Analysis of Air Quality Data

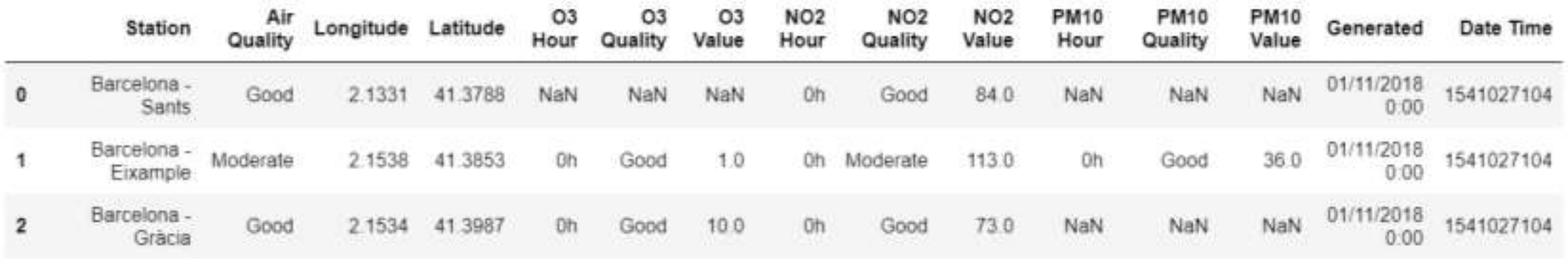
Project in IDS (UE18CS203)

Sem 3, Sec ‘D’

PES1201800002	Aronya Baksy
PES1201800046	Ruchika S
PES1201800275	Ansh Sarkar

Data Set

Air quality information of the city of Barcelona. Measure data are showed of O3 (tropospheric Ozone), NO2 (Nitrogen dioxide), PM10 (Suspended particles), as measured by 8 stations measured across the city



	Station	Air Quality	Longitude	Latitude	O3 Hour	O3 Quality	O3 Value	NO2 Hour	NO2 Quality	NO2 Value	PM10 Hour	PM10 Quality	PM10 Value	Generated	Date Time
0	Barcelona - Sants	Good	2.1331	41.3788	NaN	NaN	NaN	0h	Good	84.0	NaN	NaN	NaN	01/11/2018 0:00	1541027104
1	Barcelona - Eixample	Moderate	2.1538	41.3853	0h	Good	1.0	0h	Moderate	113.0	0h	Good	36.0	01/11/2018 0:00	1541027104
2	Barcelona - Gràcia	Good	2.1534	41.3987	0h	Good	10.0	0h	Good	73.0	NaN	NaN	NaN	01/11/2018 0:00	1541027104

COLUMNS WITH TIME DATA

5 COLUMNS

O3 hour, NO2 Hour, PM10 Hour – Time of the day

Generated Date, Generated Time

NUMERICAL DATA

6 COLUMNS

O3 Value, NO2 Value, PM10 Value concentration in ppm

Longitude and Latitude

CATEGORICAL DATA

4 COLUMNS

Station – Different places in Barcelona

Air Quality, O3 Quality, NO2 Quality, PM10 Quality – Good / Moderate

NO. OF ROWS

5745

Around 13% of data for each Station – Ciutadella, Eixample, Gracia, Observ Fabra, Palau Reial, Poblenou, Sants, Vall Hebron

NO. OF COLUMNS

15

- ➢ Greater than 10 Attributes
- ➢ 6 Numerical Columns
- ➢ 5 Categorical Columns

NaN s

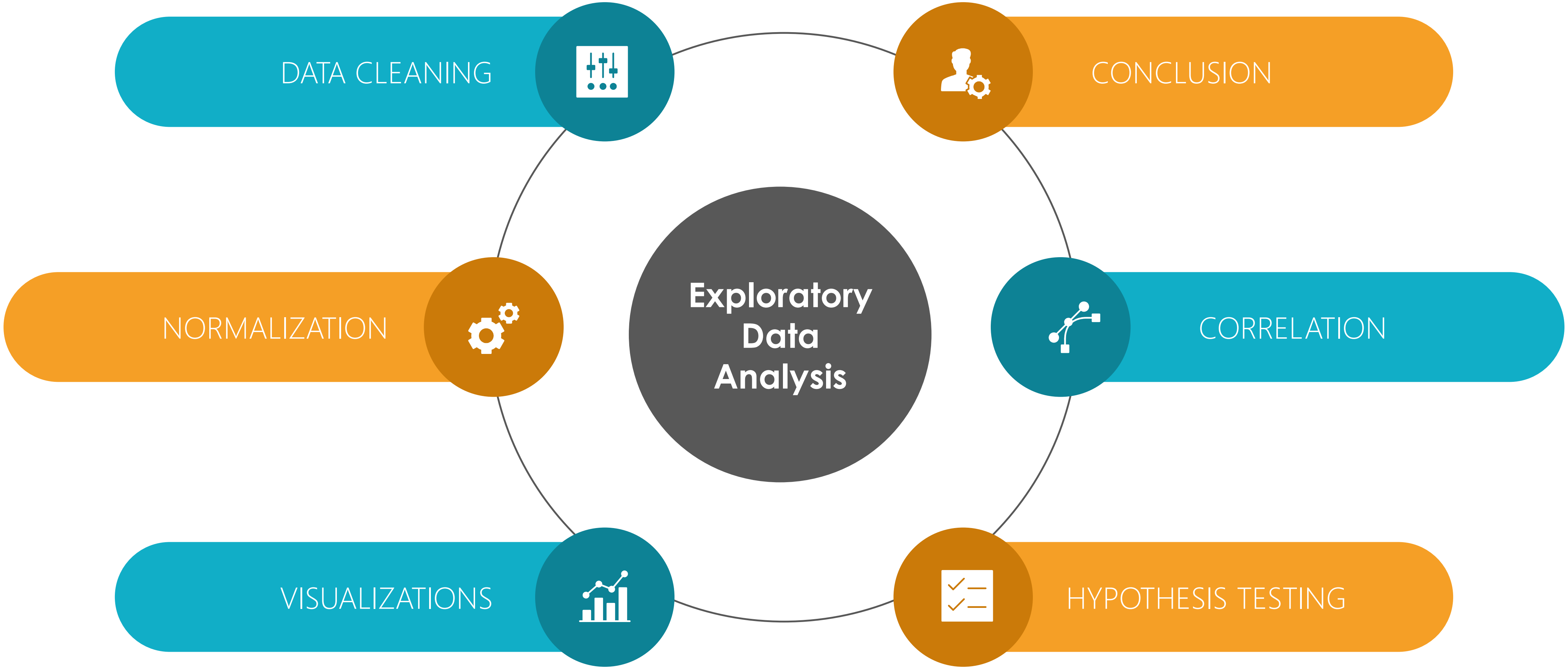
2- 5%

Missing data in columns:

Concentration Columns

Quality Columns

Project Analysis



Exploratory Data Analysis

DATA CLEANING

CONCLUSION

NORMALIZATION

CORRELATION

VISUALIZATIONS

HYPOTHESIS TESTING

Data Cleaning

Time Series Data Cleaning

Converted dd/mm/yy to values in seconds in Date Time Column, made the Generated Column Data with proper Date Time format.

To find the mean of values per station

Grouped the data according to the station where it was measured and finds the mean of all the numeric columns.

Replacing the missing values with the means

Each missing value is replaced with the average of that field for that particular station.

