Exploratory Data Analysis on the Automobiles Dataset

Report

Date: January 2026

---

1. Introduction: Description of the Dataset

The dataset selected for this analysis is the "Automobiles" dataset. It consists of technical specifications for various vehicles, including attributes such as physical dimensions (length, width, height), engine properties (engine type, size, horsepower), and performance metrics (city and highway MPG, price).

The objective of this analysis is to sanitise the raw data and identify key trends, specifically identifying which manufacturers prioritise fuel efficiency versus engine performance, and assessing the variety of models available in the market.

2. Data Cleaning

The raw data contained several formatting inconsistencies that required cleaning before analysis could begin. Using Python (Pandas), the following steps were taken:

- Column Removal: The columns `normalized-losses` and `symboling` were removed as they were not relevant to the analysis of vehicle performance and pricing.
- Duplicate Removal: I performed a check for duplicate entries. If any duplicate rows were found, they were removed to ensure statistical accuracy.
- Data Type Conversion: Several numerical columns (including `price`, `horsepower`, and `peak-rpm`) were initially read as objects (text) due to the presence of special characters. I converted these to numeric types (`int64` and `float`) to allow for mathematical aggregation.

3. Missing Data

- Identification: The dataset used the ? character to denote missing values, which prevented standard analysis.
- Handling: I programmatically replaced all instances of ? with standard NaN (Not a Number) markers.
- Resolution: Instead of imputing (guessing) the missing values, I adopted a strict approach and dropped the rows containing missing data. This resulted in a cleaner dataset consisting only of vehicles with complete, verified specifications.

4. Data Stories and Visualisations

Story 1: The Most Expensive Cars

To understand the upper end of the market, I sorted the data by price.

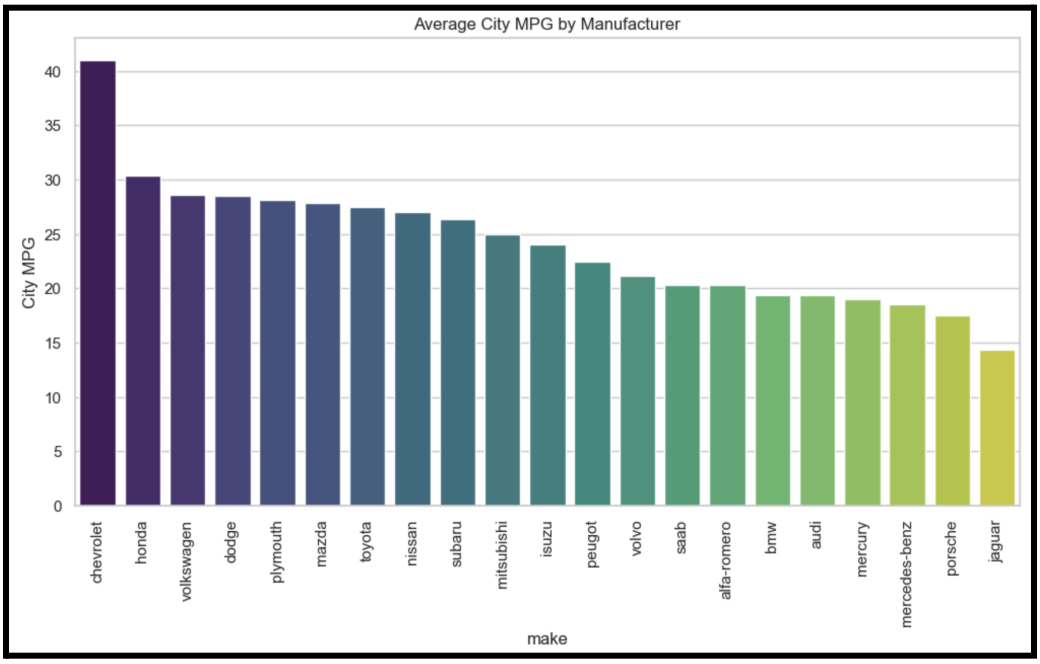
Top 5 Most Expensive Cars:

	make	body-style	price	horsepower
74	mercedes-benz	hardtop	45400	184
16	bmw	sedan	41315	182
73	mercedes-benz	sedan	40960	184
128	porsche	convertible	37028	207
17	bmw	sedan	36880	182

Finding: The analysis reveals that Mercedes-Benz and BMW dominate the high-end market. The most expensive vehicle in the dataset is a Mercedes-Benz Hardtop, priced at over 45,000. This confirms that German luxury manufacturers hold the highest price points in this sample.

