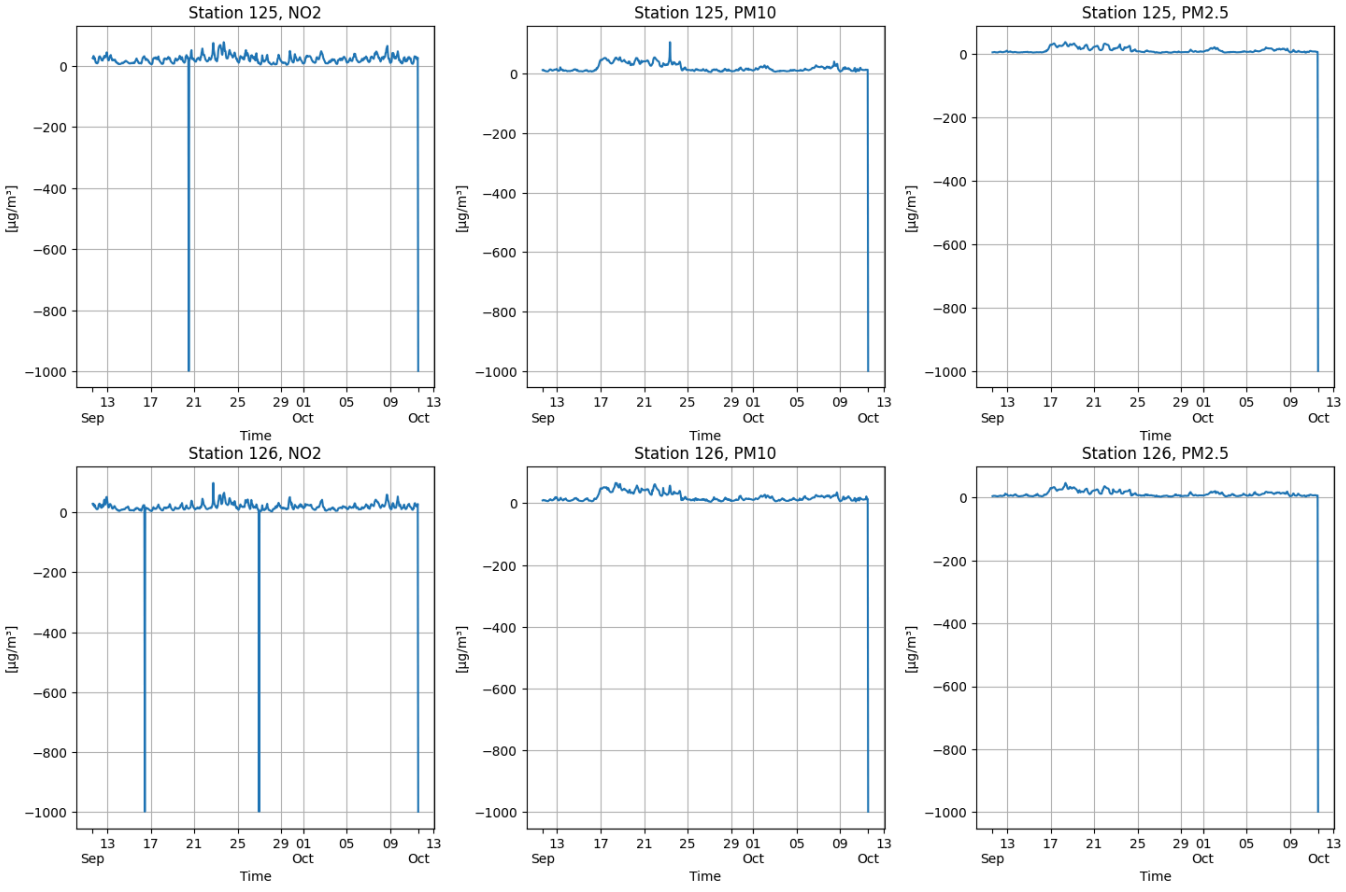
Kasra Eskandarizanjani 023012264f
Zahra Hashemi 0231846534

Introduction

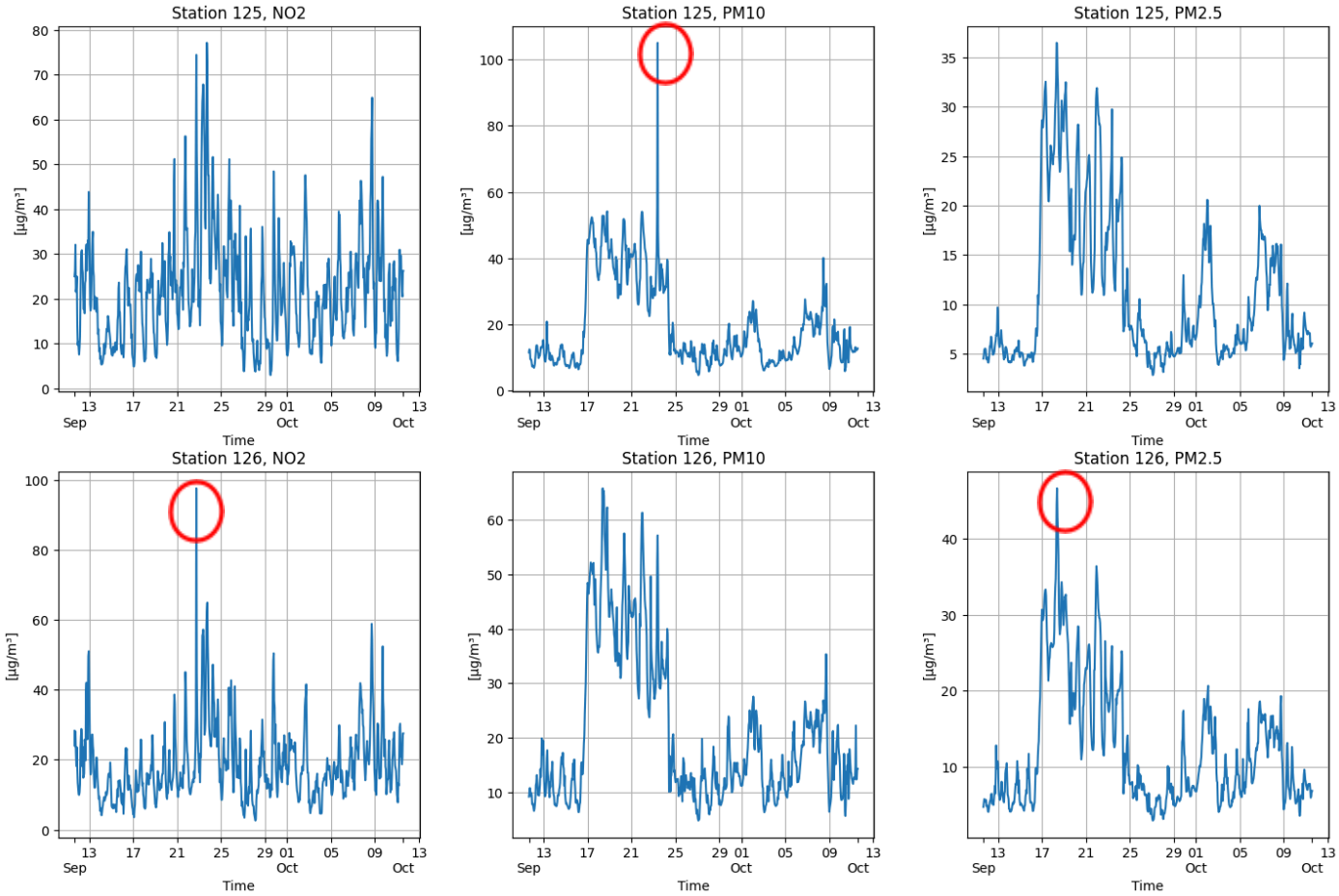
This project is an Air Quality Analysis in Bruxelles city (The data was collected from the European Environment Agency). This is an hourly dataset that has been measured for the last month.
Here we are going to explore three air quality parameters (NO2, PM10, PM2.5) captured from the monitoring stations number 125 and 126.
In the following sections, we are going to report the analysis we've done to explore these parameters measurements through the two stations.

