# DATE - 29/10/2023

# PHASE - III

# TEAM ID - 719

# PROJECT TITLE - AIR QUALITY ANALYSIS IN TAMIL NADU

# IMPORTING MODULES

In [47]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import tkinter as tk
import random
import requests
import scipy
import xgboost
```

In [5]:
```
dataset = pd.read_csv("datafile.csv")
```

In [2]:
```
import os
print("Current working directory:", os.getcwd())

file_path = 'datafile.csv'
if os.path.exists(file_path):
    print("The file exists.")
else:
    print("The file does not exist at the specified path.")
```

```
Current working directory: C:\Users\VIJAYRAJ R
The file exists.
```

# IMPORT THE DATA SET

In [6]:
```
dataset
```

Out[6]:

| | Stn Code | Sampling Date | State | City/Town/Village/Area | Location of Monitoring Station | Agency | Type of Location | SO2 | NO2 |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 38 | 01-02-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 11.0 | 17.( |
| **1** | 38 | 01-07-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 13.0 | 17.( |
| **2** | 38 | 21-01-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 12.0 | 18.( |
| **3** | 38 | 23-01-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 15.0 | 16.( |
| **4** | 38 | 28-01-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 13.0 | 14.( |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | .. |
| **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.( |
| **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.( |
| **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.( |
| **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.( |
| **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.( |

2879 rows × 11 columns

In [15]:
```
dataset.head()
```

Out[15]:

| | Stn Code | Sampling Date | State | City/Town/Village/Area | Location of Monitoring Station | Agency | Type of Location | SO2 | NO2 | RS |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 38 | 01-02-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 11.0 | 17.0 | |
| **1** | 38 | 01-07-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 13.0 | 17.0 | |
| **2** | 38 | 21-01-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 12.0 | 18.0 | |
| **3** | 38 | 23-01-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 15.0 | 16.0 | |
| **4** | 38 | 28-01-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 13.0 | 14.0 | |

In [16]:
```
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
```

In [17]:
```
dataset.describe()
```

Out[17]:

| | 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 |

In [58]:
```
print(data.isna())
```

```
      Stn Code  Sampling Date  State  City/Town/Village/Area  \
0        False          False  False                   False
1        False          False  False                   False
2        False          False  False                   False
3        False          False  False                   False
4        False          False  False                   False
...        ...            ...    ...                     ...
2874     False          False  False                   False
2875     False          False  False                   False
2876     False          False  False                   False
2877     False          False  False                   False
2878     False          False  False                   False

      Location of Monitoring Station  Agency  Type of Location    SO2    NO2  \
0                              False   False             False  False  False
1                              False   False             False  False  False
2                              False   False             False  False  False
3                              False   False             False  False  False
4                              False   False             False  False  False
...                              ...     ...               ...    ...    ...
2874                           False   False             False  False  False
2875                           False   False             False  False  False
2876                           False   False             False  False  False
2877                           False   False             False  False  False
2878                           False   False             False  False  False

      RSPM/PM10  PM 2.5
0         False    True
1         False    True
2         False    True
3         False    True
4         False    True
...         ...     ...
2874      False    True
2875      False    True
2876      False    True
2877      False    True
2878      False    True

[2879 rows x 11 columns]
```

In [59]:
```
print(data.isna().any())
```

```
Stn Code                          False
Sampling Date                     False
State                             False
City/Town/Village/Area            False
Location of Monitoring Station    False
Agency                            False
Type of Location                  False
SO2                                True
NO2                                True
RSPM/PM10                          True
PM 2.5                             True
dtype: bool
```

In [25]:
```
import pandas as pd
dataset = pd.read_csv('datafile.csv')
numeric_dataset = dataset.select_dtypes(include=[np.number])
correlation_matrix = numeric_dataset.corr()

