Project : Air Quality Assessment TN

Phase 5:Documentation And Submission

Project Definition and Design Thinking

Project Definition:

The project aims to analyse and visualize air quality data from monitoring stations in TamilNadu.The objective is 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, selective visualization techniques and creating a predictive model using python and relevant libraries.

Design Thinking:

1.Project Objectives:

a.Analyzing Air Quality Trends: This objective involves studying and analyzing historical data on air quality parameters like RSPM and PM10 levels. The aim is to identify any patterns,seasonal variations, or long-term trends in air quality.

b.Identifying Pollution Hotspots: This objective focuses on pinpointing specific areas or regions with consistently high levels of air pollution. By analyzing the spatial distribution of air quality data, we can identify pollution hotspots and prioritize targeted pollution control measures.

c.Building a Predictive Model for RSPM/PM10 Levels: This objective involves developing a model that can forecast RSPM/PM10 levels in a given area. By analyzing historical air quality data and considering relevant factors like meteorological data, industrial activities, and traffic density, the model can provide predictions and early warnings for potential pollution episodes.

2.Analysis Approach:

a. Data Loading: Load the air quality data into a suitable data analysis tool or programming environment, such as Python or R. Use appropriate libraries or modules to read the data files or connect to the datafeeds.

b. **Data Preprocessing:** Clean and preprocess the data to remove any inconsistencies, missing values, or outliers. This may involve techniques such as data imputation, normalization, or data aggregation.

c. **Data Analysis:** Perform exploratory data analysis to understand the characteristics of the air quality data. Calculate summary statistics, identify trends or patterns, and explore relationships between different variables.

d. Data Visualization: Create visualizations to effectively communicate the analysis results. Use charts, graphs, maps, or interactive visualizations to present the air quality trends, pollution hotspots, and model predictions.

3.Visualization Selection:

To effectively represent air quality trends and pollution levels, here are some visualization techniques that can be used:

a. Line Charts: Line charts are useful for showing the trend of air quality parameters over time. Plotting the RSPM/PM10 levels on the y-axis and time on the x-axis, line charts can clearly illustrate the fluctuations, seasonal patterns, and long-term trends in air quality.

b. Heatmaps: Heatmaps can be used to visualize the spatial distribution of pollution levels. By mapping the RSPM/PM10 concentrations onto a geographic map, heatmaps provide a visual representation of pollution hotspots and areas with high pollutant concentrations. The intensity of the colors can be used to represent the magnitude of the pollution levels.

Innovation

Innovation:

In this phase we are going to put our design into Innovation and to incorporate machine learning algorithm to solve the problem.

Machine Learning Algorithms:

Let us discuss some of the machine learning algorithms that is more appropriate and used to build efficient predictive model for air quality assessment or analysis.

1.Linear Regression Model:

Linear Regression is a data analysis technique that predicts the value of unknown data by using another related and known data value . It is the relationship between dependent variable and the independent variable is a linear one.

=>Linear Regression is used to relate our data attributes such as SO2 , NO2 and RSPM whether they are linearly related or not.

=>It is used find the linear equation that best describes the correlation of the explanatory variables(SO2,NO2) with the dependent variable(RSPM).

2.Random Forest Regression:

Random forest is a supervised ML , Ensemble technique that combines the prediction from other models that by increasing the accuracy of our predictions.

=>Our air quality analysis dataset is divided into many different subsets and by using the subset different decision tress were created.The prediction is made by aggregating the results of many decision trees and then outputs the most optimal solution.

=>By using the Random forest classifier in our project we could improve our accuracy and can produce effective solution for our problem.

3.AdaBoost Algorithm:

Boosting refers to the algorithm which converts the weak learner into strong learner and do the prediction.

=>Adaboost Techinque gives the most accurate result as it changes it weights to get better prediction of our problem.

=>The more accuarte classifier will have more contibution to the final answer.

=>Adaboost has a stronger capability to explai the complex features contained In air quality data.

