**AIR QUALITY ANALYSIS**

**TEAM MEMBER**

**311521205025: S JITHENDHARAN**

**Phase 5 Submission Document**

**Project:**Air Quality Analysis using data analytics

**AIR QUALITY ANALYSIS IN TAMILNADU**

**PROBLEM STATEMENT:**

Air pollution is a major problem in TamilNadu, India. It is estimated that air pollution causes over 10,000 deaths in the state each year. Air pollution can also cause a range of other health problems, such as respiratory infections, heart disease, and cancer.

There is a need for a comprehensive system for air quality analysis in Tamil Nadu. This system should be able to collect, process, model, visualize, and forecast air quality data. The information generated by the system can be used to protect the health and well-being of the state's residents.

**DESIGN THINKING:**

Design thinking is a human-centered approach to innovation that can be used to solve complex problems. It is a non-linear, iterative process that involves the following five steps:

1. Empathize: The first step is to understand the needs and experiences of the people who will be affected by the problem you are trying to solve. This can be done through interviews, focus groups, and surveys.

2. Define: Once you have a good understanding of the problem, you need to define it clearly and concisely. This will help you to focus your efforts and develop effective solutions.

3. Ideate: The third step is to generate a wide range of ideas for solving the problem. This can be done through brainstorming, sketching, and prototyping.

4. Prototype: Once you have a number of ideas, you need to develop prototypes of them. This will allow you to test your ideas and get feedback from others.

5. Test: The final step is to test your prototypes with the people who will be affected by the problem. This feedback will help you to refine your solutions and ensure that they are effective.

**APPLYING DESIGN THINKING FOR AIR QUALITY ANALYSIS:**

A diagram of a computer and a machine

Description automatically generated with medium confidenceAir quality analysis in Tamil Nadu is important for protecting the health and well-being of the state's residents. Air pollution can cause a range of health problems, including respiratory infections, heart disease, and cancer. It can also damage the environment.

**ABSTRACT:**

This document proposes a framework for air quality analysis in Tamil Nadu. The framework consists of the following modules:

1. Data collection and preprocessing: This module will collect data from various sources, such as government monitoring stations, satellite imagery, and citizen sensors. The data will then be preprocessed to ensure that it is consistent and of high quality.
2. Air quality modeling: This module will use the preprocessed data to develop air quality models. The models will be used to predict air quality levels at different locations and times.
3. Air quality visualization: This module will visualize the air quality data and model predictions in a user-friendly way. The visualizations will be used to communicate air quality information to the public and policymakers.
4. Air quality forecasting: This module will use the air quality models to forecast air quality levels for future days. The forecasts will be used to help people plan their activities and protect themselves from air pollution.

**INTRODUCTION:**

Air quality is a major public health and environmental concern. According to the World Health Organization, air pollution is responsible for about 7 million deaths each year. Air pollution can cause a variety of health problems, including asthma, heart disease, stroke, and cancer. It can also damage ecosystems and reduce agricultural yields.

Air quality analysis is important for understanding the sources and distribution of air pollution, assessing the health risks posed by air pollution, and developing strategies for reducing air pollution. Data analytics can play a vital role in air quality analysis.

**DATASET:**

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

**AIR QUALITY ANALYSIS PROJECT FOR A GIVEN DATASET**

This project will use data analytics to analyze a given air quality dataset. The dataset may contain data on a variety of air pollutants, such as particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. The dataset may also contain data on meteorological factors, such as temperature, wind speed, and wind direction.

The goal of the project is to gain insights into the air quality data using data analytics techniques. The project may focus on the following tasks:

**DATA CLEANING AND PREPROCESSING**:

The first step is to clean and preprocess the data. This may involve removing outliers, filling in missing values, and converting the data into a consistent format.

**EXPLORATORY DATA ANALYSIS**:

Exploratory data analysis is used to identify patterns and trends in the data. This can be done using a variety of visualization and statistical techniques.

**FEATURE ENGINEERING:**

Feature engineering is the process of creating new features from existing features. This can be done to improve the performance of machine learning models.

**MACHINE LEARNING:**

Machine learning can be used to develop models that can predict air quality levels or identify the sources of air pollution.

