**Exhaustive Study of Air Quality in Indian Cities (2015-2020)**

*Contributors:* Sarthak Niwate, Rishabh Choubey, Perumalla Naga Sai Venkatesh, Krupananda M T

Post Graduate Diploma in Data Science and Engineering

Great Learning, Bangalore

Great Lakes Institute of Management, Chennai

*Mentor:* Mr. Jayveer Nanda

Lead Data Scientist, Data Science and Business Analytics Mentor

1. **SUMMARY OF PROBLEM STATEMENT, DATA AND FINDINGS**

**1.1 Abstract**

India is one of the most polluted countries in the world, where several major cities are facing serious environmental issues with rapid pollution growth. The objective of this project is to, predict the air quality index of Indian cities and to analyse air pollution trends with respect to various cities in India, in order to have a global view of the damage caused, so that appropriate actions can be developed in the future to prevent air pollution. In this regard, the polluted database was established based on the data provided by the Central Pollution Control Board, India. The results show will explore the three main cities under the radar of severe air pollution. And the model to cluster the cities with different AQI and predicting the AQI.

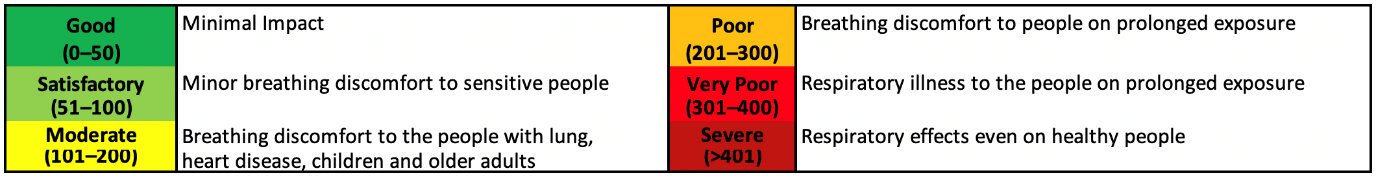
**1.2 Introduction**

In India, the atmospheric has damaged due to urbanization, industrial development, lack of awareness, poor maintenance of motor vehicles and poor road conditions *(Shrinivas.J 2011)*. For developing countries like India air pollution has tremendous impact on human vigour, agricultural practices, climatic variations and overall changes in ecosystem.

Every year about six lakh citizens of India die due to effects and happenings of air pollution which has become fifth leading cause of death across the country after other causes like water pollution, nuclear pollution etc. Out of these, almost 35,000 death occur in Delhi, rest 15,000 deaths are recorded in each industrial area *(Choking cities article, 2016)*. Though almost complete attention given to national capital due to which other cities and towns are suffering. Other cities that are suffering due to extensive threat of pollution are Lucknow, Patna and many more. Almost all the cities are suffering due to increase in concentration of Particulate Matter (PM) in air along with gaseous pollutants like oxides of nitrogen, sulphur along with other toxic materials that are already causing serious damage to environment.

In terms of air pollution, Sources of the PM secondary particles are gases with a physicochemical transformation, like sulphate and nitrate *(Limaye and Salvi 2010).* Basically, PM has been categorized into the following three groups *(Kim et al. 2015):* PM 2.5 (diameter < 2.5 μm), PM 2.5–10 (coarse particles), and PM 10 (diameter < 10 μm). The primary particles of PM consist of the carbon emitted from trucks, cars, stone crushing, and burning waste. PM particles are from natural sources, which include volcanoes, oceans, and ground *(Limaye and Salvi 2010)*. WHO explained that PM particles can penetrate into the respiratory tract with a high penetrability. They can easily enter into the lungs and can affect the functioning. In general, PM is the cause of many major diseases, and is more dangerous to human health and life than other pollutants. The studies conducted in the past are bound to a limited number of places and are not up-to-date.

AQI is a tool, introduced by Environmental Protection agency (EPA) in USA to measure the levels of pollution due to major air pollutants. An AQI is defined as an overall scheme that transforms weighted values of individual air pollution related parameters into a single number or set of numbers *(Mukesh Sharma, 2003).* In the present study the AQI was calculated using IND-AQI specified by CPCB. The index has been developed based on the dose-response relationship of various pollutants. AQI concept transforms weighted values of individual air pollutants into a single number or set of numbers which may be widely used for air quality communication and decision making. This IND-AQI has 6 categories.



**1.3 Materials and Data Resources**

**Air Pollutants**

Major pollutants present in air are basically categorized into following types:

Carbon Compounds, Sulphur Compounds, Chlorofluorocarbons, Hydrocarbons, Metallic Pollutants, Photochemical pollutants, Particulate matter.

**Materials**

This study considered seven critically polluted locations in India, as shown in Table 1. The pollution trends were derived based on data from 2015, 2016, 2017, 2018, 2019 and 2020. The parameters considered for this research were SO2, NOx, and PM 2.5, Hydrocarbons and Carbon Compounds. For the concerned locations, we collected the geography and some of the related characteristics, incorporating some of the mega cities of India.

The data files are collected from sources deployed by the Central Pollution Control Board, India. The monitoring of pollutants is carried out on a yearly basis, with a twice a week frequency. In India, the Central Pollution Control Board (CPCB) monitors the complete air quality monitoring networks and sets the standard that needs to be followed by every pollution control unit. It is noted that there are 703 air quality stations across 307 cities in India. In our data files we have got access to 110 stations and 26 cities. This program is completely managed by the Central Pollution Control Board, in collaboration with the state pollution control board *(Pant et al. 2018).* Most of the air quality stations are deployed in urban areas *(Balakrishnan et al. 2014).*

**Data Dictionary**

We have 5 csv files from CPCB dataset, city-hour.csv, city-day.csv, station-hour.csv, station-day.csv

City-day and station-day datasets has columns as city/station, date, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, AQI\_Bucket

City-hour and station-hour datasets has columns as city/station, datetime, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, AQI\_Bucket

|  |  |
| --- | --- |
| Columns | Interpretations |
| City | Cities in India for which that corresponding record belongs to |
| Station | Air stations for which that corresponding record belongs to |
| Date | Date on which it is recorded |
| Datetime | Date and time on which it is recorded |
| PM2.5 | Fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller |
| PM10 | Fine inhalable particles, with diameters that are generally 10 micrometers and smaller |
| NO | Nitrogen monoxide (NO) is a colorless gas and one of the principal oxides of nitrogen |
| NO2 | Nitrogen dioxide (NO2) is a reddish-brown gas with a pungent, acrid odor and one of the several oxides of nitrogen |
| NOx | Nitrogen oxides (NOx) is a collective term used to refer to nitrogen monoxide (nitric oxide or NO) and nitrogen dioxide (NO2) |
| NH3 | A compound of nitrogen and hydrogen which is a byproduct of agriculture and industry. |
| CO | A colorless, odorless, tasteless, and toxic air pollutant—is produced in the incomplete combustion of carbon-containing fuels, such as gasoline, natural gas, oil, coal, and wood. |
| SO2 | A colorless, bad-smelling, toxic gas, are emitted by the burning of fossil fuels, coal, oil, and diesel. |
| O3 | A gas that can form and react under the action of light and that is present in two layers of the atmosphere. |
| Benzene | Comes from products that contain benzene such as glues, paints, furniture wax, and detergents. |
| Toluene | It evaporates when exposed to air. It also evaporates from water. |
| Xylene | Motor vehicle emissions are the predominant source of xylene in the urban air environment. |
| AQI | An index for reporting air quality |
| AQI\_Bucket | Categorized on the basis of AQI (Good, Satisfactory, Moderate, Poor, Very Poor, Severe) |

1. **OVERVIEW OF THE FINAL PROCESS**
   1. **Objectives**
2. Clustering of Cities using 12 pollutants.
3. Prediction of Air Quality Index of different clusters formed.
4. Top 3 cities to implement solutions immediately.

**2.2 Methodology**

1. **Understanding the Datasets**

We have 5 csv files from CPCB dataset, city-hour.csv, city-day.csv, station-hour.csv, station-day.csv The data has been made publicly available by the Central Pollution Control Board: <https://cpcb.nic.in/> which is the official portal of Government of India.

Note: We have five csv files from which, we will be using all the files for EDA and only one file which is Station\_hour.csv for modeling purpose.

The additional dataset and other websites are used for this analysis of 3 cities which are prone to severe air pollution. The resource websites are given in the references.

