Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

Summary of the data set

The automobile dataset encompasses diverse information on car attributes. Key features include the car's symboling, normalized-losses, make, fuel type, aspiration, number of doors, body style, drive wheels, engine location, wheel base, engine type, fuel system, bore, stroke, compression ratio, horsepower, peak rpm, city and highway fuel efficiency (in mpg), and the corresponding prices. This dataset provides a comprehensive overview of various car specifications, facilitating exploratory data analysis and insights into the automotive industry.

DATA CLEANING

- **Loading the Dataset:**
    - Imported the dataset into a Pandas DataFrame using **pd.read_csv("automobile.txt", na_values='?')**.
- **Handling Missing Values:**
    - Replaced '?' with NaN using **automobile_df.replace('?', np.nan)** to standardize missing values.
- **Identifying Constant Columns:**
    - Detected and identified constant columns using **constant_columns = automobile_df.columns[automobile_df.nunique() == 1]**.
- **Identifying Highly Correlated Columns:**
    - Calculated the correlation matrix for numerical columns and identified highly correlated pairs (correlation > 0.8) using nested loops.
- **Displaying Identified Columns:**
    - Printed the names of constant and highly correlated columns to the console for reference.
- **Removing Rows with Missing Data:**
    - Removed rows with any missing values using **automobile_df = automobile_df.dropna()**.
- **Displaying Cleaned DataFrame:**
    - Printed the first few rows of the DataFrame after removing missing data to confirm the cleaning process.

These methods ensured a cleaner and more standardized dataset, preparing it for further analysis. The identification and handling of missing values, constant columns, and highly correlated features contribute to a more robust and reliable dataset for subsequent exploratory data analysis.

MISSING DATA

The code snippet **automobile_df = automobile_df.dropna()** is used to handle missing data in the automobile dataset. Specifically, it removes rows containing any missing values (NaN) across all columns. The subsequent **print(automobile_df.head())** displays the first few rows of the DataFrame after the removal process.

In simpler terms, any rows that have incomplete information are excluded, ensuring that the dataset is more complete for subsequent analysis. This approach is effective when the impact of missing data on the overall dataset is considered acceptable or when imputation is not preferred. The printed DataFrame provides a glimpse of the cleaned dataset, demonstrating the immediate effects of the removal on the first few rows.

DATA STORIES AND VISUALISATIONS

Visualization 1: Distribution of Car Prices

Insights:

- **Price Concentration:** The histogram suggests a concentration of cars within a specific price range, indicating a prevailing pricing trend.
- **Outliers:** Potential outliers at extreme price values may represent luxury or budget vehicles, warranting further investigation.

Potential Stories and Assumptions:

- **Price Segmentation:** Explore if there are distinct price segments in the market.
- **Luxury or Budget Market Presence:** Identify the role of outliers in representing luxury or budget car categories.

Visualization 2: Scatter Plot of Engine Size vs. Horsepower

Insights:

- **Correlation:** The scatter plot reveals a positive correlation between engine size and horsepower, indicating that larger engines generally have higher horsepower.
- **Engine Efficiency:** Evaluate whether certain engine sizes consistently deliver higher horsepower or if there are notable outliers.

Potential Stories and Assumptions:

- **Performance Trends:** Analyze if cars with larger engines consistently deliver superior performance.

- **Outliers Identification:** Investigate any cars with unexpected horsepower values based on their engine size.

Visualization 3: Categorical Bar Chart of Car Types

Insights:

- **Popular Body Styles:** The bar chart displays the count of cars based on body style, revealing which styles are more prevalent in the dataset.
- **Market Preferences:** Assess whether certain body styles are more popular, providing insights into consumer preferences.

Potential Stories and Assumptions:

- **Body Style Popularity:** Identify the dominant body styles in the dataset.
- **Consumer Preferences:** Explore whether market trends favor specific body styles over others.

These visualizations lay the foundation for a comprehensive exploratory data analysis, offering initial insights into the distribution of car prices, the relationship between engine size and horsepower, and the prevalence of different body styles in the dataset.

THIS REPORT WAS WRITTEN BY : Mbali Matches