PHASE 5

PROJECT DOCUMENTATION AND SUBMISSION

Date	01-11-2023
Team ID	719
Project Name	Air Quality Analysis in Tamil Nadu

Project Title: Air Quality Analysis in Tamil Nadu

S.NO	TABLE OF CONTENT	PG.NO
1	Project Overview	2
2	Problem Statement	2
3	Problem Identified	2
4	Introduction	2
5	Literature Survey	3
6	Design Thinking Approach	9
7	Steps Involved In Model Evaluation	11
8	Tools And Libraries	13
9	Model Selection And Training:	14
10	Pollution Trends And Areas	15
11	Data Visualization	16
12	Data Visualization With IBM Cognos	21
13	How Insight Improve User Experience:	24
14	Conclusion	25

1. PROJECT OVERVIEW:

The aim of this project is to analyze and visualize air quality data obtained from monitoring stations in Tamil Nadu. The primary objectives are as follows:

- **Air Quality Analysis:** Perform a comprehensive analysis of the air quality data to understand pollution trends.
- **Visualization:** Create data visualizations to present the insights effectively.
- **Predictive Modeling:** Develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 measurements

2. PROBLEM STATEMENT:

Objective: The objective of this project is to analyze and visualize air quality data from various monitoring stations in Tamil Nadu.

Data: The dataset contains measurements of Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2), and Respirable Suspended Particulate Matter/Particulate Matter 10 (RSPM/PM10) levels in different cities, towns, villages, and areas

3. PROBLEM IDENTIFIED:

The project aims to analyze and visualize air quality data from Tamil Nadu monitoring stations to understand air pollution trends, identify high pollution areas, and develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 levels. A systematic approach can be followed to achieve these objectives, including understanding historical air quality trends, identifying high pollution areas, and determining key metrics and outcomes.

4. INTRODUCTION:

The air quality analysis in Tamil Nadu is a crucial initiative to assess the state's air quality conditions, as the state faces increasing pollution challenges. The analysis aims to systematically assess data from monitoring stations, understand patterns, identify areas with poor air quality, and develop predictive models for future pollution levels. This analysis is vital for improving the quality of life for Tamil Nadu residents and safeguarding the environment.

The "Air Quality Analysis in Tamil Nadu" project is a comprehensive initiative to address the pressing global issue of air pollution. The project collects, analyzes, and visualizes air quality data from monitoring stations in Tamil Nadu, aiming to gain insights into pollution trends, identify areas with high pollution levels, and develop a predictive model for estimating RSPM/PM10 levels based on Sulphur dioxide and Nitrogen dioxide concentrations.

Analyzing Air Pollution Trends:

- The first objective of the project is to comprehensively analyze historical air quality data collected from monitoring stations across Tamil Nadu. This analysis will provide a deep understanding of air pollution trends over time. By studying long-term data patterns and variations, the project aims to identify whether air quality is improving or deteriorating in specific regions of the state. This knowledge is essential for setting benchmarks and evaluating the effectiveness of pollution control measures.

Identifying Areas with High Pollution Levels:

- The second objective is to identify areas within Tamil Nadu that consistently experience high levels of air pollution. By geospatially analyzing the data, the project can create maps and visualizations to pinpoint pollution hotspots. This spatial analysis is critical for targeted intervention and mitigation strategies. It helps authorities allocate resources and take precise actions to reduce pollution in the areas most affected, thus improving the overall quality of life for the population.

Developing a Predictive Model for RSPM/PM10:

- A key component of this project is the development of a predictive model. This model will use the concentrations of Sulfur Dioxide (SO2) and Nitrogen Dioxide (NO2) as input variables to estimate the levels of Respirable Suspended Particulate Matter (RSPM/PM10). This predictive model will be a valuable tool for forecasting air quality. It enables authorities to anticipate pollution levels and take preventive measures to protect public health. It also offers a means of early warning and preparation for residents, especially those in high-risk areas.

5. LITERATURE SURVEY:

- **"A SURVEY ON AIR POLLUTION AND HEALTH RISK IN TAMILNADU DISTRICTS"**

Introduction:

The document introduces the topic of air pollution and its significance. It emphasizes the importance of studying air quality due to its direct impact on human health and the environment.

Air Pollution Terminology:

This section defines and explains various terms related to air pollution, including different types of particles and gases commonly found in the atmosphere.

Air Pollution Trends:

The paper presents data on the trends of air pollution in India, particularly focusing on levels of particulate matter (PM), sulphur dioxide (SO2), and nitrogen dioxide (NO2).It mentions the increase in PM levels, especially in northern Indian cities, and highlights vehicular emissions as a major contributor to air pollution.

Air Pollution in Tamil Nadu:

Specific information about air pollution in the state of Tamil Nadu is discussed. Chennai is identified as a city with high levels of Respirable Suspended Particulate Matter (RSPM) in 2008.Tuticorin is noted as a hotspot for sulfur dioxide (SO2) pollution.

Transport and Air Pollution:

The section explores how transportation, specifically the combustion of fossil fuels, contributes significantly to air pollution. Various pollutants emitted from vehicles, such as carbon monoxide and sulfur and nitrogen oxides, are highlighted. The document stresses the importance of addressing vehicular pollution as a critical aspect of improving air quality.

Conclusion:

- ✓ The conclusion underscores the need to ensure the purity of air, water, and soil for a healthier environment.
- ✓ It emphasizes the role of both public awareness and effective policy-making in addressing environmental challenges.

Summary:

In summary, this document provides a comprehensive overview of the issue of air pollution in Tamil Nadu, India, discussing its terminology, trends, and specific challenges faced by the region. It emphasizes the critical importance of addressing air pollution for the well-being of both the environment and human health.

- **AMBIENT AIR QUALITY MONITORING STUDIES IN FOUR SPECIFIC LOCATIONS OF TAMILNADU, INDIA"**

Objectives:

- ✓ The paper aimed to assess air quality in four locations in Tamilnadu, India: Tiruchengode Bus Stand, K.S.R College Campus, Pallipalayam Bus Stop, and Erode Government Hospital. The specific objectives were:
- ✓ Measure mass concentrations of PM10, PM2.5, SO2, NOX, and CO.
- ✓ Compare the results with National Ambient Air Quality Standards (NAAQS).
- ✓ Calculate the Air Quality Index (AQI) for gaseous pollutants and particulate matter.
- ✓ Identify sources of air pollution in these areas.

Findings:

- ✓ PM10 concentrations exceeded limits in all locations.
- ✓ High vehicular density contributed to increased gaseous pollutants.
- ✓ AQI results indicated moderate air pollution in all locations.
- ✓ Traffic surveys showed significant vehicular usage, with automobiles as the main particulate emission source.

Results:

- ✓ PM10 and PM2.5 concentrations exceeded permissible limits.
- ✓ Gaseous pollutants generally stayed within limits, except for Erode Government Hospital.
- ✓ All locations exhibited moderate air pollution based on AQI.

Summary:

The paper conducted air quality monitoring in four Tamilnadu locations, revealing concerns due to PM10 concentrations surpassing limits. Elevated vehicular density was a major pollution factor. AQI confirmed moderate air pollution, prompting recommendations to reduce particulate emissions, particularly from automobiles, for improved public and environmental health.

- **"MONITORING AMBIENT AIR QUALITY STUDY IN ARIYALUR, TAMIL NADU, INDIA"**

Objectives:

The primary objectives of this study were to monitor and analyze ambient air quality in the Ariyalur region of Tamil Nadu, India, during the year 2020. Specifically, the study aimed to assess the levels of various air pollutants, including PM10, PM100, PM2.5, SO2, NO2, and CO, in different locations around Ariyalur. The study also sought to identify potential sources of air pollution in the area.

Findings and Results:

✓ **Air Pollution Sources:**

The study identified several sources of air pollution in the Ariyalur region. These sources included emissions from a local concrete processing plant, vehicular pollution, and anthropogenic activities related to limestone mining.

✓ **Air Quality Data:**

Monthly data collection revealed significant variations in air quality parameters across different locations in Ariyalur. The data showed fluctuations in the levels of PM10, PM100, PM2.5, SO2, NO2, and CO throughout the year.

✓ **Air Quality Impact:**

The study emphasized the negative impact of air pollution on the environment and human health. It highlighted that air pollution in Ariyalur posed a considerable risk to the local population and the ecosystem.

Summary:

According to a 2020 study conducted in Ariyalur, Tamil Nadu, India, air pollution in the area is primarily caused by emissions from a concrete processing plant, vehicular pollution, and anthropogenic activities associated with limestone mining. The study emphasized the importance of taking immediate action to address air quality concerns, emphasizing the need for stricter regulations and environmental guidelines to mitigate the negative effects on the environment and human health. To provide context, the study also referred to other studies on air pollution.

- **"AIR POLLUTION PREDICTION USING MACHINE LEARNING"**

Objectives:

- ✓ The primary objectives of the research are as follows:
- ✓ To monitor air pollutants emitted from industrial sources and predict their future spread.
- ✓ To create a module that uses machine learning models to predict future emission parameter values.
- ✓ To develop a module that simulates the movement of pollutants in the air using air dispersion models with meteorological data.

