EE798Q – Assignment 1

INTRODUCTION

Singrauli, a region located in Madhya Pradesh, India, is known for its extensive coal mining activities, including open-pit blasting operations at the Surya Kiran Bhawan Dudhichua site. **What is Open-pit blasting?** Open-pit blasting is a fundamental process employed in extracting coal from the earth's surface. At Dudhichua, this technique involves the controlled use of explosives to break down layers of soil, rock, and other materials covering the coal seams. Through strategic drilling, loading, and initiation of controlled explosions, the overlying layers are fragmented, enabling access to the coal reserves beneath.

What precautions were taken to prevent pollution? The open-pit blasting operations at Surya Kiran Bhawan Dudhichua adhere to stringent safety protocols and environmental regulations to ensure the well-being of workers and minimize any adverse impacts on the environment. Dust suppression measures are implemented to control air pollution and mitigate the release of harmful particulate matter during blasting. Additionally, the management of water resources and waste materials is carefully monitored to prevent contamination and promote responsible disposal practices.

Importance of Singrauli's coal mining industry? Singrauli's coal mining industry, including the open-pit blasting operations at Surya Kiran Bhawan Dudhichua, plays a crucial role in meeting India's energy demands. The coal extracted from this region serves as a valuable resource for power generation, contributing to the nation's energy security. However, it is essential to prioritize sustainable mining practices, considering the region's social, environmental, and economic aspects. By balancing resource extraction with environmental preservation, Singrauli can ensure continued development and progress while safeguarding the local communities and the natural environment.

What are the main pollutants? Particulate matter 10, particulate matter 2.5, Nitrogen Oxides, carbon monoxide, sulphur dioxide, ammonia, ozone, benzene.

DATASET

We have a dataset spanning 90 days of every 15-minute interval, which is around 8640 data values from the Singrauli, Surya Kiran Bhawan Dudhichua site. The dataset columns namely:

1. Singrauli, Surya Kiran Bhawan Dudhichua PM10 (µg/m3)
2. Singrauli, Surya Kiran Bhawan Dudhichua PM2.5 (µg/m3)
3. Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)
4. Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)
5. Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)
6. Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)
7. Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)

8. Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)
9. Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3)
10. Singrauli, Surya Kiran Bhawan Dudhichua Benzene (µg/m3)

COAL INDIA OPEN-PIT BLASTING

The two main air pollutants in NCL coal fields are suspended particulate matter (SPM) and reparable particulate matter (RPM). Air quality monitoring is regularly carried out at both dust-generating and non-generating locations in the vicinity to evaluate the particulate pollution in and around the opencast mining projects of the Singrauli Coalfield. SPM and RPM concentrations predominate at coal working surfaces, coal yards, coal handling facilities, and haul roads used to transport coal, as well as close to drilling sites, in overburden, and on such haul roads. Air pollution measurements available via multi-sensory systems are PM10, PM2.5, SO2, NO2, NOx, CO, NH3, O3, and BENZENE.

Q1. How can we plot the time series?

