Semester: Fall 2020 - 2021

Course Name and Code: CSE5003- Database Systems: Design and Implementation

Slot: E1

Database Systems: Design and Implementation J Component

Review 3

Date: 24th October 2020

Team number: 9

Team members: Harpal Kaur Dhindsa (20MCB0008)

Gargi Nitin Gore (20MCB0014)

Topic: Air Quality Database Management System and Prediction using Machine Learning

Introduction:

Today, for the fruitful execution of the work of monitoring air by checking air quality data framework is fundamental. Before some 30-odd years, when in the initial screening of air quality estimation, results were composed on paper in the table. Expanding the utilization of computer data has begun to digitize [1]. "Information base" or tables were first composed on the magnetic tape, at that point on floppy disks, CDs, and DVDs. Today, air quality checking is difficult to perform without the utilization of PCs and data frameworks.

Air pollution which is detrimental to people's health is a widespread problem across many countries all around the world. For people living in countries or cities with higher rates of air pollution, it becomes essential for them to be aware of the air quality around them [1]. It does become a part of their routine to check the air quality in the morning before leaving the house, similar to checking the weather or the traffic. Developing better air quality forecast approaches is a significant exploration issue. Existing techniques frequently center around the expectation of air contamination fixations, which isn't as natural to general society as the air quality levels [2]. For analysing the patterns of air quality, air zones can be defined as areas that typically exhibit similar air quality characteristics, issues, and trends, and are the foundation for monitoring, reporting, and taking action.

The system should be able to ensure that air quality data is available when required and where required and to defined quality standards. The system should enable the user to implement best practice quality assurance procedures, which provides value and confidence in data – critical where air quality data is to be used to support decision-making or establish regulatory compliance.

There are a few key problems that have been identified in the current air quality database management scenario, given as following. Lack of assembly in the architecture of the overall system of air quality and other datasets. The inequality of the datasets in terms of the standard formats. And the lack of or unpredictable metadata [3].

Air quality management is a multifaceted issue. Unlike the numerous government capacities, similar to tax assessment or foundation improvement, there is no noteworthy government record accessible to execute the air quality upgrade projects, and the executives of air quality requires association between various government regions, (for example, transportation, water assets, resources, metropolitan planning) [2].

Furthermore, the science of air pollution formation is composite and now and then has counterintuitive descriptions, and new bewildering issues seem to produce frequently. The system has sometimes led to overlook the key emission sources and accounting for cultural trends. The intricacy of pollutant foundation, interactions, and health impacts, needs to be observed. Accordingly planning tools and predictive abilities could be improved [4].

Computational capacities can permit us to show definite frameworks nearly continuously, and investigate the future with more precision. A significant number of these tools are accessible for creating to numerous nations, as they are starting their endeavours at building up an air quality the board framework [4]. Associations in most nations are creating such systems have at some level an air checking network, and the administrations, furthermore, universities all through the world can run complex meteorological modelling for expectation of air quality levels. All things reflected, there are a portion of a similar key inadequacies of these creating and existing projects that still block our capacity to assemble a deductively solid air quality management plan [5].

For example, Chongqing, Guadalajara, Sao Paulo, and Shanghai, places which have primarily point source inventories with little ability to understand overall sources of pollution or predict future trends [5]. They have no way to update or progress their pollution-related information, and when a new pollutant of significance enters the depiction, a whole separate inventory or plan is set up without satisfactory integration [5]. Often satisfactory as a first–look analysis are too slender in their design and become accountabilities when trying to develop an inclusive plan for the individual region.

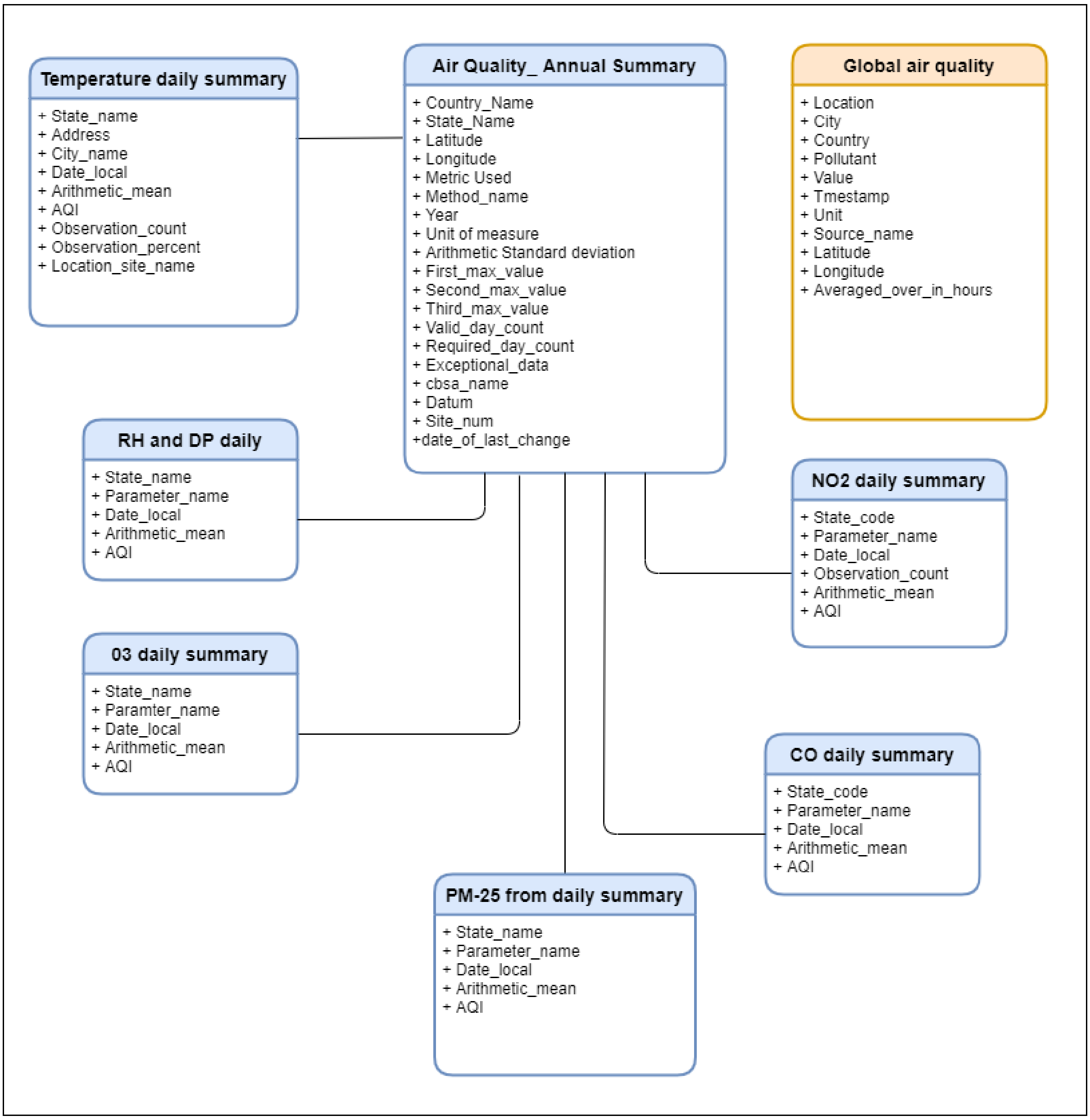
To reduce the pollution level, development of a multi-sensor figures fusion model that detects and predicts the most severe gas can be done. An effective algorithm to cluster the multiple sensors data to be used to group the data and partition the data [5].

