Exploratory Data Analysis on the [Automobile.txt] Dataset

Report

Introduction

This Exploratory Data Analysis (EDA) report examines a dataset of automobile characteristics to uncover insights into vehicle pricing, fuel economy, engine capacity, and manufacturer diversity. The dataset, sourced from 'automobile.txt', contains 205 records of vehicles, each described by 26 features, including numerical attributes (e.g., wheelbase, length, width, height, curb-weight, engine-size, compression-ratio, city-mpg, highway-mpg, price, bore, stroke, peak-rpm, horsepower) and categorical attributes (e.g., make, body style, fuel type). After preprocessing - removing duplicates, replacing missing value placeholders ('?') with NaN, dropping rows with missing values, and converting numerical columns to integers the dataset provides a robust foundation for analysis.

The objective of this EDA is to explore relationships between key variables, such as price, fuel efficiency, and engine size, across manufacturers and vehicle types. Through visualizations, including bar plots of the most expensive cars, average fuel economy, and model counts, this report highlights trends, such as which manufacturers offer the most models or the best fuel economy, and evaluates whether high-priced vehicles justify their cost through performance metrics like engine size or MPG. This analysis aims to inform decisions in automotive market studies or consumer purchasing by revealing patterns in vehicle characteristics.

Data cleaning

The automobiles dataset, originally containing 205 records and 26 features, was pre-processed to ensure consistency and suitability for exploratory data analysis. The following cleaning steps were applied to prepare the data for visualizations on pricing, fuel economy, engine capacity, and manufacturer diversity:

• **Column Removal**: Dropped normalized-losses and symboling columns as they were irrelevant to the analysis, reducing the dataset to 24 features.

• **Duplicate Removal**: Checked for duplicate rows with duplicated().sum(), which revealed no duplicates (0 rows). Applied drop_duplicates() to confirm, maintaining all 205 records, verified with shape[0]. The index was reset with reset_index(drop=True) for clean row numbering, ensuring unique vehicle records for model counts and price analyses.

• **Data Type Standardization**: Converted numerical columns (wheelbase, length, width, height, curb-weight, engine-size, compression-ratio, city-mpg, highway-mpg, price, bore, stroke, peak-rpm, horsepower) from object or float64 to int64. This ensured consistent data types for sorting and calculations in visualizations.

Missing data

The automobiles dataset was examined for missing data to ensure reliability. The following steps were taken to identify and handle missing values:

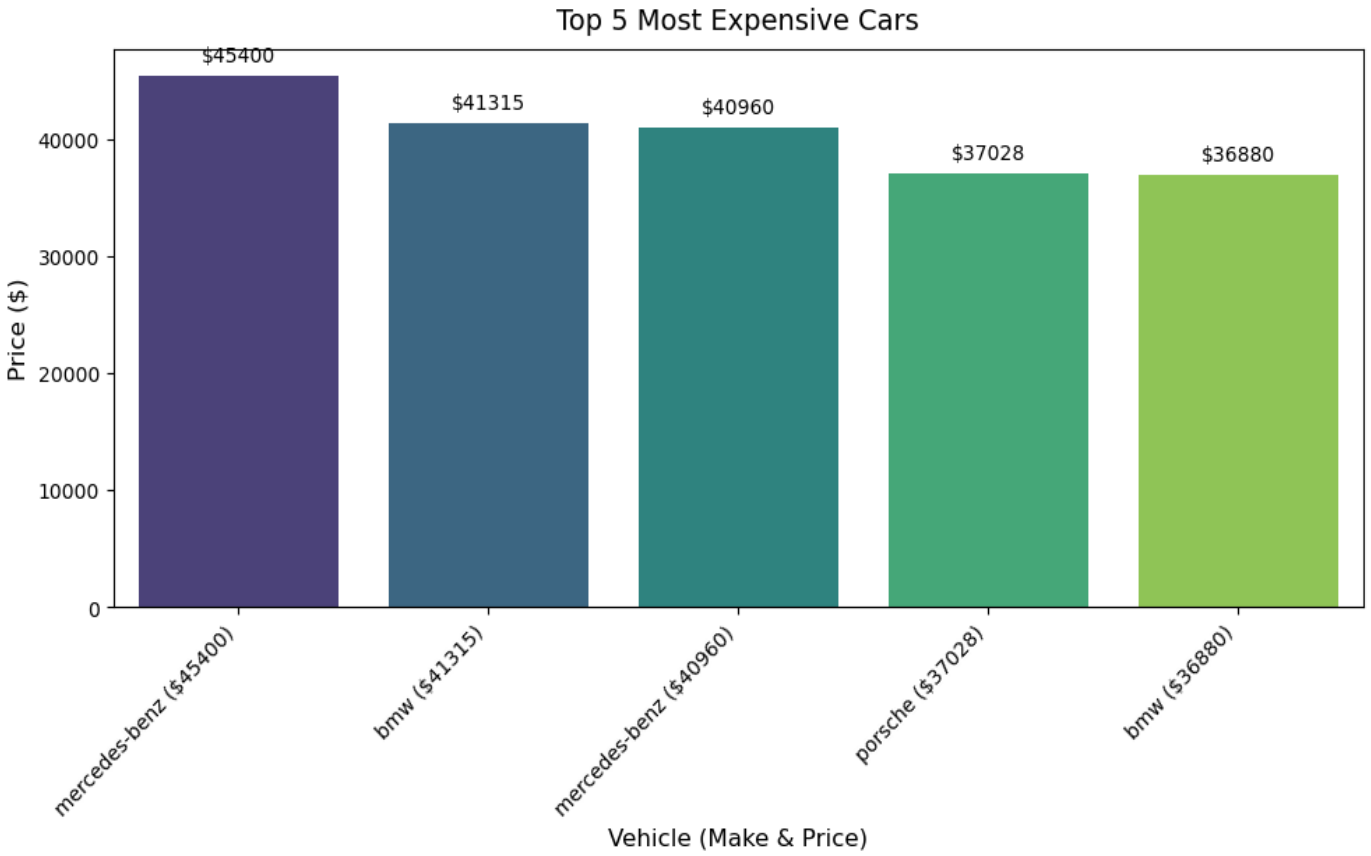
• **Identification of Missing Values**: Inspected the dataset for missing data, focusing on placeholder '?' values in columns. Used automobiles_df.isin(['?']).sum() to detect '?' occurrences and isnull().sum() to check for NaN values, revealing missing data in key analytical columns.

• **Placeholder Replacement**: Replaced all '?' values with NaN using replace('?', np.nan) to standardize missing data, enabling consistent handling.

• **Removal of Missing Data**: Dropped all rows with NaN values. This ensured no missing data in visualizations.