Methodology

After exploring the data, for **cleaning data**, some outliers have been shown up (with the value of -999).



To clean them, we took the “linear interpolation” approach. First, the outliers have been replaced with NaN, then with pandas interpolation, they were filled by calculating the straight line between available data points around the missing value.



After cleaning the data, you can see that the outliers are fixed. However, if some data points, which are shown in red circles, didn’t have those higher values (compared to the other station), we probably would have had more similar summary statistics between both stations.

Results

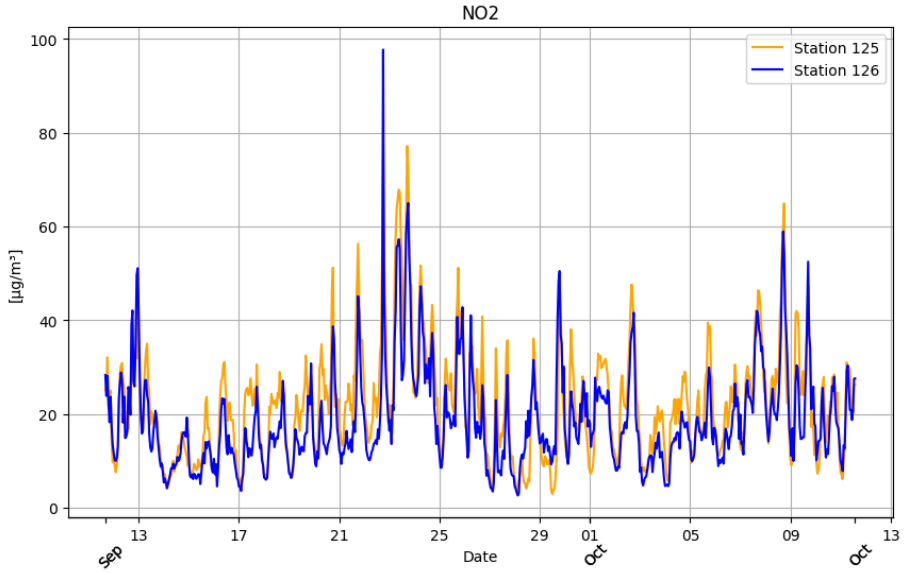
For **summary statistics**, you can find the below tables summarizing the stats, regarding each quality parameter at both stations.

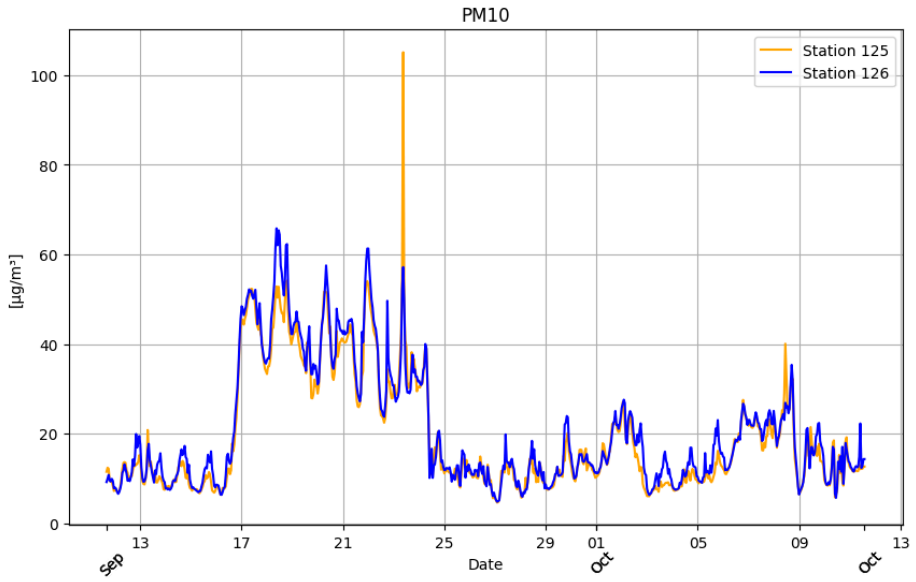
NO2	Station 125	Station 126
Mean	21.99	18.74
Median	21.02	16.04
Sd	11.97	10.70
Variance	143.29	114.48
Min	2.98	2.60
Max	77.12	97.69