Machine Learning gives one methodology that can offer new open doors for the forecast of air contamination. There are anyway various Machine Learning draws near and distinguishing the best one for the current issue is frequently testing. The fundamental spotlight will be on investigating the appropriate ML strategies that will help in better determining of the contamination fixation.

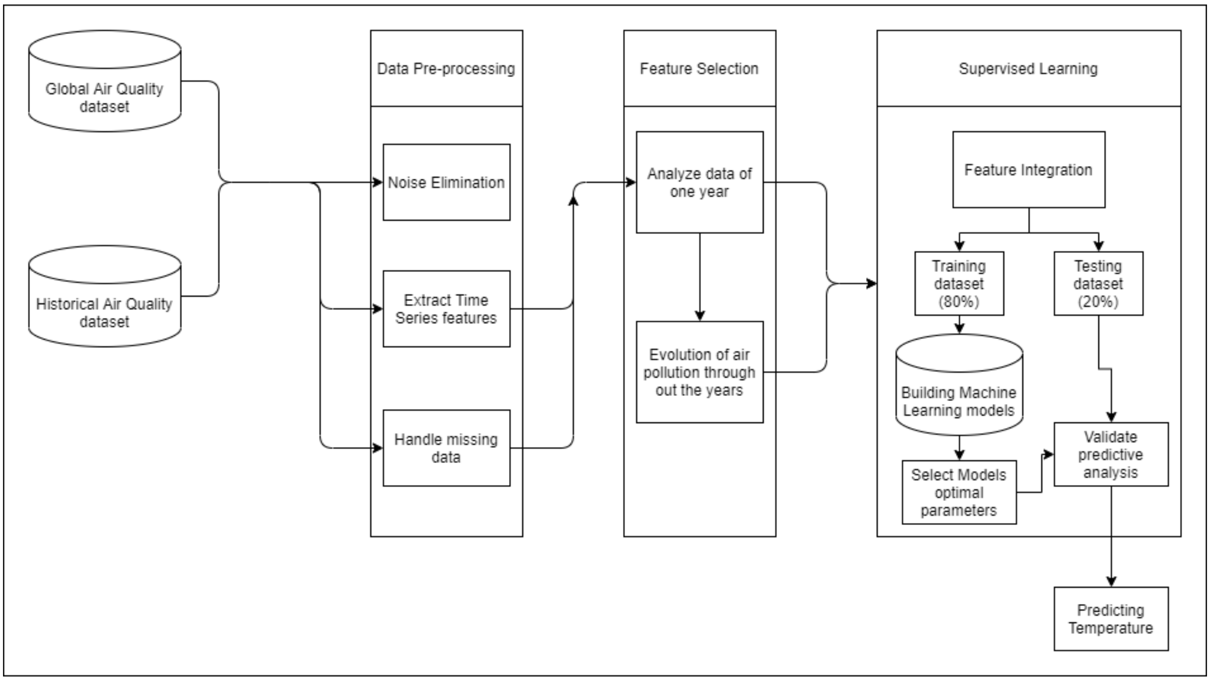
The focus on prediction of the air pollution level of a particular region could be by using certain parameters like PM10, PM 2.5, SO2, NO2, Benzene, CO and O3(ozone). The data analysis could be done by using following machine learning techniques. Algorithms like Multi-linear Regression, SVM (support vector machines) and Random Forest [6]. Which could serve as a vital reference for local government agencies in evaluating present and making future air pollution policies. The consideration of factors to predict the pollution of air like temperature, minimum temperature, maximum temperature, wind speed and relative humidity and several other features have been made [6].

To foresee the concentrations of PM2.5 from wind (speed and bearing) and precipitation levels, the arrangement model can be utilized [7]. In earlier exploration it has come about with a high dependability in the arrangement of low (<10 μg/m3) against high (>25 μg/m3) and low (<10 μg/m3) against moderate (10–25 μg/m3) convergences of PM2.5. Though, a relapse investigation of regression analysis proposes a superior expectation of PM2.5 when the climatic conditions are getting more extraordinary (solid winds or significant levels of precipitation) [7]. It is seen that having high connection amongst assessed and genuine information for a period arrangement examination during the monsoon season. Consequently, the utilization of measurable models dependent on ML is applicable to anticipate PM2.5 contemplations from meteorological information [7].

Schema diagram:



Block diagram:


Global Air Quality dataset
Historical Air Quality dataset
Data Pre-processing
Noise Elimination
Extract Time Series features
Handle missing data
Feature Selection
Analyze data of one year
Evolution of air pollution through out the years
Supervised Learning
Feature Integration
Training dataset (80%)
Testing dataset (20%)
Building Machine Learning models
Select Models optimal parameters
Validate predictive analysis
Predicting Temperature

Taking the Historical Air Quality records and Global Air Quality records as input to form the dataset.

The AQS Data Mart is a database containing all of the information from AQS. It consists of every measured rate the EPA has recorded via the national ambient air monitoring program. It also comprises of the related aggregate values calculated by EPA (by 8-hour, by daily, by annual, etc.). The AQS Data Mart is a replica of AQS prepared once per week and is accessible to the public through web-based applications. The projected users of the Data Mart consist of air quality data analysts in the supervisory, theoretical, and medical research communities.

The records consist of following aggregation levels of data (and key metrics in each) are:

- Sample Values comprises of 2.4 billion values starting from 1957, national uniformity initiates in 1980, data for 500 elements is regularly collected

- The sample value transformed to standard units of measurement, generally averages as reported to EPA on hourly basis or sometimes by daily average basis
- The date and time is recorded as per Local Standard Time (LST) and GMT timestamps
- Which Measurement method is opted
- Each monitor calculates the Daily Summary Values each day
- Number of observations counted
- The arithmetic mean of the observations
- The AQI (air quality index) where relevant
- Each monitor calculates the Annual Summary Values each year
- Arithmetic Mean and Standard Deviation
- The measured parameter's code
- POC (Parameter Occurrence Code) to distinguish from different samplers at the same site
- Latitude
- Longitude
- Measurement method information

OpenAQ is an open-source venture to apparent live, real-time air quality data from all around the world. The data embraces air quality measurements from 5490 different locations ranging over 47 countries. Scientists, scholars, inventors, and other people can make use of this data to comprehend the quality of air around them at the given moment. The dataset only takes account of the most present-day measurement accessible for the given location, no historic data.

Implementations and Outputs:

I. Analysis

Including BigQuery to the Jupyter kernels, it allows the users to explores massive databeses or datasets which gives endless possibilities and discoveries to be obtained.

In this first part, we tried to understand how polluted was the air all over the US during 2016.

To do so, we extract the average AQI for each pollutant for each county, which translates to a groupby(county) in SQL or pandas. Since every table comprises the statistics about a particular pollutant, we used JOIN.

To retrieve some useful insights clustering is the obvious answer in the context. Hence, making use t-SNE. The output obtained are 4 clusters, on the plot 4 discrete parts (essentially top, bottom, left, right), the cluster at the top has bit of a complicated shape and to extract we divided it into two parts and applied join on them.

