Name : Dev Gaonkar **Div : D15C** **Roll No : 12**

Experiment No. - 2 :

Aim :

Perform following data visualization and exploration on your selected dataset.
1. Create bar graph, contingency table using any 2 features.
2. Plot Scatter plot, box plot, Heatmap using seaborn.
3. Create histogram and normalized Histogram.
4. Describe what this graph and table indicates.
5. Handle outlier using box plot and Inter quartile range.

Problem Statement : Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn.

Introduction :

Data Visualization with Matplotlib

Matplotlib is a comprehensive Python library for creating static, animated, and interactive visualizations. It provides control over plot elements like axes, labels, and colors. Key features include:

- **Line, bar, scatter, and histogram plots**
- **Customizable visual styles** (e.g., figure size, colors, titles)
- **Subplot capabilities** for multi-plot grids
  Matplotlib is ideal for creating basic and detailed visual representations of data.

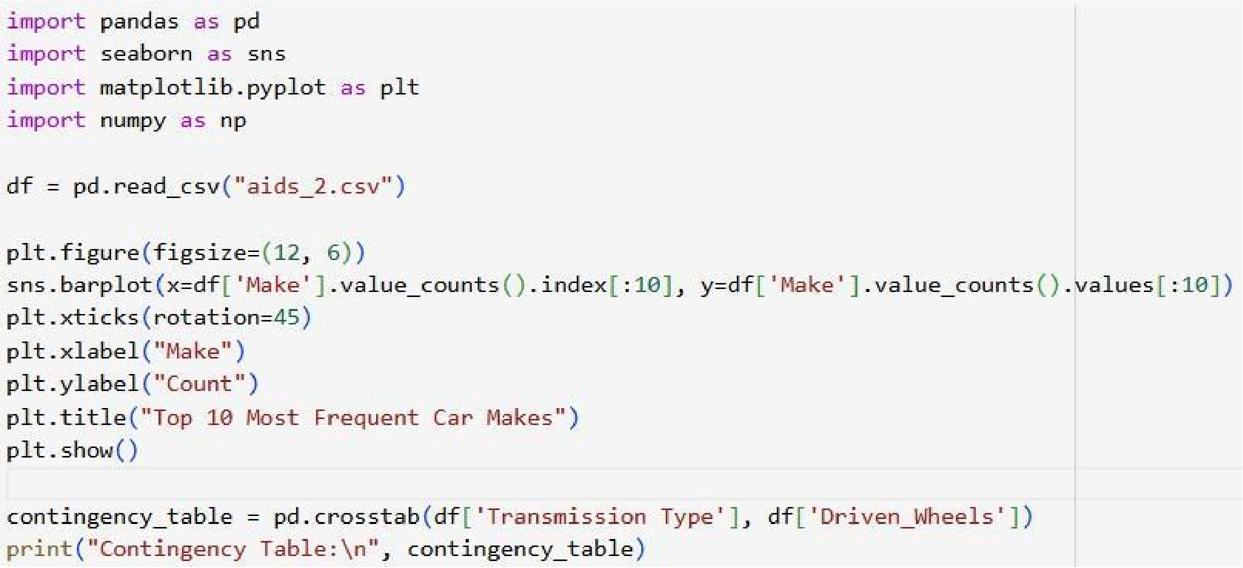
Exploratory Data Analysis (EDA) with Seaborn

Seaborn is built on top of Matplotlib and provides a high-level interface for creating visually appealing and informative statistical graphics. Key capabilities include:

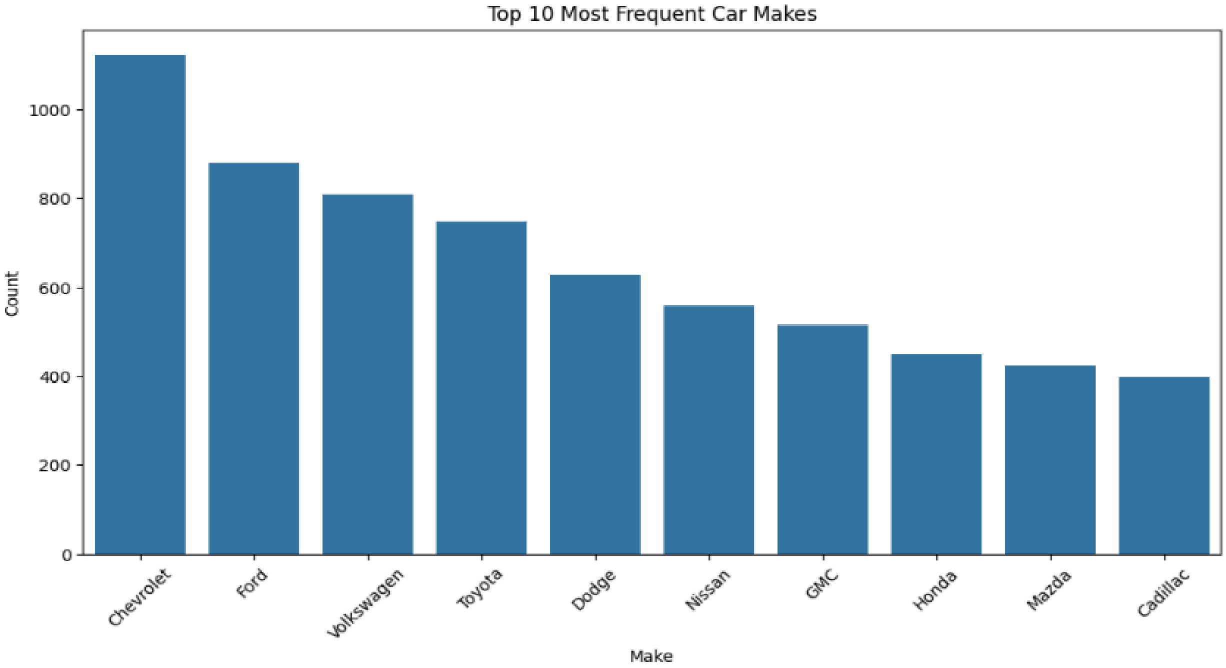
- **Visualizing distributions** (e.g., histograms, box plots, violin plots)
- **Correlation and relationships** (e.g., scatter plots, pair plots)
- **Easy integration with Pandas DataFrames**
  Seaborn simplifies the process of exploring data relationships and distributions, making it a powerful tool for EDA.

1.Bar Graph (top 10 car makes vs count of vehicles for make)

This graph provides a clear representation of the distribution of car brands within the dataset. By plotting the **Make** column's value counts, we can see which car brands are most prevalent in the dataset. The x-axis shows the top 10 car makes, while the y-axis indicates the count of vehicles for each make.

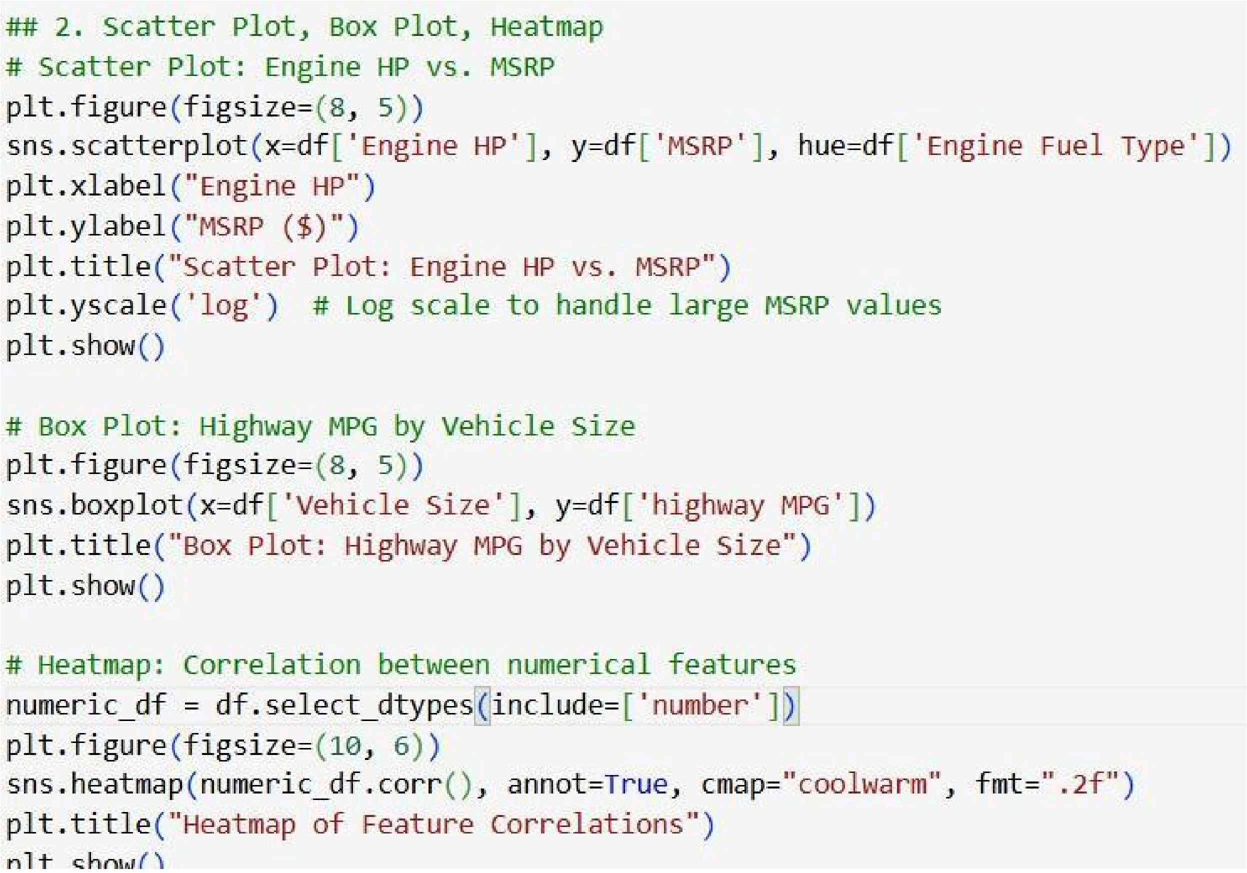


```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

df = pd.read_csv("aids_2.csv")

plt.figure(figsize=(12, 6))
sns.barplot(x=df['Make'].value_counts().index[:10], y=df['Make'].value_counts().values[:10])
plt.xticks(rotation=45)
plt.xlabel("Make")
plt.ylabel("Count")
plt.title("Top 10 Most Frequent Car Makes")
plt.show()

contingency_table = pd.crosstab(df['Transmission Type'], df['Driven_Wheels'])
print("Contingency Table:\n", contingency_table)
```



```
Contingency Table:
 Driven_Wheels     all wheel drive  four wheel drive  front wheel drive  \
Transmission Type
AUTOMATED_MANUAL               198                 0                304
AUTOMATIC                     1940              1056               3056
DIRECT_DRIVE                    11                 0                 43
MANUAL                         204               345               1380
UNKNOWN                          0                 2                  4

Driven_Wheels      rear wheel drive
Transmission Type
AUTOMATED_MANUAL                124
AUTOMATIC                      2214
DIRECT_DRIVE                     14
MANUAL                         1006
UNKNOWN                          13
```

