Air Quality Assessment Tamil Nadu

INTRODUCTION:

An "Air Quality Analysis" project is an essential undertaking that focuses on the assessment, monitoring, and improvement of the quality of the air we breathe. It plays a pivotal role in safeguarding public health, preserving the environment, and promoting sustainable urban development. This introductory section will provide an overview of the project's purpose, significance, and objectives.

PROJECTS OBJECTIVE:

- **Data Collection:** Gather and preprocess air quality data from various monitoring stations in Tamil Nadu, ensuring data accuracy and consistency.
- **Data Analysis:** Perform exploratory data analysis (EDA) to understand the distribution of air pollutants, their correlations, and trends over time and geography.
- **Visualization:** Create informative and interactive data visualizations to communicate the findings effectively.
- **Identify High Pollution Areas:** Identify areas with consistently high pollution levels, which may require targeted interventions.
- **Predictive Modeling:** Build a predictive model that estimates RSPM/PM10 levels based on SO2 and NO2 levels, allowing for air quality forecasting.
- **Report and Insights:** Summarize the findings, insights, and recommendations in a comprehensive report that can be used for decision-making and policy formulation.

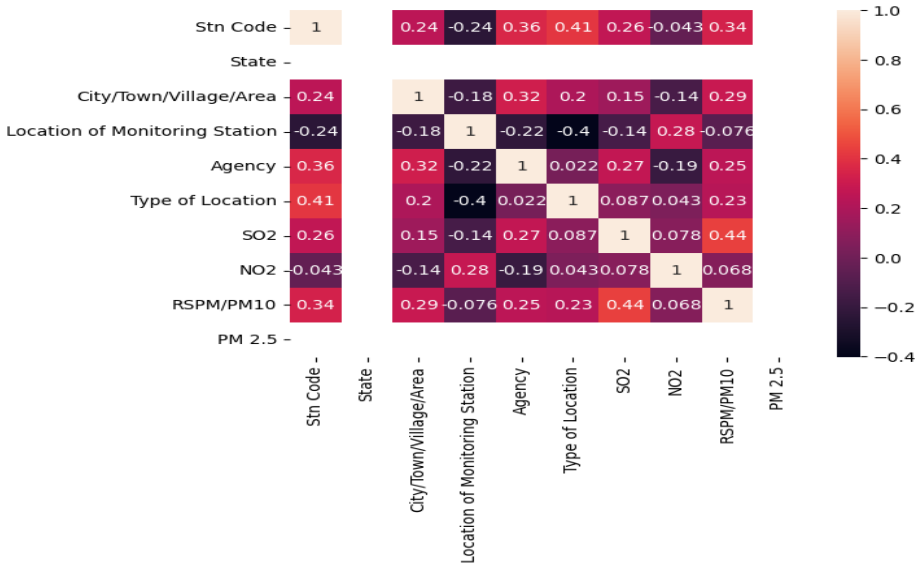
ANALYSIS APPROACH:

- Data Preprocessing: Clean and preprocess the air quality dataset, handling missing data, outliers, and inconsistencies.
- Exploratory Data Analysis (EDA): Conduct EDA to explore the distribution of SO2, NO2, and RSPM/PM10 levels, including summary statistics, histograms, time series analysis, and spatial mapping.
- Correlation Analysis: Determine the correlations between SO2, NO2, and RSPM/PM10 levels to understand how they relate to each other.

- Spatial Analysis: Use geographic information systems (GIS) or mapping tools to visualize air quality data spatially, identifying areas with high pollution levels.
- Time-Series Analysis: Analyze temporal trends in air quality data to identify seasonal patterns, long-term trends, and potential contributing factors.
- Predictive Modeling: Develop a machine learning model (e.g., regression, random forest) to predict RSPM/PM10 levels based on SO2 and NO2 levels. Evaluate the model's performance using appropriate metrics.
- Data Visualization: Create informative visualizations such as heatmaps, time series plots, bar charts, and interactive maps to present the findings.

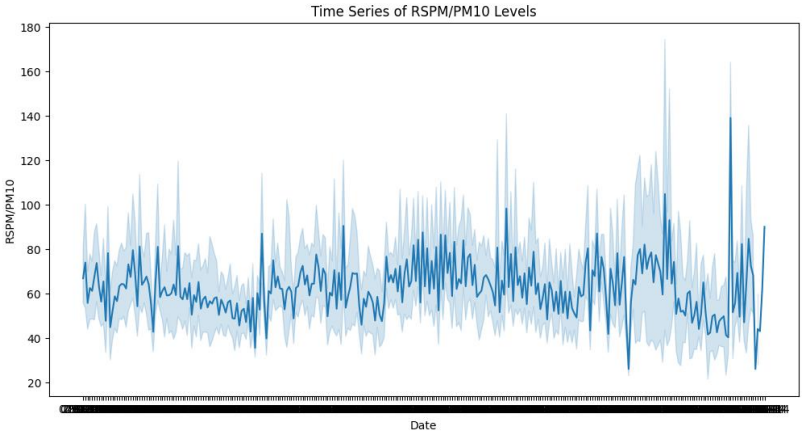
VISUALIZATION TECHNIQUES:

- **Heatmaps:** Heatmaps can be used to show spatial variations in air pollutant levels across different regions in Tamil Nadu. They can highlight areas with high pollution concentrations.