print(correlation_matrix)
```

```
           Stn Code       SO2       NO2  RSPM/PM10  PM 2.5
Stn Code   1.000000  0.263537 -0.043257   0.336190     NaN
SO2        0.263537  1.000000  0.078246   0.445152     NaN
NO2       -0.043257  0.078246  1.000000   0.068277     NaN
RSPM/PM10  0.336190  0.445152  0.068277   1.000000     NaN
PM 2.5          NaN       NaN       NaN        NaN     NaN
```

In [35]:
```
import pandas as pd
data = pd.read_csv('datafile.csv')
data = data.drop(columns=['PM 2.5'])
```

In [36]:
```
data
```

Out[36]:

| | Stn Code | Sampling Date | State | City/Town/Village/Area | Location of Monitoring Station | Agency | Type of Location | SO2 | NO2 |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 38 | 01-02-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 11.0 | 17.0 |
| **1** | 38 | 01-07-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 13.0 | 17.0 |
| **2** | 38 | 21-01-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 12.0 | 18.0 |
| **3** | 38 | 23-01-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 15.0 | 16.0 |
| **4** | 38 | 28-01-14 | Tamil Nadu | Chennai | Kathivakkam, Municipal Kalyana Mandapam, Chennai | Tamilnadu State Pollution Control Board | Industrial Area | 13.0 | 14.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | .. |
| **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 |
| **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 |
