# Mathematical Foundation for Data Science

Mini Project Report

Project No. 4

**(Air Quality in Beijing)**

**Krishna Sah Teli**

GE22M018

*Cyber-Physical Systems (I2MP)*

*Indian Institute of Technology (IITM)*

Madras - 600036, India

ge22m018@smail.iitm.ac.in

For the partial fulfilment of requirements for CH5019

January 4, 2023

## Abstracts

**In this report, Beijing's air quality has been explored using data visualization tools and analysed statistically by implementing hypothesis testing. A clustering algorithm has been implemented to segregate the data into effective groups to study its different features and correlations. It explains how the size of particulate matter (PM) changes due to the change in the concentration of air pollutants. It describes the correlation between the different data attributes visually and the variation in concentration over the period of time. It also briefs the impact of air pollutants on the greenhouse effect by correlating with temperature, humidity and rain.**


# 1 Introduction

With modernization and industrialization, human life is becoming more comfortable and luxurious. However, in another hand, it is also bringing more devastating consciousness by producing toxic substances as by-products and destroying our nature which finally imbalances the ecosystem and invites different natural calamities such as earthquakes, unseasonal rain, heavy drought, the greenhouse effect and so on. Industrialization causes air pollution by producing toxic gases in the air, land pollution with poisonous and non-degradable waste materials and water pollution with its toxic liquid by-products. Similarly, this paper explains how Beijing's air quality has got degraded because of different air pollutants and their impacts on the greenhouse.

The tiny particles present in the air which form due to the chemical reaction between different chemical air pollutants such as SO2, NO2, CO, etc., have been classified into two categories called PM2.5 and PM10. PM2.5 is the fine particulate matter of diameter less than 2.5 micrometres where as PM10 is of diameter 10 micrometres. Since the size of PM2.5 is very less, hence it can easily penetrate into our lungs and causes different respiratory diseases.

With the help of data visualization tools, the correlation between the data attributes has been studied and also their concentration variations (i.e yearly, monthly, weekly and hourly) have been explored graphically. Hypothesis testing has been implemented to study the difference in pollutants between different sites in the city. To cluster the data into meaningful groups, the K-Means clustering algorithm has been used and the group features have been studied graphically.

The rest of the paper has been organized as follows: section 2 explains the descriptive analysis of the data whereas the correlation between the attributes has been discussed in section 3. Similarly, section 4 depicts the clustering model and its finding and section 5 called hypothesis testing explain which sites of the city are closely impacted due to air pollutants.

# 2 Discriptive Analysis

The data on air quality contains 17 data attributes including information on time periods such as year, month, day and hour. It contains 6 main air pollutants i.e PM2.5, PM10, SO2, NO2, CO and 03. Besides this, it also gives meteorological information in 6 different variables i.e Temperature, Pressure, DewPoint, Rain, Wind direction, and Station.

## 2.1 Yearly Analysis

Here first data has been explored yearly as shown in fig.1. It has been found that air pollutants concentration has increased first in year 2013 and again decreased drastically from 2014 to 2016. It

may be because during 2013 China would be mainly focusing on industrialization by ignoring its consciousness. Later may be by seeing its disastrous impact on air quality, strict rule may have been adopted to cut out the air affecting industrial gaseous.

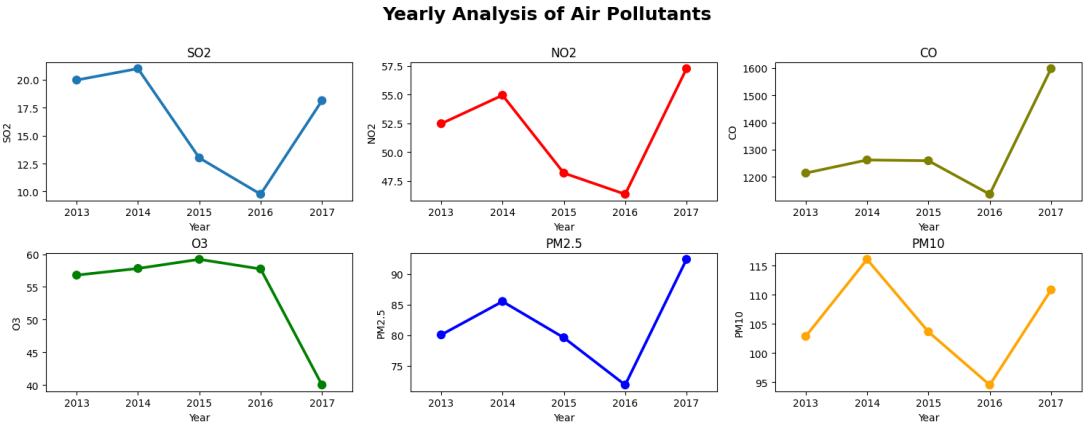


Figure 1: Yearly Analysis

## 2.2 Monthly Analysis

Similarly as shown in fig.2-4, data has been analysed monthly for each year. In the year 2013, the air pollutants concentration decreased in may except for O3. The concentration of SO2 fell down in control till august but it started increasing slowly afterwards in the year 2013. Similarly for NO2 and CO whereas PM concentration has fluctuated and eventually started falling down after January.

