Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

Third Year CSE (2020–21)

DWDM

Trimester VIII

Mini Project Manuscript

Under The Guidance Of -Prof.Sheetal Girase

Made By-

- PA 09 Anand Venkataraman
- PA 14 Amar Rakh
- PA 28 Vighnesh Sairaman
- PA 43 Kundan Walunj

Abstract-

Examining and protecting air quality in this world has become one of the essential activities for every human in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution. With this increasing air pollution, we need to implement models that will record information about concentrations of air pollutants. The deposition of these harmful gases in the air is affecting the quality of people's lives by altering their health, especially in urban areas.Our solution helps by contributing a small part to the decrease in pollution by regulation of its levels.

INDEX

Page Number	Title
3	Abstract & Introduction
4	Objectives, Related Work & Tools used
5	Literature Survey
10	Dataset Information and Features
12	Data Preprocessing
13	Architecture Methods and Learning Algorithms
17	Experiments and Inferences
26	Conclusion
27	Bibliography & Citations

Problem Statement-

To use the data set containing pollution levels of various industrial areas and to predict future pollution levels which will help to safeguard the environment by alerting necessary authorities if and when required.

Introduction-

Air of cities, especially in the developing parts of the world is turning into a serious environmental concern. The air pollution is because of a complex interaction of dispersion and emission of toxic pollutants from manufacturing Industries. Air pollution caused due to the introduction of dust particles, gasses, and smoke into the atmosphere exceeds the air quality levels. Air pollutants are the precursor of photochemical smog and acid rain that causes the asthmatic problems leading into serious illness of lung cancer, depletes the stratospheric ozone, and contributes to global warming. In the present industrial economy era, air pollution is an unavoidable product that cannot be completely removed but stern actions can reduce it. Pollution can be reduced through collective as well as individual contributions. There are multiple sources of air pollution, which are industries, fossil fuels, agro waste, and vehicular emissions. Industrial processes upgradation, energy efficiency, agricultural waste burning control, and fuel conversion are important aspects to reducing pollutants which create the industrial air pollution. Mitigations are necessary to reduce the threat of air pollution using the various applicable technologies like CO2 sequestering, industrial energy efficiency, improving the combustion processes of the vehicular engines, and reducing the gas production from agricultural cultivations.

Objectives-

- Identification of appropriate dataset for our problem statement.
- Application of preprocessing techniques on Dataset.
- Studying the dataset as well as different types of algorithms.
- Implementation of a handful of above algorithms to our Dataset.
- To analyze the inferences from the results of our algorithms and derivation of a conclusion.
- Visualization of the Results from our Dataset using a third-party software.

Related Work-

Air Quality Prediction Through Regression Model by A.Aarthi, P.Gayathri, N.R.Gomathi , S.Kalaiselvi , Dr.V.Gomathi[2]

Performance Analysis of KMeans and KMediods Algorithms in Air Pollution Prediction by- S. Suganya, T. Meyyappan, S. Santhosh Kumar ;International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5, January 2020[1]

Our Analysis precedes the latter in terms of the analytical and mathematical approach but far succeeds the the latter in terms of graphical approach and visualization.This makes it easier to understand for any individual and operate the dataset to fulfill our problem statement.

Our method of using code to show that makes it easier to understand regression and to implement it.

Tools used-

- Google Colab.
- Tableau Visualization Software.
- Domo Visualization Software.

Literature Survey-

India is a rapidly developing economy with interrelated air quality, sustainable development, and climate change mitigation goals. There are unique challenges to achieving each of these goals as well as potential trade offs among them.

- India has some of the worst air quality in the world, making its improvement a high priority for the Indian government as evidenced by the launch of the National Clean Air Programme in January 2019
- Hundreds of millions of people in India are continually exposed to toxic air: they inhale, for example, a 24-hour average of up to 25 micrograms/cubic metre of air of the deadly, microscopic pollutant, PM 2.5—far above the World Health Organization's (WHO) limit of 10 micrograms/cubic metre. With long-term exposure, this particulate matter goes deep into the lungs and on to other organs and systems, gradually defeating the body's defense mechanism. Repeated exposure to toxic air causes cardiovascular and respiratory diseases, lung and other cancers, strokes, preterm birth, type-2 diabetes, and other illnesses.
- Since February 2014, the government of India has been monitoring industrial emissions and effluents in rivers and lakes across the country. The monitoring is done through what is called the Online Continuous Emissions/Effluents Monitoring Systems (OCEMS). The 17 categories of industrial units that are required to have OCEMS—these include power plants; aluminum, zinc, copper plants; cement plants; distilleries; fertilizers, iron and steel plants; oil refineries; petrochemicals; and tanneries. The emissions monitored under the OCEMS regulations include particulate matter (PM), CO (carbon monoxide), NOx (nitrogen oxide), SO2 (sulfur dioxide), and fluoride.

- In mid-March 2020, discussions in Parliament indicated that there were some 3,700 OCEMS installed in different industrial locations across the country. A month earlier, Parliament was informed by the Union government that the total number of targeted units was 4,245.
- Meanwhile, there are only 234 continuous air pollution monitors (also known as CAAQMS – Continuous Ambient Air Quality Monitoring Systems) in the country as of October 2020; the data from these monitors serve as the basis for the AQI or national Air Quality Index. By this yardstick, it is apparent that the scale of monitoring of pollutants is bigger in the country's industrial sector. The OCEMS network is regulated by the same regulatory body, the Central Pollution Control Board (CPCB), and monitors similar parameters as those covered by the CAAQMS. However, in the OCEMS, the commissioning and operations of the monitoring systems is left to the same industries which are themselves being monitored for their emissions.
- While the installation and operations of CEMS equipment is highly technical, the understanding of the data from the CEMS network does not need to be as complex. The nuance of this is best understood by identifying who are the people most affected by the pollution from these industrial units where the CEMS equipment is installed. Both CAAQMS and CEMS are designed to benefit the people who are living in the vicinity of that monitoring system.
- However, the data on industrial emissions is almost impossible to access for the public. Each industrial unit's operator has access to the data as well as the Central and state pollution control boards. The stated aim of OCEMS is for monitoring and self-regulation. But here is where the CPCB appears to sidestep transparency of OCEMS data.[4]

Types of industries contributing to the higher pollution levels-

- **Agriculture**

Agricultural activities produce emissions, which have the potential to pollute the environment. Ammonia (NH3) and nitrous oxide (N2O) are the key pollutants released from agricultural activities. The other agricultural emissions include methane emissions from enteric fermentation processes, nitrogen excretions from animal manure, such as CH4, N2O, and NH3, methane emissions from wetlands, and nitrogen emissions from agricultural soils (N2O, NOX, and NH3) due to the addition of fertilizers and other residues to the soil (Gurjar, Aardenne, Lelieveld. et al 2004). Agricultural processes, such as 'slash and burn' are prime reasons for photochemical smog resulting from the smoke generated during the process. Crop residue burning is another process that results in toxic pollutant emissions. This is how neighboring cities of Delhi contribute to the agricultural pollution load. This is an example of how external sources contribute to the menace of air pollution in the city (Nagpure, Gurjar, Kumar, et al. 2016).