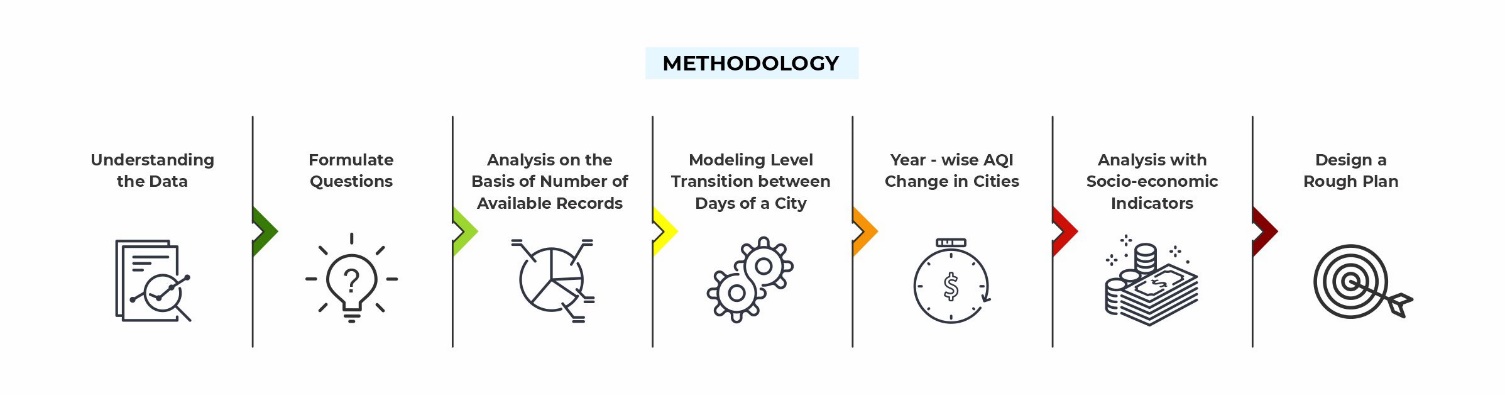
1. **Formulate Questions**
2. How we can get groups of cities using 12 pollutants? (Clustering)
3. How we can predict the Air Quality Index for those groups? (Regression)
4. Which Top 3 cities to implement solutions immediately and what solutions? (Exploratory Analysis)
5. **Methodology for Question 1**

To extract different groups of records based on 12 pollutants, we have used clustering method. KMeans clustering is applied on the data and with the help of Elbow Plot we have identified the number of clusters which can be created from the data.

1. **Methodology for Question 2**

To predict the Air Quality Index of formed clusters, we have used Linear Regression as base model. Also, for best results and predictions we have applied the model using Random Forest Regressor, Light GBM Regressor and XGBoost Regressor.

1. **Methodology for Question 3 (Specially for Objective 3)**

****

As the station\_hour.csv have the most 2589083 records for each station as per hour.

Choosing the methodology to extract the cities prone to severe air pollution was a difficult task. We have used 5 different methodologies in hierarchical manner to filter the three cities from twenty-six cities. Each methodology is explained in following points and some detailed information of sources for these is also given here only.

1. **3-Class Categorization**

We have divided the six categories of AQI into 3 essential parts for the simplicity of filtering out the cities which are mostly polluted. The categories are divided as follows:

If the AQI bucket is Good or Satisfactory, it's put into the **Acceptable** label: Means it does not harm people too much. If the AQI bucket is Moderate, Poor, Very Poor or Severe, it's put into the **Unacceptable** label: Means it can cause harm to a healthy population. **Missing (Not Available)** is a new label that takes into account the missing or null values of AQI buckets => Missing data is a red flag as it indicates poor administration or faulty apparatus.

1. **Using a State Transition Idea to Prioritize Cities based on Air Pollution Levels**

The AQI levels for a city is not uniform throughout the last 5 years. Sometimes, the level has been bad for human respiration and at other times, harmless. However, what if we hypothesized that the **AQI level of a city on day d depends on the AQI level of the same city on day (d-1).** This does follow the Markovian property (though it is not completely based on it).

AQI of a city on any given day is a result of several factors like:

* The population
* The vehicles on the road
* The industries
* and others...

Most of these factors will be common between days. Hence, if a particular city on a given day is on a given AQI level, there is a high chance it would remain in the same the next day. This helps to know a city is **bad** if this level is bad. But it tells nothing about the **volatility of a city's AQI level**. Simply put, we want an ideal city that can change into an AQI level the next day if it has a bad one today.

The working of this method:

1. Convert the AQI levels in the dataset into 4 levels –
   1. Good and Satisfactory: Level 1
   2. Moderate and Poor: Level 2
   3. Very Poor and severe: Level 3
   4. Not Available => Level 4
2. Create a transition matrix showing the probability of a day in a city to change to a Level i if its i's currently in Level j
3. Order the cities based on the probability for each city to transition from one level i to the next level j - If i > j, it is an improvement - If i < j, it is a deterioration.
4. Cities that have high probabilities of deterioration and low probabilities of improvement are the ones that need most focus.
5. **Unacceptable AQI Levels and Indeterminable AQI Levels - New metrics**

From a previous classification, there were 2 categories:

* Acceptable (Good and Satisfactory AQI levels)
* Unacceptable (Moderate, Poor, Very Poor and Severe levels)
* Not Available (Missing AQI levels)

**NOTE:** The **Not Available** classification is better called as **Indeterminable** levels. It means that even if data is collected for individual pollutant levels, the collected data did not conform to the requirements to generate a final AQI bucket for the day.

**METRICS CREATED**

For this section, two key metrics are used:

* Unacceptable AQI Level Percentage (UALP) = (Number of Unacceptable AQI Levels / Number of Records) \* 100
* Indeterminable AQI Level Percentage (IALP) = (Number of Indeterminable AQI Levels / Number of Records) \* 100

Record: A record is registered when there is an entry for a given day in the dataset. A record for a given day does not indicate that the day has a determinable AQI level.

Ideally, a city that is high priority on the AQI index will have a high UALP. Having a high IALP indicates problems with the data collection of the city.

1. **Comparing our top 5 cities on the basis of socio-economic factors**

The analysis has so far focused on the specific AQI data available. However, to choose the most relevant cities to be provided the monetary investment, there is a need for comparing the filtered cities across other indicators. For this purpose, we are comparing the cities across 2 other indicators:

* Per Capita Income of the City (In INR)
* Population of Children under 6 years in the City

**RELEVANCE OF THESE METRICS**

Direct consequences of air pollution are respiratory troubles and health disorders. Naturally, treating health issues is an expenditure. These treatments can at times be extra costs (or burden) on families that are not well off. Therefore, the **per capita income of a city** is a factor in deciding which city needs the investment to reduce air pollution and subsequently decrease medical expenditure due to respiratory problems.

Children are the most vulnerable of all age groups. Also, children have always been the most protected groups across human civilization. It is also a fact that children spend time outdoors playing games with their friends and tend to be in direct contact with air more. Therefore, cities with a larger **child population (under 6 years)** are to be given some consideration.

