Master in Data Science

Workshop I - Environmental Data Analytics

Winter Semester 2024

Air Quality Analysis in Delhi

Instructor:
Mohammad Mahdi Rajabi

Students:
Clara Duchossois
Patrick Silva

Contents

1 Data Collection and Data Cleaning **2**
1.1 Data Collection 2
1.2 Data Cleaning 2

2 Summary Statistics **4**

3 Time Series Plots and Histograms **6**
3.1 Time Series Plots 6
3.1.1 Daily Data 6
3.1.2 Monthly Aggregated Data 9
3.2 Histograms 12

4 Correlation Analyses **16**

5 Trend Analysis **20**

6 Final Discussions **22**

1 Data Collection and Data Cleaning

1.1 Data Collection

The dataset used comes from **Kaggle**. The dataset contains hourly and daily observation of the air quality in various cities of India.
This dataset has 5 files. 4 of these files list the observations done in the different cities or stations. 2 files are for the cities and 2 files are for the stations. There are 2 files because one file lists hourly observations, the other lists daily observations. This means we have: **station-hour, station-day, city-hour, city-day**. The last file shows what stations are in what city.
Not all files are needed. We opted for daily over hourly data. This removes 2 files out of the picture. Using the 5th file, the stations and cities were joined. It's not important whether you start with the file station-day or city-day, you get the same.
This gives us 1 file which contains all observations from 2015 to 2020 in several cities with several stations per city to work with.

1.2 Data Cleaning

Right now, there is more data than necessary. We want to limit the data to one city only and at least 2 stations in that city are needed. To reach that target, various filtering steps will be applied:

1. Filter the columns which are not needed out
2. Drop rows with null values
3. Filter the stations with less than 500 observations out
4. Filter every city with less than 3 stations out

After checking the number of observations per station, we stick to Delhi and its 4 stations with the highest number of rows. The stations are called: **DL002, DL019, DL028, DL031**. The air quality columns kept are: **PM2.5, PM10, O_3, NO_x, CO, SO_2** . For the plots following in the next sections, some additional date manipulations have been done. We will be looking at the observations from November 2017 until July 2020 (not included).

StationId	City	Date	Status	PM2.5	PM10	NOx	CO	SO2	O3	Month	Year	Day	Year-Month
str	str	date	str	f64	f64	f64	f64	f64	f64	i8	i32	i8	str
"DL002"	"Delhi"	2017-11-01	"Active"	156.0	310.99	49.29	2.17	22.72	20.72	11	2017	1	"2017-11"
"DL019"	"Delhi"	2017-11-01	"Active"	146.91	257.08	46.54	1.45	17.06	3.45	11	2017	1	"2017-11"
"DL028"	"Delhi"	2017-11-01	"Active"	182.43	182.43	106.56	1.53	29.66	370.85	11	2017	1	"2017-11"
"DL031"	"Delhi"	2017-11-01	"Active"	185.31	354.88	95.16	1.68	44.56	49.79	11	2017	1	"2017-11"
"DL002"	"Delhi"	2017-11-02	"Active"	181.25	329.0	128.12	1.82	20.86	32.85	11	2017	2	"2017-11"
…	…	…	…	…	…	…	…	…	…	…	…	…	…
"DL031"	"Delhi"	2020-06-29	"Active"	36.88	134.58	19.87	0.76	12.78	9.26	6	2020	29	"2020-6"
"DL002"	"Delhi"	2020-06-30	"Active"	44.5	75.04	16.59	1.11	12.06	80.42	6	2020	30	"2020-6"
"DL019"	"Delhi"	2020-06-30	"Active"	41.84	61.49	11.2	0.53	20.01	16.28	6	2020	30	"2020-6"
"DL028"	"Delhi"	2020-06-30	"Active"	39.86	91.6	20.21	0.82	15.97	22.47	6	2020	30	"2020-6"
"DL031"	"Delhi"	2020-06-30	"Active"	21.58	70.21	11.27	0.8	13.39	9.73	6	2020	30	"2020-6"

Figure 1: The dataset after cleaning

2 Summary Statistics

The statistics chosen for the air quality metrics are: **Mean, Median, Standard Deviation, Variance, Min, Max**.

StationId	CO_mean	CO_median	CO_std	CO_var	CO_min	CO_max
str	f64	f64	f64	f64	f64	f64
DL002	2.297219	2.08	0.946339	0.895557	0.04	6.16
DL019	1.293923	1.005	1.036416	1.074159	0.09	17.76
DL028	1.563828	1.43	0.847917	0.718964	0.19	5.69
DL031	1.496404	1.41	0.829889	0.688716	0.01	4.97

Figure 2: CO stats

CO is a gas that comes from things like cars or stoves. The CO stats are very good. The only high value is the maximum at DL019. This value seems to be an exception.

StationId	NOx_mean	NOx_median	NOx_std	NOx_var	NOx_min	NOx_max
str	f64	f64	f64	f64	f64	f64
DL002	114.179077	105.29	80.231417	6437.080303	0.0	368.9
DL019	52.375672	39.235	44.479584	1978.433424	0.0	274.7
DL028	65.026495	46.405	55.640561	3095.872032	0.0	337.42
DL031	74.771787	53.625	63.209881	3995.489019	0.0	297.09

Figure 3: NO_x stats

NO_x is a group of gases that come from car engines and factories. The NO_x stats are very bad. This may come from the high amount of car usage in Delhi. The minimum values may come from the experiments conducted in Delhi to reduce car usage.

StationId	O3_mean	O3_median	O3_std	O3_var	O3_min	O3_max
str	f64	f64	f64	f64	f64	f64
DL002	34.446639	27.68	21.573524	465.416917	1.2	141.11
DL019	18.921002	16.66	10.671434	113.879505	1.59	82.52
DL028	221.254763	196.065	143.470313	20583.730727	5.63	963.0
DL031	55.677708	59.845	16.150207	260.829176	6.83	94.93

Figure 4: O_3 stats

O_3 is measured near the ground. If the value is too high, breathing gets difficult. The values are quite low except for the one at station DL028. That station's values are really high. The maximum values are also good except for the one at DL028 which is absurdly high.

StationId	PM10_mean	PM10_median	PM10_std	PM10_var	PM10_min	PM10_max
str	f64	f64	f64	f64	f64	f64
DL002	303.445112	277.72	162.985754	26564.356135	26.36	955.6
DL019	196.910533	177.6	109.184834	11921.32796	28.52	826.8
DL028	126.848581	97.36	98.790394	9759.54199	6.95	789.88
DL031	227.578684	205.695	134.618023	18122.012139	10.42	888.83

Figure 5: PM10 stats

PM10 are dust or smoke particles that can get into your lungs when you breathe. Delhi has a serious problem with particles in the air. The minimum value indicates that there are days in which the values are good but the mean and median values are way too high.

StationId	PM2.5_mean	PM2.5_median	PM2.5_std	PM2.5_var	PM2.5_min	PM2.5_max
str	f64	f64	f64	f64	f64	f64
DL002	145.855929	111.12	110.401205	12188.426033	9.54	624.07
DL019	107.997996	83.12	86.343814	7455.254259	9.62	742.0
DL028	122.012774	89.05	98.782835	9758.04849	6.95	789.88
DL031	109.6775	78.035	86.419532	7468.335518	6.94	675.35

Figure 6: PM2.5 stats

PM2.5 is similar to PM10 but the particles are smaller. The PM2.5 statistics shows the same as the previous table. Delhi has a real issue with particles in its air.

StationId	SO2_mean	SO2_median	SO2_std	SO2_var	SO2_min	SO2_max
str	f64	f64	f64	f64	f64	f64
DL002	14.572651	13.26	7.928821	62.866201	0.94	55.12
DL019	11.052047	8.93	7.739613	59.901615	1.78	48.98
DL028	20.144667	17.675	12.721023	161.824425	2.4	129.84
DL031	16.264134	12.905	11.101624	123.246062	0.68	63.98

Figure 7: SO_2 stats

SO_2 comes from burning coal or oil. This statistics show good and healthy values. The maximum values indicates there exist days with extreme values but the mean stays pretty close to the median which lets the reader think that this extreme values are quite rare.

3 Time Series Plots and Histograms

3.1 Time Series Plots

3.1.1 Daily Data

In order to generate time series plots for each studied parameter at the 4 different stations we ended up choosing, we first looked at all the data we had at hand, which was collected on a daily basis. That means that we had very detailed and complete plots, but it also impacted the readability and accuracy of said plots. Indeed, the time series plots are rather hard to read, especially because we plotted the data for 4 stations at a time, and the overwhelming amount of data meant that we were taking into account potentially meaningless, or random, variations in our data, rather than overall patterns and trends. We want to point out an overall trend across all parameters and all stations, occurring around May 2020, where all the measures seem to significantly decrease. This pattern would coincide with the COVID-19 lockdown, and a remarkable stop to car usage and industrial activities.

