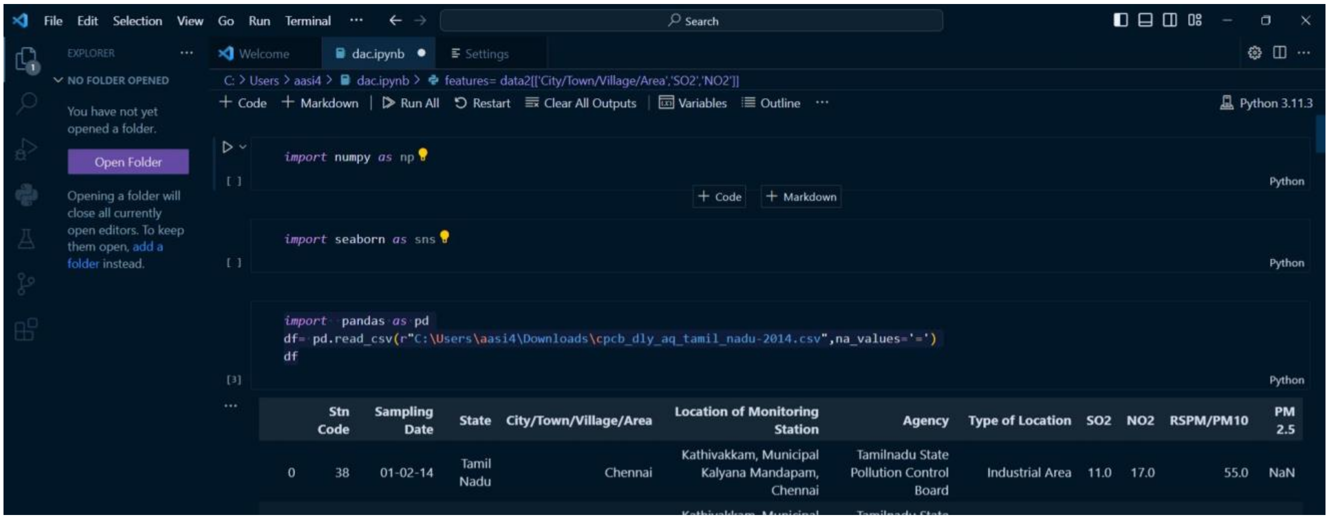
PHASE 3- DEVELOPMENT PART-1
AIR QUALITY ANALYSIS IN TAMILNADU

Import Libraries:

In this step, we import the necessary Python libraries, including pandas for data manipulation,pandas is a common library used in data analysis and Jupyter Notebook environments. If you have 'pandas' installed and are using it in your Jupyter Notebook, upgrading 'nbformat' is an independent step to ensure that you can render content properly, such as plots or visualizations, which might be related to other libraries like 'matplotlib' or 'plotly.'


File Edit Selection View Go Run Terminal ···
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
Welcome
dac.ipynb
Settings
C: > Users > aasi4 > dac.ipynb > features= data2[['City/Town/Village/Area','SO2','NO2']]
+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ···
Python 3.11.3
Python

```
[ ]
import numpy as np
```

