**Assignment 1- EE 798Q**

**Name**- Rahul Singh

**Roll Number**- 210807

**Problem Statement**- Analyse the provided time-series data on India's coal production and its environmental impact, with a focus on open-pit mining. The objective is to conduct a comprehensive statistical analysis to gain insights into the descriptive, exploratory, and inferential aspects of the data.

The data will be used to address the following key questions:

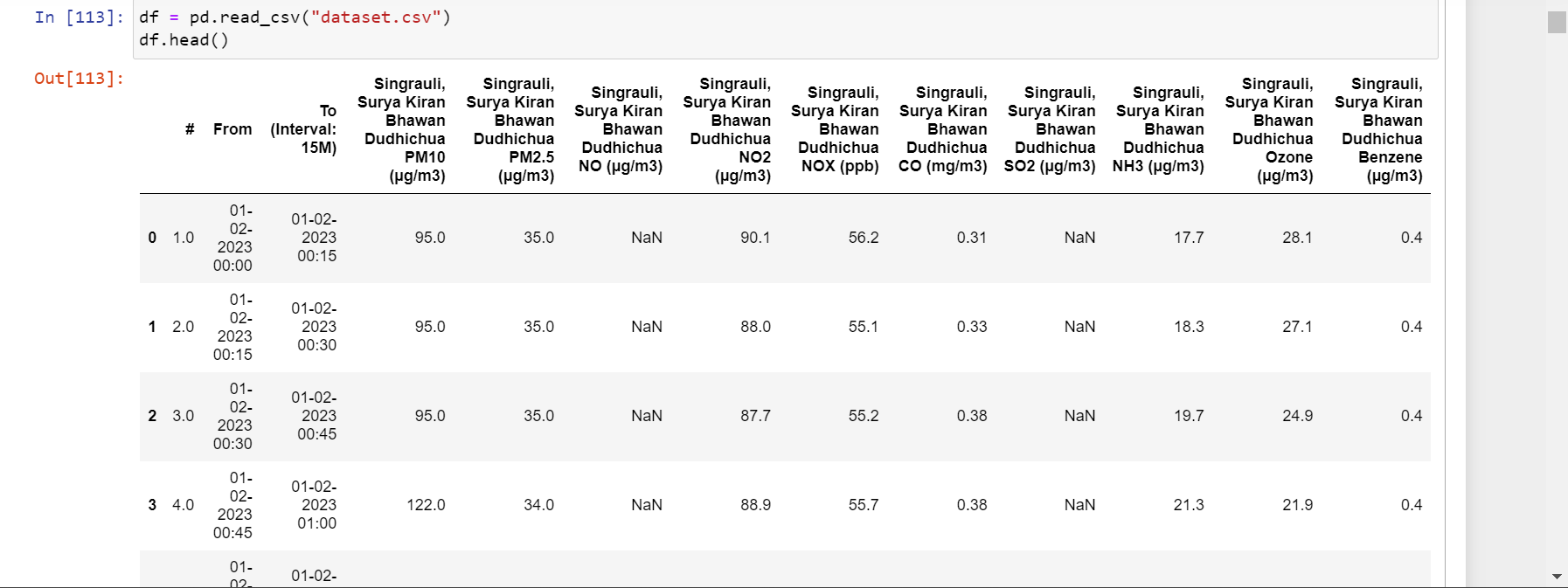
* **Preprocessing:** Due to reasons like sensor failure, sensor-to-central-hub communication link failure, data packet loss etc., there will be some missing sensory data for certain duration of the time. With what values can we replace the NA values? What will be the approach for choosing those values?
* **Identifying distribution Pattern-** How to plot the histogram of this blast trigger times across all months of data. What kind of distribution it is following? Check from QQ plot whether is its normal distribution or not? What will be the frequency, central tendency, dispersion?
* **Classification & Exploratory analysis:** Classify the time-series into mainly two categories (1) **Stock** time series. (2) **Flow** time series. What are the inferences that we can draw from each of them? Highlight the main characteristics of the time series air pollution, usually in a visual format.
* **Descriptive analysis**: What are the patterns in time series data at the time of coal India open-pit blasting effect, in coal India blasting effect time is 13:45 pm to 14:45 pm? Attempt to understand the air pollution data and the relationships within it. What is the effect on air pollution at the time of blasting? Find trends like cycles, seasonal variation.
* **Modelling**: Analyse the time series methods used for forecasting like Autoregression (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA). Generate autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Check the presence of autocorrelation and determine the order of autoregressive (AR) and moving average (MA) components in a time series model.
* **Train and Testing the model & Forecasting**: Estimate the parameters to minimize the difference between the observed data and the predicted values to find the best fir curve. Predict the future data. How can we use historical trends of air pollution data set to predict future data? How will we be using the historical data as a model for future data, predicting scenarios that could happen along future plot points? What model does our time series data follow?

In this assignment, I have used different python libraries for data analysis. I will be providing code snippets, graphs and inference drawn from the data and graphs.

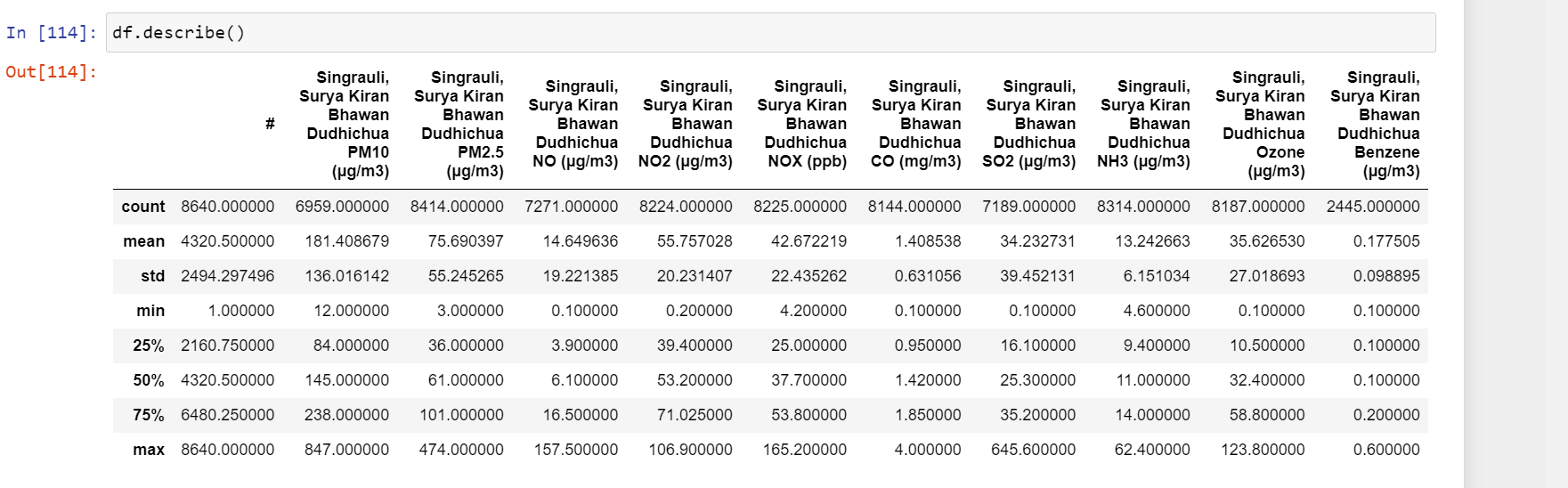
* **Preprocessing-**
* Importing various libraries-



* Reading the dataset-

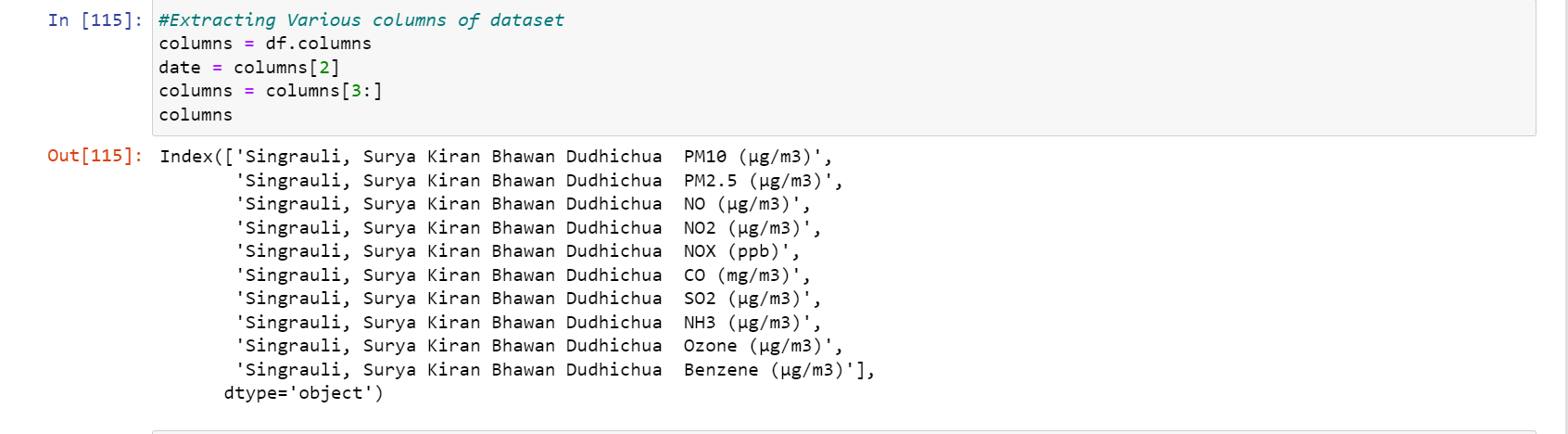


* As we can observe, some datapoints are missing in our dataset. To replace nan values, we are going to study the frequency distribution of various pollutants.

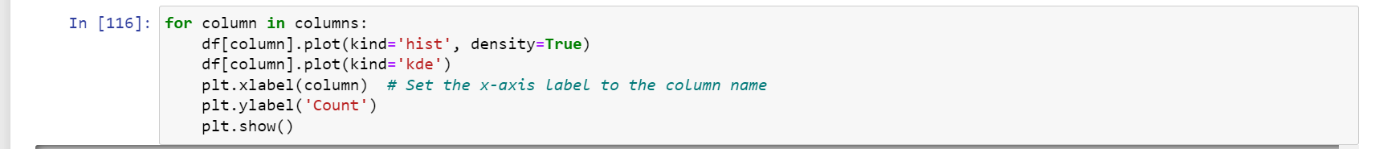


Descriptive statistics of the numerical columns in the Data Frame before data preprocessing>

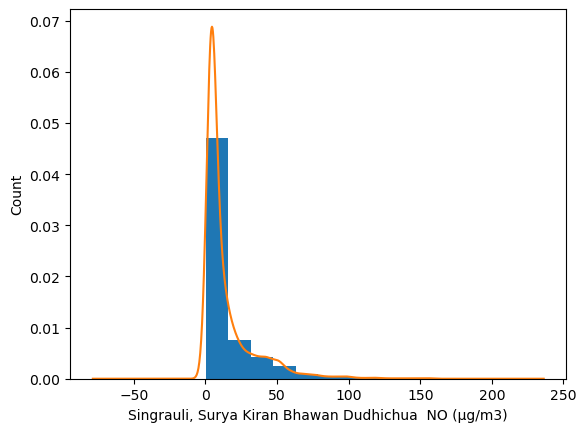
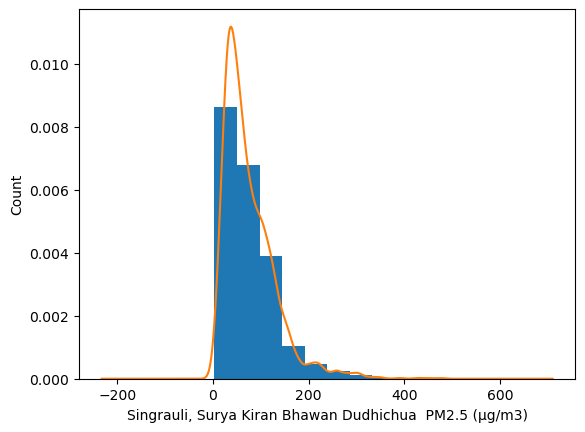
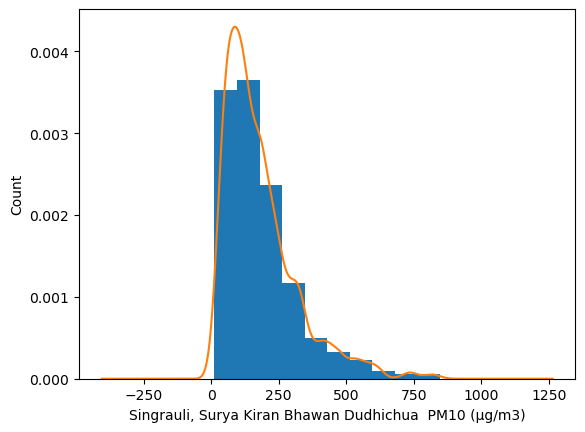
* Count: The number of non-null values in each column.
* Mean: The average value of each column.
* Standard Deviation: The measure of the spread or dispersion of values in each column.
* Minimum: The minimum value in each column.
* 25th Percentile (or First Quartile): The value below which 25% of the data falls.
* 50th Percentile (or Second Quartile or Median): The value below which 50% of the data falls.
* 75th Percentile (or Third Quartile): The value below which 75% of the data falls.
* Maximum: The maximum value in each column.
* For this we will first extract the columns.

