IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 3, Ser. IV (May - June 2020), PP 10-17
www.iosrjournals.org

Application of Data Science on Air Pollution in India

Vishal Dwivedi[1], Dr. Anuradha Misra[2], Dr.Sheenu Rizvi[3]
[1](Department of Computer Science & Engineering, ASET / Amity University, India)
[2](Department of Computer Science & Engineering, ASET / Amity University, India)
[3](Department of Computer Science & Engineering, ASET / Amity University, India)

Abstract: *Air Pollution suggests the solvents of toxins in the air and cause negative to someone's health and breathe related issues. Now a day's air pollution has been one of the serious issues inside the country. South Asia stands in number six in this scenario. It doesn't commonly comprehend the ruinous effects of an issue if a person has not experienced it regardless. Taking instance of Delhi for examples, we all in all have reading in newspaper articles, social media platform and all other sources of information about the effect on weather in after Diwali. Inhabitants of Delhi were urged not to come out from their houses and were drawn closer to wear covers at whatever point heading outside. Looking out from their home window made them feel like they are living in gas filled box. Lower intelligible are some of the hazardous effect of Air pollution inside the country. In this paper I would like to focus on some of the key points which are responsible for this result and the data analytical survey of the same.*

Key Word*:: Environment, Ambient Air Quality, Data Analysis, Air Pollution*

Date of Submission: 06-06-2020 Date of Acceptance: 22-06-2020

I. Introduction

The data is collected from TMEF&CPCB(Central Pollution Control Board of India) database, this is just an approximate value for the use data analysis.[1][2] The database has following attributes given as-
STATION CODE: A particular code that is given to every station for organizing data.
SAMPLING DATE: The date along with other sampling details at the time data was recorded in reality.
THE STATE: It speaks to the name of the Indian states where air quality information is anticipated.
THE AGENCY**:** Name of the organization that estimates the database.
THE TYPE: The type of region where the amount was made on.
THE LOCATION_MONITORING STATION**:** It simply demonstrates the location of the area to be monitored under.
THE DATE**:** It demonstrates date of recording the data.
Here is the sample of some data recorded with the given attributes and some data in format.

stn_code	sampling_date	state	location	iger	type	SO2	NO2	rsi	spi	n	m	date
150	Feb-M021990	Andra pradesh	Hyderabad		Residential, Rural and Other areas	4.8	17.2					02-Jan
151	feb-M021990	Andra pradesh	Hyderabad		Industrial area	3.1	7					02-Jan
152	feb-M021990	Andra pradesh	Hyderabad		Residential, Rural and Other areas	6.2	28.5					02-Jan
150	Mar-M031990	Andra pradesh	Hyderabad		Residential, Rural and Other areas	6.3	14.7					03-Jan
151	Mar-M031990	Andra pradesh	Hyderabad		Industrial area	4.7	7.5					03-Jan
152	Mar-M031990	Andra pradesh	Hyderabad		Residential, Rural and Other areas	6.4	25.7					03-Jan
150	Apr-M041990	Andra pradesh	Hyderabad		Residential, Rural and Other areas	5.4	17.1					04-Jan
151	Apr-M041990	Andra pradesh	Hyderabad		Industrial area	4.7	8.7					04-Jan
152	Apr-M041990	Andra pradesh	Hyderabad		Residential, Rural and Other areas	4.2	23					04-Jan

Fig 1: A sample of several attributes to be used in Air quality measure and their records

II. Feature Explanation

SO_2: Abbreviation of Sulphur Dioxide and it's a harmful gas. The contribution of this gas results in increase of air pollutant and its very hazardous to the health of human being, consequences are sometimes very tough to handle.