Ans - To plot a time series, we can use various Python libraries such as matplotlib, pandas, or seaborn.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from math import sqrt
from sklearn.metrics import mean_squared_error
```

Q2. How NA values are interfering with plotting?

Ans - When plotting a time series, NA (missing) values can interfere with the plot in different ways. Here are some scenarios:

1. **Gaps in the Time Series:** The presence of NA values in our time series can lead to breaks or interruptions in the plot, resulting in a disjointed or fragmented appearance with disconnected data points. This can pose difficulties in perceiving the overall pattern or continuity of the time series.
2. **Line Plots:** When utilizing line plots to depict the time series, the presence of NA values can cause disruptions in the lines. If there is an NA value at a specific point, it will result in a gap in the plot, interrupting the line's continuity. This can distort the accurate representation of the data and potentially lead to misleading interpretations of the time series

Q3. Can we just replace NA with 0 values?

Ans - Replacing NA values with 0 in time series analysis is a possible approach, but it should be done with caution and careful consideration of the specific context and requirements of the analysis. Here are some points to consider:

1. **Impact on Analysis:** By substituting NA values with 0 in a time series, the analysis outcomes may be affected as zero values are incorporated into the data. This has the

potential to distort statistical measures like means, variances, and correlations, thereby resulting in biased or misleading interpretations.

2. **Artificial Trends:** When missing observations are replaced with 0 values, there is a risk of introducing fabricated trends or patterns in the time series. This can inaccurately portray the underlying characteristics of the data, potentially leading to erroneous conclusions.
3. **Loss of Information:** By replacing NA values with 0, the information regarding missingness is disregarded. However, it is crucial to recognize that the presence of missing data can hold significance and offer valuable insights into the underlying processes or data generation mechanism. Therefore, it is essential to carefully consider the reasons for missingness and evaluate whether imputing with 0 is appropriate within the context of our specific analysis.

STATISTICAL INFERENCE

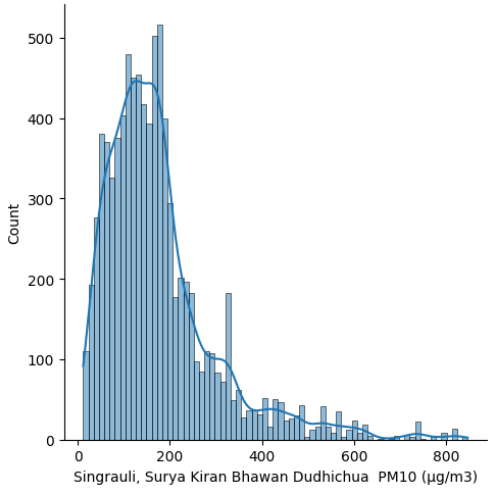
What is Time Series Analysis? Time series analysis is a statistical technique used to analyze and interpret data points collected over time. It focuses on understanding the patterns, trends, and relationships within the data, as well as making predictions or forecasting future values.

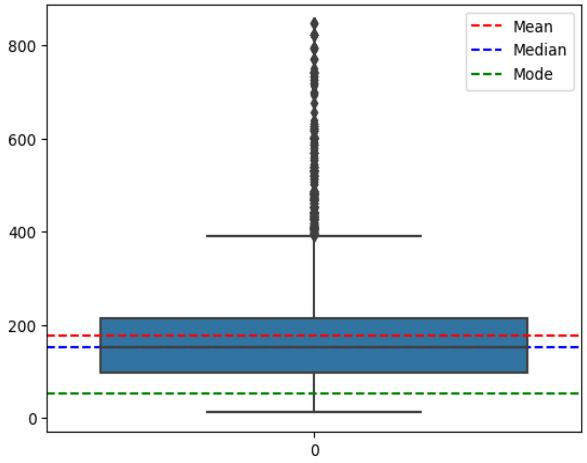
What is Statistical Inference? Statistical inference refers to the process of drawing conclusions or making decisions about a population based on sample data. It involves using statistical techniques to analyze the sample data and make inferences or generalizations about the larger population from which the sample was drawn.

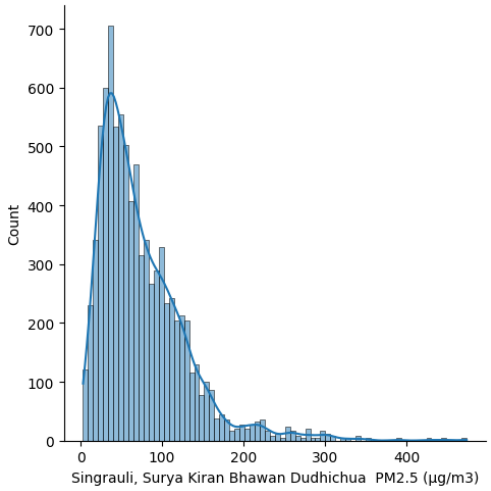
Q1. Can you plot the histogram of this blast trigger times across all months of data. What kind of distribution it is following? Can you infer from QQ plot whether is Normal distribution or not?

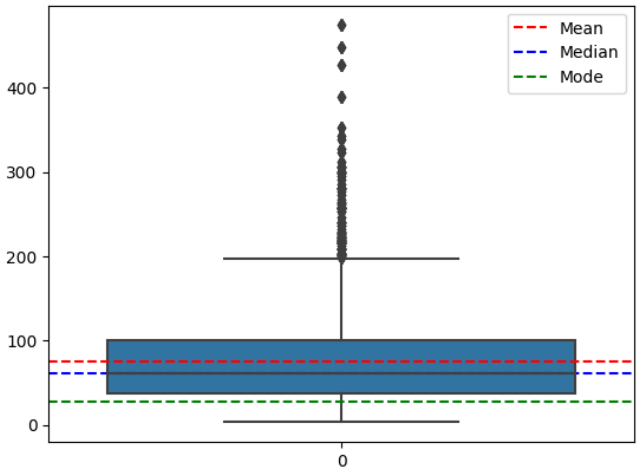
Ans –

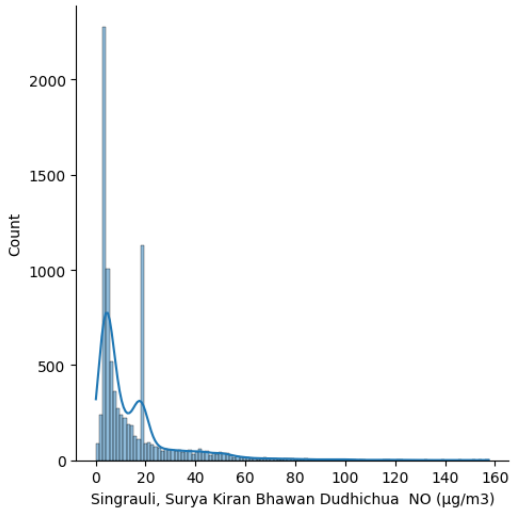
```
import statistics
for column in columns:
    sns.displot(dataSet[column], kde = True)
    plt.show()
    mean = np.mean(dataSet[column])
    median = np.median(dataSet[column])
    mode = float(statistics.mode(dataSet[column]))
    print("Mean = %f" %mean)
    print("Median = %f" %median)
    print("Mode = %f" %mode)
    # Create a boxplot with mean, median, and mode annotations
    sns.boxplot(data=dataSet[column])
    plt.axhline(mean, color='red', linestyle='--', label='Mean')
    plt.axhline(median, color='blue', linestyle='--', label='Median')
    plt.axhline(mode, color='green', linestyle='--', label='Mode')
    plt.legend()
    plt.show()
```

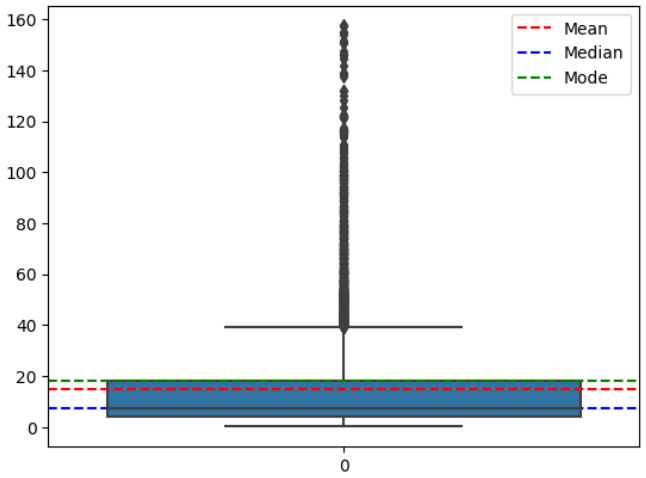

Count
Singrauli, Surya Kiran Bhawan Dudhichua PM10 (µg/m3)

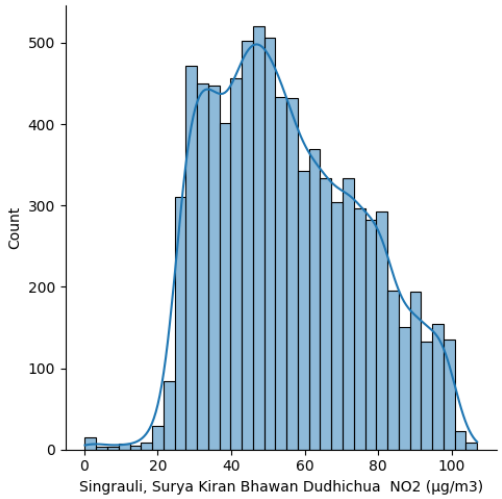

Mean
Median
Mode
0

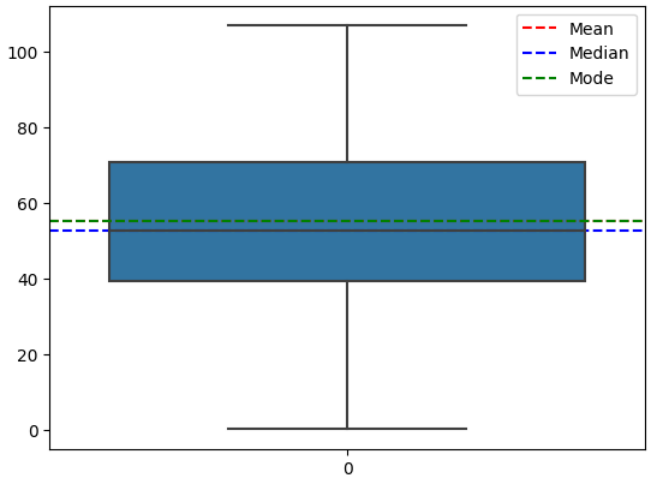

Count
Singrauli, Surya Kiran Bhawan Dudhichua PM2.5 (µg/m3)

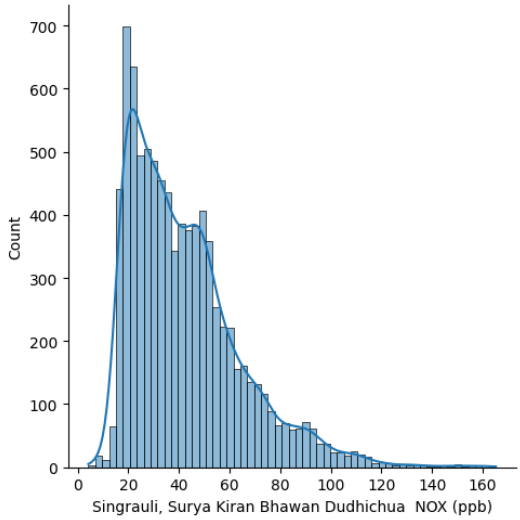

Mean
Median
Mode
0

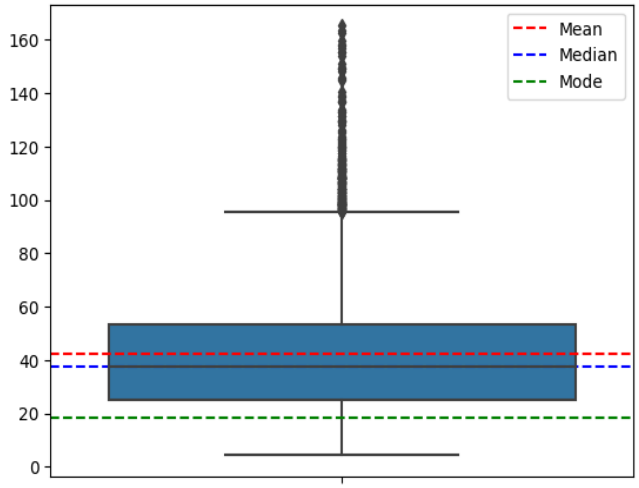

Count
Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)

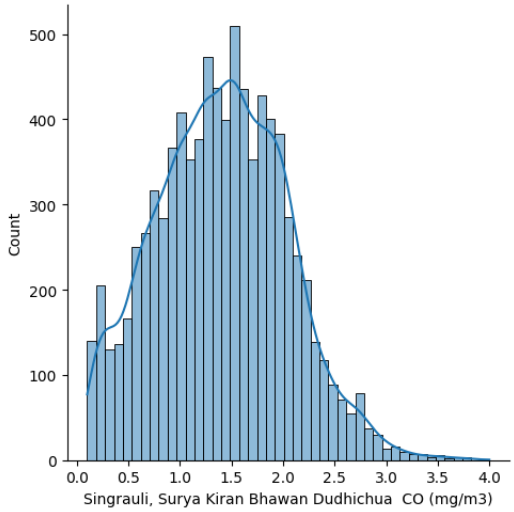

Mean
Median
Mode
0

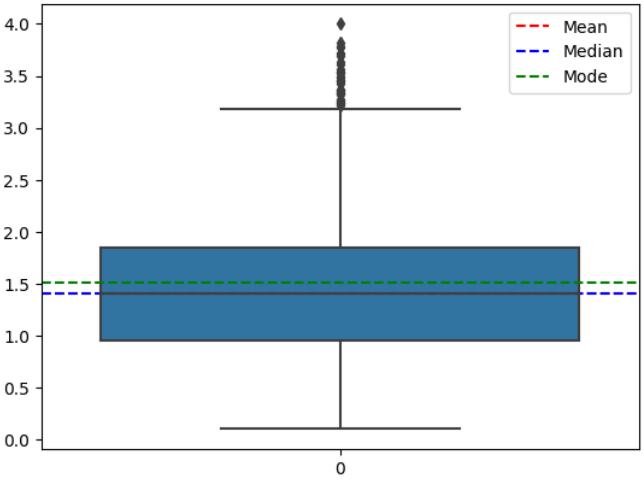

Count
Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)

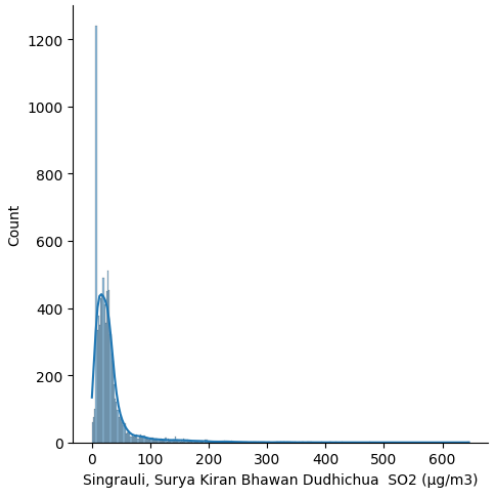

Mean
Median
Mode
0

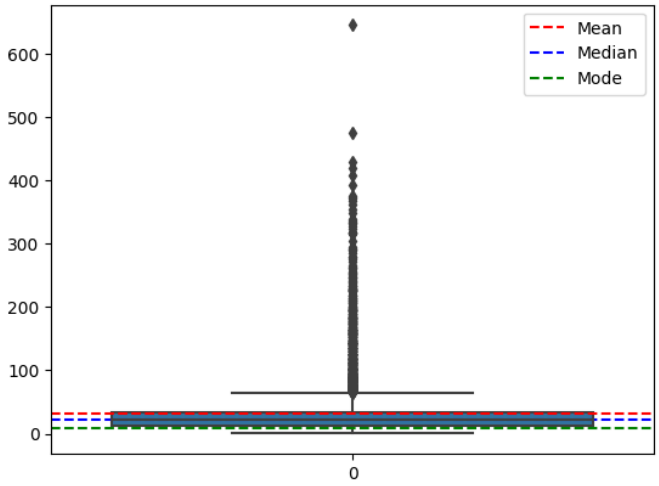

Count
Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)

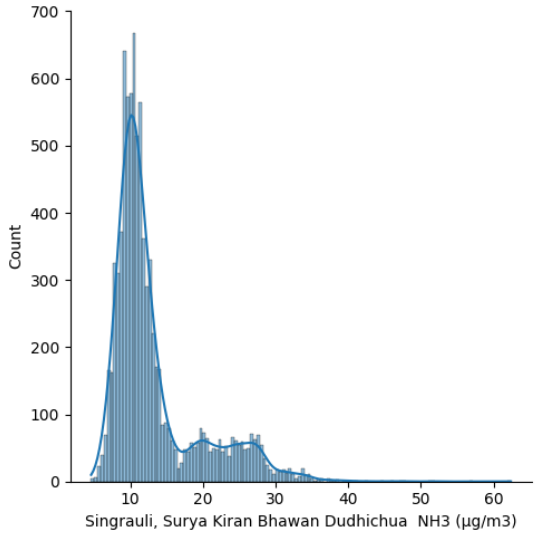

Mean
Median
Mode
0

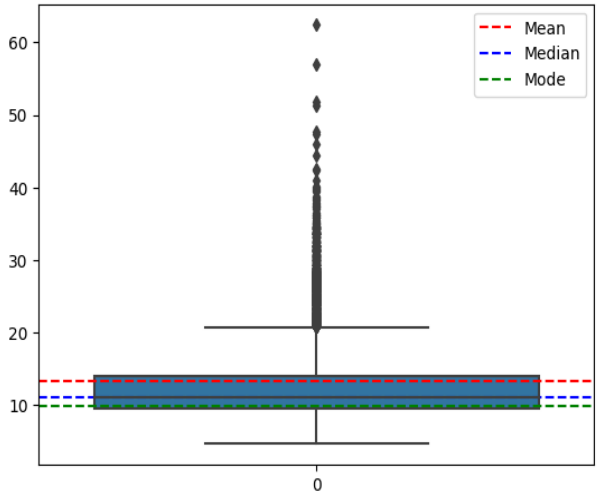

Count
Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)

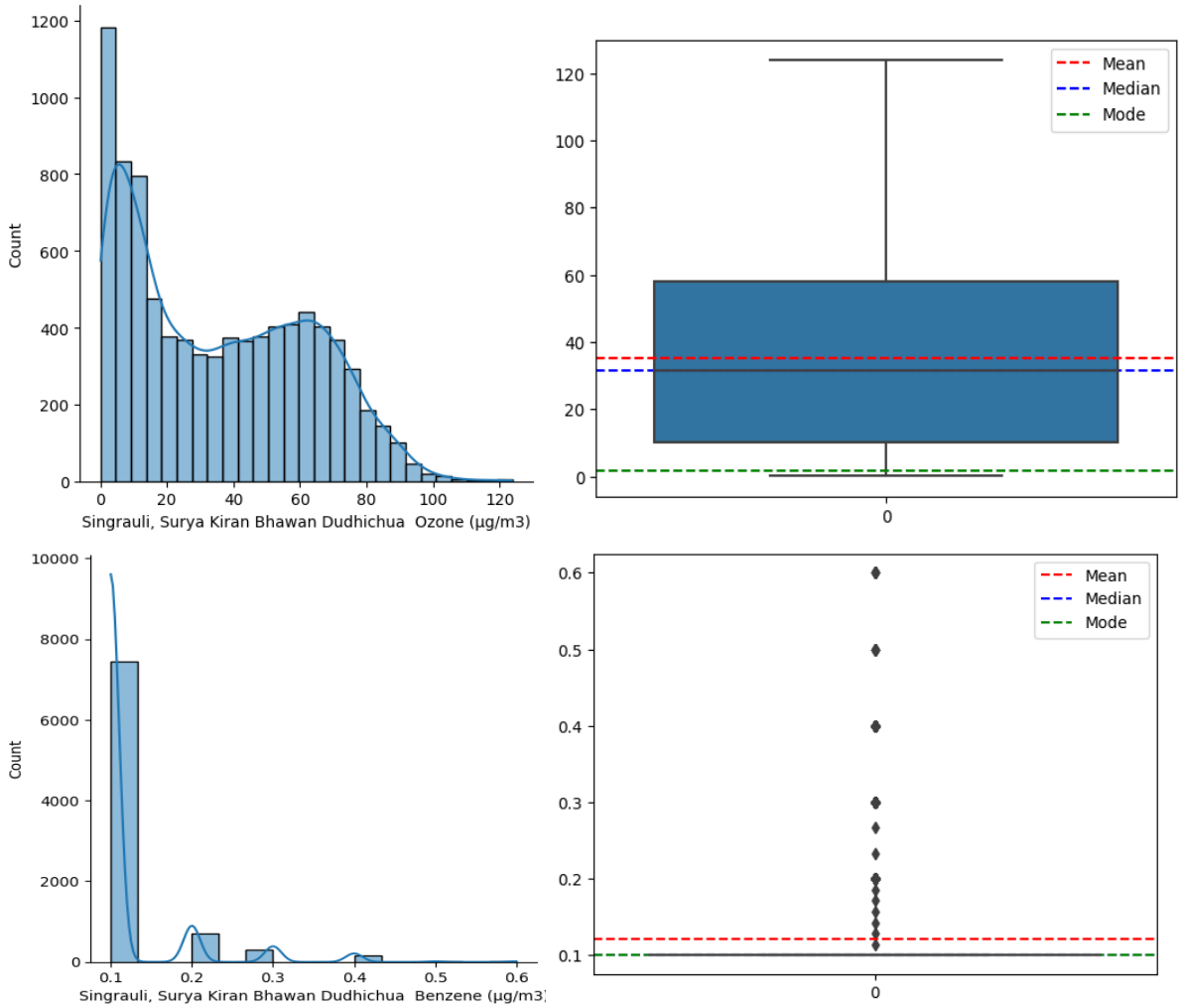

Mean
Median
Mode
0


Count
Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)

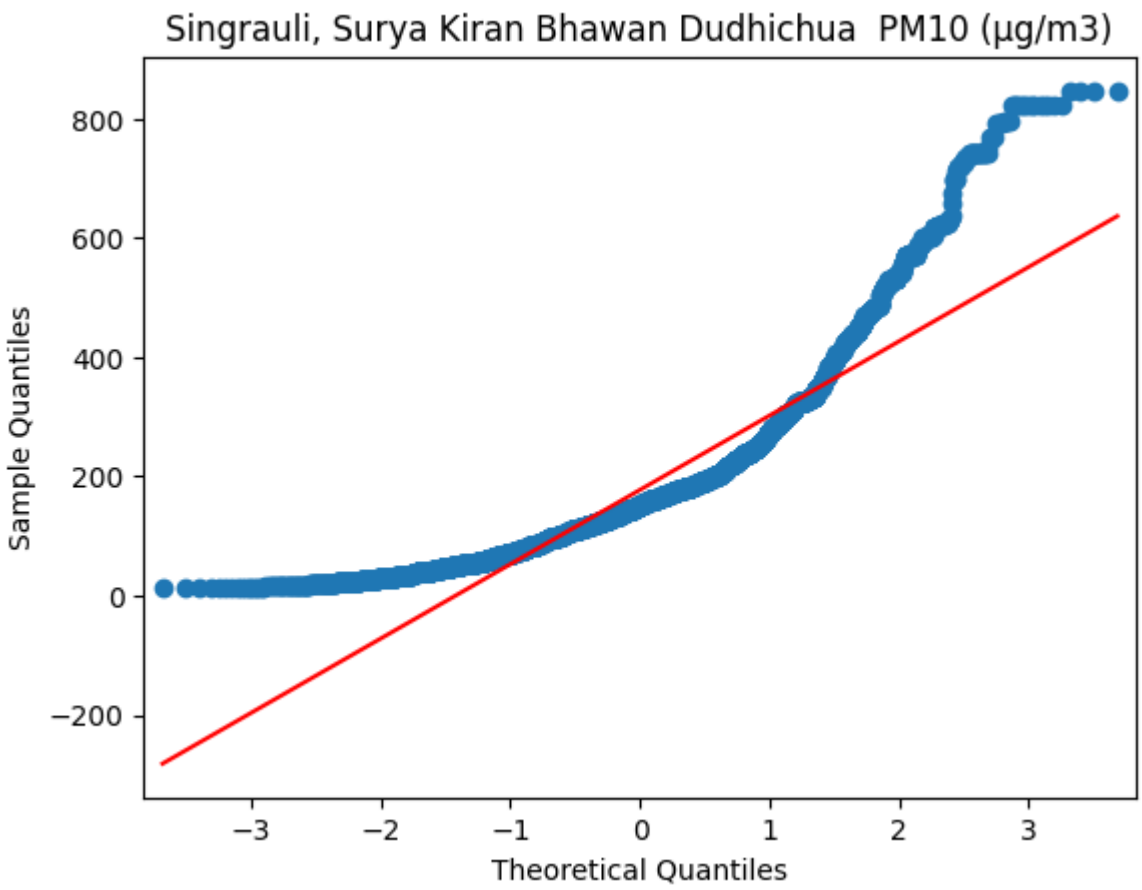

Mean
Median
Mode
0

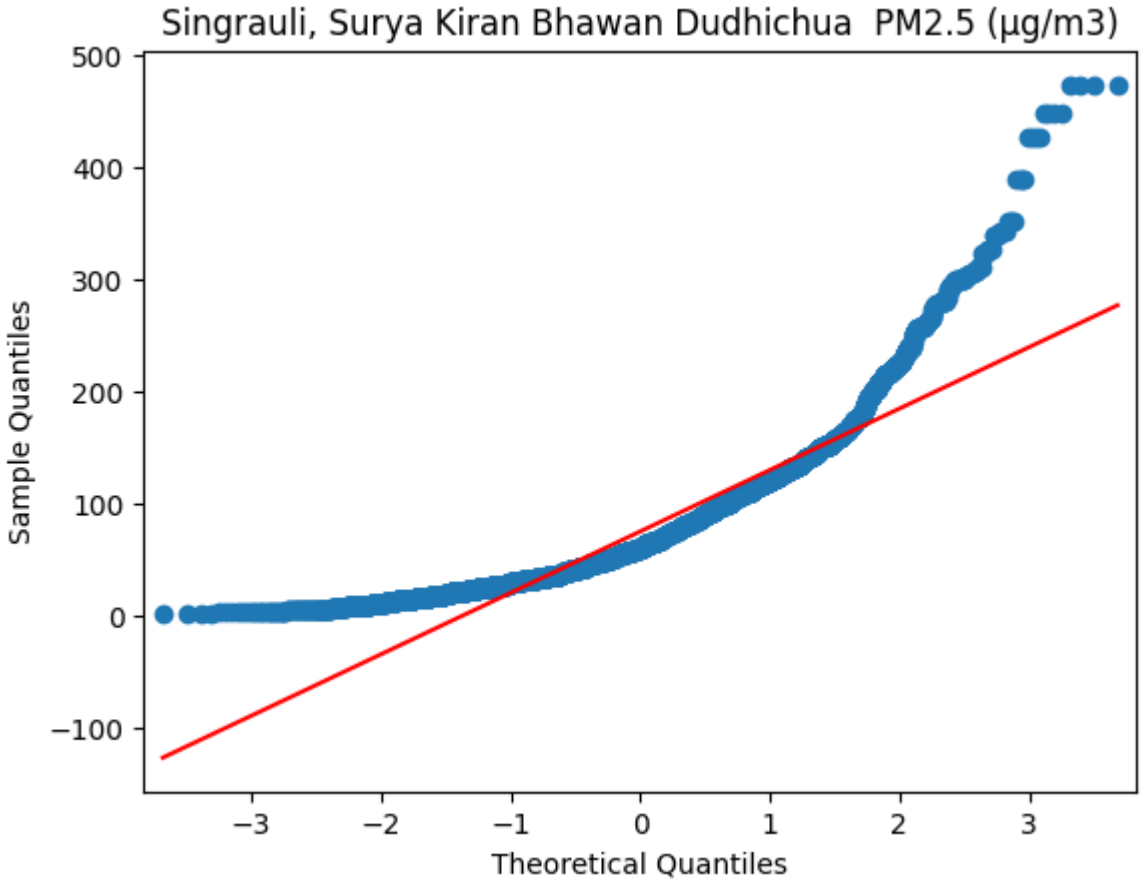

Count
Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)

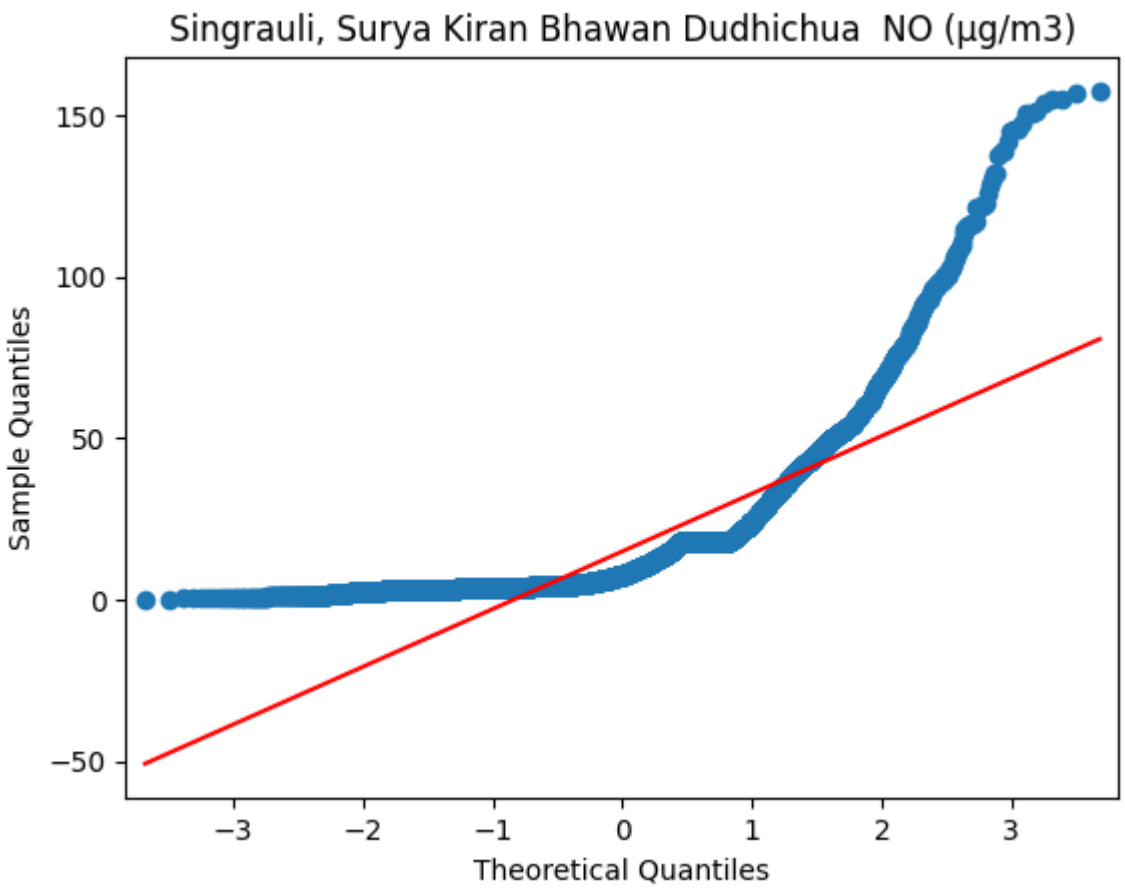

Mean
Median
Mode
0

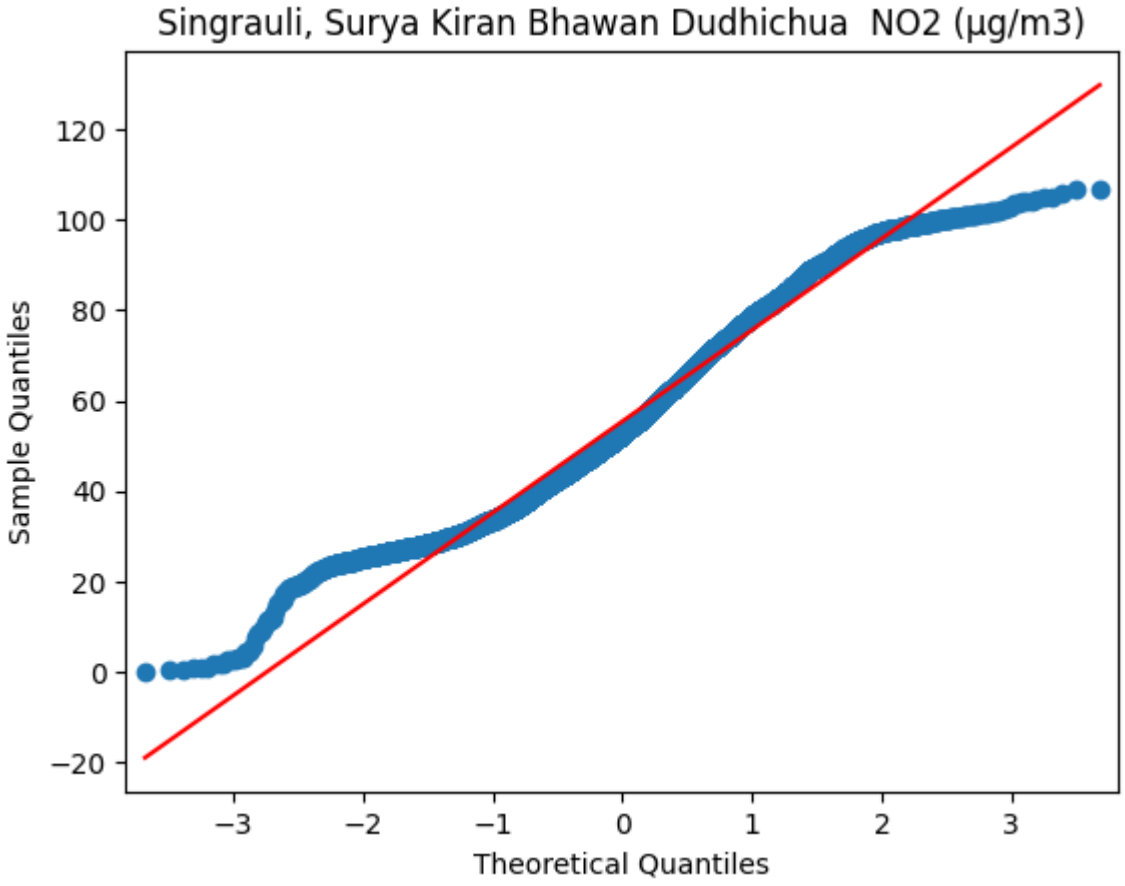


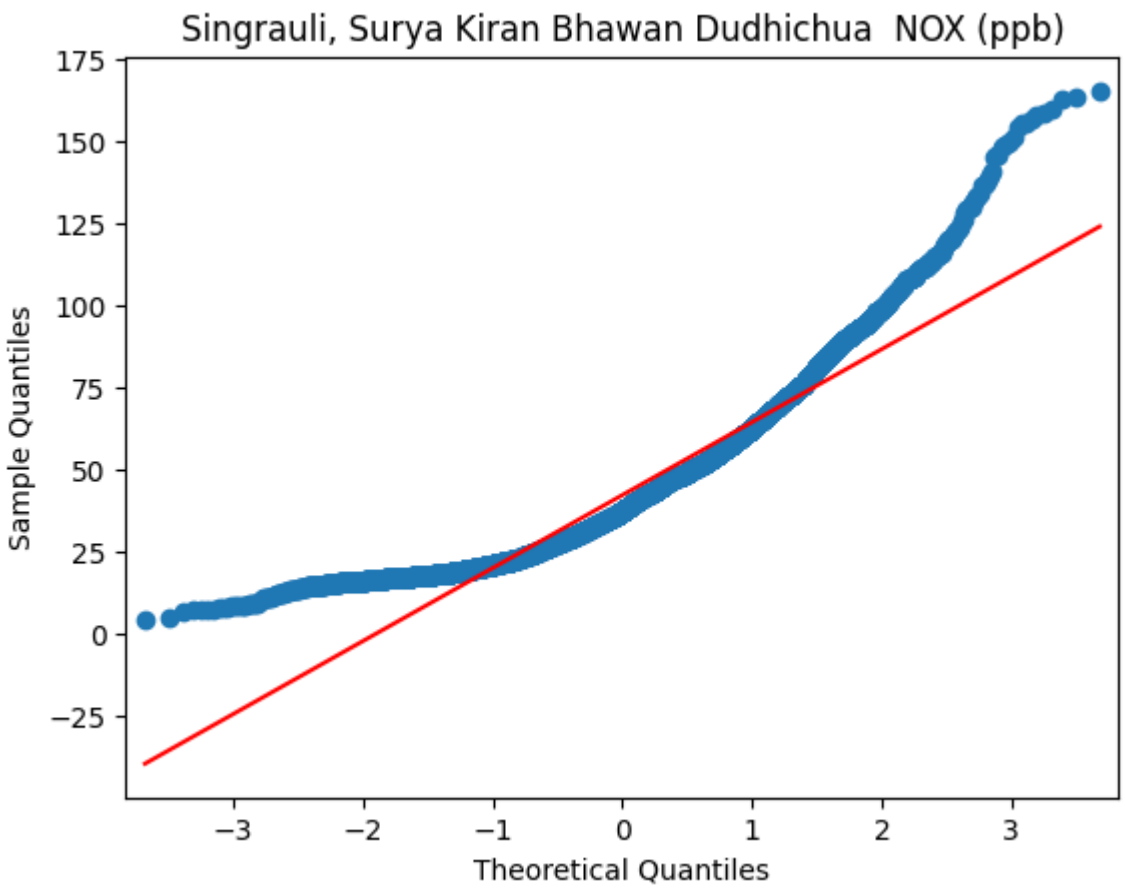
By examining the aforementioned histograms and box plots, it becomes evident that none of the graphs display a normal distribution. However, certain graphs, such as the ones depicting SO2 and CO, do exhibit a normal distribution pattern. The majority of the graphs demonstrate positive skewness, which is also evident from the Q-Q plots.

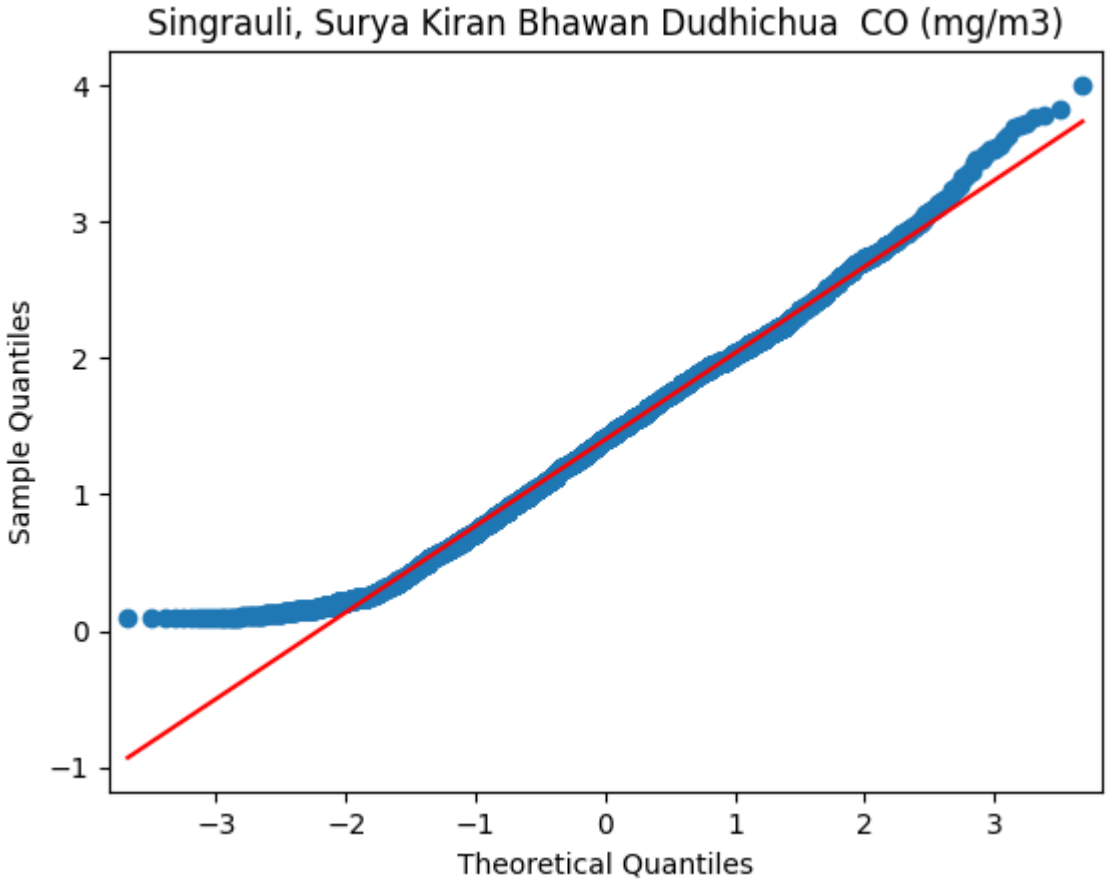

Singrauli, Surya Kiran Bhawan Dudhichua PM10 (µg/m3)
Sample Quantiles
Theoretical Quantiles
800
600
400
200
0
−200
−3
−2
−1
0
1
2
3

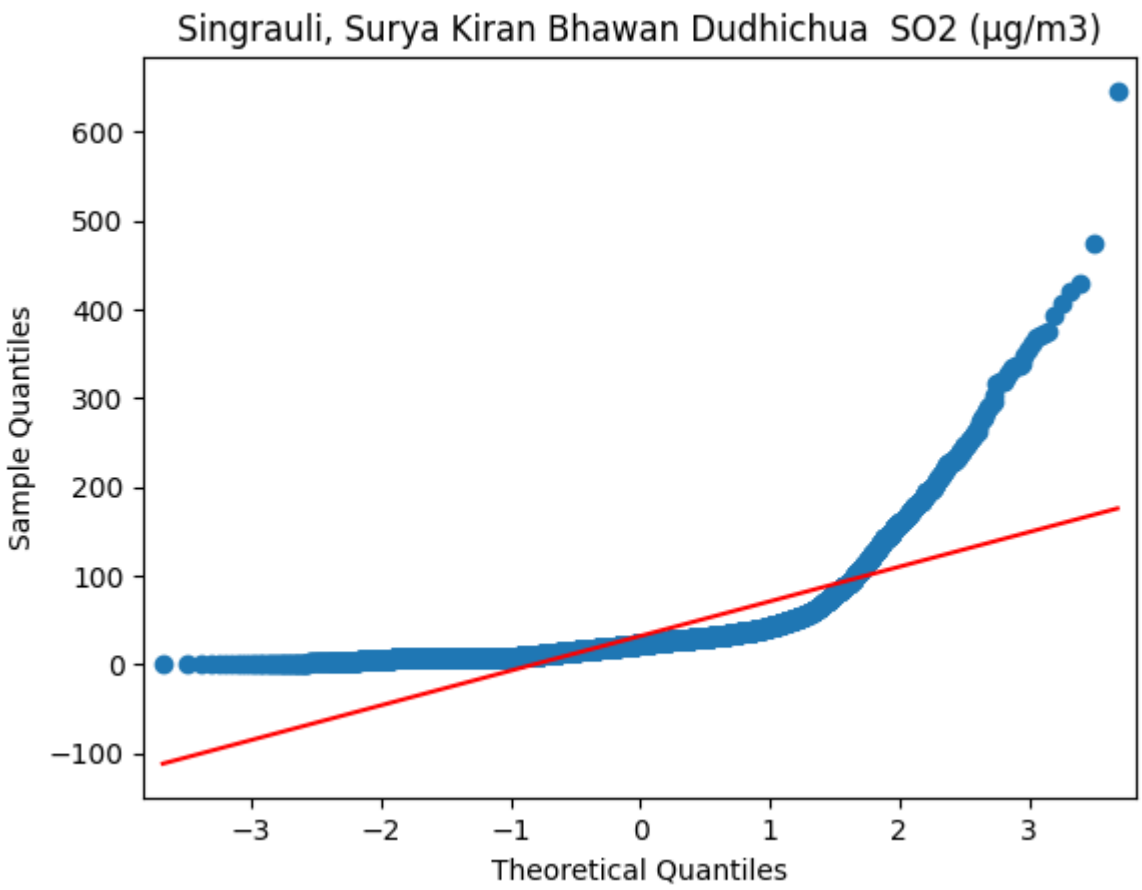

Singrauli, Surya Kiran Bhawan Dudhichua PM2.5 (µg/m3)
Sample Quantiles
Theoretical Quantiles
500
400
300
200
100
0
−100
−3
−2
−1
0
1
2
3

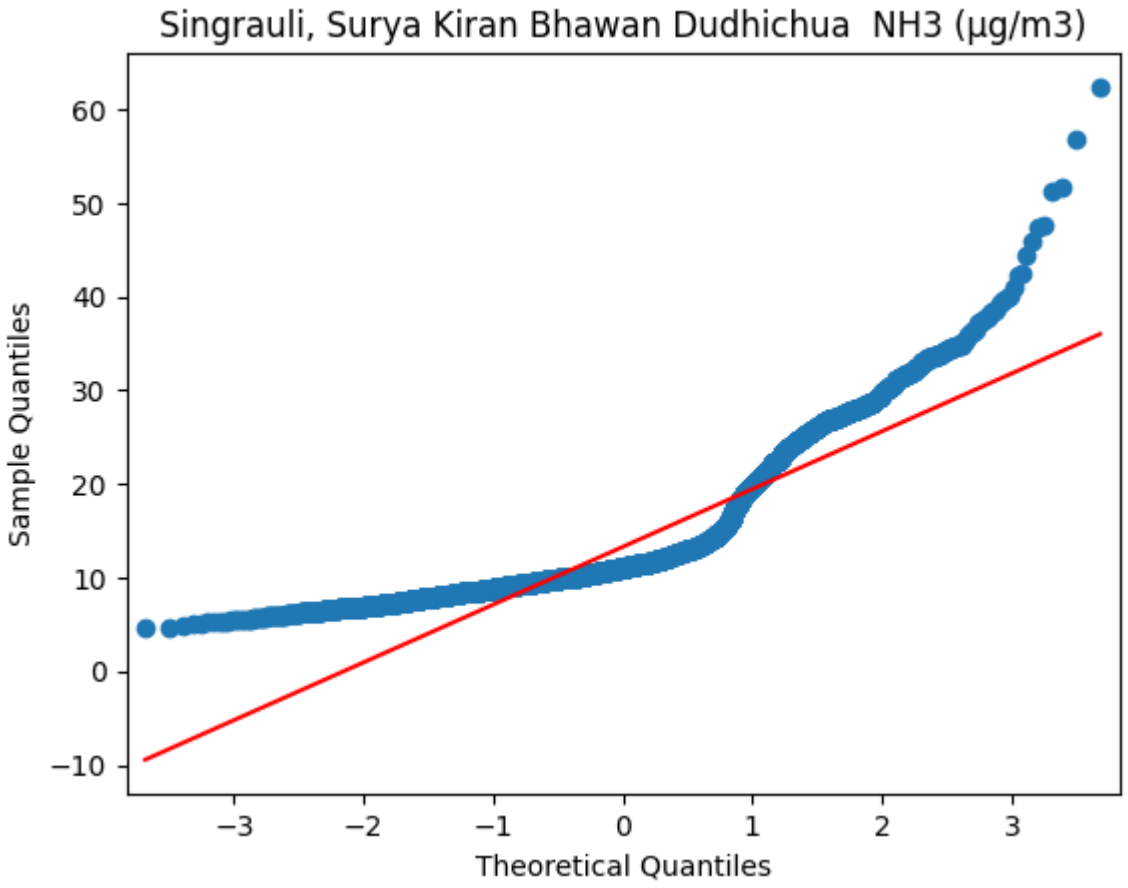

Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)
Sample Quantiles
Theoretical Quantiles
