COAL INDIA OPEN-PIT BLASTING

Inferential Statistics and Automation

INSTRUCTED BY: **PROF. TUSHAR SANDHAN**

Name:	**Deeksha Rawat**
Department:	Economic Sciences
Course:	EE798Q
Date of Sub:	25 June, 2023

1 COAL INDIA OPEN-PIT BLASTING

The Python Files for the below analysis can be found in the given link
https://drive.google.com/drive/folders/1DDNK5uUVF1xwY3YuZiaKQjS-rUhiGDE0?usp=sharing
The two main air pollutants in NCL coal fields are :

1. **Suspended particulate matter (SPM)**
2. **Reparable particulate matter (RPM)**

SPM and RPM concentrations are 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 [1] measurements available via multi-sensory system are PM10, PM2.5, SO2, NO2, NOx, CO, NH3, O3 and BENZENE.
The air pollution data set is obtained from the Singrauli Coalfield Pollution Control Board for coal India's (Singrauli Coalfield). The pollution is monitored during open-pit blasting. There are 13 columns overall in the air pollution data collection of pollutants that are available at intervals of 15 minutes.
It is clearly evident from Figure 1 that due to reasons like sensor failure, sensor-to-central-hub communication link failure, data packet loss etc., there are some missing sensory data for certain duration of the time. **The broken points represent the missing values.**
Let's check the total number of values missing for each column or feature.

```
0s
df.isnull().sum()

#                          0
From                       0
To (Interval: 15M)         0
PM10 (µg/m3)            1681
PM2.5 (µg/m3)            226
NO (µg/m3)              1369
NO2 (µg/m3)              416
NOX (ppb)                415
CO (mg/m3)               496
SO2 (µg/m3)             1451
NH3 (µg/m3)              326
Ozone (µg/m3)            453
Benzene (µg/m3)         6195
Time                       0
Date                       0
dtype: int64
```

Figure 1

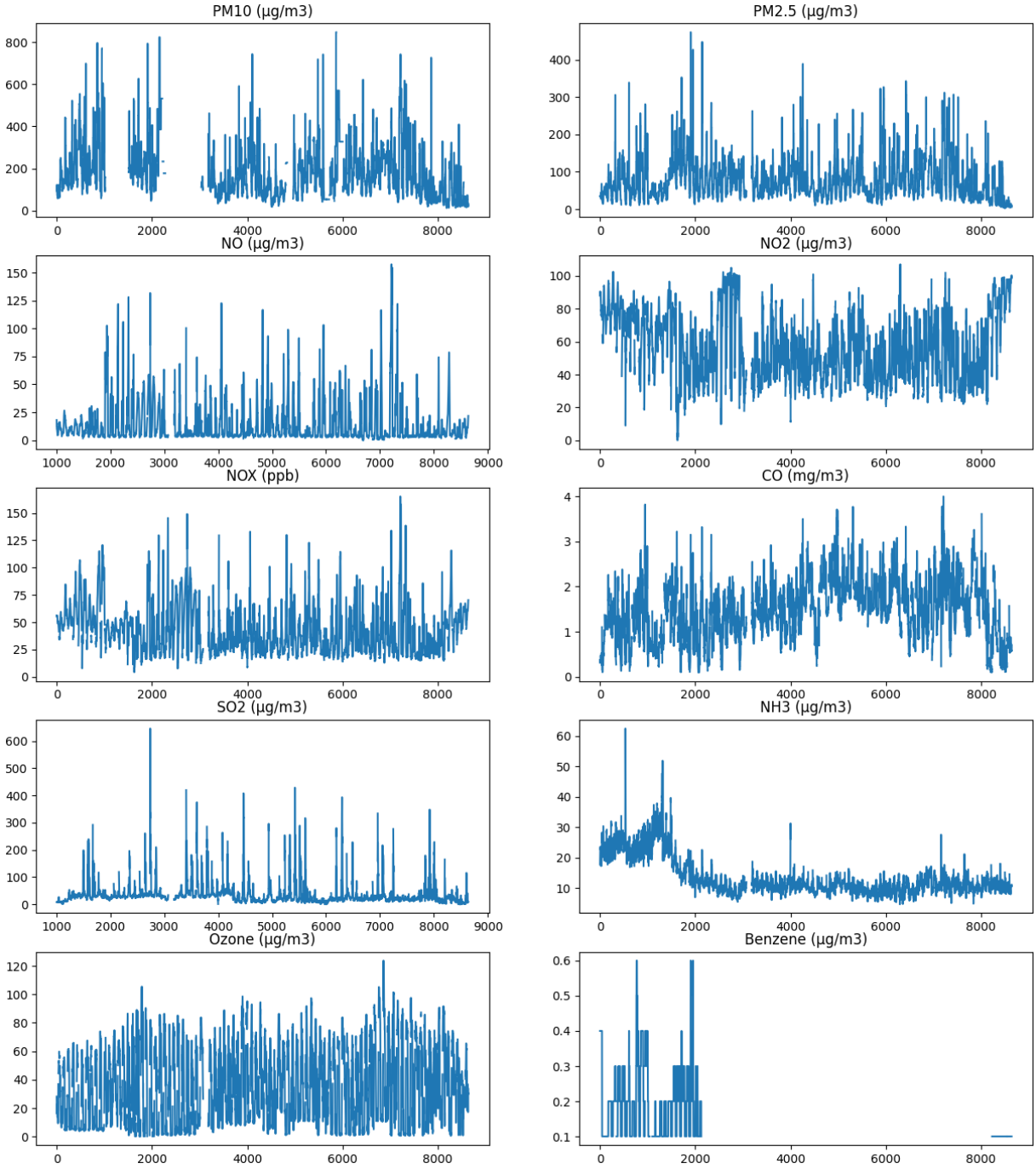


Figure 2

There are various approaches to handle the missing values. Few of them are mentioned below:

1. Deletion : This strategy entails eliminating any rows that contain missing values. Although straightforward to execute, if the missing data isn't Missing Completely at Random (MCAR), this approach may result in valuable information loss.

2. Constant Imputation : This technique substitutes all missing values with a constant, which might be a common value like zero or an unusual one that effectively establishes a new category for missing values. Although it's simple to apply, it can skew the data's distribution and possibly yield biased estimates. Unless there's a compelling reason to select a particular constant, this method is usually not recommended.

3. Interpolation : By cutting a slick curve through the time(ti) i = 1, 2, 3, ..., we are able to estimate PM10 and other column of data for any given time(t). Interpolation is used when the intended time(t) falls between the greatest and smallest of the time
(a) Linear Interpolation: This is basically like connecting two points in a dataset by drawing a line between them.
(b) Cubical Interpolation: It offers true continuity between the segments. As such it requires more than just the two endpoints of the segment but also the two points on either side of them.
(c) Spline Interpolation: Low-degree polynomials are used in each of the intervals in spline interpolation, which is similar to polynomial interpolation in that it selects the polynomial parts to fit together smoothly. The outcome is a function known as a spline.

Types of Missing Data: Understanding the reasons why data are missing is important for handling the remaining data correctly. If values are missing completely at random, the data sample is likely still representative of the population. But if the values are missing systematically, analysis may be biased.
i) Missing completely at random(MCAR): Values in a data set are missing completely at random (MCAR) if the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random.In other words, no systematic differences exist between participants with missing data and those with complete data.In these instances, the missing data reduce the analyzable population of the study and consequently, the statistical power, but do not introduce bias: when data are MCAR, the data which remain can be considered a simple random sample of the full data set of interest. MCAR is generally regarded as a strong and often unrealistic assumption.
Missing-ness is independent of the data:

$$P(R_i|Y_i) = P(R_i)$$

where R_i : missing data indicator, Y_i : p variables for unit i

$$Y_i = (Y_{i,o}, Y_{i,m})$$

ii) Missing at random(MAR): When data are MAR, the fact that the data are missing is systematically related to the observed but not the unobserved data.Complete case analyses, which are based on only observations for which all relevant data are present and no fields are missing, of a data set containing MAR data may or may not result in bias. If the complete case analysis is biased, however, proper accounting for the known factors can produce unbiased results in analysis.
Missing-ness depends only on the observed data:

$$P(R_i|Y_i) = P(R_i|Y_{i,o})$$

iii) Missing not at random (MNAR): When data are MNAR, the fact that the data are missing is systematically related to the unobserved data, that is, the missingness is related to events or factors which are not measured by the researcher. As with MAR data, complete case analysis of a data set containing MNAR data may or may not result in bias; if the complete case analysis is biased, however, the fact that the sources of missing data are themselves unmeasured means that (in general) this issue cannot be addressed in analysis and the estimate of effect will likely be biased.
Missing-ness depends only on the missing data:

$$P(R_i|Y_i) \neq P(R_i|Y_{i,o})$$

2 VISUALISING THE MISSING VALUES

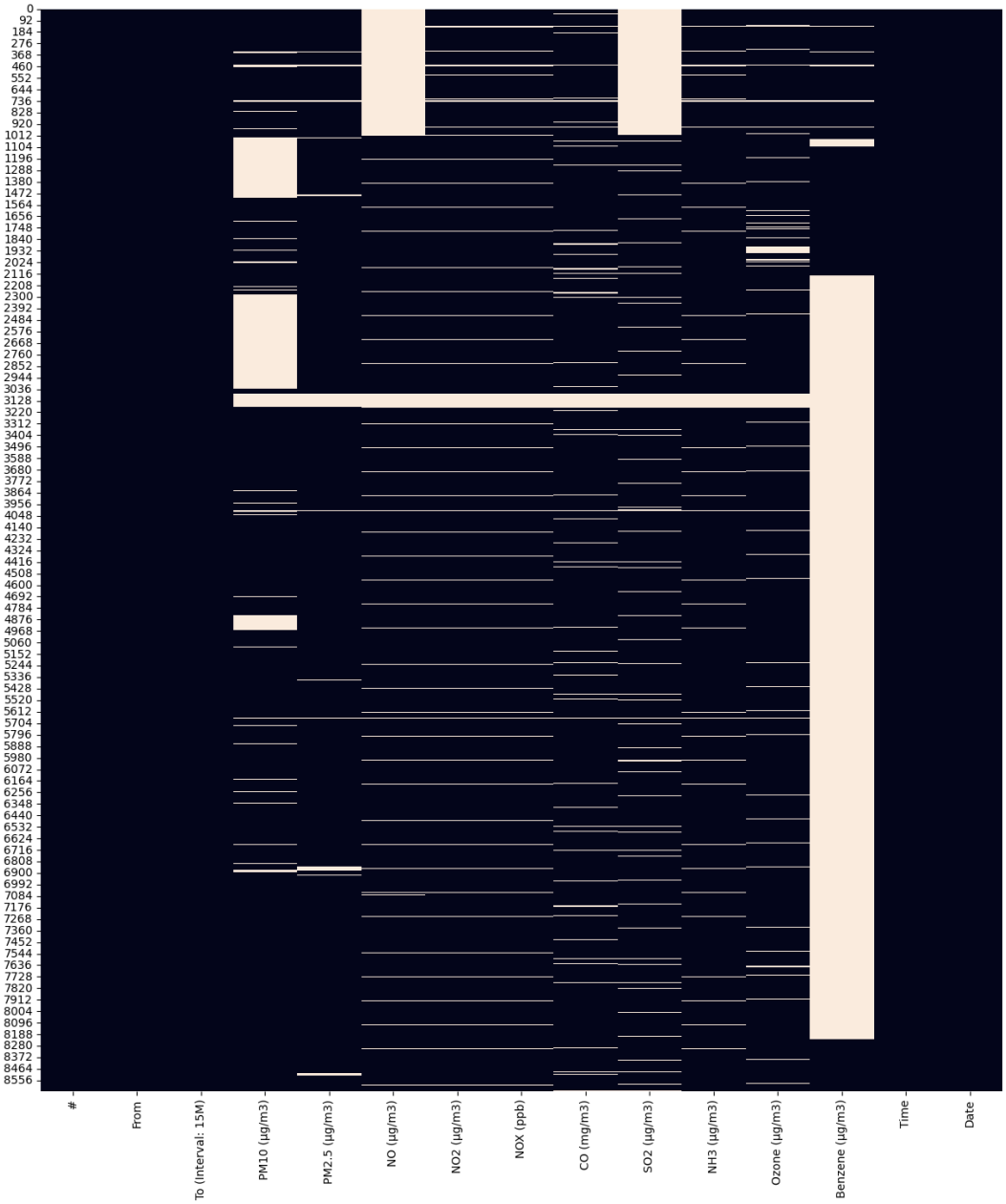


Figure 3

It is clearly evident from the heatmap that **Benzene** has the largest amount of data missing.

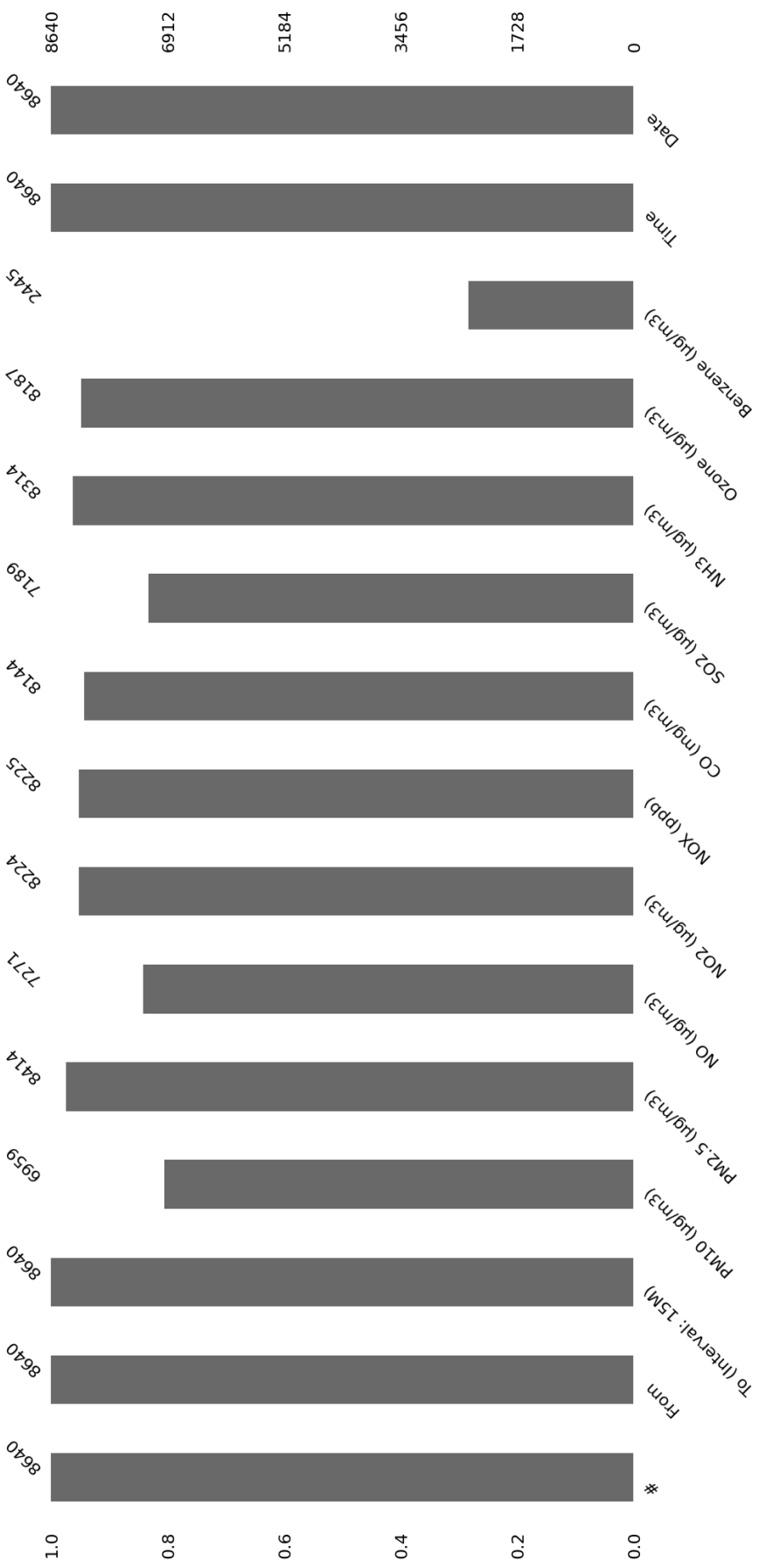

8640
6912
5184
3456
1728
0
8640
Date
8640
Time
2445
Benzene (ug/m3)
8187
Ozone (ug/m3)
8314
NH3 (ug/m3)
7189
SO2 (ug/m3)
8144
CO (mg/m3)
8225
NOX (ppb)
8224
NO2 (ug/m3)
7271
NO (ug/m3)
8414
PM2.5 (ug/m3)
6959
PM10 (ug/m3)
8640
To (interval: 15M)
8640
From
8640
*
1.0
0.8
0.6
0.4
0.2
0.0

Figure 4

Deletion of the rows with missing data might lead to throwing away the useful information. There are around 13,000 data points which is quite a large number.
Imputation with Zero : Upon replacing the missing values with zero, and then applying the ARIMA model, we observed that the data distribution got skewed. The R2 score was a highly negative value which is definitely not good for better predictions.
i) Checked for the stationarity of each column using Augmented Dickey fuller Test.
if p-value > 0.05 , the data is non-stationary.
if p-value ≤ 0.05 , the data is stationary.
ii) To find the total number of significant lags, we plotted the ACF and PACF plots.

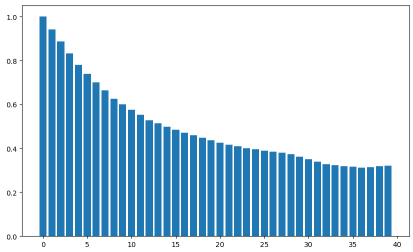


Figure 5: ACF Plot for PM10

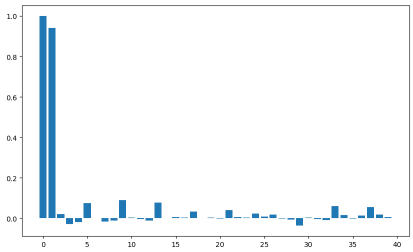


Figure 6: PACF Plot for PM10

However to get a more exact order of the ARIMA model, we used the auto-arima tool. This tool iterates over different lags and finds the order for which the AIC value is minimum.

```
[41] stepwise_fit = auto_arima(new_df['PM2.5 (µg/m3)'],trace=True, suppress_warnings=True)

Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=inf, Time=51.45 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=76986.389, Time=0.45 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=76988.345, Time=0.40 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=76988.346, Time=1.23 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=76984.390, Time=0.24 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=76990.335, Time=1.97 sec

Best model:  ARIMA(0,1,0)(0,0,0)[0]
Total fit time: 55.786 seconds
```

Figure 7

iii) Build an ARIMA model on the given data and predicted on the test data. Plots are obtained are shown in figure 8. The R2 Score of the model came out to be negative(-0.17) which means, it is not a good model

