.

Description of the Dataset

The dataset includes both **numerical** and **categorical** features such as:

- **Numerical:** engine-size, horsepower, city-mpg, highway-mpg, price, curb-weight
- **Categorical:** make, fuel-type, body-style, drive-wheels, num-of-doors

It helps to analyze:

- Price based on engine size, horsepower, fuel type, number of doors
- Most fuel-efficient manufacturers
- Most expensive and cheapest cars
- Which cars have the biggest engines
- Distribution of body styles

Data cleaning

Steps Taken:

1. **Removed unnecessary columns:**
    - normalized-losses, symbolling
    - These were not useful for this analysis.
2. **Removed duplicate rows** to ensure clean and accurate data.
3. **Removed rows with missing values** using dropna(), and confirmed no missing values remained.
4. **Replaced '?' with NaN**, converted object-type numeric values to numeric, and filled missing data with **median**, which is more robust than mean.
5. **Converted all numeric columns to int64** for consistency.

Missing data

Before performing this analysis, several steps were taken to inspect and prepare the dataset for accuracy, completeness, and usability.

To check for missing values, I used the following methods:

- automobiles_df.isnull().sum()
- automobiles_df.isnull().any(axis=1).sum()

Results:

- Total rows with missing values: 0
- Rows before cleaning: 205
- Rows after cleaning: 205
- Total rows removed due to missing data: 0

The detailed output also showed that every column has 0 missing values, including key features such as:

- price, horsepower, city-mpg, highway-mpg, engine-size, bore, stroke, compression-ratio, curb-weight, etc.

Since no missing data was present in any column, no data imputation (mean/median), no replacement was needed.
Therefore, the dataset is complete and ready for analysis without any missing value treatment.

Data stories and visualizations

1. Relationship between Numerical Features

We used a correlation heatmap to identify how features relate to each other.

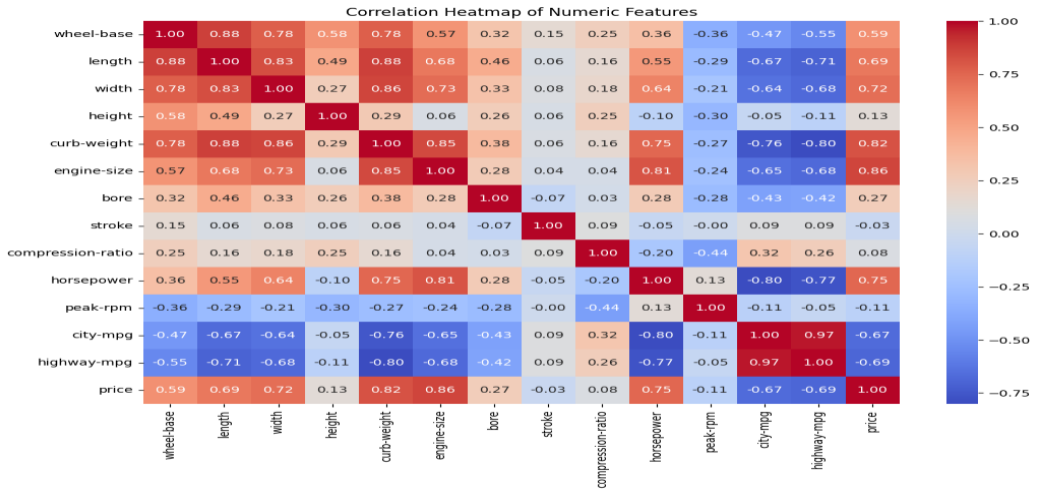

Figure 1: Correlation Heatmap

Key Insights:

- Strong positive correlation between **engine-size and price**.
- Horsepower also has a high correlation with price.
- City-mpg and highway-mpg are strongly positively correlated.

