Project Component 1

Himanshu, MDS202327

2023-10-01

Introduction

In this report we try to discover patterns and make inferences about the pollution level in stations in and around Delhi. We also look at the relation between the pollutant parameters and try to find which parameter are related to each other and see if there is a pattern across all ten stations. We specifically study PM2.5 pollutant, which is known to be a good indicator of air health.

Data Description

The data contains 10,600 rows and 9 columns, namely, siteName, siteCode, Date and six air pollution parameters i.e. PM2.5, PM10, NO_2, NH_3, SO_2, Ozone for ten stations in New Delhi, collected from CPCB website from 08-02-2018 to 02-01-2021 on daily basis. There are 1060 entries for each station, one for all the dates between 08-02-2018 and 02-01-2021 (both inclusive). The data for the parameters is average of 24 hour data collected every 15 minutes. The units for all the parameters in the data are $\frac{\text{ug}}{\text{m}^3}$ that represents micrograms(one-millionth of a gram) of a gaseous pollutant per cubic meter of air.

Data file `delhi.csv` available in the repository for this project here.

siteName	**siteCode**	**Date**	**PM2.5**	**PM10**	**NO_2**	**NH_3**	**SO_2**	**Ozone**
<chr>	<int>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Sonia Vihar	1432	2019-09-19	17.62	65.71	13.18	26.37	12.64	36.09
Jahangirpuri	1423	2020-03-01	51.20	120.17	72.40	36.34	2.04	12.23
Wazirpur	1434	2020-04-12	44.46	85.50	32.24	23.36	14.07	52.15
Najafgarh	1427	2018-05-19	100.06	287.78	28.60	46.65	7.63	73.52
Patparganj	1431	2018-10-27	189.89	384.89	63.65	85.26	4.39	18.85

Table 1: A glimpse of random sample of the data.

The names of all ten stations with their respective site codes are displayed in the table below.

Site Name	Ashok Vihar	Dwarka-Sector	Jahangirpuri	Najafgarh	Narela
Site Code	1420	1422	1423	1427	1426
Site Name	Patparganj	Rohini	Sonia Vihar	Vivek Vihar	Wazirpur
Site Code	1431	1430	1432	1435	1434

Table 2: Site Names and corresponding Site Codes

Exploratory Data Analysis

Since our data has a date column, we would want to exploit it to our use to plot some time-series plots and analysis. As displayed in table 1, the data column has type chr, so we must first convert it to date type.

```r
#Changing Date data type from chr to date
df$Date <- as.Date(df$Date)
```

siteName	**siteCode**	**Date**	**PM2.5**	**PM10**	**NO_2**	**NH_3**	**SO_2**	**Ozone**
<chr>	<int>	<date>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Sonia Vihar	1432	2019-09-19	17.62	65.71	13.18	26.37	12.64	36.09
Jahangirpuri	1423	2020-03-01	51.20	120.17	72.40	36.34	2.04	12.23
Wazirpur	1434	2020-04-12	44.46	85.50	32.24	23.36	14.07	52.15
Najafgarh	1427	2018-05-19	100.06	287.78	28.60	46.65	7.63	73.52
Patparganj	1431	2018-10-27	189.89	384.89	63.65	85.26	4.39	18.85

Table 3: A glimpse of random sample of the data after chnaging type of Date column.

A Histogram

We would naturally want to see the range of the values attained by all the parameters and the density with which these are attained. Hence, histograms come handy.