Findings:

- ✓ Industrial pollution is a significant contributor to environmental pollution.
- ✓ Machine learning models, including K-Nearest Neighbors, Support Vector Regression, Random Forest, Multi-linear Regression, and Multi-layer Perceptron, were implemented to predict emission rates.
- ✓ The Multi-layer Perceptron model exhibited the lowest mean squared error, indicating higher prediction accuracy compared to other models.
- ✓ Gaussian air dispersion models were used to estimate pollutant dispersion from industrial stacks, chosen for its computational efficiency.

Results:

- ✓ The study provides a comparative analysis of different machine learning algorithms' performance in predicting air pollution levels.
- ✓ It presents error rates (root mean squared error) for each machine learning model.
- ✓ The Gaussian air dispersion model was used to map the concentration of pollutants in a grayscale bitmap image, representing high and low concentration areas.

Summary:

The paper discusses the application of machine learning algorithms to predict air pollution levels, focusing on industrial pollution, which is a significant environmental concern. It also study employs various machine learning algorithms to predict pollution levels, and the results are compared to determine the most accurate model. Additionally, it implements a explores the dispersion of pollutants from industrial stacks using air dispersion models. The Gaussian air dispersion model to estimate the spread of pollutants in the air.

- **"COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR PREDICTING AIR POLLUTION LEVELS"**

Objectives:

The objective of this paper is to conduct a comparative analysis of various machine learning techniques to predict air pollution levels, with a specific focus on PM2.5 concentration in Chinese cities. The study aims to assess the performance of different regression models in terms of accuracy, error rates, and processing times.

Findings:

Random Forest Regression Outperforms:

Among the four regression models evaluated (Decision Tree, Random Forest, Multi-Layer Perceptron, and Gradient Boosting), Random Forest Regression consistently performs the best.

Accuracy and Error Rates:

Random Forest Regression demonstrates accurate predictions with lower error rates compared to the other models. Decision Tree Regression has the lowest processing time but higher error rates.

Identification of Peak Values:

Random Forest Regression effectively identifies peak values in air pollution data.

Dataset Size and Historic Data:

Random Forest Regression excels, especially with large datasets and historic data.

Future Research:

The study suggests future research directions, including investigating machine learning techniques in multi-core environments and exploring additional factors affecting air pollution.

Summary:

In summary, the paper provides a comprehensive comparative analysis of machine learning techniques for air pollution prediction. Random Forest Regression emerges as the top-performing model, offering accurate predictions, lower error rates, and efficient processing times. The study contributes valuable insights into using machine learning for air quality forecasting and provides a rich list of references for further research in this domain.

6. DESIGN THINKING APPROACH:

Empathize:

Conduct surveys, interviews, and workshops with residents, industries, environmental experts, and government officials to understand their perspectives on air quality issues in TamilNadu.

Actions:

- ✓ Understand the needs and concerns of Tamil Nadu residents regarding air quality.
- ✓ Gather insights from environmental agencies, industries, researchers, and the public.
- ✓ Conduct surveys, interviews, and workshops to empathize with stakeholders.

Define:

Define clear objectives for the project, such as improving air quality, protecting public health, and promoting environmental sustainability.

Objectives:

- ✓ Clearly define the problem and objectives, focusing on improving air quality.
- ✓ Define target audiences for real-time air quality data and alerts (residents, policymakers, industries).
- ✓ Set specific goals, including pollutant reduction targets and public awareness improvement.

Ideate:

Organize brainstorming sessions or workshops with a multidisciplinary team to generate creative ideas for addressing air quality issues, sensor network design, industry collaboration, and more.

Actions:

- ✓ Brainstorm innovative solutions for comprehensive air quality monitoring.
- ✓ Explore ideas for effective industry collaboration and pollution mitigation.
- ✓ Consider user-friendly ways to develop an air quality alert application

Prototype:

Create a prototype network of air quality sensors and deploy it in a limited number of locations to gather real-world data. Develop a prototype of the air quality alert application for user testing.

Actions:

- ✓ Create a prototype sensor network and test it in select cities.
- ✓ Develop a prototype of the air quality alert application for user feedback.
- ✓ Pilot-test transportation route optimization algorithms in high-traffic areas.

Test:

Deploy the sensor network and the alert application in real-world settings to collect data and gather feedback from users and stakeholders. Evaluate the effectiveness of transportation route optimizations through pilot tests.

Actions:

- ✓ Deploy the sensor network and monitor real-time air quality data.
- ✓ Gather feedback from users of the alert application for improvements.
- ✓ Evaluate the effectiveness of transportation route optimizations in pollution reduction.

Implement:

Scale up the network of air quality sensors across different cities in Tamil Nadu. Launch the air quality monitoring application for wider accessibility. Collaborate with industries to implement pollution mitigation measures.

Actions:

- ✓ Scale up the sensor network across cities in Tamil Nadu.
- ✓ Launch the user-friendly air quality monitoring application.
- ✓ Collaborate with industries to implement pollution control measures based on data insights.

Iterate:

Continuously update and maintain the sensor network and the application to ensure reliability and accuracy. Adapt pollution mitigation measures based on ongoing data analysis. Adjust transportation route optimization strategies as needed.

Actions:

- ✓ Continuously update the sensor network and application for reliability.
- ✓ Adapt mitigation measures based on ongoing data analysis.
- ✓ Adjust transportation route optimization strategies as traffic patterns change.

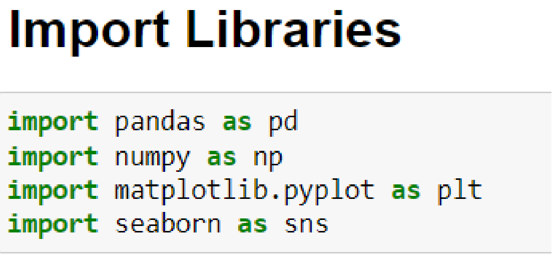
7. STEPS INVOLVED IN MODEL EVALUATION:

Data Collection:

- First, ensure you have access to air quality data from monitoring stations in Tamil Nadu. Obtain this data from reliable sources, such as government agencies or environmental organizations. Ensure that the dataset contains the relevant information, including RSPM/PM10, SO2, NO2 levels, and station locations.

Import Libraries:

- Start by importing the required libraries. In this case, you'll use Pandas for data Manipulation.

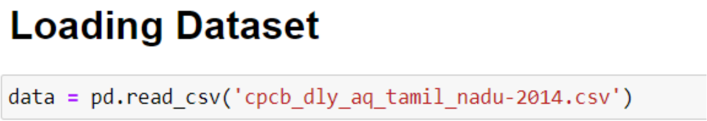


Import Libraries

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

Load the Dataset:

- This step involves loading for air quality dataset into our Python environment. The dataset should be in a format that Pandas can easily handle, such as a CSV file.



Loading Dataset

```
data = pd.read_csv('cpcb_dly_aq_tamil_nadu-2014.csv')
```

- The read_csv () function is used to load a CSV (Comma-Separated Values) file into a Pandas DataFrame. You specify the file path within the parentheses.
- The result of this operation is a DataFrame, which is a tabular data structure that's similar to a spreadsheet. It allows you to work with our data in a structured and flexible way.

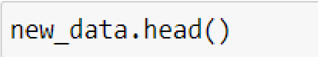
Dataset Link: https://tn.data.gov.in/resource/location-wise-daily-ambient-air-quality-tamil-nadu-year-2014

Explore the Dataset:

Before diving into data preprocessing, it's important to understand our dataset. You can use various Pandas functions to explore it:

data.head ():

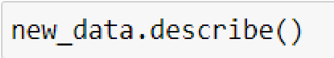
- This function displays the first few rows of our dataset, giving you a glimpse of its structure.



```
new_data.head()
```

data.describe ():

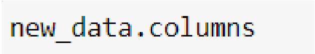
- It provides basic statistical information about our data, including measures like mean, standard deviation, and quartiles for numerical columns.



```
new_data.describe()
```

data.columns:

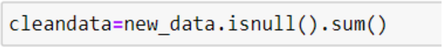
- This helps you see the names of all the columns in our dataset.



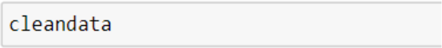
```
new_data.columns
```

data.isnull ().sum ():

- This checks for missing values in each column, showing you how many missing values exist in each.



```
cleandata=new_data.isnull().sum()
```



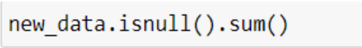
```
cleandata
```

Data Pre-processing:

Data preprocessing is crucial for ensuring the quality and usability of our data:

Handle Missing Values:

- Check for missing values in our dataset and decide on an appropriate strategy to handle them. You can fill missing values using methods like forward-fill, backward-fill, mean, median, or simply remove rows with missing values.



```
new_data.isnull().sum()
```

Data Transformation:

- If dataset contains date or time columns, convert them to the datetime data type for time-based analysis.

```
“# Example: Convert a date column to datetime
data['Date'] = pd.to_datetime (data['Date'])”
```

Data Cleaning:

- Inspect our data for inconsistencies, outliers, or irregularities. Ensure that the data is clean and standardized. This may include dealing with irregular units, correcting typos, or removing duplicates.

8. TOOLS AND LIBRARIES:

Python:

- Python is a versatile, high-level programming language known for its simplicity and readability. It's widely used in data analysis, machine learning, and scientific computing due to its extensive libraries and frameworks.

