Business Intelligence Report for

ITL601 - BI Lab
Semester VI

Name : Shubham Talawadekar

Roll no : 65

Seat No : 3665

Bachelor's Degree in Information Technology

Don Bosco Institute of Technology, Mumbai 400070 (Affiliated to the University of Mumbai) Department of Information Technology

Academic Year 2022 – 2023

Experiment No : 10

Date : 11/03/23

Title : MINI – PROJECT - BI Report

Problem Definition

The problem at hand is to develop a system to analyze pollutant data collected hourly over the past three years from three locations in Mumbai - Chakala, Kurla, and the Airport. The dataset includes various pollutants such as PM2.5, PM10, CO, NO2, SO2, and more. The goal is to learn patterns in the data and understand the correlation between pollutants and meteorological data, as well as between different pollutants themselves. This involves developing a robust and accurate methodology to analyze the data, taking into account various factors such as pollutant levels, weather patterns, and other environmental factors. The end objective is to identify trends and patterns in the data that can help in analysis.

Software Used :

VSCode : Visual Studio Code (VSCode) is a popular code editor that supports Jupyter Notebooks through an extension. The Jupyter extension for VSCode provides a streamlined environment for creating and editing notebooks, as well as running and debugging code cells.

Description :

a) Data Mining Task

Our system uses a dataset that includes various pollutants such as PM2.5, PM10, CO, NO2, SO2, and more. The data mining task that is needed to be performed here is the **regression**.
The problem at hand is to learn patterns in pollutants data. We need to follow a specific procedure.

Firstly, we need to gather the hourly pollutant data from the three locations in Mumbai for the past three years. Once the data is collected, we need to preprocess it by cleaning and formatting it, ensuring that it is consistent and free from errors.

Next, we need to perform exploratory data analysis (EDA) to understand the data better. This will involve analyzing and visualizing the data to identify trends, patterns, and anomalies in the data. We will need to identify any missing values, outliers, or data inconsistencies, which may require further data cleaning or imputation.

Once the EDA is complete, we can move onto building statistical models to understand the correlation between pollutants and meteorological data. We can use various statistical techniques such as regression analysis to identify the relationship between the pollutants and meteorological data.

We can also use machine learning algorithms like regression to analyze the data and identify patterns and relationships.

b) Dataset Used

Website:- CCR (cpcb.gov.in)

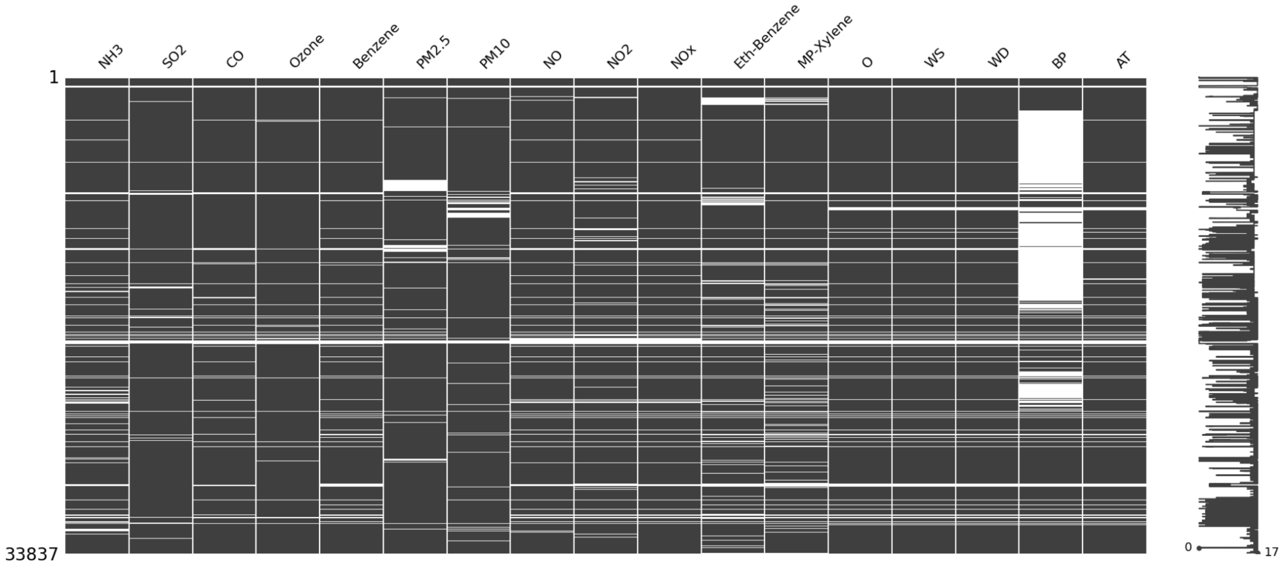
Datasets:-
kurla.csv
airport.csv
chakala.csv

The data is collected hourly and includes pollutants such as PM2.5, PM10, CO, NO2, SO2, etc. The data is collected from three locations in Mumbai, namely Chakala, Kurla, and the airport, for the past three years. The goal is to understand the correlation between pollutants and meteorological data, as well as the correlation between different pollutants.
To solve this problem,
Time Interval : 3 Years

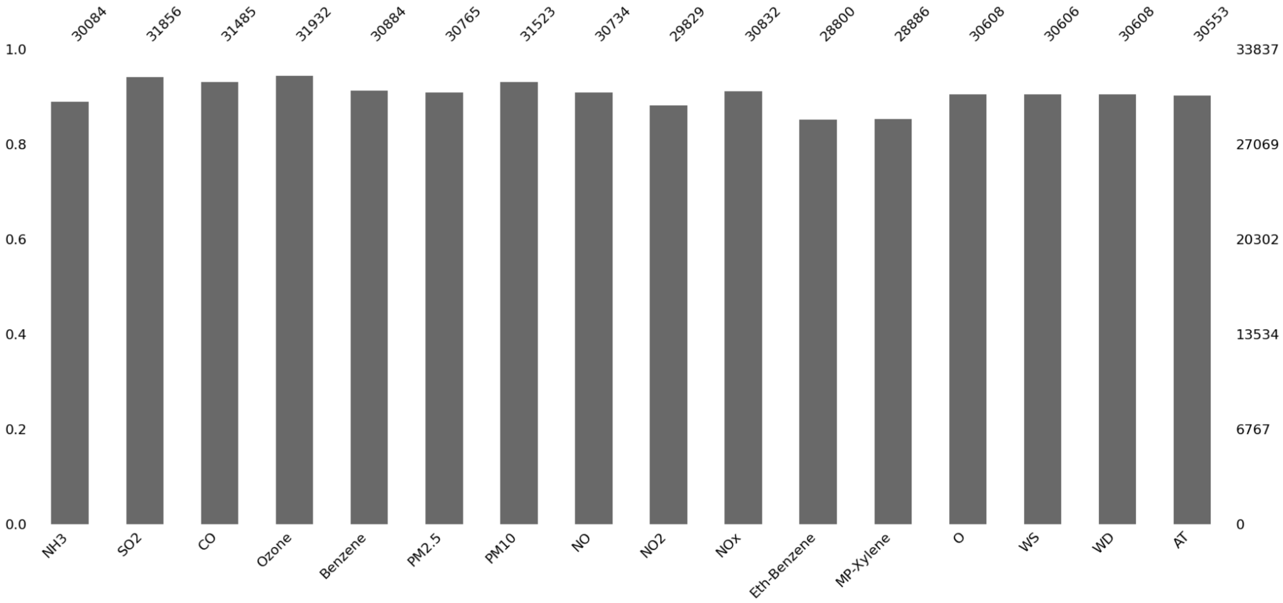
c) Snapshots showing implementation of the data mining algorithm of your choice and the results

i) Missing Values

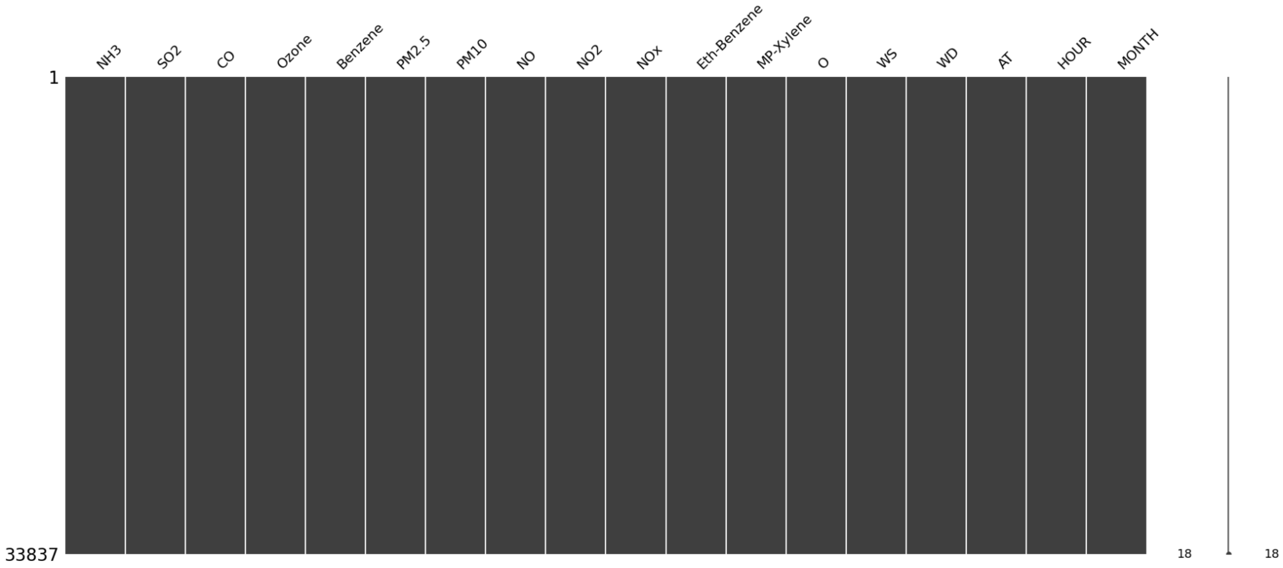
```
msno.matrix(df.iloc[:,1:])
plt.show()
```



```
msno.bar(df.iloc[:,1:])
```