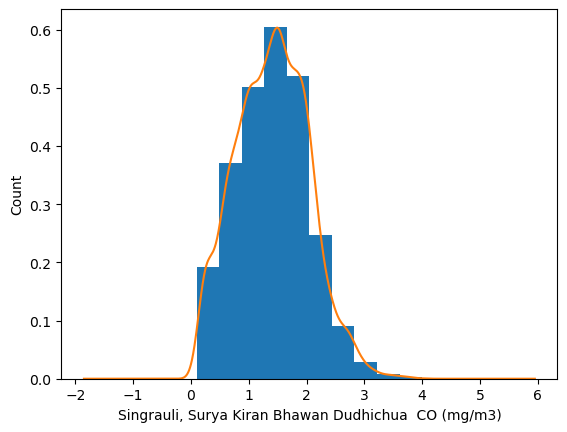
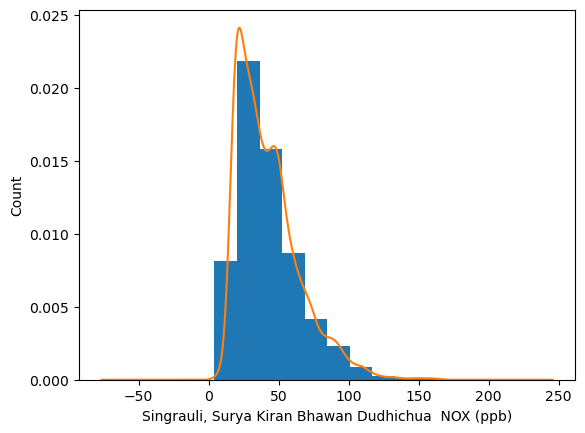
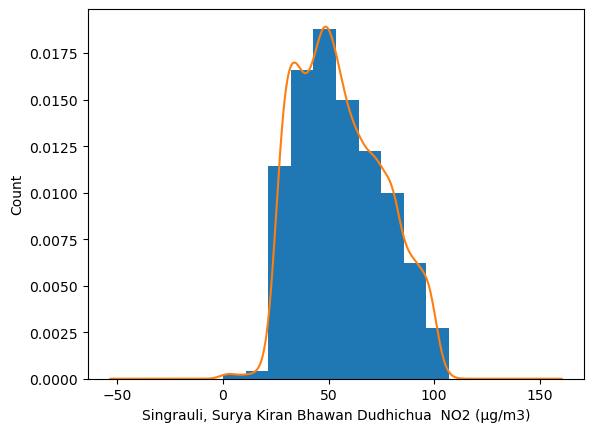


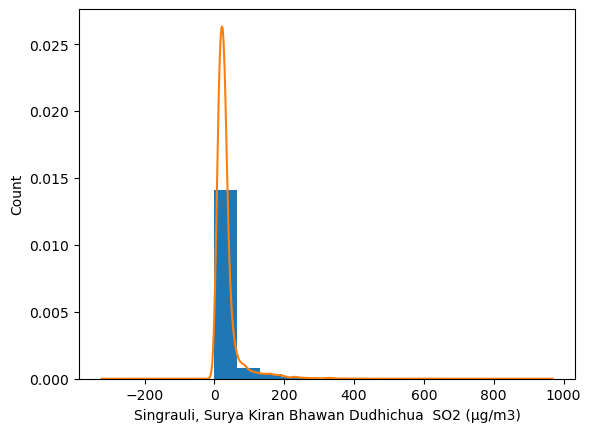
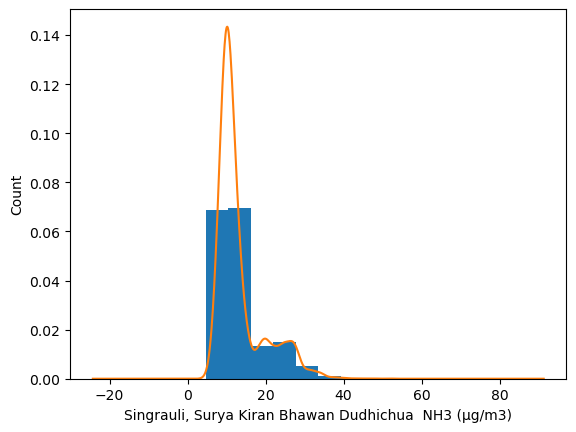
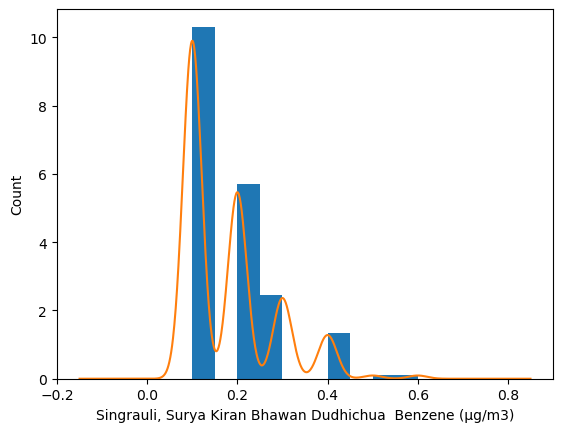
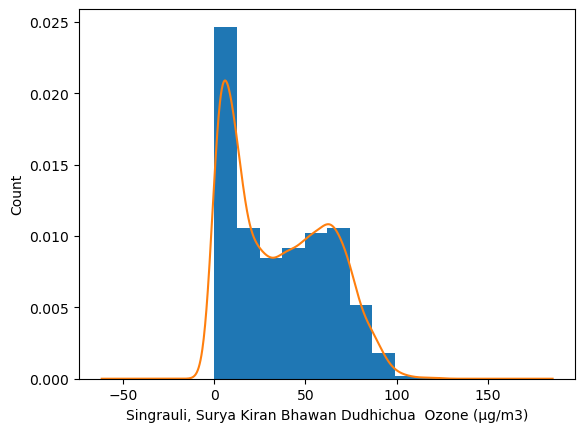
* Then, we will plot their frequency distribution using the following function->



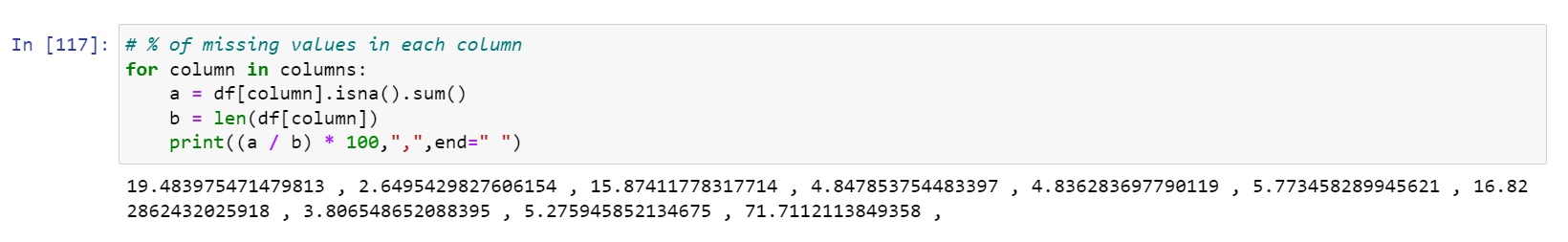
* Frequency-distribution before data-preprocessing-



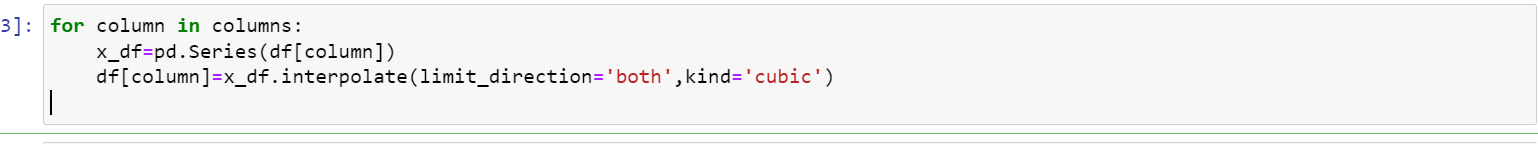


* We observe here that-
* Graphs of CO, NO2, NOX count distributions are approximately symmetric.
* PM10, PM2.5, NO, SO2, NH3, Ozone, Benzene count distributions are right skewed.
* The following function returns the percentage of NA values in each column-

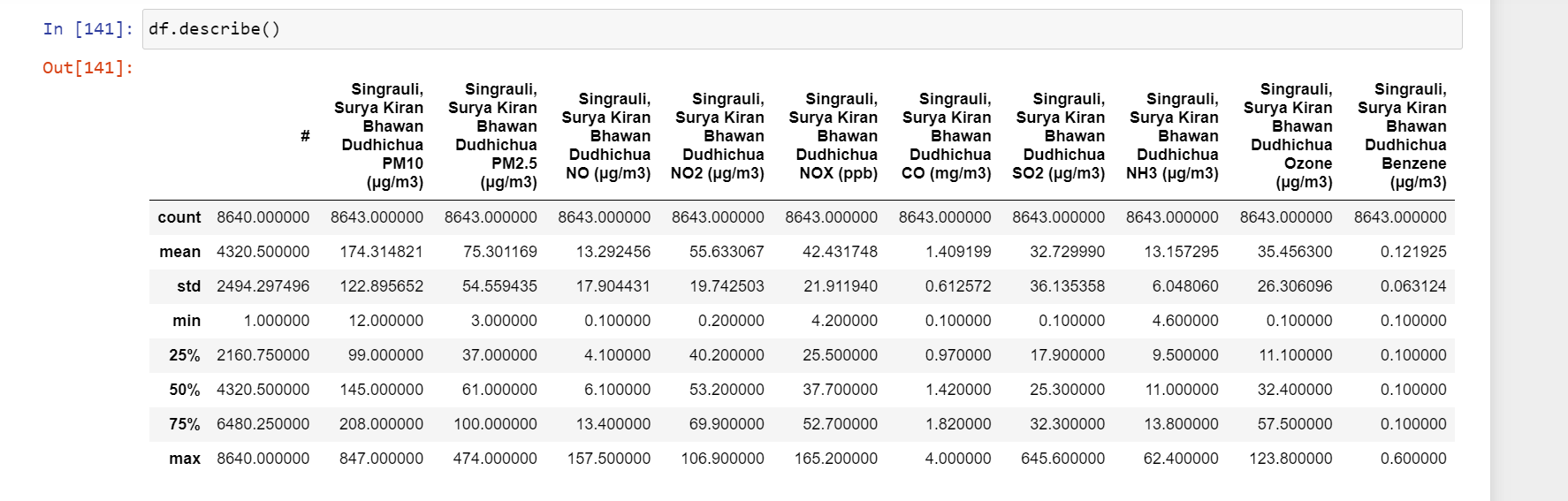


* Since most of columns have skewed distribution, using the mean to fill NA values may not be right approach to fill the NA values.
* Here we are filling NA values using cubic interpolation. This ensures that the summary statistics, such as the mean and standard deviation, are less affected compared to filling with other values like zero or the mean. This helps to maintain the integrity of the statistical measures.

