AIR QUALITY ANALYSIS IN TAMIL NADU

INTRODUCTION:

- Air quality analysis is a crucial field dedicated to assessing the composition of the air we breathe.
- It involves monitoring various pollutants and particulate matter in the atmosphere,such as PM2.5,PM10,carbon monoxide,sulfur dioxide,nitrogen oxides,and voltaile organic compounds.
- This analysis aims to understand air quality's impact on human health and the environment.Data is collected from monitoring stations.
- With the growing concern over air pollution's detrimental effects,air quality analysis plays a pivotal role in fostering cleaner and healthier environments.

GIVEN DATA SET:

Stn Code	Sampling Date	State	City/Town/Villag	Location of Moni	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
38	01-02-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	11	17	55	NA
38	01-07-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	13	17	45	NA
38	21-01-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	12	18	50	NA
38	23-01-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	15	16	46	NA
38	28-01-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	13	14	42	NA
38	30-01-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	14	18	43	NA
38	02-04-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	12	17	51	NA
38	02-06-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	13	16	46	NA
38	02-11-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	10	19	50	NA
38	13-02-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	15	14	48	NA
38	18-02-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	14	16	32	NA
38	20-02-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	14	14	29	NA
38	25-02-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	13	17	17	NA
38	27-02-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	15	16	44	NA
38	03-04-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	12	17	25	NA
38	03-06-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	13	16	29	NA
38	03-11-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	11	18	29	NA
38	13-03-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	15	16	41	NA
38	18-03-14	Tamil Nadu	Chennai	Kathivakkam, Mu	Tamilnadu State	Industrial Area	14	17	43	NA

Some of the techniques that can be used in this process:

1. Data collection:

Obtain the air quality data from the given dataset.This data may include parameters like PM 2.5,PM 10,SO2,NO2.

2. Loading and Preprocessing Data:

- Use a programming language like Python for data analysis. Popular libraries for this task include NumPy, Pandas, and Matplotlib/Seaborn for visualization.
- Load your dataset into a Pandas DataFrame.
- Explore the dataset to understand its structure and quality.
- Handle missing data and outliers as needed.

3. Data Exploration:

- Visualize the data to gain insights into air quality trends, seasonal variations, and correlations between different pollutants.

4. Air Quality Metrics:

- Calculate air quality metrics like AQI (Air Quality Index) if they are not provided in the dataset.

5. Machine Learning:

- Using machine learning techniques to forecast air quality or detect anomalies.

PROGRAM:

Import Necessary Libraries:

```
python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
```

Load the Dataset:

```
import pandas as pd
# Replace dataset with the path to dataset file
df = pd.read_excel('dataset.excel')
```

Explore the dataset:

```
python
print(dataset.head()) # Display the first few rows of the dataset
print(dataset.info()) # Get information about the dataset
print(dataset.describe()) # Get summary statistics
```

Handle missing values:

```
python
dataset.isna().sum() # Check for missing values
```

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)
```

Model building:

model = RandomForestRegressor(n_estimators=100, random_state=42)

model.fit(X_train, y_train)

LOADING AND PREPROCESSING THE DATASET:

LOADING:

PROGRAM:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_excel('/content/cpcb_dly_aq_tamil_nadu-2014 (1).xlsx')
```

```
df.head()
```

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
0	38	41641	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN
1	38	41646	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0	NaN
2	38	21-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0	NaN
3	38	23-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0	NaN
4	38	28-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0	NaN

```
print(dataset.info())
```

```
<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
```

```
print(dataset.describe())
```

	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

```
dataset.isna().sum()
```

```
Stn Code                          0
Sampling Date                     0
State                             0
City/Town/Village/Area            0
Location of Monitoring Station    0
Agency                            0
Type of Location                  0
SO2                              11
NO2                              13
RSPM/PM10                         4
PM 2.5                         2879
dtype: int64
```

DATA PREPROCESSING:

Methods:

ONE -HOT SCALING:

Loading&Perform the dataset:

```
[15] import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     %matplotlib inline
     import seaborn as sns
```

```
[18] dataset=pd.read_excel('/content/cpcb_dly_aq_tamil_nadu-2014 (1) (2).xlsx')
```

```
dataset.tail(5)
```

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
2874	773	2014-12-03 00:00:00	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	18.0	102.0	NaN
2875	773	2014-12-10 00:00:00	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	12.0	14.0	91.0	NaN
2876	773	17-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	19.0	22.0	100.0	NaN
2877	773	24-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	17.0	95.0	NaN
2878	773	31-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	14.0	16.0	94.0	NaN

```
dataset_encoded=pd.get_dummies(dataset,columns=['Location of Monitoring Station'])
```

OUTPUT:

```
[26] dataset_encoded=pd.get_dummies(dataset,columns=['Location of Monitoring Station'])
```

```
print(dataset_encoded.head())
```

```
   Stn Code       Sampling Date       State City/Town/Village/Area  \
0        38  2014-01-02 00:00:00  Tamil Nadu               Chennai
1        38  2014-01-07 00:00:00  Tamil Nadu               Chennai
2        38             21-01-14  Tamil Nadu               Chennai
3        38             23-01-14  Tamil Nadu               Chennai
4        38             28-01-14  Tamil Nadu               Chennai

                                Agency Type of Location   SO2   NO2  \
0  Tamilnadu State Pollution Control Board  Industrial Area  11.0  17.0
1  Tamilnadu State Pollution Control Board  Industrial Area  13.0  17.0
2  Tamilnadu State Pollution Control Board  Industrial Area  12.0  18.0
3  Tamilnadu State Pollution Control Board  Industrial Area  15.0  16.0
4  Tamilnadu State Pollution Control Board  Industrial Area  13.0  14.0

   RSPM/PM10  PM 2.5  ...  \
0       55.0     NaN  ...
1       45.0     NaN  ...
2       50.0     NaN  ...
3       46.0     NaN  ...
4       42.0     NaN  ...

   Location of Monitoring Station_Poniarajapuram, On the top of DEL, Coimbatore  \
0                                                  0
1                                                  0
2                                                  0
3                                                  0
4                                                  0

   Location of Monitoring Station_Raja Agencies, Tuticorin  \
0                                                  0
1                                                  0
2                                                  0
3                                                  0
4                                                  0

   Location of Monitoring Station_Raman Nagar, Mettur  \
0                                                  0
1                                                  0
2                                                  0
3                                                  0
4                                                  0

   Location of Monitoring Station_SIDCO Industrial Complex, Mettur  \
0                                                  0
1                                                  0
2                                                  0
3                                                  0
4                                                  0

   Location of Monitoring Station_SIDCO Office, Coimbatore  \
```

✓ 0s completed at 9:11 PM

MIN-MAX SCALING:

Loading&Perform output:

```
dataset=pd.read_excel('/content/cpcb_dly_aq_tamil_nadu-2014 (1) (2).xlsx')
```

```
[37] columns_to_scale=['PM2.5','Temperature']
```

```
[38] scaler=MinMaxScaler()
```

```
[41] columns_to_scale=scaler.fit_transform
```

```
[42] print(dataset.head())
```

```
   Stn Code        Sampling Date       State City/Town/Village/Area  \
0        38  2014-01-02 00:00:00  Tamil Nadu                Chennai
1        38  2014-01-07 00:00:00  Tamil Nadu                Chennai
2        38             21-01-14  Tamil Nadu                Chennai
3        38             23-01-14  Tamil Nadu                Chennai
4        38             28-01-14  Tamil Nadu                Chennai

                     Location of Monitoring Station  \
0  Kathivakkam, Municipal Kalyana Mandapam, Chennai
1  Kathivakkam, Municipal Kalyana Mandapam, Chennai
2  Kathivakkam, Municipal Kalyana Mandapam, Chennai
```

