Name: Ganesh Gupta D15C Roll No: 13

Experiment 2

Aim: Data Visualization / Exploratory Data Analysis using Matplotlib and Seaborn

Introduction

Exploratory Data Analysis (EDA) is a crucial step in the data analysis pipeline. It involves visually and statistically exploring the data to gain insights and understand its underlying patterns, distributions, and relationships. In this section, we will use Matplotlib and Seaborn, two popular Python libraries, to create visualizations that help in uncovering these insights.

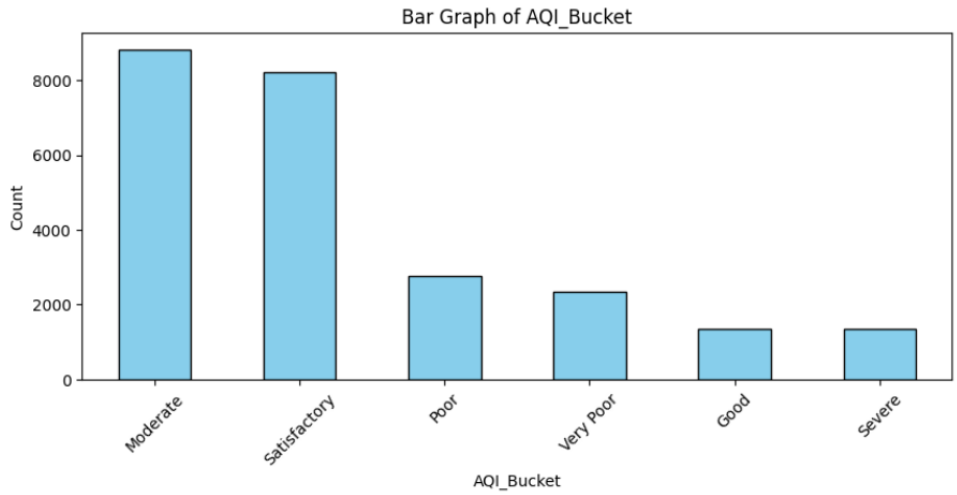
Matplotlib: A comprehensive library for creating static, animated, and interactive visualizations in Python.

Seaborn: Built on top of Matplotlib, Seaborn provides a high-level interface for drawing attractive and informative statistical graphics.

The primary objective of this analysis is to explore and visualize key patterns in the vehicle dataset, focusing on understanding relationships between different attributes and identifying significant trends.

Bar Graph and Contingency Table

```
# Bar Graph
plt.figure(figsize=(10, 4))
df[feature1].value_counts().plot(kind='bar', color='skyblue', edgecolor='black')
plt.xlabel(feature1)
plt.ylabel("Count")
plt.title(f"Bar Graph of {feature1}")
plt.xticks(rotation=45)
plt.show()
```



Observation: The majority of air quality readings fall under "Moderate" and "Satisfactory" categories, indicating generally acceptable pollution levels, but the presence of "Poor" and "Severe" levels suggests occasional hazardous conditions.

Contingency Table

A contingency table helps analyze the relationship between PM2.5 (particulate matter) and AQI_Bucket.

```
feature1 = "PM2.5"
feature2 = "AQI_Bucket"

contingency_table = pd.crosstab(df[feature1], df[feature2])
print("\nContingency Table:\n", contingency_table)
```

```
Contingency Table:
 AQI_Bucket  Good  Moderate  Poor  Satisfactory  Severe  Very Poor
PM2.5
0.04           0         0     0             1       0          0
0.16           1         0     0             0       0          0
0.24           1         0     0             0       0          0
0.28           1         0     0             0       0          0
0.98           1         0     0             0       0          0
```