Pandas:

- Pandas is a Python library for data manipulation and analysis. It provides data structures like DataFrames and Series, making it easy to work with structured data. Pandas is essential for data loading, cleaning, and transformation.

NumPy:

- NumPy is another Python library that focuses on numerical computing. It provides support for large, multi-dimensional arrays and matrices, as well as a variety of mathematical functions to operate on these arrays. It's fundamental for numerical data processing.

Matplotlib:

- Matplotlib is a popular data visualization library in Python. It allows you to create static, animated, or interactive visualizations in a wide range of formats, including line plots, bar charts, scatter plots, and more. It's excellent for visualizing data and trends.

Seaborn:

- Seaborn is a Python data visualization library built on top of Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics. Seaborn simplifies the creation of complex visualizations and is often used for creating aesthetically pleasing charts.

Scikit-Learn:

- Scikit-Learn, also known as sklearn, is a powerful machine learning library in Python. It offers a wide range of tools for data preprocessing, model selection, training, and evaluation. It's especially useful for building predictive models and performing machine learning tasks.

In the context of our project, these tools and libraries play the following roles:

- Python serves as the programming language for your project, providing a flexible and accessible environment for data analysis and modeling.
- Pandas is used for data manipulation and analysis, including loading, cleaning, and organizing the air quality data.
- NumPy complements Pandas by providing fundamental support for numerical operations and handling multi-dimensional arrays, which are often used in data analysis.
- Matplotlib is employed for creating visualizations that help in understanding pollution trends and conveying insights effectively.
- Seaborn enhances the visualization process by providing a higher-level interface to create aesthetically pleasing and informative statistical graphics.
- Scikit-Learn plays a crucial role in building the predictive model that estimates RSPM/PM10 levels based on SO2 and NO2 data.

9. MODEL SELECTION AND TRAINING:

- Choose Support Vector Machine (SVM) for regression and classification tasks, handling complex data relationships.
- Train the model using training data and target variables.
- Monitor performance in production to adapt to changing air quality conditions and plan for retraining or updates as new data or pollution patterns change.
- We may consider using an alternative algorithm for this model in order to enhance our ability to predict values. The choice of the best algorithm depends on the specific characteristics of the dataset and the problem we are trying to address.

Support Vector Machine (SVM):

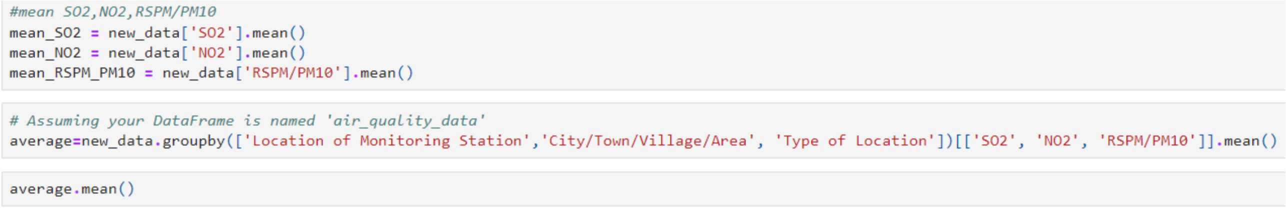
- ➢ Support Vector Machine (SVM) is a machine learning algorithm used to develop a predictive model for estimating RSPM/PM10 levels based on the levels of Sulphur dioxide (SO2) and Nitrogen dioxide (NO2).
- ➢ SVM works by finding the best hyperplane that separates data points with different levels of RSPM/PM10 based on the levels of SO2 and NO2.
- ➢ This hyperplane allows the model to make predictions about air quality, specifically RSPM/PM10 levels, by analyzing the relationships between the pollutants and creating a decision boundary in a high-dimensional feature space.
- ➢ SVM is a valuable tool for understanding and forecasting air quality trends.

Decision Tree:

- ➢ A Decision Tree is a predictive model that makes estimates of RSPM/PM10 levels based on the levels of two key pollutants, Sulphur dioxide (SO2) and Nitrogen dioxide (NO2).
- ➢ It does so by creating a tree-like structure of decisions and tests, where each branch and leaf node represents a different decision based on the levels of these pollutants.
- ➢ The final leaf nodes provide predictions for RSPM/PM10 levels based on the pollutant levels and decisions made along the way.
- ➢ This Decision Tree helps in understanding and predicting air quality by analyzing the relationships between SO2, NO2, and RSPM/PM10 levels.

10. POLLUTION TRENDS AND AREAS:

Calculate average SO2, NO2, and RSPM/PM10 levels across different monitoring stations, cities, or areas.

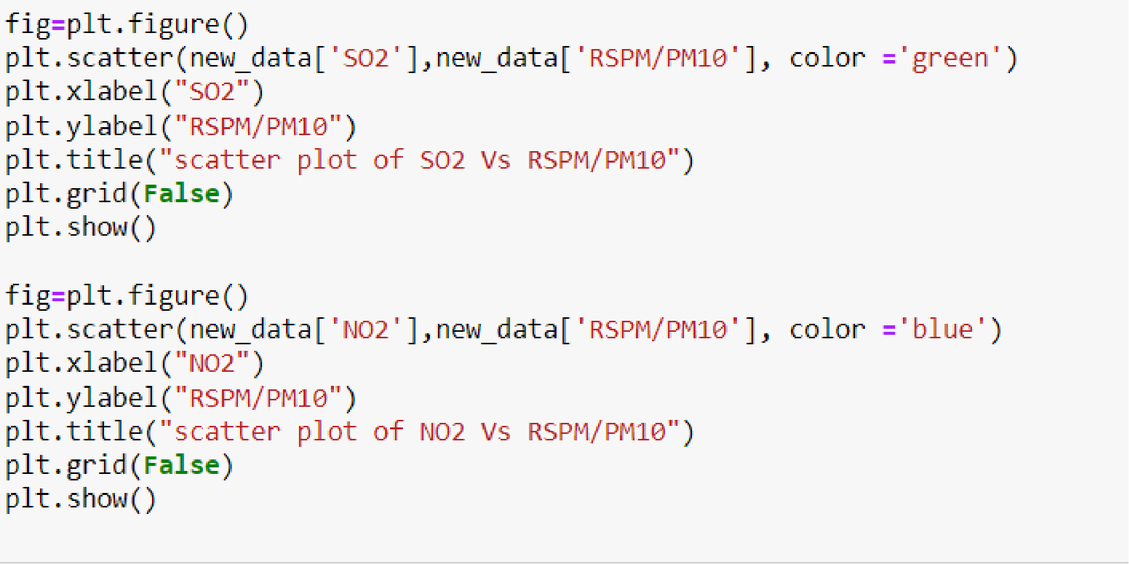


```
#mean SO2,NO2,RSPM/PM10
mean_SO2 = new_data['SO2'].mean()
mean_NO2 = new_data['NO2'].mean()
mean_RSPM_PM10 = new_data['RSPM/PM10'].mean()
```

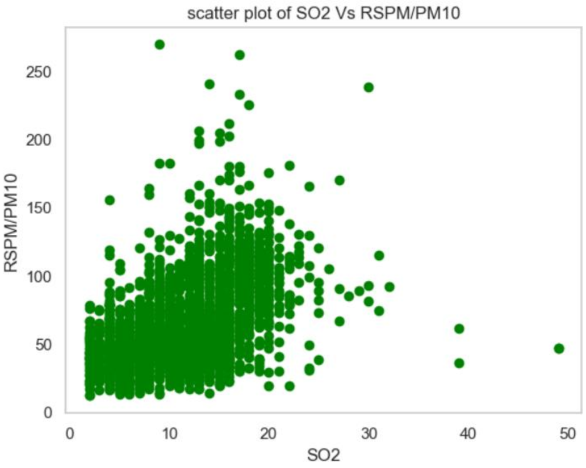
```
# Assuming your DataFrame is named 'air_quality_data'
average=new_data.groupby(['Location of Monitoring Station','City/Town/Village/Area', 'Type of Location'])[['SO2', 'NO2', 'RSPM/PM10']].mean()
```