The following figure illustrates the methodology involved in the air quality analysis project:

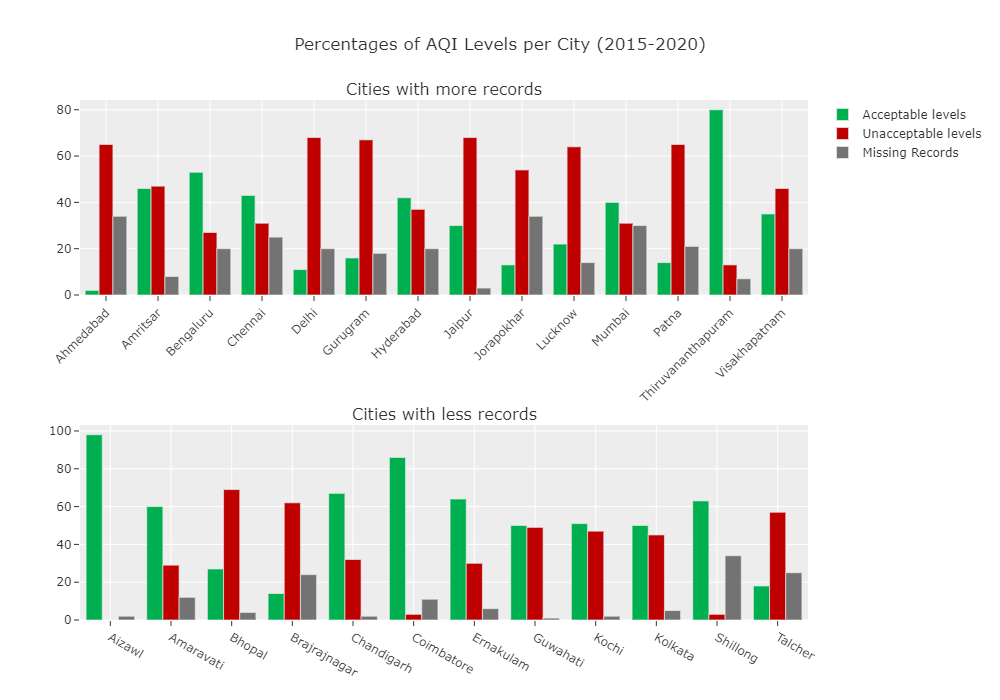
A diagram of a scientific process

Description automatically generated with medium confidence

**BENEFITS OF DATA ANALYTICS FOR AIR QUALITY ANALYSIS**

Data analytics can provide a number of benefits for air quality analysis, including:

**IDENTIFYING PATTERNS AND TRENDS:**

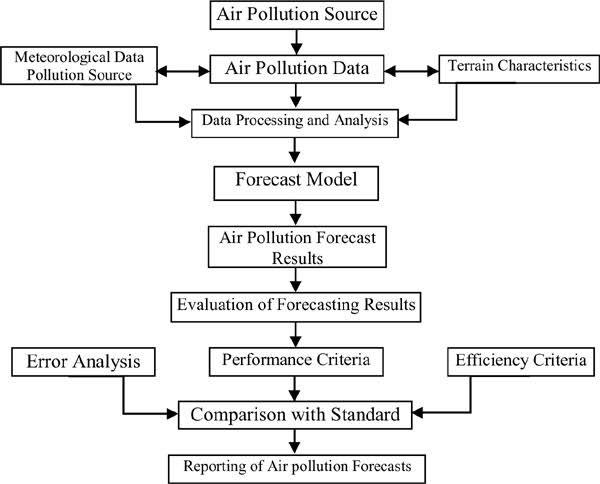
Data analytics can be used to identify patterns and trends in air quality data that would be difficult to see with the naked eye. This information can be used to develop a better understanding of the factors that contribute to air pollution and to predict future air quality conditions.

**QUANTIFYING THE HEALTH AND ENVIRONMENTAL IMPACTS OF AIR POLLUTION:**

Data analytics can be used to quantify the health and environmental impacts of air pollution. This information can be used to make informed decisions about air pollution control policies and regulations.