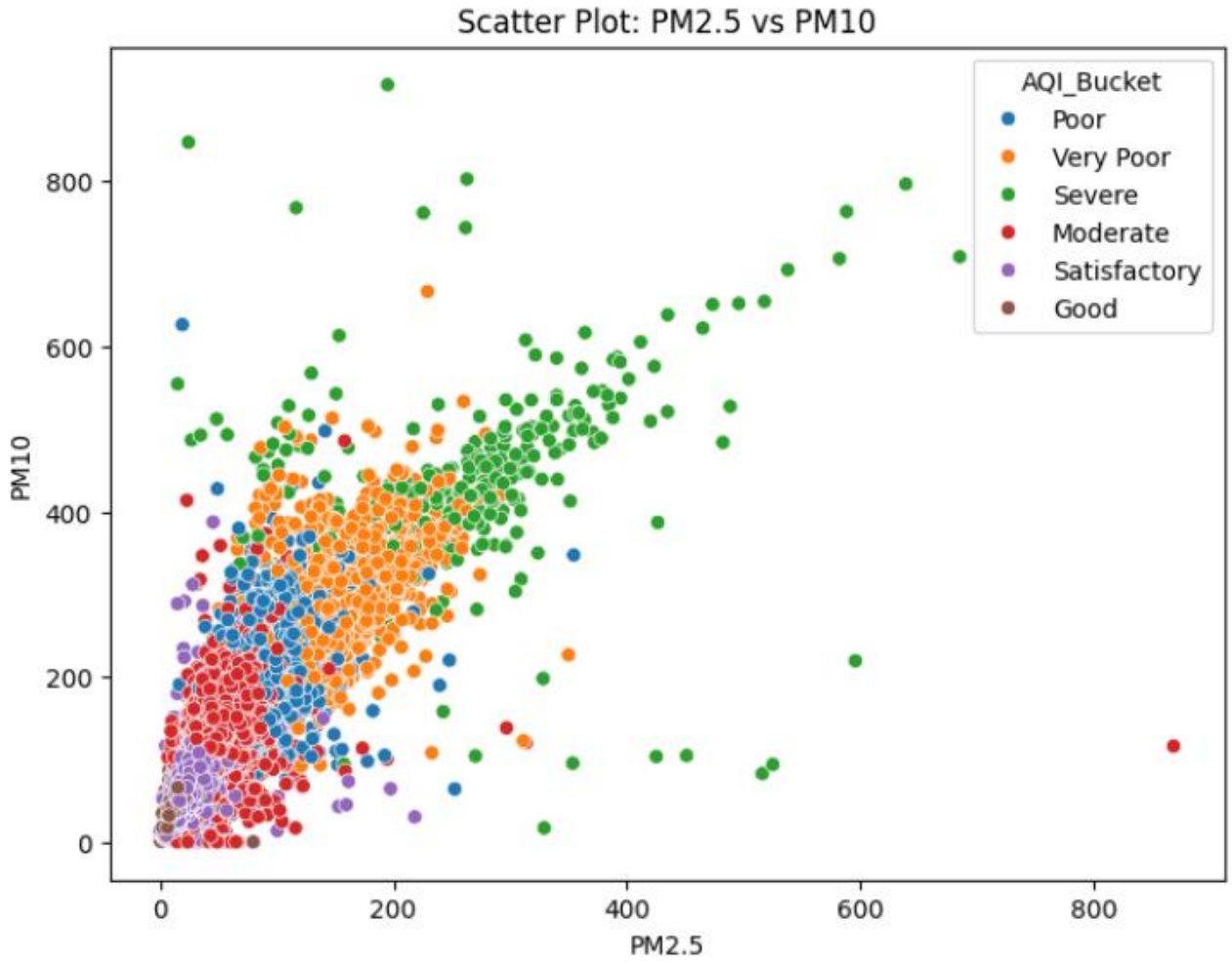
Observation: The contingency table shows the distribution of PM2.5 values across different AQI categories, indicating a relationship between particulate matter concentration and air quality levels. If dust matter is finer (0.16 > 0.04), the AQI reduces.