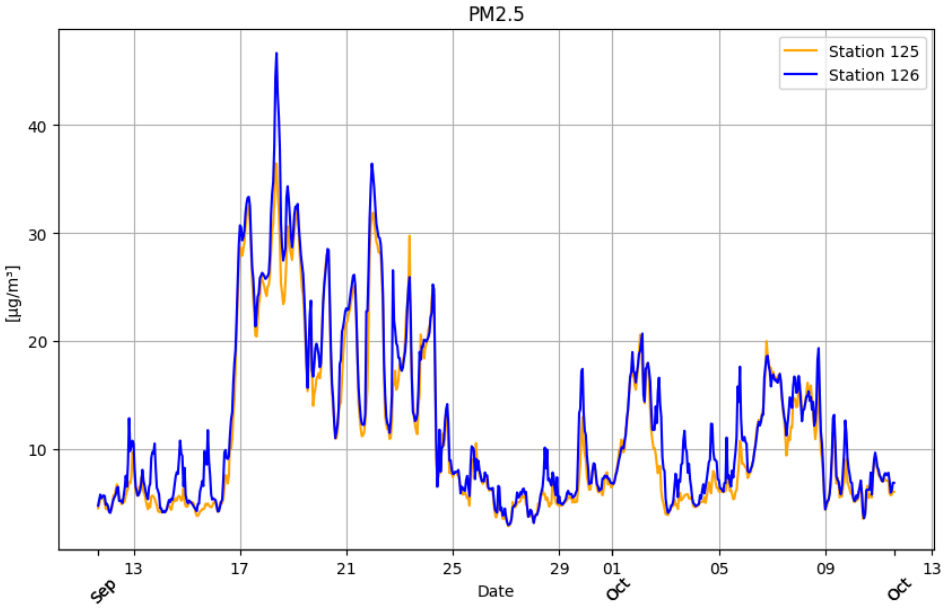
PM2.5	Station 125	Station 126
Mean	11.20	12.39
Median	7.49	9.03
Sd	7.76	8.23
Variance	60.16	67.71
Min	2.82	2.94
Max	36.47	46.68

PM10	Station 125	Station 126
Mean	19.46	20.80
Median	13.18	15.05
Sd	13.31	13.79
Variance	177.17	190.17
Min	4.60	4.80
Max	105.10	65.79

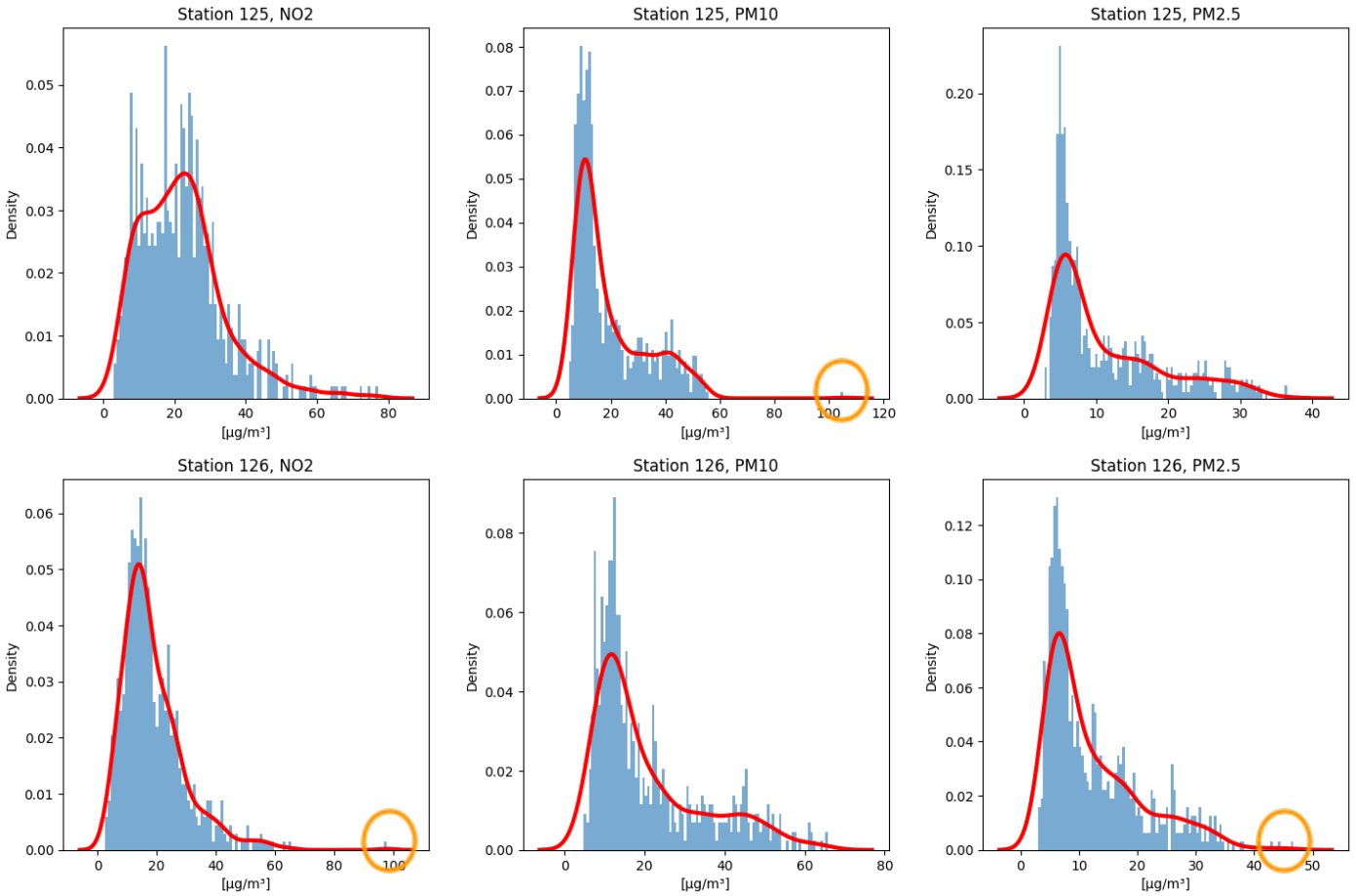
For **time series plots**, since our dataset is hourly for one month, the x-label will show the data points for each day.



