

**Mini Project Report**

***on* :**

**“INDIAN AIR QUALITY DATA ANALYSIS”**

**Submitted by**

Rishikesh Mate

**Under the Guidance of**

**Prof. Shilpa Sonawani**

**At**

**School of Computer Engineering and Technology**

**ABSTRACT**

Examining and protecting air quality has become one of the most essential activities for the government 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 are in need of implementing models which will record information about concentrations of air pollutants(so2,no2,etc). The deposition of these harmful gases in the air is affecting the quality of people’s lives, especially in urban areas. Lately, many researchers began to use the Big Data Analytics approach as there are environmental sensing networks and sensor data available. In this report, machine learning techniques are used to predict the concentration of so2 in the environment. Sulphur dioxide irritates the skin and mucous membranes of the eyes, nose, throat, and lungs. Models in time series are employed to predict the so2 readings in nearing years or months.

**CONTENTS**

**1. Introduction……......……………………………………………………….................. 3**

1.1Motivation .………………………………………………………….…..... 3

1.2 Problem Definition...……………………………………………................ 4

1.3Objectives .……………………………………….………………….…..... 4

**2. Tools and Technologies…….……………………………………………………….…. 5**

**3. System Architecture..……..………………………………………………………....... 6**

**4. Dataset Schema…..……...……………………………………………………….......... 7**

**5. Data Preprocessing tasks performed …………..…………………...……………… 9**

**6. Exploratory data analysis….………………………………………………………... 13**

**8. Model building using ML Algorithms ….……...…………………………………... 18**

**9. Output Analysis…………………………….......………..…………………………... 20**

**10. Visualizations ………………………..…….…………………………………………21**

**11. Conclusion & Future Work…………………………………………..……………... 24**

**12. References in IEEE Format….…………………………………….………………... 25**

1. **INTRODUCTION**

Air pollution occurs when harmful or excessive quantities of substances including gases, particles, and biological molecules are introduced into the Earth's atmosphere. Air pollution in India is a serious issue, ranking higher than smoking, high blood pressure, child and maternal malnutrition, and risk factors for diabetes. At least 140 million people breathe air 10 times or more over the WHO safe limit and 13 of the world's 20 cities with the highest annual levels of air pollution are in India. Air pollution contributes to the premature deaths of 2 million Indians every year. In urban areas, most emissions come from vehicles and industry, whereas in rural areas, much of the pollution stems from biomass burning for cooking and keeping warm. In autumn and winter months, large scale crop residue burning in agriculture fields – a low cost alternative to mechanical tilling – is a major source of smoke, smog and particulate pollution.

**1.1 MOTIVATION**

As the largest growing industrial nation, India is producing record amount of pollutants specifically Co2, pm2.5 etc and other harmful aerial contaminants. Air quality of a particular state or a country is a measure on the effect of pollutants on the respected regions, as per the Indian air quality standard pollutants are indexed in terms of their scale, these air quality indexes indicates the levels of major pollutants on the atmosphere. There are various atmospheric gases which causes pollution on our environment. Each pollution has individual index and scales at different levels. The major pollutants Such as (no2, so2, rspm, spm) indexes AQI is acquired, with this individual AQI, the data can be categorized based on the limits. We collected the data from the Indian government database, which contains pollutant concentration occurring at various places across India. We start by calculating the individual index of the pollutant for every available data point and find their respective AQI for the region. We have designed a model to predict the air quality index of every available data point in the dataset, our model is capable of forecasting the air quality of India in any given area.

**1.2 PROBLEM DEFINITION**

In developing countries like India, the rapid increase in population and economic upswing in cities have led to environmental problems such as air pollution, water pollution, noise pollution and many more. India has made it to the headlines when it comes to air pollution. The growing air pollution in the country has been one of the major concerns for both the government and the citizens. Most of the population is not even aware of the colossal harm caused by the air they breathe everyday.

Air pollution in India is a serious issue, ranking higher than smoking, high blood pressure, child and maternal malnutrition, and risk factors for diabetes. There has been increased public awareness about the same in our country. Global warming, acid rains, and an increase in the number of asthma patients are some of the long-term consequences of air pollution.

Precise air quality forecasting can reduce the effect of maximal pollution on the humans and biosphere as well. Hence, enhancing air quality forecasting is one of the prime targets for the society. We thus propose a solution to help analyze the trend of AQI across India and also predicting AQI for future years.