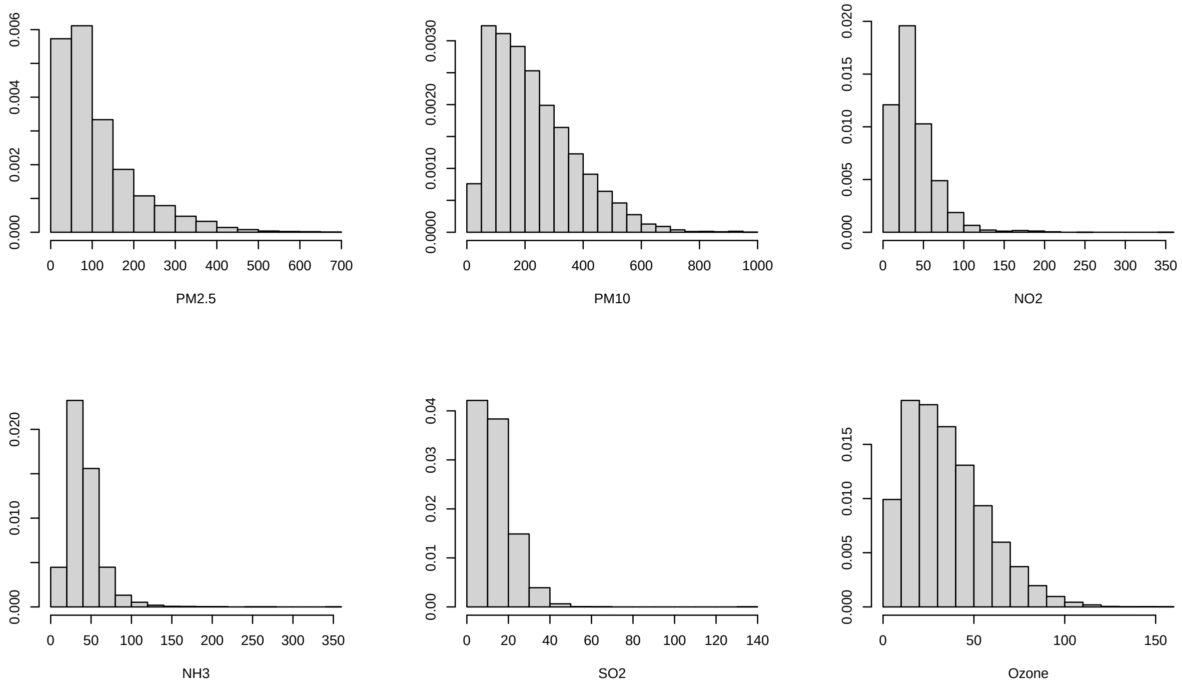


Figure 1: The above graph shows the histogram of values of the six pollutant parameters.

A Pair Plot

An important part of exploratory data analysis is to find out which parameters are correlated, positive or negative. This helps one decide on which parameters to include in their prediction model. Since there are 10,600 data points in our data and also since all the stations are present in Delhi, without the loss of generality, we will look at the pair plot of parameter values for the station Rohini, station code 1430.