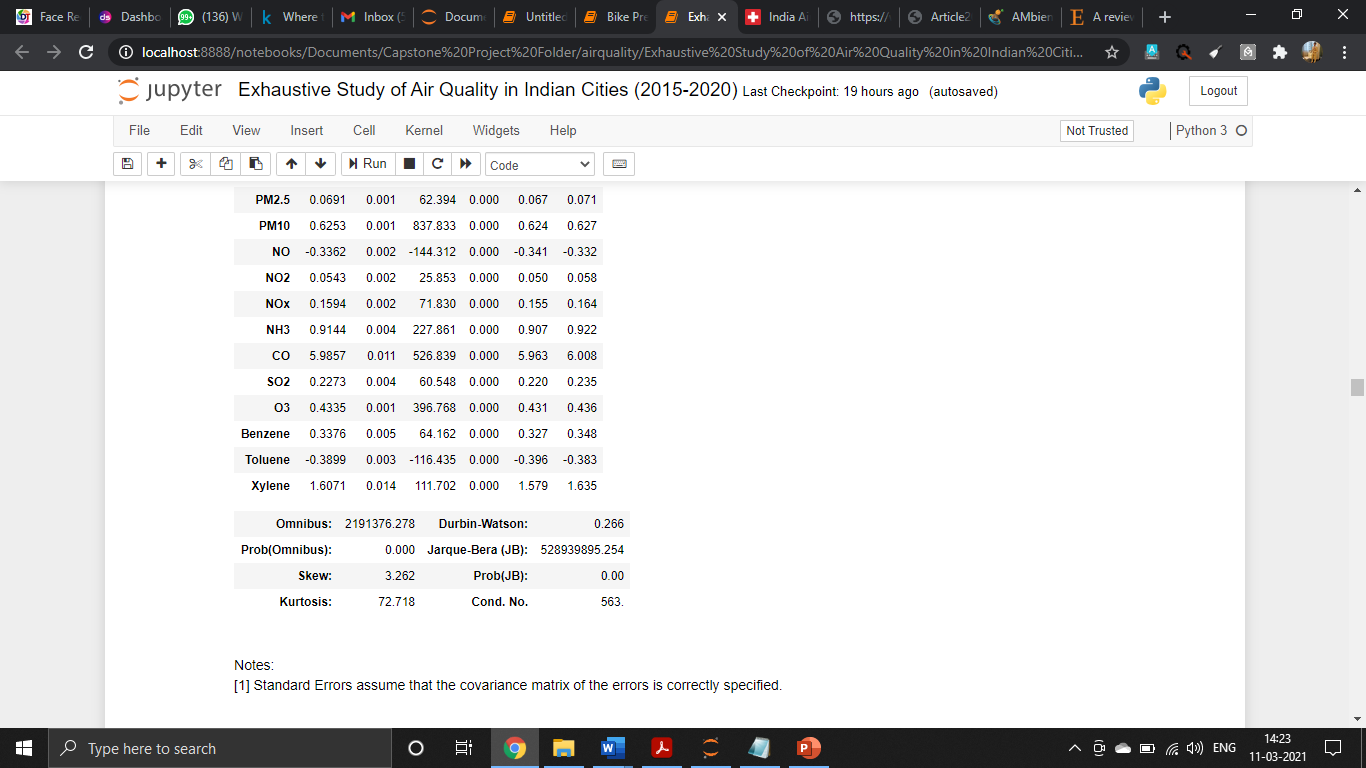
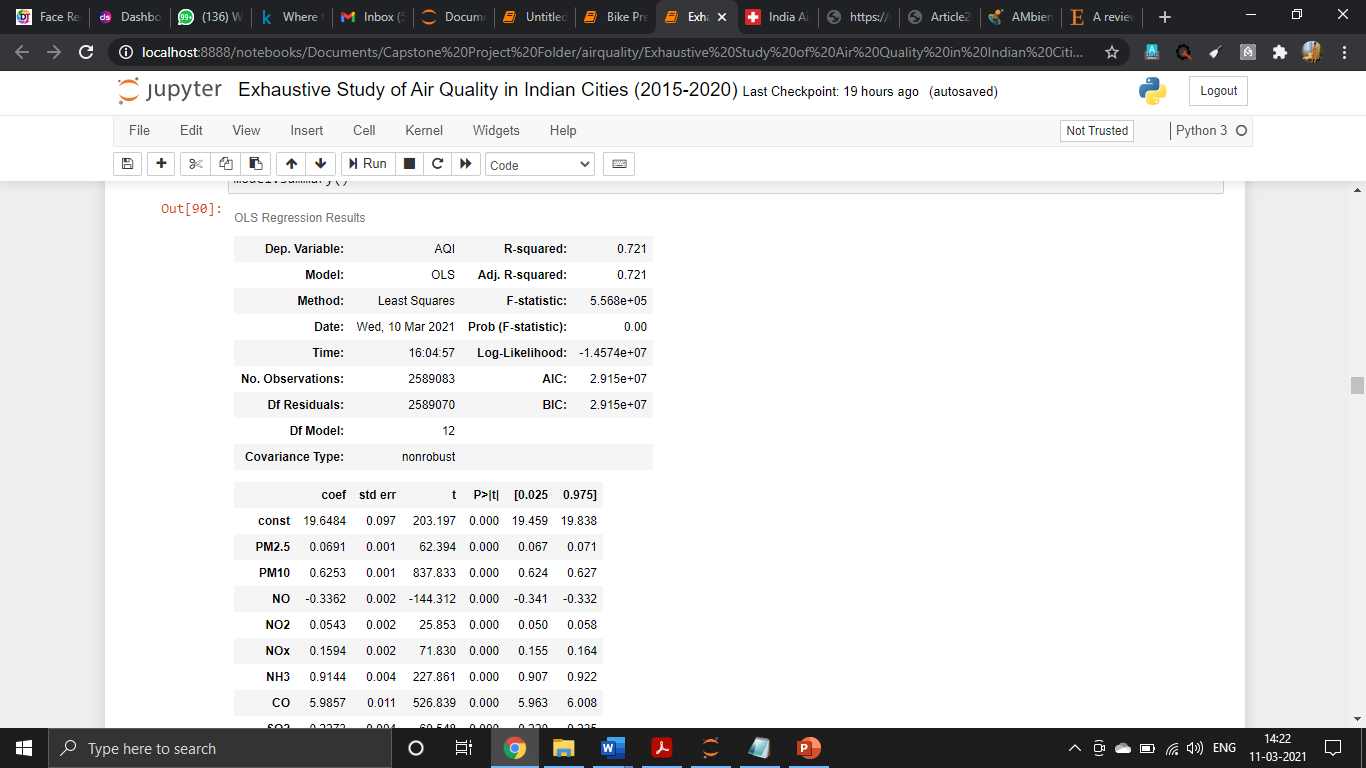
Finally, cities with low per capita income and high child population is to be given a very high priority as these fall under a *region of urgency*.

1. **STEP BY STEP WALKTHROUGH OF THE SOLUTION**

After the study of data files and some exploratory data analysis, we came to a conclusion that for modelling we should use the data files which is more exposed to records. The data file station\_hour.csv is more of records for total 110 stations across the India and the data is available as per hour.

**3.1 OLS Model on Full Data**

To understand the significance of the features, we have applied a OLS model which is clearly showing that, all the features are significant for this model. Before, applying the model we have also checked that our training data for this model is representation of a population from which it is taken.



Observations from the model:

1. The R2 score is showing that, the model on the raw data is working fine. The 0.721 variation in ‘AQI` (dependant variable) is explained by all the ‘pollutants` (independent variables) comes out to be good.
2. Prob (Omnibus): A test of the skewness and kurtosis of the residual. The Prob (Omnibus) performs a statistical test indicating the probability that the residuals are normally distributed. We have got value 0, means the data is normal.
3. Durbin-Watson: A test for homoscedasticity. The value less than 2 means, variance of residuals is constant across the data.
   1. **Checking the assumptions of Linear Regression on the Data**

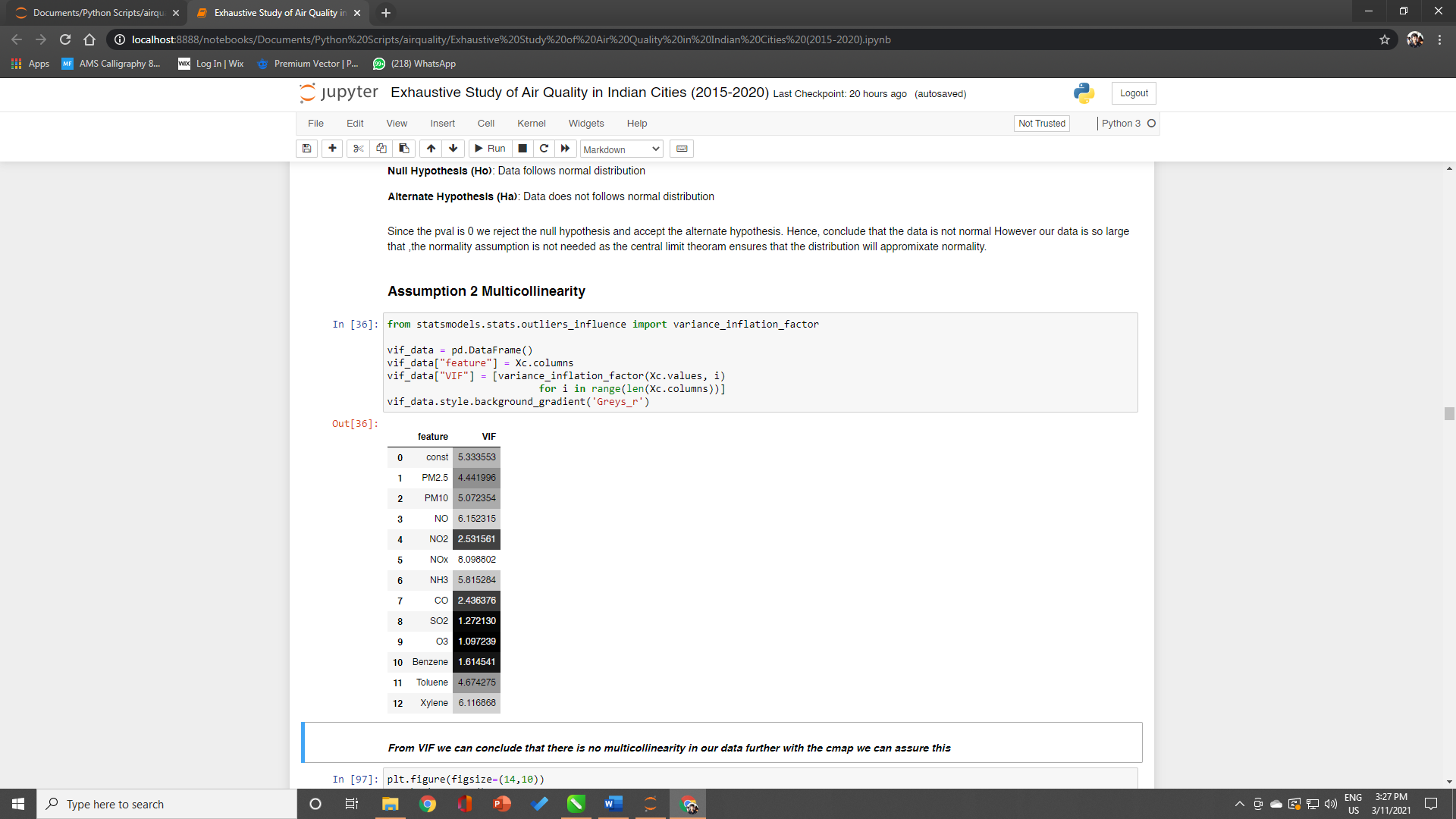
**3.2.1 Tests of Linearity**

Residual density plot: shows the correct distribution of residuals.

QQ plot: shows that the residuals are close to the normality around the means. It deviates from normality at extreme values

Jarque-Bera Test: Since the p-value is 0 we reject the null hypothesis and accept the alternate hypothesis. Hence, conclude that the data is not normal. However, our data is so large that, the normality assumption is not needed as the central limit theorem ensures that the distribution will approximate normality.

