Indian Institute of Technology Gandhinagar

Summer Research Internship Program 2023

Final Project Report

Project: Air Quality Prediction

Tanvi Jain

Intern Id: 21303438

L.D College of Engineering, Ahmedabad

B.E Computer Engineering

Abstract

- Air pollution in India is estimated to kill about 2 million people every year; it is the fifth largest killer in India.
- India has the world's highest death rate from chronic respiratory diseases and asthma, according to the WHO.
- In Delhi, poor quality air irreversibly damages the lungs of 2.2 million or 50 percent of all children.

- Air pollution is a global issue of paramount importance, posing significant risks to public health. Accurate and timely forecasting of air pollution levels is crucial for effective implementation of mitigation measures.

Problem Statement

- However, air quality prediction is often challenging, and conventional models may struggle to outperform baseline predictions, particularly in the context of Delhi's air quality. To overcome this hurdle, it is essential to gain deep insights into the nature of air pollution before developing robust prediction models.
- Recognizing this need, we have developed the comprehensive data analysis and visualization package, "**Vayu**". This innovative tool empowers researchers by providing them with a range of powerful analytical capabilities to delve into air quality data.

Aim of Vayu

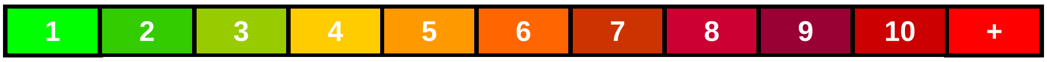
- By exploring the **spatial, temporal, and seasonal patterns** of air pollution, researchers can gain a holistic understanding of the factors contributing to poor air quality.
- By combining insights from Vayu with advanced modelling techniques, our approach **enables evidence-based decision-making** to develop precise strategies for combating air pollution.

Domain Knowledge

- **Air Quality Index:** AQI is a numerical representation of complex data set of criteria pollutants (NOx, CO, O_3, SOx, $PM_{2.5}$, PM_{10}) to understand the air quality situation for public information purposes at local and regional scale.

Figure 1: Air Quality Index

AQI category (range)	Good (0–50)	Satisfactory (51–100)	Moderate (101–200)	Poor (201–300)	Severe (301–400)	Hazardous (401–500)
Color	Deep Green	Light Green	Yellow	Orange	Red	Maroon



1	2	3	4	5	6	7	8	9	10	+

Risk: Low **(1–3)** Moderate **(4–6)** High **(7–10)** Very high **(above 10)**

- National Ambient Air Quality Standard for criteria pollutants along with their sources, prescribed limits and impacts on human health.

Table 1: Pollutant its source and impact

Pollutant	Source	Impact
CO (mg/m^3)	Incomplete combustion	Risk for cardiovascular
NO_x ($\mu g/m^3$)	Lightening, high temperature burning, biomass burning, fossils fuel burning	Cause of pulmonary disorder, increase the risk of respiratory diseases
SO_x ($\mu g/m^3$)	Coal burning, volcanic emission, petrochemical burning	Affect upper respiratory system, affect nose mucous, bronchoconstriction
O_3 ($\mu g/m^3$)	Photochemical reaction	Irritation in respiratory track, cause of cough, mild chest pain, sensitive toward allergens
Pb ($\mu g/m^3$)	Vehicular exhaust, coal burning, waste burning	Acute effect on central nervous system, affect soft tissue and bone, anemia
$PM_{2.5}$ ($\mu g/m^3$)	Windblown dust, construction activity, vehicular, and industrial emission	Premature death, chronic respiratory diseases, asthma, toxic metal
PM_{10} ($\mu g/m^3$)	Dust storm, construction activity, sea salt spray	Breathing problem, respiratory symptom

Details of air quality index along with range of concentrations of criteria pollutant (PM2.5, PM10, NOX, SOX, O3, CO, Pb) and NH3.

1	2	3	4	5	6	7	8	9	10	+

Risk: Low **(1–3)** Moderate **(4–6)** High **(7–10)** Very high **(above 10)**

Table 2: Concentrations of criteria pollutant in AQI

PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)	NH_3 (24hr)	Pb (24hr)
0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.5–1.0
101–250	61–90	81–180	101–168	2.1–10	81–380	401–800	1.1–2.0
251–350	91–120	181–280	169–208	10–17	381–800	801–1200	2.1–3.0
351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
430+	250+	400+	748+	34+	1600+	1800+	3.5+

How other factors can contribute?

- **Wind Speed & Direction**: Understanding wind speed and direction helps to pinpoint the sources and forecast the trends of air pollution in a given area. When wind carries air pollution away from its source it can appear that there are lower levels of pollution coming from this source, but this pollution has just been moved elsewhere, affecting air quality in a different location

- **Rainfall:** Rain has a 'scavenging' effect when it washes particulate matter out of the atmosphere and dissolves gaseous pollutants. Removing particles improves visibility. Where there is frequent high rainfall, air quality is generally better.

- **Humidity:** Like temperature and solar radiation, water vapour plays an important role in many thermal and photochemical reactions in the atmosphere. As water molecules are small and highly polar, they can bind strongly to many substances. If attached to particles suspended in the air they can significantly increase the amount of light scattered by the particles (measuring visibility). If the water molecules attach to corrosive gases, such as sulfur dioxide, the gas will dissolve in the water and form an acid solution that can damage health and property.

Table 3: Summary of variation in the emissions inventory for PM10 and PM2.5 in five key sectors in Delhi

Sector	Variation	
	PM 2.5(%)	**PM 10(%)**
Transport	5.5-19.0	17.9-39.2
Industries	1.3-18.3	2.3-28.9
Power Plants	2.5-17.0	3.1-11.0
Road Dust	35.6-65.9	18.1-37.8
Construction	3.6-21.0	2.2-8.4

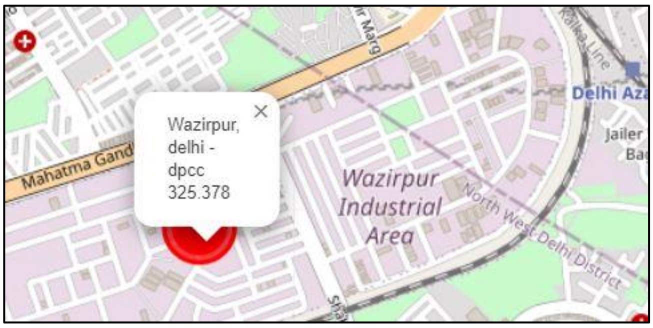
Interpretability and Insights:

Data Source:

The air quality data used in this analysis is sourced from the **Wazirpur Delhi-DPCC station,** which is a monitoring site equipped to measure various air pollutants. The dataset includes hourly measurements of pollutants such as PM2.5, PM10, CO, NOX, O3, NO2, NH3, and SO2, along with meteorological variables like wind speed, wind direction, temperature, rainfall, total rainfall and relative humidity.