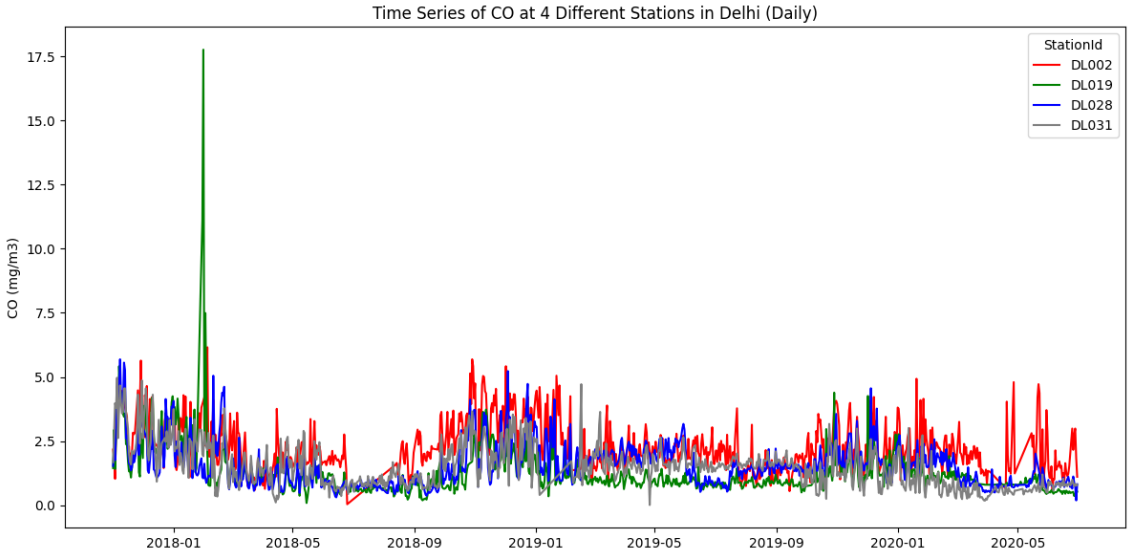

Figure 8: Time Series of CO in Delhi (Daily)

Besides one outlier value measured at Station DL019 in 2018, the CO levels throughout Delhi are overall following a similar trend.

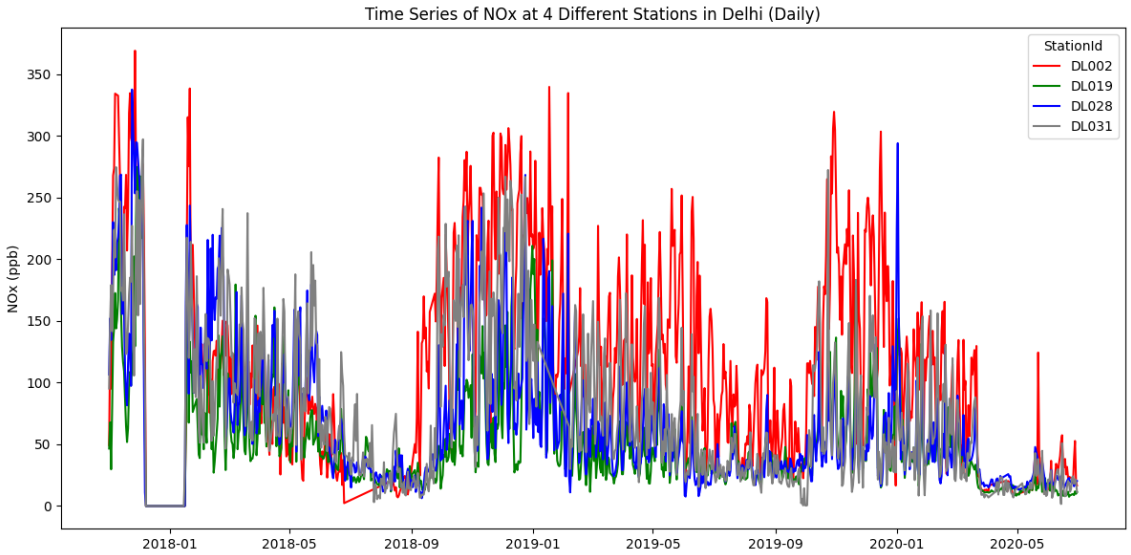

Figure 9: Time Series of NO_x in Delhi (Daily)

Regarding NO_x , measured levels are globally higher at Station DL002, but they do seem to follow similar patterns across all stations.

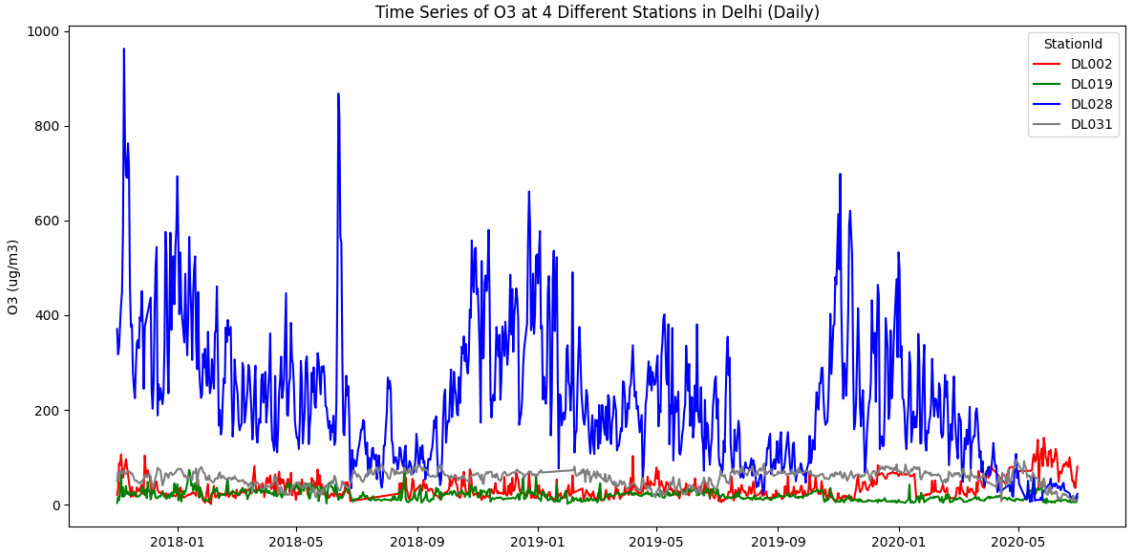

Figure 10: Time Series of O_3 in Delhi (Daily)

O_3 is the one measure that has one clear outlier among our chosen stations. Values measured at Station DL028 are remarkably higher than at the 3 other stations. We can hypothesize the presence of industrial facilities around this specific station, which would explain this discrepancy.

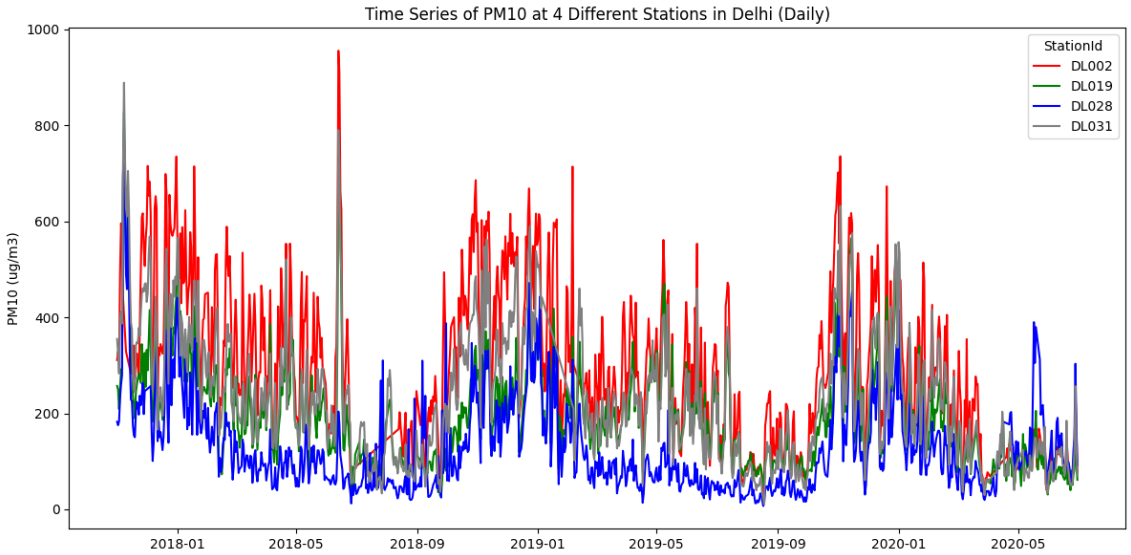

Figure 11: Time Series of PM10 in Delhi (Daily)

For PM10 levels, the values collected at Station DL002 are slightly higher than the rest, while values collected at Station DL0028 seem to be overall lower than at the all the other stations. Every year, the overall PM10 values seem to be reaching a yearly minimum around August/September.