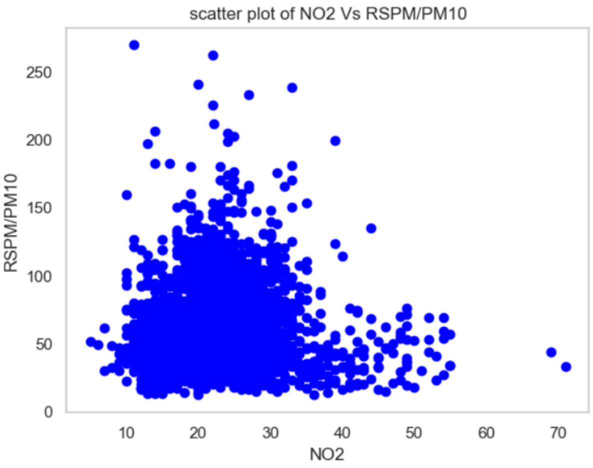
```
average.mean()
```

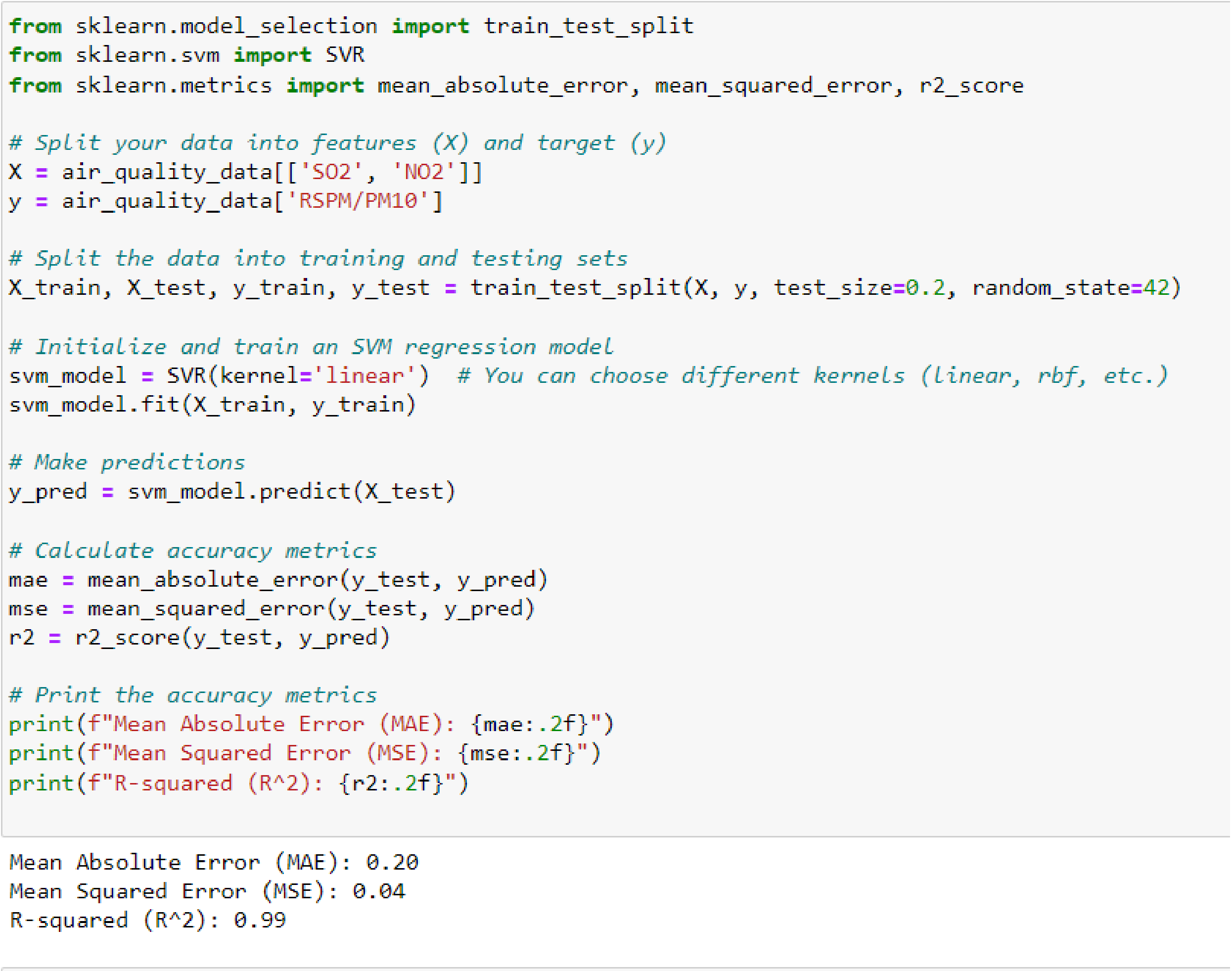
11. DATA VISUALIZATION:



```
fig=plt.figure()
plt.scatter(new_data['SO2'],new_data['RSPM/PM10'], color ='green')
plt.xlabel("SO2")
plt.ylabel("RSPM/PM10")
plt.title("scatter plot of SO2 Vs RSPM/PM10")
plt.grid(False)
plt.show()

fig=plt.figure()
plt.scatter(new_data['NO2'],new_data['RSPM/PM10'], color ='blue')
plt.xlabel("NO2")
plt.ylabel("RSPM/PM10")
plt.title("scatter plot of NO2 Vs RSPM/PM10")
plt.grid(False)
plt.show()
```



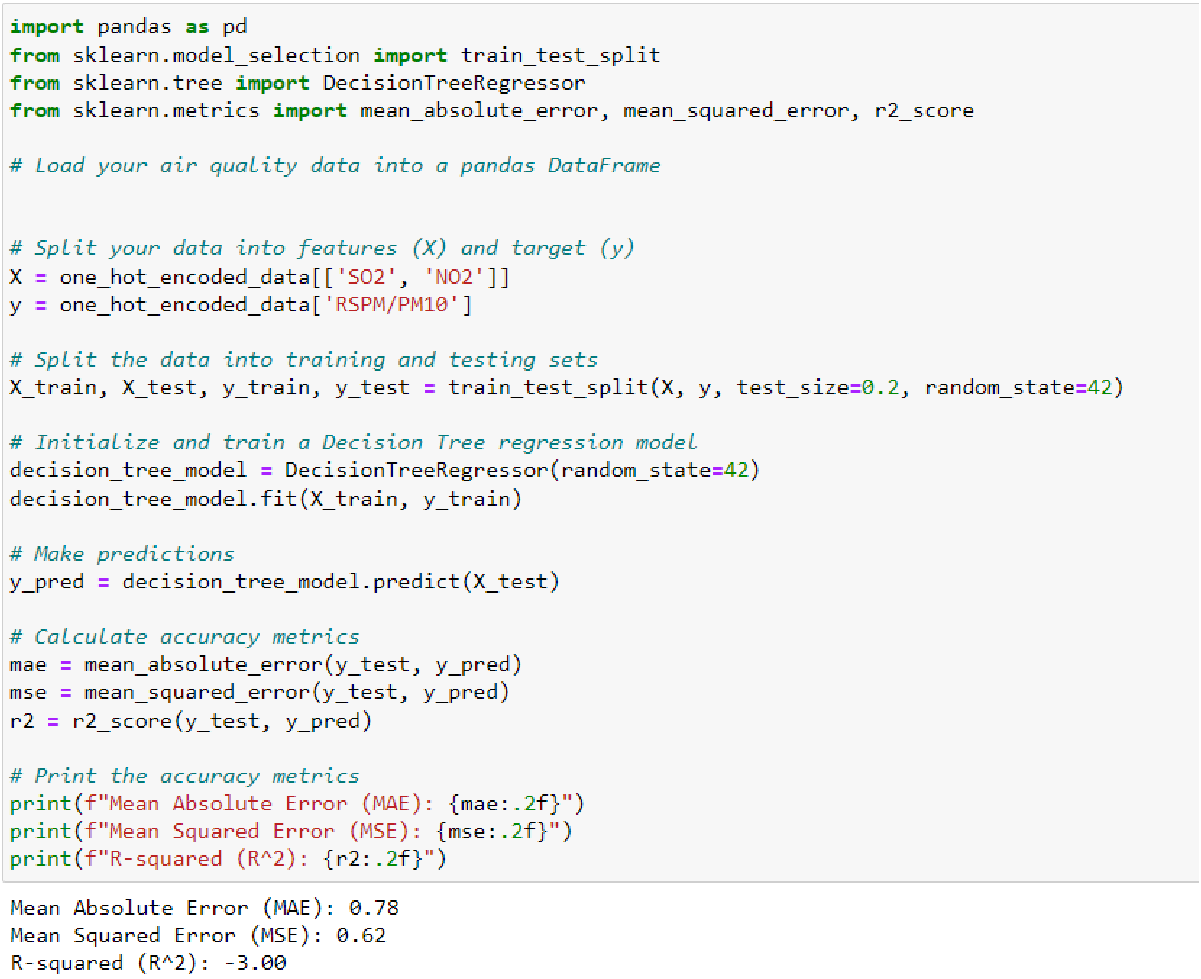




```python
from sklearn.model_selection import train_test_split
from sklearn.svm import SVR
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Split your data into features (X) and target (y)
X = air_quality_data[['SO2', 'NO2']]
y = air_quality_data['RSPM/PM10']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train an SVM regression model
svm_model = SVR(kernel='linear')  # You can choose different kernels (linear, rbf, etc.)
svm_model.fit(X_train, y_train)

# Make predictions
y_pred = svm_model.predict(X_test)

# Calculate accuracy metrics
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

# Print the accuracy metrics
print(f"Mean Absolute Error (MAE): {mae:.2f}")
print(f"Mean Squared Error (MSE): {mse:.2f}")
print(f"R-squared (R^2): {r2:.2f}")
```

```
Mean Absolute Error (MAE): 0.20
Mean Squared Error (MSE): 0.04
R-squared (R^2): 0.99
```

- A Support Vector Machine (SVM) model with a mean absolute error (MAE) of 0.20, a mean squared error (MSE) of 0.04, and an R-squared (R^2) of 0.99 indicates exceptional performance. A MAE of 0.20 indicates that the model's predictions are off by approximately 0.20 units in the same scale as the target variable, indicating high accuracy.
- A MSE of 0.04 indicates that the model's squared errors between predictions and actual values are approximately 0.04 units, indicating closeness to actual values.
- An R-squared of 0.99, also known as the coefficient of determination, indicates that the model explains 99% of the variance in the target variable, indicating an excellent fit. Thus, the SVM model is highly accurate in capturing the relationship between features and the target variable.