Replacing the Nans for Categorical Data

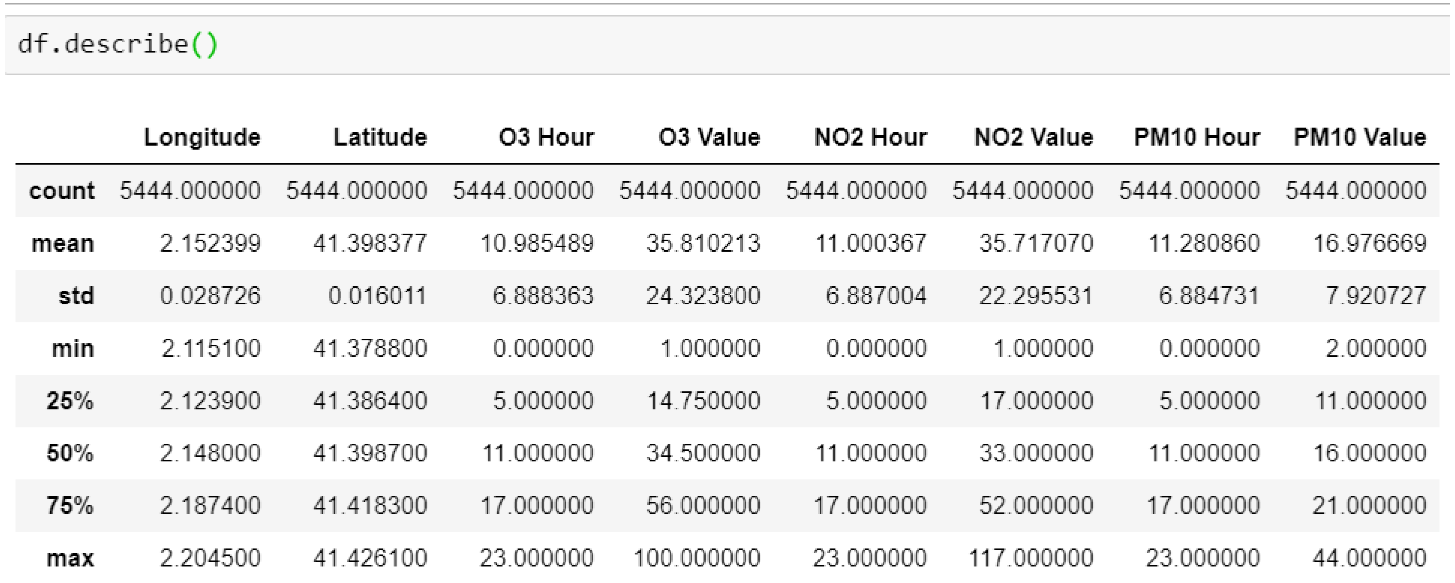
Each missing value in the columns such as O3 Hour, O3 Quality is replaced with it's previous row values.

Latitude Data Cleaning

A few of the values in Latitude Column were off by a scle of 10^4 for every 90 intervals, so these values were brought down to the right scale.

Descriptive Statistics

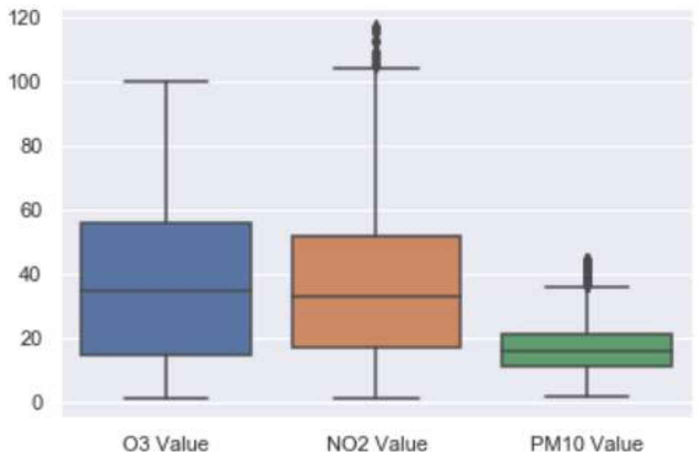
Generated a descriptive statistics after cleaning, that summarizes the central tendency, dispersion and shape of a dataset's distribution. For numeric data, the result index includes count, mean, std, min, max, as well as lower, 50 and upper percentiles. Here, the lower percentile is 25 and upper percentile is 75. The 50th percentile is the same as the median.

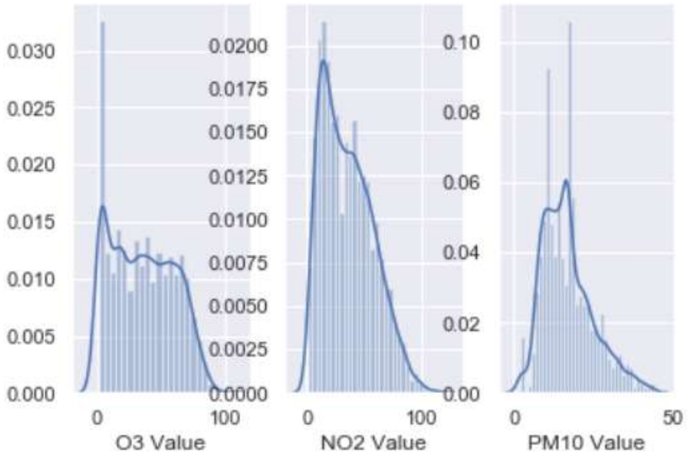


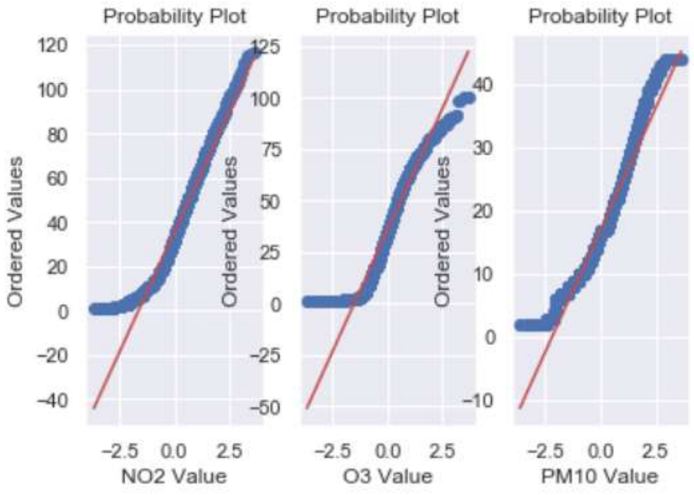
```
df.describe()
```

	Longitude	Latitude	O3 Hour	O3 Value	NO2 Hour	NO2 Value	PM10 Hour	PM10 Value
count	5444.000000	5444.000000	5444.000000	5444.000000	5444.000000	5444.000000	5444.000000	5444.000000
mean	2.152399	41.398377	10.985489	35.810213	11.000367	35.717070	11.280860	16.976669
std	0.028726	0.016011	6.888363	24.323800	6.887004	22.295531	6.884731	7.920727
min	2.115100	41.378800	0.000000	1.000000	0.000000	1.000000	0.000000	2.000000
25%	2.123900	41.386400	5.000000	14.750000	5.000000	17.000000	5.000000	11.000000
50%	2.148000	41.398700	11.000000	34.500000	11.000000	33.000000	11.000000	16.000000
75%	2.187400	41.418300	17.000000	56.000000	17.000000	52.000000	17.000000	21.000000
max	2.204500	41.426100	23.000000	100.000000	23.000000	117.000000	23.000000	44.000000

