PHASE 3 - AIR QUALITY ANALYSIS AND PREDICTION IN TAMIL NADU

The project aims to analyze and visualize air quality data from monitoring stations in Tamil Nadu. The objective is to gain insights into air pollution trends, identify areas with high pollution levels, and develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 levels. This project involves defining objectives, designing the analysis approach, selecting visualization techniques, and creating a predictive model using Python and relevant libraries.

DEVELOPMENT PART 1:

Step 1: Data Loading

Data loading is the process of bringing external data into a format suitable for analysis. In this case, we've imported data in CSV format by utilizing the Pandas library and subsequently printed it to confirm the successful loading of the data.

Step 2: Explore the data

Exploring the data using the head() and info() function is a process of initially examining a dataset to understand its structure, content, and quality.

head() - This function displays the first few rows of the dataset.

info() - It displays information about the data types of each column, the number of non-null entries, and the memory usage.

Step 3: Data cleaning

To address the issue of missing values in the provided dataset, we can resolve it by filling those missing values with zeros.

- Check whether the data set contain any missing values
- Replace the missing values with zeros
- Save the preprocessed data to a new file
- Check missing values again to verify they are handled

Step 4: Data Analysis

This analysis aims to visually assess patterns and variations in SO2 levels across different locations (City/Town/Village/Area). It helps identify areas with notably high or low SO2 pollution levels, providing insights into air quality variations across different areas.

Step 5: Scatter Plot

It creates the scatter plot with the specified data, axis labels, color, size, and title.The plot visually represents the relationship between SO2, NO2, and RSPM/PM10 levels, with

color and marker size indicating RSPM/PM10 levels, making it easy to observer patterns and associations between these variables.

Code and Output:

```
import pandas as pd
data = pd.read_csv('/content/Air_quality_TN_Dataset.csv')
df = pd.DataFrame(data)
print(df)
```

```
Stn Code Sampling Date      State City/Town/Village/Area  \
0        38    01-02-14  Tamil Nadu           Chennai
1        38    01-07-14  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
...     ...       ...       ...              ...
2874     773   12-03-14  Tamil Nadu           Trichy
2875     773   12-10-14  Tamil Nadu           Trichy
2876     773   17-12-14  Tamil Nadu           Trichy
2877     773   24-12-14  Tamil Nadu           Trichy
2878     773   31-12-14  Tamil Nadu           Trichy

                Location of Monitoring Station  \
0     Kathivakkam, Municipal Kalyana Mandapam, Chennai
1     Kathivakkam, Municipal Kalyana Mandapam, Chennai
2     Kathivakkam, Municipal Kalyana Mandapam, Chennai
3     Kathivakkam, Municipal Kalyana Mandapam, Chennai
4     Kathivakkam, Municipal Kalyana Mandapam, Chennai
...                                    ...
2874                  Central Bus Stand, Trichy
2875                  Central Bus Stand, Trichy
2876                  Central Bus Stand, Trichy
2877                  Central Bus Stand, Trichy
2878                  Central Bus Stand, Trichy

                          Agency  \
0     Tamilnadu State Pollution Control Board
1     Tamilnadu State Pollution Control Board
2     Tamilnadu State Pollution Control Board
3     Tamilnadu State Pollution Control Board
4     Tamilnadu State Pollution Control Board
...                                      ...
2874  Tamilnadu State Pollution Control Board
2875  Tamilnadu State Pollution Control Board
2876  Tamilnadu State Pollution Control Board
2877  Tamilnadu State Pollution Control Board
2878  Tamilnadu State Pollution Control Board

                 Type of Location   SO2   NO2  RSPM/PM10  PM 2.5
0                  Industrial Area  11.0  17.0       55.0    NaN
1                  Industrial Area  13.0  17.0       45.0    NaN
2                  Industrial Area  12.0  18.0       50.0    NaN
3                  Industrial Area  15.0  16.0       46.0    NaN
4                  Industrial Area  13.0  14.0       42.0    NaN
...                            ...   ...   ...        ...    ...
2874  Residential, Rural and other Areas  15.0  18.0      102.0    NaN
```

```
2875  Residential, Rural and other Areas  12.0  14.0      91.0    NaN
2876  Residential, Rural and other Areas  19.0  22.0     100.0    NaN
2877  Residential, Rural and other Areas  15.0  17.0      95.0    NaN
2878  Residential, Rural and other Areas  14.0  16.0      94.0    NaN

[2879 rows x 11 columns]
```

---

[6]

0s

```
print(data.head())
print(data.info())
```

output

```
  Stn Code Sampling Date       State City/Town/Village/Area  \
0       38     01-02-14  Tamil Nadu            Chennai
1       38     01-07-14  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
3  Kathivakkam, Municipal Kalyana Mandapam, Chennai
4  Kathivakkam, Municipal Kalyana Mandapam, 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

<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
```