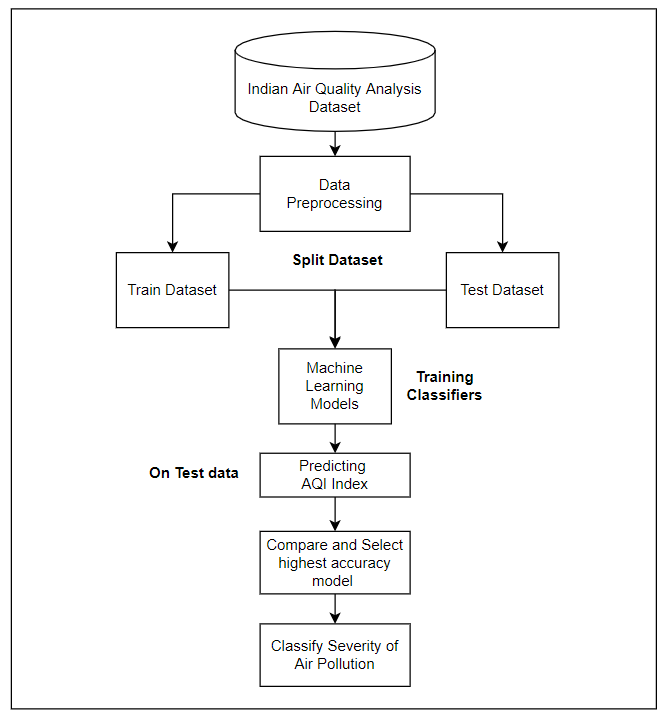
**1.3 OBJECTIVE**

By predicting the air quality index, we can backtrack the major pollution causing pollutant and the location affected seriously by the pollutant across India. With this forecasting model, various knowledge about the data is extracted using various techniques to obtain heavily affected regions on a particular region(cluster). This gives more information and knowledge about the cause and seniority of the pollutants.To calculate a location's Air Quality Index and determine the severity of air pollution in that region.

**TOOLS AND TECHNOLOGIES**

* WINDOWS 10
  + Jupyter
  + Tableau
  + seaborn
  + matplotlib
  + sklearn

1. **SYSTEM ARCHITECTURE**



**4. DATASET SCHEMA**

This data is a combined(across the years and states) version of the [Historical Daily Ambient Air Quality Data](https://data.gov.in/catalog/historical-daily-ambient-air-quality-data) released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP).

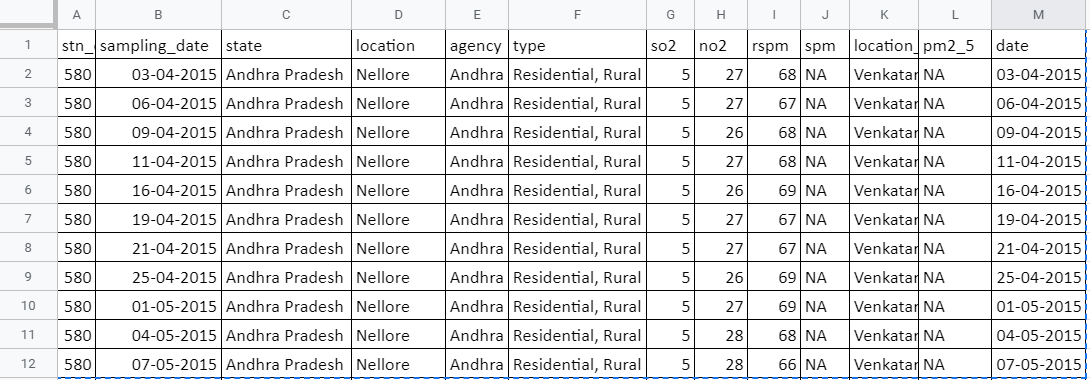
[data.gov.in](https://data.gov.in/)

**ROWS :** 435743 **COLUMNS :**13

The dataset contains the following features :

1. stn\_code : Station code. A code given to each station that made the measurements.
2. sampling\_date : The date when the data was recorded.
3. state : It represents the states whose air quality data is measured.
4. location : It represents the city whose air quality data is measured.
5. agency : Name of the agency that measured the data.
6. type : The type of area where the measurement was made.
7. so2 : The amount of Sulphur Dioxide measured.
8. no2 : The amount of Nitrogen Dioxide measured.
9. rspm : Respirable Suspended Particulate Matter measured.
10. spm : Suspended Particulate Matter measured.
11. location\_monitoring\_station : It indicates the location of the monitoring area.
12. pm2\_5 : It represents the value of particulate matter measured.
13. date : It represents the date of recording (It is cleaner version of 'sampling\_date' feature)

The fragment of dataset is shown in the Fig. 1 as follows :