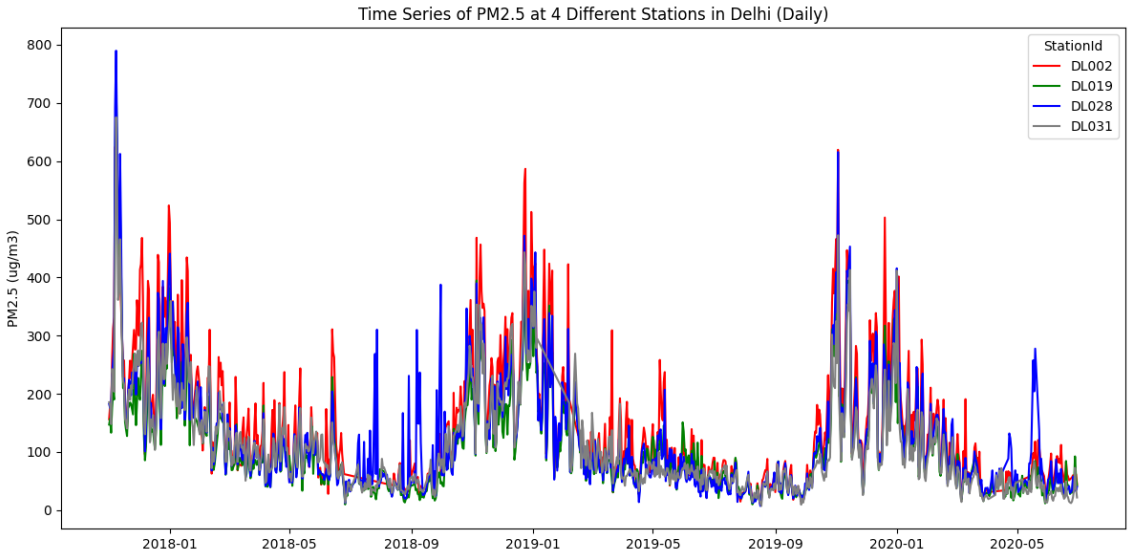

Figure 12: Time Series of PM2.5 in Delhi (Daily)

Collected PM2.5 levels are extremely similar across all 4 stations, and seem to follow a slightly decreasing trend over the years, with a seasonal pattern peaking every year around January.

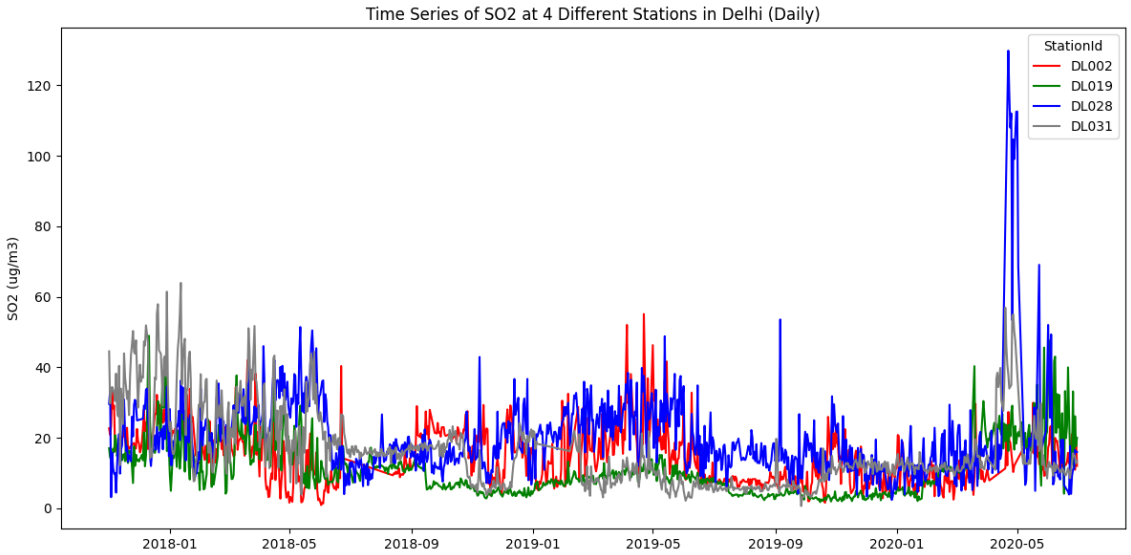

Figure 13: Time Series of SO_2 in Delhi (Daily)

The collected values of SO_2 notably increase around May 2020, especially at Station DL028. Otherwise, the SO_2 levels seem to be slightly lower at Station DL019.

3.1.2 Monthly Aggregated Data

To counter the readability and accuracy issues of our previous time series plots, we decided to aggregate the data on a monthly basis. Thus, we took the average of the measures within each month for each metric and each station. The obtained plots are indeed easier to interpret, and show much clearer trends and patterns in our data.

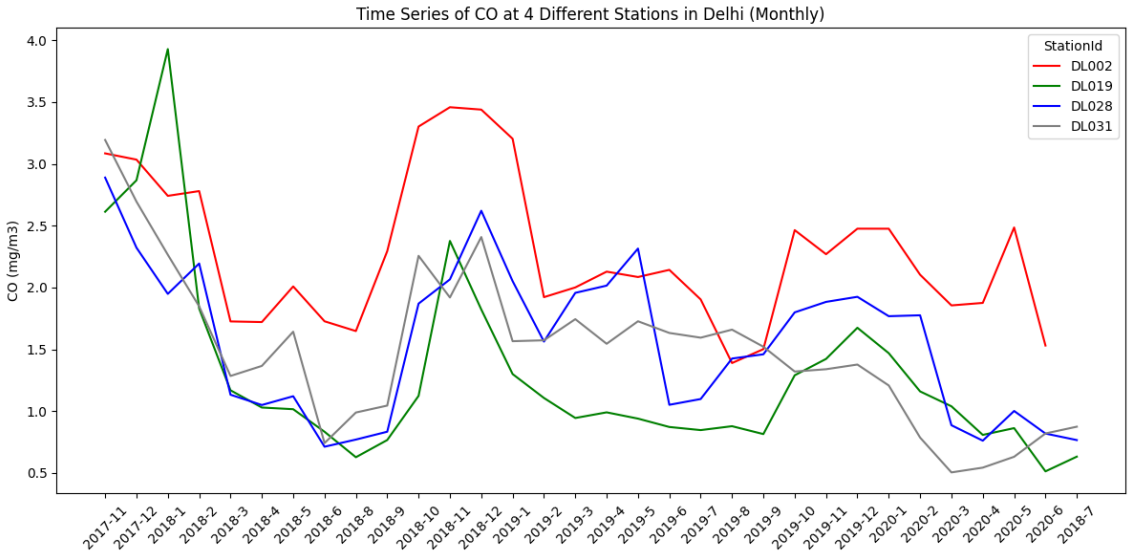

Figure 14: Time Series of CO in Delhi (Monthly)

By aggregating the data, we got rid of the previously present outlier value in 2018, which

allows to distinguish the time series of the different stations better, by improving the y-axis scale, since their overall values were really close. We can now see that the values collected are slightly higher at Station DL002.

Regarding NO_x , O_3 , PM10, and PM2.5, the observed trends do not differ from the observations made with the daily data.