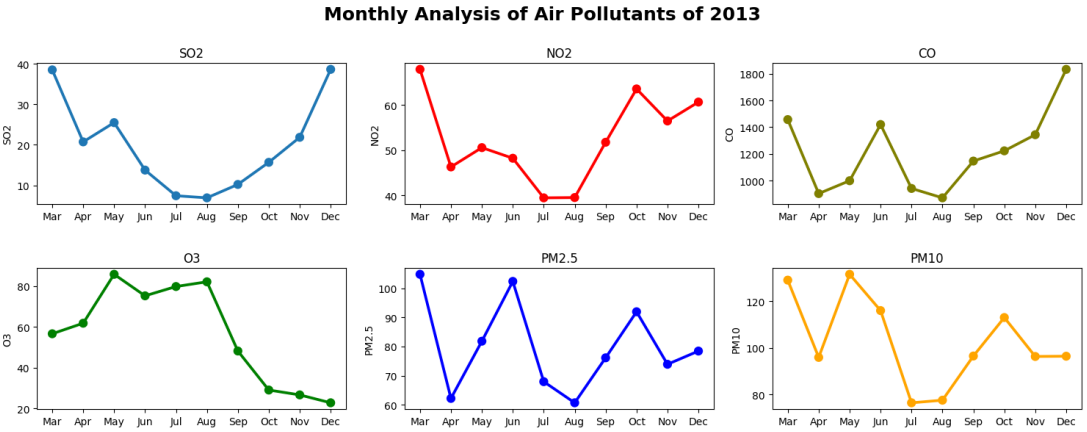


Figure 2: Monthly Analysis of 2013

Likewise in year 2013, as shown in fig.3 the concentration of pollutants decreased gradually except O3. In case of NO2, it has decreased till July and again started surging drastically whereas O3 has increased first gradually till August and then started decreasing exponentially. Eventually, air pollutants concentration has fallen down in comparison to the previous year.

As shown in fig. 4, in the year 2016, the pollutants remained as it was or decreased gradually till the month of August-September. After this, it started increasing slowly. However, this pattern is opposite in the case of ozone gas which increased at the first till June and started decreasing and fell at its min.

If we analyse the concentration of air pollutants overall monthly as shown in fig.5, we find at the beginning of the year, concentration starts decreasing till September but starts increasing afterwards and eventually becomes the same as it was at the beginning. However, it is the opposite in the case of O3 but the beginning and end points are close only.