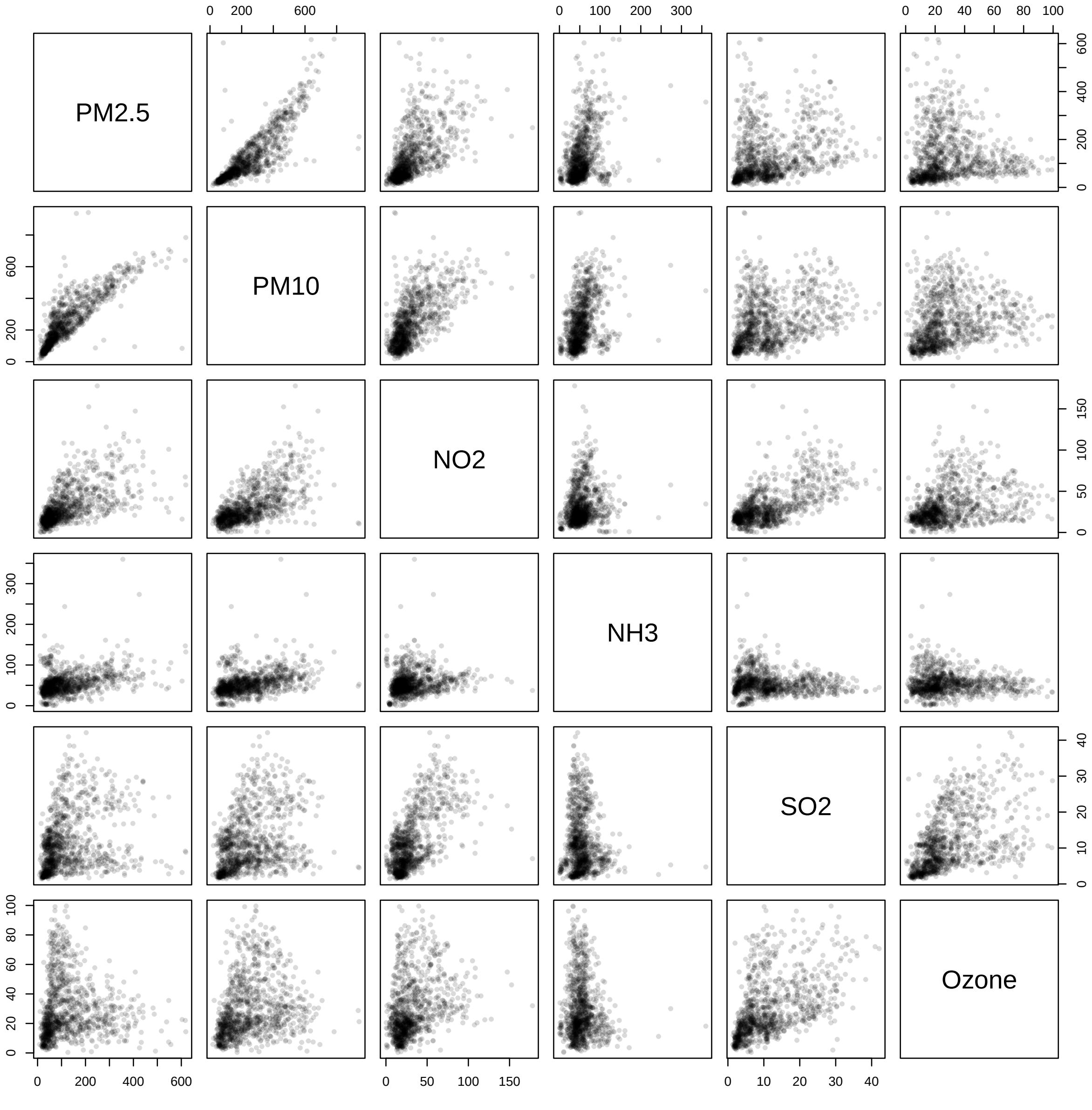
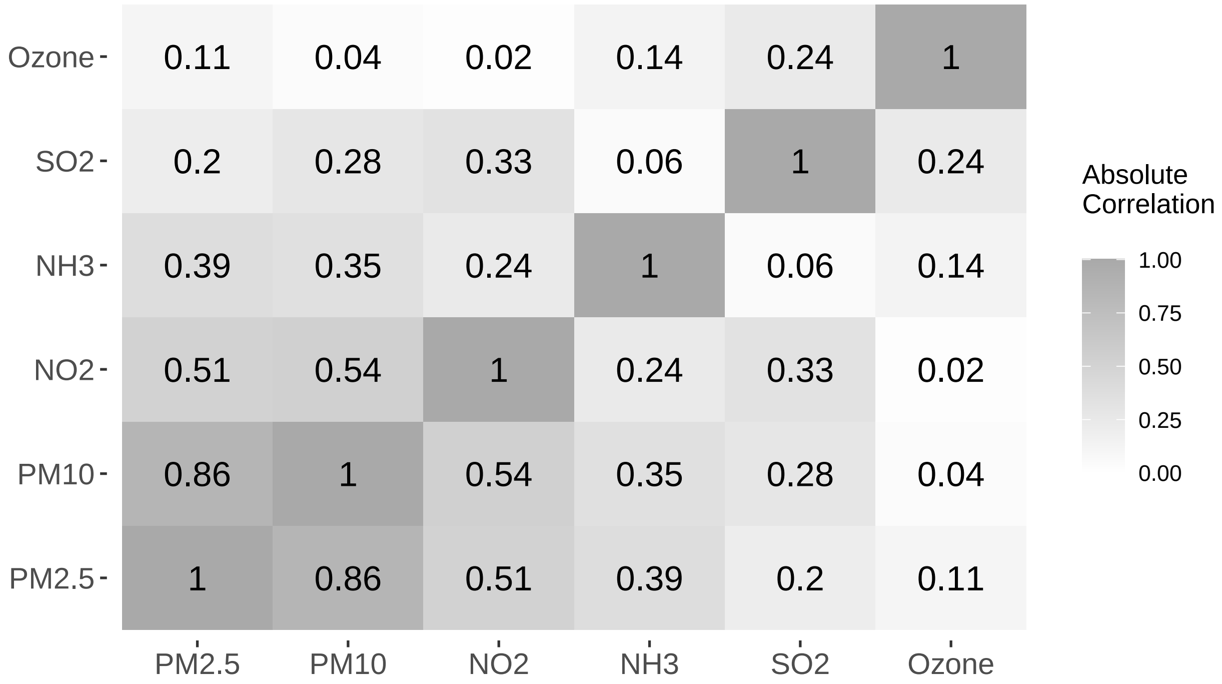


Figure 2: The above graph shows the pair plot for all the six pollutant parameters.