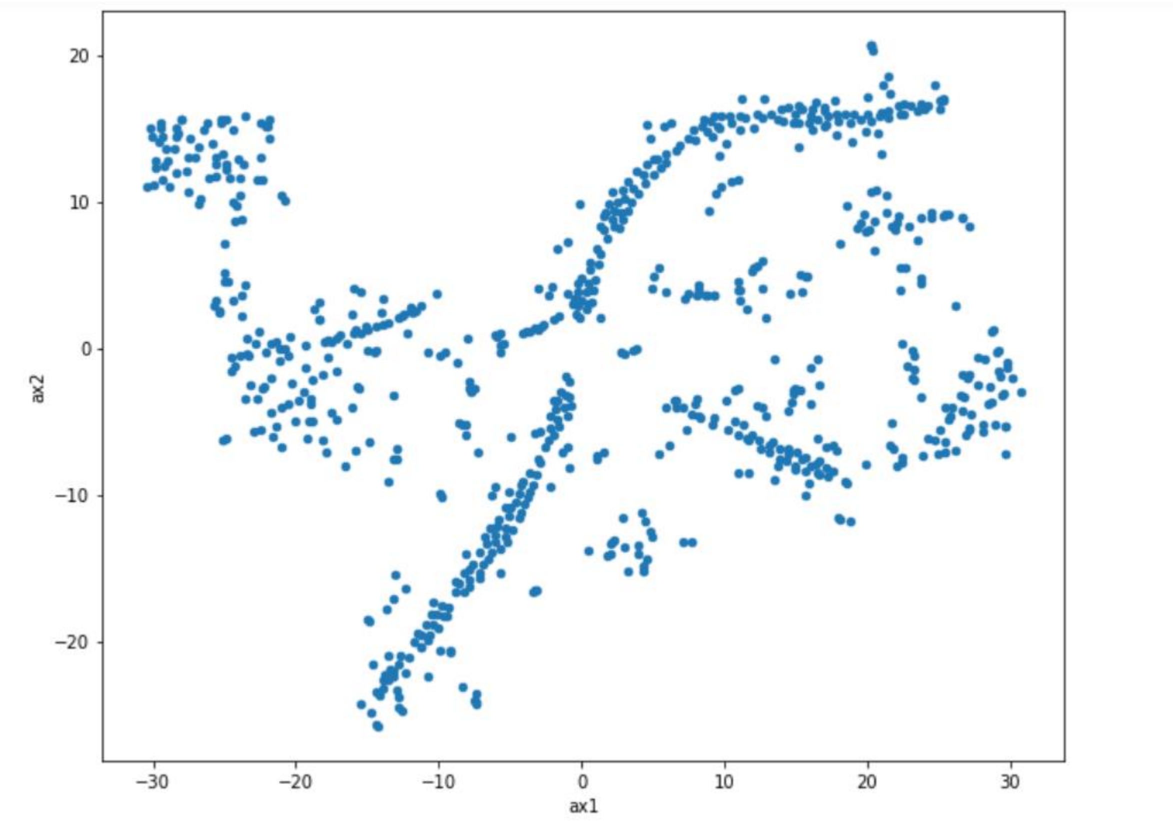
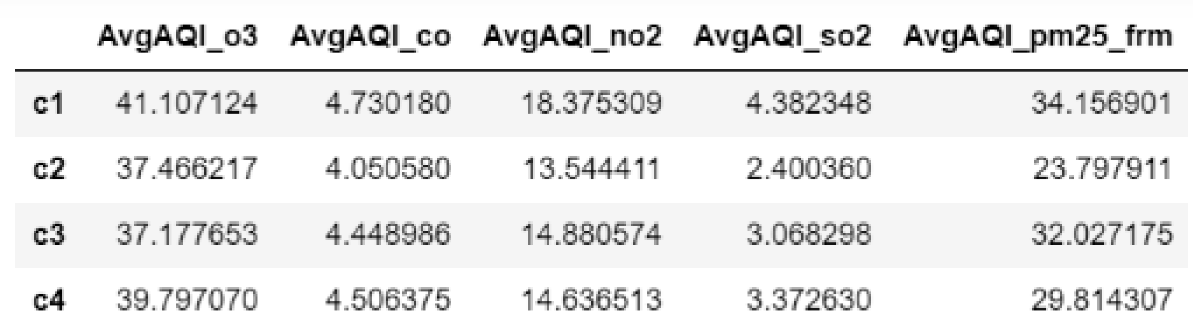


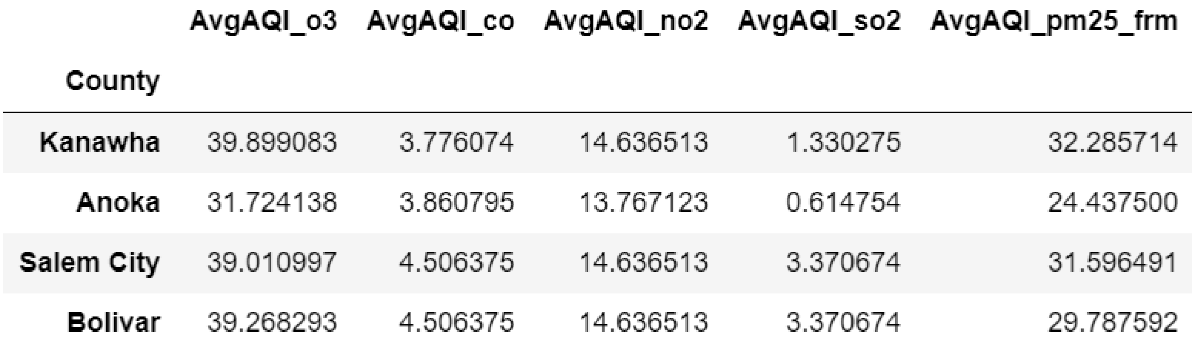
Figure 1. t-SNE output

To extract all the clusters, checked the means for each and every pollutant in each cluster and obtaining the properties counties of each cluster have in mutual.

It appears that moving from right to left on the plot, the air gets cleaner and the average AQI for PM2.5 drops greatly. On the other hand, moving from bottom to top on the plot, the biggest change appears for the average AQI for O3.



	AvgAQI_o3	AvgAQI_co	AvgAQI_no2	AvgAQI_so2	AvgAQI_pm25_frm
c1	41.107124	4.730180	18.375309	4.382348	34.156901
c2	37.466217	4.050580	13.544411	2.400360	23.797911
c3	37.177653	4.448986	14.880574	3.068298	32.027175
c4	39.797070	4.506375	14.636513	3.372630	29.814307



County	AvgAQI_o3	AvgAQI_co	AvgAQI_no2	AvgAQI_so2	AvgAQI_pm25_frm
Kanawha	39.899083	3.776074	14.636513	1.330275	32.285714
Anoka	31.724138	3.860795	13.767123	0.614754	24.437500
Salem City	39.010997	4.506375	14.636513	3.370674	31.596491
Bolivar	39.268293	4.506375	14.636513	3.370674	29.787592