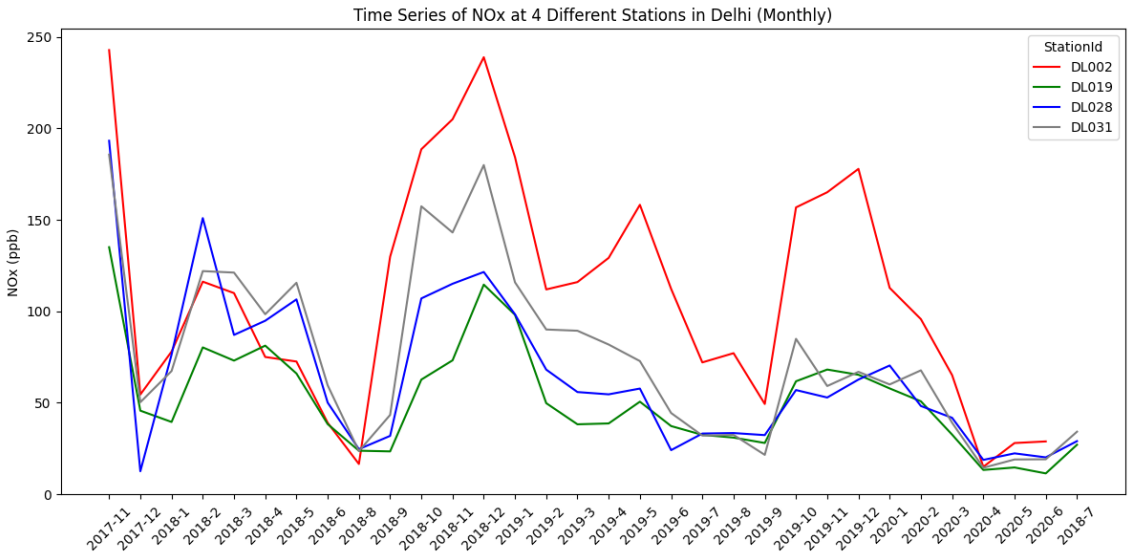

Figure 15: Time Series of NO_x in Delhi (Monthly)

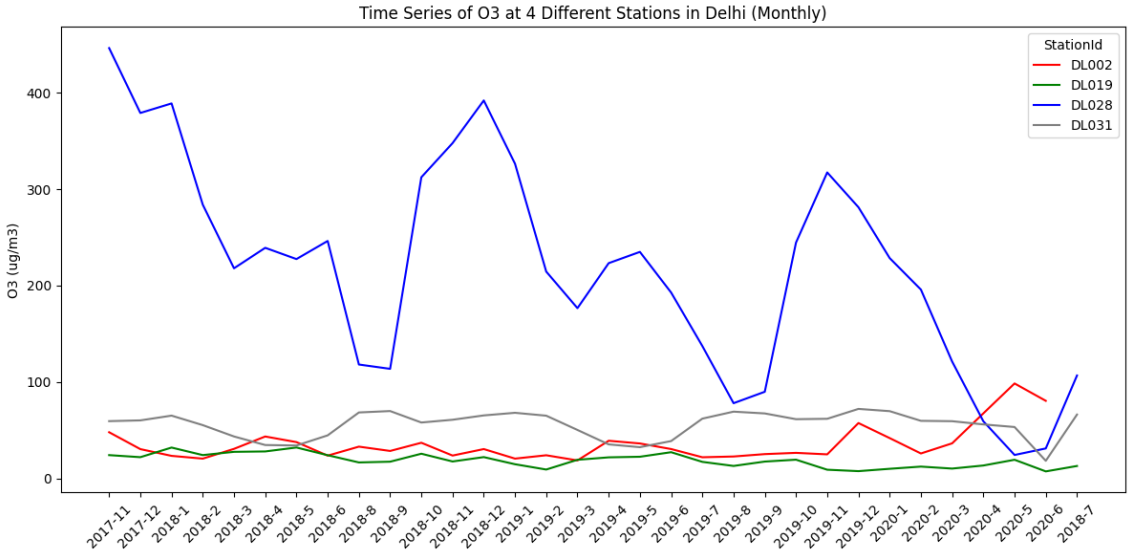

Figure 16: Time Series of O_3 in Delhi (Monthly)

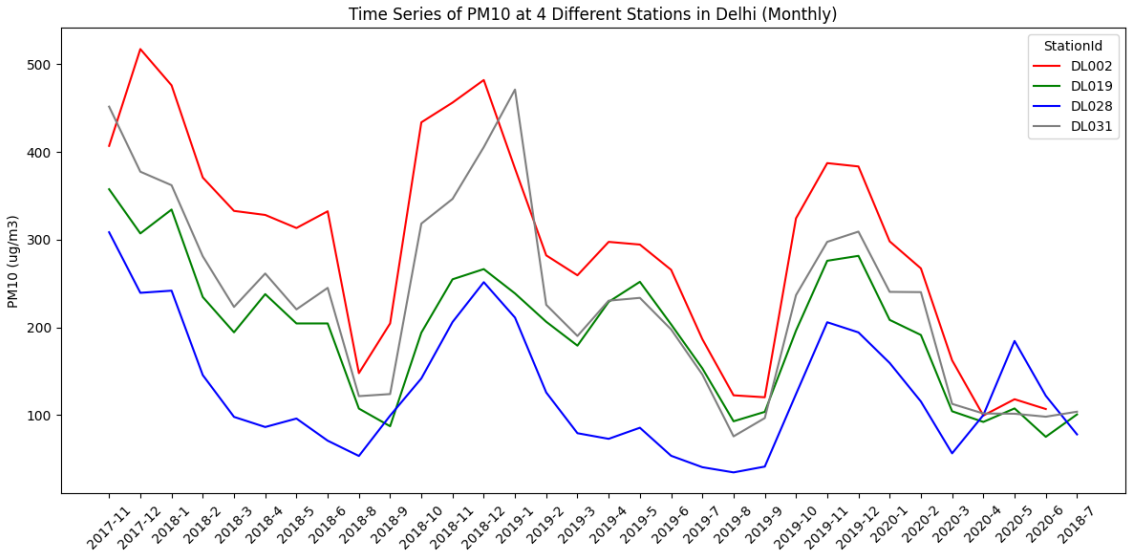

Figure 17: Time Series of PM10 in Delhi (Monthly)

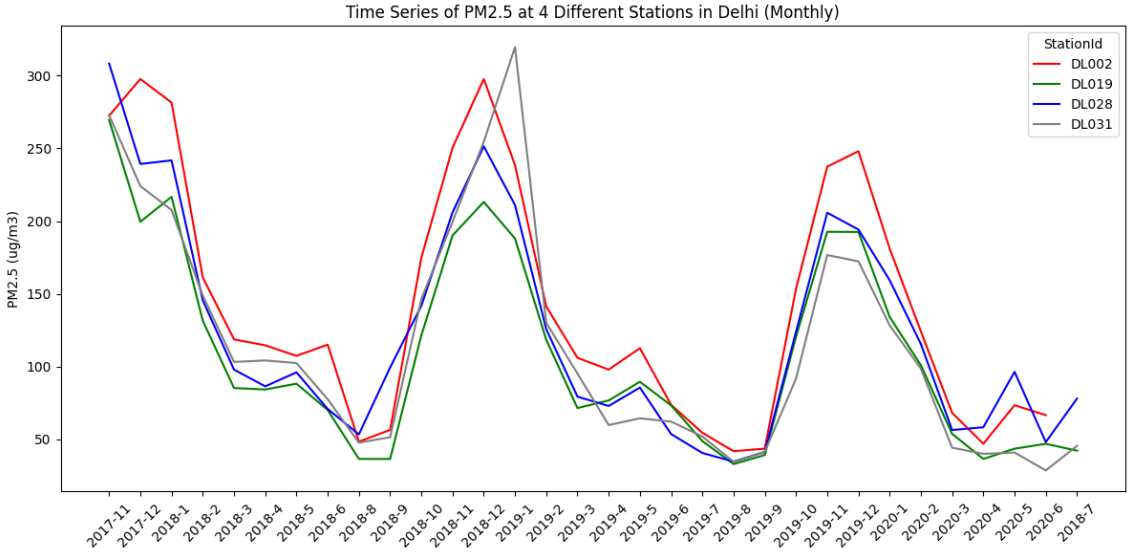

Figure 18: Time Series of PM2.5 in Delhi (Monthly)