Visualizations





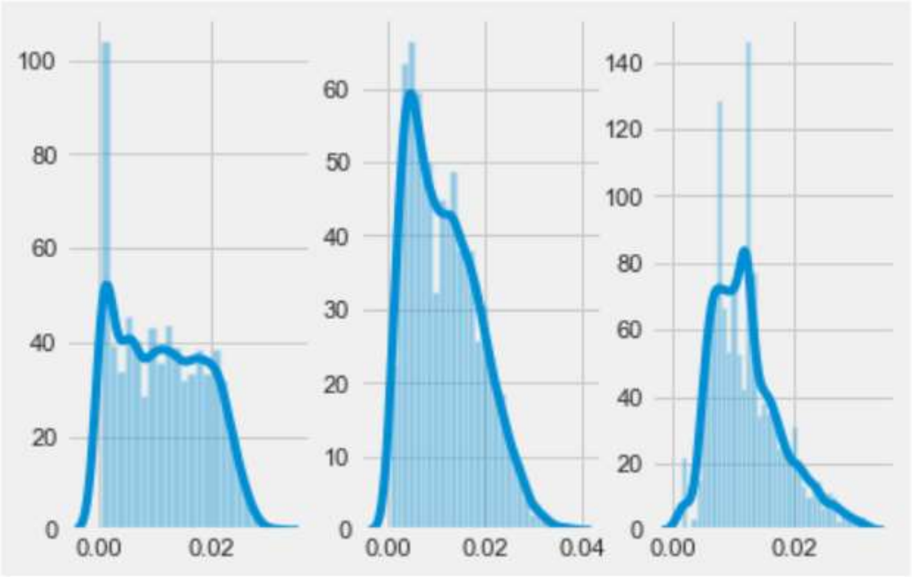


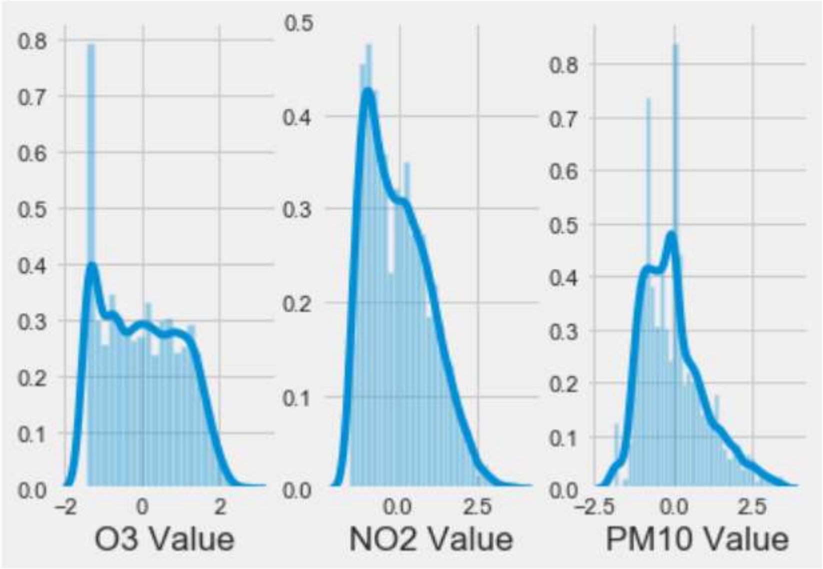
Insights: (Shape and distribution)

These plots describe the shape and distribution of the numerical columns in the data. The boxplots indicate the spread and the presence of outliers in the data. The distplot combines a histogram and the kernel density estimate that help to describe the shape of the data. The normal probability plot (quantile-quantile plot) describes the normal behaviour of the data.

Inference:
O3 Concentration shows few outliers and a skewed distribution as per the boxplot. However there are significant skewness and outliers in both NO2 and PM10 concentrations. All three columns show very significant deviation from normal behaviour.

Normalization & Standardization



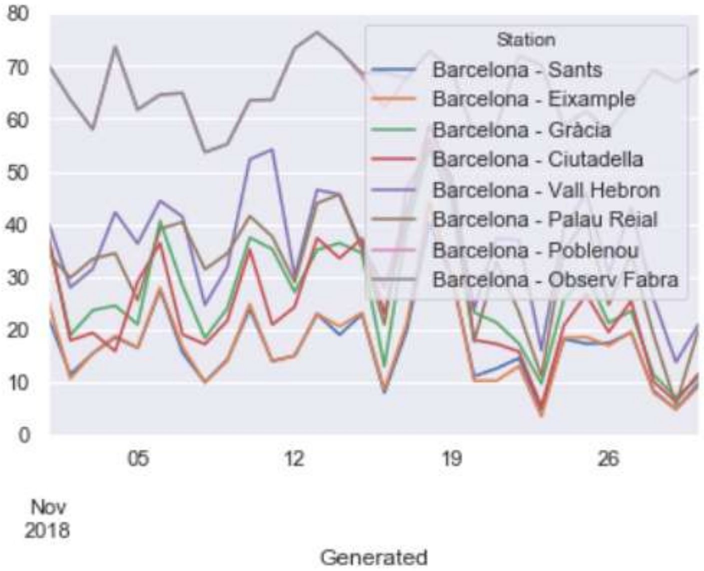


Normalized all the numerical values for the whole dataset. The above plot was obtained after normalizing the O3, NO2, PM10 Values by using sklearn.preprocessing.normalize( ) method for the column data.

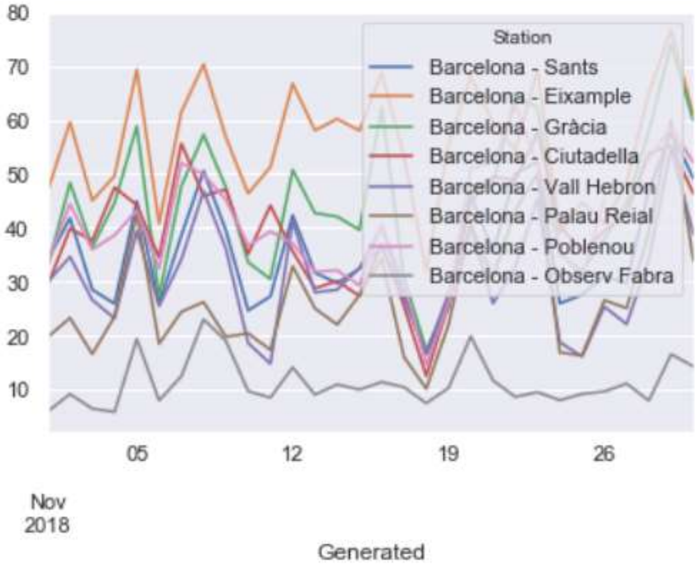
Standardized all the numerical values for the whole dataset. The above plot was obtained after normalizing the O3, NO2, PM10 Values by using sklearn.preprocessing.StandardScaler.fit_transform() method for the column data.