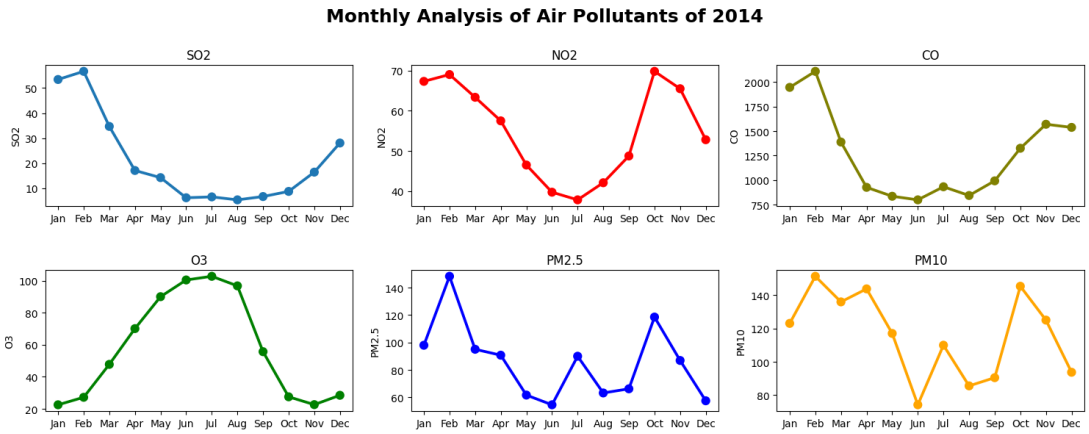


Figure 3: Monthly Analysis of 2014

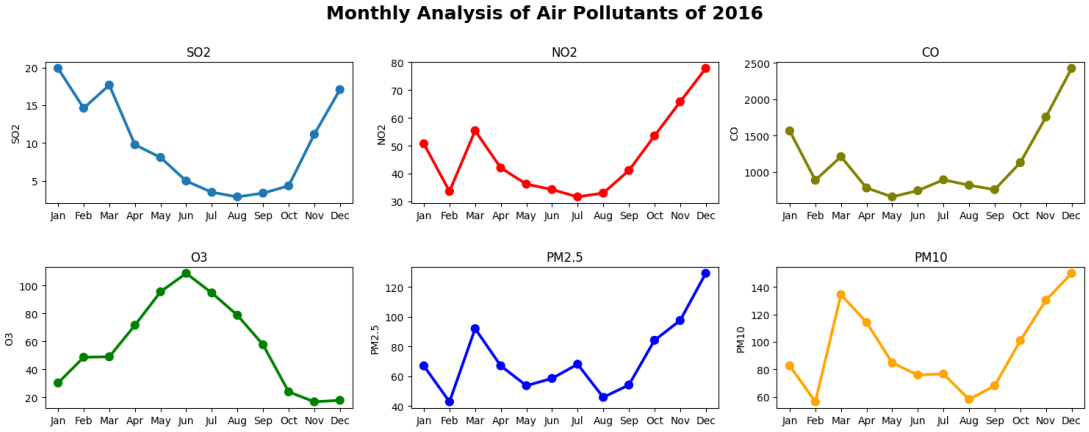


Figure 4: Monthly Analysis of 2016

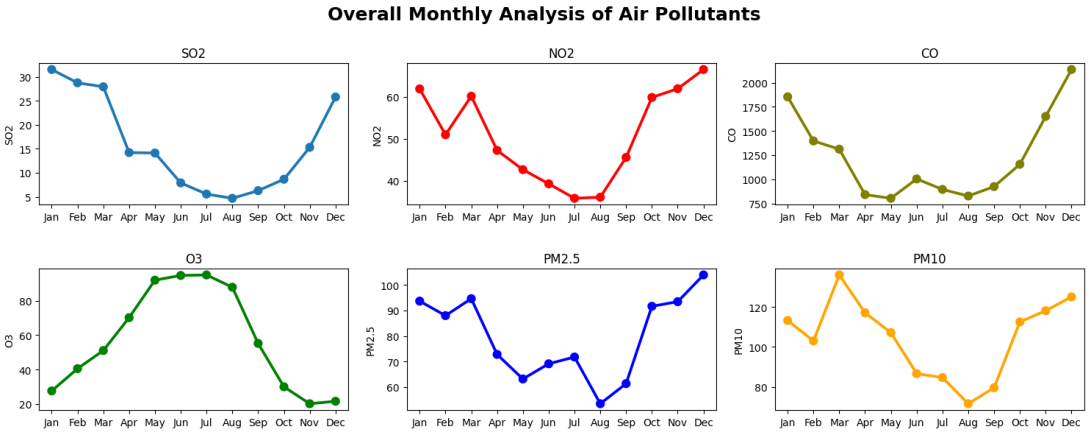


Figure 5: Overall Monthly Analysis

## 2.3 Weekly Analysis

Here data has been more precisely analysed in smaller units i.e weekly to see the pattern of changing the concentration of air pollutants in a different year. As shown in fig.6, the concentration of pollutants fluctuates and falls down on the day of Wednesday and Thursday and starts increasing. So, by the end of the week, the concentration except for O3 is found to be more than earlier.

Similarly, if we see the graph of the year 2014 as shown in fig.7, the concentration increases till Saturday and eventually falls down exponentially on the day of Sunday. It may be because Sunday is a holiday i.e no working day thus the emission from the vehicles reduces. However, O3 continues on increasing throughout the weeks. The same pattern is followed in the year 2015 also.

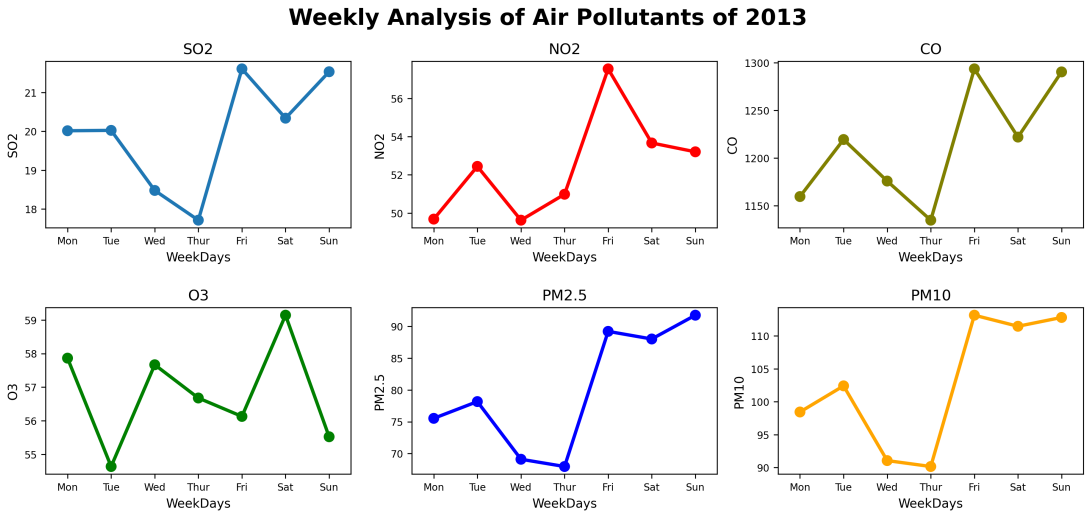


Figure 6: Weekly Analysis of 2013

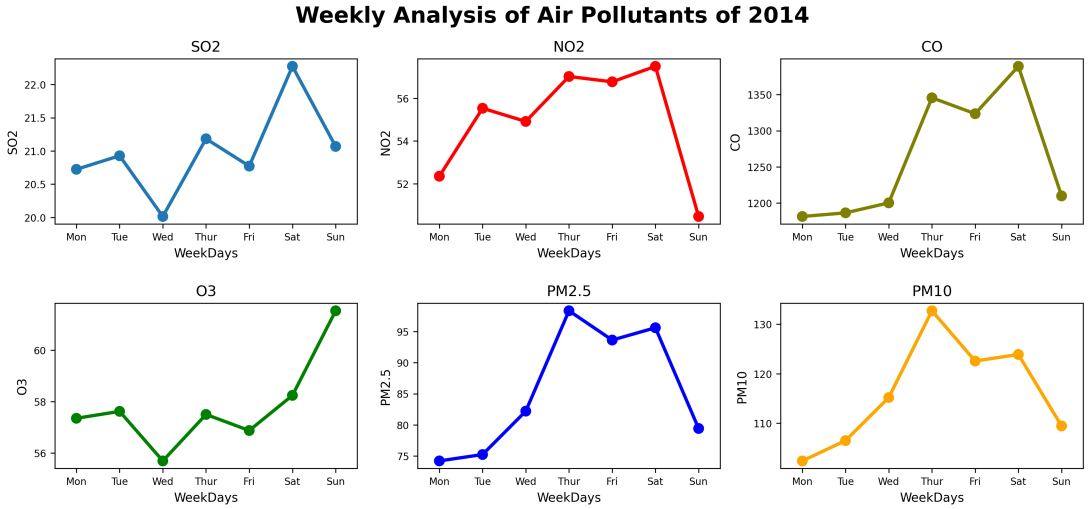


Figure 7: Weekly Analysis of 2014

In the year 2016, the concentration keeps on hitting the top on the day of Wednesday and starts decreasing slowly. However, the concentration exceeds the beginning concentration i.e it slowly increases in the year 2016. The O3 fist falls down drastically on the first day and starts increasing and reaches somewhat close to the beginning point.