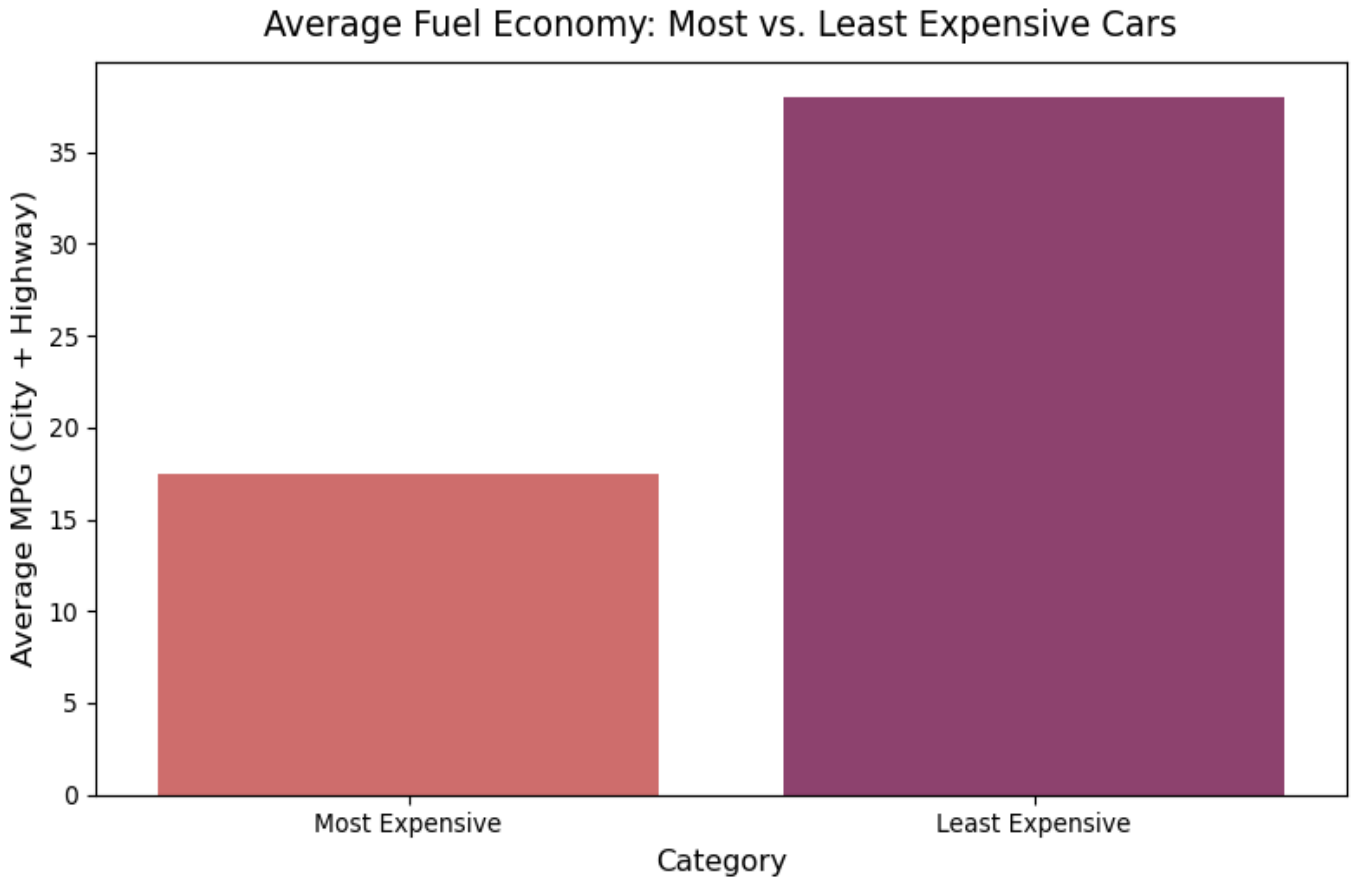
• **Validation**: Verified post-cleaning dataset size with shape[0], confirming that rows with missing records were reduced and isnull().sum() to confirm no remaining NaN in the columns, ensuring data integrity for all analyses. These steps eliminated missing data in columns, preparing the dataset for accurate visualizations and analyses.