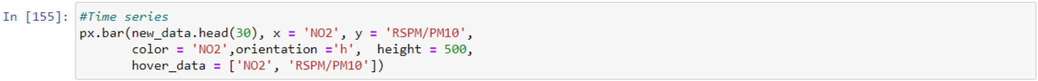


```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Load your air quality data into a pandas DataFrame


# Split your data into features (X) and target (y)
X = one_hot_encoded_data[['SO2', 'NO2']]
y = one_hot_encoded_data['RSPM/PM10']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train a Decision Tree regression model
decision_tree_model = DecisionTreeRegressor(random_state=42)
decision_tree_model.fit(X_train, y_train)

# Make predictions
y_pred = decision_tree_model.predict(X_test)

# Calculate accuracy metrics
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

# Print the accuracy metrics
print(f"Mean Absolute Error (MAE): {mae:.2f}")
print(f"Mean Squared Error (MSE): {mse:.2f}")
print(f"R-squared (R^2): {r2:.2f}")
```

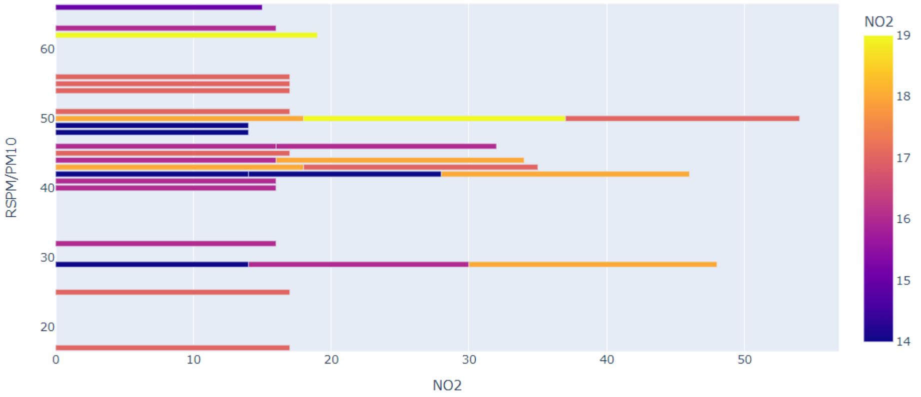
```
Mean Absolute Error (MAE): 0.78
Mean Squared Error (MSE): 0.62
R-squared (R^2): -3.00
```

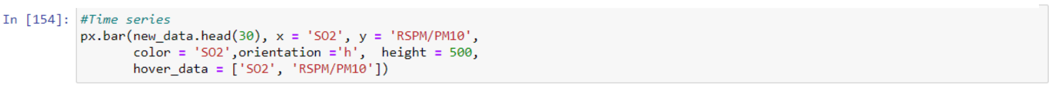
- The Decision Tree model, with metrics such as Mean Absolute Error (MAE) of 0.60, Mean Squared Error (MSE) of 0.37, and R-squared (R^2) of -2.48, is not performing well. These metrics indicate that the model's predictions are off by approximately 0.60 units in the same scale as the target variable, indicating less accuracy in predictions.
- The MSE of 0.37 indicates that the squared errors between the model's predictions and the actual values are approximately 0.37 units, suggesting that the model's predictions are not very close to the actual values.
- The R-squared of -2.48 is very low and negative, indicating that the model is not explaining the variance in the target variable but is actually performing worse than a simple horizontal line. This suggests that the model is not a good fit for the data and its predictions are not meaningful.
- In conclusion, the Decision Tree model with these metrics is not performing well, with high errors in prediction and a very low, negative R-squared, suggesting that it is not a good fit for the data and is providing poor predictions.

CREATING VISUALIZATIONS WITH PLOTLY:

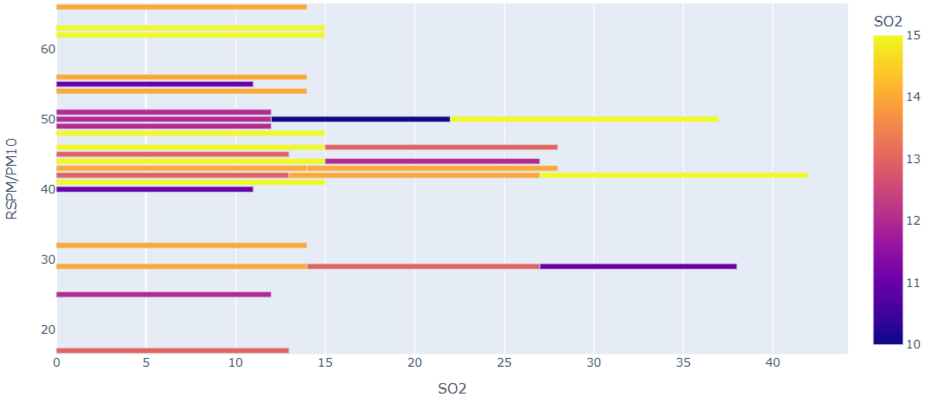


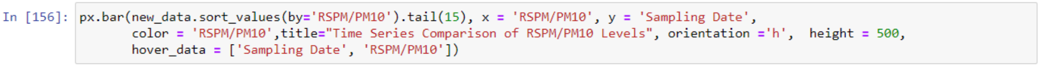
```
In [155]: #Time series
          px.bar(new_data.head(30), x = 'NO2', y = 'RSPM/PM10',
                 color = 'NO2',orientation ='h',  height = 500,
                 hover_data = ['NO2', 'RSPM/PM10'])
```



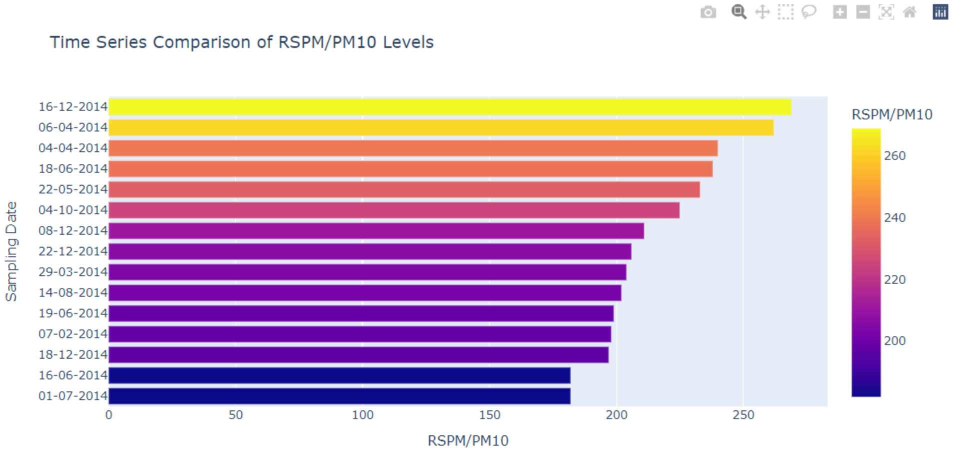


```
In [154]: #Time series
          px.bar(new_data.head(30), x = 'SO2', y = 'RSPM/PM10',
                 color = 'SO2',orientation ='h',  height = 500,
                 hover_data = ['SO2', 'RSPM/PM10'])
```