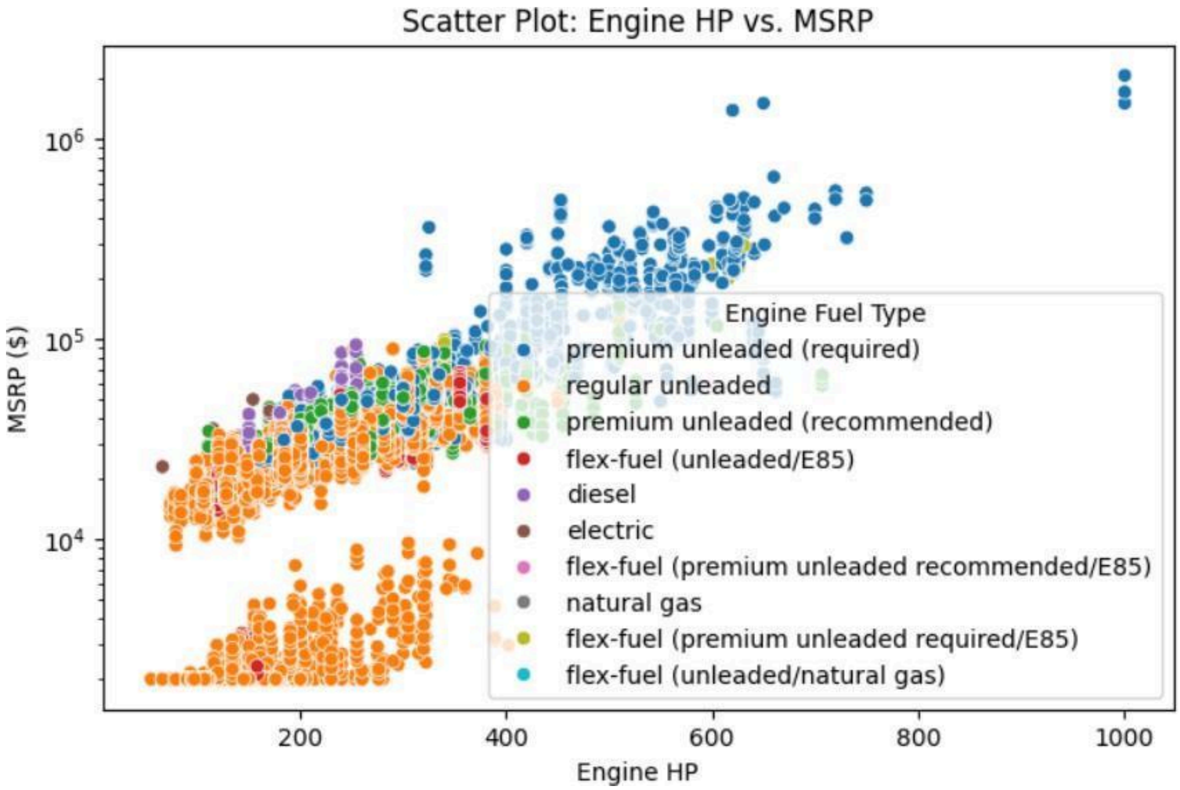
2.Scatter Plot,box plot, Heatmap using seaborn.



```
## 2. Scatter Plot, Box Plot, Heatmap
# Scatter Plot: Engine HP vs. MSRP
plt.figure(figsize=(8, 5))
sns.scatterplot(x=df['Engine HP'], y=df['MSRP'], hue=df['Engine Fuel Type'])
plt.xlabel("Engine HP")
plt.ylabel("MSRP ($)")
plt.title("Scatter Plot: Engine HP vs. MSRP")
plt.yscale('log')  # Log scale to handle large MSRP values
plt.show()

# Box Plot: Highway MPG by Vehicle Size
plt.figure(figsize=(8, 5))
sns.boxplot(x=df['Vehicle Size'], y=df['highway MPG'])
plt.title("Box Plot: Highway MPG by Vehicle Size")
plt.show()

# Heatmap: Correlation between numerical features
numeric_df = df.select_dtypes(include=['number'])
plt.figure(figsize=(10, 6))
sns.heatmap(numeric_df.corr(), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Heatmap of Feature Correlations")
plt.show()
```