Visualizations

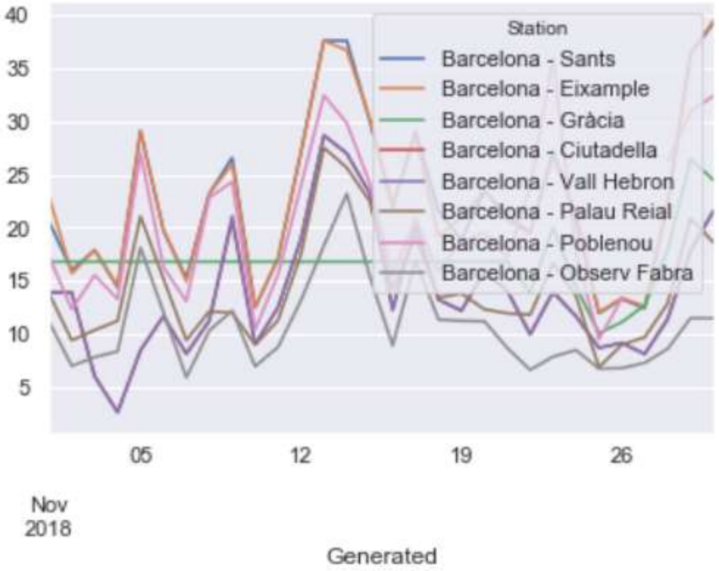
O3 Value



NO2 Value



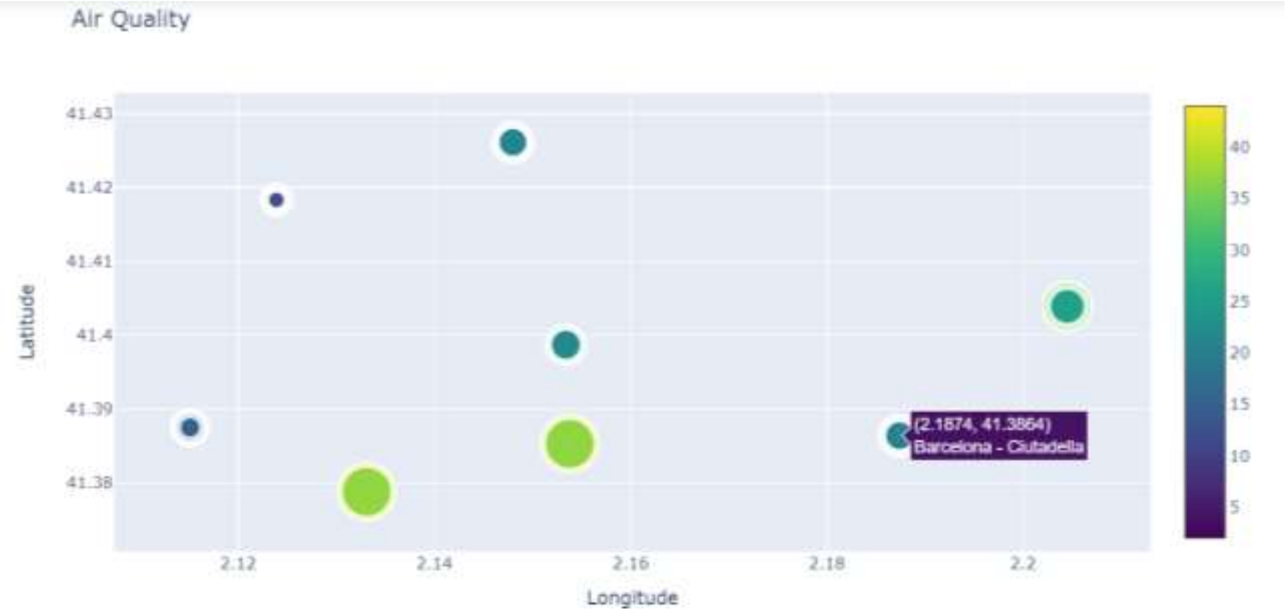
PM10 Value



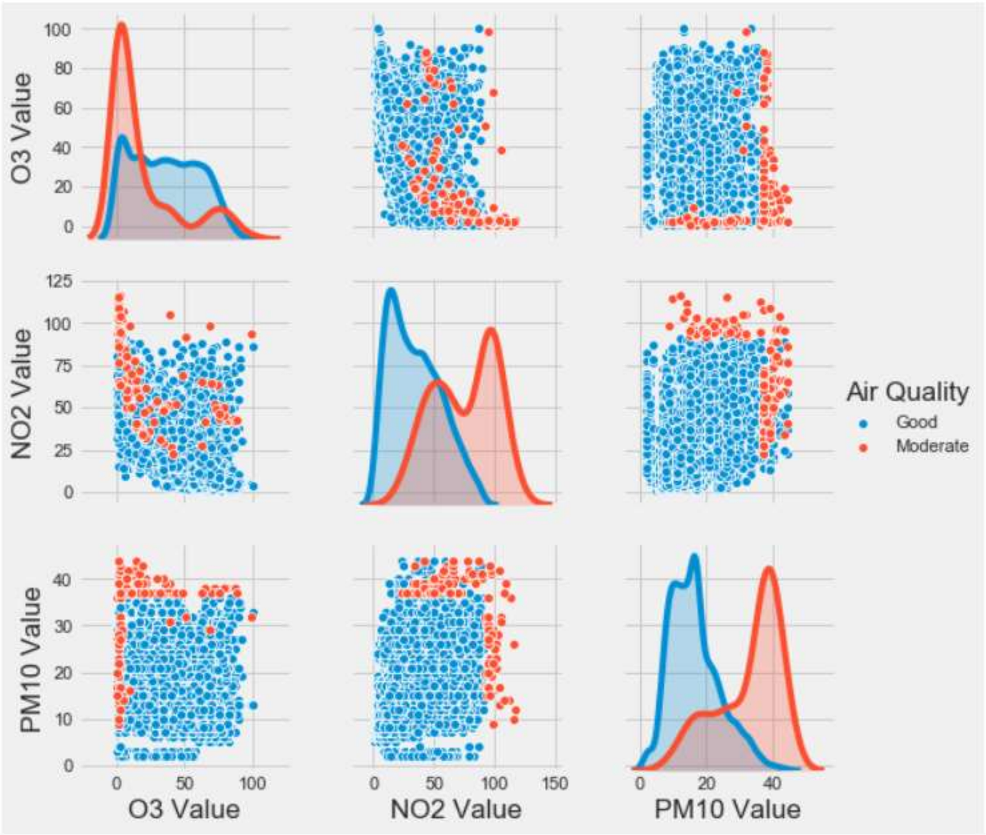
Insights: (Variation with time)

Visualization of the date-wise trends in concentration of O3, NO2 and PM10 over the given timeframe (November 2018) for each of the 8 stations in the city. The concentration of PM10 is clearly uniform across all the stations, while the values of NO2 and O3 concentrations follow no fixed pattern. But the rise and fall in concentrations of all three gases is uniform across all 8 stations.

Visualizations



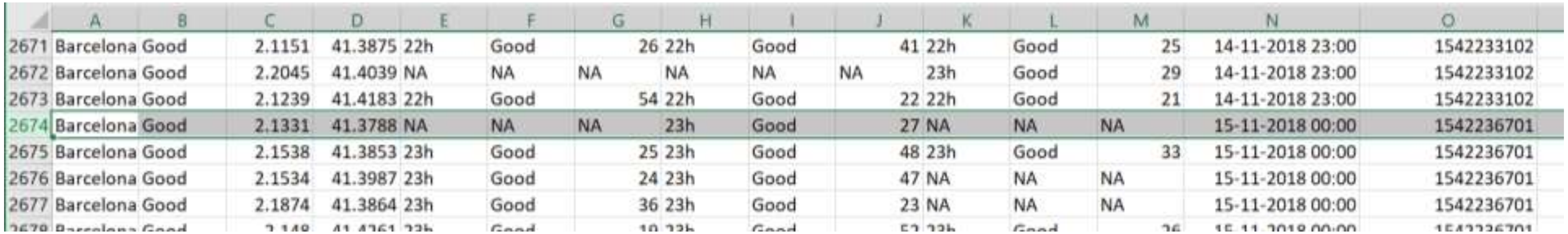
This plot shows the marked values of O3 Values, NO2 Values and PM10 Values at different stations in Barcelona.



This visualization is a Pairplot for different concentrations of O3, NO2, PM10 Values against the hue of Quality of Air.

Sampling the Dataset

We sampled the dataset for better visualizations and predictions to make a statistical inferences about the population from a small set of observations. We sampled it for a Time duration from 15Th Nov to 20Th Nov.

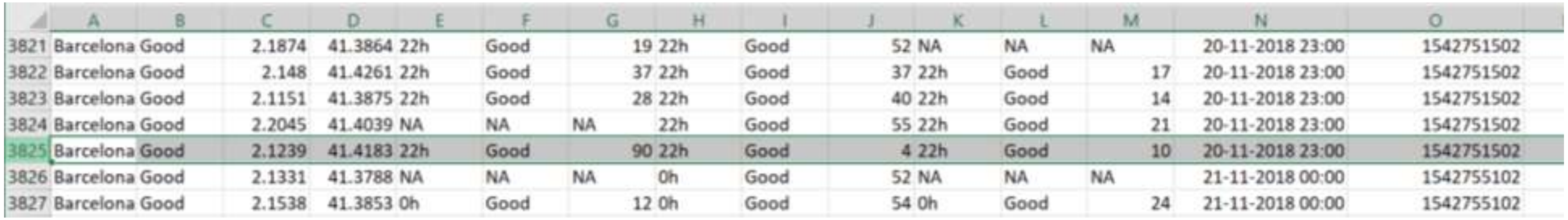


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
2671	Barcelona	Good	2.1151	41.3875	22h	Good	26	22h	Good	41	22h	Good	25	14-11-2018 23:00	1542233102
2672	Barcelona	Good	2.2045	41.4039	NA	NA	NA	NA	NA	NA	23h	Good	29	14-11-2018 23:00	1542233102
2673	Barcelona	Good	2.1239	41.4183	22h	Good	54	22h	Good	22	22h	Good	21	14-11-2018 23:00	1542233102
2674	Barcelona	Good	2.1331	41.3788	NA	NA	NA	23h	Good	27	NA	NA	NA	15-11-2018 00:00	1542236701
2675	Barcelona	Good	2.1538	41.3853	23h	Good	25	23h	Good	48	23h	Good	33	15-11-2018 00:00	1542236701
2676	Barcelona	Good	2.1534	41.3987	23h	Good	24	23h	Good	47	NA	NA	NA	15-11-2018 00:00	1542236701
2677	Barcelona	Good	2.1874	41.3864	23h	Good	36	23h	Good	23	NA	NA	NA	15-11-2018 00:00	1542236701
2678	Barcelona	Good	2.148	41.4261	23h	Good	19	23h	Good	52	23h	Good	26	15-11-2018 00:00	1542236701

.

.