2. Engine Size vs Price & Horsepower vs Price

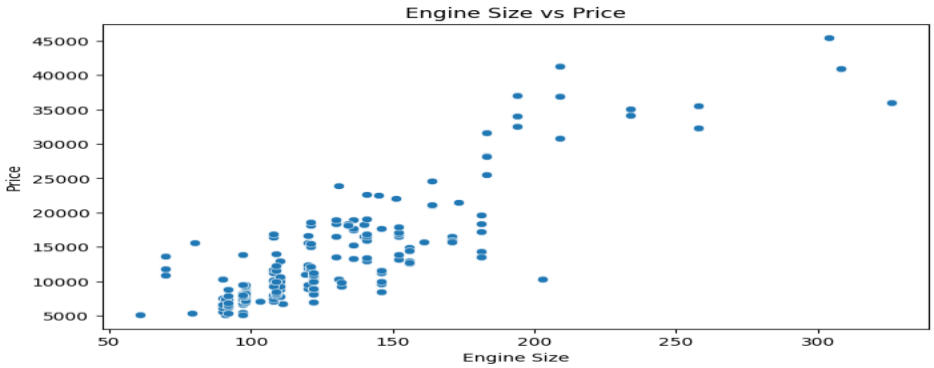

Figure 2: Engine Size vs Price Scatter Plot

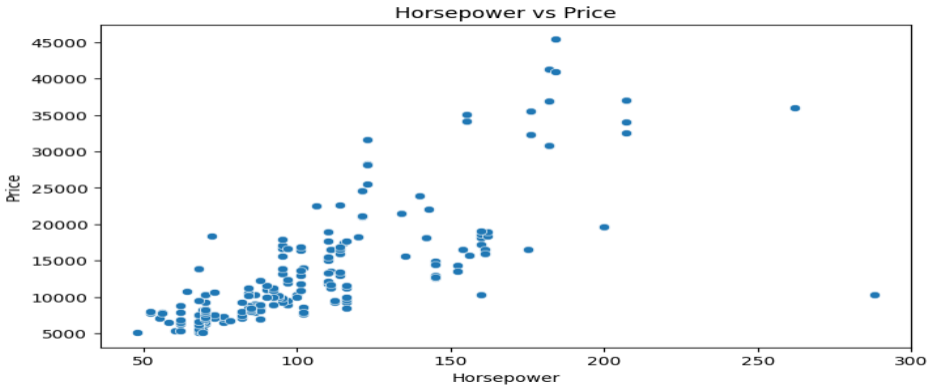

Figure 3: Horsepower vs Price Scatter Plot

Observations:

- Cars with larger engine sizes and higher horsepower are more expensive.
- High-performance cars command premium prices.