A Correlation Plot

A numerical view to pair plot is a correlation plot. The patterns between pollutant parameters we observed in the pair plot are expressed as a number between -1 and 1 called correlation coefficient. Here, we are not differentiating between positive correlation and negative correlation. We just look if they are correlated be it positive or negative. Hence we plot the absolute value of the correlation coefficient between parameters.



	PM2.5	PM10	NO2	NH3	SO2	Ozone
Ozone	0.11	0.04	0.02	0.14	0.24	1
SO2	0.2	0.28	0.33	0.06	1	0.24
NH3	0.39	0.35	0.24	1	0.06	0.14
NO2	0.51	0.54	1	0.24	0.33	0.02
PM10	0.86	1	0.54	0.35	0.28	0.04
PM2.5	1	0.86	0.51	0.39	0.2	0.11

Figure 3: The above graph shows the correlation plot for all six pollutants in the data.

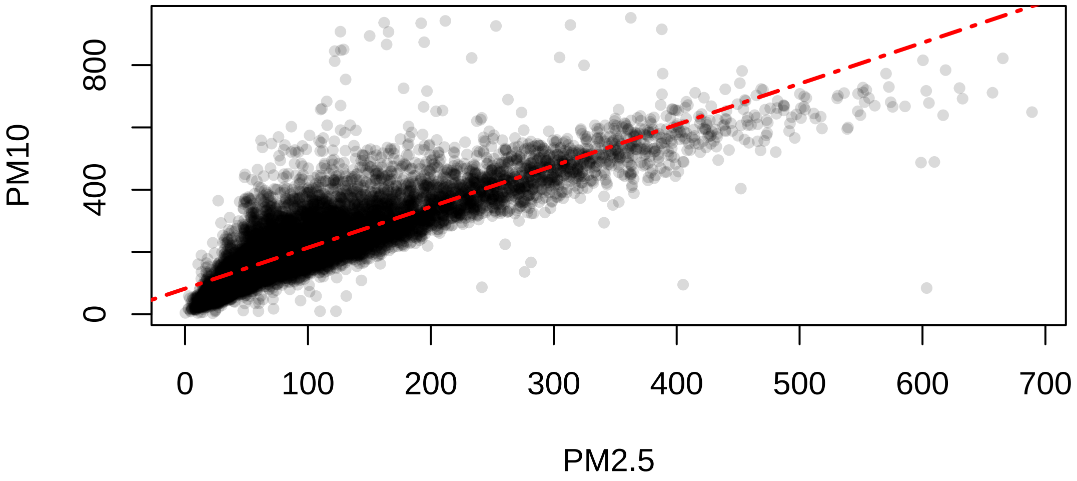


Figure 4: Scatter plot of highest correlated parameters PM2.5 and PM10.

A Time-Series Plot

Since we have a time-stamped data, one of the basic analysis in time-series is to find the trends across time. For the sake of simplicity we will show time series plots for only one parameter i.e. PM2.5 for all the stations. The y-axis has been adjusted to same scale for ease of analysis.

