Seminar on Managing Industrial Air Emissions

21st Feb, 2018

Organized by: EMC & SDG Foundation

Supported by : Environmental SA & TSI

AIR QUALITY ANALYTICS DASHBOARD

Dr. Prasad Modak

Structure of the presentation

- About environmentalanalytics.co
- Air Quality Analytics Dashboard

What is environmentalanalytics.co ?

A platform will feature tools for analysing and visualizing air quality, stack emissions, water quality, wastewater and noise related BIG data.

Many of these tools have been conceived and developed based on Dr. Modak's extensive research and consulting experience.

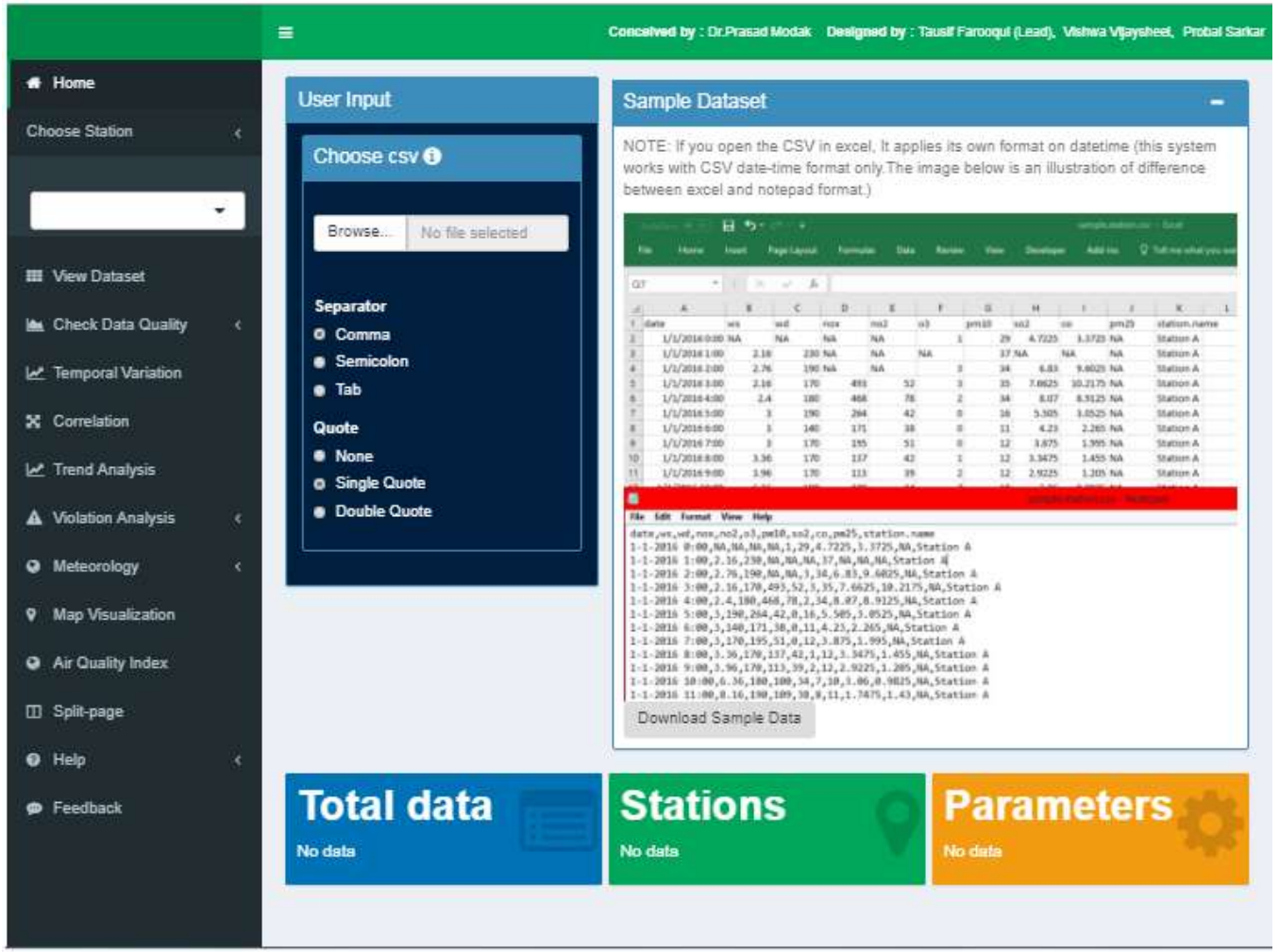
Air Quality Analytics Dashboard

This platform will feature tools for analysing and visualizing air quality, many of these tools have been conceived and developed based on Dr. Modak's extensive research and consulting experience.

The Dashboard

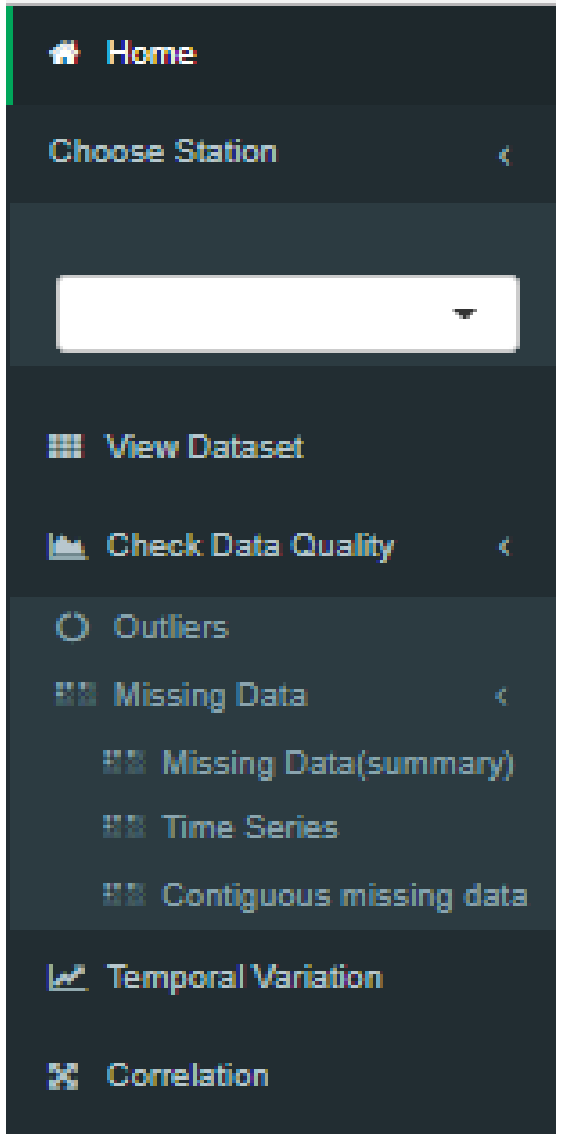
This is how the opening page of the dashboard looks like.

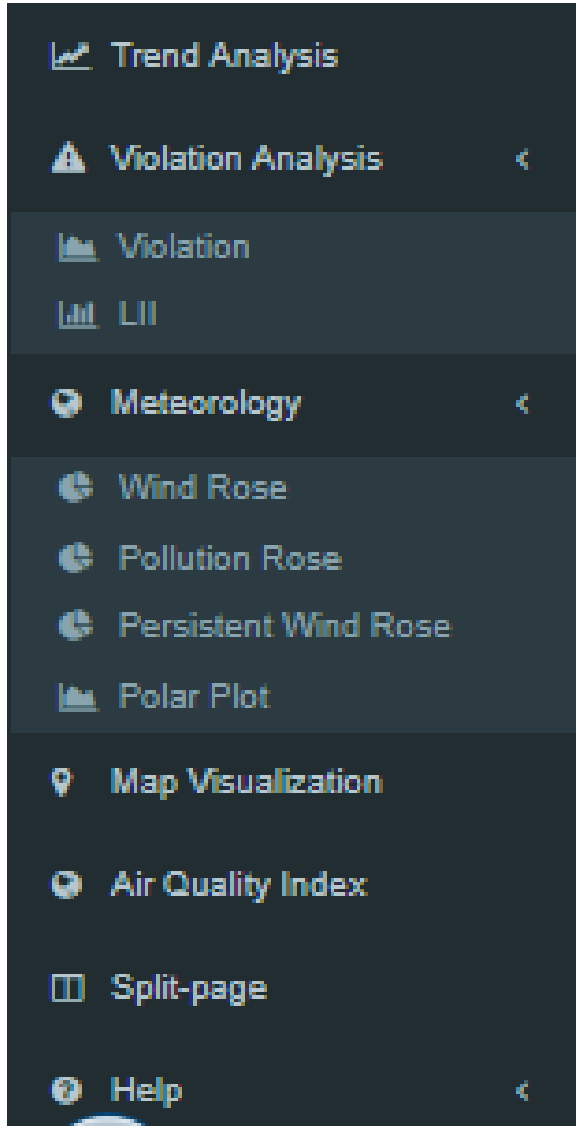
You will see a list of various tabs on the sidebar which when you select will open that page on the right and lets you perform various functions on the dataset accordingly.



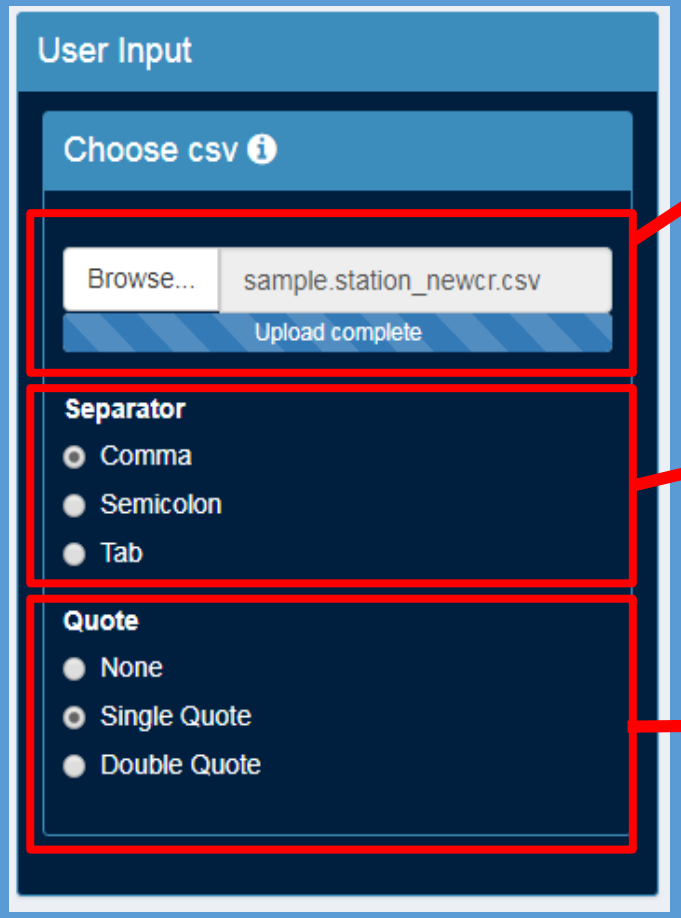
List of tools

- Outliers analysis
- Missing data analysis
- Trend analysis
- Location Importance Index (LII)
- Meteorological data analysis
- Map Visualizations
- Violation Analysis
- Temporal Variation





Home Page



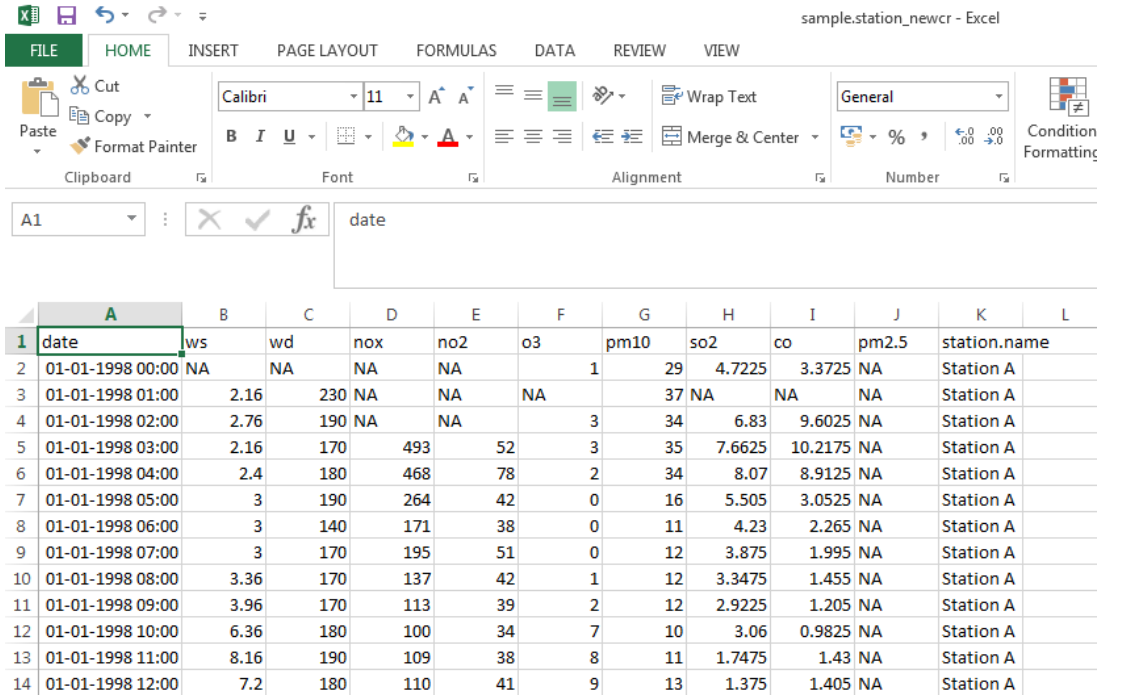
Choose CSV – You can choose a CSV file from your directory, this 'Browse 'button enables you to choose a file from your directory. Make sure the file data is in the specified format.

Separator – the field separator character. Values on each line of the file are separated by this character.

Quote – The set of quoting characters. Quoting is only considered for columns read as character, which is all of them unless

EXCEL

When you open the file in **'Excel'** the system automatically allocates different columns to different variables and hence when you open a CSV file in excel you will see something like this.

