EE798 FISA: Assignment

Air Pollution Forecasting

Using Time Series

April Samad

200178

Indian Institute of Technology Kanpur

Mentor-Prof. Tushar Sandhan

Introduction

India’s coal production scenario is dominated by open-pit mining. Future coal demand is expected to be very strong. However, environmental problems, such as worsening air quality brought on by the emission of particulate matter and other gaseous pollutants from diverse mining operations, would restrict the use of coal. We will be analyzing real-world recent time-series data, for descriptive as well as inferential statistics.

COAL INDIA OPEN-PIT BLASTING

The two main air pollutants in NCL coal fields are suspended particulate matter (SPM) and respirable particulate matter (RPM). Air quality monitoring is regularly carried out at both dust generating and non-generating locations in the vicinity in order to evaluate the particulate pollution in and around the opencast mining projects of the Singrauli coalfield. SPM and RPM concentrations are predominant at coal working surfaces, coal yards, coal handling facilities, andhaul roads used to transport coal, as well as close to drilling sites, in overburden, and on such haul roads. Air pollution [1] measurements available via multi-sensory systems are PM10, PM2.5, SO2, NO2, NOx, CO, NH3, O3 and BENZENE.

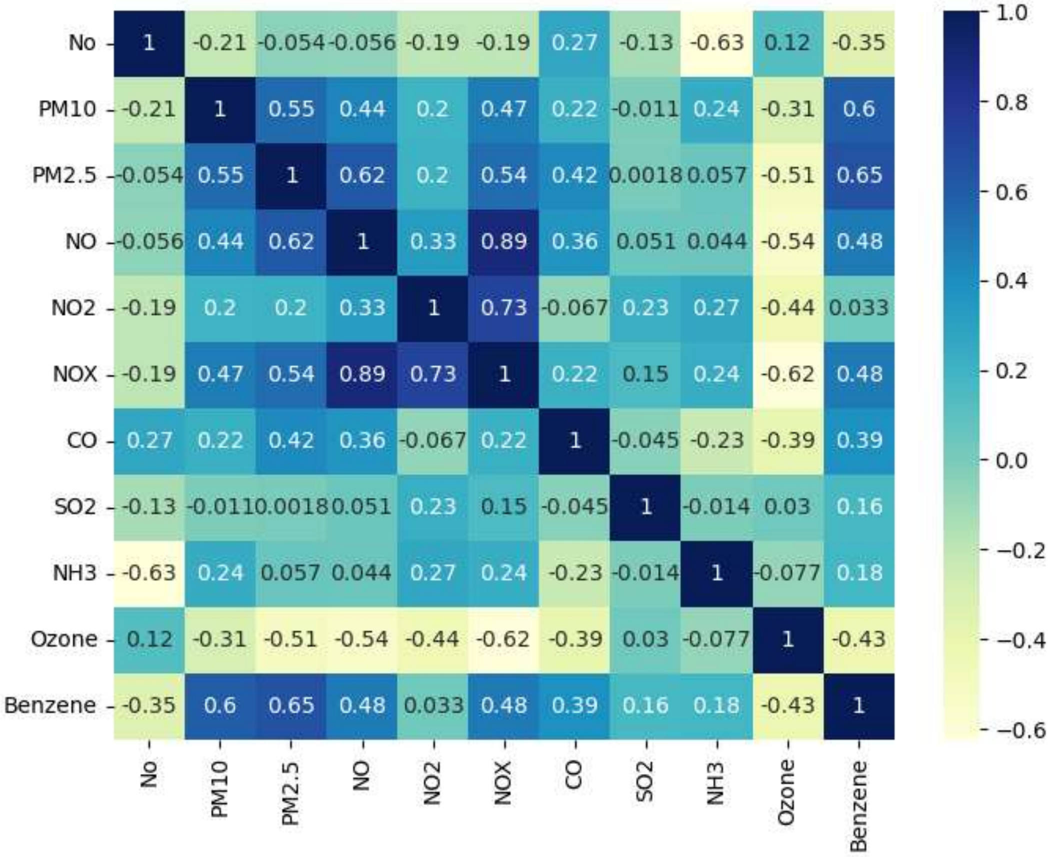
Due to reasons like sensor failure, sensor-to-central-hub communication link failure, data packet loss etc., there will be some missing sensory data for a certain duration of the time. Entire sensory array link failure renders missing values in entire rows, whereas individual sensor mishap causes few entries missing from the column.

Methodology

Data Collection and Understanding :

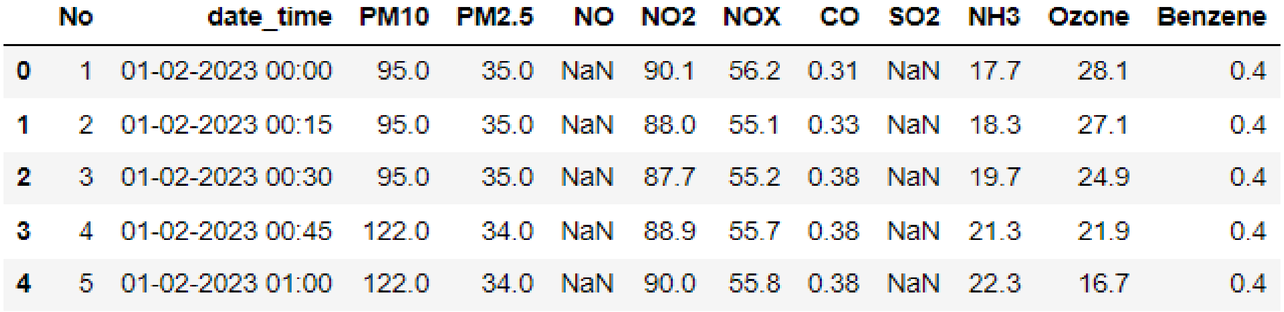
The data for this project was gathered from opencast mining projects of the Singrauli coalfield. The data ranges from January 2nd, 2023 to may 1st, 2023 and contains 15 minute concentrations of air pollutants in Singrauli coalfield , India.
A minimum of three pollutant data points to calculate the Air Quality Index (AQI), one of which must be either PM10 or PM2.5. This was taken into account when gathering the data.

Since Singrauli coalfield is a small town and the air quality monitoring stations have only recently been installed, there was not enough meteorological data available for this project. However, by plotting and analyzing the gathered pollution data, it was found that PM2.5 accounted for 99% of the AQI values. This is because PM2.5 is the most responsible pollutant affecting air pollution.



Data Preparation

Preprocessing refers to the steps taken to prepare data for modeling. It involves various tasks like handling missing values, dealing with white spaces, performing calculations, and splitting the data into training and testing sets. In a specific project, preprocessing was the most time-consuming aspect because the data required cleaning and transformation before it could be used for modeling.
The dataset used in the project had 10 columns namedPM10, PM2.5, SO2, NO2, NOx, CO, NH3, O3(ozone) and BENZENE. All the columns were then converted to numeric format, and any values that were null were replaced with interpolation values.