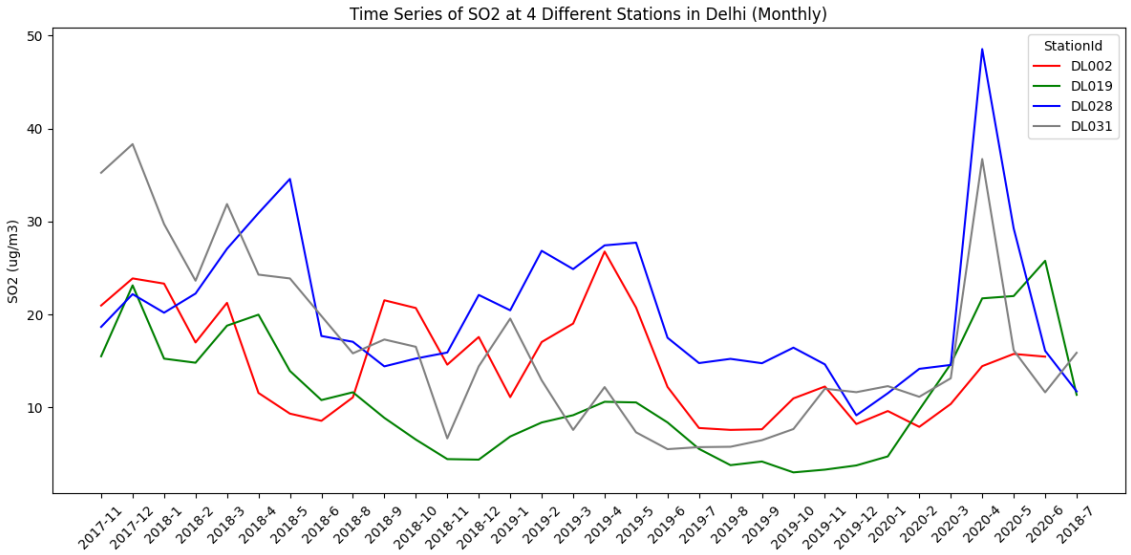

Figure 19: Time Series of SO_2 in Delhi (Monthly)

In the same manner as for the CO time series, by aggregating our SO_2 data to a monthly format, we smoothed out the outlier which previously appeared for Station DL028 around May 2020. We can still observe a peak in SO_2 levels around that time for all stations. Other than that, the levels do not seem to follow a specific trend, and seem to remain rather constant over time.

3.2 Histograms

With the following histograms, the same information as for the time series can be read but in a clearer way as the stations are now well separated. There is a clear pattern in the different air quality observations. Around October-November each year, the activity reaches a peak. Probably due to temperature and habitants trying to act against it. For the rest of the year it just continues going down. This repeats every year but with a decreasing peak. This lets us believe that the air quality is slowly becoming better over time.

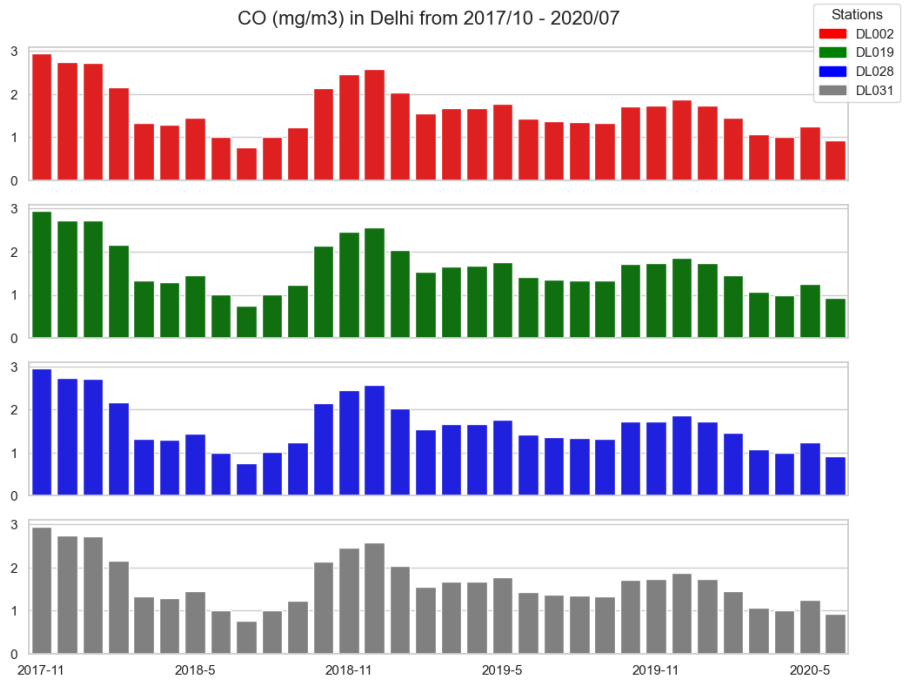

Figure 20: CO in Delhi (Histogram)

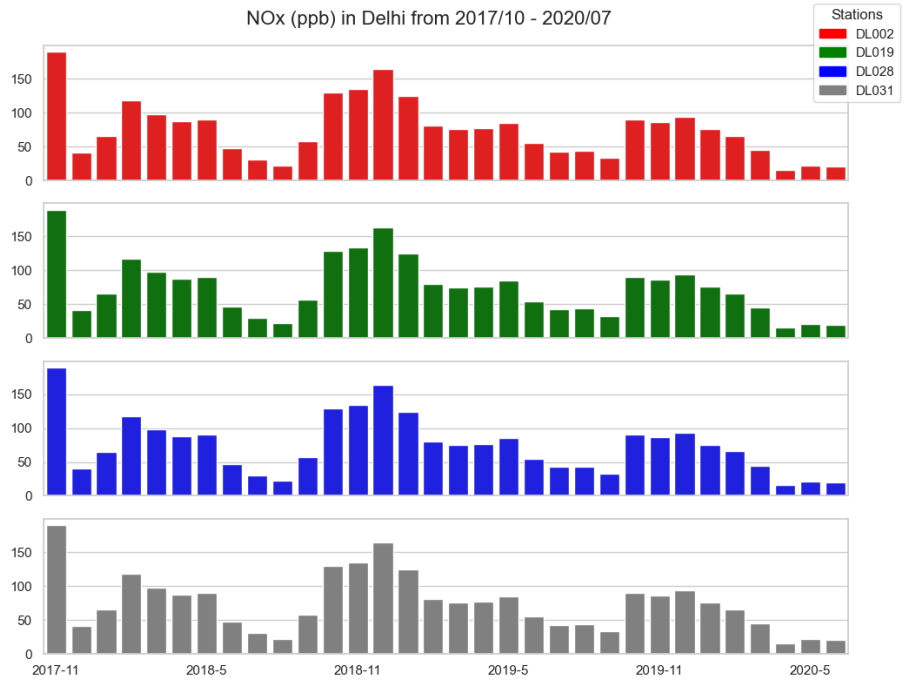

Figure 21: NO_x in Delhi (Histogram)

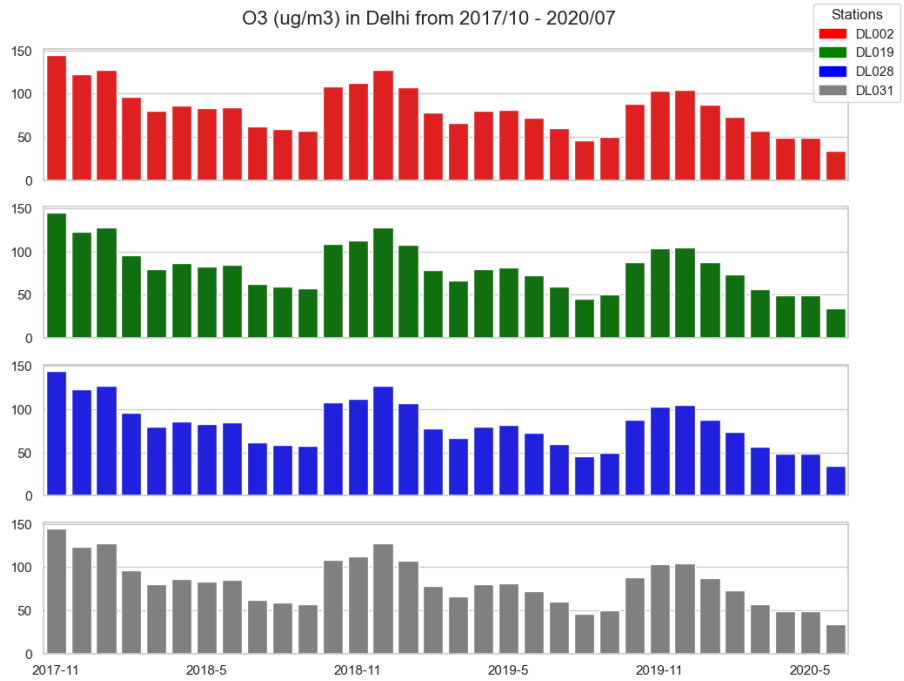

Figure 22: O_3 in Delhi (Histogram)

