EDA Report – Automobile Dataset

Introduction

This exploratory data analysis (EDA) investigates the Automobile dataset, which has specifications and pricing information for different vehicle models. The aim is to clean and analyze the dataset to reveal key insights about automobile pricing, efficiency, and relationships between technical features.

Data Cleaning Summary

- Replaced placeholder values ('?') with missing values (NaN).

- Converted numeric-like columns (e.g., horsepower, price, engine-size) to numeric data types.

- Removed duplicate rows to ensure data integrity.

- Discarded entries with missing values to create a complete dataset for analysis.

- Saved the cleaned dataset as automobile_cleaned.csv.

Handling Missing Data

The dataset contained missing values across several columns, such as normalized losses and horsepower. Instead of imputing, all rows containing any missing value were removed, ensuring only complete cases remained for EDA.

Data Stories & Visualizations

1. Price Distribution
   A histogram shows prices are right-skewed, most vehicles are in the low-to-mid price range, with a few luxury outliers.

2. Horsepower vs Price
   A positive trend appears higher horsepower generally corresponds to higher prices. Powerful vehicles tend to be more expensive.

3. Price by Body Style
   A boxplot comparison shows convertibles and hardtops have the highest median prices, while sedans and hatchbacks are more affordable.

4. Correlation Matrix
   A heatmap reveals strong positive correlations between engine-size, curb-weight, horsepower, and price, indicating these features jointly influence vehicle cost.

5. Fuel Efficiency by Manufacturer
   By averaging city and highway MPG, the most fuel-efficient brands emerge at the top of the bar chart,

giving insights into economic models versus performance-focused ones.

Findings & Insights

Price drivers: Horsepower, engine size, and curb weight are the most influential predictors of vehicle price.

Body style impact: Luxury body types (convertible, hardtop) command premium pricing.

Fuel efficiency trade-off: Economical manufacturers achieve higher MPG but generally lower prices.

Data quality: After cleaning, the dataset is complete and ready for modelling or predictive analysis.