| **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 |
| **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 |
| **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 |

2879 rows × 10 columns

In [38]:
```
mean_SO2 = data['SO2'].mean()
mean_NO2 = data['NO2'].mean()
mean_RSPM_PM10 = data['RSPM/PM10'].mean()
```

In [39]:
```
mean_SO2,mean_NO2, mean_RSPM_PM10
```

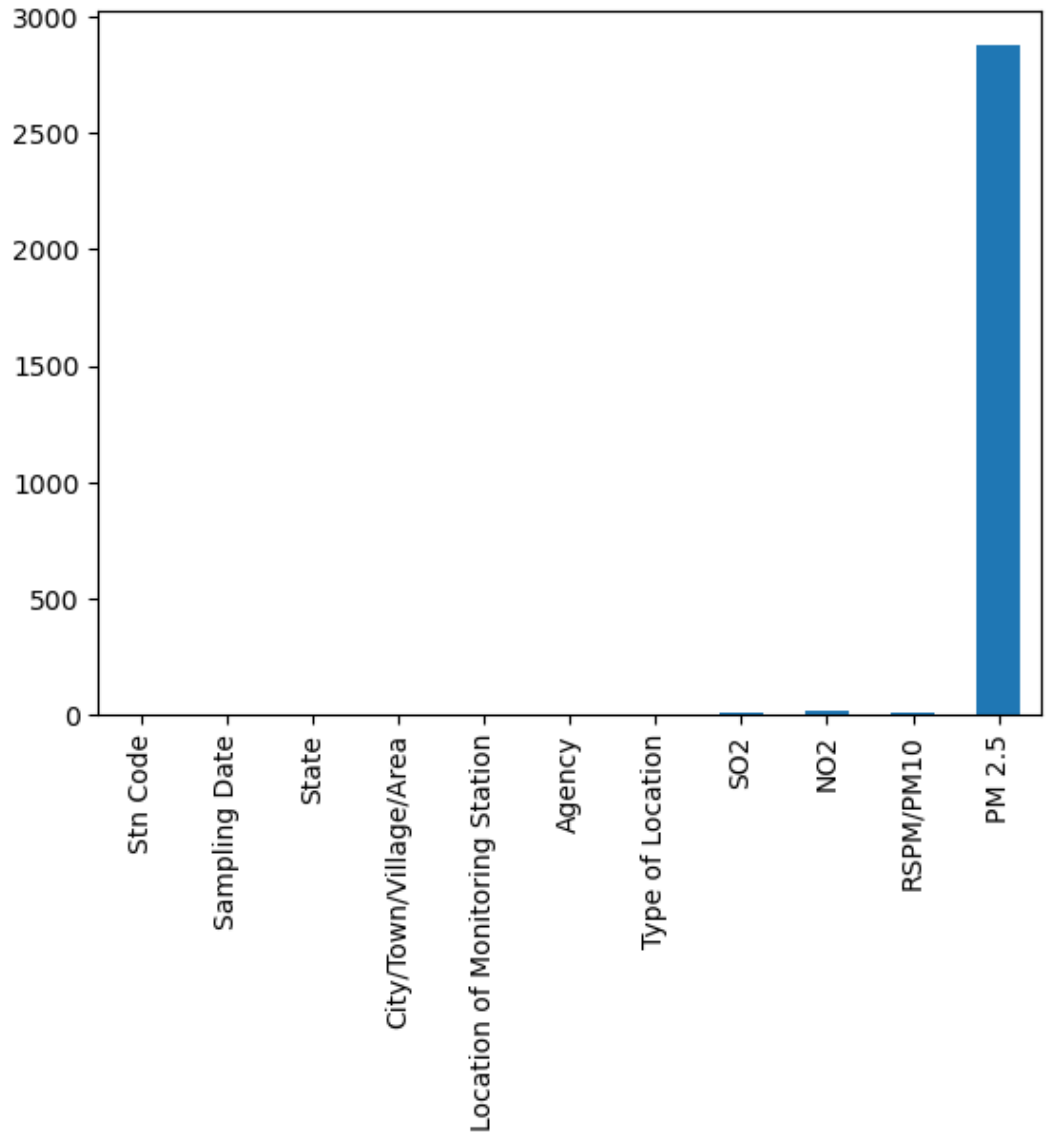
Out[39]:
```
(11.503138075313808, 22.136775994417306, 62.494260869565224)
```

In [40]:
```
data['SO2'].fillna(value=mean_SO2,inplace=True)
data['NO2'].fillna(value=mean_NO2,inplace=True)
data['RSPM/PM10'].fillna(value=mean_RSPM_PM10,inplace=True)
```