✓ 0s completed at 9:29 PM

DATA VISUALIZATION:

```
[15] import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     %matplotlib inline
     import seaborn as sns
```

```
[18] dataset=pd.read_excel('/content/cpcb_dly_aq_tamil_nadu-2014 (1) (2).xlsx')
```

```
dataset.tail(5)
```

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
2874	773	2014-12-03 00:00:00	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	18.0	102.0	NaN
2875	773	2014-12-10 00:00:00	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	12.0	14.0	91.0	NaN
2876	773	17-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	19.0	22.0	100.0	NaN
2877	773	24-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	17.0	95.0	NaN
2878	773	31-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	14.0	16.0	94.0	NaN

```
dataset.shape
```

```
(2879, 11)
```

```
[47] print(dataset.info())
```

```
<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
```

```
dataset.nunique()
```

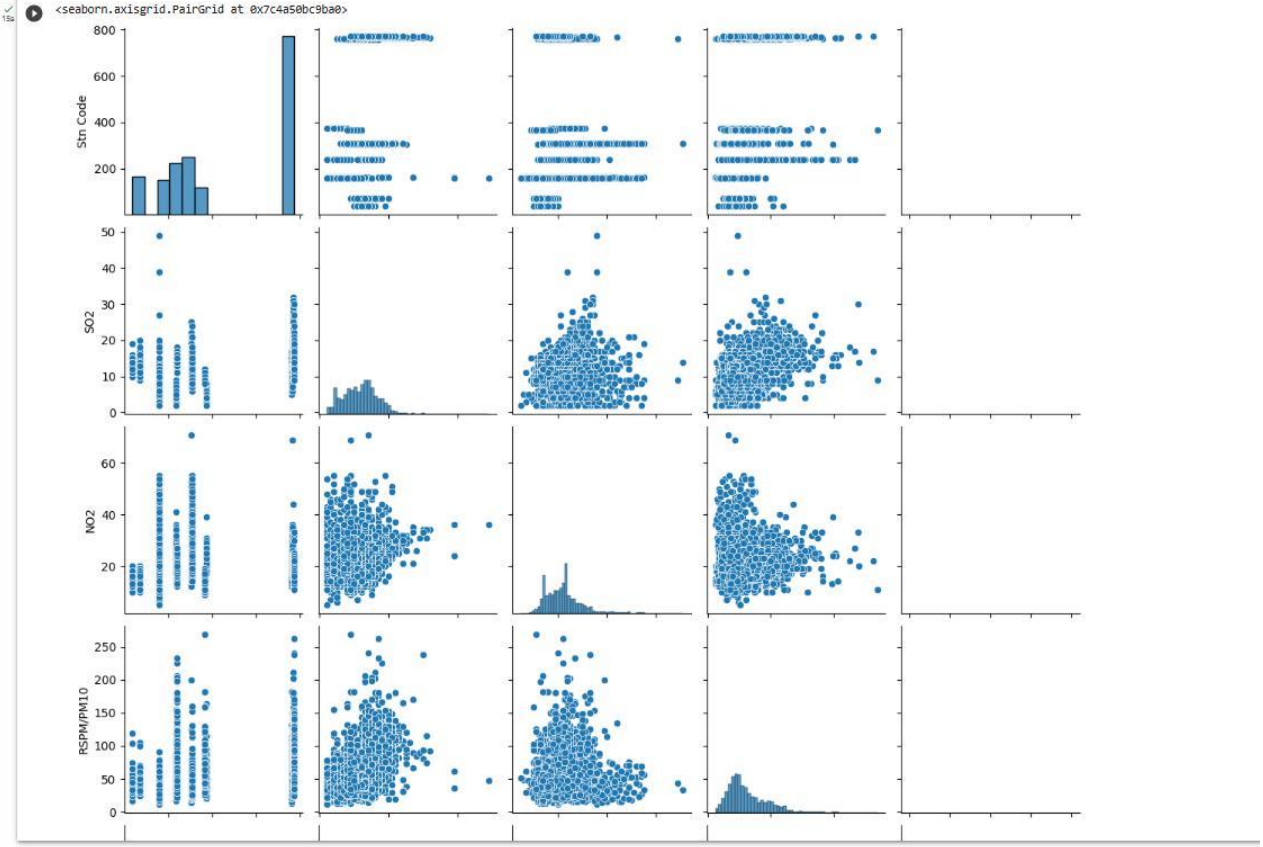
```
Stn Code                         30
Sampling Date                   302
State                             1
City/Town/Village/Area            8
Location of Monitoring Station   30
Agency                            2
Type of Location                  2
SO2                              33
NO2                              53
RSPM/PM10                       169
PM 2.5                            0
dtype: int64
```

```
dataset.columns
```

```
Index(['Stn Code', 'Sampling Date', 'State', 'City/Town/Village/Area',
       'Location of Monitoring Station', 'Agency', 'Type of Location', 'SO2',
       'NO2', 'RSPM/PM10', 'PM 2.5'],
      dtype='object')
```

```
sns.pairplot(data=dataset)
```