Google Maps showing Air station

Figure 2: Visualizing Air Station



PM10 value for Wazirpur Air Station.

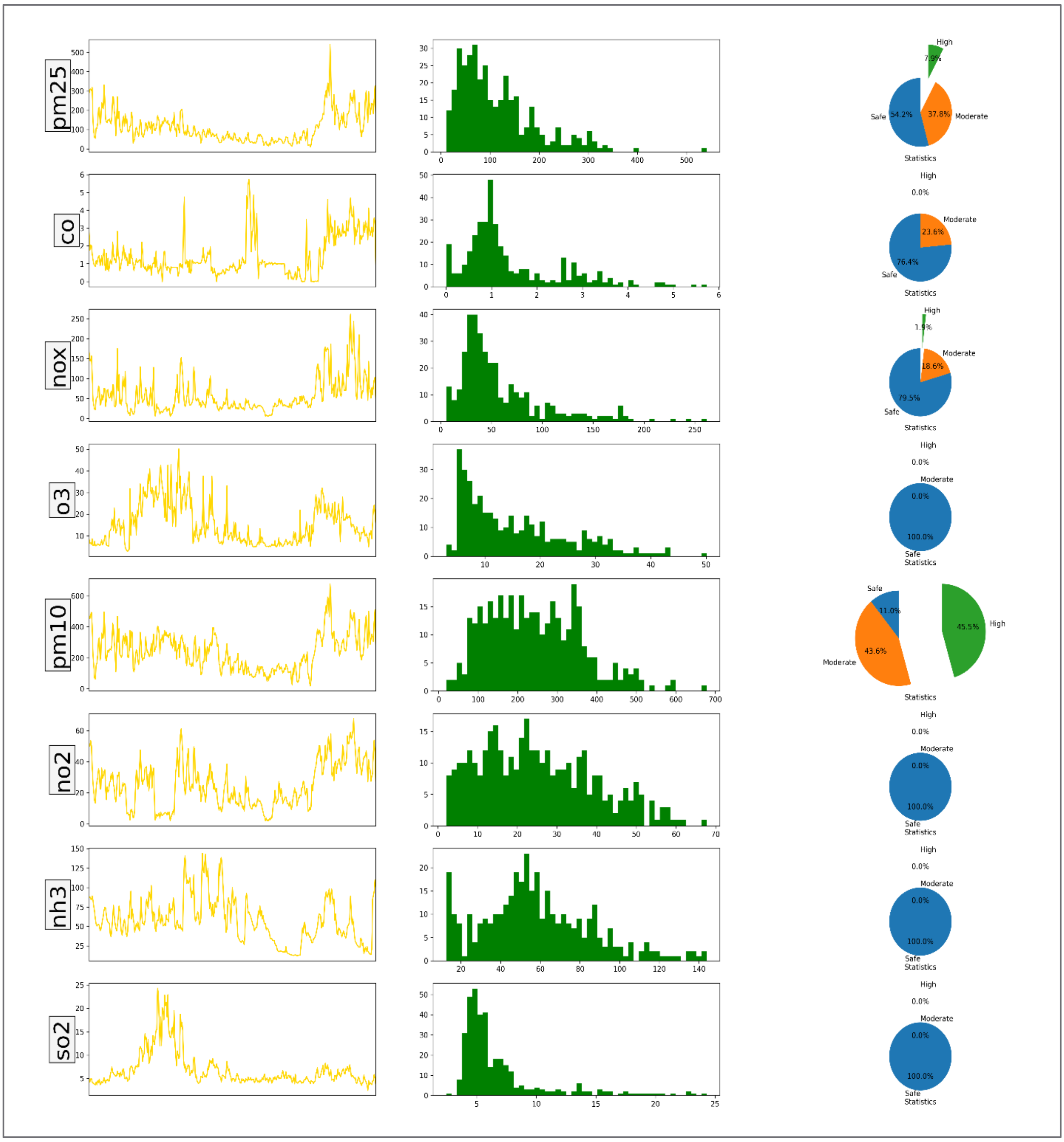
Industrial Surroundings: Using Google Maps reveals that Wazirpur Air Station is surrounded by industrial areas.
Industrial activities are known to be significant contributors to air pollution, releasing various pollutants into the air, including PM2.5. Combustion processes, emissions from factories, and industrial waste can all contribute to elevated PM2.5 levels. Hence 2.3-28.9% of PM10 arises from Industrial areas. Refer table-2

Evaluating vicinity: If air quality improvement projects have been implemented in certain areas, Google Maps can be used to track changes in the vicinity as well as air quality over time. This helps assess the effectiveness of pollution control measures.

Summary Plot

The summary plot with three subplots for each pollutant, including a time plot, histogram, and pie chart for high, low, and moderate values.

Figure 3: Summary of All pollutants



Observation: Analyzing time plot its clear to see that all the pollutants are increasing in the end months, and just before rising they were low compare to its previous values. To know much about it we should observe some seasonal plots [refer figure 13].