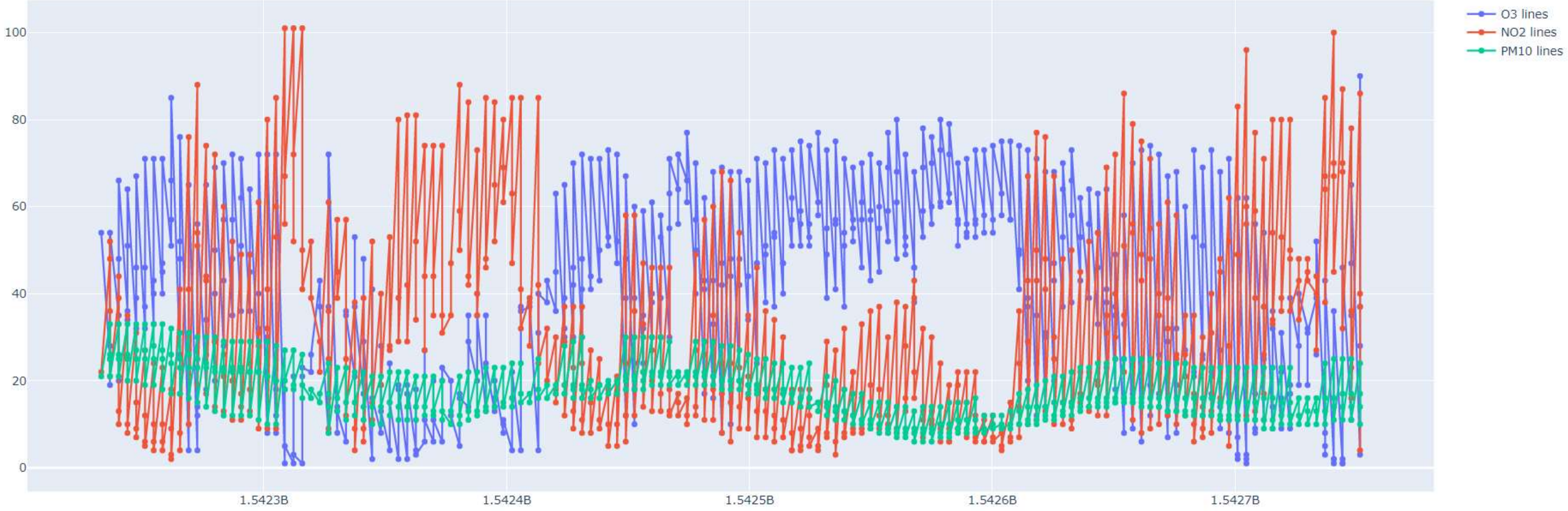
.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
3821	Barcelona	Good	2.1874	41.3864	22h	Good	19	22h	Good	52	NA	NA	NA	20-11-2018 23:00	1542751502
3822	Barcelona	Good	2.148	41.4261	22h	Good	37	22h	Good	37	22h	Good	17	20-11-2018 23:00	1542751502
3823	Barcelona	Good	2.1151	41.3875	22h	Good	28	22h	Good	40	22h	Good	14	20-11-2018 23:00	1542751502
3824	Barcelona	Good	2.2045	41.4039	NA	NA	NA	22h	Good	55	22h	Good	21	20-11-2018 23:00	1542751502
3825	Barcelona	Good	2.1239	41.4183	22h	Good	90	22h	Good	4	22h	Good	10	20-11-2018 23:00	1542751502
3826	Barcelona	Good	2.1331	41.3788	NA	NA	NA	0h	Good	52	NA	NA	NA	21-11-2018 00:00	1542755102
3827	Barcelona	Good	2.1538	41.3853	0h	Good	12	0h	Good	54	0h	Good	24	21-11-2018 00:00	1542755102

Visualizations

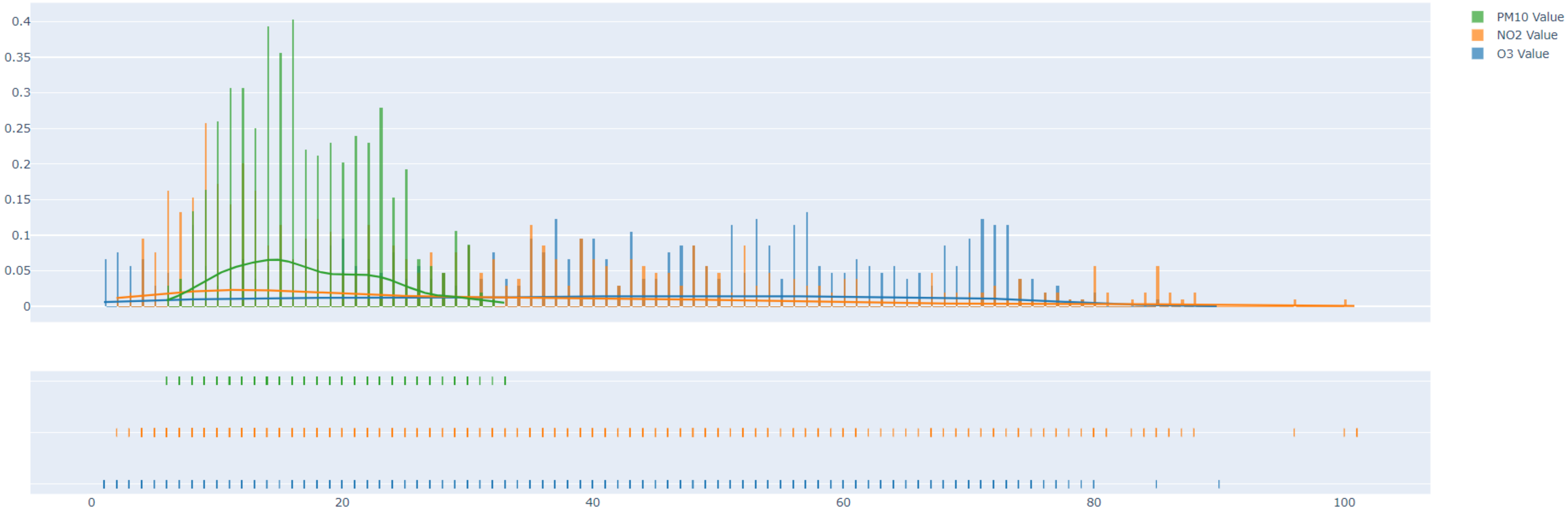
Analysis of the concentration of all three compounds values wrt time in seconds



Selecting a particular gas value in the graph we can basically identify the time interval weather forecasts can be predicted. The variation in O3 levels gives us the insight for about the UV radiation .

Visualizations

Plotted a histogram to show the probabilities of the values of the gases wrt time

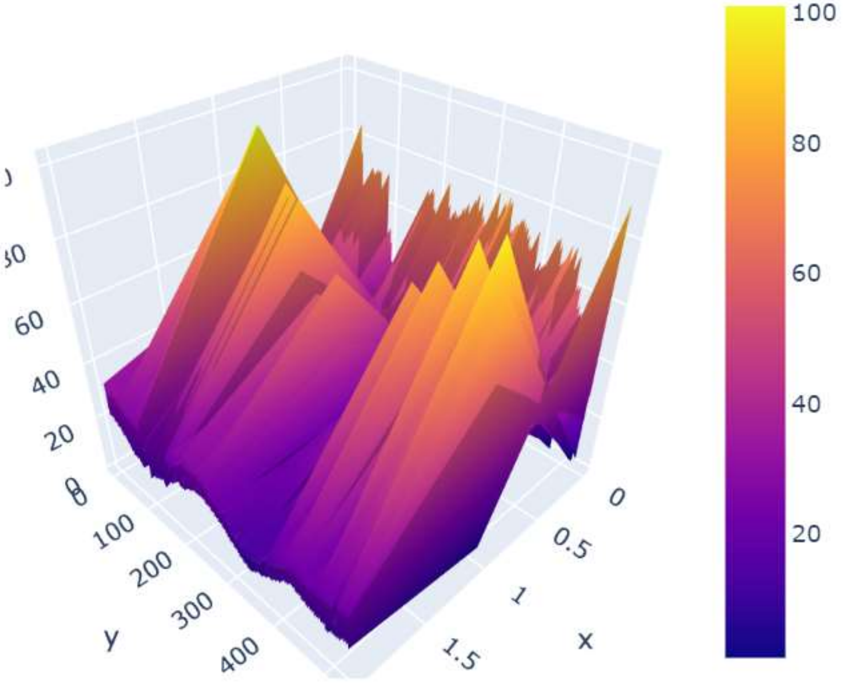


Analysing the above visualisation we can see which kind of distribution is followed for the sample ,