# PREPROCSSING DATA

In [33]:
```
dataset.isna().sum().plot(kind='bar')
```

Out[33]:
```
<Axes: >
```

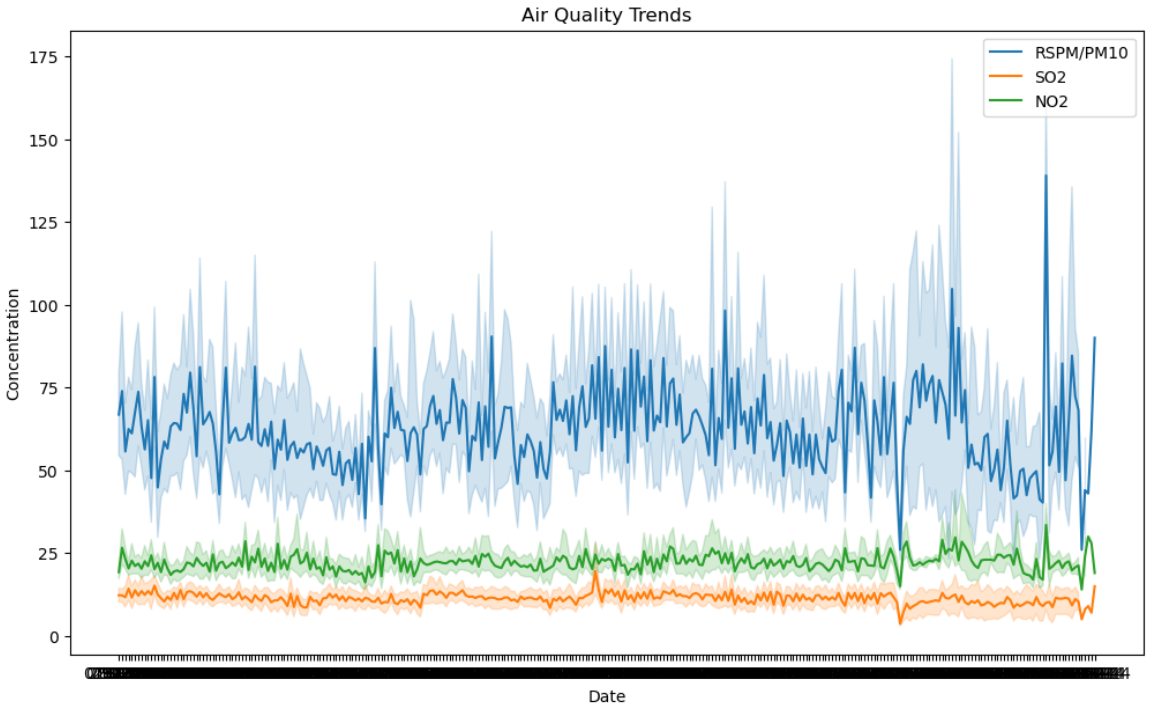



In [41]:
```
cleandata=new_data.isnull().sum()
```

In [42]:
```
cleandata
```

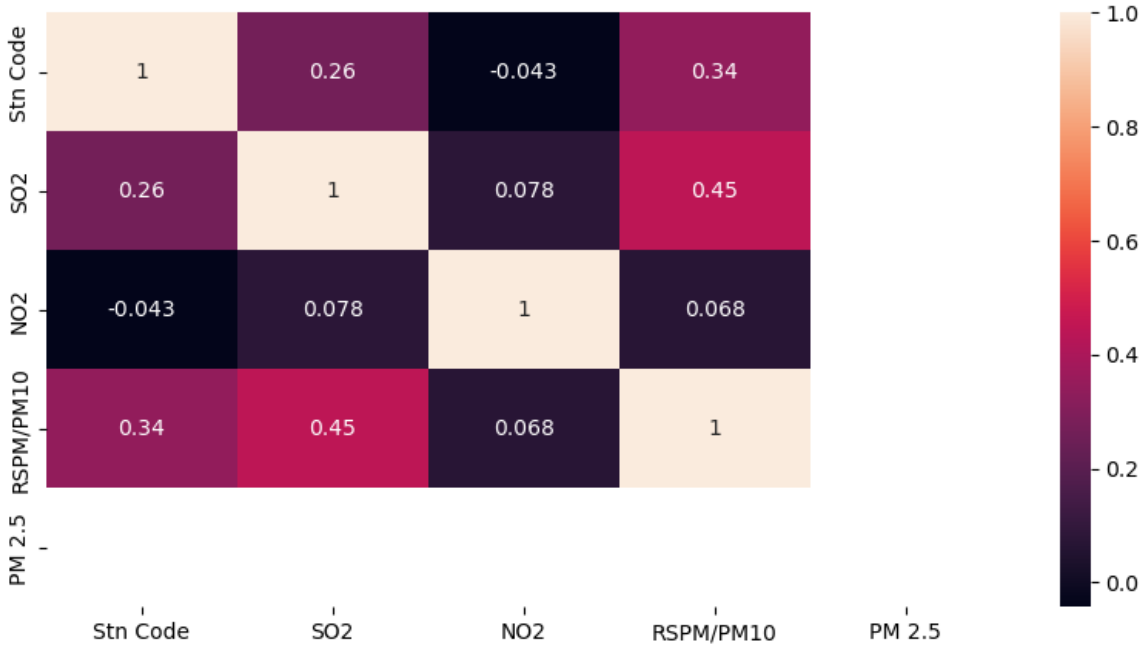
Out[42]:
```
Stn Code                          0
Sampling Date                     0
State                             0
City/Town/Village/Area            0
Location of Monitoring Station    0
Agency                            0
Type of Location                  0
SO2                               0
NO2                               0
RSPM/PM10                         0
dtype: int64
```

In [43]:
```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(12, 7))