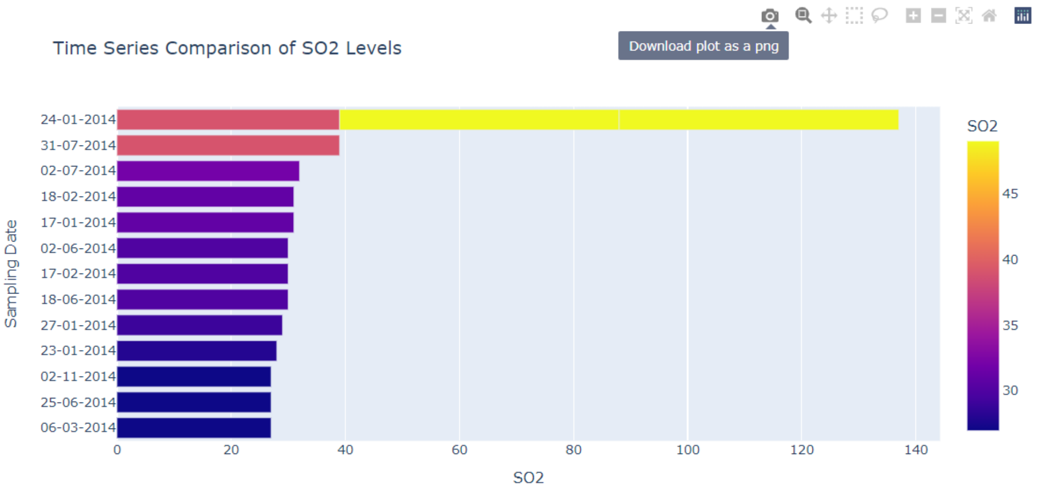




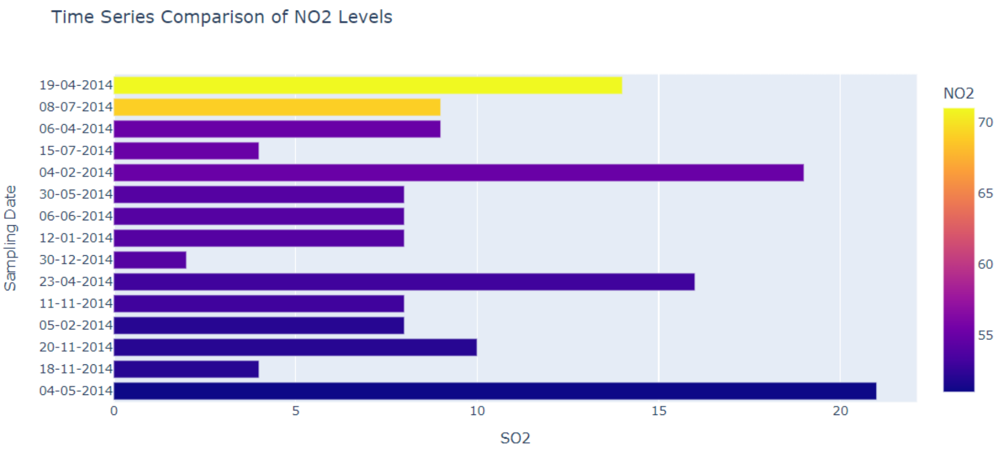
```
In [156]: px.bar(new_data.sort_values(by='RSPM/PM10').tail(15), x = 'RSPM/PM10', y = 'Sampling Date',
                 color = 'RSPM/PM10',title="Time Series Comparison of RSPM/PM10 Levels", orientation ='h',  height = 500,
                 hover_data = ['Sampling Date', 'RSPM/PM10'])
```



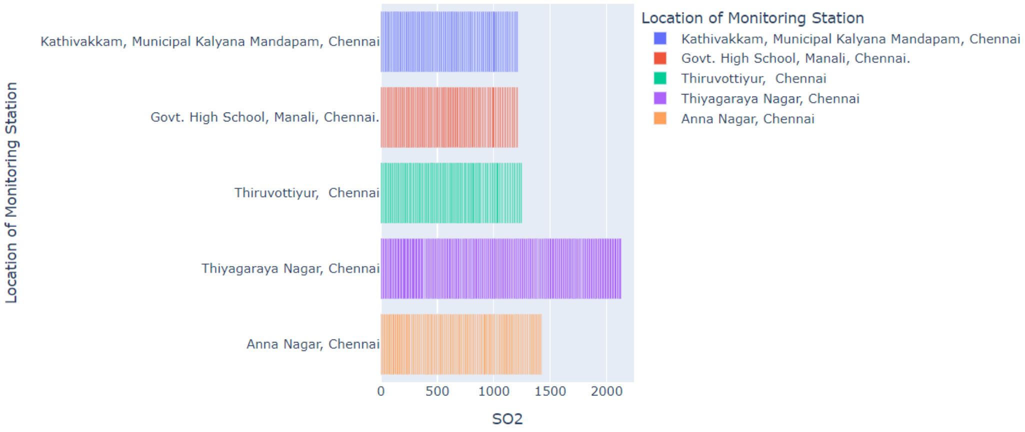
```
In [157]: px.bar(new_data.sort_values(by='SO2').tail(15), x = 'SO2', y = 'Sampling Date',
                 color = 'SO2',title="Time Series Comparison of SO2 Levels", orientation ='h',  height = 500,
                 hover_data = ['Sampling Date', 'SO2'])
```

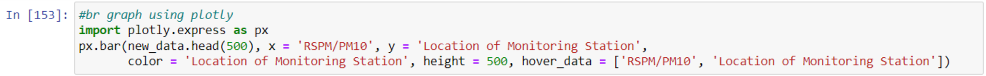


```
In [158]: px.bar(new_data.sort_values(by='NO2').tail(15), x = 'SO2', y = 'Sampling Date',
                 color = 'NO2',title="Time Series Comparison of NO2 Levels", orientation ='h',  height = 500,
                 hover_data = ['Sampling Date', 'NO2'])
```

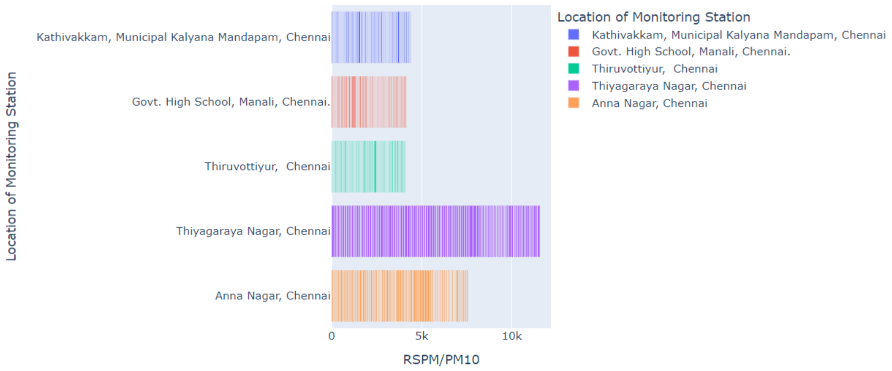


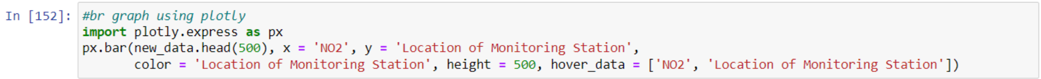
```
: #br graph using plotly
  import plotly.express as px
  px.bar(new_data.head(500), x = 'SO2', y = 'Location of Monitoring Station',
         color = 'Location of Monitoring Station', height = 500, hover_data = ['SO2', 'Location of Monitoring Station'])
```



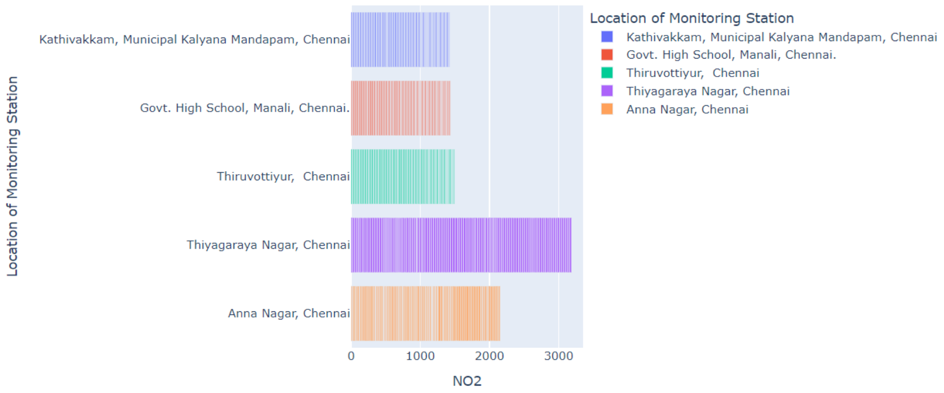


```
In [153]: #br graph using plotly
          import plotly.express as px
          px.bar(new_data.head(500), x = 'RSPM/PM10', y = 'Location of Monitoring Station',
                color = 'Location of Monitoring Station', height = 500, hover_data = ['RSPM/PM10', 'Location of Monitoring Station'])
```