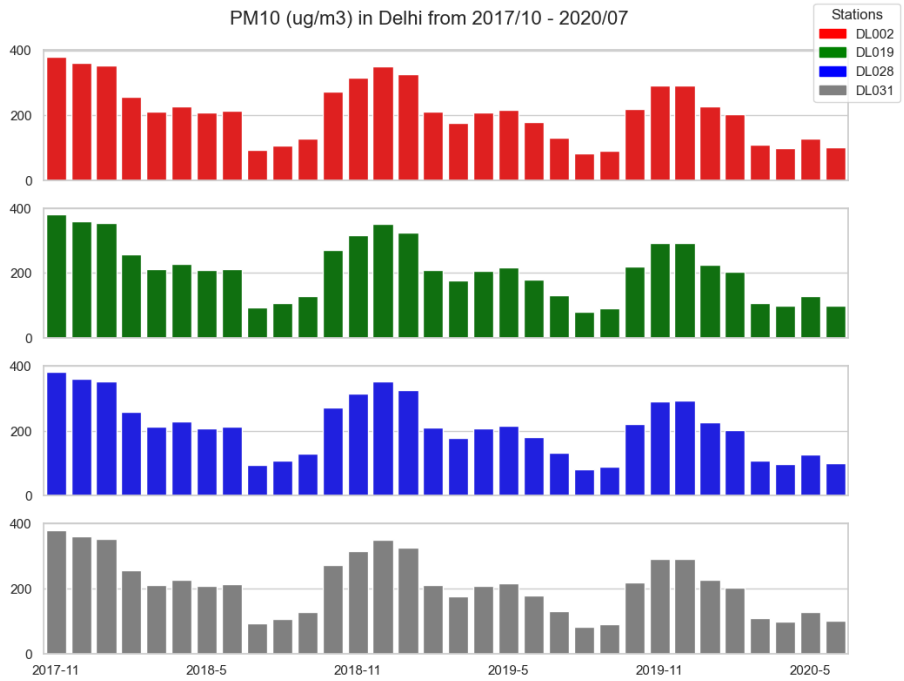

Figure 23: PM10 in Delhi (Histogram)

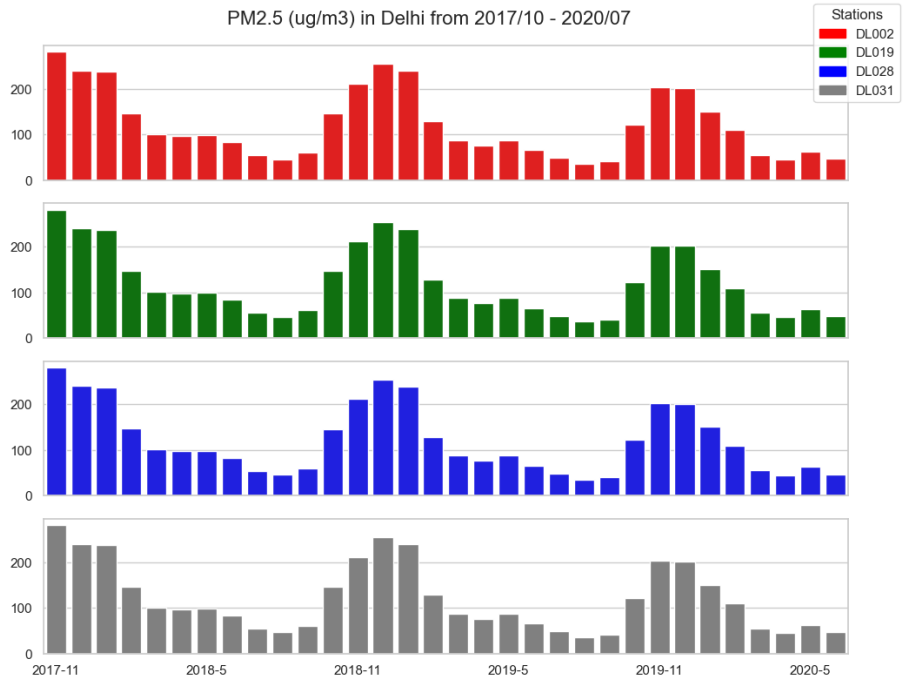

Figure 24: PM2.5 in Delhi (Histogram)

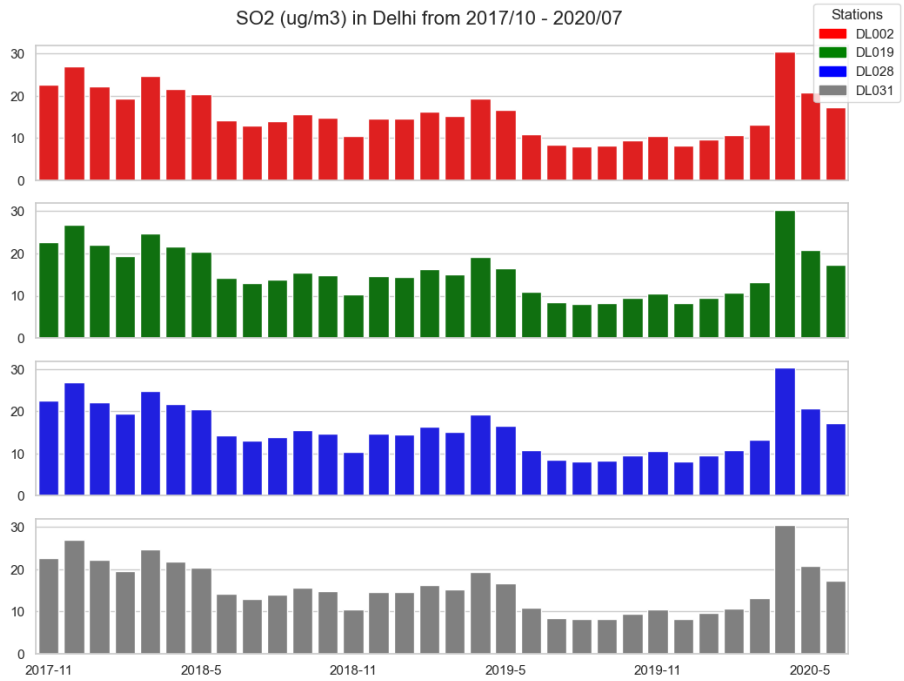

Figure 25: SO_2 in Delhi (Histogram)

The SO_2 histogram is the only histogram who does not respect the aforementioned pattern. However, there is nothing to worry. The values shown by this histogram show a non-dangerous amount of SO_2 in breathing air.

4 Correlation Analyses

Now, let's start the correlation analyse between the air quality metrics and between the stations.

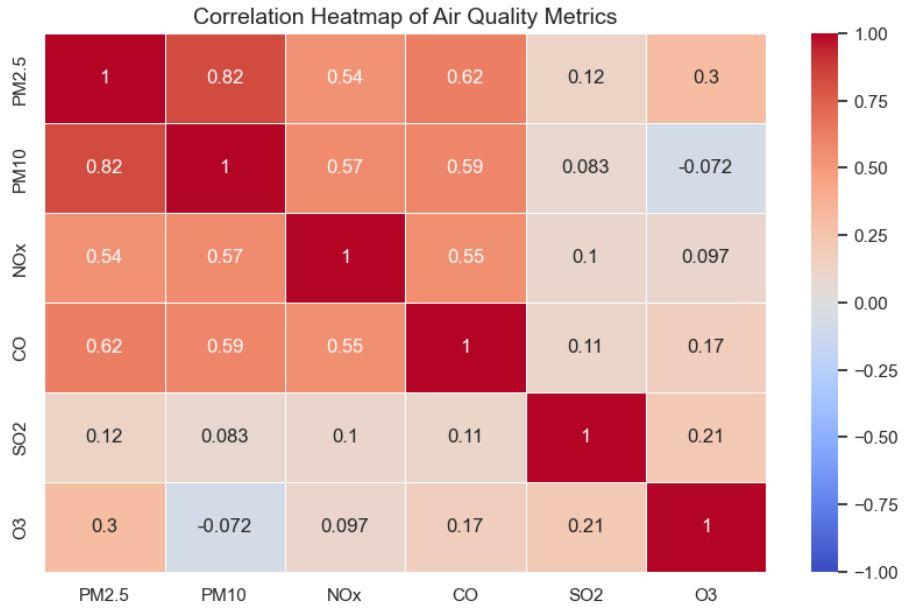

Figure 26: Correlation between Air Quality Metrics

So far, there is nothing unexpected. PM2.5 and PM10 were expected to correlate as both measure particles in the air. NO_x and CO check the gases produced by engines. Not the same gases, but the gases come from similar sources. O_3 and SO2 are a bit on their own and it shows on the heatmap above.