150
100
50
0
−50
−3
−2
−1
0
1
2
3


Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)
Sample Quantiles
Theoretical Quantiles
120
100
80
60
40
20
0
−20
−3
−2
−1
0
1
2
3

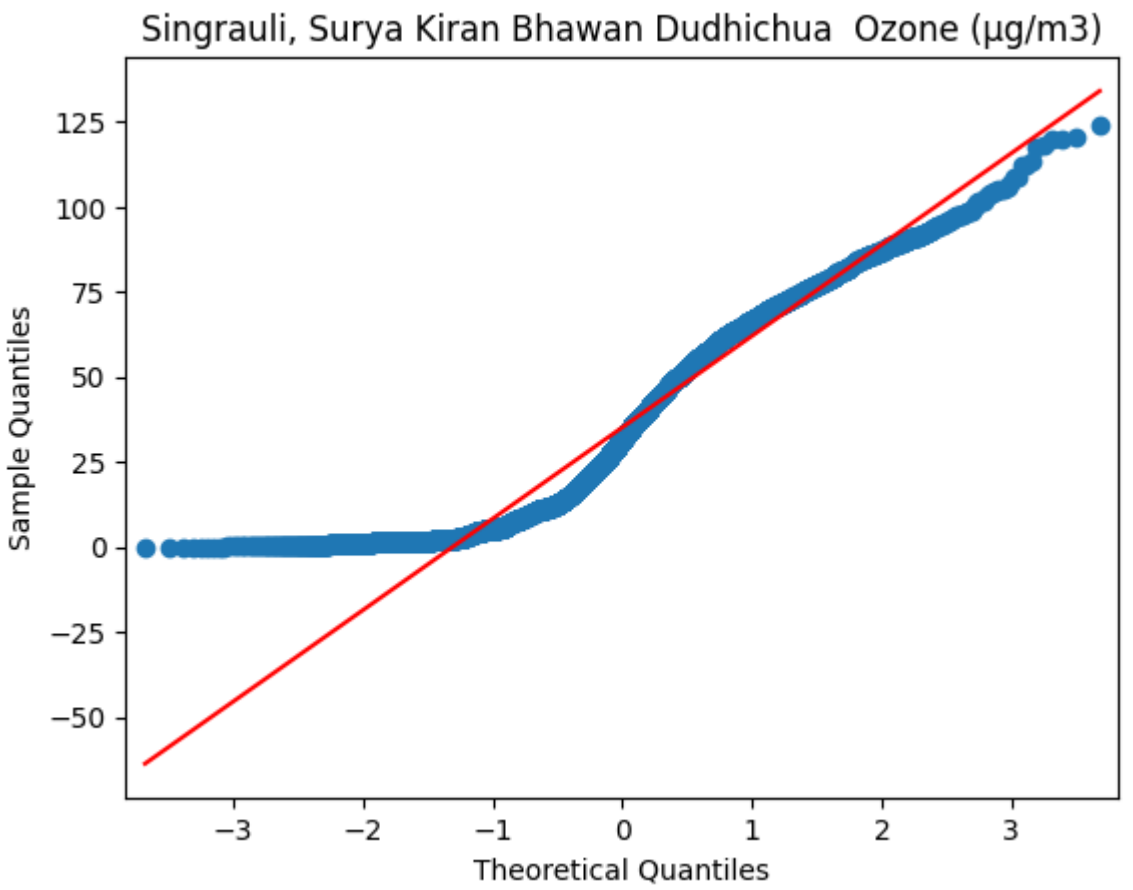

Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)
Sample Quantiles
Theoretical Quantiles
175
150
125
100
75
50
25
0
−25
−3
−2
−1
0
1
2
3

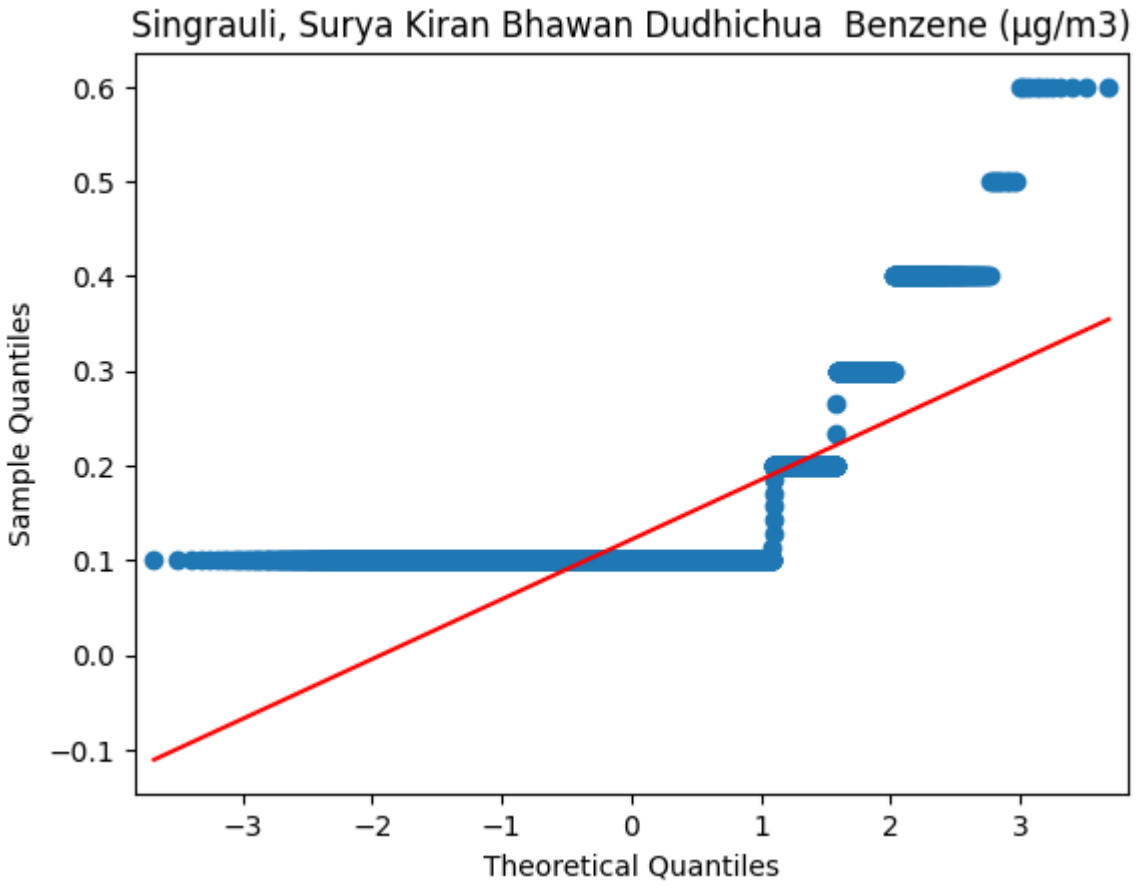

Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)
Sample Quantiles
Theoretical Quantiles
4
3
2
1
0
−1
−3
−2
−1
0
1
2
3


Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)
Sample Quantiles
Theoretical Quantiles
600
500
400
300
200
100
0
−100
−3
−2
−1
0
1
2
3


Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)
Sample Quantiles
Theoretical Quantiles
60
50
40
30
20
10
0
−10
−3
−2
−1
0
1
2
3

v


Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3)
Sample Quantiles
Theoretical Quantiles
125
100
75
50
25
0
−25
−50
−3
−2
−1
0
1
2
3


Singrauli, Surya Kiran Bhawan Dudhichua Benzene (µg/m3)
Sample Quantiles
Theoretical Quantiles
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−3
−2
−1
0
1
2
3

PROBLEM SETTING AND PREDICTION

Q1. What is stock time series data?

Ans - Stock time series data refers to a type of time series data that represents the historical prices and other relevant information about a particular stock or financial instrument. It captures the price movements and trading activity of a stock over a specific period, typically recorded at regular intervals such as daily, weekly, or monthly.

Stock time series data typically includes the following information:

- **Date or Time:** The specific date or time stamp when the price or trading activity was recorded.
- **Open Price:** The price at which the stock started trading at the beginning of the time interval.
- **High Price:** The highest price reached by the stock during the time interval.
- **Low Price:** The lowest price reached by the stock during the time interval.
- **Close Price:** The price at which the stock ended trading at the end of the time interval.
- **Volume:** The number of shares or contracts traded during the time interval.
- **Adjusted Close:** The closing price is adjusted for factors such as stock splits, dividends, or other corporate actions.

Stock time series data is widely used in financial analysis, investment strategies, and risk management. It allows analysts and investors to analyze historical price trends, identify patterns, and make informed decisions regarding buying or selling stocks. Various statistical and mathematical techniques can be applied to stock time series data, including trend analysis, volatility modeling, and forecasting methods such as ARIMA models.

Q2. What is Flow time series data?

Ans - Flow time series data refers to a type of time series data that captures the measurement or observation of a variable over a continuous period, typically at regular intervals. Unlike stock time series data that represents characteristics at a specific moment, flow time series data focuses on measuring the activity or flow of the variable over time.

Flow time series data often represents the cumulative or aggregated values of a variable within a specified time interval. It is commonly used to track the flow of quantities or events that accumulate or change over time.

Examples of flow time series data include:

- **Sales Volume:** Tracking the daily, weekly, or monthly sales volume of a product or service.
- **Website Traffic:** Monitoring the number of visitors or page views on a website over time.
- **Electricity Consumption:** Measuring the energy consumption of a household or building over regular intervals.

- **Population Growth:** Tracking the increase or decrease in population size over time.
- **Water Flow Rate:** Monitoring the rate of water flow in a river or pipeline over specific time intervals.

Flow time series data is useful for analyzing trends, seasonality, and long-term patterns in the flow of a variable. It can be visualized through line charts, bar graphs, or area plots to understand the overall flow pattern and detect any anomalies or changes over time. Analytical techniques such as time series decomposition, forecasting, and trend analysis are commonly applied to flow time series data to gain insights and make informed decisions.

Q3. To which category of time series data does this dataset belong?

Ans - Air pollution data is typically considered a type of flow time series data. It represents the activity or variation of pollutant concentrations over a specific period, such as hourly, daily, monthly, or yearly measurements. The data reflects the changes in pollutant levels and provides information about the temporal patterns, trends, and fluctuations in air quality.

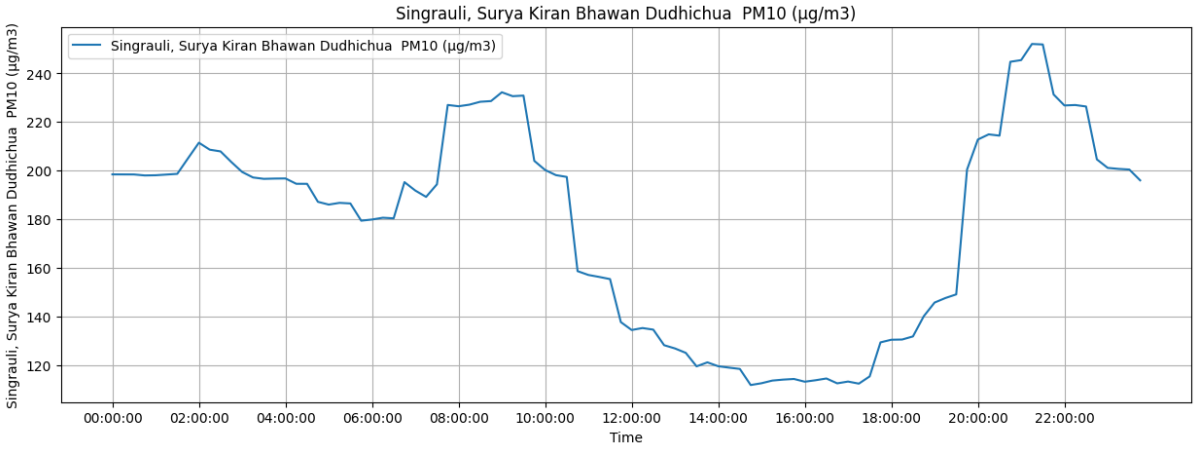
To classify air pollution data, we can consider different attributes, such as pollutant concentrations (e.g., PM10, PM2.5, SO2, NO2), meteorological factors (e.g., temperature, humidity, wind speed), and temporal features (e.g., time of day, day of the week, season). Various statistical and machine learning techniques can be applied to analyze and classify air pollution data, including clustering, time series analysis, and supervised or unsupervised classification algorithms.

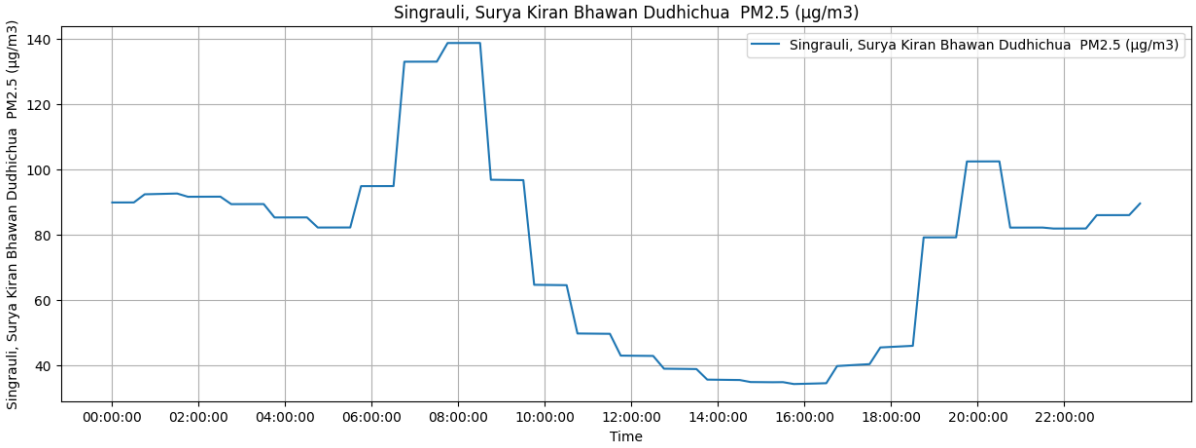
It's worth noting that air pollution data classification can serve different purposes, such as identifying pollution events, assessing the severity of air quality, identifying pollution sources, or predicting future pollution levels. The specific classification approach would depend on the objectives and the available data.

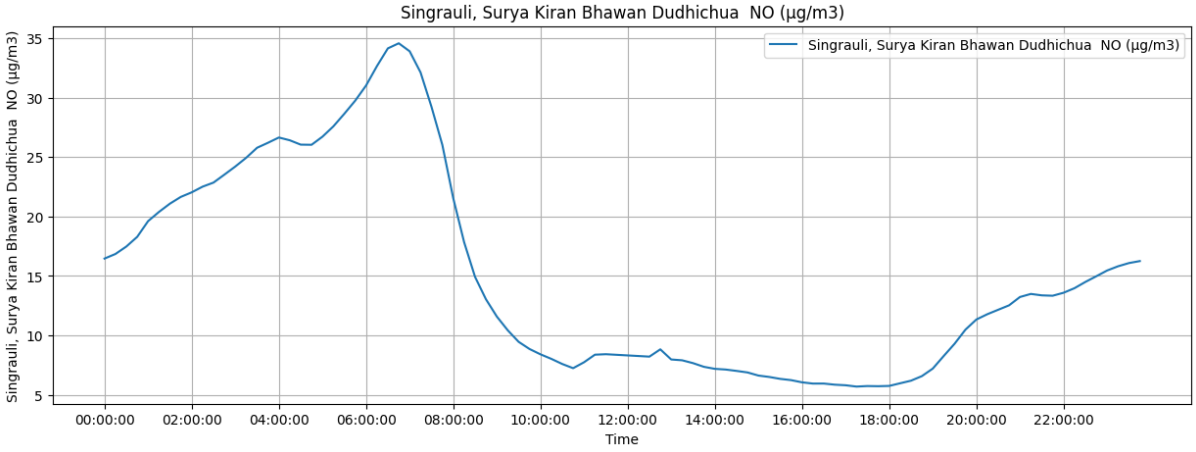
Q4. Descriptive Analysis: Identifies patterns in time series data at the time of coal India open-pit blasting effect, in coal India blasting effect time is 13:45 to 14:45, finds trends like cycles, or seasonal variation. Can Descriptive analysis be categorized into four types which are measures of frequency, central tendency, dispersion or variation, and position of air pollution data?

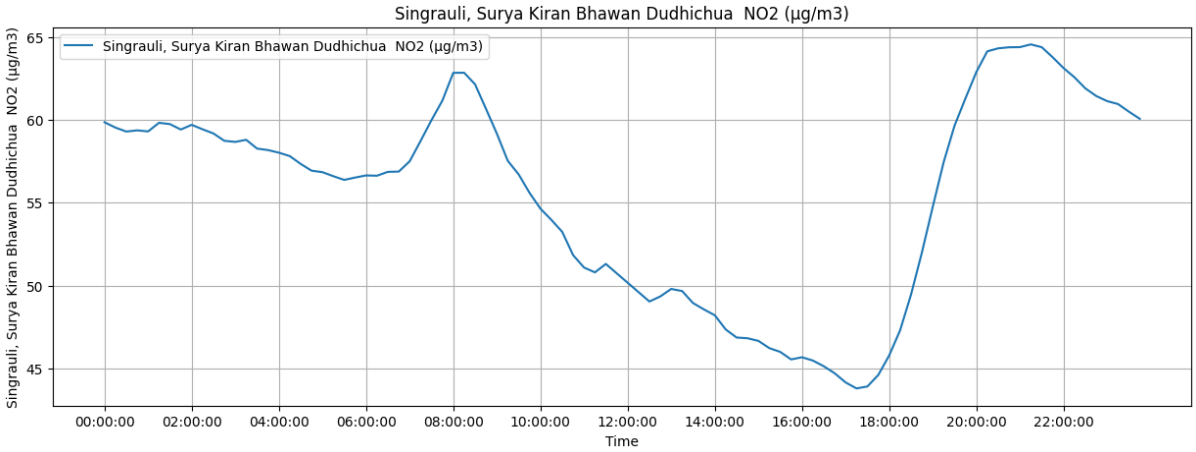
Ans –

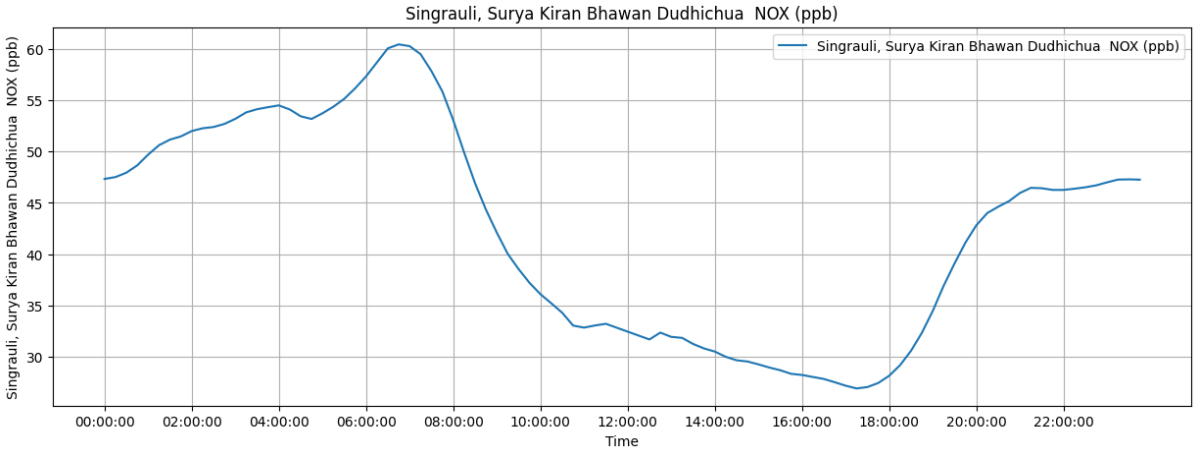
```
new_df = dataSet.copy()
new_df['From'] = pd.to_datetime(new_df['From'][:8640], format="mixed")
new_df['Time'] = new_df['From'].dt.strftime('%H:%M:%S')