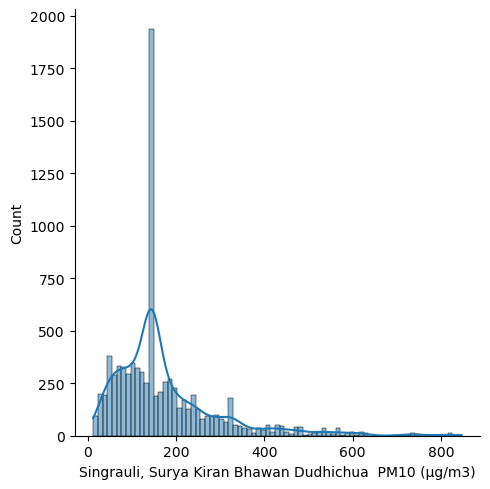
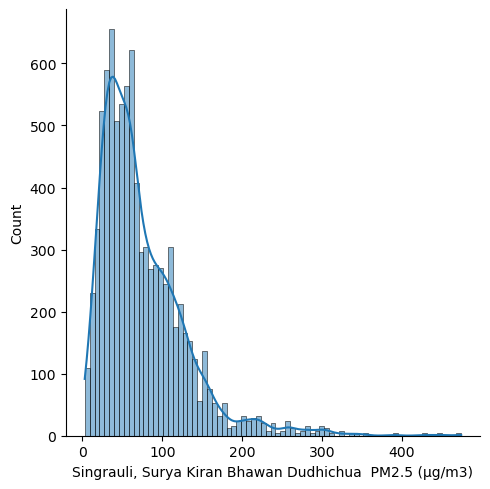
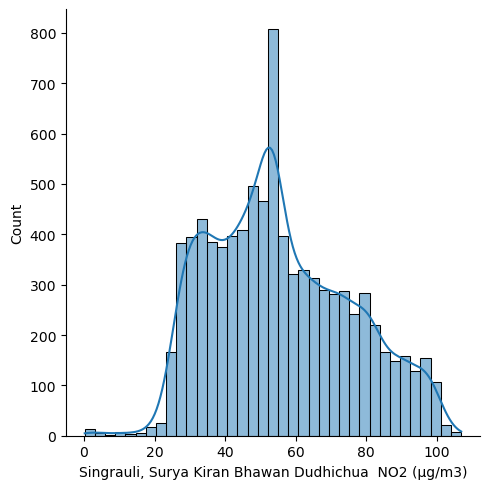
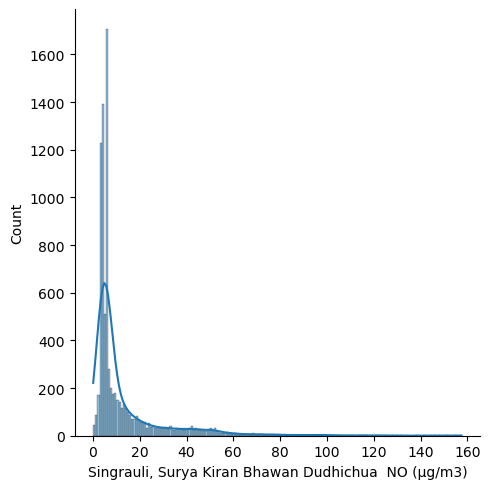
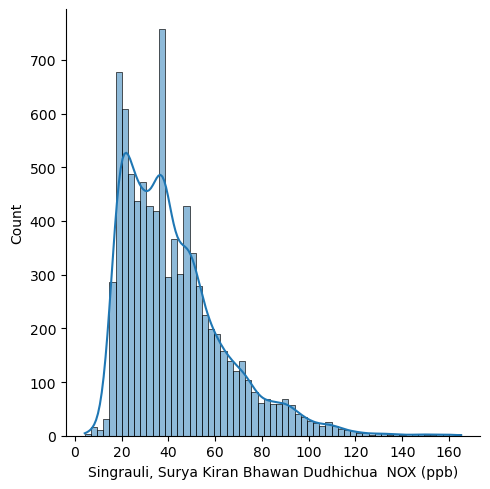
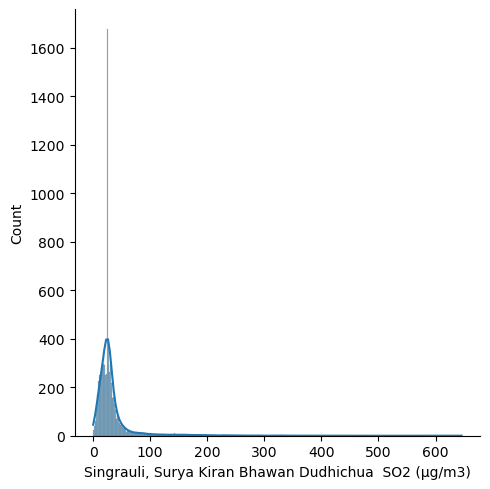
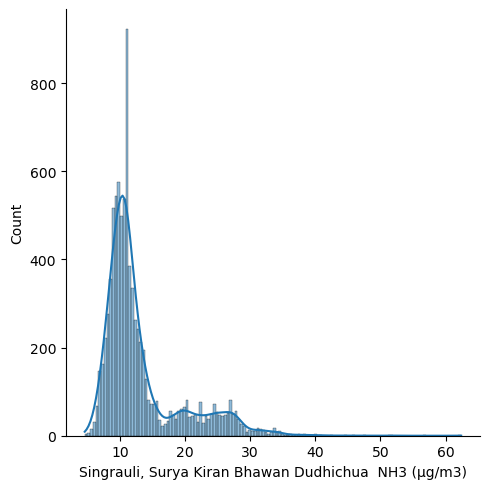
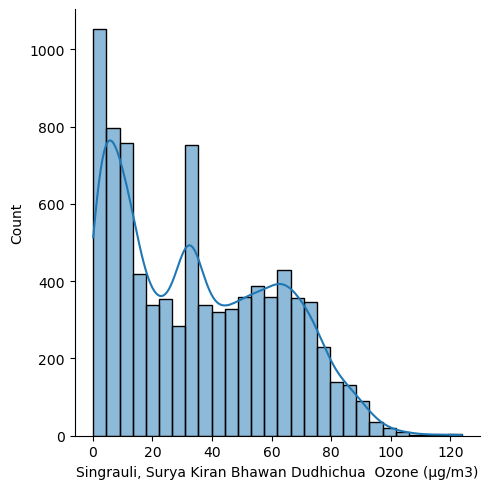
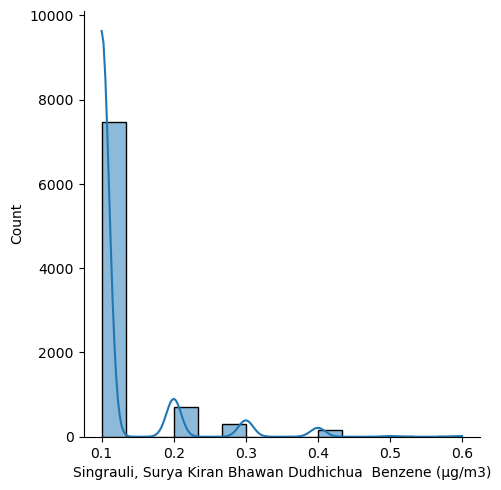




* As we can observe, the NA values are replaced interpolated values in each column.
* **Identifying Distribution Pattern-**
* Describing data after preprocessing-



* Frequency Distribution of time series after preprocessing-

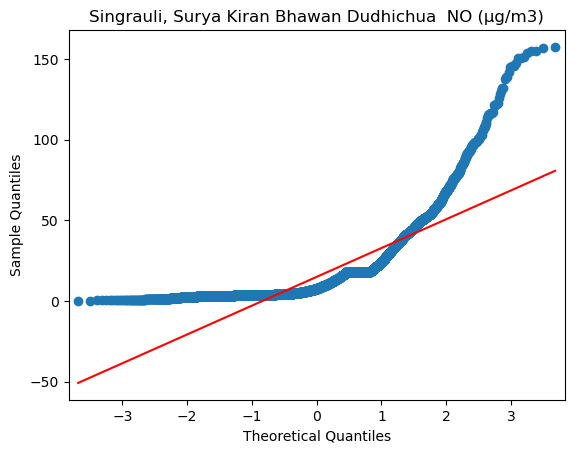
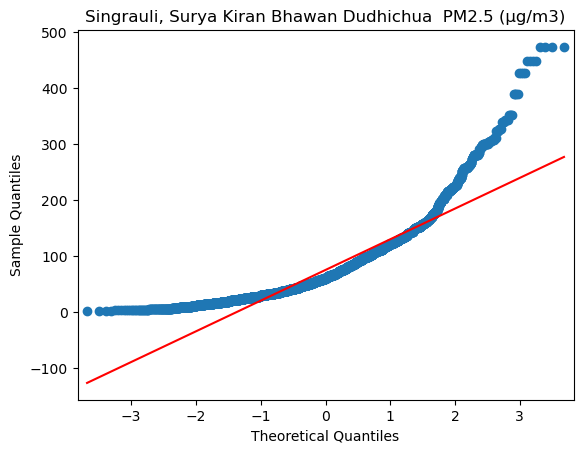
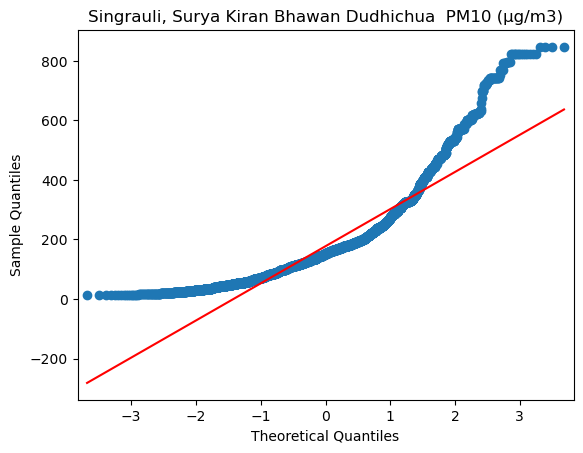
        

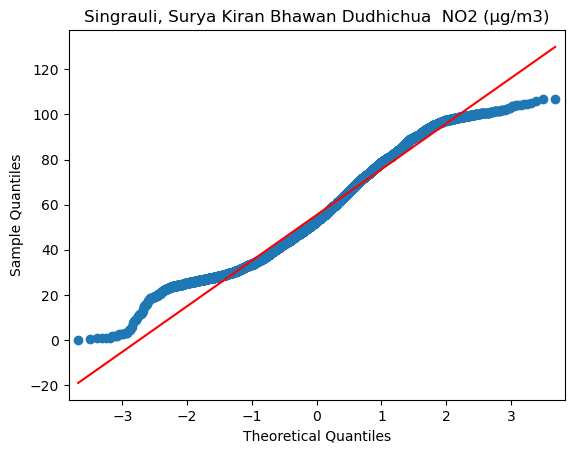
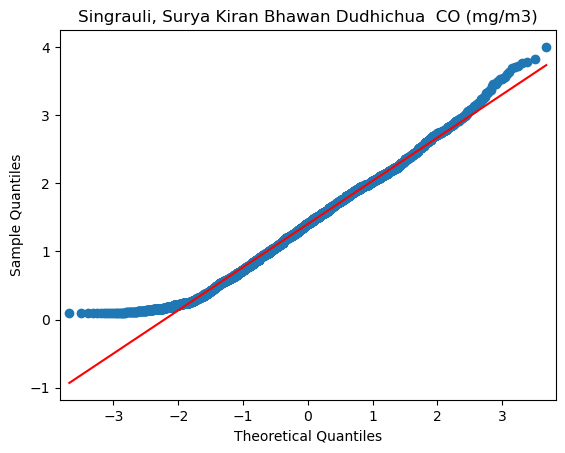
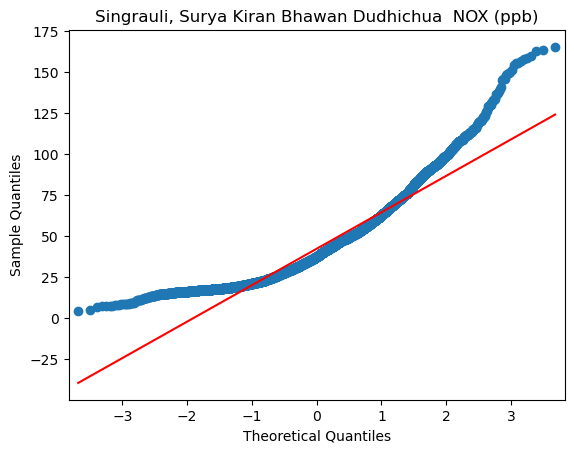
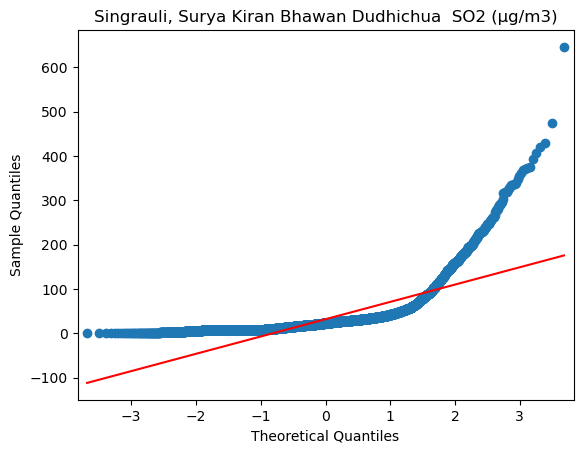
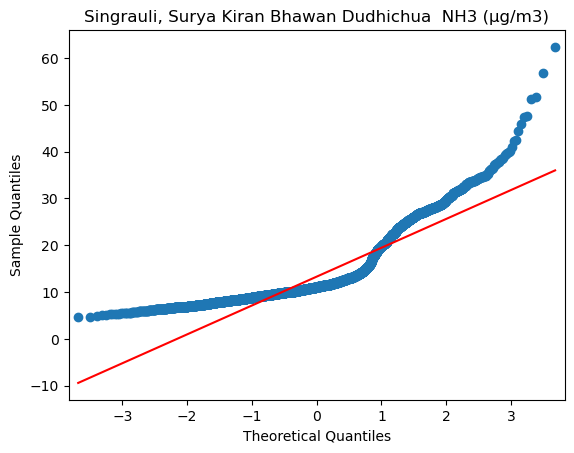
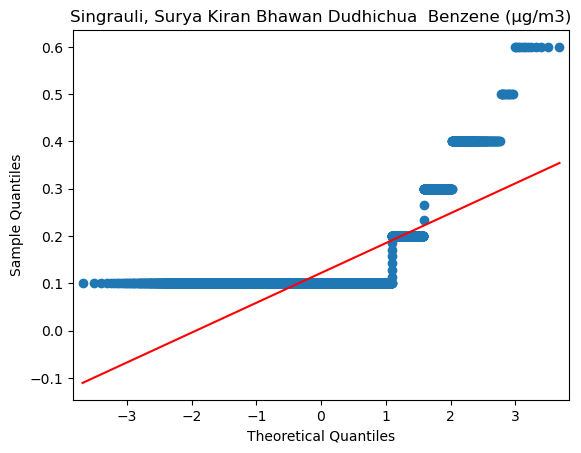
* Inference drawn from these distributions-
* **Central Tendency**- Since most of the columns are right skewed, the central tendency of our time-series data of each column is **median.** In a right-skewed frequency distribution, the central tendency is typically represented by the median. The median is the value that divides the distribution into two equal halves, with 50% of the data points below it and 50% above it.
* **Dispersion**- Nearly all the frequency distributions here are right-skewed distribution, the data tends to have a tail that extends towards the right side. This means that the majority of the data points are concentrated on the left side of the distribution, with a few extreme values pulling the tail towards the right.
* **Distribution pattern of the time series data-**

#### In the following code, the sm.qqplot() function is used to generate QQ plots for each column in the DataFrame. The line='s' argument specifies that a standardized line should be drawn on the plot, representing the theoretical quantiles of the standard normal distribution.