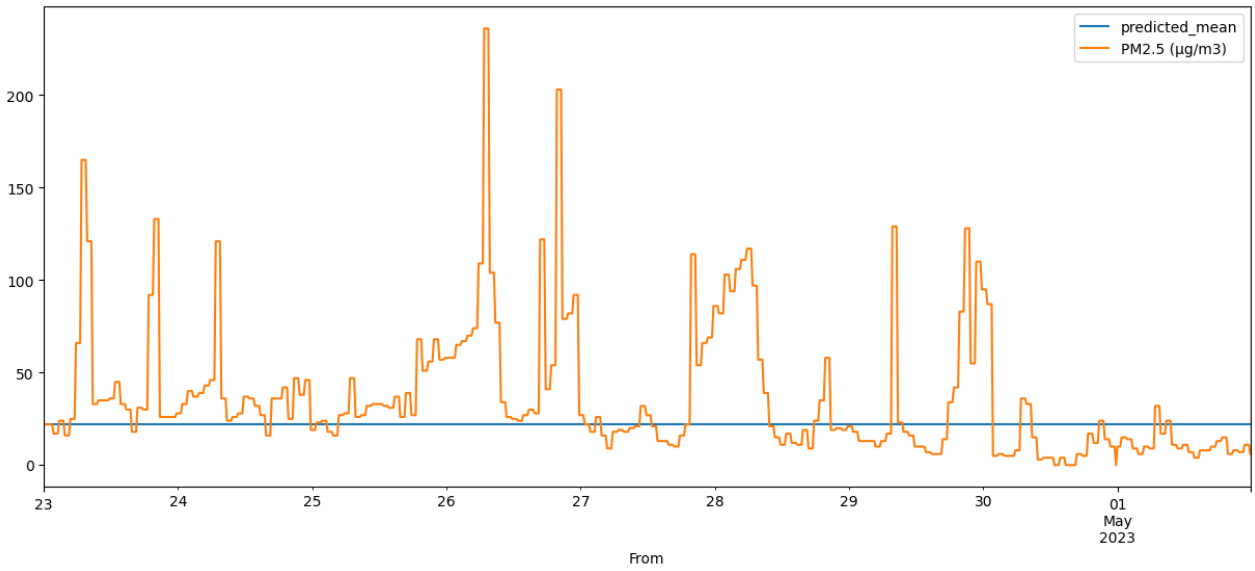


Figure 8

Interpolation over the entire data set won't be very useful, since data is missing in a significant amount. Upon using different methods of interpolation, we got the following results.

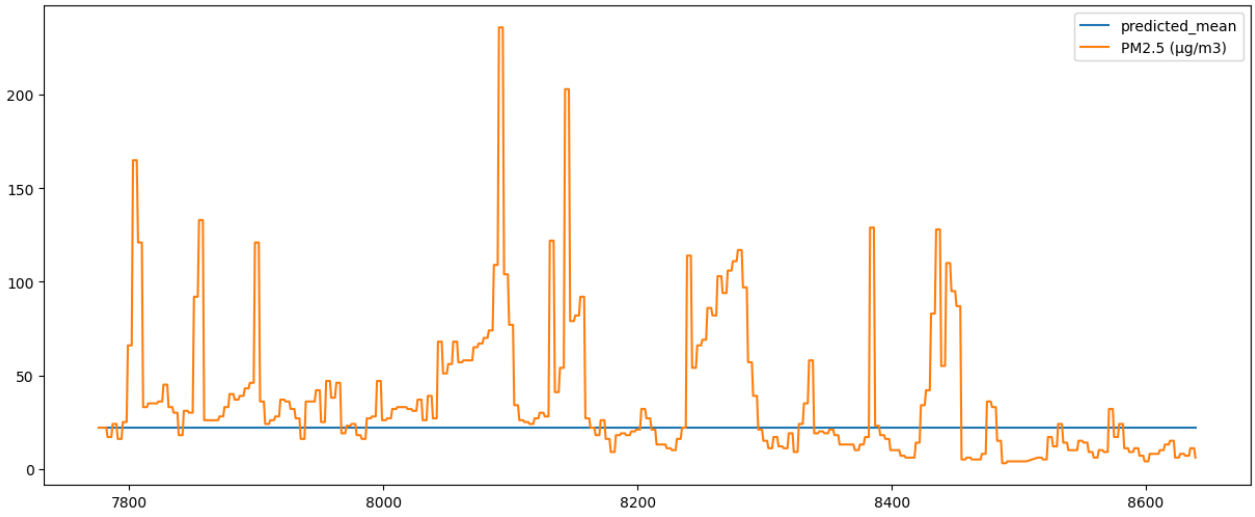


Figure 9

The R2 score for this model came out to be -0.18, which was also not that good.

3 Building a model to fill in the Missing Values :

Coal Blasting activity is expected to increase or decrease during particular time of the day. Therefore, it is logically fair to predict the missing values by extracting the values of concentration of gases during each time stamp and then , building an ARIMA model over those values. We considered the past 2 values to evaluate the missing values

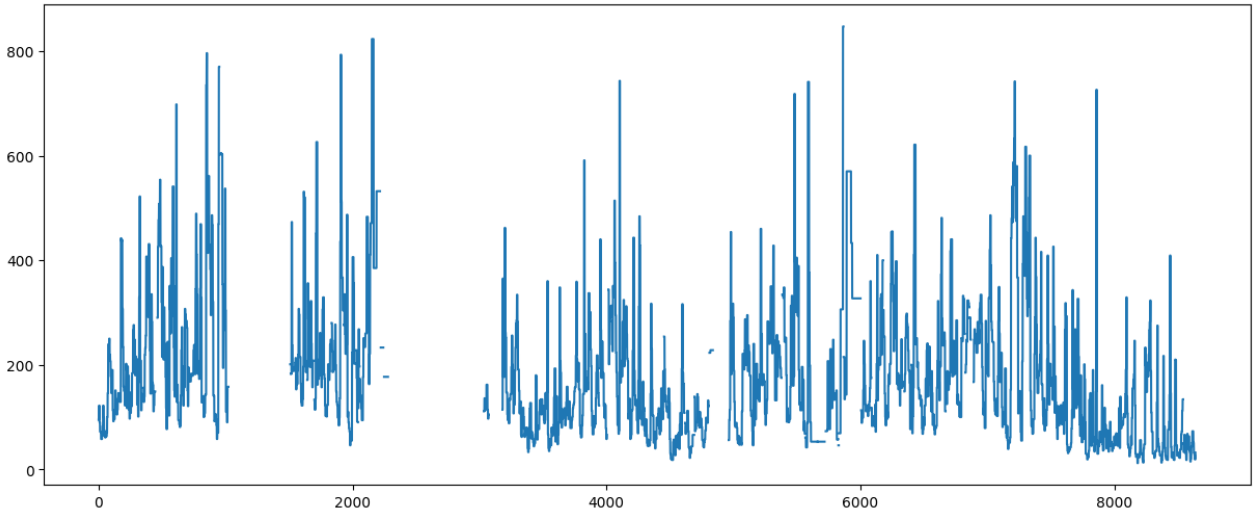


Figure 10

After applying the ARIMA model, we get the following picture.

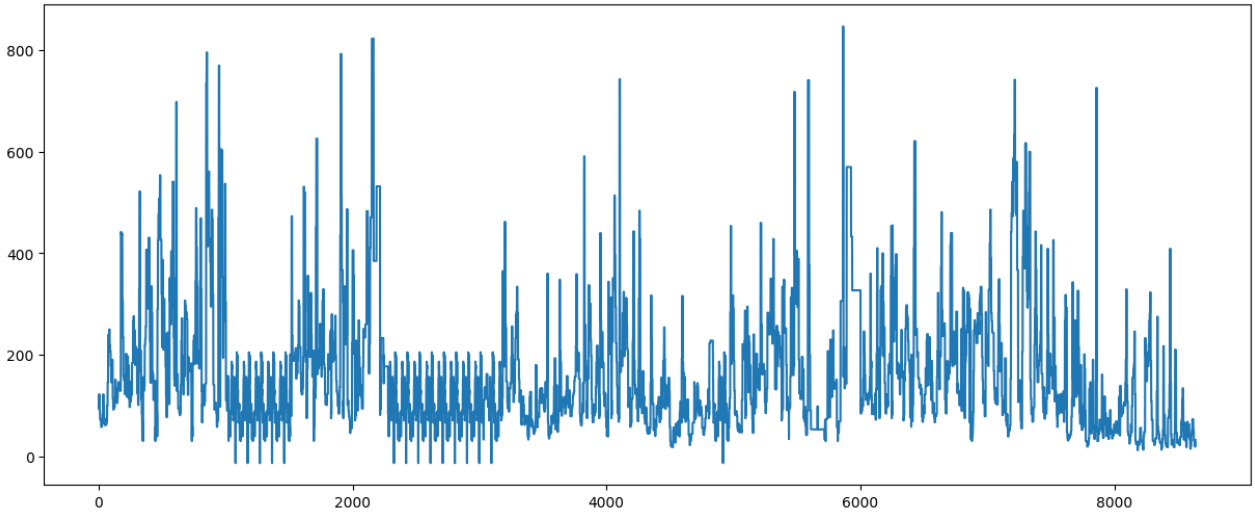


Figure 11

To get a clear picture, consider a particular timestamp , say 16:45:00. When the values are not filled, the gaps are clearly observable as shown in figure12.