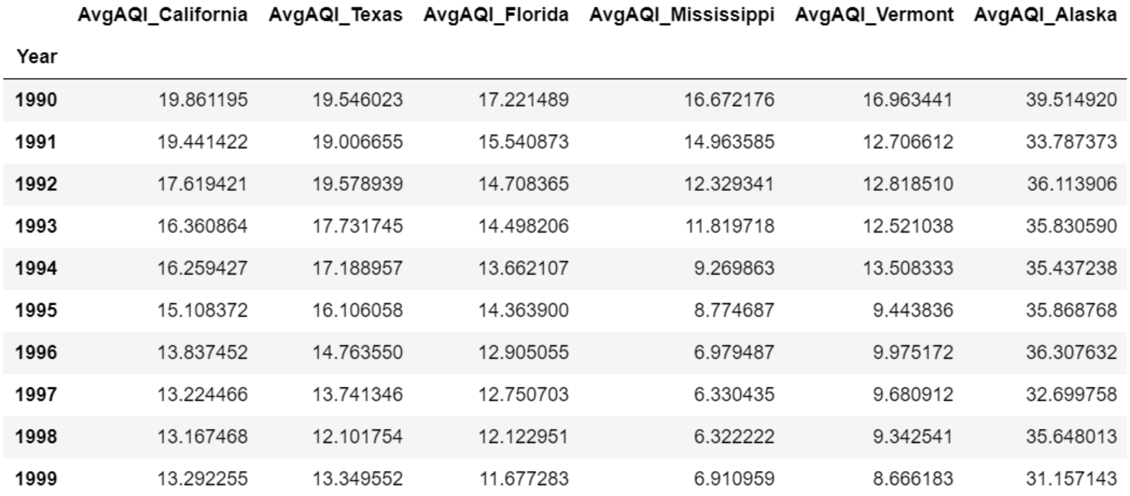
Rural AQI compared with Urban AQI:

We took 6 states:

- Urban (the top three populated states): California, Texas, Florida
- Rural states: Mississippi, Vermont, Alaska

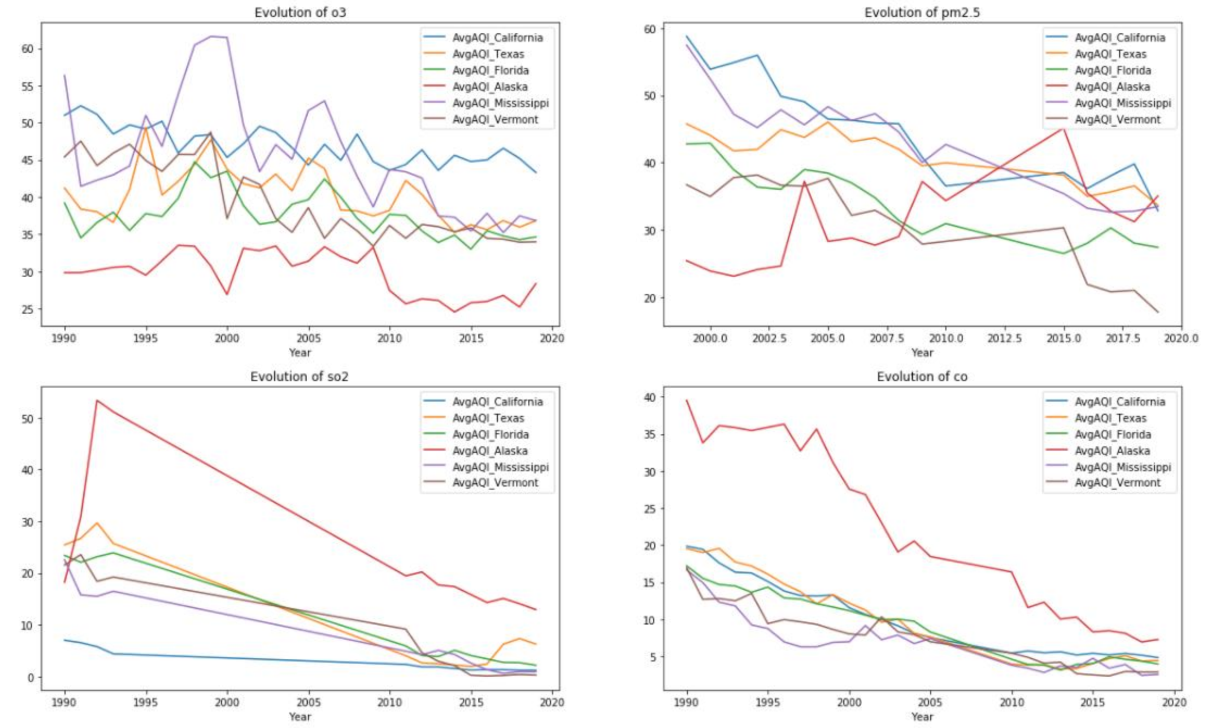
For each of these states, we studied the evolution of the yearly AQI of several pollutants, namely: o3, co, so2, pm2.5 and we tried to find whether the urban states are cleaner or rural states.

We build a dictionary where the keys are the pollutants and the values are the dataframes giving the average AQI over the years, for the given pollutant for the selected six states.



Year	AvgAQI_California	AvgAQI_Texas	AvgAQI_Florida	AvgAQI_Mississippi	AvgAQI_Vermont	AvgAQI_Alaska
1990	19.861195	19.546023	17.221489	16.672176	16.963441	39.514920
1991	19.441422	19.006655	15.540873	14.963585	12.706612	33.787373
1992	17.619421	19.578939	14.708365	12.329341	12.818510	36.113906
1993	16.360864	17.731745	14.498206	11.819718	12.521038	35.830590
1994	16.259427	17.188957	13.662107	9.269863	13.508333	35.437238
1995	15.108372	16.106058	14.363900	8.774687	9.443836	35.868768
1996	13.837452	14.763550	12.905055	6.979487	9.975172	36.307632
1997	13.224466	13.741346	12.750703	6.330435	9.680912	32.699758
1998	13.167468	12.101754	12.122951	6.322222	9.342541	35.648013
1999	13.292255	13.349552	11.677283	6.910959	8.666183	31.157143

On plotting the results obtained, we get:



For so2 and co, all states have approximately the same average AQI except Alaska which is a bit higher. For PM2.5, the rural states, that is, Vermont and Mississippi, they have a better AQI than the urban states. For o3, California comes first by a large margin and the rural states fair better but not with a considerable margin. The overall Air Quality seemed to be better in the rural states considering Ozone and Particulate Matter.

Considering temperature and humidity:

The database also contains tables about temperature and humidity. To analyse how air quality affects or is affected by the weather. We selected California state, as it not only the most populated but also one of the most polluted states.

We extracted the following features:

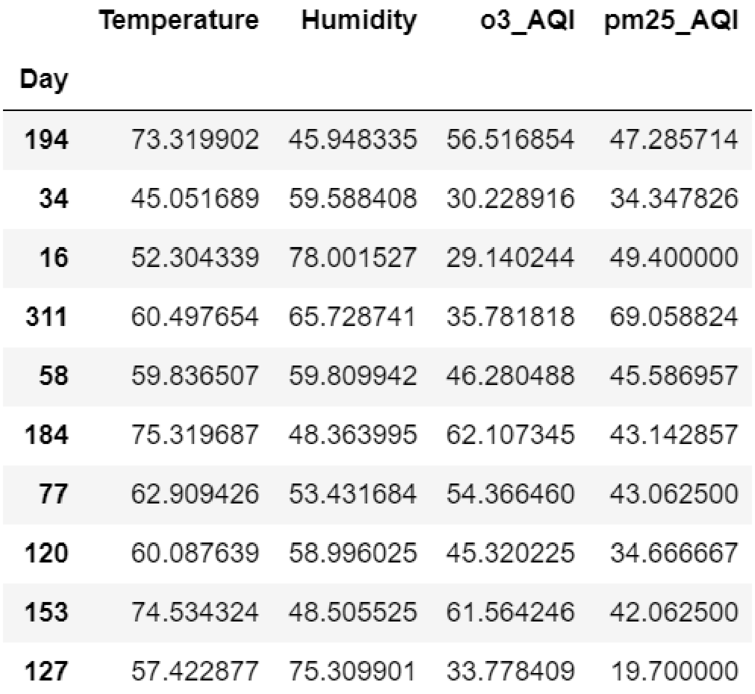
The average temperature in California for the year 2016 and form a dataframe