3. Price Distribution by Categorical Features

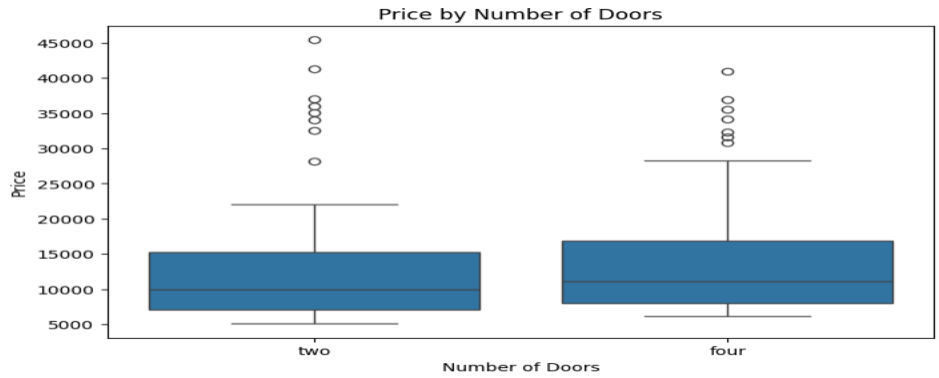

Figure 4: Price by Number of Doors Boxplot

Figure 5: Price by Drive-Wheels Boxplot

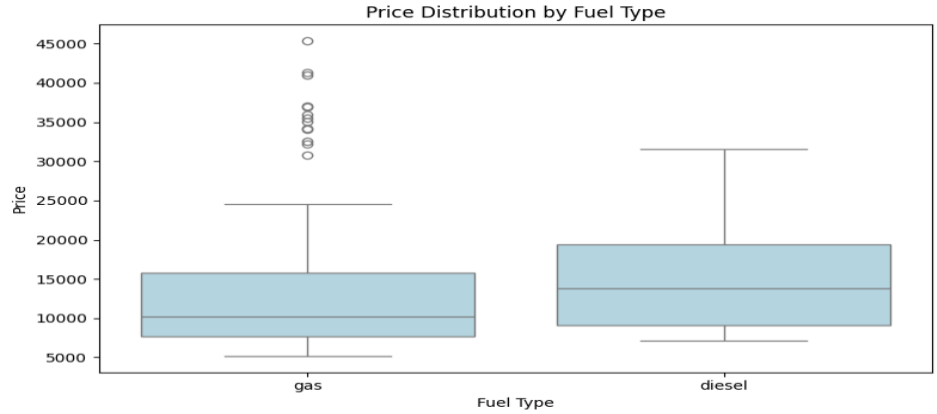

Figure 6: Price by Fuel Type Boxplot

Findings:

- **Two-door cars are generally more expensive**, often sporty or luxury.
- **Rear-wheel drive cars (RWD) show higher price variation**-common in premium cars.
- **Diesel cars tend to have higher prices** than gasoline cars.

4. Engine Size vs Horsepower

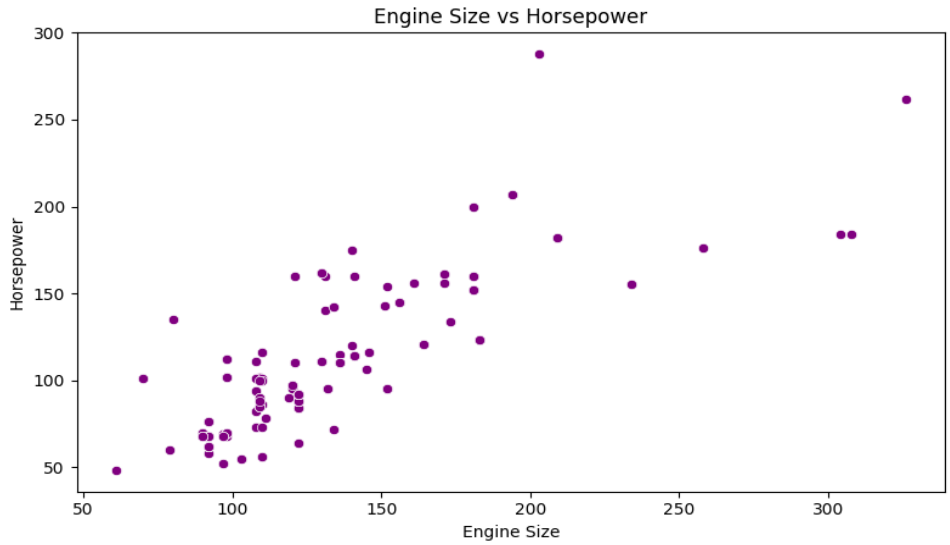

Figure 7: Engine Size vs Horsepower Scatter Plot

Insight: Larger engine capacity results in higher horsepower. Clear positive relationship.

5. Fuel Type: City MPG vs Highway MPG

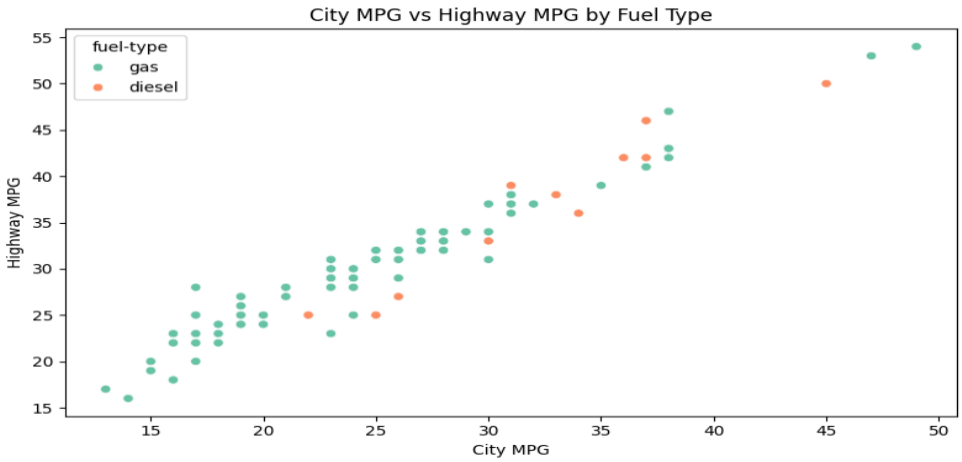

Figure 8: City MPG vs Highway MPG Scatter

Observation:

- Diesel cars are more fuel efficient.
- City and highway MPG are strongly correlated for both fuel types.

6. Distribution of Cars by Body Style

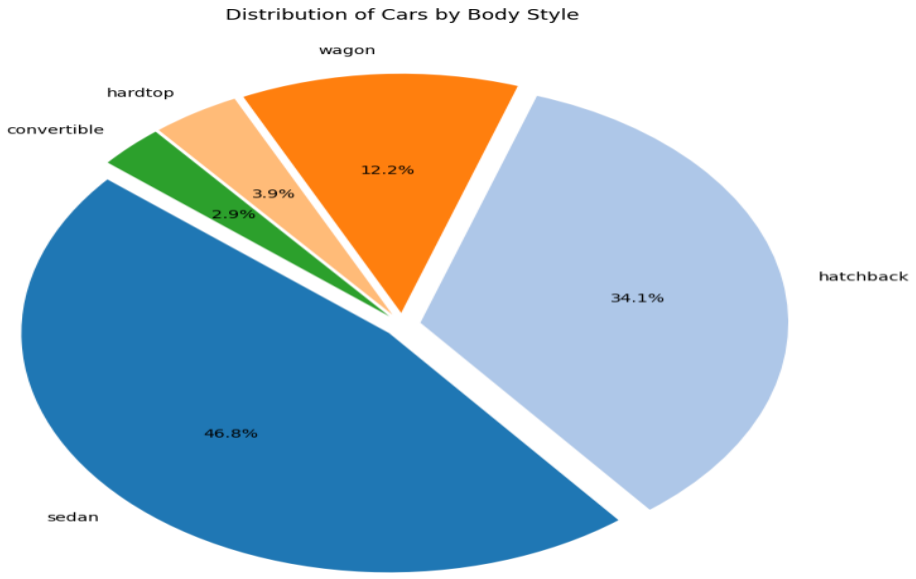

Figure 9: Pie Chart of Body-Style Distribution

Observation:

- Sedans and hatchbacks are the most common.
- Convertibles and hardtops are the least common.