4.Artificial Neural Network(ANN):

A neural network is method in AI that teaches computer to process data in a way that is inspired by the human brain and it is also known as deep learning.

=>ANN has multilayered perceptron and the first input layer contains the input variable. The hidden layer is used optimize the ANN performace.The output layer cosist of target variable.Here SO2 ad NO2 are used as output variables.

=>ANN produces best prediction for air quality analysis compared to other models.

Development Part 1

Aim:

To start building our project by loading and preprocessing the dataset.

Loading of dataset:

Load air quality data set using python and data manipulation libraries like pandas.

Preprocessing of Data:

It is the process of converting raw data into clean data.Preprocessing includes the following steps,

=>Drop columns that are not useful

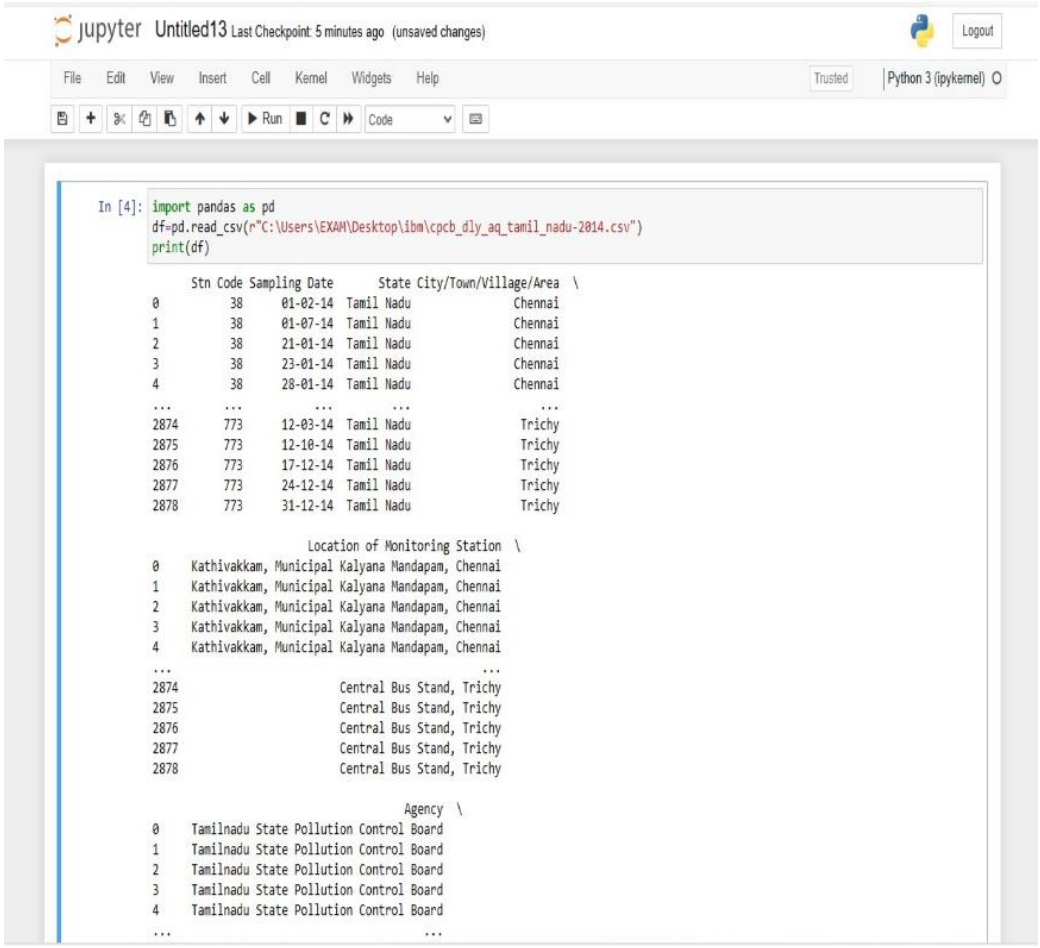
=>Drop rows with missing values