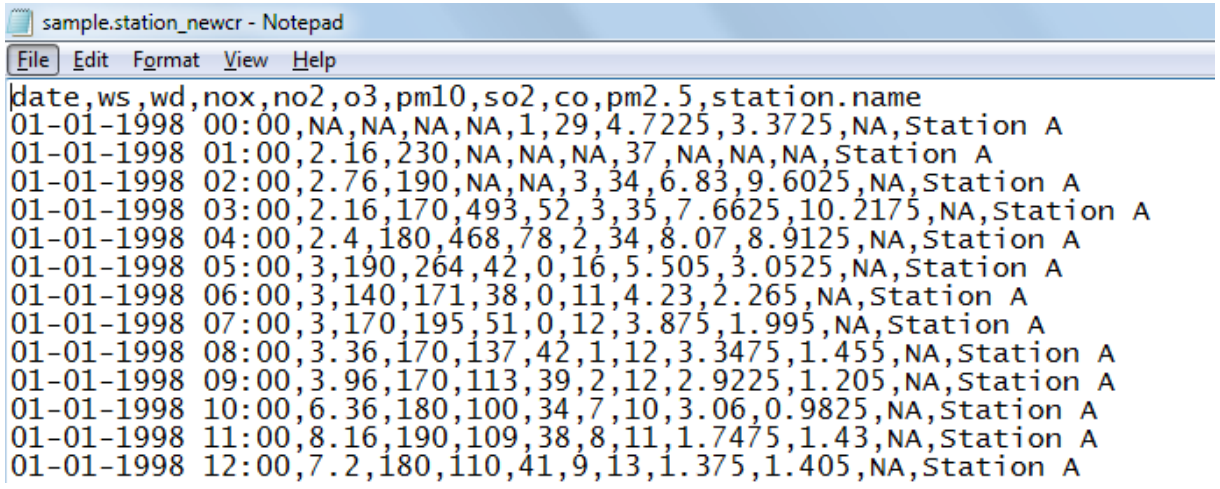


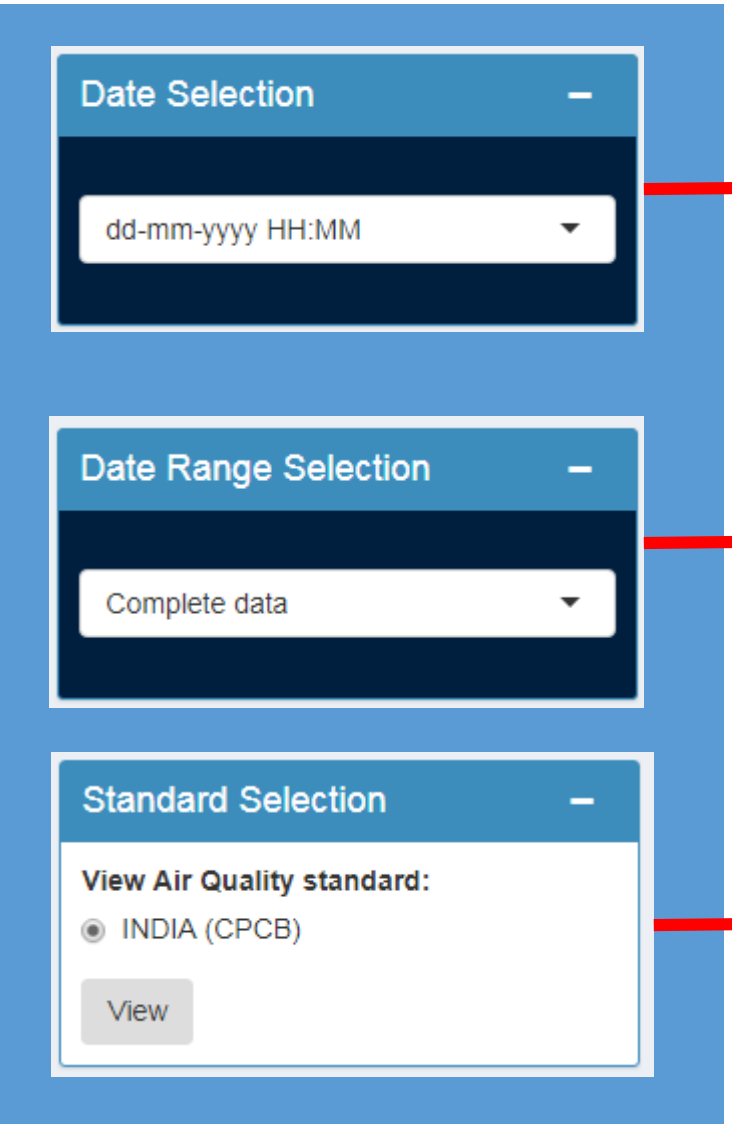
	A	B	C	D	E	F	G	H	I	J	K	L
1	date	ws	wd	nox	no2	o3	pm10	so2	co	pm2.5	station.name	
2	01-01-1998 00:00	NA	NA	NA	NA	1	29	4.7225	3.3725	NA	Station A	
3	01-01-1998 01:00	2.16	230	NA	NA	NA	37	NA	NA	NA	Station A	
4	01-01-1998 02:00	2.76	190	NA	NA	3	34	6.83	9.6025	NA	Station A	
5	01-01-1998 03:00	2.16	170	493	52	3	35	7.6625	10.2175	NA	Station A	
6	01-01-1998 04:00	2.4	180	468	78	2	34	8.07	8.9125	NA	Station A	
7	01-01-1998 05:00	3	190	264	42	0	16	5.505	3.0525	NA	Station A	
8	01-01-1998 06:00	3	140	171	38	0	11	4.23	2.265	NA	Station A	
9	01-01-1998 07:00	3	170	195	51	0	12	3.875	1.995	NA	Station A	
10	01-01-1998 08:00	3.36	170	137	42	1	12	3.3475	1.455	NA	Station A	
11	01-01-1998 09:00	3.96	170	113	39	2	12	2.9225	1.205	NA	Station A	
12	01-01-1998 10:00	6.36	180	100	34	7	10	3.06	0.9825	NA	Station A	
13	01-01-1998 11:00	8.16	190	109	38	8	11	1.7475	1.43	NA	Station A	
14	01-01-1998 12:00	7.2	180	110	41	9	13	1.375	1.405	NA	Station A	

NOTEPAD

When you open the file in **'Notepad'** the system is divides the variables on the basis of their **format like number or character or date and separate them using a 'Separator' like comma/Semi colon/Tab.**

The character variables are stored using 'Quotes' , single quotes/double quotes/None, depending on your system.





Date Selection –The **'date' column** format changes, make sure you select the right date format from the option given. If the date format is wrong the application will not process any results. You can either correct the date format in the original file or choose the correct date format.

Date Range Selection – The following tab will allow you to select a date range from within the dataset, If you to analyse data within a certain range from your dataset, you might want to use this option.

Standard Selection – This shows the standard selected for calculations in the coming pages like 'Violation analysis'. You can have a look at the standard by clicking the view button

View Dataset

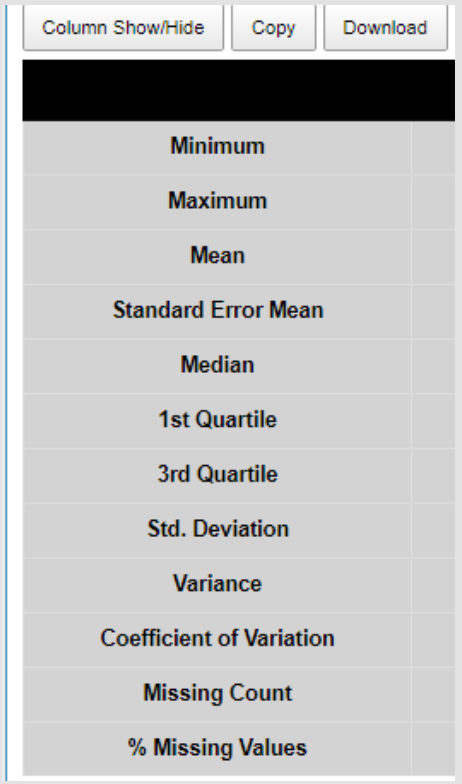
1st Quartile : The first quartile (also called the lower quartile) is the number below which lies the 25 % of the bottom data points.

3rd Quartile : The third quartile (also called the upper quartile) has 75 % of the data below it.

Standard deviation : Standard Deviation is a measure of the mean distance of values in a data set from their mean.

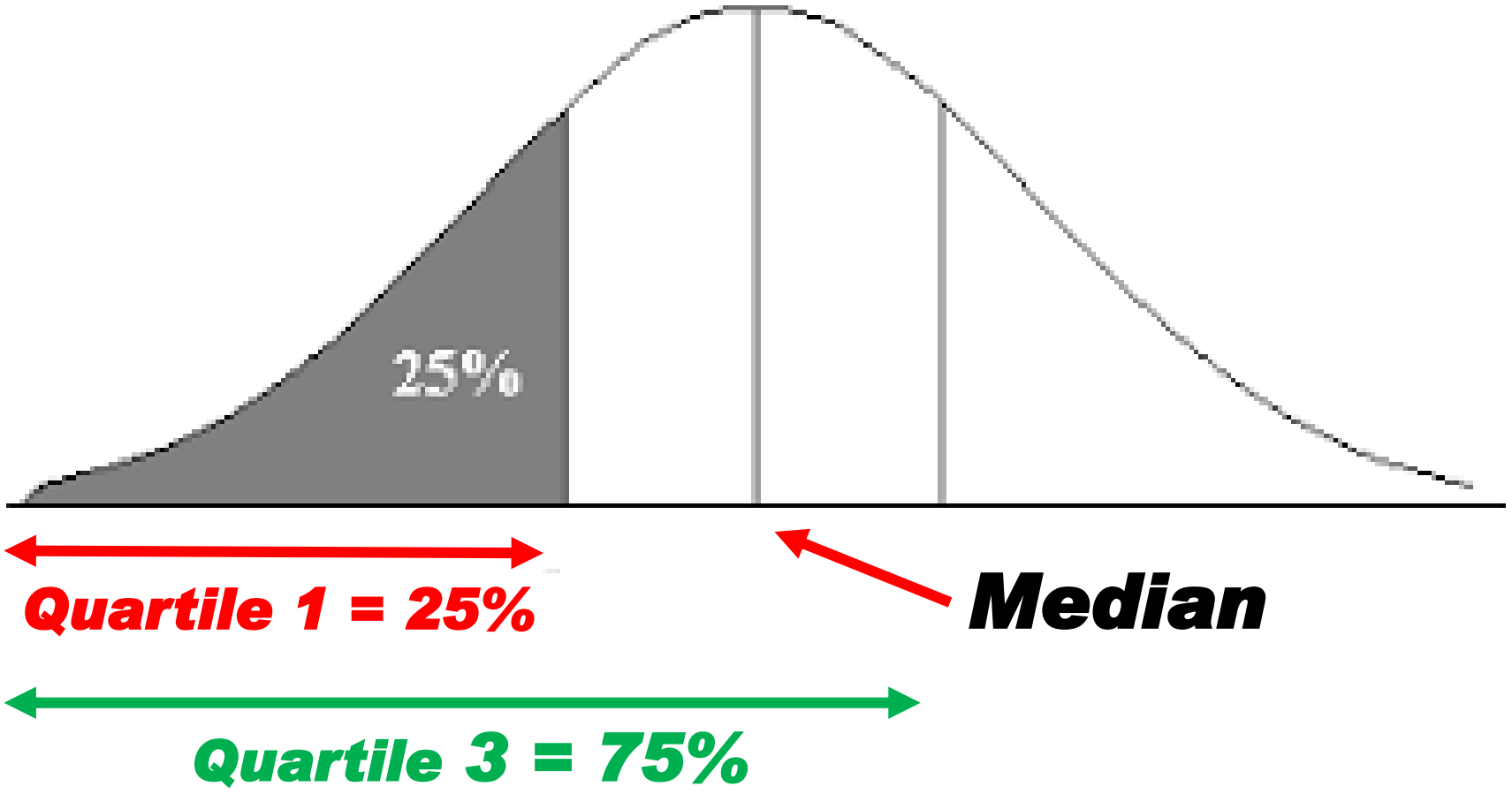
Variance : It tells how the data is spread around the mean..

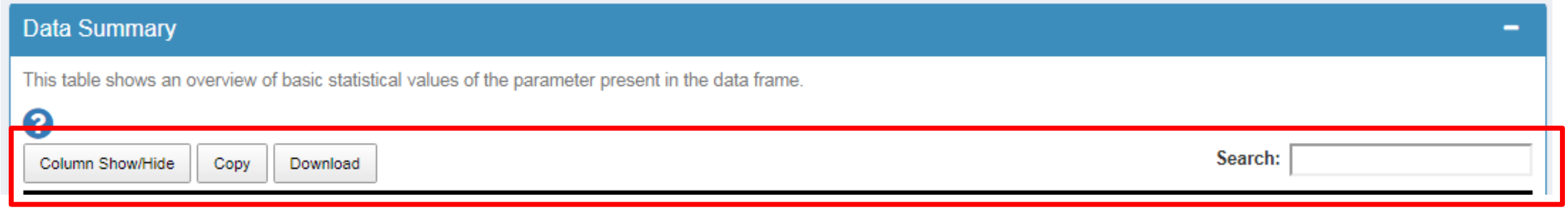
Coefficient of variation : It is often expressed as a percentage, and is defined as the ratio of the standard deviation to the mean.



A view of the Data summary tab on Dataset page

A normal distribution curve depicting **Quartiles** and **Median**





Column Show/Hide : This button is used to select the columns to show or hide in the table.

Copy : This button is used to copy data to the clipboard.

Download : This button is used to download table in different formats, CSV, Excel or PDF.

Search – You can search for data relevant terms or values within the table using the search bar.

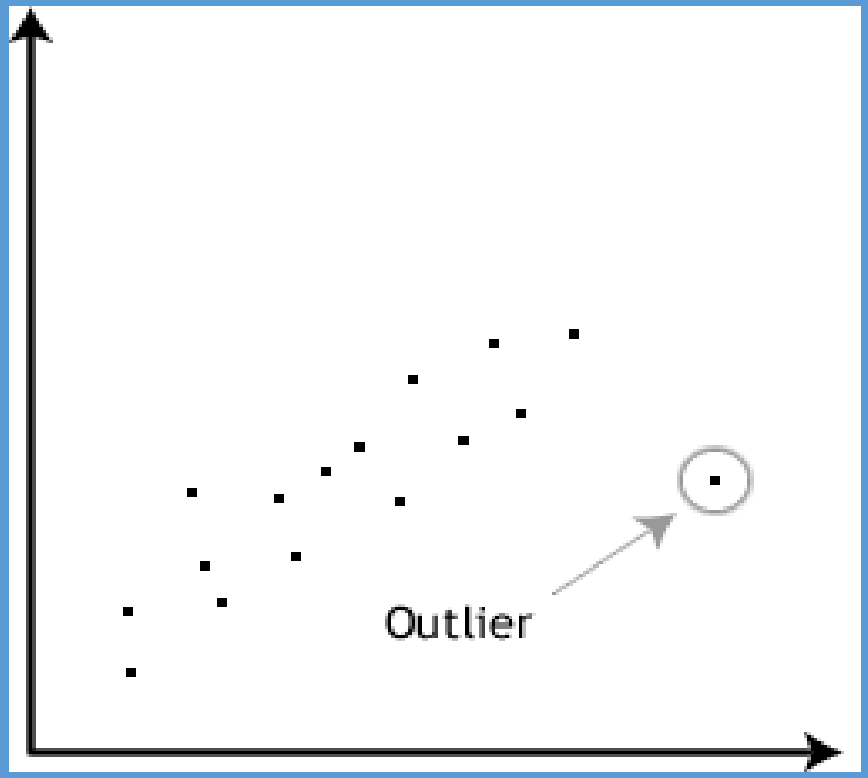
Value filter tab : These buttons are used to sort the columns.

Column Sorting : This button filter data based on column values.

What are Outliers ?

Statistical data which is extremely different from the others in the same sample.

Mathematically :
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

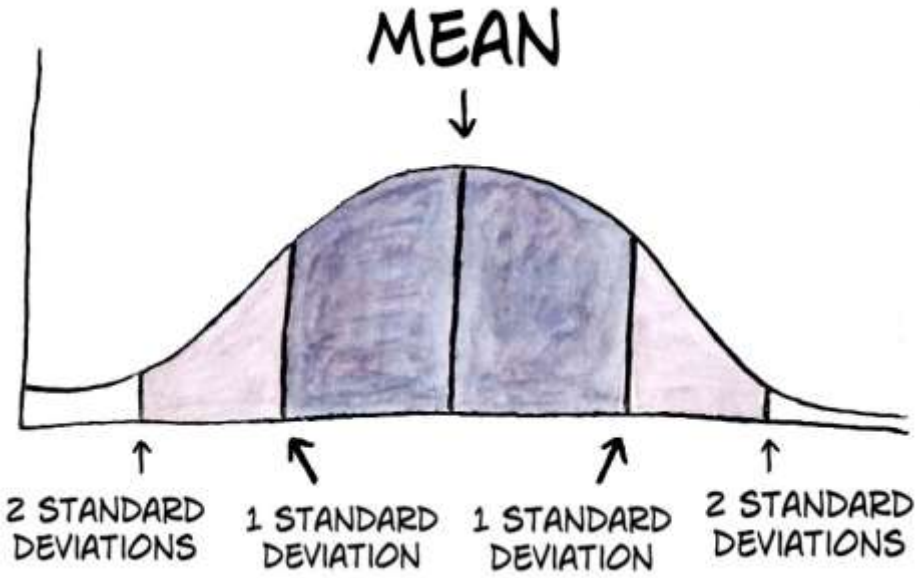


Outlier calculation

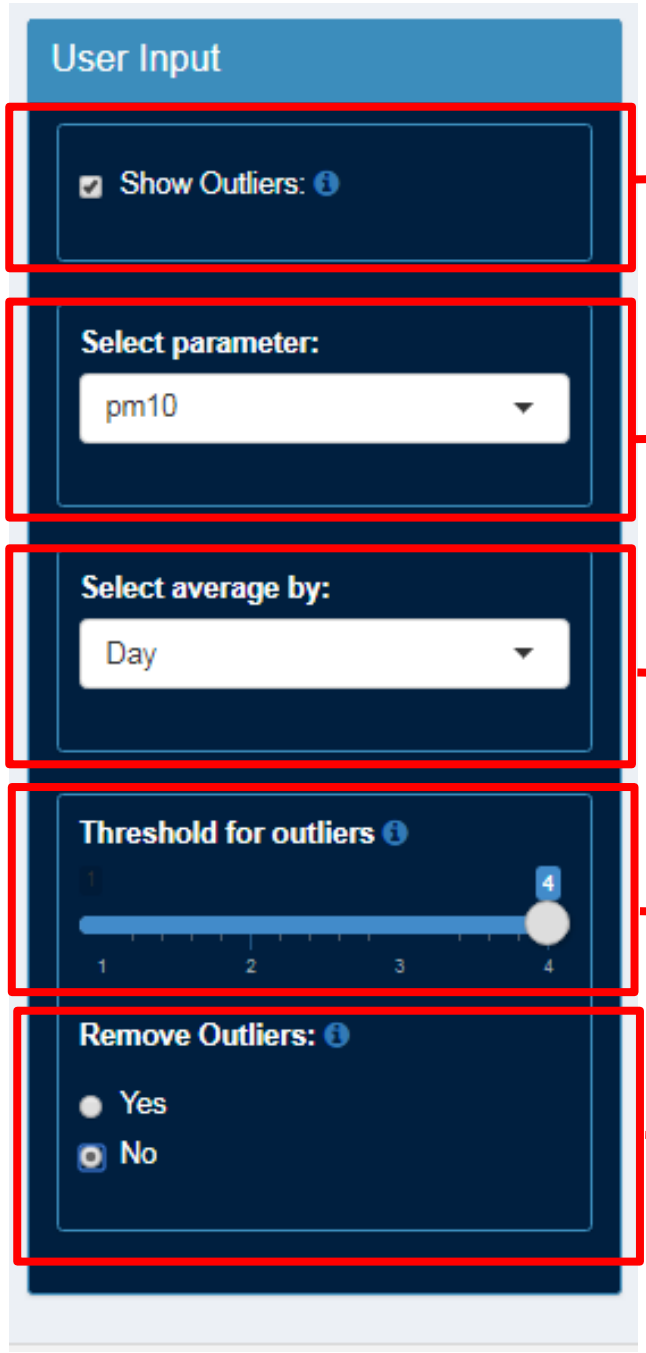
Outlier for every parameter is individually calculated on the basis of the average and standard deviation calculation of that parameter. Mean value and the standard deviation of each parameter is calculated from the data,

Mean = Sum of parameter values/ Total number of parameter values
SD is calculated by the formula below:
x = Parameter values
x = Mean
n = Total number of parameter values

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$



Outlier values are the values which are : less the (Mean - SD) and greater than (Mean + SD)



Show outliers – This option will show outlier values in a different colour on the graph.