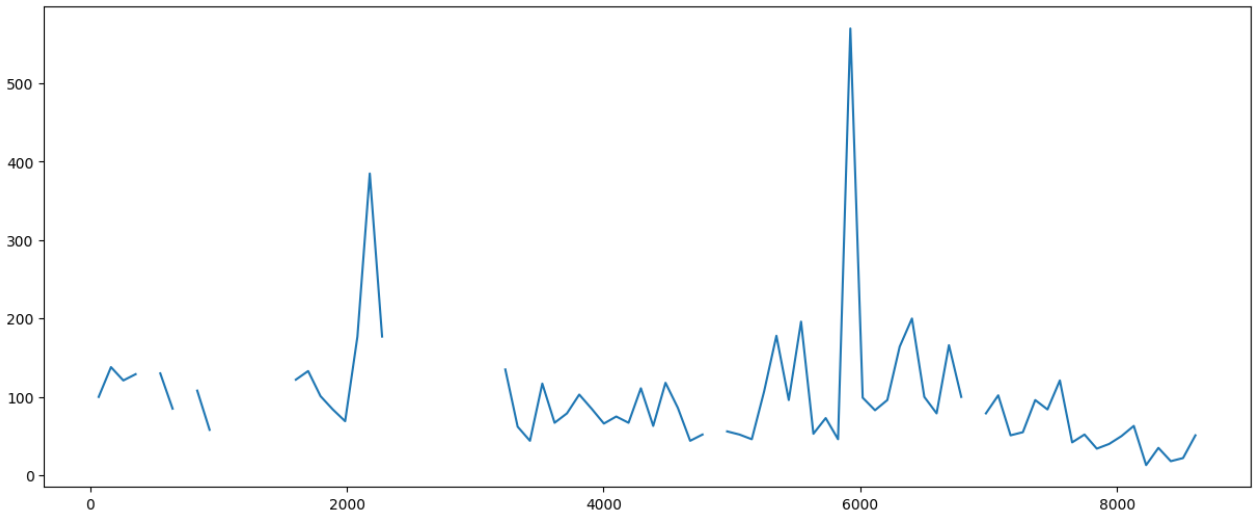


Figure 12

Using the ARIMA model, we get the missing values filled, as shown in figure13.

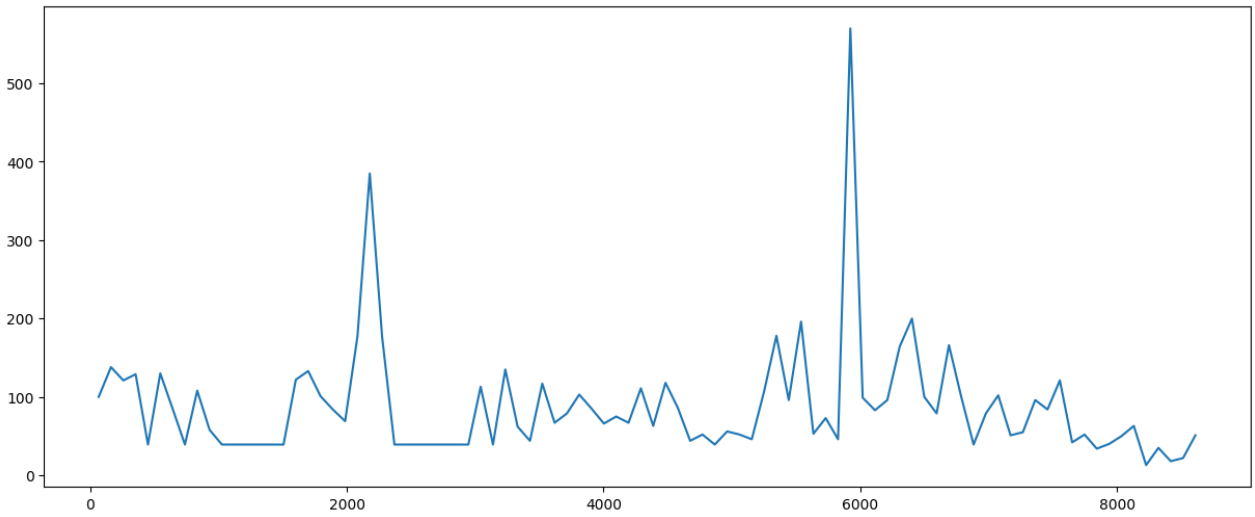


Figure 13

However, Benzene has a huge chunk of data missing, filling those missing values using ARIMA model is time consuming and hence leading to Run Time Error. Therefore, for Benzene we substitute the Null values with mean. After filling in all the missing values, the following graphs are obtained.

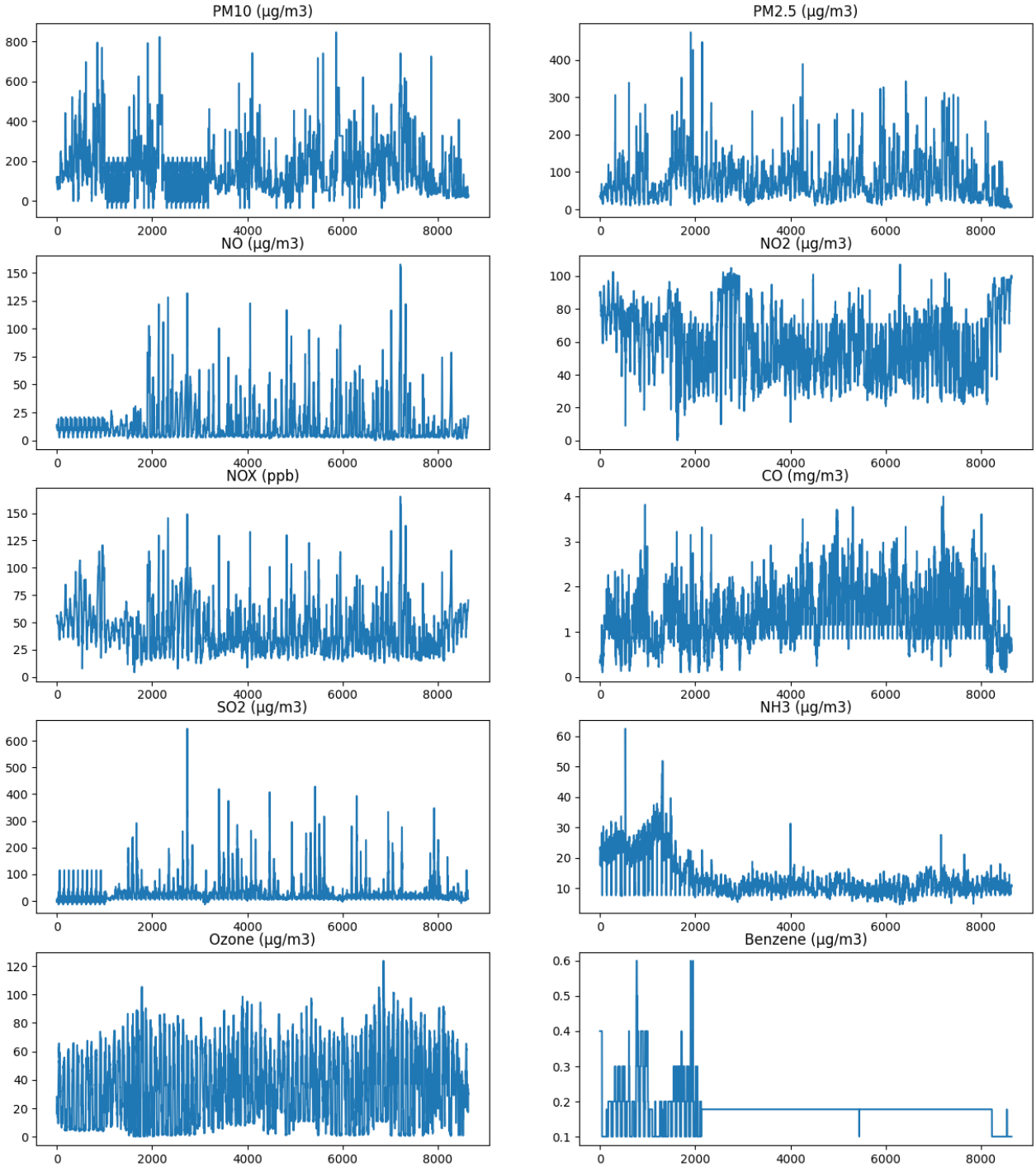


Figure 14

4 Statistical Inference :

In order to determine the combined effect of the various pollutants, we use the Air Quality Index (AQI) measure.

O_3 (ppb)	O_3 (ppb)	$PM_{2.5}$ (µg/m^3)	PM_{10} (µg/m^3)	CO (ppm)	SO_2 (ppb)	NO_2 (ppb)	AQI	AQI
C_{low} – C_{high} (avg)							I_{low} – I_{high}	Category
0–54 (8-hr)	—	0.0–12.0 (24-hr)	0–54 (24-hr)	0.0–4.4 (8-hr)	0–35 (1-hr)	0–53 (1-hr)	0–50	Good
55–70 (8-hr)	—	12.1–35.4 (24-hr)	55–154 (24-hr)	4.5–9.4 (8-hr)	36–75 (1-hr)	54–100 (1-hr)	51–100	Moderate
71–85 (8-hr)	125–164 (1-hr)	35.5–55.4 (24-hr)	155–254 (24-hr)	9.5–12.4 (8-hr)	76–185 (1-hr)	101–360 (1-hr)	101–150	Unhealthy for sensitive groups
86–105 (8-hr)	165–204 (1-hr)	55.5–150.4 (24-hr)	255–354 (24-hr)	12.5–15.4 (8-hr)	186–304 (1-hr)	361–649 (1-hr)	151–200	Unhealthy
106–200 (8-hr)	205–404 (1-hr)	150.5–250.4 (24-hr)	355–424 (24-hr)	15.5–30.4 (8-hr)	305–604 (24-hr)	650–1249 (1-hr)	201–300	Very unhealthy
—	405–504 (1-hr)	250.5–350.4 (24-hr)	425–504 (24-hr)	30.5–40.4 (8-hr)	605–804 (24-hr)	1250–1649 (1-hr)	301–400	Hazardous
—	505–604 (1-hr)	350.5–500.4 (24-hr)	505–604 (24-hr)	40.5–50.4 (8-hr)	805–1004 (24-hr)	1650–2049 (1-hr)	401–500	