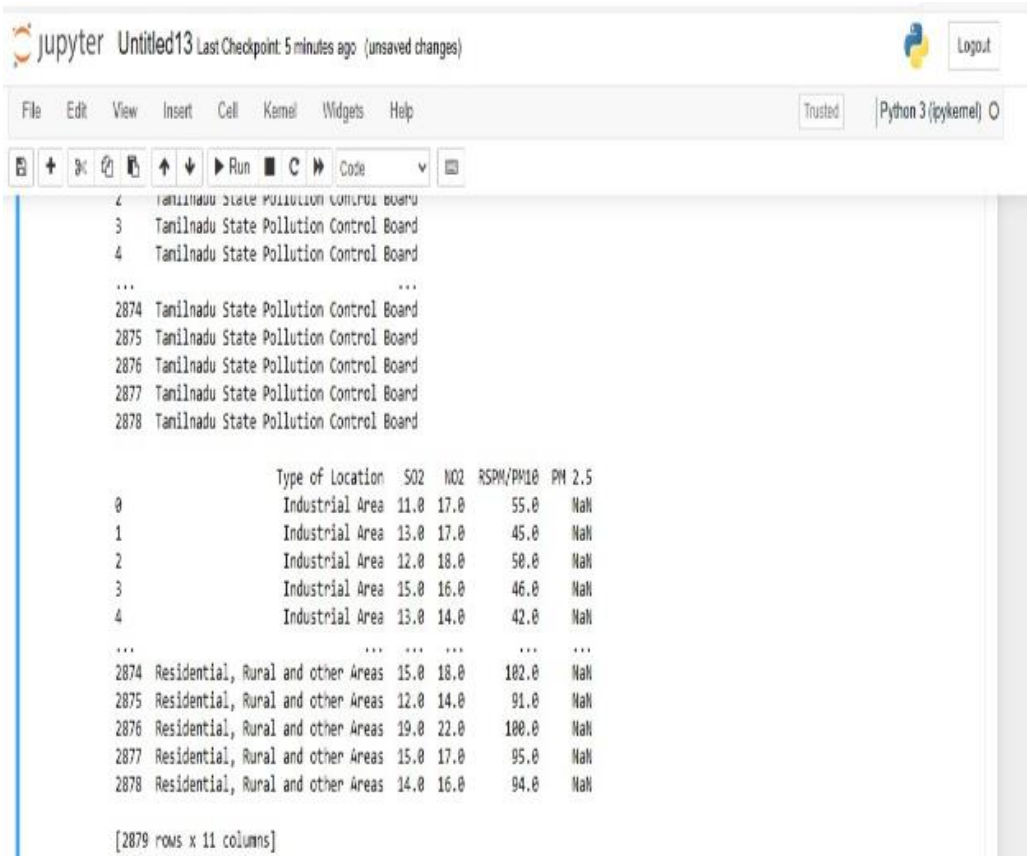
=>Take care of missing data

Code:

```
#Loading   dataset
import pandas as pd
df =pd.read_csv(r”C\Users\Exam\Desktop\ibm\
    cpcb_dly_aq_tamil_nadu-2014.csv”
```

print(df)


jupyter Untitled13 Last Checkpoint: 5 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [4]:
import pandas as pd
df=pd.read_csv(r"C:\Users\EXAM\Desktop\ibm\cpcb_dly_aq_tamil_nadu-2014.csv")
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
... ...




```
2    Tanilnadu State Pollution Control Board
3    Tanilnadu State Pollution Control Board
4    Tanilnadu State Pollution Control Board
...                                      ...
2874 Tanilnadu State Pollution Control Board
2875 Tanilnadu State Pollution Control Board
2876 Tanilnadu State Pollution Control Board
2877 Tanilnadu State Pollution Control Board
2878 Tanilnadu 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]
```