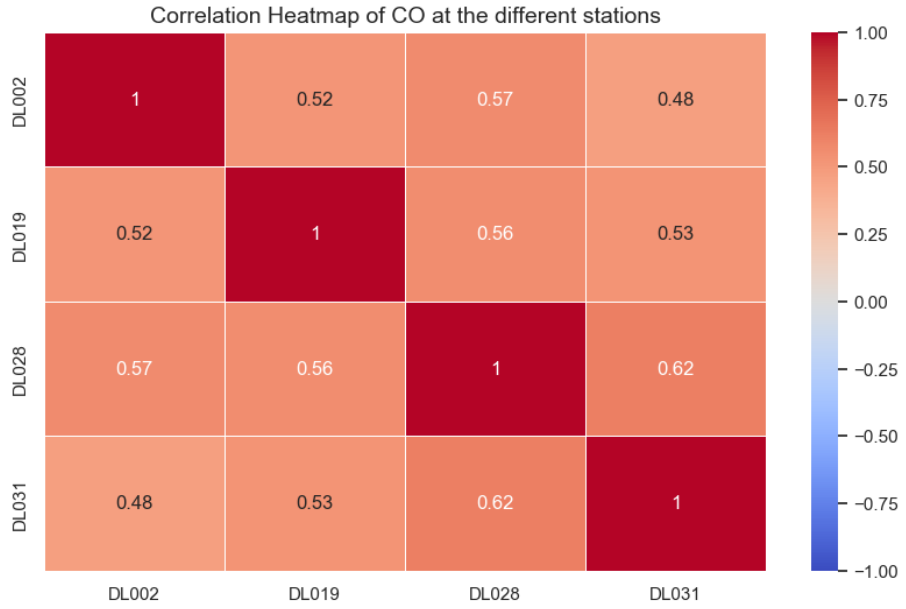

Figure 27: Correlation of CO between different stations

This heatmap does not share much with us. Each station seems to have different observations.

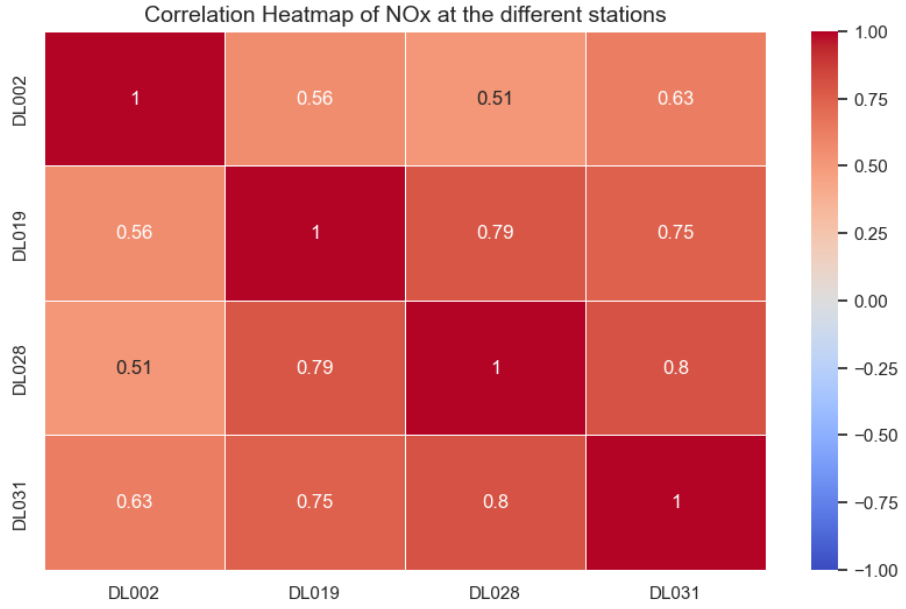

Figure 28: Correlation of NOx between different stations

DL019, DL028, and DL031 seem have similar car and factory activity around them. DL002 is a bit on its own.

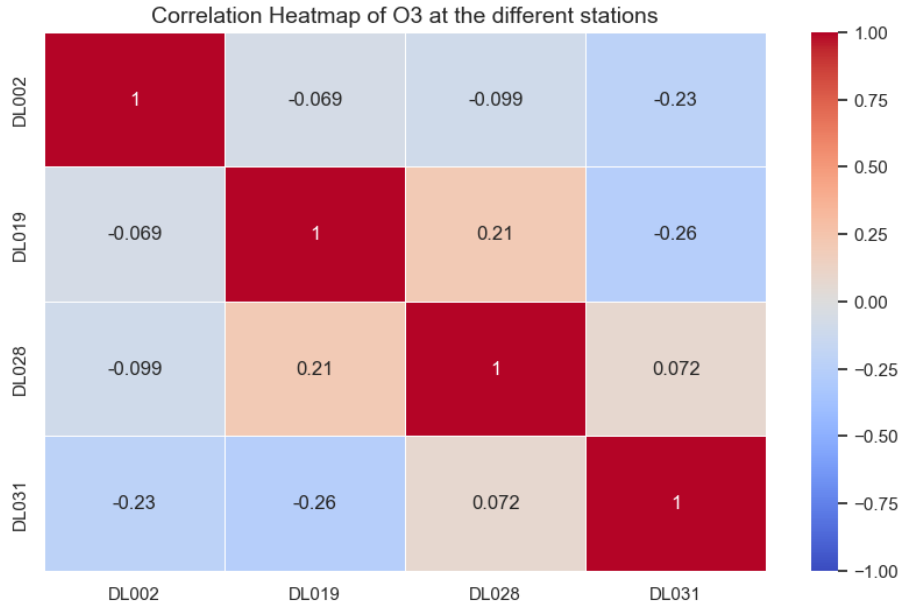

Figure 29: Correlation of O_3 between different stations

The O_3 heatmap is interesting. There is no correlation between any station. We assume that the ground levels around the different stations have different altitudes.

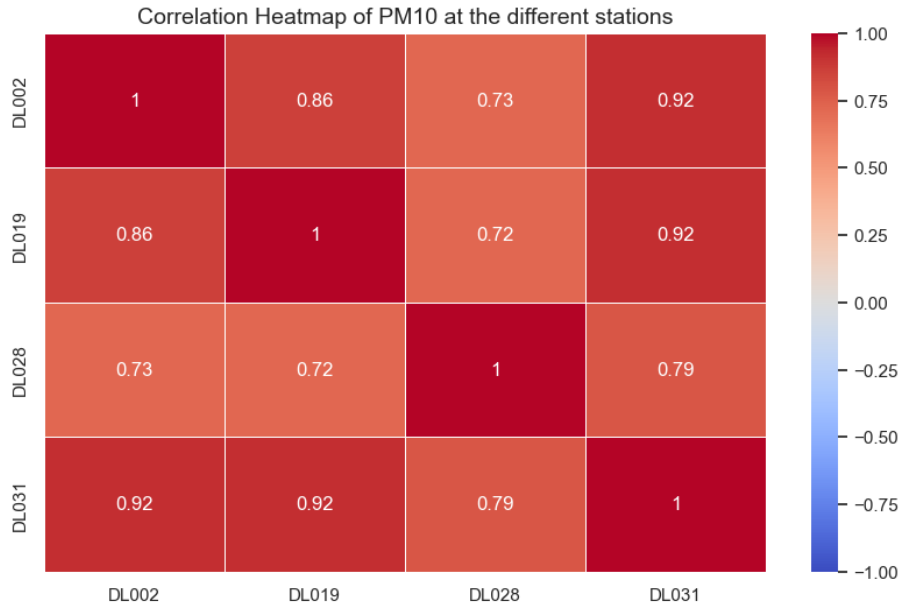

Figure 30: Correlation of PM10 between different stations

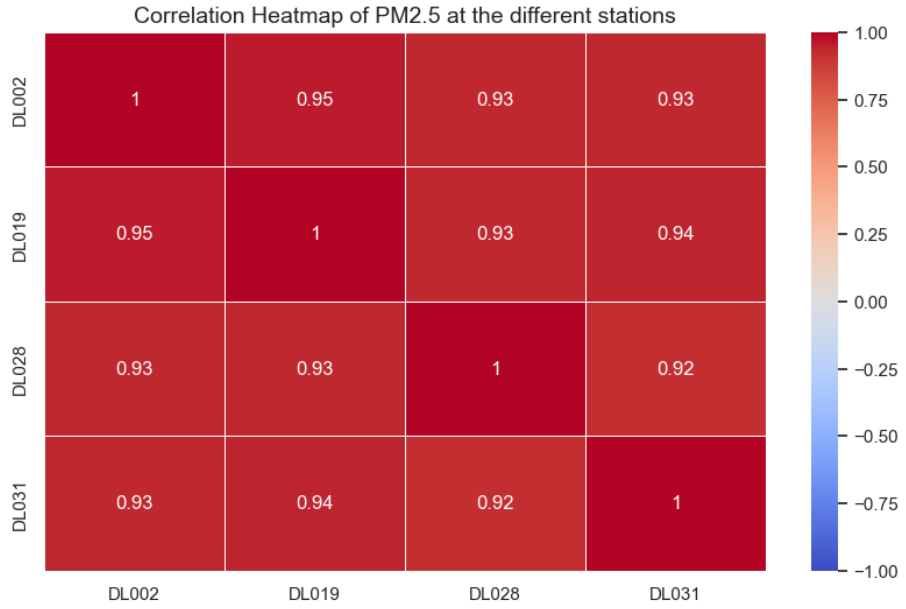

Figure 31: Correlation of PM2.5 between different stations

The PM2.5 and PM10 correlations between stations are expected. Delhi has a problem with particles in the air. In the summary table of these air quality metrics, the values were similar between stations and quite high.