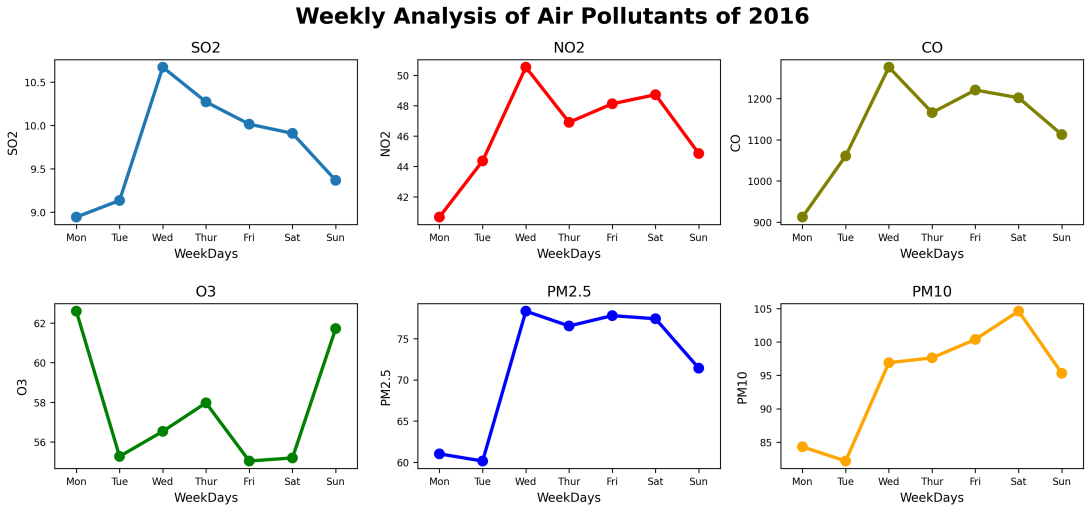


Figure 8: Weekly Analysis of 2016

By seeing the overall variation of concentration over the week as shown in fig.9, the concentration has increased till the day, Saturday and has fallen down yet exceeded the beginning conc. Overall, the concentration has increased gradually. It seems repeating the same pattern.

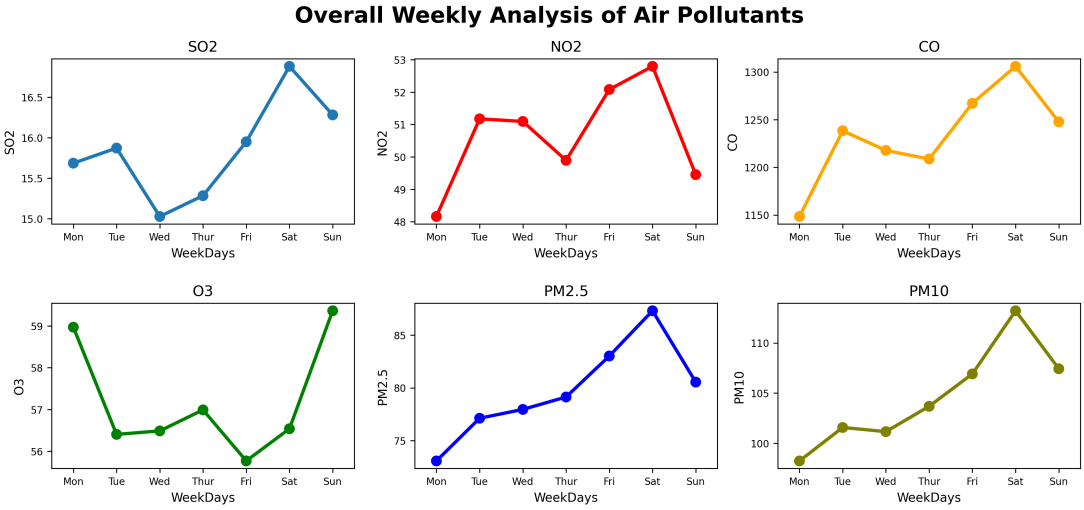


Figure 9: Overall Weekly Analysis

## 2.4 Hourly Analysis

It describes the data in hourly time series. With the help of fig. 10, we can analyse how the concentration of pollutants changes over the 24 hours of a day. The air pollutant SO2 falls down in the morning time and again starts increasing after 6 am and touches the peak at noon time and again starts slowly falling. The same pattern is also followed by O3. However, others are first falling down till 3-4 pm (evening time) and start increasing drastically.

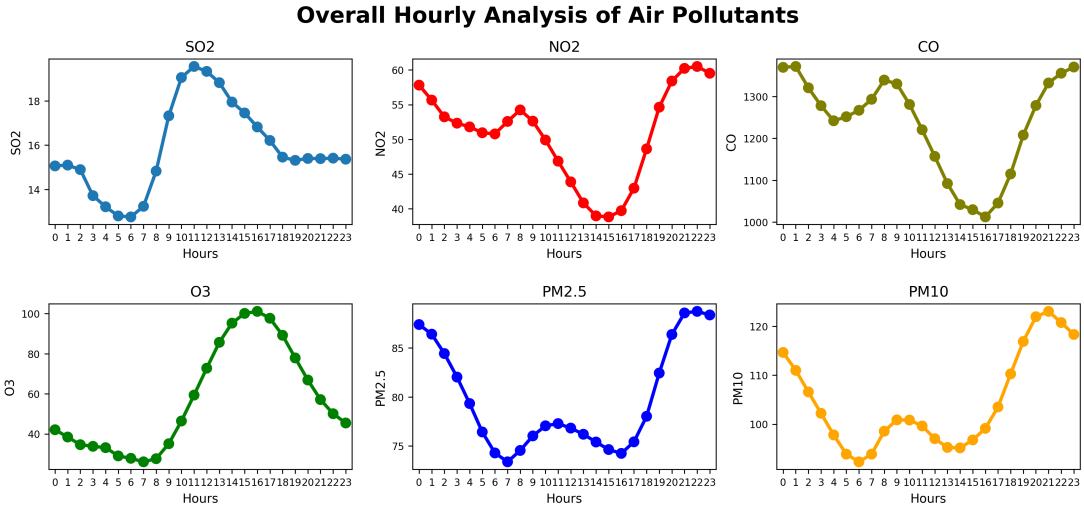


Figure 10: Overall Hourly Analysis of Air Pollutants

## 2.5 Station-wise Analysis

This illustrates the variation of air pollutants concentration variation over the period of time of different stations in Beijing city. As shown in fig.11 which depicts a variation of conc for PM2.5. It is interestingly found that concentration in each station has increased in the year 2014 and started decreasing. We see the concentration has surged in 2017. It is because there is complete data for 2017. So here we neglect 2017. The other pollutants have followed somewhat the same pattern.

