Team Mavericks

Members:

Rushabh Wakekar CSE(DS)

Pranav Bhusari CSE(DS)

Problem Statement:

Air Quality Index Prediction

Description: Work with environmental datasets to predict the air quality index (AQI) in specific regions. Include visualizations of pollution trends and health recommendations.

Proposed Solution:

To address the problem, we first categorized the regions based on their Air Quality Index (AQI) values using a bucketing approach, effectively grouping them into predefined buckets such as Satisfactory, Good, Moderate, Poor, Very Poor and Severe. This classification enabled us to understand the severity of air pollution in each region. Building on this, we introduced two new columns: Health Advisory, which provides tailored guidance for public safety based on the AQI bucket (e.g., precautions like wearing masks or limiting outdoor activities), and Suggested Solution, which recommends region-specific actions to improve air quality (e.g., reducing vehicle emissions or banning waste burning). This comprehensive approach combines prediction with actionable insights to address the air quality challenge.

1. Health Advisory –
   This column provides a health-related advisory based on the predicted Air Quality Index (AQI) bucket. The advisories are tailored to inform people about the possible health effects associated with the AQI category, such as "Wear a mask" for poor air quality or "No precautions needed" for good air quality.
2. Suggested Solution –
   This column lists actionable solutions to improve air quality in the respective regions. Solutions are based on the AQI bucket and may include recommendations such as "Reduce vehicle emissions," "Plant more trees," or "Avoid burning waste." These suggestions align with the severity of pollution levels.

Progress Until Now:

We began by cleaning the dataset to ensure data consistency and reliability. Missing values were handled using Python functions such as

- fillna() for imputing missing data with appropriate statistics (e.g., mean or median)
- dropna() for removing rows with extensive missing information.

We also removed irrelevant columns such as:

- Benzene
- Toluene
- Xylene

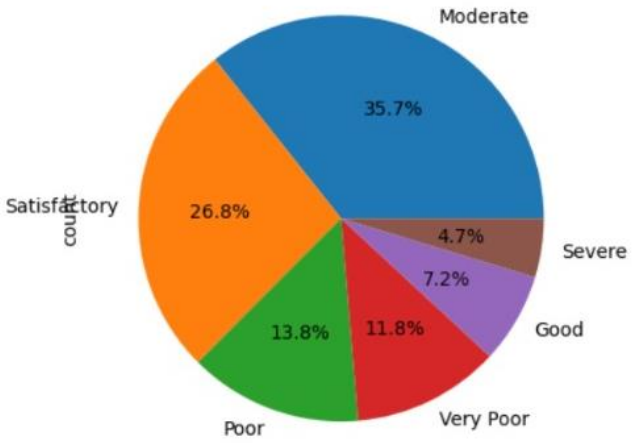
as they were either redundant or not critical to our analysis.

Visualisations on Google Colab:

- Pie Chart of the AQI_Bucket, which shows what percent of various AQI is present in the dataset

```
df['AQI_Bucket'].value_counts().plot.pie(autopct="%1.1f%%")
```