Figure 4: Types of Distribution

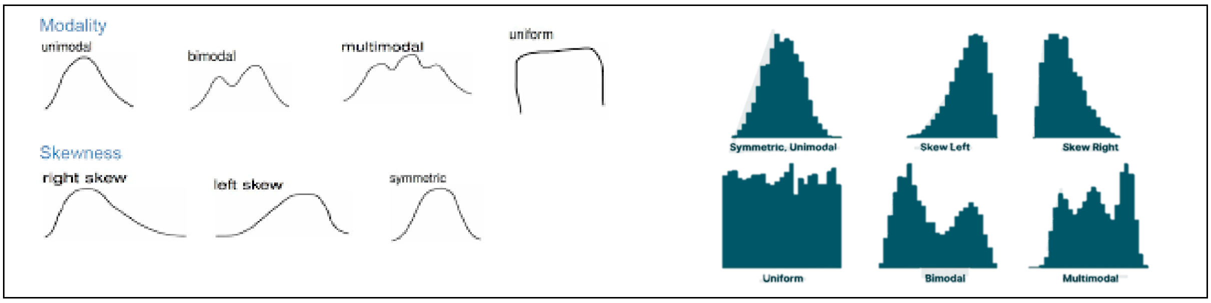
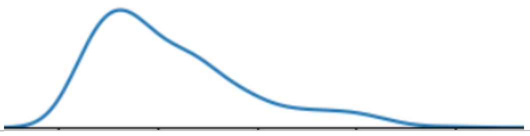
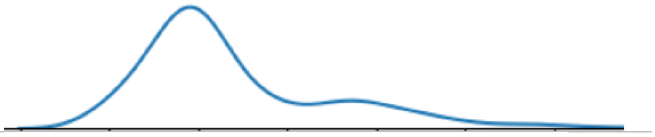
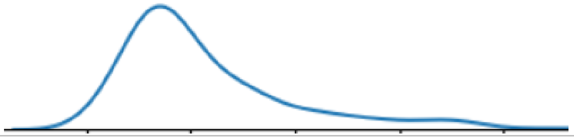
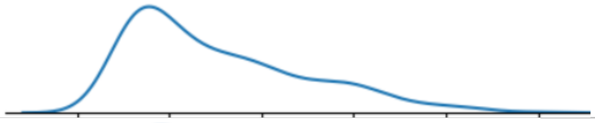
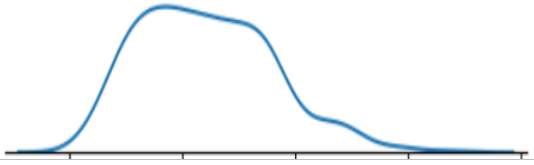
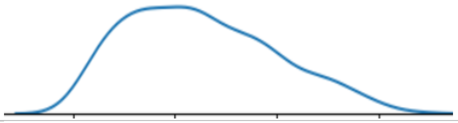
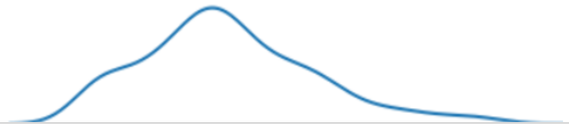
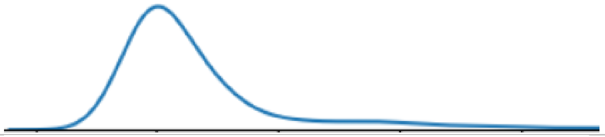
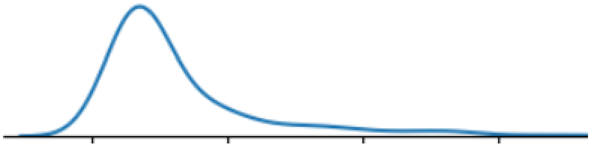


Table 4: Distribution and Variability of all pollutants

Pollutant	Distribution	
PM25	Right skewed Moderate variability	
CO	Right skewed	
NOX	Unimodal	
O3	Right skewed, High variability	
PM10	Bimodal High variability	
NO2	Right skewed High variability	
NH3	Multimodal High variability	
SO2	Unimodal	
NO	Unimodal	

Observation: After going thoroughly, we can observe that most of the pollutant are either right skewed or unimodal.

Right-skewed distributions have a long tail on the right side, indicating that there are fewer high values and more low values.

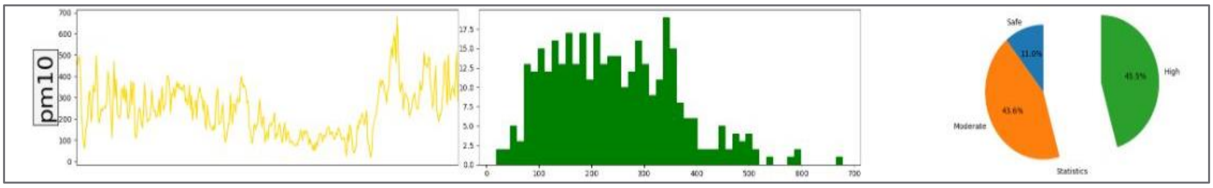
Pie Chart (High, Low, and Moderate):

Percentage of Pollution Levels: The pie chart categorizes the pollutant data into three categories: high, low, and moderate. It shows the proportion of time the pollutant falls into each category.

Table 5: Percentage Distribution of Pollutant Levels

Pollutant	High (%)	Moderate (%)	Low (%)
PM25	7.9	37.8	54.2
CO	-	23.6	74.6
NOX	1.9	18.6	79.5
O3	-	-	100
PM10	45.5	43.6	11
NO2	-	-	100
NH3	-	-	100
SO2	-	-	100

Figure 5: Observation of PM10



From table-4 we can observe that PM10 has bimodal distribution with high variability.

- **Selection of Analytical Methods:** When dealing with bimodal data, it may be necessary to apply different analytical methods or models for each mode. Using a single model for the entire data set could lead to biased or inaccurate results. Researchers and analysts may need to develop separate models for each mode or perform subgroup analysis to account for the differences between the two groups.

- **Decision Boundaries**: In machine learning and classification tasks, bimodal data can affect the choice of decision boundaries. Classifying data points in the overlapping region between the two modes can be challenging, and specific techniques like ensemble methods or Bayesian approaches might be required.

Table 6: Statistics of all pollutants

Pollutant	Min	Max	Missing	Mean	Median	Std Deviation	95th per
PM25	10.55	540.96	0	114.23	93.21	79.22	280.98
CO	0.00	5.74	0	1.41	1.02	1.09	3.56
NOX	5.45	261.68	0	58.03	44.14	43.48	157.58
O3	2.98	50.11	0	15.42	12.47	9.54	33.09
PM10	18.67	678.25	0	242.54	234.26	118.25	465.12
NO2	2.06	67.77	0	25.18	23.23	14.39	50.87
NH3	12.50	143.88	0	58.19	54.71	27.96	111.99
SO2	2.43	24.29	0	6.82	5.49	3.79	2.43

- **Minimum (Min):** The minimum value of PM10 recorded in the dataset is 18.67. This indicates the lowest concentration of particulate matter with a diameter of 10 micrometers or smaller observed at the Wazirpur Delhi-DPCC station.
- **Maximum (Max):** The maximum value of PM10 recorded in the dataset is 678.25. This represents the highest concentration of particulate matter with a diameter of 10 micrometers or smaller observed at the station.