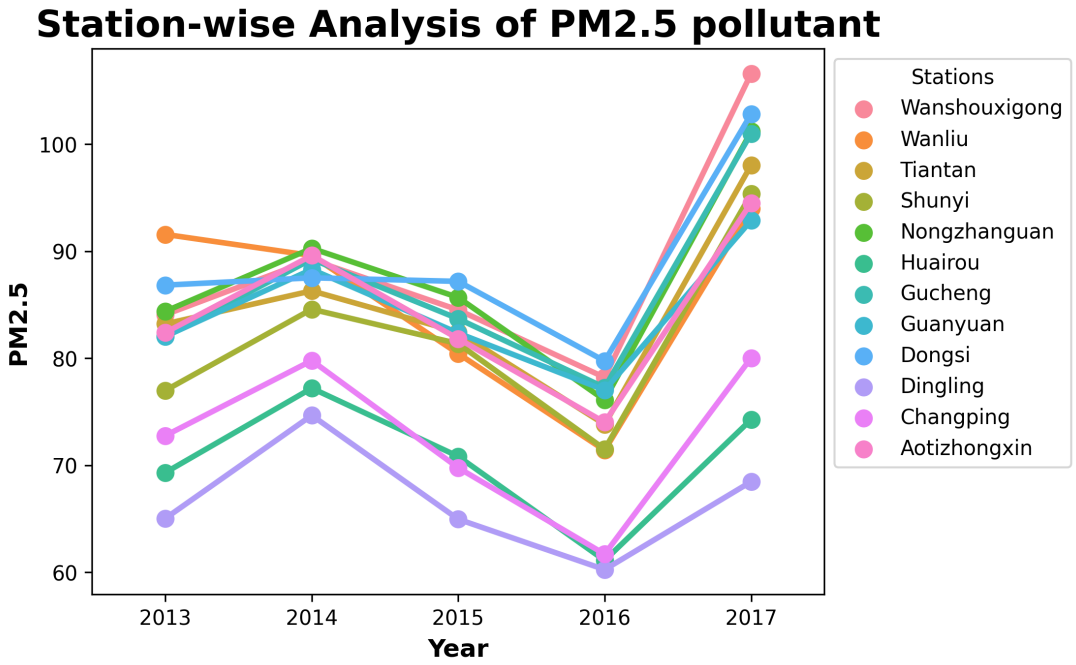


Figure 11: Station-wise Analysis of PM2.5

Similarly, we can see the concentration variation pattern of SO2 over the years. It gradually decreased till the year 2016 and then started increasing in each station as shown in fig.12.

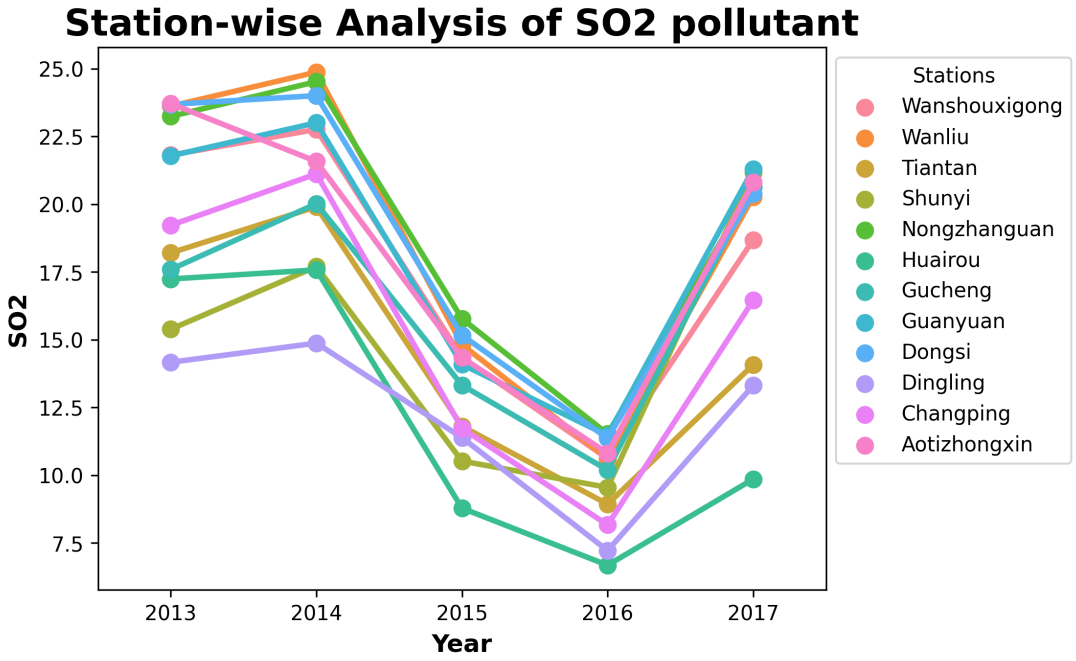


Figure 12: Station-wise Analysis of SO2

## 2.6 Greenhouse Effect

We know there is a great impact of air pollutants on the greenhouse effect as some of them absorb the sunlight which eventually increases the atmospheric temperature. As a result, the sea surface level starts increasing causing the drowning of islands. As shown in fig. 13, ozone gas (O3) directly impacts the greenhouse. Another causes the severe cold thus causing the ecological imbalance.