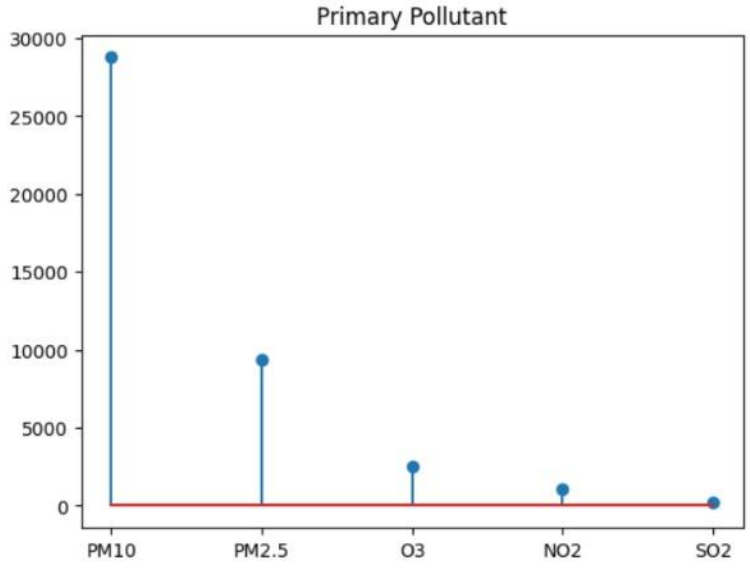
```
<Axes: ylabel='count'>
```



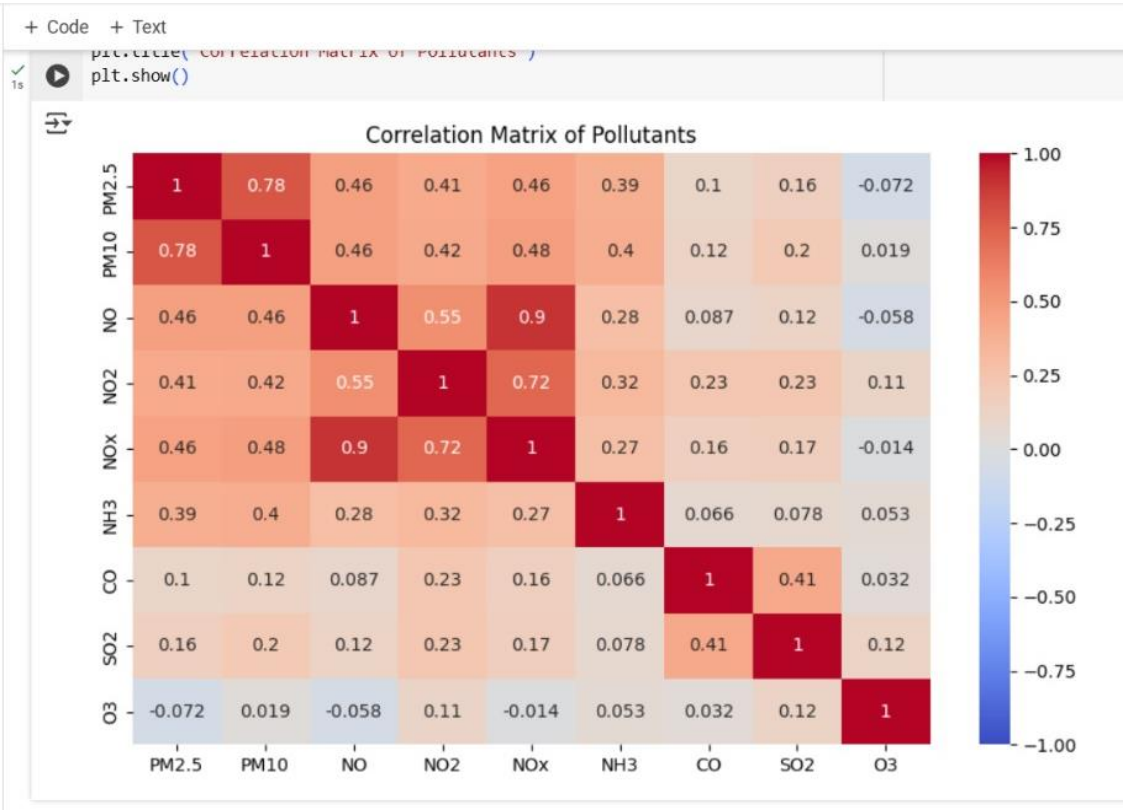
- Stem plot of Primary Pollutant

```
plt.stem(a.index, a)
plt.title("Primary Pollutant")
```