**DEVELOPING AND EVALUATING AIR POLLUTION CONTROL STRATEGIES:**

Data analytics can be used to develop and evaluate air pollution control strategies. For example, data analytics can be used to identify the most cost-effective ways to reduce air pollution from different sources.

**STEPS INVOLVED IN AIR QUALITY ANALYSIS:**

**AIR QUALITY ANALYSIS PROJECT FOR A GIVEN DATASET**

This project will use data analytics to analyze a given air quality dataset. The dataset may contain data on a variety of air pollutants, such as particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. The dataset may also contain data on meteorological factors, such as temperature, wind speed, and wind direction.

The goal of the project is to gain insights into the air quality data using data analytics techniques. The project may focus on the following tasks:

**DATA CLEANING AND PREPROCESSING**:

The first step is to clean and preprocess the data. This may involve removing outliers, filling in missing values, and converting the data into a consistent format.

**EXPLORATORY DATA ANALYSIS**:

Exploratory data analysis is used to identify patterns and trends in the data. This can be done using a variety of visualization and statistical techniques.

**FEATURE ENGINEERING:**

Feature engineering is the process of creating new features from existing features. This can be done to improve the performance of machine learning models.

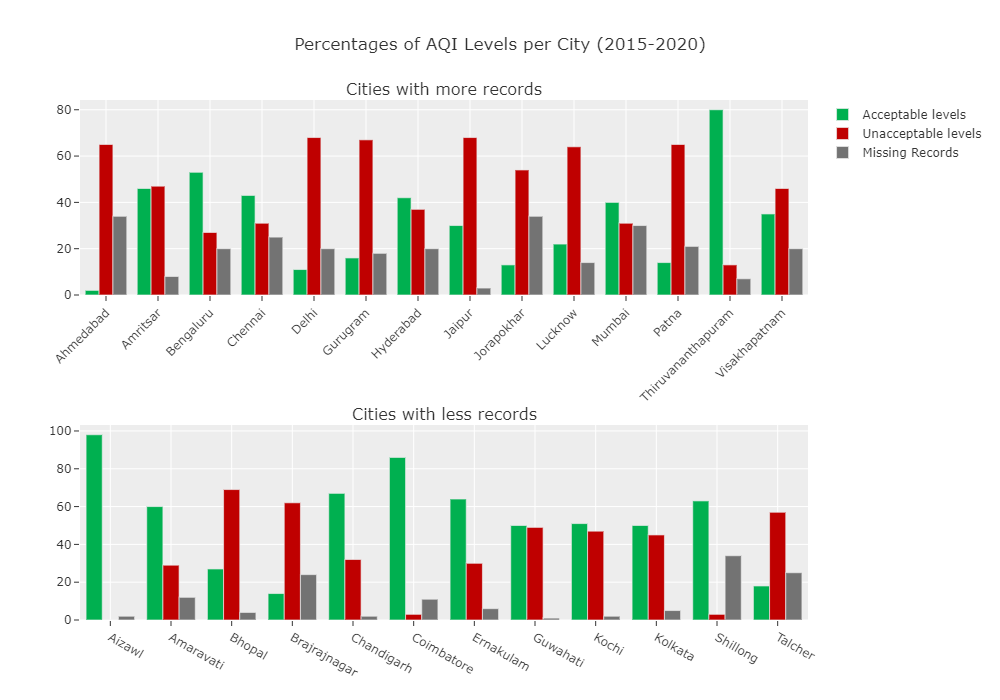
**MACHINE LEARNING:**

Machine learning can be used to develop models that can predict air quality levels or identify the sources of air pollution.

**BENEFITS OF DATA ANALYTICS FOR AIR QUALITY ANALYSIS**

Data analytics can provide a number of benefits for air quality analysis, including:

**IDENTIFYING PATTERNS AND TRENDS:**

Data analytics can be used to identify patterns and trends in air quality data that would be difficult to see with the naked eye. This information can be used to develop a better understanding of the factors that contribute to air pollution and to predict future air quality conditions.