Select parameter – You can select the parameter for which you to analyse the data.

Select average by : This option will tell you choose on how the mean value of a parameter you want to calculate. It will then average out the data on your selected option like 'Day', 'Month' or 'Year'.

Sigma – This will define the variation within which you want to see the outliers from the mean value. Higher sigma will give you less outlier values.

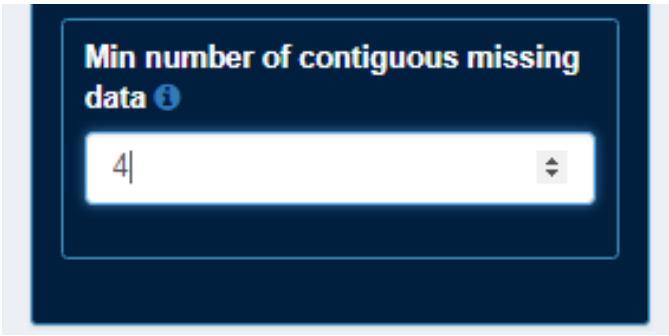
Remove outliers – You can select whether you want to remove outlier values from your dataset or not.

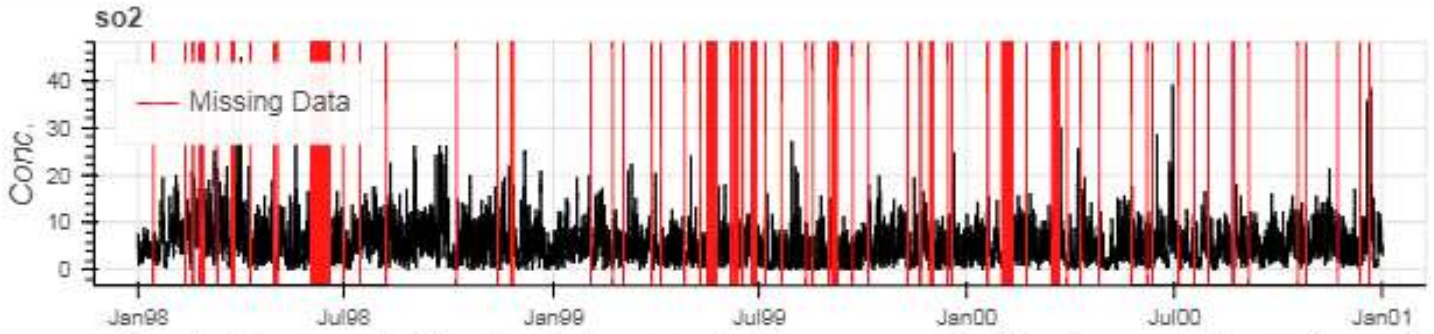
Missing Data Analysis

Missing data, or **Missing values**, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

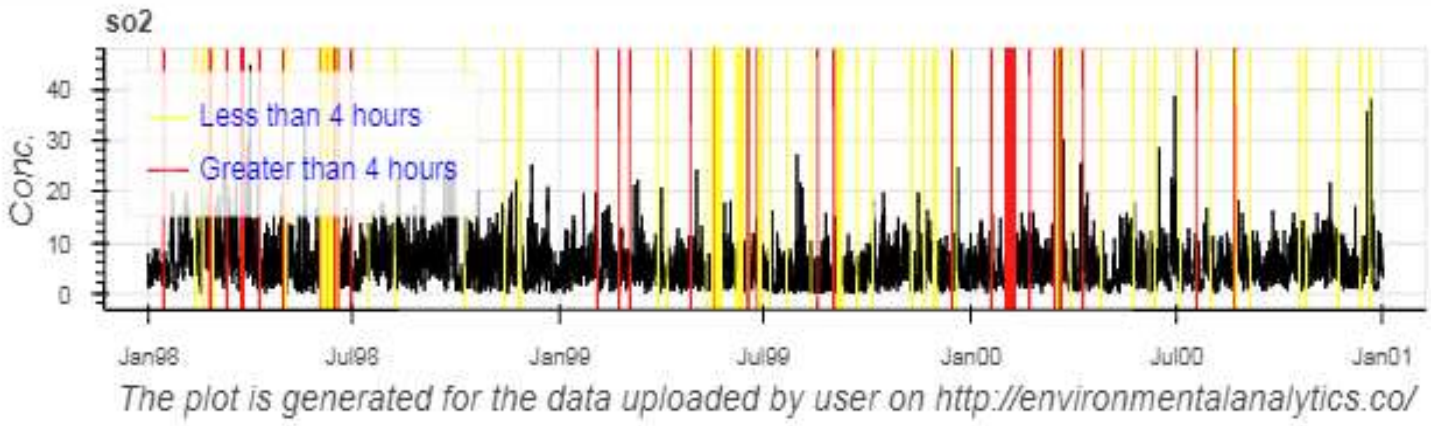
In **Air Quality Management,** missing data analysis tells us the efficiency of the instrument. If there is a missing data point, It means that the data was not recorded for that hour, which directly tells us for the instances in which the instrument was working or not.

Time series, this graph shows you the time series analysis of the missing data in your dataset, the 'Red' lines represent the exact missing data point on the given date and time in your dataset. The 'Yellow' lines represent the exact missing data point below the below the 'Contiguous value'.





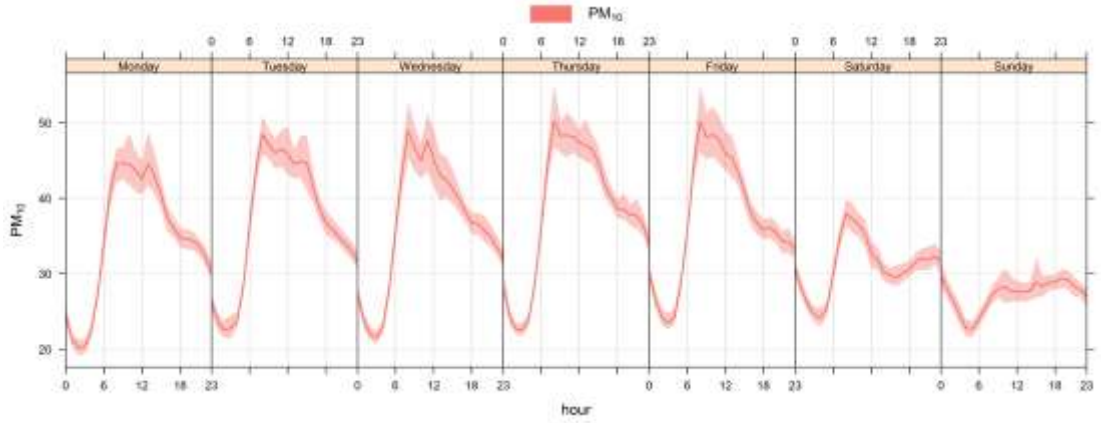
The purpose of setting 'Contiguous value' is for clarity: with long time series it is difficult to see where individual missing hours are. Setting it to 4, for example would show where there are at least 4 hours of continuous missing data.



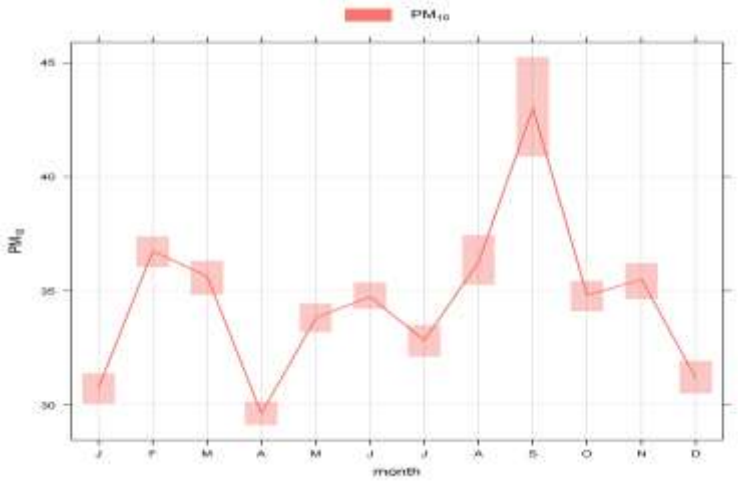
Temporal Variation

In air pollution, the variation of a parameter by time of day and day of week can reveal useful information concerning the likely sources. For example, road vehicle emissions tend to follow very regular patterns both on a daily and weekly basis. By contrast some industrial emissions or parameters from natural sources (e.g. sea salt aerosol) may well have very different patterns. There are four plots generated : day of the week variation, mean hour of day variation and a combined hour of day - day of week plot and a monthly plot. Also shown on the plots is the 95 % confidence interval in the mean. These uncertainty limits can be helpful when trying to determine whether one candidate source is different from another.

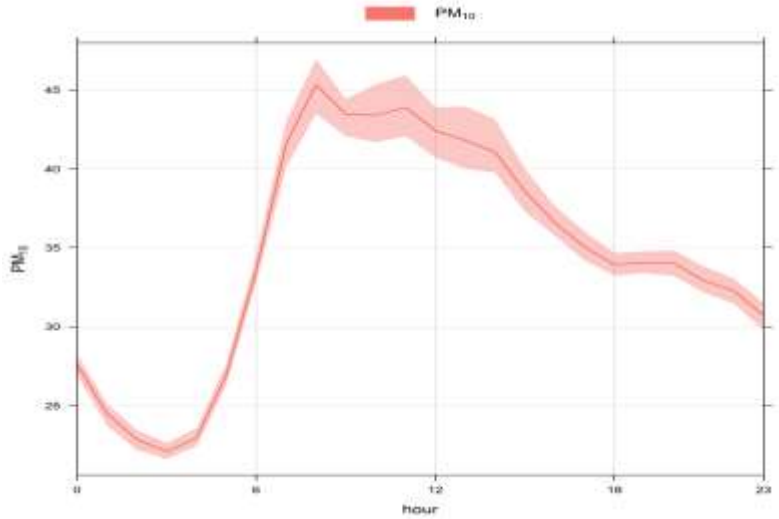
WEEKDAY AND HOURS PLOT



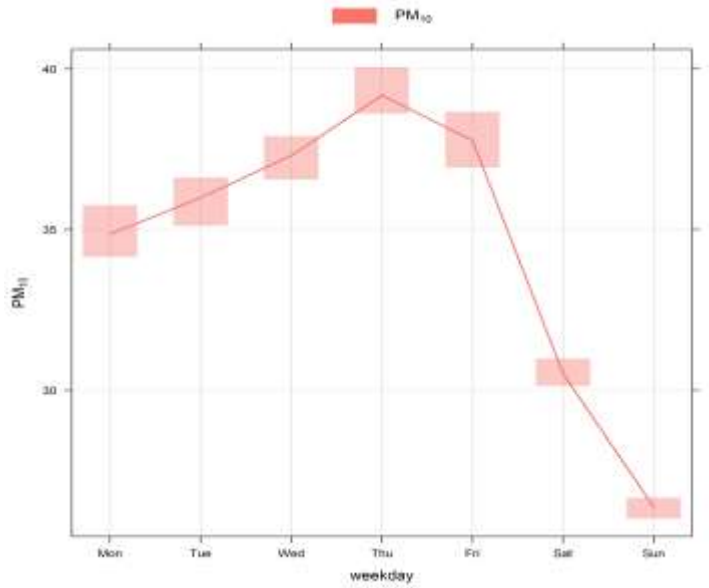
MONTHLY PLOT



DIURNALPLOT



WEEKDAY PLOT



What is Correlation?

Degree and type of relationship between any two or more quantities (variables) in which they vary together over a period; for example, variation in the level of expenditure or savings with variation in the level of income. A positive correlation exists where the high values of one variable are associated with the high values of the other variable(s). A 'negative correlation' means association of high values of one with the low values of the other(s). Correlation can vary from +1 to -1. Values close to +1 indicate a high-degree of positive correlation, and values close to -1 indicate a high degree of negative correlation.

Examples of the 'colour' visualization method

Visualization method

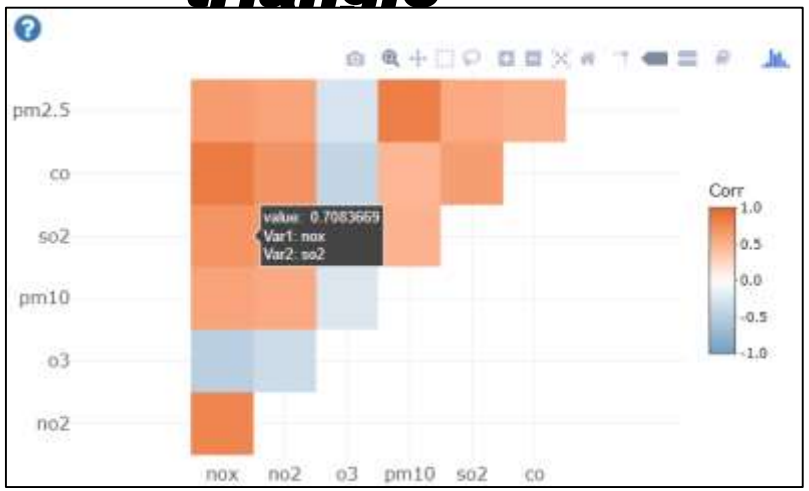
'Number' shows the correlation coefficient values. 'Colour' shows the correlation matrix in shades of different colours with the colour code depending on the extent of correlation

Choose Layout

'Upper triangle' displays the upper triangular matrix.
'Full' displays the full correlation matrix.

Full

Upper triangle



Trend Analysis

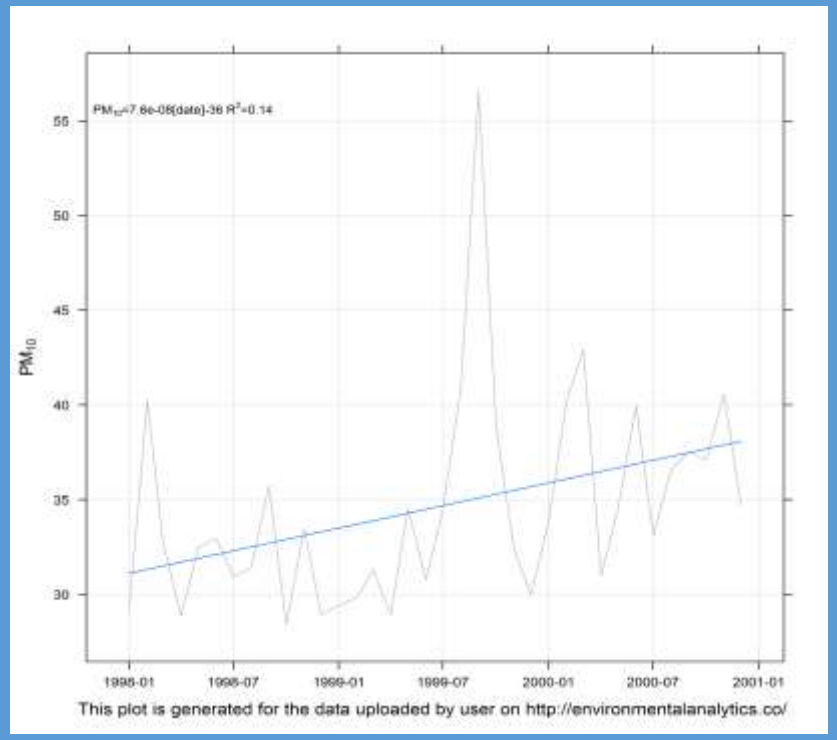
Calculating trends for air parameters is one of the most important and common tasks. Sometimes it is useful to have a general idea about how concentrations might have changed.