The average relative humidity in California for the year 2016 and form a dataframe

The average AQI for Ozone in California for the year 2016 and form a dataframe

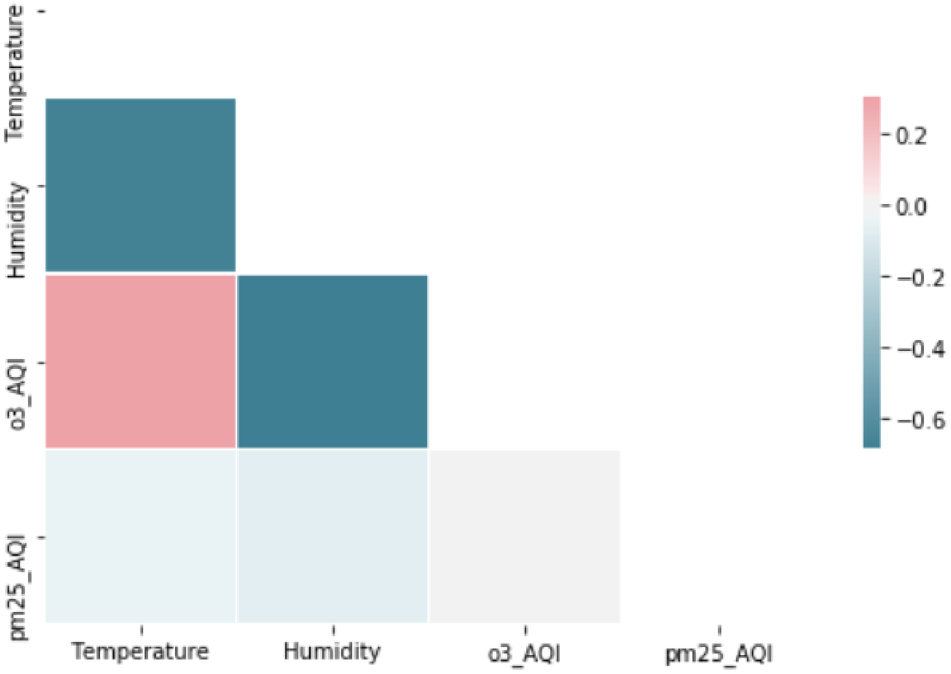
The average AQI for PM_2.5 in California for the year 2016 and form a dataframe

On concatenate all these dataframes we get:

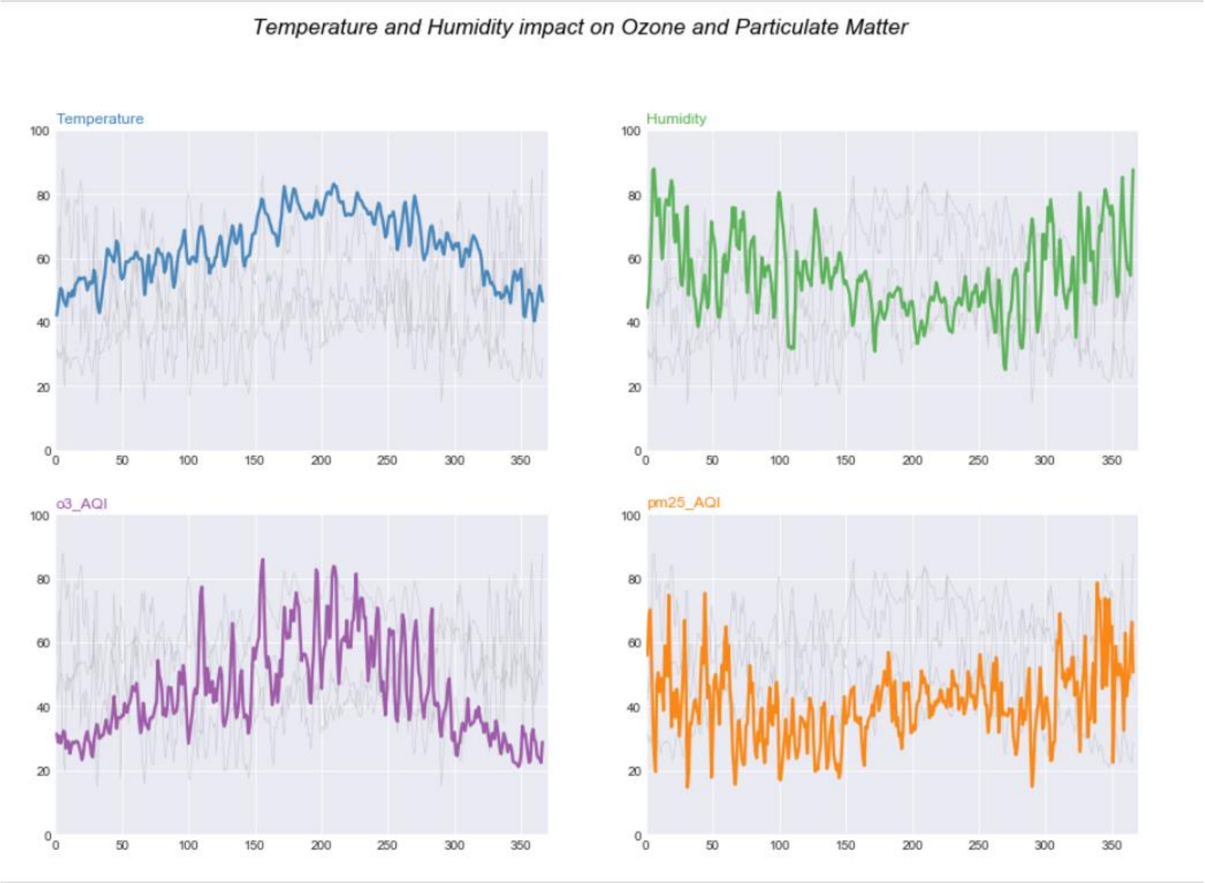


	Temperature	Humidity	o3_AQI	pm25_AQI
Day				
194	73.319902	45.948335	56.516854	47.285714
34	45.051689	59.588408	30.228916	34.347826
16	52.304339	78.001527	29.140244	49.400000
311	60.497654	65.728741	35.781818	69.058824
58	59.836507	59.809942	46.280488	45.586957
184	75.319687	48.363995	62.107345	43.142857
77	62.909426	53.431684	54.366460	43.062500
120	60.087639	58.996025	45.320225	34.666667
153	74.534324	48.505525	61.564246	42.062500
127	57.422877	75.309901	33.778409	19.700000

On plotting a correlation matrix to be able to get some understanding about our features, we get:

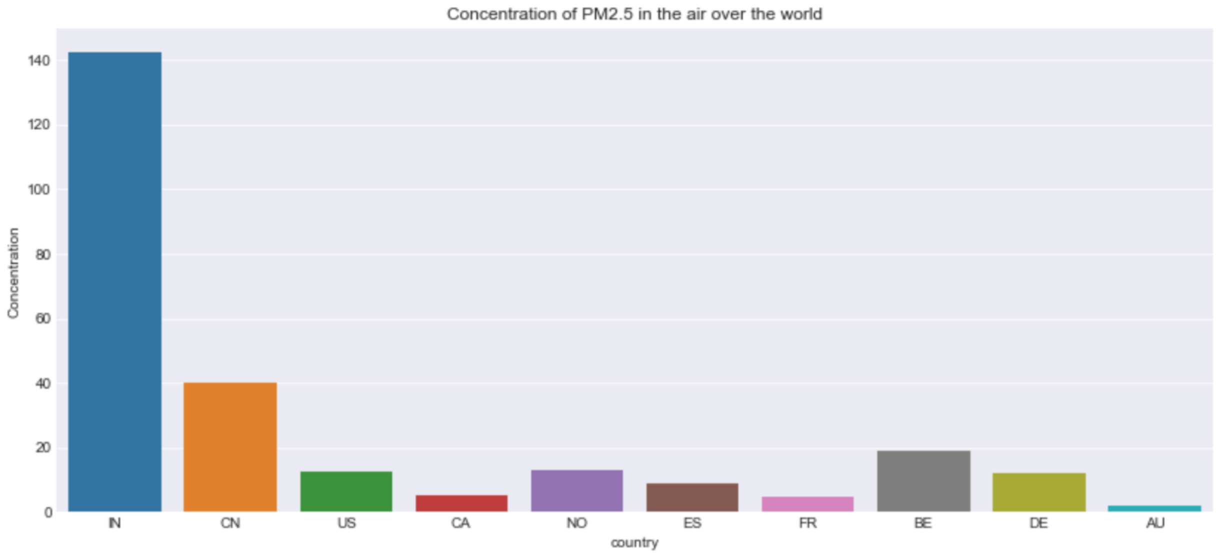


On plotting lines for the same we get:



There is an impact of Temperature as seen from the plot on the Ozone level. It is noticeable that the hottest days have the highest Ozone level.

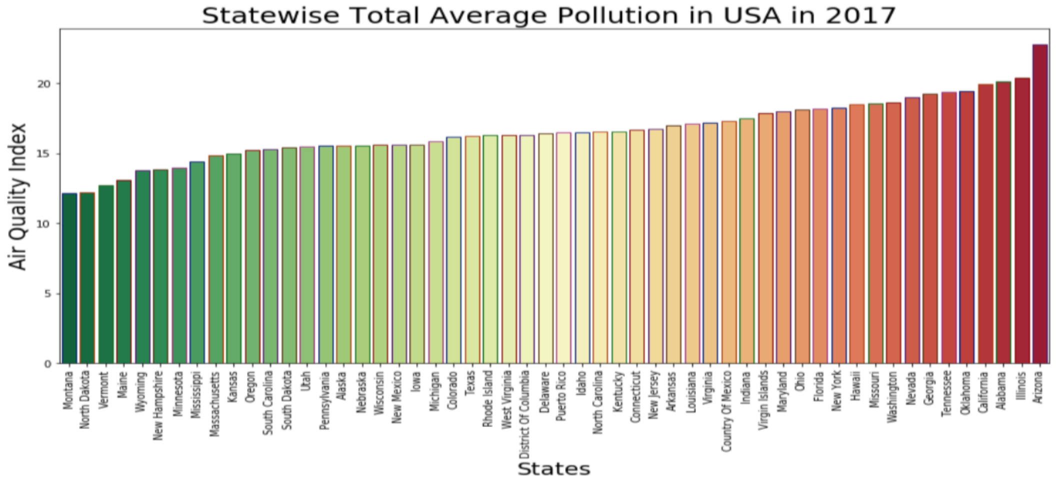
On plotting a graph for the AQI all over the world for 2018, we get:



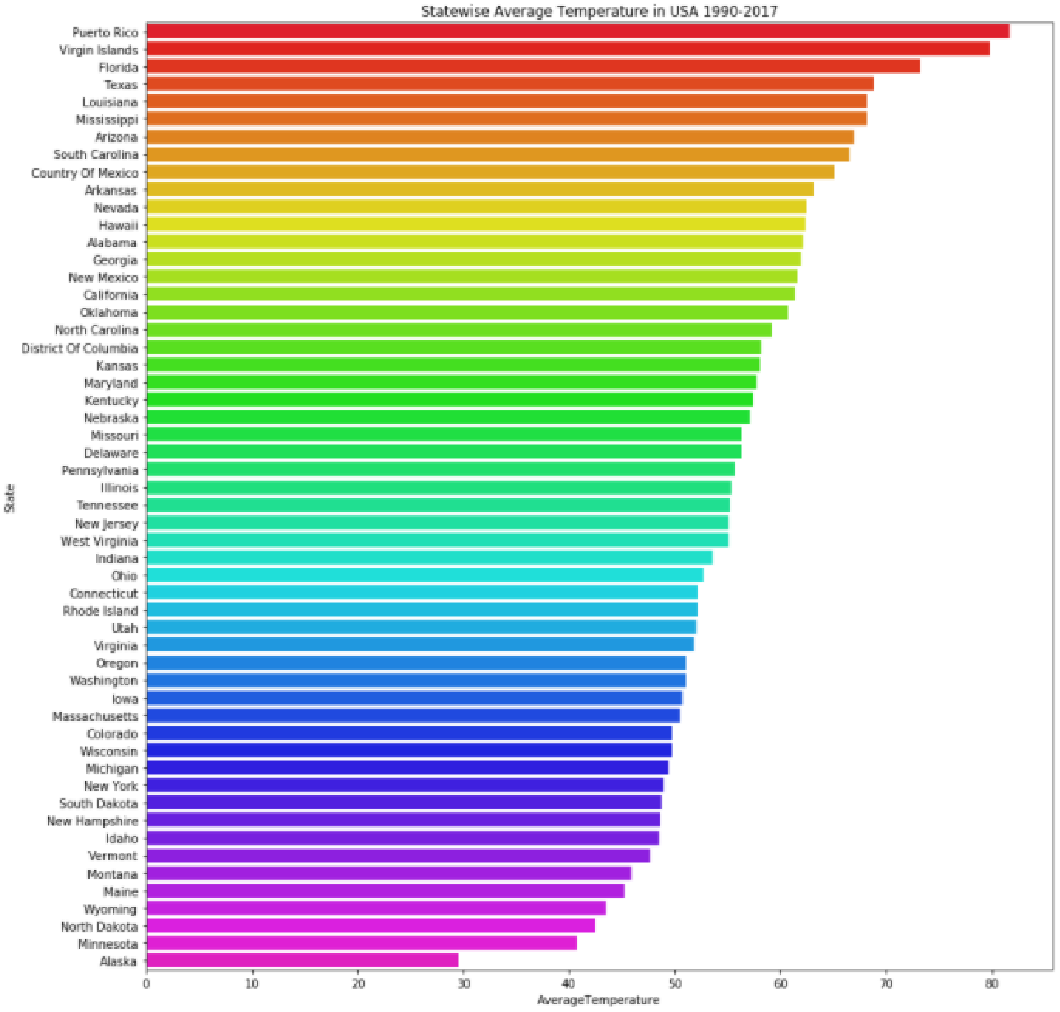
From the plot we can see that the most polluted air in 2018 is in India and the values reached are critical and clearly unhealthy. On the contrary, Australia seems to have the cleanest air.

