Exploratory Data Analysis on the Automobile Dataset

Report

Introduction

This dataset provides specifications for various automobiles, including attributes like engine size, fuel efficiency, horsepower, and price.

It is useful for analyzing manufacturer trends, performance metrics, and relationships between technical specifications and pricing.

The goals of this analysis are to:

- Identify which manufacturers produce the largest or smallest engines
- Determine which vehicles are most or least fuel-efficient
- Explore pricing trends and technical specifications across vehicle models

Data Cleaning

- Loaded the dataset from automobile.txt into a pandas DataFrame

- Removed all duplicate entries using df.drop_duplicates()

- Dropped rows with missing data using df.dropna()

- Converted numerical columns (e.g., engine-size, horsepower, price) to integer data types using astype(np.int64)

- Renamed column headers to be consistent and lowercase for easier manipulation (df.columns = df.columns.str.lower().str.replace('-', '_'))

Missing Data

Some columns such as horsepower and price had missing values.

These rows were removed because the number was small and filling in fake values could affect accuracy — especially when analyzing performance or pricing.

df.dropna(inplace=True)

Insights and Visualizations

1. Most Car Models by Manufacturer

Grouped by make and counted the number of unique models per brand.

Finding:

Toyota had the highest number of models, followed by **Nissan** and **Mazda**.

2. Vehicles with the Largest Engine Sizes

Sorted the dataset by engine_size and selected the top 10 vehicles.

Finding:

Jaguar and **Mercedes-Benz** appeared to be the highest among the largest engines.

3. Least Fuel-Efficient Manufacturers

Sorted by highway_mpg in ascending order to identify cars with the worst fuel economy.

Finding:

Manufacturers like **Porsche**, **Mercedes-Benz**, and **Jaguar** produce vehicles with high power but lower fuel efficiency — showing a clear tradeoff between performance and economy.

4. Price vs. Fuel Efficiency

To compare the price of cars with their MPG, I used a scatter plot.

This was better than a box plot because it shows the relationship between two numeric variables.

Finding:

The scatter plot showed that expensive cars usually use more fuel, while cheaper cars are more fuel-efficient.

This helped reveal the tradeoff between vehicle cost and economy.

Conclusion

This analysis gave insight into how car specs like engine size, price, and fuel efficiency vary across manufacturers.

It also highlighted trends between performance and affordability that can help both consumers and researchers better understand the automobile market.

This report was written by: Veli Nhlapo