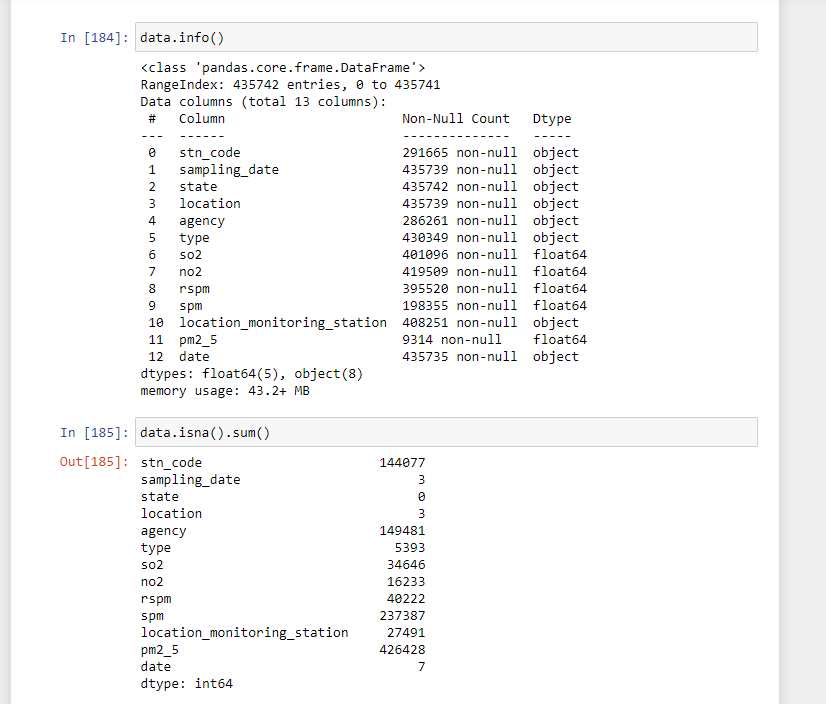


**5. DATA PREPROCESSING**

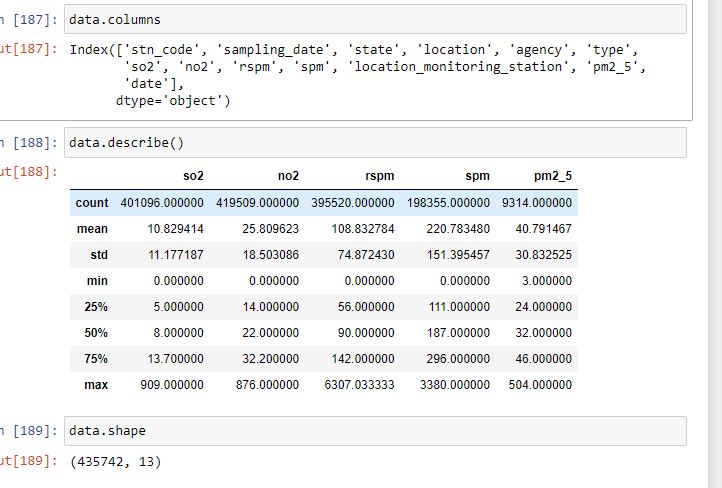
Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

In the real world data are generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. Noisy: containing errors or outliers. Inconsistent: containing discrepancies in codes or names.

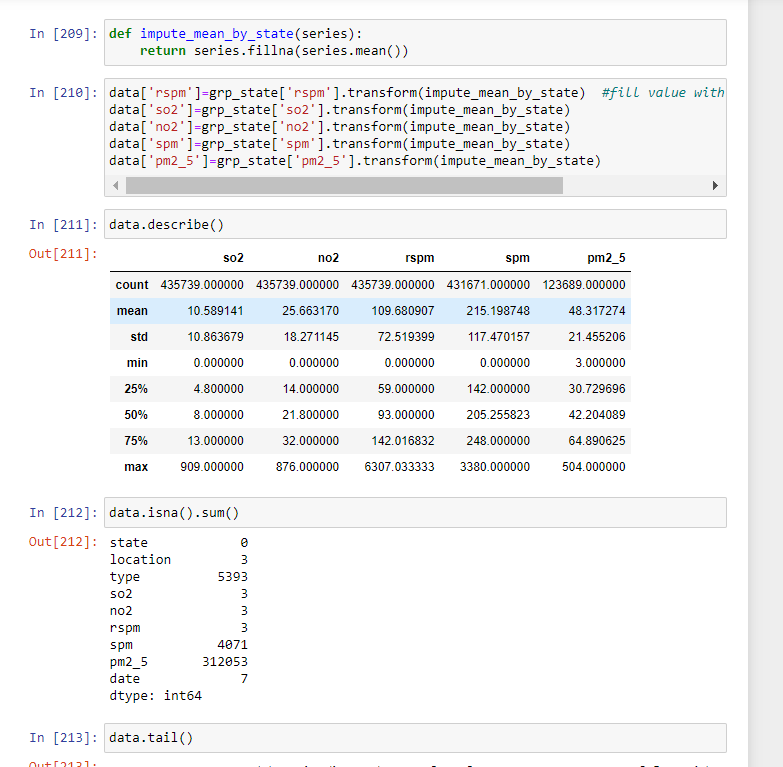
* **DATA INFO AND MISSING VALUES CHECK**