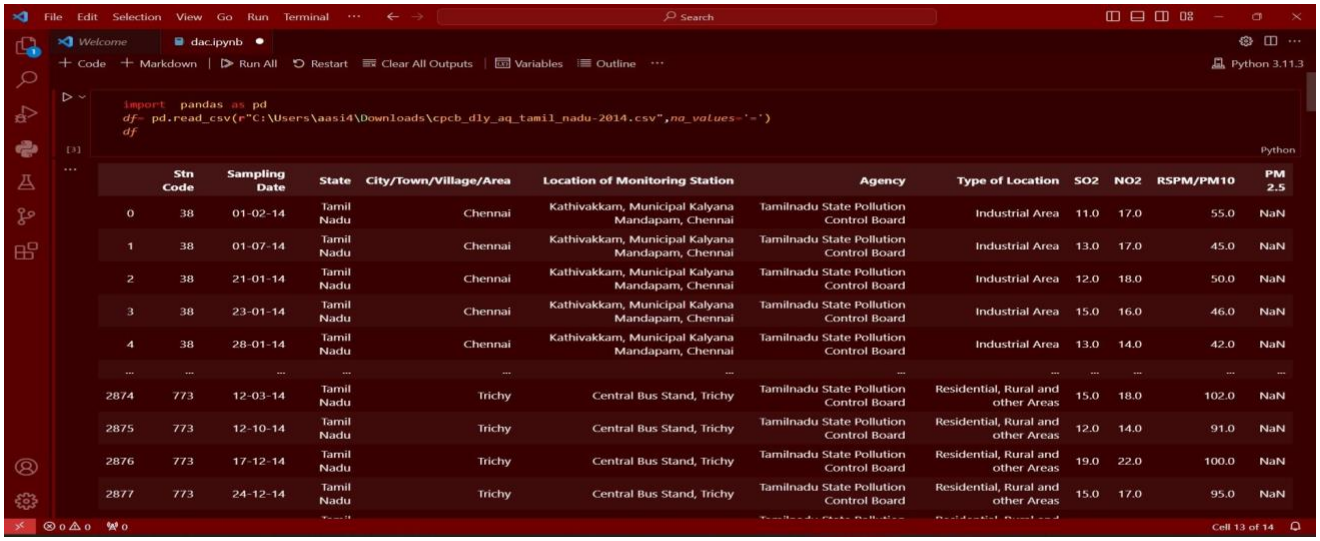
```
[ ]
import seaborn as sns
```

```
[3]
import  pandas as pd
df= pd.read_csv(r"C:\Users\aasi4\Downloads\cpcb_dly_aq_tamil_nadu-2014.csv",na_values='=')
df
```

	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	01-02-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN

Load the Dataset:

Once pandas is imported, you can load your dataset. You typically do this by providing the path to the dataset file (usually a CSV file) .in a CSV file, into a pandas DataFrame. Replace `"your_dataset.csv"` with the actual file path of your dataset.df is the name of the pandas DataFrame that will hold your dataset.pd.read_csv() is a pandas function designed to read CSV files and load them into a DataFrame."my_dataset.csv" should be replaced with the actual file path or URL of your dataset.



```
import  pandas as pd
df= pd.read_csv(r"C:\Users\aasi4\Downloads\cpcb_dly_aq_tamil_nadu-2014.csv",na_values='=')
df
```

[3]

	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	01-02-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN
1	38	01-07-14	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
...	...	...	...	...	...	...	...	...	...	...	...
2874	773	12-03-14	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	12-10-14	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

Explore the Dataset:

Exploring the Loaded Dataset:

After loading the dataset, it's a good practice to explore it and get a better understanding of its structure. You can use various pandas functions to achieve this:

Display the First Few Rows:

You can use df.head() to display the first few rows of your dataset. This helps you get an initial sense of the data's content.

Check Column Names and Data Types:

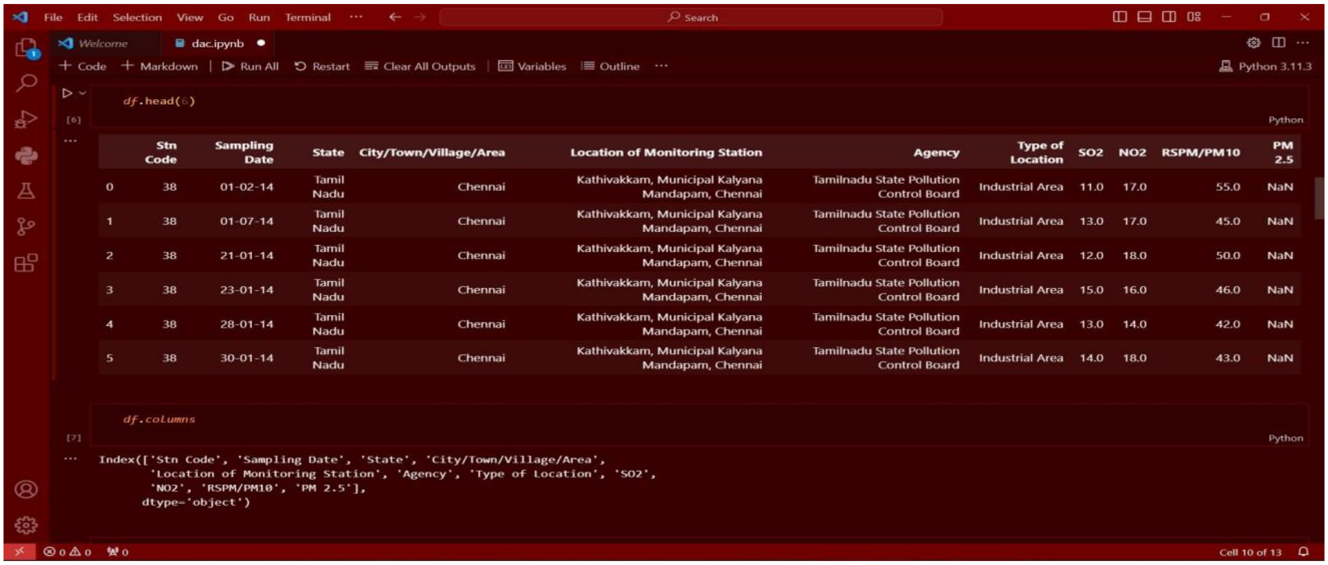
Use df.info() to check the column names, data types, and non-null counts for each column. This is useful for understanding the dataset's structure.

