```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

```
# We are reading our data
df = pd.read_csv('/content/India Air Quality Data (2).csv')
```

```
# First 5 rows of our data
df.head(10)
```

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	s
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	N
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	N
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	N
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	N
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	N
5	152.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.4	25.7	NaN	N
6	150.0	April - M041990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	5.4	17.1	NaN	N
7	151.0	April - M041990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	8.7	NaN	N
8	152.0	April - M041990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and	4.2	23.0	NaN	N

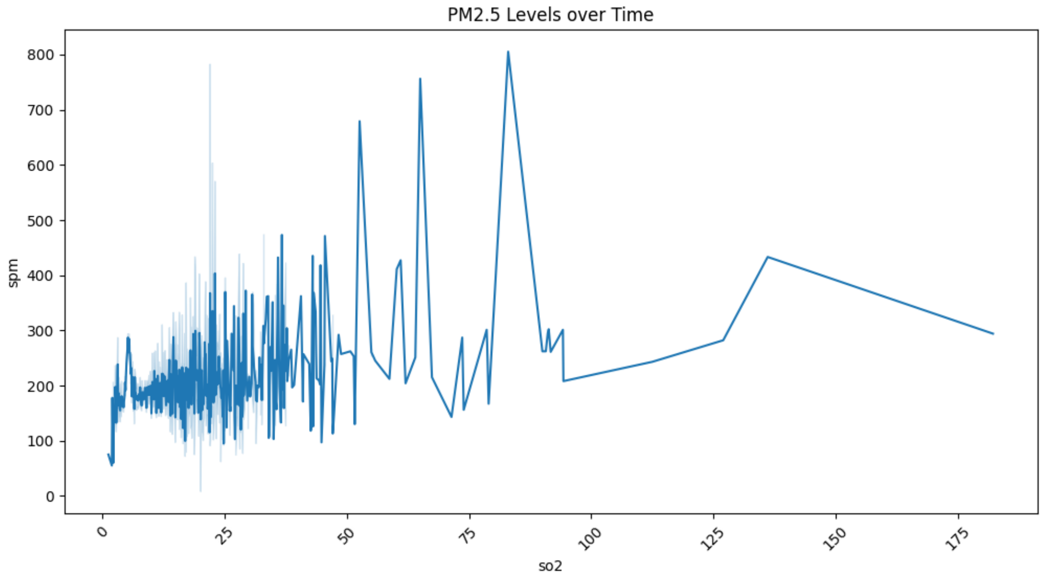
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21513 entries, 0 to 21512
Data columns (total 13 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   stn_code                     12441 non-null  float64
 1   sampling_date                21513 non-null  object
 2   state                        21513 non-null  object
 3   location                     21513 non-null  object
 4   agency                       12015 non-null  object
 5   type                         20902 non-null  object
 6   so2                          20976 non-null  float64
 7   no2                          21112 non-null  float64
 8   rspm                         20372 non-null  float64
 9   spm                          11789 non-null  float64
 10  location_monitoring_station  20476 non-null  object
 11  pm2_5                        0 non-null      float64
 12  date                         21512 non-null  object
dtypes: float64(6), object(7)
memory usage: 2.1+ MB
```

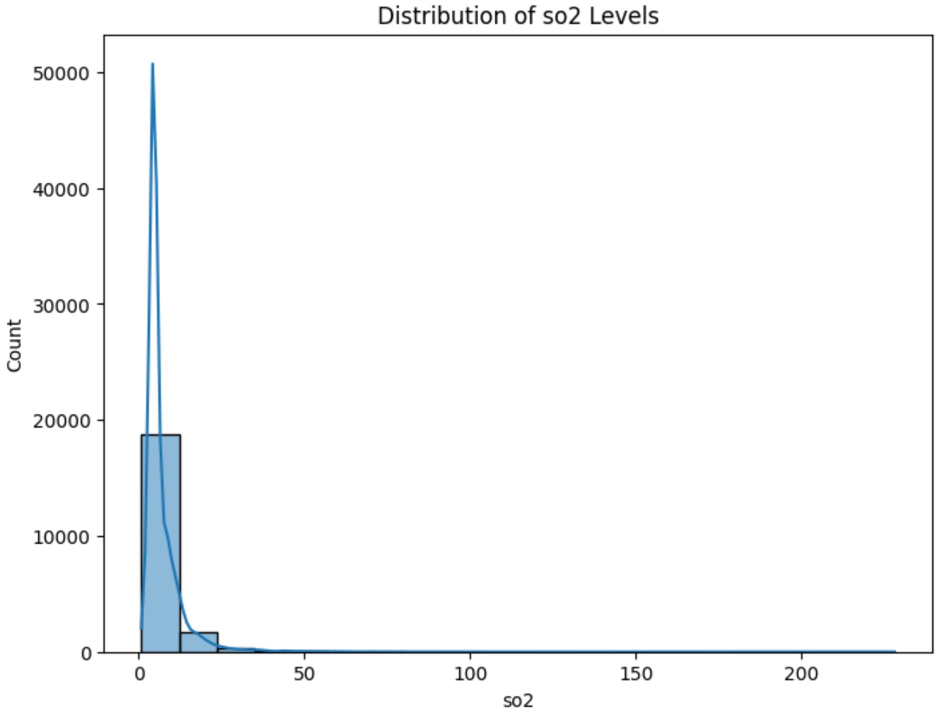
```
df.describe()
```

	stn_code	so2	no2	rspm	spm	pm2_5
count	12441.000000	20976.000000	21112.000000	20372.000000	11789.000000	0.0