It impacts human being prosperity when it is taken in. It results in irritation of the throat, nose and aeronautics courses to cause hacking, quickness of breathing, or a tight looking about chest. Those most in peril of creating issues if they're introduced to Sulfur Dioxide are people with "Asthma" or relative conditions. In like manner, the gathering of SO_2 in the earth can affect the regular surroundings sensibility for the systems of plant, similarly as the life of animal. Breathing in this hazardous gas is related with expanded respiratory indications, sickness, and trouble in breathing. The consequences overall are not good for the health of human being or animals.

NO_2: It is a ruddy dark colored gas with an impactful, bitter smell. It can cause bronchoconstriction, irritation, diminished insusceptible reaction, and many more impact on the heart to. Straight introduction to the coating of skin can cause consumes and aggravations.[3]

The accompanying gives an unpleasant thought of nitrogen dioxide's effect upon wellbeing:

10–20 ppm can cause mellow bothering of the nose and throat
25–50 ppm can cause edema prompting bronchitis or pneumonia
Levels over 100 ppm can cause passing because of suffocation from liquid in the lungs.

Significant levels of NO_2 can affect vegetation extreme negatively, including diminished development and leaf harm. It can make plant life progressively helpless and then leads to sickness. The effect of this hazardous gas is seen in peoples as well as nature too.

III. Particulates

Particulates are recognized as Atmospheric vaporized particles, atmospheric PM (Particulate Matter). These are minute strong or fluid issue suspended in the air. Particulates are the deadliest type of air contamination because of their capacity to infiltrate profound into the lungs and circulation systems unfiltered, causing perpetual DNA changes, cardiovascular failures, respiratory infection, and sudden passing. [5]

"Overall presentation to Particulate Matter 2.5 added to approx. four million passing from coronary illness and cardiac stroke, lungs malignant growth, endless lung malady, and the respiratory contaminations in year 2016. In general, encompassing particulate issue positions as the 6th driving danger factor for the unexpected passing all over the globe". We have a lot of information about this harmful gas on internet and other sources, and subsequently it creates them a basic factor being dissected and thought about when talking about air pollution. Pm2_5 is the term used here to represent Particulates matter 2.5 .

IV. Data Exploration

Here are some of the entries which with all the components like station code, sampling date, location, agency and many more on the basis of which I am going to study the data analysis: -

stn_code	291665
sampling_date	435739
state	435739
location	435739
agency	282661
type	430349
SO2	401096
NO2	419509
rspm	395520
spm	198355
location_mo_station	408251
pm2_5	9314
date	435735

Fig 2: A sample with all the attributes involves in analysis.

From the above data taken from the database, we see that we have 435742 entries there are some lesser values present for pm2_5. Now we will learn for the null values.

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

Fig 3: Dataset with null values in the attributes

It appears that we have a ton of invalid qualities in certain sections. Taking a glance at the figure-3, we see that particulate matter 25 have less non-invalid qualities and it perhaps won't have the option to contribute a lot.

Station code, office, spm additionally are loaded up with invalid qualities. On the off chance that I need to break down the air pollution information of India, what amount polluted are the state. Likewise, station code is additionally unnecessary.

It is provided in the info portrayal that date is a fresher portrayal of *sampling date* trait thus we will dispense with the recurrence by vacating the last mentioned. It appears that we have a great deal of invalid qualities in certain sections. Taking a glance at the figure 3, we see that particulate matter v2.5 have less non-invalid qualities and it probably won't have the option to contribute a lot. Station code, organization, suspended particulates matter additionally are loaded up with invalid qualities.Similarly, station code is also unnecessary attribute for the next study.

Locationmonitoringstation: This attribute is yet unnecessary as it contains the location of monitoring station that we do not need to consider for the analysis. Removing unnecessary column will help to study the data analysis process easily and the dataset can be more precise and accurate.