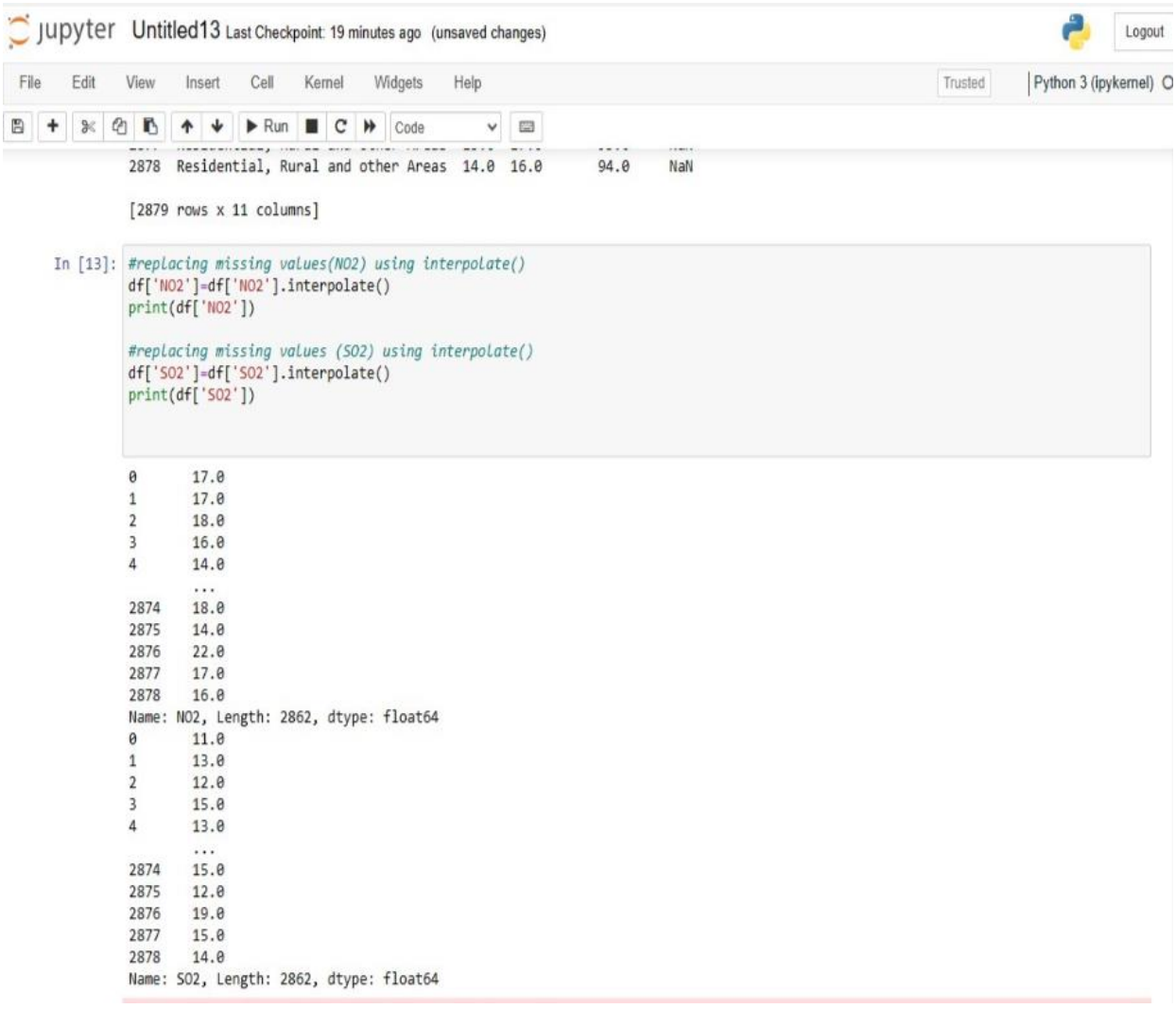
Prepcessing of dataset:

```
#Drop the useless columns
df=df.drop('PM 2.5',axis=1)

#Take care of missing data
df['NO2']=df['NO2'].interpolate()
print(df['NO2'])

df['SO2']=df['SO2'].interpolate
```

print(df[‘SO2’])

#Drop the Null values

df=df.dropna()