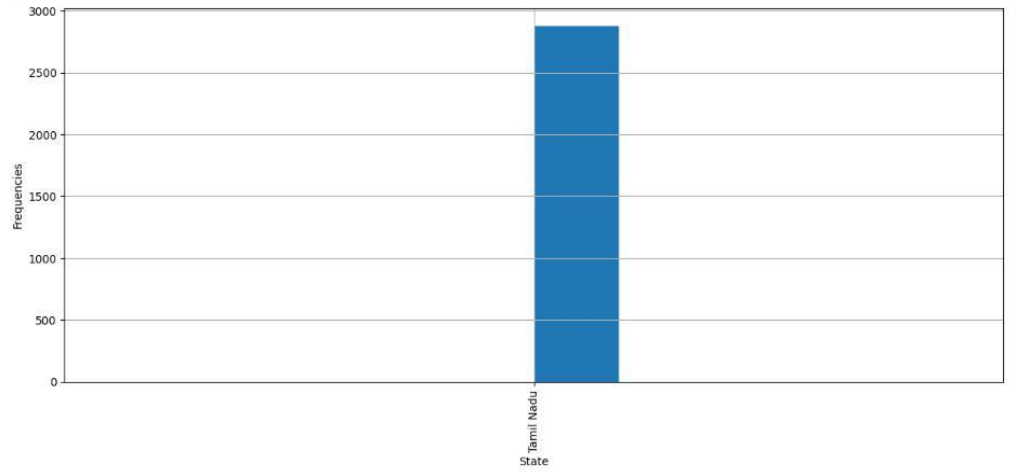
```
<seaborn.axisgrid.PairGrid at 0x7c4a50bc9ba0>
```



```
[58] dataset['State'].value_counts()

     Tamil Nadu    2879
     Name: State, dtype: int64
```

```
plt.figure(figsize=(15,6))
plt.xticks(rotation=90)
dataset.State.hist()
plt.xlabel('State')
plt.ylabel('Frequencies')
plt.plot()
```