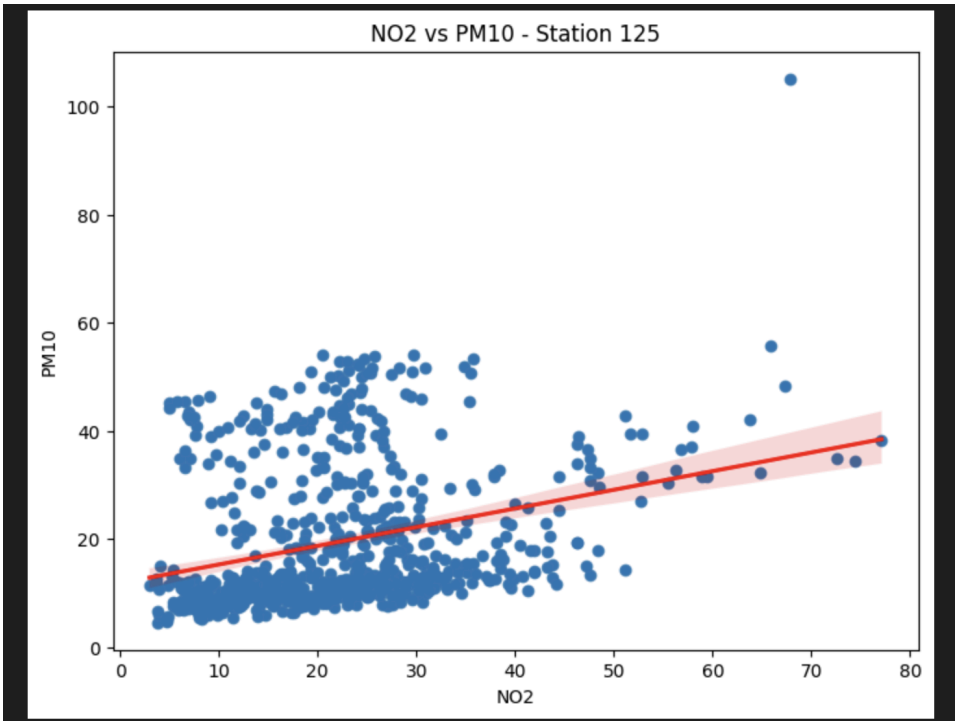




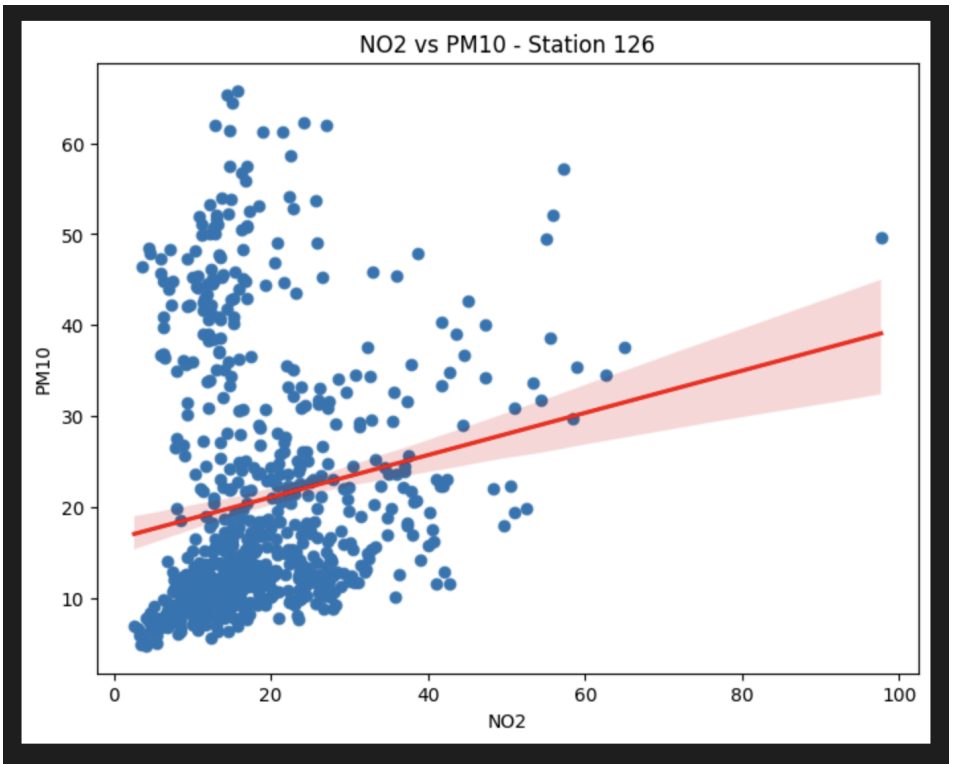
In the **histogram plots**, the distribution of the values have been normalized and the KDE also has been plotted.
The density curves for each parameter look pretty much the same, except the fact that if we didn't have those yellow values, we would have had the same range of values for each parameter at both stations. Besides, in the histogram plots for NO2, we can see different densities and different curve for the range (0, 40)



Correlation Analysis between different air quality parameters within the same station