Development Part 4

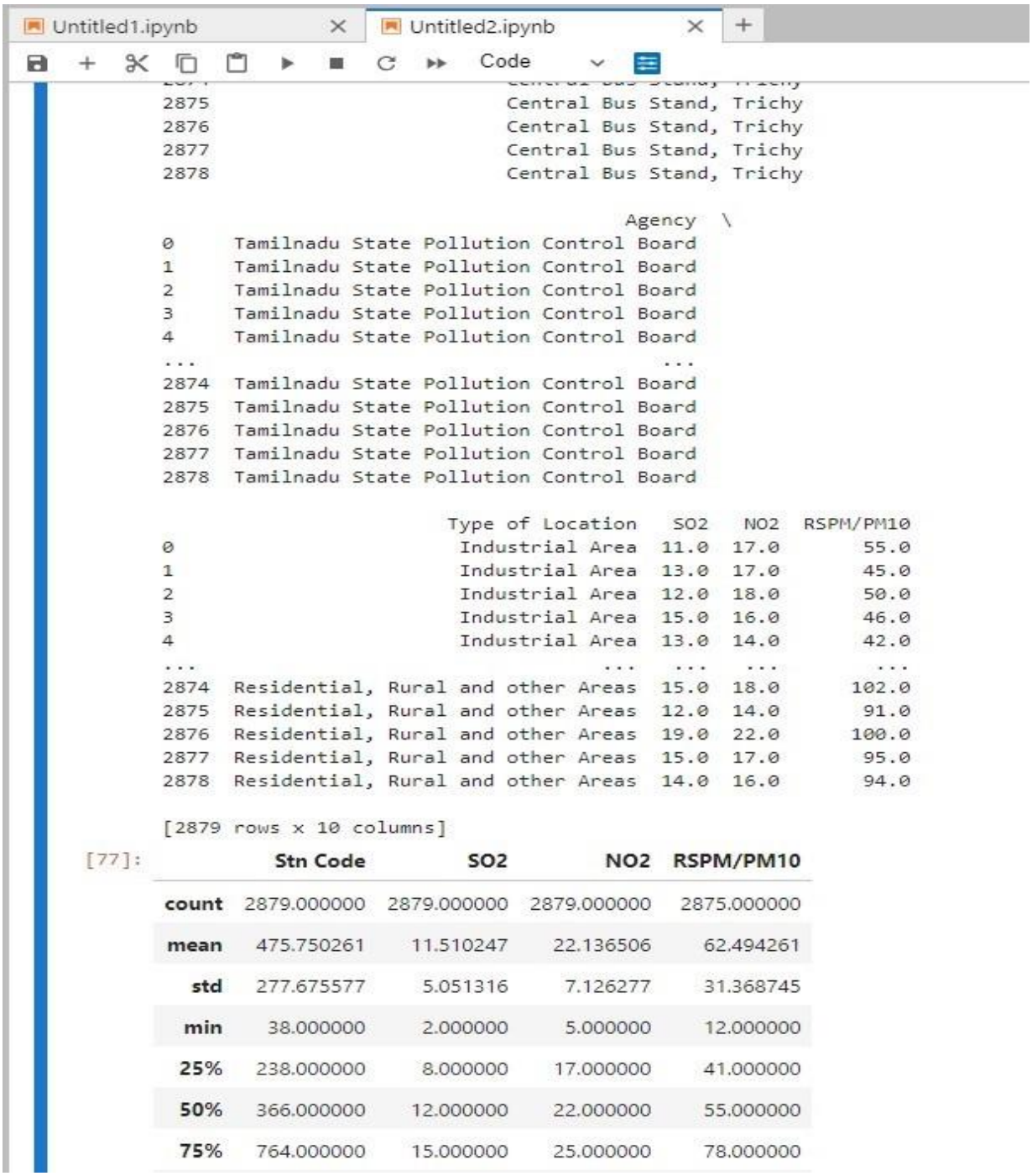
Aim:

To Calculate the average SO2,NO2,RSPM/ PM10 Levels across different monitoring stations ,cities, States.To identify the pollution trends and areas With high pollution levels.To create visualization Using Data Visualization libraries like Seaborn , Matplotlib,etc.