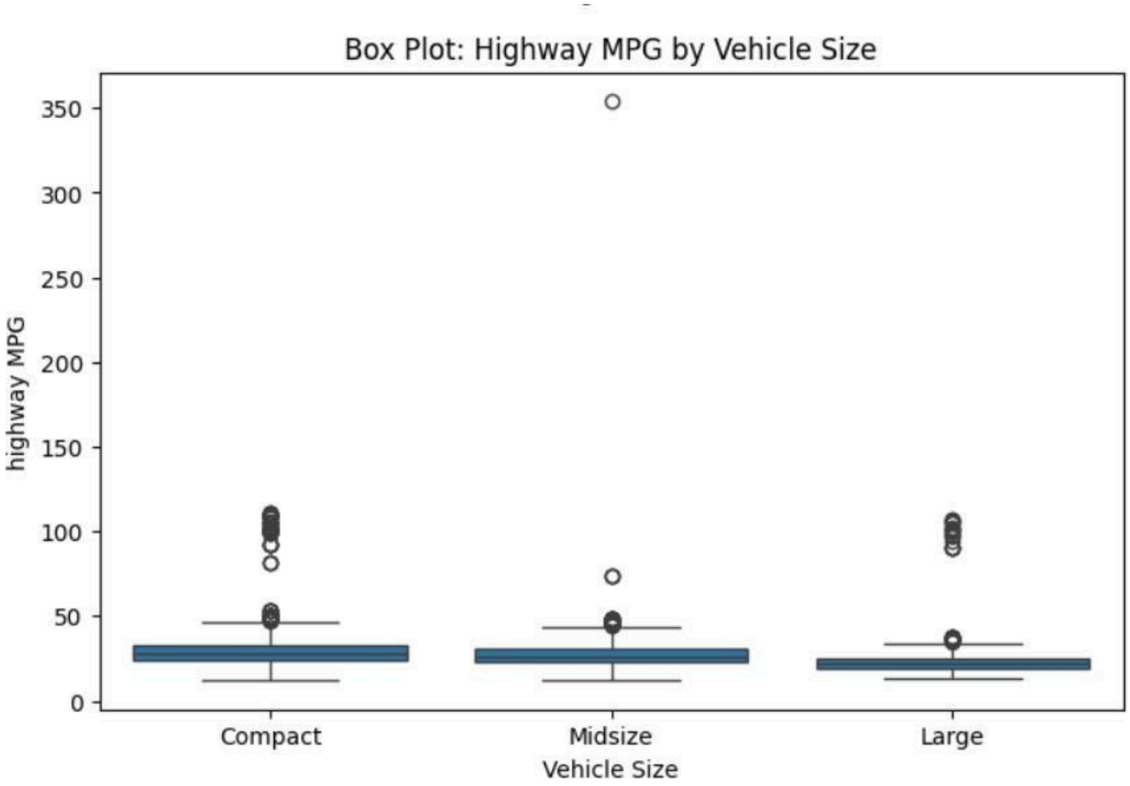
A **scatter plot** is a graph used to visualize the relationship between two continuous variables. It helps to identify correlations and trends in data. In this dataset, the scatter plot visualizes the relationship between Engine HP and MSRP, showing whether more powerful engines are associated with higher prices.

The plot uses color to represent Engine Fuel Type, adding another layer of insight. A logarithmic scale on the y-axis helps manage the wide range of MSRP values, making patterns clearer, especially in higher price ranges.