Check for Missing Values:

To identify missing values in your dataset, use df.isnull().sum(). This will show the count of missing values in each column.By loading and exploring your dataset, you set the foundation for data analysis, cleaning, and manipulation. Understanding the structure and content of your data is essential for making informed decisions and preparing it for further analysis.

Import Visualization Libraries:

First, you need to import the data visualization libraries you plan to use. Depending on your choice of library, you can import Matplotlib, Seaborn, or any other visualization tool.


File Edit Selection View Go Run Terminal ···
Search
Welcome
dac.ipynb
+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ···
Python 3.11.3
Python
Cell 10 of 13

```
df.head(6)
```

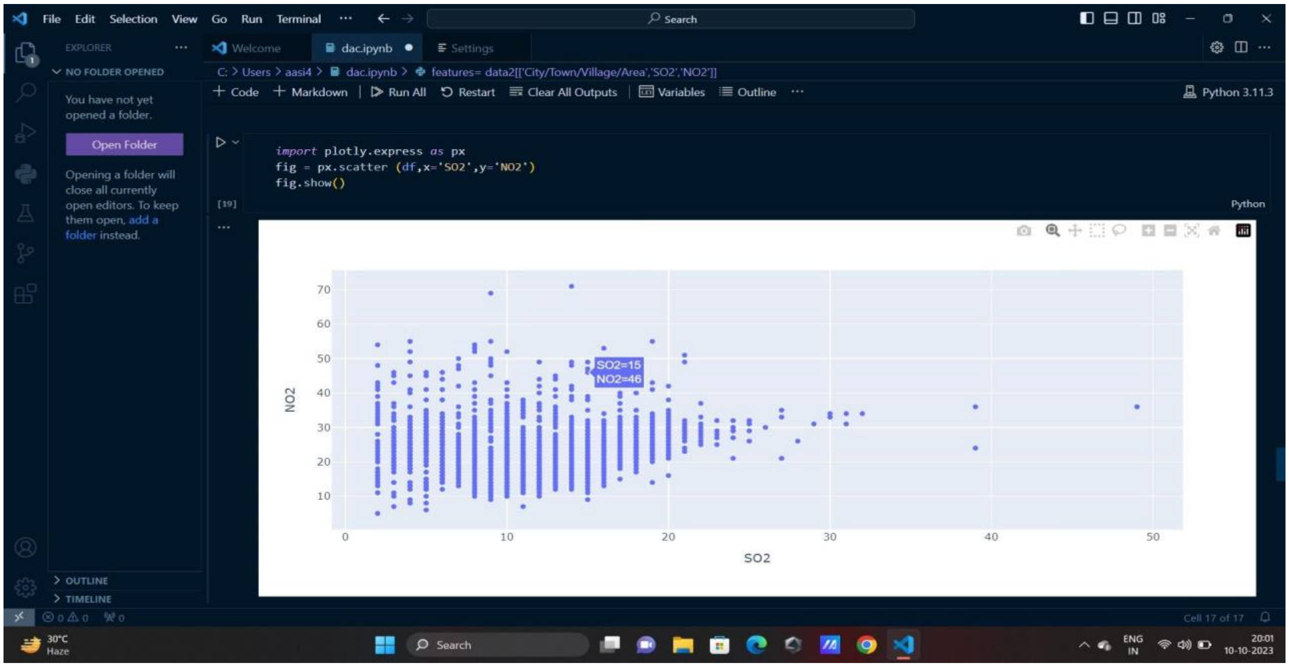
[6]