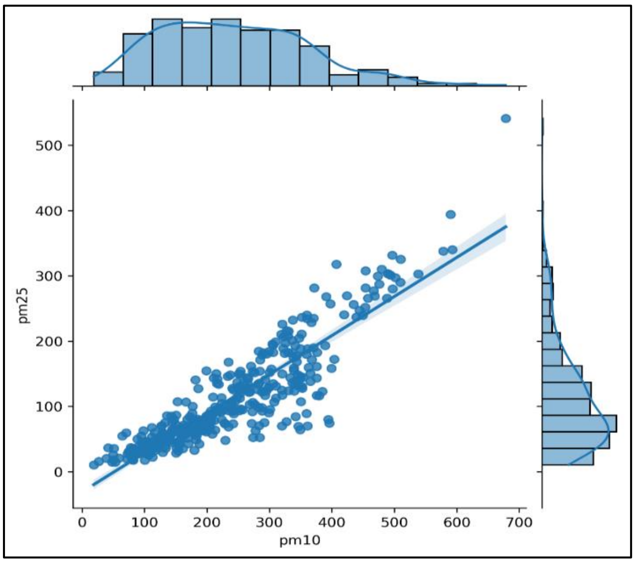
- **Missing:** 0, There are no missing values in the dataset. All data points for PM10 are present, which ensures data's completeness and reliability.
- **Mean**: The mean value of PM10 is 242.54. This is the average value, which gives an idea of the central tendency of the data.
- **Median:** The median value of 234.26 µg/m^3, which is lower than the mean, suggests that on most days, the PM10 concentration is around this value. However, this value is less affected by occasional high pollution events, such as wildfires or dust storms, which could significantly increase the mean but have less impact on the median. Thus, the median provides a more robust estimate of the typical daily PM10 concentration experienced by the population.
- **Standard deviation:** of PM10 is 118.25. This measures the spread or dispersion of data points around the mean. A higher standard deviation indicates greater variability in the PM10 concentrations.
    - Low variability: Standard deviation less than approximately 1/4 of the mean.
    - Moderate variability: Standard deviation between approximately 1/4 to 1/2 of the mean.
    - High variability: Standard deviation greater than approximately 1/2 of the mean.
- **95th percentile:** represents the value below which 95% of the data points fall, for PM10 its 465.65. This provides insight into the higher end of the distribution, indicating that 95% of the time, PM10 concentrations are below 465.65 value, but 5% of the time, they exceed it.

Observation: From Table 4 it's clear that most of the pollutant are right skewed or Unimodal, only PM10 is bimodal and nh3 looks like a multimodal. PM10 which is 45% high and rest of the pollutants are Low/Safe comparatively. Also the standard deviation of PM10 is 118.25 far from ½ of its mean which is 234, this says that PM10 is high in variability and the values goes high and low.

Bivariate Plot

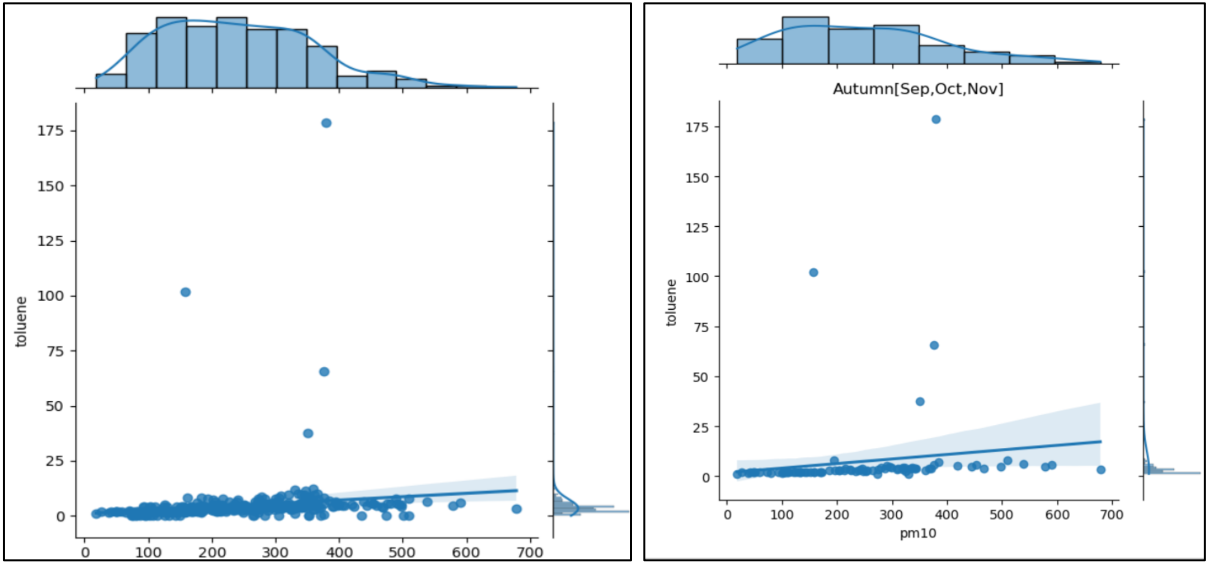
It is used for visualizing the relationship between two variables in a dataset. It creates a scatter plot of the two variables with histograms along the sides, allowing us to simultaneously examine the distributions of each variable individually and their joint relationship.

Figure 6: Comparison of PM10 with PM25



Linear Relation: If the data points form a pattern that roughly follows a straight line, it indicates a linear relationship between PM10 and PM25. This means that as PM10 concentrations increase (or decrease), there is a corresponding increase (or decrease) in PM25 concentrations. The strength of the linear relationship can be gauged by how tightly the points cluster around the line.

Figure 7: Comparison of PM10 with Toluene



Toluene also has linear relation with PM10. There are some extreme values that significantly deviate from the typical concentration ranges in Toluene. After viewing seasonal plot it can be seen that in Autumn season the value for Toluene is increasing highly than any other season**.**

Seasonal Patterns: Each mode in the distribution might represent different seasonal patterns of pollution. Some pollutants might have higher concentrations during certain seasons, which can be observe