sns.lineplot(data=new_data, x='Sampling Date', y='RSPM/PM10', label='RSPM/PM10')
sns.lineplot(data=new_data, x='Sampling Date', y='SO2', label='SO2')
sns.lineplot(data=new_data, x='Sampling Date', y='NO2', label='NO2')
plt.title('Air Quality Trends')
plt.xlabel('Date')
plt.ylabel('Concentration')
plt.legend()
plt.show()
```



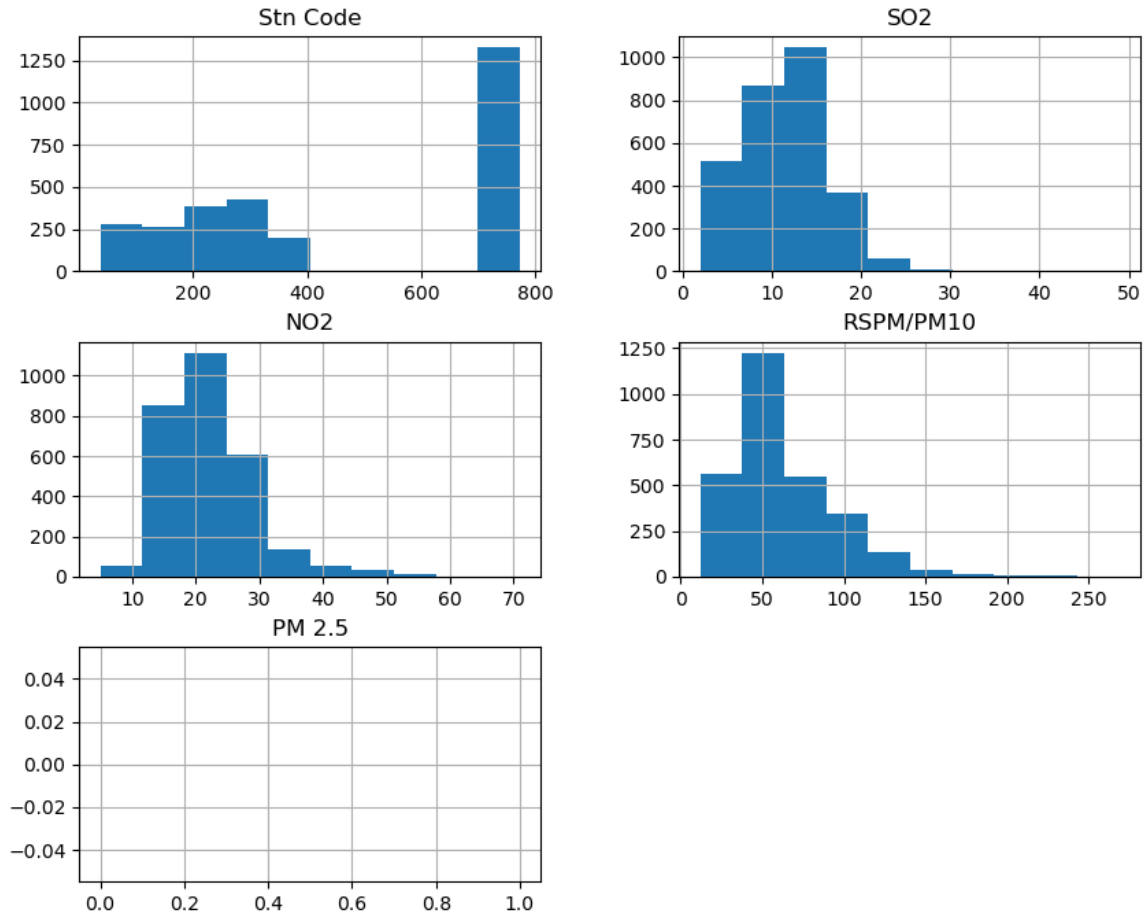
In [28]:
```
plt.figure(figsize=(10,5))
sns.heatmap(dataset.corr(numeric_only = True), annot=True)
```

Out[28]: <Axes: >



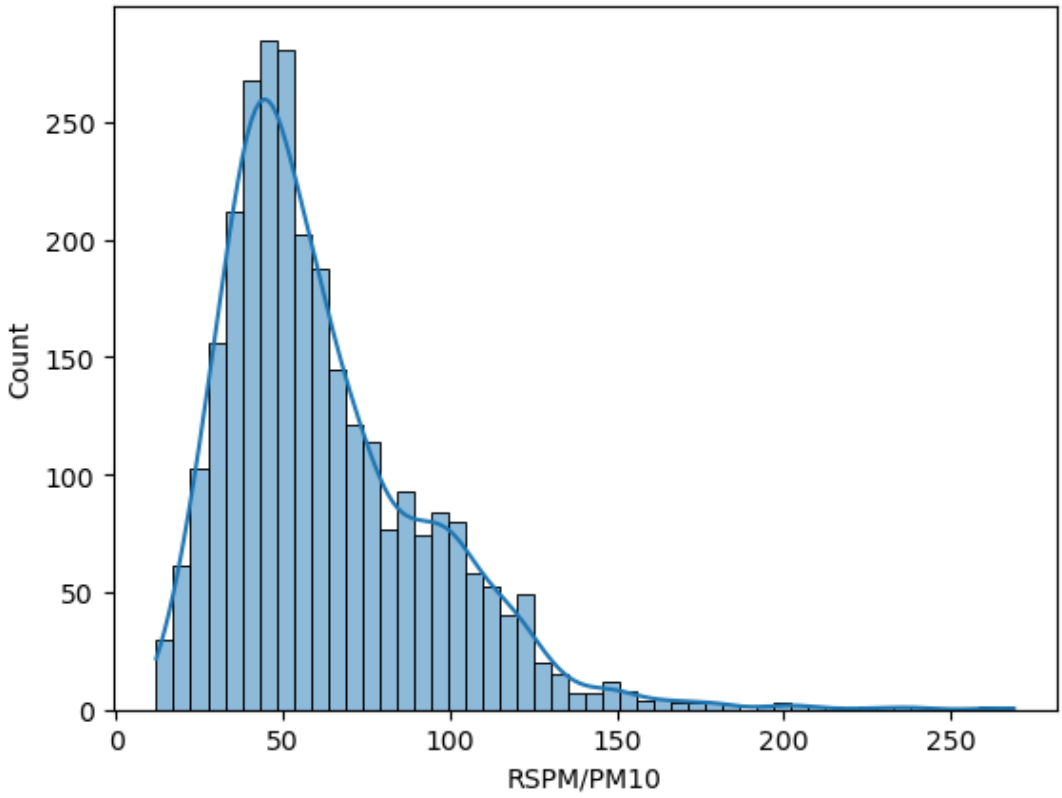

In [29]:
```
dataset.hist(figsize=(10,8))
```

Out[29]:
```
array([[<Axes: title={'center': 'Stn Code'}>,
        <Axes: title={'center': 'SO2'}>],