II. Prediction using RandomForestRegressor

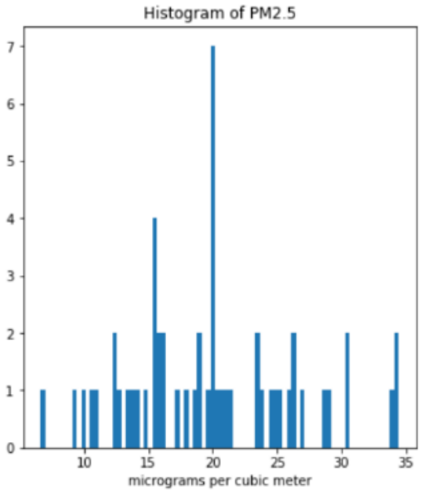
We first find the top 5 most polluted states, which are – Hawaii, Arizona, Illinois, Oklahoma and California



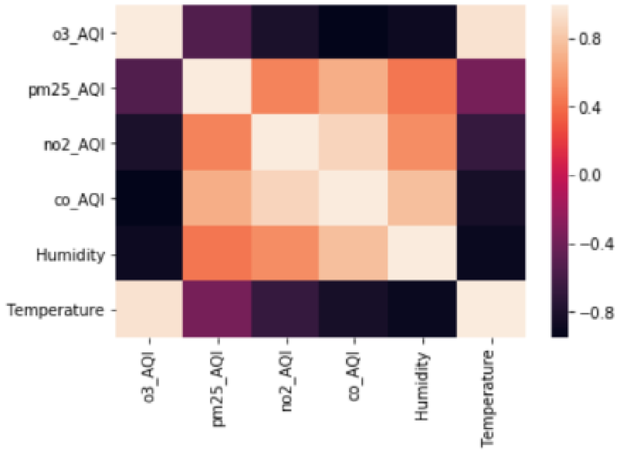
On plotting the graph of evolution of the elements taken into consideration – o3, co, no2, so2, pm25_frm and pm25_nonfrm, for the selected five states. The concluding graph being plotted is:



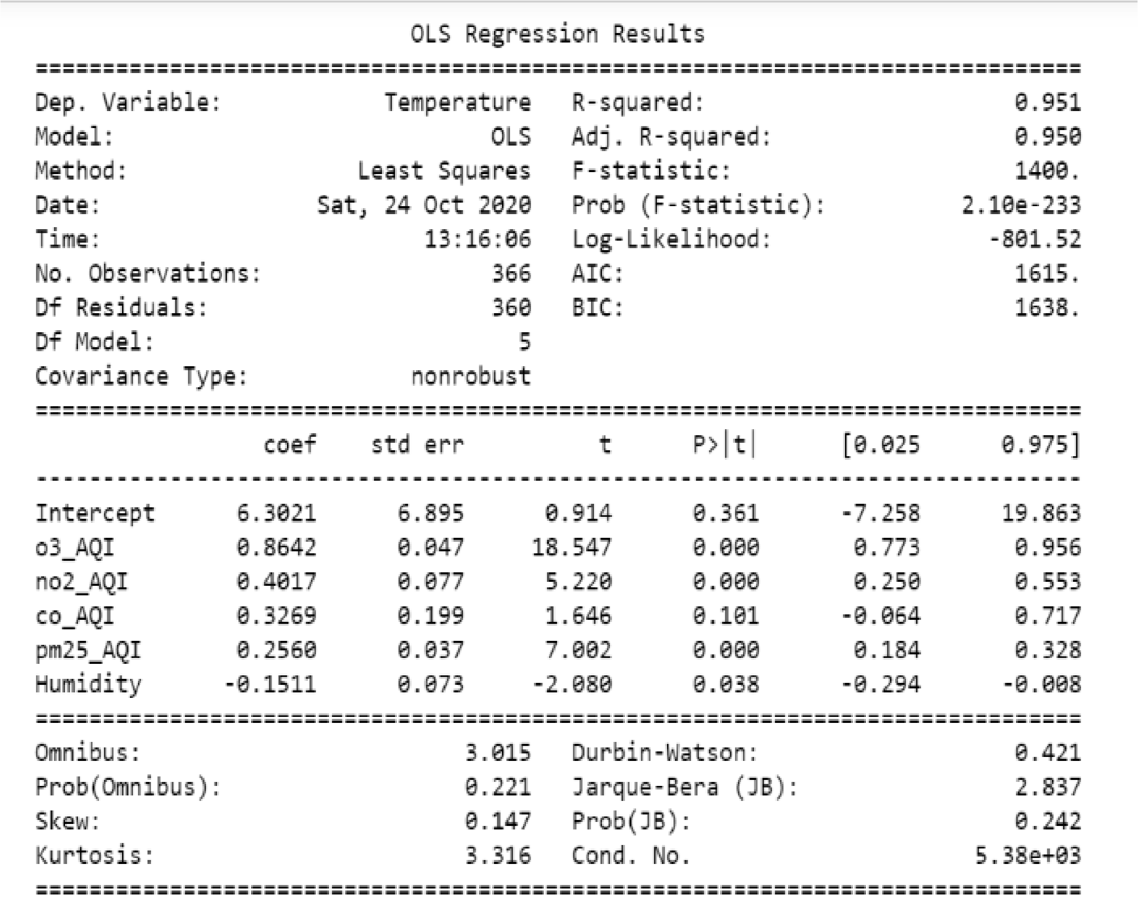
On plotting histogram for PM2.5, we get:



On plotting heat map, we get:

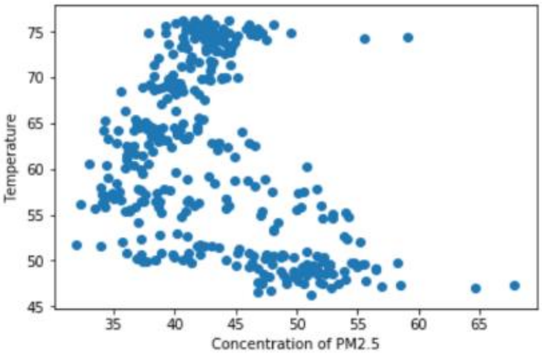


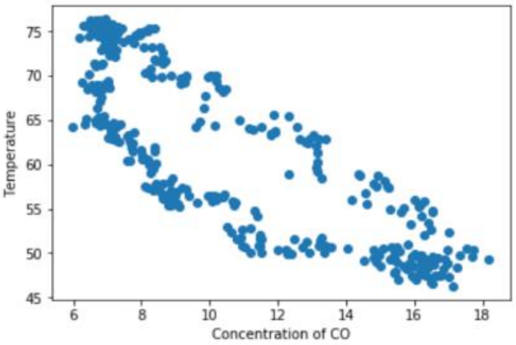
The OLS (Ordinary Least Squares) Regressor is used to get the following outputs:

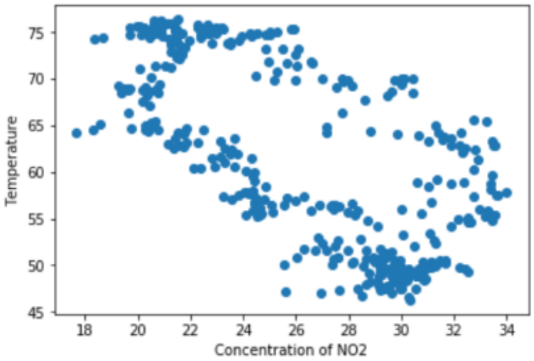


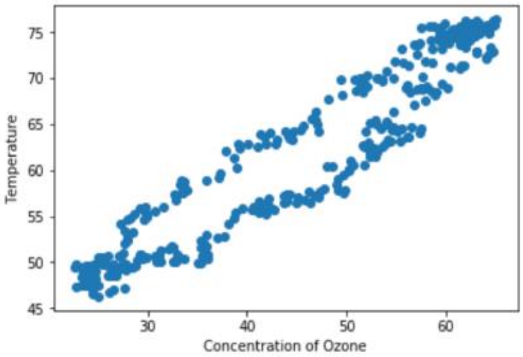
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Temperature   R-squared:                       0.951
Model:                            OLS   Adj. R-squared:                  0.950
Method:                 Least Squares   F-statistic:                     1400.
Date:                Sat, 24 Oct 2020   Prob (F-statistic):          2.10e-233
Time:                        13:16:06   Log-Likelihood:                -801.52
No. Observations:                 366   AIC:                             1615.
Df Residuals:                     360   BIC:                             1638.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.3021      6.895      0.914      0.361      -7.258      19.863
o3_AQI         0.8642      0.047     18.547      0.000       0.773       0.956
no2_AQI        0.4017      0.077      5.220      0.000       0.250       0.553
co_AQI         0.3269      0.199      1.646      0.101      -0.064       0.717
pm25_AQI       0.2560      0.037      7.002      0.000       0.184       0.328
Humidity      -0.1511      0.073     -2.080      0.038      -0.294      -0.008
==============================================================================
Omnibus:                        3.015   Durbin-Watson:                   0.421
Prob(Omnibus):                  0.221   Jarque-Bera (JB):                2.837
Skew:                           0.147   Prob(JB):                        0.242
Kurtosis:                       3.316   Cond. No.                     5.38e+03
==============================================================================
```

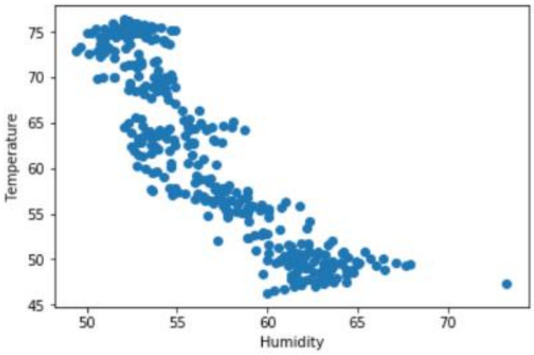
The scatter plots for the temperature vs the six elements are:









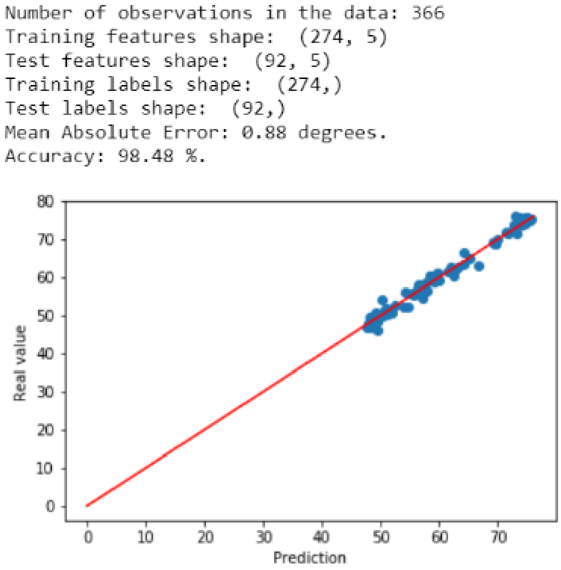


Since, it is time-series based database we opted to use RandomForestRegressor. For the RandomForestRegression, we first dropped humidity and temperature. Then divided the dataframe into training and testing parts.

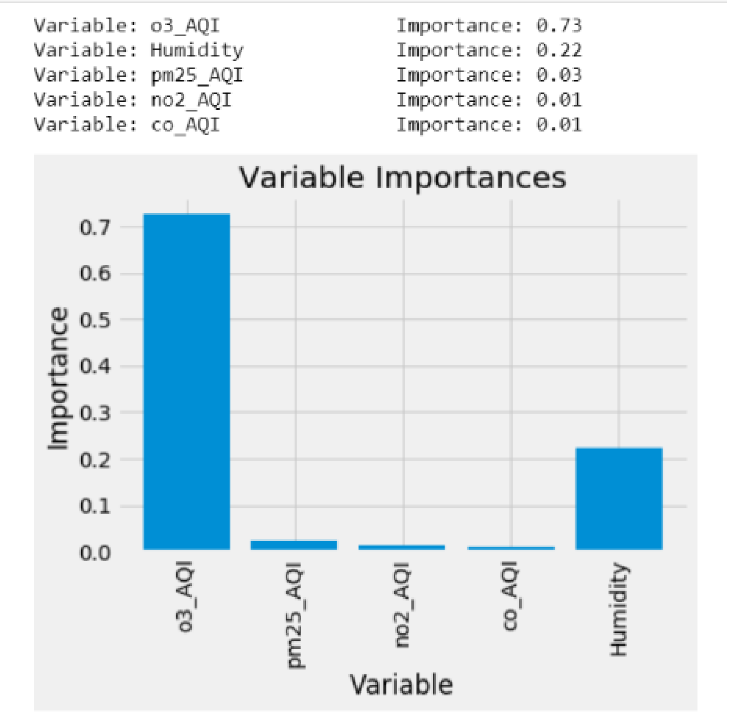
The steps are as follows:

1. Take the samples repetitively from the training dataset so that each data point has an equal possibility of getting designated, every sample has the same size as given for the original training set.

2. The model is trained as each sample is drawn and the prediction is recorded for each sample.
3. Each sample prediction is calculated by taking an average of the predictions from the trees producing the final prediction.



```
Number of observations in the data: 366
Training features shape:  (274, 5)
Test features shape:  (92, 5)
Training labels shape:  (274,)
Test labels shape:  (92,)
Mean Absolute Error: 0.88 degrees.
Accuracy: 98.48 %.
```



```
Variable: o3_AQI              Importance: 0.73
Variable: Humidity            Importance: 0.22
Variable: pm25_AQI            Importance: 0.03
Variable: no2_AQI             Importance: 0.01
Variable: co_AQI              Importance: 0.01
```

We get 98.48 % accuracy on using RandomForestRegression, showing that the model trained properly and gave correct predictions.

Background study with citation and references:

Citations and References:

[1] Air Quality Data Management and Integration System Scoping Study, Defra and the Devolved Administrations, Unrestricted ED46602, Issue 1.2 AEAT

[2] Air quality data management system an integration of data acquisition, deposition and analysis, H. T. sieh', Y. hen' and D. pepper, Nevada Center for Advanced Computational Methods (NCACM), Department of Mechanical Engineering, University of Nevada, Las Vegas

[3] Article: Air quality data management, https://ee.ricardo.com/air-quality/air-quality-measurements

[4] Article: Air quality databases, https://app.croneri.co.uk/feature-articles/air-quality-databases

[5] "Air Pollution Detection and Prediction Using Multi Sensor Data Fusion", E. Brumancia, S. Justin Samuel, L. Mary Gladence, Intelligent Computing and Control Systems (ICICCS) 2020 4th International Conference.

[6] "Air Pollution Prediction Using Machine Learning", Maghvendra Singh, Harshit Saran, Mrs.Sapna Yadav, International Journal of Innovations in Engineering and Science, Vol. 4, No.6, 2019.

[7] "Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters", Jan Kleine Deters, Rasa Zalakeviciute, Mario Gonzalez, Yves Rybarczyk, Journal of Electrical and Computer Engineering, vol. 2017.