Figure 9: Seasonal plot of PM10 with PM2.5

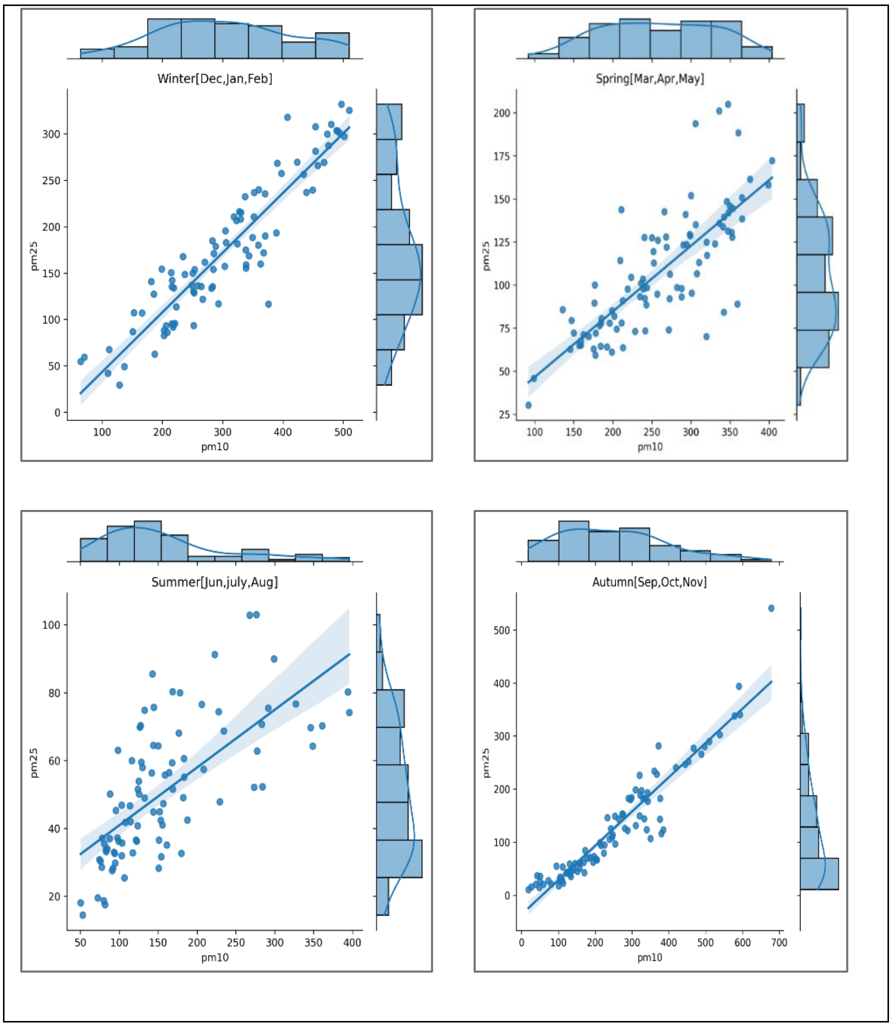


Table 7: Seasonal Analysis of PM10 with PM25

Season	Datapoints	PM25	PM10
Winter	Packed	Initial	Initial
Spring	Moderately Spread	Decreasing	Decreasing
Summer	Spread	Decreasing	Constant
Autmn	Packed	Increasing	Increasing

Observation: Here, in Summer [Jun, Jul, Aug] the datapoints are highly spread. Note: that in figure-13 both were in contrast and same month is also present here i.e. June-July.

Time Variation:

In these plots x-axis are hours and y-axis are the pollution level. Each plot shows the average value of each week day for PM10 and PM25.

Figure 10: Average hour plot for each weekday

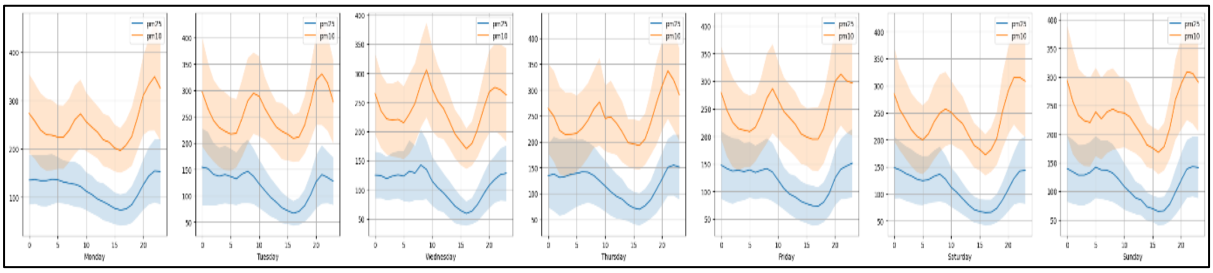


Figure 11: Annual Average hour plot of a year

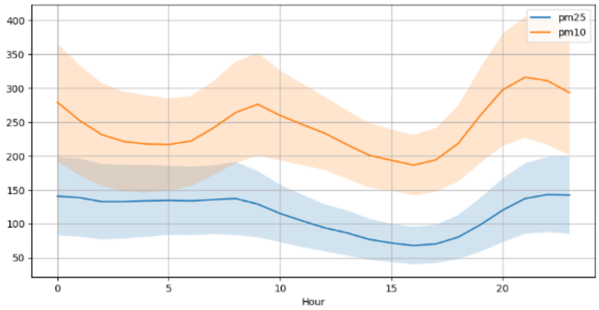


Figure 12: Annual plot for each weekday

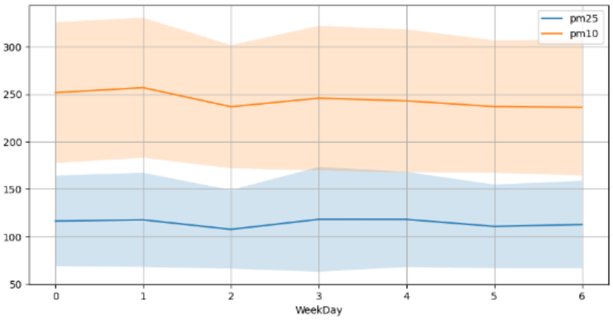


Table 8: Hourly Variation of PM2.5 and PM10 Concentrations

Hours	PM25	PM10
0-3	Constant	Decreases
3-5	Constant	Constant
5-10	Inc	Inc
10-16	Dec	Dec
16-22	Inc	Inc
22-24	Constant	Dec

Observation: If we thoroughly observe the week chart we can see that from 0 to 3 hours on x-axis, both the pollutant opposes each other PM2.5 increases and PM10 decreases. In all the other time both has linear relationship.