**QUANTIFYING THE HEALTH AND ENVIRONMENTAL IMPACTS OF AIR POLLUTION:**

Data analytics can be used to quantify the health and environmental impacts of air pollution. This information can be used to make informed decisions about air pollution control policies and regulations.

**DEVELOPING AND EVALUATING AIR POLLUTION CONTROL STRATEGIES:**

Data analytics can be used to develop and evaluate air pollution control strategies. For example, data analytics can be used to identify the most cost-effective ways to reduce air pollution from different sources.

**ARCHITECTURE OF AIR QUALITY ANALYSIS IN TN:**

**A diagram of a process

Description automatically generated**

**PROGRAM:**

import pandas as pd

df = pd.read\_csv("air\_quality.csv")

print(df.head())

print(df.info())

print(df.describe())

df = df.dropna()

df = df.reset\_index(drop=True)

import matplotlib.pyplot as plt

df = df.dropna(subset=["RSPM/PM10"])

plt.hist(df["RSPM/PM10"], bins=20, edgecolor="k")

plt.xlabel("RSPM/PM10 Concentration")

plt.ylabel("Frequency")

plt.title("Histogram of RSPM/PM10 Concentration")

plt.show()

grouped\_data = df.groupby('Stn Code')['SO2'].mean()

correlation\_matrix = df.corr(numeric\_only=True)

correlation\_matrix = correlation\_matrix.fillna(0)

import seaborn as sns

import matplotlib.pyplot as plt

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', linewidths=0.5)

plt.show()

**OUTPUT:**

Stn Code Sampling Date State ... NO2 RSPM/PM10 PM 2.5

0 38 01-02-14 Tamil Nadu ... 17.0 55.0 NaN

1 38 01-07-14 Tamil Nadu ... 17.0 45.0 NaN

2 38 21-01-14 Tamil Nadu ... 18.0 50.0 NaN

3 38 23-01-14 Tamil Nadu ... 16.0 46.0 NaN

4 38 28-01-14 Tamil Nadu ... 14.0 42.0 NaN

[5 rows x 11 columns]

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2879 entries, 0 to 2878

Data columns (total 11 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Stn Code 2879 non-null int64

1 Sampling Date 2879 non-null object

2 State 2879 non-null object

3 City/Town/Village/Area 2879 non-null object

4 Location of Monitoring Station 2879 non-null object

5 Agency 2879 non-null object

6 Type of Location 2879 non-null object

7 SO2 2868 non-null float64

8 NO2 2866 non-null float64

9 RSPM/PM10 2875 non-null float64

10 PM 2.5 0 non-null float64

dtypes: float64(4), int64(1), object(6)

memory usage: 247.5+ KB

None

Stn Code SO2 NO2 RSPM/PM10 PM 2.5

count 2879.000000 2868.000000 2866.000000 2875.000000 0.0

mean 475.750261 11.503138 22.136776 62.494261 NaN

std 277.675577 5.051702 7.128694 31.368745 NaN

min 38.000000 2.000000 5.000000 12.000000 NaN

25% 238.000000 8.000000 17.000000 41.000000 NaN

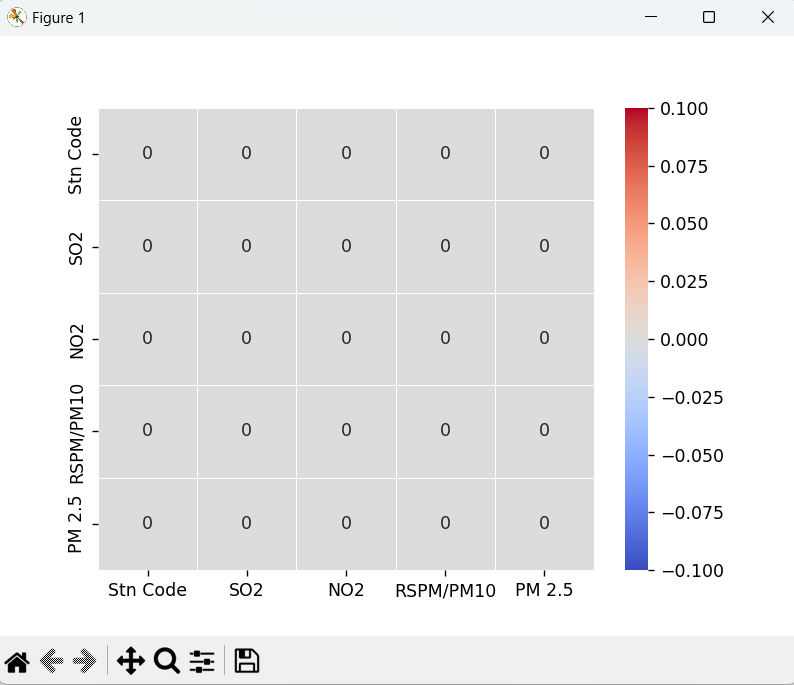
50% 366.000000 12.000000 22.000000 55.000000 NaN

