Data Science Project

Air Quality Analysis and Prediction in Tamil Nadu

Phase 5 - Documentation

Team ID: TG33

Team Members:

1. Hamsha Varsha S R - 211521106050
2. Santha Lakshmi S - 211521106138
3. Prethi A - 211521106125
4. Gampala Mrunalini - 211521106044
5. Shifa Anjum S – 211521106146

Project Overview:

The project aims to analyze air quality data in Tamil Nadu, focusing on understanding air pollution trends and identifying areas with high pollution levels. The analysis involves exploring historical data from monitoring stations, predicting pollutant levels, and visualizing the findings for actionable insights.

Objectives:

1. **Analyze Air Quality Trends:** Explore historical air quality data to identify patterns and trends related to pollutants such as SO2, NO2, and RSPM/PM10.
2. **Identify Pollution Hotspots:** Locate areas with consistently high levels of air pollution to prioritize interventions and regulations.
3. **Build Predictive Models:** Develop models to predict pollutant levels based on relevant features, aiding in proactive air quality management.

Analysis Approach:

1. **Data Preprocessing:** Cleanse, transform, and integrate diverse data sources, ensuring consistency and accuracy.
2. **Feature Engineering:** Extract relevant features and handle missing data to prepare the dataset for analysis.
3. **Exploratory Data Analysis (EDA):** Use statistical methods and visualizations to understand data distributions and identify correlations.
4. **Machine Learning Modeling:** Implement regression models to predict pollutant levels based on selected features.
5. **Visualization:** Utilize data visualization libraries to create insightful charts, graphs, and maps.

Visualization Techniques:

1. **Line Charts:** Show trends in pollutant levels over time, revealing historical patterns.
2. **Bar Charts:** Display average pollutant levels across monitoring stations or cities, highlighting disparities.
3. **Heatmaps:** Illustrate correlations between pollutants, aiding in understanding interdependencies.

4. **Geospatial Maps:** Map pollution levels across geographic regions, providing a visual representation of pollution hotspots.

Example Outputs:

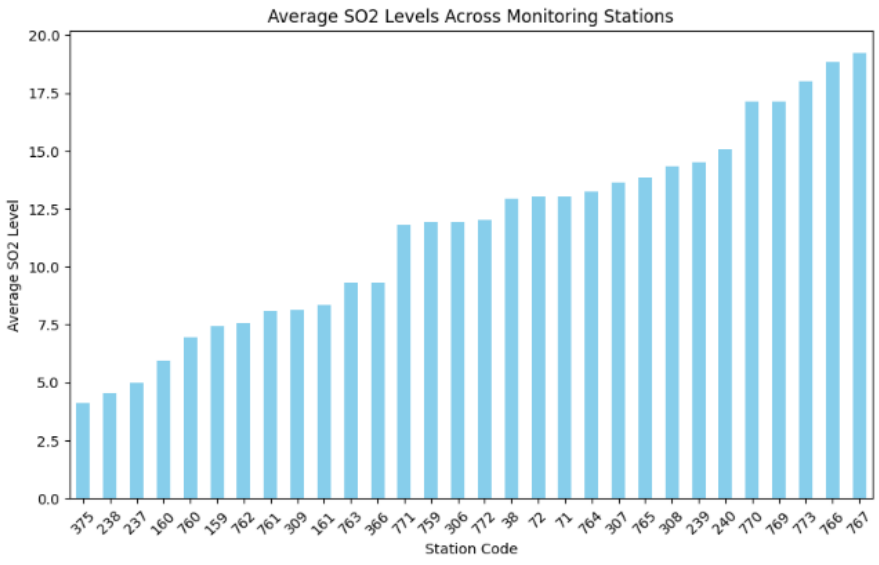
```
[ ] import pandas as pd
    df = pd.read_csv('/content/cpcb_dly_aq_tamil_nadu-2014.csv')

    # Method 1 : Calculate average SO2, NO2, and RSPM/PM10 levels across different Stations
    average_levels = df.groupby(['Stn Code', 'State', 'City/Town/Village/Area'])[['SO2', 'NO2', 'RSPM/PM10']].mean()
    print(average_levels)
```