- **Power Plants**

The contribution of power plants to air emissions in India is both immense and worrisome. The thermal power plants manufacture around 74% of the total power generated in India (Gurjar, Ravindra, and Nagpure 2016). According to The Energy and Resources Institute (TERI), the emissions of SO2, NOX, and PM increased over 50 times from 1947 to 1997. Thermal power plants are the main sources of SO2 and TSP emissions (Gurjar, Aardenne, Lelieveld, et al. 2004), thereby contributing significantly to the emission inventories. In Delhi, power plants contributed 68% of SO2 emissions and 80% of PM10 concentrations over a period from 1990 to 2000 (Gurjar, Aardenne, Lelieveld, et al. 2004). Thus, there is an urgent need to adopt alternative power sources including green and renewable resources for meeting the energy requirements.

- **Waste Treatment and Biomass Burning**

In India, about 80% of municipal solid waste (MSW) is still discarded into open dumping yards and landfills, which leads to various GHG emissions apart from the issues of foul odour and poor water quality at nearby localities. The lack of proper treatment of MSW and biomass burning has been responsible in causing air pollution in urban cities. In Delhi alone, around 5300 tonne of PM10 and 7550 tonne of PM2.5 are generated every year from the burning of garbage and other MSW (Nagpure, Gurjar, Kumar, et al. 2016).

Methane (CH_4) is the major pollutant released from landfills and wastewater treatment plants. Ammonia (NH_3) is another by-product, which is released from the composting process. The open burning of wastes, including plastic, produces toxic and carcinogenic emissions, which are a grave concern (Gurjar, Aardenne, Lelieveld, et al. 2004).

- **Domestic Sector**

Households are identified as a major contributor of air pollution in India. The emissions from fossil fuel burning, stoves or generators come under this sector, thereby affecting the overall air quality. Domestic energy is powered by fuels, such as cooking gas, kerosene, wood, crop wastes or cow dung cakes (Gurjar, Aardenne, Lelieveld, et al. 2004).

Though liquefied petroleum gas (LPG) is used as an alternative source of cooking fuel in most urban cities, the major share of India's rural population continues to rely on cow dung cakes, biomass, charcoal or wood as the primary fuel for cooking and other energy purposes and demands. These lead to severe implications on air quality, especially the indoor air quality, which could directly affect human health. According to HEI (2019), about 60% of India's population was exposed to household pollution in 2017.

- **Construction and Demolition Waste**

Another major source of air pollution in India is waste, which is an outcome of construction and demolition activities. Guttikunda and Goel (2013) inferred from their study that around 10,750 tonnes of construction waste is generated in Delhi every year. Even after the construction phase, these buildings have the potential to be the major contributors of GHG emissions. Nowadays, the increasing interest in green building technologies and the application of green infrastructure and materials during construction could tackle this issue to a large extent, thereby preserving our biodiversity and maintaining cleaner air quality[7]

Dataset and Features

In the proposed system, the air quality dataset is downloaded, which is available in CSV format. The comma separated value data format can easily be processed and analyzed fast using a computer and the data utilized for various purposes. It is imported to the project by using a panda package available in google colab and Jupyter software. The dataset contains 10 important attributes that help in air quality prediction. Initially, the dataset is preprocessed with suitable techniques to remove the inconsistent and missing valued data, and the needed features from the dataset are selected for better results.

The air quality dataset for this project is collected from the UCI repository. The dataset is available in CSV format. It is downloaded and imported to the project by mentioning the location of a downloaded dataset using the panda package available in Jupyter and Colab. The dataset contains data of average hourly responses of different elements in the air for nearly one year from 1990 to 2015. Dataset consists of 253076 rows × 13 columns.

Column	Description
STN_CODE	Serial Code for the Location
Sampling Date	Date
State	Name of the state
City	Name of the city
Location	The type of area where pollution data was recorded
{Pollutant} SO_2	Value of that pollutant(SO_2)
{Pollutant} NO_2	Value of that pollutant(NO_2)
{Pollutant} RSPM	Value of that pollutant(RSPM)
{Pollutant} SPM	Value of that pollutant(SPM)
Date	Formatted Date

In the Dataset pollutant gasses of SO_2,NO_2,RSPM,SPM (PM is Particulate matter) have been mentioned of various regions of states of the Indian subcontinent.On this matter pre-processing has been done to prepare the data for analysis.

Data Preprocessing

Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), and missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running any analysis. Often, data preprocessing is the most important phase of a machine learning project, especially in computational biology.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take a considerable amount of processing time. Examples of data preprocessing include cleaning, instance selection, normalization, one hot encoding, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set.

Data preprocessing may affect the way in which outcomes of the final data processing can be interpreted. This aspect should be carefully considered when interpretation of the results is a key point, such in the multivariate processing of chemical data (chemometrics).

Architecture Methods and Learning Algorithms

1) K Means:
   - K-Means is the individual of the majority famous unsupervised learning algorithms.
   - K-Means clustering system is group the similar objects or data points based on the selected centroid point of the data
   - Data points in the K-means are grouped into K clusters. Here K- indicates the centroid value of each cluster

Algorithmic steps for k-Means Clustering:

Let X be the collected works of data point and Y be the beginning point to the near centroid.

1. Select two random points as centroid in the graph.

2. Create clusters and assign points that are closer to the particular centroid.

3. Further, for creating accurate clusters, join the centroids with a line; draw a line perpendicular to that line.

4. Create accurate clusters depending on position of the points with respect to the perpendicular line.

5. Repeat steps 2-4 until the centroid data points remain in the same cluster and the centroid doesn't change.[1]

$$J = \sum_{i=1}^{m} \sum_{k=1}^{m} w_{ik} \, \|x_i - c_k\|^2$$

$w_{ik} = 0$ *if the data point does not belong to the cluster*
$w_{ik} = 1$ *if the data point belongs to the cluster*

2) Decision Tree

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node.
- Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

Algorithmic steps for Decision Tree Clustering:

1. Begin the tree with the root node, says S, which contains the complete dataset.
2. Find the best attribute in the dataset using **Attribute Selection Measure (ASM)**
3. Divide the S into subsets that contain possible values for the best attributes.
4. Generate the decision tree node, which contains the best attribute.
5. Recursively make new decision trees using the subsets of the dataset created in step -3
6. Continue this process until a stage is reached where you cannot further classify the nodes and call the final node as a leaf node.[3]

3)Linear Regression

- Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).
- When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.
- The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.
- The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

- In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).
- It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

- When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.[5]
- Air quality is predicted using the R squared value.

$$R^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

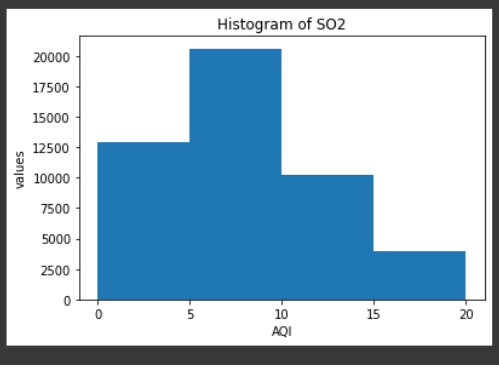
- R square determines the proportion of variance in the dependent variable of the system that can be explained by the independent variable. It is a statistical measure in a regression model. It is also called a coefficient of determination.
- The predicted R square values indicate how well a regression model predicts responses for the given observations. R square value generally lies between -1 to +1.
- In this project, R square value for training and test dataset is calculated using four different regression models. Here in this project R square value of the training dataset is always greater than the test data. If the R square value is near to 1, then the regression model is better for than dataset.
- Root Mean Square Error is the standard deviation (SD) of the prediction errors. Residuals are the measure of how far from the regression line data points are; RMSE tells you how the data is concentrated around the best fit line or a measure of how the residuals are spread out. It is commonly used in forecasting, climatology, and regression analysis to verify experimental results.[2]