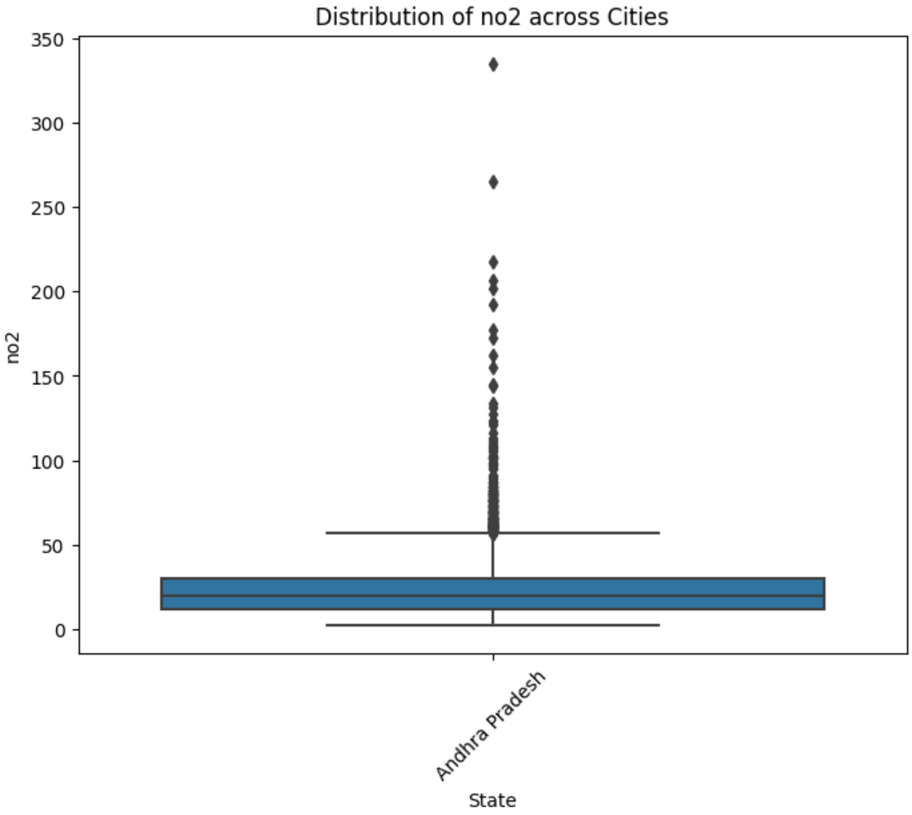
```
plt.figure(figsize=(12, 6))
sns.lineplot(data=df, x='so2', y='spm')
plt.xlabel('so2')
plt.ylabel('spm')
plt.title('PM2.5 Levels over Time')
plt.xticks(rotation=45)
plt.show()
```



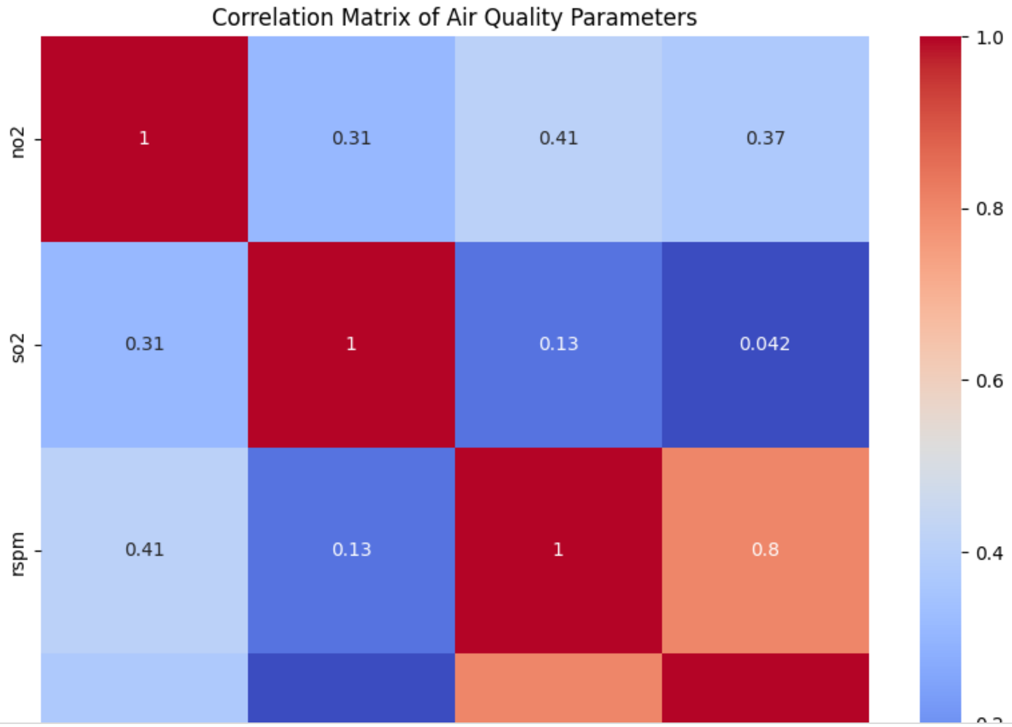
```
plt.figure(figsize=(8, 6))
sns.histplot(data=df, x='so2', bins=20, kde=True)
plt.xlabel('so2')
plt.ylabel('Count')
plt.title('Distribution of so2 Levels')
plt.show()
```



```
plt.figure(figsize=(8, 6))
sns.boxplot(data=df, x='state', y='no2')
plt.xlabel('State')
plt.ylabel('no2')
plt.title('Distribution of no2 across Cities')
plt.xticks(rotation=45)
plt.show()
```



```
plt.figure(figsize=(10, 8))
corr_matrix = df[[ 'no2', 'so2', 'rspm', 'spm']].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix of Air Quality Parameters')
plt.show()
```



✓ 1s completed at 1:30 PM