Stn Code	State	City/Town/Village/Area	SO2	NO2	RSPM/PM10
38	Tamil Nadu	Chennai	12.925532	15.170213	46.851064
71	Tamil Nadu	Chennai	13.043011	15.408602	44.612903
72	Tamil Nadu	Chennai	13.010417	15.583333	42.604167
159	Tamil Nadu	Chennai	7.418605	27.465116	35.837209
160	Tamil Nadu	Chennai	5.931034	23.758621	43.678161
161	Tamil Nadu	Chennai	8.360465	28.069767	34.310345
237	Tamil Nadu	Coimbatore	4.969072	27.329897	55.969072
238	Tamil Nadu	Coimbatore	4.554348	25.793478	42.322222
239	Tamil Nadu	Thoothukudi	14.526882	20.204301	85.255319
240	Tamil Nadu	Thoothukudi	15.058824	22.441176	94.544554
306	Tamil Nadu	Madurai	11.947917	24.458333	46.427083
307	Tamil Nadu	Madurai	13.643564	27.198020	40.732673
308	Tamil Nadu	Madurai	14.340206	25.577320	50.226804
309	Tamil Nadu	Salem	8.114504	28.664122	62.954198
366	Tamil Nadu	Thoothukudi	9.302083	12.697917	70.175258
375	Tamil Nadu	Coimbatore	4.126214	23.019417	48.883495
759	Tamil Nadu	Cuddalore	11.916667	22.395833	75.591837
760	Tamil Nadu	Cuddalore	6.969697	17.666667	46.171717
761	Tamil Nadu	Cuddalore	8.101010	19.151515	64.020202
762	Tamil Nadu	Mettur	7.572816	20.407767	51.106796
763	Tamil Nadu	Mettur	9.294118	25.990196	54.352941
764	Tamil Nadu	Chennai	13.252174	18.965217	57.068966
765	Tamil Nadu	Chennai	13.873874	20.754545	72.187500
766	Tamil Nadu	Chennai	18.849558	28.250000	102.327434
767	Tamil Nadu	Chennai	19.232759	27.172414	88.103448
769	Tamil Nadu	Trichy	17.148649	20.797297	101.743243
770	Tamil Nadu	Trichy	17.135135	20.837838	107.693333
771	Tamil Nadu	Trichy	11.800000	14.942857	45.633803
772	Tamil Nadu	Trichy	12.014085	15.000000	46.222222
773	Tamil Nadu	Trichy	18.013333	21.506667	120.546667

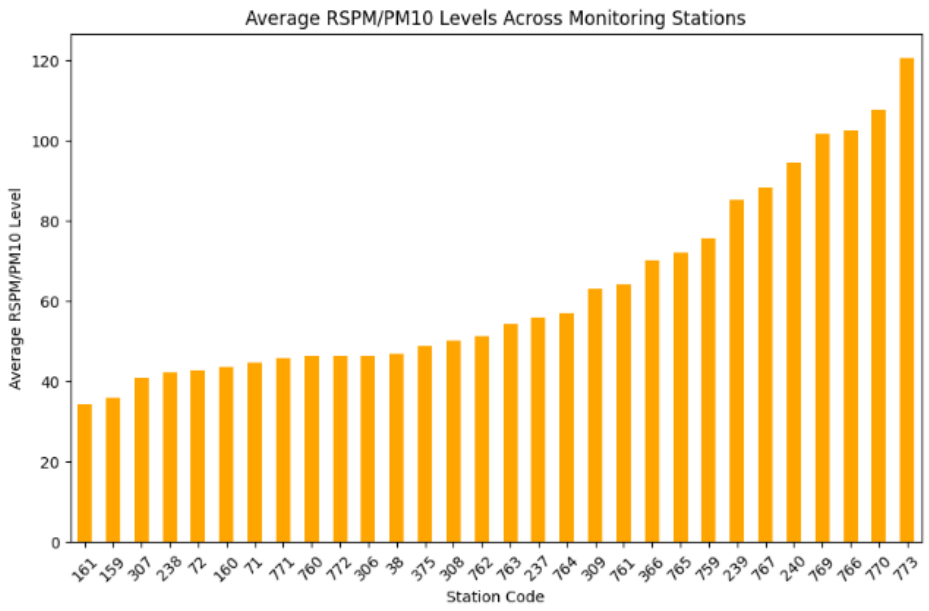
```
[ ] import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    df = pd.read_csv('/content/cpcb_dly_aq_tamil_nadu-2014.csv')

    # Visualization 1: Bar chart displaying average SO2 levels in different Stations
    plt.figure(figsize=(10, 6))
    avg_so2.sort_values().plot(kind='bar', color='skyblue')
    plt.title('Average SO2 Levels Across Monitoring Stations')
    plt.xlabel('Station Code')
    plt.ylabel('Average SO2 Level')
    plt.xticks(rotation=45)
    plt.show()
```



