The mpg dataset contains information on various automobile models. The dataset includes multiple numerical and categorical variables representing attributes such as fuel efficiency (mpg), engine characteristics (cylinders, displacement, horsepower), vehicle weight, acceleration, model year, country of origin, and car name. The analysis involves handling missing values, removing duplicates, detecting outliers, and performing various techniques to understand the data.

Data Cleaning

Loading the Dataset

- The dataset was loaded into a Pandas DataFrame.
- The first few rows were displayed to understand its structure.
- Data types of each column were checked to identify categorical and numerical variables.

Handling Missing Values

- The dataset was inspected for missing values.
- The horsepower column had 6 missing values, which were imputed using the median.

Removing Duplicate Records

- The dataset was checked for duplicate rows.
- Any duplicate records found were removed to ensure data integrity.

Detecting and Treating Outliers

- Box plots were used to visualize outliers in numerical variables.
- The Z-score method was applied to detect extreme values in mpg, horsepower, and weight.
- Outliers were either removed or transformed using appropriate scaling techniques.

Standardizing Categorical Values

- Categorical variables were inspected for inconsistencies.
- The origin column was standardized to have uniform values (USA, Europe, Japan).
- Car names were formatted to maintain consistency.

Exploratory Data Analysis (EDA)

Univariate Analysis

- Summary Statistics: Mean, median, mode, variance, and skewness were calculated for numerical variables.
- Frequency Distribution: The count of each unique value in categorical variables was determined.
- Histogram: Used to visualize the distribution of mpg.
- Box Plot: Used to detect outliers in horsepower.

Bivariate Analysis

- Correlation Matrix: Identified relationships between numerical variables.
- Scatter Plot: Explored the relationship between mpg and weight.
- Box Plot: Compared mpg across different cylinders categories.
- Violin Plot: Examined variations in mpg based on origin.

Multivariate Analysis

- Pair Plot: Analyzed multiple relationships among mpg, horsepower, weight, and displacement.
- Heatmap: Visualized correlations between all numerical variables.
- Line Plot: Tracked fuel efficiency trends over different model_years.

Key Insights

- Mpg and weight have a strong negative correlation (-0.83), indicating that heavier cars tend to have lower fuel efficiency.
- Horsepower and displacement are positively correlated, meaning that larger engines produce more power.
- Japanese cars generally have higher fuel efficiency compared to American and European cars.
- Cars with 4 cylinders are the most fuel-efficient, whereas 8-cylinder cars consume significantly more fuel.
- The automotive industry has shown improvements in fuel efficiency over time, with newer model years having higher mpg.

- The distribution of mpg is right-skewed, indicating that while most cars have moderate fuel efficiency, a few have very high values.
- Weight is positively skewed, meaning a few vehicles are significantly heavier than the rest.
- Outliers in horsepower suggest the presence of high-performance sports cars in the dataset.
- Acceleration has a weak correlation with mpg, implying that faster cars do not necessarily have better or worse fuel efficiency.
- Cars from the 1970s have a more diverse range of mpg values compared to later years, possibly due to varying fuel efficiency regulations and technology improvements.