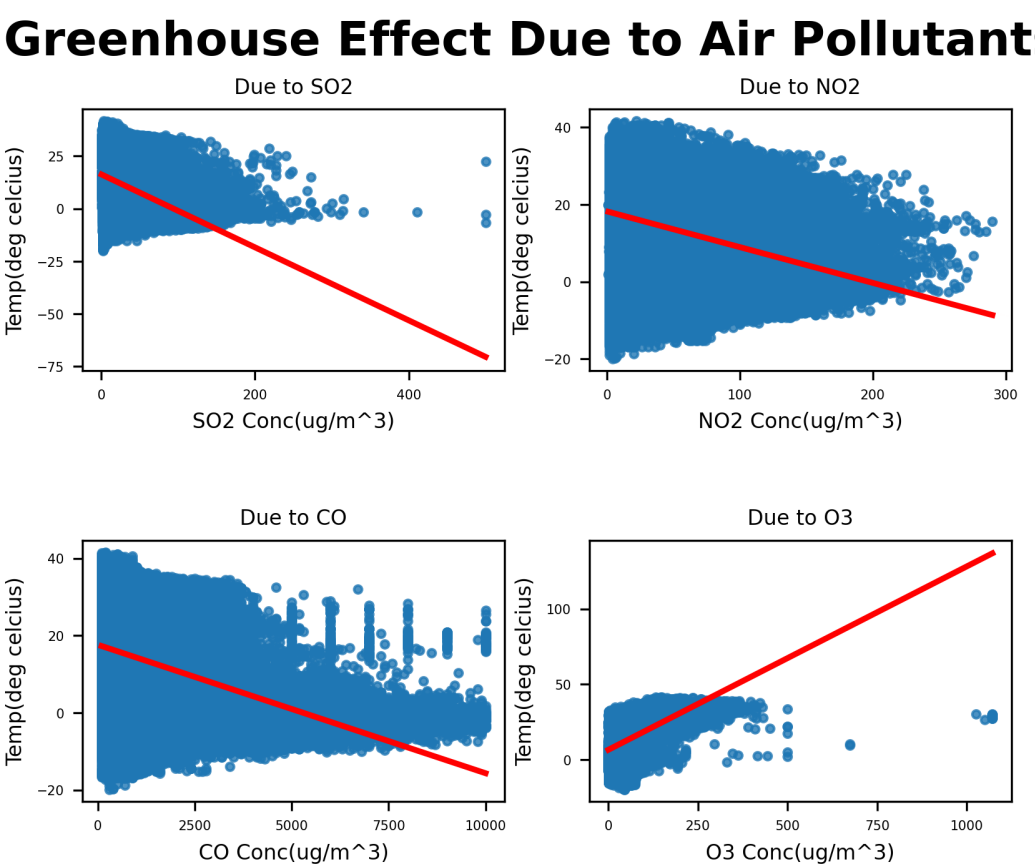


Figure 13: Greenhouse Effect of Pollutants

## 3 Correlation

With the help of data analysis and visualization tools, the correlation between the attributes has been explored and plotted visual graphs for effective and easy study. Particulate matter (PM) is a fine particle which can easily penetrate into the lungs and causes lung diseases, is formed due to the reaction of other air pollutants such as SO2, NO2, CO, O3, etc. So, here we find to what extent air pollutants are responsible for this.

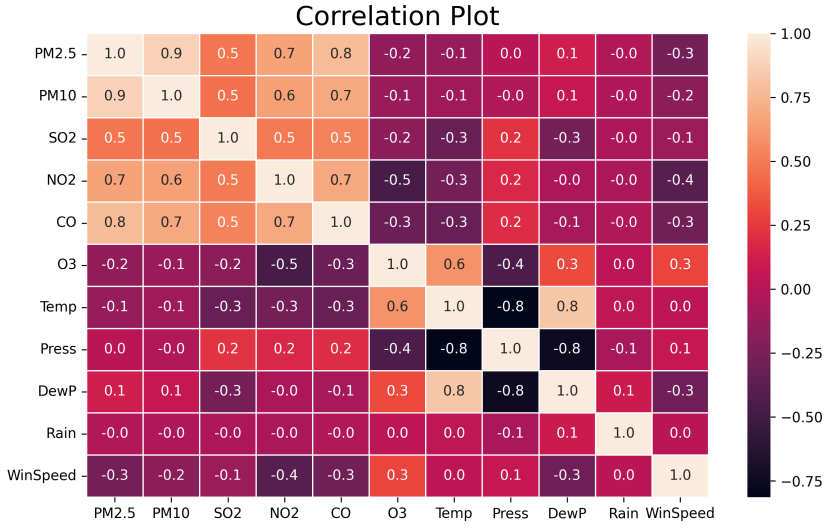


Figure 14: Correlation Graph

We have two categories of PM i.e PM10 (particle of size less than 10 mm) and PM2.5 (particle of size less than 2.5 mm). As shown in fig.14, PM2.5 is positively correlated with SO2 by 0.5, NO2 by

0.7, and CO by 0.8 whereas negatively correlated with O3 by 0.2, temperature by 0.1 and windspeed by 0.3. Similarly, we can similar impact of pollutants on PM10. It is seen that PM is more affected due to the emission of carbon monoxide (CO) gas whose main source is smoke. The prime source of smoke is gasoline vehicles, and industries. So, the more emission of CO, the more PM forms and eventually more people suffer from lung diseases. The greenhouse is highly correlated with pollutants O3 positively by 0.6, and other pollutants negatively.

If we try to visualize the correlation, we can see in fig.15 where PMs are plotted against pollutants. The figure depicts that PMs concentration increases with an increase in the concentration of SO2, NO2 and CO whereas it decreases with an increase in O3 slowly. It also shows the temperature and dew point have no such impact on PM as graph is parallel with x-axis.