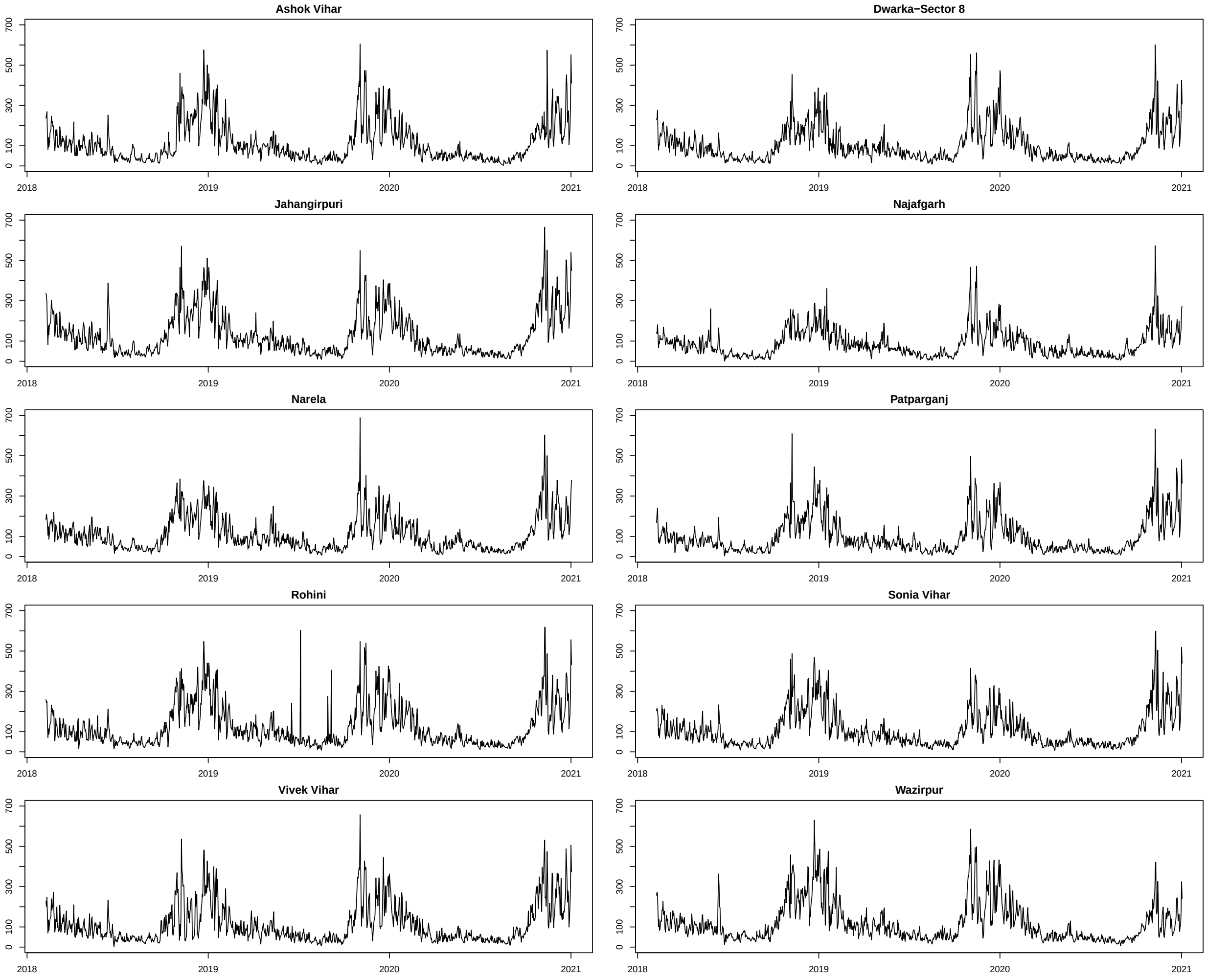


Figure 5: The above graph shows the time series plot of PM2.5 parameter for all 10 stations in the data.

Another Time-Series Plot

As stated previously, our main objective is to look for trends. This plot is same as the previous plot but added with educated guess of ticks where there is a trend. A regular jump in values of pollutant parameter PM2.5 between the month of October and February every year. The yellow coloured region is from Oct 15 to Feb 15 every year. The red coloured region is the part where PM2.5 value is greater than 300, which is termed as **hazardous**. This trend can be observed uniformly at all ten stations. But the red coloured region is predominant in yellow region i.e. between October and February.