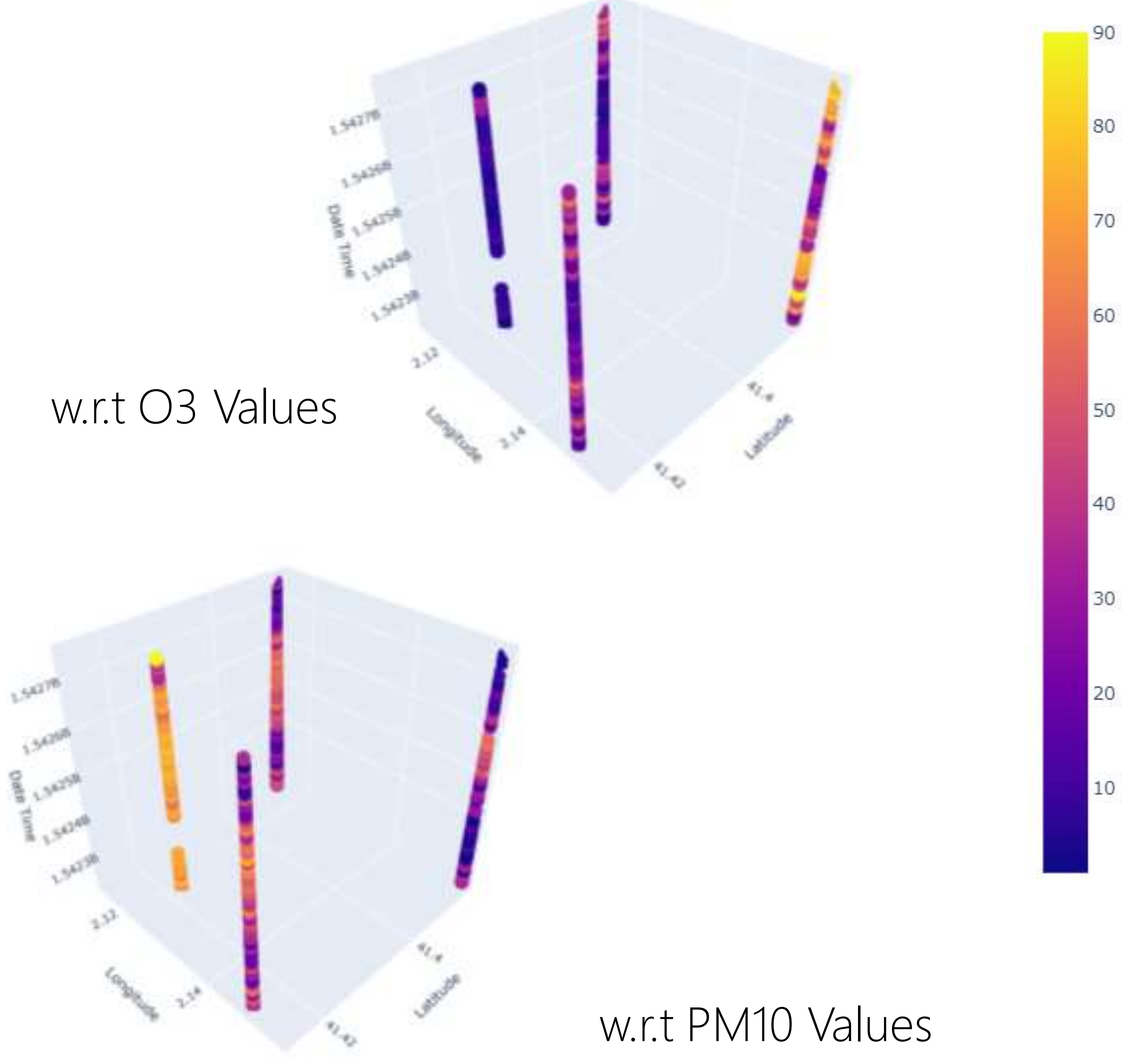
Visualizations

Time Series Plot

Realtion between O3 , NO2 and PM10 Values



Taking time into account and plotting variation of NO2 Values

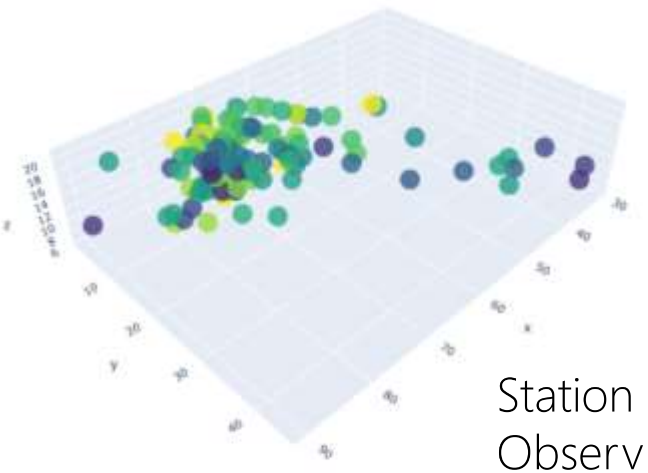


w.r.t O3 Values

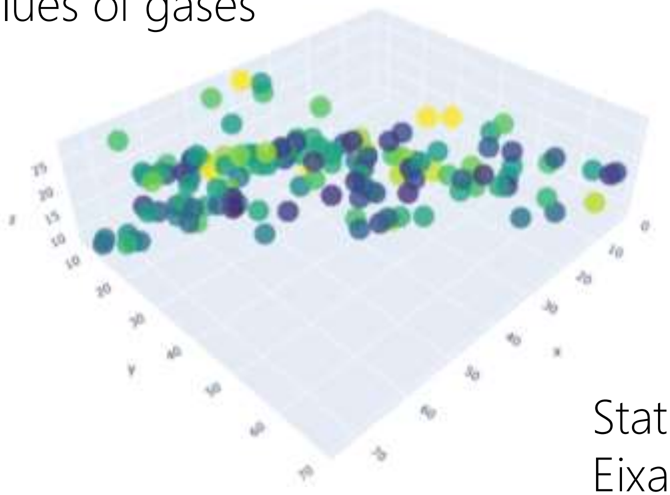
w.r.t PM10 Values

Visualizations

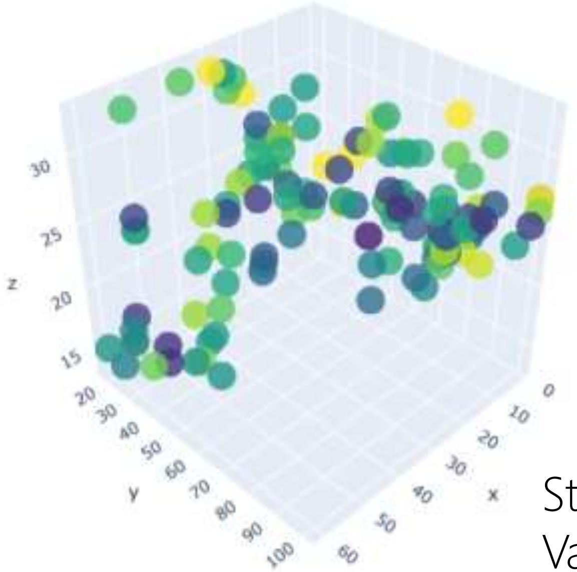
Station wise the distribution of the values of gases