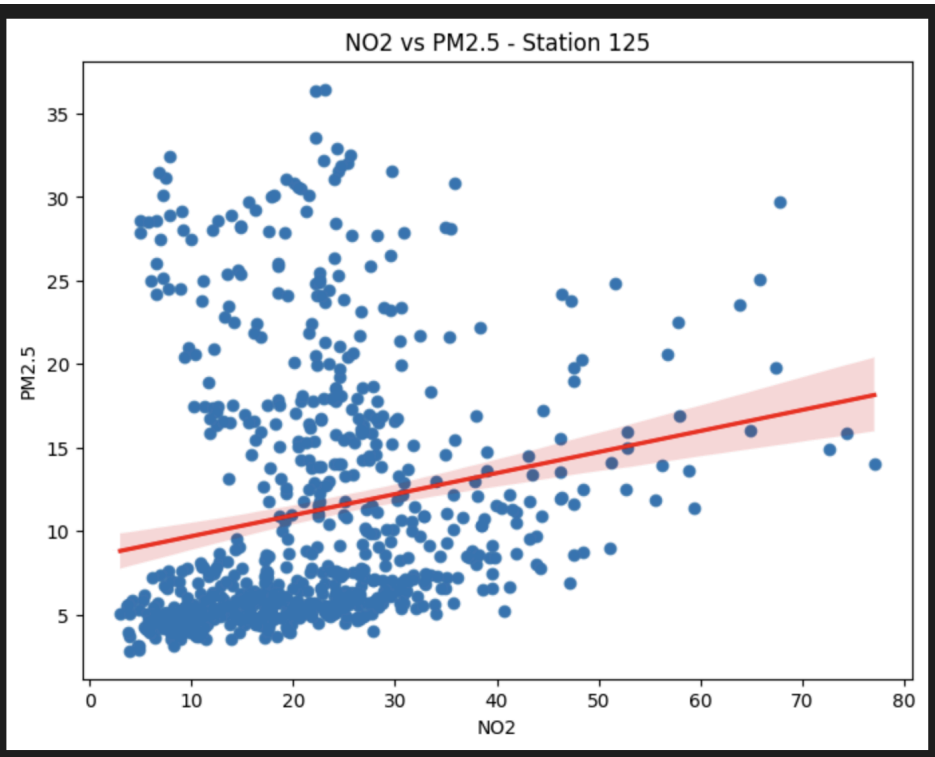



```
Correlation coefficient between NO2 and PM10: 0.31034201182513527
P-value: 1.619212655000087e-17
The correlation is statistically significant.
```

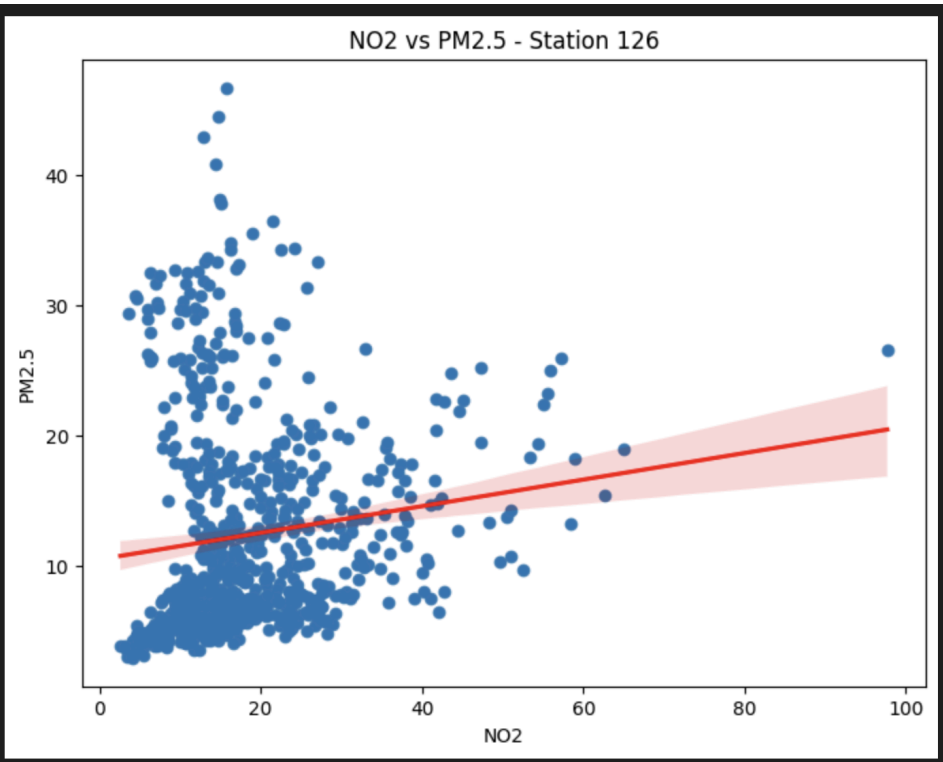



```
Correlation coefficient between NO2 and PM10: 0.17982918489874744
P-value: 1.2148565525461215e-06
The correlation is statistically significant.
```

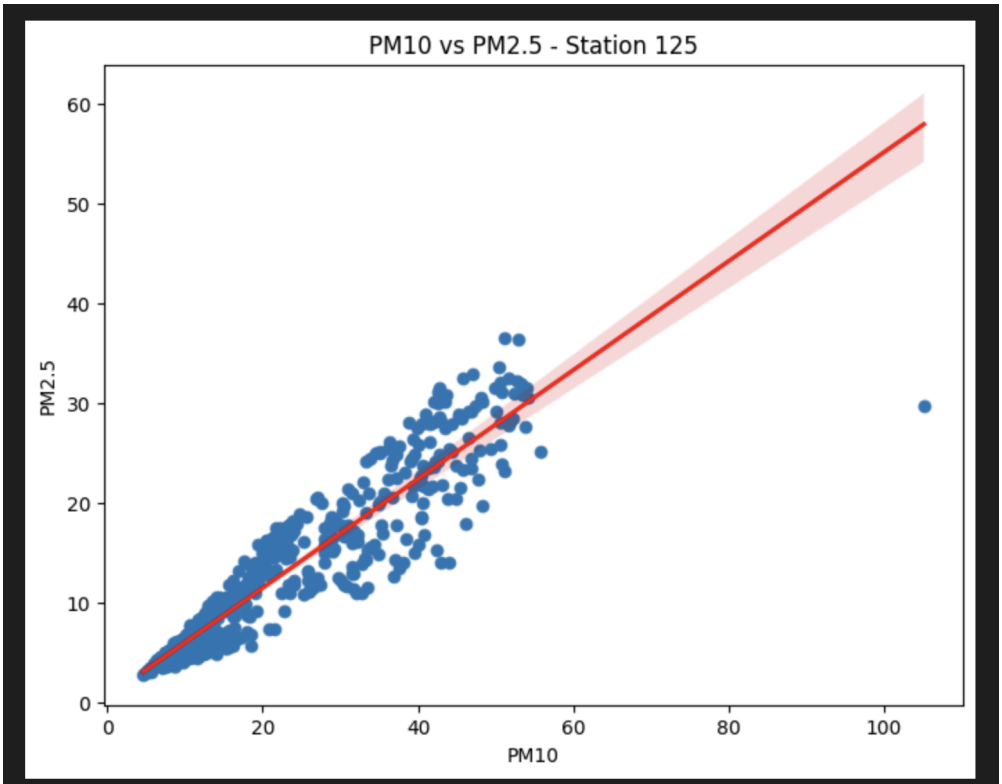
The statistical approach we took for calculating the significance of correlation is measuring the p-value, where the significance level defined is 0.05. If the p-value is less than that, the correlation is statistically significant.



