

Being a data analysis and data science enthusiast, I decided to analyze the air quality data of my own country to find some underlying principles or patterns which might give me an insight into how severe the problem it is

THE APPROACH

The following data analysis is carried out in python

DATA PREPROCESSING

All data pre-processing are handled using Python. When dealing with air quality historical data, we need to replace all unknown value to zero or some meaningful value. Here, the original data use special characters to denote the value is invalid. For example,

means the value is detected as invalid by instruments.

- Means the value is detected as invalid by computer programs.

X means the value is detected as invalid by human beings.

Blank means the value is missing. For simplicity, all invalid data are currently replaced by 0.

DATASET

The dataset contains the following features :

Stn_code : Station code. A code is given to each station that recorded the data.

Sampling_date: The date when the data was recorded.

State: It represents the states whose air quality data is measured.

Location: It represents the city whose air quality data is measured.

Type: The type of area where the measurement was made.

So2: The amount of Sulphur Dioxide measured.

No2: The amount of Nitrogen Dioxide measured

Rspm: Respirable Suspended Particulate Matter measured.

Spm: Suspended Particulate Matter measured.

Location_monitoring_station: It indicates the location of the monitoring area.

Pm2_5: It represents the value of particulate matter measured.

Date: It represents the date of recording (It is a cleaner version of 'sampling_date' feature)

FEATURES PREDICTED WHILE TESTING AIR QUALITY

SO₂: Sulphur Dioxide is a gas. It is one of the major pollutants present in the air.

It is colourless and has a nasty, sharp smell.

It combines effortlessly with other chemicals to form harmful substances like sulphuric acid, sulfurous acid, etc.

Sulfur dioxide affects human health when it is inhaled. It irritates the nose, throat, and airways to cause coughing, wheezing, shortness of breath, or a tight feeling around the chest. Those most at risk of developing problems if they are exposed to sulfur dioxide are people with asthma or similar conditions. Also, the concentration of sulfur dioxide in the atmosphere can influence the habitat suitability for plant communities, as well as animal life.

Inhaling sulfur dioxide is associated with increased respiratory symptoms and disease, difficulty in breathing, and premature death.

It also causes acid rain.

NO₂: Nitrogen Dioxide is a reddish-brown gas with a pungent, acrid odour.

It can cause bronchoconstriction, inflammation, reduced immune response, and may have effects on the heart. Direct exposure to the skin can cause irritations and burns.

The following gives a rough idea of nitrogen dioxide's impact on health :

10–20 ppm can cause mild irritation of the nose and throat

25–50 ppm can cause oedema leading to bronchitis or pneumonia

Levels above 100 ppm can cause death due to asphyxiation from fluid in the lungs.

High levels of NO₂ can harm vegetation, including leaf damage and reduced growth. It can make vegetation more susceptible to disease and frost damage.

Longer exposures to elevated concentrations of NO₂ may contribute to the development of asthma and potentially increase susceptibility to respiratory infections.

Particulates: These are also known as Atmospheric aerosol particles, atmospheric particulate matter, particulate matter (PM) or suspended particulate matter (SPM).

These are microscopic solid or liquid matter suspended in the atmosphere.

Particulates are the deadliest form of air pollution due to their ability to penetrate deep into the lungs and bloodstreams unfiltered, causing permanent DNA mutations, heart attacks, respiratory disease, and premature death.

Worldwide exposure to PM 2.5 contributed to 4.1 million deaths from heart disease and stroke, lung cancer, chronic lung disease, and respiratory infections in 2016. Overall, ambient particulate matter ranks as the sixth leading risk factor for premature death globally.

The internet is filled with the harmful effects of the above pollutants, and hence it makes them an essential factor to be analyzed and considered when discussing air pollution.

Coming back to the analysis.

OBJECTIVES

The objectives for analysing air quality trends are characterizing pollution patterns, protecting the public health, and determining compliance with air quality standards.

POLLUTION HOTSPOTS

These are the areas where air pollution emissions expose individuals to increased negative health effects. Hotspots denote areas in which a population's exposure to pollution and estimated health risks are high.