Figure 13: Average hour plot for each Month

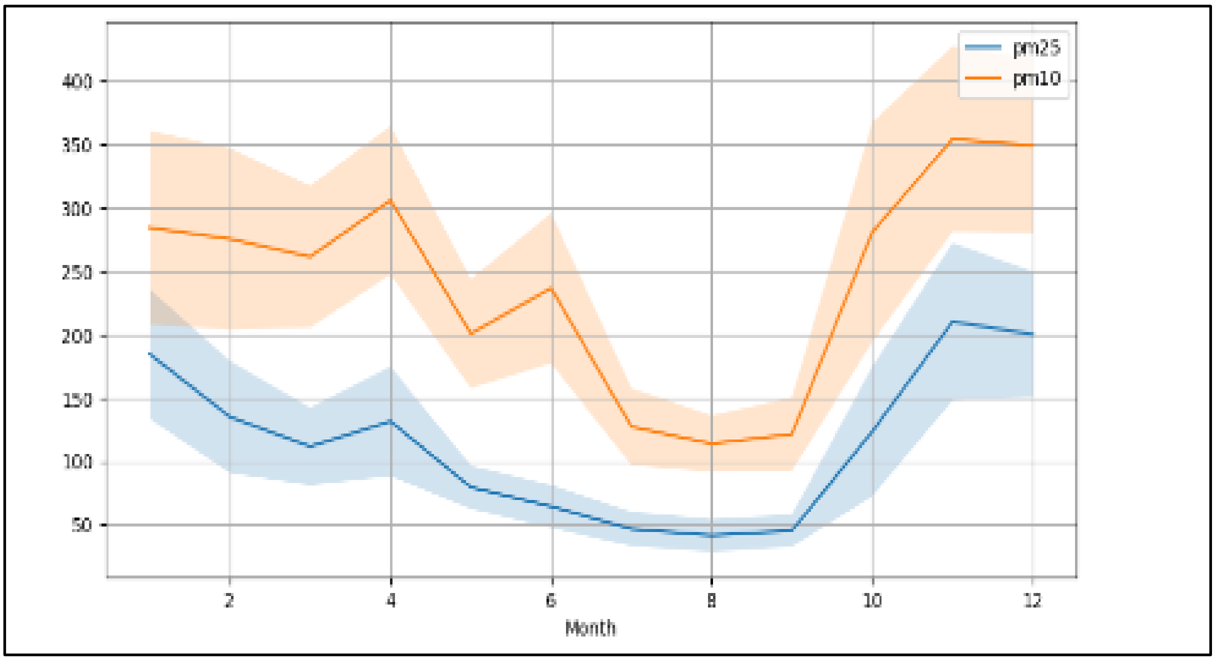


Table 9: Monthly Variation of PM2.5 and PM10 concentrations

Months	PM25	PM10
1-3	Dec	Dec
3-4	Inc	Inc
4-5	Dec	Dec
5-6	Dec	Inc
6-7	Dec	Dec
7-8	Dec	Constant
8-9	Constant	Constant
9-11	Inc	Inc
11-12	Constant	Constant

Observation: Here also both the pollutant follows the same trend except in the **month of Jun-July, PM2.5 decreases whereas PM10 increases.**

Trendlevel

The following figure shows months on x-axis and hours on y-axis, It then creates a heatmap showing each hours value averaged of particular month, for all months, The aim of this plot is to identify which hour of which month is more polluted and least polluted, inorder to find the trend.

Figure 14: Hour plot for each month of PM2.5

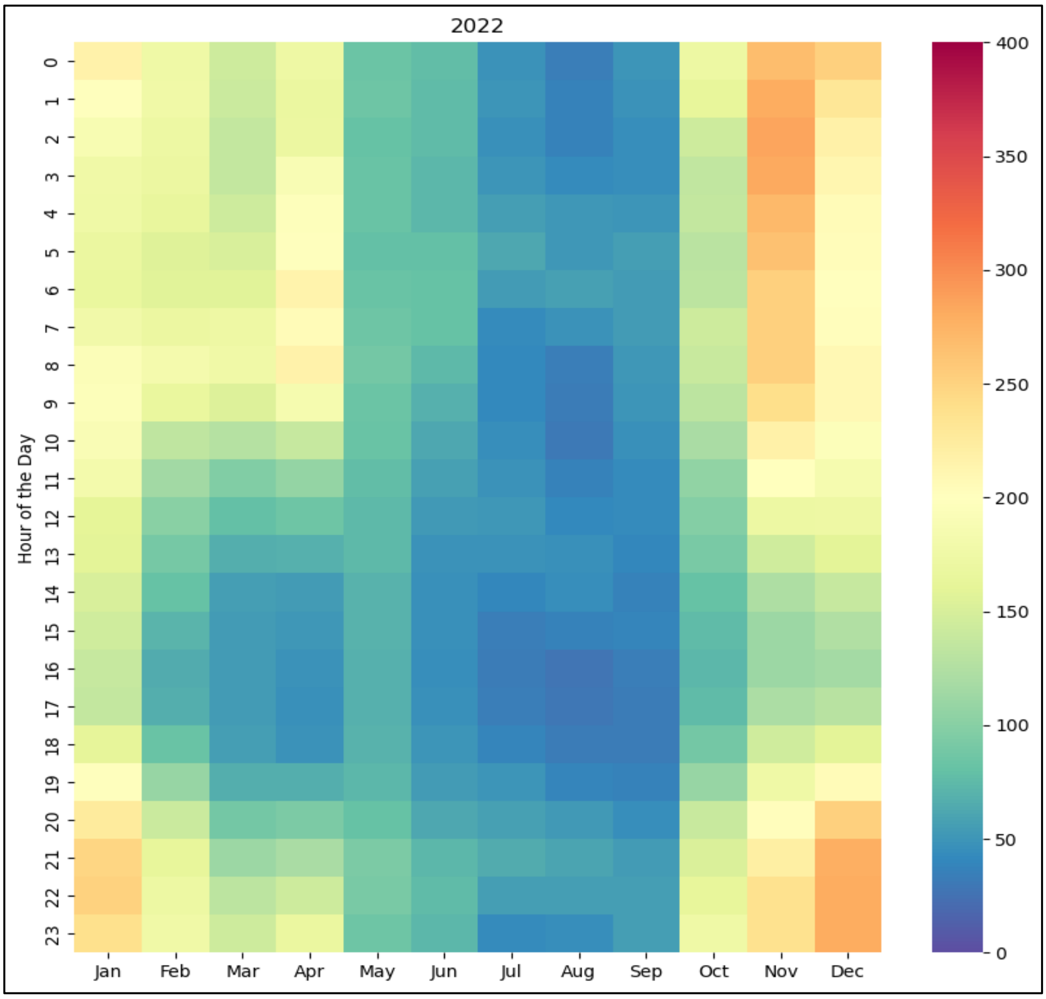


Table 10: Monthly Observation of PM2.5

Months	PM25
Jan, Nov, Dec	High
Feb, Mar, Apr	Trend
May	Constant
June, July Aug, Sept	Low
Oct	Increasing

Table 11: Hourly Variation of PM2.5

Month	Hour	Trend
Jan	20 to 0	High
Feb, Mar, Apr	10 to 21	Low
Nov	22 to 9	High
Dec	20 to 0	High