Figure 15

The AQI for each pollutant is calculated using the given formula equation :

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low})$$

Analysis for the blasting time

Finding the blasting trigger time in the time interval 13:45:00 to 14:45:00, we get the following results -

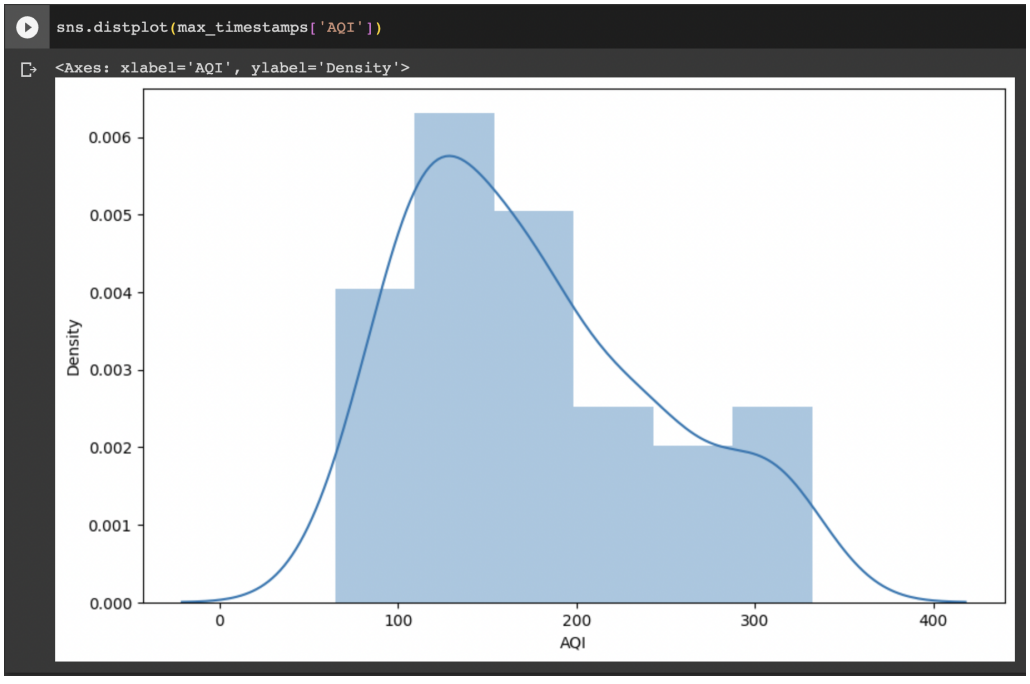


Figure 16

During the Blasting time, most of the days have their maximum AQIs in the range of 100 to 200. Let us plot the QQ plot if it follows normal distribution or not.

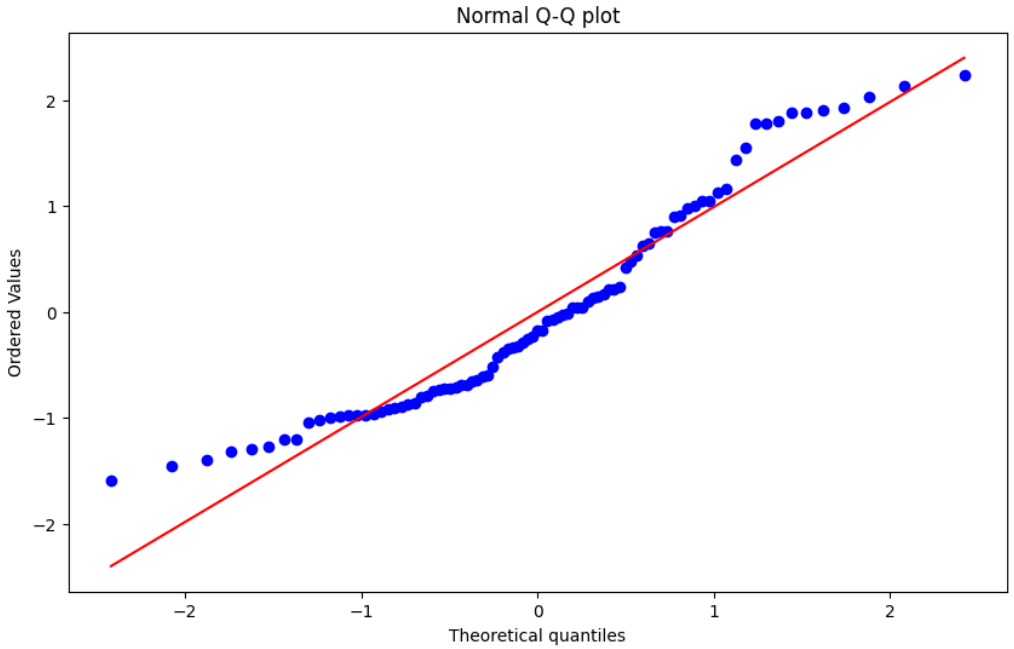
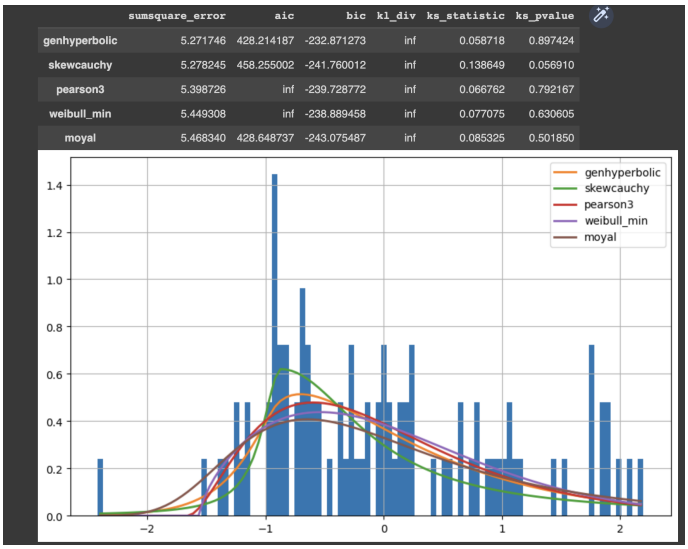


Figure 17

It does not follow strictly normal distribution pattern. Using Fitter, we find what distribution is followed by the blast trigger time during blasting time of the day.



	sumsquare_error	aic	bic	kl_div	ks_statistic	ks_pvalue
genhyperbolic	5.271746	428.214187	-232.871273	inf	0.058718	0.897424
skewcauchy	5.278245	458.255002	-241.760012	inf	0.138649	0.056910
pearson3	5.398726	inf	-239.728772	inf	0.066762	0.792167
weibull_min	5.449308	inf	-238.889458	inf	0.077075	0.630605
moyal	5.468340	428.648737	-243.075487	inf	0.085325	0.501850

Figure 18

The Blast Trigger times AQI follows **Generalised hyperbolic** Probability Distribution Function.

Generalised hyperbolic

Parameters	λ (real) α (real) β asymmetry parameter (real) δ scale parameter (real) μ location (real) $\gamma = \sqrt{\alpha^2 - \beta^2}$
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{(\gamma/\delta)^\lambda}{\sqrt{2\pi}K_\lambda(\delta\gamma)} e^{\beta(x-\mu)} \times \frac{K_{\lambda-1/2}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)}{\left(\sqrt{\delta^2+(x-\mu)^2}/\alpha\right)^{1/2-\lambda}}$
Mean	$\mu + \frac{\delta\beta K_{\lambda+1}(\delta\gamma)}{\gamma K_\lambda(\delta\gamma)}$
Variance	$\frac{\delta K_{\lambda+1}(\delta\gamma)}{\gamma K_\lambda(\delta\gamma)} + \frac{\beta^2\delta^2}{\gamma^2}\left(\frac{K_{\lambda+2}(\delta\gamma)}{K_\lambda(\delta\gamma)} - \frac{K^2_{\lambda+1}(\delta\gamma)}{K^2_\lambda(\delta\gamma)}\right)$
MGF	$\frac{e^{\mu z}\gamma^\lambda}{(\sqrt{\alpha^2-(\beta+z)^2})^\lambda} \frac{K_\lambda(\delta\sqrt{\alpha^2-(\beta+z)^2})}{K_\lambda(\delta\gamma)}$

Figure 19

Analysis over each Timestamp

The Overall AQI is the maximum of all the AQIs calculated for each pollutant.
Using the above information, the AQI value for all the months at each and every timestamp is calculated. The given plots are obtained :