75% 764.000000 15.000000 25.000000 78.000000 NaN

max 773.000000 49.000000 71.000000 269.000000 NaN

A screenshot of a computer screen

Description automatically generated



**PROGRAM:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv("air\_quality.csv")

average\_pollution = df.groupby('Location of Monitoring Station')[['NO2', 'RSPM/PM10']].mean()

plt.figure(figsize=(12, 6))

plt.bar(average\_pollution.index, average\_pollution['NO2'])

plt.xlabel('Area')

plt.ylabel('Average NO2 Level')

plt.title('Average NO2 Levels Across Areas')

plt.xticks(rotation=90)

plt.show()

plt.figure(figsize=(12, 6))

plt.bar(average\_pollution.index, average\_pollution['RSPM/PM10'])

plt.xlabel('Area')

plt.ylabel('Average RSPM/PM10 Level')

plt.title('Average RSPM/PM10 Levels Across Areas')

plt.xticks(rotation=90)

plt.show()

high\_no2\_threshold = 30

high\_no2\_areas = df[df['NO2'] > high\_no2\_threshold]

plt.figure(figsize=(12, 6))

sns.countplot(data=high\_no2\_areas, x='Location of Monitoring Station', order=high\_no2\_areas['Location of Monitoring Station'].value\_counts().index)

plt.xlabel('Area')

plt.ylabel('Count of High NO2 Readings')

plt.title(f'Areas with NO2 Levels Above {high\_no2\_threshold}')

plt.xticks(rotation=90)

plt.show()

plt.figure(figsize=(10, 2))

sns.scatterplot(data=df, x='NO2', y='RSPM/PM10', hue='Location of Monitoring Station')

plt.xlabel('NO2 Level')

plt.ylabel('RSPM/PM10 Level')

plt.title('Scatter Plot of NO2 vs. RSPM/PM10')

plt.show()

correlation\_matrix = df[['NO2', 'RSPM/PM10']].corr()

plt.figure(figsize=(6, 4))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Correlation Heatmap: NO2 vs. RSPM/PM10')

plt.show()

average\_pollution = df.groupby('Stn Code')[['SO2', 'NO2', 'RSPM/PM10']].mean()

import matplotlib.pyplot as plt

plt.figure(figsize=(12, 6))

plt.bar(average\_pollution.index, average\_pollution['SO2'])