1.Python code:

```
import pandas as pd
df=pd.read_csv(“C:User\admin\Desktop\ibm\air.csv”)
df['SO2']=df['SO2'].interpolate()
df['NO2']=df['NO2'].interpolate()
df=df.drop(columns=df.columns[-1],
axis=1,inplace=False)
print(df)
df.describe()
```

Output:




```
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
0                          Industrial Area  11.0  17.0       55.0
1                          Industrial Area  13.0  17.0       45.0
2                          Industrial Area  12.0  18.0       50.0
3                          Industrial Area  15.0  16.0       46.0
4                          Industrial Area  13.0  14.0       42.0
...                                    ...   ...   ...        ...
2874  Residential, Rural and other Areas  15.0  18.0      102.0
2875  Residential, Rural and other Areas  12.0  14.0       91.0
2876  Residential, Rural and other Areas  19.0  22.0      100.0
2877  Residential, Rural and other Areas  15.0  17.0       95.0
2878  Residential, Rural and other Areas  14.0  16.0       94.0

[2879 rows x 10 columns]
```

[77]:

	Stn Code	SO2	NO2	RSPM/PM10
count	2879.000000	2879.000000	2879.000000	2875.000000
mean	475.750261	11.510247	22.136506	62.494261
std	277.675577	5.051316	7.126277	31.368745
min	38.000000	2.000000	5.000000	12.000000
25%	238.000000	8.000000	17.000000	41.000000
50%	366.000000	12.000000	22.000000	55.000000
75%	764.000000	15.000000	25.000000	78.000000

2.Finding average for SO2,NO2,RSPM/PM10 on different Locations.

Code:

#Average levels in chennai

chennai = df.loc[df['City/Town/Village/Area'] == 'Chennai']

avg_so2_chenn=chennai['SO2'].mean()

avg_no2_chenn=chennai['NO2'].mean()

avg_rspm_chenn=chennai['RSPM/PM10'].mean()

print(avg_so2_chenn)

print(avg_no2_chenn)

print(avg_rspm_chenn)

Output:

```
import matplotlib.pyplot as plt
import seaborn as sns
chennai = df.loc[df['City/Town/Village/Area'] == 'Chennai']
avg_so2_chenn=chennai['SO2'].mean()
avg_no2_chenn=chennai['NO2'].mean()
avg_rspm_chenn=chennai['RSPM/PM10'].mean()
print("Average SO2 in chennai",avg_so2_chenn)
print("Average NO2 in chennai",avg_no2_chenn)
print("Average RSPM in chennai",avg_rspm_chenn)
```

```
Average SO2 in chennai 13.025
Average NO2 in chennai 22.1035
Average RSPM in chennai 58.998
```

#Average levels in Trichy:

```
import matplotlib.pyplot as plt
import seaborn as sns
Trichy = df.loc[df['City/Town/Village/Area'] == 'Trichy']
avg_so2_tri=Trichy['SO2'].mean()
avg_no2_tri=Trichy['NO2'].mean()
avg_rspm_tri=Trichy['RSPM/PM10'].mean()
print("Average SO2 in Trichy",avg_so2_tri)
print("Average NO2 in Trichy",avg_no2_tri)
print("Average RSPM in Trichy",avg_rspm_tri)
```

Output:

```
Average SO2 in Trichy 15.279291553133515
Average NO2 in Trichy 18.682561307901906
Average RSPM in Trichy 85.05449591280654
```

#Average levels in Coimbatore

```
import matplotlib.pyplot as plt
import seaborn as sns
coim = df.loc[df['City/Town/Village/Area'] == 'Coimbatore']
avg_so2_coi=coim['SO2'].mean()
avg_no2_coi=coim['NO2'].mean()
avg_rspm_coi=coim['RSPM/PM10'].mean()
print("Average SO2 in Coimbatore ",avg_so2_coi)
print("Average NO2 in Coimbatore",avg_no2_coi)
print("Average RSPM in Coimbatore",avg_rspm_coi)
```