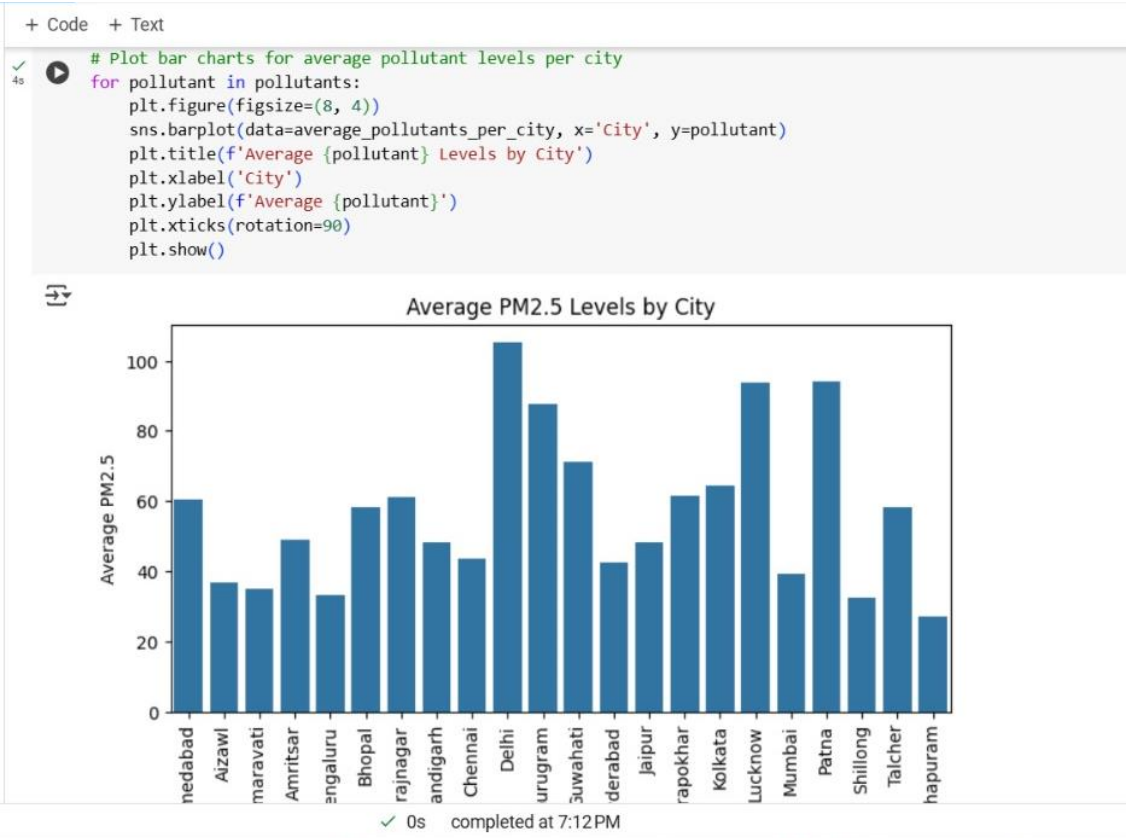
```
Text(0.5, 1.0, 'Primary Pollutant')
```



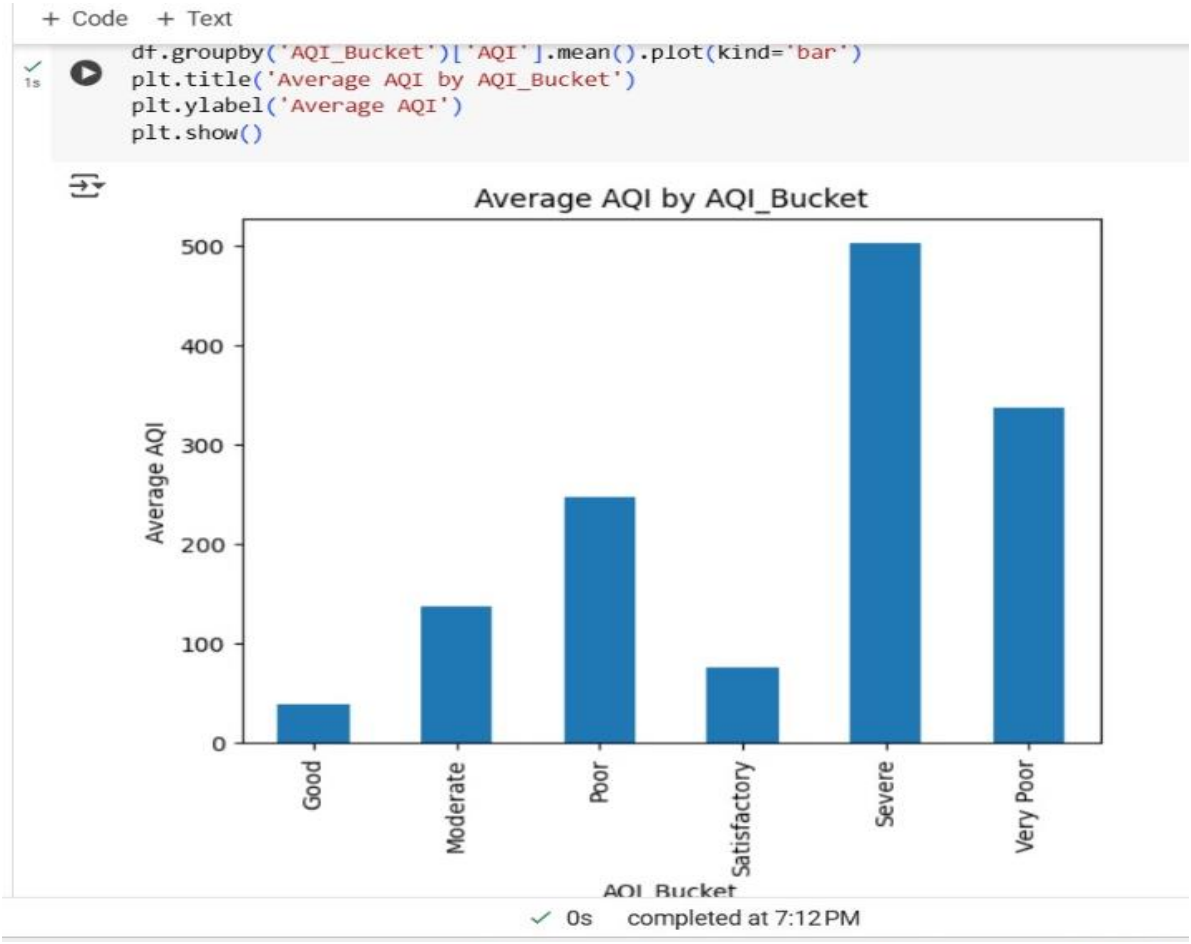
- Confusion Matrix showing the Correlation between Pollutants