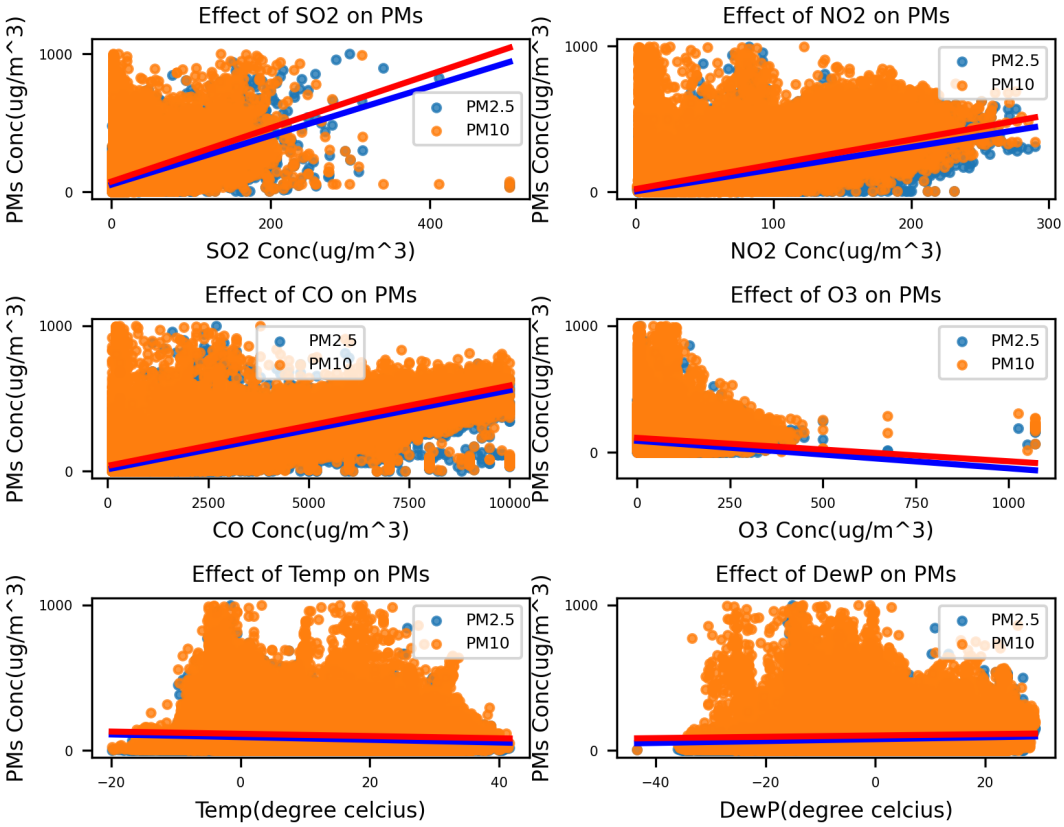


Figure 15: Effect of Air Pollutants on PM2.5

## 4 Clustering

By using a clustering algorithm (K-means), data has been segregated into four effective groups. To find the optimal number of clusters, Elbow method has been used where wcs (within cluster of the sum of squares value) or inertia of different models is analysed to get an idea about the optimal number as shown in fig.16. From the figure, the wcs has decreased drastically as the number of clusters has increased and after 4 clusters, it has started in steady state. Thus, we have taken 4 as an optimal number of clusters. Before applying K-Means, PCA (Particle Components Analysis) helps in retaining the important information and thus helps in proper visualization of clustering. It also helps in dimension reduction by combining similar attributes.

For the K-Means, only highly correlated attributes have been chosen. The attributes are PMs, SO2, NO2, CO, O3, Temperature, Dew Point, and Wind speed. After initializing the KMeans model with 4 clusters, data is fed to segregate into groups. The clustering has been shown in fig. 17. In the

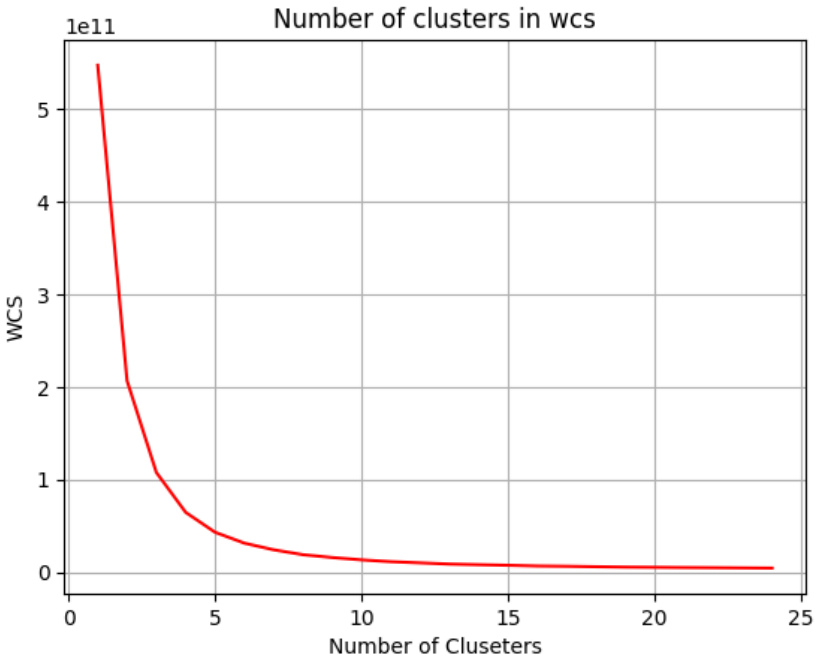


Figure 16: Number of Clusters using Elbow method

figure, red marks represent the centroids of each cluster.