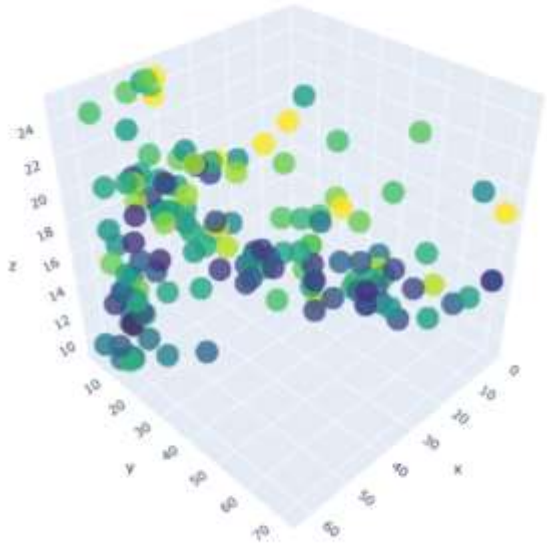
Station 1 :
Observ Fabra



Station 2:
Eixample



Station 3:
Vall Hebron



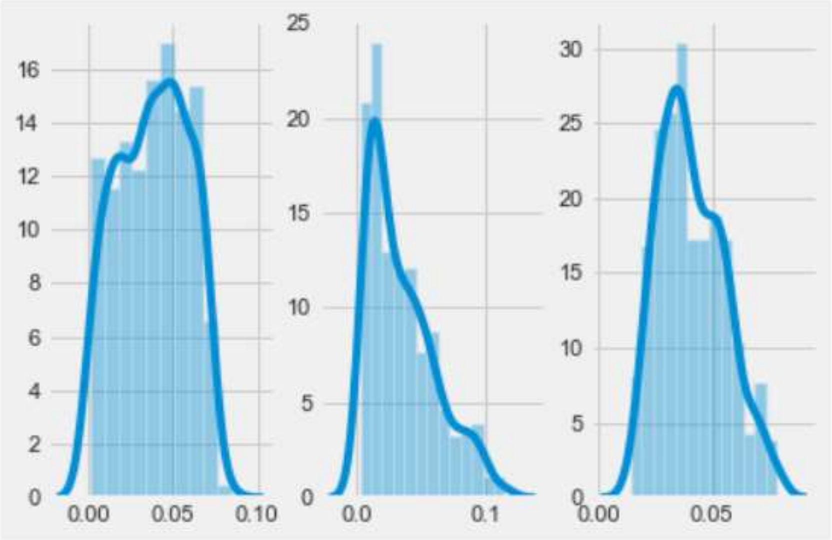
Station 4:
Palau Reial

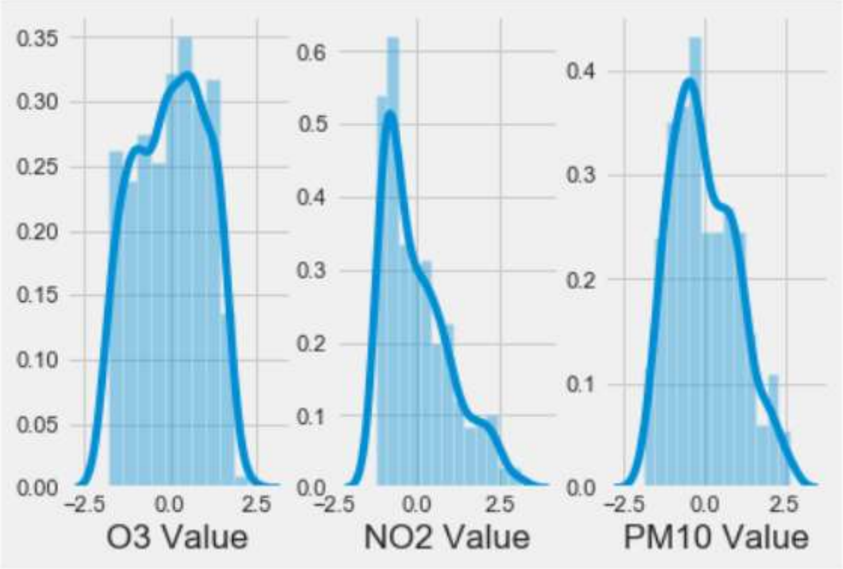
X: O3 Value

Y: NO2 Value

Z: PM10 Value

Normalization & Standardization



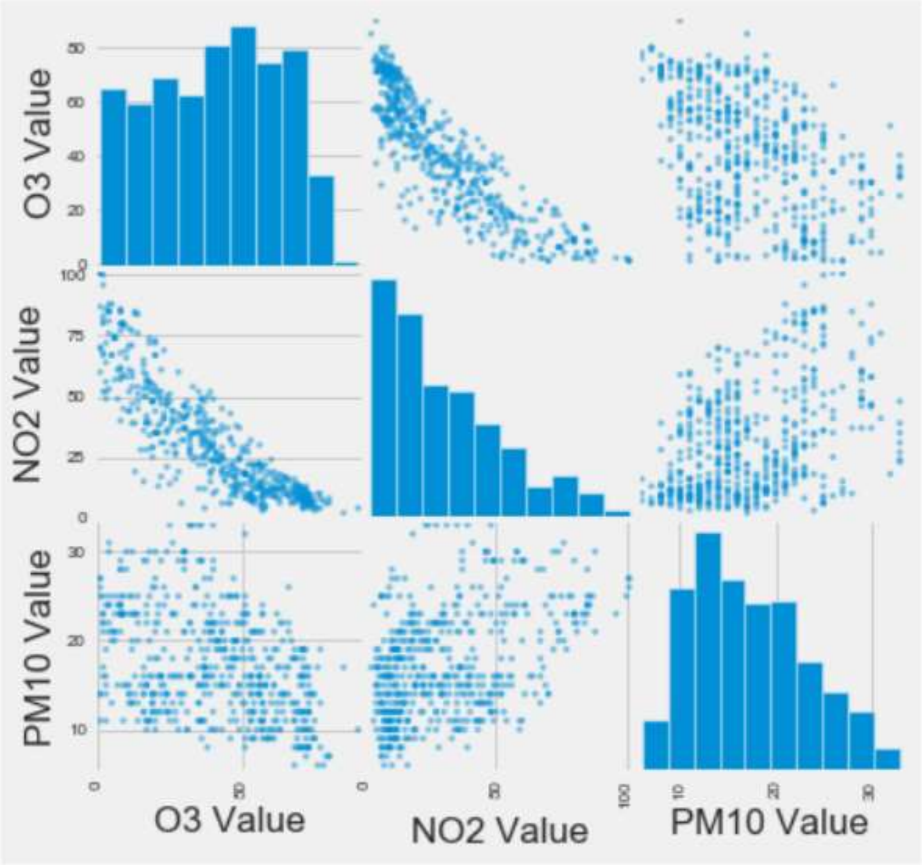


Normalized all the numerical values for the sampled dataset. The above plot was obtained after normalizing the O3, NO2, PM10 Values by using sklearn.preprocessing.normalize( ) method for the column data.

Standardized all the numerical values for the sampled dataset. The above plot was obtained after normalizing the O3, NO2, PM10 Values by using sklearn.preprocessing.StandardScaler.fit_transform() method for the column data.

Inference: The plots after sampling fit more towards a Gaussian / Bell shaped curve, than all the values in the dataset before sampling

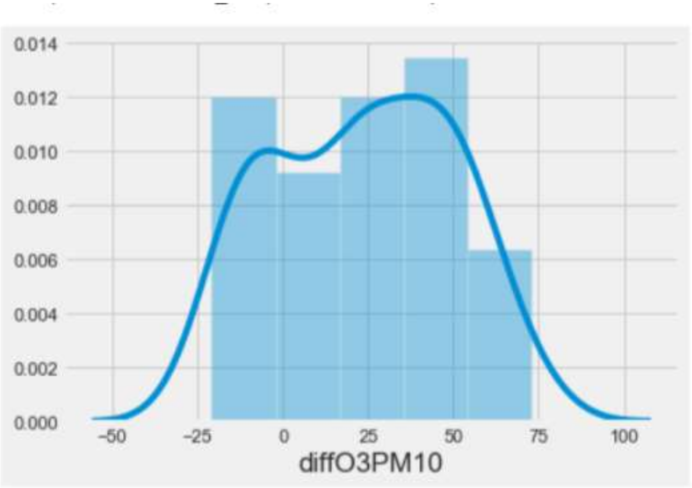
Visualizations



This visualization is a Pairplot for different concentrations of O3, NO2, PM10 Values against the hue of Good Quality of Air.