Scatter Plot, Box Plot, and Heatmap

A scatter plot helps in visualizing the relationship between particulate matter of varying micro-meters and AQI.

```
# Scatter Plot
plt.figure(figsize=(8, 6))
sns.scatterplot(x=df["PM2.5"], y=df["PM10"], hue=df["AQI_Bucket"])
plt.title("Scatter Plot: PM2.5 vs PM10")
plt.xlabel("PM2.5")
plt.ylabel("PM10")
plt.show()
```

Name: Ganesh Gupta D15C Roll No: 13



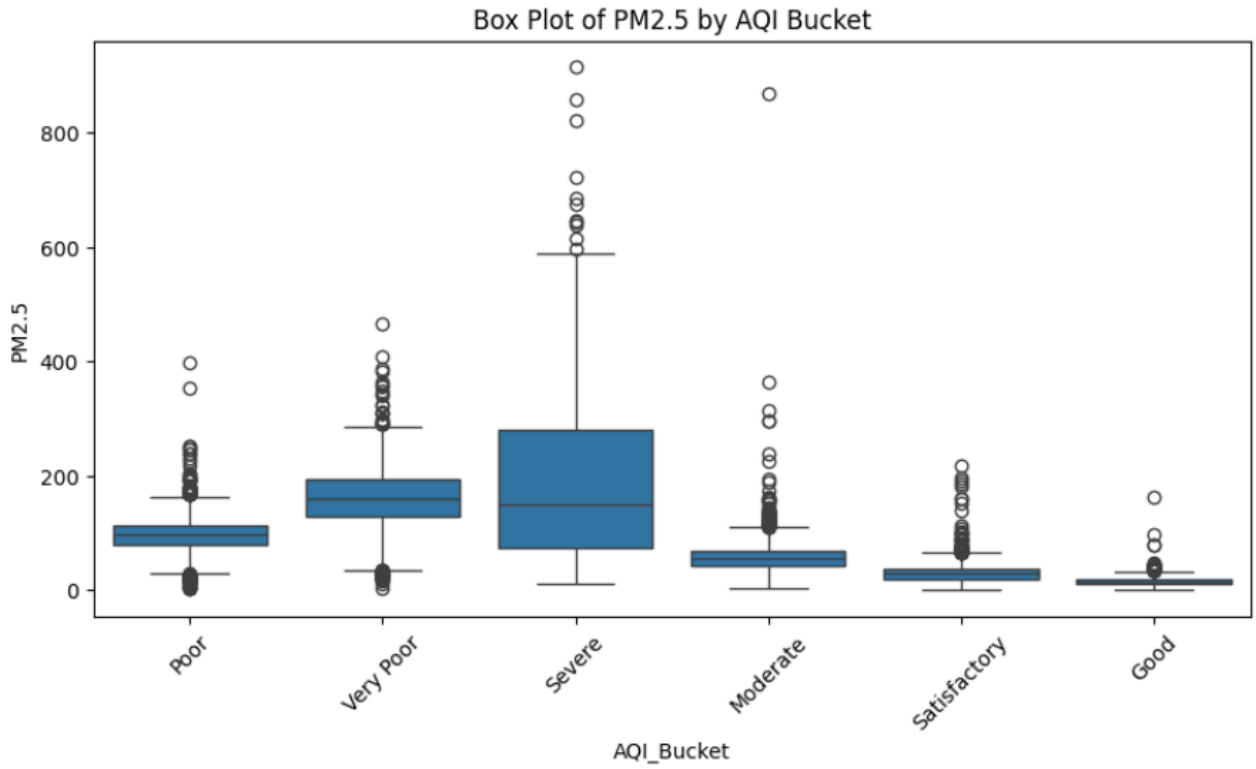
Observation: The scatter plot shows a positive correlation between PM2.5 and PM10, with higher pollutant levels corresponding to worse AQI categories, confirming that fewer particulates result in better air quality. Dots t the edge of the plot can be considered as outliers.