	No	date_time	PM10	PM2.5	NO	NO2	NOX	CO	SO2	NH3	Ozone	Benzene
0	1	01-02-2023 00:00	95.0	35.0	NaN	90.1	56.2	0.31	NaN	17.7	28.1	0.4
1	2	01-02-2023 00:15	95.0	35.0	NaN	88.0	55.1	0.33	NaN	18.3	27.1	0.4
2	3	01-02-2023 00:30	95.0	35.0	NaN	87.7	55.2	0.38	NaN	19.7	24.9	0.4
3	4	01-02-2023 00:45	122.0	34.0	NaN	88.9	55.7	0.38	NaN	21.3	21.9	0.4
4	5	01-02-2023 01:00	122.0	34.0	NaN	90.0	55.8	0.38	NaN	22.3	16.7	0.4

If the missing values are located at the beginning of the time series and there are no preceding values available, interpolation may not be the most suitable approach. Interpolation methods typically rely on having neighboring values on both sides of the missing data to estimate and fill in the gaps.

In such cases, it might be more appropriate to consider other techniques for handling missing data. Some alternatives include:

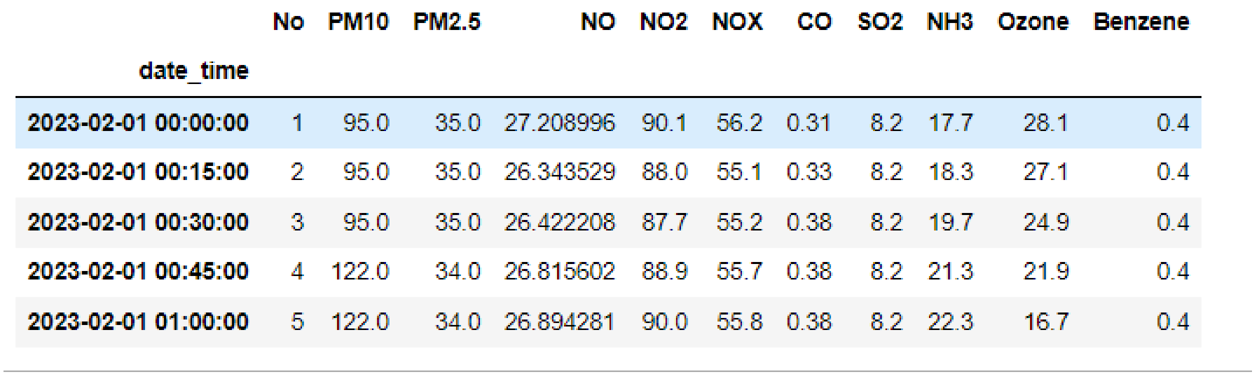
1. Forward Fill (or Last Observation Carried Forward): Propagate the last observed value forward to fill in the missing values. This assumes that the missing values continue to follow the same pattern as the last observed value.

2. Mean or Median Imputation: Replace the missing values with the mean or median of the available data. This assumes that the missing values are similar to the overall distribution of the data.

3. Seasonal Imputation: If there are seasonal patterns in the data, you can impute the missing values by taking the average value of the corresponding season in previous years or periods.

4. Time-based Interpolation: If you have additional time-related information or variables, you can use regression or time series modeling techniques to estimate the missing values based on other correlated factors.

It is important to consider the nature of the data and the specific context before choosing the most appropriate method for handling missing values.

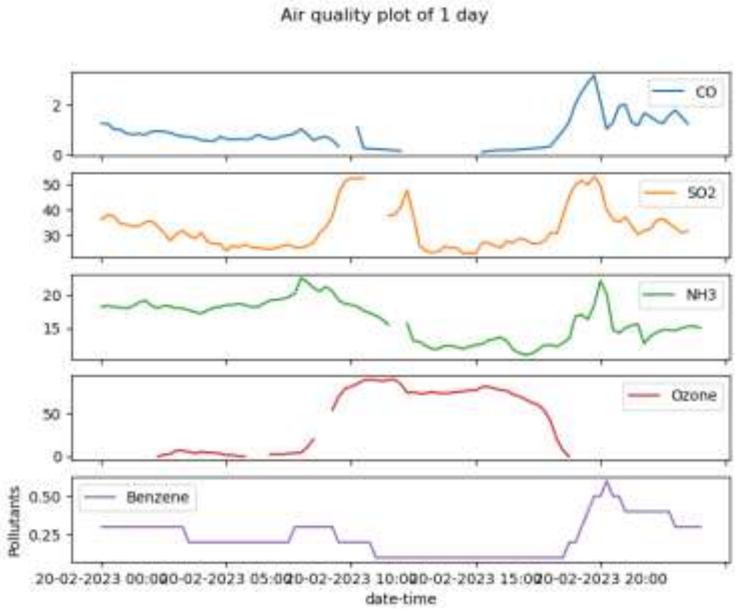


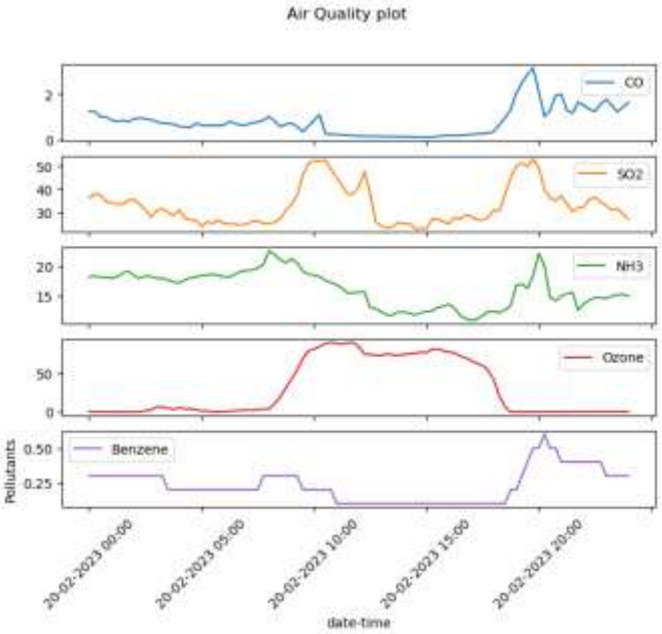
	No	PM10	PM2.5	NO	NO2	NOX	CO	SO2	NH3	Ozone	Benzene
date_time											
2023-02-01 00:00:00	1	95.0	35.0	27.208996	90.1	56.2	0.31	8.2	17.7	28.1	0.4
2023-02-01 00:15:00	2	95.0	35.0	26.343529	88.0	55.1	0.33	8.2	18.3	27.1	0.4
2023-02-01 00:30:00	3	95.0	35.0	26.422208	87.7	55.2	0.38	8.2	19.7	24.9	0.4
2023-02-01 00:45:00	4	122.0	34.0	26.815602	88.9	55.7	0.38	8.2	21.3	21.9	0.4
2023-02-01 01:00:00	5	122.0	34.0	26.894281	90.0	55.8	0.38	8.2	22.3	16.7	0.4

Fig:after interpolation and using linear regression for NO and using backward fill of SO2.