	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	01-02-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN
1	38	01-07-14	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
5	38	30-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	14.0	18.0	43.0	NaN

```
df.columns
```

[7]

```
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')
```

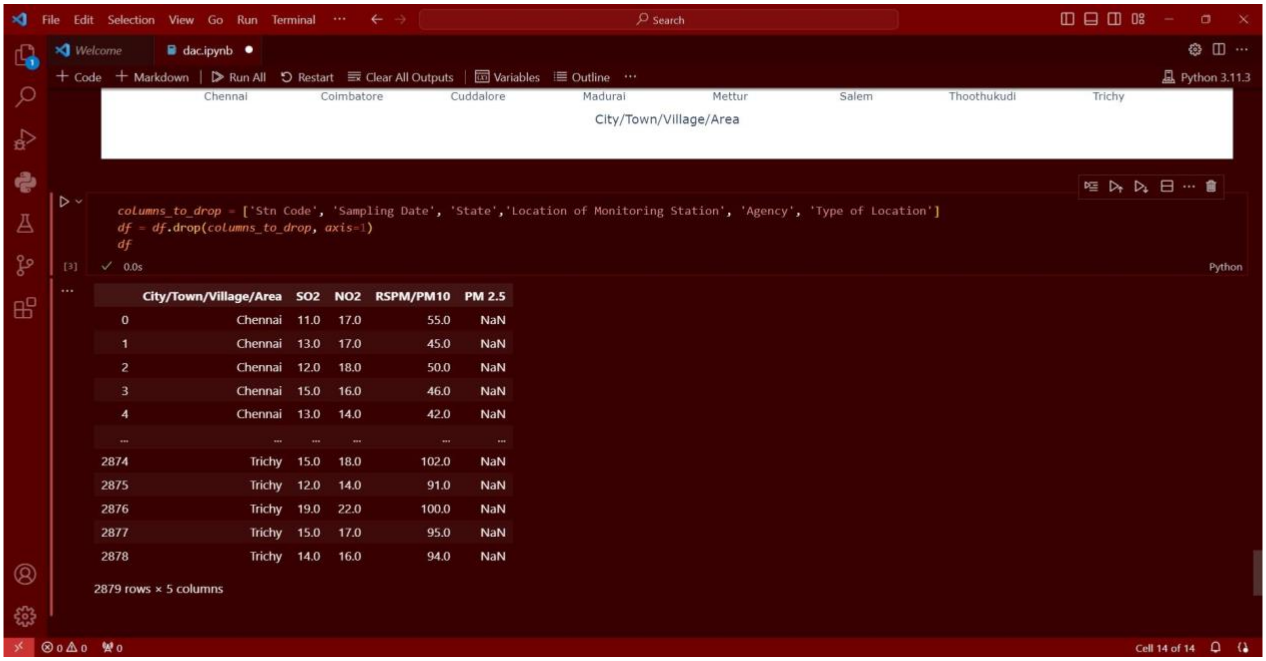

File Edit Selection View Go Run Terminal ···
Search
EXPLORER
NO FOLDER OPENED
Welcome
dac.ipynb
Settings
C: > Users > aasi4 > dac.ipynb > features= data2[['City/Town/Village/Area','SO2','NO2']]
+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ···
Python 3.11.3
Open Folder
OUTLINE
TIMELINE
Python
SO2=15
NO2=46
70
60
50
40
30
20
10
NO2
0
10
20
30
40
50
SO2
Cell 17 of 17
30°C
Haze
Search
ENG
IN
20:01
10-10-2023

You have not yet opened a folder.