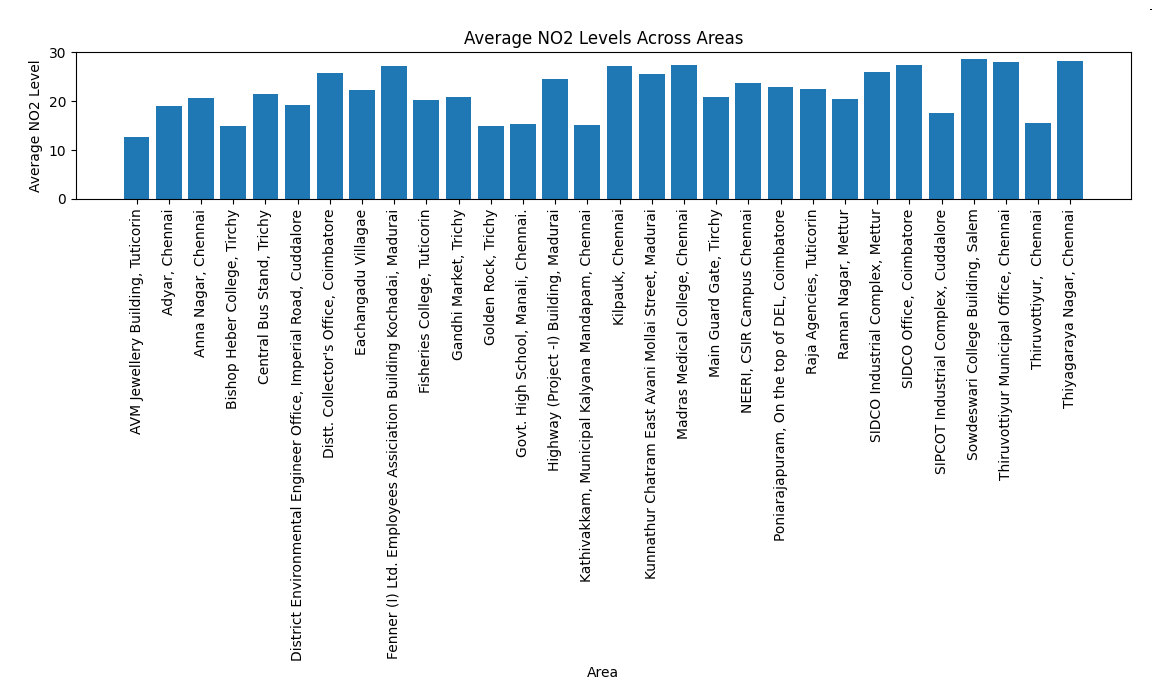
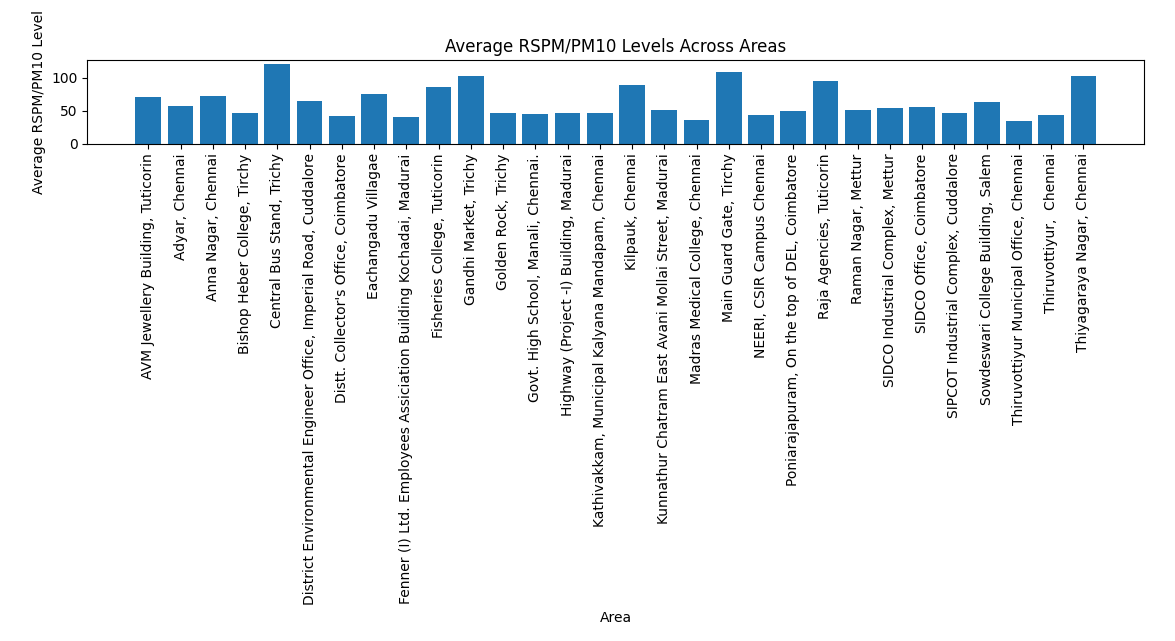
plt.xlabel('Monitoring Station')