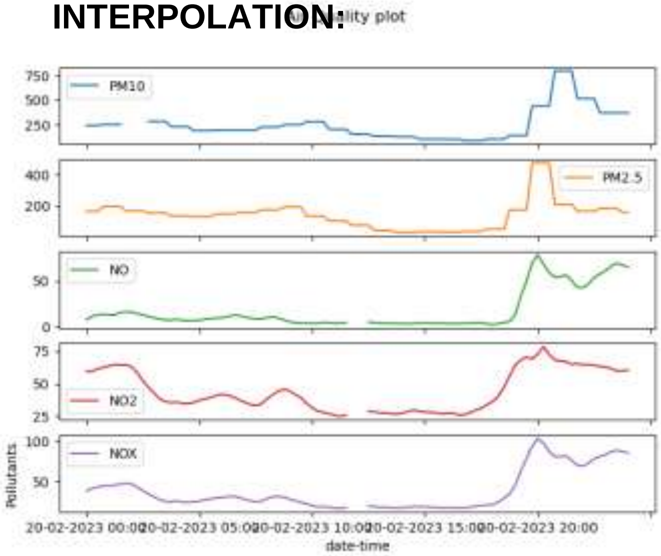
BEFORE INTERPOLATION :

AFTER INTERPOLATION:





BEFORE INTERPOLATION : **AFTER**
INTERPOLATION:



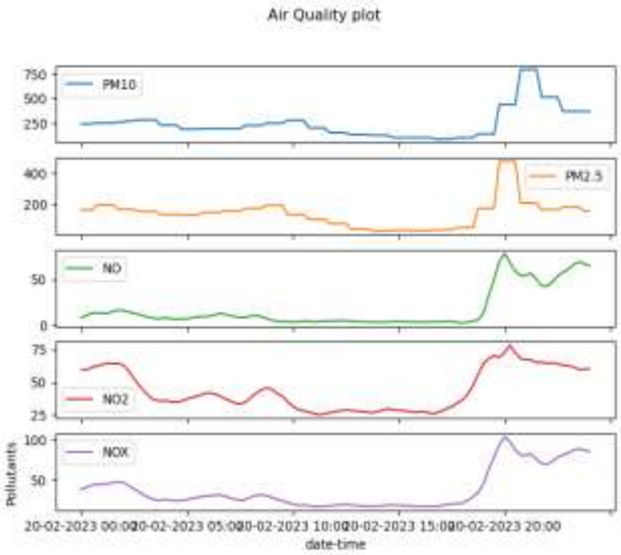
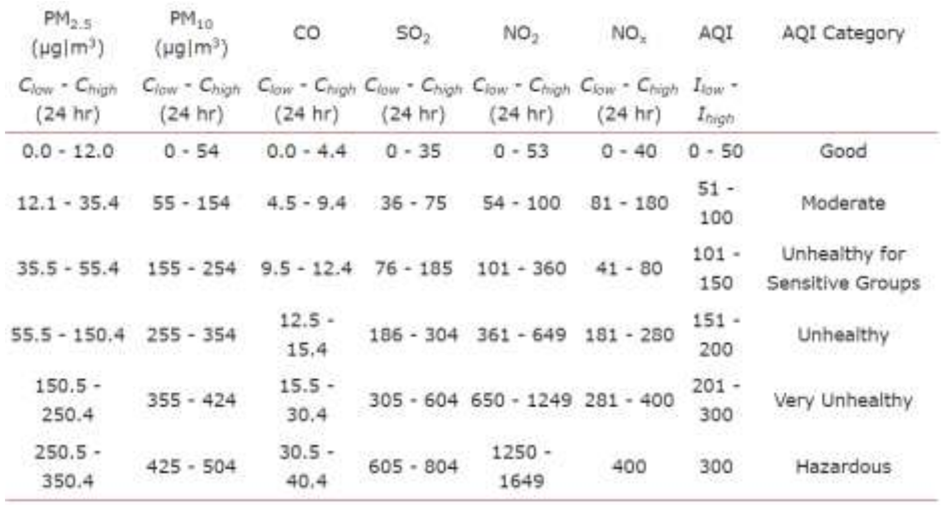


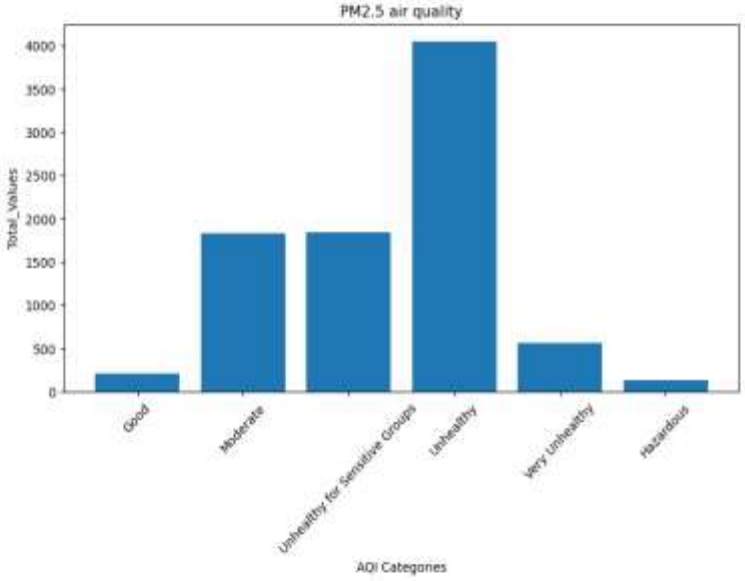
Fig: interpolated data of 1 day from 20-02-2023 00:00 to 21-02-2023 00:00

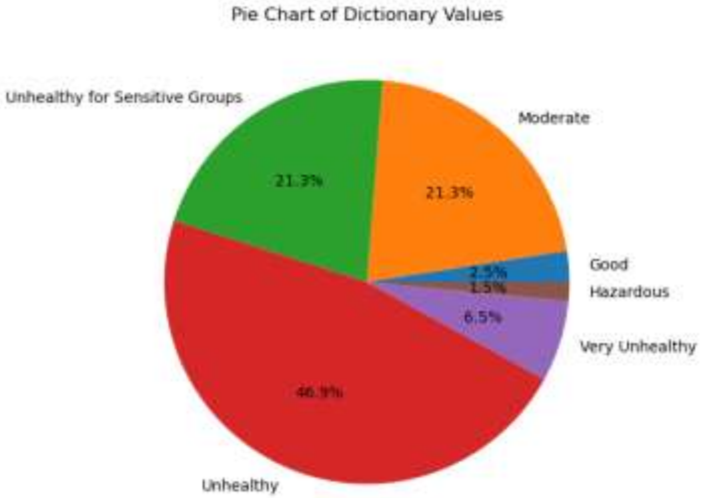
Descriptive data of measured air quality.