new_df_means = new_df.groupby('Time').mean(numeric_only = "numeric_only").reset_index()
new_df_means

for column in columns:
    plt.figure(figsize=(15, 5))
    plt.plot(new_df_means['Time'] ,new_df_means[column] , label = column)
    tick_frequency = 8
    x_ticks = range(0, len(new_df_means['Time']), tick_frequency)
    x_labels = new_df_means['Time'].iloc[x_ticks]
    plt.xticks(x_ticks, x_labels)

    plt.title(column)
    plt.xlabel('Time')
    plt.ylabel(column)
    plt.legend()
    plt.grid()
    plt.show()
```

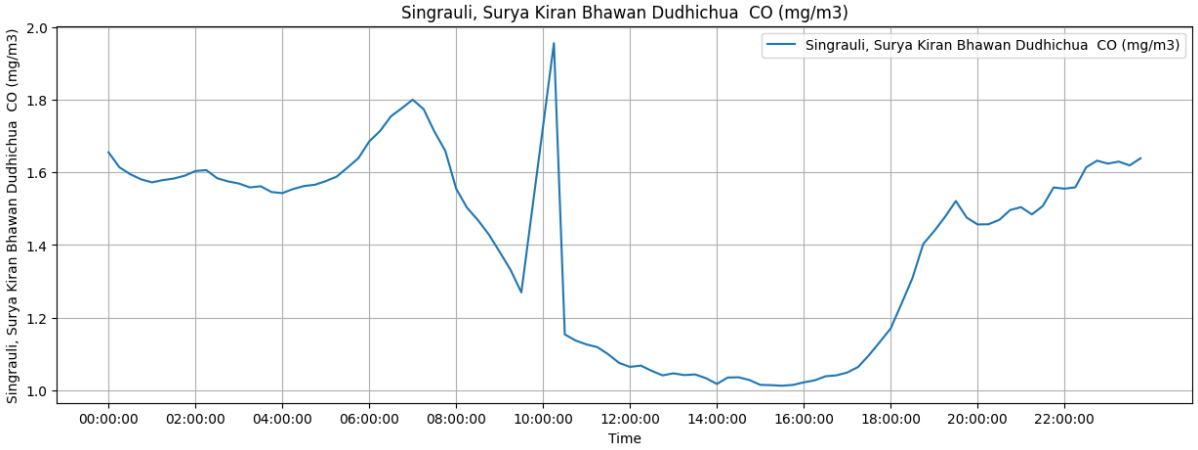


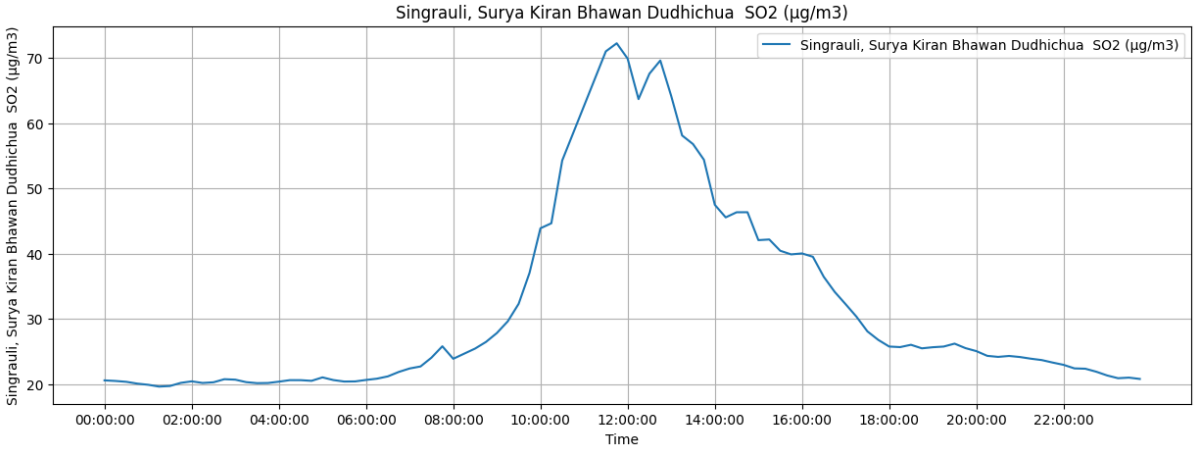


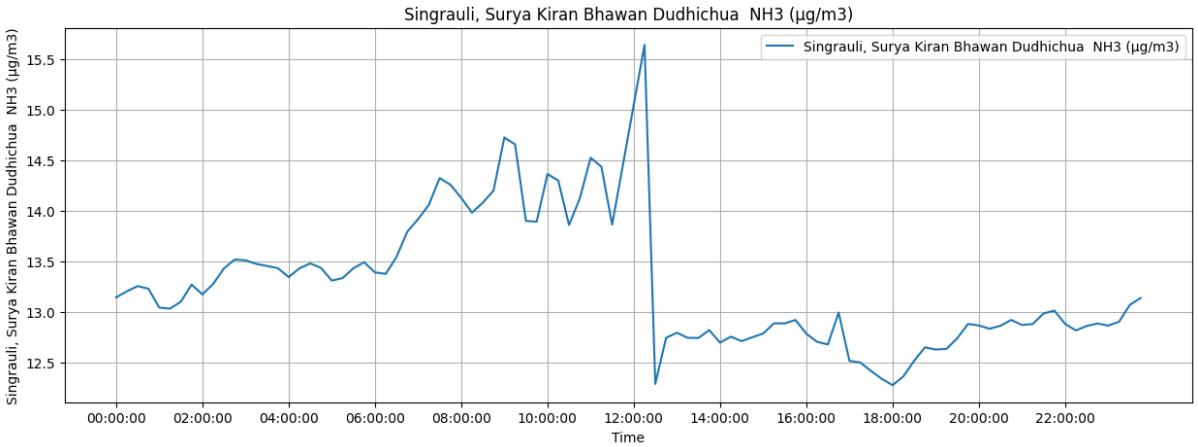

Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)
35
30
25
20
15
10
5
00:00:00
02:00:00
04:00:00
06:00:00
08:00:00
10:00:00
12:00:00
14:00:00
16:00:00
18:00:00
20:00:00
22:00:00
Time

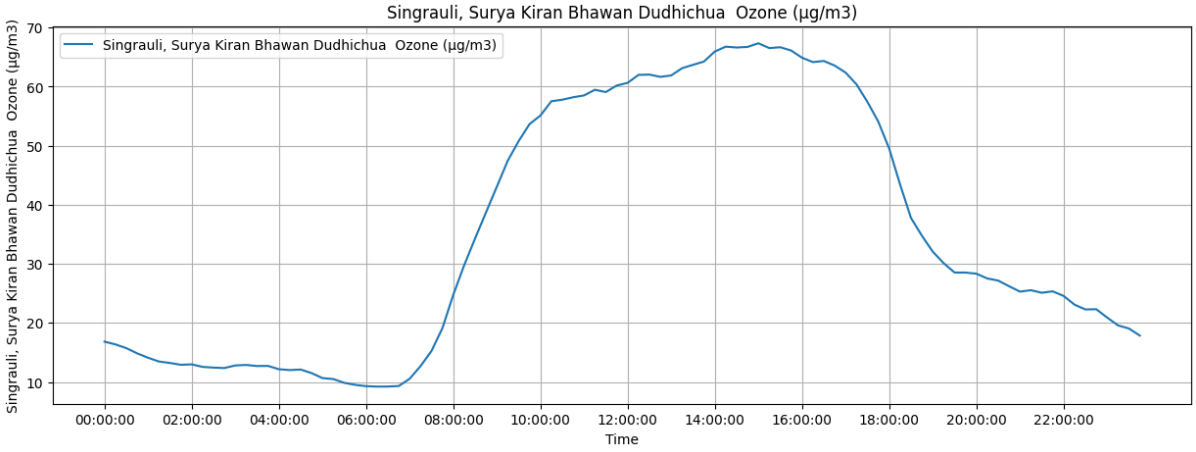

Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)
65
60
55
50
45
00:00:00
02:00:00
04:00:00
06:00:00
08:00:00
10:00:00
12:00:00
14:00:00
16:00:00
18:00:00
20:00:00
22:00:00
Time

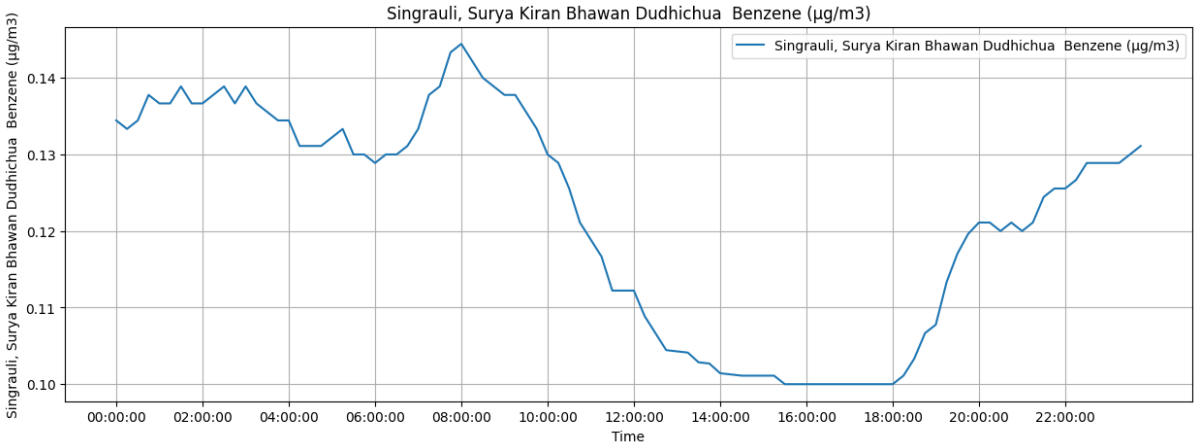

Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)
Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)
Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)
60
55
50
45
40
35
30
00:00:00
02:00:00
04:00:00
06:00:00
08:00:00
10:00:00
12:00:00
14:00:00
16:00:00
18:00:00
20:00:00
22:00:00
Time

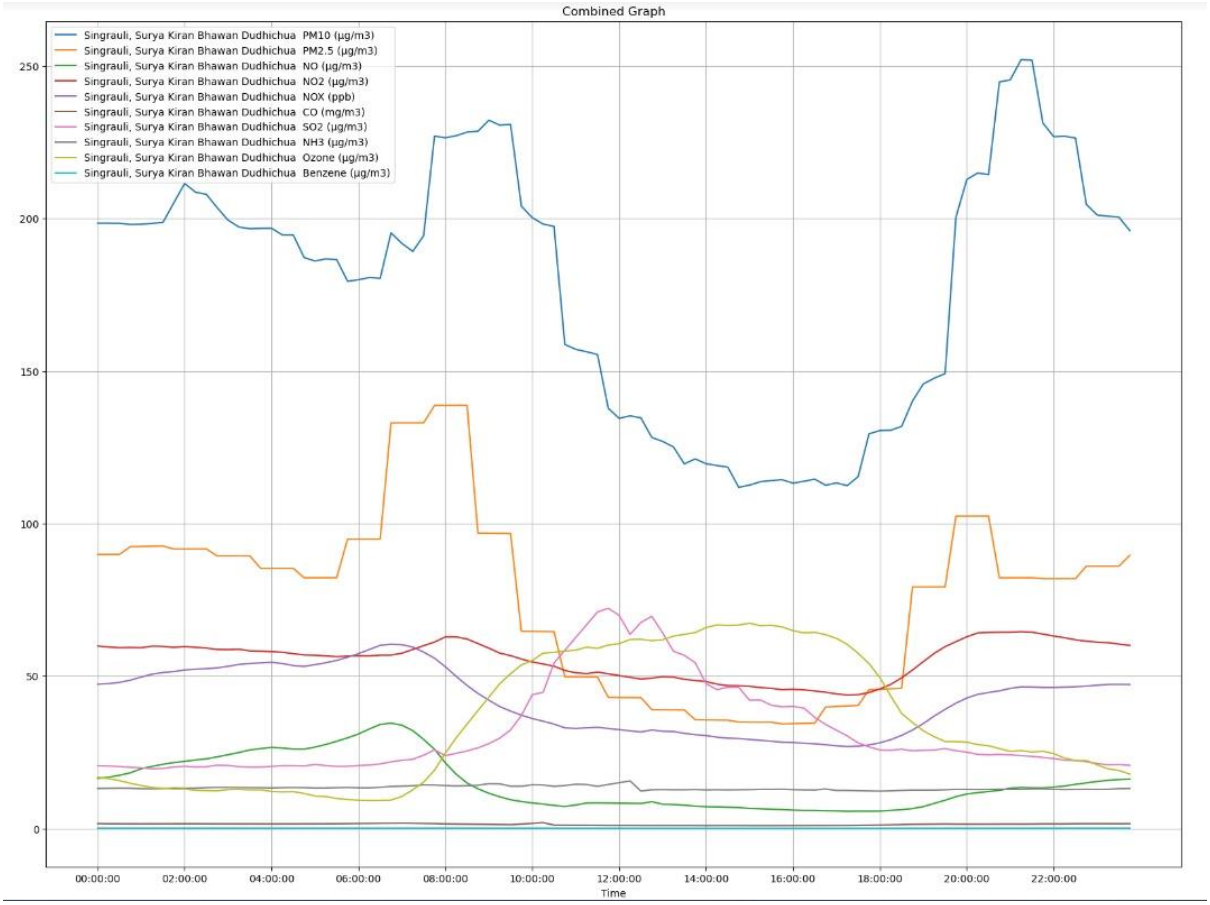

Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)
2.0
1.8
1.6
1.4
1.2
1.0
00:00:00
02:00:00
04:00:00
06:00:00
08:00:00
10:00:00
12:00:00
14:00:00
16:00:00
18:00:00
20:00:00
22:00:00
Time


Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)
70
60
50
40
30
20
00:00:00
02:00:00
04:00:00
06:00:00
08:00:00
10:00:00
12:00:00
14:00:00
16:00:00
18:00:00
20:00:00
22:00:00
Time


Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)
15.5
15.0
14.5
14.0
13.5
13.0
12.5
00:00:00
02:00:00
04:00:00
06:00:00
08:00:00
10:00:00
12:00:00
14:00:00
16:00:00
18:00:00
20:00:00
22:00:00
Time


Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3)
Time


Singrauli, Surya Kiran Bhawan Dudhichua Benzene (µg/m3)
Singrauli, Surya Kiran Bhawan Dudhichua Benzene (µg/m3)
Time



INFERENCE FOR BLASTING TIME

Upon analyzing the aforementioned graphs, several patterns and trends can be observed. At 8 o'clock, the majority of the pollutants exhibit a downward slope, indicating a decrease in their concentrations. This decline is likely attributed to the addition of dust suppressants to the coalfield during that time. Dust suppressants are substances applied to mitigate the dispersion of pollutants, resulting in a reduction in their quantities in the air.

Between 14 o'clock and 17 o'clock, the graphs display a relatively stable trend. This period coincides with the occurrence of blasting activities. During blasting, the release of pollutants from the process overrides the effects of the dust suppressants, rendering their impact less significant. As a result, the pollutant concentrations remain relatively constant during this time frame.

After 18 o'clock, the graphs exhibit an increasing slope, indicating a rise in pollutant levels. This can be attributed to the settled dust particles becoming unsettled and mobile once again due to factors such as wind and demographic changes. The re-suspended dust, combined with other pollutants, leads to an increase in their concentrations.

In summary, the observations from the graphs illustrate the temporal variations in pollutant levels resulting from the combined influences of dust suppressants, blasting activities, and other environmental factors. The understanding of these temporal patterns is crucial for comprehending the dynamics of air pollution in the area. Such insights aid in identifying the

factors that contribute to pollution levels and facilitate the development of appropriate mitigation strategies to address the environmental impact of the coalfield operations.

SEASONALITY