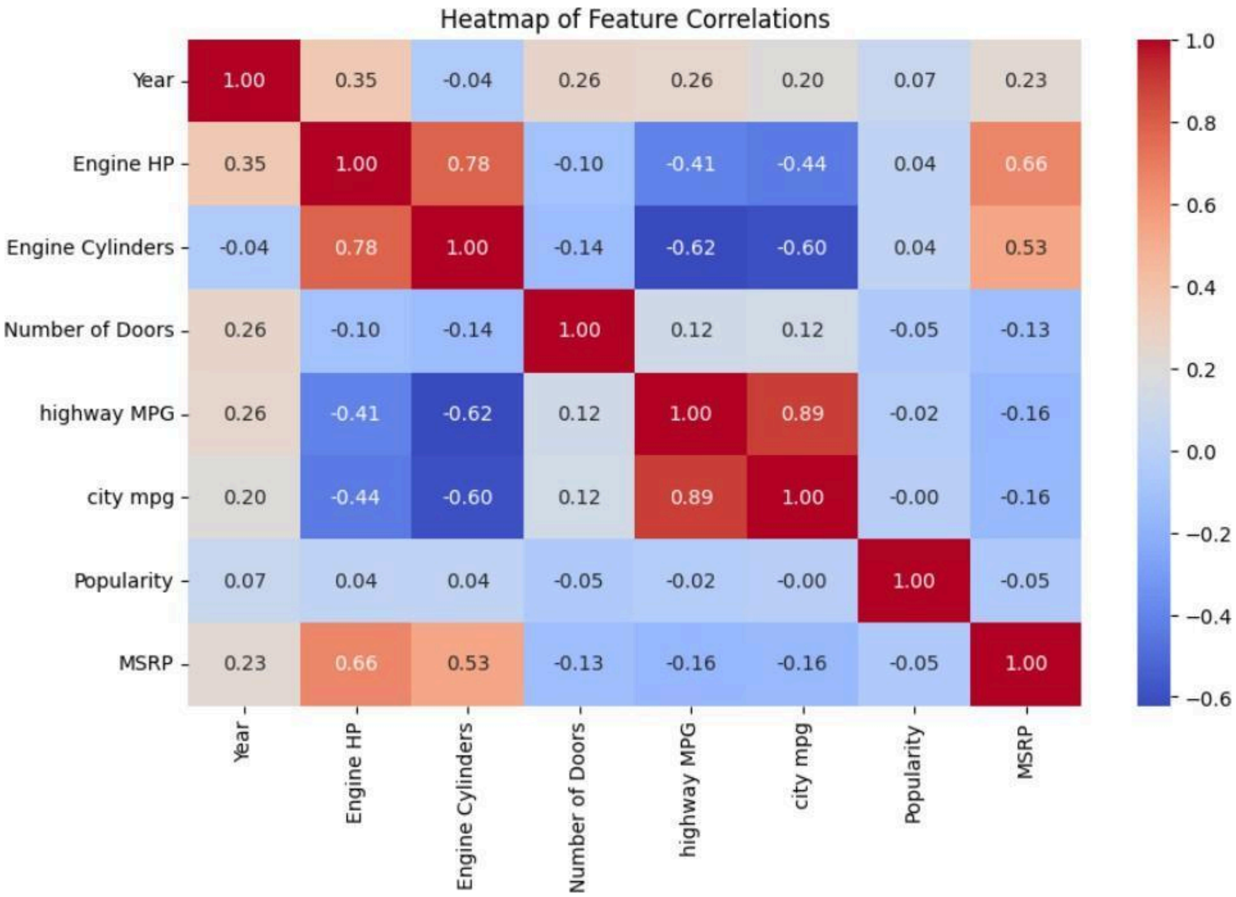
A **box plot** is used to show the distribution of data based on quartiles, highlighting the median, spread, and potential outliers.

In the dataset, the box plot visualizes the distribution of Highway MPG by Vehicle Size. It helps identify if larger vehicles tend to have lower fuel efficiency and points out any extreme outliers in highway MPG for different vehicle categories.