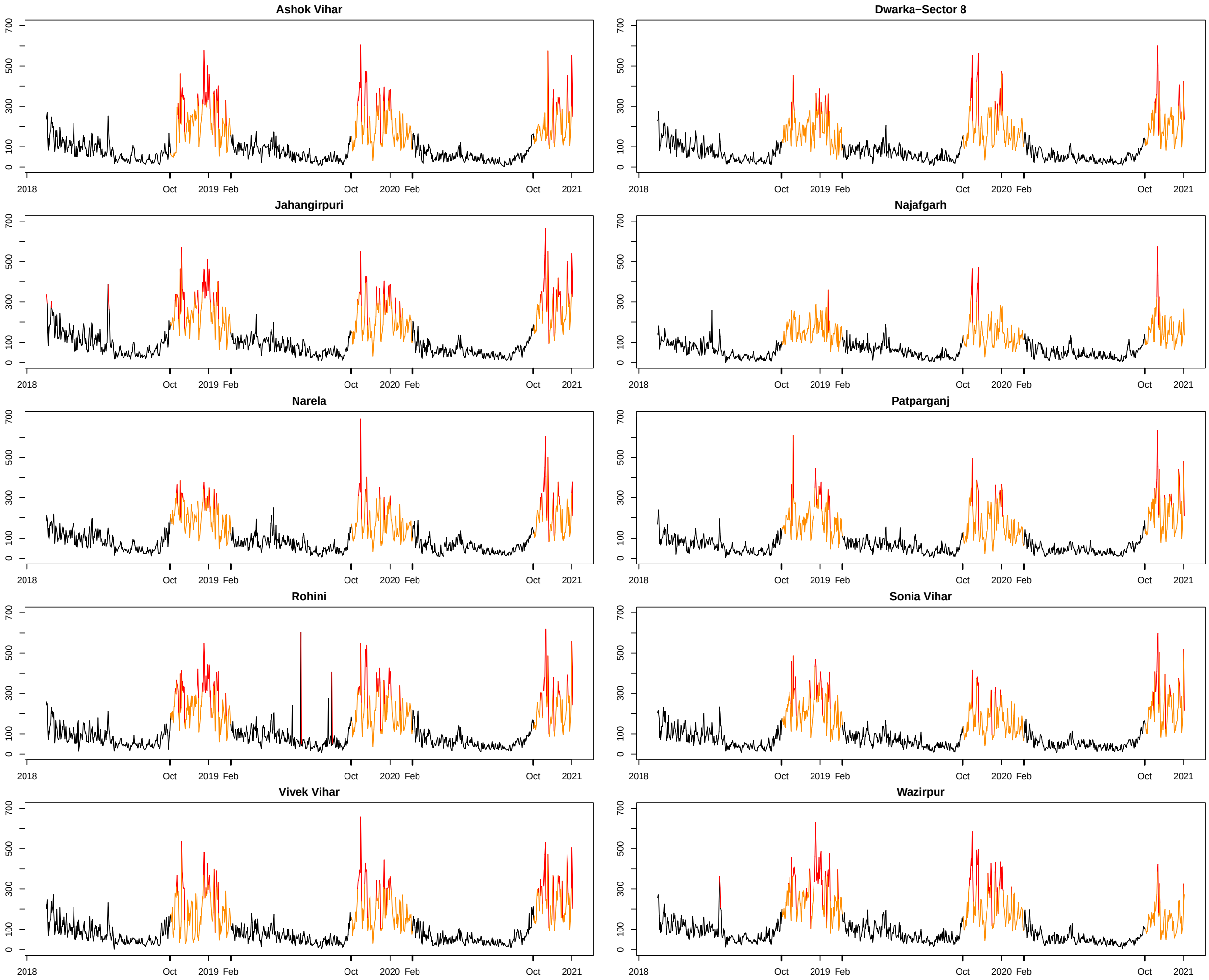


Figure 6: The above graph shows the time series plot of PM2.5 parameter for all 10 stations in the data highlighted with yellow the interval where there is a jump in PM2.5 values and in red the region where PM2.5 level is above 300.

Results

- From Figure 1, we see that PM2.5 resembles exponential and PM10 and Ozone resembles as left-skewed normal.
- From Figure 2, most of the parameters seem to have slightly correlation with each other, while we suspect PM2.5 and PM10 to be highly correlated as all the particles in PM2.5 are included in PM10.
- From Figure 3, we see that in fact our guess from Figure 2 is correct. PM2.5 and PM10 do actually have a strong correlation of 0.89 while other are slightly correlated
- From Figure 4, we see that both PM2.5 and PM10 have linear relation.
- From Figure 5, we see that there are subtle jumps in value of PM2.5 parameter across all the stations and repeats at equal interval.
- From Figure 6, we see that in fact the jumps are between the month of October and February. The level of PM2.5 crosses 300 mainly in these months.

Conclusion

Recalling our problem statement,

In this report we try to discover patterns and make inferences about the pollution level in stations in and around Delhi. We also look at the relation between the pollutant parameters and try to find which parameter are related to each other and see if there is a pattern across all ten stations. We specifically study PM2.5 pollutant, which is known to be a good indicator of air health.

We do conclude that:

In result of factors like, in no particular order, stubble burning in large amount in states surrounding Delhi in the month of October, end of raining season in North India, increase in North-West wind flowing leading to dust storms and other minor reasons like vehicular pollution and smog creating an air bubble, the pollution level starts to increase from October and stays till January or February. The increase in PM2.5 level validates the phenomenon.