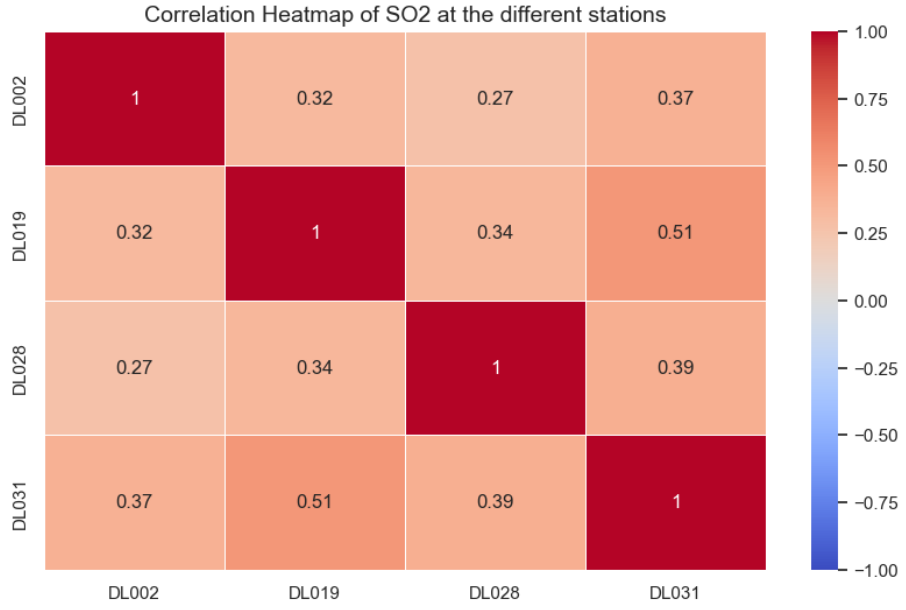

Figure 32: Correlation of SO_2 between different stations

SO2 comes from burning coal. The amount of coal burning is probably not the same everywhere and this heatmap lets us believe the same.

5 Trend Analysis

For our trend analysis, we chose to study the evolution of the **PM2.5** levels at **Station DL028** over time (from November 2017 until July 2020). In order to verify whether the parameter shows a significant increasing or decreasing trend over time, we decided build a linear regression model of the form $Y_i = b_0 + b_1 \times t_i + \varepsilon_i$, where the Y_i represent the PM2.5 levels after i days. We then tested (using **Student's t-test**) the following null hypothesis: $H_0 : b_1 = 0$ against $H_1 : b_1 \neq 0$, that is, the slope of the linear regression fitted to the data is constant, versus the slope is strictly monotonic. In other words, the null hypothesis H_0 states that the PM2.5 levels at Station DL028 have not increased or decreased over time. We chose to test our null hypothesis at significance level $\alpha = 0.01$.
We obtain the following scatter plot and linear regression for the given parameter:

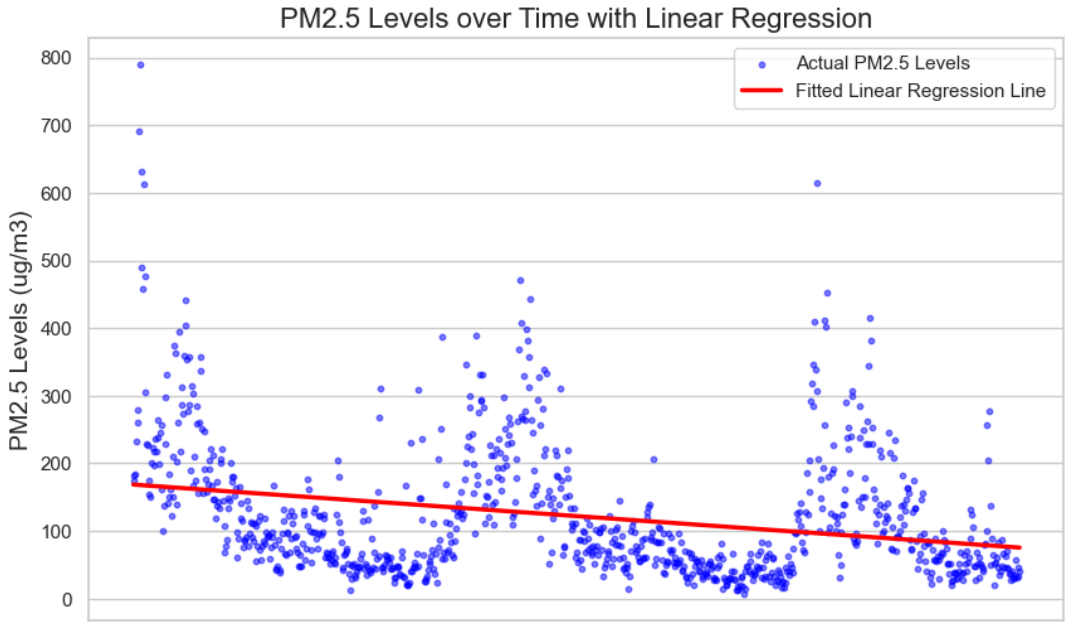

Figure 33: PM2.5 levels at Station DL028 over time, and fitted Linear Regression

The estimated coefficients of our linear model are given by:

- b_0 (intercept) $= 168.59$,
- b_1 (slope) $= -0.10$

In order to apply Student's t-test, we then compute the **t-statistic**, as follows:

$$t_{b_1} = \frac{b_1}{S.E.(b_1)},$$

where $S.E.(b_1)$ is the standard error of the estimated coefficient b_1. It is given by:

$$S.E.(b_1) = \sqrt{\frac{1}{n-2} \times \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2}},$$

where:

- Y_i is the PM2.5 level actually measured on the i-th day,
- $\hat{Y_i}$ is the value of Y_i predicted by our linear regression model,
- t_i is the number of days passed,
- $\bar{t}$ is the mean value of the t_is.

As a result, we get a t-statistic equal to -8.63 and with a corresponding p-value of 0.0 (after rounding). Since the p-value obtained is lower than the value of α we had set for our significance level, we reject the null hypothesis, and conclude that the PM2.5 levels at Station DL028 have significantly decreased over time.
The reason why we chose to study this specific parameter, is because it has the highest correlation coefficient between other air quality metrics, as seen in Figure 26. This means that it should be (one of) the best indicators of overall air quality at hand.

6 Final Discussions

Ultimately, we have found that Delhi seems to have a distinct problem with its Particulate Matter (PM10 and PM2.5) levels, which are overall extremely high. Most studies recommend a concentration below $15\mu g/m^3$ as a healthy threshold for PM2.5, while the observed mean values in Delhi vary around $120\mu g/m^3$. For PM10, the healthy threshold is usually set at and $210\mu g/m^3$. Both of these values can be considered as unhealthy, and expose the public to potential health issues, mostly respiratory. This is especially true for chronically diseased and sensitive people, as well as children and elder.
We also want to point out that, regardless, Delhi exhibited healthy amounts of SO_2 , with daily concentrations usually under $50\mu g/m^3$.
Nevertheless, we come to the conclusion that the overall air quality in Delhi improved over time. Indeed, we have shown that the PM2.5 levels at Station DL028 have significantly decreased over the years, and given that this metric is highly correlated to most other air quality metrics taken into account (regardless of the station), we can impute that this improving trend is a global one.
We also feel the need to mention that the collected data we worked with stops during the COVID-19 lockdown, and that there is a possibility that the different air metrics concentrations have started increasing again after the return to normal activities. We, however, can only extrapolate on this matter.