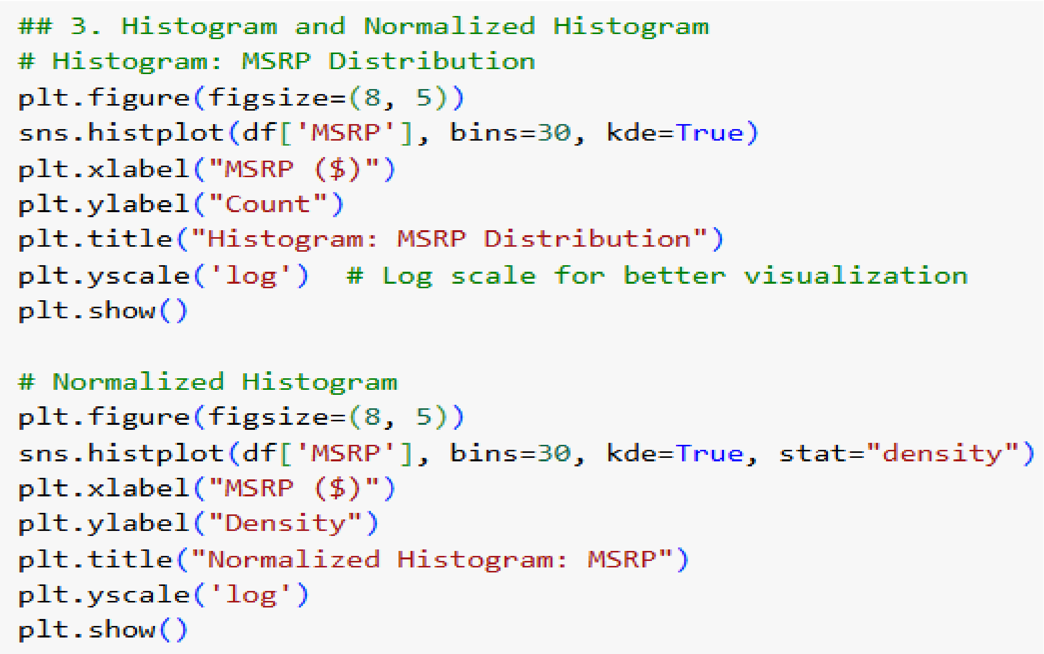


A **heatmap** is a graphical representation of data where values are represented by color, often used to show correlations between variables.

In the dataset, the heatmap visualizes the correlation between numerical features like Engine HP, MSRP, and Highway MPG. The color intensity indicates the strength of relationships, helping identify which variables are strongly correlated, such as whether Engine HP and MSRP are positively correlated.



3. Histogram and normalized Histogram.



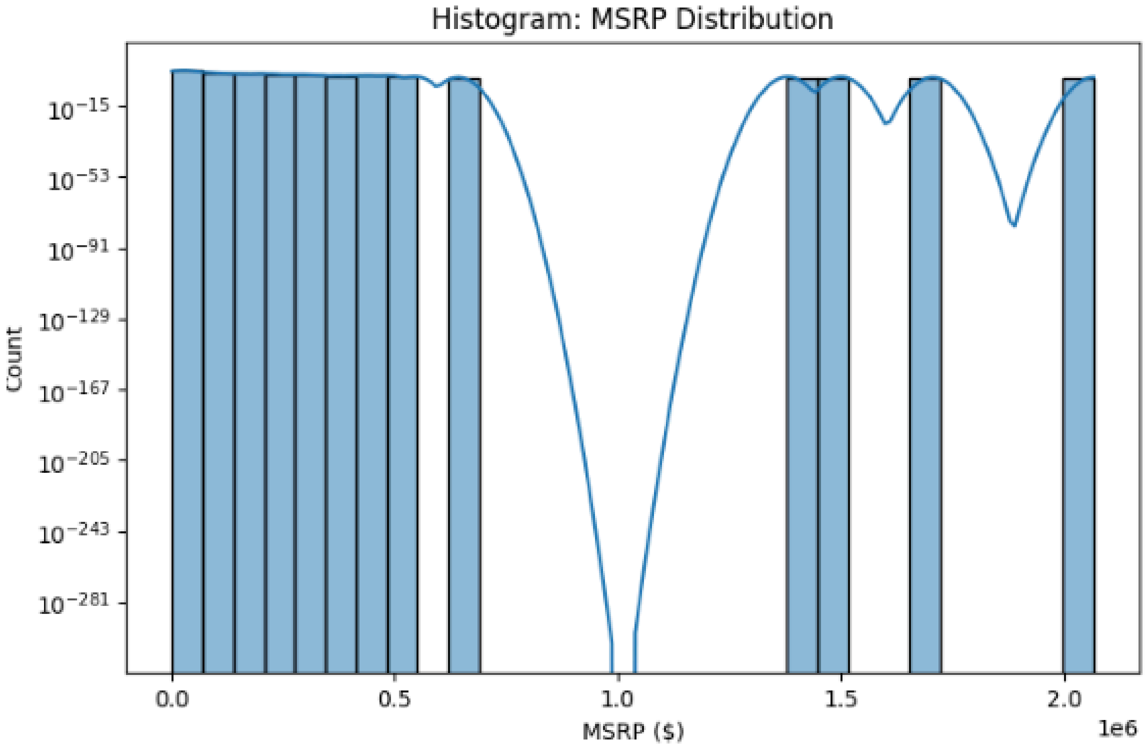
```
## 3. Histogram and Normalized Histogram
# Histogram: MSRP Distribution
plt.figure(figsize=(8, 5))
sns.histplot(df['MSRP'], bins=30, kde=True)
plt.xlabel("MSRP ($)")
plt.ylabel("Count")
plt.title("Histogram: MSRP Distribution")
plt.yscale('log')  # Log scale for better visualization
plt.show()

# Normalized Histogram
plt.figure(figsize=(8, 5))
sns.histplot(df['MSRP'], bins=30, kde=True, stat="density")
plt.xlabel("MSRP ($)")
plt.ylabel("Density")
plt.title("Normalized Histogram: MSRP")
plt.yscale('log')
plt.show()
```

A histogram is a graphical representation that shows the distribution of a dataset by grouping data into bins. It is particularly useful for visualizing the frequency of values within a continuous range.