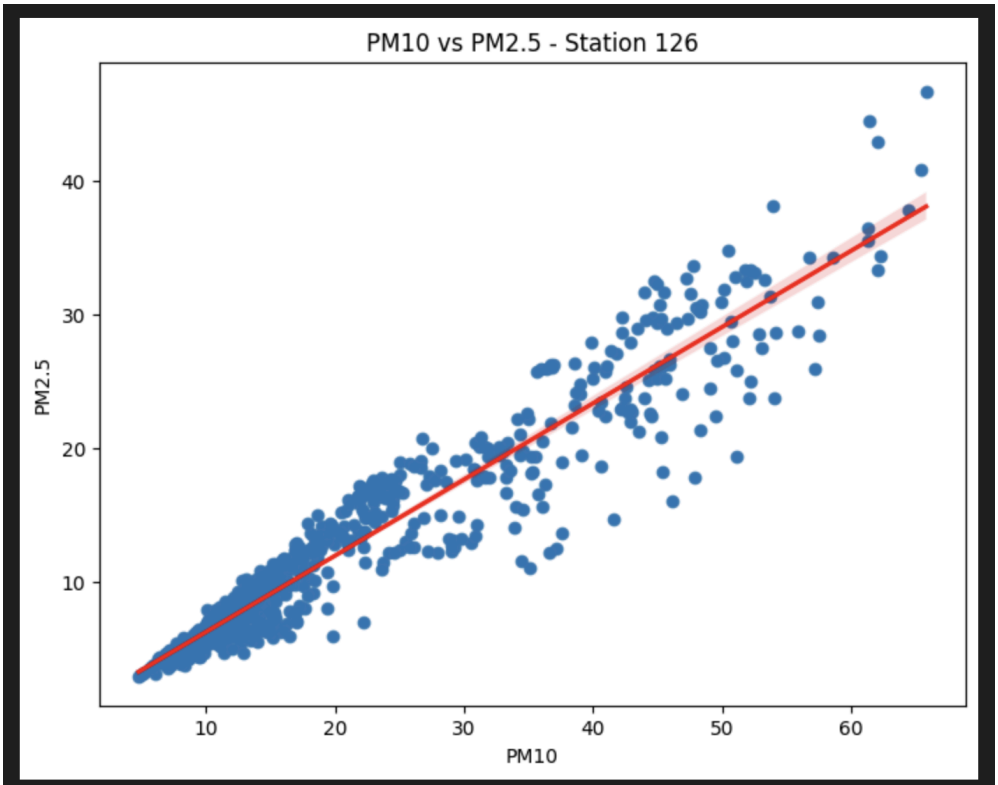

```
Correlation coefficient between NO2 and PM2.5: 0.19448969037697023
P-value: 1.46999399267122e-07
The correlation is statistically significant.
```




```
Correlation coefficient between NO2 and PM2.5: 0.13266761250986503
P-value: 0.00036114449581694207
The correlation is statistically significant.
```



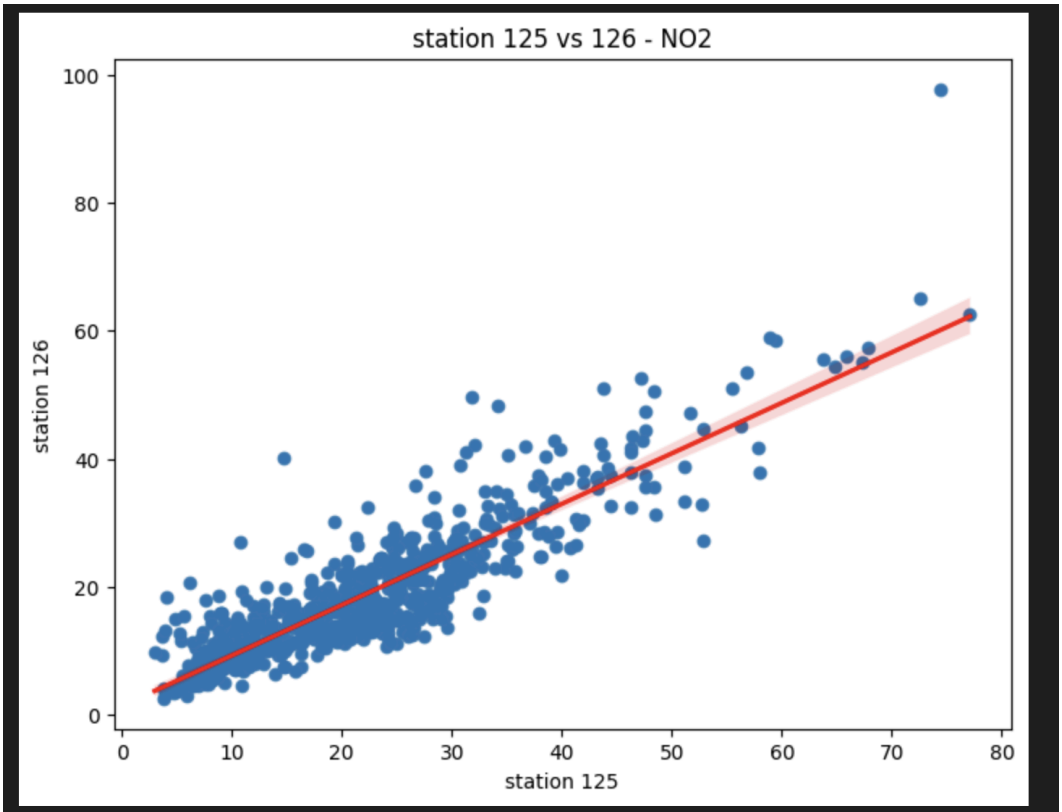

```
Correlation coefficient between PM10 and PM2.5: 0.9369157609067829
P-value: 0.0
The correlation is statistically significant.
```