Theil-Sen estimate of the slope is the median of all these slopes. The advantage of the using the Theil-Sen estimator is that it tends to yield accurate confidence intervals even with non-normal data and heteroscedasticity (non-constant error variance). It is also resistant to outliers - both characteristics can be important in air pollution.

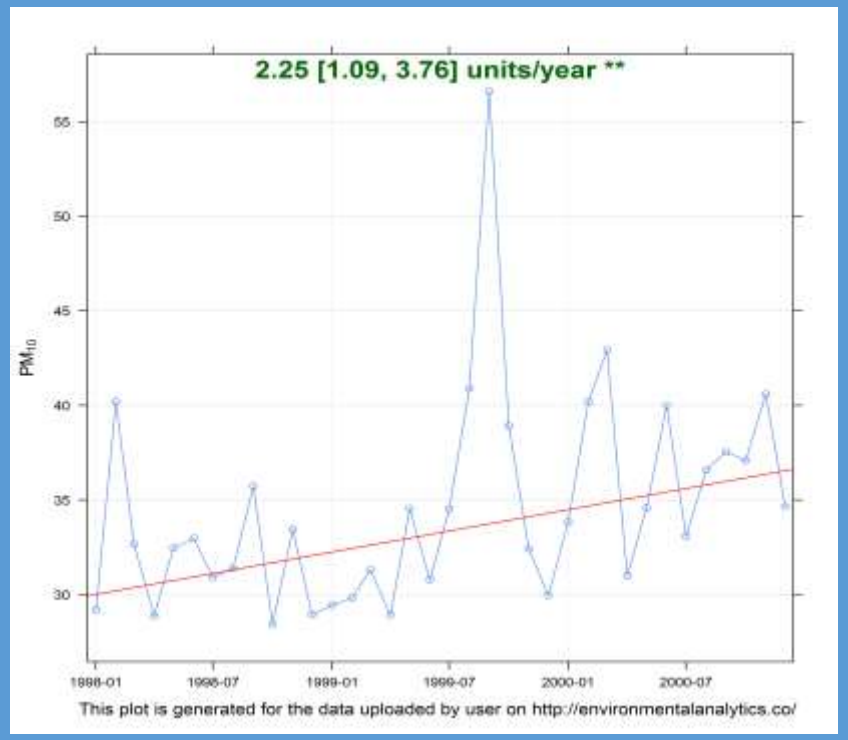
Scatter plots with conditioning and three main approaches: conventional scatter plot, hexagonal binning and kernel density estimates. It is flexible enough to consider lots of conditioning variables and takes care of fitting smooth or linear relationships to the data.

Examples of trend plots using different methods

Scatter plot method



Theil-Sen method

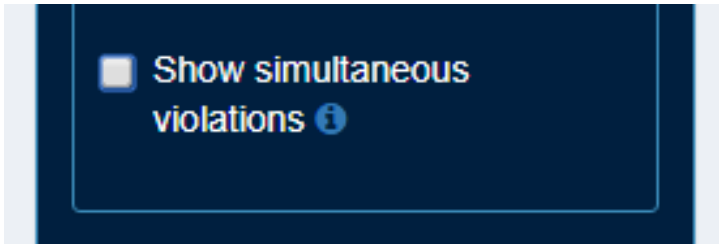


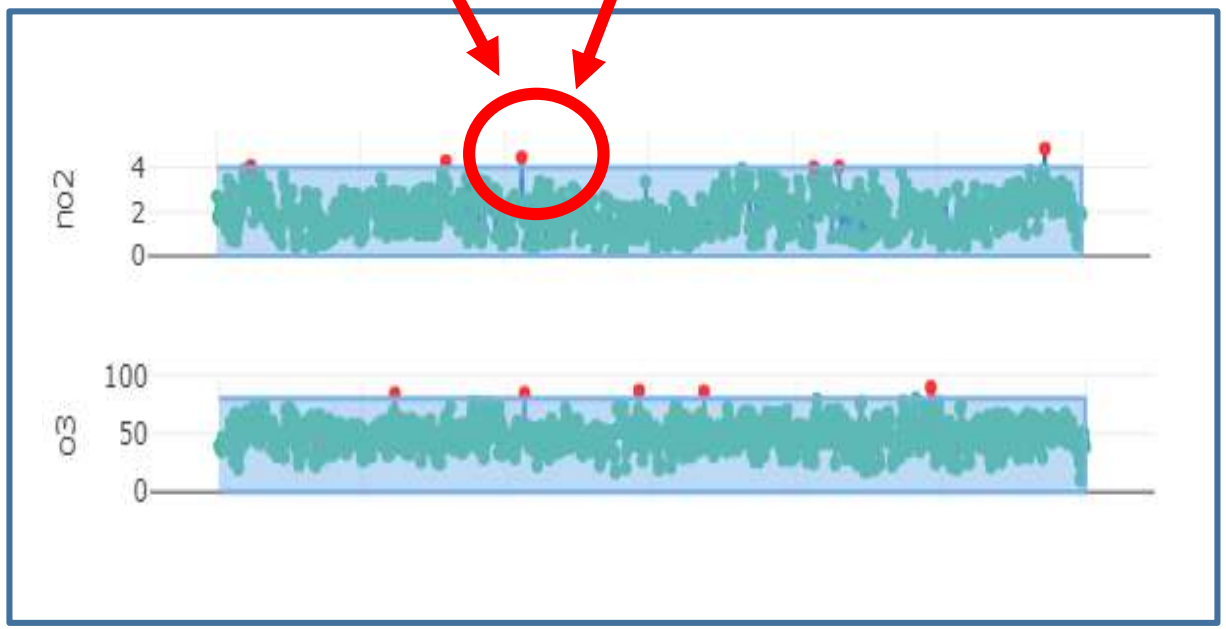
Why is Violation Analysis important?

Violated data points are those which lie above the defined average standard values. The standard range is different for every parameter. The standards considered for our calculation are defined by the Central Pollution Control Board (CPCB) standards.

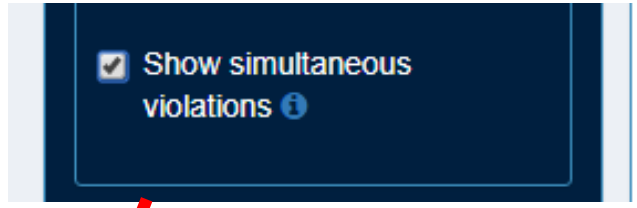
Simultaneous violations : A simultaneous violation option is also given to check for simultaneous violations of the selected parameters

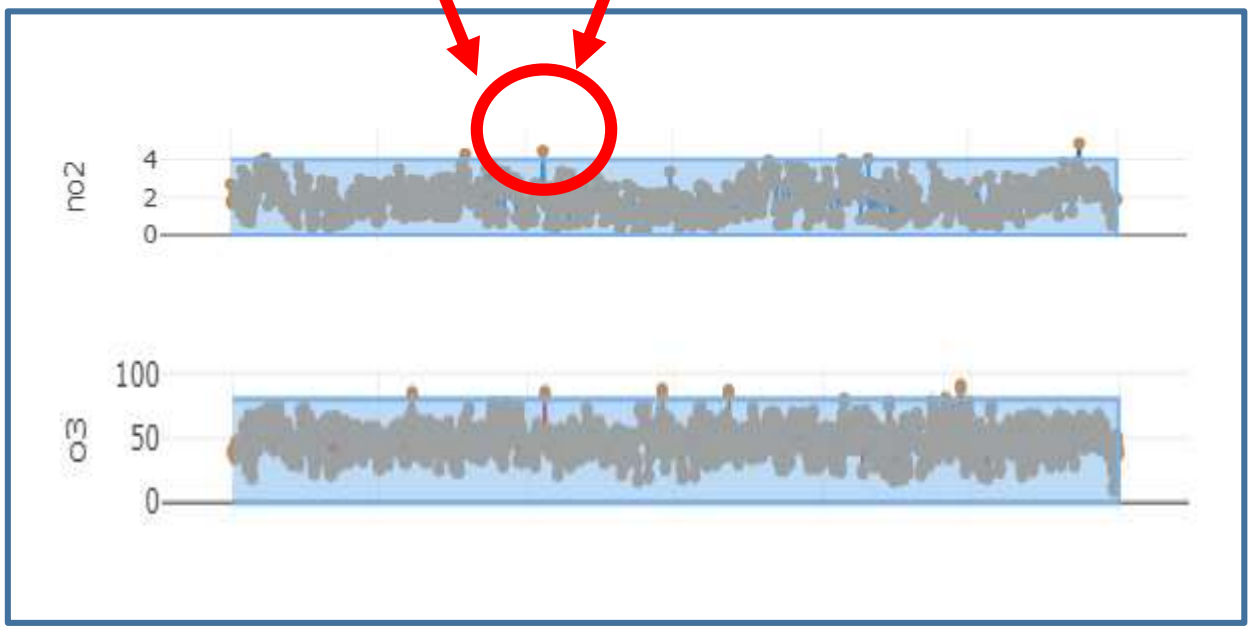
When simultaneous violations are not selected ,Violations are checked for each variable separately





However ,It is not the same case in simultaneous violations, If any other variable is not violated, all the other violations are crossed out.



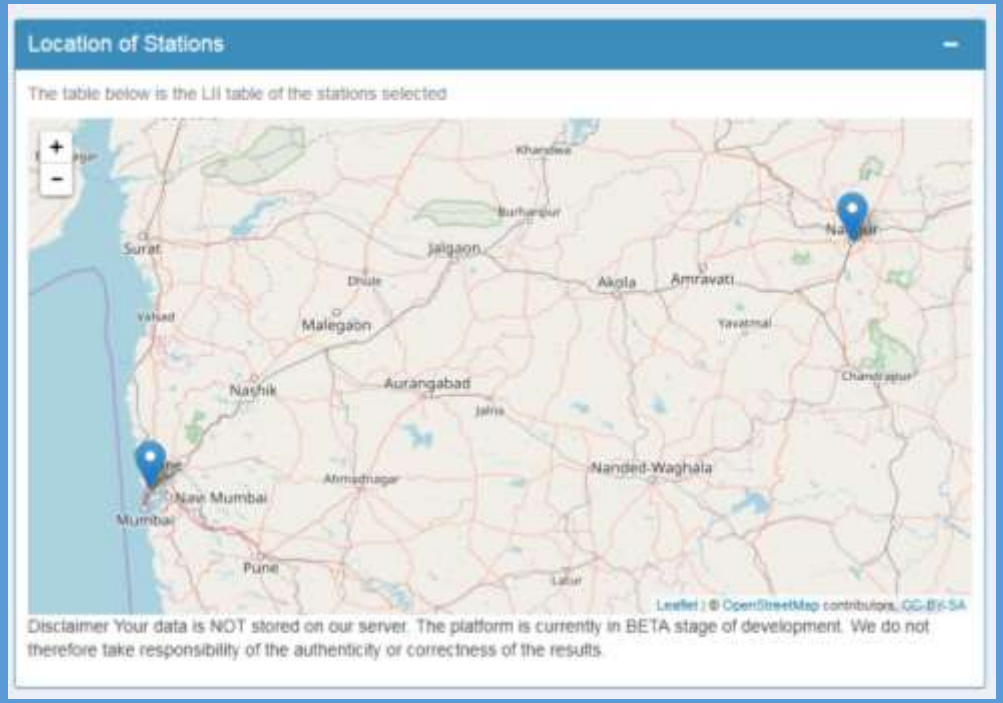


Location Importance Index (LII)

Location Importance Index or LII is a technique to measure significance of location of monitoring site by comparison of two or more sites on the performance of a

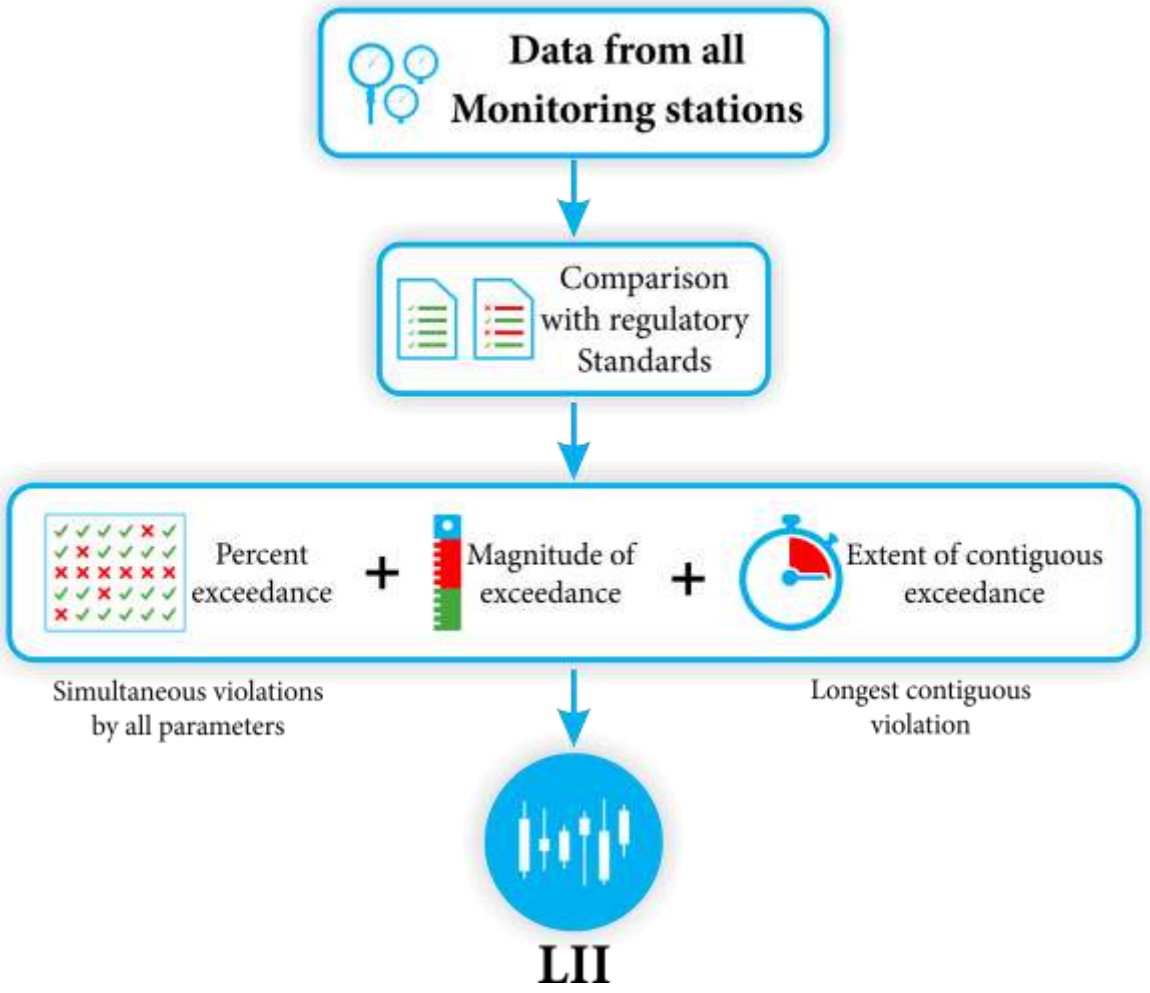
1) single parameter at a time and
2) two or more parameters simultaneously.

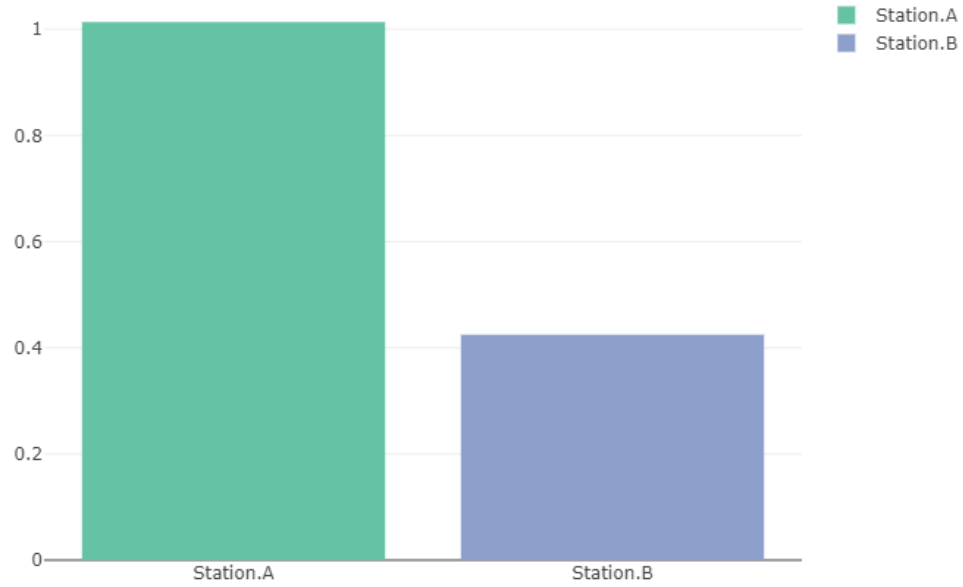
Missing percentage are only calculated for Contiguous exceedance**. Contiguous exceedance** is the number of continuous exceeded data points.