Experiments and Inferences

- Redundancy and Correlation Analysis

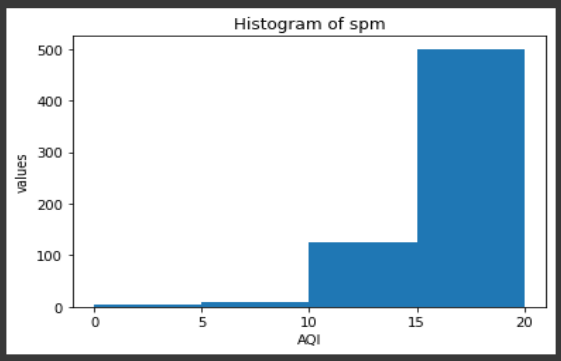
Histogram Of SO2

```
[190] from matplotlib import pyplot as plt
      import numpy as np
      fig,ax =plt.subplots(1,1)

      ax.hist(df['so2'],bins=[0,5,10,15,20])
      ax.set_title("Histogram of SO2")
      ax.set_xticks([0,5,10,15,20])
      ax.set_xlabel("AQI")
      ax.set_ylabel("values")
      plt.show()
```



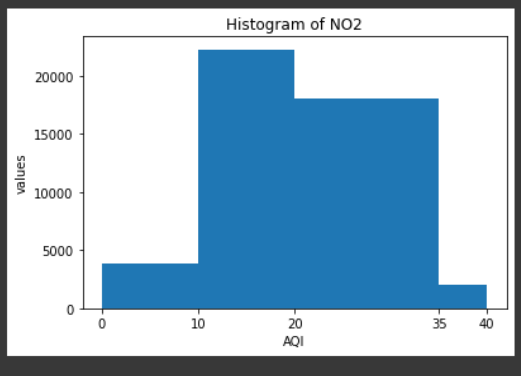
Histogram of spm

```
[192] from matplotlib import pyplot as plt
      import numpy as np
      fig,ax =plt.subplots(1,1)

      ax.hist(df['spm'],bins=[0,5,10,15,20])
      ax.set_title("Histogram of spm")
      ax.set_xticks([0,5,10,15,20])
      ax.set_xlabel("AQI")
      ax.set_ylabel("values")
      plt.show()
```



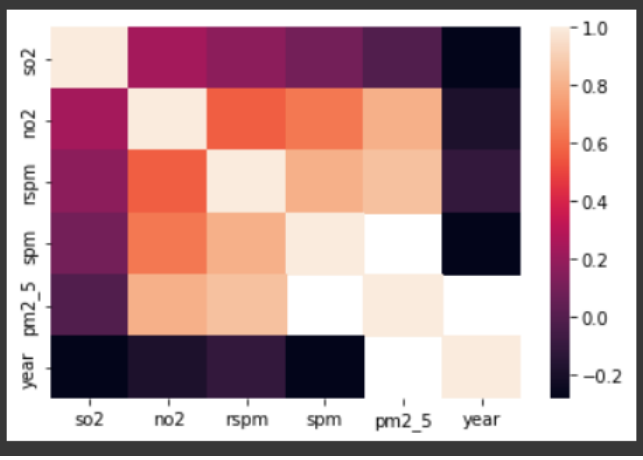
Histogram of NO2

```
[191] # Histogram for NO2
      from matplotlib import pyplot as plt
      import numpy as np
      fig,ax =plt.subplots(1,1)

      ax.hist(df['no2'],bins=[0,10,20,35,40])
      ax.set_title("Histogram of NO2")
      ax.set_xticks([0,10,20,35,40])
      ax.set_xlabel("AQI")
      ax.set_ylabel("values")
      plt.show()
```



Heatmap-

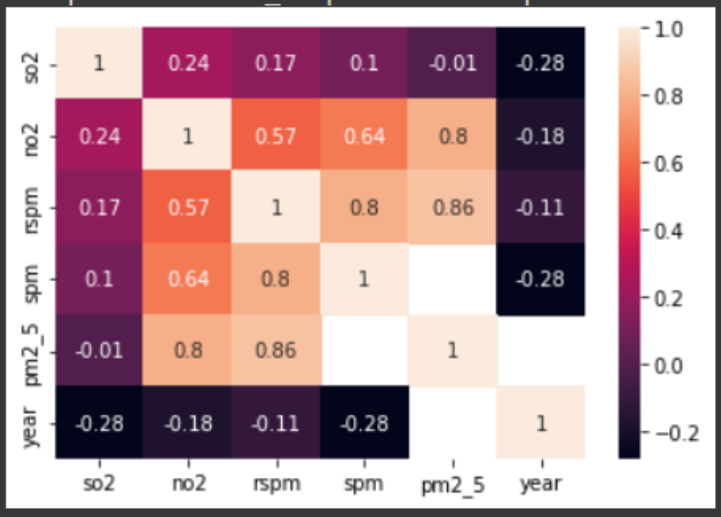
```
sns.heatmap(df.corr(method='pearson'))
plt.show()
plt.savefig("heatmap_pearson.png")
plt.clf()
plt.close()
```



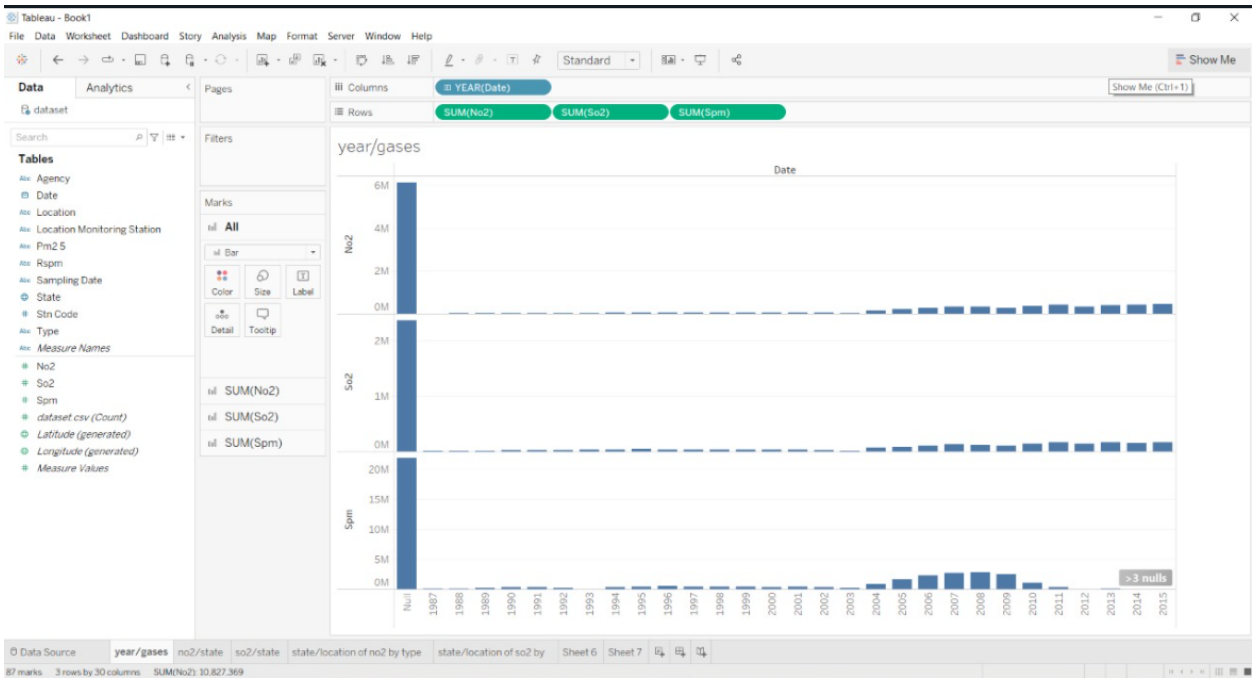
Correlation Matrix-

```
[198] correlation_matrix=df.corr().round(2)
      sns.heatmap(data=correlation_matrix,annot=True)
```