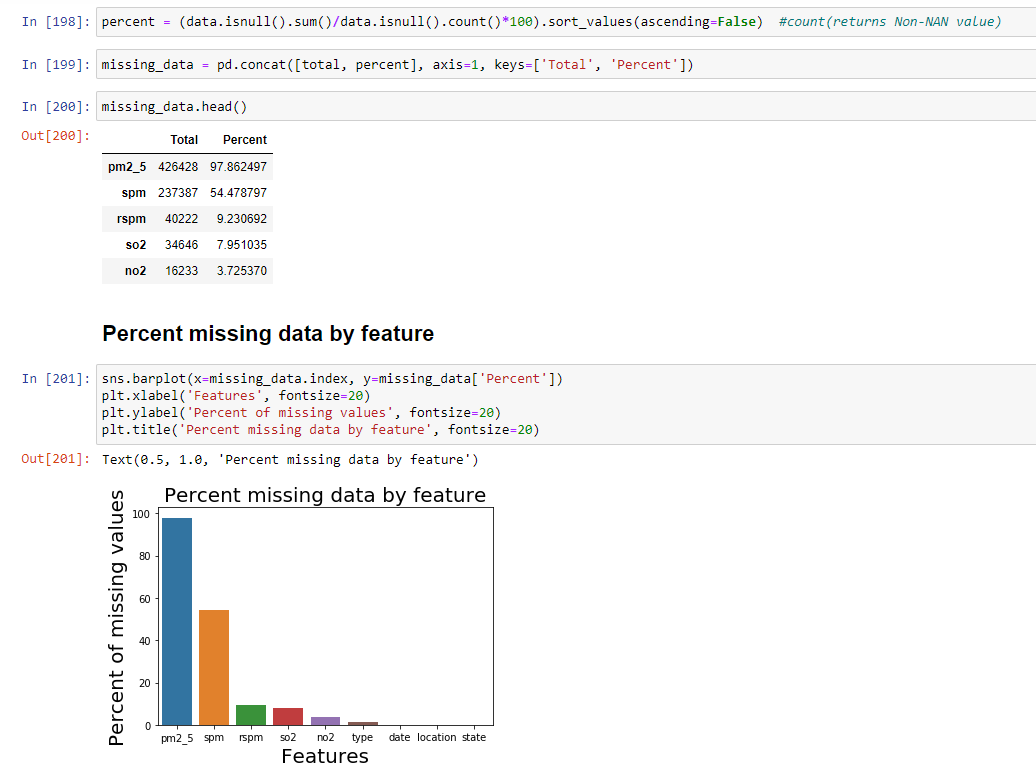
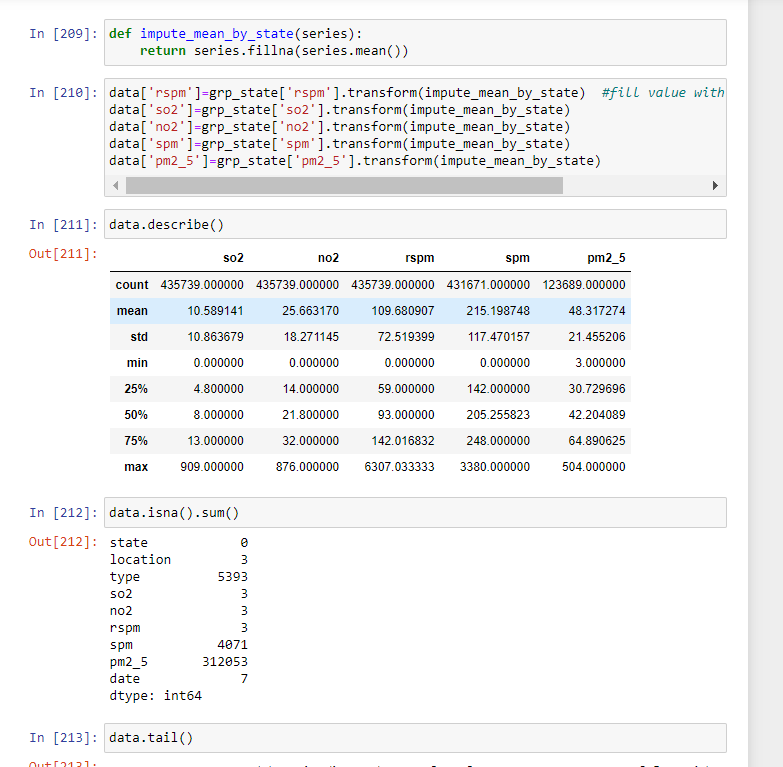
****