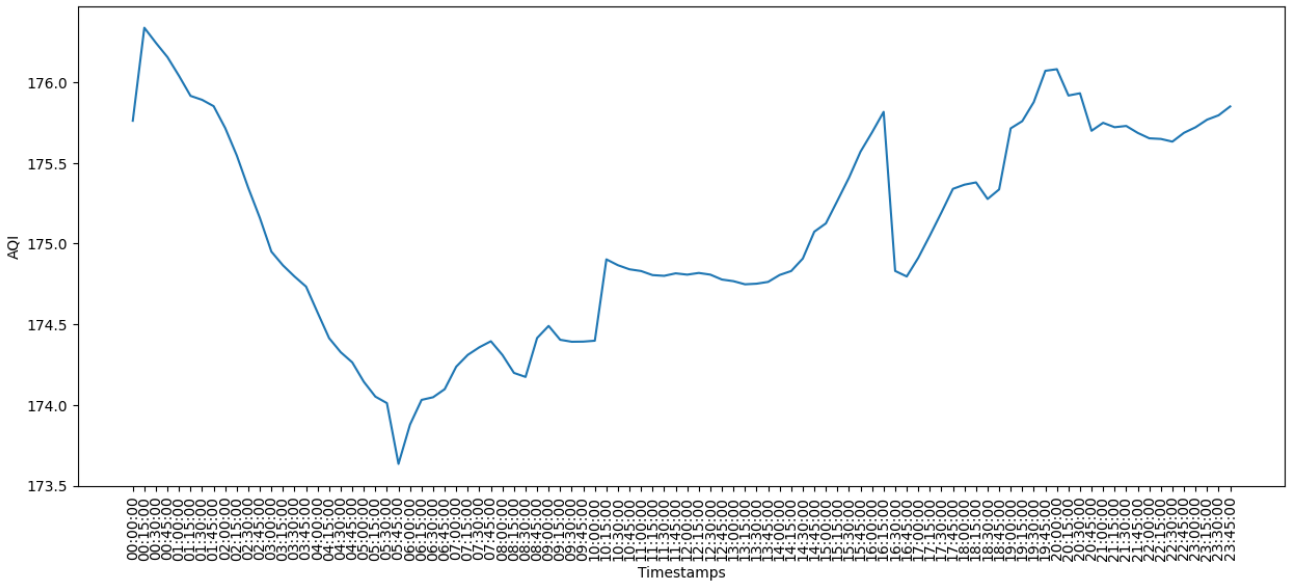


Figure 20

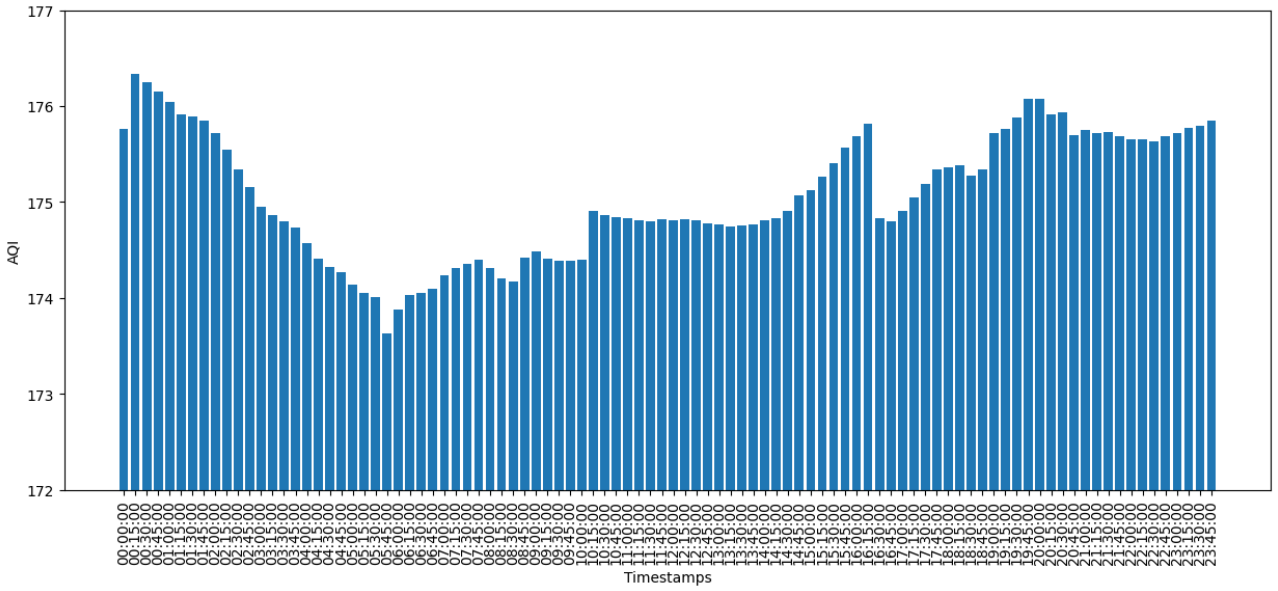


Figure 21

Blasting time in coal India is 13:45 pm to 14:45 pm major effect on air pollution. According to the plots, there is a rise observed during this time interval. The sensors detect the Pollutants during peak blasting hrs, leading to rise in AQI and hence Air pollution. **Therefore, it is validated from the above plots that the peak blasting time is 13:45 pm to 14:45 pm.** The rise is not obsereved at that exact time because the sensors detect the pollutants after some time. **Calculating the probability of the blast to occur in the time interval 14:15:00 to 14:30:00**
Since, the sensors don't detect the pollution instantly, so, the window to assess the pollution is increased to the time interval 14:00:00 to 17:30:00. In 89 days, 5 days have blasting time in this interval. Therefore, Poisson parameter for the above case is

$$\frac{5}{89} \approx 0.05618$$

Therefore, the probability of having at least one blast in this time interval is :

$$1 - P(X = 0) = 1 - \frac{e^{-0.05617}(-0.05617)^0}{0!} = 0.056179$$

The Probability distribution function followed by the Air Pollution measure is **Generalized normal distribution**.

	sumsquare_error	aic	bic	kl_div	ks_statistic	ks_pvalue
gennorm	10.176142	327.230211	-201.759968	inf	0.091415	0.375983
dweibull	10.257288	339.051903	-200.997484	inf	0.101563	0.257384
johnsonsb	10.309876	316.725248	-195.942220	inf	0.104463	0.229119
gengamma	10.324528	314.410217	-195.805877	inf	0.123036	0.100379
gausshyper	10.330638	320.888025	-186.620388	inf	0.095928	0.319432

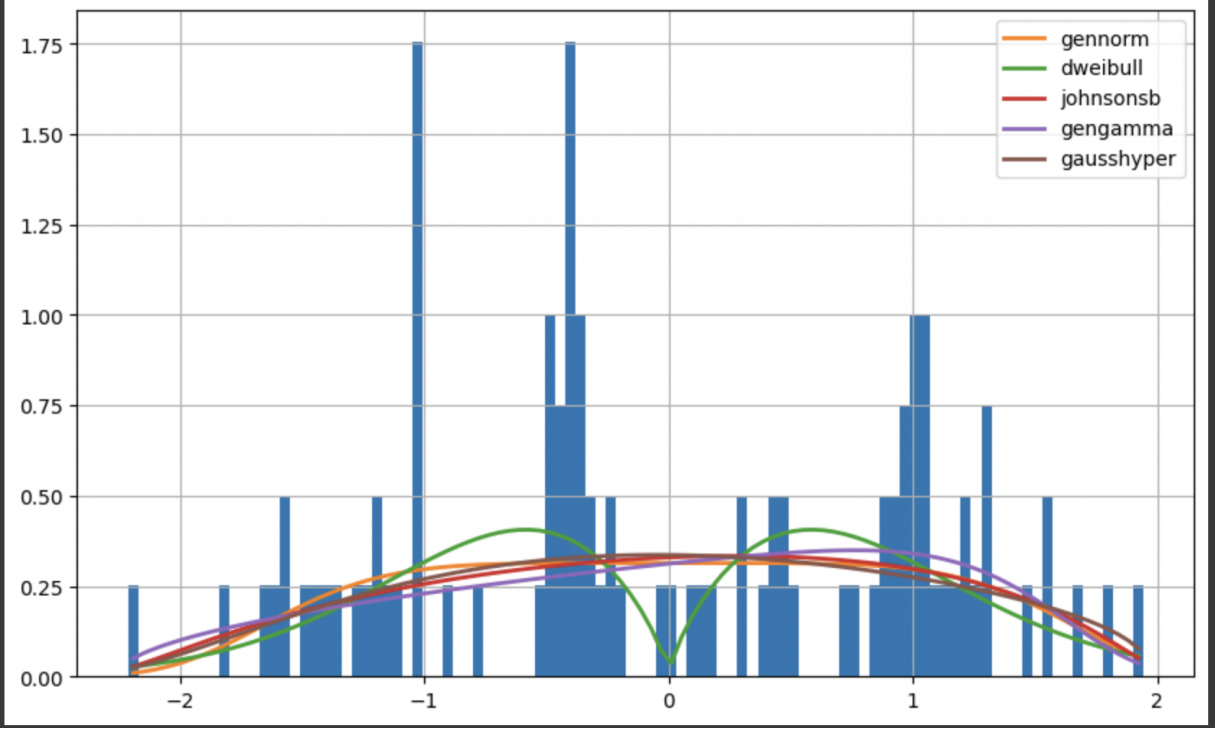


Figure 22

The QQ-plot for the AQI data is shown in the figure below. It clearly depicts that it does not follow Normal Distribution :

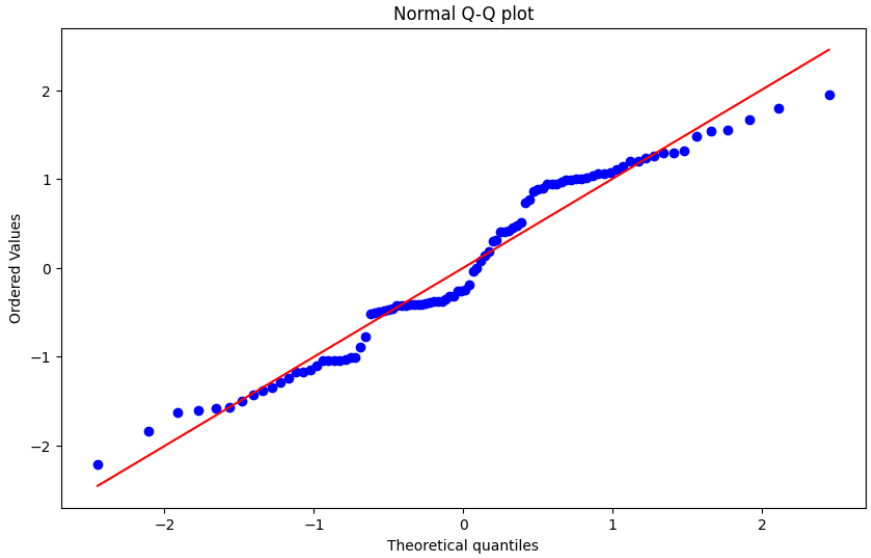


Figure 23

What is Generalized Normal Distribution?