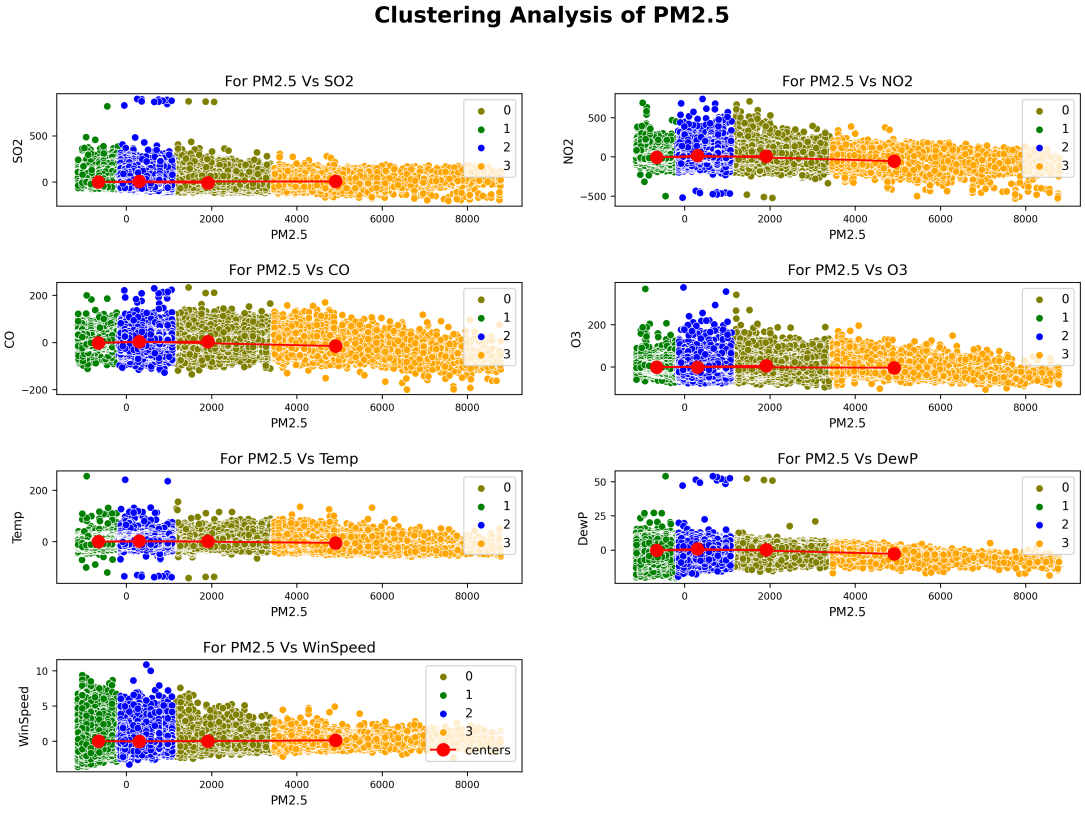


Figure 17: Clustering based on PM2.5

## 5 Hypothesis Testing

Hypothesis testing is a methodology of analyzing a statement or assumptions statistically. Similarly, we will create an assumption and based on sample statistics, we will analyse the assumptions. The hypothesis for our problem is to study the difference between the two sites in a city.

Null Hypothesis, $H_0$ : Mean value of pollutant (PM2.5 OR PM10) two sites of a city is equal

Alternate Hypothesis, $H_1$ : Mean value of pollutant (PM2.5 OR PM10) two sites of a city is not equal

For this testing, the T-statistic has been used. The procedure for this test as follows:

I. The data is normalized using PowerTransformer method of sklearn.preprocessing.

II. The unique elements are extracted to reduce the redundancy in data.

III. Their sample sizes are stored ie. n1 and n2.

IV. The sample means of the samples i.e mean1 and mean2 are computed from the unique data sample.

V. Their variances are computed from the unique data sample.

VI. Based on 95% confidence, the interval is calculated

$T\_val = T.Inv.2RT(0.05, min[n1, n2])$

$CI = (mean1 - mean2) \pm T\_val * sqrt(var1/n1 + var2/n2)$

VII. If the interval includes zero then the null hypothesis is accepted otherwise rejected.

Here we can see some results of the hypothesis testing in fig. 18.

```
[Aotizhongxin, Changping]--->[0.0619, 0.5593]
Since the interval does not include, hence it fails to accept null hypothesis.

[Aotizhongxin, Dingling]--->[0.1449, 0.6479]
Since the interval does not include, hence it fails to accept null hypothesis.

[Aotizhongxin, Dongsi]--->[-0.1373, 0.3510]
Since the interval includes 0, hence it accepts null hypothesis.

[Aotizhongxin, Guanyuan]--->[-0.2132, 0.2738]
Since the interval includes 0, hence it accepts null hypothesis.

[Aotizhongxin, Gucheng]--->[-0.1570, 0.3320]
Since the interval includes 0, hence it accepts null hypothesis.

[Aotizhongxin, Huairou]--->[0.1110, 0.6071]
Since the interval does not include, hence it fails to accept null hypothesis.

[Aotizhongxin, Nongzhanguan]--->[-0.1632, 0.3242]
Since the interval includes 0, hence it accepts null hypothesis.

[Aotizhongxin, Shunyi]--->[-0.2367, 0.2552]
Since the interval includes 0, hence it accepts null hypothesis.
```

Figure 18: The Some Results of Hypothesis Testing

## Conclusion

In this report, the given time series data has been explored year-wise, month-wise, week-wise, hour-wise, and station-wise. Interestingly, it has been found that air pollutants first increased in 2013 and then started decreasing drastically. It may be because the government might have brought new rules against gas emissions. Similarly, it first decreases till August-September and then starts increasing suddenly and exceeds the initial value whereas it is the opposite in the case of O3. However, the fall in the concentration of pollutants is seen on Sunday and it is at its peak on Saturday except for O3. The greenhouse effect has been analysed and SO2, NO2, and CO have been found to be the most negatively impacting pollutants whereas O3 is positive. The clustering using Kmeans has been done in four clusters and its results have been discussed along with other methods such as Elbow and PCA. Eventually, hypothesis testing has been implemented to study the difference in pollutants of two sites in a city.