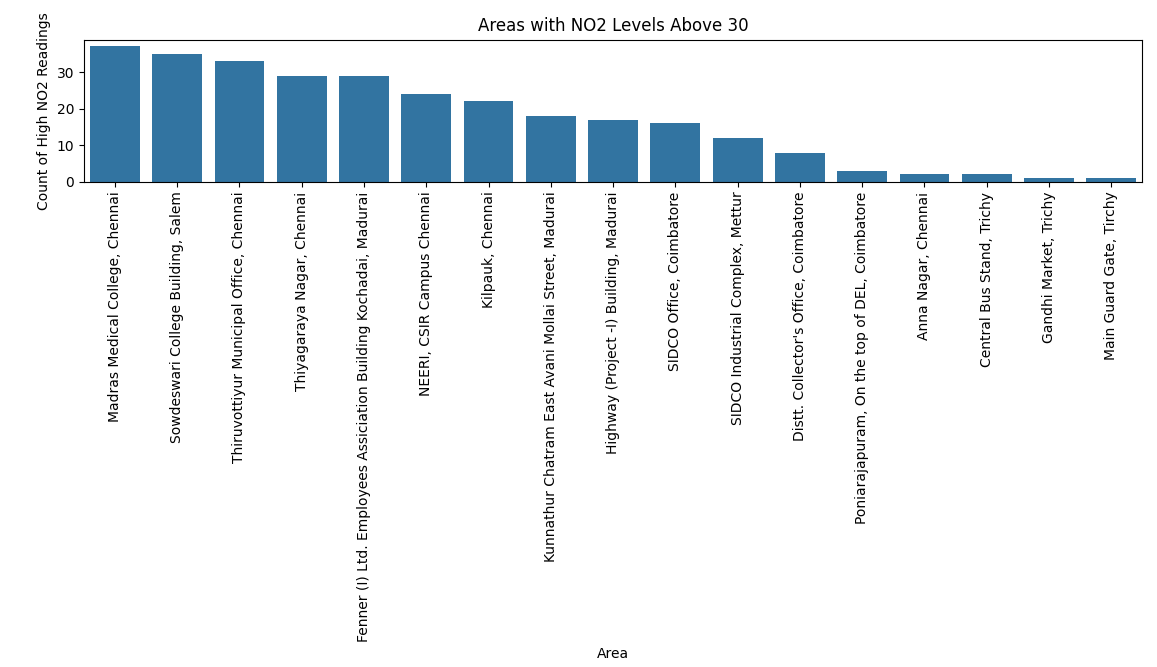
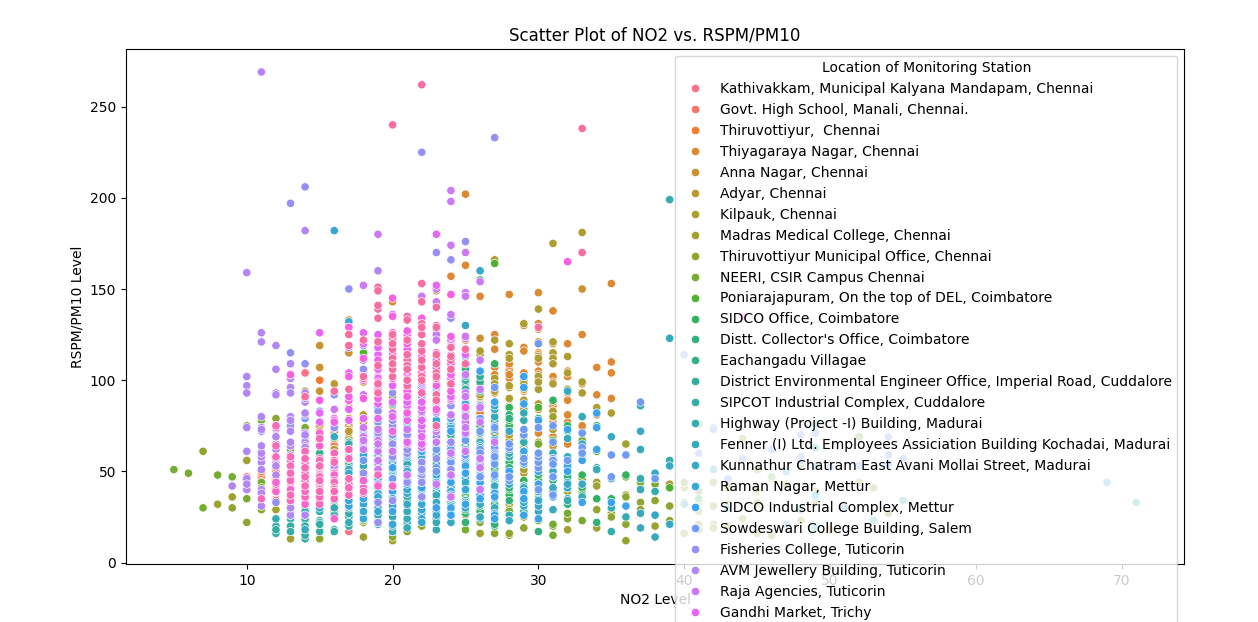
plt.ylabel('Average SO2 Level')