Hence, I have removed these three features from our dataset and the table for same will be containing attributes like agency, station code,*sampling_date and locationmonitoringstation. T*he study without three attributes are shown in Figure 4.

state	location	type	SO2	NO2	date
Andra Pradesh	Hyderabad	Residential, Rural and others	4.8	17.2	02-Jan
Andra Pradesh	Hyderabad	Industrial	3.1	7	02-Jan
Andra Pradesh	Hyderabad	Residential, Rural and others	6.2	28.5	02-Jan
Andra Pradesh	Hyderabad	Residential, Rural and others	6.3	14.7	03-Jan
Andra Pradesh	Hyderabad	Industrial	4.7	7.5	03-Jan
Andra Pradesh	Hyderabad	Residential, Rural and others	6.4	25.7	03-Jan
Andra Pradesh	Hyderabad	Residential, Rural and others	5.4	17.1	04-Jan
Andra Pradesh	Hyderabad	Industrial	4.7	8.7	04-Jan
Andra Pradesh	Hyderabad	Residential, Rural and others	4.2	23	04-Jan

Fig 4: Dataset with some required attributes.

It commonly represents the area type where the data was collected or recorded such as, Residential area, industrial areas and others. Because the presence of unwanted particle defers from area to are and from location to location. Industrial area produces more hazardous and dangerous gases from their outlet chimneys which harms the air quality, respectively with Residential and then other one. Since people are conscious about their health and surroundings, the amount of pollution is very less in residential areas. Let us study these three areas that is taken in consideration. It appears that we have repetitive sorts. Taking a gander at the mentioned figure 5, it tends being said that a mentioned zone can be ordered in 3 classes/types i.e., Residential, Industrial and other.

Here, I am going to plot the variance in counts of the threepreviously mentioned classes. Type characteristic in the wake of changing the classes:

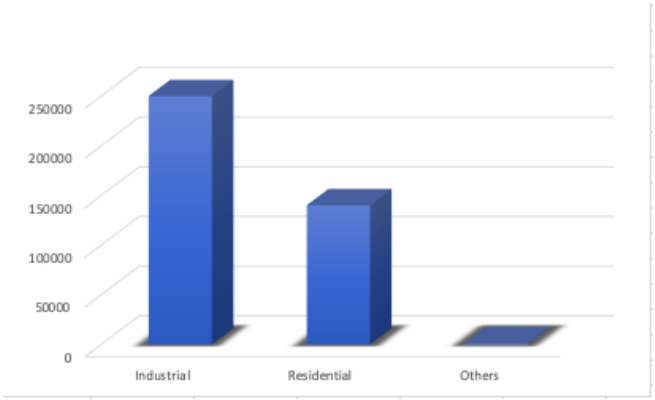

Fig 5: - Plot of Number of entries with respect to Categories (Residential Industrial and others) in column.

V. Data visualization

Data Visualization is the realistic portrayal of information. It includes creating pictures that convey connections among the spoke to information to watchers of the pictures. This correspondence is accomplished using an orderly mapping between realistic imprints and information esteems in the production of the perception.[4]