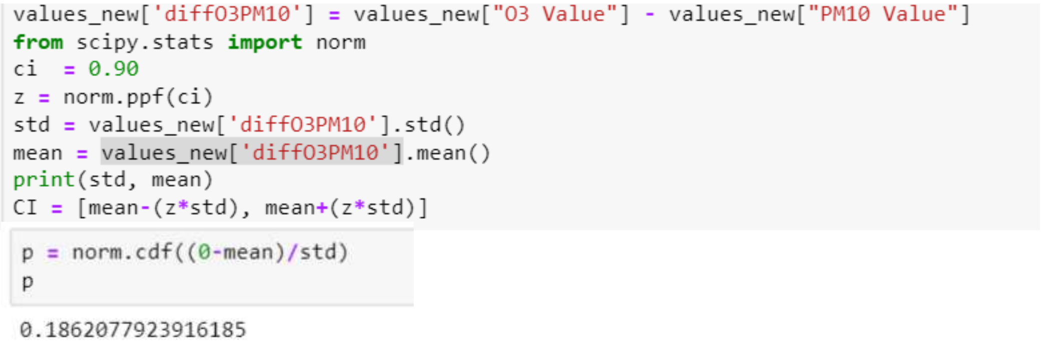
The Pairplot is much more evenly distributed and normal in nature than compared to the Pairplot w.r.t the whole dataset

Hypothesis Testing



H0: O3 Values and PM10 values are equal
H1: O3 Values and PM10 Values are not equal
p-value < alpha

Hence, the Null Hypothesis can be rejected.

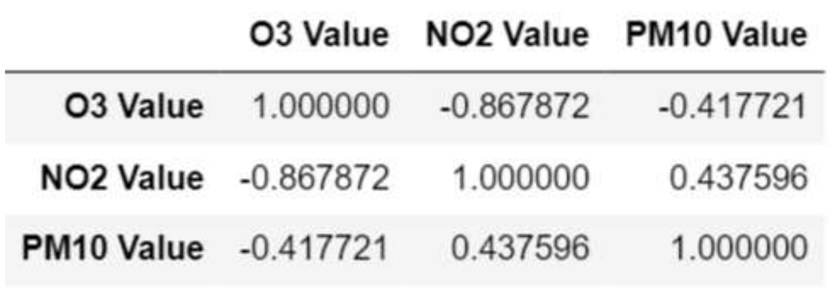


```
values_new['diffO3PM10'] = values_new["O3 Value"] - values_new["PM10 Value"]
from scipy.stats import norm
ci  = 0.90
z = norm.ppf(ci)
std = values_new['diffO3PM10'].std()
mean = values_new['diffO3PM10'].mean()
print(std, mean)
CI = [mean-(z*std), mean+(z*std)]
```

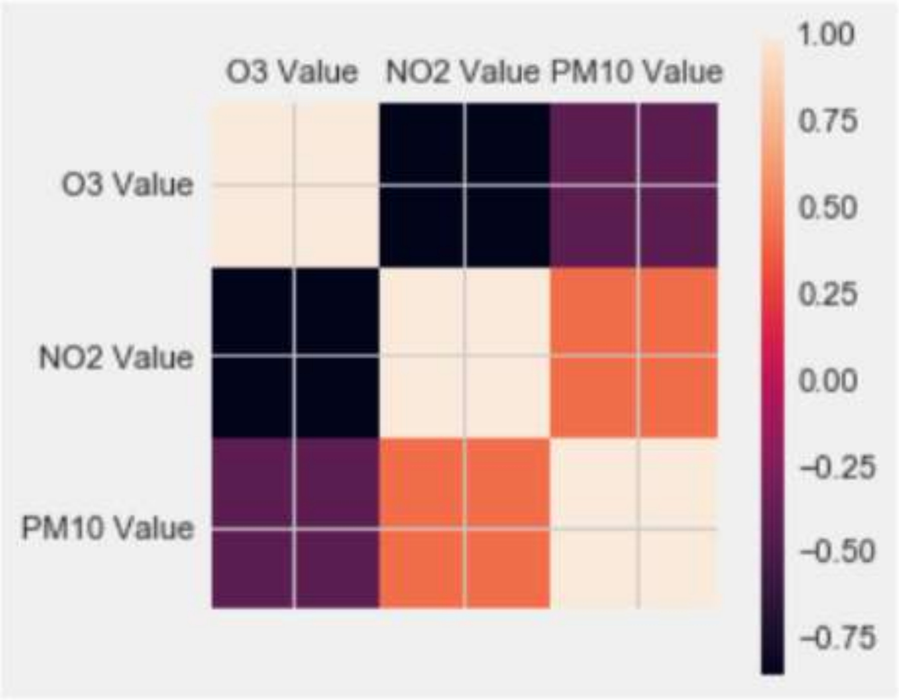
```
p = norm.cdf((0-mean)/std)
p
```

```
0.1862077923916185
```

Correlations



	O3 Value	NO2 Value	PM10 Value
O3 Value	1.000000	-0.867872	-0.417721
NO2 Value	-0.867872	1.000000	0.437596
PM10 Value	-0.417721	0.437596	1.000000



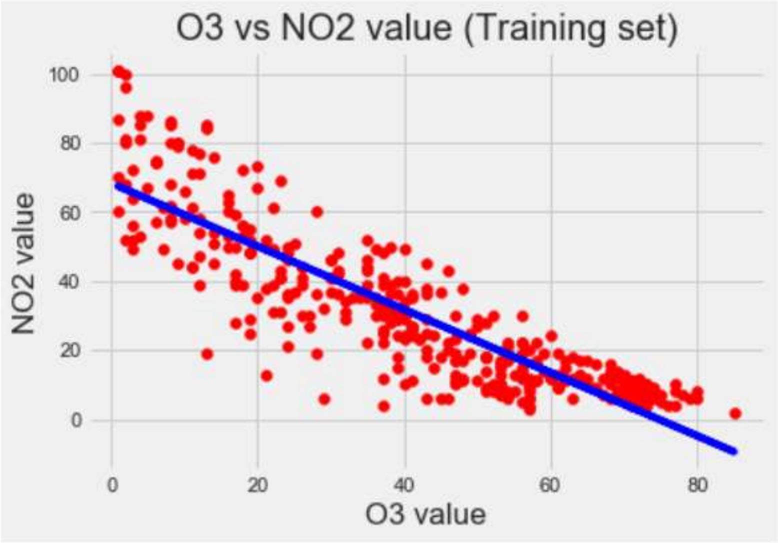
The cells with the value 1 show that the two column values are correlated highly and positive values show that those 2 columns are related and some insights can be inferred.
The negative values show that the two columns are somewhat strongly correlated (negative correlation)

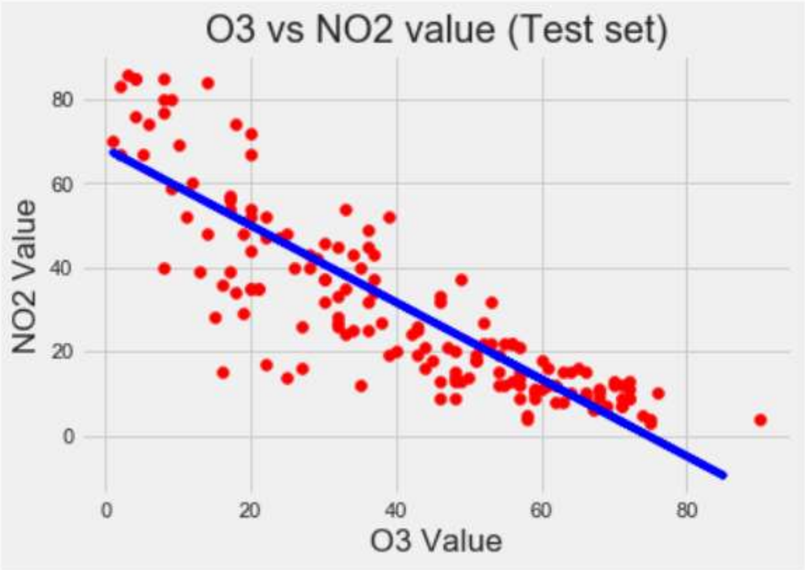
The intensity of the colour in the map indicates the strength of the correlation

We cannot, however, conclude that the variations in O3 concentration are caused by the variations in NO2 concentration

Linear Regression

We sampled the dataset for better visualizations and predictions to make a statistical inferences about the population from a small set of observations. We sampled it for a Time duration from 15Th Nov to 20Th Nov.





We also performed Linear Regression, to find the correlation to see how well our data was best fit to the regressed line. The above plot shows the result being visualized. We see a somewhat strong negative correlation between O3 and NO2 values, in both training and testings (r = -0.86)

Conclusion

As can be expected from these kind of measurements, we observe that most of the analysis behind this dataset involves effectively cleaning it at first.

After this, we perform exploratory data analysis techniques on our numerical data.

Initially we observe that our data shows significant deviation from normal behaviour and is not symmetric at all. but we can derive insights from our data by carefully sampling the given data.

Thank You