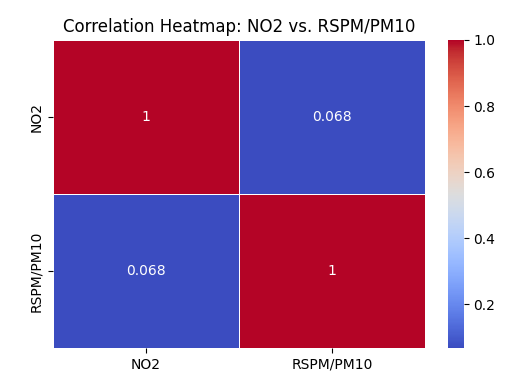
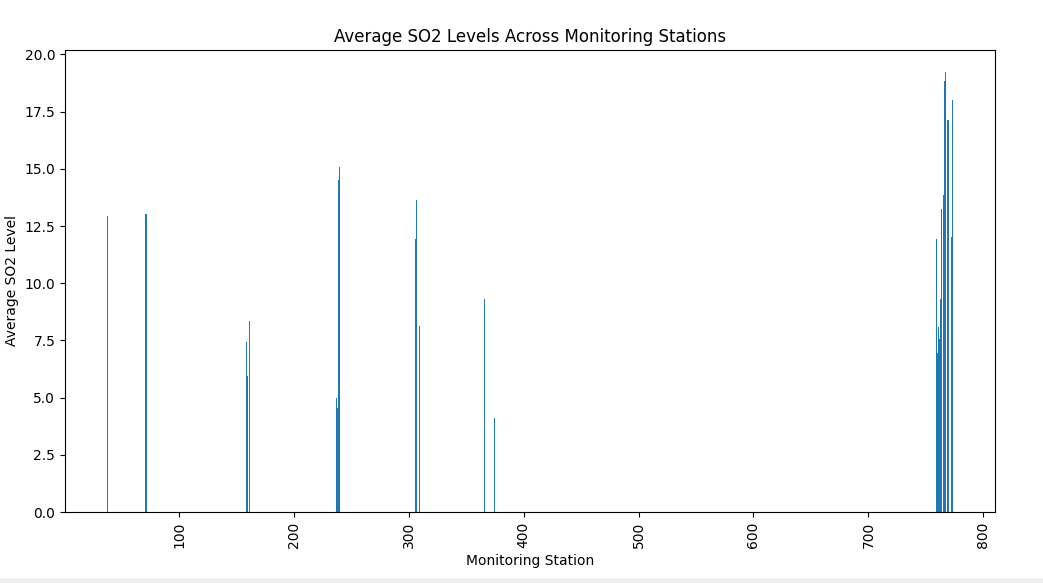
plt.title('Average SO2 Levels Across Monitoring Stations')

plt.xticks(rotation=90)

plt.show()

**OUTPUT:**





**CONCLUSION:**

Data analytics can play a vital role in air quality analysis. By using data analytics to analyze air quality data, we can gain insights into the sources and distribution of air pollution, assess the health risks posed by air pollution, and develop strategies for reducing air pollution.