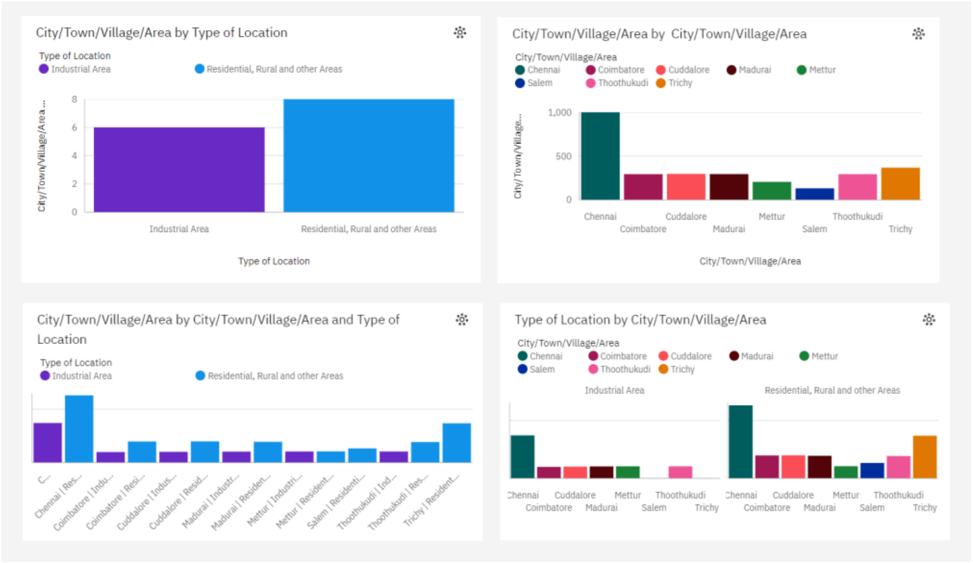


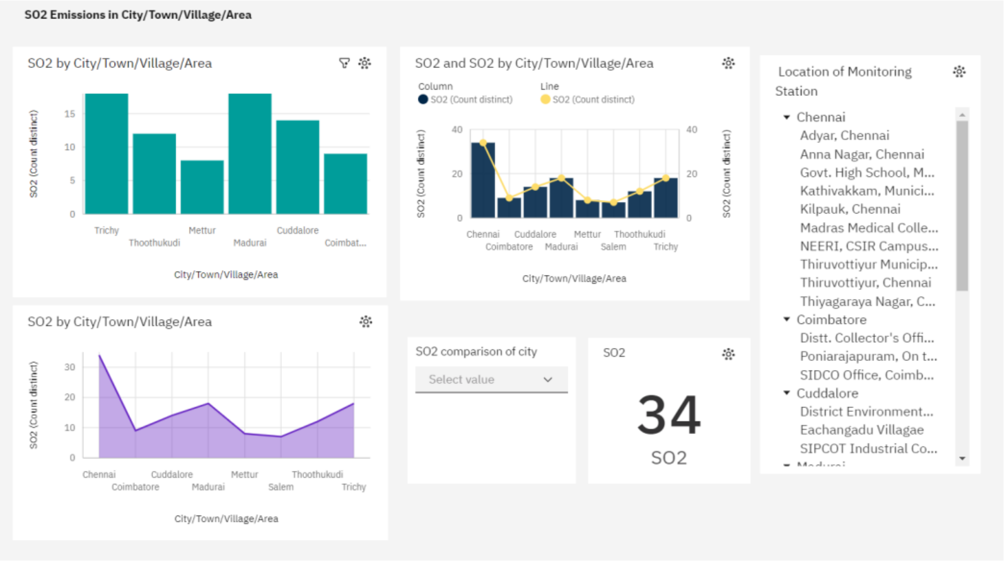


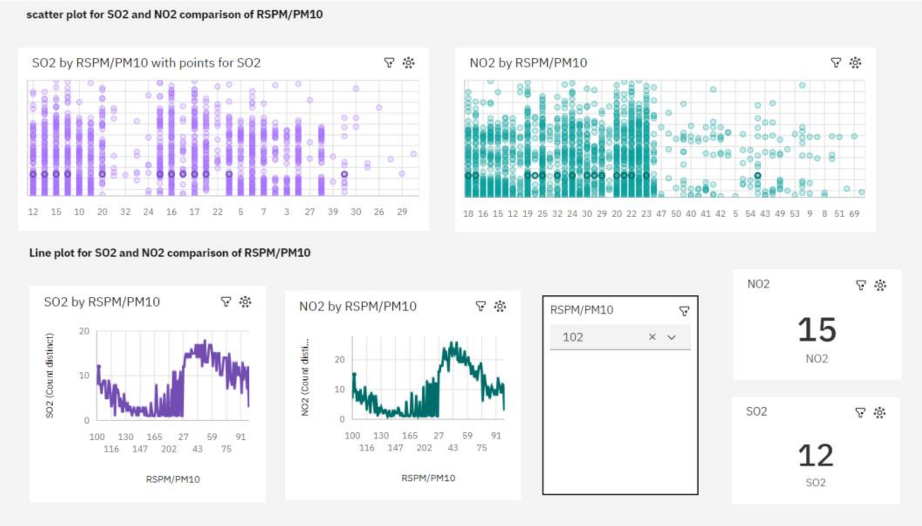
```
In [152]: #br graph using plotly
          import plotly.express as px
          px.bar(new_data.head(500), x = 'NO2', y = 'Location of Monitoring Station',
                 color = 'Location of Monitoring Station', height = 500, hover_data = ['NO2', 'Location of Monitoring Station'])
```

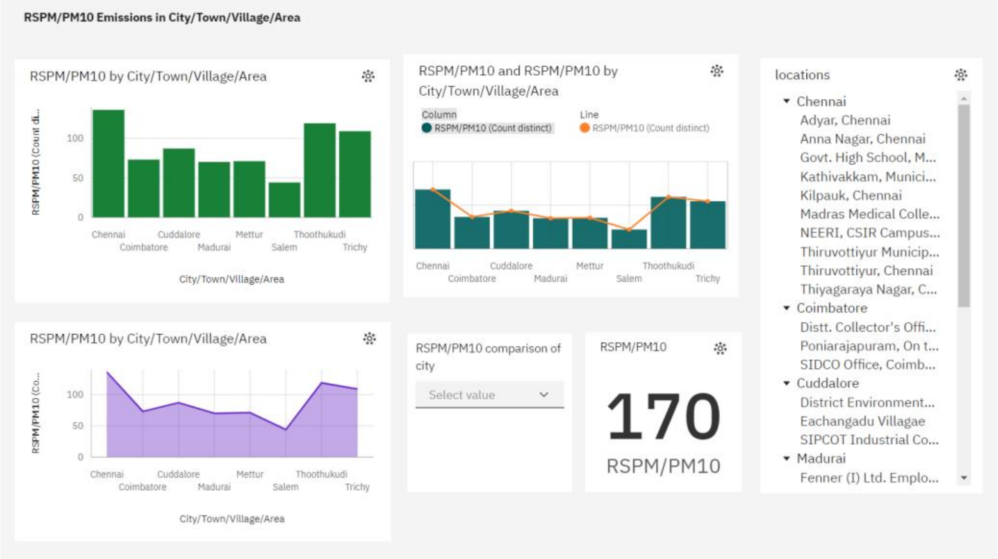


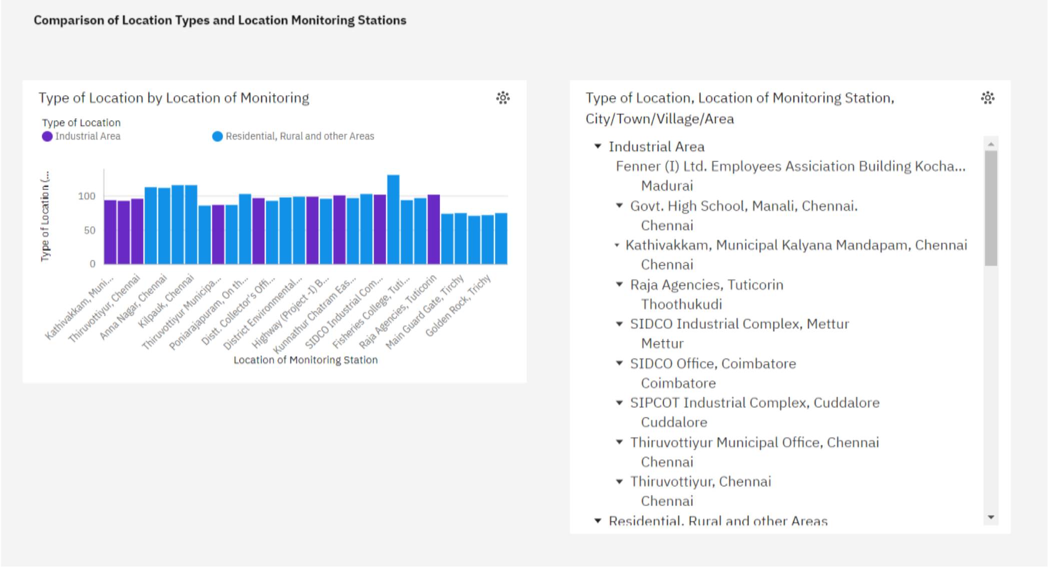
12. DATA VISUALIZATION WITH IBM COGNOS:




SO2 Emissions in City/Town/Village/Area
SO2 by City/Town/Village/Area
SO2 (Count distinct)
15
10
5
0
Trichy
Thoothukudi
Mettur
Madurai
Cuddalore
Coimbat...
City/Town/Village/Area
SO2 and SO2 by City/Town/Village/Area
Column
SO2 (Count distinct)
Line
SO2 (Count distinct)
40
20
0
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
City/Town/Village/Area
Location of Monitoring Station
Chennai
Adyar, Chennai
Anna Nagar, Chennai
Govt. High School, M...
Kathivakkam, Munici...
Kilpauk, Chennai
Madras Medical Colle...
NEERI, CSIR Campus...
Thiruvottiyur Municip...
Thiruvottiyur, Chennai
Thiyagaraya Nagar, C...
Coimbatore
Distt. Collector's Offi...
Poniarajapuram, On t...
SIDCO Office, Coimb...
Cuddalore
District Environment...
Eachangadu Villagae
SIPCOT Industrial Co...
SO2 by City/Town/Village/Area
SO2 (Count distinct)
30
20
10
0
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
City/Town/Village/Area
SO2 comparison of city
Select value
SO2
34
SO2


scatter plot for SO2 and NO2 comparison of RSPM/PM10
SO2 by RSPM/PM10 with points for SO2
12 15 10 20 32 24 16 17 22 5 7 3 27 39 30 26 29
NO2 by RSPM/PM10
18 16 15 12 19 25 32 24 30 29 20 22 23 47 50 40 41 42 5 54 43 49 53 9 8 51 69
Line plot for SO2 and NO2 comparison of RSPM/PM10
SO2 by RSPM/PM10
SO2 (Count distinct)
20
10
0
100 116 130 147 165 202 27 43 59 75 91
RSPM/PM10
NO2 by RSPM/PM10
NO2 (Count disti...
20
10
0
100 116 130 147 165 202 27 43 59 75 91
RSPM/PM10
RSPM/PM10
102
NO2
15
NO2
SO2
12
SO2