DATA EXPLORATION

```

In [5]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
stn_code                291665 non-null object
sampling_date           435739 non-null object
state                  435742 non-null object
location               435739 non-null object
agency                 286261 non-null object
type                   430349 non-null object
so2                    401096 non-null float64
no2                    419509 non-null float64
rspm                   395520 non-null float64
spm                    198355 non-null float64
location_monitoring_station 408251 non-null object
pm2_5                  9314 non-null float64
date                   435735 non-null object
dtypes: float64(5), object(8)

```

Now, let us check the null values

```

In [6]: df.isnull().sum()
Out[6]:
stn_code                144077
sampling_date              3
state                      0
location                  3
agency                 149481
type                     5393
so2                     34646
no2                     16233
rspm                     40222
spm                     237387
location_monitoring_station 27491
pm2_5                   426428
date                      7
dtype: int64

```

SAMPLE DATASET IN DIFFERENT AREAS

[illegible]

It represents the type of area where the data was recorded like industrial, residential, etc.

Let us see how many types of area were considered :

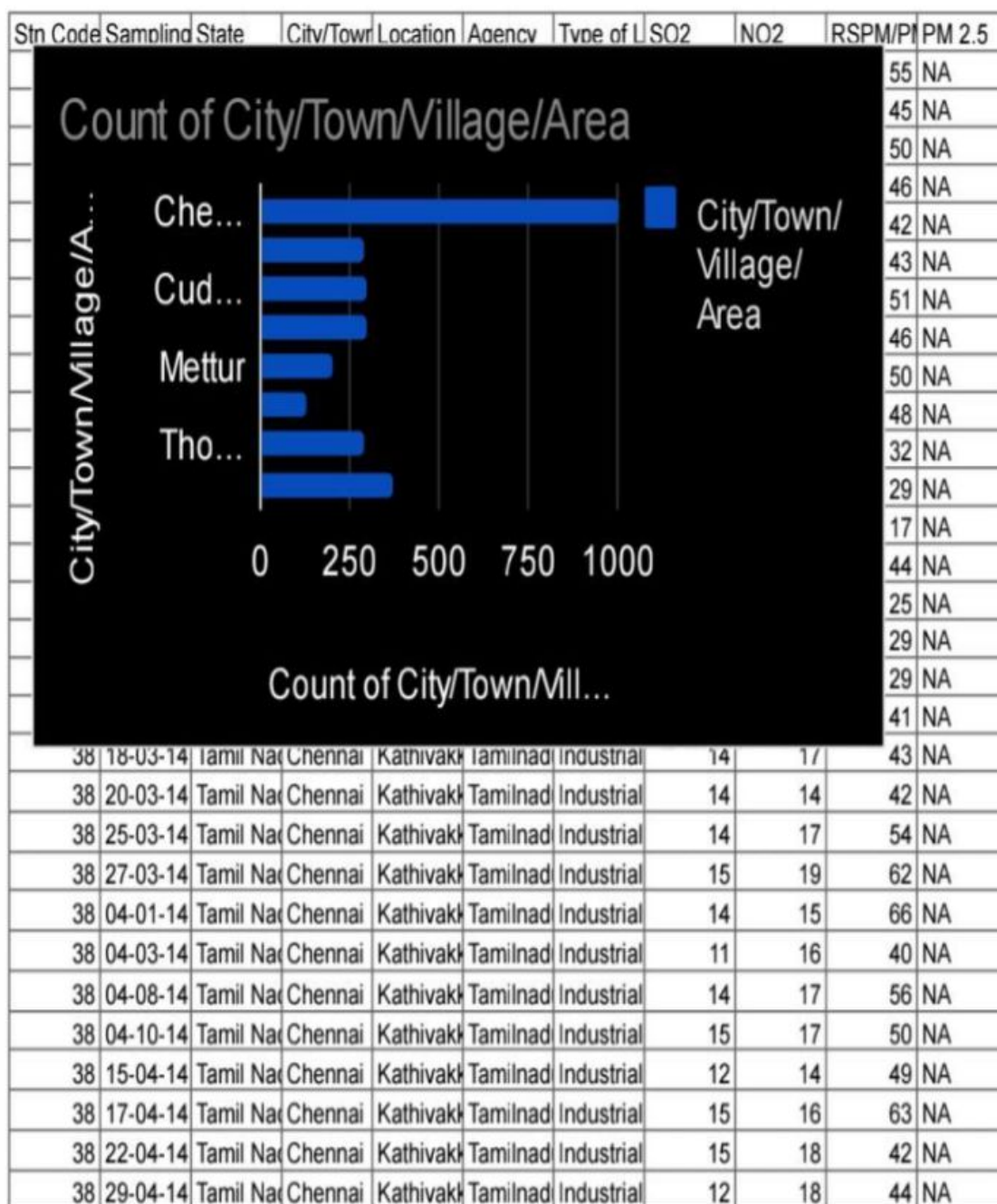
```
In [4]: df['type'].value_counts()
```

```
Out[4]:
```