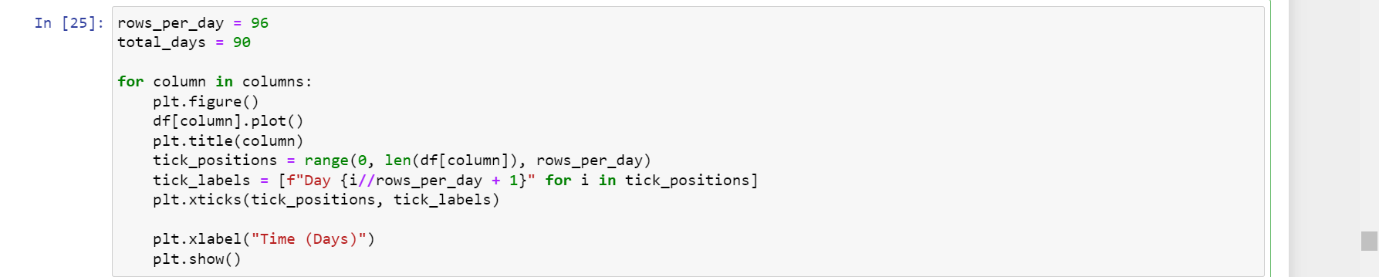


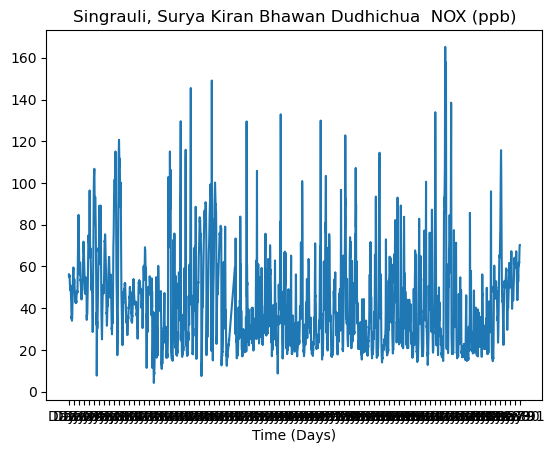
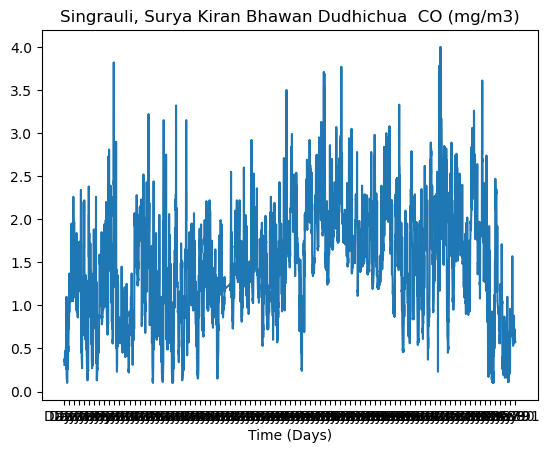
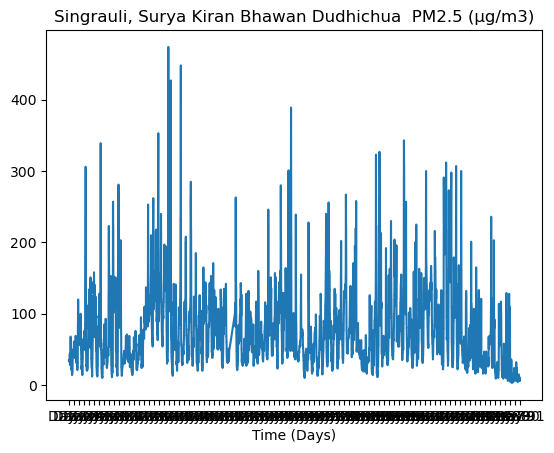
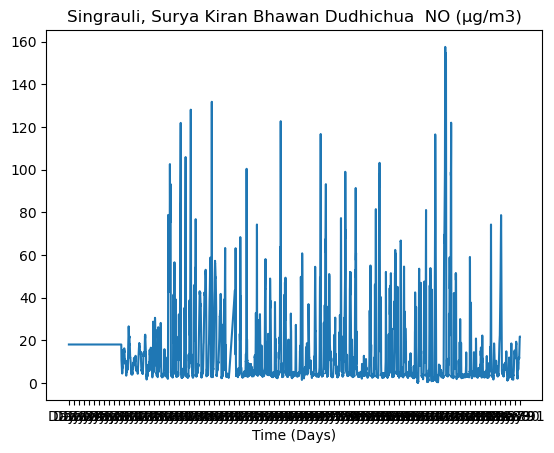
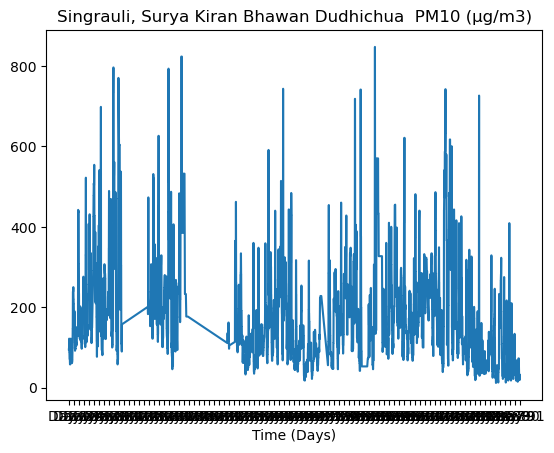
     

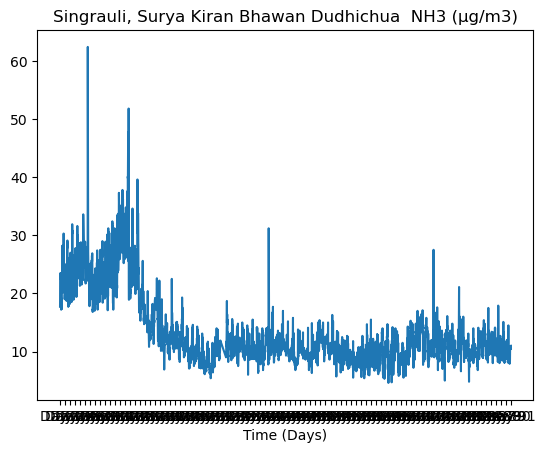
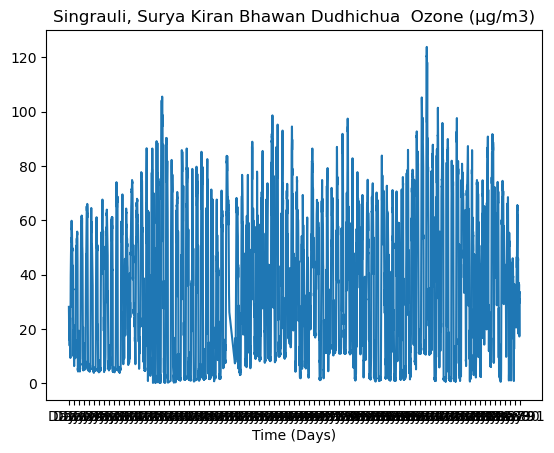
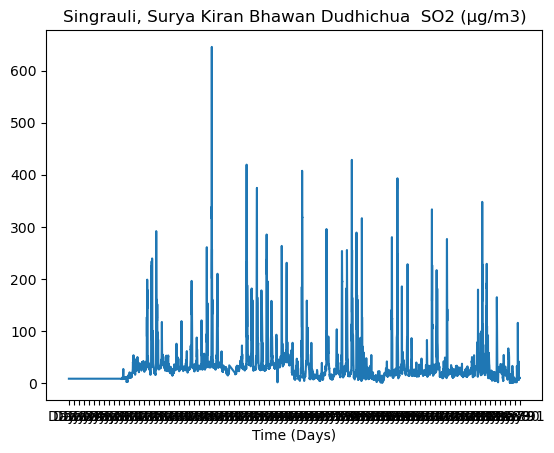
* Inference drawn from QQ plot of the time series data of pollutants-
* From the above graphs we infer that nearly all the graphs are **positively skewed** and hence they do not follow normal distribution.
* **Classification & Exploratory analysis-**

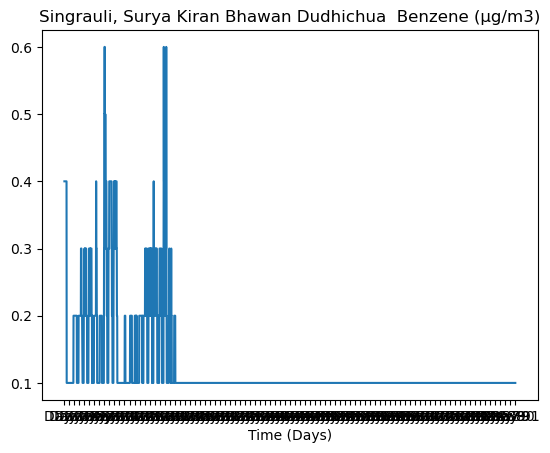
**Stock Time Series Analysis-**

* Using the below code-snippet, I did the plotting of **stock time-series** to check the level of every pollutant in air on each day for entire 90 days.









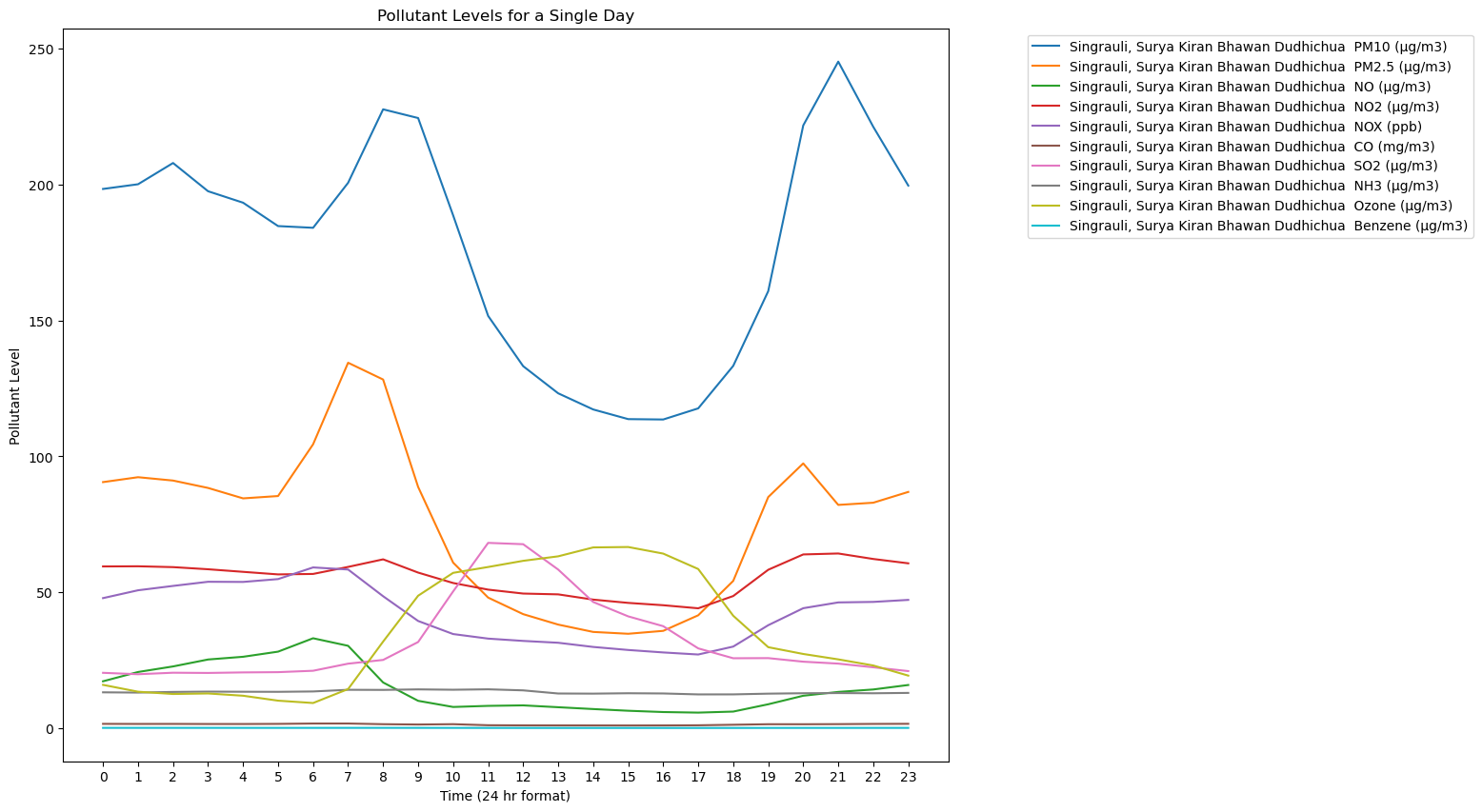
* Inference drawn from stock-time series-
* After carefully examining the stock time series data of pollutants over a 90-day interval, I have not been able to identify any significant inferences.