A map depicting the location of stations on the LII page from the dashboard.

How is LII calculated?

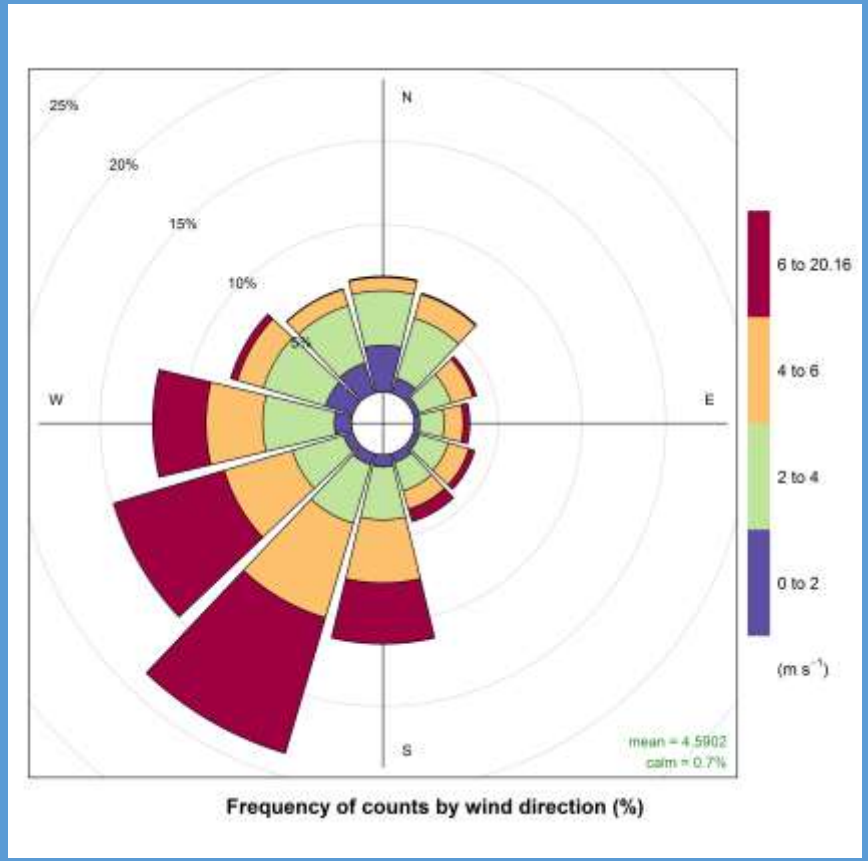




An output of comparison of LII values graphically on the dashboard

WIND ROSE

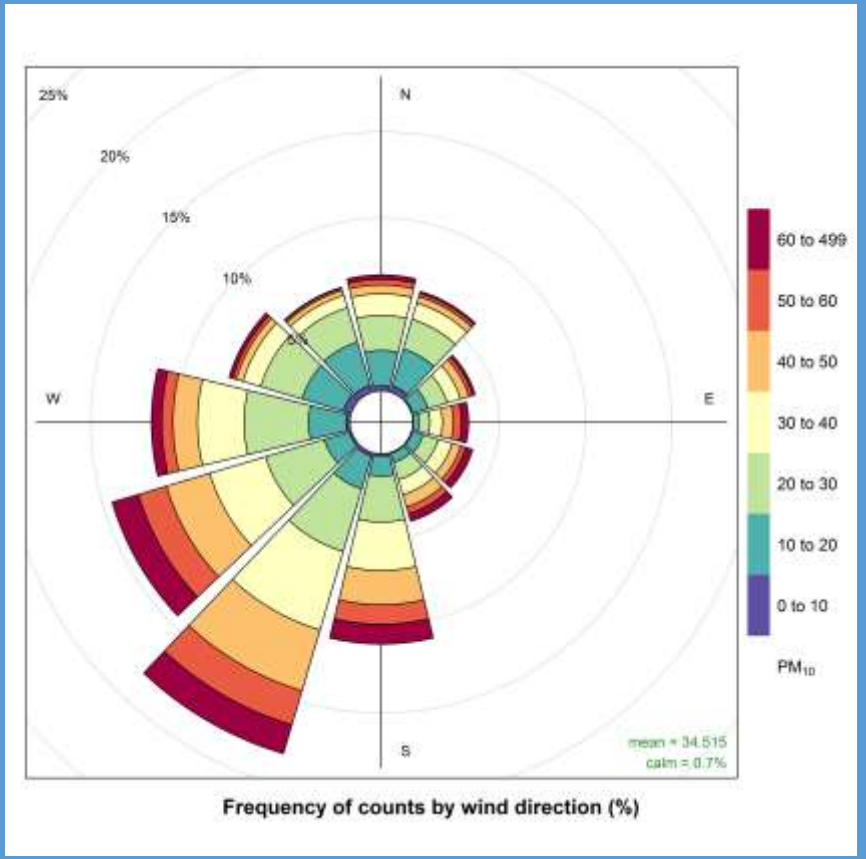
The wind rose is a very useful way of summarising meteorological data. It is particularly useful for showing how wind speed and wind direction conditions vary by year. Plots are generated for wind roses in a variety of ways: summarising all available wind speed and wind direction data, plotting individual wind roses by year, and also by month. The latter is useful for considering how meteorological conditions vary by season.



Example of Wind Rose plot generated on the dashboard

POLLUTION ROSE

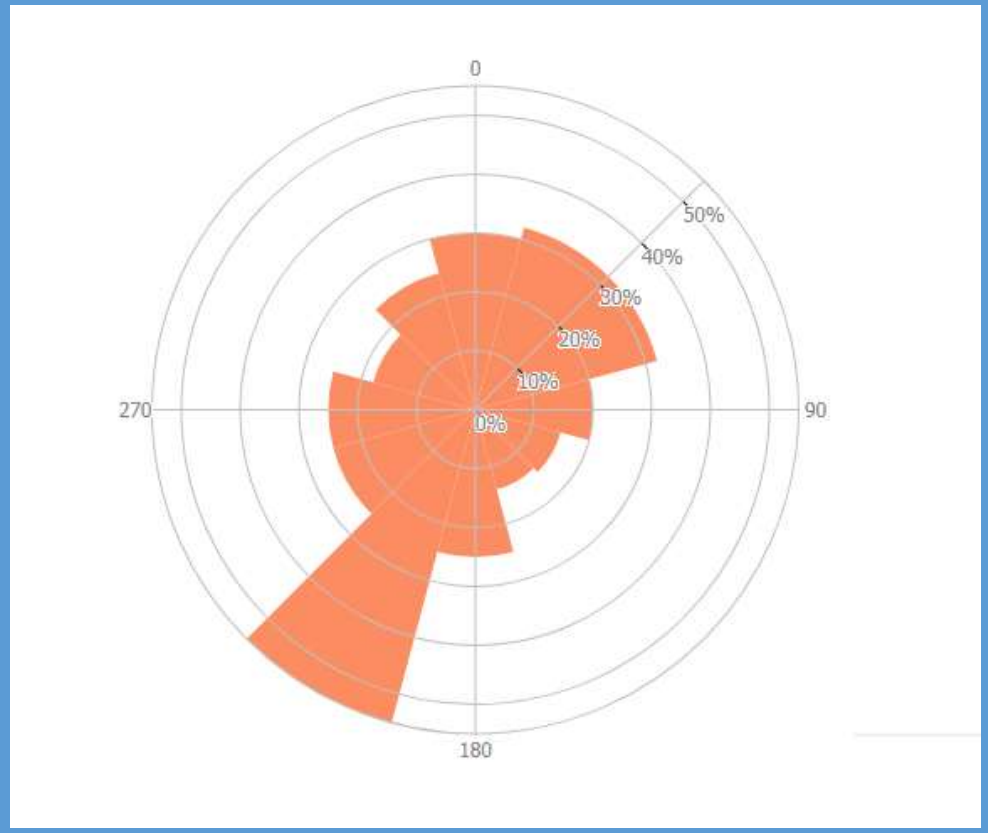
Pollution Rose is a variant of Wind Rose that is useful for considering parameter concentrations by wind direction, or more specifically the percentage time the concentration is in a particular range. Pollution Rose can also usefully be used to show which wind directions dominate the overall concentrations.



Example of Pollution Rose plot generated on the dashboard

PERSISTENT WIND ROSE

This plot gives information of maximum duration of continuous wind direction.
This plot helps to detect flaws in the measuring instrument. For example, if a particular wind bin shows a value of 70 hours for a particular direction, it can signal to faulty instruments as the data looks suspicious.

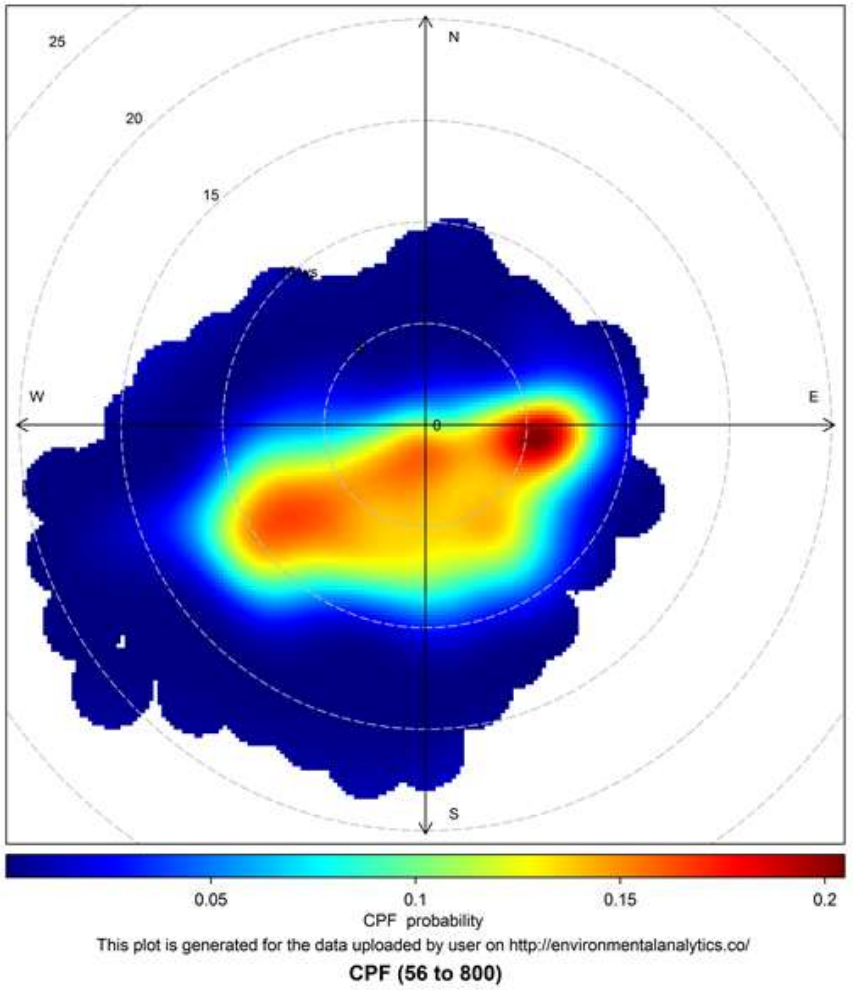


Example of Persistent Wind Rose plot generated on the dashboard

POLAR PLOT

The Polar Plot is bivariate. It is designed to illuminate the effect of both wind speed and wind direction on the movement of air pollution. This plot shows a continuous series of tones that reflect the mean concentrations by changing hue. The colour and intensity at a given location on the plot reflects the concentration.

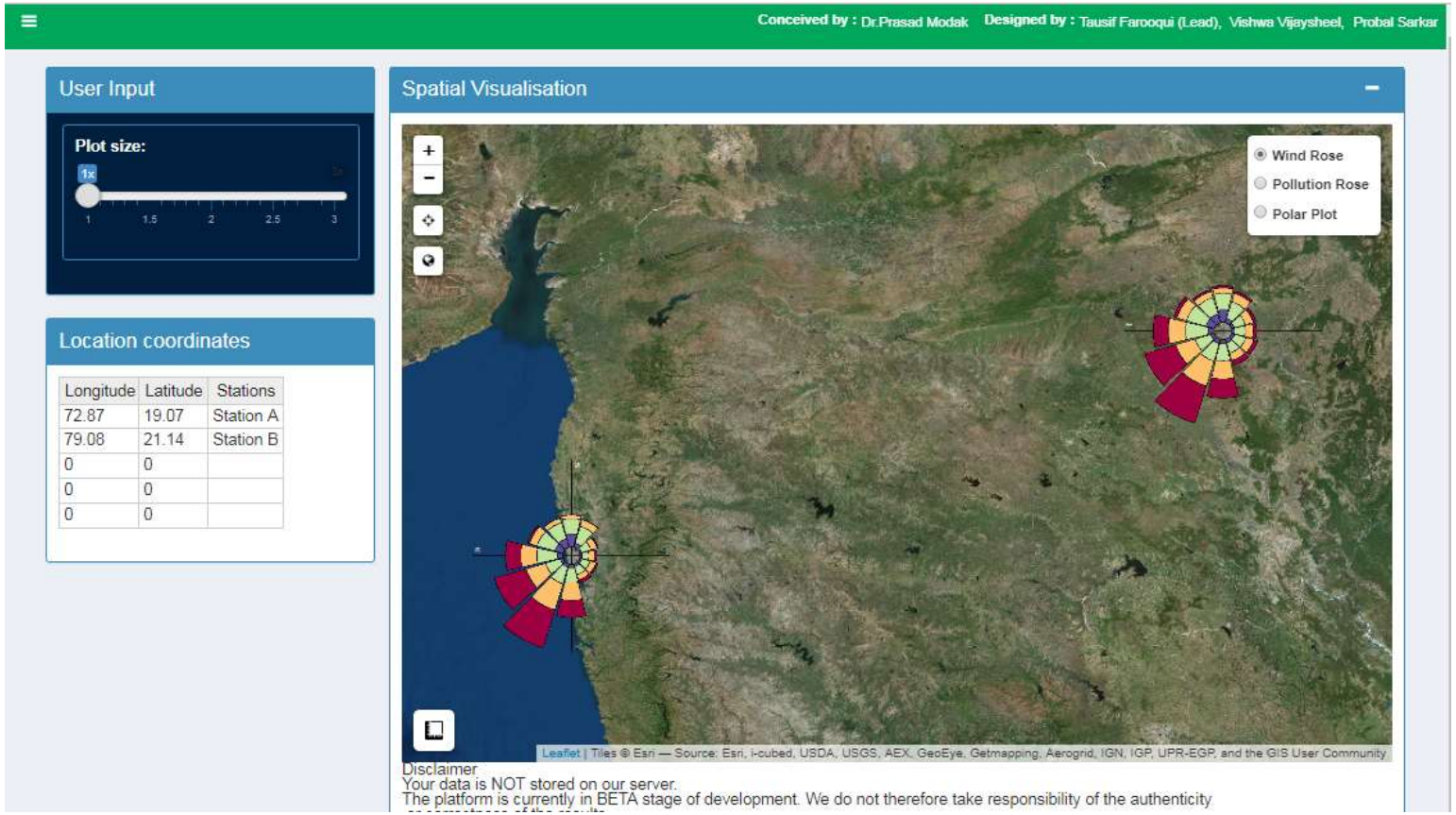
Polar Plot illustrates the localized concentration mean based on wind direction and speed (how much is present at a specific regions of the plot based on the wind direction and speed and other factors)



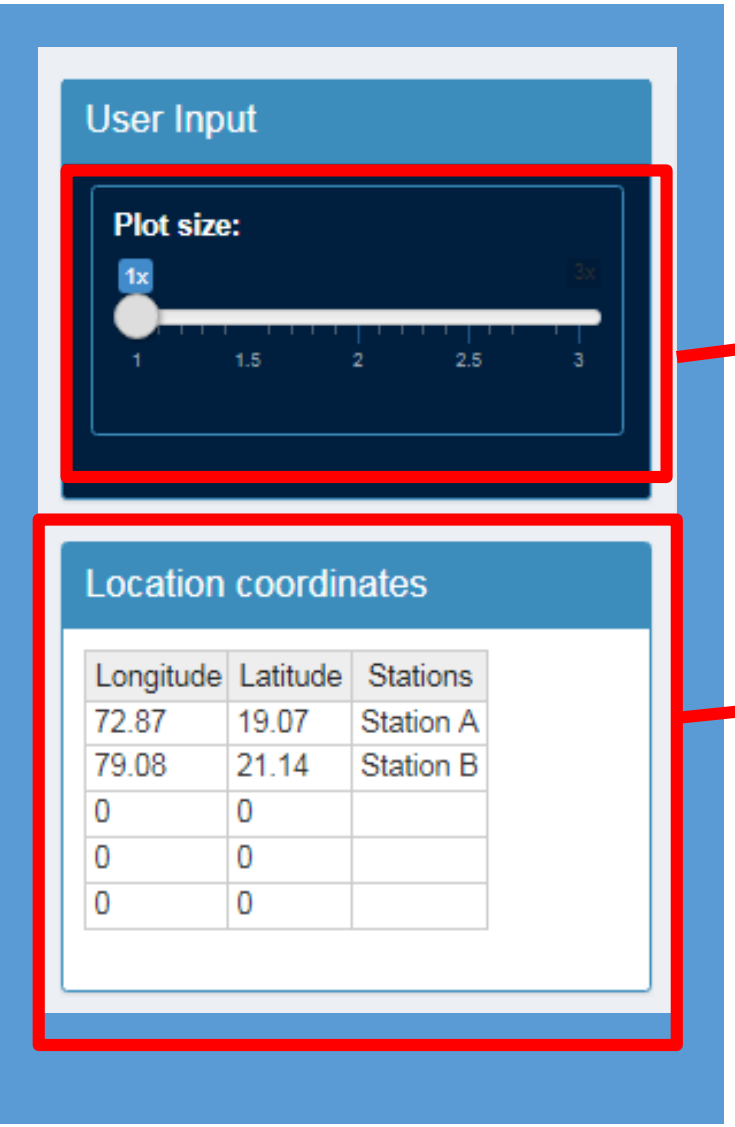
Example of **polar plot** generated on the dashboard

Map Visualizations

This tab allows to visualises the Wind rose, Pollution rose and the Polar plots for all the stations in maps. The plots generated will automatically get fetched and one can control the visibility of the plots in the control panel.