Data stories and visualisations



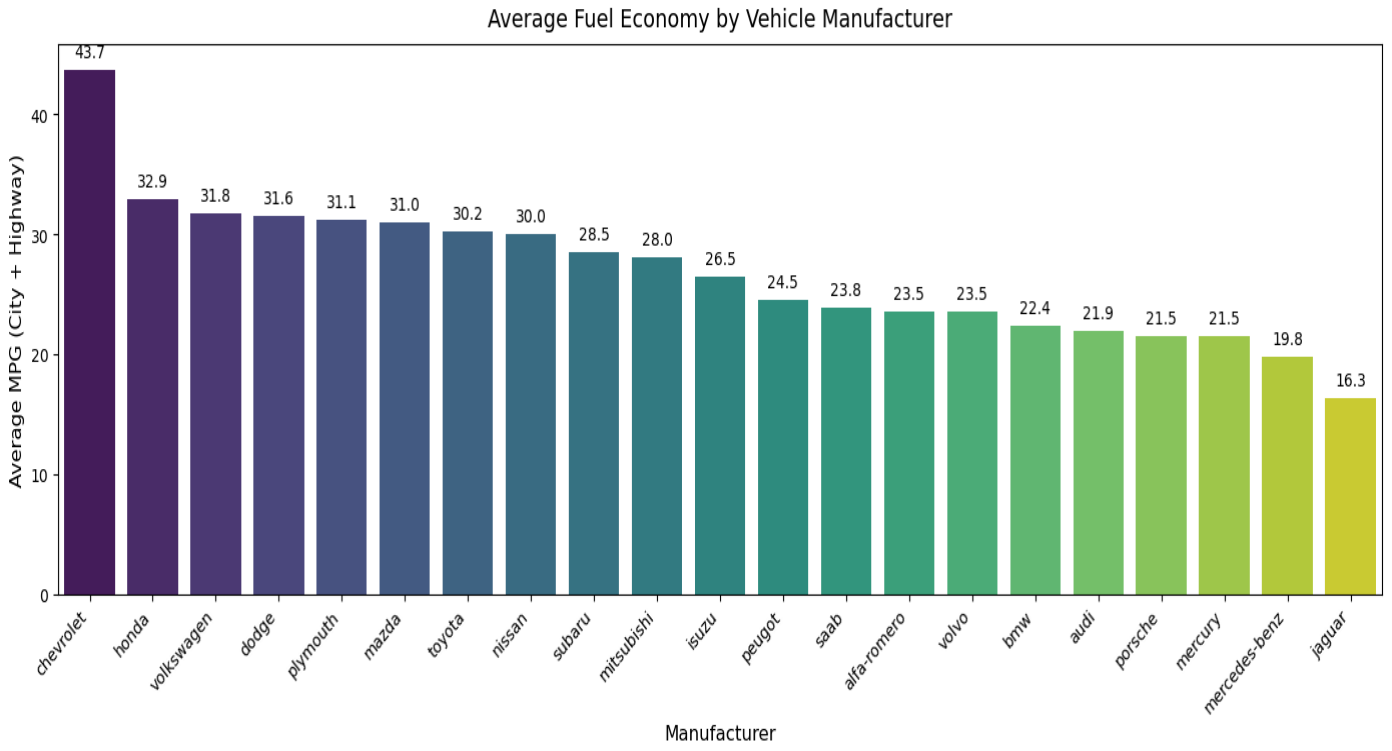
The bar plot above, titled “Top 5 Most Expensive Cars,” showcases the highest-priced vehicles from the automobile dataset. The x-axis lists the top 5 cars, identified by make and model (e.g., Mercedes-Benz, BMW), while the y-axis represents their prices in dollars.

The analysis highlights that the most expensive car is a Mercedes-Benz, priced at $45,400, setting the benchmark for luxury in the dataset. Notably, both Mercedes-Benz and BMW dominate the top 5, each appearing twice. This concentration suggests these premium brands lead in high-end pricing, likely due to advanced features and larger engines.