**Flow time series analysis-**

* Analysis of pollution level of pollutants throughout a day
* **Strategy**- I have plotted the time series data of each pollutant on an average day by taking average of time series data of 90 days.



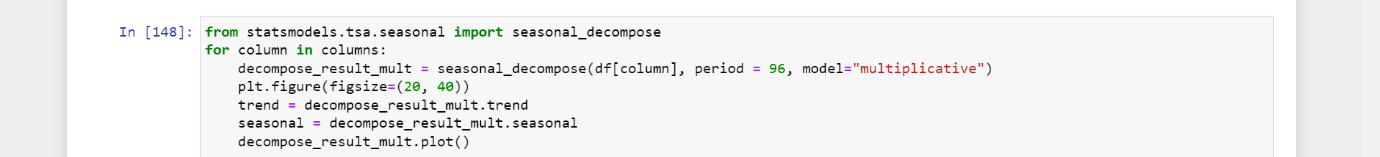
* The above code calculates the level of pollutant by taking average of entire 90 days. Then it plots the graph in which each pollutant level is shown with different coloured line though out a day.



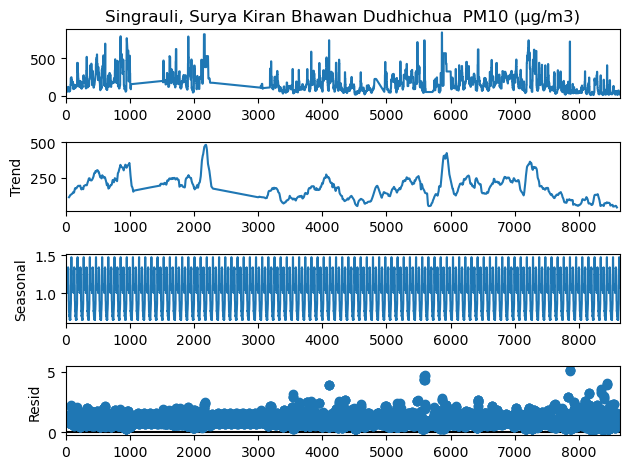
* **Inferences Drawn from flow time-series**- After the open-pit blasting at 14:45, there is a gradual increase in the level of pollutants in the air, which eventually peaks.
* **Descriptive Analysis-**

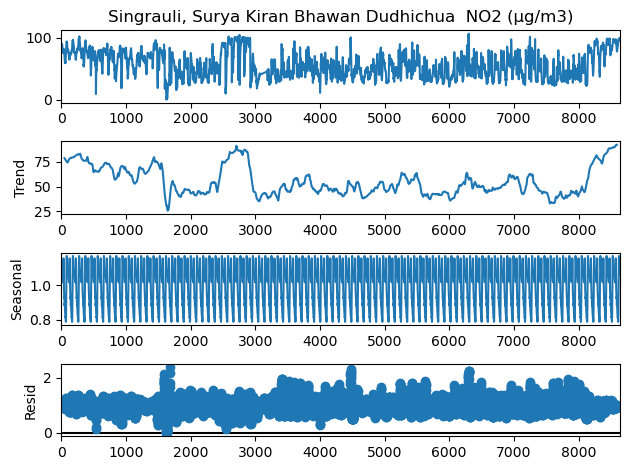
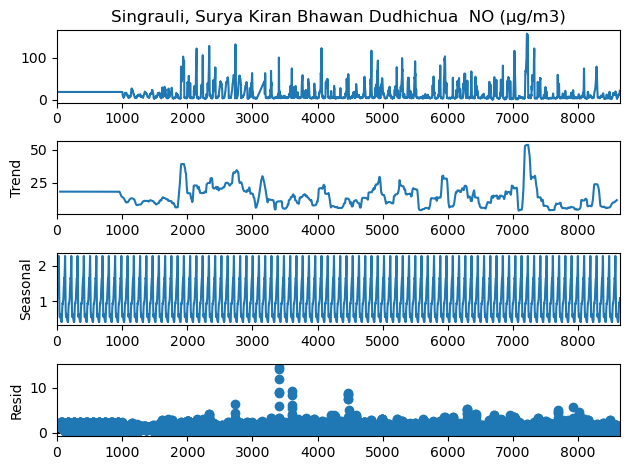
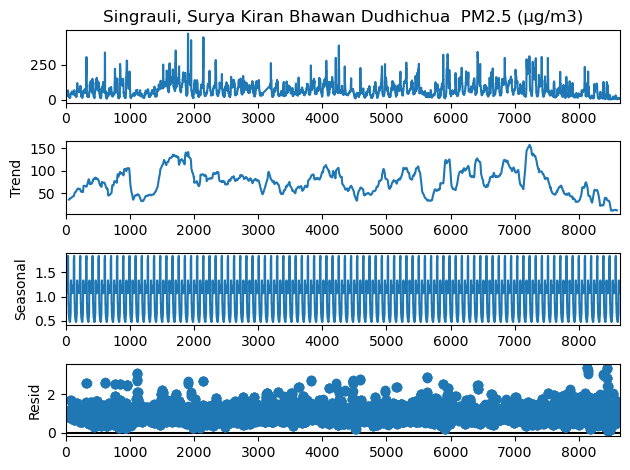
Identifying patterns in time series data at the time of coal India open-pit blasting effect-

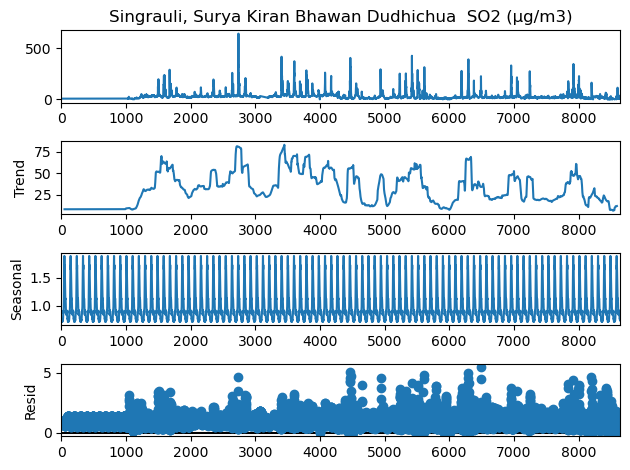
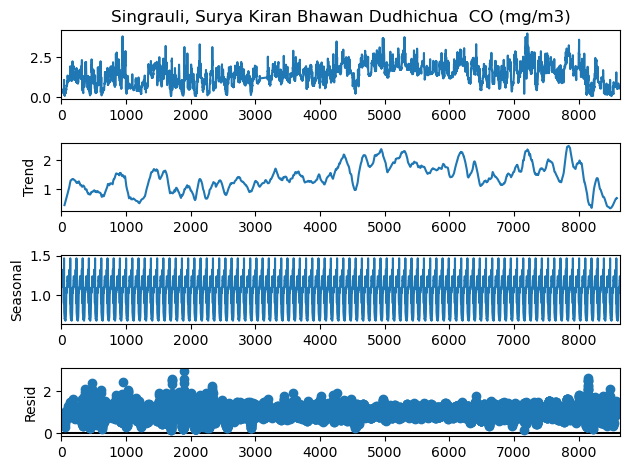
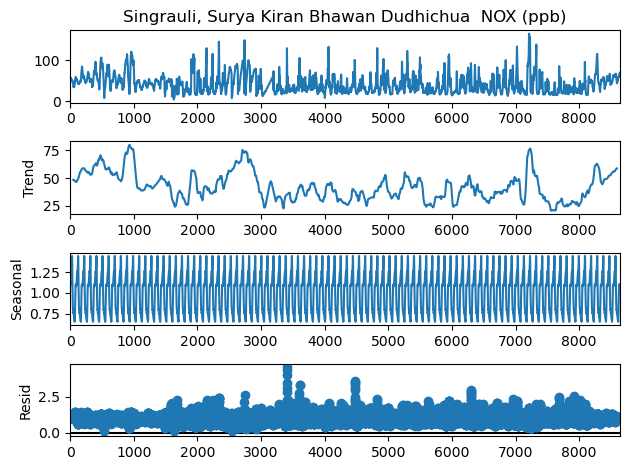
* The level of PM10 and PM2.5 particles in the air is highest throughout a day.
* **Effect of open pit-blasting on air pollution**- We can clearly infer from the graph that after pit-blasting time i.e., after period of 13:45 to 14:45 the level of pollutants in the air spike, particularly the level of PM10 particles spike tremendously resulting in increase in air pollution.
* One more thing that can be observed that the level of ozone decreases shortly after open-pit-blasting. We can infer from this observation that open-pit-blasting may lead to ozone layer depletion.
* **Checking Trend and Seasonality-**

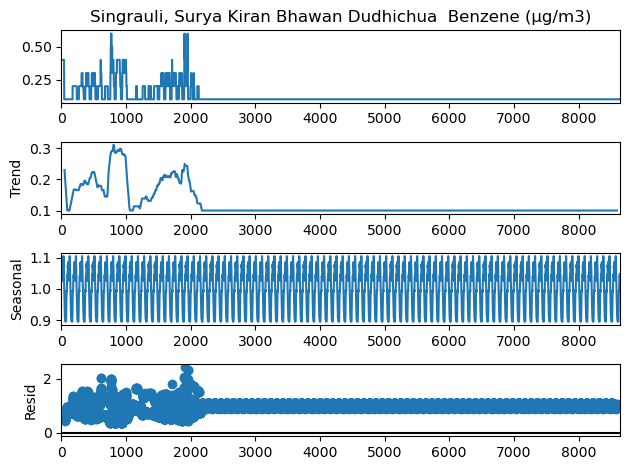
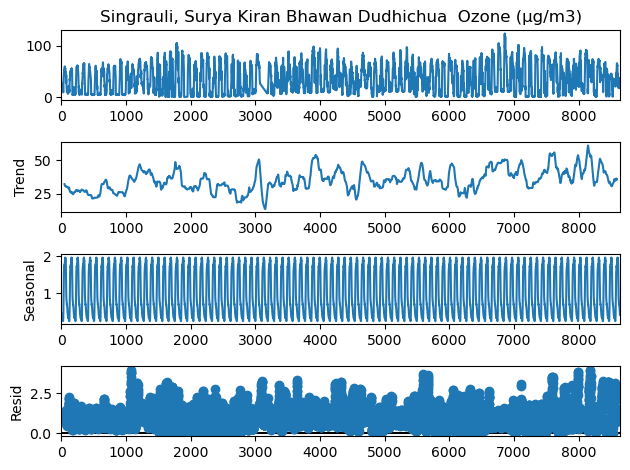
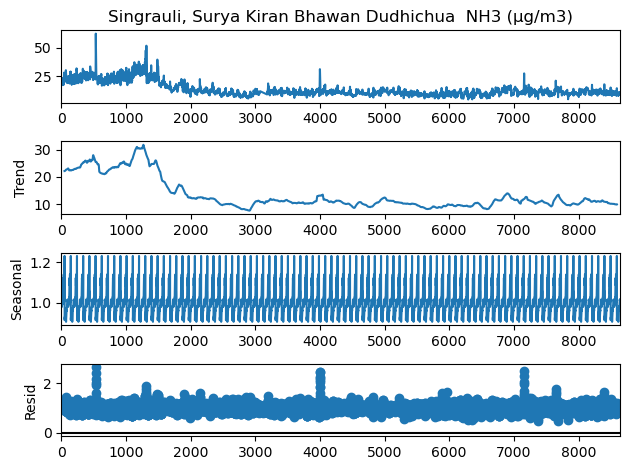


* After decomposing the time series data using the seasonal\_decompose function from statsmodels, we can analyze the resulting plots to draw conclusions about seasonality and trend
* The trend component represents the long-term pattern or direction of the time series. It captures the overall upward or downward movement of the data over time.
* The seasonal component represents the repetitive patterns or cycles that occur within the data. It captures regular fluctuations that repeat at fixed intervals.