Box Plot

A box plot is used to summarize the distribution of the Data_Value column, helping to identify outliers, the median, and the interquartile range.

```
# Box Plot
plt.figure(figsize=(10, 5))
sns.boxplot(x=df["AQI_Bucket"], y=df["PM2.5"])
plt.title("Box Plot of PM2.5 by AQI Bucket")
plt.xticks(rotation=45)
plt.show()
```

Name: Ganesh Gupta D15C Roll No: 13



The lines on every plot represent the upper bound, Q3 (75%ile), Q2 (50%ile), Q1 (25%ile) and the lower bound. Circles represent outliers.

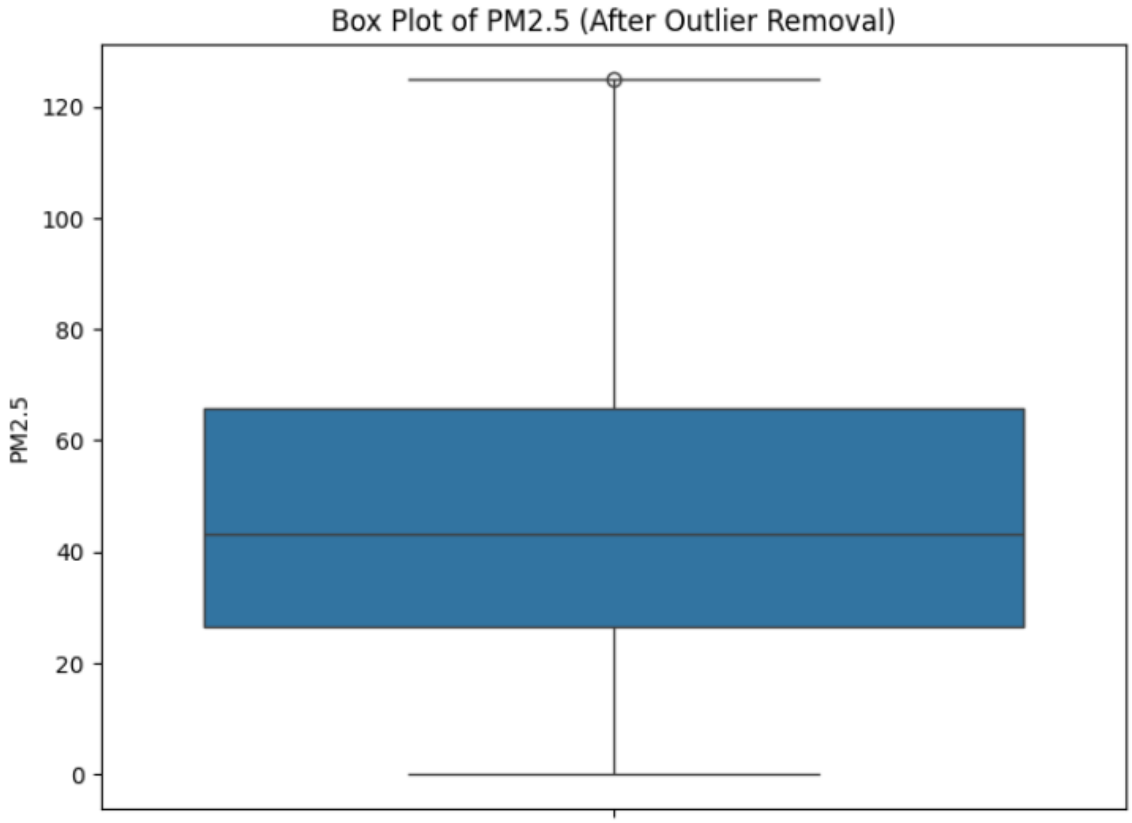
After removing outliers using IQR (Inter Quartile Range Method)

```
Q1 = df["PM2.5"].quantile(0.25)
Q3 = df["PM2.5"].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

outliers = df_cleaned_5[(df["PM2.5"] < lower_bound) | (df["PM2.5"] > upper_bound)]
#print("\nOutliers in PM2.5:\n", outliers)

df_cleaned = df[(df["PM2.5"] >= lower_bound) & (df["PM2.5"] <= upper_bound)]

plt.figure(figsize=(8, 6))
sns.boxplot(y=df_cleaned["PM2.5"])
plt.title("Box Plot of PM2.5 (After Outlier Removal)")
plt.show()
```



Observation: The box plot of PM2.5 (after outlier removal) shows that most values are concentrated within a reasonable range, but a few high values still exist as minor outliers, indicating occasional spikes in pollution levels.