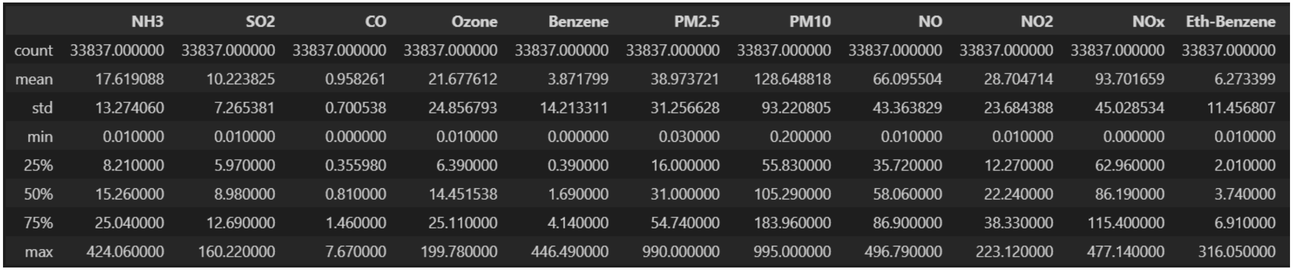


- After filling missing values



ii) Identifying the spread, min/max, avg, etc

```
df.describe()
```



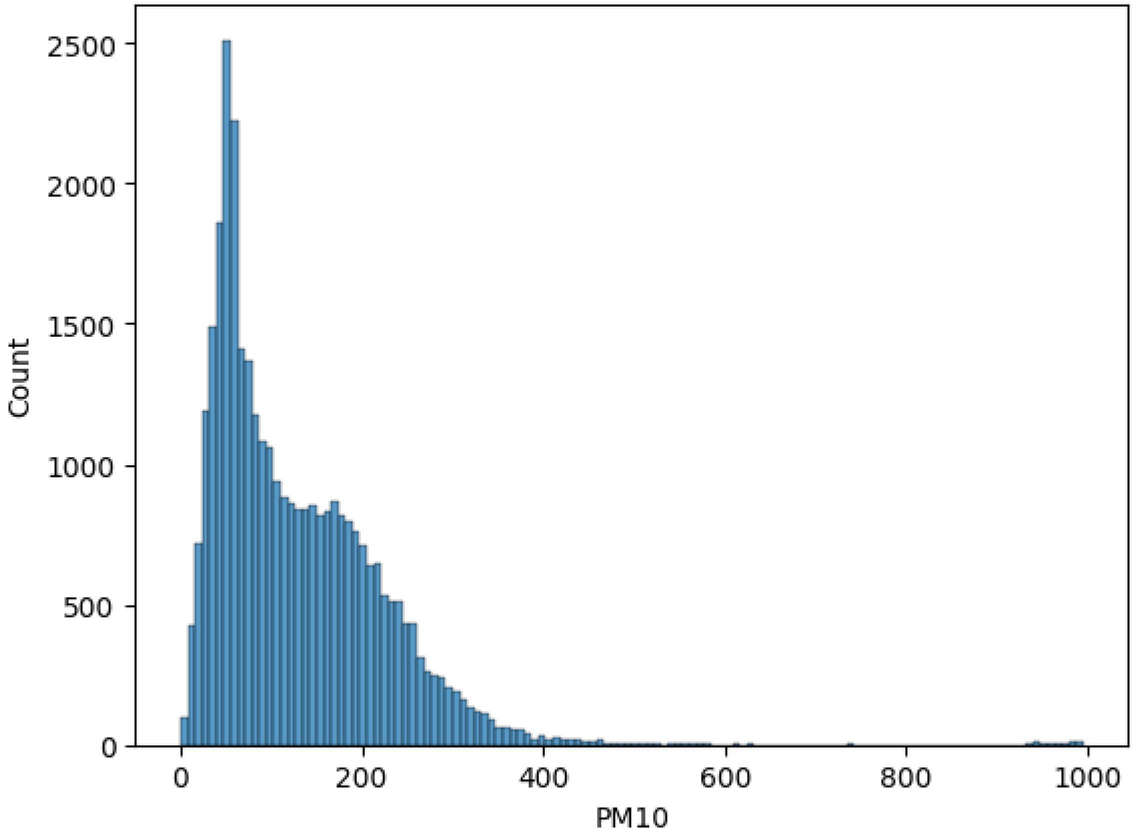
	NH3	SO2	CO	Ozone	Benzene	PM2.5	PM10	NO	NO2	NOx	Eth-Benzene
count	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000
mean	17.619088	10.223825	0.958261	21.677612	3.871799	38.973721	128.648818	66.095504	28.704714	93.701659	6.273399
std	13.274060	7.265381	0.700538	24.856793	14.213311	31.256628	93.220805	43.363829	23.684388	45.028534	11.456807
min	0.010000	0.010000	0.000000	0.010000	0.000000	0.030000	0.200000	0.010000	0.010000	0.000000	0.010000
25%	8.210000	5.970000	0.355980	6.390000	0.390000	16.000000	55.830000	35.720000	12.270000	62.960000	2.010000
50%	15.260000	8.980000	0.810000	14.451538	1.690000	31.000000	105.290000	58.060000	22.240000	86.190000	3.740000
75%	25.040000	12.690000	1.460000	25.110000	4.140000	54.740000	183.960000	86.900000	38.330000	115.400000	6.910000
max	424.060000	160.220000	7.670000	199.780000	446.490000	990.000000	995.000000	496.790000	223.120000	477.140000	316.050000

iii) Identifying the skewness.

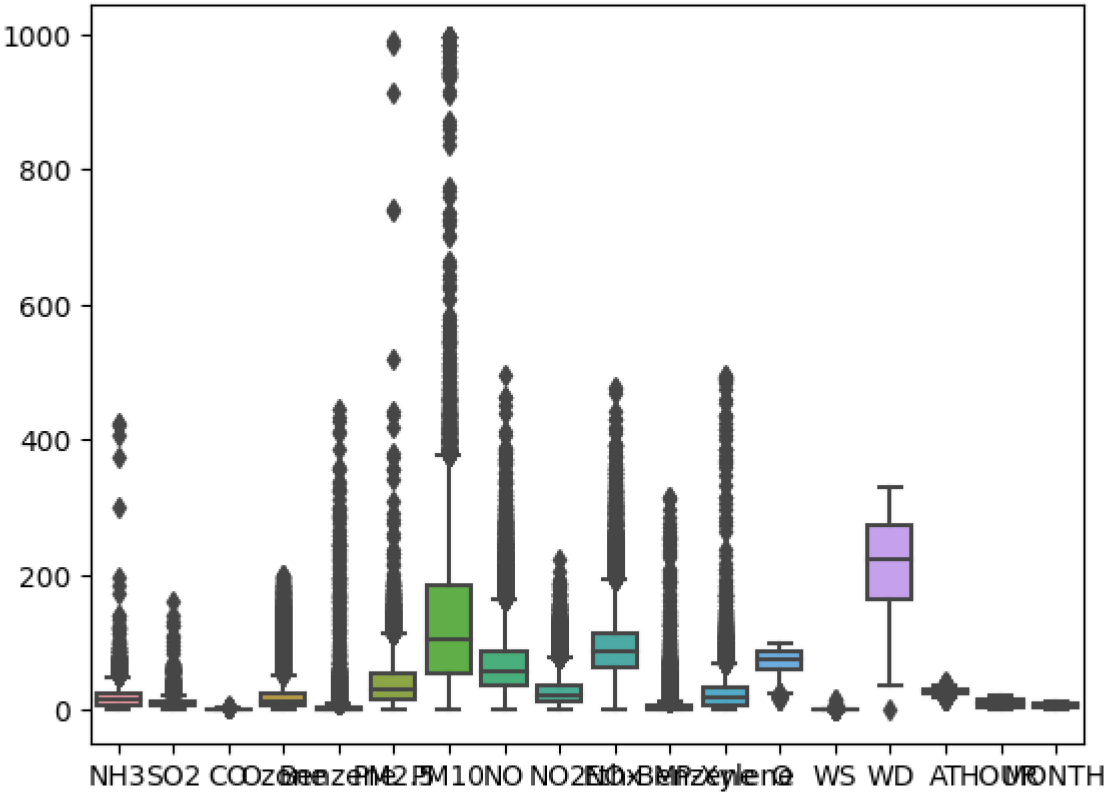
```
display(df.skew())
```

```
NH3                 4.642032
SO2                 3.035418
CO                  0.848790
```

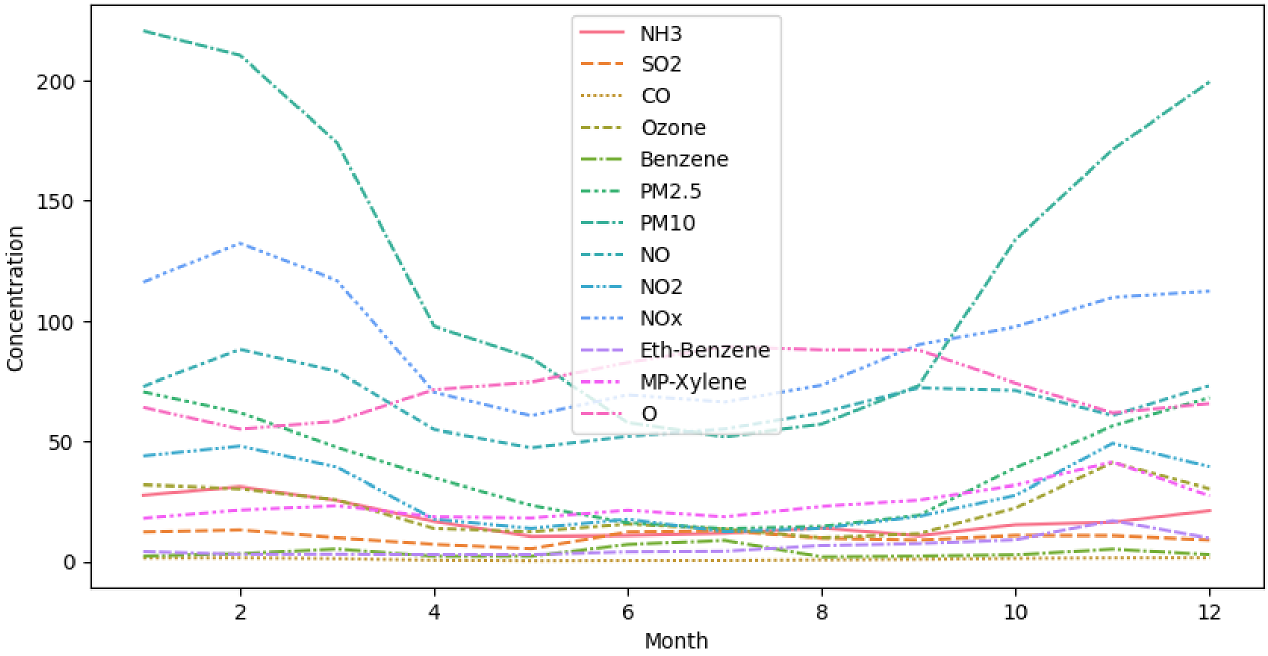
```
Ozone            2.822309
Benzene         16.931789
PM2.5            4.446476
PM10             1.981250
NO               1.715760
NO2              1.669439
NOx              1.454079
Eth-Benzene     11.561283
MP-Xylene        5.747266
O               -0.548206
WS               4.717906
WD              -0.242565
AT               0.172617
HOUR             0.000088
MONTH           -0.061831
```