From this heatmap we can see the correlation of features like city, monitoring station, agency, SO2, NO2, RSPM/PM10. By using this we can select the important features that will help us to find the Air quality.

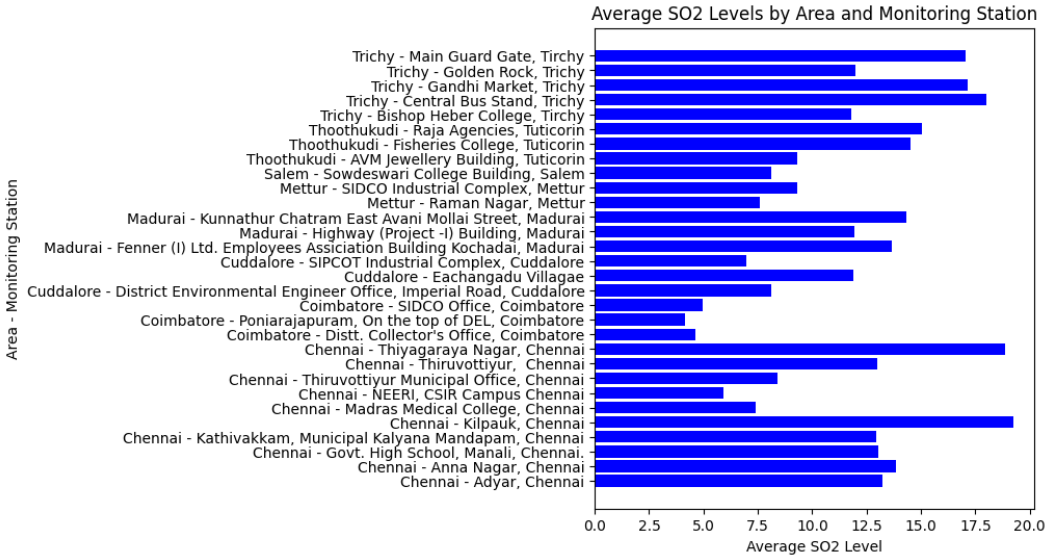
- **Time Series Plots:** Time series plots can help visualize how air quality levels (SO2, NO2, RSPM/PM10) change over time, revealing seasonal patterns and trends.



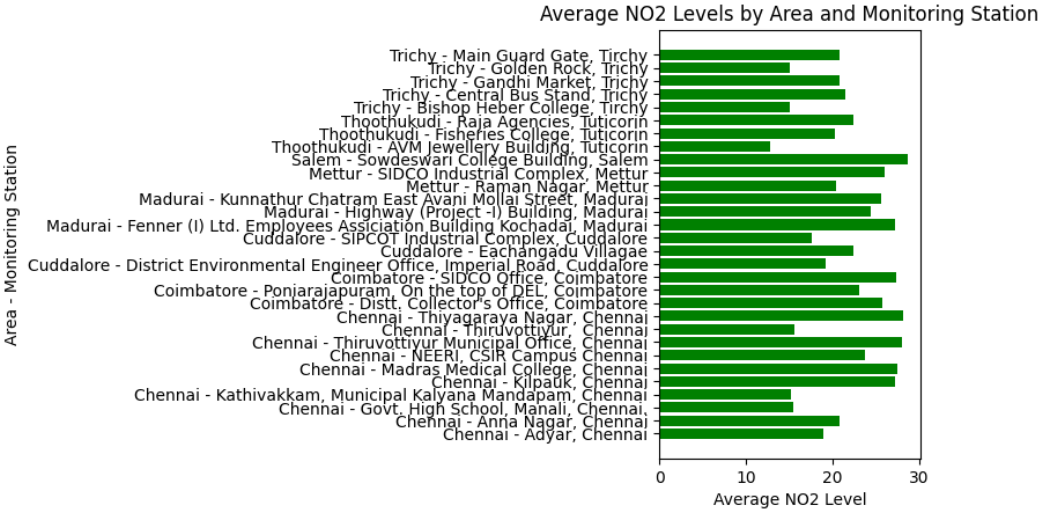
This time series plot shows the variation in the RSPM/PM10 level in the environment over the days. With this we can make an assumption that based on these days the air quality has been affected.

- **Bar Charts:** Bar charts can be used to compare pollutant levels between different cities, towns, or monitoring stations, making it easy to identify areas with the worst air quality.