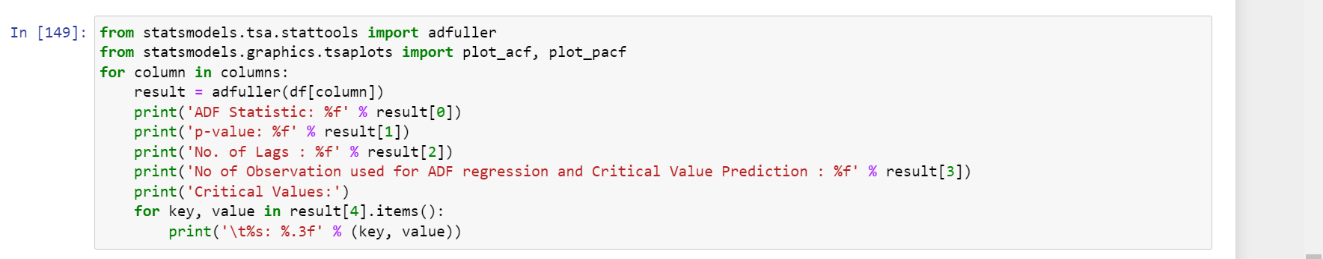


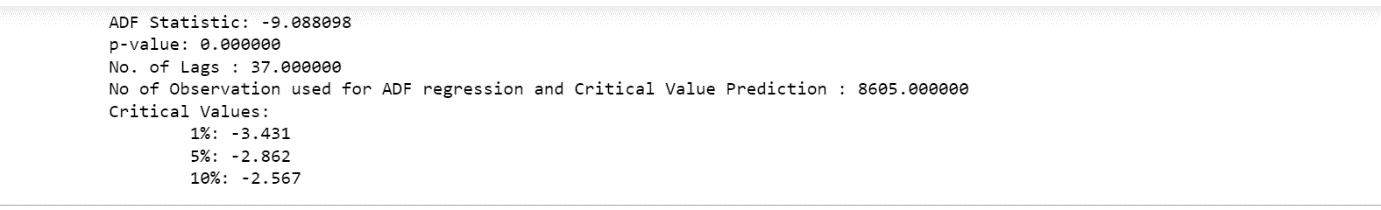
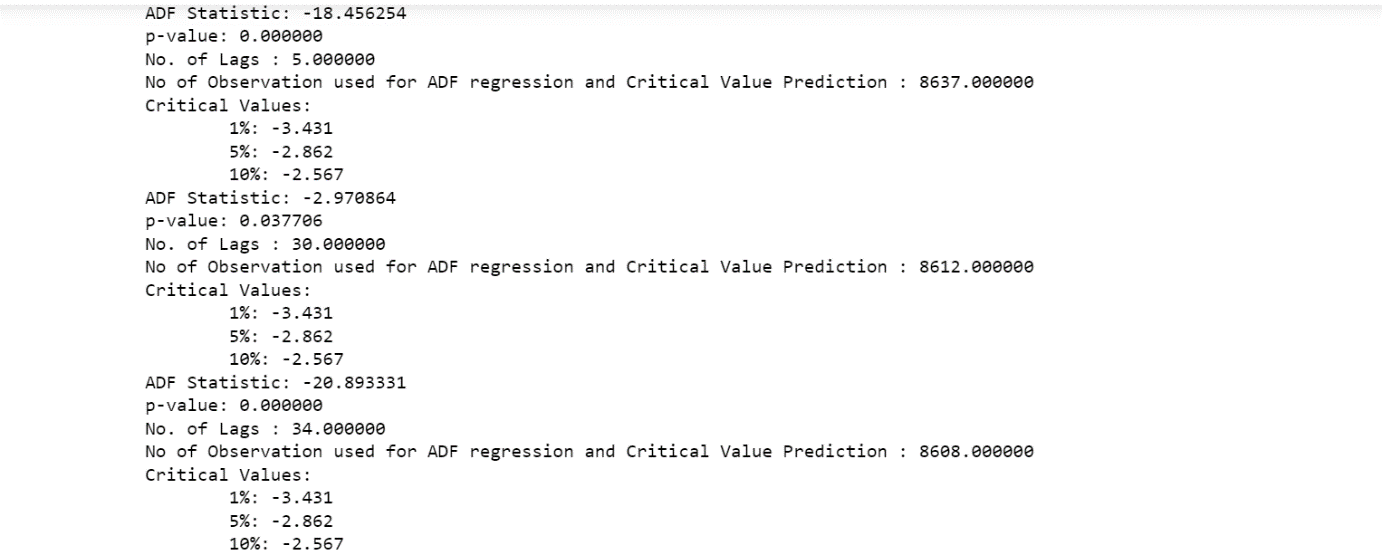
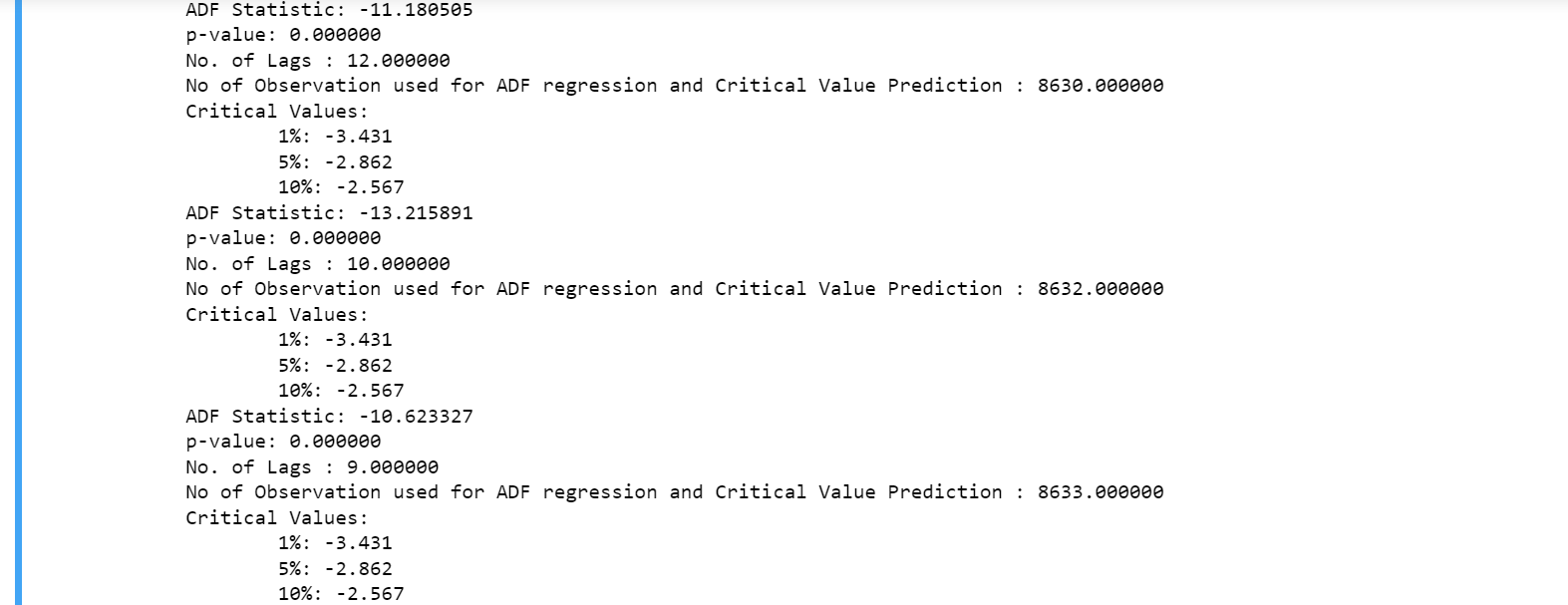
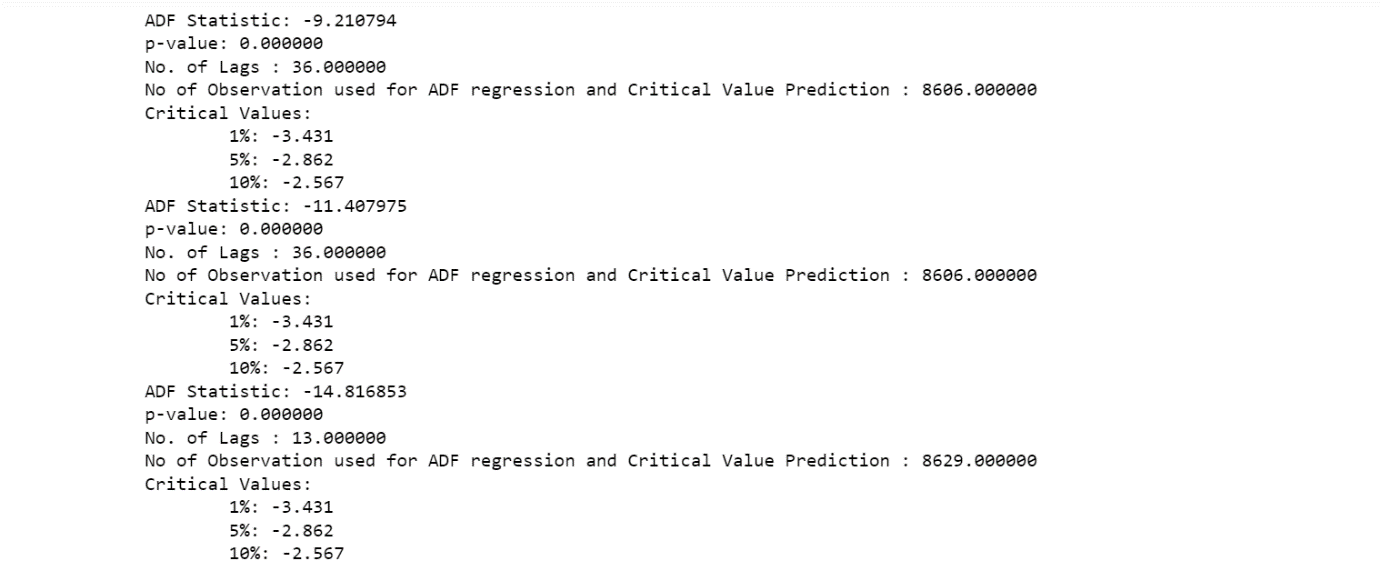






* Inference from the graphs of trends and seasonality-
* From the above time series data, no trend is observed.
* From the graph we observe the time interval of seasonal graph of every data is very small. Thus, we observe no such seasonality in our time series data.
* **Test for Trend and Seasonality-**
* **ADFULLER TEST**-
* The adfuller function from the **statsmodels.tsa.stattools** module is used to perform the Augmented Dickey-Fuller (ADF) test on each column in the DataFrame. The ADF test is a statistical test used to determine if a time series is stationary or not.
* The ADF test examines whether a time series is non-stationary. The null hypothesis of the ADF test is that the time series is non-stationary. The alternate hypothesis is that the time series is stationary. If the time series data is stationary, we can say that, there is no such trend and seasonality in our data.