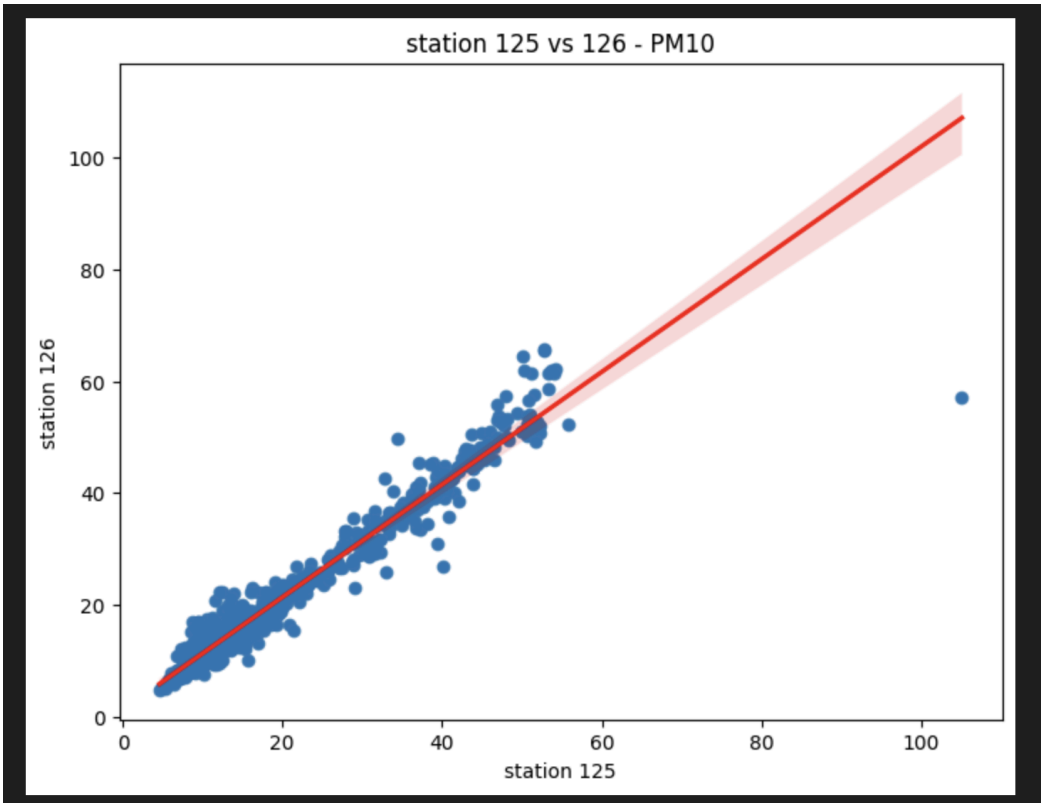

```
Correlation coefficient between PM10 and PM2.5: 0.9564067306855296
P-value: 0.0
The correlation is statistically significant.
```

Here, between NO2 and PM2.5, we can see a more significant correlation compared to other two correlations.

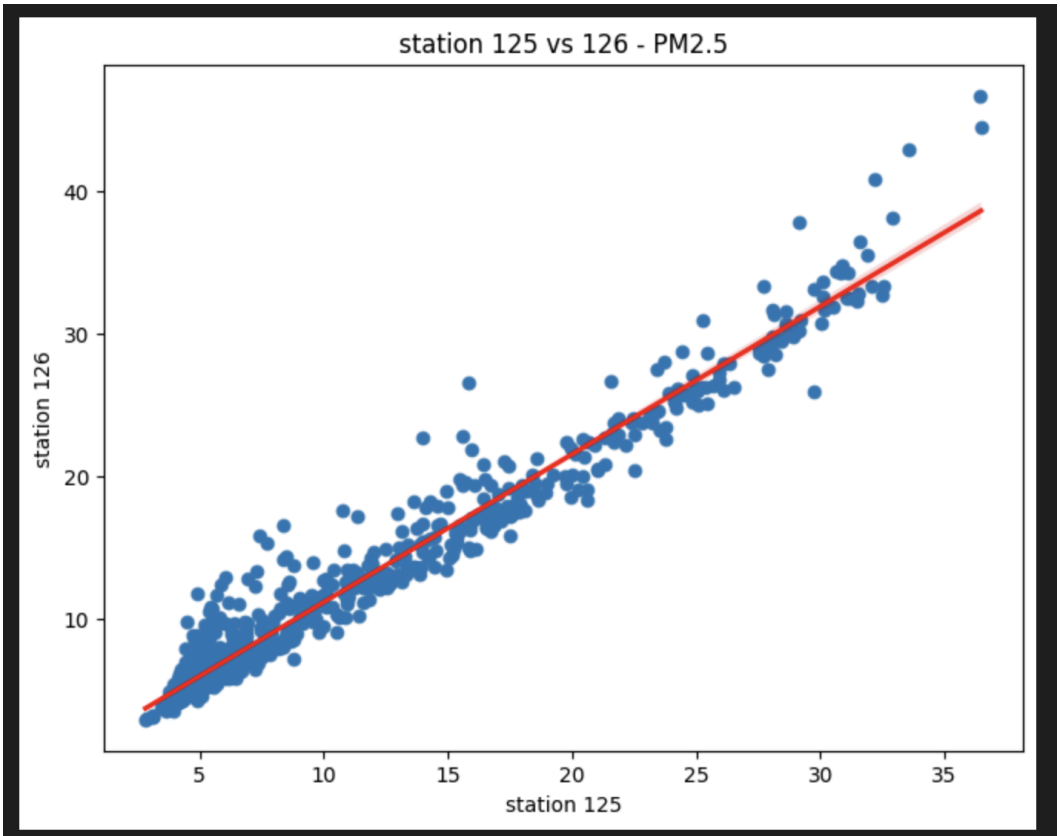
Correlation analysis between the two stations for each parameter




```
Correlation coefficient between None and None: 0.8821305108364673
P-value: 1.2236481638434159e-236
The correlation is statistically significant.
```