Figure 15: Averaged Hour plot for each month of PM2.5

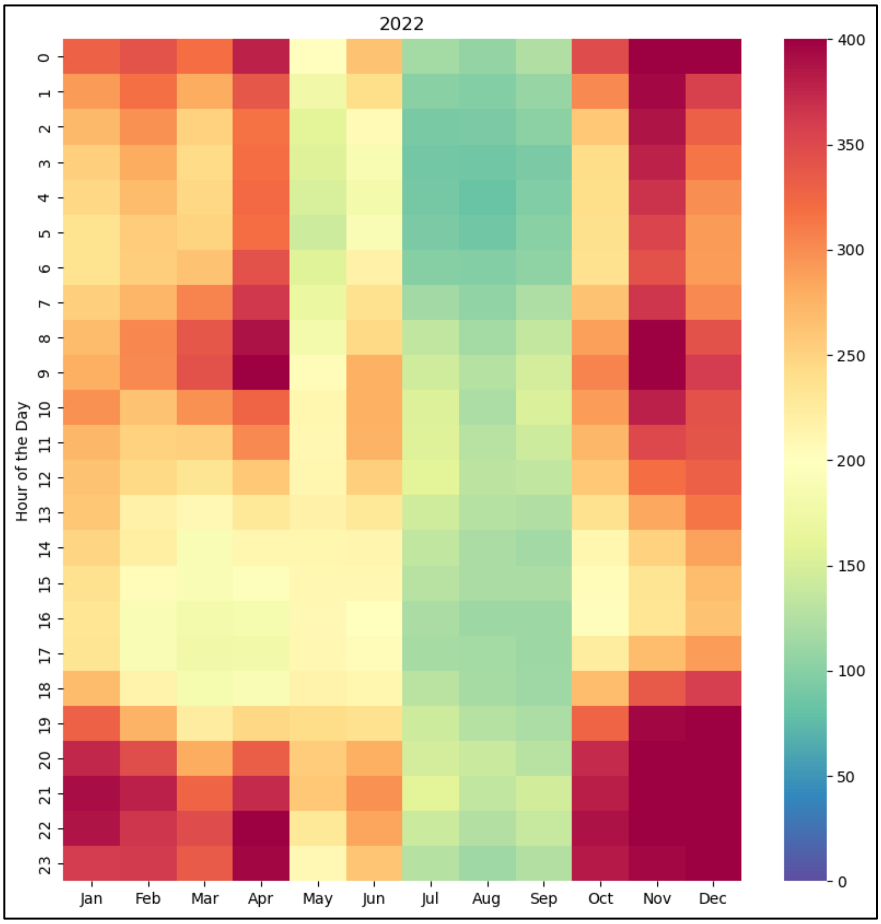


Table 12: Monthly Variation of PM10 concentrations

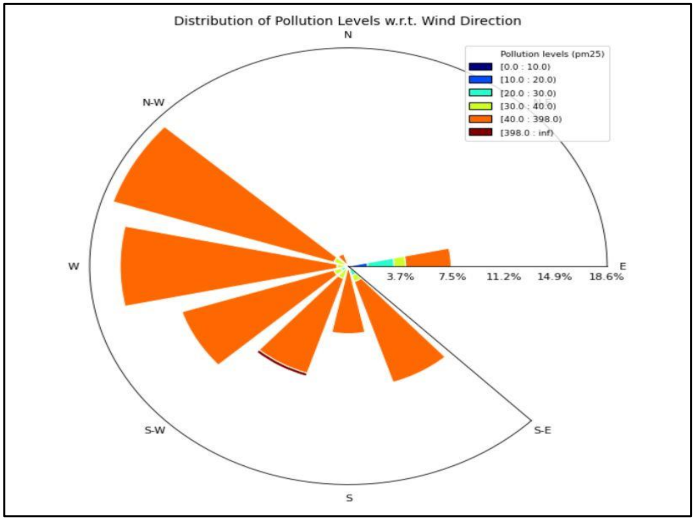
Months	PM10
Jan, Nov, Dec	High
Feb, Mar, Apr	High(trend between 13 to 19)
May	Low
Jun	High
July Aug, Sept	Moderate
Oct	Increasing

Table 13: Hourly Variation of PM10 in Months

Month	Hour	Trend
Jan,Nov,Dec	19 to 0	High
Feb, Mar, Apr, May, Jun	13 to 19	Constant

It can be clearly seen that in June, July, August PM2.5, PM10 has decrement which can be due to rainfall, hence in these month rainfall might play a good role for prediction. In Jan, Oct, Nov, Dec both the pollutant are increasing which can be due to festive seasons in India like: Navratri, Dusshera, Diwali, and finally New Year.

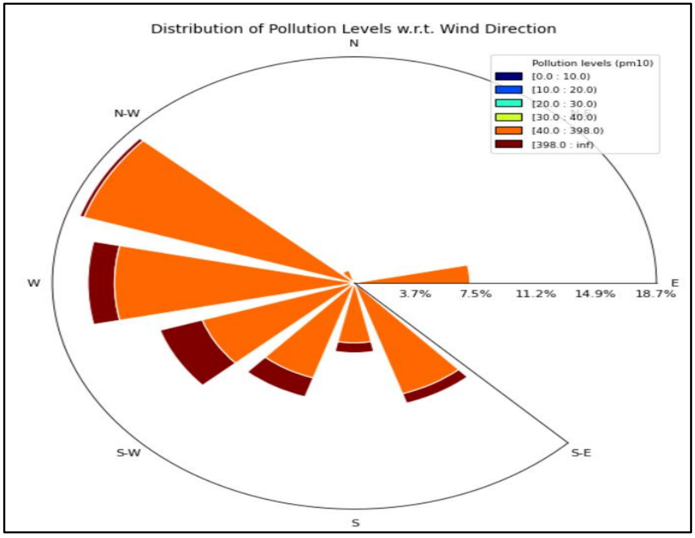
Figure 16: PM2.5 Pollution rose with respect to direction



This chart specifically tailored to visualize pollutant levels with respect to wind direction. The windrose plot helps to understand how the distribution of pollutant levels varies depending on the direction of the wind. The color of the bars or rectangles in the plot indicates different pollutant level ranges.

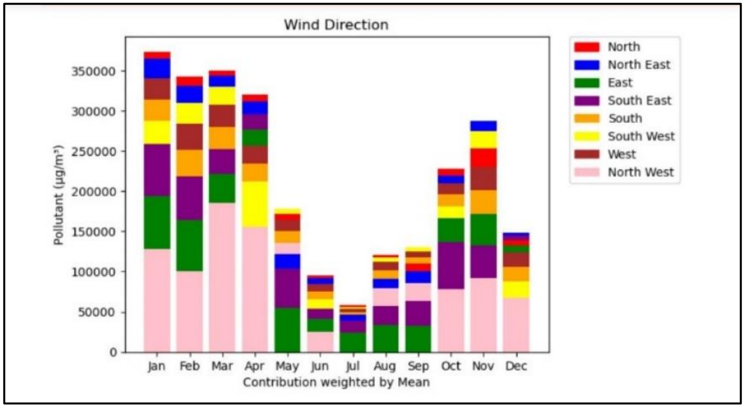
PM25 increases from south-east to north-west and mostly spread over West and North-West direction.