PM$_{2.5}$ (µg|m^3)	PM$_{10}$ (µg|m^3)	CO	SO$_2$	NO$_2$	NO$_x$	AQI	AQI Category
C_{low} - C_{high} (24 hr)	C_{low} - C_{high} (24 hr)	C_{low} - C_{high} (24 hr)	C_{low} - C_{high} (24 hr)	C_{low} - C_{high} (24 hr)	C_{low} - C_{high} (24 hr)	I_{low} - I_{high}	
0.0 - 12.0	0 - 54	0.0 - 4.4	0 - 35	0 - 53	0 - 40	0 - 50	Good
12.1 - 35.4	55 - 154	4.5 - 9.4	36 - 75	54 - 100	81 - 180	51 - 100	Moderate
35.5 - 55.4	155 - 254	9.5 - 12.4	76 - 185	101 - 360	41 - 80	101 - 150	Unhealthy for Sensitive Groups
55.5 - 150.4	255 - 354	12.5 - 15.4	186 - 304	361 - 649	181 - 280	151 - 200	Unhealthy
150.5 - 250.4	355 - 424	15.5 - 30.4	305 - 604	650 - 1249	281 - 400	201 - 300	Very Unhealthy
250.5 - 350.4	425 - 504	30.5 - 40.4	605 - 804	1250 - 1649	400	300	Hazardous

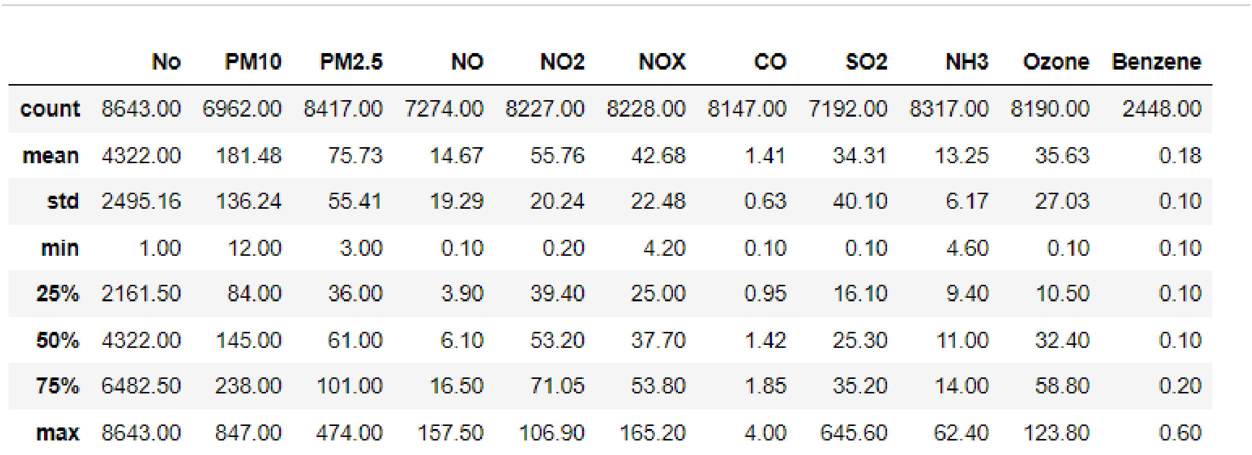
Table 4. EPA's AQI values.





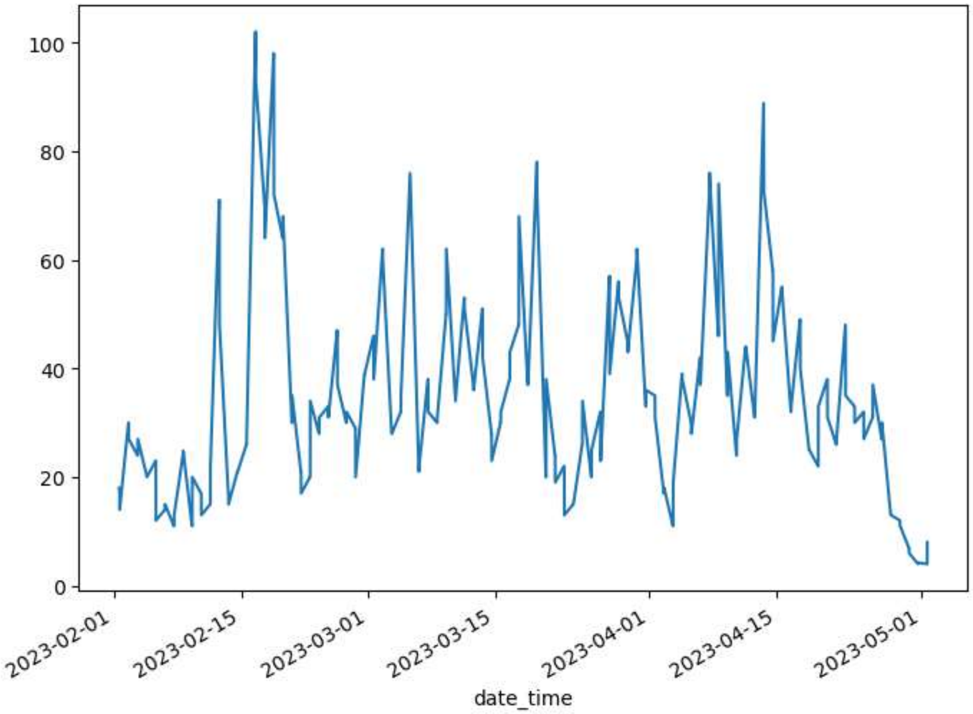
We notice that (2.5%) out of total days are leveled as good, (21%) leveled as moderate, (46%) as unhealthy, (21%) as unhealthy for sensitive, and (1.5) as hazardous .

Descriptive analysis:

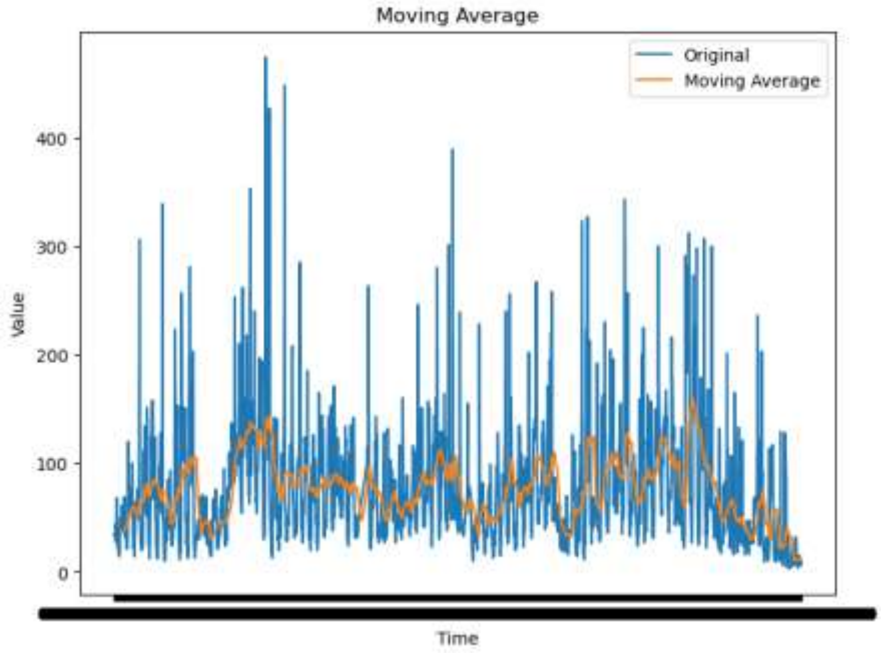


	No	PM10	PM2.5	NO	NO2	NOX	CO	SO2	NH3	Ozone	Benzene
count	8643.00	6962.00	8417.00	7274.00	8227.00	8228.00	8147.00	7192.00	8317.00	8190.00	2448.00
mean	4322.00	181.48	75.73	14.67	55.76	42.68	1.41	34.31	13.25	35.63	0.18
std	2495.16	136.24	55.41	19.29	20.24	22.48	0.63	40.10	6.17	27.03	0.10
min	1.00	12.00	3.00	0.10	0.20	4.20	0.10	0.10	4.60	0.10	0.10
25%	2161.50	84.00	36.00	3.90	39.40	25.00	0.95	16.10	9.40	10.50	0.10
50%	4322.00	145.00	61.00	6.10	53.20	37.70	1.42	25.30	11.00	32.40	0.10
75%	6482.50	238.00	101.00	16.50	71.05	53.80	1.85	35.20	14.00	58.80	0.20
max	8643.00	847.00	474.00	157.50	106.90	165.20	4.00	645.60	62.40	123.80	0.60

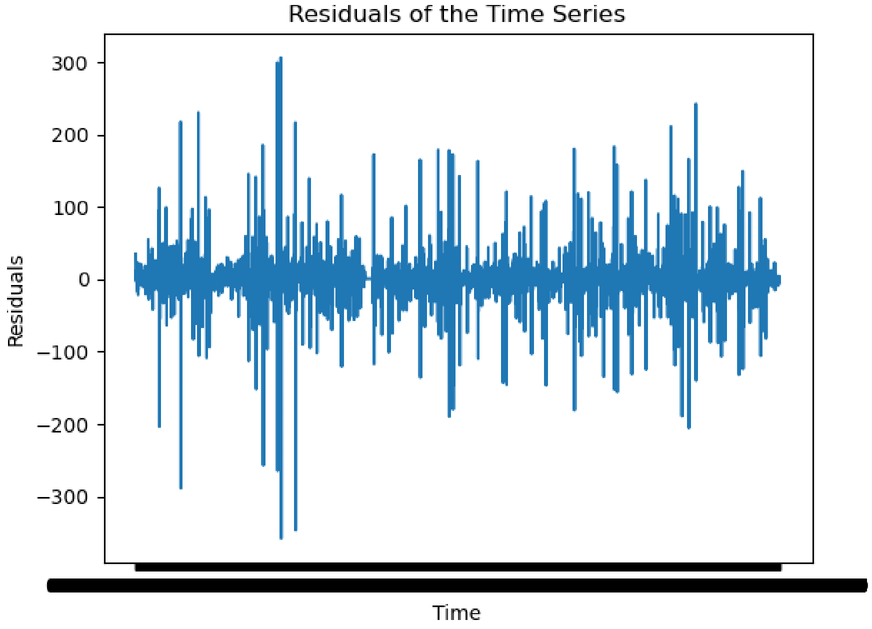
Coal india blasting effect time is 13:45 pm to 14:45 pm plot



Moving Average:



Residual of Time series:



Data Modeling

Test of stationarity:

Before applying time series forecasting models to the data, several tests were conducted to ensure the suitability of the dataset. One crucial aspect that distinguishes time series data from other types of data is its stationarity. A stationary time series means that it lacks any specific trend or seasonality patterns. Unlike general classification or regression projects, time series data often exhibits variations related to time, such as changing pollution levels across seasons or months. Models like AR, ARMA, and ARIMA are not effective when dealing with non-stationary data. Therefore, stationarity tests were performed to determine the viability of using these models.

Autocorrelation and partial autocorrelation plots were generated to assess the presence of any patterns in the data. Statistical tests examining the variance and mean of the data across different intervals indicated variations, suggesting non-stationarity.

To further evaluate the stationarity of the dataset, the Augmented Dickey Fuller (ADF) test was employed. This test is widely used in time series analysis. The results of these tests indicated that the data is indeed stationary. Furthermore, seasonal decomposition analysis was performed, which confirmed the absence of significant seasonal patterns in the data.

Evaluation :

Mean Squared Error:

MSE is obtained by taking the average of squared forecast errors in any prediction. Due to the squared errors, large forecasting differences are converted to even bigger errors by squaring them. So outliers increase the MSE values abruptly.

IMPLEMENTATION :

Introduction:

Implementation and evaluation are two crucial components of any data analytics project, and for this particular project, Python Jupyter notebook was utilized for implementing and evaluating the models. The analysis and visualization of the results were carried out using Python. Python provides a wide range of packages that were readily available for this project, while any additional required packages were installed using the "pip install" function. The "statsmodel" package was particularly helpful as it offers various time series models. Additionally, plotting functions such as "pacf", "acf", and "lag plot" were accessed either through the "pandas" or "statsmodel" packages.

Lag Plot Correlation :

By examining the obtained lag plot, it was observed that there is a positive correlation among the time series values at different time intervals. This indicates that the time series data is suitable for forecasting future values. The presence of correlation suggests that there is some predictable pattern or relationship between the current and future values, which can be leveraged for forecasting purposes.

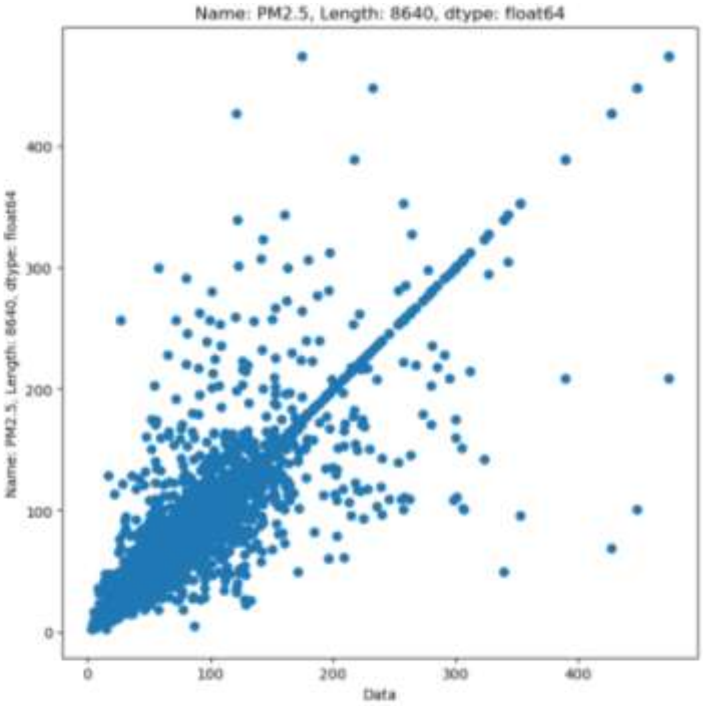


Fig: Lag Plot

Auto-correlation and Partial-autocorrelation:

PACF and ACF plots are valuable tools in time series analysis to examine the correlation between a time series and its lagged values. ACF calculates the correlation between the current value and its lagged values, while PACF measures the autocorrelation with the residuals.