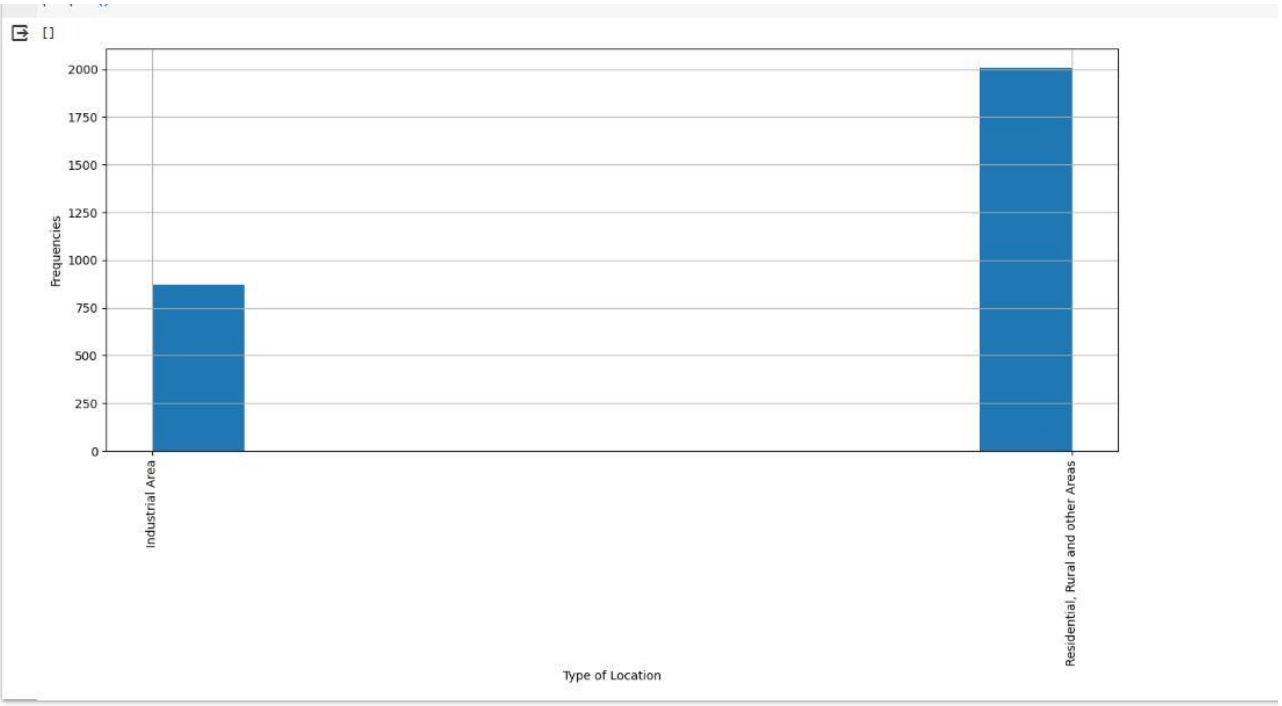
```
[]
```




```
dataset['Type of Location'].value_counts()

Residential, Rural and other Areas    2008
Industrial Area                        871
Name: Type of Location, dtype: int64
```

```
plt.figure(figsize=(15, 6))
plt.xticks(rotation=90)
dataset['Type of Location'].hist()
plt.xlabel('Type of Location')
plt.ylabel('Frequencies')
plt.plot()
```