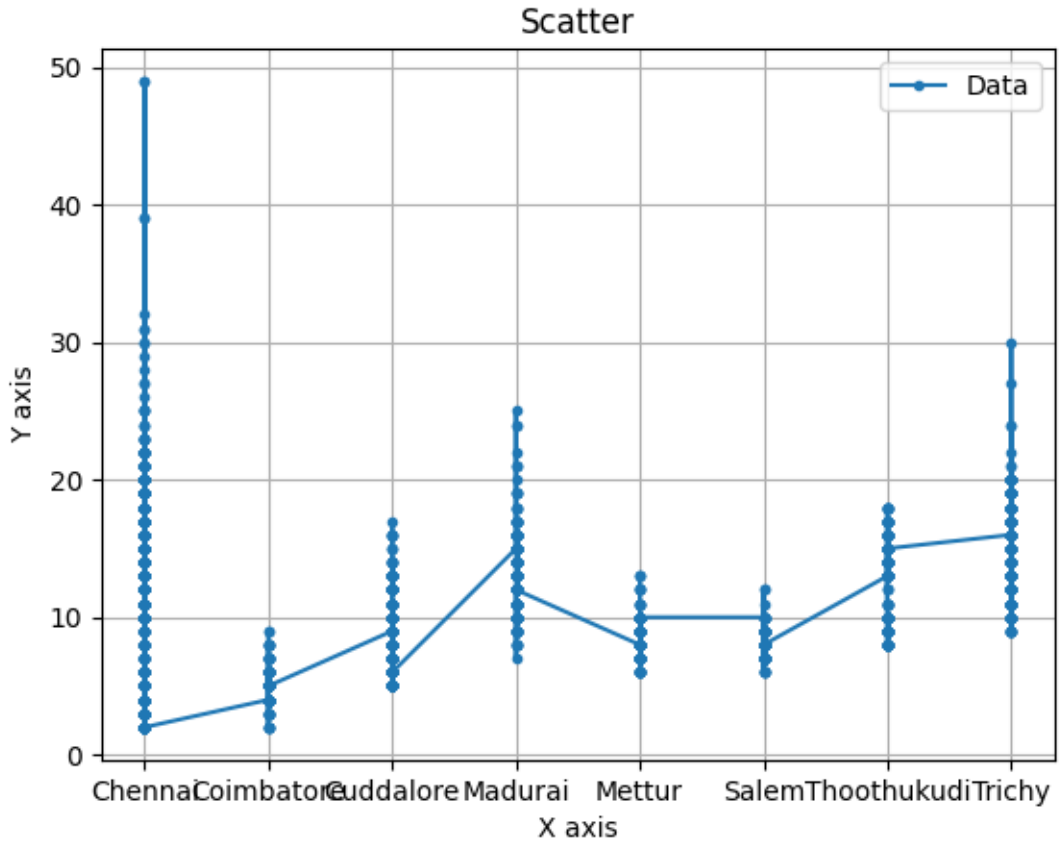
None

0s

```
df.dropna()
```

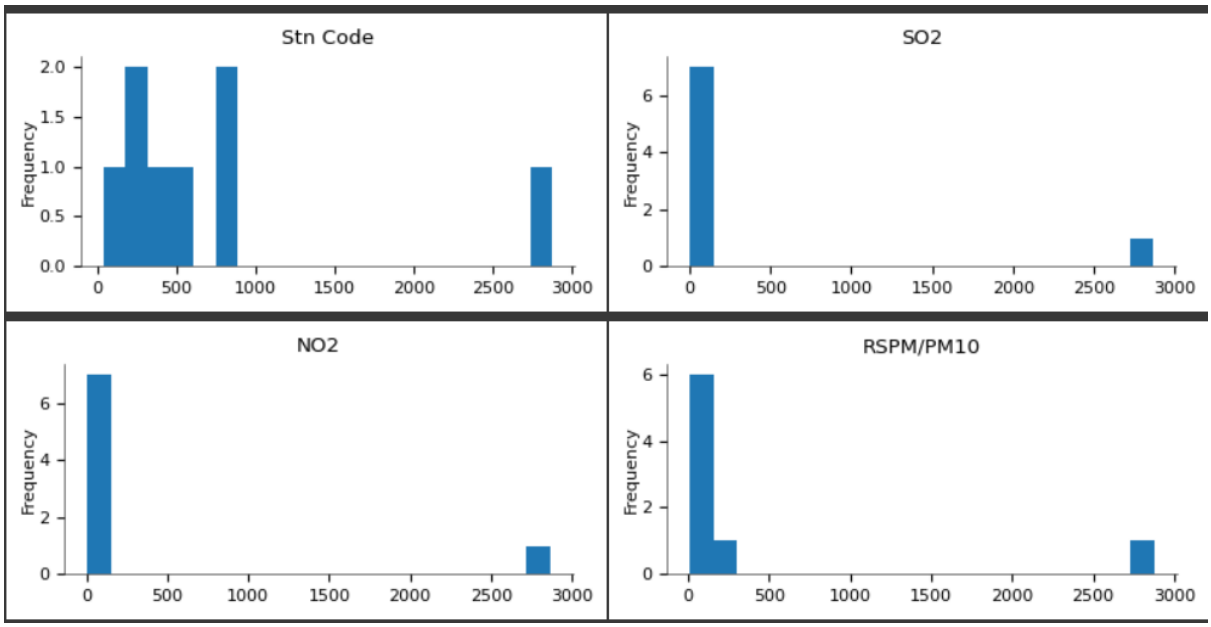
Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5

```
import matplotlib.pyplot as plt
x=data['City/Town/Village/Area']
y=data['SO2']
plt.plot(x,y,marker='.',linestyle='-',label='Data')
plt.xlabel("X axis")
plt.ylabel("Y axis")
plt.title("Scatter")
plt.legend()
plt.grid(True)
plt.show()
```