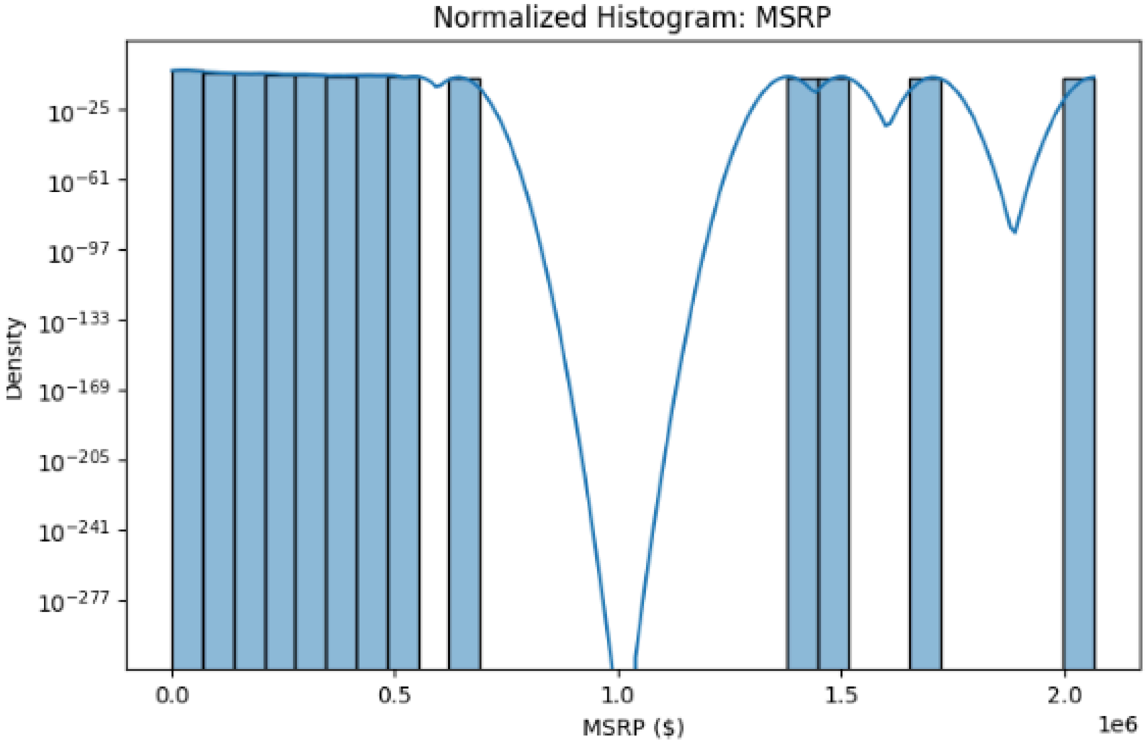
In this dataset, the histogram is used to display the distribution of MSRP (Manufacturer's Suggested Retail Price). The x-axis represents the range of MSRP values, while the y-axis shows the count of vehicles within each bin. The log scale on the y-axis helps visualize the distribution more effectively, especially with large values of MSRP, making it easier to observe the frequency of cars in different price ranges.

A normalized histogram represents the relative density (probability) rather than the raw count, showing how the distribution of data is spread out across the range of values.

In this dataset, the normalized histogram of MSRP helps to understand the probability distribution of car prices. The y-axis now represents density instead of count, providing a clearer view of the distribution's shape and allowing for easier comparison between different datasets or features. Like the histogram, the log scale is applied to the y-axis to help manage the skewed distribution of MSRP values.