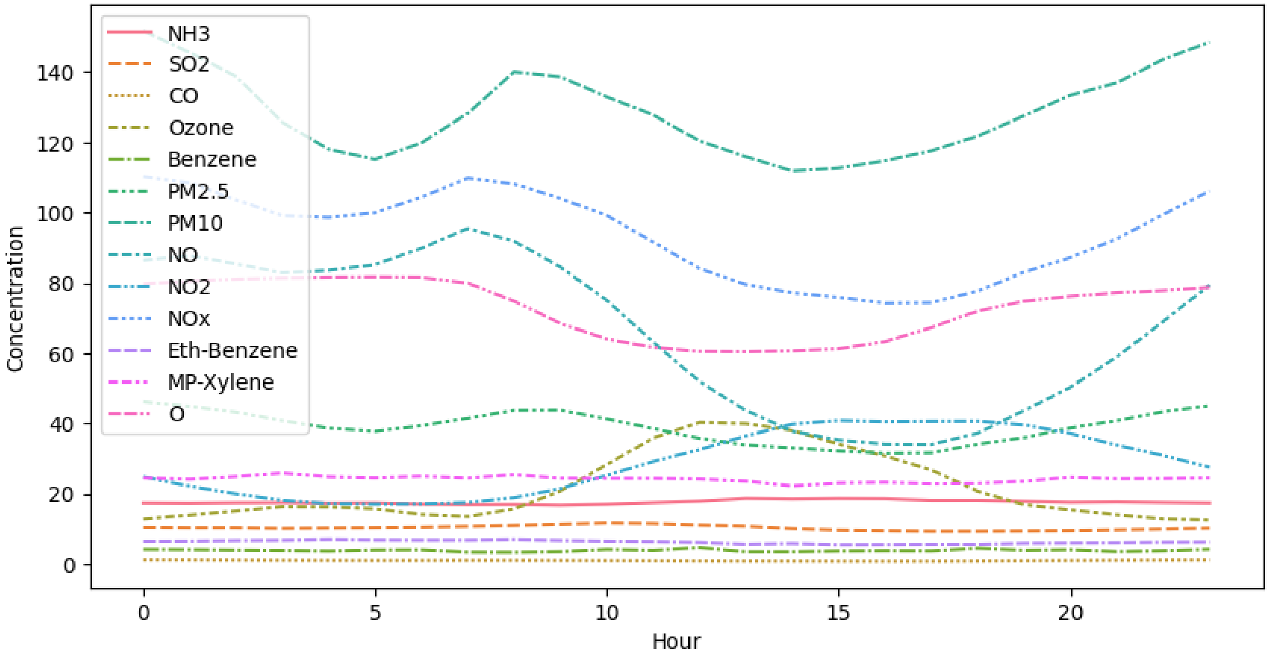


iv) Identifying the outliers using boxplot

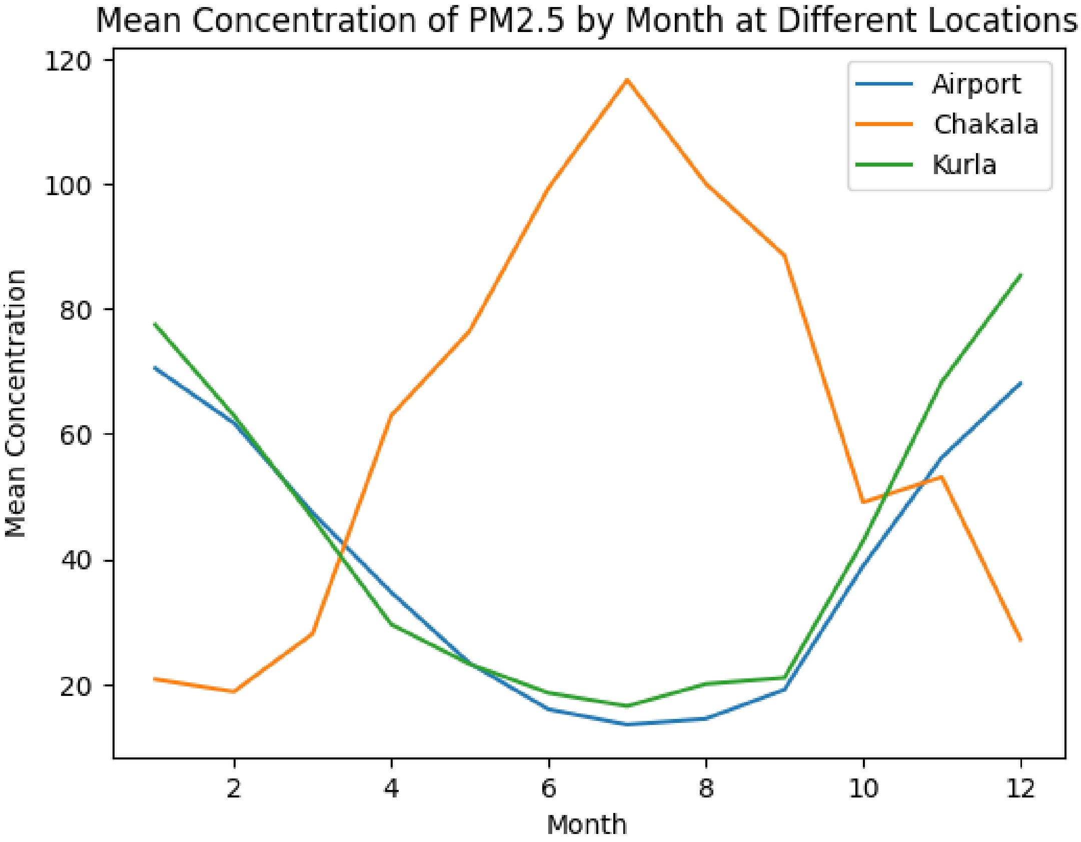


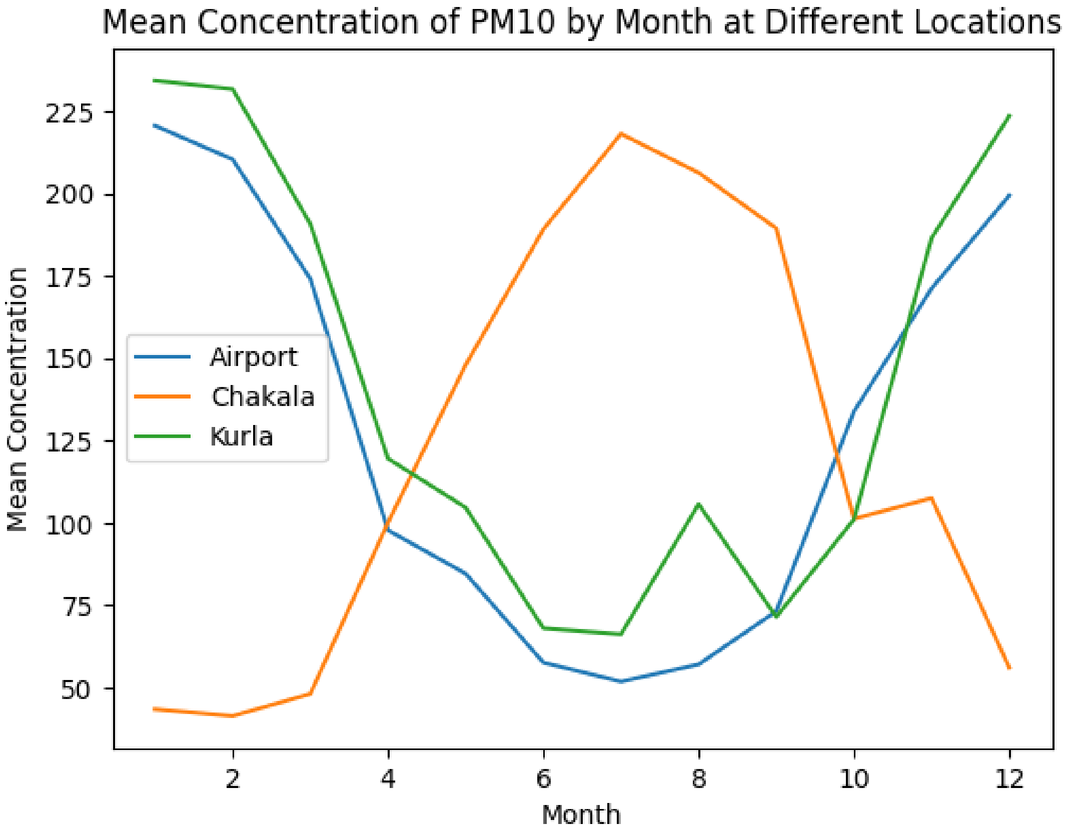
v) Pattern of pollutants vs time

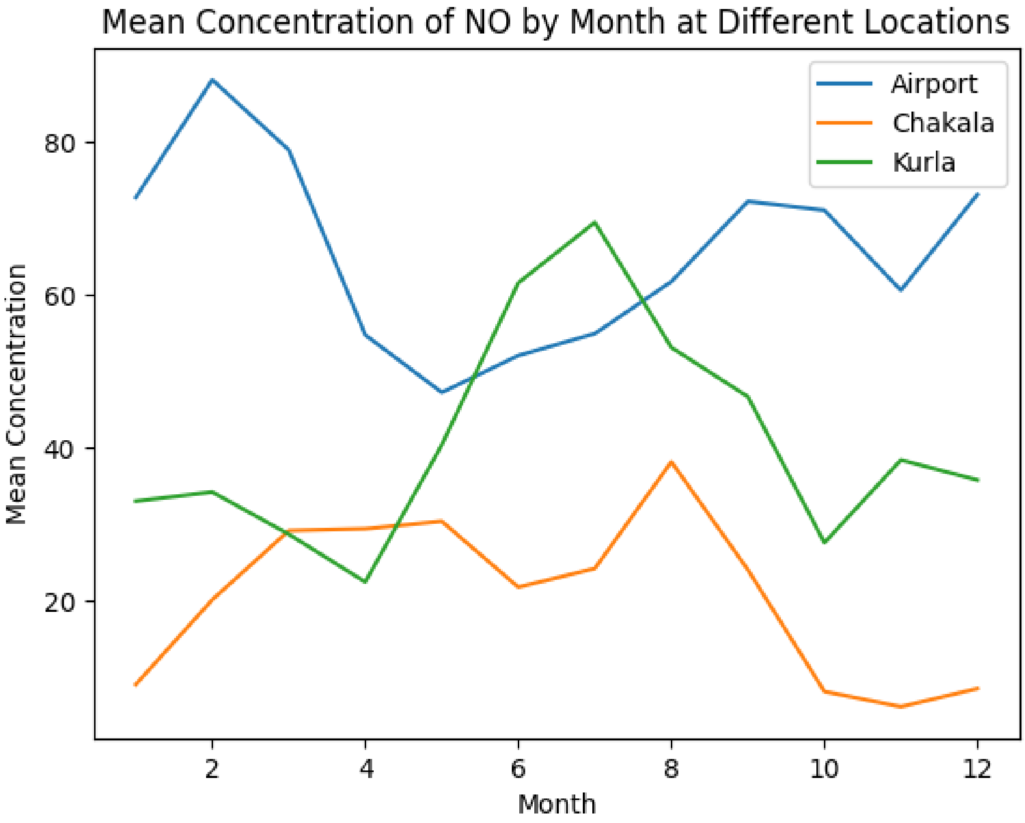


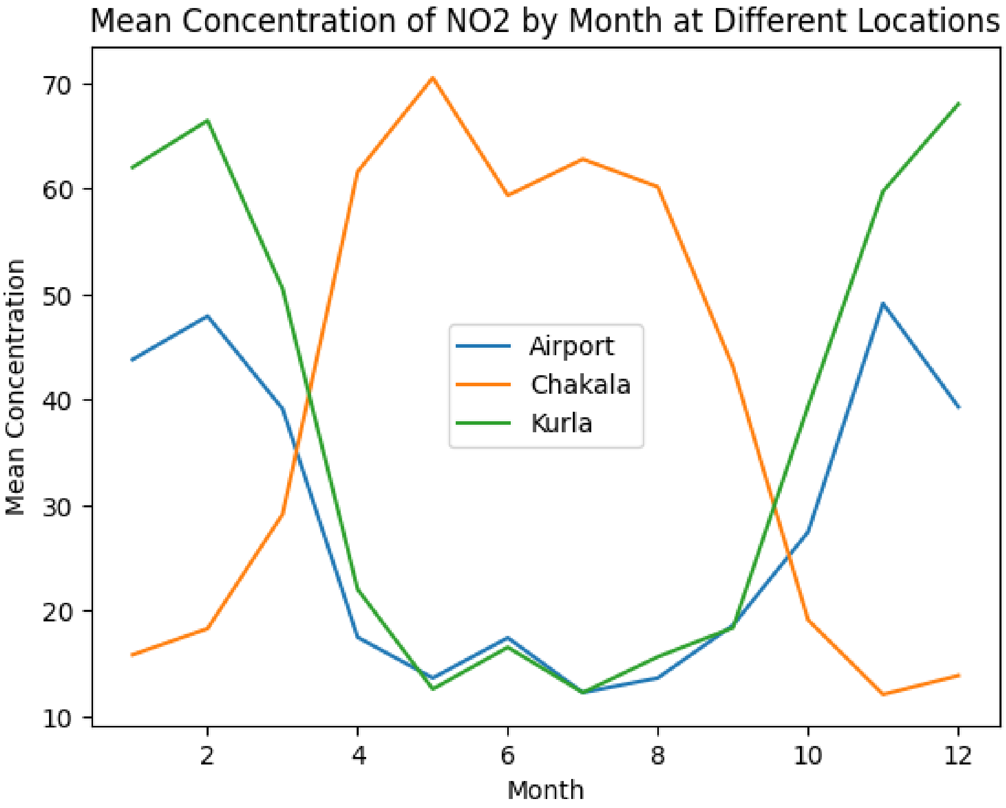


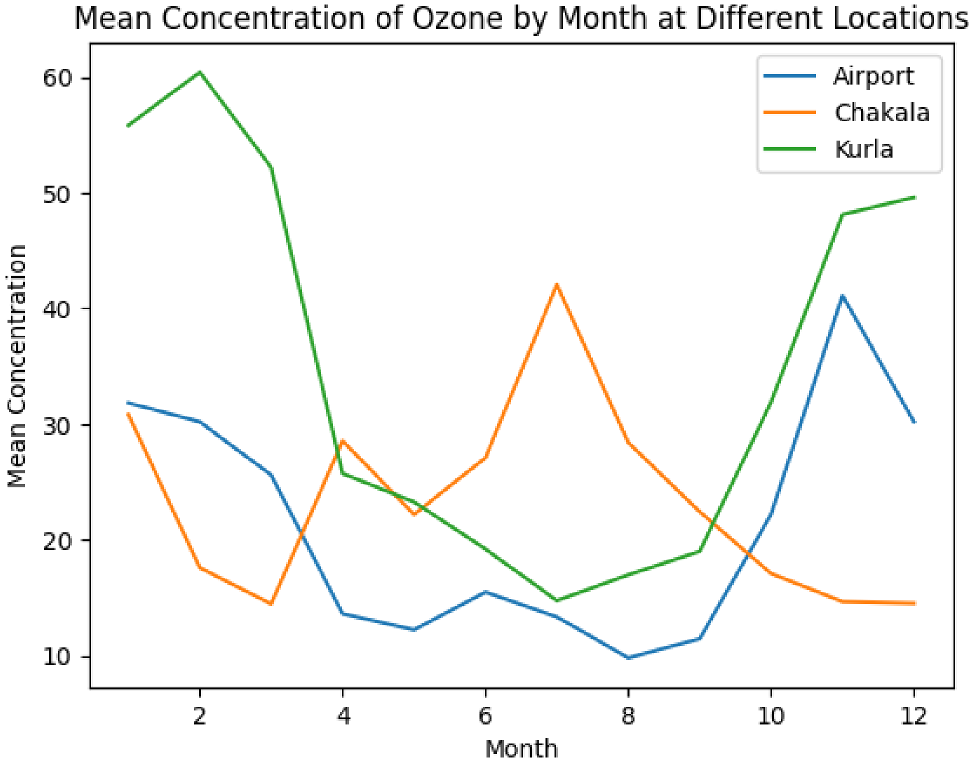
vi)Comparison of different locations

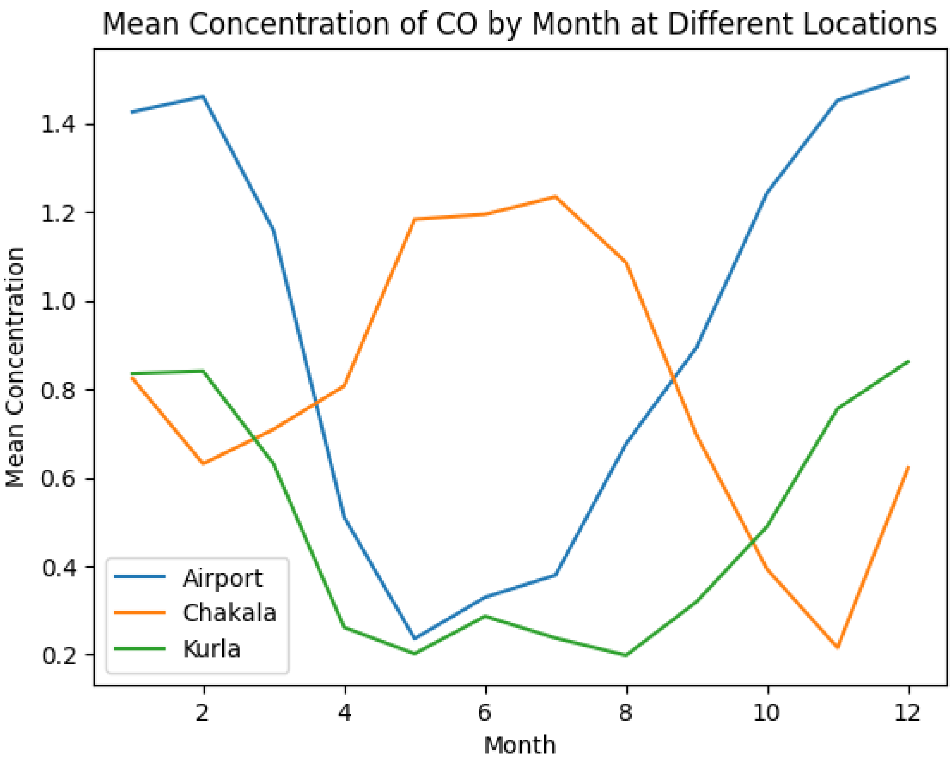


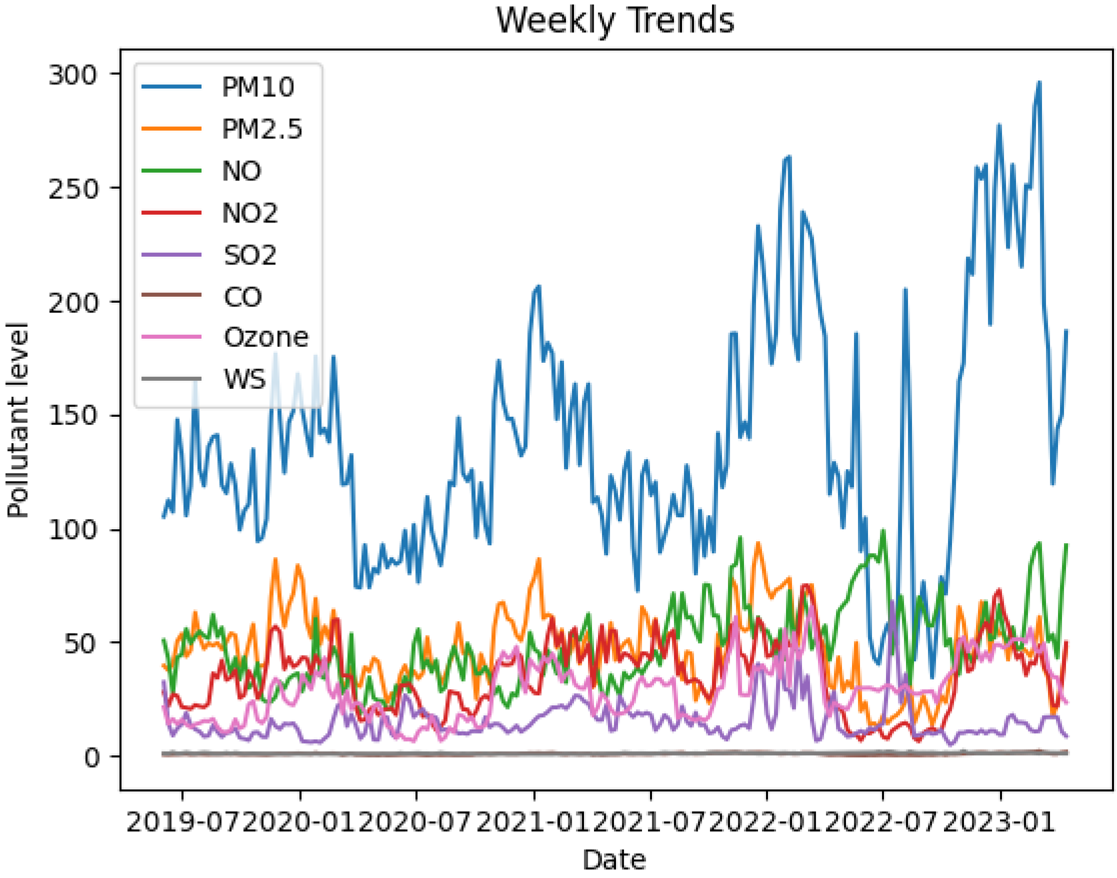

Mean Concentration of PM10 by Month at Different Locations