Heatmap for Correlation Analysis

A heatmap is used to visualize the contingency table, which represents the relationship between two categorical variables.

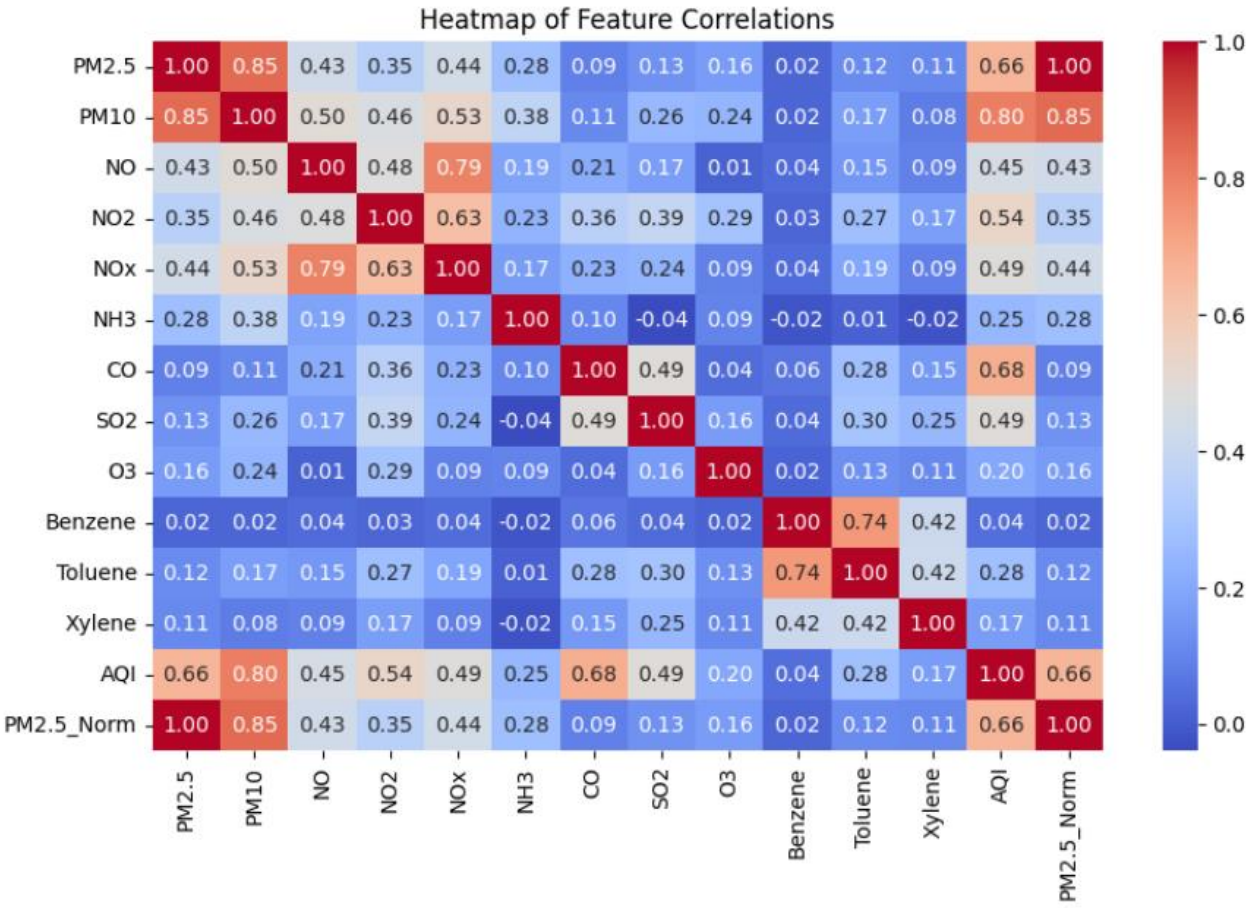
```
# Heatmap
plt.figure(figsize=(10, 6))

numeric_df = df.select_dtypes(include=['number'])

sns.heatmap(numeric_df.corr(), annot=True, cmap="coolwarm", fmt=".2f")

plt.title("Heatmap of Feature Correlations")
plt.show()
```