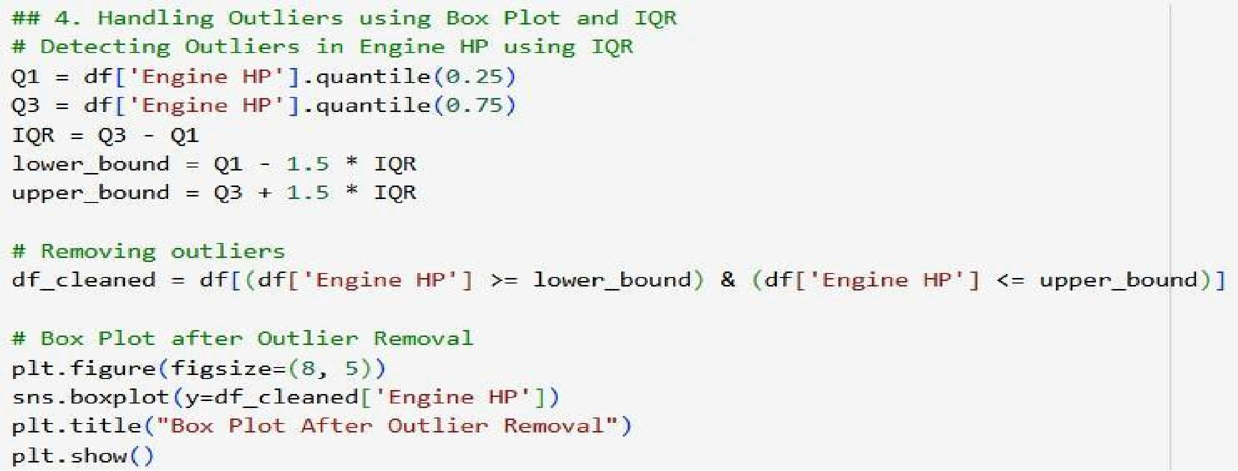


Histogram: MSRP Distribution



Normalized Histogram: MSRP

4.Outlier using box plot and Inter quartile range.



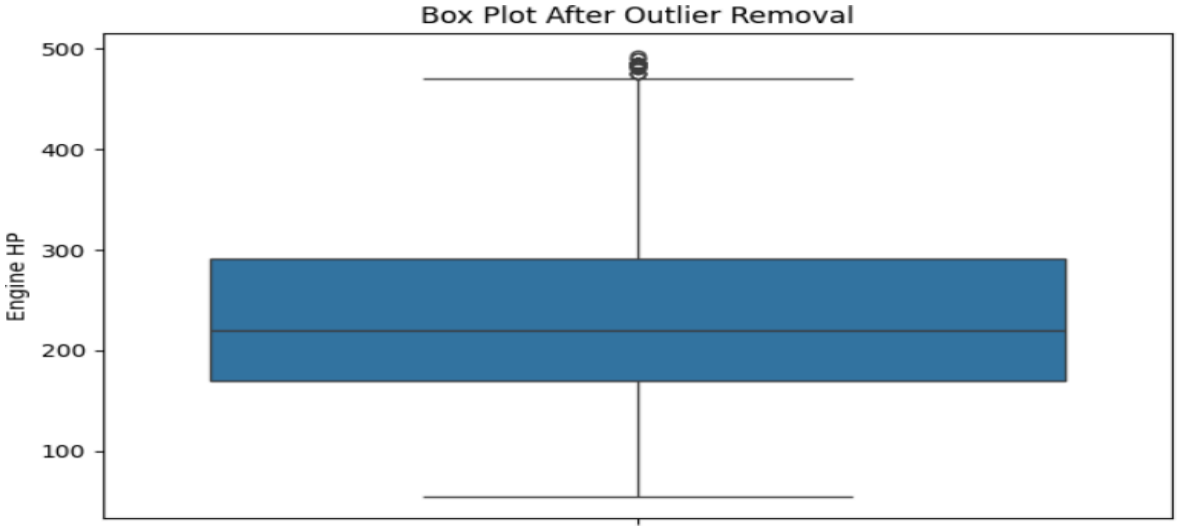
```
## 4. Handling Outliers using Box Plot and IQR
# Detecting Outliers in Engine HP using IQR
Q1 = df['Engine HP'].quantile(0.25)
Q3 = df['Engine HP'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Removing outliers
df_cleaned = df[(df['Engine HP'] >= lower_bound) & (df['Engine HP'] <= upper_bound)]

# Box Plot after Outlier Removal
plt.figure(figsize=(8, 5))
sns.boxplot(y=df_cleaned['Engine HP'])
plt.title("Box Plot After Outlier Removal")
plt.show()
```

A box plot is a useful tool for visualizing the distribution of a dataset and detecting outliers. It displays the median, quartiles, and potential outliers, providing a clear overview of the data's spread and central tendency.

In this dataset, the box plot is applied to the Engine HP feature, helping identify any extreme values or outliers in engine horsepower. The whiskers of the box plot indicate the range of data within the interquartile range (IQR), while points outside this range are considered outliers.

To handle outliers in Engine HP, the Interquartile Range (IQR) method is used. The first step is to calculate the IQR by subtracting the 25th percentile (Q1) from the 75th percentile (Q3). Outliers are typically defined as values outside the range of 1.5 times the IQR above Q3 or below Q1.

In this case, any Engine HP values outside the calculated bounds (lower and upper) are considered outliers and are removed from the dataset. The box plot is then regenerated to display the distribution of Engine HP after removing the outliers, allowing for a cleaner view of the data without extreme values skewing the results.



Conclusion :

In conclusion, using Matplotlib and Seaborn for Exploratory Data Analysis (EDA) helps uncover key insights in a dataset. Bar graphs reveal the distribution of categorical features, while contingency tables show relationships between them. Scatter plots identify correlations between continuous variables, and box plots detect outliers, which are handled using the IQR method. Heatmaps visualize feature correlations, and histograms provide insights into variable distributions. These tools are essential for understanding data patterns, ensuring clean datasets, and guiding further analysis or modeling.