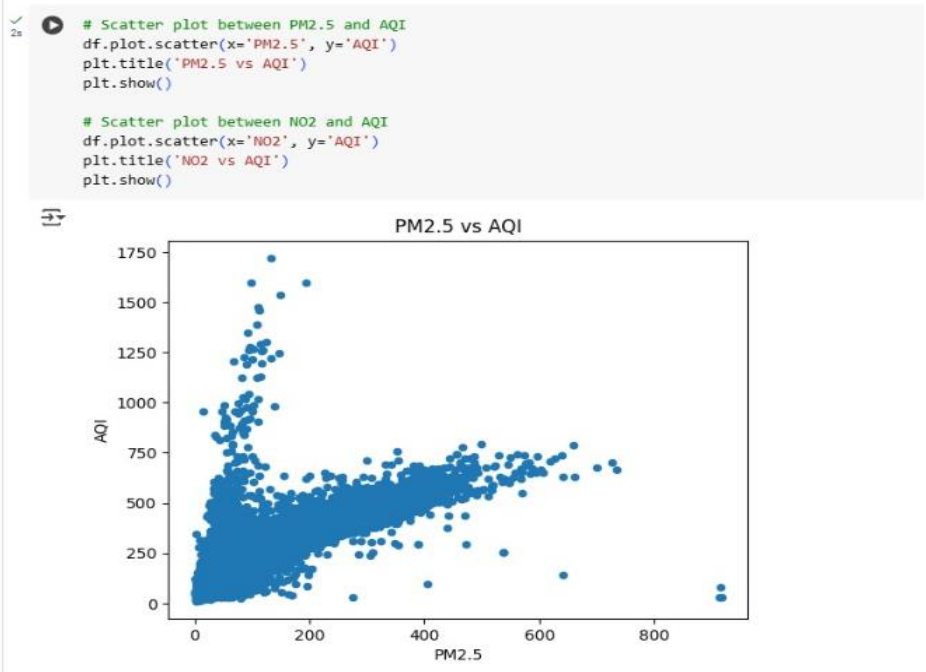
- Bar Chart showing Average PM levels by City