A view of the **Map Visualization** page on the dashboard



Plot size – Will let you choose the size of the plots to be made on the maps. The plots made will be the generated plots on the made in the respective tabs.

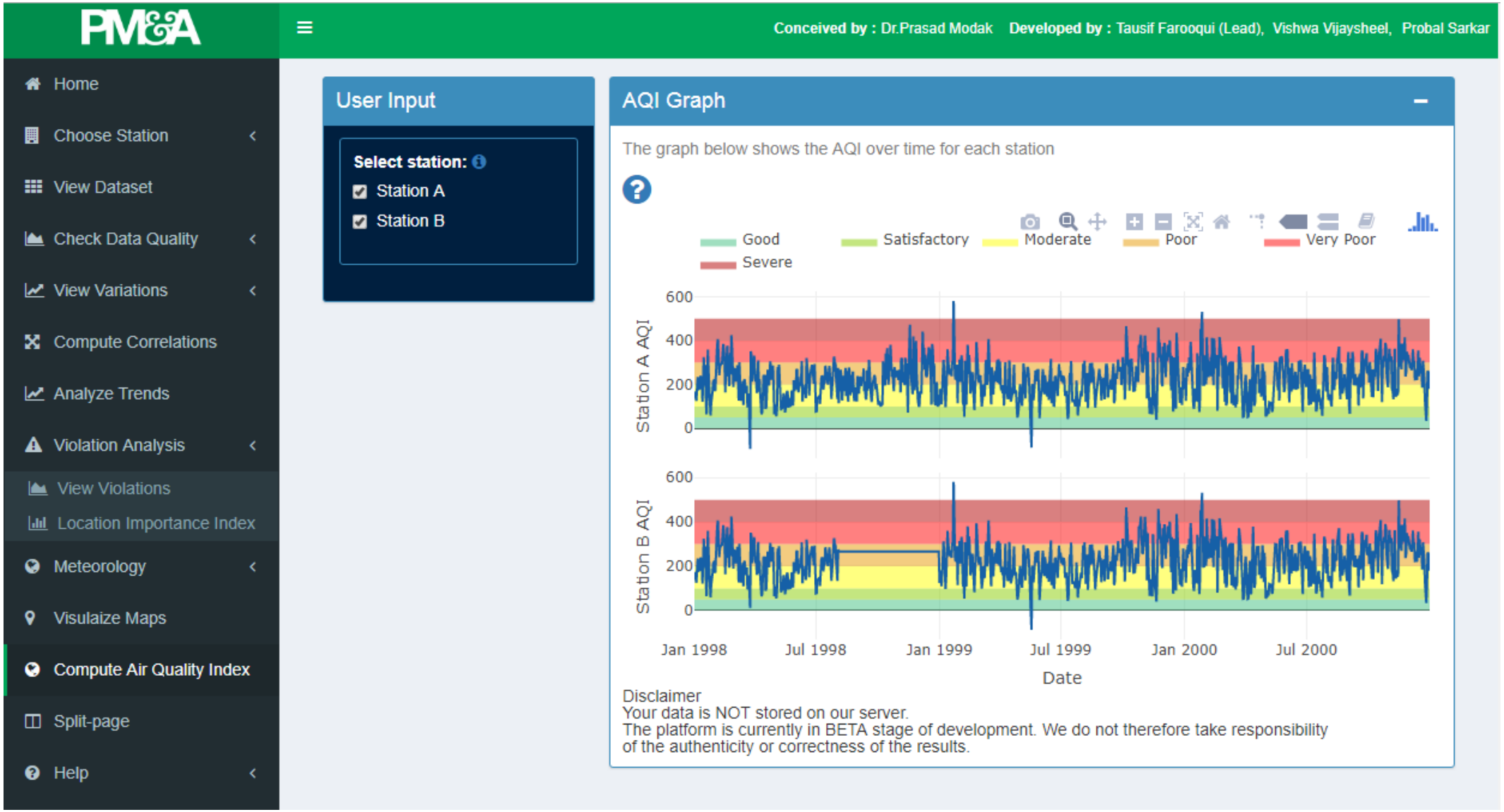
Location coordinates– You must specify the coordinates of station when analysing multiple stations.

Air Quality Index

Air Quality Index is a tool for effective communication of air quality status to people in terms, which are easy to understand. It transforms complex air quality data of various pollutants into a single number (index value), nomenclature and colour.

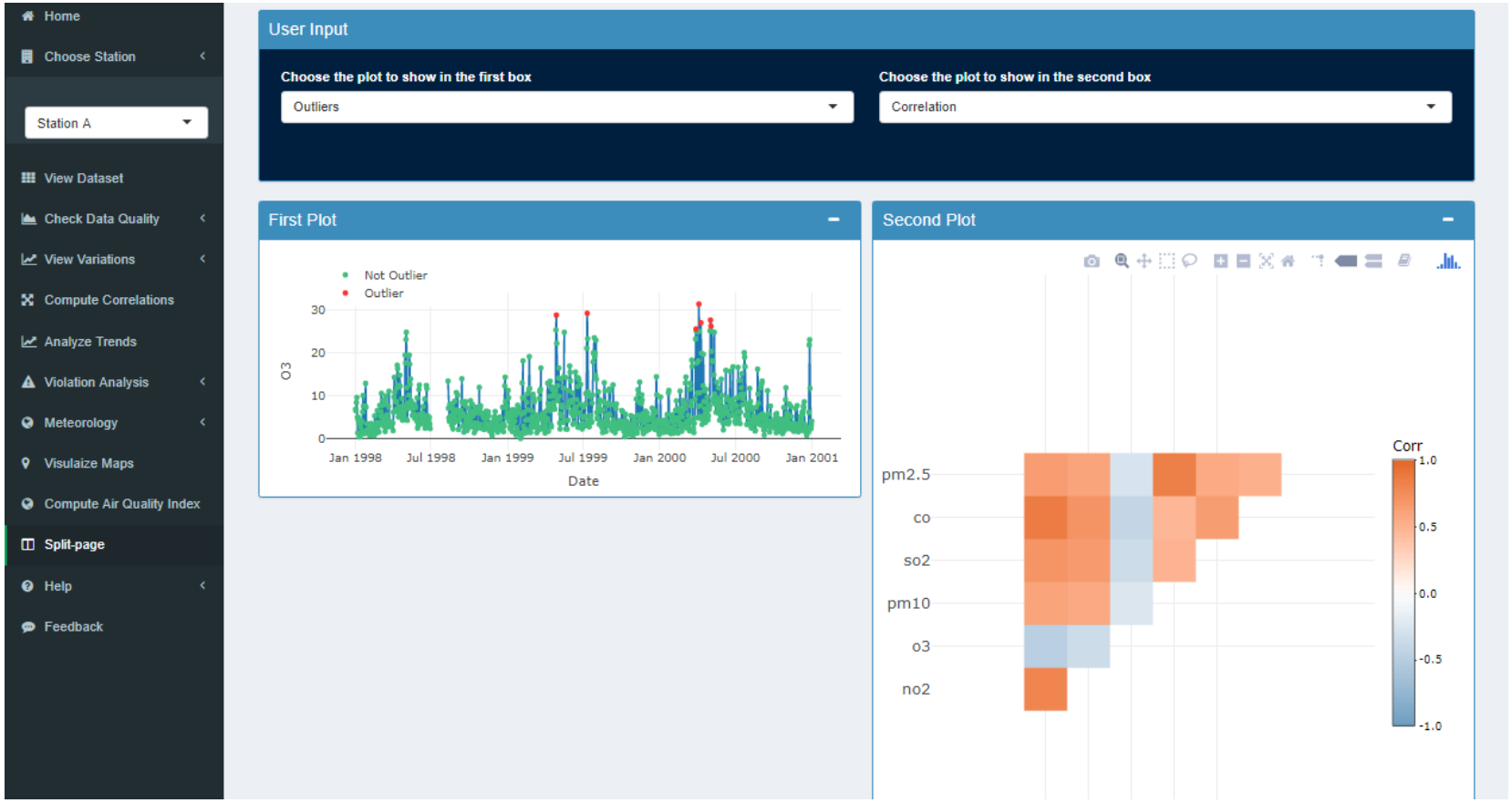
Based on the measured ambient concentrations of a pollutant, sub-index is calculated, which is a linear function of concentration (e.g. the sub-index for PM2.5 will be 51 at concentration 31 µg/m3 , 100 at concentration 60 µg/m3 , and 75 at concentration of 45 µg/m3).

A view of the **Air Quality Index plots** page on the dashboard



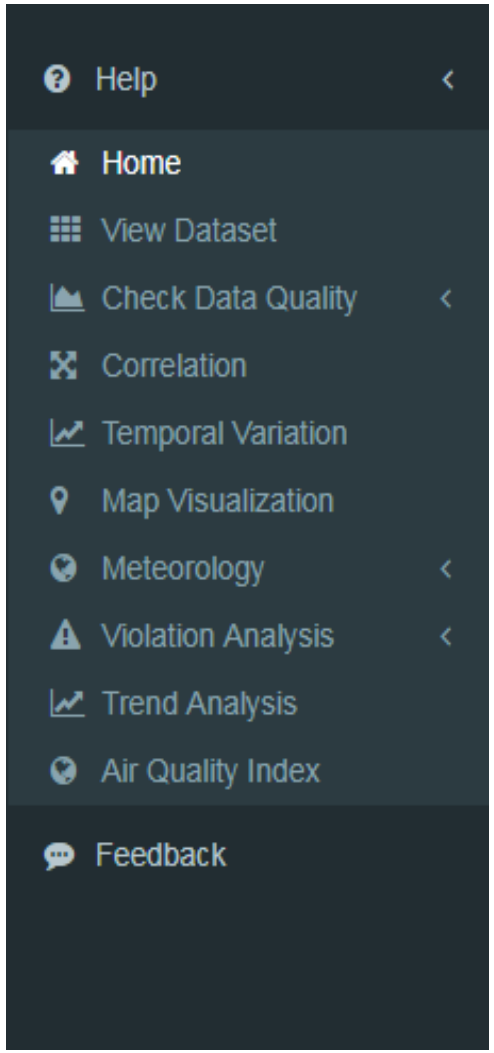
Split Page

This tab will allow you to see and compare two plot outputs on the same page. The plots that were generated on the subsequent pages earlier are list in the list and given as options to choose from.

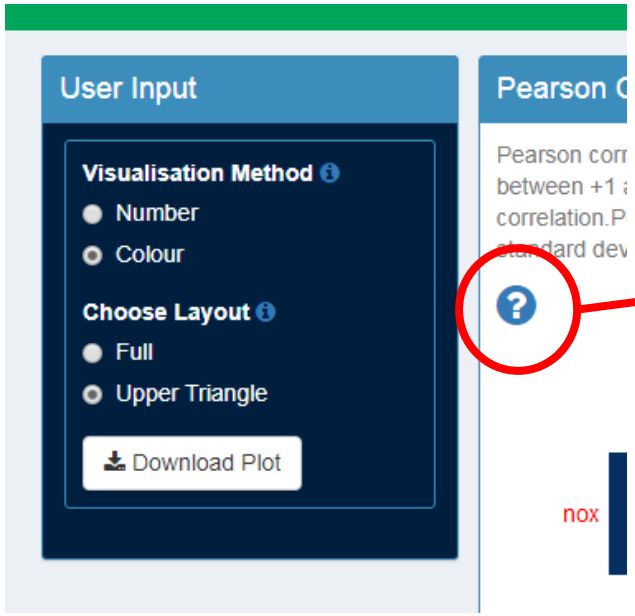


A view of the **Map Visualization** page on the dashboard

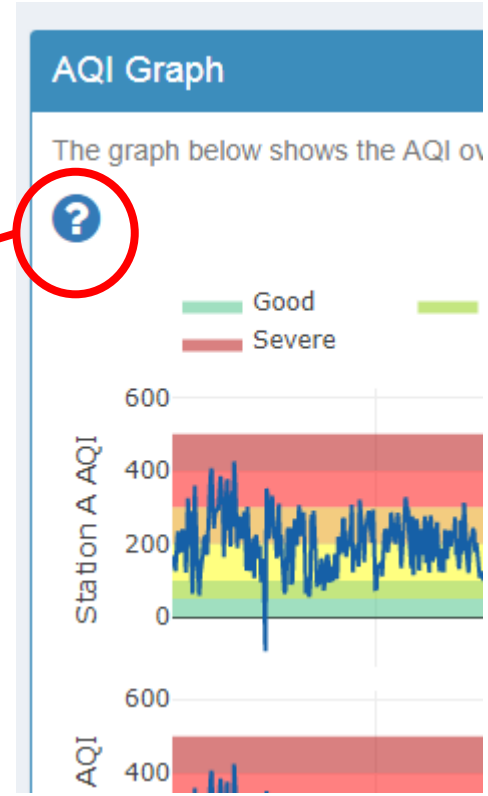
Help Page



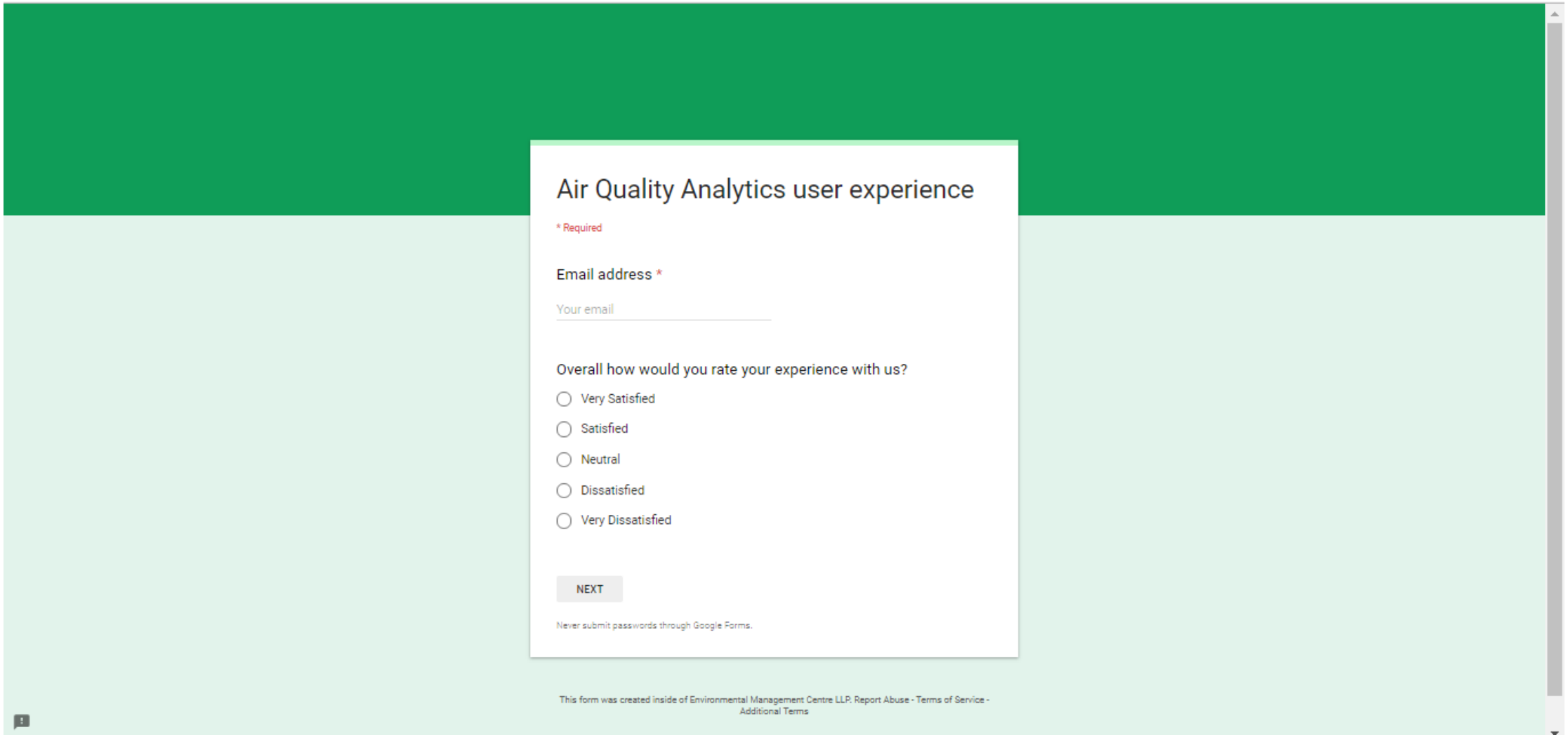
Help page – A list of help pages are given for all the various list of tools in the dashboard. You can directly access the help page of a particular tool under the **'Help Tab'**