Residential, Rural and other Areas	179014
Industrial Area	96091
Residential and others	86791
Industrial Areas	51747
Sensitive Area	8980
Sensitive Areas	5536
RIRUO	1304
Sensitive	495
Industrial	233
Residential	158

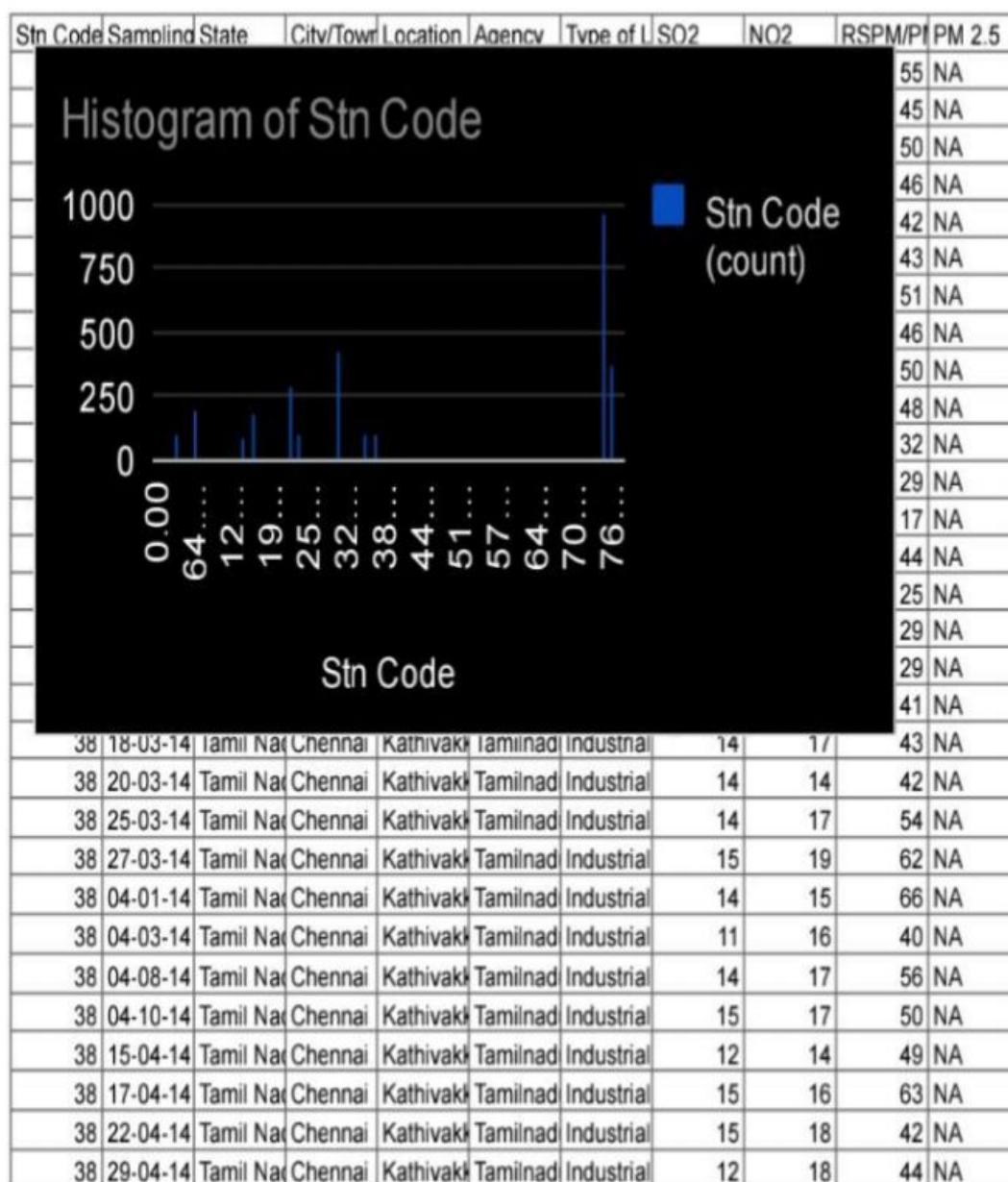
Name: type, dtype: int64

DATA VISUALIZATION



HISTOGRAM OF GATHERED DATA

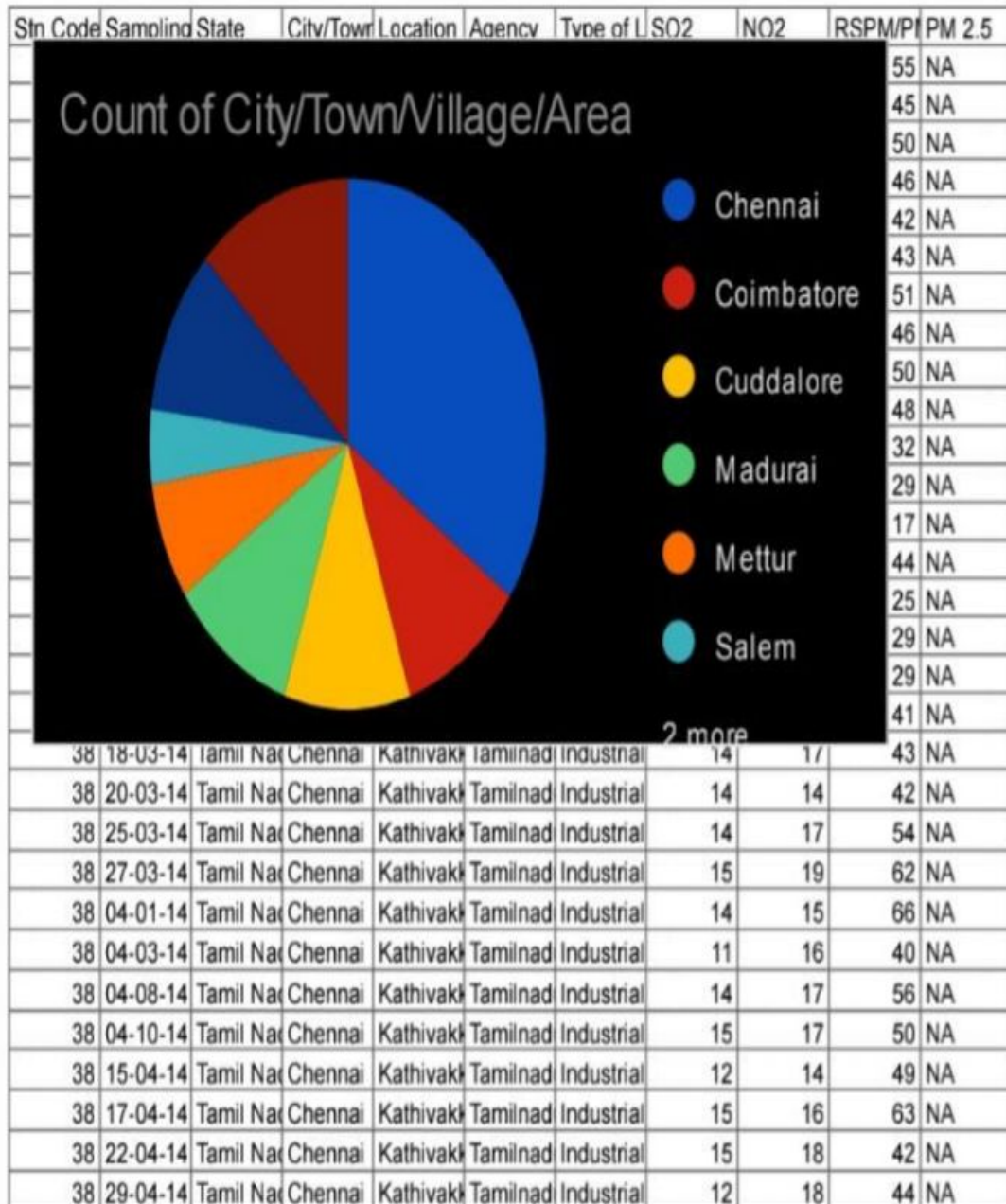
The historical line chart displays information as a series of data points connected by straight line segments. we show all the historical data of one specific item at one station. Generally speaking, a line chart is often useful to visualize a trend in data over time. Here, this histogram is represented with the station code which is given to each district to record the air quality analysis data