* **DATA DESCRIPTION - INTERQUARTILE RANGE**

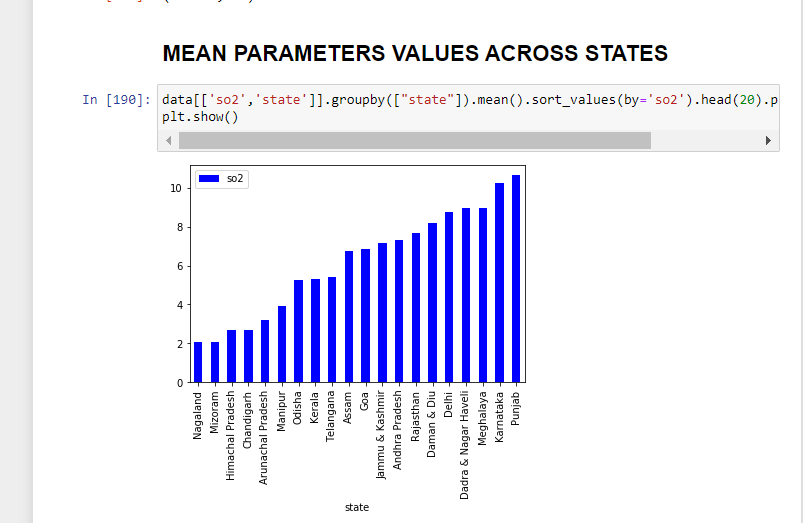
****

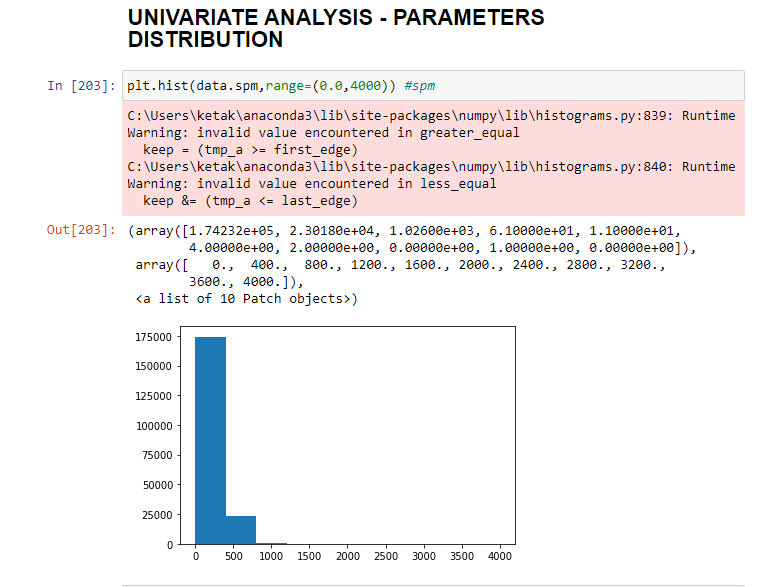
* **HANDLING MISSING VALUES**

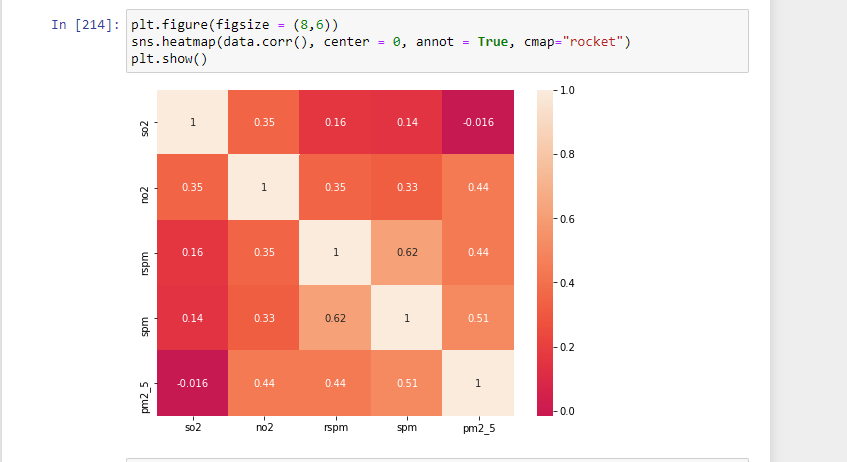
****

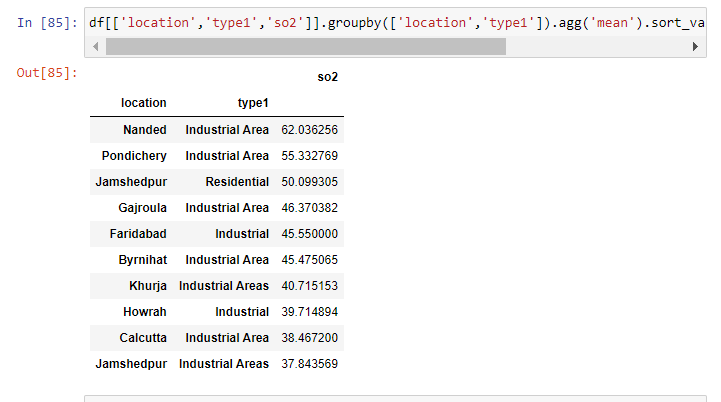
****

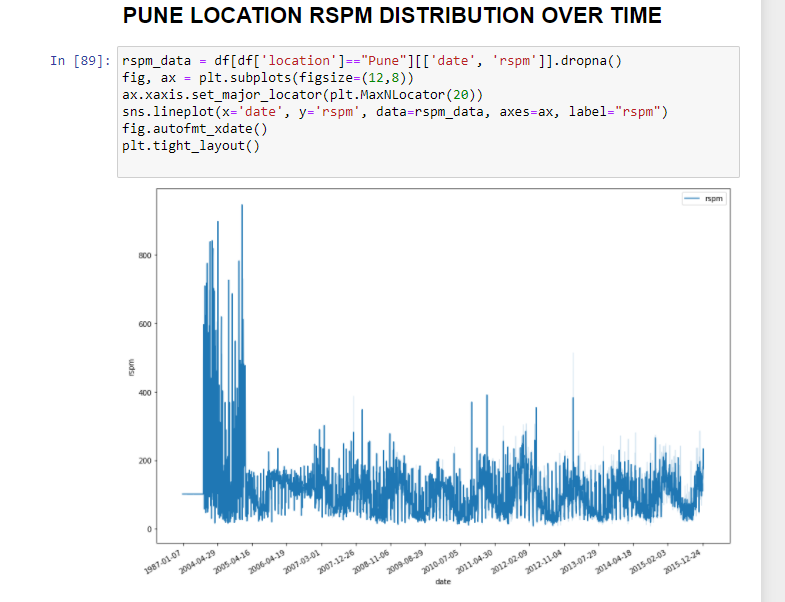
**6. EXPLORATORY DATA ANALYSIS**

****

****

****

****

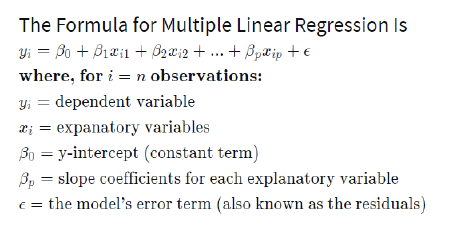
****

**8. MODEL BUILDING USING ML ALGORITHMS**

**8.1 LINEAR REGRESSION**

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.



**8.2 LOGISTIC REGRESSION**

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression)

**8.3 RANDOM FOREST CLASSIFIER**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees

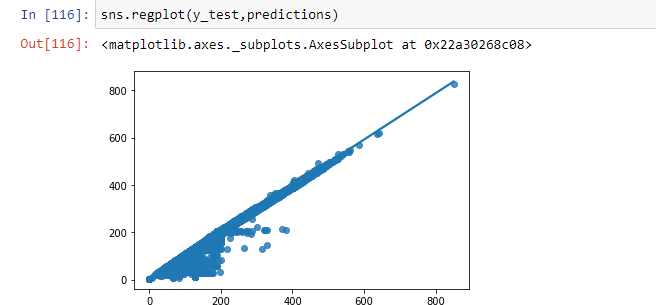
**9. OUTPUT ANALYSIS**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Linear Regression | 97.91 |
| Logistic Regression | 70.13 |
| Random Forest | 99.97 |

Training Features: SOi, NOi, RSPMi, SPMi (the calculated pollution indices)