Name: Ganesh Gupta D15C Roll No: 13



Observation: The heatmap of feature correlations shows that PM2.5 has a strong positive correlation with PM10 (0.85) and AQI (0.66), indicating that higher partiicle levels are associated with poorer air quality.

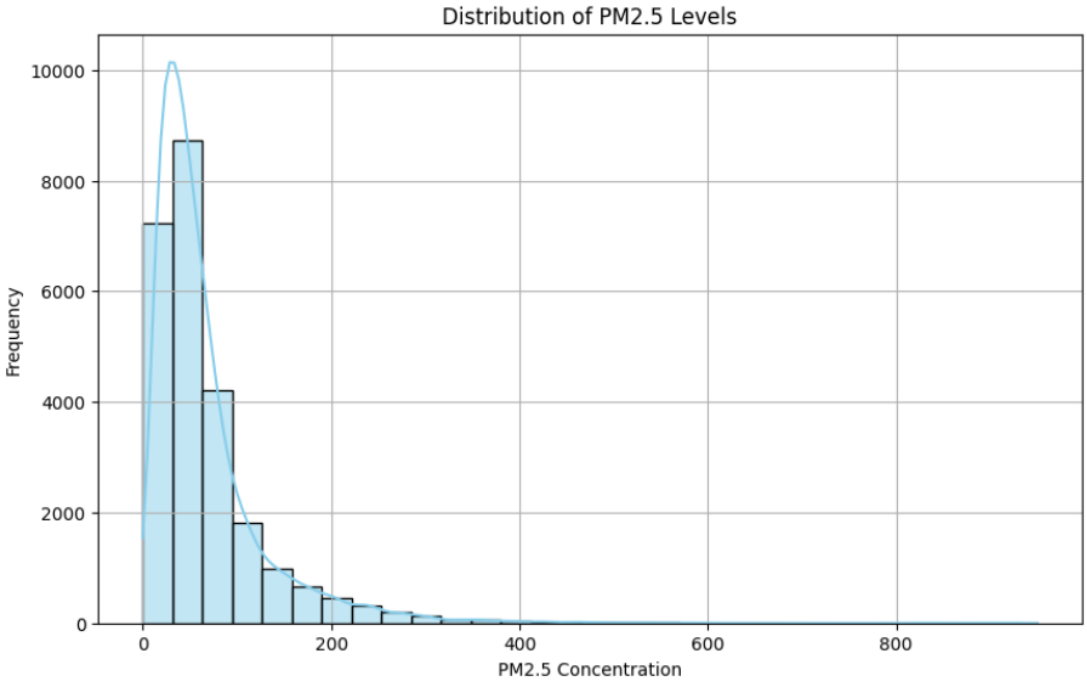
Histogram and Normalized Histogram

A histogram is used to visualize the distribution of Data_Value, showing how frequently different values occur.