```
from statsmodels.tsa.stattools import adfuller
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
for column in columns:
    result = adfuller(dataSet[column])
    print('ADF Statistic: %f' % result[0])
    print('p-value: %f' % result[1])
    print('No. of Lags : %f' % result[2])
    print('No of Observation used for ADF regression and Critical Value Prediction : %f' % result[3])
    print('Critical Values:')
    for key, value in result[4].items():
        print('\t%s: %.3f' % (key, value))
```

DATA RESULT

```
ADF Statistic: -9.023314
p-value: 0.000000
No. of Lags: 36.000000
No of Observation used for ADF regression and Critical Value Prediction:
8603.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
ADF Statistic: -11.159054
p-value: 0.000000
No. of Lags: 36.000000
No of Observation used for ADF regression and Critical Value Prediction:
8603.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
ADF Statistic: -14.799960
p-value: 0.000000
No. of Lags: 11.000000
No Observation was used for ADF regression and Critical Value Prediction:
8628.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
ADF Statistic: -9.181725
p-value: 0.000000
No. of Lags: 21.000000
No Observation was used for ADF regression and Critical Value Prediction:
8618.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
ADF Statistic: -12.734220
p-value: 0.000000
No. of Lags: 24.000000
```

```
No Observation was used for ADF regression and Critical Value Prediction:
8615.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
ADF Statistic: -9.976113
p-value: 0.000000
No. of Lags: 8.000000
No Observation was used for ADF regression and Critical Value Prediction:
8631.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
ADF Statistic: -14.141881
p-value: 0.000000
No. of Lags: 20.000000
No Observation was used for ADF regression and Critical Value Prediction:
8619.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
ADF Statistic: -3.059899
p-value: 0.029667
No. of Lags: 33.000000
No of Observation used for ADF regression and Critical Value Prediction:
8606.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
ADF Statistic: -20.958946
p-value: 0.000000
No. of Lags: 34.000000
No of Observation used for ADF regression and Critical Value Prediction:
8605.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
ADF Statistic: -9.037111
p-value: 0.000000
No. of Lags: 37.000000
No of Observation used for ADF regression and Critical Value Prediction:
8602.000000
Critical Values:
      1%: -3.431
      5%: -2.862
      10%: -2.567
```

ADF TEST

In the Augmented Dickey-Fuller (ADF) test, the p-value is a crucial measure that helps determine the significance of the test results. The p-value represents the probability of

obtaining test results as extreme as, or more extreme than, the observed results if the null hypothesis is true.

In the context of the ADF test, the null hypothesis assumes the presence of a unit root, indicating that the time series data is non-stationary. The alternative hypothesis assumes the absence of a unit root, indicating that the time series data is stationary.

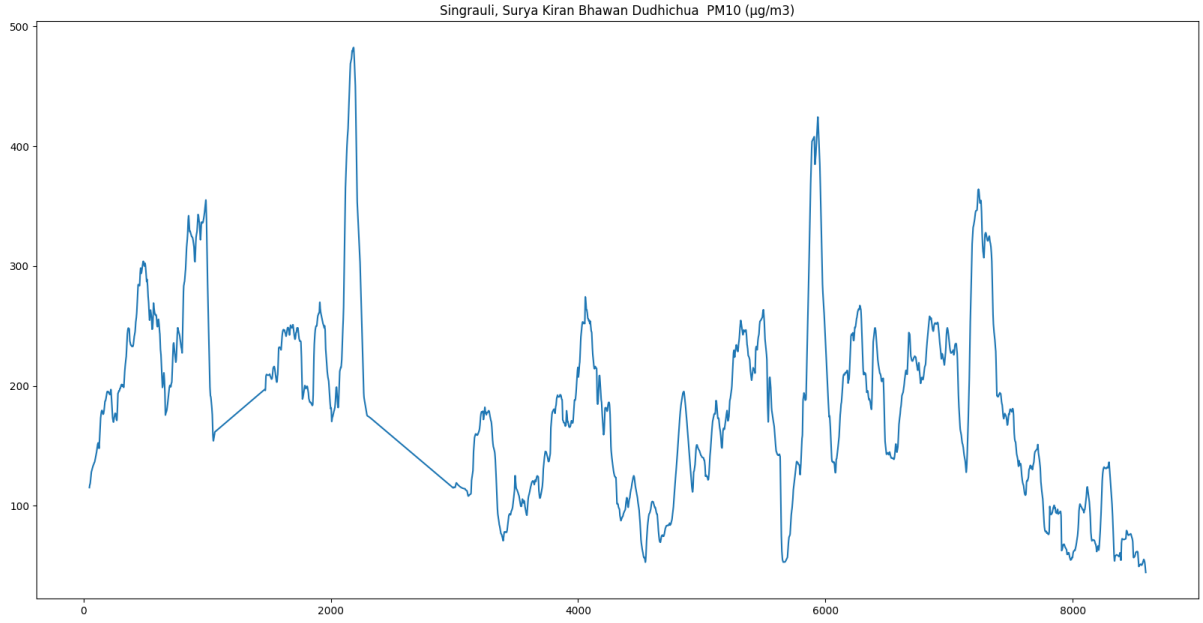
The interpretation of the p-value in the ADF test is as follows:

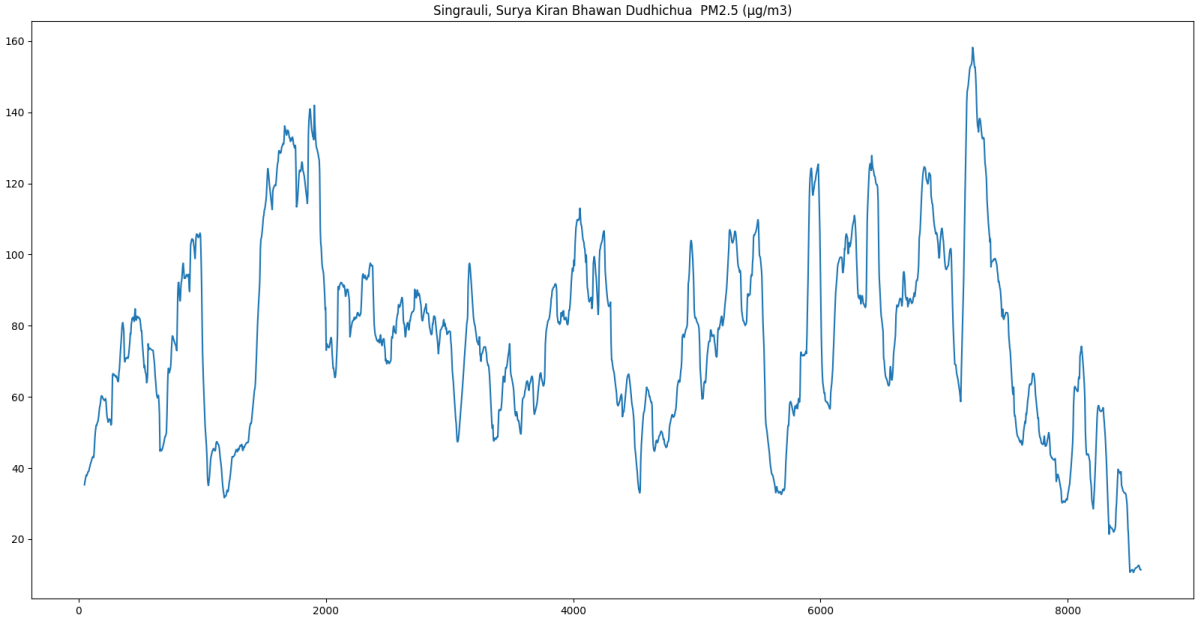
1. p-value > 0.05: If the p-value is greater than the chosen significance level (e.g., 0.05 or 5%), it indicates that there is insufficient evidence to reject the null hypothesis. In this case, the data is considered non-stationary, and the presence of a unit root cannot be ruled out. This suggests that the data may exhibit a trend or other non-stationary behaviour.
2. p-value ≤ 0.05: If the p-value is less than or equal to the chosen significance level, it provides evidence to reject the null hypothesis. This suggests that the data is stationary, and there is evidence against the presence of a unit root. A smaller p-value indicates stronger evidence against the null hypothesis.

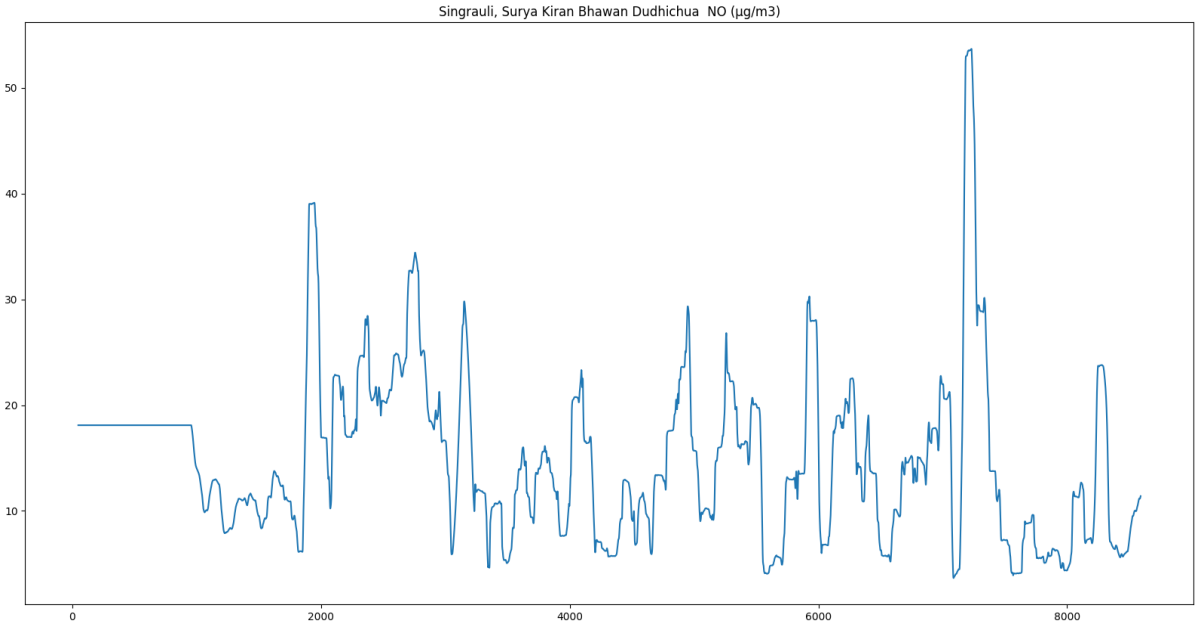
In summary, a smaller p-value in the ADF test indicates stronger evidence in favour of stationarity and against the presence of a unit root. When the p-value is below the chosen significance level (e.g., 0.05), it suggests that the data is likely stationary, and the null hypothesis of non-stationarity can be rejected.

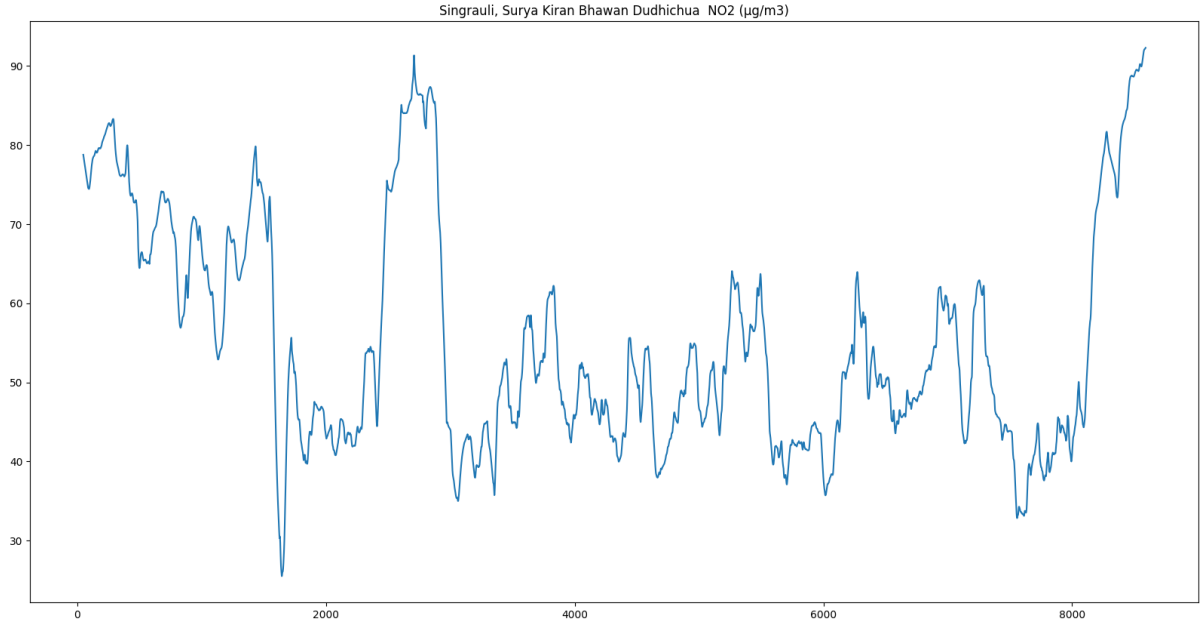
Hence from this test, we can simply say that our data is stationary and have non-seasonality.

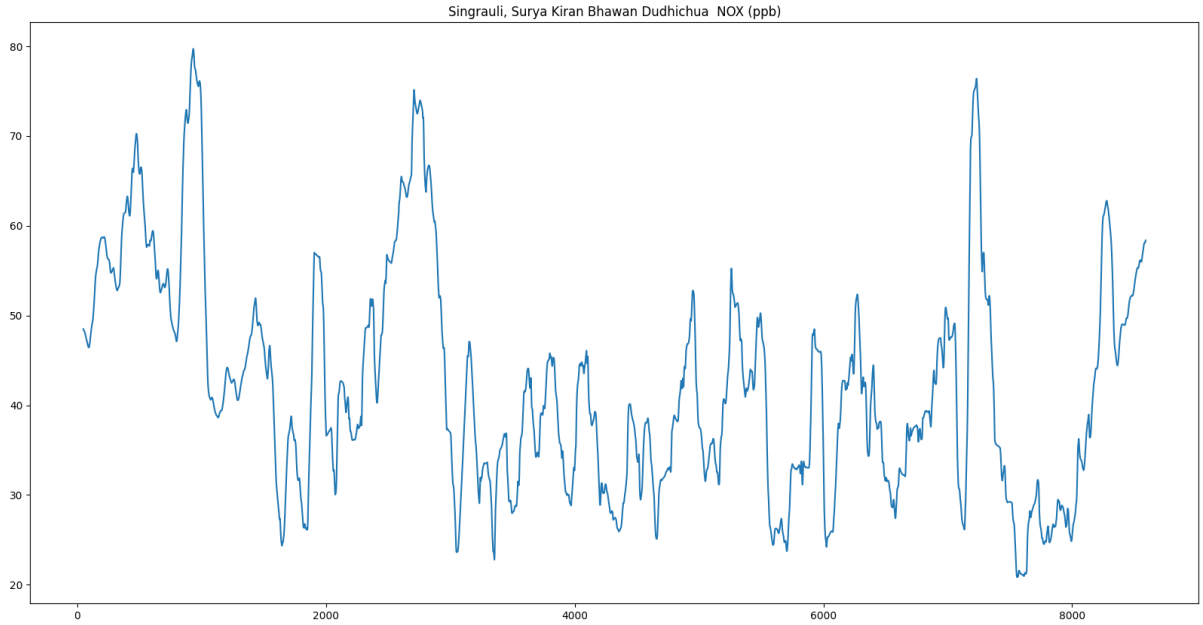
TREND

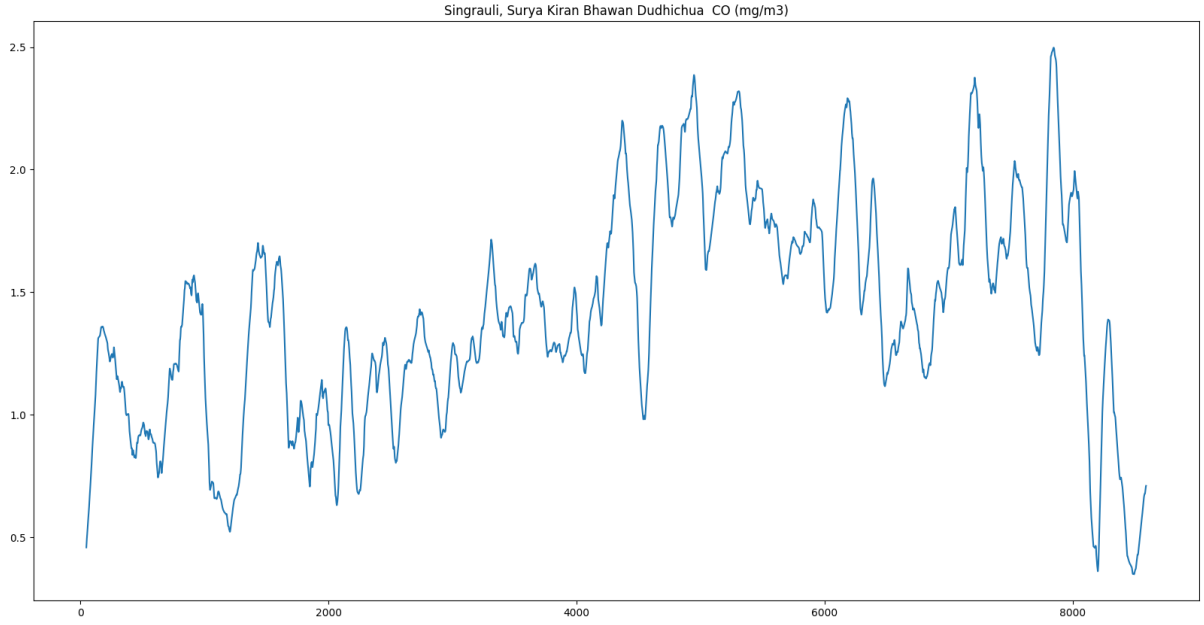


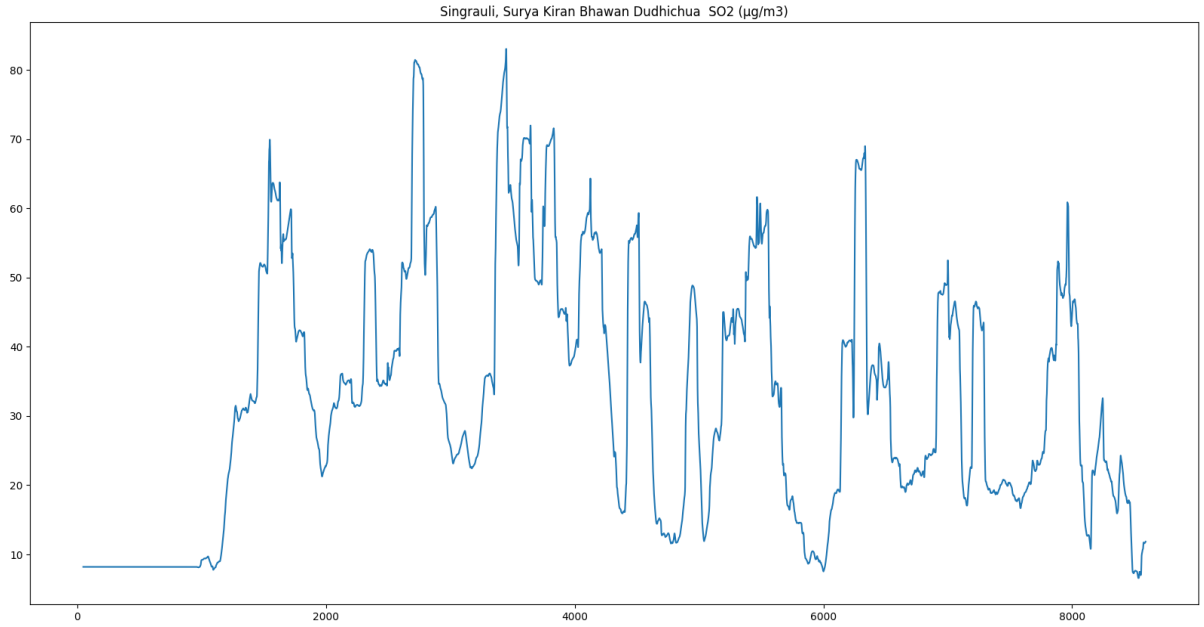

Singrauli, Surya Kiran Bhawan Dudhichua PM2.5 (µg/m3)
160
140
120
100
80
60
40
20
0
2000
4000
6000
8000

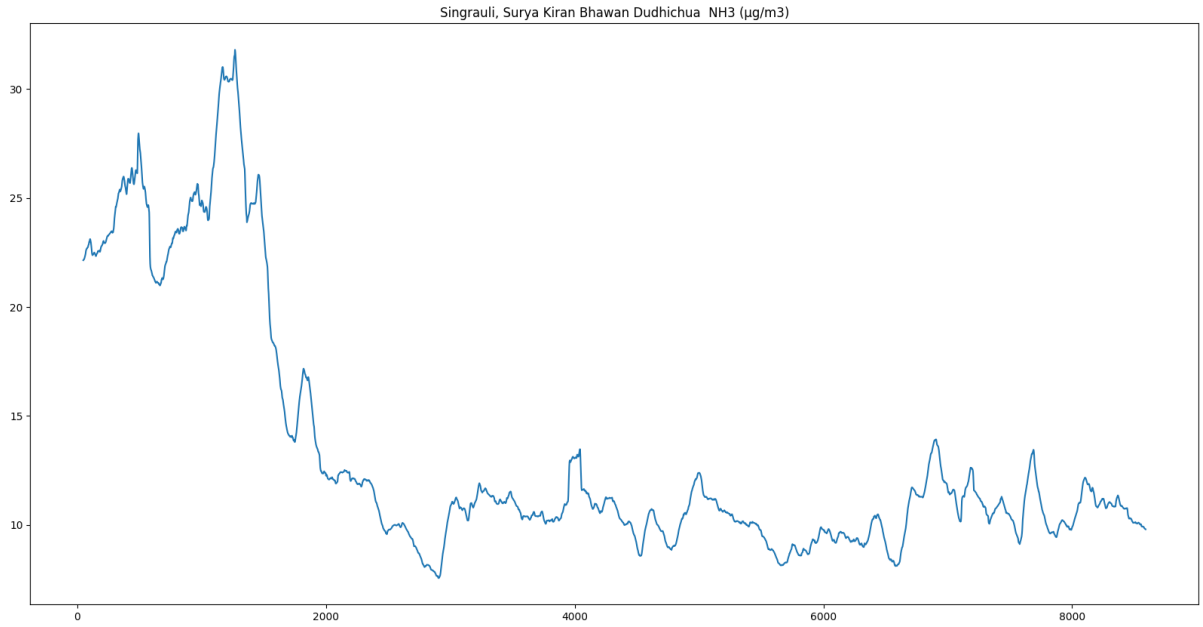

Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)
50
40
30
20
10
0
2000
4000
6000
8000

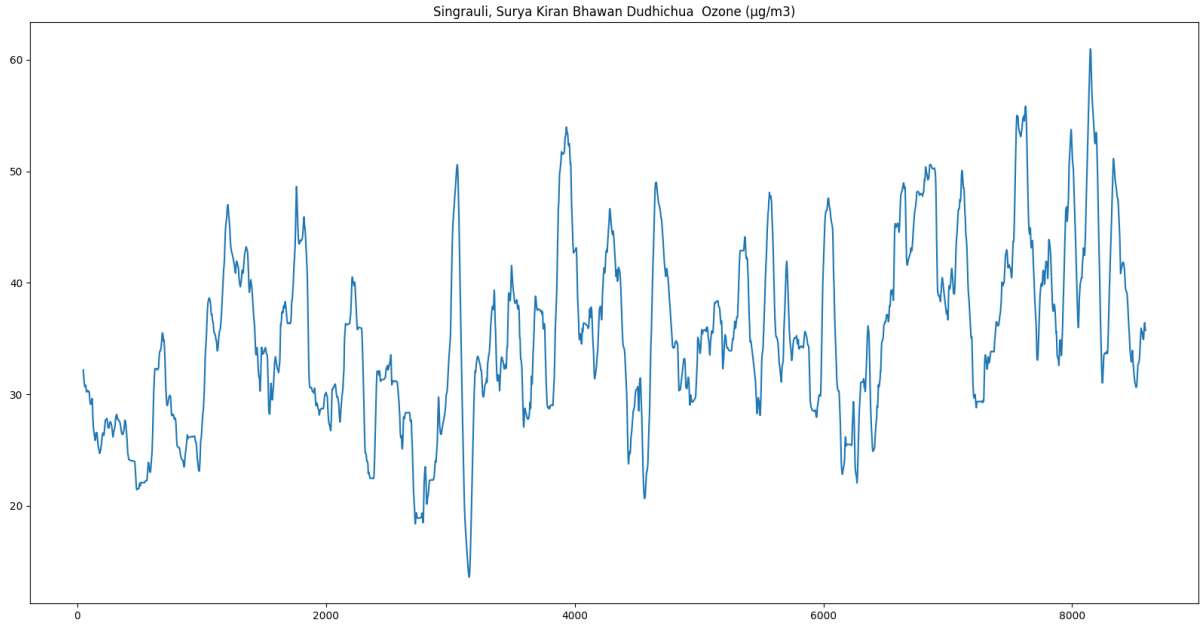

Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)
90
80
70
60
50
40
30
0
2000
4000
6000
8000

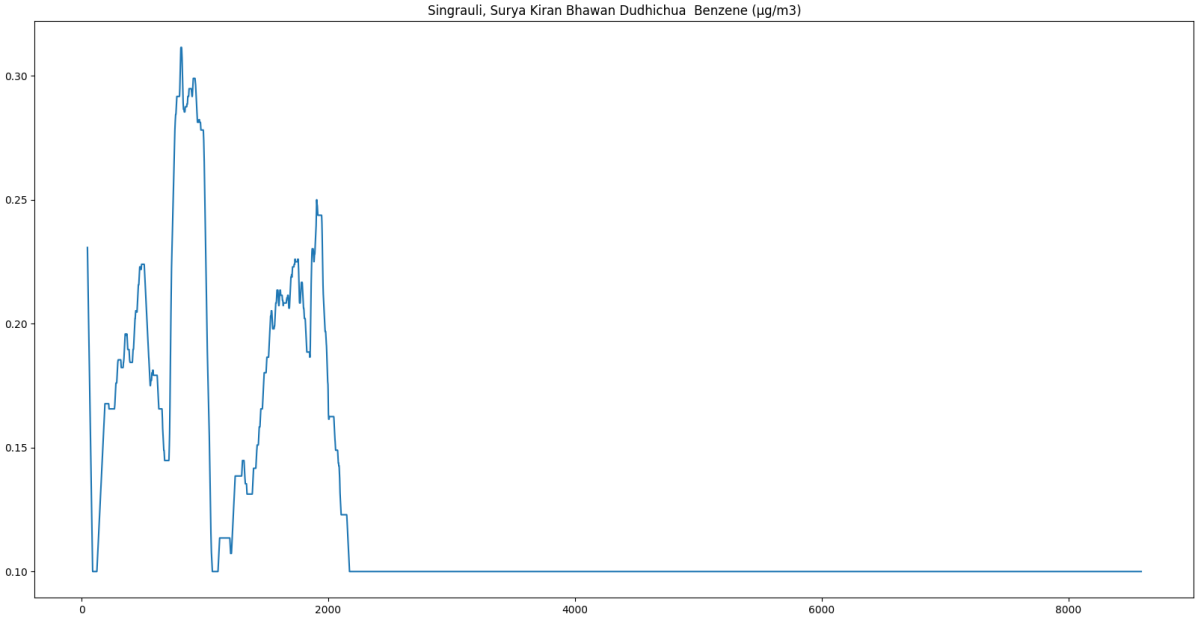

Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)
80
70
60
50
40
30
20
0
2000
4000
6000
8000


Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)
2.5
2.0
1.5
1.0
0.5
0
2000
4000
6000
8000


Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)
80
70
60
50
40
30
20
10
0
2000
4000
6000
8000


Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)
30
25
20
15
10
0
2000
4000
6000
8000


Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3)
60
50
40
30
20
0
2000
4000
6000
8000



INFERENCE

The trend component extracted from the decomposition captures the long-term behaviour of the time series without considering seasonality. By examining the trend component, we can gain insights into the general direction and pattern of the underlying trend in the data. When the trend component appears relatively flat, it indicates a stable or stationary series. Conversely, if the trend component shows a discernible upward or downward pattern, it suggests a systematic trend in the corresponding direction. However, upon analyzing the graph provided, it is evident that there is no apparent upward or downward trend. The graph displays a random pattern, indicating the absence of a discernible trend in our data.