Parameters	μ location (real) α scale (positive, real) β shape (positive, real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\beta}{2\alpha\Gamma(1/\beta)}\, e^{-(\lvert x-\mu\rvert/\alpha)^{\beta}}$ Γ denotes the gamma function

Figure 24: PDF of Generalised Normal Distribution

Most of the timestamps have the AQIs in the range of 174 to 176. This can be seen from the histogram below -

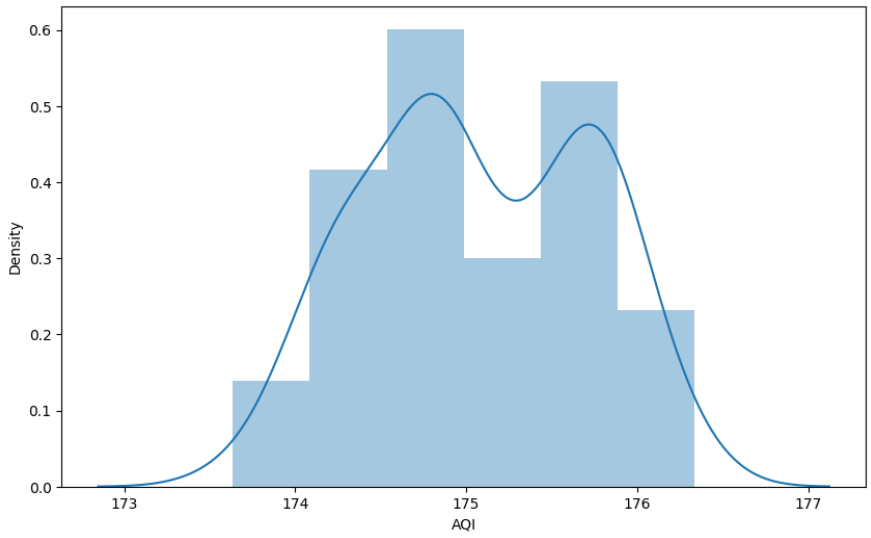


Figure 25: PDF of Generalised Normal Distribution

5 Problem setting and prediction

Explanatory analysis: Attempts to understand the air pollution data and the relationships within it, as well as cause and effect air pollution coal India at the time of blasting?
Let's make a pairplot to understand which factors contribute the most to the AQI measure. The plot is shown in figure 23.

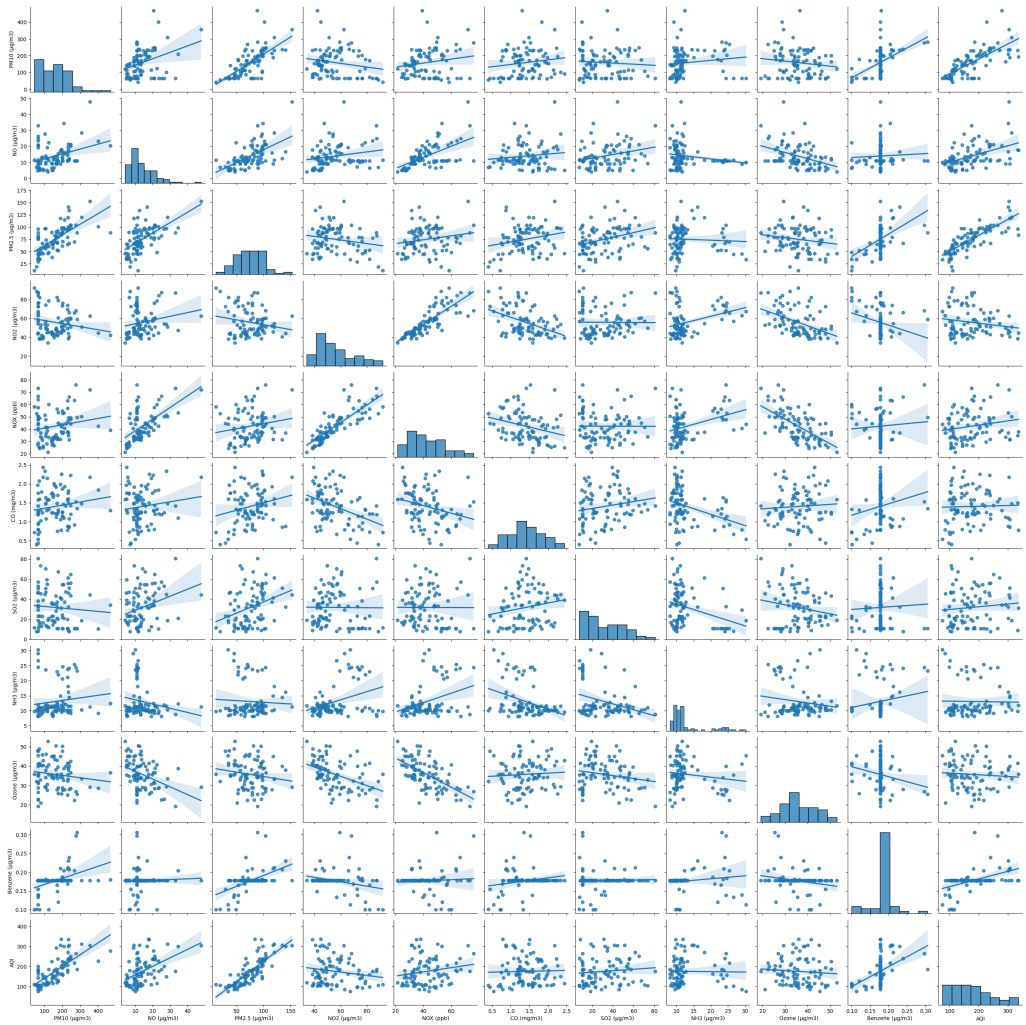


Figure 26

A Regression model is used to assign weights to different pollutants. The results from the linear Regression model are quite satisfactory with an R2 score of 0.723.

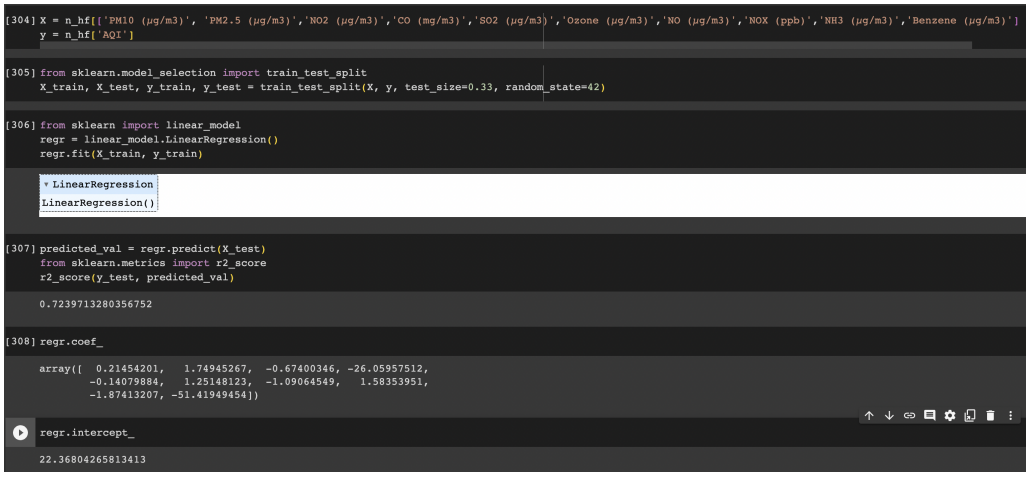


Figure 27

Hence, the following weights are provided to different pollutants:

Pollutant	Weight
PM10 (µg/m3)	0.21454201
PM2.5 (µg/m3)	1.74945267
NO2 (µg/m3)	-0.67400346
CO (mg/m3)	-26.05957512
SO2 (µg/m3)	-0.14079884
Ozone (µg/m3)	1.25148123
NO (µg/m3)	-1.09064549
NOX (ppb)	1.58353951
NH3 (µg/m3)	-1.87413207
Benzene (µg/m3)	-51.41949454

The Bias to this model is **22.368**.
Based on the weights, it can be concluded that PM10, NOX, Ozone, PM2.5 are a good indicator of AQI whereas Benzene, CO are not very helpful in indicating the Air Pollution. Benzene had a lot of missing data points, therefore it cannot give better results as far as AQI calculation is concerned.
For a better visualisation, heatmaps for the correlation matrix is shown below :