Figure 17: PM10 Pollution rose with respect to direction



PM10 is highly spread over West and North-West direction. Also in all the South direction i.e. South-West, South-East direction PM10 value is high.

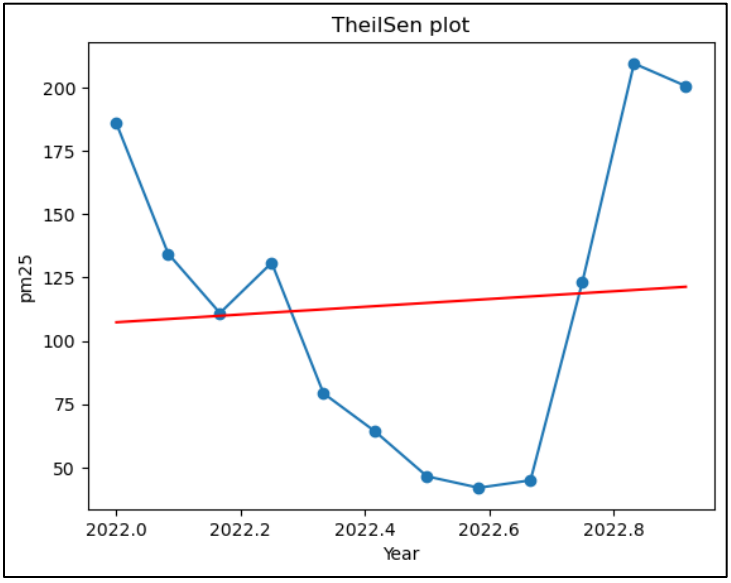
Figure 18: PM10 Pollution Proportional to wind direction



Plot a stacked bar graph of all data based on frequency of wind direction in compass directions. Takes the average of every 30 days in each bar. The height of the bar is value of the pollutant in that 30day period. The bars are binned proportionally based on the overall value of the pollutant.

Where most of the pollutant is in North-West, East and South East.

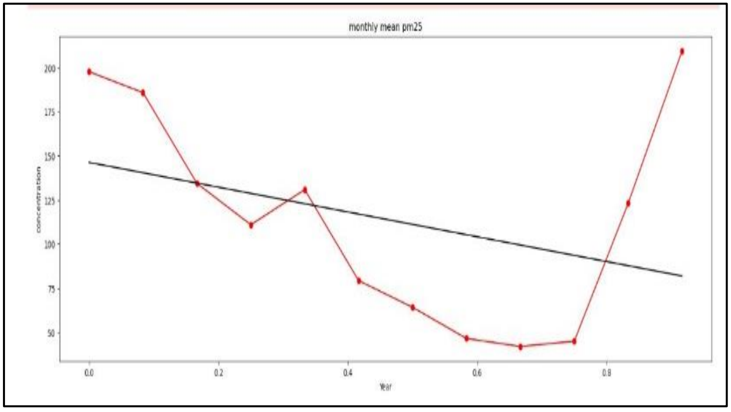
Figure 19: Theilsen Plot



Plots a connected scatter plot of the average value of the pollutant every month of every year. Then plots a line of best fit through the plot showing the user the overall trend of the pollutant through the years.

Here we can observe that the best fit line is near the mean.

Figure 20: Smooth line over dataset



Plots a connected scatter plot of the average value of the pollutant every month of every year. Then plots a smooth line of best fit through the plot showing the user the overall trend of the pollutant through the years. As we can observe that there is no smooth trend for one year

Conclusion

1. The vicinity of Wazirpur is an industrial area, which could be a significant source of pollution in the region.
2. From Table 4, we observed that most pollutants have a right-skewed or unimodal distribution, except for PM10, which exhibits a bimodal distribution. NH3 shows a potentially multimodal distribution.
3. PM10 stands out with 45% of its data points in the high category, indicating higher pollution levels compared to other pollutants.
4. The standard deviation of PM10 (118.25) is far from half of its mean (234), indicating high variability and fluctuation in PM10 concentrations.
5. The linear relationship between PM10 and toluene, as seen in the bivariate plot, suggests a potential connection between these pollutants.
6. During the summer months (June, July, August), there is a notable spread of data points, indicating a wider range of pollutant concentrations during this period.
7. In the bivariate plot for PM2.5 and PM10, both pollutants follow a similar trend, except for June-July, where PM2.5 decreases while PM10 increases. This anomaly might be influenced by specific local factors during that period.
8. From February to June, there is a significant decrease in pollutant levels between 13 to 19 hours, suggesting potential relief from pollution during these hours.
9. During June, July, and August, both PM2.5 and PM10 exhibit a decrement in values, possibly due to rainfall, which can play a crucial role in reducing pollution during these months.
10. In January, October, November, and December, both pollutants show increasing trends, possibly influenced by festive seasons in India, such as Navratri, Dusshera, Diwali, and the New Year.
11. PM25 and PM10 increases from south-east to north-west and mostly spread over West and North-West direction in increasing fashion.

Future Aim

1. Implement Baseline models for PM10 and PM25 to establish a benchmark for comparison with advanced models. This will aid in evaluating the effectiveness of more complex modelling approaches.
2. Due to the bimodal nature of PM10, explore advanced modelling techniques like Bayesian methods, sqrt transformation, and log transformation to capture the complexity of pollutant behaviour.
3. Develop a user-friendly web application to visualize and interact with the air quality data, making the insights accessible to a wider audience
4. Enhance Trend level by incorporating daily and yearly plots to provide a analysis of pollutant trends over different time scales.
5. Implement AQI color coding in all charts to improve visual understanding. Utilize colors such as red for high pollution levels, green for safe levels, and blue for moderate pollution l.
6. Categorize the trend data into three distinct levels: low, moderate, and high, aligned with the concentration criteria of each pollutant. This will provide clearer insights and avoid confusion from too many colors.
7. Conduct analysis of pollution levels in different districts to identify safe and high-risk areas for air quality. This can aid in prioritizing pollution control measures in specific regions.
8. Rank the months based on their pollution levels, highlighting which months are most polluted. This information can guide policymakers in formulating targeted pollution control strategies.

Acknowledgement

I am deeply grateful to my esteemed Advisor, Professor Nipun Batra from IIT Gandhinagar, and my co-advisor, Zeel Patel [PhD, IITGandhinagar], for their exceptional guidance and support throughout my internship at the Sustainability Lab.

Working with such a talented and supportive teacher has been an incredible learning experience. I am thankful for the valuable insights and knowledge shared by my advisors, which significantly enriched my understanding of the subject matter.