Q5. Do Descriptive analysis can be categorized into four types which are measures of frequency, central tendency, dispersion or variation, and position of air pollution data?

Ans - Yes, descriptive analysis can indeed be categorized into four types based on the measures used to summarize and describe air pollution data. These four types are:

1. **Measures of Frequency:** This type of descriptive analysis focuses on determining the frequency or count of different values or categories within the air pollution data. It provides information on how often certain values or categories occur. Examples of measures of frequency include counts, proportions, percentages, and histograms.
2. **Measures of Central Tendency:** These measures aim to identify the central or typical value around which the air pollution data tend to cluster. The most commonly used measures of central tendency are the mean, median, and mode. The mean represents the average value, the median represents the middle value when the data is arranged in ascending or descending order, and the mode represents the value that occurs most frequently.
3. **Measures of Dispersion or Variation:** This type of descriptive analysis focuses on quantifying the spread, variability, or dispersion of the air pollution data. It provides insights into how much the data points deviate from the central tendency. Measures

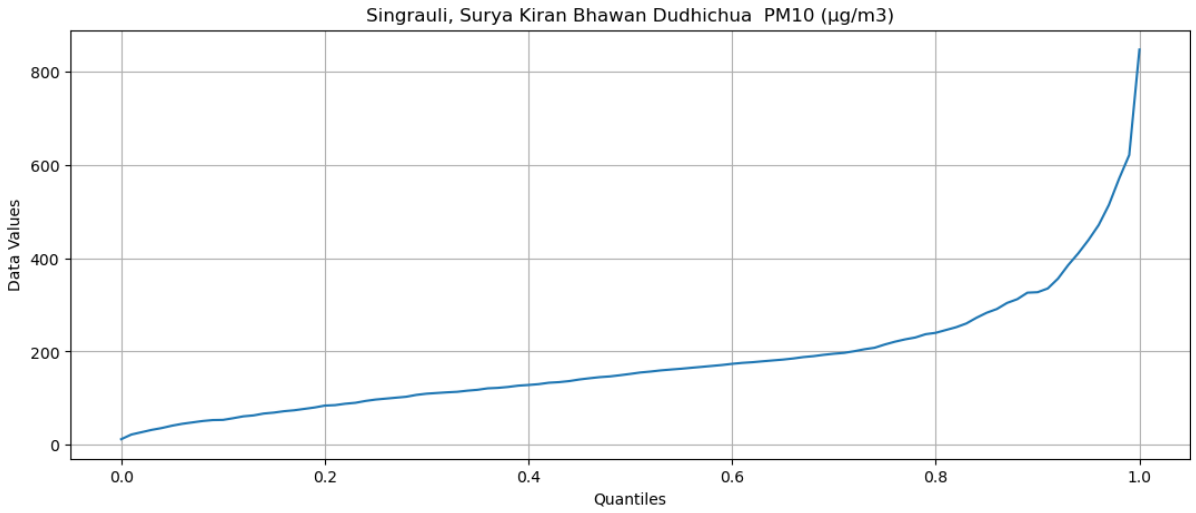
of dispersion include the range, variance, standard deviation, and interquartile range. A larger value for these measures indicates greater variability in the data.

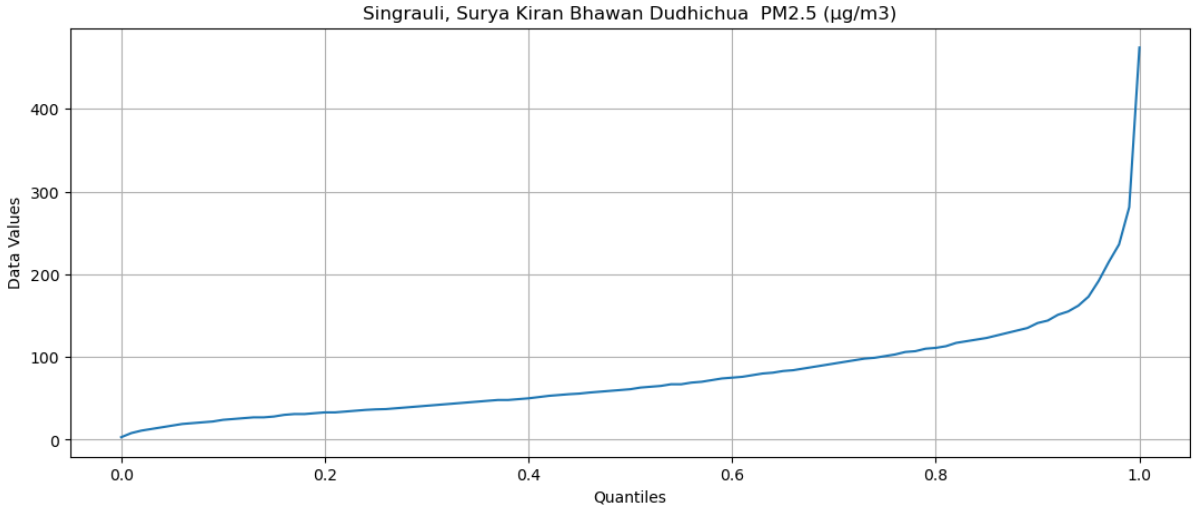
```
dataSet.describe()
✓ 0.0s                                                                    MagicPython
```

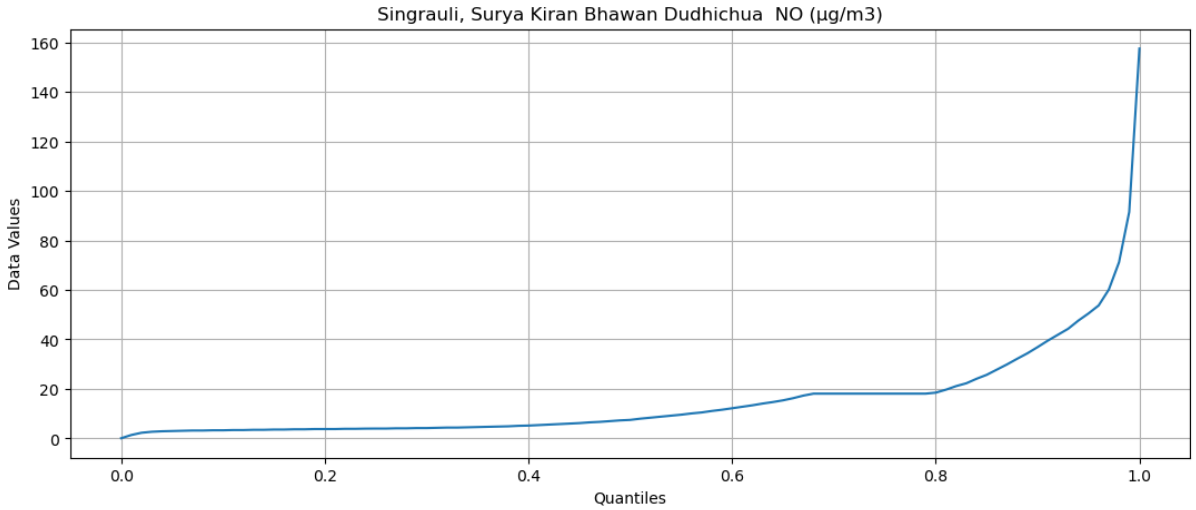
	#	Singrauli, Surya Kiran Bhawan Dudhichua PM10 (µg/m3)	Singrauli, Surya Kiran Bhawan Dudhichua PM2.5 (µg/m3)	Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)	Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)	Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)	Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)	Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)	Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)	Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3)	Singrauli, Surya Kiran Bhawan Dudhichua Benzene (µg/m3)
count	8640.000000	6959.000000	8414.000000	7271.000000	8224.000000	8225.000000	8144.000000	7189.000000	8314.000000	8187.000000	2445.000000
mean	4320.500000	181.408679	75.690397	14.649636	55.757028	42.672219	1.408538	34.232731	13.242663	35.626530	0.177505
std	2494.297496	136.016142	55.245265	19.221385	20.231407	22.435262	0.631056	39.452131	6.151034	27.018693	0.098895
min	1.000000	12.000000	3.000000	0.100000	0.200000	4.200000	0.100000	0.100000	4.600000	0.100000	0.100000
25%	2160.750000	84.000000	36.000000	3.900000	39.400000	25.000000	0.950000	16.100000	9.400000	10.500000	0.100000
50%	4320.500000	145.000000	61.000000	6.100000	53.200000	37.700000	1.420000	25.300000	11.000000	32.400000	0.100000
75%	6480.250000	238.000000	101.000000	16.500000	71.025000	53.800000	1.850000	35.200000	14.000000	58.800000	0.200000
max	8640.000000	847.000000	474.000000	157.500000	106.900000	165.200000	4.000000	645.600000	62.400000	123.800000	0.600000

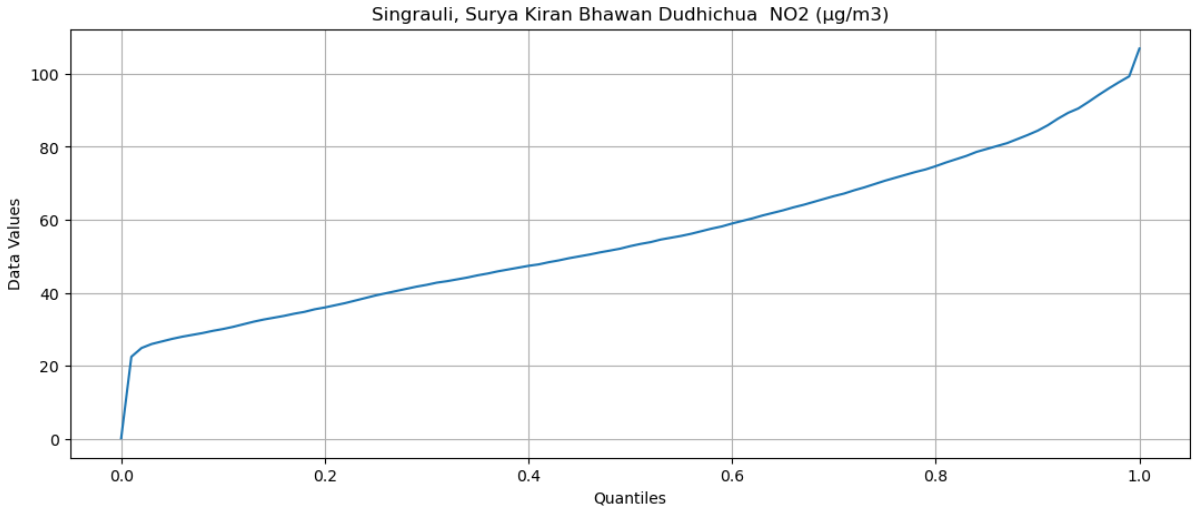
4. **Measures of Position:** This category includes measures that identify specific positions or percentiles within the air pollution data. For example, quartiles divide the data into four equal parts, with the first quartile representing the 25th percentile, the second quartile representing the 50th percentile (which is equivalent to the median), and the third quartile representing the 75th percentile. Measures of position help understand the relative location of data points within the overall distribution.

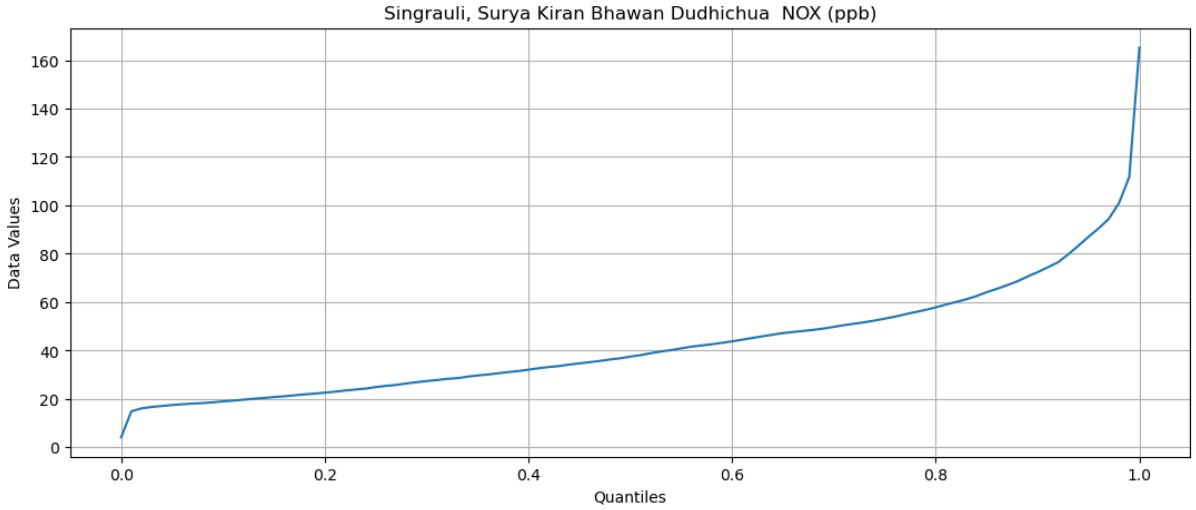
Quantile Plot

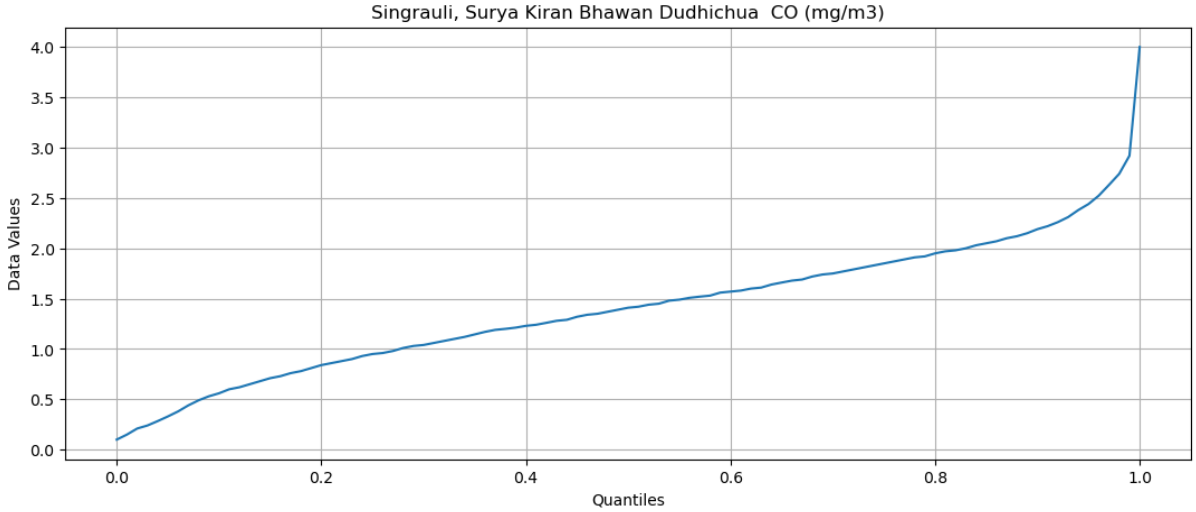


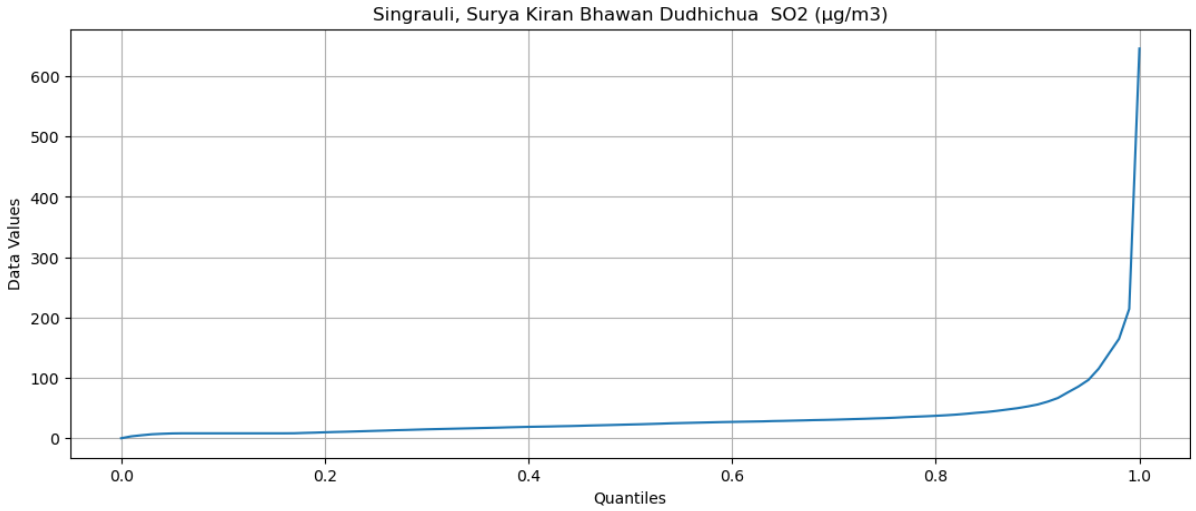

Singrauli, Surya Kiran Bhawan Dudhichua PM2.5 (µg/m3)
Data Values
Quantiles
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0

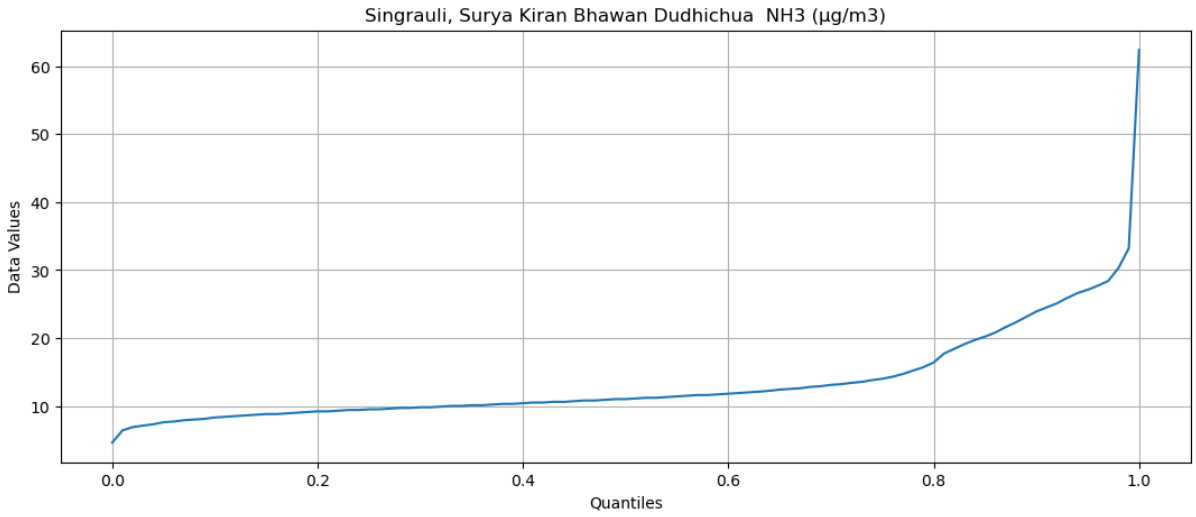

Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)
Data Values
Quantiles
0
20
40
60
80
100
120
140
160
0.0
0.2
0.4
0.6
0.8
1.0

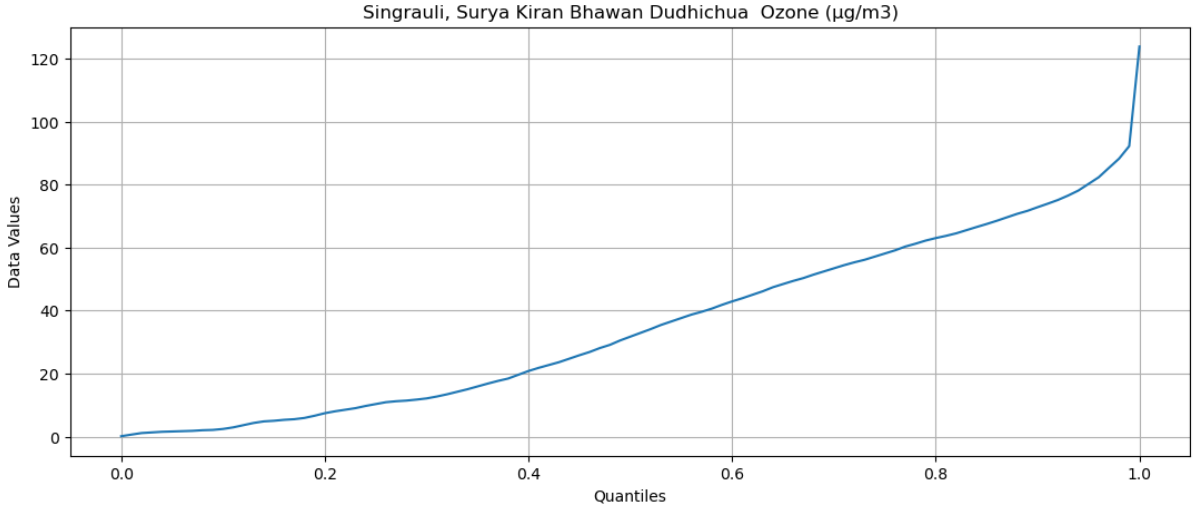

Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)
Data Values
Quantiles
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0

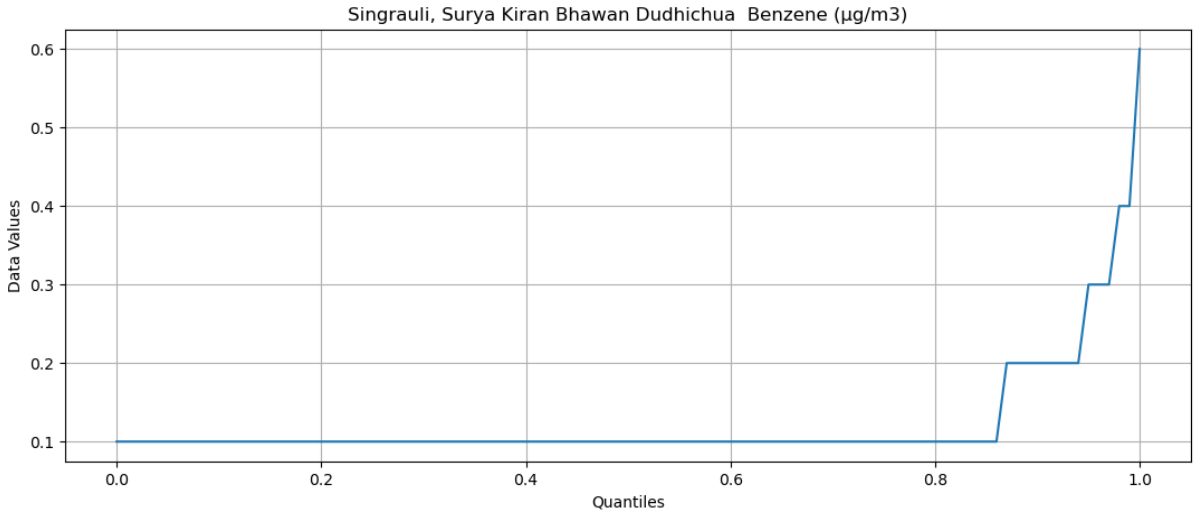

Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)
Data Values
Quantiles
0
20
40
60
80
100
120
140
160
0.0
0.2
0.4
0.6
0.8
1.0


Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)
Data Values
Quantiles
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0.0
0.2
0.4
0.6
0.8
1.0


Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)
Data Values
Quantiles
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0





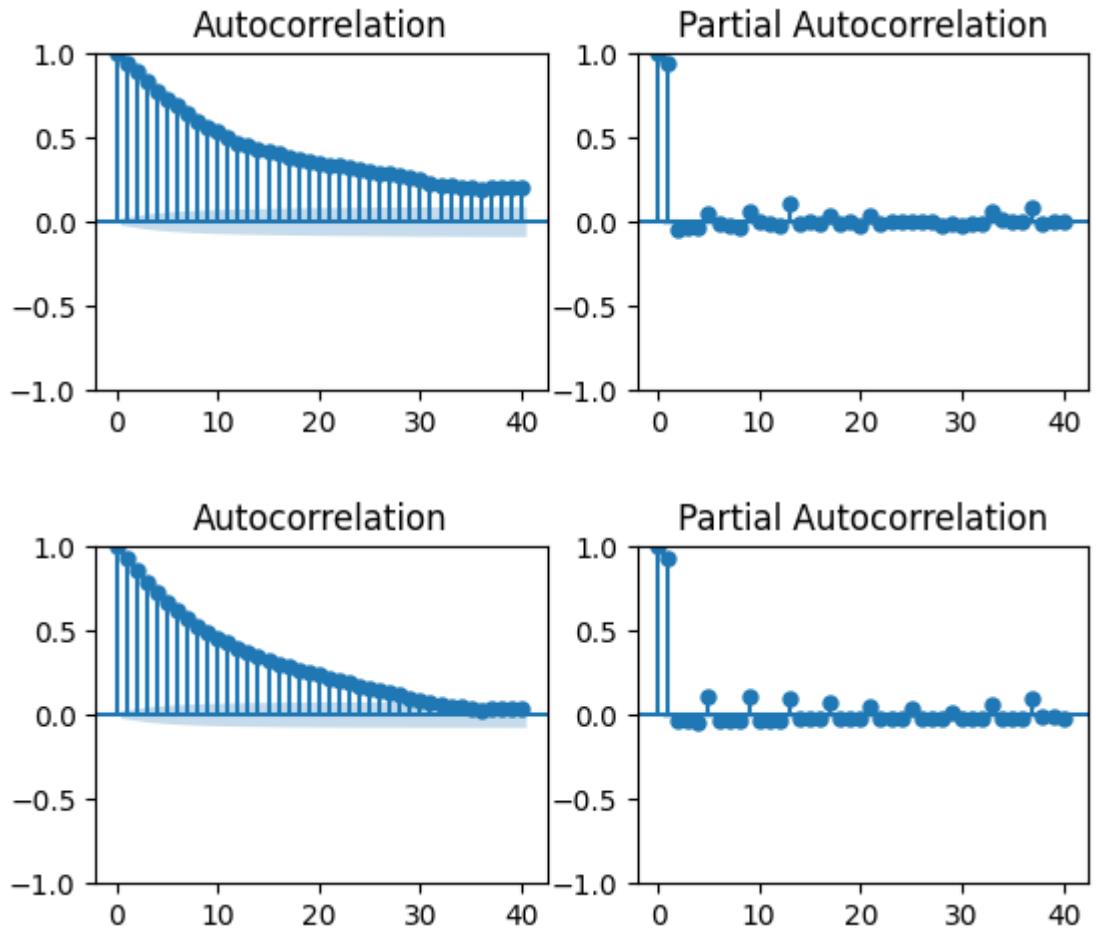


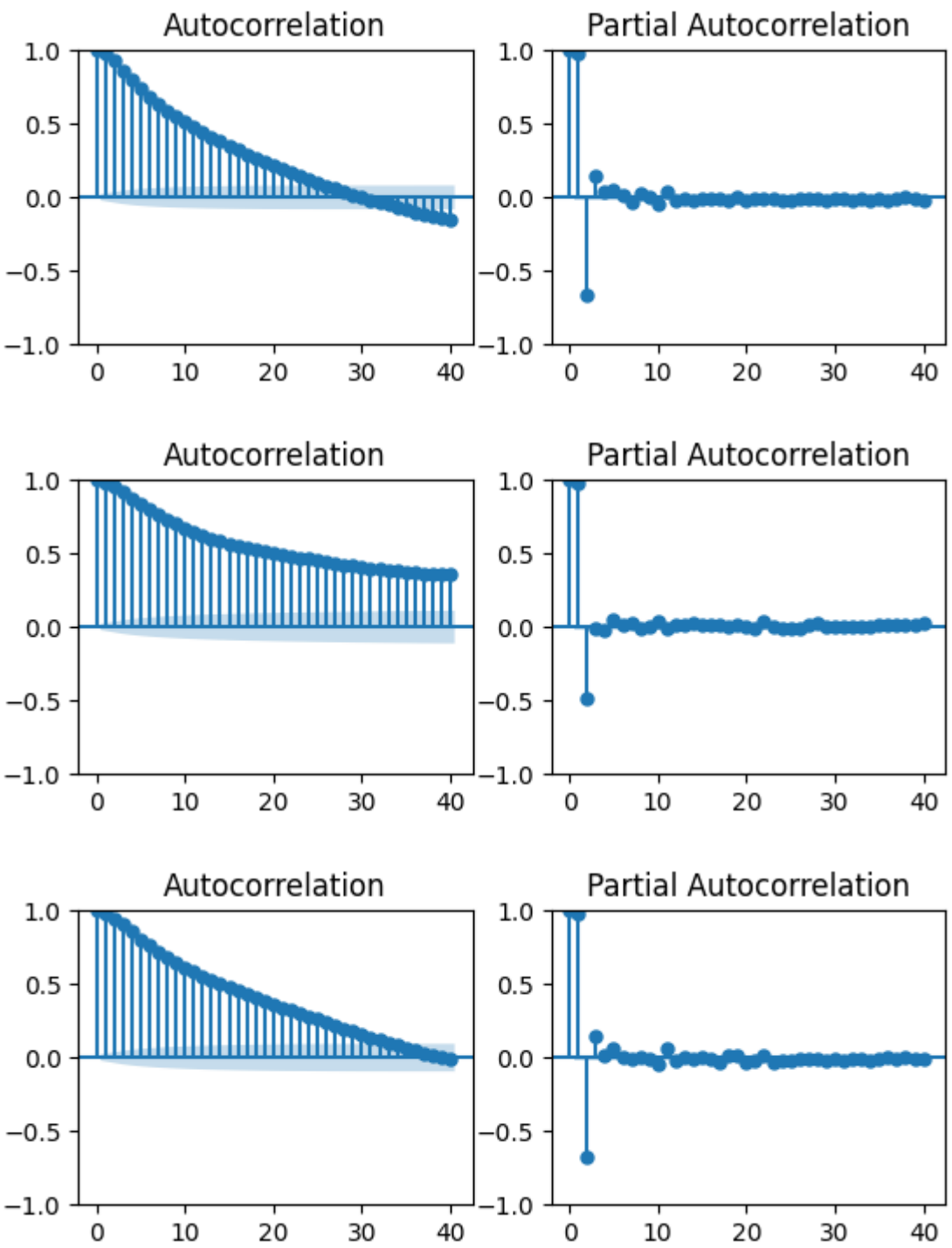
By employing these four types of descriptive analysis, we can gain a comprehensive understanding of the frequency, central tendency, dispersion, and position of air pollution data, enabling us to summarize and interpret the characteristics of the data effectively.

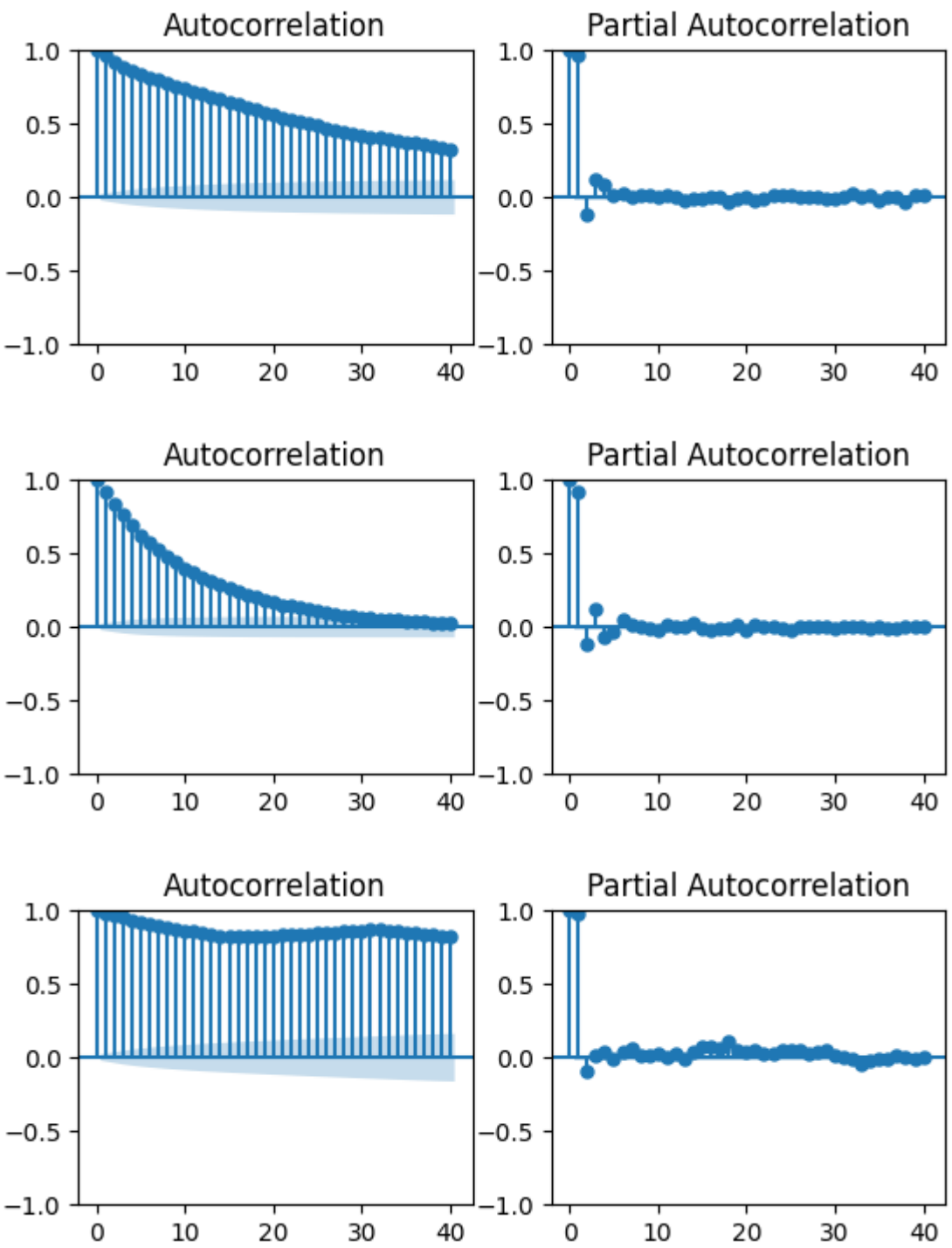
Q6. Forecasting: Predicts future data. This type is based on historical trends of air pollution data set. It uses the historical data as a model for future data, predicting scenarios that could happen along future plot points. Analyse the time series methods used for forecasting are Autoregression (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA)?

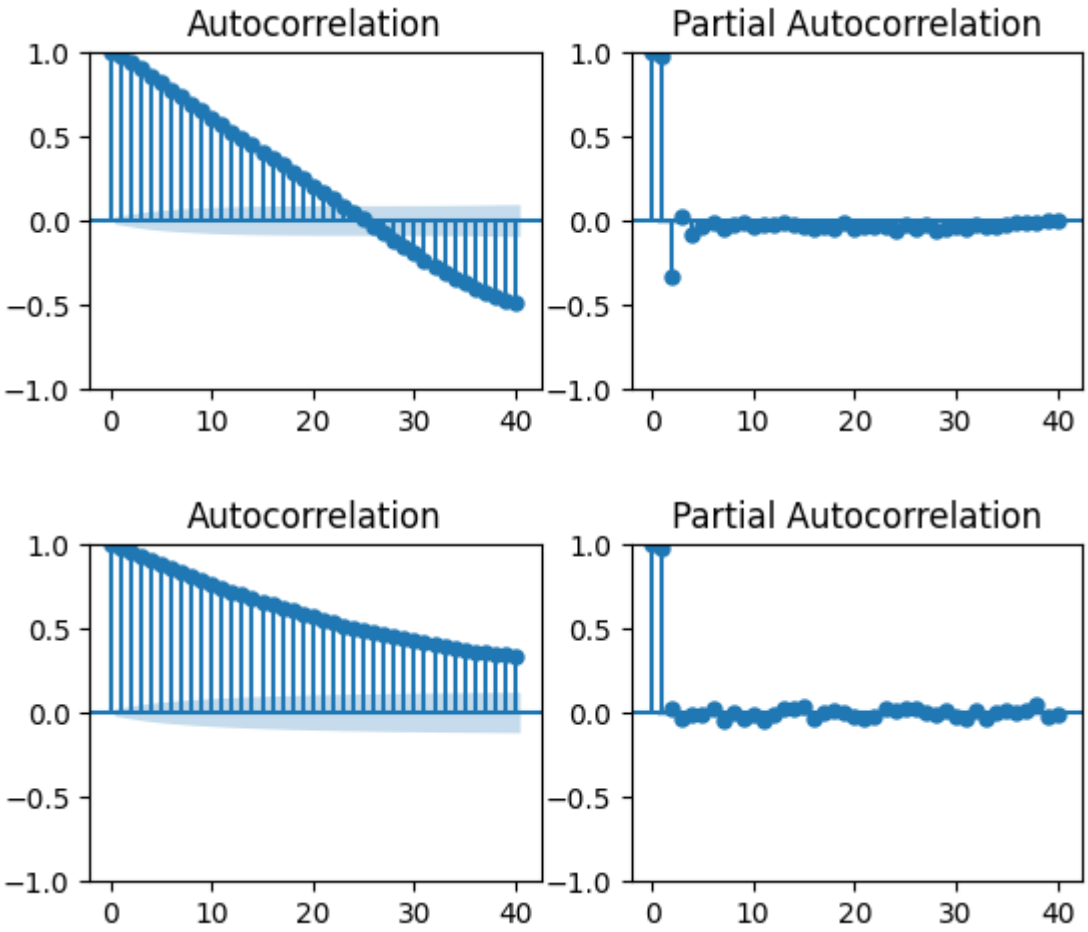
Ans - Forecasting in time series analysis refers to the process of predicting future values or patterns in a time series based on historical data. Time series forecasting is a widely used technique in various domains, including finance, economics, sales, weather forecasting, and more.