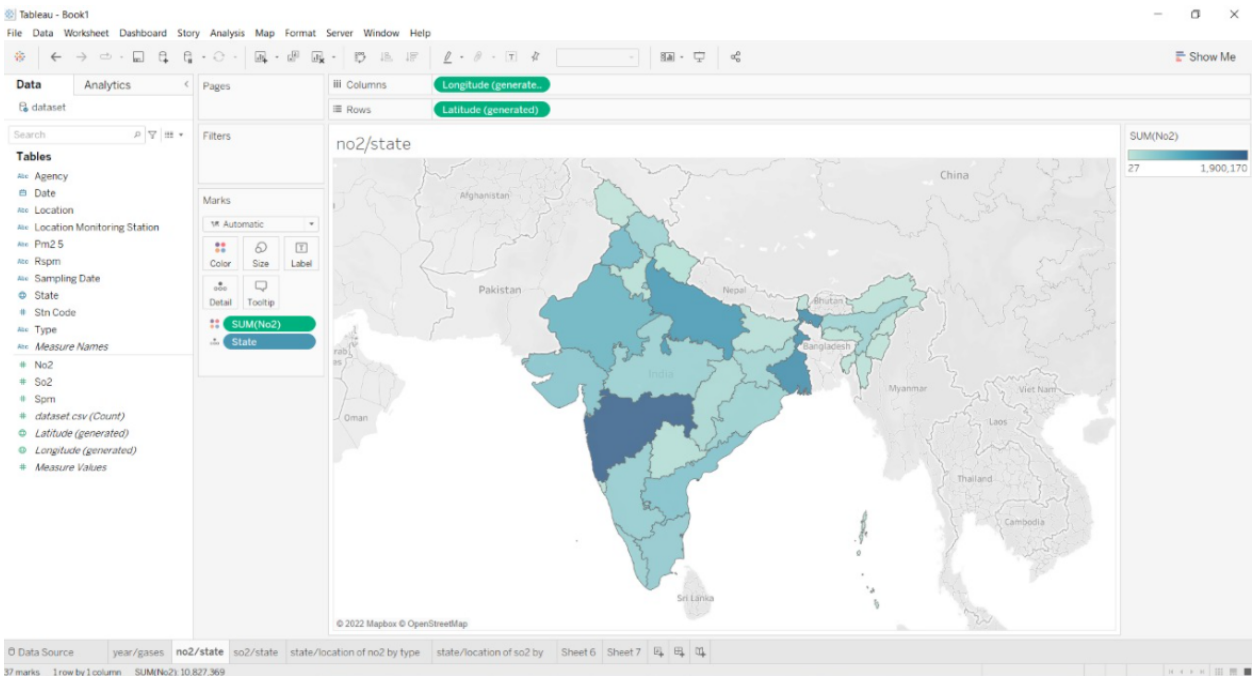
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f81b8d186d0>
```



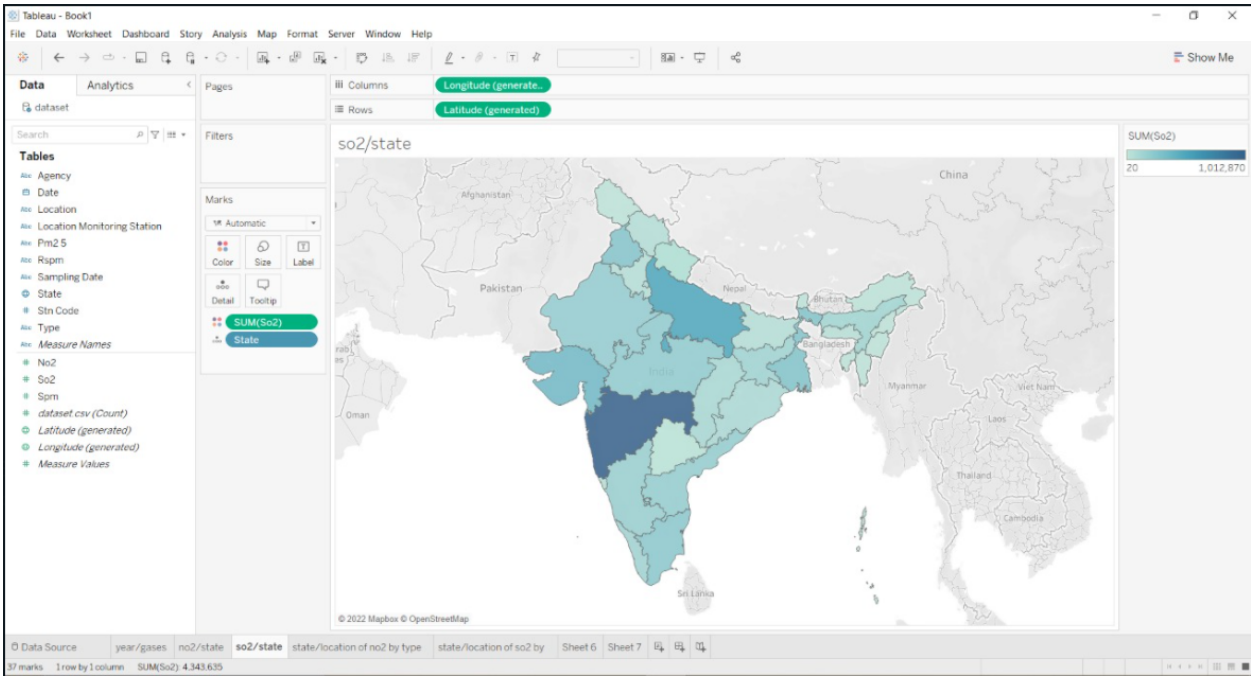
Graphs Made Using Tableau



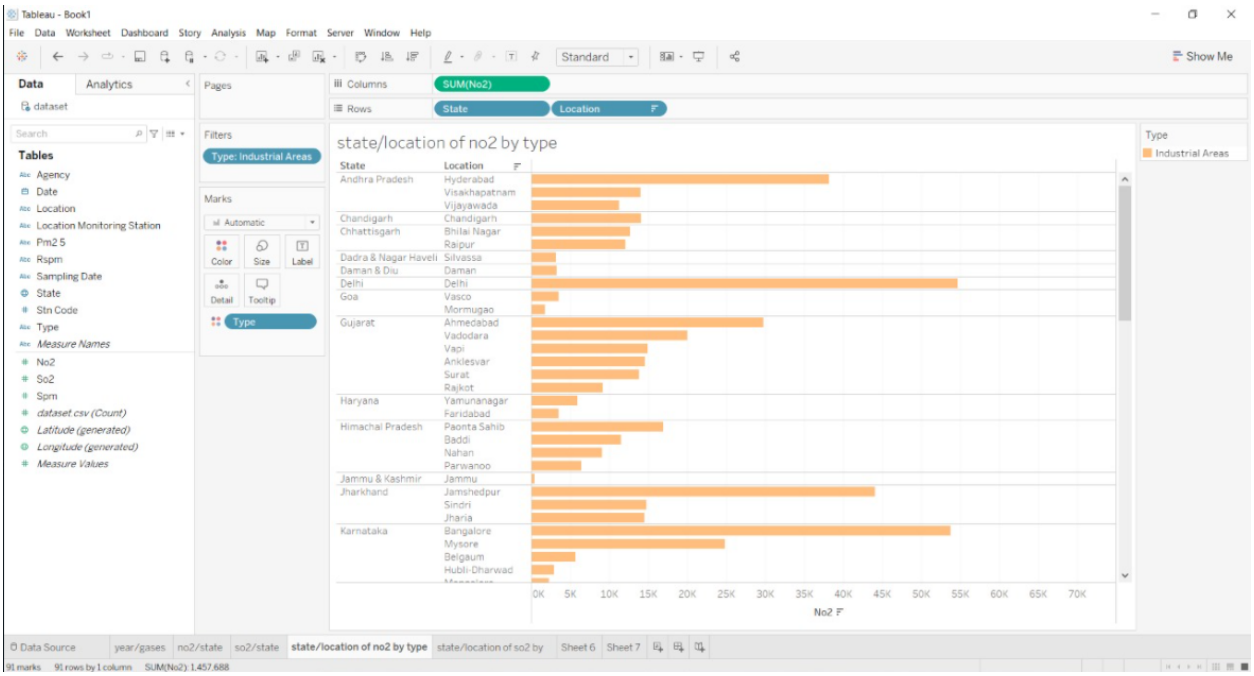
The above graph represents the absolute cumulative amount of gasses in the entire Indian subcontinent



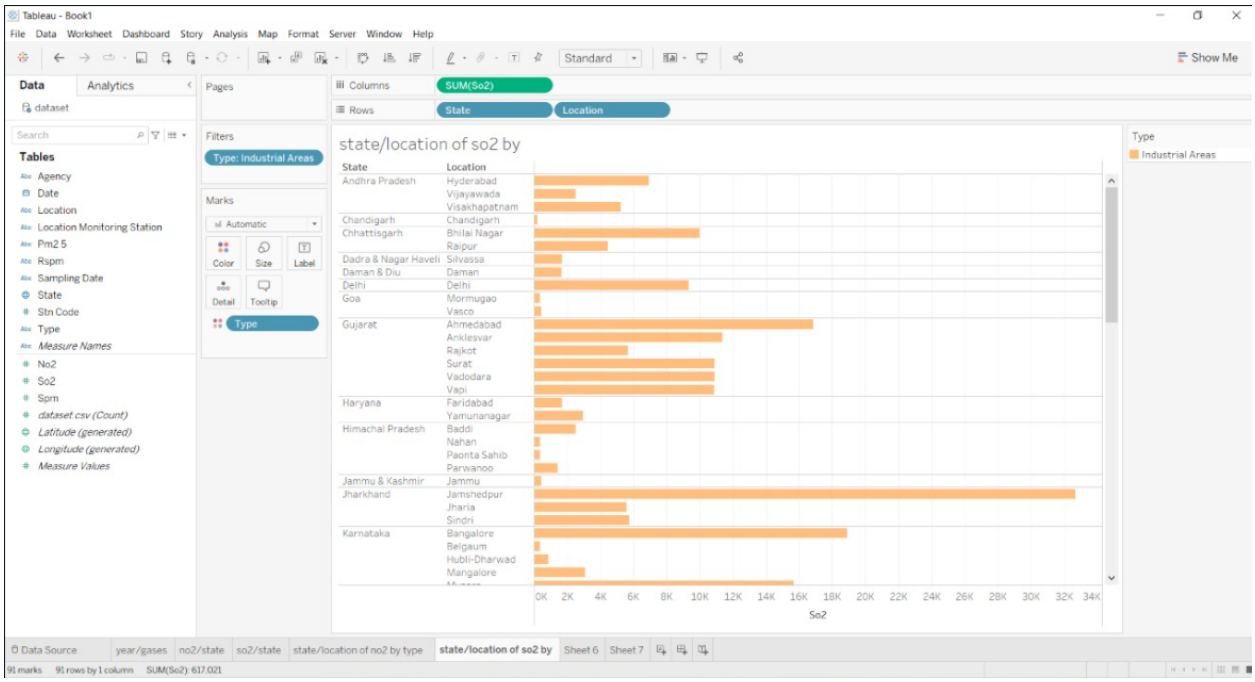
The above graph represents the gross amount of pollution(NO_2) in various states.Darker shade indicates more pollution.



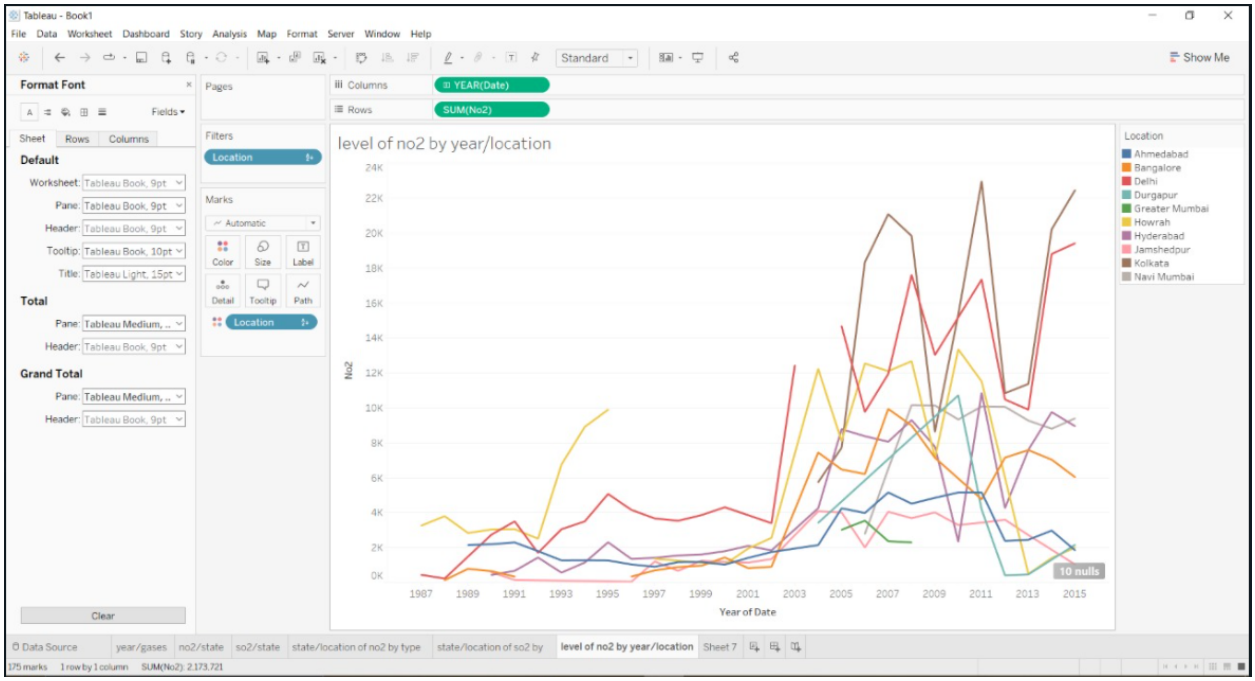
The above graph represents the gross amount of pollution(SO_2) in various states.Darker shade indicates more pollution.



The above graph represents the sum of AQI (NO_2) of various cities in various states of India.(Values are in larger number as it is the absolute sum of AQI but still are approximately very close to the real averages when divided by 1000).

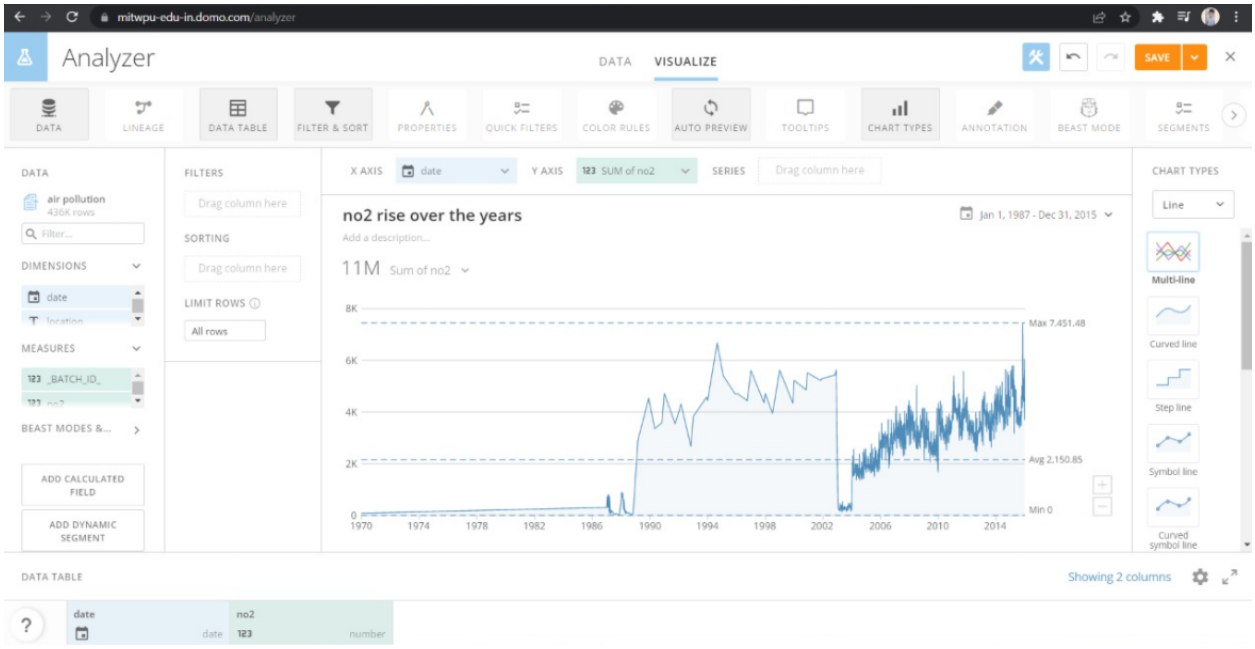