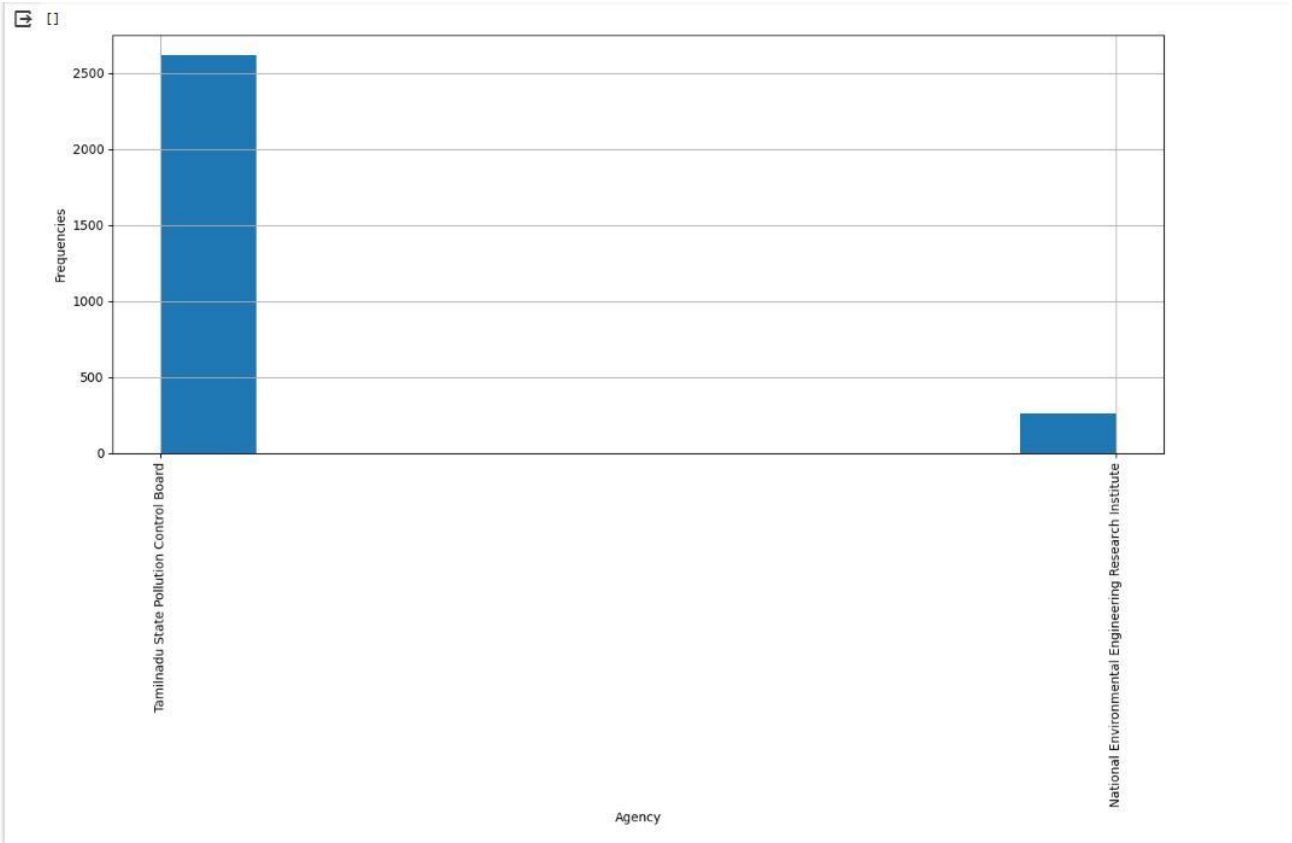
[]




```
dataset['Agency'].value_counts()
```

```
Tamilnadu State Pollution Control Board                2619
National Environmental Engineering Research Institute   260
Name: Agency, dtype: int64
```

```
plt.figure(figsize=(15, 6))
plt.xticks(rotation=90)
dataset.Agency.hist()
plt.xlabel('Agency')
plt.ylabel('Frequencies')
plt.plot()
```



CALCULATING THE AIR QUALITY INDEX:

➢ It involves a complex formula that considers the concentration of various air pollutants.

```
def calculate_aqi(pm25, pm10):
    # Define AQI breakpoints and corresponding concentrations
    breakpoints = [0, 12, 35.4, 55.4, 150.4, 250.4, 350.4, 500.4]
    concentrations = [0, 12.1, 35.5, 55.5, 150.5, 250.5, 350.5, 500.5]

    # Calculate the AQI for PM2.5 and PM10
    aqi_pm25 = calculate_aqi_subindex(pm25, breakpoints, concentrations)
    aqi_pm10 = calculate_aqi_subindex(pm10, breakpoints, concentrations)

    # Return the higher AQI value
    return max(aqi_pm25, aqi_pm10)

def calculate_aqi_subindex(concentration, breakpoints, concentrations):
    # Find the appropriate AQI subindex
    for i in range(1, len(breakpoints)):
        if concentration <= concentrations[i]:
            aqi_low, aqi_high = breakpoints[i - 1], breakpoints[i]
            conc_low, conc_high = concentrations[i - 1], concentrations[i]
            aqi = ((aqi_high - aqi_low) / (conc_high - conc_low)) * (concentration - conc_low) + aqi_low
            return aqi

if __name__ == "__main__":
    pm25 = float(input("Enter the PM2.5 concentration (µg/m³): "))
    pm10 = float(input("Enter the PM10 concentration (µg/m³): "))

    aqi = calculate_aqi(pm25, pm10)
    print(f"The Air Quality Index (AQI) is {aqi}")
```