```
for column in columns:
    fig = plt.figure()
    ax1 = fig.add_subplot(221)
    ax2 = fig.add_subplot(222)
    plot_acf(dataSet[column], ax = ax1)
    plot_pacf(dataSet[column], ax = ax2)
    plt.show()
```




Autocorrelation
Partial Autocorrelation
Autocorrelation
Partial Autocorrelation
Autocorrelation
Partial Autocorrelation
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40


Autocorrelation
Partial Autocorrelation
Autocorrelation
Partial Autocorrelation
Autocorrelation
Partial Autocorrelation
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40



INFERENCE

Analyzing the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots can provide insights into identifying whether the model is an AR (Autoregressive), MA (Moving Average), ARMA (Autoregressive Moving Average), or ARIMA (Autoregressive Integrated Moving Average) model. Here are some general properties:

- **AR Model:**
  - **ACF:** In an AR model, the ACF plot will show a gradual decrease as the lag increases. The autocorrelation values will slowly decrease and approach zero.
  - **PACF:** The PACF plot will have significant spikes at the lags corresponding to the order of the autoregressive component, and the remaining spikes will generally decrease gradually or become insignificant.
- **MA Model:**
  - **ACF:** In an MA model, the ACF plot will have significant spikes at the lags corresponding to the order of the moving average component. The autocorrelation values will cut off abruptly after the lag corresponding to the order of the MA component.
  - **PACF:** The PACF plot will gradually decrease, potentially with some significant spikes at the beginning that then tail off.
- **ARMA Model:**

- **ACF:** In an ARMA model, both the ACF and PACF plots will have significant spikes at the lags corresponding to the orders of both the autoregressive and moving average components. The autocorrelation values may decay exponentially or gradually decrease.
- **PACF:** The PACF plot may show significant spikes at the beginning that then tail off or gradually decrease.

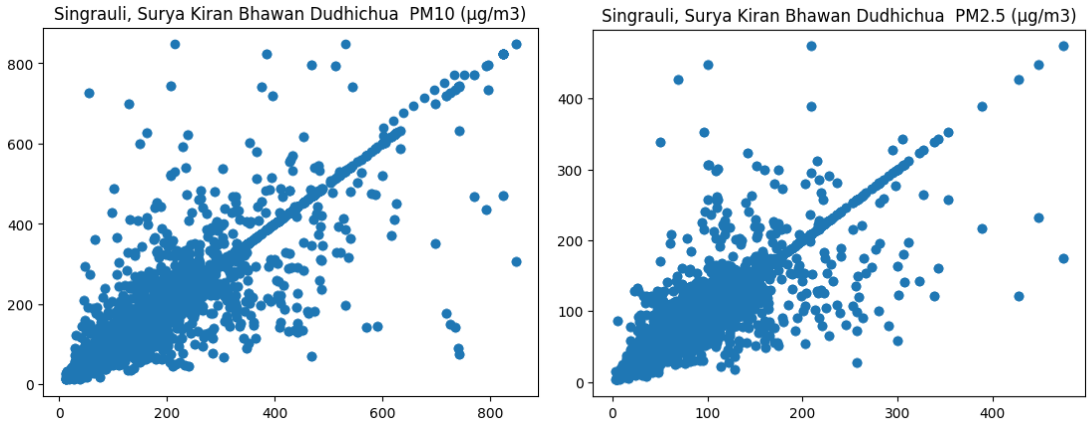
- **ARIMA Model:**
  - **ACF:** In an ARIMA model, the differencing component is also considered. If the ACF plot shows a gradual decrease and the autocorrelation values remain significant even after multiple lags, it suggests the need for differencing (indicating the presence of a unit root).
  - **PACF:** The PACF plot may show a significant spike at lag 1 and then tail off, indicating the autoregressive component. The order of differencing can be determined by examining the number of times differencing is required to achieve stationarity.

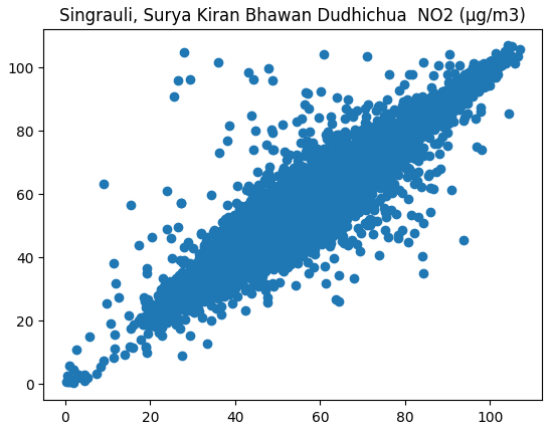
By examining the aforementioned graphs, a clear pattern emerges. The autocorrelation function (ACF) plot shows a gradual decrease as the lag increases, indicating a diminishing correlation between observations as they become more distant in time. On the other hand, the partial autocorrelation function (PACF) plot exhibits a notable spike at the second lag, suggesting a strong correlation specifically at that point. However, beyond the second lag, the correlation decreases and becomes statistically insignificant.

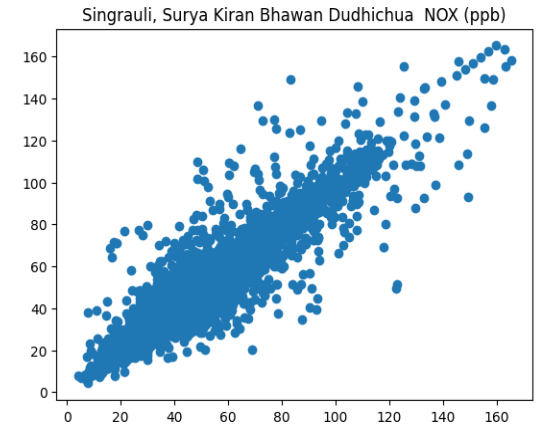
Based on these findings, we can conclude that the most suitable model for the given data is an Auto Regressive (AR) model with a lag value, denoted as "p," equal to 2. This means that the current observation is dependent on the two previous observations. Therefore, we select the **AR(2)** model as the most appropriate choice to capture the underlying dynamics of the data.

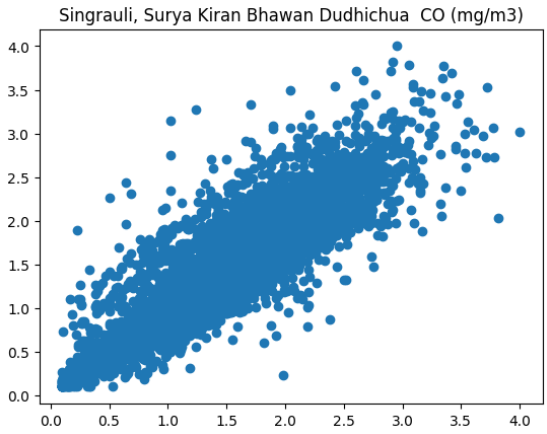
$$X_t = \beta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t$$

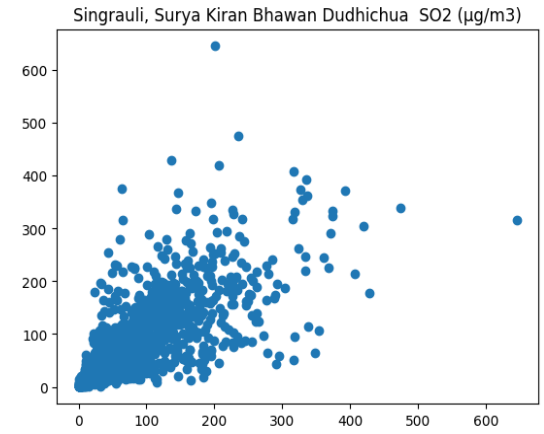
The lag graphs provide evidence that the data exhibits a high level of correlation for a value of p equal to 2. This observation strongly supports the reasonableness of selecting p = 2 as an appropriate choice.

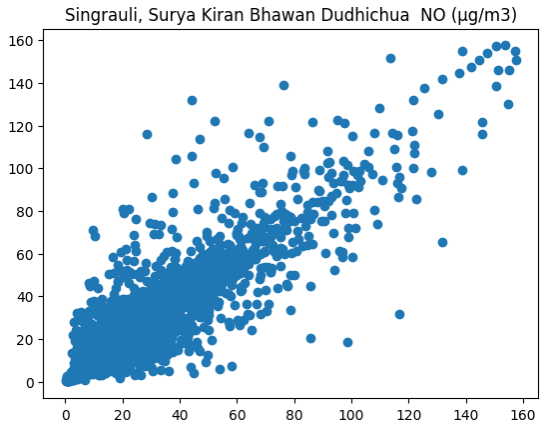


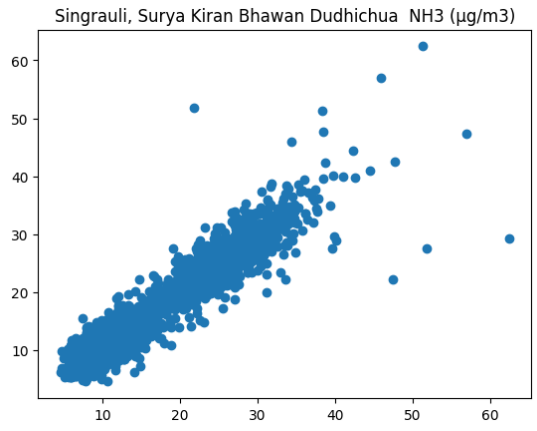

Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)

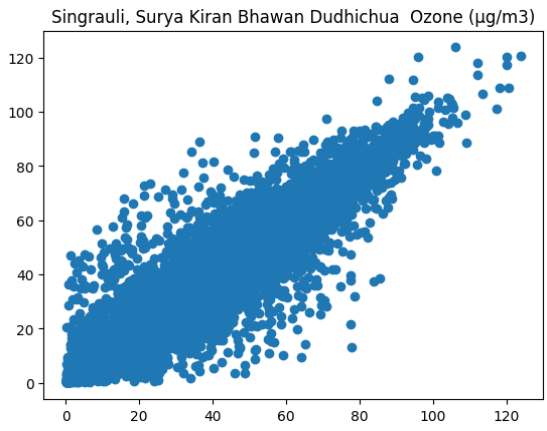

Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)

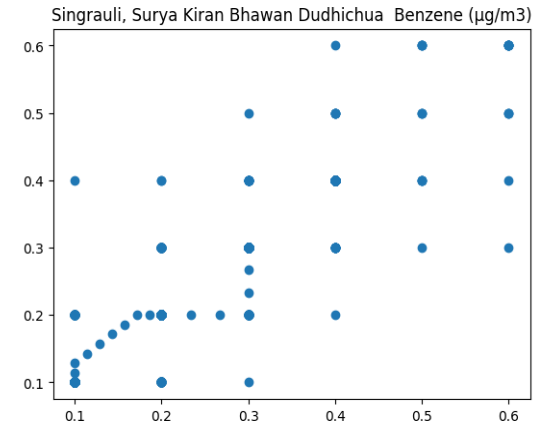

Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)


Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)


Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)


Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)


Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3)


Singrauli, Surya Kiran Bhawan Dudhichua Benzene (µg/m3)

MODEL FITTING