Mean Concentration
225
200
175
150
125
100
75
50
Airport
Chakala
Kurla
2
4
6
8
10
12
Month


Mean Concentration of NO by Month at Different Locations
Mean Concentration
80
60
40
20
Airport
Chakala
Kurla
2
4
6
8
10
12
Month


Mean Concentration of NO2 by Month at Different Locations
Mean Concentration
70
60
50
40
30
20
10
Month
2
4
6
8
10
12
Airport
Chakala
Kurla


Mean Concentration of Ozone by Month at Different Locations
Mean Concentration
60
50
40
30
20
10
Month
2
4
6
8
10
12
Airport
Chakala
Kurla

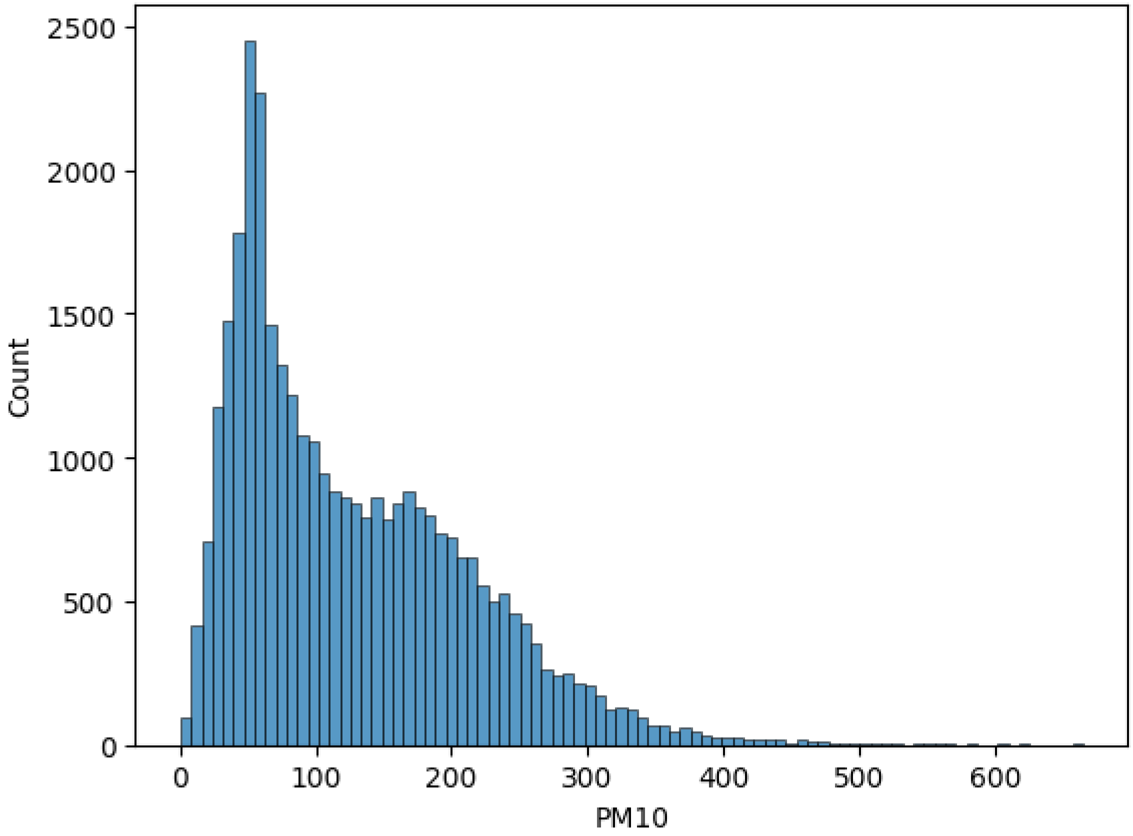

Mean Concentration of CO by Month at Different Locations
Mean Concentration
1.4
1.2
1.0
0.8
0.6
0.4
0.2
2
4
6
8
10
12
Month
Airport
Chakala
Kurla

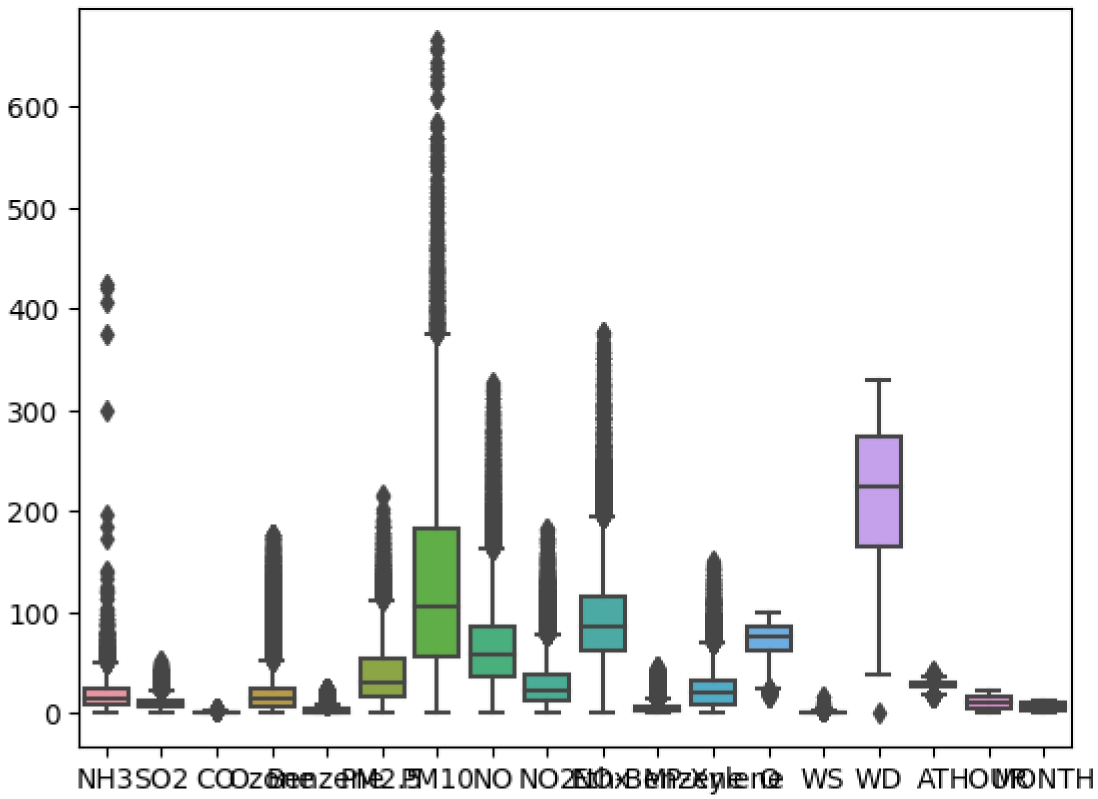

Weekly Trends
Pollutant level
300
250
200
150
100
50
0
PM10
PM2.5
NO
NO2
SO2
CO
Ozone
WS
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
Date

d) Interpretation and visualize the results

i)Missing Values:
5-10% missing values in each Columns
More than 70% missing Value in BP Column