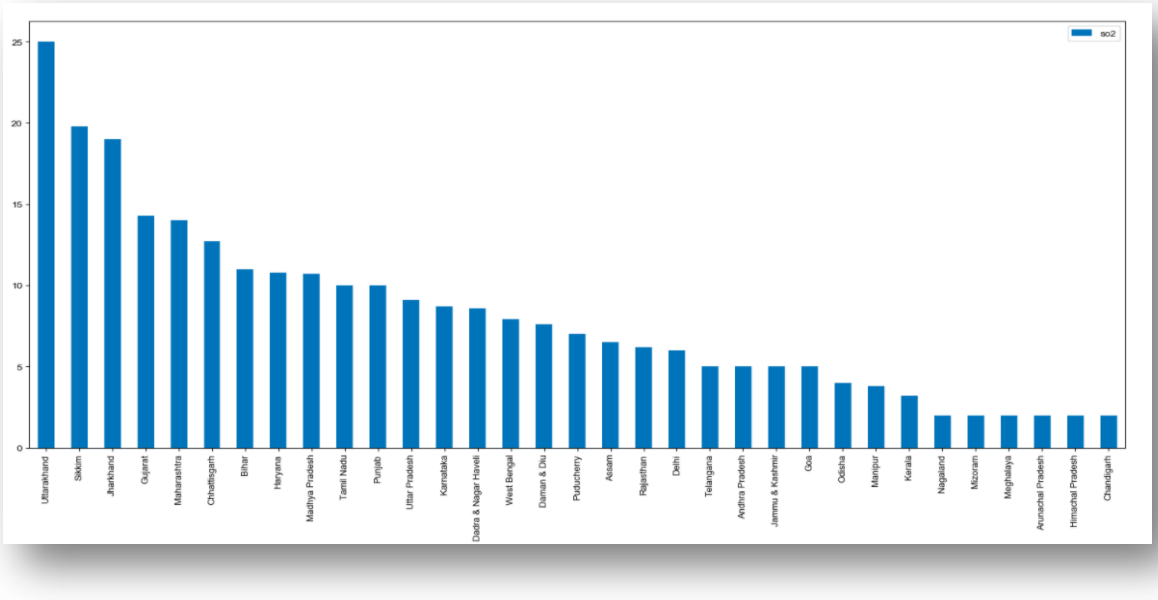

Fig 6: The plot of SO2 with respect to the states.

From the above stats we can observe that SO2 obsession is generally raised in the state of Uttarakhand and the least in the state of Chandigarh. Uttarakhand, Gujarat, Maharashtra, Sikkim, Jharkhand, and Chhattisgarh — the legislature should make a move against the developing SO_2 focus in these states.

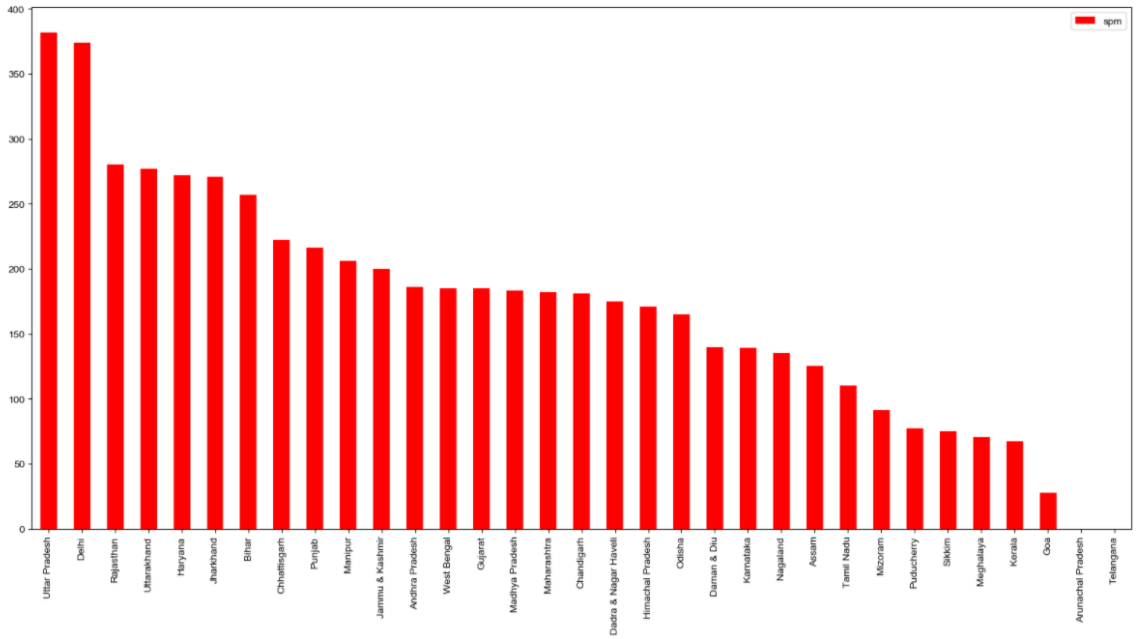

Fig 7: The plot *spm* with respect to different locations in India

It's being clear from the above mention's statistics that Meerut (in the state of UP) is the dirtiest city as far as SPM pursued by Kharja and Ghaziabad. The partition of state to the cities will clearly demonstrate the affect respectively Believe it or not the fundamental 7 urban zones to be explicit, Firozabad, Noida, Ghaziabad, Kanpur, Meerut, Kharja, and Allahabad are orchestrated in Uttar Pradesh.