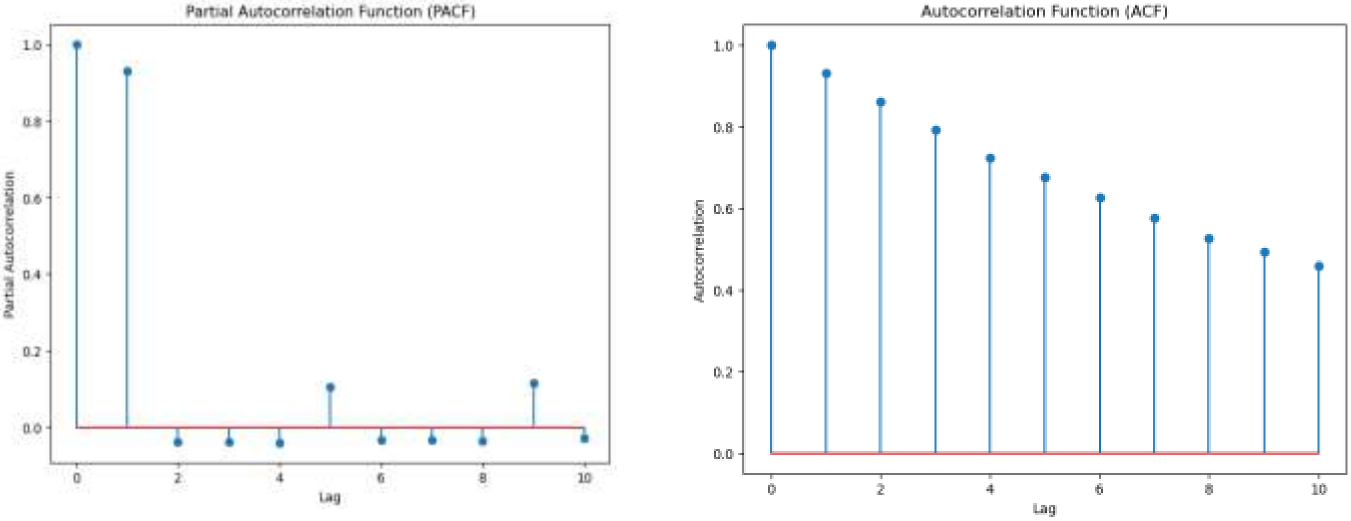


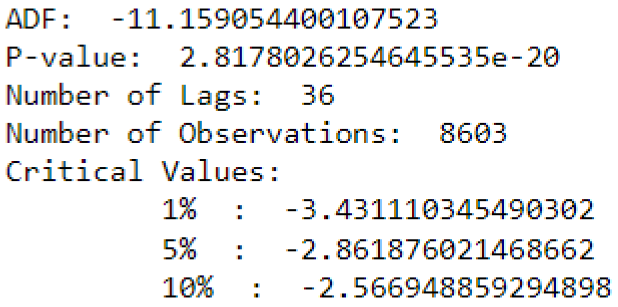
Fig: PACF and ACF

Stationarity Tests:

Augmented Dickey Fuller Test:

To ensure the accuracy of the analysis and avoid misinterpretation, the stationarity of the time series data was further confirmed using statistical tests. The Augmented Dickey-Fuller (ADF) tests were employed for this purpose.

The ADF test examines the null hypothesis that the time series possesses a unit root, indicating non-stationarity. This test provides complementary information, allowing us to assess the stationarity characteristics of the data accurately.



```
ADF:  -11.159054400107523
P-value:  2.8178026254645535e-20
Number of Lags:  36
Number of Observations:  8603
Critical Values:
         1%  :  -3.431110345490302
         5%  :  -2.861876021468662
         10%  :  -2.566948859294898
```

Fig : ADF test

Model Implementation and Evaluation :

1. Evaluation of Auto-Regressive(AR) Forecasting Model

The autoregressive (AR) model is a time series model that incorporates the relationship between the current state of the series and its past values. The AR model calculates the optimal number of past values, known as lags, to predict the current values. In this project, the AR model was evaluated using three different lag values: 1, 2, and 10. Through testing, it was determined that a lag value of 1 produced the best performance.