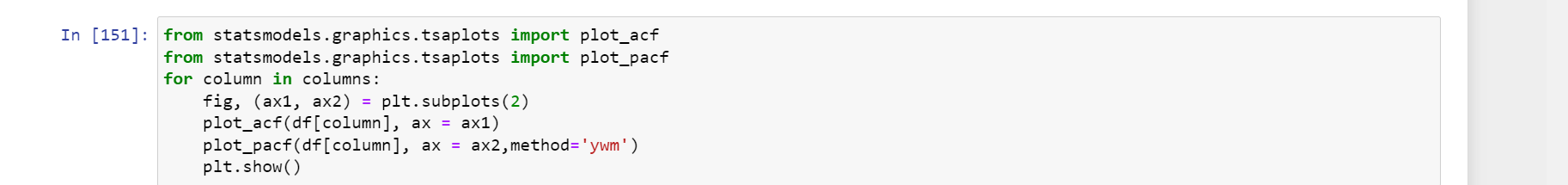


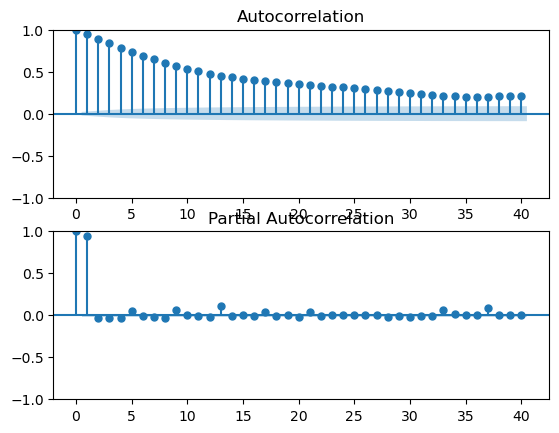
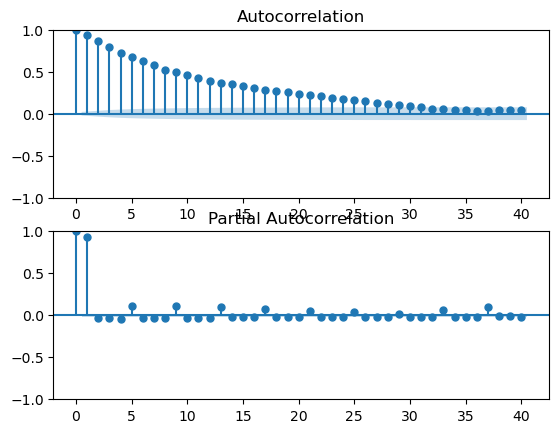
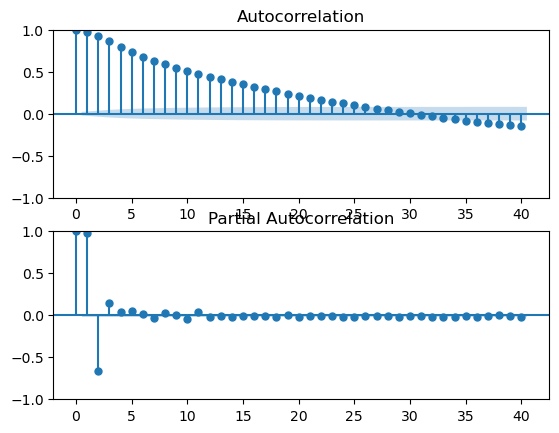
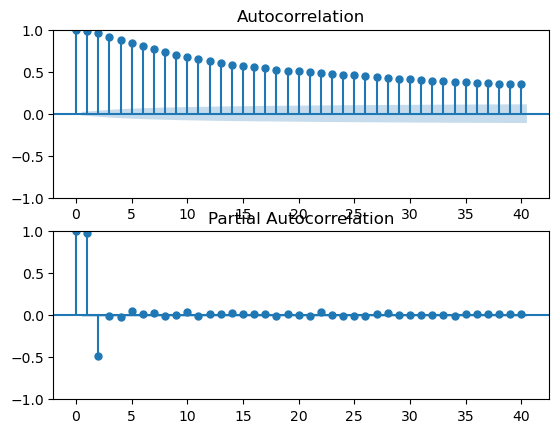
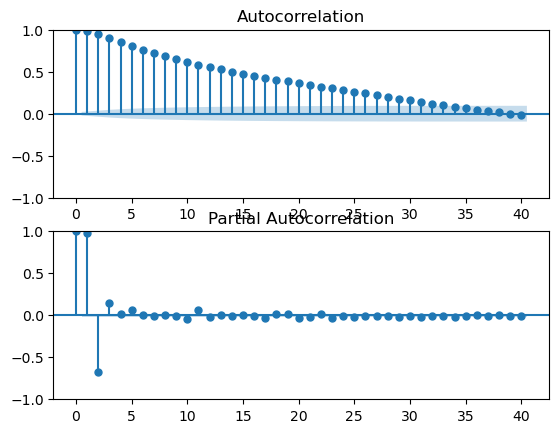
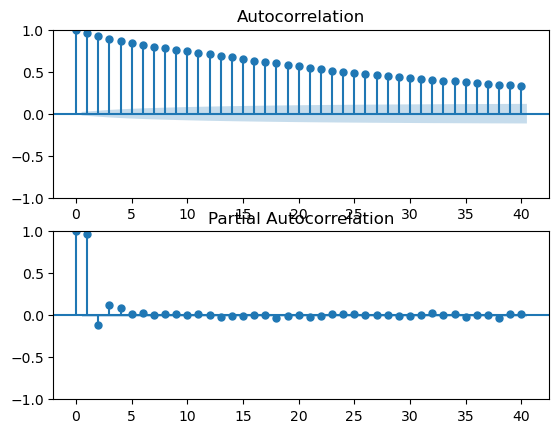
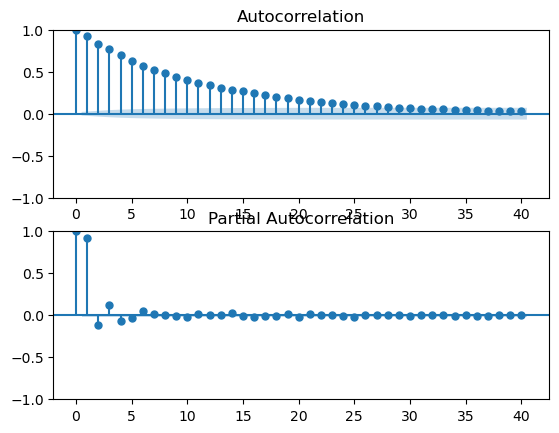
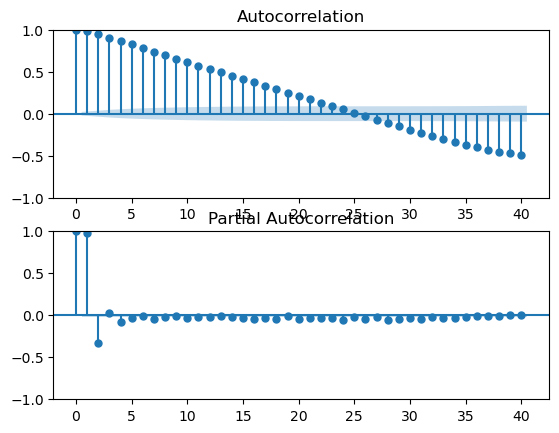
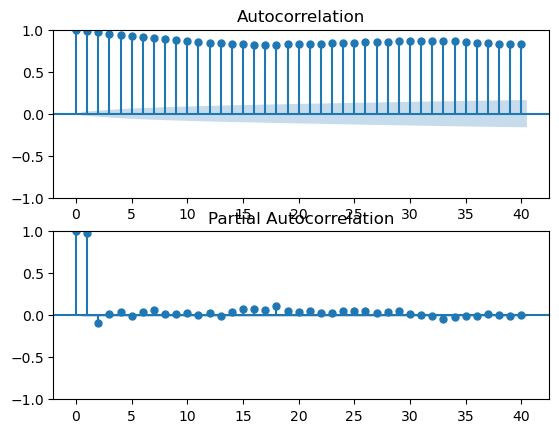
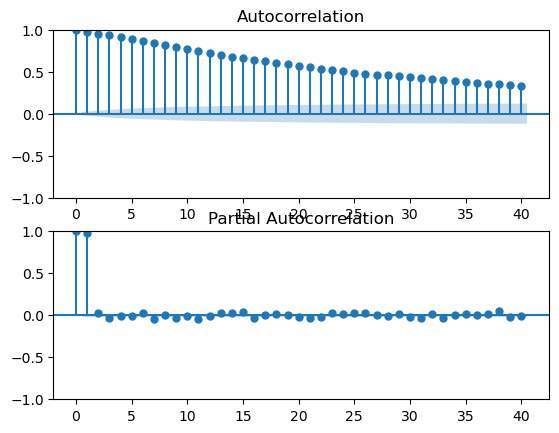


* **Inference drawn from ADFULLER TEST-**

From the above values we conclude the following ->

* The p-value is zero and the ADF statistic is more negative than the critical values at a chosen level of significance, it suggests strong evidence to reject the null hypothesis of non-stationarity. In this case, we can conclude that the time series is stationary.
* Regarding trend and seasonality, here the time series is stationary, it implies that there is no significant trend or seasonality present in the data.
* Overall, a zero p-value and negative critical values in the ADF test suggest that there is no systematic trend or seasonality observed in the time series.
* **Modelling –**
* **Generating Autocorrelation and Partial correlation plots-**
* In the following code we are using **statsmodels.graphics.tsaplots** module to generate autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. The ACF and PACF plots are commonly used in time series analysis to identify the presence of autocorrelation and determine the order of autoregressive (AR) and moving average (MA) components in a time series model.



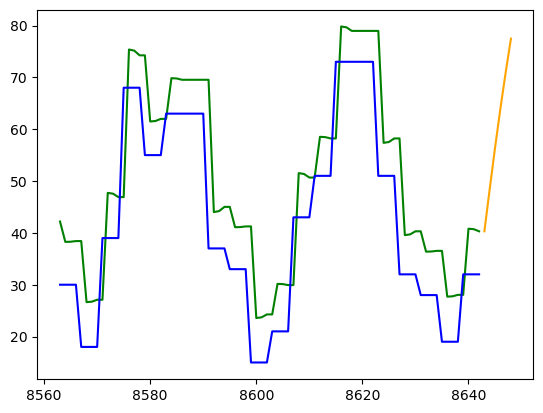
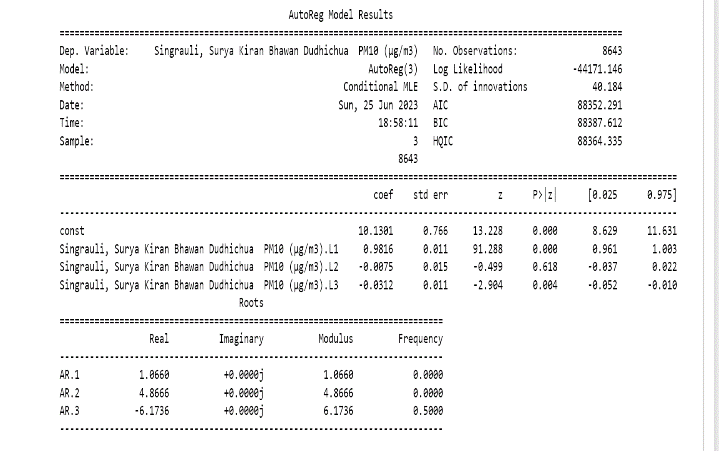
        

* **Inference from ACF and PACF graphs-**
* **ACF**: We observe a gradual decline in autocorrelation graph of most of the time series data.
* **PACF**: We observe significant spikes at lags and subsequent decay. If the PACF plot exhibits significant spikes at lags 'k' and subsequent decay, it suggests an AR component of order p = k.
* From above we infer that time series data follows Autoregressive (AR) model.
* **Training and Testing the Autoregressive (AR) Model-**
* **Autoregressive (AR) model of the provided time series data-**
* Now that we have observed that the time series data follows AR model. We are using it to train and test our data.
* In the subsequent code snippet, we are performing **Autoregressive** modelling on each column in the DataFrame i.e., for each pollutant. The code fits an autoregressive model to the data, generate predictions, and evaluate the model's performance using the RMSE. It also displays the model summary and plots the predicted values against the test data for visual inspection.
* **Following shows the results of Autoregressive Model and shows the best fit curve-**

****

* Results->

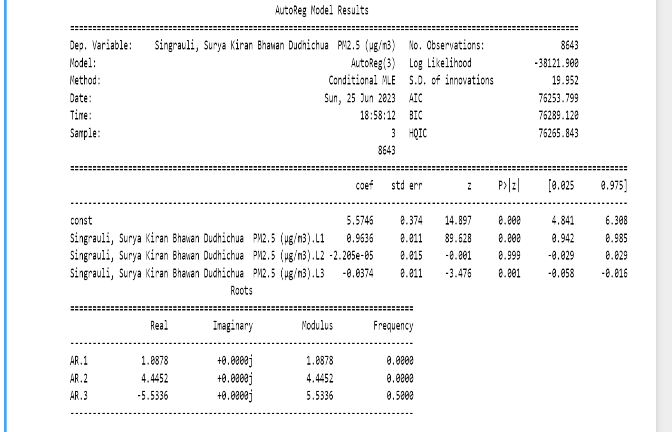
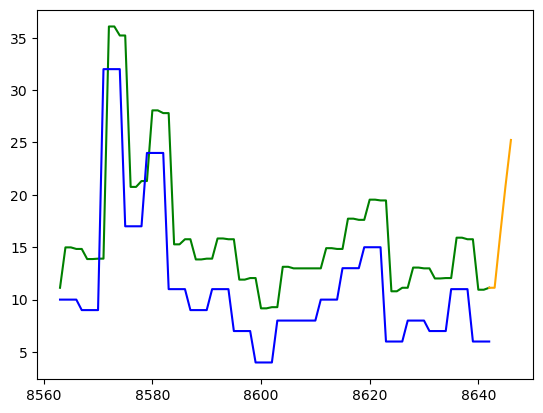
PM10



Mean: 177.412588

Root Mean Squared Error: 10.879073

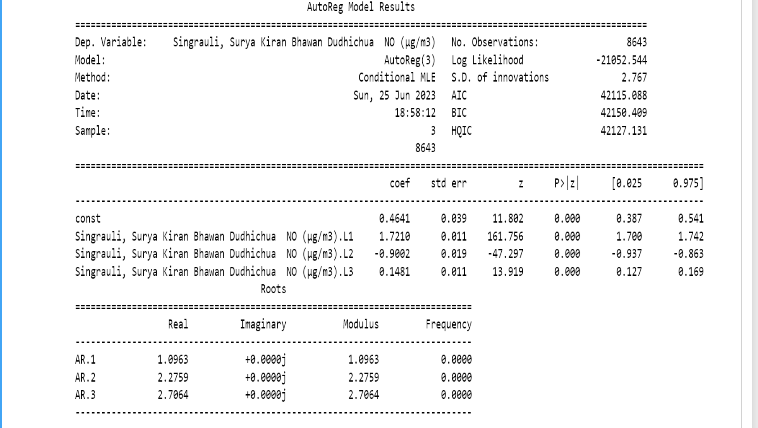
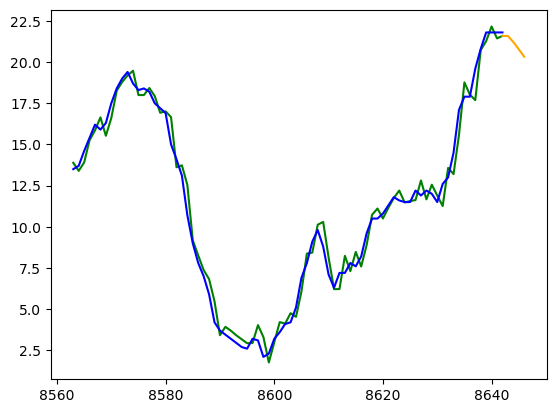
PM2.5