**3.2.2 Test for Multicollinearity**

****

From VIF we can conclude that there is no multicollinearity in our data further with the cmap we can assure this more definitely. It implies that, no independent variables in this case our pollutants are not correlated with each other.

**3.2.3 Test for Homoscedasticity**

Gold-Feld Quant Test:

Null Hypothesis (Ho): variance is constant across the range of data

Alternate Hypothesis (Ha): Variance is not constant across the range of data

Since p value is greater than 0.05, Ho is accepted to conclude that variance of residuals is constant across the data

**3.2.4 Test for Linearity in Parameters**

Rainbow Test:

Null Hypothesis (Ho): The parameters are linear

Alternate Hypothesis (Ha): The parameters are not linear

Since, p-value >0.05 we strongly accept the null hypothesis that states our parameters are linear.

**3.3 Clustering of Data using KMeans**

For this as discussed, we have taken up the station\_hour.csv as over data. As the data is so much versatile and there are 26 lakhs rows in the data. So, we decide to take the data from Year 2018 to 2020 so that our systems can take up the pressure of complex models on this data.

**3.3.1 Null-value Imputation of data:**

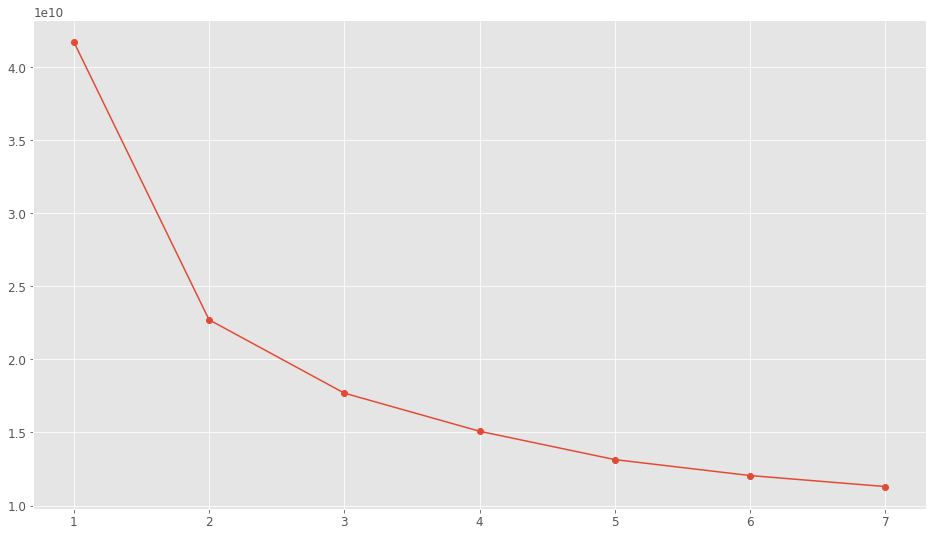
There were so much missing values seen in the data. We have imputed all the missing values with help of Iterative Imputer from sklearn pre-processing module. It is a multivariate imputer that estimates each feature based on the other features.

The AQI for all the records was available, to use Unsupervised Learning algorithm. We have dropped the column of AQI. And then we divided out data into training and validation.

**3.3.2 KMeans Clustering:**

1. Optimal value of k:

We applied the Kmeans clustering algorithm on the training data in a loop for n\_clusters from range 1 to 8. Then by using inertia and cluster errors (Elbow Plot), we got the optimal k value as 2.

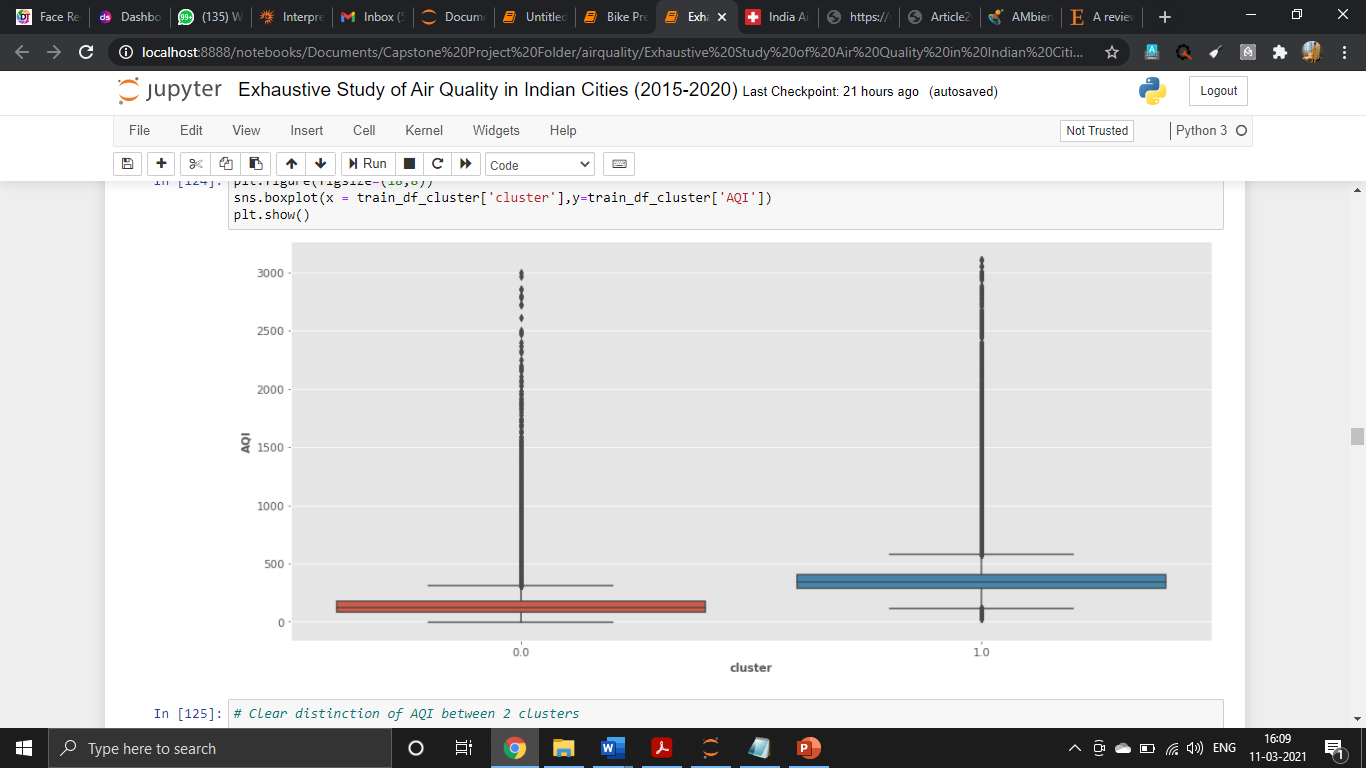


1. Kmeans Clustering Algorithm:

We applied the n\_clusters = 2 in Kmeans clustering algorithm on the training data. It gave the results as label 0 and 1. Also, it gave the cluster centers for each feature in label 0 and 1.

1. Evaluation of Kmeans Clustering

We again appended the AQI feature to the data and now with one added feature of labels predicted using Kmeans. From 12,97,979 records in this data. The label 0 is contaning 1102010 records and label 1 is contaning 1,95,969 records from whole data. We have got clear distinction in the two clusters, is shown below in the boxplot.



Most of the values of the cluster 0 and 1 lies between 0-500 which is practically correct. The label distinction clearly says that, we can divide the six categories of air quality index in 3 prime and main categories.

As per as we have divided clusters, we also calculated the mean AQI for two labels for label 0 it was 140.26 and label 1 it was 358.47. Now, it is more understanding with simple thoughts. Cities with good AQI are in label 0 and cities with poor AQI are in label 1.