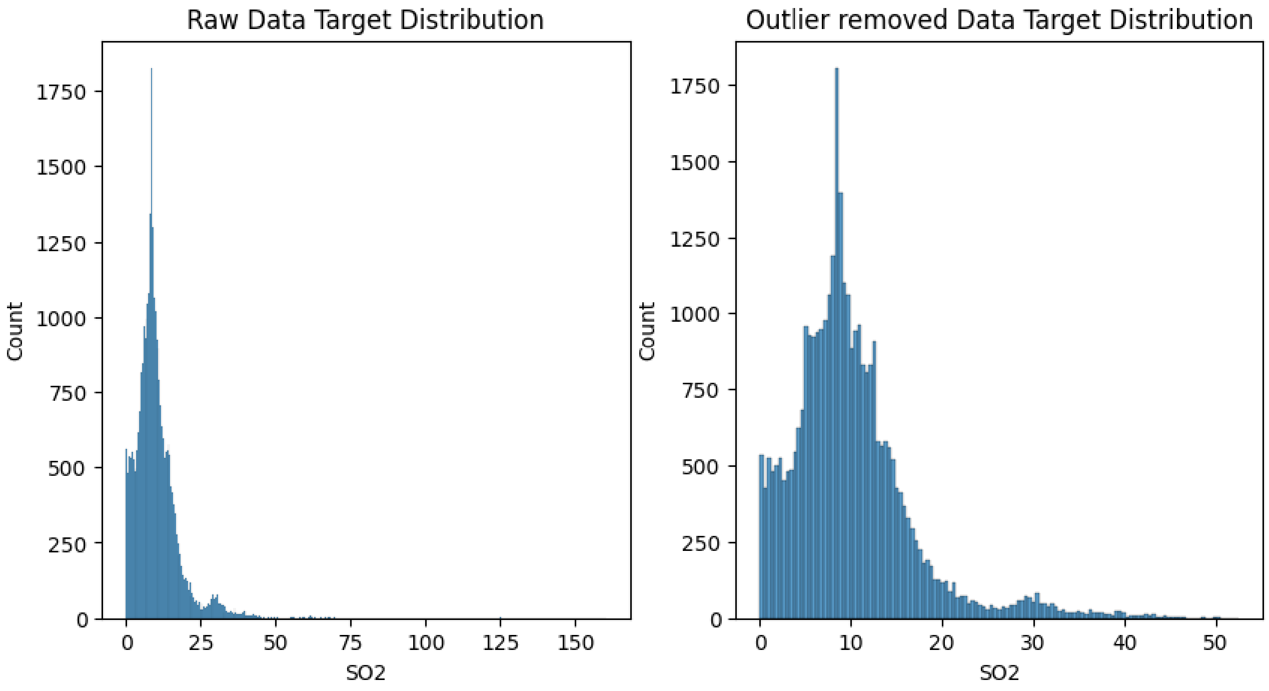
ii)After Handling Outliers

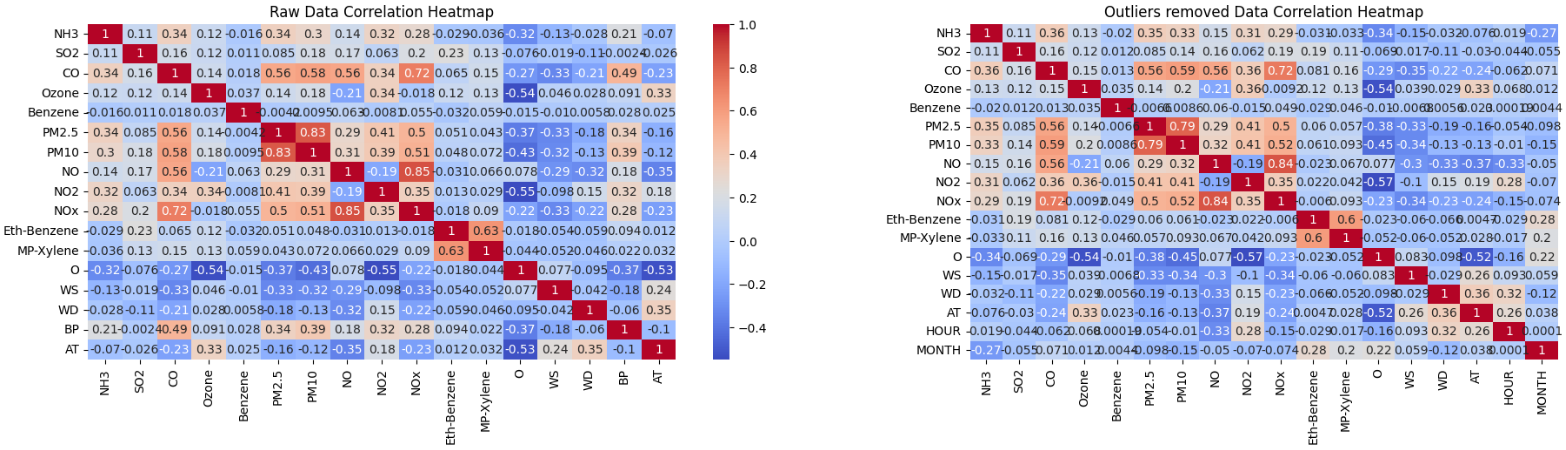




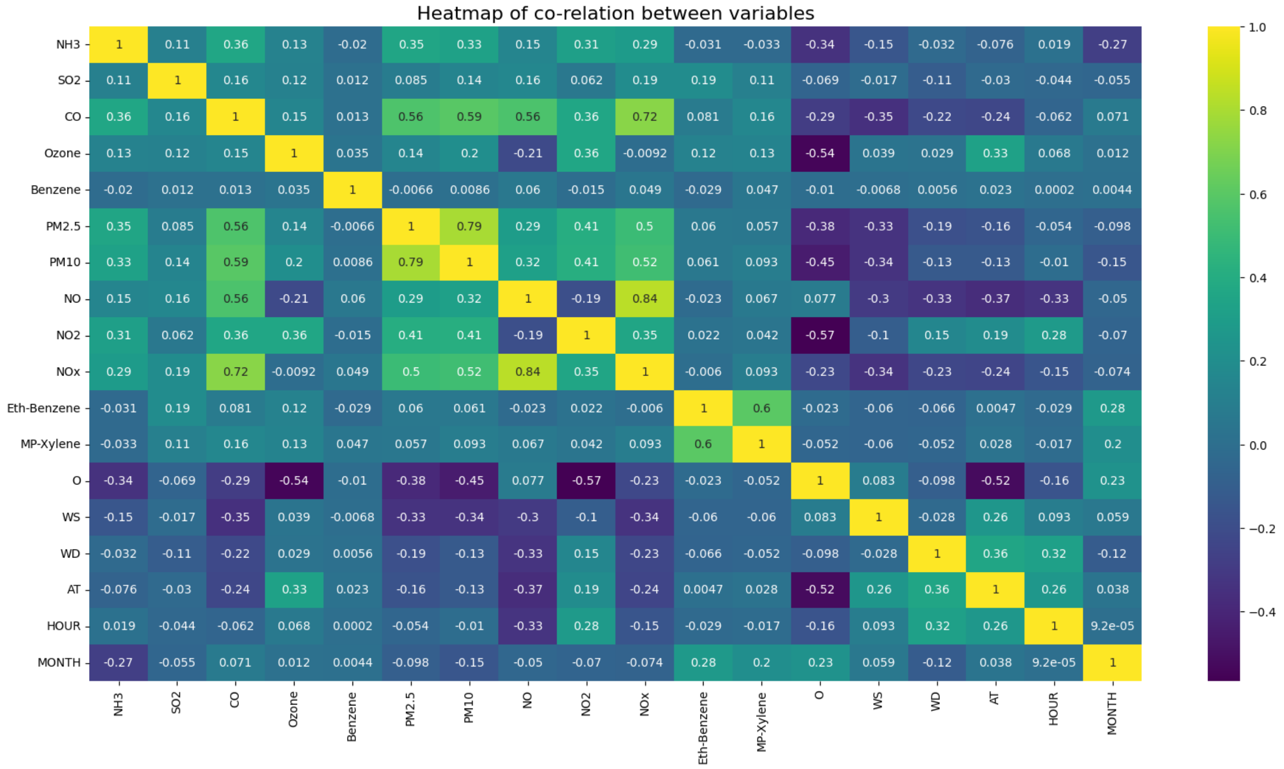
iii)Skewness

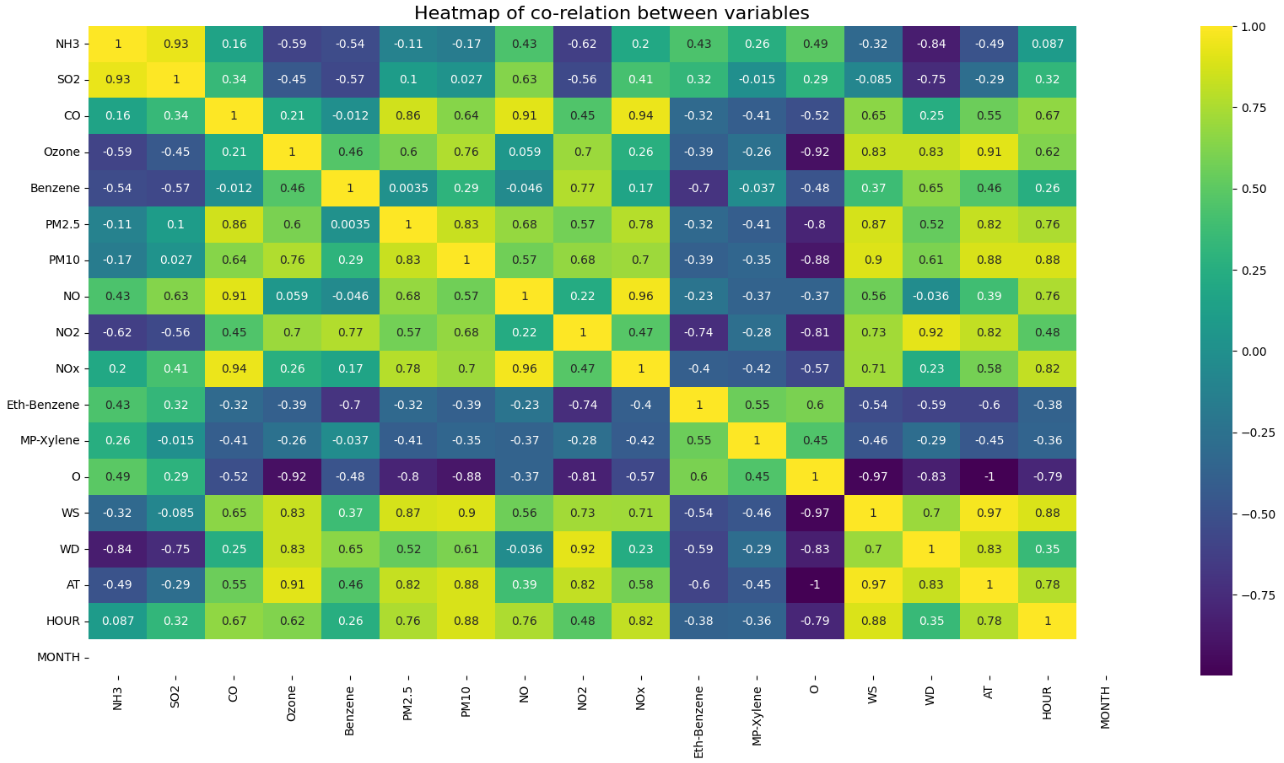
```
NH3            4.642032
SO2            1.652023
CO             0.827032
Ozone          2.682223
Benzene        2.336922
PM2.5          1.220802
PM10           1.025759
NO             1.466077
NO2            1.606635
NOx            1.327905
Eth-Benzene    2.937981
MP-Xylene      1.411164
O             -0.548206
WS             4.717906
WD            -0.242565
AT             0.172617
HOUR           0.000088
MONTH         -0.061831
```