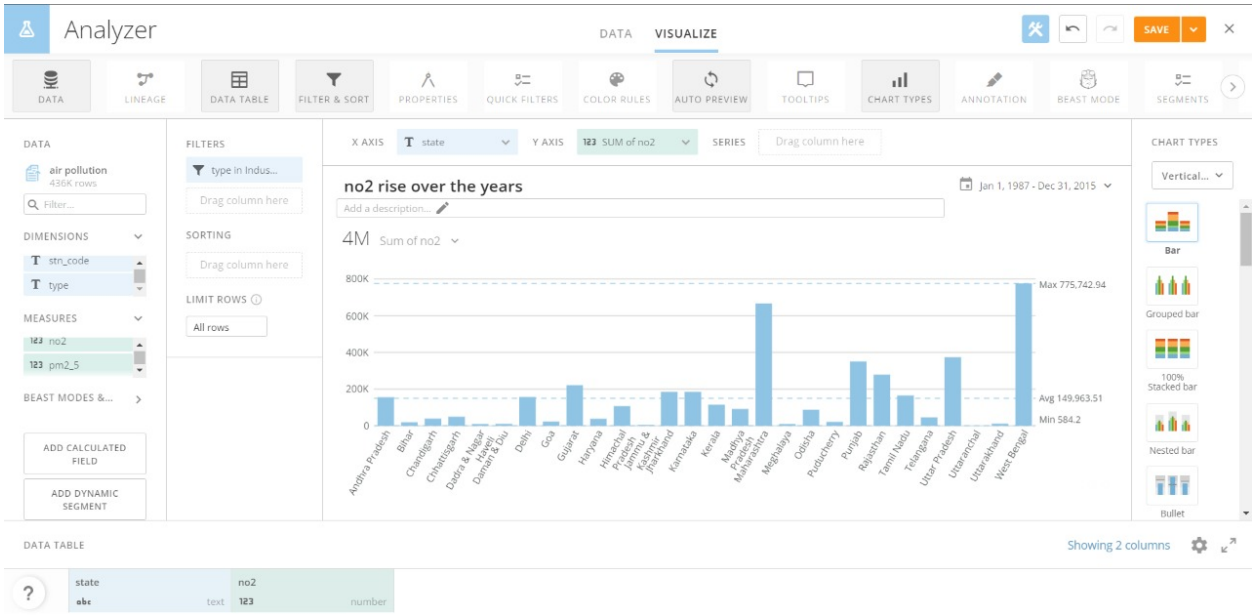
The above graph represents the sum of AQI (SO_2) of various cities in various states of India.(Values are in larger number as it is the absolute sum of AQI but still are approximately very close to the real averages when divided by 1000).



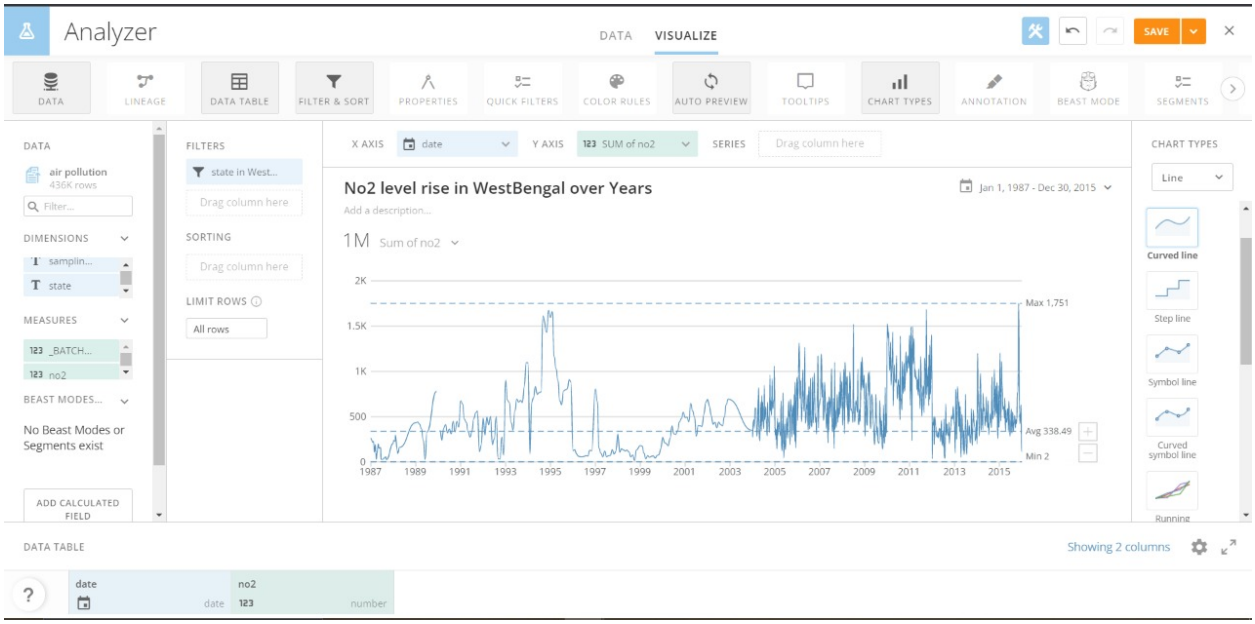
The above graph represents the total level of NO2 in various major cities in India by year

Visual Representation Using a new Third Party Software (DOMO)

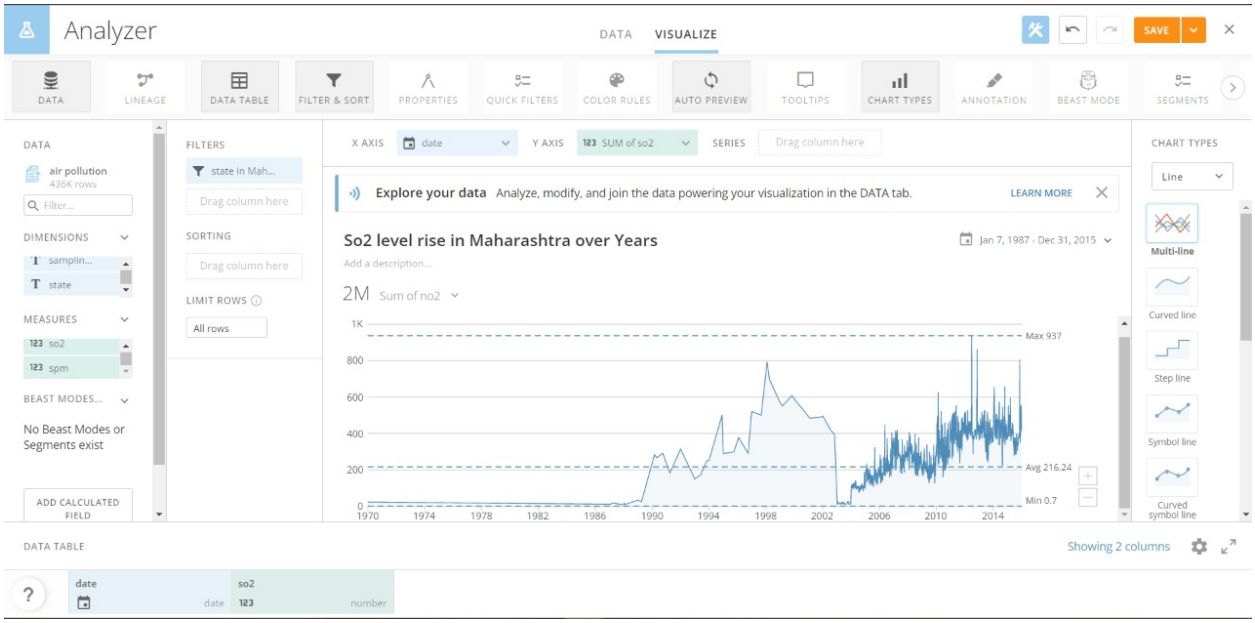




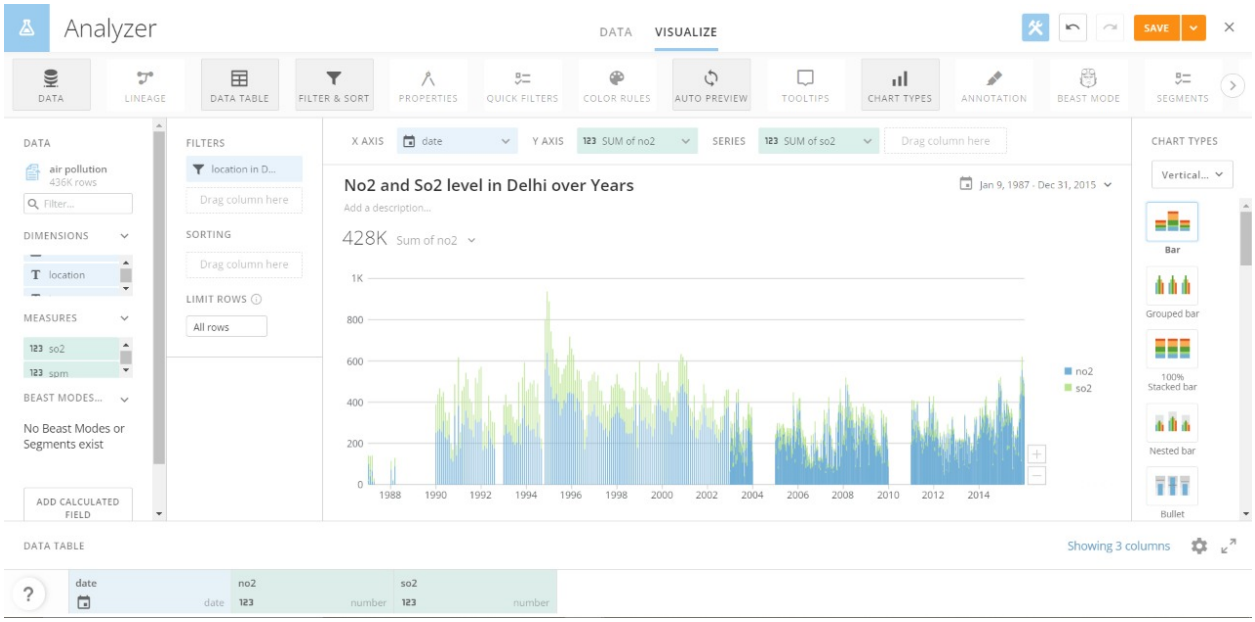
The above graph represents pollution in industrial areas(NO_2)



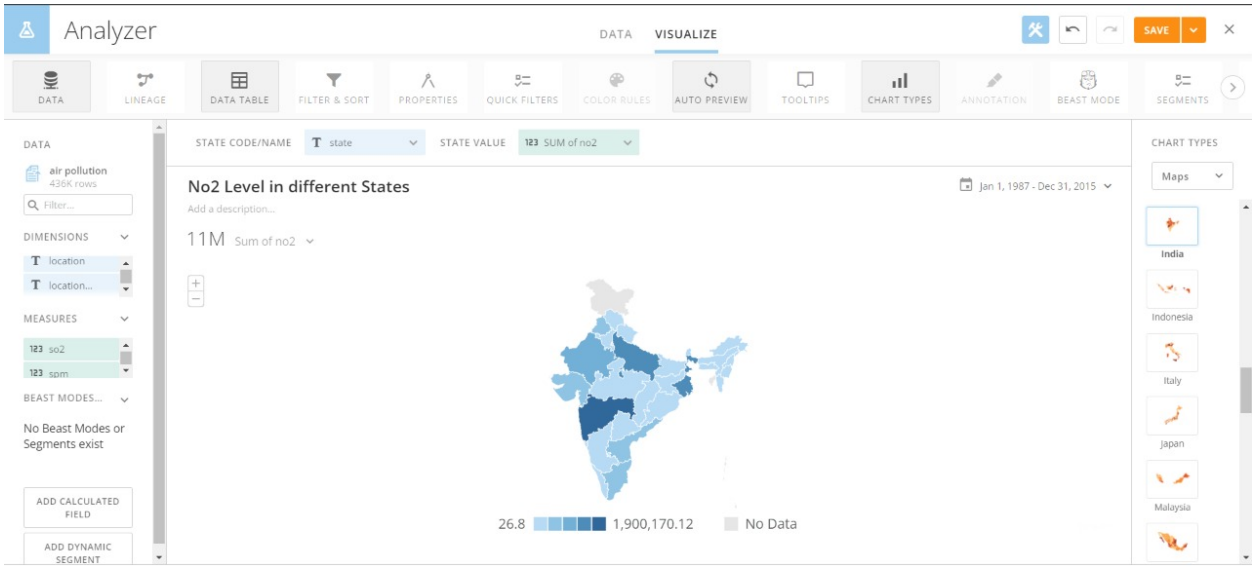
The above graph represents pollution in West Bengal(NO_2)



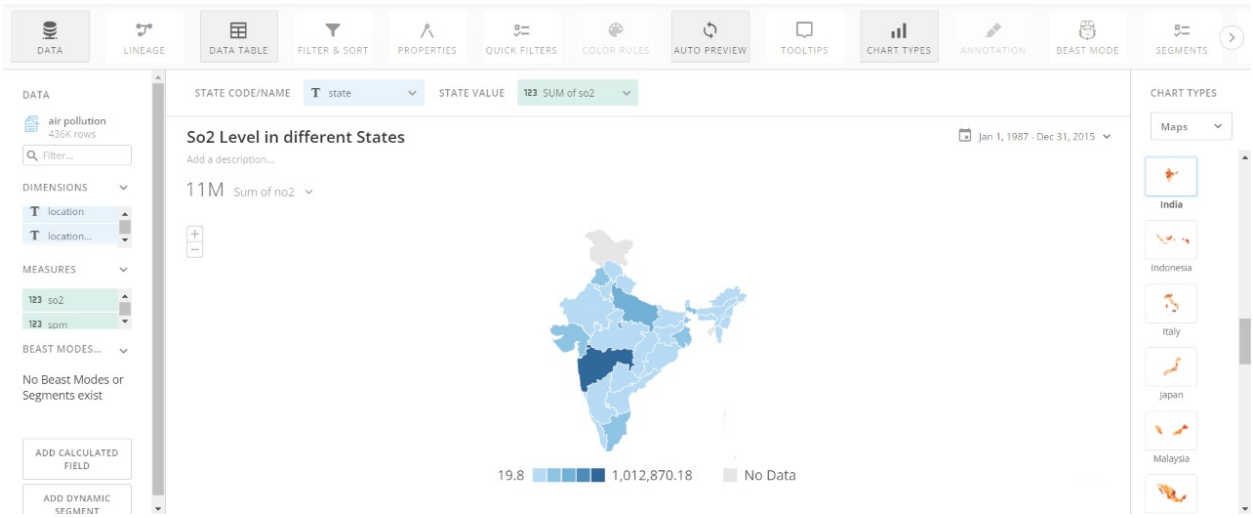
The above graph represents pollution in Maharashtra(SO_2)



The above graph represents pollution in Delhi(NO_2 & SO_2)