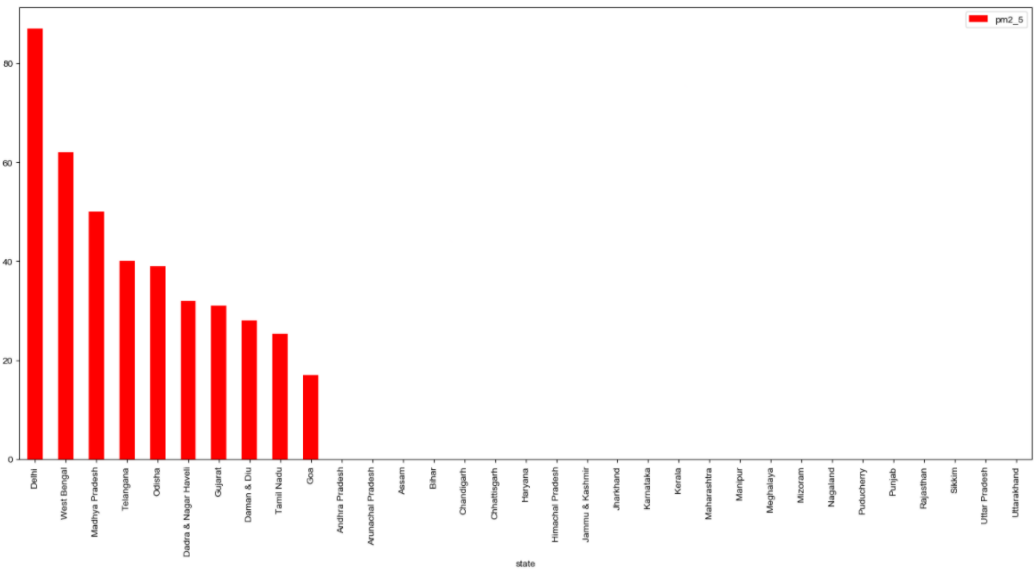

Fig 8: Bar plot of pm2_5 and the location for non-zero values

We got to know that the city of Delhi is on the top of list by Talcher and city of Odisha and Gwalior in MP.