Opening a folder will close all currently open editors. To keep them open, add a folder instead.

```
import plotly.express as px
fig = px.scatter (df,x='SO2',y='NO2')
fig.show()
```

[19]


File Edit Selection View Go Run Terminal ···
Search
Welcome
dac.ipynb
+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ···
Python 3.11.3
Chennai Coimbatore Cuddalore Madurai Mettur Salem Thoothukudi Trichy
City/Town/Village/Area
Python
Cell 14 of 14

```
columns_to_drop = ['Stn Code', 'Sampling Date', 'State','Location of Monitoring Station', 'Agency', 'Type of Location']
df = df.drop(columns_to_drop, axis=1)
df
```

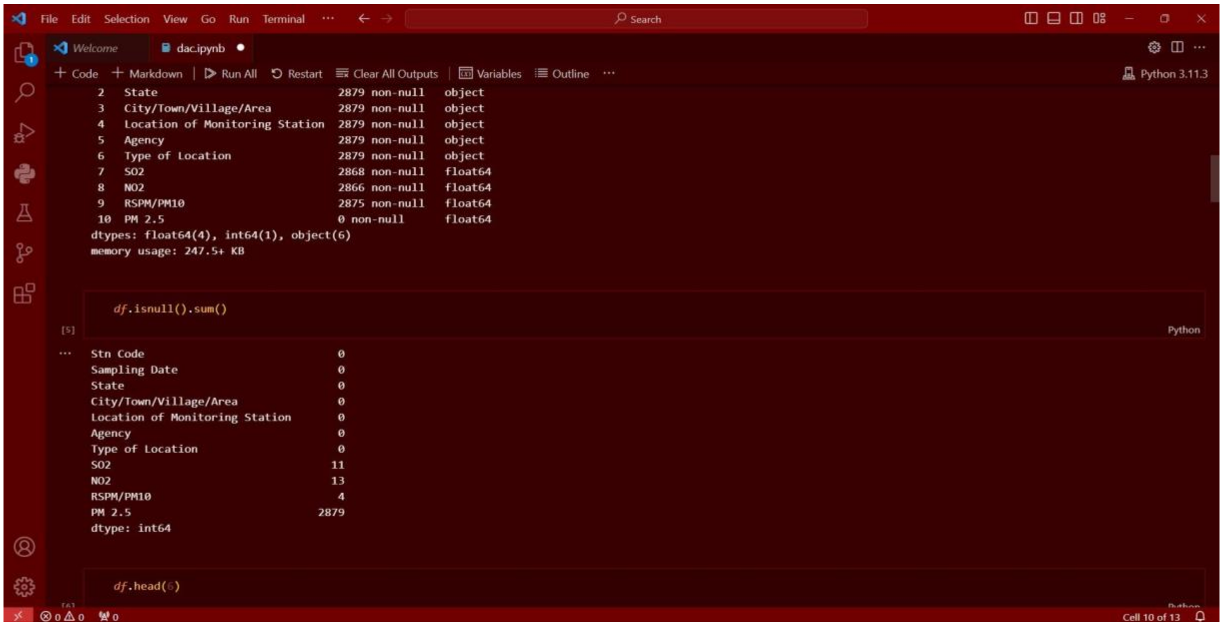
[3] ✓ 0.0s

	City/Town/Village/Area	SO2	NO2	RSPM/PM10	PM 2.5
0	Chennai	11.0	17.0	55.0	NaN
1	Chennai	13.0	17.0	45.0	NaN
2	Chennai	12.0	18.0	50.0	NaN
3	Chennai	15.0	16.0	46.0	NaN
4	Chennai	13.0	14.0	42.0	NaN
...	...	...	...	...	...
2874	Trichy	15.0	18.0	102.0	NaN
2875	Trichy	12.0	14.0	91.0	NaN
2876	Trichy	19.0	22.0	100.0	NaN
2877	Trichy	15.0	17.0	95.0	NaN
2878	Trichy	14.0	16.0	94.0	NaN

2879 rows × 5 columns

Handle Missing Values:

If there are missing values in your dataset, you'll need to decide how to handle them. Common strategies include removing rows with missing values, filling them with mean or median values, or


File Edit Selection View Go Run Terminal ···
Search
Welcome
dac.ipynb
+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ···
Python 3.11.3
Python
Cell 10 of 13

```
 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
```

```
df.isnull().sum()
```

[5]