Mean: 75.532206

Root Mean Squared Error: 6.028429

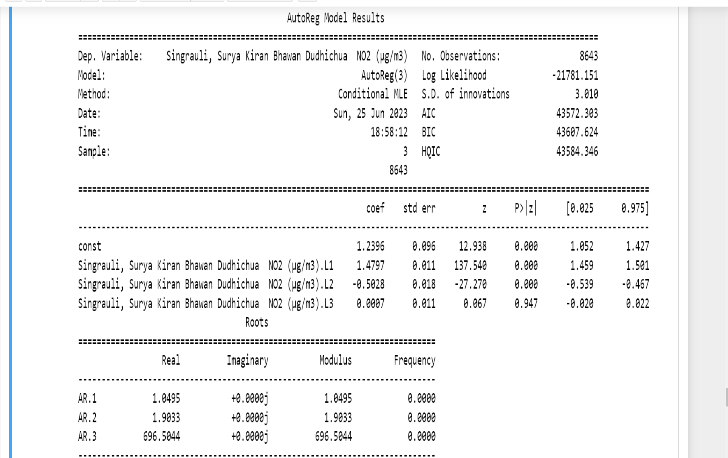
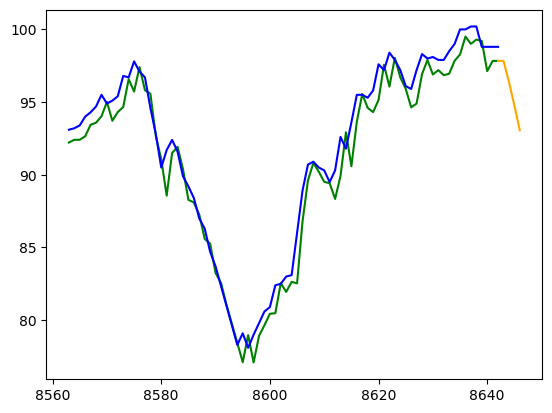
NO

Mean: 14.942589

Root Mean Squared Error: 0.718244

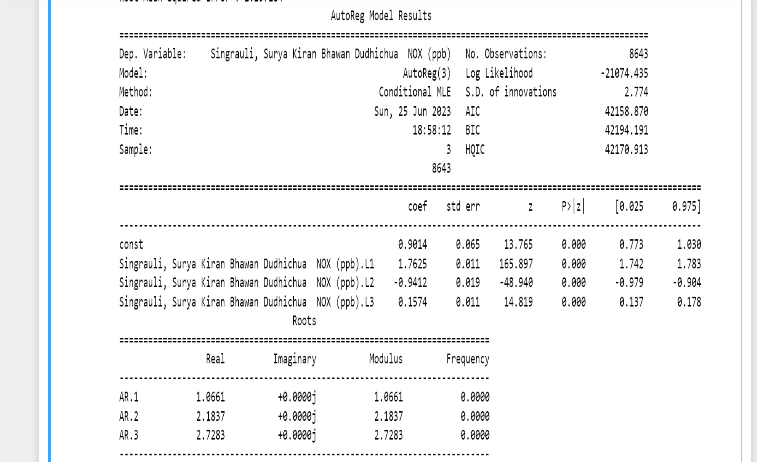
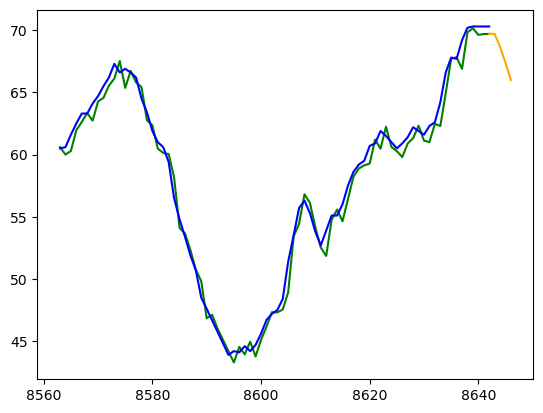
NO2

Mean: 55.445742

Root Mean Squared Error: 1.297184

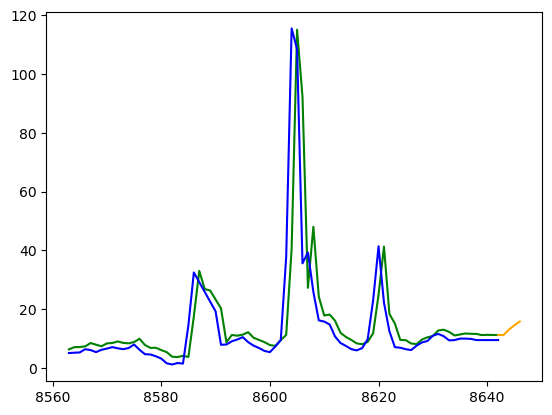
NOX

Mean: 42.338511

Root Mean Squared Error: 0.886471

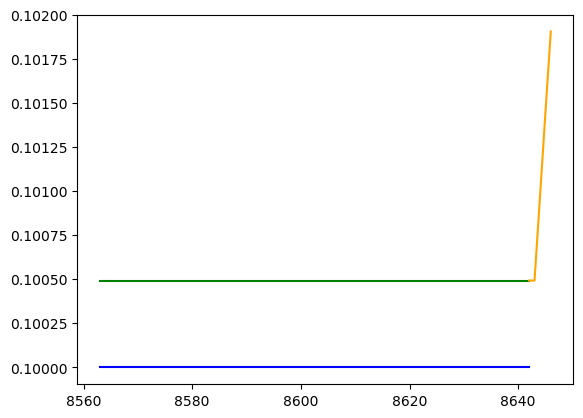
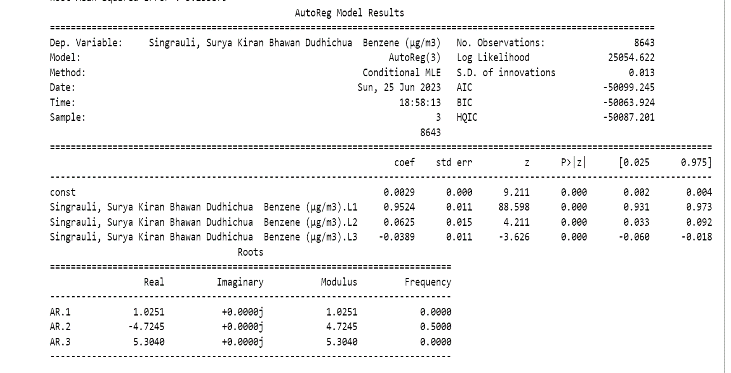
SO2



Mean: 31.915487

Root Mean Squared Error: 12.247141

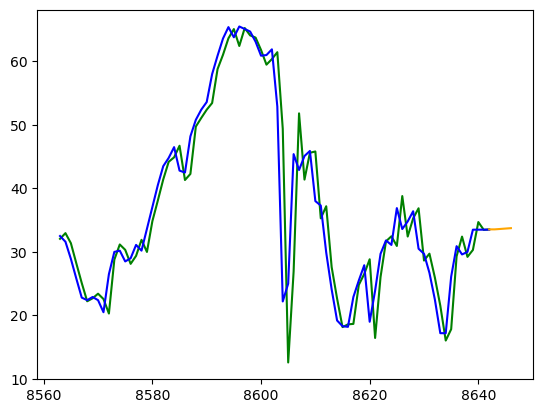
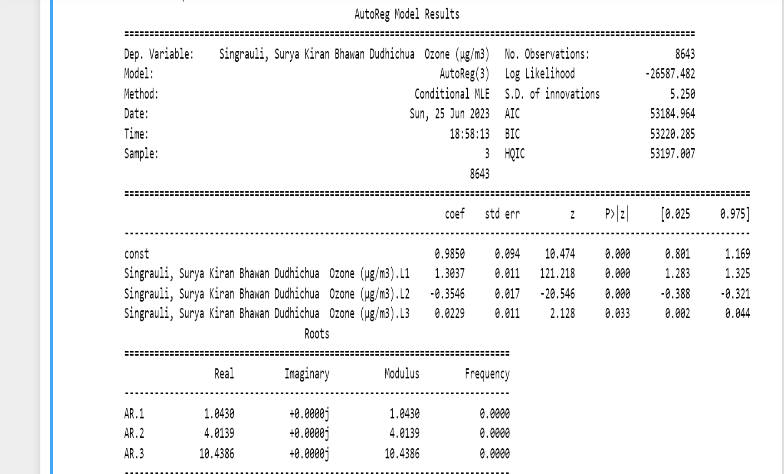
Benzene



Mean: 0.121995

Root Mean Squared Error: 0.000493

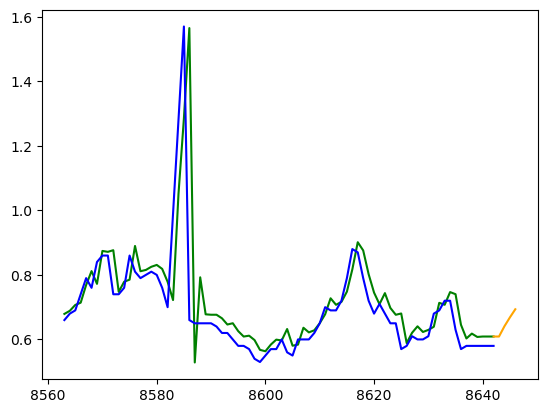
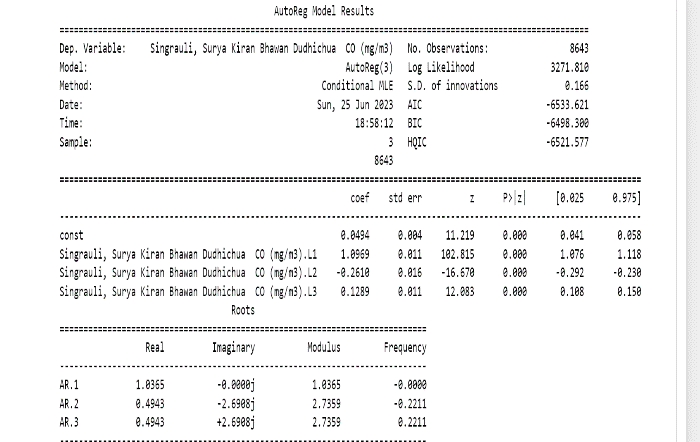
Ozone



Mean: 35.193382

Root Mean Squared Error: 5.233879

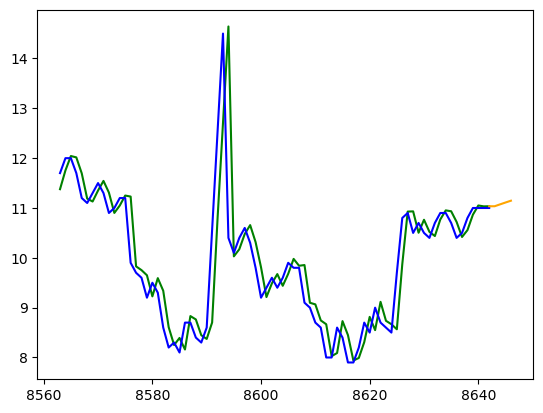
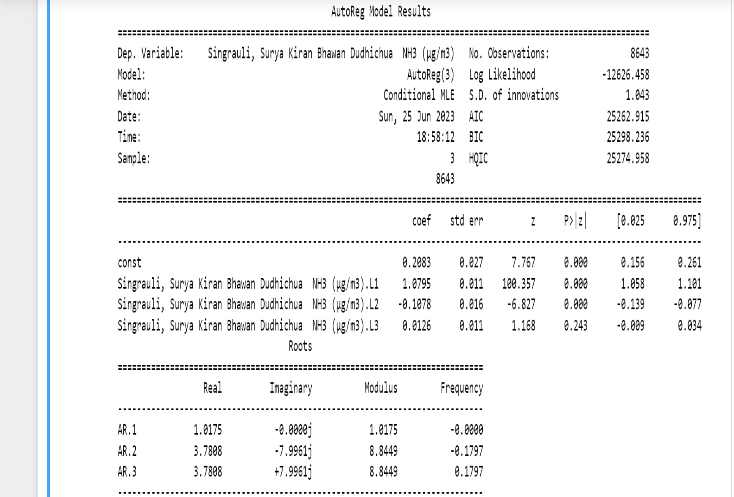
CO



Mean: 1.401641

Root Mean Squared Error: 0.123193

NH3



Mean: 13.286162

Root Mean Squared Error: 0.693484

* **Results-**
* From the above, we observe that using AR modelling, the curves for the time series data for NO, NO2, CO, NOX, Ozone are accurately fitted, while for PM10, PM2.5 curves are approximately fitted.
* Blue lines denote actual time series data while green line shows the predicted test data.
* RMSE values are very low in nearly every case showing our curve is accurately fitted and can be used for forecasting.
* **Forecasting-**
* In the fitted curve graph, we have an orange line which is extrapolating the green line, the one used for prediction. This extrapolated orange line in the curve fitting graph provides a basis for forecasting the future levels of pollutants