You can access the help page of a particular tool under the by clicking the **help button** on that particular page as well



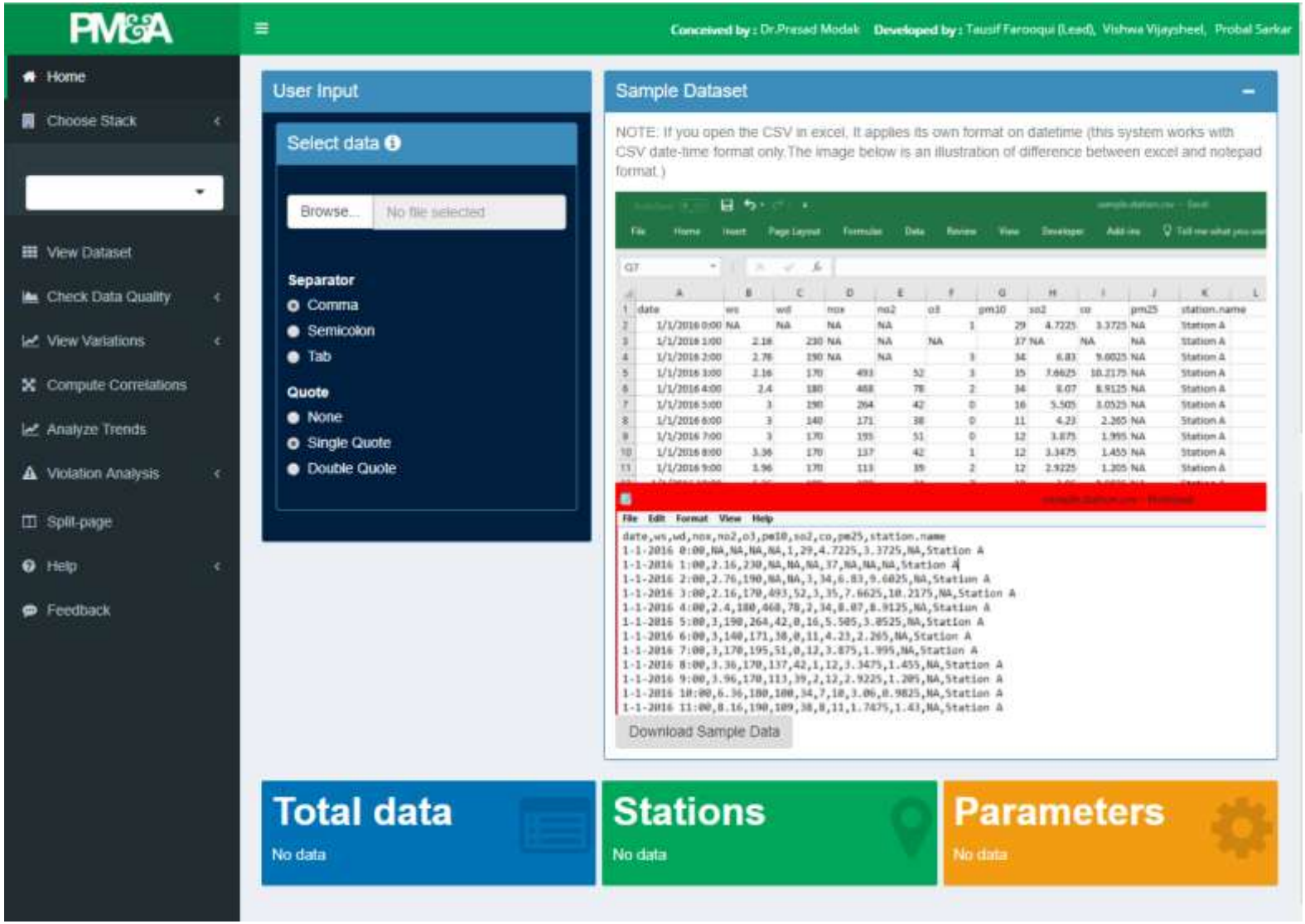
Your may also give a feedback on how your user experience



Air Quality Analytics Dashboard –STACK EMSSIONS

This is how the opening page of the Stack Emissions dashboard looks like.

The Stack Emissions Dashboard offers the same set of functions as there are for the Ambient Air Quality dashboard.



Real-Time Air Quality Analytics Dashboard

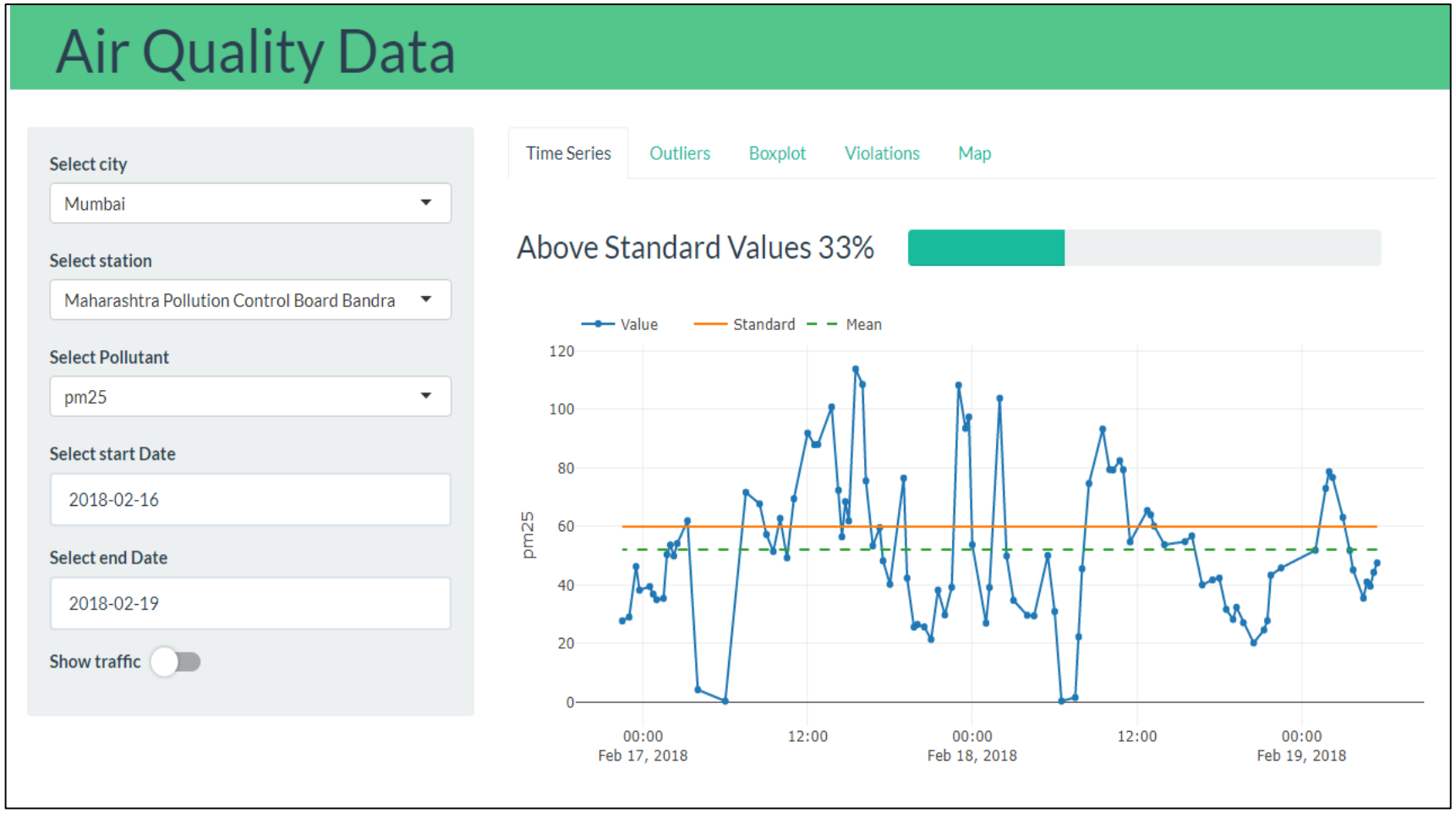
The Dashboard visualizes real-time Air Quality data showing you the air quality levels from the current date till the last 3 days. The dashboard fetches data from an API (https://docs.openaq.org/) and visual analysis is done.

The dashboard captures data on a real-time basis for various stations of Mumbai and Delhi and for a list of 4 pollutants i.e. pm25, pm10, so2 and o3. Time-Series analysis, Outlier Analysis and Violation Analysis are few exaples of the graphical methods to analyse the data.

Time Series

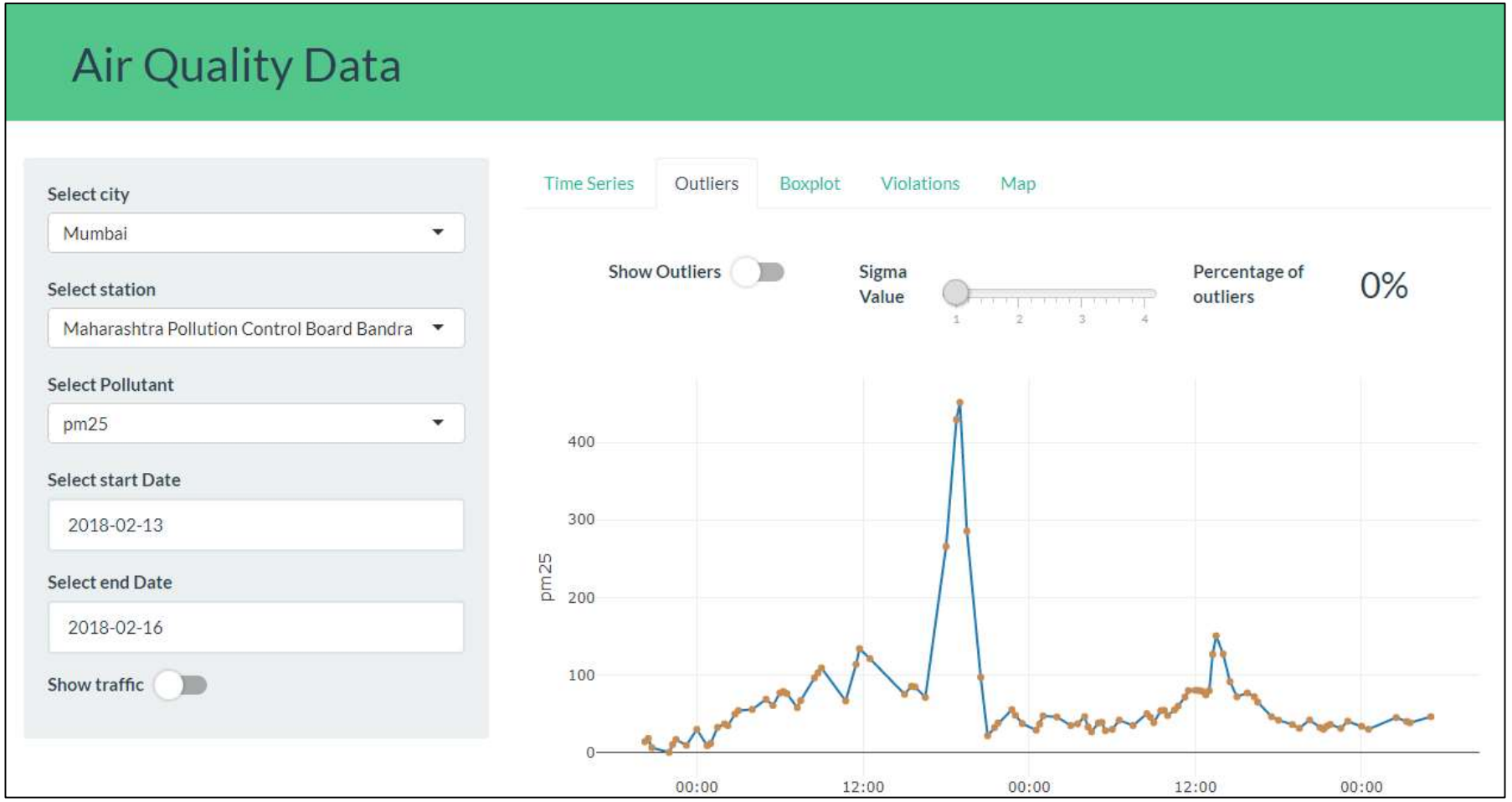
ASV (Above standard Values) – It tells you the percentage of data points that have violated the standard within the last 3 days.

The green dotted line and the yellow line show the ‘Mean’ and the ‘Standard Value’ respectively.



Outliers

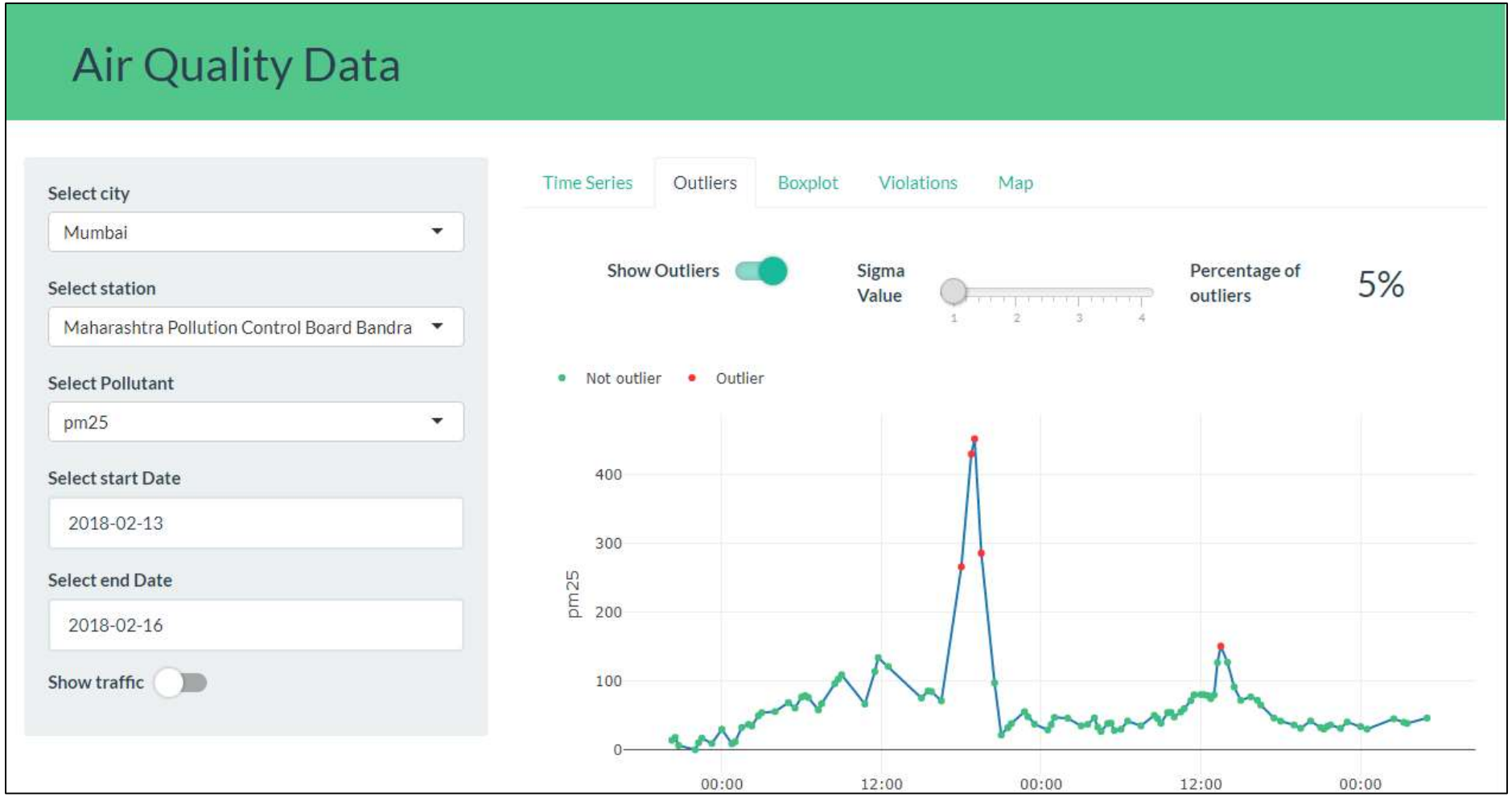
Outlier Analysis – It shows you the outlier values for the time series of the captured data. The outliers are recalculated for each new data point fetched from the server.