The above graph represents the gross amount of pollution(NO_2) in various states.Darker shade indicates more pollution.



The above graph represents the gross amount of pollution(SO_2) in various states.Darker shade indicates more pollution.

Conclusion

Our analysis and inference show efficiently represents the pollution levels in various areas in India.By moving forward with our project if it scaled to a larger extent it will create awareness among people about the air quality degradation and its health effects. We need to support environmentalists and help the government to implement better air quality standards and regulations based on issues of toxic and pathogenic air exposure and health-related hazards for human welfare.

Workload Distribution:	
Preprocessing	Kundan W Anand V Jighnesh S
Data Algorithm	KMeans-Vighnesh S,Anand V Decision tree-Kundan W, Vighnesh S Linear Regression-Amar R Data Visualization(Tableau & DOMO)-Kundan W
Report	Vighnesh S Amar R Anand V

Bibliography

- https://www.orfonline.org/research/tackling-industrial-pollution-in-india-where-is-the-data/
- https://www.teriin.org/article/air-pollution-india-major-issues-and-challenges
- NAMP Air Pollution Data | Kaggle
- Air Pollution in India: Major Issues and Challenges | TERI (teriin.org)
- Industrial air pollution and mortality in the Taranto area, Southern Italy: A difference-in-differences approach - ScienceDirect
- https://www.activesustainability.com/environment/effects-air-pollution-human-health/
- Top 10 Machine Learning Algorithms for Beginners | Built In

Citations(IEEE)

[1]Performance Analysis of KMeans and KMediods Algorithms in Air Pollution Prediction by- S. Suganya, T. Meyyappan, S. Santhosh Kumar ;International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5, January 2020

[2] Air Quality Prediction Through Regression Model By.Aarthi, P.Gayathri, N.R.Gomathi , S.Kalaiselvi , Dr.V.Gomathi;INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 03, MARCH 2020

[3]Study and Analysis of Decision Tree Based Classification Algorithms October 2018; INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING by Harsh Patel,Purvi Prajapati

[4]GnanaSoundari.A Mtech, (Ph.D.), Mrs.J.GnanaJeslin M.E, (Ph.D.), Akshaya A.C. "IndianAir Quality Prediction And Analysis Using MachineLearning." ISSN 0973-4562 Volume 14, Number 11, 2017

[5]McCollister G.M. and Wilson K.R. (2008), "Linear regression model for forecasting daily maxima and hourly concentrations of air pollutants,"Atmospheric Environment.

[6]Rao, S.T., and Zurbenko, I.G.(2014)."Detecting And Tracking Changes in Air Quality using regression analysis". J. Air Waste Manage. Assoc.44: 1089–1092.

[7]RuchiRaturi, Dr. J.R. Prasad "Recognition OfFuture Air Quality Index Using Regression andArtificial Neural Network" IRJET .e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03Mar-2018