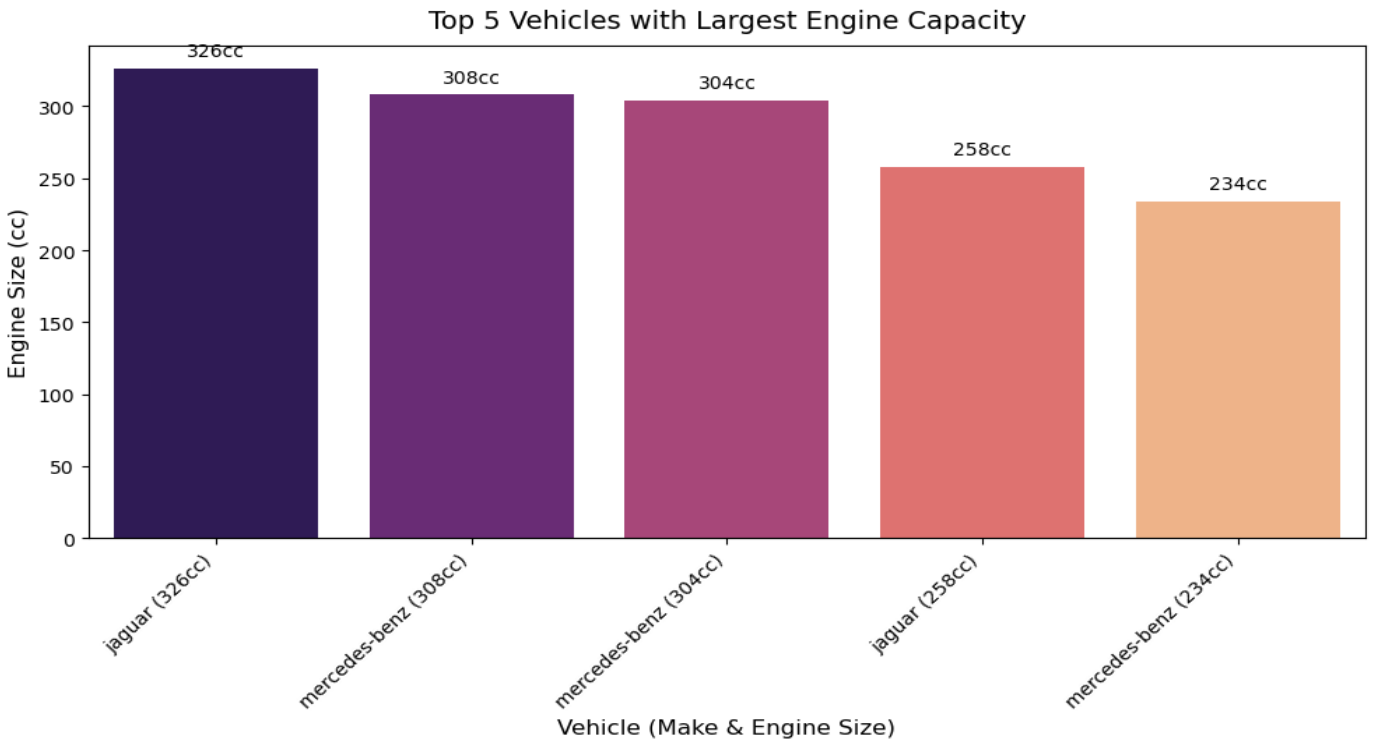
The bar plot “Average Fuel Economy: Most vs. Least Expensive Cars” above compares the average fuel economy, measured in miles per gallon (MPG), between the top 5 most expensive and the bottom 5 least expensive vehicles in the dataset. The x-axis categorizes the data into two groups: “Most Expensive” and “Least Expensive,” representing the top 5 and bottom 5 cars by price, respectively, with each group’s average MPG calculated as the mean of city and highway MPG. The y-axis displays these average MPG values.

The analysis shows that the most expensive cars average a lower MPG (17.50 MPG), reflecting their larger engines. Conversely, the least expensive cars, typically economy models achieve a higher average MPG (38 MPG), indicating better fuel efficiency. This disparity suggests that high-priced vehicles may not justify their cost through fuel economy, as their MPG is notably lower than cheaper alternatives, highlighting a trade-off between luxury and efficiency.



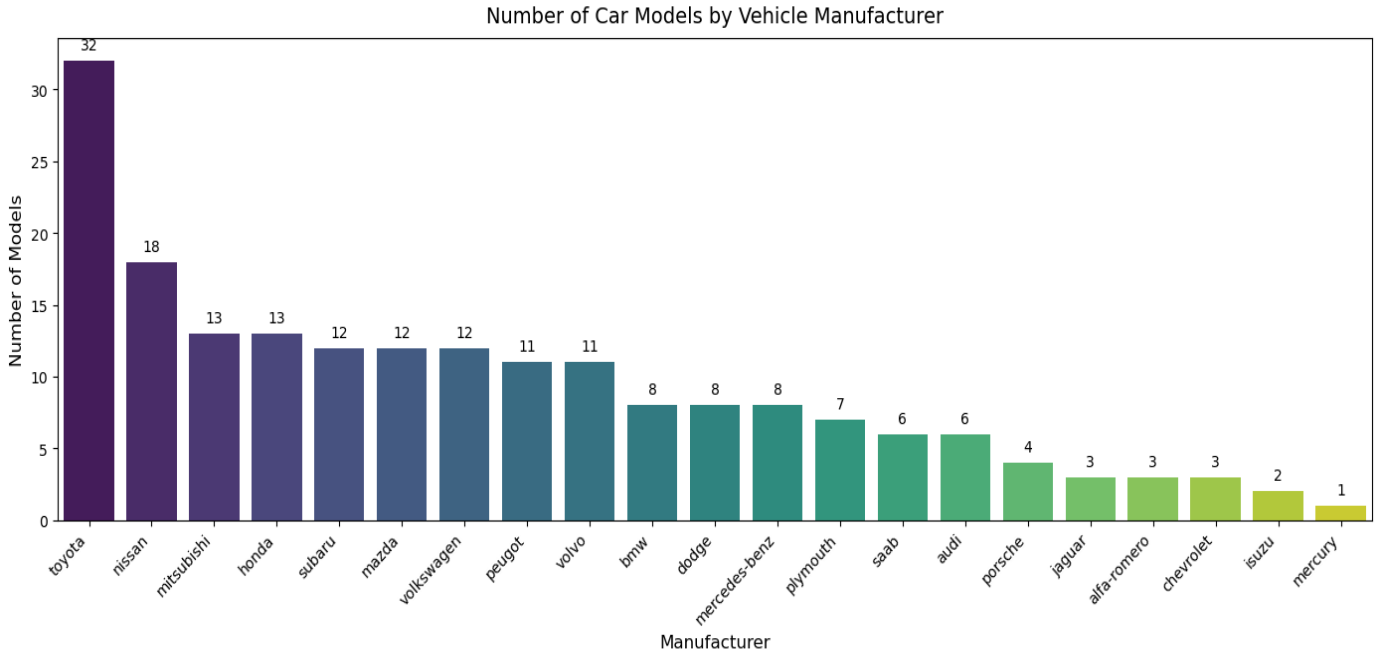
The bar plot “Average Fuel Economy by Manufacturer” above compares the average fuel economy, measured in miles per gallon (MPG), across various vehicle manufacturers in the dataset. The x-axis lists manufacturers in descending order of average MPG, while the y-axis represents the mean MPG, calculated as the average of city and highway values for each brand.

The analysis reveals that Chevrolet leads with the highest average MPG at 43.7, indicating it builds the most fuel-efficient vehicles, likely due to its focus on economy models. In contrast, Jaguar trails with the lowest average at 16.3 MPG, reflecting its emphasis on luxury and performance with larger engines. Other efficient manufacturers include Honda (32.9 MPG) and Volkswagen (31.8 MPG), while luxury brands like Mercedes-Benz (19.8 MPG) and BMW (22.4 MPG) rank lower. This trend suggests that fuel efficiency is strongest among economy-focused manufacturers, offering insights into vehicle design trade-offs between efficiency and performance.



The bar plot “Top 5 Vehicles with the Largest Engine Capacity” above highlights the vehicles with the biggest engine sizes, measured in cubic centimetres (cc), within the dataset. The x-axis lists the top 5 vehicles, identified by make and engine capacity (e.g., Jaguar 326 cc, Mercedes-Benz 308 cc), while the y-axis represents engine size in cc.

The analysis reveals that Jaguar dominates with the largest engine at 326 cc, followed by Mercedes-Benz models at 308 cc and 304 cc. The presence of multiple Mercedes-Benz entries and two Jaguars indicates these brands prioritize high-performance, luxury vehicles with larger engines. This suggests that vehicles with engine capacities exceeding 250 cc are typically from premium manufacturers, likely offering enhanced power at the cost of fuel efficiency, providing insight into the trade-offs of engine size in the automotive market.



The bar plot above, titled “Number of Car Models by Manufacturer,” displays the count of car models produced by each vehicle manufacturer in the dataset. The x-axis lists the manufacturers, while the y-axis represents the number of car models.

The analysis shows that Toyota leads with the highest number of car models at 32, underscoring its prominent status as a manufacturer known for reliable and fuel-efficient vehicles. This dominance suggests Toyota’s broad market presence and appeal. In contrast, Mercury trails with the fewest models at 1, highlighting its limited representation and relative unfamiliarity in the dataset.

This report was written by : [Xolani]