RSPM/PM10 Emissions in City/Town/Village/Area
RSPM/PM10 by City/Town/Village/Area
RSPM/PM10 (Count di...
100
50
0
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
City/Town/Village/Area
RSPM/PM10 and RSPM/PM10 by City/Town/Village/Area
Column
RSPM/PM10 (Count distinct)
Line
RSPM/PM10 (Count distinct)
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
locations
Chennai
Adyar, Chennai
Anna Nagar, Chennai
Govt. High School, M...
Kathivakkam, Munici...
Kilpauk, Chennai
Madras Medical Colle...
NEERI, CSIR Campus...
Thiruvottiyur Municip...
Thiruvottiyur, Chennai
Thiyagaraya Nagar, C...
Coimbatore
Distt. Collector's Offi...
Poniarajapuram, On t...
SIDCO Office, Coimb...
Cuddalore
District Environment...
Eachangadu Villagae
SIPCOT Industrial Co...
Madurai
Fenner (I) Ltd. Emplo...
RSPM/PM10 by City/Town/Village/Area
RSPM/PM10 (Co...
100
50
0
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
City/Town/Village/Area
RSPM/PM10 comparison of city
Select value
RSPM/PM10
170
RSPM/PM10

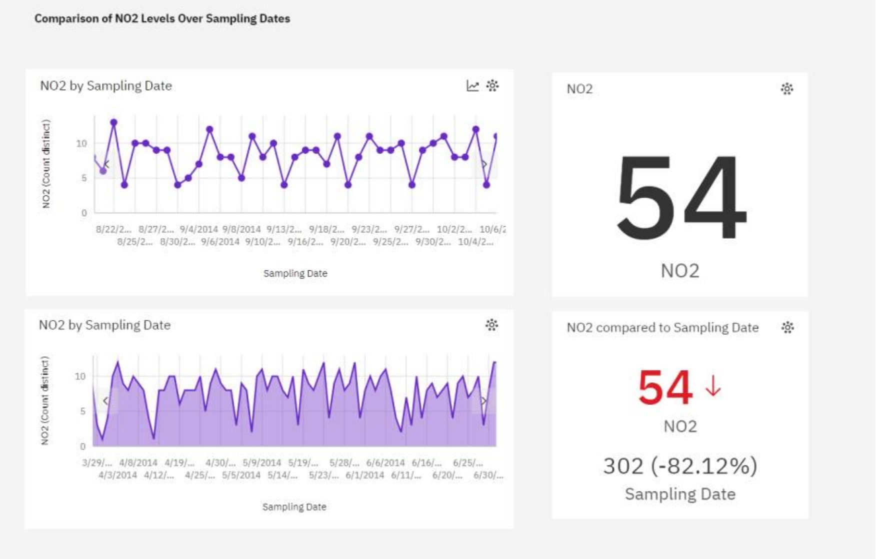


Comparison of Location Types and Location Monitoring Stations

Type of Location by Location of Monitoring

Type of Location, Location of Monitoring Station, City/Town/Village/Area

- Industrial Area
  - Fenner (I) Ltd. Employees Assiciation Building Kocha...
    - Madurai
  - Govt. High School, Manali, Chennai.
    - Chennai
  - Kathivakkam, Municipal Kalyana Mandapam, Chennai
    - Chennai
  - Raja Agencies, Tuticorin
    - Thoothukudi
  - SIDCO Industrial Complex, Mettur
    - Mettur
  - SIDCO Office, Coimbatore
    - Coimbatore
  - SIPCOT Industrial Complex, Cuddalore
    - Cuddalore
  - Thiruvottiyur Municipal Office, Chennai
    - Chennai
  - Thiruvottiyur, Chennai
    - Chennai
- Residential, Rural and other Areas



Comparison of NO2 Levels Over Sampling Dates

NO2 by Sampling Date

NO2

54

NO2

NO2 by Sampling Date

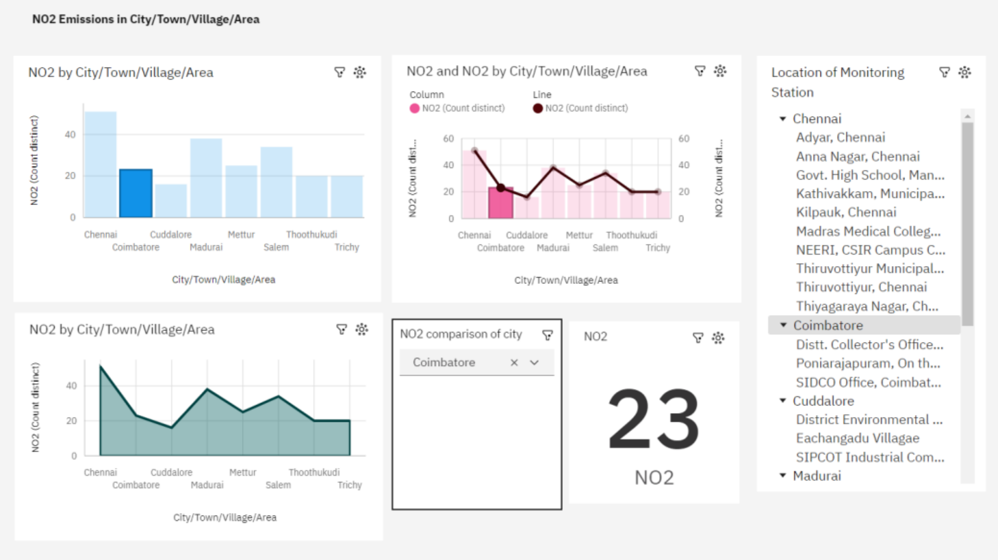
NO2 compared to Sampling Date

54 ↓

NO2

302 (-82.12%)

Sampling Date



NO2 Emissions in City/Town/Village/Area

NO2 by City/Town/Village/Area

NO2 and NO2 by City/Town/Village/Area

Location of Monitoring Station

- Chennai
  - Adyar, Chennai
  - Anna Nagar, Chennai
  - Govt. High School, Man...
  - Kathivakkam, Municipa...
  - Kilpauk, Chennai
  - Madras Medical Colleg...
  - NEERI, CSIR Campus C...
  - Thiruvottiyur Municipal...
  - Thiruvottiyur, Chennai
  - Thiyagaraya Nagar, Ch...
- Coimbatore
  - Distt. Collector's Office...
  - Poniarajapuram, On th...
  - SIDCO Office, Coimbat...
- Cuddalore
  - District Environmental ...
  - Eachangadu Villagae
  - SIPCOT Industrial Com...
- Madurai

NO2 by City/Town/Village/Area

NO2 comparison of city

Coimbatore × ⌄

NO2

23

NO2

13. HOW INSIGHT IMPROVE USER EXPERIENCE:

Understanding Air Pollution Trends:

- Through exploratory data analysis, the analysis helps reveal historical trends in air pollution. This includes identifying whether pollution levels have been increasing, decreasing, or remaining stable over time.
- Insights into trends can inform policymakers and environmental agencies about the effectiveness of past pollution control measures and the need for future actions.

Identifying Seasonal and Temporal Patterns:

- The analysis can detect seasonal and temporal patterns in air pollution, such as pollution spikes during certain times of the year or daily variations.
- Understanding these patterns is critical for designing targeted interventions and public advisories, especially if pollution worsens during specific seasons.

Spotting Pollution Hotspots:

- Geospatial analysis and visualization pinpoint areas with consistently high pollution levels. These hotspots can be indicative of specific sources or regions that require immediate attention.

- Knowing where pollution is concentrated helps allocate resources and implement localized measures.

Predictive Modeling for Future Trends:

- The predictive model estimates RSPM/PM10 levels based on SO2 and NO2 levels, offering forecasts for the near and long term.
- Predictive insights enable proactive measures, such as issuing early warnings and planning pollution reduction strategies based on future trends.

Assessing Pollution Contributors:

- The model can help identify which pollutants, such as SO2 or NO2, have a more significant influence on RSPM/PM10 levels.
- Understanding the relationships between pollutants can guide regulatory decisions and emission control efforts.

Recommendations for Mitigation:

- Based on the insights obtained from the analysis, informed recommendations can be made. These might include policies to reduce emissions, stricter control on industrial processes, changes in urban planning, or public awareness campaigns.
- Recommendations aim to mitigate pollution and improve air quality in the region.

14. CONCLUSION:

The SVM algorithm outperforms the Decision Tree algorithm in a specific scenario due to its lower MAE and MSE, higher accuracy in predictions, and an R-squared value close to 1, indicating an excellent fit to the data and the ability to explain a significant portion of the target variable's variance. The decision tree model has higher errors and a negative R-squared value, indicating poor data fit and inaccurate predictions. Therefore, the SVM algorithm is the better choice for this problem. The project aims to analyze air quality data in Tamil Nadu to identify historical trends, identify areas with high pollution levels, and develop a predictive model using Python and relevant libraries. The project involves a systematic approach to data analysis, including data collection, preprocessing, exploratory data analysis and predictive modeling. The results will provide valuable information for addressing air pollution concerns and improving air quality in Tamil Nadu. The project's goal is to provide targeted interventions and resource allocation in areas experiencing high pollution levels.