Output:

```
Average SO2 in Coimbatore  4.546075085324232
Average NO2 in Coimbatore 25.339590443686006
Average RSPM in Coimbatore 49.217241379310344
```

#Average levels in Mettur

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
mettur= df.loc[df['City/Town/Village/Area'] == 'Mettur']
avg_so2_mett=mettur['SO2'].mean()
avg_no2_mett=mettur['NO2'].mean()
avg_rspm_mett=mettur['RSPM/PM10'].mean()
print("Average SO2 in Mettur ",avg_so2_mett)
print("Average NO2 in Mettur",avg_no2_mett)
print("Average RSPM in Mettur",avg_rspm_mett)
```

Output:

#Average levels in Thoothukudi

```
import matplotlib.pyplot as plt
import seaborn as sns
thoo= df.loc[df['City/Town/Village/Area'] == 'Thoothukudi']
avg_so2_thoo=thoo['SO2'].mean()
avg_no2_thoo=thoo['NO2'].mean()
avg_rspm_thoo=thoo['RSPM/PM10'].mean()
print("Average SO2 in Thoothukudi ",avg_so2_thoo)
print("Average NO2 in Thoothukudir",avg_no2_thoo)
print("Average RSPM in Thoothukudi",avg_rspm_thoo)
```

Output:

```
import matplotlib.pyplot as plt
import seaborn as sns
thoo= df.loc[df['City/Town/Village/Area'] == 'Thoothukudi']
avg_so2_thoo=thoo['SO2'].mean()
avg_no2_thoo=thoo['NO2'].mean()
avg_rspm_thoo=thoo['RSPM/PM10'].mean()
print("Average SO2 in Thoothukudi ",avg_so2_thoo)
print("Average NO2 in Thoothukudir",avg_no2_thoo)
print("Average RSPM in Thoothukudi",avg_rspm_thoo)
```

```
Average SO2 in Thoothukudi  12.988054607508532
Average NO2 in Thoothukudir 18.503412969283275
Average RSPM in Thoothukudi 83.45890410958904
```

Visualization:

#For SO2

import matplotlib.pyplot as plt

data=[avg_so2_tri,avg_so2_coi,avg_so2_mett,avg_so2_thoo,avg_so2_chennai]

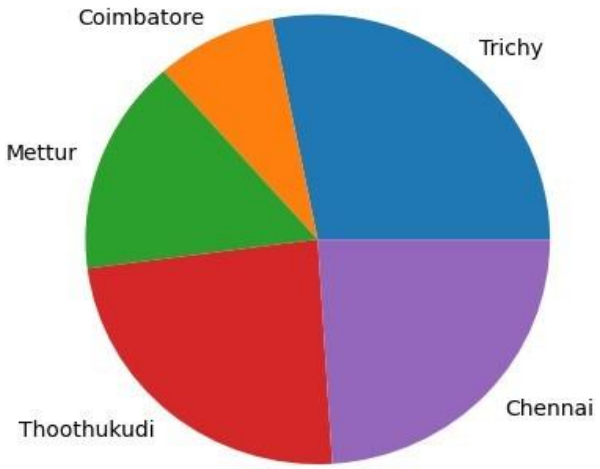
mylabels=["Trichy","Coimbatore","Mettur", "Thoothukudi","Chennai"]

plt.pie(data,labels=mylabels)

plt.show()

Output:

```
import matplotlib.pyplot as plt
data=[avg_so2_tri,avg_so2_coi,avg_so2_mett,avg_so2_thoo,avg_so2_chenn]
mylabels=["Trichy","Coimbatore","Mettur","Thoothukudi","Chennai"]
plt.pie(data,labels=mylabels)
plt.show()
```



#For NO2

import matplotlib.pyplot as plt

data=[avg_no2_tri,avg_no2_coi,avg_no2_mett,

avg_no2_thoo,avg_no2_chenn]

mylabels=["Trichy","Coimbatore","Mettur",

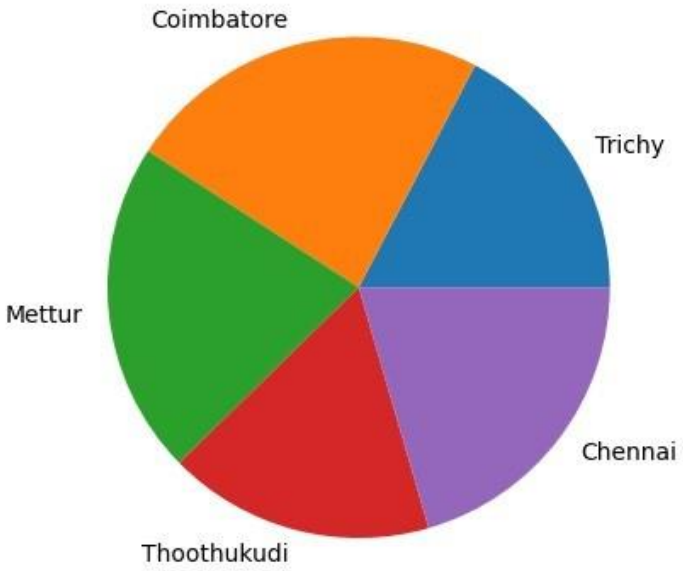
"Thoothukudi","Chennai"]

plt.pie(data,labels=mylabels)

plt.show()

Output:

```
import matplotlib.pyplot as plt
data=[avg_no2_tri,avg_no2_coi,avg_no2_mett,avg_no2_thoo,avg_no2_chenn]
mylabels=["Trichy","Coimbatore","Mettur","Thoothukudi","Chennai"]
plt.pie(data,labels=mylabels)
plt.show()
```



#For RSPM/PM10

import matplotlib.pyplot as plt

Data=[avg_rspm_tri,avg_rspm_coi,avg_rspm_mett,avg_rspm_thoo,avg_rspm_chenn]

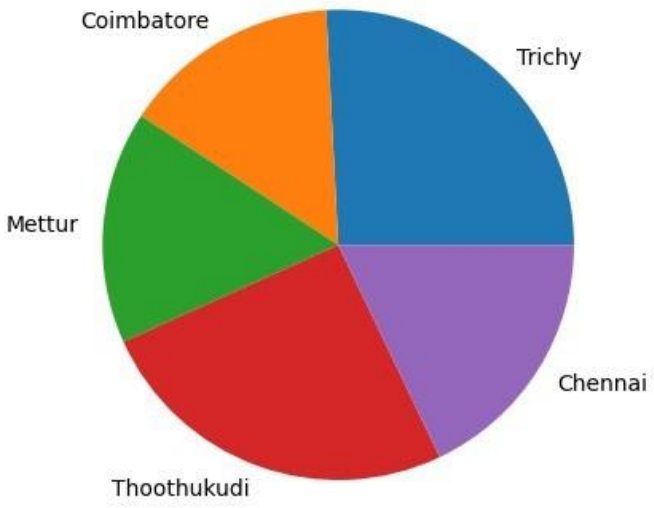
Mylabels=["Trichy","Coimbatore","Mettur","Thoothukudi","Chennai"]

plt.pie(data,labels=mylabels)

plt.show()

Output:

```
import matplotlib.pyplot as plt
data=[avg_rspm_tri,avg_rspm_coi,avg_rspm_mett,avg_rspm_thoo,avg_rspm_chenn]
mylabels=["Trichy","Coimbatore","Mettur","Thoothukudi","Chennai"]
plt.pie(data,labels=mylabels)
plt.show()
```



Conclusion :

=> By refering to the above pie charts the average RSPM/PM10 level of Trichy and Thoothukudi is nearly same and it is considered as High pollution areas,

=>Also mettur has Low level of RSPM Level And it is considered as Low pollution area.