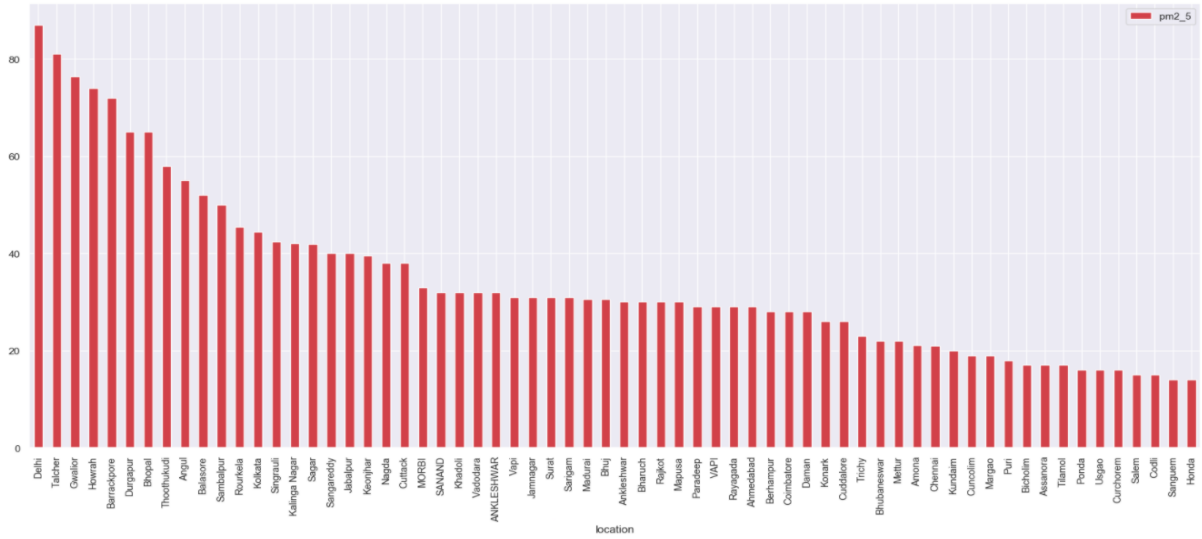

Fig 9: Most polluted states in the list of overall with respect to pm2_5.

We can see that you've invalid qualities for the greater part of Indian states that is very obvious as talked about over that particulate matter 2_5 had the maximum number of invalid values (approx. 97.86% of invalid qualities). From whatever we've, we see that the city of Delhi again best the neglected, trailed by Madhya Pradesh and West Bengal. Very little is talked about for pm2_5 as it has enormous figure of invalid qualities.

VI. Statistical Analysis of Data

Statistical analysis is an occurrence of data analytics. The process includes collecting and analyzing every data sample in the set of items from which statistics can be drawn very easily. As a matter of first importance, I won't comment on the association among pm2_5 and some other component basically in light of the fact that pm2_5 have colossal measures of invalid regards and therefore its genuine centrality is low maybe even irrelevant.[6]

We found out that NO_2 and SO_2 have somewhere practically the same model with respect to dissimilar highlights. It very well may be said that RSPM & SPM share direct relationship quite genuinely, rest all features aren't actually related.

For a more profound investigation let us take a gander at the relationship network

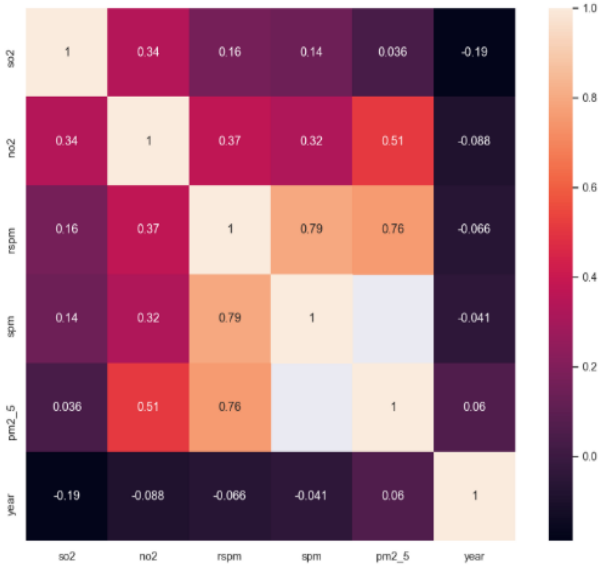

Fig11: The Co-relation Matrix for the Dataset contains attributes

It is noticeable from the mentioned relationship grid that we have some connection amid RSPM and SPM that underpins our scatter the plot analysis. There is a petite connection among different highlights.[9]

VII. Data Features

Date comprise connotes the date when the details was recorded. Let's be adept and devise another feature (mention the year here) from the date feature. This is on the basis that we're keen on the yearly collisions of air contamination. The information appears like trailing and making the section of year.