Story 2: Fuel Efficiency Leaders

I investigated which manufacturers prioritize economy by calculating the average City MPG (Miles Per Gallon) for each brand.



Finding: The visualization clearly demonstrates a divide in the market. Chevrolet is the market leader in fuel efficiency, followed by other economy brands like Honda and Dodge. Conversely, luxury brands like Jaguar and Mercedes-Benz appear at the bottom of the chart. This highlights a clear trade-off: high-performance luxury vehicles sacrifice fuel economy for power.

Story 3: Engine Capacity Heavyweights

This investigation looked for the vehicles with the largest engines.

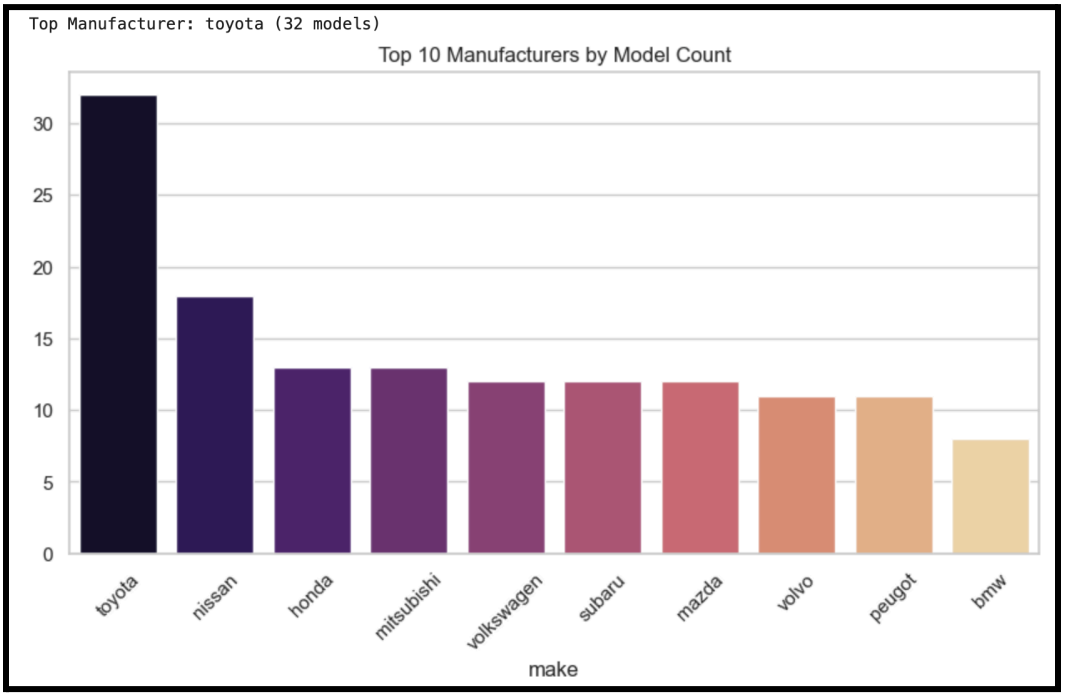
Vehicles with Largest Engines:

	make	engine-size	horsepower
49	jaguar	326	262
73	mercedes-benz	308	184
74	mercedes-benz	304	184
48	jaguar	258	176
47	jaguar	258	176

Finding: Consistent with the fuel efficiency findings, the vehicles with the largest engines (over 300 size) are exclusively from Jaguar and Mercedes-Benz. This corroborates the previous finding that large engine displacement is a key characteristic of the luxury segment in this dataset.

Story 4: Market Dominance (Model Variety)

Finally, I analyzed which manufacturer offers the widest variety of car models.



Finding: Toyota is the dominant manufacturer in this dataset, with over 30 unique models represented. This is significantly higher than competitors like Nissan or Mazda. This suggests that Toyota targets a much broader segment of the market, offering a wider range of trim levels and body styles than niche luxury manufacturers.

---

5. Conclusion

The Exploratory Data Analysis successfully categorized the automotive market into two distinct segments:

1. Economy Segment: Led by Toyota (variety) and Chevrolet (efficiency), focusing on practicality and lower running costs.
2. Luxury Segment: Led by Mercedes-Benz and Jaguar, focusing on high prices and large, powerful engines at the expense of fuel economy.

The data cleaning process was essential, as the initial formatting (missing values masked as ?) would have prevented accurate price and engine analysis.

This report was written by: Gert Bester