7. Most Expensive vs Cheapest Cars

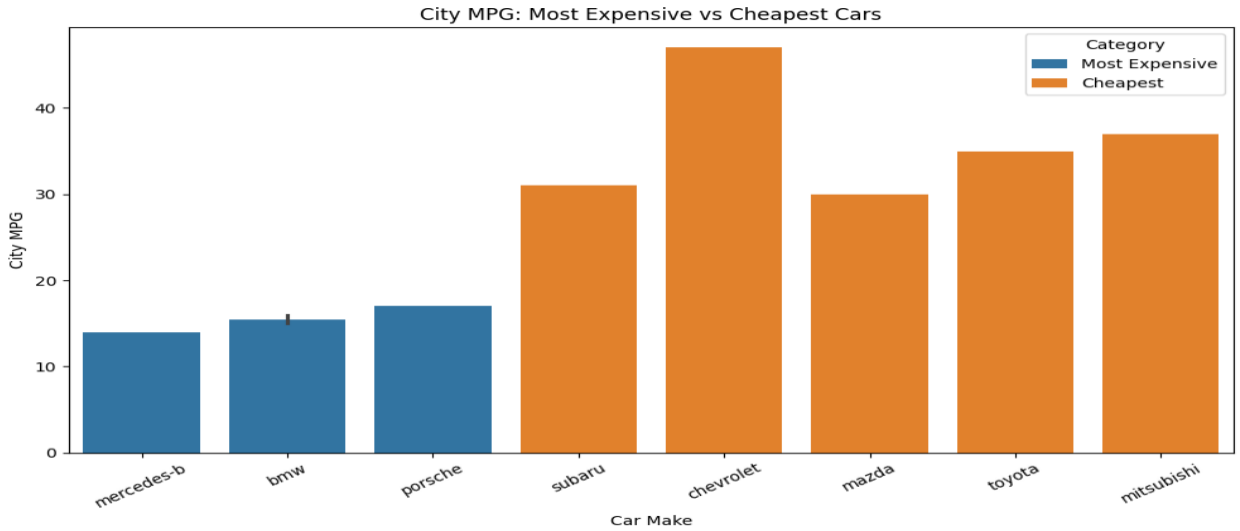

Figure 10: Bar Plots: City MPG Comparisons

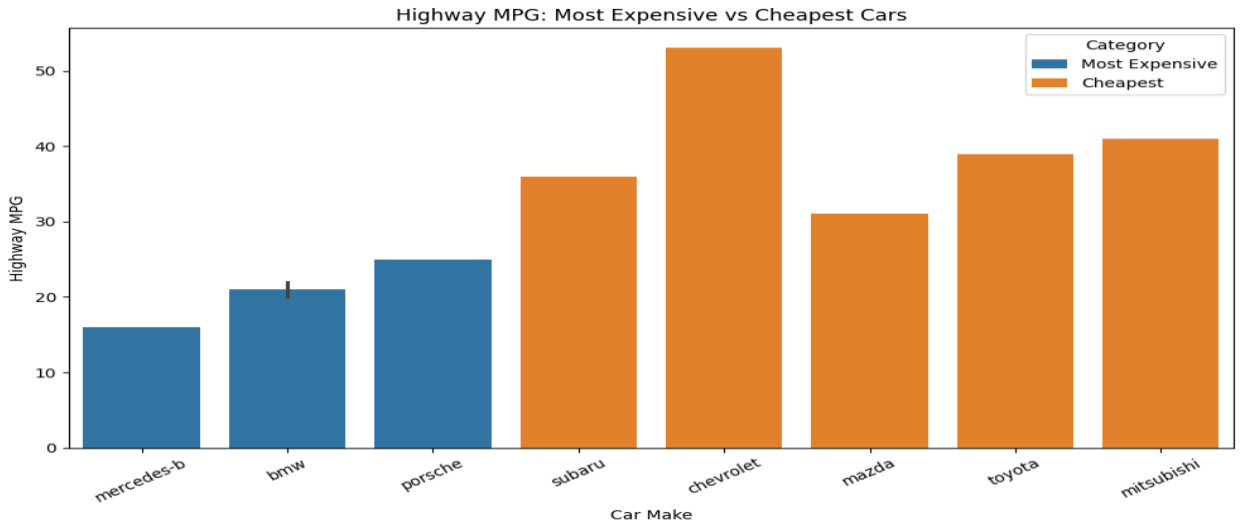

Figure 11: Bar Plots: Highway MPG Comparisons

Insights:

- Expensive cars focus on **luxury, performance, and brand prestige**, but have **poor fuel economy**.
- Cheaper cars are **fuel-efficient and budget-friendly**.