index	state	location	type	so2	no2	rspm	spm	PM2.5	date	year
0	Andra Pradesh	Hyderabd	Residential	4.8	17.4	nan	nan	nan	01/02/90 0:00	1990
1	Andra Pradesh	Hyderabd	Industrial	3.1	7	nan	nan	nan	02/02/90 0:00	1990
2	Andra Pradesh	Hyderabd	Residential	6.2	28.5	nan	nan	nan	03/02/90 0:00	1990
3	Andra Pradesh	Hyderabd	Residential	6.3	14.7	nan	nan	nan	04/02/90 0:00	1990
4	Andra Pradesh	Hyderabd	Industrial	4.7	7.5	nan	nan	nan	05/02/90 0:00	1990
5	Andra Pradesh	Hyderabd	Residential	6.4	25.7	nan	nan	nan	06/02/90 0:00	1990
6	Andra Pradesh	Hyderabd	Residential	5.4	17.1	nan	nan	nan	07/02/90 0:00	1990
7	Andra Pradesh	Hyderabd	Industrial	4.7	8.7	nan	nan	nan	08/02/90 0:00	1990
8	Andra Pradesh	Hyderabd	Residential	4.2	23	nan	nan	nan	09/02/90 0:00	1990
9	Andra Pradesh	Hyderabd	Industrial	4	8.9	nan	133	nan	10/02/90 0:00	1990
10	Andra Pradesh	Hyderabd	Residential	3.6	18.6	nan	82	nan	11/02/90 0:00	1990
11	Andra Pradesh	Hyderabd	Residential	3.9	14.1	nan	111	nan	12/02/90 0:00	1990
12	Andra Pradesh	Hyderabd	Industrial	5.6	11.8	nan	118	nan	13/02/90 0:00	1990
13	Andra Pradesh	Hyderabd	Residential	3.3	19.3	nan	135	nan	14/02/90 0:00	1990
14	Andra Pradesh	Hyderabd	Residential	3.9	8.2	nan	80	nan	15/02/90 0:00	1990
15	Andra Pradesh	Hyderabd	Residential	3.5	12.1	nan	179	nan	16/02/90 0:00	1990

Fig12: Data set after the plan of years

Since we have made multiyear segment so we can examine the information every year.

SO2 examination utilizing heatmap Give us a chance to plot the heatmap for SO2 which contains.*Push*: state quality Segment: year trait Esteem: SO2 property It is clear from the statistics that there has been a reasonable augmentation of SO_2 obsession in the state of Bihar from year 1987 to 1999. In this way, there's been high centralization of SO_2 in Gujarat in 1995. In the state of Haryana in like manner we can see that obsession of SO_2 has been high around in year 1987 and has been dependably high till the year 2003. In Karnataka, furthermore, there's a consistent addition in SO_2 center from the years 1987 to 2000. The union territory of Pondicherry in like manner has seen enhanced estimation of SO_2 obsession in 1996. Rajasthan, similarly, have experienced high gathering of SO_2 in 1987. Uttarakhand has been experiencing SO_2 obsession with raised degree from 2004 till the date. I found the associated news story that strengthens our decision and the examination of heatmap.[6]

Information discharged by the Aura satellite of NASA raises the doubt regarding the reality of CPCB (Central Pollution Control Board's) guarantee made in the year 2012 that the mean SO_2 emanations in India reduced in 2010 when contrasted with the level in 2001. A portion of the states such as Sikkim, Uttarakhand, Jharkhand, etc., on still experience widely elevated levels of SO_2 focus. The accomplishment showed results and the decreased level of SO2.The NO_2 concentration has reduced yearly in some states such Rajasthan, whereas in Bihar, Delhi, etc., it has augmented.We see that Uttar Pradesh, Delhi, Haryana, Jharkhand, Punjab, have suffered from very high levels of the pollutants like RSPM.[10].We can see that Delhi, Uttar Pradesh, Haryana, and Punjab have been the major suffering states from extreme high concentration of pollutants of SPM.