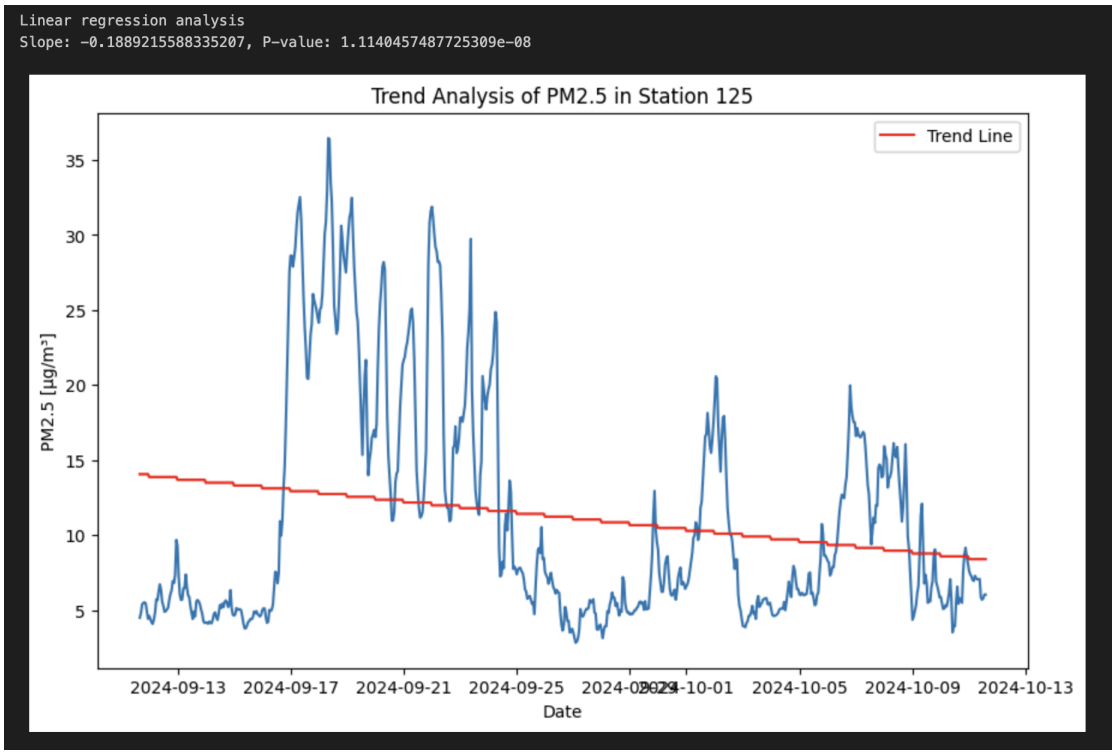



```
Correlation coefficient between None and None: 0.9728503238391462
P-value: 0.0
The correlation is statistically significant.
```

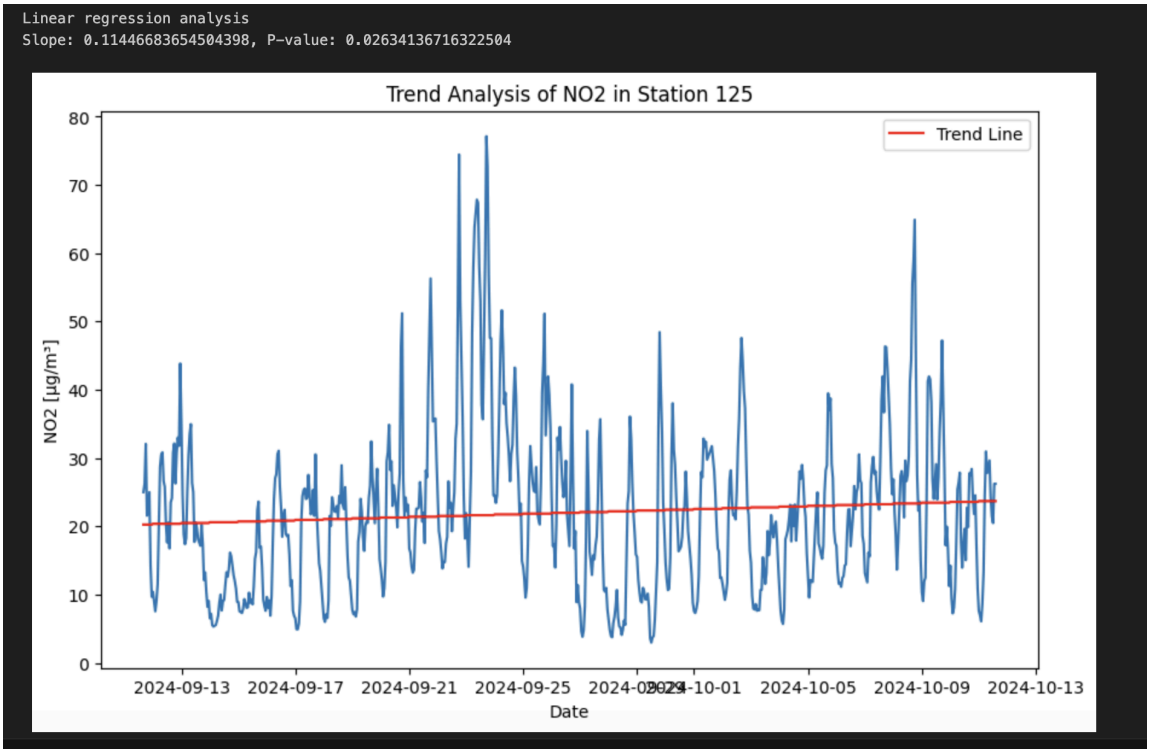


We took the same approach for identifying the significant correlations.
Here, the scatter plots as well show that the parameters captured by those two stations have high correlations, meaning that they are almost close to each other.

Trend Analysis



We chose to analyze the trend of PM2.5 in station 125. As you can see the decreasing red trend line indicates that the PM2.5 level of concentration at Station 125 has been reducing over time. Besides this trend is statistically significant because of the very small value of p-value.



Discussion

In this project, we analyzed three air quality parameters measured in two stations in the Bruxelles city.

We can say that there is a negative trend in pollutant PM2.5, meaning that they are being reduced over time, which is good news.

Stations measuring the pollutants are having a good correlation with each other, meaning that their measured values for each pollutant are almost close to each other and having the same trend in time series plots. This shows that except for some outliers, they seem logically correct captured.

Besides, the trend of NO2 is positive over time, which is not environmentally good and needs to be considered by the government.