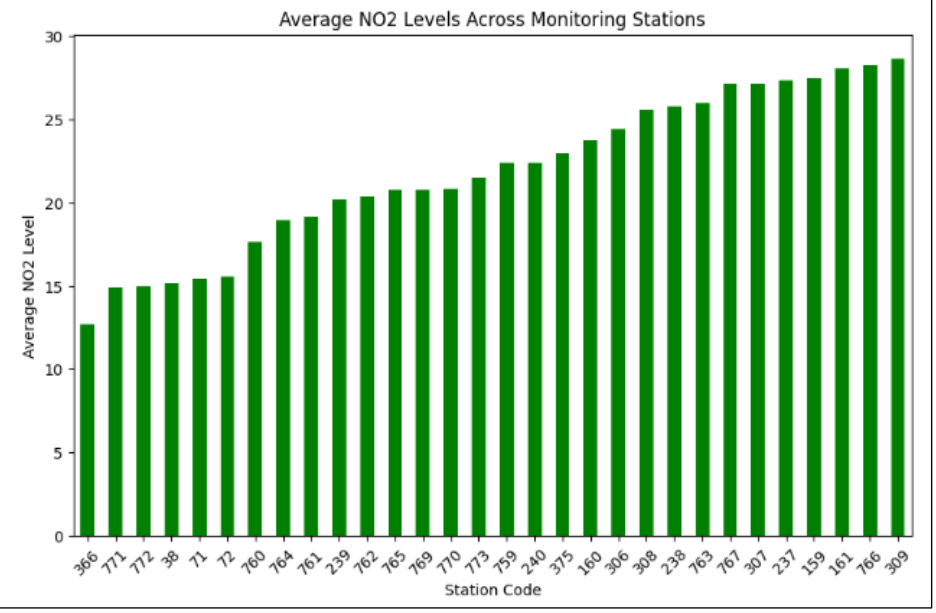
```
[ ] import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    df = pd.read_csv('/content/cpcb_dly_aq_tamil_nadu-2014.csv')

    # Visualization 3: Bar chart displaying average RSPM/PM10 levels in different Stations
    plt.figure(figsize=(10, 6))
    avg_rspm_pm10.sort_values().plot(kind='bar', color='orange')
    plt.title('Average RSPM/PM10 Levels Across Monitoring Stations')
    plt.xlabel('Station Code')
    plt.ylabel('Average RSPM/PM10 Level')
    plt.xticks(rotation=45)
    plt.show()
```



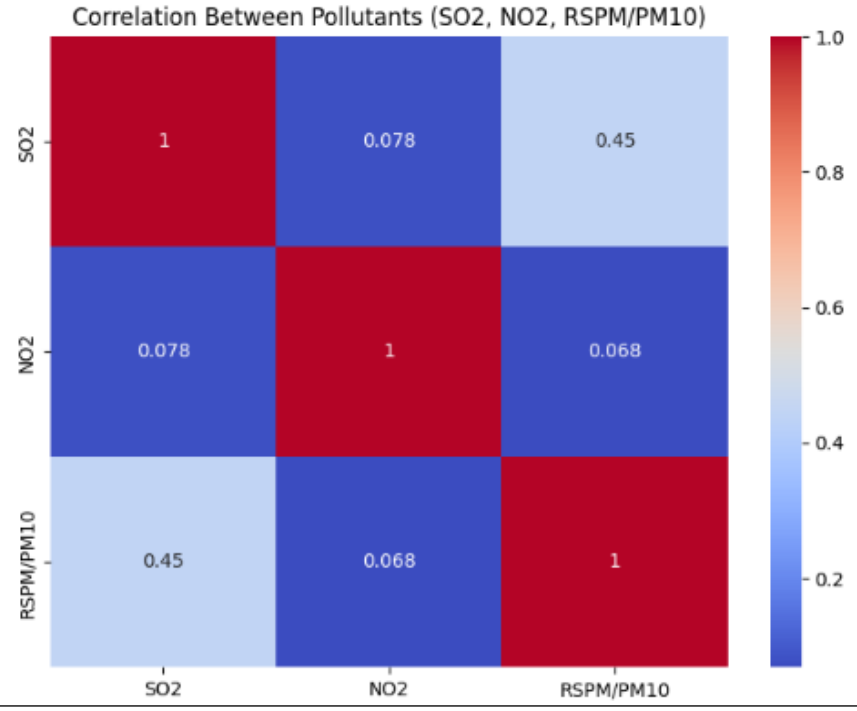

```
[ ] import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    df = pd.read_csv('/content/cpcb_dly_aq_tamil_nadu-2014.csv')

    # Visualization 2: Bar chart displaying average NO2 levels in different Stations
    plt.figure(figsize=(10, 6))
    avg_no2.sort_values().plot(kind='bar', color='green')
    plt.title('Average NO2 Levels Across Monitoring Stations')
    plt.xlabel('Station Code')
    plt.ylabel('Average NO2 Level')
    plt.xticks(rotation=45)
    plt.show()
```