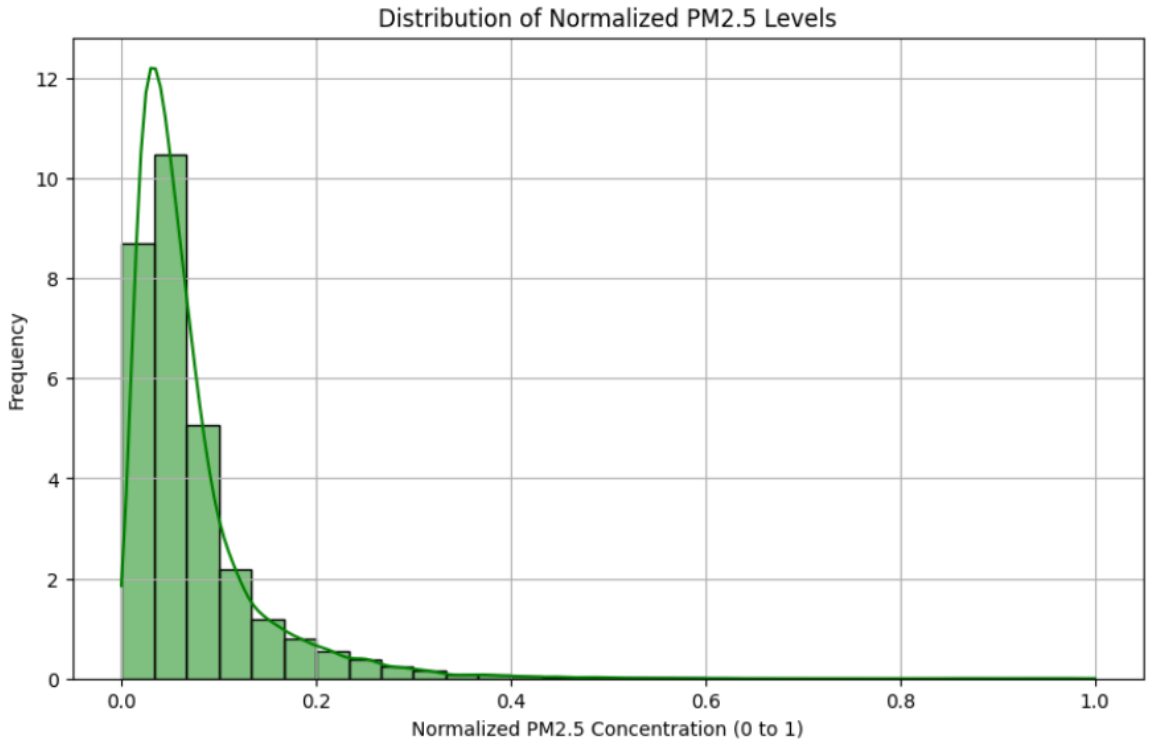
```
# Plot histogram for PM2.5
plt.figure(figsize=(10, 6))
sns.histplot(df["PM2.5"], bins=30, kde=True, color="skyblue")

plt.xlabel("PM2.5 Concentration")
plt.ylabel("Frequency")
plt.title("Distribution of PM2.5 Levels")
plt.grid(True)
plt.show()
```

Name: Ganesh Gupta D15C Roll No: 13



A normalized histogram represents data in terms of density instead of raw frequency, ensuring that the total area under the bars equals 1. The stat='density' parameter in sns.histplot normalizes the counts. The bin width is calculated, and the total area of the histogram is verified to confirm normalization.



Conclusion

This experiment provided valuable insights into the dataset through various visualizations. The contingency table highlighted the relationship between PM2.5 levels and AQI categories, demonstrating that lower particulate concentrations correspond to better air quality. The scatter plot revealed a strong correlation between PM2.5 and PM10, reinforcing the idea that both pollutants contribute significantly to air quality degradation. The heatmap further confirmed these correlations, showing strong associations between PM2.5, PM10, and AQI. The box plot helped detect and visualize outliers, emphasizing the need for proper handling to ensure accurate statistical analysis. Identifying and managing these extreme values is crucial for maintaining data integrity. Overall, this study reinforced the importance of exploratory data analysis (EDA) in understanding air pollution trends, detecting patterns, and ensuring data-driven decision-making for effective environmental monitoring.