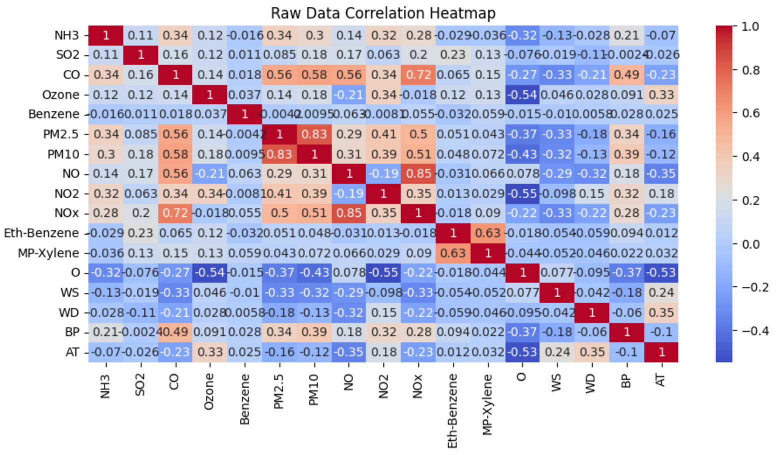


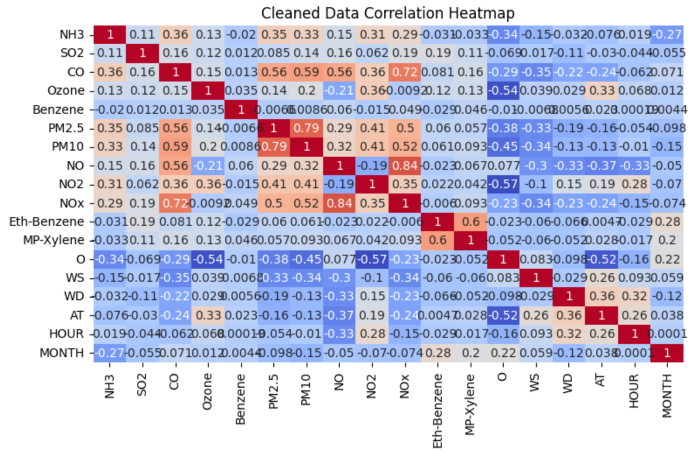


iv) Correlation heatmap before and after handling outliers

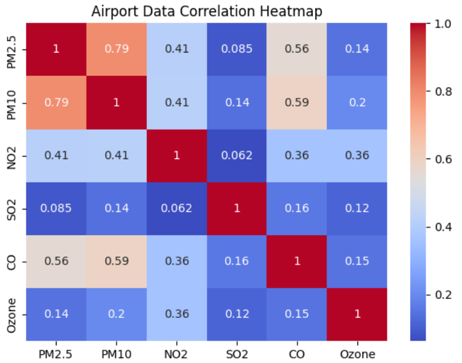

Heatmap of co-relation between variables
NH3 SO2 CO Ozone Benzene PM2.5 PM10 NO NO2 NOx Eth-Benzene MP-Xylene O WS WD AT HOUR MONTH
NH3 1 0.11 0.36 0.13 -0.02 0.35 0.33 0.15 0.31 0.29 -0.031 -0.033 -0.34 -0.15 -0.032 -0.076 0.019 -0.27
SO2 0.11 1 0.16 0.12 0.012 0.085 0.14 0.16 0.062 0.19 0.19 0.11 -0.069 -0.017 -0.11 -0.03 -0.044 -0.055
CO 0.36 0.16 1 0.15 0.013 0.56 0.59 0.56 0.36 0.72 0.081 0.16 -0.29 -0.35 -0.22 -0.24 -0.062 0.071
Ozone 0.13 0.12 0.15 1 0.035 0.14 0.2 -0.21 0.36 -0.0092 0.12 0.13 -0.54 0.039 0.029 0.33 0.068 0.012
Benzene -0.02 0.012 0.013 0.035 1 -0.0066 0.0086 0.06 -0.015 0.049 -0.029 0.047 -0.01 -0.0068 0.0056 0.023 0.0002 0.0044
PM2.5 0.35 0.085 0.56 0.14 -0.0066 1 0.79 0.29 0.41 0.5 0.06 0.057 -0.38 -0.33 -0.19 -0.16 -0.054 -0.098
PM10 0.33 0.14 0.59 0.2 0.0086 0.79 1 0.32 0.41 0.52 0.061 0.093 -0.45 -0.34 -0.13 -0.13 -0.01 -0.15
NO 0.15 0.16 0.56 -0.21 0.06 0.29 0.32 1 -0.19 0.84 -0.023 0.067 0.077 -0.3 -0.33 -0.37 -0.33 -0.05
NO2 0.31 0.062 0.36 0.36 -0.015 0.41 0.41 -0.19 1 0.35 0.022 0.042 -0.57 -0.1 0.15 0.19 0.28 -0.07
NOx 0.29 0.19 0.72 -0.0092 0.049 0.5 0.52 0.84 0.35 1 -0.006 0.093 -0.23 -0.34 -0.23 -0.24 -0.15 -0.074
Eth-Benzene -0.031 0.19 0.081 0.12 -0.029 0.06 0.061 -0.023 0.022 -0.006 1 0.6 -0.023 -0.06 -0.066 0.0047 -0.029 0.28
MP-Xylene -0.033 0.11 0.16 0.13 0.047 0.057 0.093 0.067 0.042 0.093 0.6 1 -0.052 -0.06 -0.052 0.028 -0.017 0.2
O -0.34 -0.069 -0.29 -0.54 -0.01 -0.38 -0.45 0.077 -0.57 -0.23 -0.023 -0.052 1 0.083 -0.098 -0.52 -0.16 0.23
WS -0.15 -0.017 -0.35 0.039 -0.0068 -0.33 -0.34 -0.3 -0.1 -0.34 -0.06 -0.06 0.083 1 -0.028 0.26 0.093 0.059
WD -0.032 -0.11 -0.22 0.029 0.0056 -0.19 -0.13 -0.33 0.15 -0.23 -0.066 -0.052 -0.098 -0.028 1 0.36 0.32 -0.12
AT -0.076 -0.03 -0.24 0.33 0.023 -0.16 -0.13 -0.37 0.19 -0.24 0.0047 0.028 -0.52 0.26 0.36 1 0.26 0.038
HOUR 0.019 -0.044 -0.062 0.068 0.0002 -0.054 -0.01 -0.33 0.28 -0.15 -0.029 -0.017 -0.16 0.093 0.32 0.26 1 9.2e-05
MONTH -0.27 -0.055 0.071 0.012 0.0044 -0.098 -0.15 -0.05 -0.07 -0.074 0.28 0.2 0.23 0.059 -0.12 0.038 9.2e-05 1
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

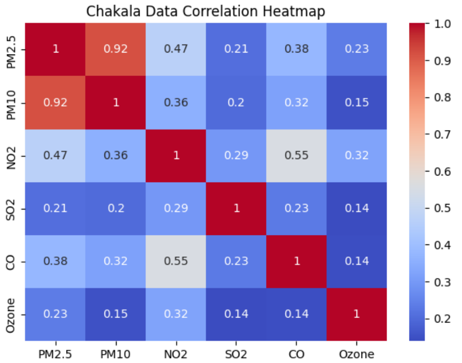

Heatmap of co-relation between variables
NH3 SO2 CO Ozone Benzene PM2.5 PM10 NO NO2 NOx Eth-Benzene MP-Xylene O WS WD AT HOUR MONTH
NH3 1 0.93 0.16 -0.59 -0.54 -0.11 -0.17 0.43 -0.62 0.2 0.43 0.26 0.49 -0.32 -0.84 -0.49 0.087
SO2 0.93 1 0.34 -0.45 -0.57 0.1 0.027 0.63 -0.56 0.41 0.32 -0.015 0.29 -0.085 -0.75 -0.29 0.32
CO 0.16 0.34 1 0.21 -0.012 0.86 0.64 0.91 0.45 0.94 -0.32 -0.41 -0.52 0.65 0.25 0.55 0.67
Ozone -0.59 -0.45 0.21 1 0.46 0.6 0.76 0.059 0.7 0.26 -0.39 -0.26 -0.92 0.83 0.83 0.91 0.62
Benzene -0.54 -0.57 -0.012 0.46 1 0.0035 0.29 -0.046 0.77 0.17 -0.7 -0.037 -0.48 0.37 0.65 0.46 0.26
PM2.5 -0.11 0.1 0.86 0.6 0.0035 1 0.83 0.68 0.57 0.78 -0.32 -0.41 -0.8 0.87 0.52 0.82 0.76
PM10 -0.17 0.027 0.64 0.76 0.29 0.83 1 0.57 0.68 0.7 -0.39 -0.35 -0.88 0.9 0.61 0.88 0.88
NO 0.43 0.63 0.91 0.059 -0.046 0.68 0.57 1 0.22 0.96 -0.23 -0.37 -0.37 0.56 -0.036 0.39 0.76
NO2 -0.62 -0.56 0.45 0.7 0.77 0.57 0.68 0.22 1 0.47 -0.74 -0.28 -0.81 0.73 0.92 0.82 0.48
NOx 0.2 0.41 0.94 0.26 0.17 0.78 0.7 0.96 0.47 1 -0.4 -0.42 -0.57 0.71 0.23 0.58 0.82
Eth-Benzene 0.43 0.32 -0.32 -0.39 -0.7 -0.32 -0.39 -0.23 -0.74 -0.4 1 0.55 0.6 -0.54 -0.59 -0.6 -0.38
MP-Xylene 0.26 -0.015 -0.41 -0.26 -0.037 -0.41 -0.35 -0.37 -0.28 -0.42 0.55 1 0.45 -0.46 -0.29 -0.45 -0.36
O 0.49 0.29 -0.52 -0.92 -0.48 -0.8 -0.88 -0.37 -0.81 -0.57 0.6 0.45 1 -0.97 -0.83 -1 -0.79
WS -0.32 -0.085 0.65 0.83 0.37 0.87 0.9 0.56 0.73 0.71 -0.54 -0.46 -0.97 1 0.7 0.97 0.88
WD -0.84 -0.75 0.25 0.83 0.65 0.52 0.61 -0.036 0.92 0.23 -0.59 -0.29 -0.83 0.7 1 0.83 0.35
AT -0.49 -0.29 0.55 0.91 0.46 0.82 0.88 0.39 0.82 0.58 -0.6 -0.45 -1 0.97 0.83 1 0.78
HOUR 0.087 0.32 0.67 0.62 0.26 0.76 0.88 0.76 0.48 0.82 -0.38 -0.36 -0.79 0.88 0.35 0.78 1
MONTH
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75

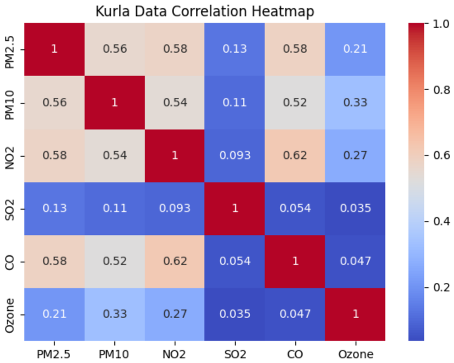




v) Correlation heatmaps between different regions/locations







e) Provide clearly the BI decision that is to be taken as a result of mining.