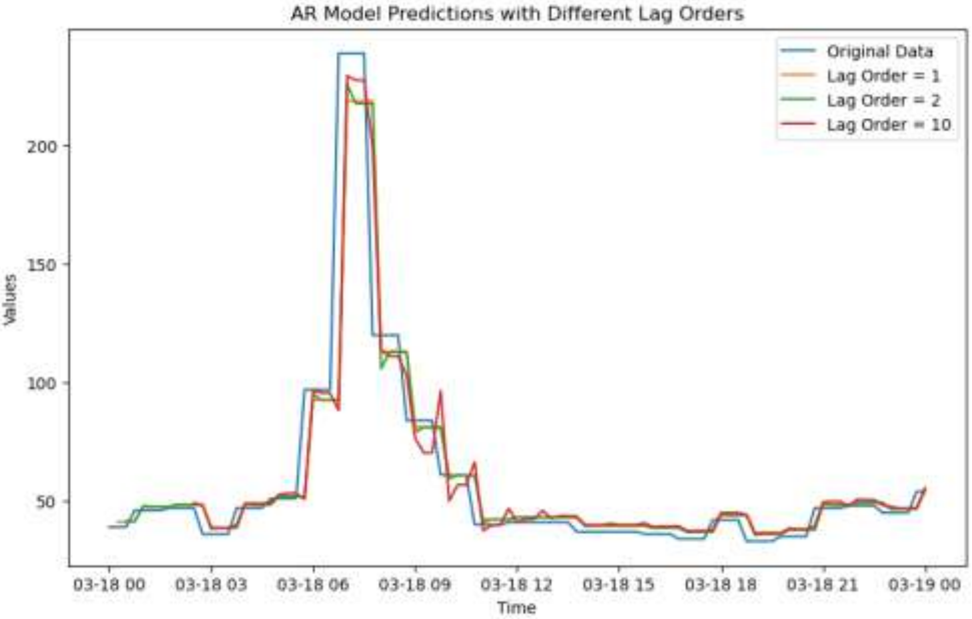


Fig : figure shows AR model of order 1,2,10

And the data is of [PM2.5] From 18-03-2023 00:00 to 19-03-2023 00:00

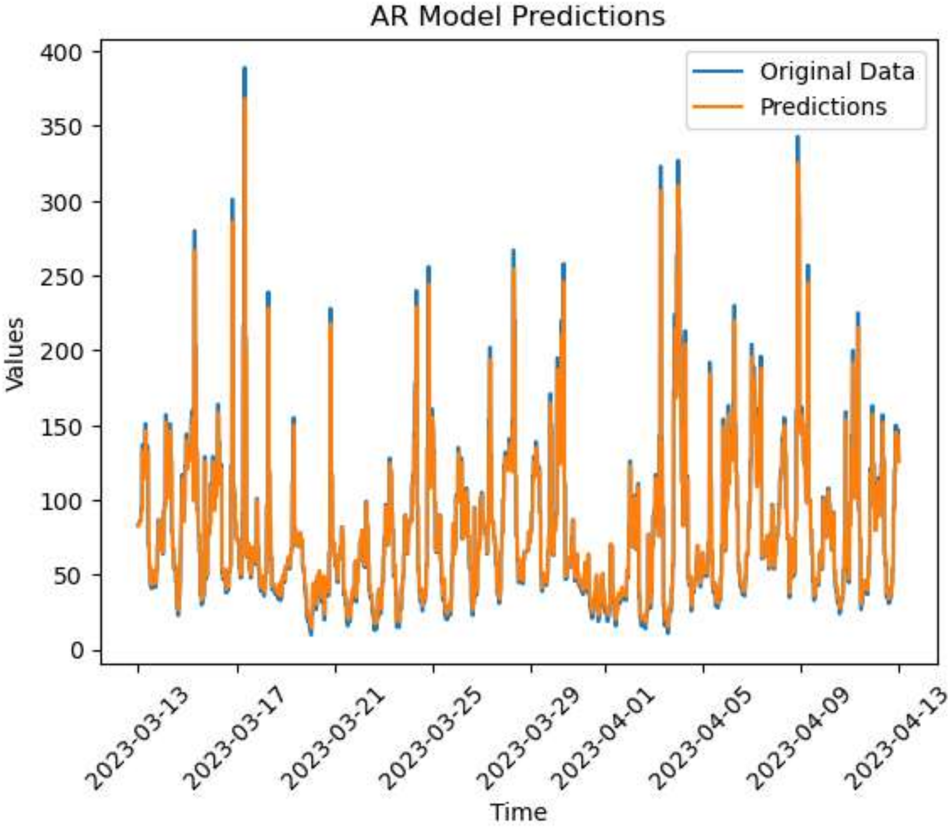


Fig : figure shows AR model of order 1 and the data is of PM2.5 in duration of 1 month from 13-03-2023 to 13-04-2023

Evaluation: AR of order 1 gave the best performance compared to AR2 and AR3. , AR also performed better for long term predictions. The overall performance of AR was not satisfactory but the execution time was impressive.

Forecasts	RMSE	Time
1 Day	4.7792	0.7245 sec
30 Days	3.5545	0.5757 sec

2. Evaluation of Moving average

Moving average is a commonly used method in time series analysis to smooth out variations and identify underlying trends or patterns in the data. It involves calculating the average of a fixed window of consecutive data points and using it as an estimate for the current value.

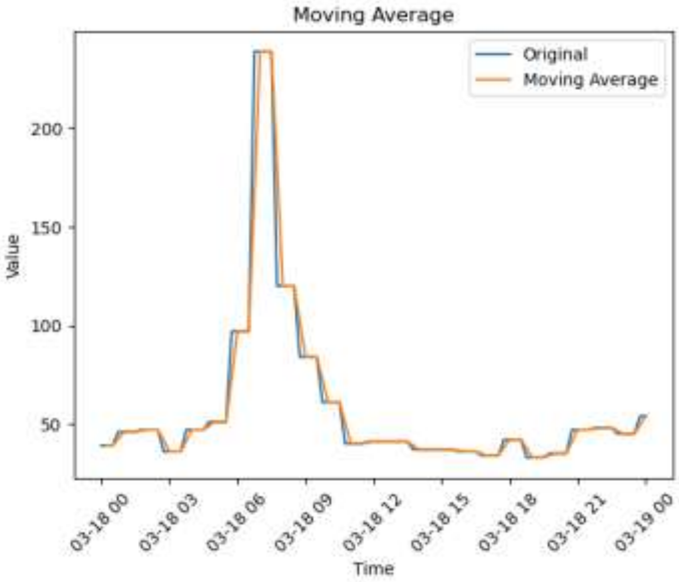


Fig: shows moving avg of PM2.5 from 13-03-23 to 14-03-23

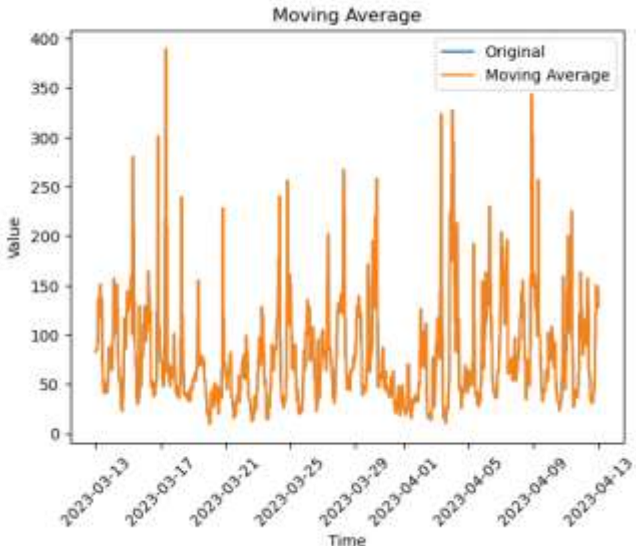


Fig: shows moving average of PM2.5 from 13-03-23 to 13-04-23.

Implementation and Evaluation of ARMA

Forecasting Model Autoregressive moving average model or ARMA is a collection of two separate processes namely autoregression and moving average. Similar to AR it uses an auto-regression component ‘p’ and combines it with an additional moving average component ‘q’ .