```
from sklearn.metrics import mean_absolute_error
for column in columns:
    data = dataSet[column][:8640]
    train_data = data[:-100]
    test_data = data[-100:]
    ar_model = AutoReg(data, lags = 2).fit()
    pred = ar_model.predict(start = len(train_data), end = len(data), dynamic=False)
    forecast = ar_model.predict(start = len(data), end = len(data)+3, dynamic=False)
    plt.figure(figsize = (15,10))
    plt.title(column)
    plt.plot(forecast , color ="red", label="Forecasted Values")
    plt.plot(pred, color = "blue", label="Predicted Data")
    plt.plot(test_data, color = "green", label="Real Data")
    plt.legend()
    plt.show()
    pred = ar_model.predict(start = len(train_data), end = len(data)-1, dynamic=False)
    rmse = sqrt(mean_squared_error(pred, test_data))
    mean = data.mean()
    print("Mean Absolute Error:", mean_absolute_error(pred,test_data))
    print("Root Mean Squared Error:",rmse)
```

For PM10:

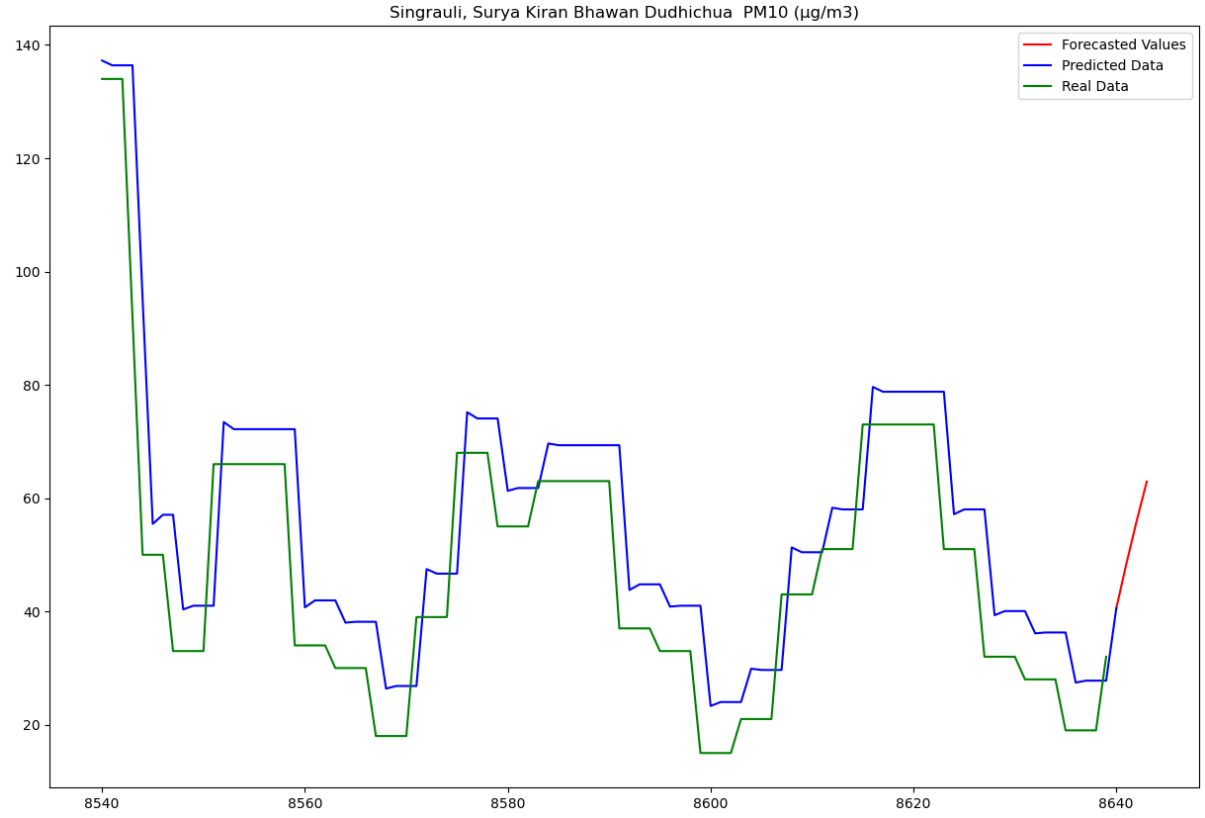
AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua PM10 (µg/m3)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	-44166.077
Method:	Conditional MLE	S.D. of innovations	40.208
Date:	Mon, 26 Jun 2023	AIC	88340.154
Time:	22:17:40	BIC	88368.410
Sample:	2	HQIC	88349.788
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	9.8291	0.759	12.947	0.000	8.341	11.317
Singrauli, Surya Kiran Bhawan Dudhichua PM10 (µg/m3).L1	0.9828	0.011	91.410	0.000	0.962	1.004
Singrauli, Surya Kiran Bhawan Dudhichua PM10 (µg/m3).L2	-0.0382	0.011	-3.556	0.000	-0.059	-0.017

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0613	+0.0000j	1.0613	0.0000
AR.2	24.6460	+0.0000j	24.6460	0.0000



Mean: 177.463079

Root Mean Squared Error: 12.807041

For PM2.5:

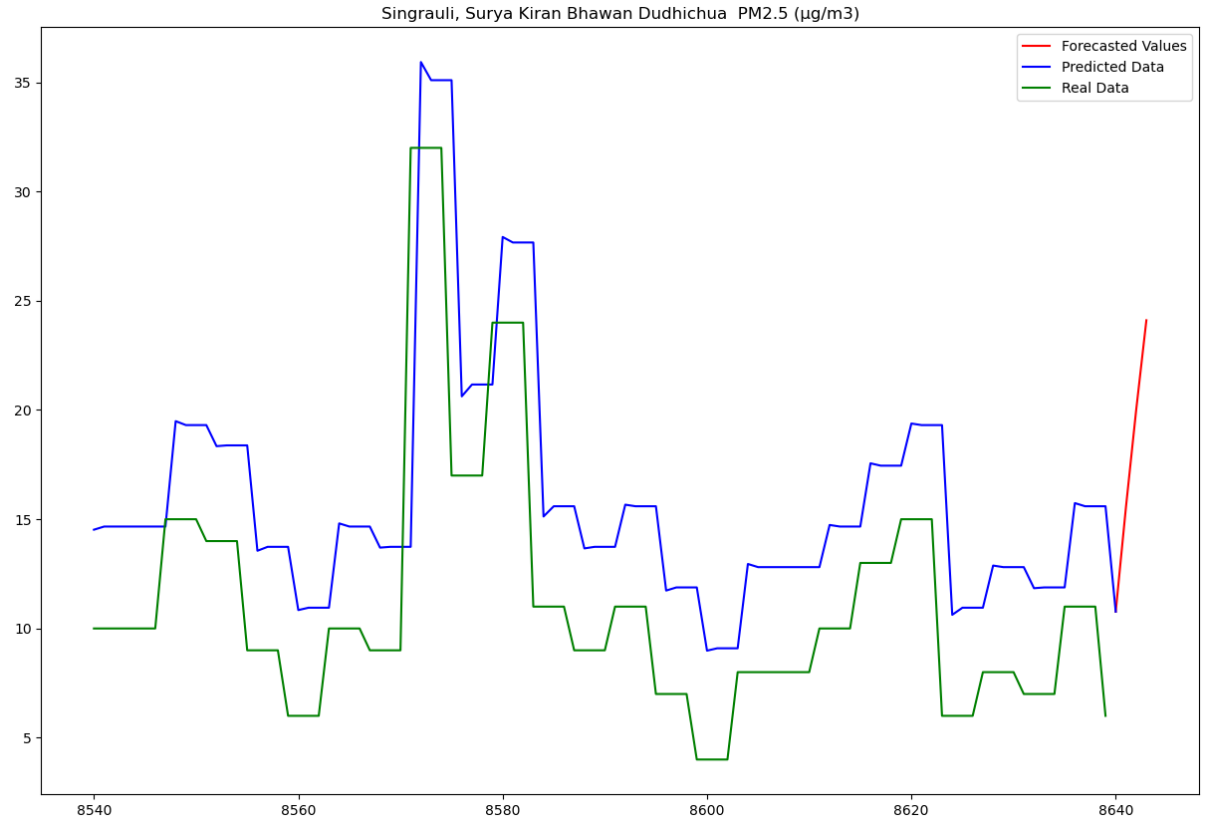
AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua PM2.5 (µg/m3)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	-38120.032
Method:	Conditional MLE	S.D. of innovations	19.968
Date:	Mon, 26 Jun 2023	AIC	76248.064
Time:	22:17:40	BIC	76276.320
Sample:	2	HQIC	76257.698
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.3773	0.370	14.528	0.000	4.652	6.103
Singrauli, Surya Kiran Bhawan Dudhichua PM2.5 (µg/m3).L1	0.9649	0.011	89.735	0.000	0.944	0.986
Singrauli, Surya Kiran Bhawan Dudhichua PM2.5 (µg/m3).L2	-0.0361	0.011	-3.356	0.001	-0.057	-0.015

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0800	+0.0000j	1.0800	0.0000
AR.2	25.6547	+0.0000j	25.6547	0.0000



Mean: 75.555370

Root Mean Squared Error: 5.739367

For NO:

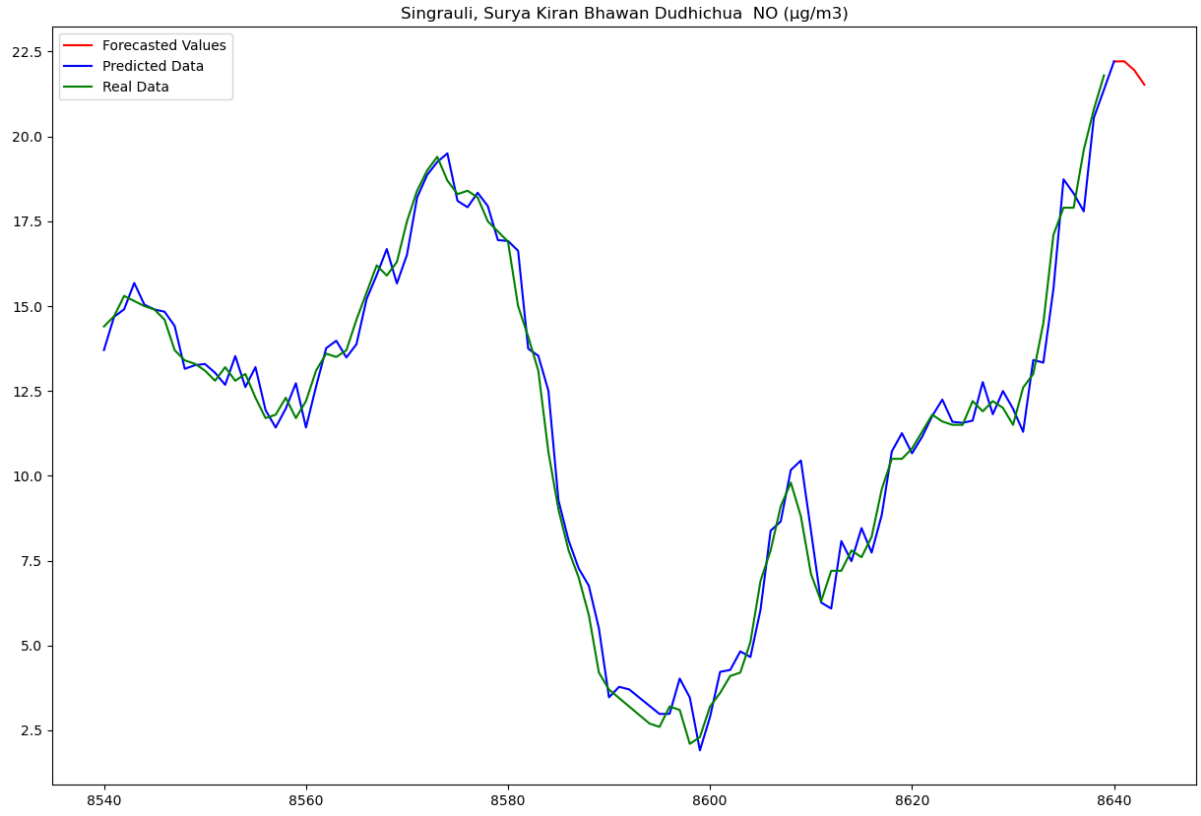
AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	-21144.423
Method:	Conditional MLE	S.D. of innovations	2.798
Date:	Mon, 26 Jun 2023	AIC	42296.847
Time:	22:17:40	BIC	42325.103
Sample:	2	HQIC	42306.481
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5448	0.039	13.850	0.000	0.468	0.622
Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3).L1	1.6233	0.008	200.773	0.000	1.607	1.639
Singrauli, Surya Kiran Bhawan Dudhichua NO (µg/m3).L2	-0.6598	0.008	-81.601	0.000	-0.676	-0.644

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.2302	-0.0475j	1.2311	-0.0061
AR.2	1.2302	+0.0475j	1.2311	0.0061



Mean: 14.940208

Root Mean Squared Error: 0.679109

For NO2:

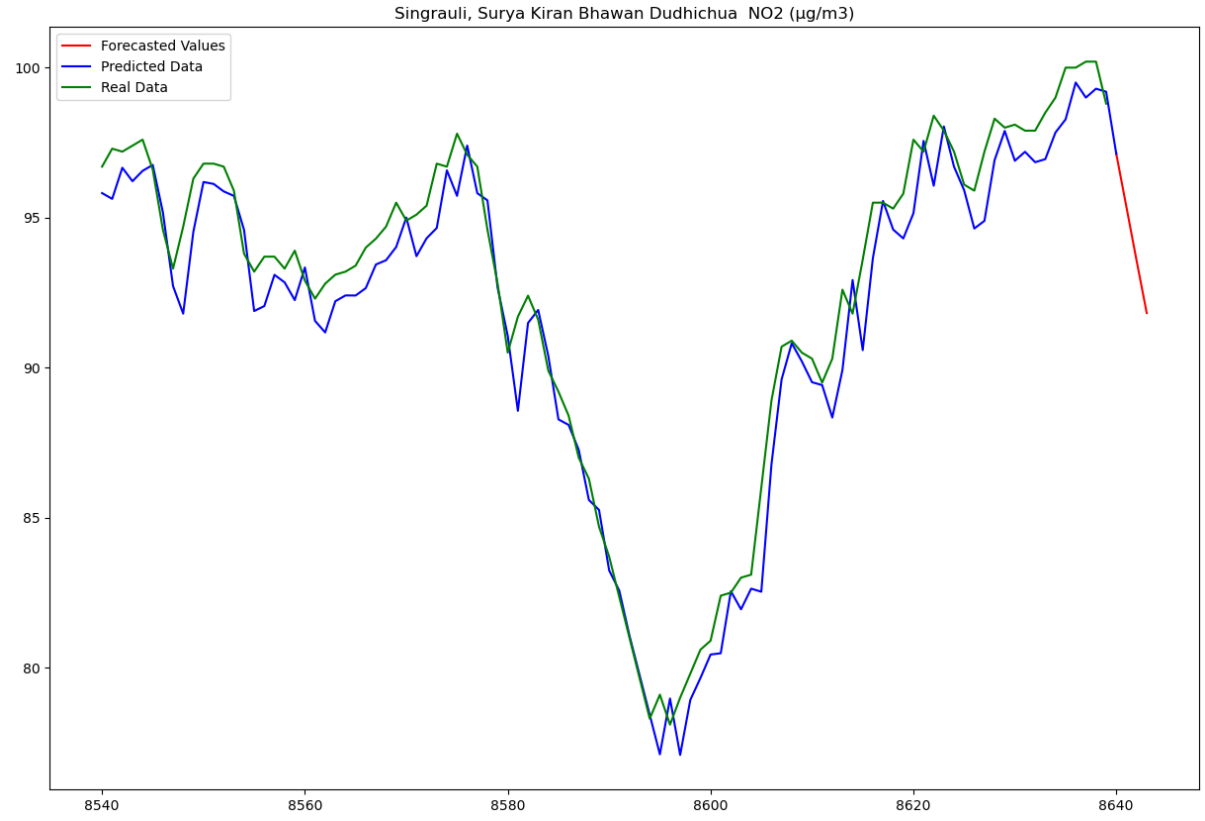
AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	-21776.975
Method:	Conditional MLE	S.D. of innovations	3.011
Date:	Mon, 26 Jun 2023	AIC	43561.949
Time:	22:17:41	BIC	43590.205
Sample:	2	HQIC	43571.584
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2420	0.095	13.084	0.000	1.056	1.428
Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3).L1	1.4793	0.009	158.933	0.000	1.461	1.498
Singrauli, Surya Kiran Bhawan Dudhichua NO2 (µg/m3).L2	-0.5017	0.009	-53.895	0.000	-0.520	-0.483

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0497	+0.0000j	1.0497	0.0000
AR.2	1.8989	+0.0000j	1.8989	0.0000



Mean: 55.430689

Root Mean Squared Error: 1.273081

For NOX:

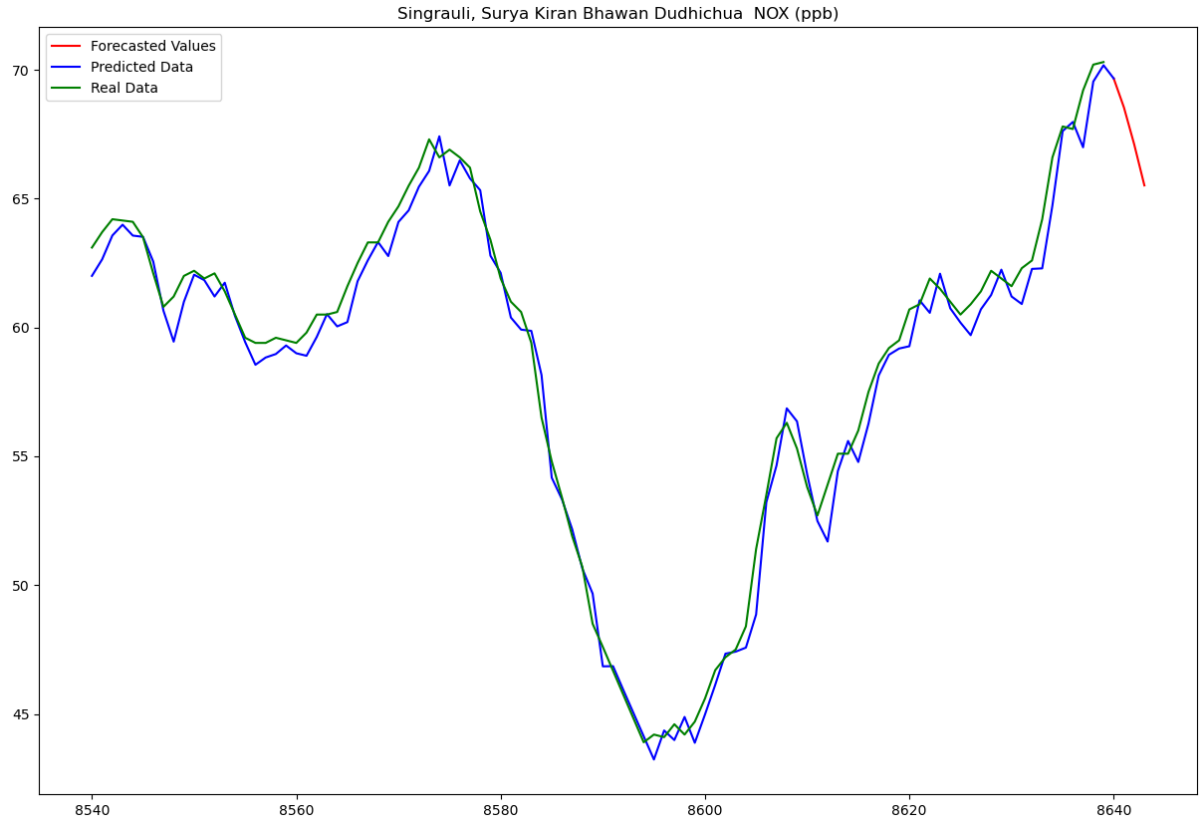
AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	-21178.954
Method:	Conditional MLE	S.D. of innovations	2.809
Date:	Mon, 26 Jun 2023	AIC	42365.907
Time:	22:17:41	BIC	42394.163
Sample:	2	HQIC	42375.542
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.0703	0.065	16.387	0.000	0.942	1.198
Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb).L1	1.6553	0.008	209.979	0.000	1.640	1.671
Singrauli, Surya Kiran Bhawan Dudhichua NOX (ppb).L2	-0.6806	0.008	-86.329	0.000	-0.696	-0.665

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1183	+0.0000j	1.1183	0.0000
AR.2	1.3138	+0.0000j	1.3138	0.0000



Mean: 42.328802

Root Mean Squared Error: 0.862718

For CO:

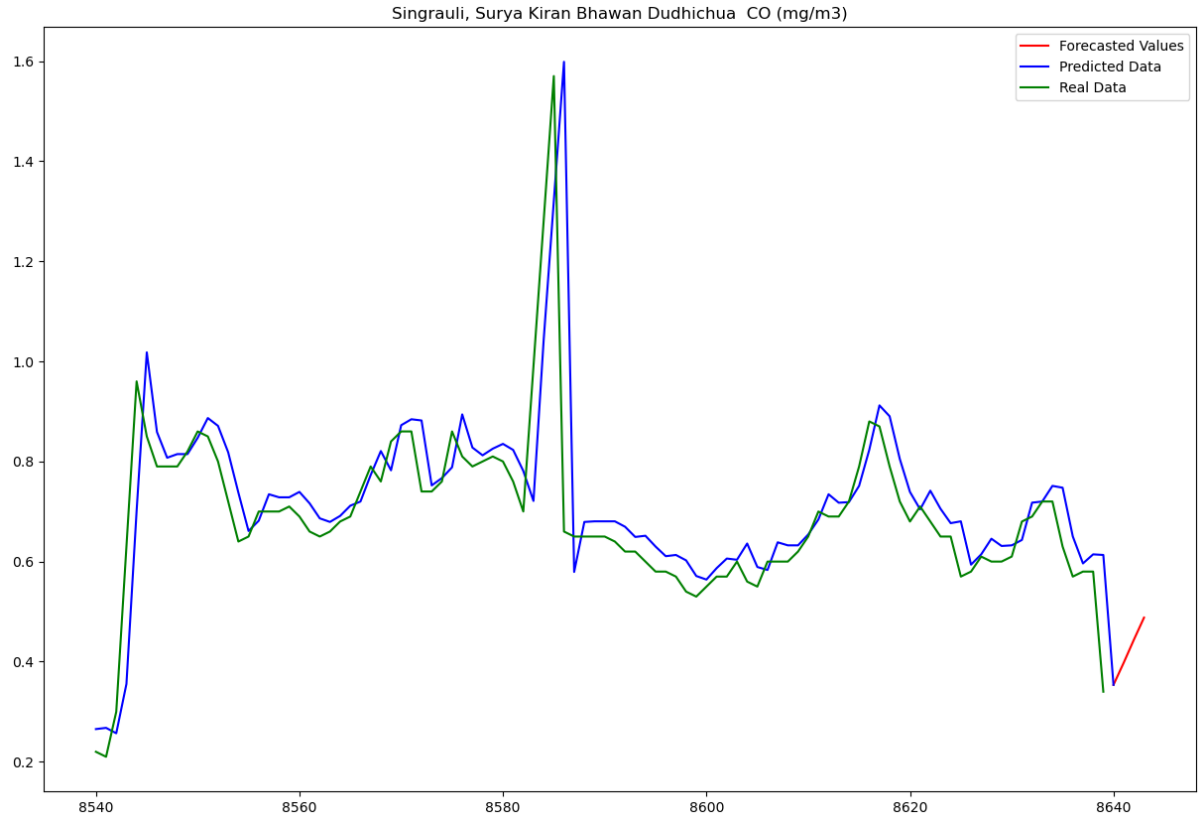
AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	3197.737
Method:	Conditional MLE	S.D. of innovations	0.167
Date:	Mon, 26 Jun 2023	AIC	-6387.475
Time:	22:17:41	BIC	-6359.219
Sample:	2	HQIC	-6377.840
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0568	0.004	12.893	0.000	0.048	0.065
Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3).L1	1.0812	0.011	101.239	0.000	1.060	1.102
Singrauli, Surya Kiran Bhawan Dudhichua CO (mg/m3).L2	-0.1217	0.011	-11.393	0.000	-0.143	-0.101

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0486	+0.0000j	1.0486	0.0000
AR.2	7.8385	+0.0000j	7.8385	0.0000



Mean: 1.401927

Root Mean Squared Error: 0.121037

For SO2:

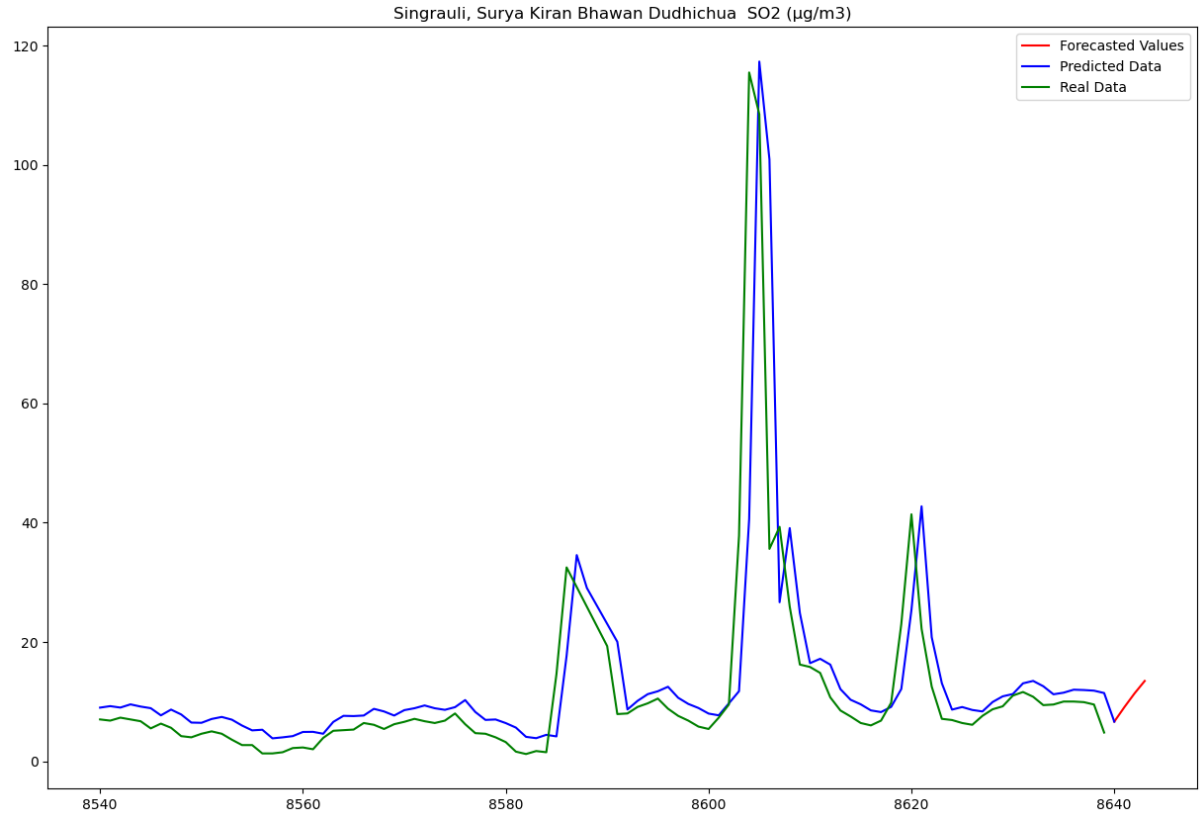
AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	-35690.757
Method:	Conditional MLE	S.D. of innovations	15.073
Date:	Mon, 26 Jun 2023	AIC	71389.514
Time:	22:17:41	BIC	71417.770
Sample:	2	HQIC	71399.148
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.8112	0.211	13.308	0.000	2.397	3.225
Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3).L1	1.0302	0.011	96.426	0.000	1.009	1.051
Singrauli, Surya Kiran Bhawan Dudhichua SO2 (µg/m3).L2	-0.1183	0.011	-11.069	0.000	-0.139	-0.097

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1128	+0.0000j	1.1128	0.0000
AR.2	7.5983	+0.0000j	7.5983	0.0000



Mean: 31.923270

Root Mean Squared Error: 11.393467

For NH3:

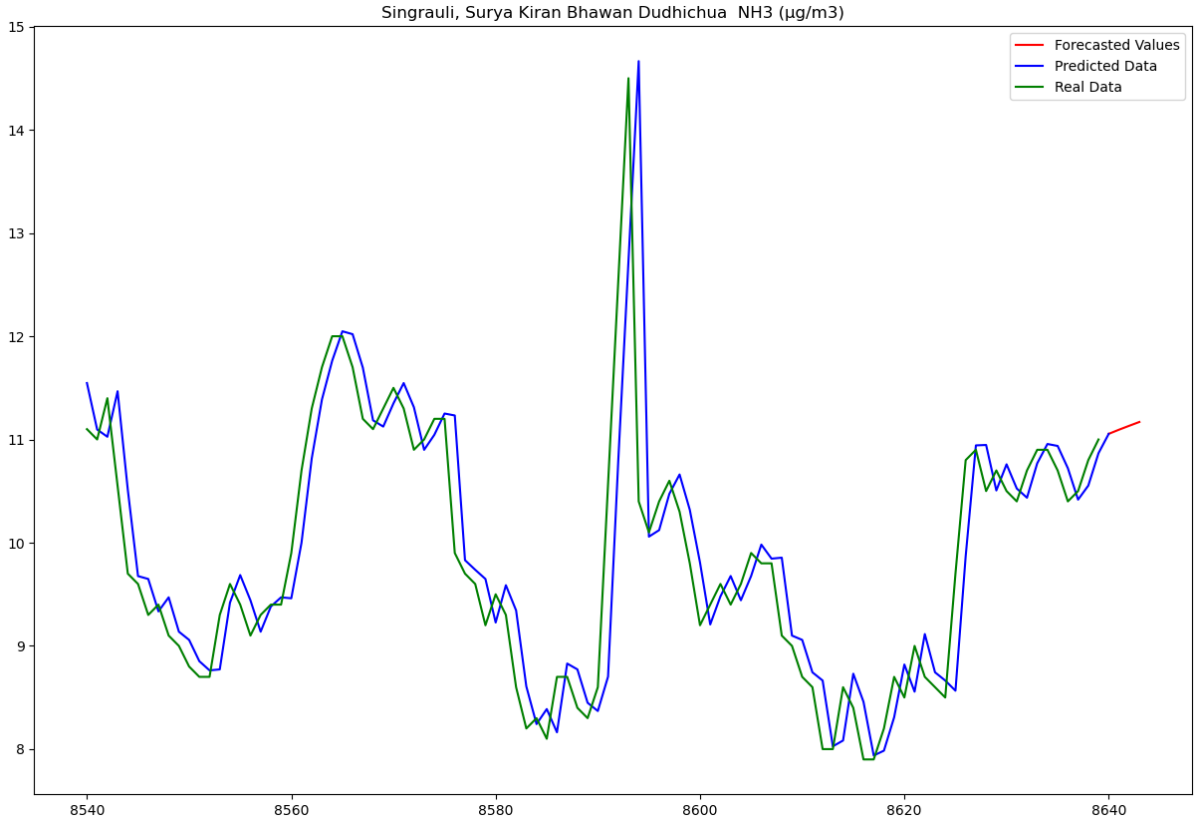
AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	-12626.145
Method:	Conditional MLE	S.D. of innovations	1.044
Date:	Mon, 26 Jun 2023	AIC	25260.290
Time:	22:17:41	BIC	25288.546
Sample:	2	HQIC	25269.925
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2109	0.027	7.888	0.000	0.159	0.263
Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3).L1	1.0784	0.011	100.676	0.000	1.057	1.099
Singrauli, Surya Kiran Bhawan Dudhichua NH3 (µg/m3).L2	-0.0943	0.011	-8.805	0.000	-0.115	-0.073

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0179	+0.0000j	1.0179	0.0000
AR.2	10.4164	+0.0000j	10.4164	0.0000



Mean: 13.286956

Root Mean Squared Error: 0.650620

For Ozone:

AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	-26584.645
Method:	Conditional MLE	S.D. of innovations	5.252
Date:	Mon, 26 Jun 2023	AIC	53177.289
Time:	22:17:41	BIC	53205.545
Sample:	2	HQIC	53186.923
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.0079	0.093	10.785	0.000	0.825	1.191
Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3).L1	1.2963	0.010	127.393	0.000	1.276	1.316
Singrauli, Surya Kiran Bhawan Dudhichua Ozone (µg/m3).L2	-0.3250	0.010	-31.934	0.000	-0.345	-0.305

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0453	+0.0000j	1.0453	0.0000
AR.2	2.9440	+0.0000j	2.9440	0.0000

Mean: 35.193970

Root Mean Squared Error: 4.775819

For Benzene:

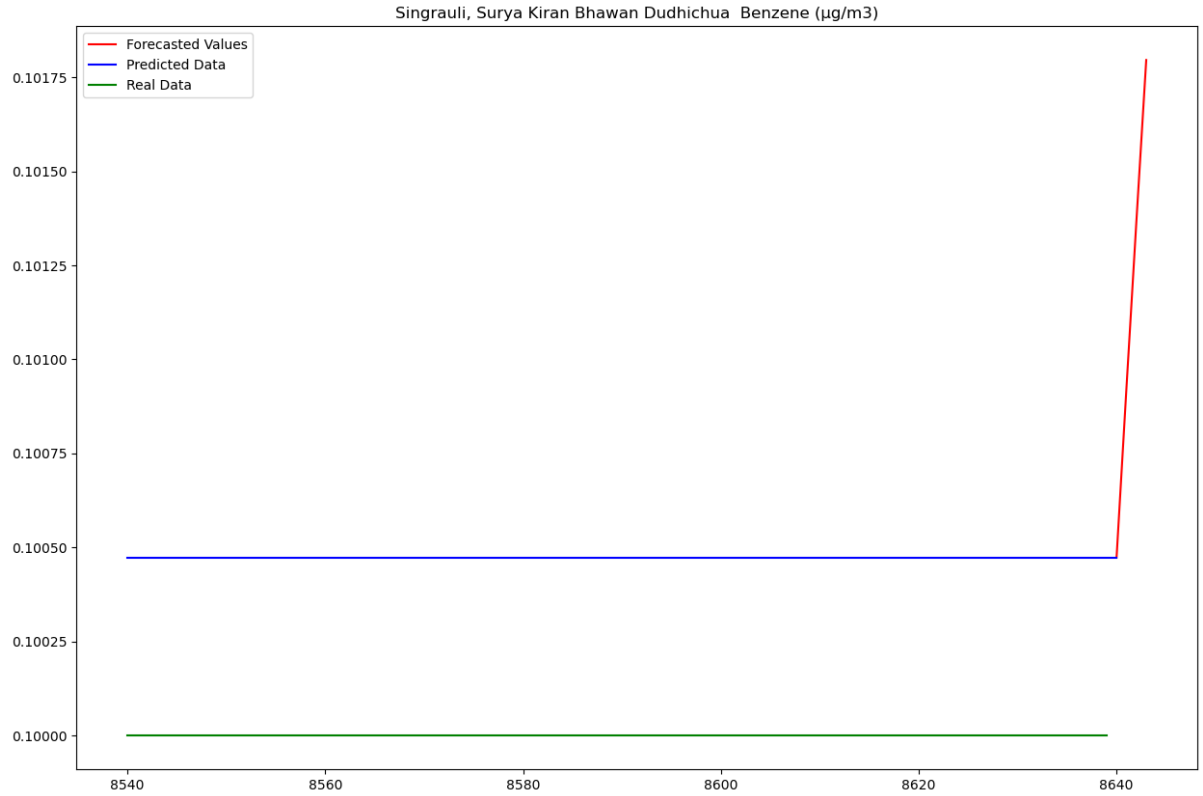
AutoReg Model Results

Dep. Variable:	Singrauli, Surya Kiran Bhawan Dudhichua Benzene (µg/m3)	No. Observations:	8640
Model:	AutoReg(2)	Log Likelihood	25041.140
Method:	Conditional MLE	S.D. of innovations	0.013
Date:	Mon, 26 Jun 2023	AIC	-50074.281
Time:	22:17:42	BIC	-50046.025
Sample:	2	HQIC	-50064.646
	8640		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0028	0.000	8.906	0.000	0.002	0.003
Singrauli, Surya Kiran Bhawan Dudhichua Benzene (µg/m3).L1	0.9515	0.011	88.459	0.000	0.930	0.973
Singrauli, Surya Kiran Bhawan Dudhichua Benzene (µg/m3).L2	0.0254	0.011	2.361	0.018	0.004	0.046

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0231	+0.0000j	1.0231	0.0000
AR.2	-38.5316	+0.0000j	38.5316	0.5000



Mean: 0.122002

Root Mean Squared Error: 0.000474

GOODNESS OF A MODEL

To assess the goodness-of-fit of an Auto Regressive (AR) model, you can analyze the model report or summary. Here are some key indicators to consider:

Coefficient significance: Examine the p-values associated with the coefficients of the lagged terms in the AR model. Lower p-values (typically less than 0.05) indicate greater statistical significance. Significant coefficients suggest that the corresponding lagged terms have a meaningful impact on the current observation.

Residual analysis: Assess the residuals (i.e., the differences between the predicted values and the actual values) of the AR model. A good AR model should have residuals that exhibit the following characteristics:

- **Mean of zero:** The average of the residuals should be close to zero, indicating that the model does not consistently over- or under-predict the data.
- **Constant variance:** The spread or dispersion of the residuals should be roughly consistent across different values of the predicted variable.
- **No autocorrelation:** The residuals should not display any systematic patterns or correlation with the lagged residuals, indicating that the model captures the temporal dependencies adequately.
- **Normality:** The residuals should follow a roughly normal distribution, as assessed through techniques such as a histogram or a Q-Q plot.

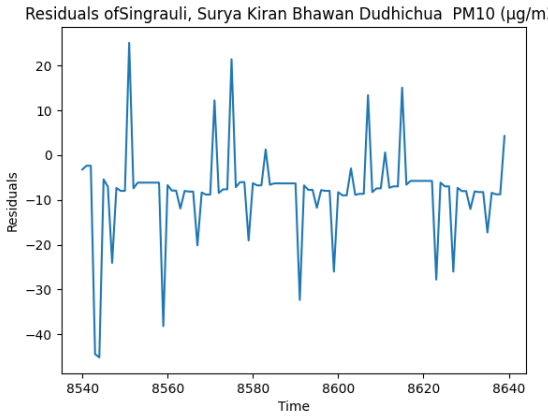
Information criteria: Consider information criteria such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). These criteria evaluate the balance between model fit and complexity, aiming to identify the model with the best trade-off. Lower AIC or BIC values indicate better model performance.

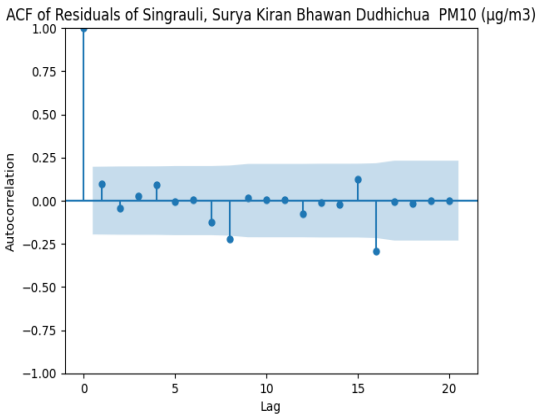
INFERENCE

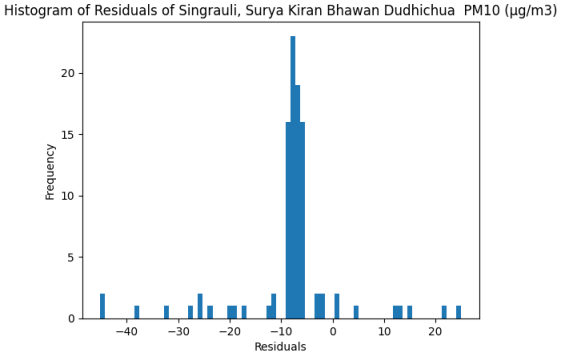
Coefficient Analysis: By examining the Auto Regressive (AR) model reports provided above, a notable observation emerges. The calculated p-values for the lagged terms are found to be less than 0.05. This indicates that the chosen lagged values indeed exert a statistically significant influence on the current observation.

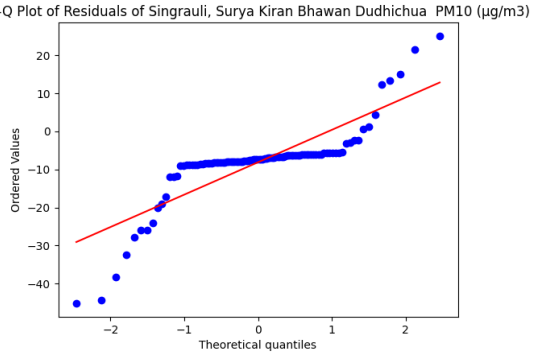
Residual Analysis:

- **For PM10:**

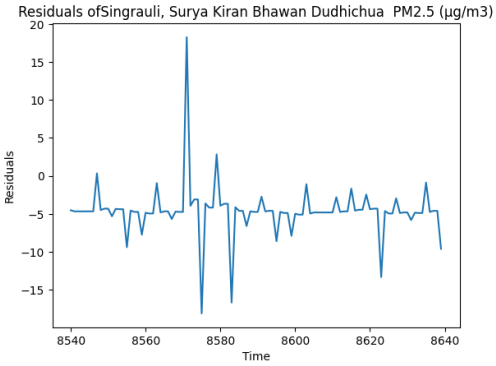


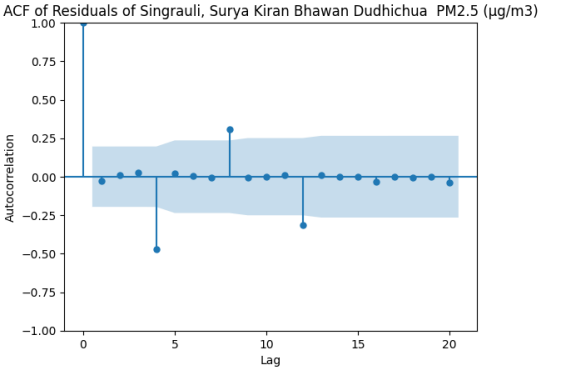


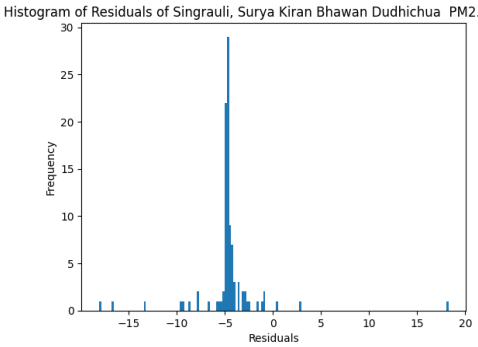


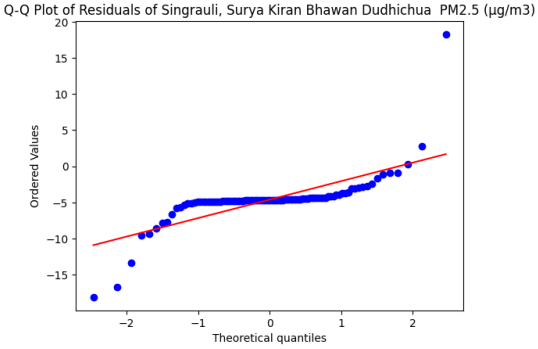


- **For PM2.5:**

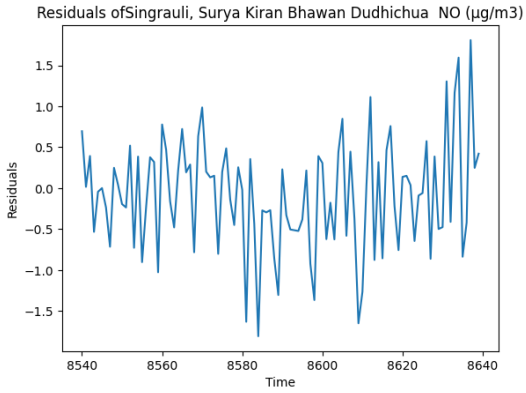


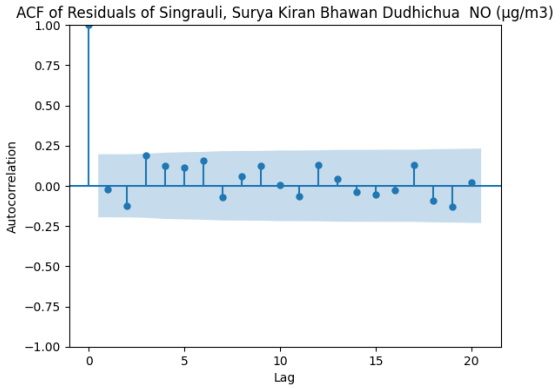


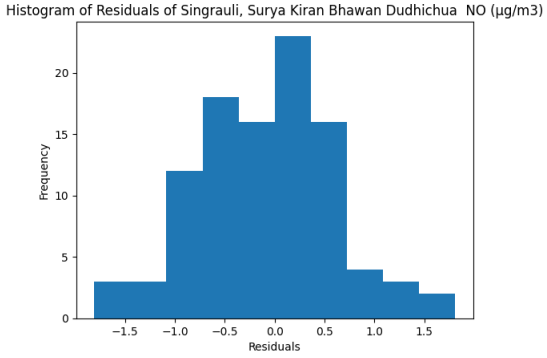


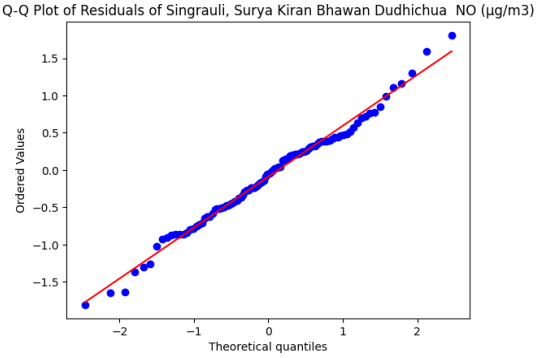


- **For NO:**

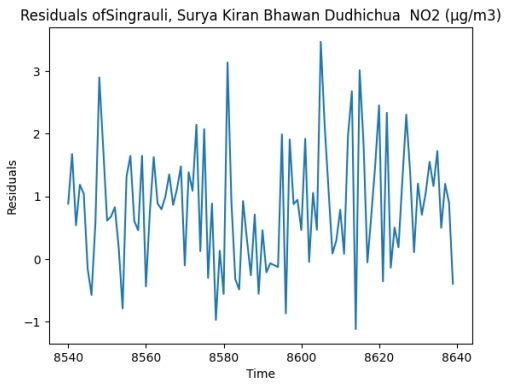


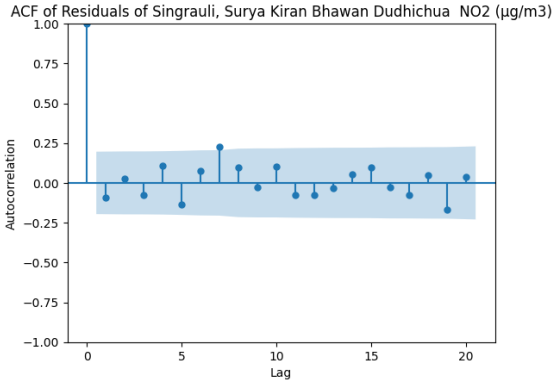


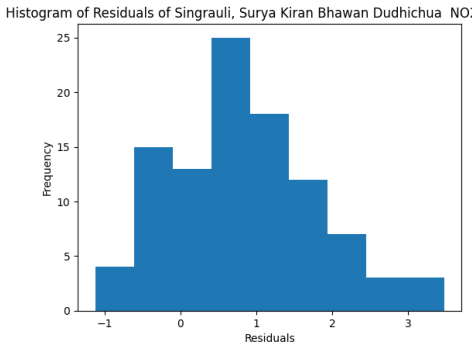


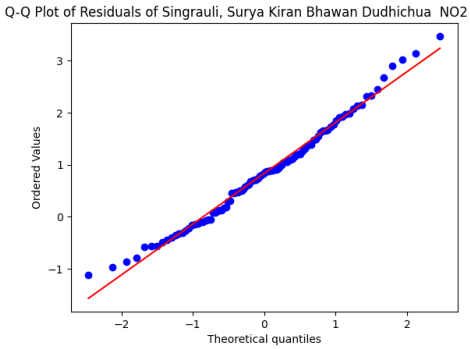


- **For NO2:**

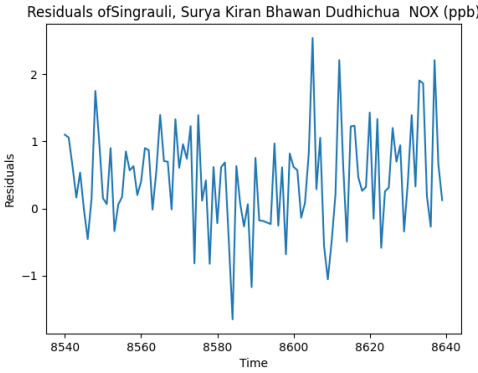


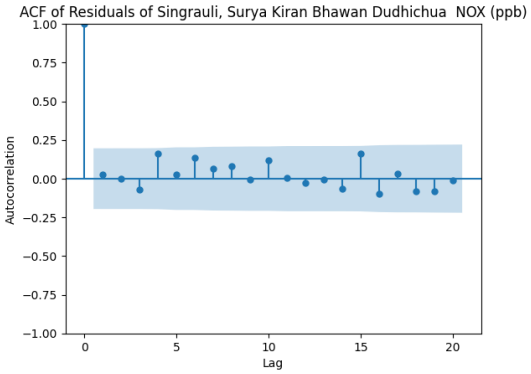


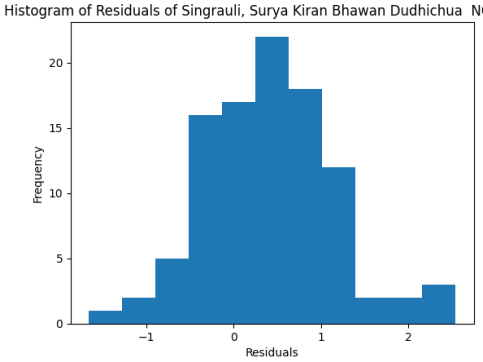


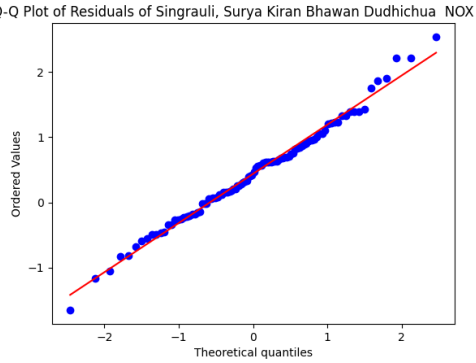


- **For NOX:**

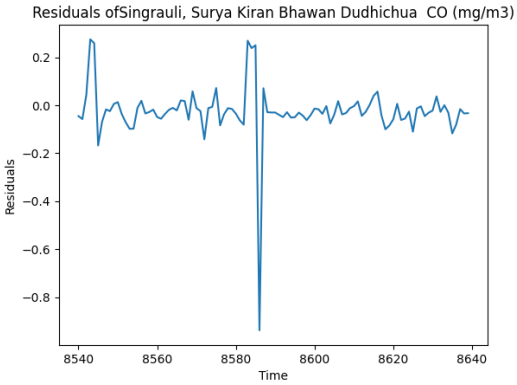


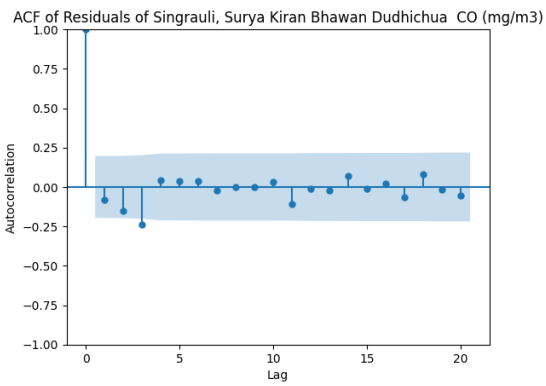


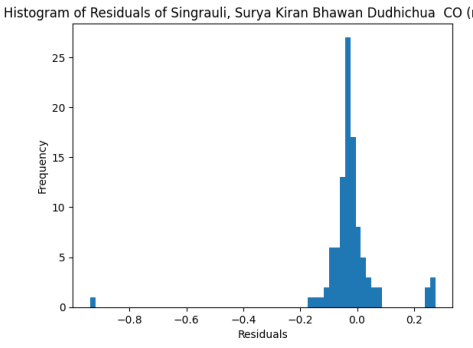


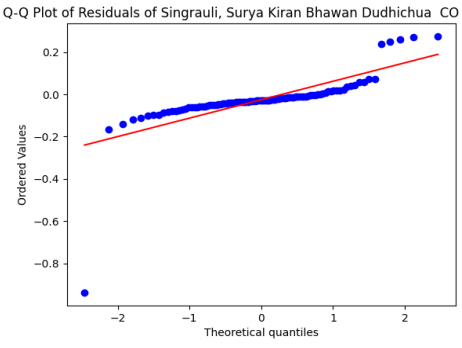


- **For CO:**

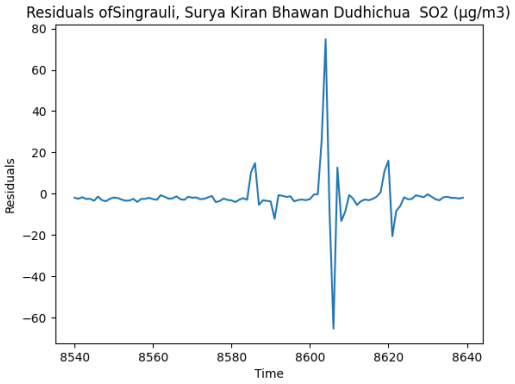


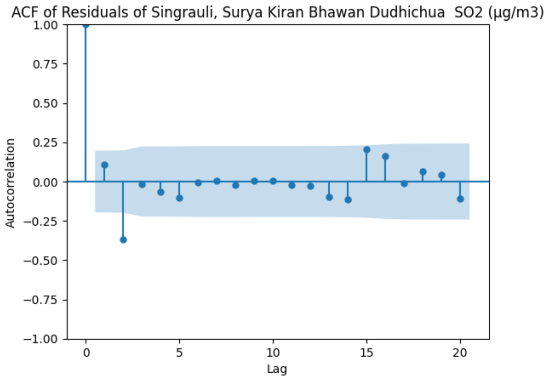


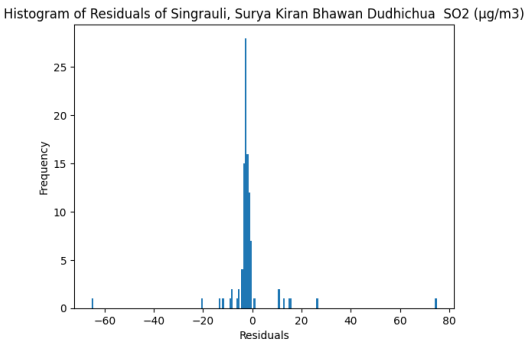


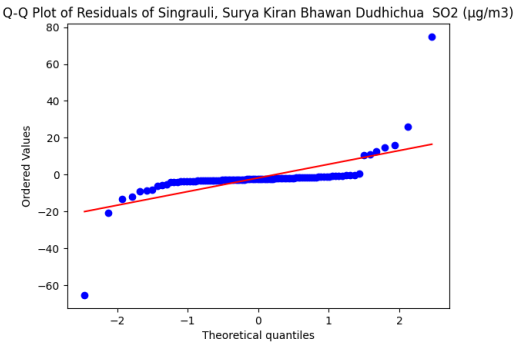


- **For SO2:**

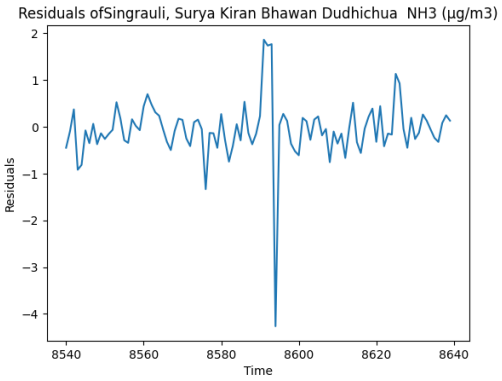


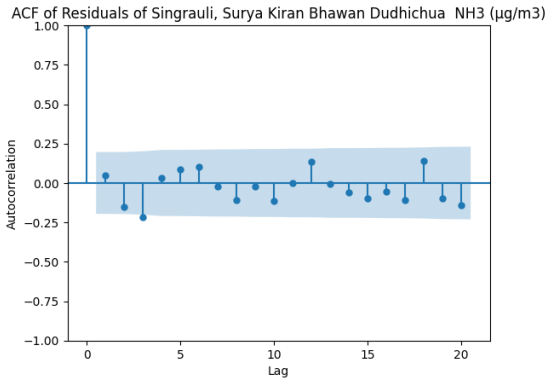


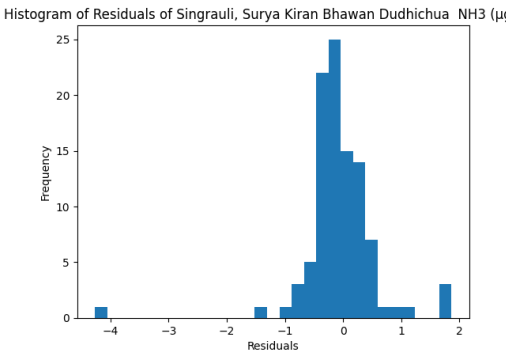


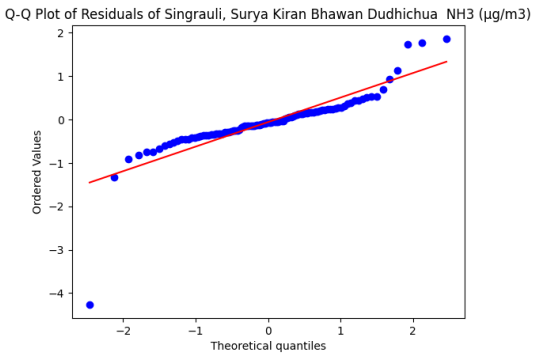


- **For NH3:**

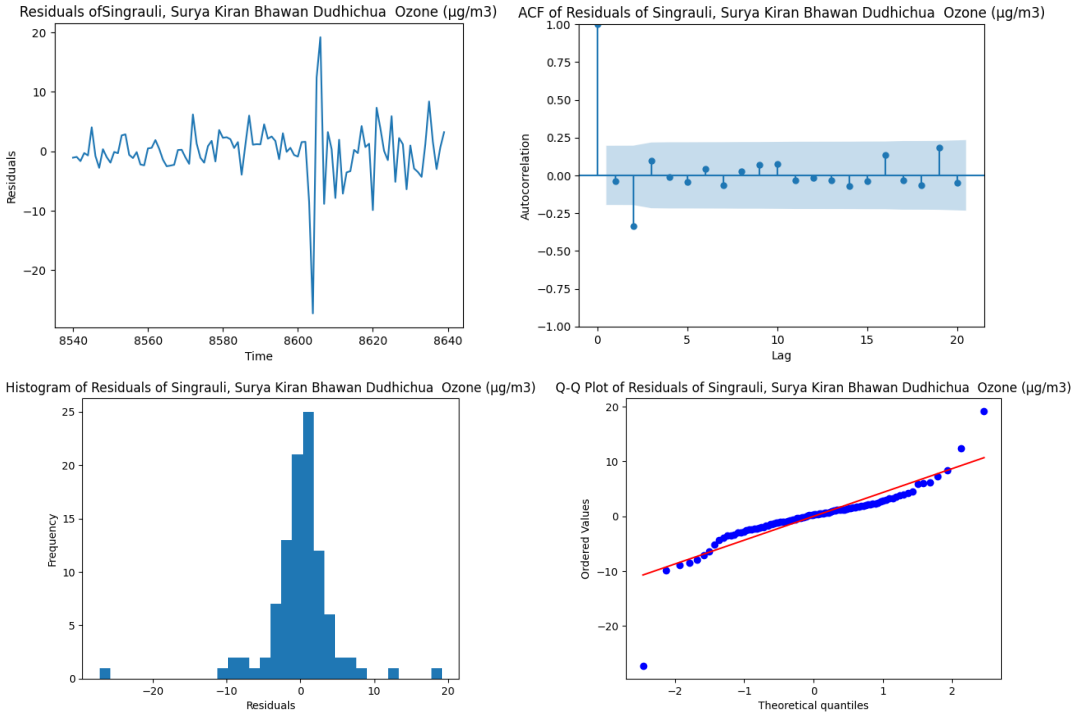








- **For Ozone:**



Based on the analysis of the above graphs:

- **Residual Graph:** The graph indicates that the mean of the residuals is in close proximity to zero, suggesting that the model captures the data's central tendency effectively.
- **Autocorrelation Graph:** The graph reveals that the residuals exhibit no significant correlation with each other, implying that the model adequately captures the temporal dependencies in the data.
- **Histogram and Q-Q Plot:** These graphs illustrate that the residuals closely align with the curve of a normal distribution, indicating that the model's residuals follow the expected distribution.

Considering these graphical representations and observations, it can be confidently concluded that the model performs well and provides a good fit for the data.

Information Criteria: Based on the AutoRegressive (AR) model report, compelling evidence emerges to support the effectiveness of the model:

- **AIC and BIC**: Both AIC and BIC exhibit exceptionally low values, indicating a superior model performance. This suggests that the AR model outperforms alternative models when considering the trade-off between goodness-of-fit and complexity.
- **Log likelihood value:** The log likelihood value is notably high, indicating a strong fit between the AR model and the observed data. This supports the notion that the model accurately captures the underlying patterns and dynamics present in the dataset.

Considering these inferences, it is evident that the AR model stands out as a superior choice among other models. The combination of low AIC and BIC values, along with a high log likelihood value, underscores the model's favorable performance.