EXPLORATORY DATA ANALYSIS ASSIGNMENT

Name: Pradyuman Singh Shekhawat
Enrolment: 23115107
Sub-batch: CS-3

Introduction

The report gives an overall idea of the methodology adopted to clean and analyse the Cars93 dataset. It points out the structured approach followed, the issues faced, important insights gained from exploratory data analysis (EDA), and learning experiences. The main aim was to preprocess the dataset efficiently and identify useful patterns in the data using different statistical and visualization methods.

Dataset Overview

The Cars93 dataset has data on various car models and their features, including price, horsepower, fuel efficiency, and engine size. It also has missing values, categorical inconsistencies, and outliers that must be handled before any useful analysis can be done.

Approach and Implementation

1. Data Loading and Initial Exploration
The dataset was loaded with Pandas.

The info() function was employed to examine the dataset structure, determining numerical and categorical variables.

head() function was called to have an initial glimpse at the data.

2. Managing Missing Values
Missing values count in each feature was printed out.

Numerical columns' missing values were replaced with the most frequent value.

Missing categorical values were replaced by the mode (most frequent category).

3. Deleting Duplicate Rows
Duplicate entries were detected and deleted to keep the data reliable.

4. Encoding Categorical Variables
Label Encoding was used to encode categorical features into numerical form.

Original values were stored in a different column for reference purposes.

5. Feature Scaling
A Min-Max Scaler was used to scale numerical features to the range 0-1.

6. Outlier Detection and Removal
The Z-score approach was used to delete extreme outliers (threshold: Z-score > 3).

7. Exploratory Data Analysis (EDA)

Univariate Analysis

Histograms were plotted for numerical features to analyze their distributions.
Boxplots were employed to determine the spread and possible outliers.
Summary statistics were produced for all numerical variables.

Bivariate Analysis

A **correlation heatmap** was produced to study correlations between numerical variables.
Pair plots were employed to display relationships between major numerical attributes such as Horsepower, MPG.city, and MPG.highway.
Violin plots and **boxplots** tested categorical-numerical interactions like Make vs. Horsepower.

Multivariate Analysis

A **3D scatter plot** was employed to study the relationship between Horsepower, MPG.city, and Price.
A new variable, power_to_weight, was designed to estimate efficiency.
An **Isolation Forest** model was used to identify and eliminate outlier points.
A **heatmap for grouped comparisons** was created in order to look for trends among numerical values against various categorical variables.

Key Findings and Observations

1. Correlation Analysis

Horsepower demonstrated a high negative correlation with MPG.city (-0.87), meaning that the more horsepower, the less efficient fuel.

Price was moderately positively correlated with Horsepower (0.61), reflecting the idea that high-powered cars are typically more costly.

Weight was positively correlated with EngineSize (0.78) and is consistent with the notion that bigger engines make heavier cars.

2. Features Distribution

The Horsepower distribution was right-skewed, as there were a greater number of automobiles with lower horsepower ratings.

MPG.city and MPG.highway distributed normally but possessed some outliers as exceptionally frugal or wasteful cars.

3. Outlier Analysis

Some vehicles had abnormally high horsepower ratings, probably sports cars or luxury vehicles.

A couple of records in MPG.city were outliers, indicating either data entry errors or extremely efficient hybrid/electric vehicles.

Anomalies identified by the Isolation Forest model primarily consisted of vehicles with outlier values in several numeric features.

4. Effect of Categorical Features

Certain car brands had much higher average horsepower, e.g., sports car manufacturers.

The cylinders_original count impacted several characteristics, including fuel consumption and cost, as evidenced from the grouped comparison heatmap.

Issues and Trial-and-Error Process

1. Missing Value Handling

Missing numerical values were first handled by imputing them with the mean, but it created an unrealistic distribution for certain characteristics. Changing to the most frequent value worked better.

For categorical features, some categories had very low frequency, and that caused label encoding problems. Keeping the original categories together with encoded values was useful for interpretation.

2. Outlier Detection

At first, outlier removal using the IQR approach (Interquartile Range) caused too much data loss. Shifting to the Z-score approach with a threshold value of 3 worked better.

The Isolation Forest model first misclassified certain normal points as anomalies and needed parameter tuning to be more accurate.

3. Scaling and Encoding

Standard Scaling first warped relationships among numerical features; hence, Min-Max Scaling was used instead.

Categorical encoding first eliminated useful interpretability; hence, original values were preserved in addition to encoded values.

Learning Outcomes

1. Data Cleaning is Imperative: Missing value handling and categorical data encoding correctly are vital to avoid analysis distortions.

2. Feature Engineering Further Insights: Developing new features such as power_to_weight gave greater insights into car efficiency.

3. EDA Facilitates Pattern Discovery: Plotting data with heatmaps, scatter plots, and boxplots uncovered concealed interdependencies among features.

4. Anomaly Detection is Tricky: Statistical and machine learning-based outlier detection needs to be finely tuned so that there is not too much loss of data.

5. Correlation Analysis Plays a Central Role: Knowledge about which variables affect one another helps improve decision-making.