**3.4 Predicting the AQI for formed clusters**

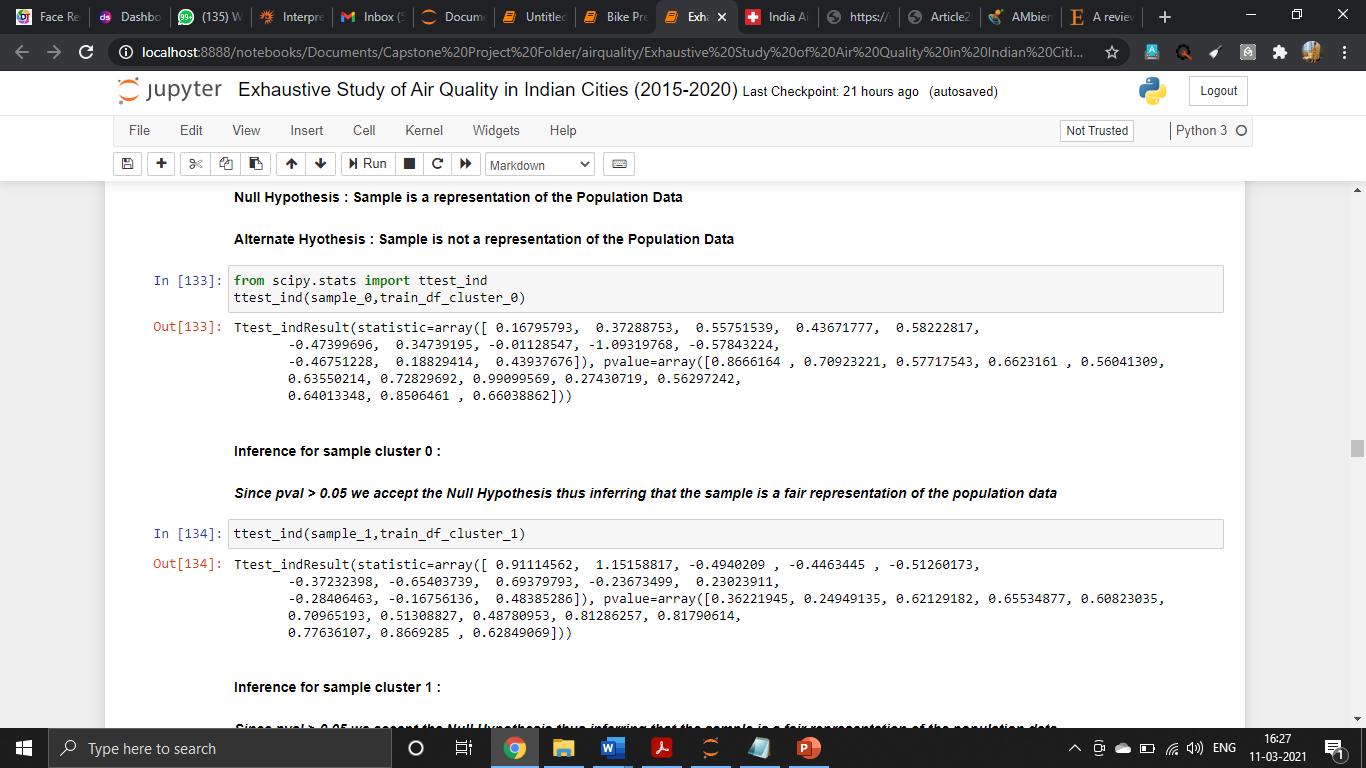
**3.4.1 Taking sample from data:**

As the data is huge, we experienced a load on our system and as per discussion with our mentor, we decided to take sample from the data. After taking the sample we checked the statistical significance of the data that it is correctly representing the population data. We have extracted out 40% of the sample from the label 0 and label1.

Null Hypothesis: Sample is a representation of the Population Data

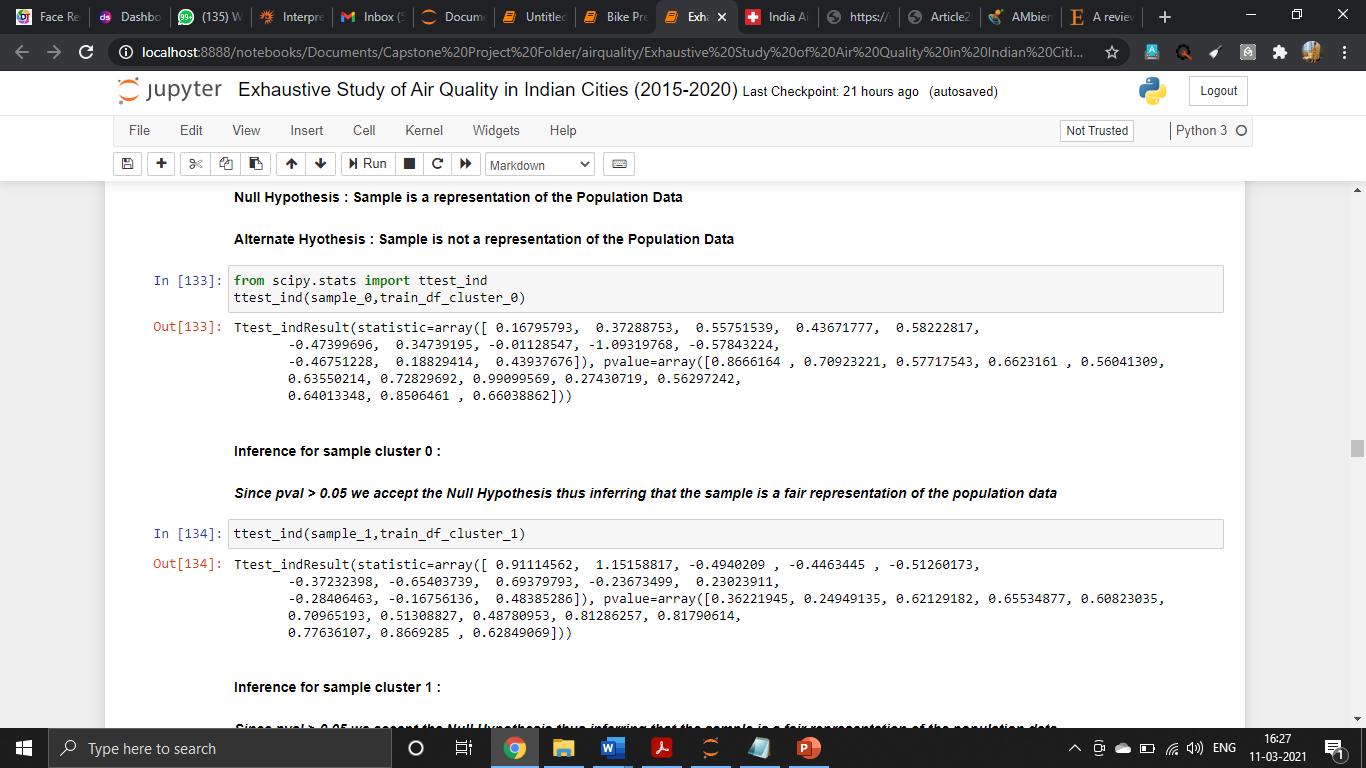
Alternate Hypothesis: Sample is not a representation of the Population Data

Ttest independent for cluster 0:



Observations: Since pvalue > 0.05 we accept the Null Hypothesis thus inferring that the sample is a fair representation of the population data

Ttest independent for cluster 1:



Observations: Since pvalue > 0.05 we accept the Null Hypothesis thus inferring that the sample is a fair representation of the population data

Same process is repeated for the test data of clusters.

**3.4.2 Base Linear Regression Model on Clusters:**

After taking the samples from the cluster 0,1 and taking data for training and validation. We applied the linear regression base model on the data. And predicted the AQI values for cluster 0 and cluster 1.

The results for this model are:

|  |  |
| --- | --- |
| Evaluation Metric | Results |
| R2 Score for cluster 0 | 0.61 |
| R2 Score for cluster 1 | 0.50 |
| RMSE Score for cluster 0 | 54.32 |
| RMSE Score for cluster 1 | 120.57 |

The results are not so good and as per industry standards we should try different regression algorithms to get better results and prediction than this base model.

**3.4.3 Random Forest Regressor**

From sklearn ensemble we applied the Random Forest Regression algorithm which is based in boosting. This time prediction and results precision increased tremendously.

The results for this model are:

|  |  |
| --- | --- |
| Evaluation Metric | Results |
| R2 Score for cluster 0 | 0.75 |
| R2 Score for cluster 1 | 0.77 |
| RMSE Score for cluster 0 | 43.13 |
| RMSE Score for cluster 1 | 78.78 |

These results were impressive and then we decided to tune some parameters of Random Forest Regressor using Randomized Search Cross Validation. After tunning the results were quite shocking. The results have fluctuated a deviated from the best to low level.

The results for **Random Forest Tunned** model are:

|  |  |
| --- | --- |
| Evaluation Metric | Results |
| R2 Score for cluster 0 | 0.68 |
| R2 Score for cluster 1 | 0.72 |
| RMSE Score for cluster 0 | 49.33 |
| RMSE Score for cluster 1 | 88.28 |

**3.4.4 Light GBM Regressor**

The light GBM regressor is a heavy algorithm and uses a histogram-based algorithm to find the optimal split point while creating a weak learner. These results were not good as Random Forest Regressor without tunned model. But the RMSE is controlled that shows that it was looking after overfitting of model.

The results for this model are:

|  |  |
| --- | --- |
| Evaluation Metric | Results |
| R2 Score for cluster 0 | 0.70 |
| R2 Score for cluster 1 | 0.74 |
| RMSE Score for cluster 0 | 47.50 |
| RMSE Score for cluster 1 | 83.92 |

Just because the overfitting was in control from the above model, we tune it got the good results from Light GBM Regressor. The RMSE decreased slightly and R2 for both the clusters increased.

The results for **Light GBM Regressor Tunned** model are:

|  |  |
| --- | --- |
| Evaluation Metric | Results |
| R2 Score for cluster 0 | 0.75 |
| R2 Score for cluster 1 | 0.77 |
| RMSE Score for cluster 0 | 43.38 |
| RMSE Score for cluster 1 | 79.78 |

**3.4.5 XGBoost Regressor**

XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. We applied the XGBoost on our data expecting increase in R2 and decrease and RMSE.

The results for **XGBoost Regressor** model are:

|  |  |
| --- | --- |
| Evaluation Metric | Results |
| R2 Score for cluster 0 | 0.72 |
| R2 Score for cluster 1 | 0.77 |
| RMSE Score for cluster 0 | 45.71 |
| RMSE Score for cluster 1 | 84.35 |

We were not quite impressed by these results and we decided to tune the model. And look at the results with better expectations.

The results **XG Boost Regressor Tunned** model are:

|  |  |
| --- | --- |
| Evaluation Metric | Results |
| R2 Score for cluster 0 | 0.75 |
| R2 Score for cluster 1 | 0.79 |
| RMSE Score for cluster 0 | 43.23 |
| RMSE Score for cluster 1 | 78.61 |

As these results were so impressive and the R2 is increased potentially and RMSE has dropped up till good level. We considered this model as our final model for the modelling on our new data.

1. **MODEL EVALUATION**

Good metrics for regression model are R2 score and RMSE.

**R2**

R2 tells us the variation in the dependent features explained by all the independent features in the data. For our final model applied on two clusters 0 and 1 using XGBoost Regressor after tunning.

We have got R2 of 0.75 for cluster 0, practically there was in and around 11 lakh records in cluster 0 for the tremendous data this R2 score is good.

For cluster 1, we have got R2 of 0.77, this cluster was having competitively less records which is good that, records with more AQI are less that is anyways good for any country. And the score obtained from the final model is very good.

**RMSE**

Root mean square error tells us the difference between the actual and predicted values.

The RMSE we have got for cluster 0 is 43.23 which is good that in spite of having lot of data, the score is good. May be just because the mean AQI in this cluster was 140, it is less.

For cluster 1, we have got RMSE 78.61 which is also good. This cluster was having mean AQI of 354, that is why we have got increased RMSE than cluster 0. But these scores are very good in practical sense as we are analysing data of 3 years.

1. **COMPARISON TO BENCHMARK**

The objective to cluster the data and predict the AQI was achieved with these above steps and we have got a good model to be called as our Final model. The base model R2 scores were near 0.50 which was too bad but, eventually with using heavy and complex regressors we achieved R2 score 0.78. The RMSE scores were also too high with respect to mean AQI values for those corresponding clusters. The RMSE for both clusters we got from the final model is good.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Evaluation Metrics** | **Linear Regression Base Model** | **Random Forest Results** | **Random Forest Tunned Results** | **Light GBM Results** | **Light GBM Tunned Results** | **XGBR Results** | **XGBR Tunned Results** |
| **R2 Score for cluster 0** | 0.61 | 0.75 | 0.68 | 0.70 | 0.75 | 0.72 | **0.75** |
| **R2 Score for cluster 1** | 0.50 | 0.77 | 0.72 | 0.74 | 0.77 | 0.77 | **0.79** |
| **RMSE Score for cluster 0** | 54.32 | 43.13 | 49.33 | 47.50 | 43.38 | 45.71 | **43.23** |
| **RMSE Score for cluster 1** | 120.57 | 78.78 | 88.28 | 83.92 | 79.78 | 84.35 | **78.61** |

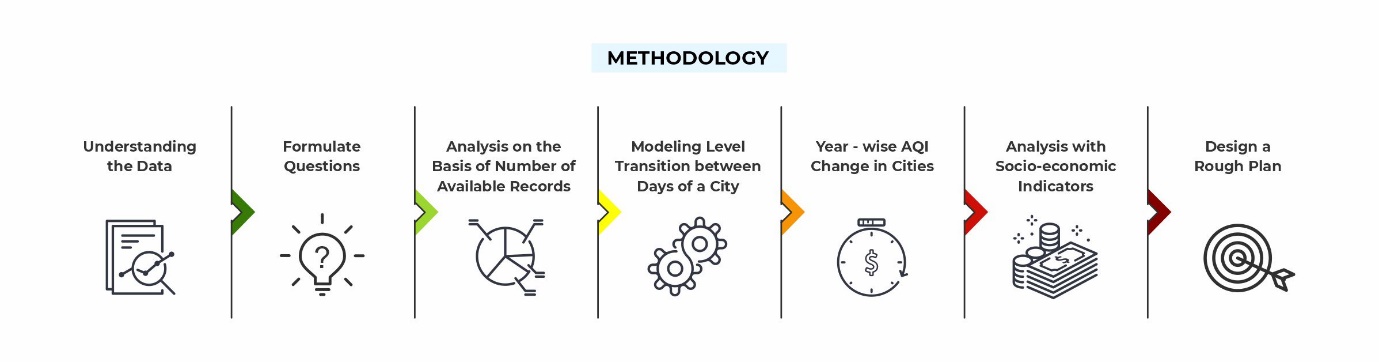
In the above table, there are all the model results comprised at one place. Here it is easy to demonstrate the importance of individual model for our data. It is seen that, XGBR Algorithm results in best and final model for project. Also, we can easily track the improvement in scores we have got across all the different algorithms.

The data of AQI is ideal for modelling if used correctly, the features of pollutants are good as the data satisfies about all the assumptions of linear regression. Right from Normality to constant variance of residuals. We have used sample from the data every time just because our systems were not that compatible to run this much data which was originally extracted from the sources. The reason for defining for defining 3 objectives out of this data was to use the data properly for many good outcomes. The data is original and premium, it is reflecting the current scenario of our country which can be good if we worked in correct way.

As this data was exposed by 3 years from 2018-2020, this model has proven the good skills to predict the AQI after clustering the data. We think that, if on systems with high and good specifications this data of total 5 years will work magnanimous.

**6. VISUALIZATIONS AND IMPLICATIONS OF THE DATA**

We have achieved the two objects of this project and now the third objective is the most crucial and important with respect to development of the country and better implementation of mitigation strategies to control air pollution in the top 3 cities. Point to note, we are using the data from 2015-2020 for this objective. We would like to now again focus upon our methodology for objective 3.



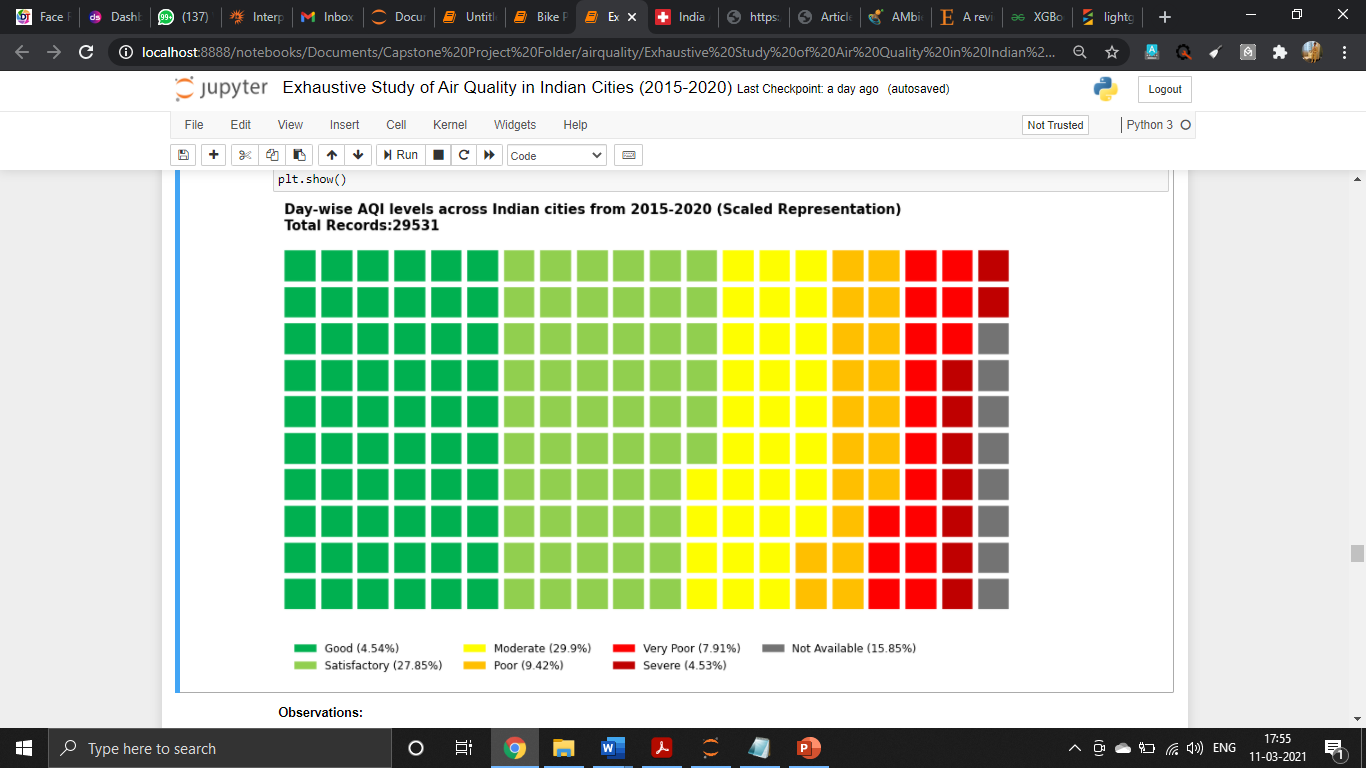
To find the top 3 cities to implement mitigation strategies and where to invest to combat air pollution! We have worked with 5 different methodologies which we have already explained in the methodology part. Here we will take look upon the results we have got by using those methodologies. Also, for the last methodology from that, to select top 3 cities from top 5 by comparing socio-economic factors we have used another data of which the source and to create some new features we sourced the data from some websites those are also mentioned here.

**Source of Data**: The two indicators used in this section have been compiled from the following sources:

* Population of Children under 6 years in the City: [Top 500 Indian Cities - Based on 2011 Census](https://www.kaggle.com/zed9941/top-500-indian-cities)
* Per Capita Income of the City (INR):
* Compiled manually from multiple sources
* Ahmedabad: <https://www.prsindia.org/sites/default/files/budget_files/State%20Budget%20Analysis%20-%20Gujarat%202020-21_Final.pdf>
* Delhi: <https://www.prsindia.org/sites/default/files/budget_files/Delhi%20Budget%20Analysis%20-%202019-20.pdf>
* Gurgaon: <https://economictimes.indiatimes.com/work-career/best-cities-to-move-into-if-you-are-starting-a-new-career/gurgaon/slideshow/55224194.cms>
* Patna: <https://patna.nic.in/economy/#:~:text=As%20of%202015%2C%20GDP%20per,rate%20is%207.29%20per%20cent>.
* Lucknow: <https://www.hindustantimes.com/lucknow/kasganj-s-per-capita-gdp-fourth-highest-in-uttar-pradesh/story-8tSlU6yrjAeF9H92835X5I.html>

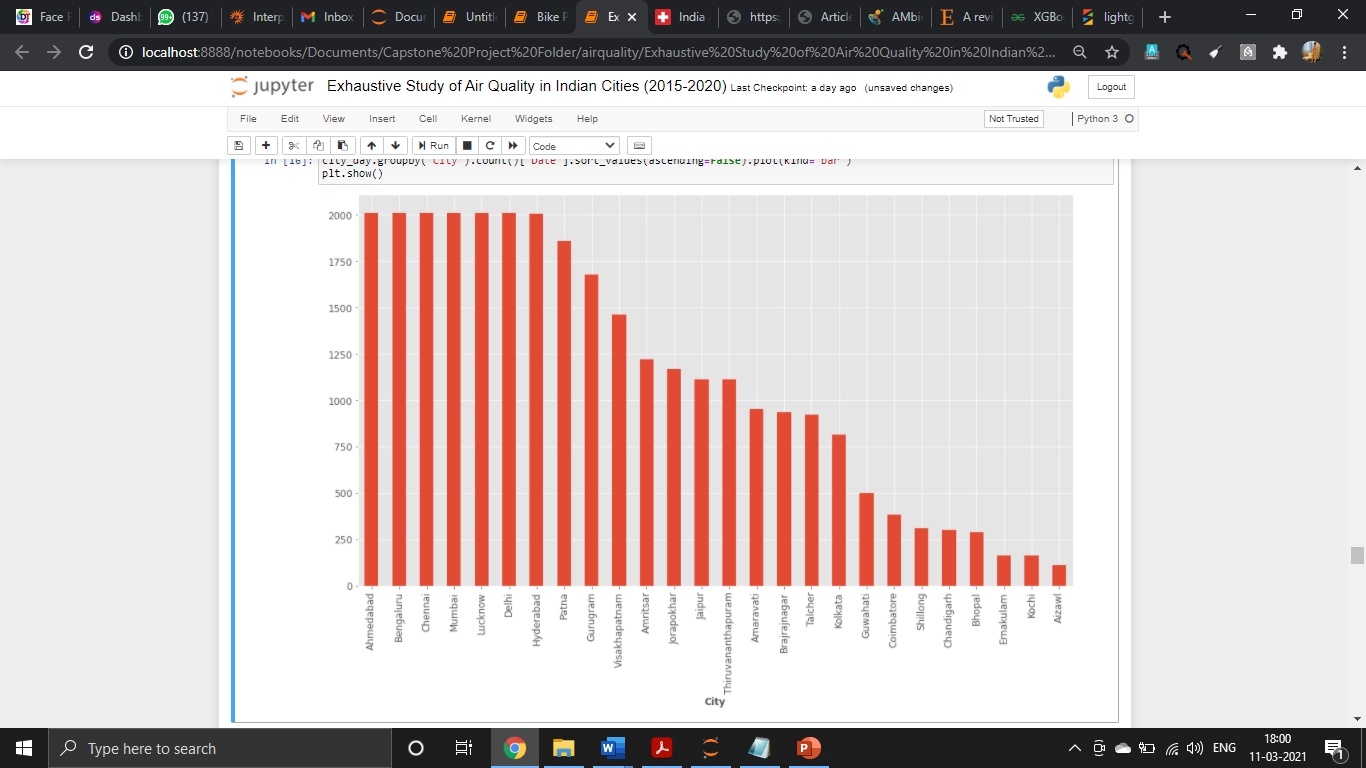
**NOTE:** The data above has been collected through multiple different sources as there was no single source to get this from. There is some difference between the years when each city's per capita income was collected was collected.

**6.1 Methodology 1: Using AQI, Day-wise AQI levels across Indian cities from 2015-2020 (Scaled Representation)**

****

This distribution of categories is shown from total 29,531 records. Only 5% of all entries recorded have an AQI that would inflict no harm on any section of the demography. Categories Poor, Very Poor and Severe (i.e., the ones that are capable of harming healthy people) contribute to around 1/5th of all entries. 16% of all entries have missing AQI levels!

**6.1.a Records Available as per city and their categorization**

****

**Observations:** The most records a city in the given dataset has is 2009. 12 out of the 26 cities have total day-wise records for 2015-2020 which is lesser than the maximum number of records available for a city i.e., 2009 for Mumbai, Delhi, Lucknow, Chennai, Bengaluru, Ahmedabad, Hyderabad

Note: We are setting a threshold of 1500 (which is 50%) to divide the cities in two groups respectively WMR and WLR

**WMR: With More Records** and **WLR: With Less Records**

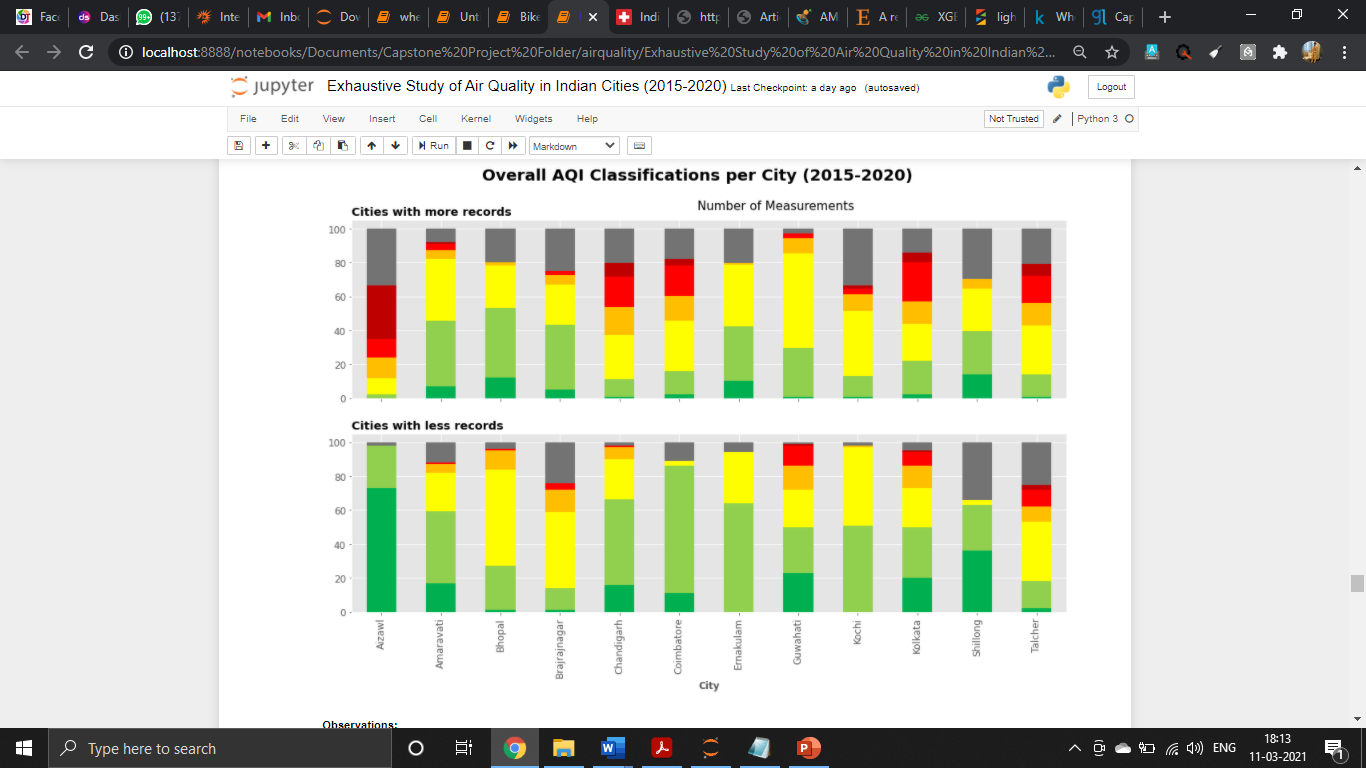
Why do the WLR states have low records?