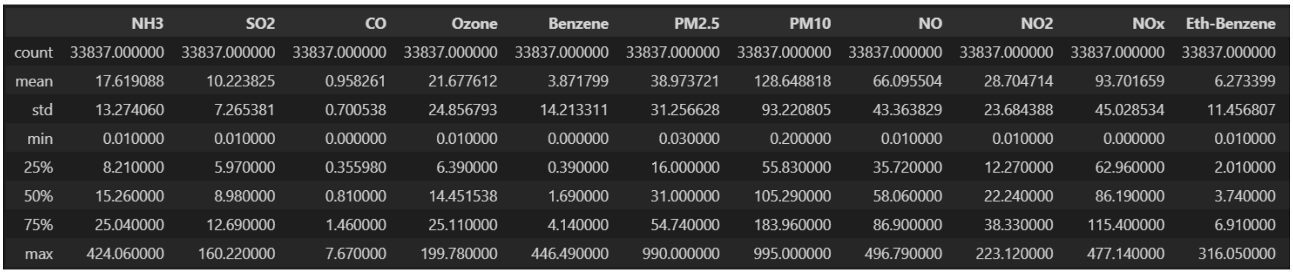
i)Missing Values:

- Filling the columns which have less than 10% missing values
- Removing column with more than 50% missing values
- Filling missing values using monthly average

```
cols=['NH3','SO2','CO','Ozone','Benzene','PM2.5','PM10','NO','NO2','NOx
','Eth-Benzene','MP-Xylene','O','WS','WD','AT']
for col in cols:

df[col]=df[col].fillna(df.groupby(['MONTH','HOUR'])[col].transform('mea
n'))
```

ii) Identifying the spread, min/max, avg, etc



	NH3	SO2	CO	Ozone	Benzene	PM2.5	PM10	NO	NO2	NOx	Eth-Benzene
count	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000	33837.000000
mean	17.619088	10.223825	0.958261	21.677612	3.871799	38.973721	128.648818	66.095504	28.704714	93.701659	6.273399
std	13.274060	7.265381	0.700538	24.856793	14.213311	31.256628	93.220805	43.363829	23.684388	45.028534	11.456807
min	0.010000	0.010000	0.000000	0.010000	0.000000	0.030000	0.200000	0.010000	0.010000	0.000000	0.010000
25%	8.210000	5.970000	0.355980	6.390000	0.390000	16.000000	55.830000	35.720000	12.270000	62.960000	2.010000
50%	15.260000	8.980000	0.810000	14.451538	1.690000	31.000000	105.290000	58.060000	22.240000	86.190000	3.740000
75%	25.040000	12.690000	1.460000	25.110000	4.140000	54.740000	183.960000	86.900000	38.330000	115.400000	6.910000
max	424.060000	160.220000	7.670000	199.780000	446.490000	990.000000	995.000000	496.790000	223.120000	477.140000	316.050000

iii) Identifying the skewness.

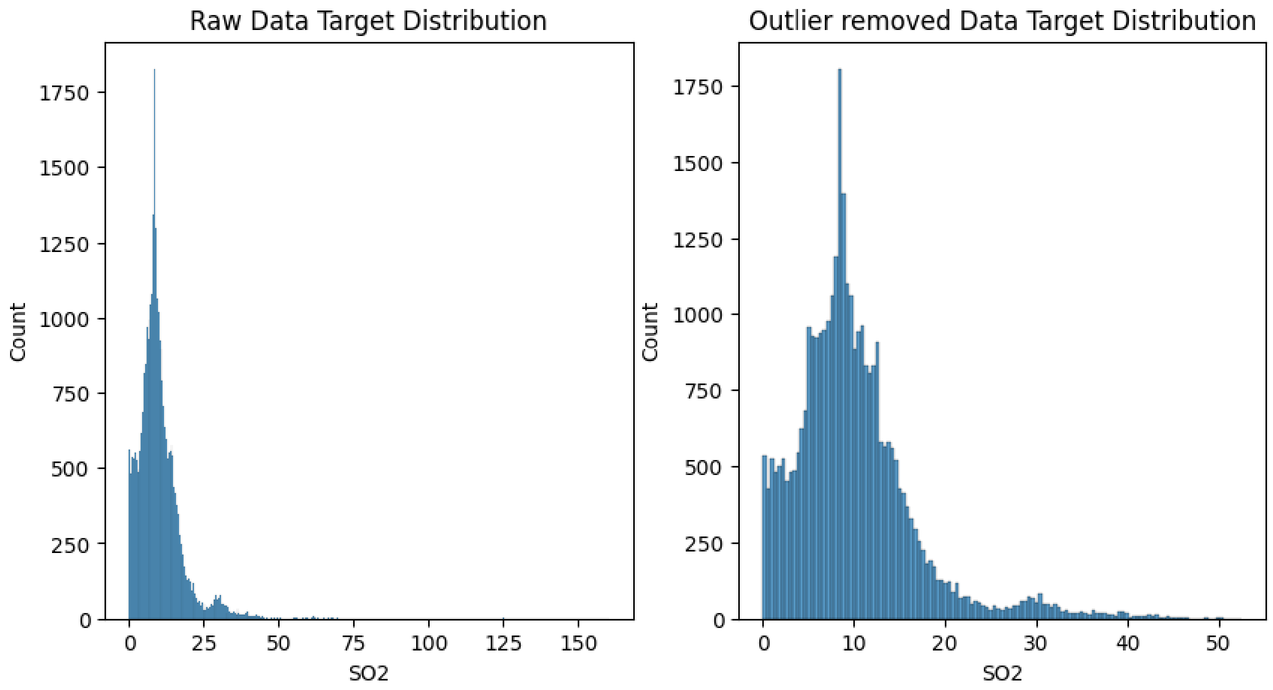
Positive skewness variables include Benzene, Eth-Benzene, MP-Xylene, NH3, and WS. Other variables have a negative skewness, indicating a left-skewed distribution, such as O and WD.

The values for CO, Ozone, NO, NO2, NOx, PM2.5, and PM10 are closer to zero, suggesting that their distributions are more symmetrical

.

iv) Identifying the outliers using boxplot

Before removing outliers data had many outliers which makes difficult to find relation between the variables so after removing outliers it became easy to find relation between variables for feature selection .



v) Pattern of pollutants vs time

For Pollutants like PM2.5 , PM10 and NOx concentrations are high in the start and end of the year and moderate in mid year .

Concentration of PM2.5 is high at starting of season but then it decreases in mid season . Except PM10 other pollutants does not show huge diversion in concentration based based on hours . But PM10 is low between 12 to 4 pm period

vi) Comparison of different locations

For pollutants PM2.5,NO2 and PM10 concentration for Chakala is indirectly proportional to locations Kurla and Airport. The graph says that whenever there is an increase of these games in Airport and Kurla there is less concentration of samegases in Chakala is High and vice versa.

For Airport data CO,NO2 and PM10 are somehow dependent on each other while for Kurla data PM10 and SO2 are features and for Chakala SO2 and Ozone are correlating .

Conclusion :

The spread, min/max, average, and skewness were identified for different pollutants, with some showing positive skewness and others showing negative skewness. Outliers were identified using boxplots, and it was found that removing outliers made it easier to find the relationship between variables for feature selection.

In terms of the pattern of pollutants vs time, it was found that concentrations for pollutants like PM2.5, PM10, and NOx were high at the start and end of the year and moderate in the mid-year. The concentration of PM2.5 was high at the beginning of the season and decreased in the mid-season. While for the comparison of different locations, it was found that the concentration of pollutants like PM2.5, NO2, and PM10 for Chakala was indirectly proportional to Kurla and Airport, and the concentrations of some pollutants were found to be dependent on each other in different locations.

Overall, the analysis provides valuable insights into the pollutant concentrations in different locations and their patterns over time, which can be used to identify potential sources of pollution and develop strategies to mitigate their effects.

References :

1. M. Al-Fuqaha, M. Al-Fuqaha and M. Guizani, "Integrating Data-Based Strategies and Advanced Technologies with Efficient Air Pollution Management in Smart Cities," Sustainability, vol. 13, no. 2, p. 740, Jan. 2021, doi: 10.3390/su13020740.
2. J. Brownlee, "A Gentle Introduction to Long Short-Term Memory Networks by the Experts - MachineLearningMastery.com Jason Brownlee," MachineLearningMastery.com, Jan. 2017. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/. [Accessed: Feb. 27, 2023].

3. J. Brownlee, "A Gentle Introduction to XGBoost for Applied Machine Learning - MachineLearningMastery.com," MachineLearningMastery.com, Apr. 2019. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/. [Accessed: Feb. 27, 2023].

4. World Health Organization, "Ambient (outdoor) air pollution," WHO, Dec. 19, 2022. [Online]. Available: https://www.who.int/news-room/q-a-detail/ambient-outdoor-air-pollution. [Accessed: Mar. 04, 2023].

5. T. Madan, S. Sagar and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms –A Review," 2021 International Conference on Computer, Information, Communication and Networks (CICN), Indore, India, 2021, pp. 1-6, doi: 10.1109/CICN52508.2021.9427116.

6. T. Chen, "XGBoost: A Scalable Tree Boosting System," arXiv:1603.02754 [cs.LG], Jun. 2016. [Online]. Available: https://arxiv.org/abs/1603.02754. [Accessed: Mar. 20, 2023].