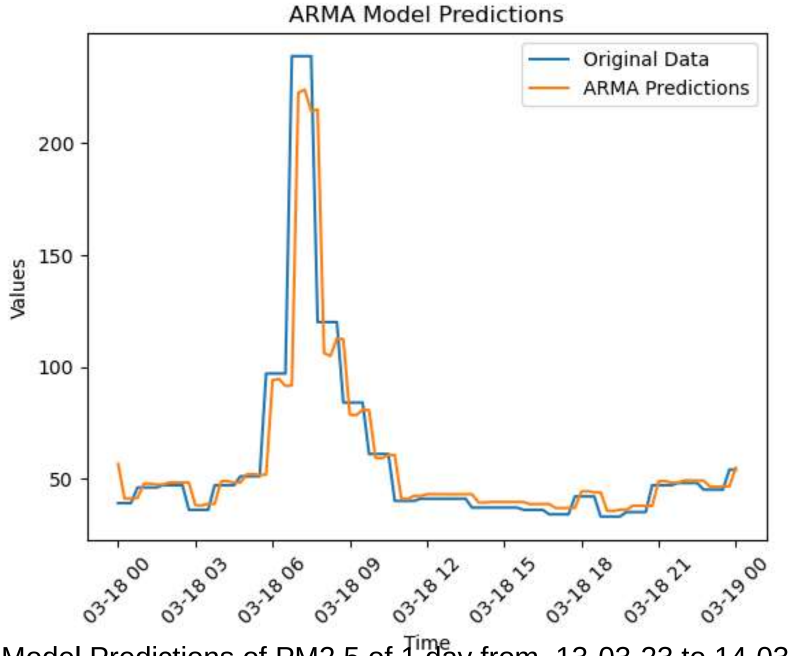


Fig : ARMA Model Predictions of PM2.5 of 1 day from 13-03-23 to 14-03-23

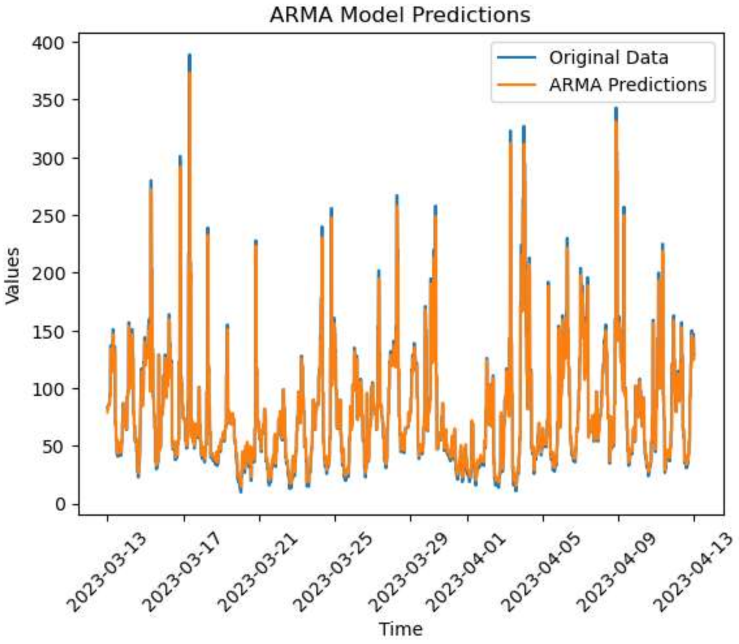


Fig : ARMA Model Predictions of PM2.5 of 1 month from 13-03-23 to 13-04-23

Evaluation:

Forecasts	RMSE	Time
1 Day	19.6044	0.8223 sec
30 Days	19.0871	1.8613 sec

Implementation and Evaluation of ARIMA Forecasting Model:

Considering the unsatisfactory performance of the ARMA model, an ARIMA model was implemented as an alternative. ARIMA (AutoRegressive Integrated Moving Average) is a widely used time series model that combines autoregressive (AR) and moving average (MA) components, along with a differencing term (d) to handle non-stationary data. The ARIMA model is represented as ARIMA(p, d, q), where p represents the order of the autoregressive component, d represents the degree of differencing, and q represents the order of the moving average component.

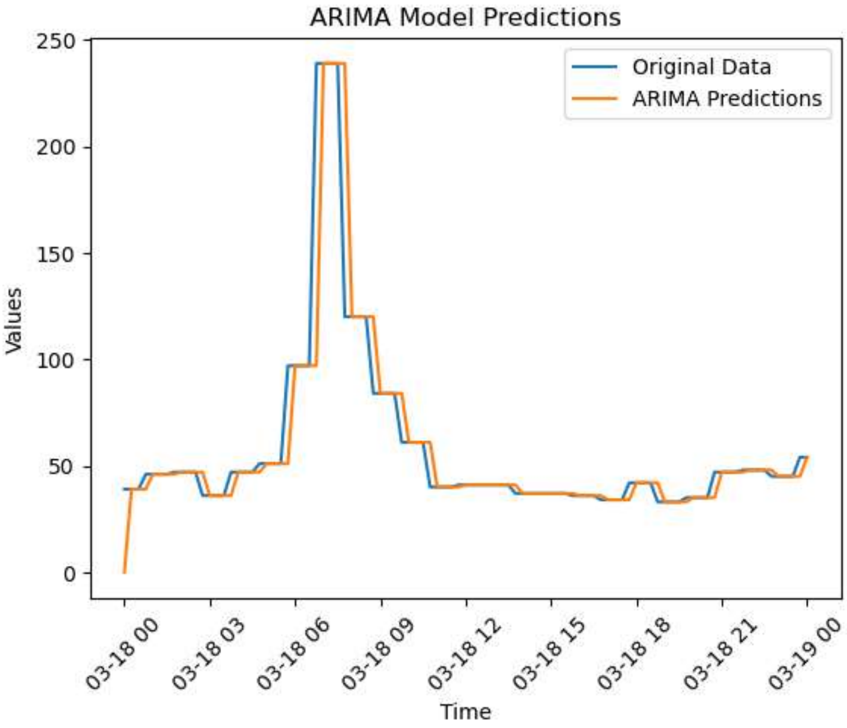


Fig: ARIMA Model Predictions of PM2.5 of 1 day from 13-03-23 to 14-03-23

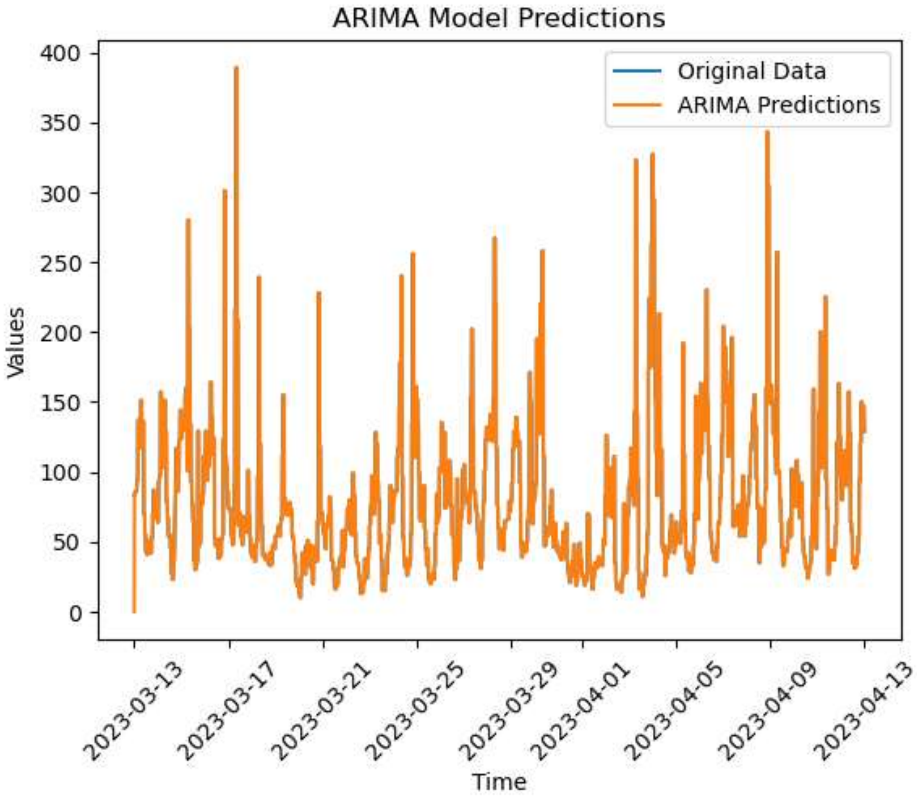


Fig : ARIMA Model Predictions of PM2.5 of 1 month from 13-03-23 to 13-04-23

Evaluation:
ARMA performed better than ARIMA for short term forecasting but the performance of ARMA was not good compared to ARIMA for 30 days forecast.

Forecasts	RMSE	Time
1 Day	20.5446	1.1081 sec
1 month	19.4982	1.0916 sec