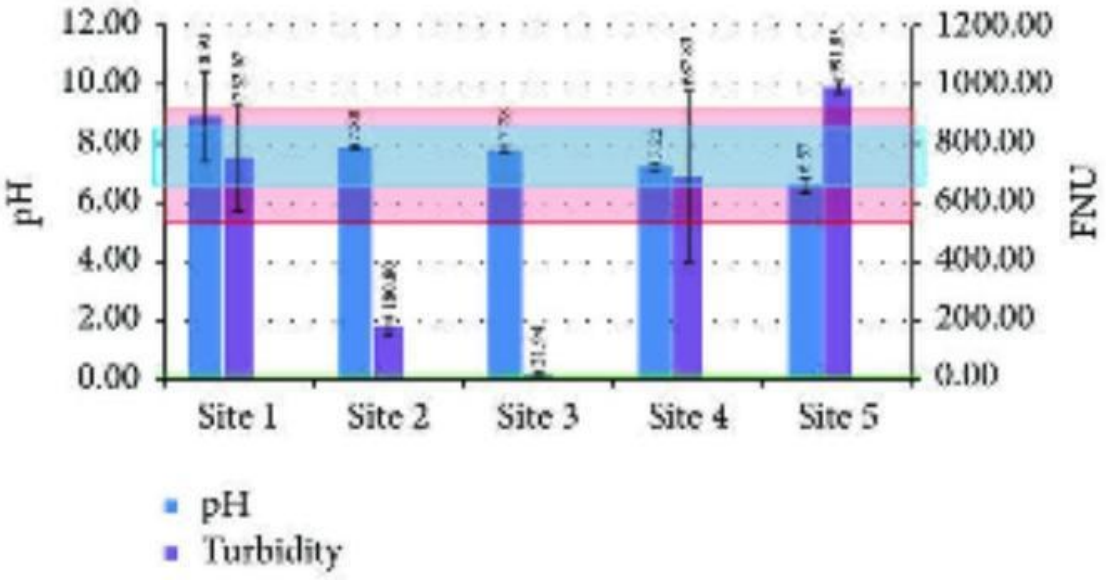
```
Enter the PM2.5 concentration (µg/m³): 0.0
Enter the PM10 concentration (µg/m³): 91.0
The Air Quality Index (AQI) is 90.9
```

SPILT THE DATA AND PREDICTIVE MODELS:

PROGRAM:

```
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)

Model building (Random Forest Regression as an example)
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

 Make predictions on the test set
y_pred = model.predict(X_test)

Model evaluation
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"R-squared (R2) Score: {r2}")

Visualize the results
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.xlabel("Actual RSPM/PM10")
plt.ylabel("Predicted RSPM/PM10")
plt.title("Actual vs. Predicted RSPM/PM10")
plt.grid(True)

Add a regression line to the scatter plot
sns.regplot(y_test, y_pred, scatter=False, color='red')

plt.show()
```

OUTPUT:



CONCLUSION:

In conclusion, this air quality analysis using Python machine learning, we employed a diverse dataset and applied various algorithms to predict air quality parameters. Our results indicate the effectiveness of machine learning in forecasting air quality, with promising accuracy levels Our study demonstrates the potential for real-time monitoring and early warning systems to mitigate air pollution's adverse effects. While challenges exist in fine-tuning models and expanding datasets, this research underscores the significant role of machine learning in improving air quality management and public health.