Since we have studied the effects of air pollution, here I am proposing the method to forecast for the future circumstances and the harmful results for the same, and the methodologies used in data science for the prediction. Here is the simple flow chart for the proposed approach. [8]

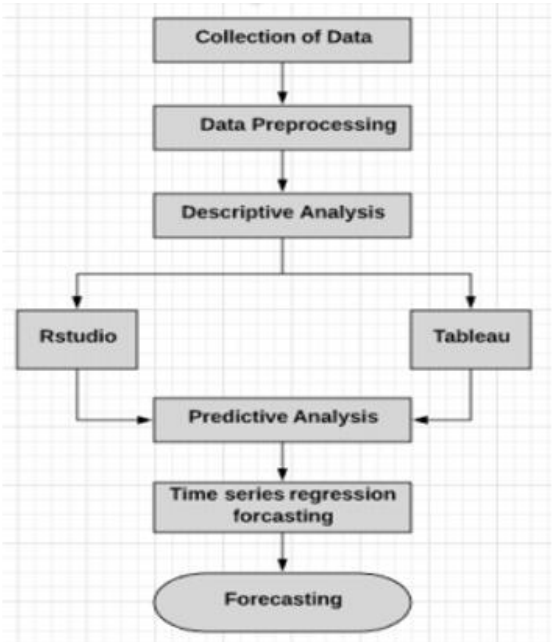

Fig 14: - A flow chart for the study of Air pollution in Future.

VIII. Conclusion

The above analysis and study of data states that, in total, influencing states of India through the issues like air pollution that have a spot with the northern region. Delhi, Uttar Pradesh, Haryana, Punjab are seriously extremely polluted and require some planned and instant action.

We also observed that paying little attention to whether a state in India had enormous degree of poisons and pollutants there were a couple of areas here those are not polluted. From the provided heatmap, we got to know that several states were tremendously polluted in the early stages simultaneously, later, were taken thought and the condition brought back to normal. More number of cities and states can be added in the charts and the study can be done even in broad way.

The explanation behind the lessening could be care in government plans and occupants. This deep data analysis helps mankind to understand the severely critical condition due to Air pollution in whole India and most of the bad effects has been seen in northern part.

References

[1]. Statistical Abstract (2016) Delhi Govt Portal, www.delhi.gov.in
[2]. "Real-time ambient quality of Air pollution in Delhi and India", DPCClink: http://www.dpccairdata.com/dpccairdata/display.
[3]. Mishra D., Goyal P. (2015) "Development of artificial intelligence based NO2 forecasting models at Taj Mahal, Agra Centre for atmospheric sciences", Atmospheric Pollution Research.
[4]. "Ambient Air Quality Data at various locations". Central Pollution Control Board. 2015-16.
[5]. Sindhwani R. (2012) "Assessment of gaseous and respirable suspended Particulate matter (PM10) emission estimates over megacity Delhi: past trends and future scenario (2000–2020)." 13th Annual CMAS Conference, Chapel Hill, NC, USA.
[6]. TanejaShweta, SharmaNidhi, OberoiKettun, NavoriaYash. (2016) "Predicting Trends in Air Pollution in Delhi using Data Mining." Information Processing (IICIP), 2016 1st India International Conference, DTU, New Delhi.
[7]. Ma Xin, Gong Wei, Zhu Zhongmin. (2016) "The study of long-term air pollution characteristic in Wuhan, China." Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International DOI: 10.1109/IGARSS.2016.7730073
[8]. Sharma Nidhi, Taneja Shweta, Sagar Vaishali, "Forecasting air pollution load in Delhi using Data Analysis tools", International conference on Computational Intelligence and Data Science(ICCIDS 2018)
[9]. David Núñez-Alonso, Luis Vicente Pérez-Arribas, Sadia Manzoor, and Jorge O. Cáceres, "Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies", Journal of Analytical Methods in Chemistry, Volume 2019, Article ID 9753927
[10]. Central Pollution Control Board report on Air pollution in Delhi; An analysis (2016).

Vishal Dwivedi, et. al. "Application of Data Science on Air Pollution in India." *IOSR Journal of Computer Engineering (IOSR-JCE),* 22(3), 2020, pp. 10-17.