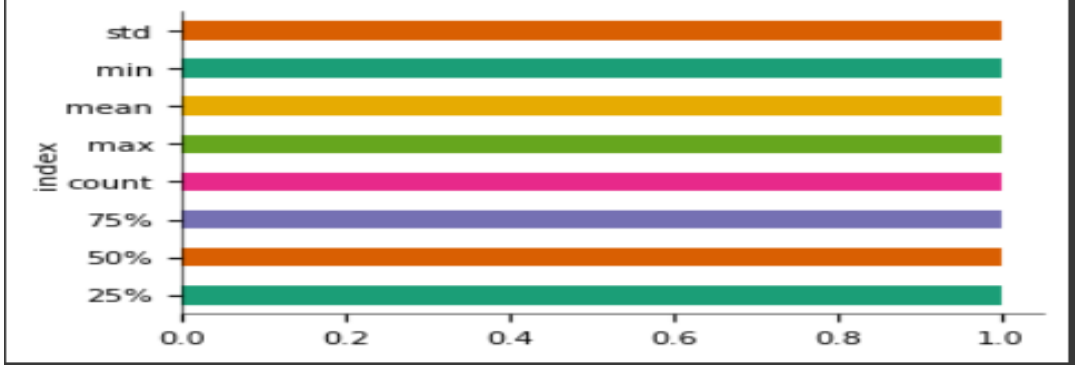



```
df.describe()
```

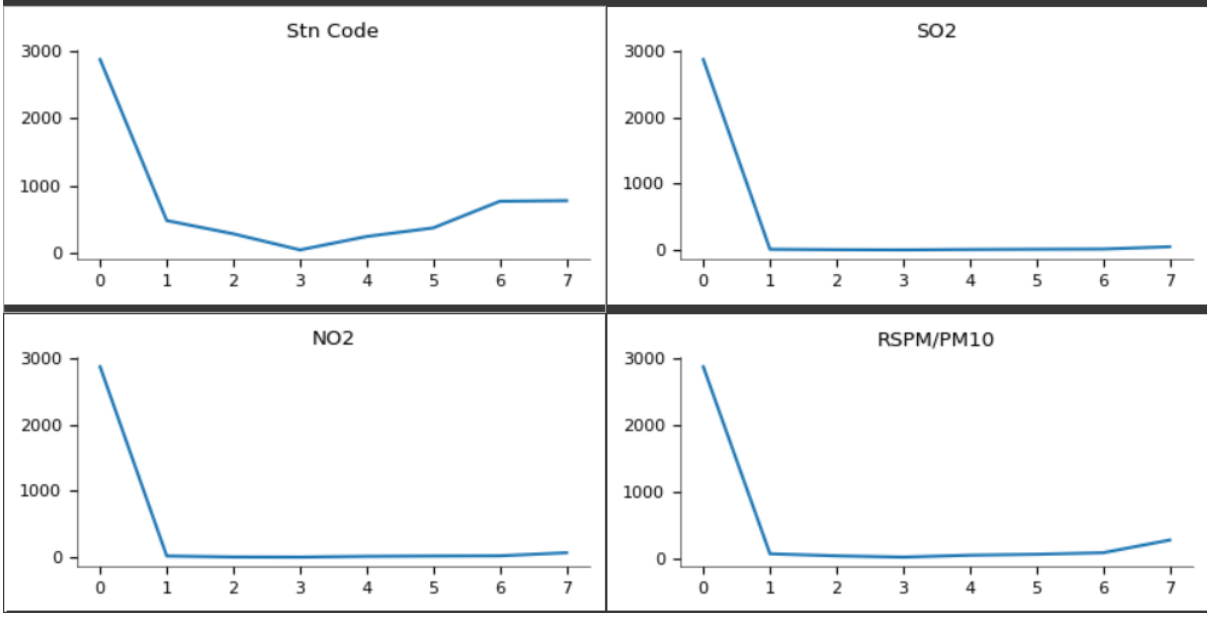
	Stn Code	SO2	NO2	RSPM/PM10	PM 2.5
count	2879.000000	2868.000000	2866.000000	2875.000000	0.0
mean	475.750261	11.503138	22.136776	62.494261	NaN
std	277.675577	5.051702	7.128694	31.368745	NaN
min	38.000000	2.000000	5.000000	12.000000	NaN
25%	238.000000	8.000000	17.000000	41.000000	NaN
50%	366.000000	12.000000	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



Categorical distributions



Values



Faceted distributions