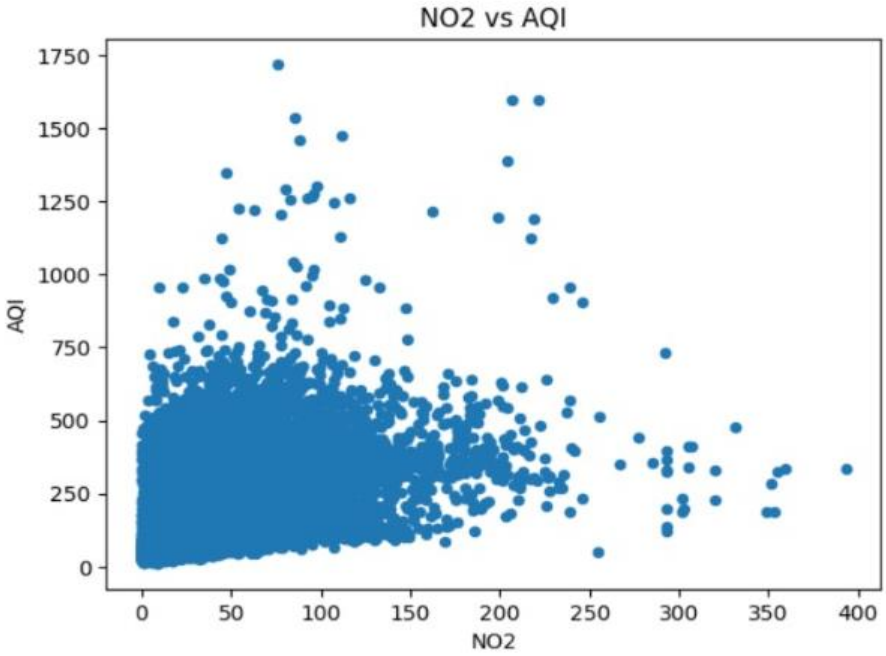


- Bar Plot showing Average AQI ain context to AQI_Bucket column.



- Scatterplots showing the PM2.5 and NO2 vs AQI.



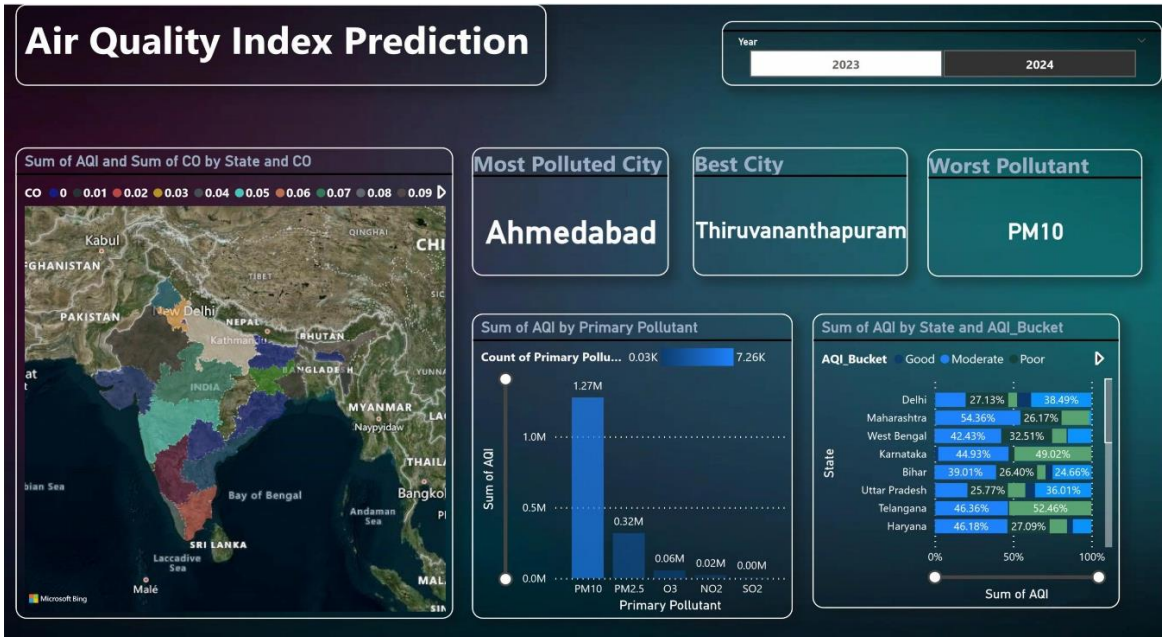


Libraries used for Data Visualisation in Colab:

The following libraries were utilized for visualizing the data:

- Matplotlib
- Seaborn
- Pandas Visualization

Visualisations on Power BI:



The above dashboard from Power BI shows the relevant visualisations such as Sum of primary pollutant and AQI bucket. It also includes the best and worst cities in context with pollution.

The above dashboard also contains various insightful visualisations such of Sum of AQI by City, Sum of O3 by Primary Pollutant.

Model Building Approach:

To build the model, we plan an approach where first, a Linear Regression model will predict the AQI values based on relevant input features such as PM2.5 and PM10 concentrations. Second, the model will be trained further using the Random Forest Algorithm to enhance accuracy and robustness in predicting AQI categories and corresponding solutions. This combined approach ensures reliable AQI prediction and actionable recommendations tailor to regional air quality conditions.