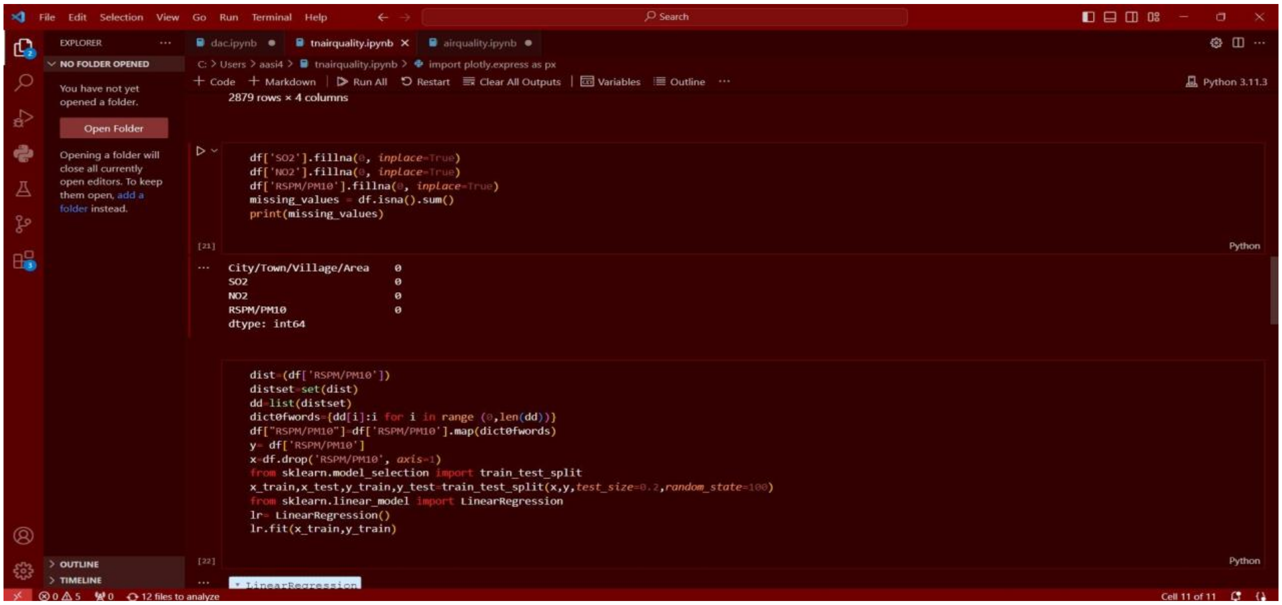
```
Stn Code                           0
Sampling Date                      0
State                              0
City/Town/Village/Area             0
Location of Monitoring Station     0
Agency                             0
Type of Location                   0
SO2                               11
NO2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

```
df.head(6)
```

using more advanced imputation techniques. Here's an example of how to fill missing values with the mean.

Data Cleaning and Transformation:

Depending on your dataset, you may need to perform additional data cleaning and transformation. For example, converting date and time columns to datetime objects, dropping irrelevant columns, or encoding categorical variables.



File Edit Selection View Go Run Terminal Help
Search
EXPLORER
NO FOLDER OPENED
You have not yet opened a folder.
Open Folder
Opening a folder will close all currently open editors. To keep them open, add a folder instead.
OUTLINE
TIMELINE
dac.ipynb
tnairquality.ipynb
airquality.ipynb
C: > Users > aasi4 > tnairquality.ipynb > import plotly.express as px
+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline
Python 3.11.3
Python
12 files to analyze
Cell 11 of 11

2879 rows × 4 columns

```
df['SO2'].fillna(0, inplace=True)
df['NO2'].fillna(0, inplace=True)
df['RSPM/PM10'].fillna(0, inplace=True)
missing_values = df.isna().sum()
print(missing_values)
```

[21]

```
City/Town/Village/Area    0
SO2                       0
NO2                       0
RSPM/PM10                 0
dtype: int64
```

```
dist=(df['RSPM/PM10'])
distset=set(dist)
dd=list(distset)
dict0fwords={dd[i]:i for i in range (0,len(dd))}
df["RSPM/PM10"]=df['RSPM/PM10'].map(dict0fwords)
y= df['RSPM/PM10']
x=df.drop('RSPM/PM10', axis=1)
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=100)
from sklearn.linear_model import LinearRegression
lr= LinearRegression()
lr.fit(x_train,y_train)
```

[22]

```
▾ LinearRegression
```

Save the Preprocessed Dataset:

Once you've loaded, cleaned, and transformed the data, it's a good practice to save the preprocessed dataset for future use. Be sure to replace **"your_dataset.csv"** with the actual file path, and adjust the preprocessing steps to match the specific characteristics of your data. Preprocessing often varies from one dataset to another, so tailor it to your project's requirements.