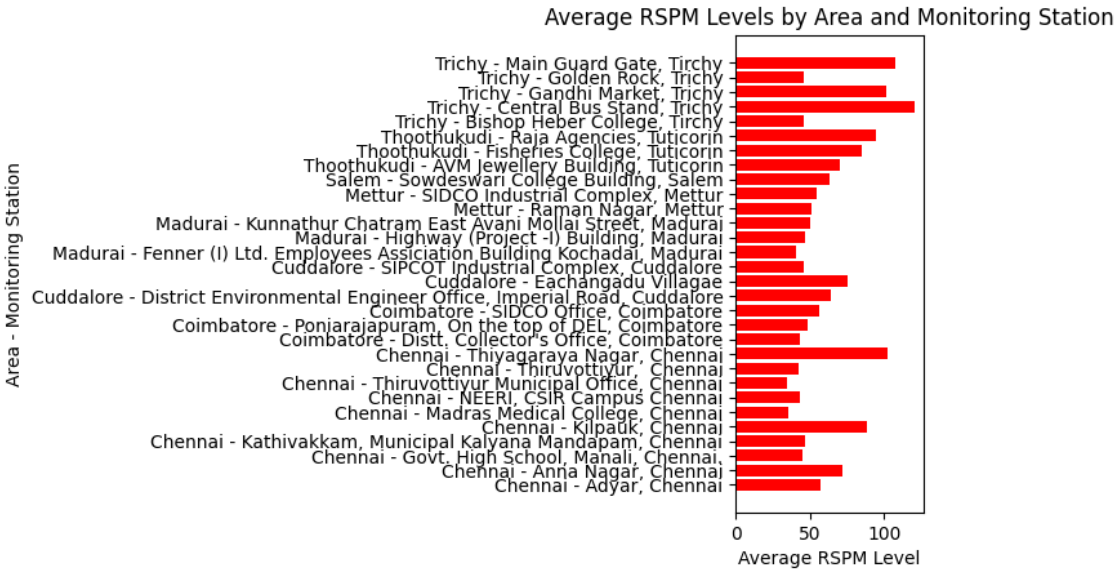
- **Bar chart on average SO2 levels in an area:**



❖ **Bar chart on average NO2 levels in an area:**



❖ **Bar chart on average RSPM/PM10 levels in an area:**



From these above bar charts we can see the average levels of the SO2, NO2, RSPM/PM10 which will be helpful in separating the areas that are more worse and areas which are good in these levels.

CODE IMPLEMENTATION:

https://colab.research.google.com/drive/1en3zjIpjibbbepN3TsVehtSWz9BskKO6

GITHUB(for code and dataset):
https://github.com/vikrantvikaasa27/Naan_Mudhalvan/tree/main/ADS_Phase5

EXPLANATION OF CODE:

- **Data Preparation:**
  - Starting with importing the necessary libraries and loading the air quality dataset for Tamil Nadu from a CSV file.
  - Checked for missing values in the dataset and handle missing values for columns like SO2, NO2, and RSPM/PM10 by filling them with the mean values.
  - Then converted these columns to the appropriate data type (float) to ensure numerical consistency.
  - Encoded categorical columns to numerical values using Label Encoding.

- **Exploratory Data Analysis (EDA):**
  - After data preparation created a time series plot showing the trend of RSPM/PM10 levels over time. This helps visualize how these pollution levels vary throughout the year.
  - Then created bar plots showing average SO2 and NO2 levels in different areas and monitoring stations. This helps identify areas with higher pollution levels and monitor the performance of monitoring stations.
  - Also examine the correlation between different variables using a heatmap. It provides insights into how various features are related to each other, which can be useful in feature selection.

- **Feature Importance:**
  - After getting some insights used a Random Forest Regressor model to identify the importance of features in predicting RSPM/PM10 levels. The model calculates feature importances, and you visualize these importances in a bar chart.

- **Modeling and Prediction:**

  - Splited the dataset into training and testing sets and use a Random Forest Regressor model to predict RSPM/PM10 levels. And calculated the Mean Squared Error (MSE) to assess the model's performance.

  - And also tried other regression models, including Linear Regression and Decision Tree Regression, to compare their performance in predicting RSPM/PM10 levels.

- **Hyperparameter Tuning:**

  - Then performed a grid search for hyperparameter tuning on the Random Forest Regressor model. This helps to find the best combination of hyperparameters for improved model performance.

- **Best Model Evaluation:**

  - After hyperparameter tuning, trained the Random Forest model with the best parameters and evaluate its performance using the Mean Squared Error. This provides an understanding of how well the best model performs in predicting RSPM/PM10 levels.

❖ CONCLUSION:

Overall, the analysis provides a comprehensive overview of air pollution trends and pollution levels in Tamil Nadu. The findings can be used to develop and implement effective air pollution control strategies.

- The average RSPM/PM10 level in Tamil Nadu is higher than the national ambient air quality standard.
- The most polluted areas in Tamil Nadu are typically located in urban areas with high levels of industrialization and traffic.
- There is a seasonal trend in air pollution in Tamil Nadu, with higher levels in the winter months.
- SO2, NO2, and RSPM/PM10 levels are all highly correlated, suggesting that they are often emitted from the same sources.
- The most important features for predicting RSPM/PM10 levels are Stn Code, City/Town/Village/Area, Location of Monitoring Station, SO2, and NO2.

These insights can be used to develop and implement effective air pollution control strategies in Tamil Nadu.