       [<Axes: title={'center': 'NO2'}>,
        <Axes: title={'center': 'RSPM/PM10'}>],
       [<Axes: title={'center': 'PM 2.5'}>, <Axes: >]], dtype=object)
```



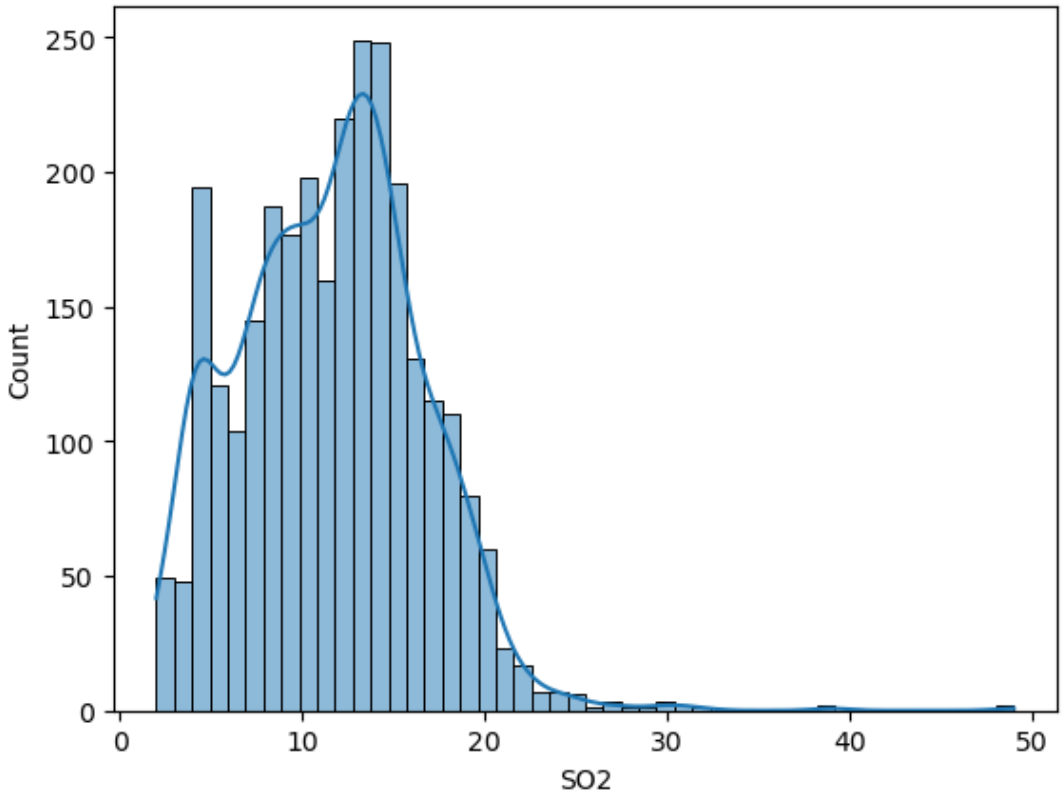
In [30]:
```
import seaborn as sns
sns.histplot(data=dataset, x="RSPM/PM10", kde=True)
```

Out[30]:
```
<Axes: xlabel='RSPM/PM10', ylabel='Count'>
```



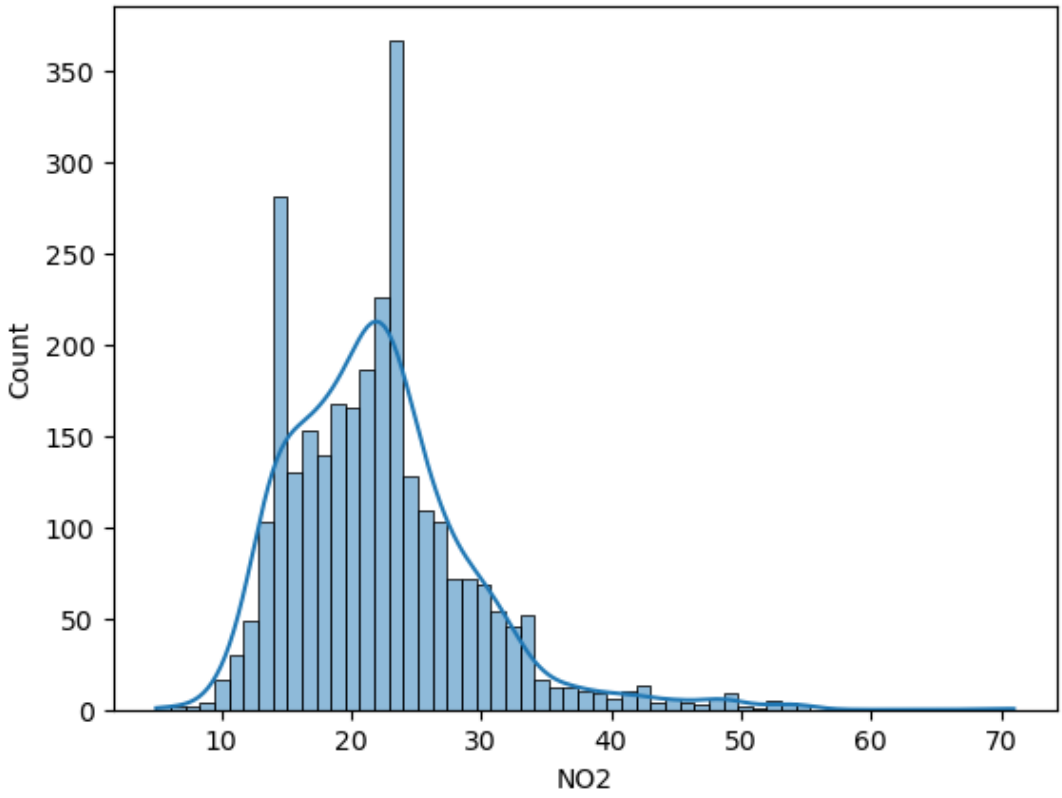

In [32]:
```
import seaborn as sns
sns.histplot(data=dataset, x="SO2", kde=True)
```

Out[32]:
```
<Axes: xlabel='SO2', ylabel='Count'>
```



In [34]:
```
import seaborn as sns
sns.histplot(data=dataset, x="NO2", kde=True)
```

Out[34]:
```
<Axes: xlabel='NO2', ylabel='Count'>
```




In [ ]: