ANALYSIS AND PREDICTION OF AIR QUALITY INDEX IN INDIA

(June-July 2021)

Submitted by:

Khushboo Gupta

Aakriti Sharma

TABLE OF CONTENTS

INTRODUCTION

Abstract:
Air Quality Index (AQI) is an index that helps to report air quality. It is a measure used to indicate the level of pollutants (SO2, NO2, NO, etc.) over a period and tells how clean or polluted the air is. Further on, using the Air Quality Index we can find the associated health hazards that might affect us. AQI Bucket is used to group the AQI values into six categories based on the value namely: Good (0–50), Satisfactory (51–100), Moderate (101–200), Poor (201–300), Very Poor (301–400), Severe (401–500).

Introduction:
India is a growing industrial nation. Industrial activities emit various pollutants in the air that affect air quality. Particulate matter 2.5 and 10, Nitrogen dioxide, Sulphur dioxide, and carbon monoxide are some of the main pollutants emitted. Other causes like stubble burning increases PM2.5 in the air. Air Quality Indexes indicate the level of major pollutants in the air.
Through this project we will give a detailed analysis of AQI and of different pollutants affecting it in various states of India. We have prepared a model for predicting the AQI bucket for a given value of AQI and a model to predict AQI on the basis of concentration of pollutants. During the Covid-19 lockdown times there was a significant change in the AQI. We will also be giving the analysis on how AQI was affected during that time.
The project is sub-divided in following sections. These are as follows:
1- Loading necessary libraries
2- Loading AQI dataset from a csv file
3- Summarizing the data for better understanding (Descriptive Statistics)
4- Data pre-processing
5- Data visualization using plots, graphs etc.
6- Applying different learning algorithms on the training dataset for predictions
7- Evaluating the performance of the model using evaluation metrics like confusion matrix etc.

DATASET

The source of this dataset is Kaggle. The data has been compiled from the Central Pollution Control Board (CPCB) website: https://cpcb.nic.in/ which is the official body of Government of India. The dataset contains air quality data and AQI (Air Quality Index) at daily level of various stations across multiple cities in India. The dataset has 16 columns and 29531 rows where each row corresponds to a new record of the city. However, in some rows the values are 'Nan'. Therefore, the data was pre-processed before using it for training models.

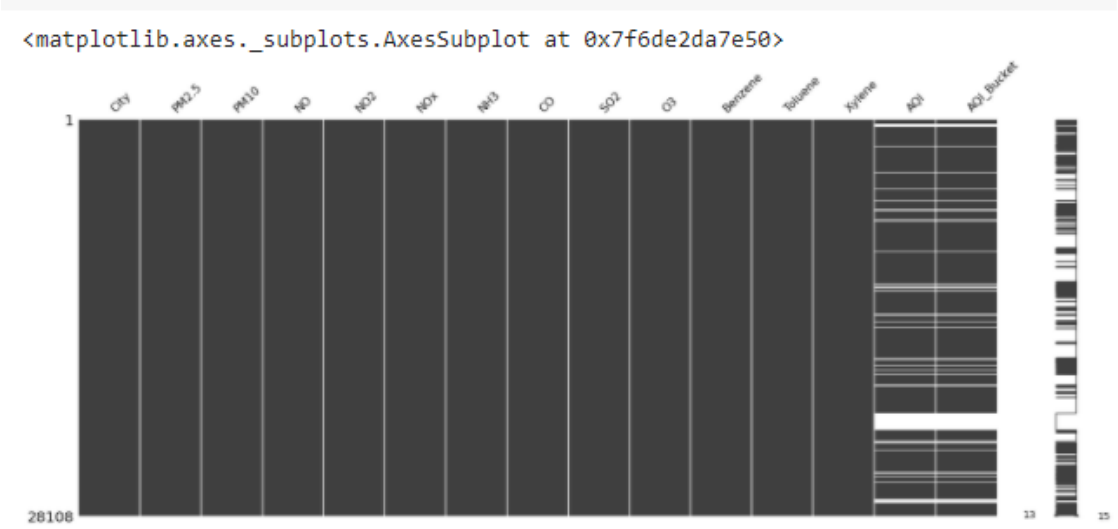
Dataset URL: https://www.kaggle.com/rohanrao/air-quality-data-in-india

Columns:

- City – Different cities of India (Ahmedabad, Aizawl, Amaravati, Amritsar, Bengaluru, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, Visakhapatnam)
- Date – The date of which the data is given. The dataset is from the date 01-01-2015 to 01-07-2020.
- PM2.5 – Particulate Matter 2.5 (pollutant)
- PM10 – Particulate Matter 10 (pollutant)
- NO – Nitric Oxide (Pollutant)
- NO2 – Nitrogen di Oxide (Pollutant)
- NOx – Other oxides of nitrogen (Pollutant)
- NH3 – Nitric Oxide (Pollutant)
- CO – Carbon Monoxide (Pollutant)
- SO2 –Sulphur di Oxide (Pollutant)
- O3 – Ozone (Pollutant)
- Benzene – (Pollutant)
- Toluene – (Pollutant)
- Xylene – (Pollutant)
- AQI – Air Quality Index
- AQI_Bucket – Categories of AQI values: Good (0–50), Satisfactory (51–100), Moderate (101–200), Poor (201–300), Very Poor (301–400), Severe (401–500).

METHODOLOGY, ANALYSIS AND RESULTS

- **Data Pre-processing-**
  The final data fed to our models has to be created by cleaning the dataset. From observing our dataset, we find that our pollutants have NaN concentration values. We fill these values with the city wise mean for each pollutant.

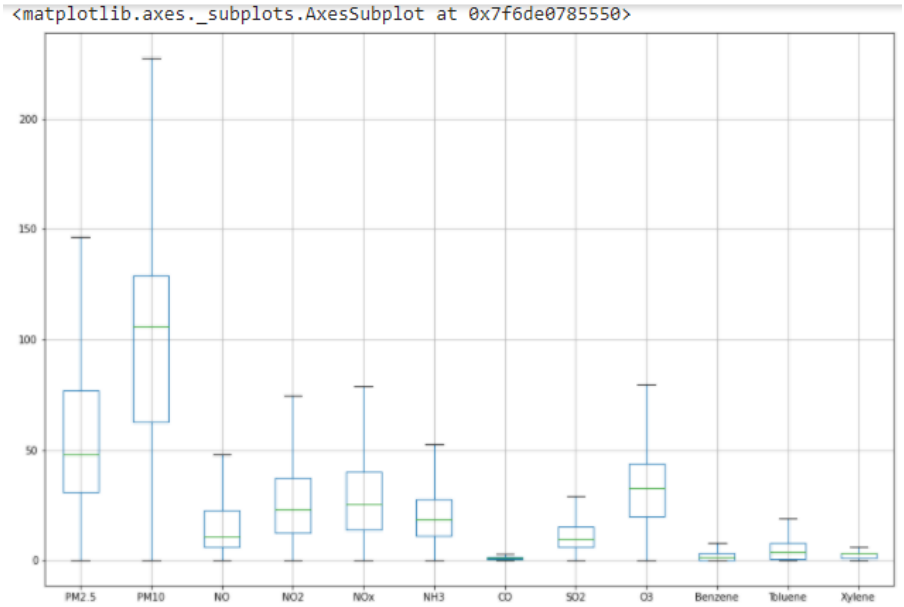


- **Data Visualization-**
  We first plot visualization of our dataset to construct initial impressions about the data.

  Summarizing concentration of pollutants in air:
  We first plot a box plot to display the summary of the set of data values of concentration of the pollutants. This also shows us the maximum of the concentration of each pollutant.

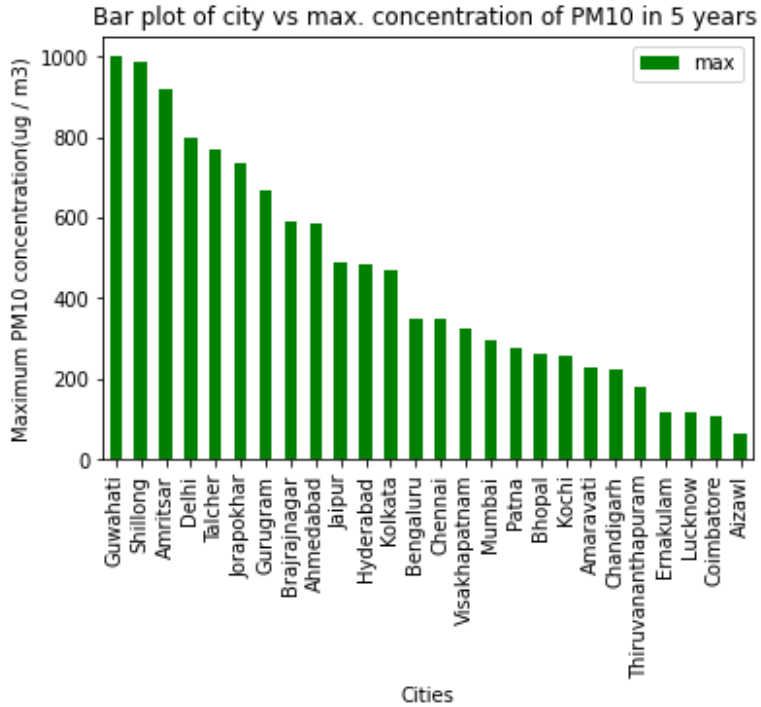
  Maximum Concentration of Pollutants with high concentrations:
  From the box plot, we can see that **PM2.5, PM10, NOx and O3** are the pollutants having maximum concentrations. Hence, we

plot individual bar graphs for these pollutants showing their maximum concentration city wise by grouping our dataset by City.



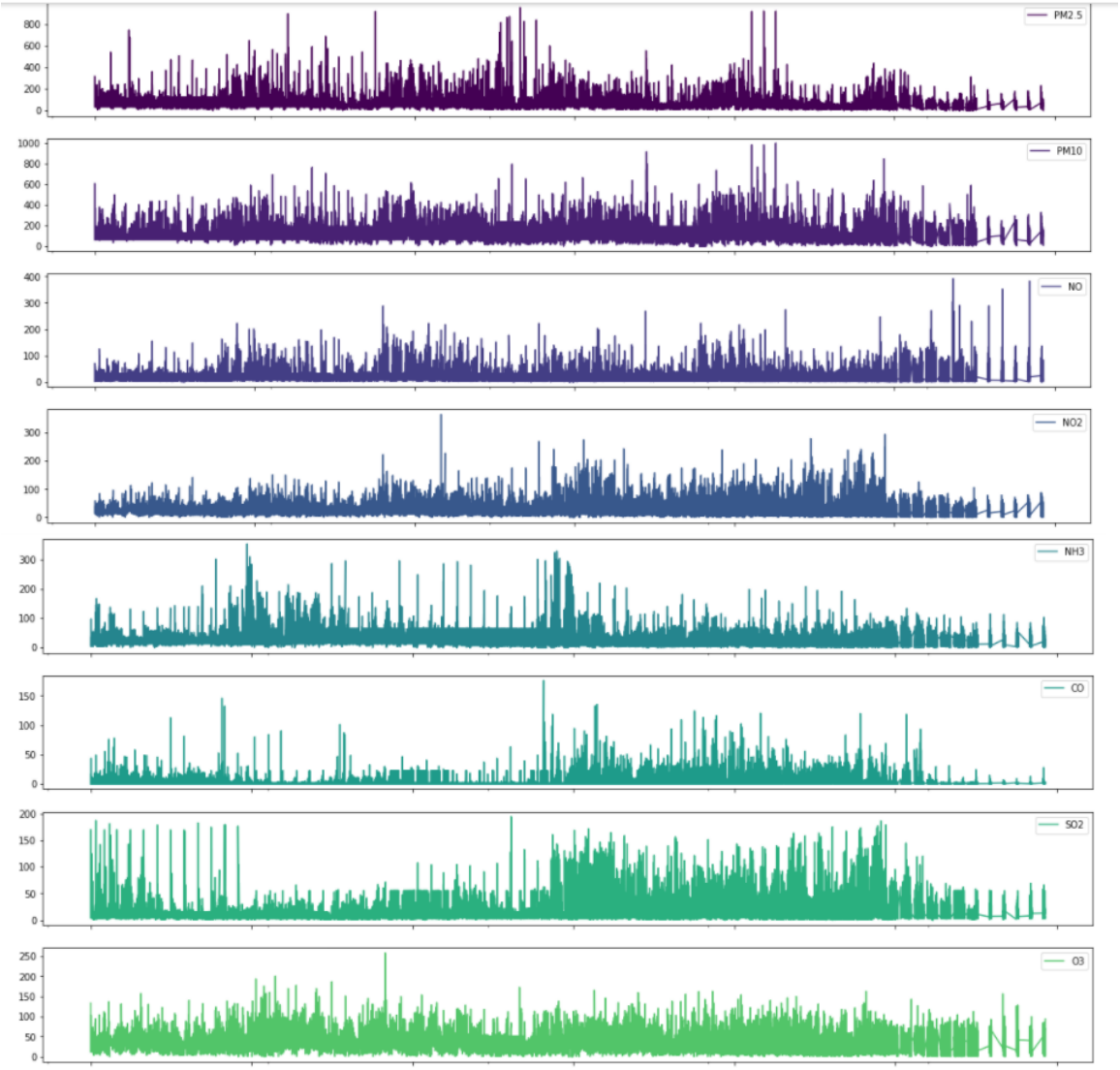
Concentrations of each pollutant in all cities:

We plot individual bar graphs to show the total concentration of all the pollutants in each city. This shows us that **the pollutant PM10 has the maximum concentration in the city Delhi** followed by PM 10 concentration in the city Lucknow.



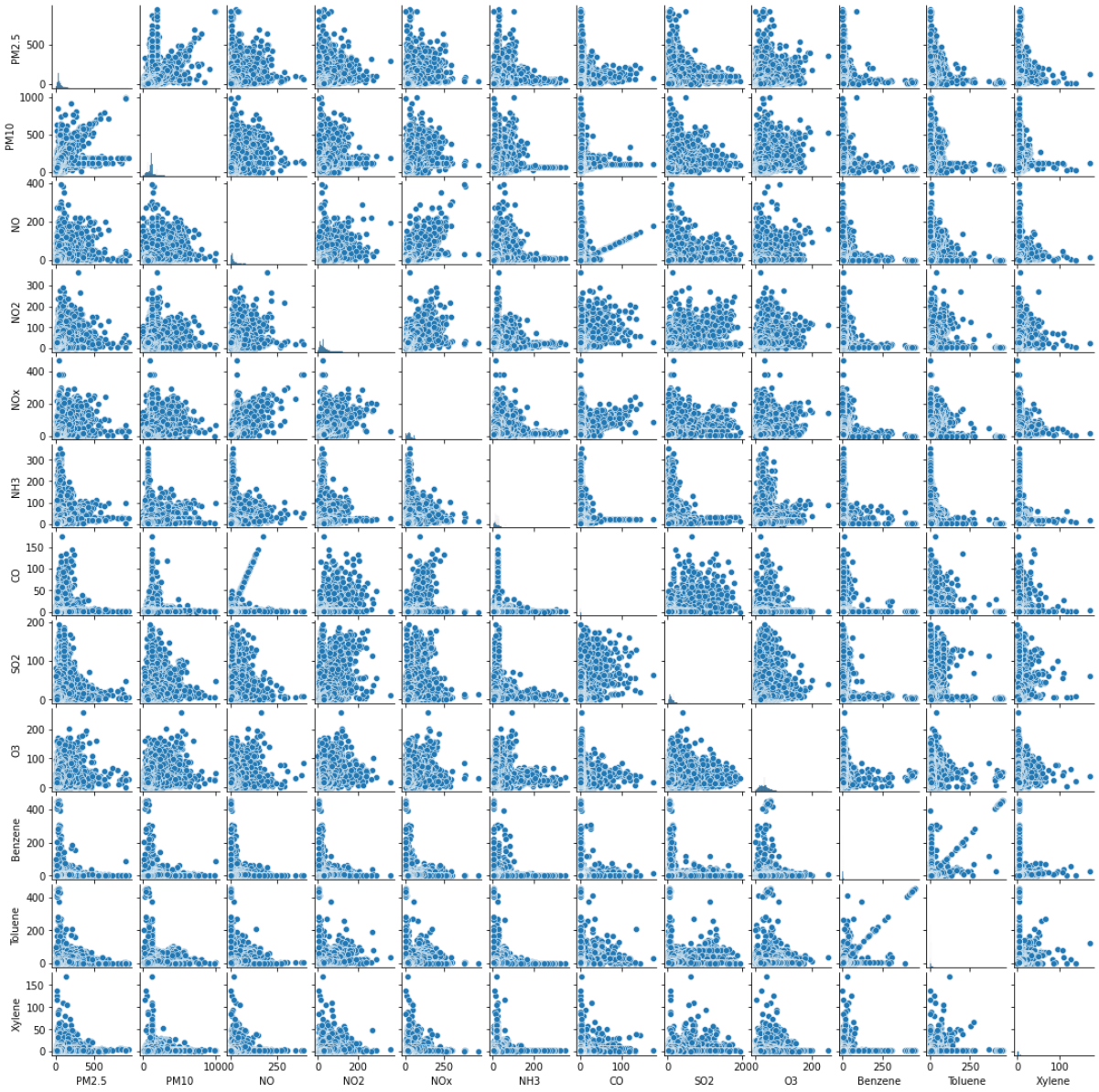
Distribution of concentration of Pollutants:

Next, we plot a line graph showing the distribution of concentration of individual pollutants over time. (from Jan 2015 to Dec 2020). The concentrations of PM2.5, PM10 and NOx and O3 are dramatically higher in the years 2018 to 2020 and rapidly decrease from the year 2020.



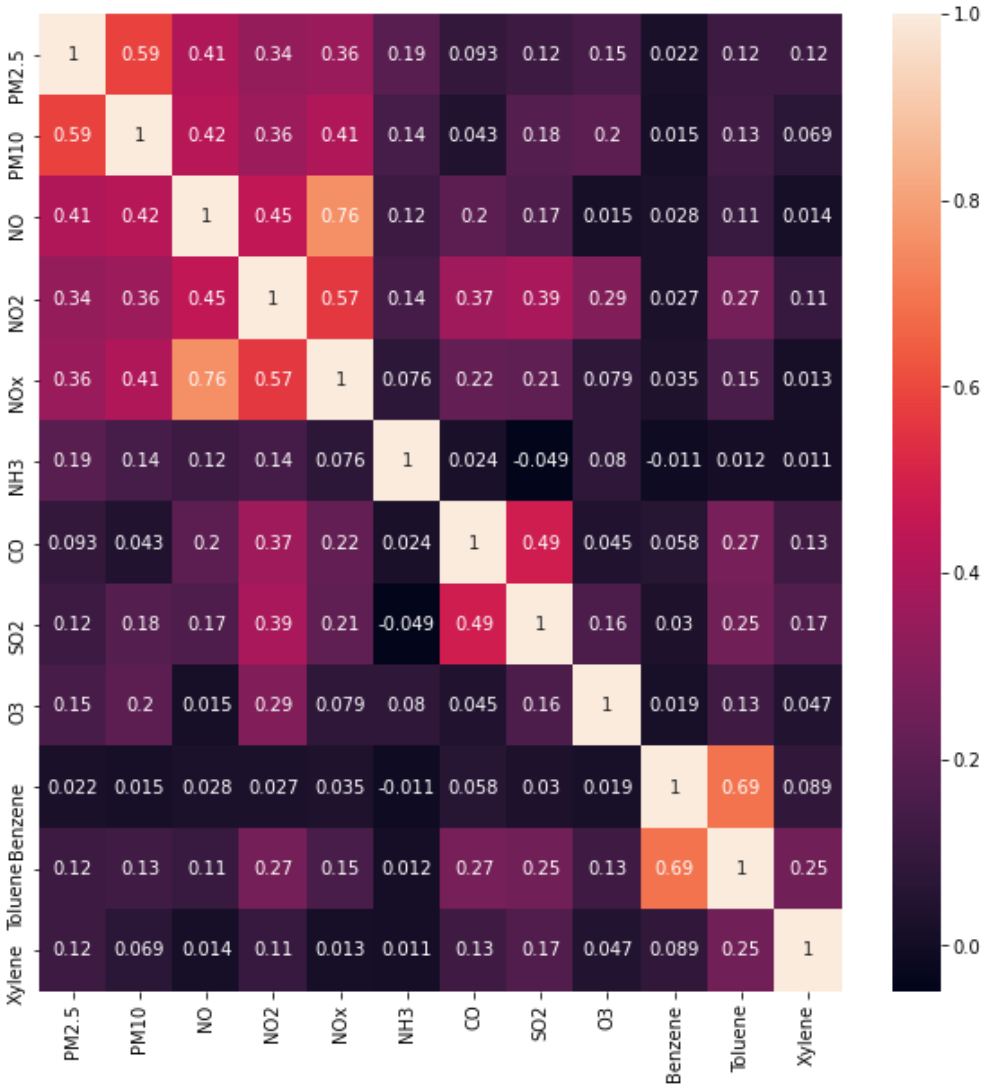
Relationship between the various pollutants:

A pairplot is plotted showing the relationships between the pollutants in our dataset.



Correlation among the various pollutants:

We plot a correlation matrix to study the correlation between the pollutants in our dataset and further supporting our relationship analysis in the earlier graph. We see that NO and NOx show the maximum correlation of 0.76.



Filling null values for AQI:

We see that our pre-processed data contains null values for AQI. These values need to be filled to pass our data to prediction models. We follow the fundamental approach of data cleaning, by observing the missing values and replacing those by other values.

We use the standard AQI calculation formula to calculate the AQI values from the 24-hourly average concentration of pollutants and fill the missing values of AQI.

The AQI calculation method used is-

$I_i = I_{i,j+1} - I_{i,j} (X_i - X_{i,j}) + I_{i,j} X_{i,j+1} - X_{i,j}$ for $X_{i,j} \leq X_i \leq X_{i,j+1}$

where X_i = Observed concentration for the ith pollutant $I_{i,j}$ = PSI value for the ith pollutant and the jth breakpoint as given in the table $I_{i,j+1}$= PSI value for the ith pollutant and the (j+1)th breakpoint as given in the table $X_{i,j}$ = Concentration for the ith pollutant and jth breakpoint as given in the table $X_{i,j+1}$= Concentration for the ith pollutant and (j+1)th breakpoint as given in the table Finally, the overall index is calculated as the maximum of sub-indices: PSI = maximum (I1, I2, I3, I4, I5, I6)

AQI Category	AQI	Concentration range*							
		PM_{10}	$PM_{2.5}$	NO_2	O_3	CO	SO_2	NH_3	Pb
Good	0 - 50	0 - 50	0 - 30	0 - 40	0 - 50	0 - 1.0	0 - 40	0 - 200	0 - 0.5
Satisfactory	51 - 100	51 - 100	31 - 60	41 - 80	51 - 100	1.1 - 2.0	41 - 80	201 - 400	0.5 - 1.0
Moderately polluted	101 - 200	101 - 250	61 - 90	81 - 180	101 - 168	2.1 - 10	81 - 380	401 - 800	1.1 - 2.0
Poor	201 - 300	251 - 350	91 - 120	181 - 280	169 - 208	10 - 17	381 - 800	801 - 1200	2.1 - 3.0
Very poor	301 – 400	351 - 430	121 - 250	281 - 400	209 - 748*	17 - 34	801 - 1600	1200 -1800	3.1 - 3.5
Severe	401 - 500	430 +	250+	400+	748+*	34+	1600+	1800+	3.5+

Example of computation-

Suppose a 24-hr PM2.5 concentration of 40 µg/m3 is observed. Based on the table, the observed concentration of Xi = 40 µg/m3 lies between 30 and 60 µg/m3 .

Therefore, the computation is carried out for the second segment (j = 2). For this segment, X1,1 = 30 µg/m3 and X1,2 = 60 µg/m3 with corresponding sub-index values of I1,1 = 50 and I1,2 = 100.

The computation is as follows:

$I_i = I_{i,j+1} - I_{i,j} (X_i - X_{i,j}) + I_{i,j} X_{i,j+1} - X_{i,j}$

$= 50 + (40 - 30) * (100-50) / (60-30)$

$= 66.67$

Therefore, the PM2.5 sub-index is 66.67. If the five other pollutant sub-indices were calculated in a similar manner from concentrations, then the overall index is reported as the maximum of these values.

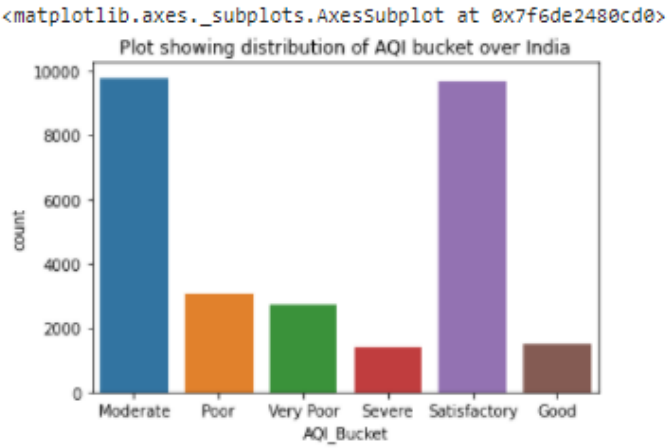
AQI Bucket Calculation -

CENTRAL POLLUTION CONTROL BOARD'S
AIR QUALITY STANDARDS

AIR QUALITY INDEX (AQI)	CATEGORY
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very Poor
401-500	Severe

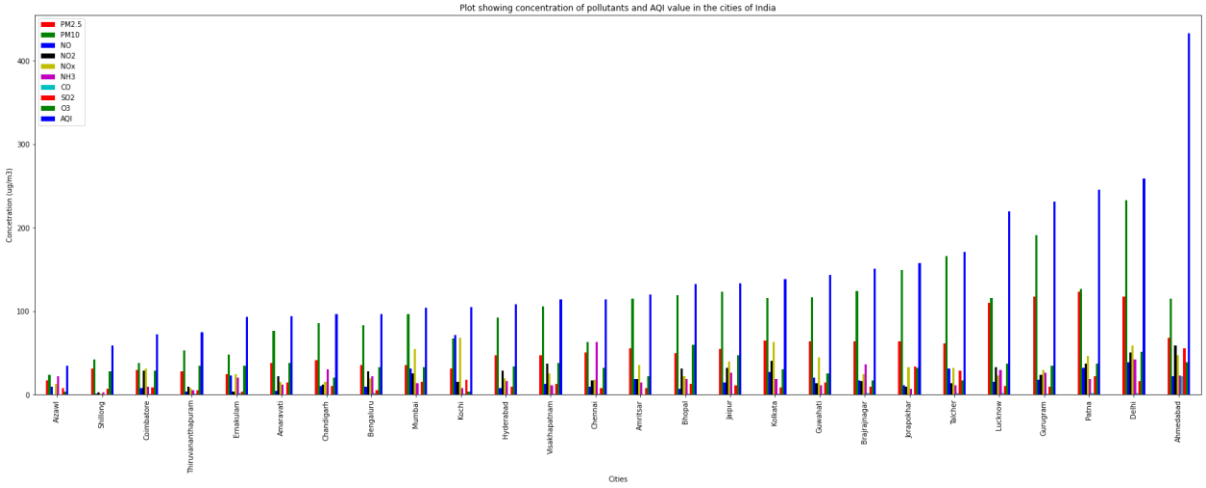
Once AQI is calculated, the null values of AQI Bucket can also be filled using the above table.
Using the above approach, the data is cleaned. Now we move further with analyzing our final cleaned data to get insights of the AQI and AQI Bucket in Indian cities.

Data Visualization of AQI Bucket-
Countplot showing the counts of each category of AQI Bucket is plotted. We see that the maximum AQI values fall under the category Moderate and Satisfactory.

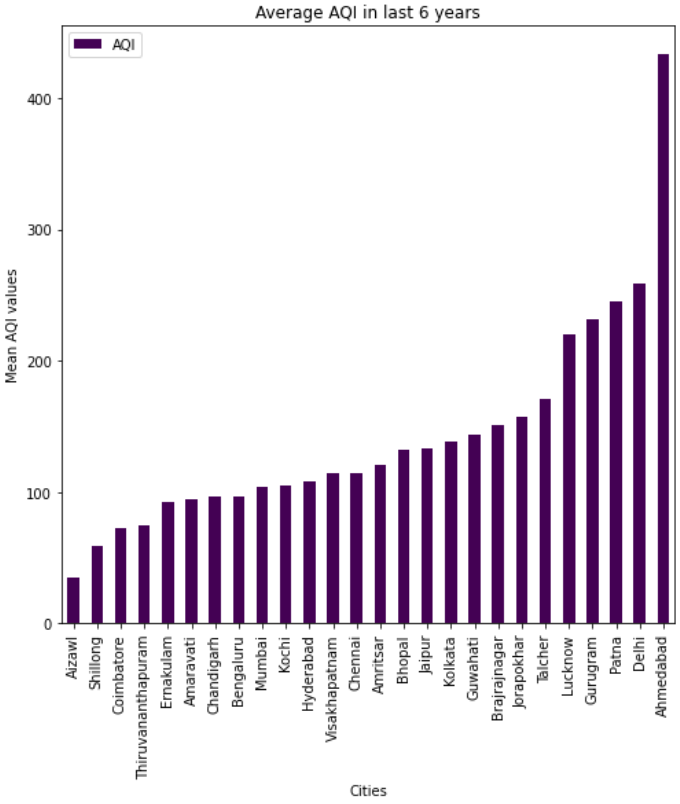


Data Visualization AQI:

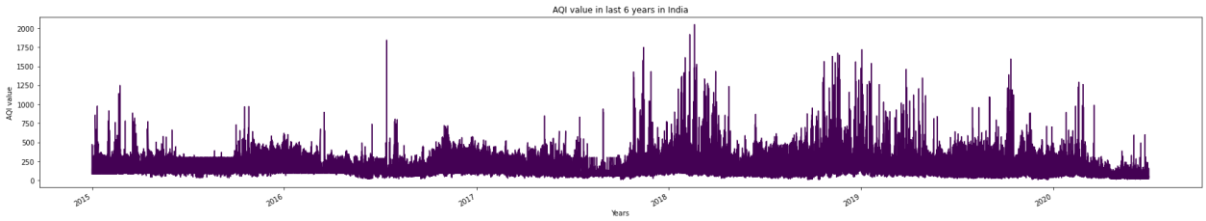
Bar plot showing the concentrations of involved pollutants against their mean AQI value is plotted for all cities of India. PM2.5 and PM10 have significant contribution in calculation of AQI and Ahmedabad shows maximum value of AQI.



Bar plot showing the average AQI values of each city is plotted again supporting our above fact showing Ahmedabad as the city with the highest AQI value.

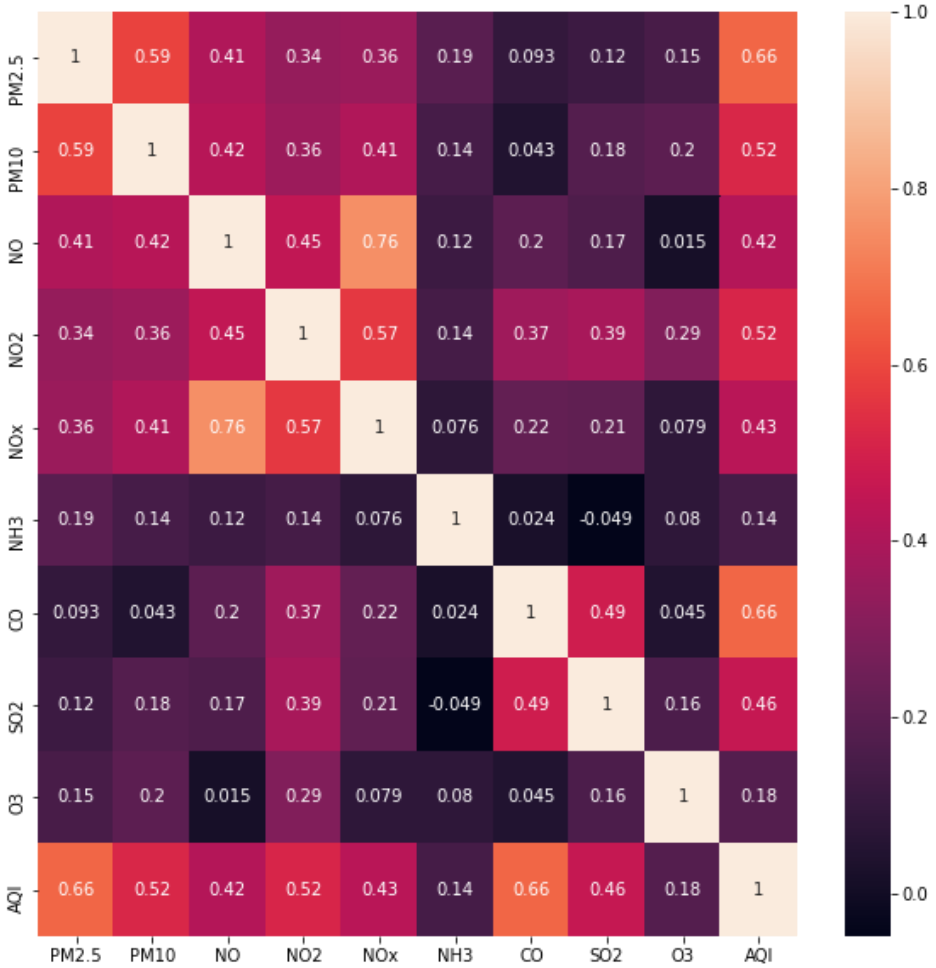


Distribution of AQI values in last 6 years is shown using line plot. The AQI values are highest during the years 2018-2020 and fall dramatically from 2020 onwards.



Correlation among AQI and other pollutants-

Correlation matrix is plotted which shows the correlation between AQI values and the pollutants involved in the calculation of AQI. We see that PM2.5 and CO have the maximum correlation with AQI and thus are the significant contributors in its calculation.



Multiple Linear Regression Model:

Multiple linear regression is a supervised machine learning model that uses multiple independent variables to predict the outcome of a response variable. It finds the linear relationship between independent and dependent variables.

In our project, we have used multiple linear regression to predict the AQI values on the basis of the past values of the concentration of the pollutants. We have used 'PM2.5','PM10', 'NO', 'NO2', 'NOx', 'NH3', 'CO', 'SO2','O3' pollutant values to predict the AQI values.

When we found the score of our model, it came out to be 0.839(approx.) which indicates that model is good and the regression line fits well.

KNN (K-Nearest Neighbor) algorithm:

This is a classification model that uses K Nearest Neighbors to predict the values of new datapoints.

We have used the AQI values to predict the AQI bucket which has categories (Good (0–50), Satisfactory (51–100), Moderate (101–200), Poor (201–300), Very Poor (301–400), Severe (401–500). The accuracy rate was found to be 1.0.

Random Forest Algorithm:

It is a machine learning algorithm based on supervised learning technique. It randomly chooses multiple subsets and makes prediction based on best results.

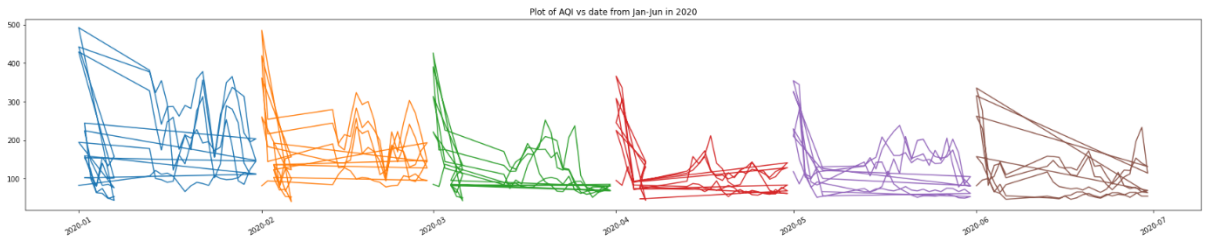
We have used the AQI values to predict the AQI bucket which has categories (Good (0–50), Satisfactory (51–100), Moderate (101–200), Poor (201–300), Very Poor (301–400), Severe (401–500). The accuracy rate was found to be 1.0.

KNN (K-Nearest Neighbor) algorithm vs Random Forest Algorithm:

In our project we have used both KNN algorithm and random forest algorithm to predict the values of AQI Bucket based on the AQI values. The accuracy rate is also the same.
The difference here lies that there is no real training time in case of KNN algorithm as KNN algorithm does not train the data so there are no computations performed while training. It only stores the data. Whereas in the case of the Random Forest algorithm, randomly many trees(subsets) are made so it takes more time.
On the other hand, KNN algorithm need very high testing time as the distance needs to be calculated between every point whereas Random Forest algorithm has low testing time as it just has to traverse down a tree.
So, overall Random Forest algorithm is faster.

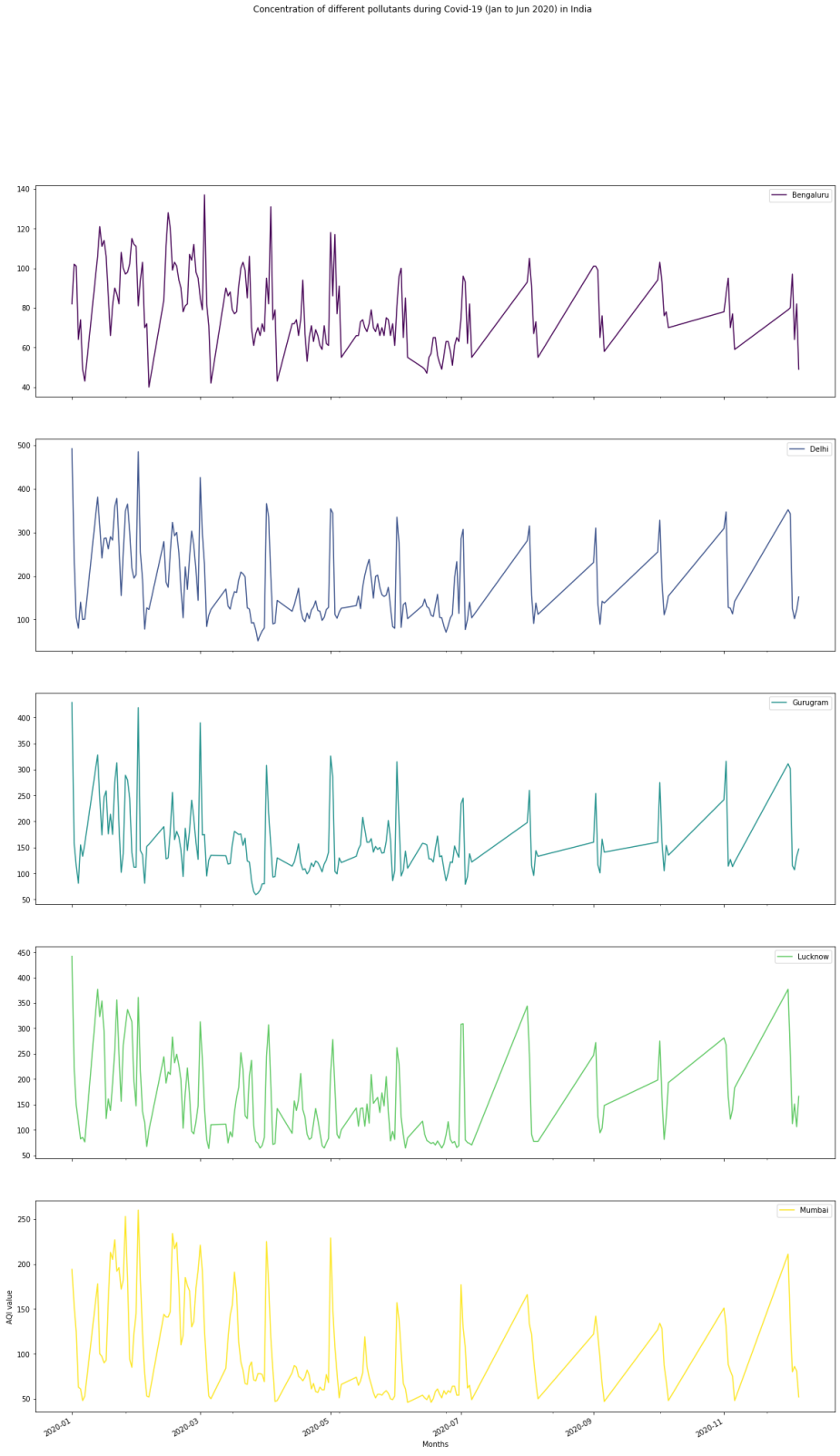
But if we want to update the dataset:
In case of KNN algorithm it will be easy and just like adding new points to existing data. However, in case of Random Forest Classifier we can't update an existing tree. We have to create a new one.

ANALYZING COVID-19 EFFECT ON AQI-
We pivot and get the filtered data from the month of Jan 2020 onwards to analyze the effect of Covid-19 on the AQI.
Plotting the AQI values month wise from 2020 Jan to 2020 June, we see that AQI is maximum in 2020 Jan and reduces gradually thereafter, being the least in May 2020.



This shows us that before the Covid-19 in March, the AQI was higher and it decreased in the months of Covid-19.

Next, supporting the same, we show plots of individual cities against their month wise AQI values. Each city follows the same pattern as stated above.



Finally, we find the mean values of AQI for a few metro cities for the months of Jan to June 2020. The mean AQI values numerically confirm our above stated visualizations.
We can see that the AQI value is maximum in Delhi in the month of Jan and minimum in Bengaluru in the month of June.

MODULES USED

Python Libraries:

- **Numpy** – NumPy (Numerical Python) is a Python package used for performing numerical computations and it works on arrays.

- **Pandas** - Pandas is an open source python library used for data analysis and high performance data manipulation. It is built on top of the NumPy package.

- **Matplotlib** - Matplotlib is a visualization Python library that helps to create 2D plots from data in arrays. It is built on NumPy arrays.

- **Seaborn** - Seaborn is a visualization Python library built on top of Matplotlib library. It is used for plotting amazing statistical plots with many styles and colour palettes. It is closely integrated with data structures from Pandas.

- **Missingno** - It is a simple python library used for data visualization of missing data values in Pandas dataframe.

- **Scikit- learn** - Scikit- learn is a useful Python library for machine learning.It has effective tools for predictive data analysis and statistical modelling which includes classification, regression etc.

Machine Learning Algorithms:

- **Multiple linear regression** - Multiple linear regression helps to make a model that represents the relationship between two or more independent features and one dependent feature by

fitting a linear equation to the observed data. R^2 score is used to evaluate how good the regression line fit i.e. how near the actual data is to the fitted regression line. This score does not tell if the line is appropriate or not. A low R^2 score indicates that the regression line does not fit well with the data whereas a high R^2 score indicates that the model is good and the regression line fits well.

- **K Nearest Neighbor**: It is a simple machine learning algorithm used to solve regression and classification problems. It considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new data point. It does not train the data. At the time of training it just stores the dataset.

- **Random Forest Classifier**: It is a supervised learning algorithm used for both regression and classification problems. It creates decision trees(subsets) on randomly selected data at training time. Then it takes predictions from each tree and selects the best result on the basis of voting (i.e. the class selected by most trees).

CONCLUSIONS

1. We used different Python libraries like Pandas, NumPy, etc. to manipulate and analyse data.

2. Using different plots like bar plot, line plot we visualized the maximum concentration of each pollutant in cities in 5 years and concentration of each pollutant in each city (city wise). Also, we visualized the distribution of AQI bucket over India and pollutants have affected the AQI value.

3. Using heat maps, we visualized the correlation between different pollutants and between pollutants and AQI.

4. With the help of Linear Regression, we were able to study the relationship between pollutants and AQI value and make predictions.

5. We also used Random Forest Classifier and KNN algorithms to create a model for prediction of the AQI bucket for the given AQI value. We also analysed the difference between these two models.

6. Lockdowns during the Covid-19 times affected the Air Quality Index. We analysed the data from January,2020 to June,2020 and we found that there was a significant drop in the AQI value due to less pollution during lockdowns.

REFERENCES

1. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

2. https://www.javatpoint.com/machine-learning-random-forest-algorithm

3. https://www.geeksforgeeks.org/linear-regression-python-implementation/

4. https://datascience.stackexchange.com/questions/9228/decision-tree-vs-knn

5. https://forum.airnowtech.org/t/the-aqi-equation/169

6. https://medium.com/@yrnigam/how-to-write-a-data-science-report-181bd49d8f4d

7. https://app.cpcbccr.com/ccr_docs/How_AQI_Calculated.pdf