Percentage of Outliers

Sigma value can be varied to check for outliers and the percentage is calculated on the number of points that are outliers with the total number of points.

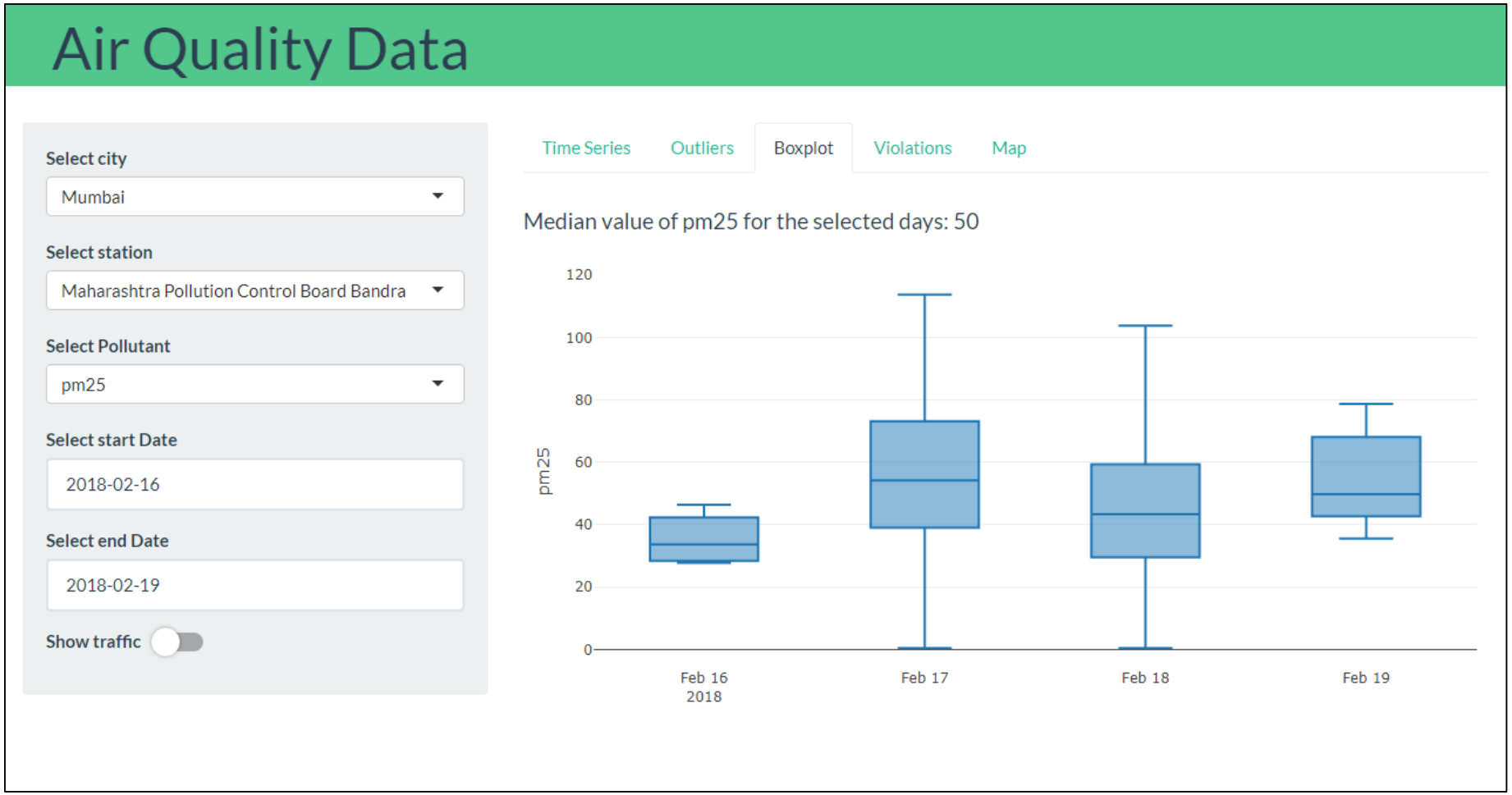
More the percentage of outliers more fluctuating the air quality levels will be.



Boxplots

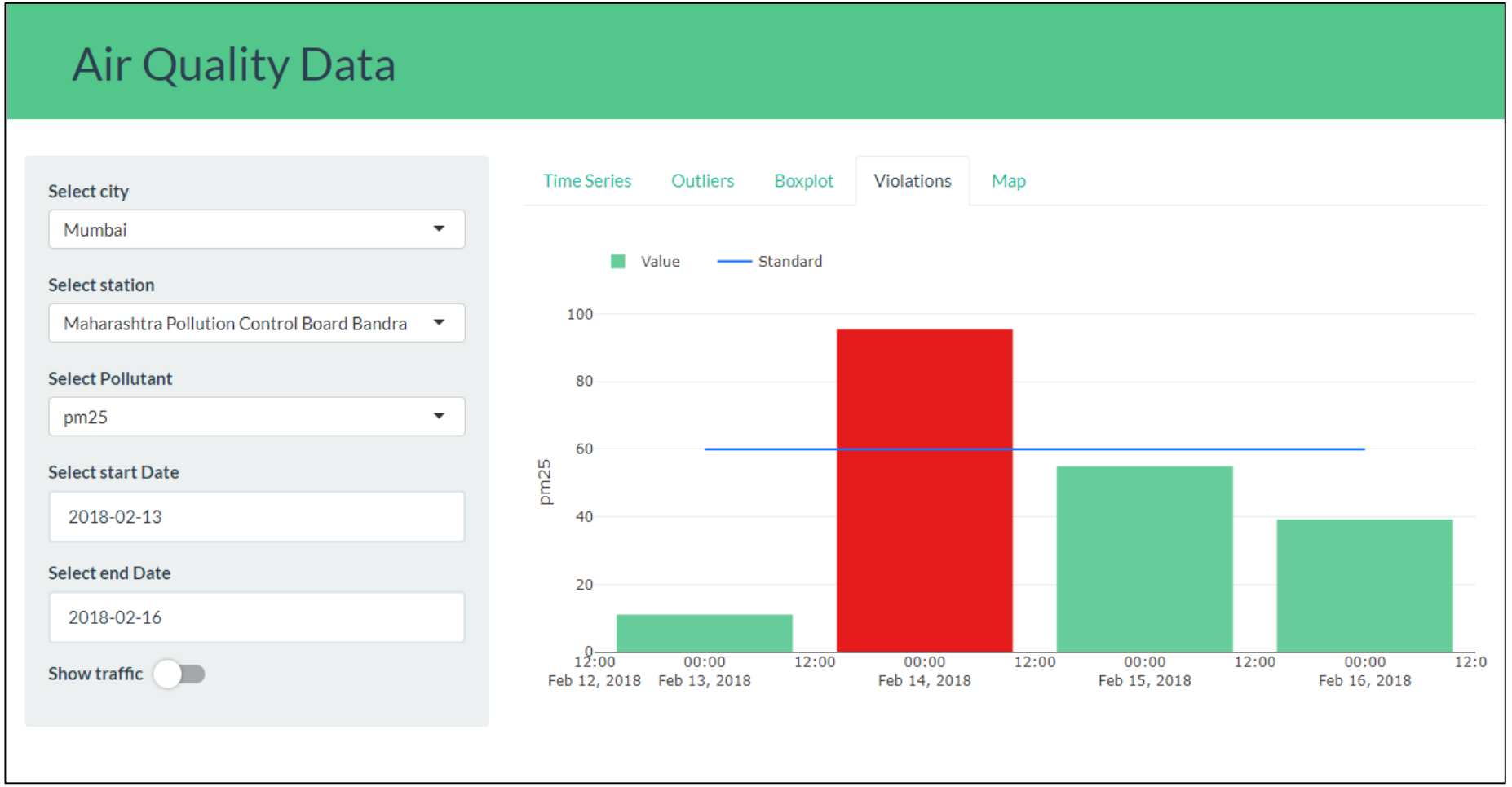
This page shows you the boxplot variation for the selected pollutant for each day.

A more comprehensive statistical analysis of the varying pollutant concentrations is done.



Violations

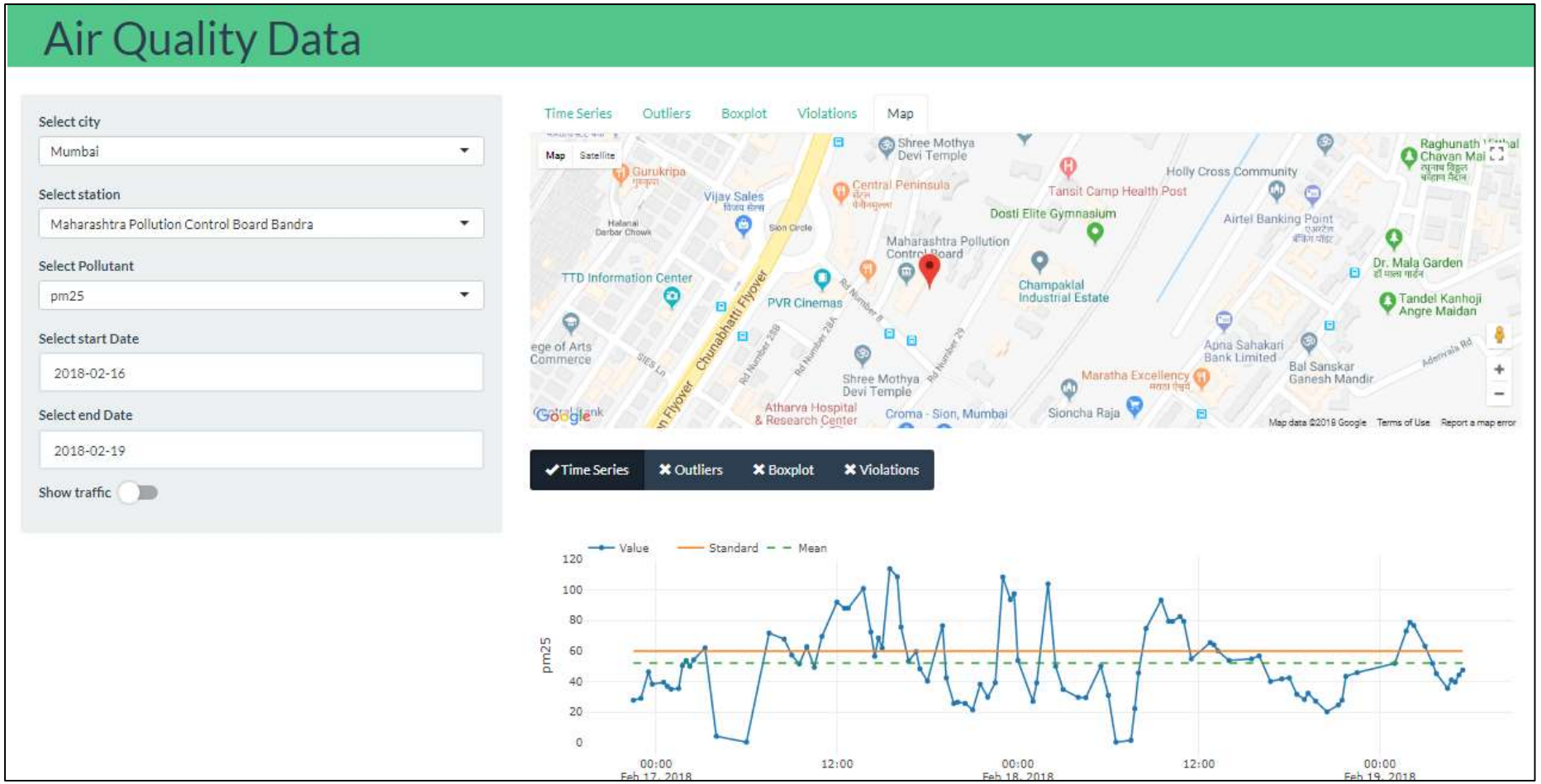
The violation graph calculates an average value of the particular pollutant for a particular day, ‘Height of the bar’ represents the value, compares it with the standard (Blue line) and tells you whether it is violated or not. ‘Red’ means violated, ‘Green’ means complaint.



Maps

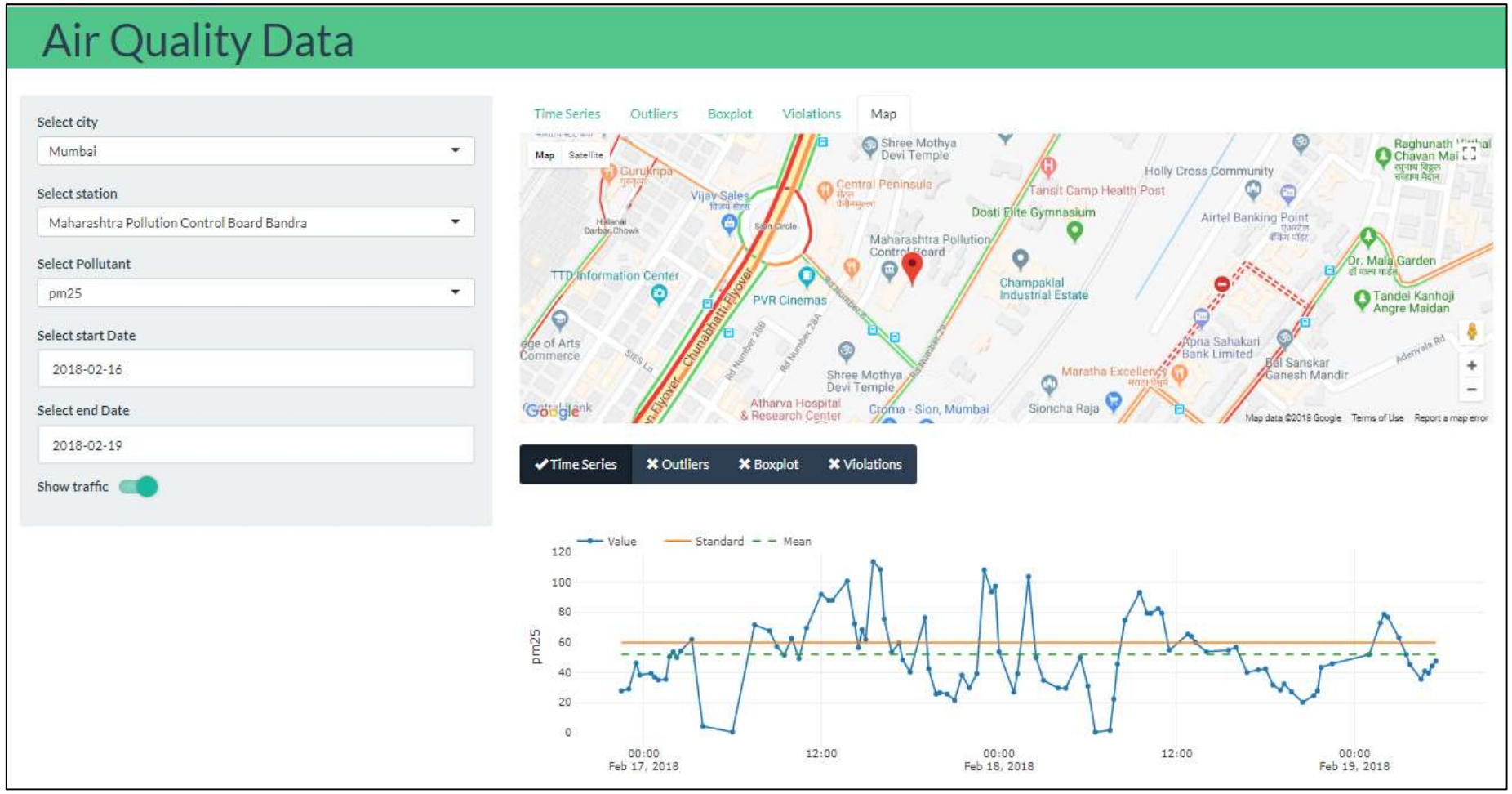
The dashboard also captures the location coordinates of the selected station and automatically plots it on a map.

There is also an option of viewing the other graphs with the maps for further analysis.



Real-Time Traffic

You can also see the Real-time traffic around the location of the station.
Instant traffic is shown at a particular point of time.



THANK YOU