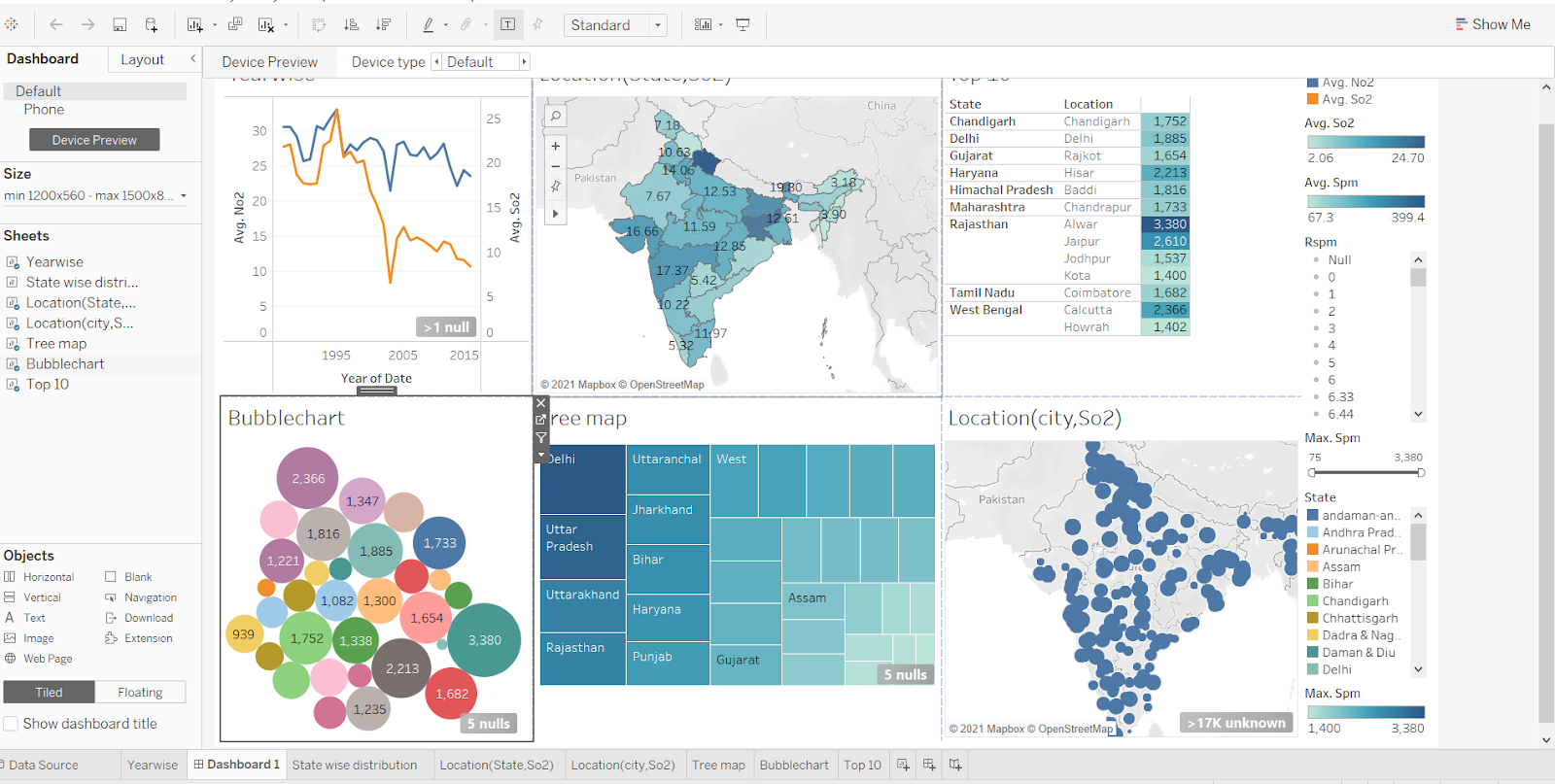
Target Feature: AQI (calculated Air Quality Index)

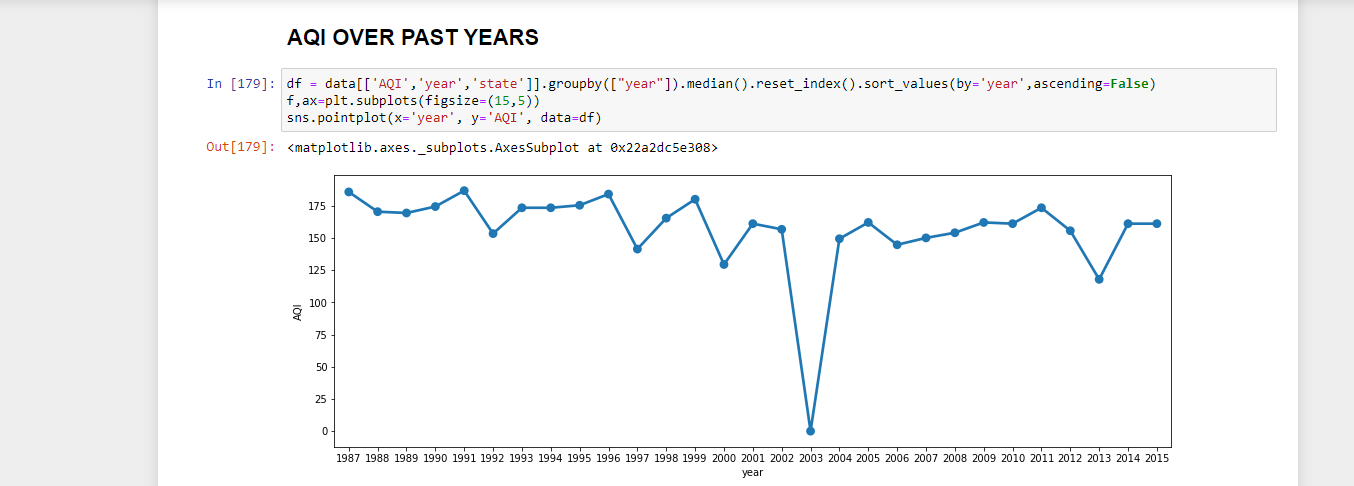
The data was split into training and testing data, where training data used 80% of the dataset and testing data used 20% of the total dataset.