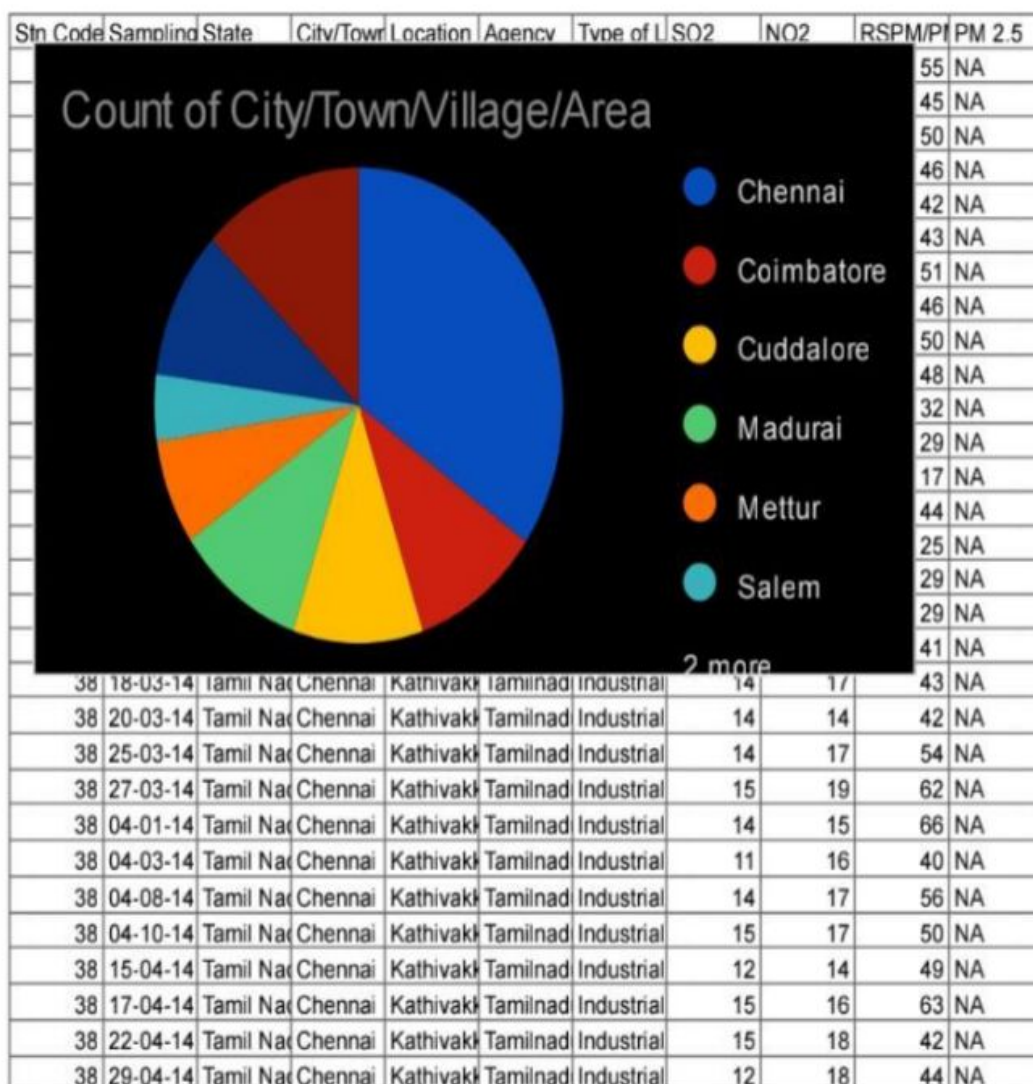


Stn Code	Sampling Date	State	City/Town	Location	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
<p>Count of Location of Monitoring Station</p> <p>Location of Monitor...</p> <p>Count of Location of M...</p>										55 NA
										45 NA
										50 NA
										46 NA
										42 NA
										43 NA
										51 NA
										46 NA
										50 NA
										48 NA
										32 NA
										29 NA
										17 NA
										44 NA
										25 NA
										29 NA
										29 NA
										41 NA
38	18-03-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	14	17	43	NA
38	20-03-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	14	14	42	NA
38	25-03-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	14	17	54	NA
38	27-03-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	15	19	62	NA
38	04-01-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	14	15	66	NA
38	04-03-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	11	16	40	NA
38	04-08-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	14	17	56	NA
38	04-10-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	15	17	50	NA
38	15-04-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	12	14	49	NA
38	17-04-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	15	16	63	NA
38	22-04-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	15	18	42	NA
38	29-04-14	Tamil Nar	Chennai	Kathivak	Tamilnad	Industrial	12	18	44	NA

CIRCULAR HEATMAP

A circular heat map is a graphical representation of data where the individual values contained in a circle shape. It displays quantitative data as an array of circular segments, colored according to value. It is a good way to display cyclic data in a well-chosen time period, in order to show a pattern as value changes





CONCLUSION

From the above data analysis approach, we conclude that data analysis is a crucial aspect for a better future. It is interesting to see how data analysis and the day to day instances are coherent and how data analysis can be used to deal with significant problems. Not only India other countries are also suffering from air pollution.