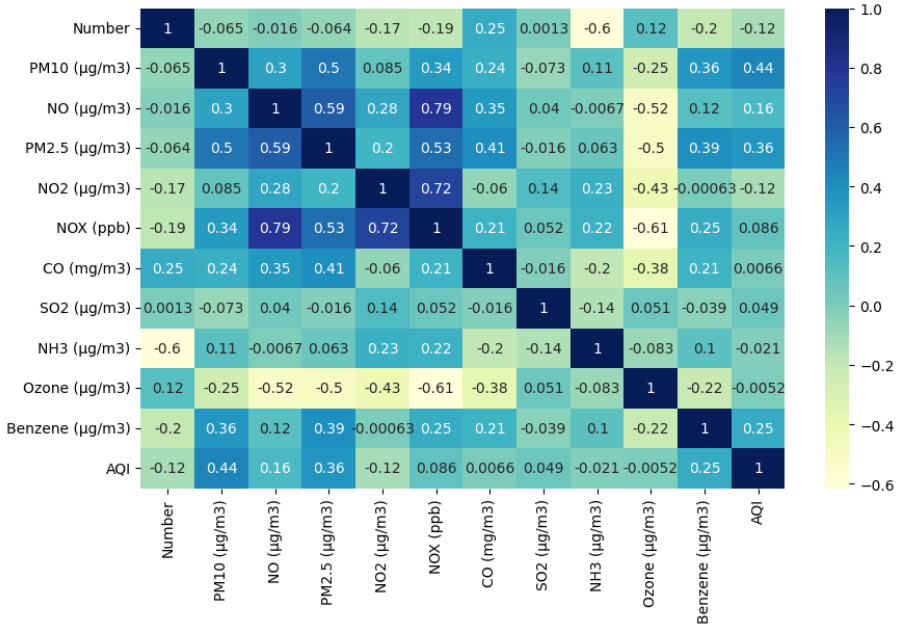


Figure 28

Exploratory Analysis: Highlight the main characteristics of the time series air pollution, usually in a visual format.
Since, AQI is used as the measure for Air Pollution. We get the following plot of air pollution.
Finding the various components of AQI :

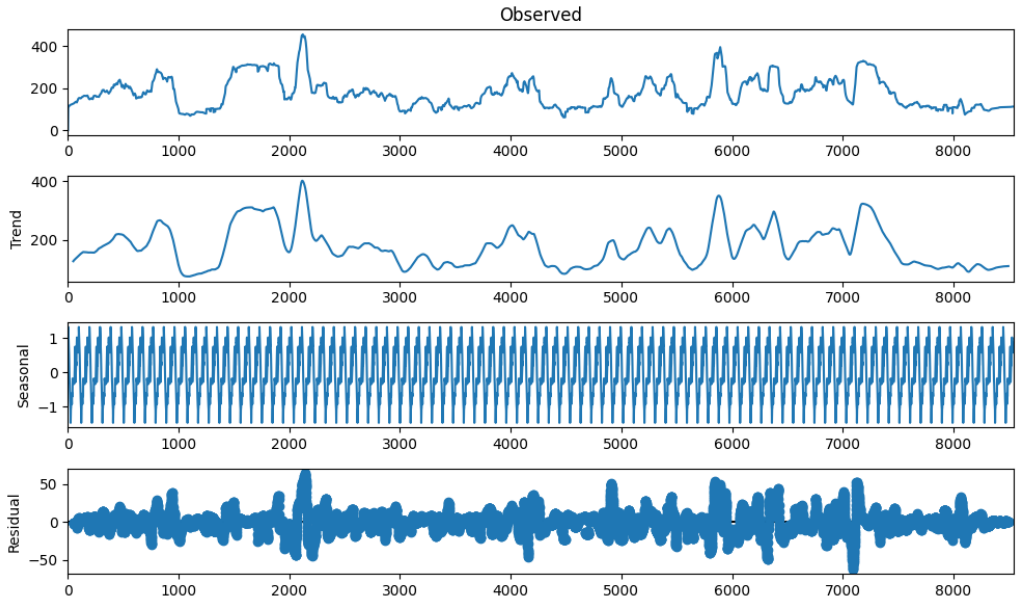


Figure 29

Based on the above plot, there is no specific trend observed. Some sort of seasonality is evident.The Autocorrelation plot is used to know if seasonality is present or not.

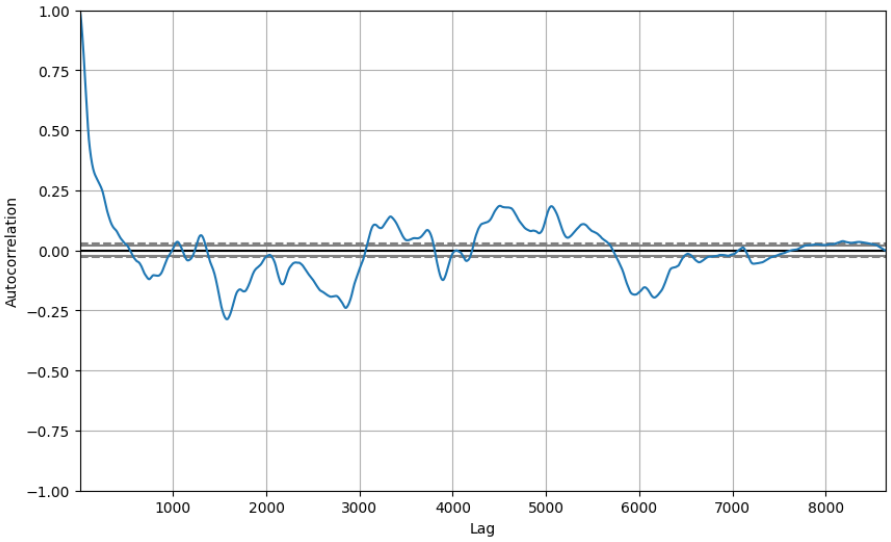


Figure 30

To check the stationarity quantitatively, Augmented Dickey Fuller Test is used.

```
from statsmodels.tsa.stattools import adfuller

def adfuller_test(aqi):
    result = adfuller(aqi)
    labels = ['ADF test statistics', 'P-value', '#Lags used', 'Number of observation used']
    for value, label in zip(result, labels):
        print(label+' : '+str(value))
    if result[1] <= 0.05:
        print('Strong evidence against the null hypothesis (Ho), Reject the null hypothesis, Data has no unit root and is stationary')
    else:
        print('Weak evidence against the null hypothesis (Ho), time series has a unit root, indicating it is non stationary. ')


adfuller_test(hi['AQI'])
```

```
ADF test statistics : -5.087759564174565
P-value : 1.4835643942896133e-05
#Lags used : 13
Number of observation used : 8626
Strong evidence against the null hypothesis (Ho), Reject the null hypothesis, Data has no unit root and is stationary
```

Figure 31

Using Augmented Dickey Fuller Test, the AQI series came out to be Stationary Series.
Forecasting We built an MA model of order 10 for forecasting. Based on the prediction model built, we get the following results-

```
rrt[['AQI','rolling_av','forecast']].plot(lw=2);
```

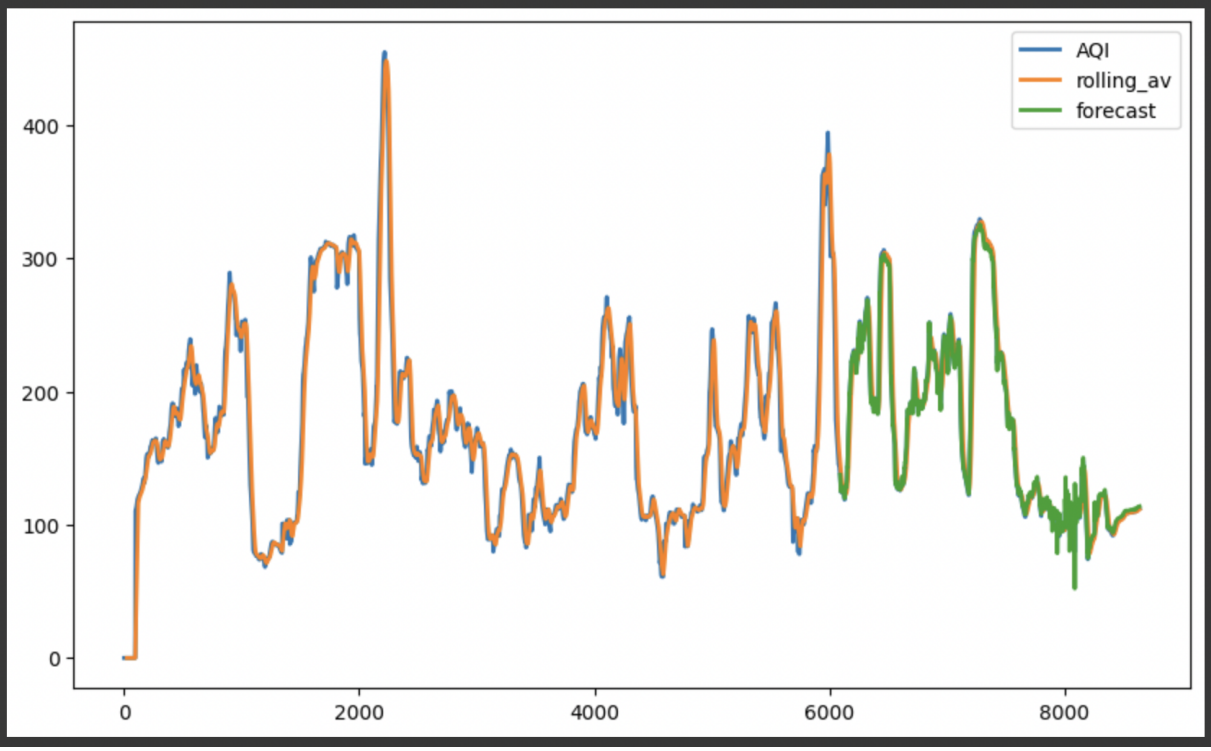


Figure 32

We get a better prediction model with an R2 score of 0.998.

```
pred = rrt['forecast'].dropna()
```

```
from sklearn.metrics import r2_score
r2_score(test_data,pred)
```

```
0.9984709590734517
```

Figure 33

6 References

- https://www.analyticsvidhya.com/blog/2021/10/end-to-end-introduction-to-handling-missing-values/
- https://en.wikipedia.org/wiki/Air_quality_index
- https://en.wikipedia.org/wiki/Open-pit_mining
- https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/
- https://pandas.pydata.org/docs/user_guide/index.html#user-guide