8. Most Fuel-Efficient Manufacturers

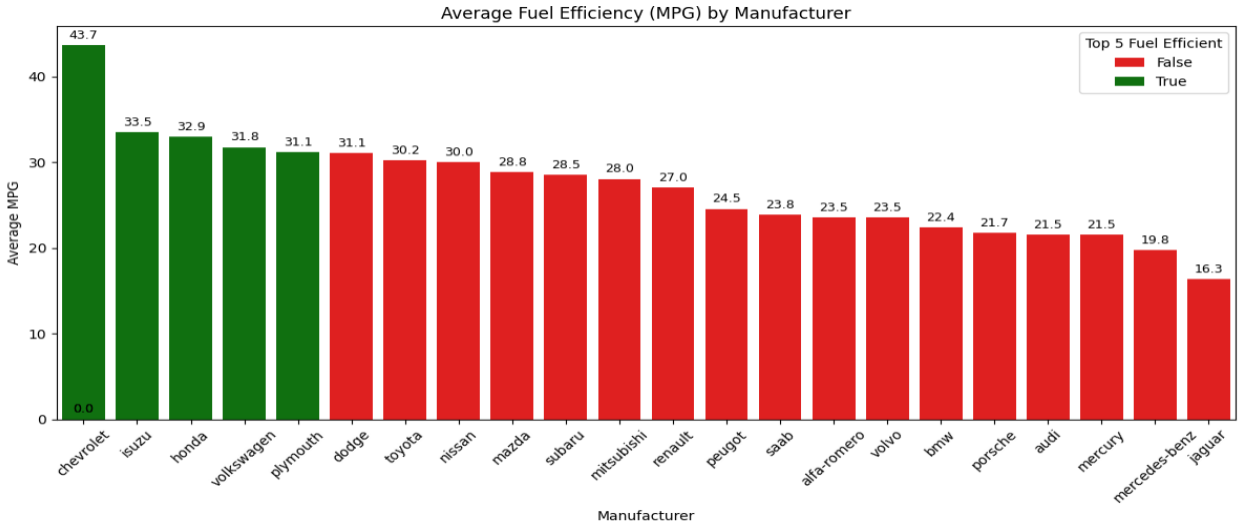

Figure 12: Bar Chart of Average MPG by Manufacturer

Top 5 Fuel-Efficient Manufacturers:

- Identified using average MPG (city + highway).
- These brands offer best cost-efficiency and eco-friendliness.

9. Cars with the Largest Engines

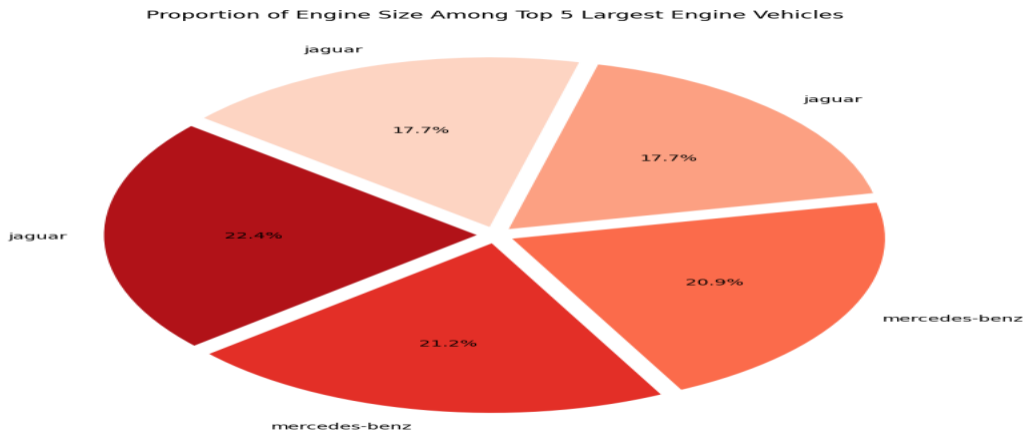

Figure 13: Table & Pie Chart of Top 5 Largest Engine Cars

Observation:

- Brands like **Jaguar, Mercedes-Benz, and Porsche** lead in engine capacity.
- Larger engines = higher price & horsepower, but poorer fuel efficiency.

10. Manufacturer with the Most Models

- Number of car models per manufacturer:

Make	Models
Toyota	32
Nissan	18
Mazda	17
Mitsubishi	13
Honda	13

Observation:

- The manufacturer with the most models in the dataset is Toyota with 32 models.
- Indicates brand diversity and market reach.

Conclusion

This EDA provided valuable insights into automotive pricing, performance, fuel efficiency, and brand characteristics.

- Engine size and horsepower are major price predictors.
- Fuel efficiency is highest in compact, affordable vehicles.
- Luxury brands dominate power, prestige, and engine size.
- Diesel vehicles offer better MPG than gas vehicles.
- Some manufacturers offer broad variety, while luxury brands focus on performance.

This report was written by: **Senzo Nelson Ncekana**