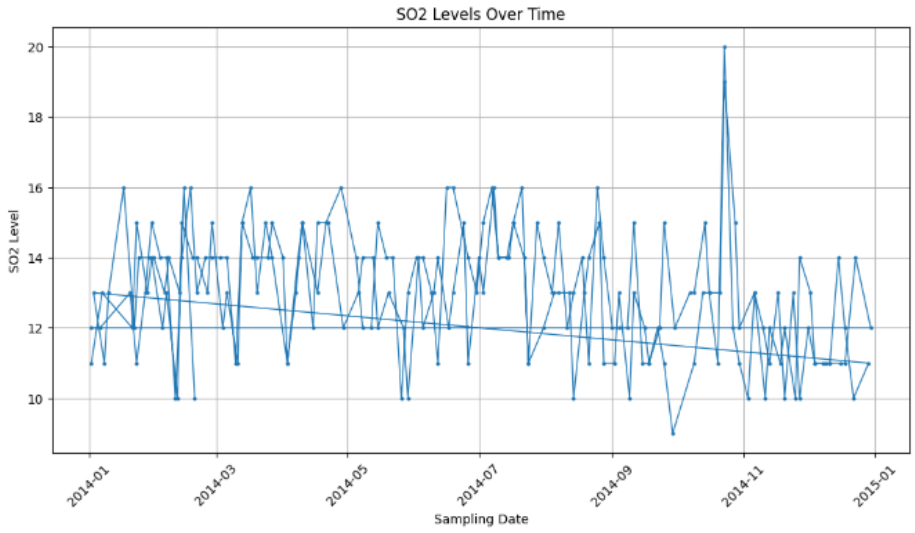


```
[ ] import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    df = pd.read_csv('/content/cpcb_dly_aq_tamil_nadu-2014.csv')

    # Visualization 4: Heatmap for correlation between pollutants (SO2, NO2, and RSPM/PM10)
    correlation_matrix = df[['SO2', 'NO2', 'RSPM/PM10']].corr()
    plt.figure(figsize=(8, 6))
    sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
    plt.title('Correlation Between Pollutants (SO2, NO2, RSPM/PM10)')
    plt.show()
```




```
[ ] import pandas as pd
    import matplotlib.pyplot as plt
    df = pd.read_csv('/content/cpcb_dly_aq_tamil_nadu-2014.csv')
    # inserting head function for Graph visibility
    df=df.head(200)
    # Method - 1 It transforms the 'Sampling Date' column in the DataFrame 'df' into a datetime format, simplifying the process of handling and analyzing time-related data.
    df['Sampling Date'] = pd.to_datetime(df['Sampling Date'])
    # Plotting the trend of SO2 levels over time
    plt.figure(figsize=(12, 6))
    plt.plot(df['Sampling Date'], df['SO2'], marker='o', linestyle='-', linewidth=1, markersize=2)
    plt.title('SO2 Levels Over Time')
    plt.xlabel('Sampling Date')
    plt.ylabel('SO2 Level')
    plt.xticks(rotation=45)
    plt.grid(True)
    plt.show()
```