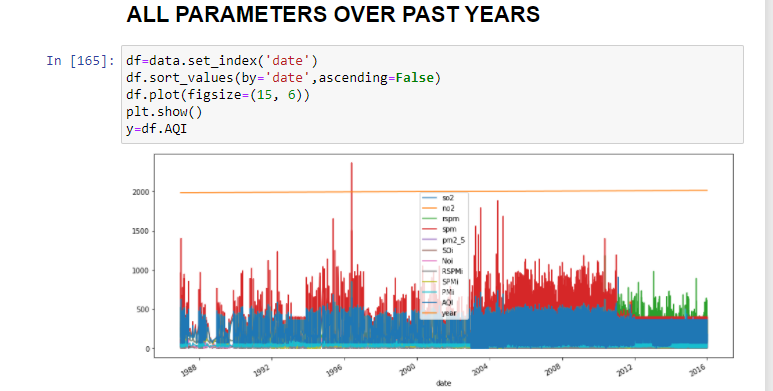


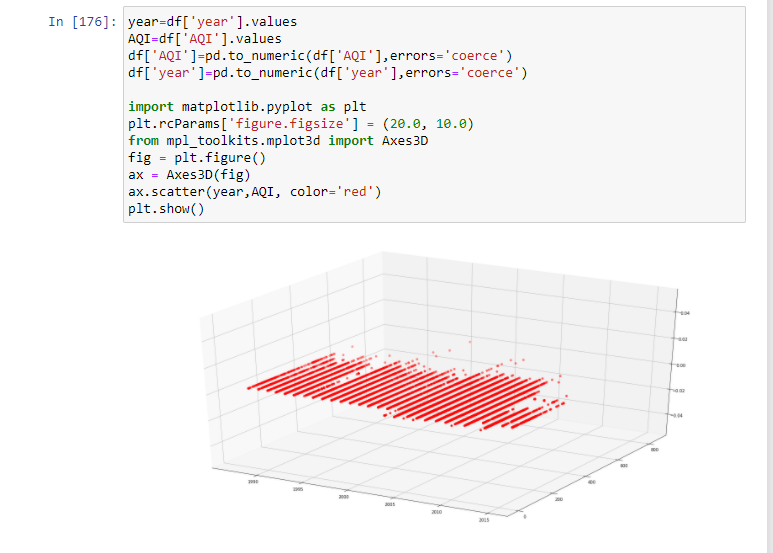
**10. VISUALIZATIONS**

**TABLEAU**

****

****

****

****

**12. CONCLUSION AND FUTURE WORK**

Since our model is capable of predicting the current data with 97.91% accuracy it will successfully predict the upcoming air quality index of any particular data within a given region. With this model we can forecast the AQI and alert the respected region of the country. Also a progressive learning model capable of tracing back to the particular location needed attention provided the time series data of every possible region needed attention.

The air quality information utilized in this report originates from the Indian air quality checking and investigation stage, and incorporates the normal every day fine particulate issue (PM2.5), inhalable particulate issue (PM10), ozone (O3), CO, SO2, NO2 fixation and air quality record(AQI). In order to predict air quality, pm2\_5 is also an important attribute. The values of this must be recorded in future as these particulates are responsible for various health effects including cardiovascular effects such as cardiac arrhythmias and heart attacks, and respiratory effects such as asthma attacks and bronchitis.

The dataset lacks information on the cities. If we predict for the entire state, it won’t be helpful. So, we can include extended parameters and use the model for further AQI forecasting and classification.

**13. REFERENCES**

1. Bhalgat, Pooja & Bhoite, Sachin & Pitare, Sejal. (2019). Air Quality Prediction using Machine Learning Algorithms. International Journal of Computer Applications Technology and Research. 8. 10.7753/IJCATR0809.1006.
2. Soundari, A. G., Jeslin, J. G., & Akshaya, A. C. (2019). INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING. *International Journal of Applied Engineering Research*, *14*(11).