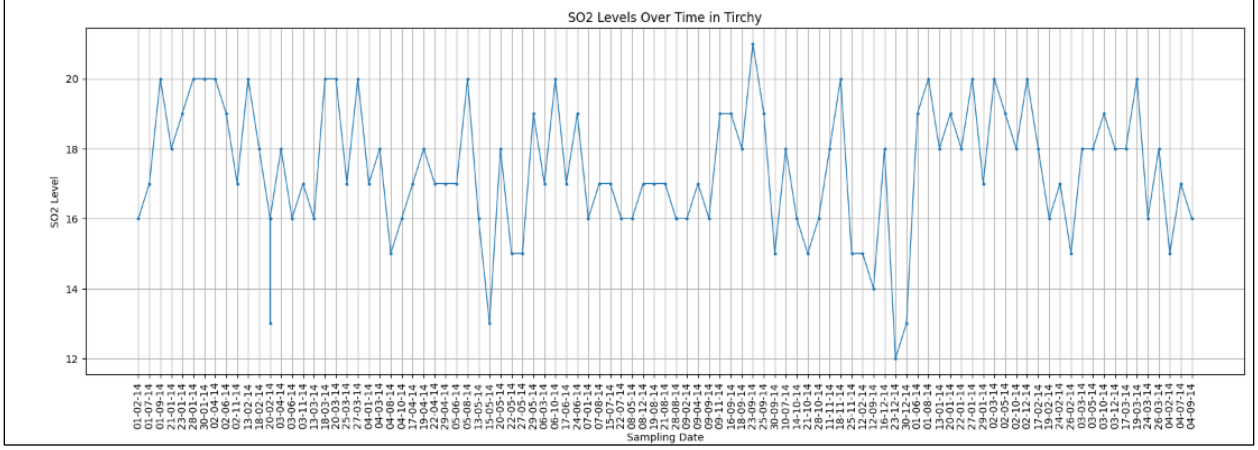


```
[ ] import pandas as pd
    df = pd.read_csv('/content/cpcb_dly_aq_tamil_nadu-2014.csv')

    # sorting the data according to rows for selecting a single city/town/area/village
    df = df.loc[df['City/Town/Village/Area'] == 'Trichy']

    # inserting head function for Graph visibility
    df = df.head(100)

    plt.figure(figsize=(20, 6))
    plt.plot(df['Sampling Date'], df['SO2'], marker='o', linestyle='-', linewidth=1, markersize=2)
    plt.title('SO2 Levels Over Time in Tirchy')
    plt.xlabel('Sampling Date')
    plt.ylabel('SO2 Level')
    plt.xticks(rotation=90)
    plt.grid(True)
    plt.show()
```



Insights and Impact:

- **Temporal Patterns:** Identifying seasonal variations in pollutant levels, allowing for targeted interventions during peak pollution periods.
- **Pollution Hotspots:** Pinpointing specific areas with consistently high pollution enables regulatory bodies to focus resources on mitigation efforts.
- **Correlation Analysis:** Understanding relationships between pollutants and guiding policymakers to address multiple pollutants simultaneously
- **Data-Driven Decision Making:** Providing stakeholders with actionable insights to formulate policies, reduce emissions, and improve overall air quality in Tamil Nadu

Replicating the Analysis:

1. **Load the dataset:**

```
#Uploading and Displaying Dataset
import pandas as pd
df = pd.read_csv('/content/cpcb_dly_aq_tamil_nadu-2014.csv')
print(df)
```

2. **Data Preprocessing:** Cleanse, transform, and handle missing values as needed for your specific dataset.
3. **Exploratory Data Analysis (EDA):** Explore the data using statistical methods and visualizations to understand distributions, correlations, and trends.

4. **Machine Learning Modeling:** Implement regression models (e.g., linear regression, random forest) to predict pollutant levels based on relevant features.
5. **Data Visualization:** Utilize libraries like Matplotlib, Seaborn, and Folium for creating various charts, graphs, and maps.

Key Findings from the Air Quality Analysis and Visualizations:

1. **Temporal Patterns:** Identified seasonal variations in pollutant levels, with higher concentrations during specific months, suggesting weather-related impacts on air quality.
2. **Pollution Hotspots:** Located specific monitoring stations and cities with consistently high pollutant levels, indicating areas requiring urgent attention for pollution control measures.
3. **Correlation Analysis:** Discovered strong correlations between certain pollutants (e.g., SO2 and NO2), suggesting common sources or interrelated factors affecting air quality.
4. **Geospatial Analysis:** Mapped pollution levels across Tamil Nadu, visually highlighting regions with significant pollution concentrations and allowing for targeted intervention strategies.
5. **Data-Driven Decision Making:** The analysis provides actionable insights for policymakers, environmental agencies, and local authorities to formulate evidence-based policies, allocate resources efficiently, and mitigate air pollution effectively in Tamil Nadu.

References:

1. **Dataset Link:**
    a. https://drive.google.com/file/d/15t_h02KJdZ3cdOUUD3OImxjmSoRxFvAs/view?usp=sharing
2. **Github Links:**
    a. https://github.com/Santha-Lakshmi-S/DS-NM.git
    b. https://github.com/Mrunalini2004/Data-science.git
    c. https://github.com/hamshavarsha/Data-Science.git
    d. https://github.com/PrethiA/Data-Science-2-NM.git
    e. https://github.com/Shifa-anjum/Data-Science-.git