**Hypothesis 1**: Maybe they don't have the resources

**Hypothesis 2**: Maybe they are not prioritized by AQI

Cities with low records may or may not be crossed out for investment on the basis of whether, they have low records because they are clean and safe. They have low records because they are not efficiently monitored.

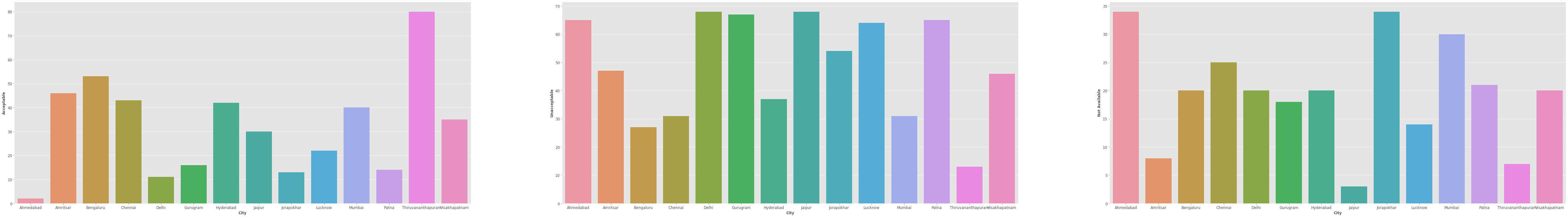
**6.1.b Overall AQI Classifications per City (2015-2020)**

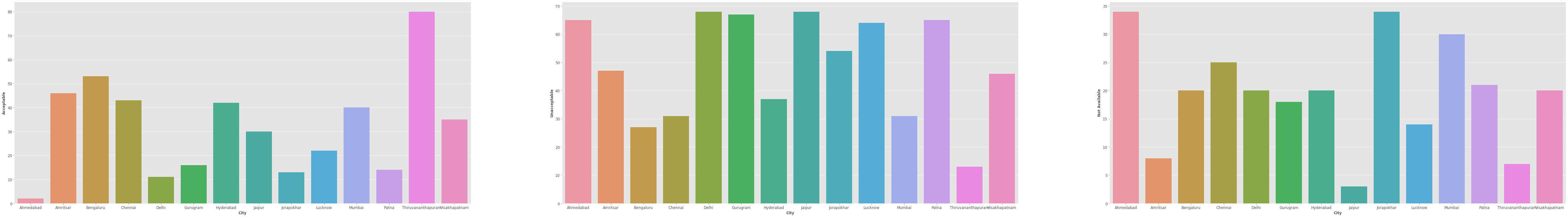


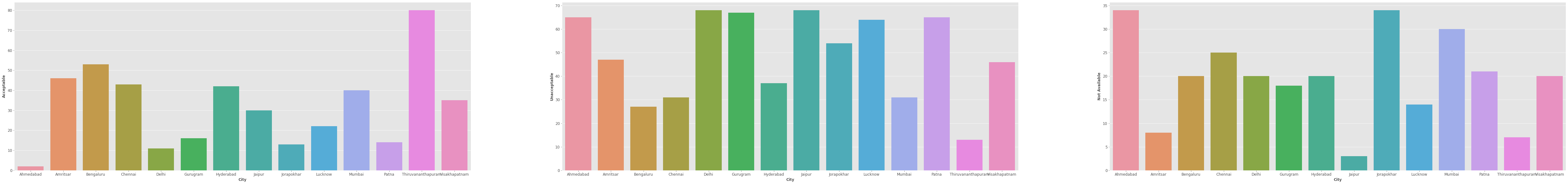
**Observations:** In a general look, it is evident that 'Cities with less records' have measured more Good or Satisfactory classifications than the 'Cities with more records'. In the first set, Thiruvananthapuram registered the highest percentage of its daily measurements as of 2015-2020 as "Good" or "Satisfactory". Ahmedabad on the other hand has less than 5% of its entries from 2015-2020 under the positive category (Good or Satisfactory). Ahmedabad also has close to 35% of its entries under the 'Not Available' status.

**6.2 Methodology 2: 3-Class Categorization**

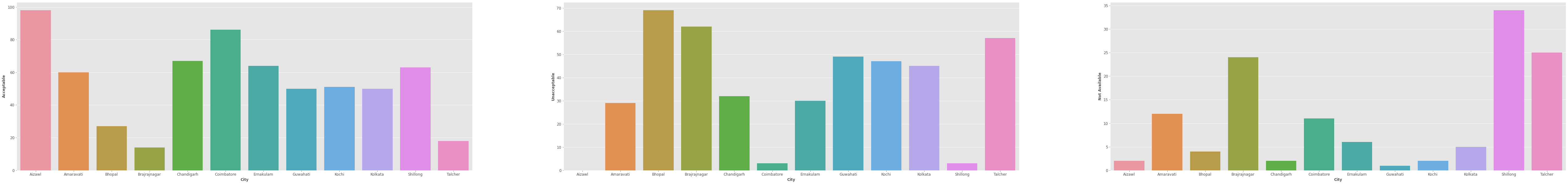
**6.2.a These is the 3-class categorization for WMR (With more Records) cities**

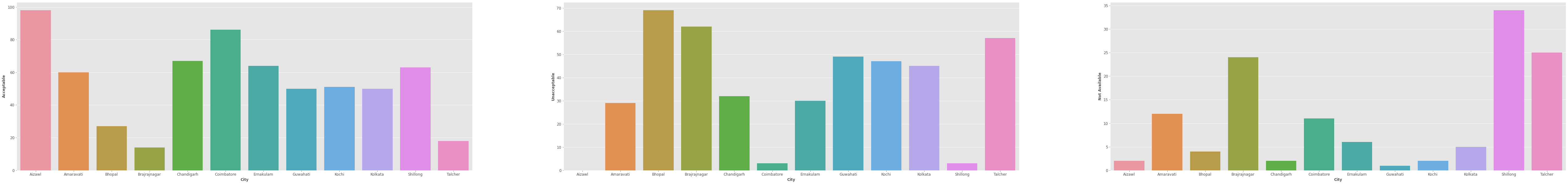


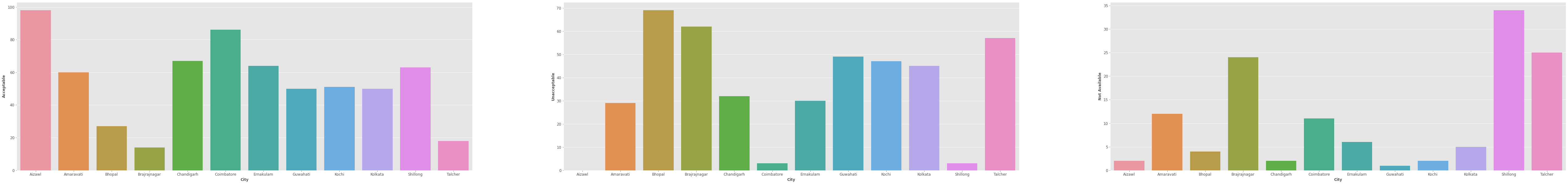




**These is the 3-class categorization for WLR (With Less Records) cities**







If the AQI bucket is Good or Satisfactory, it's put into the Acceptable label >> Means it does not harm people too much. If the AQI bucket is Moderate, Poor, Very Poor or Severe, it's put into the Unacceptable label >> Means it can cause harm to a healthy population. Missing (Not Available) is a new label that takes into account the missing or null values of AQI buckets >> Missing data is a red flag as it indicates poor administration or faulty apparatus

**Observations:** Over the last 5 years, Ahmedabad has had the lowest number of Acceptable levels, 3rd highest number of Unacceptable levels and the highest amount in terms of Missing levels in all WMR cities**.** Other cities that look troubled are Delhi, Gurugram, Patna, Lucknow, Jaipur and Jorapokhar**.** Thiruvananthapuram is a happy outlier in that top bar plot**.** Amongst the WLR cities, Bhopal looks the most troubled. Other cities in WLR that have recorded more Unacceptable days than Acceptable ones are Brajrajnagar and Talcher.

**Which cities are under the radar as of now?**

With the analysis so far, a few cities have emerged as potential recipients of the monetary funding to improve their state. These are:

Ahmedabad: A very high percentage of the days it has registered measurements show unacceptable AQI levels

Delhi: Same as Ahmedabad. Also, it's highly discussed in national and international media

Kolkata: The more stations, but a smaller number of records phenomena puts Kolkata under scrutiny for poor administration

Other cities that make it to this list are: Gurugram, Patna, Lucknow, Jaipur, Jorapokhar, Bhopal, Brajrajnagar, Talcher.

**6.3 Methodology 3: Using a State Transition Idea to Prioritize Cities based on Air Pollution Levels**

We have divided the AQI levels in the dataset into 4 levels

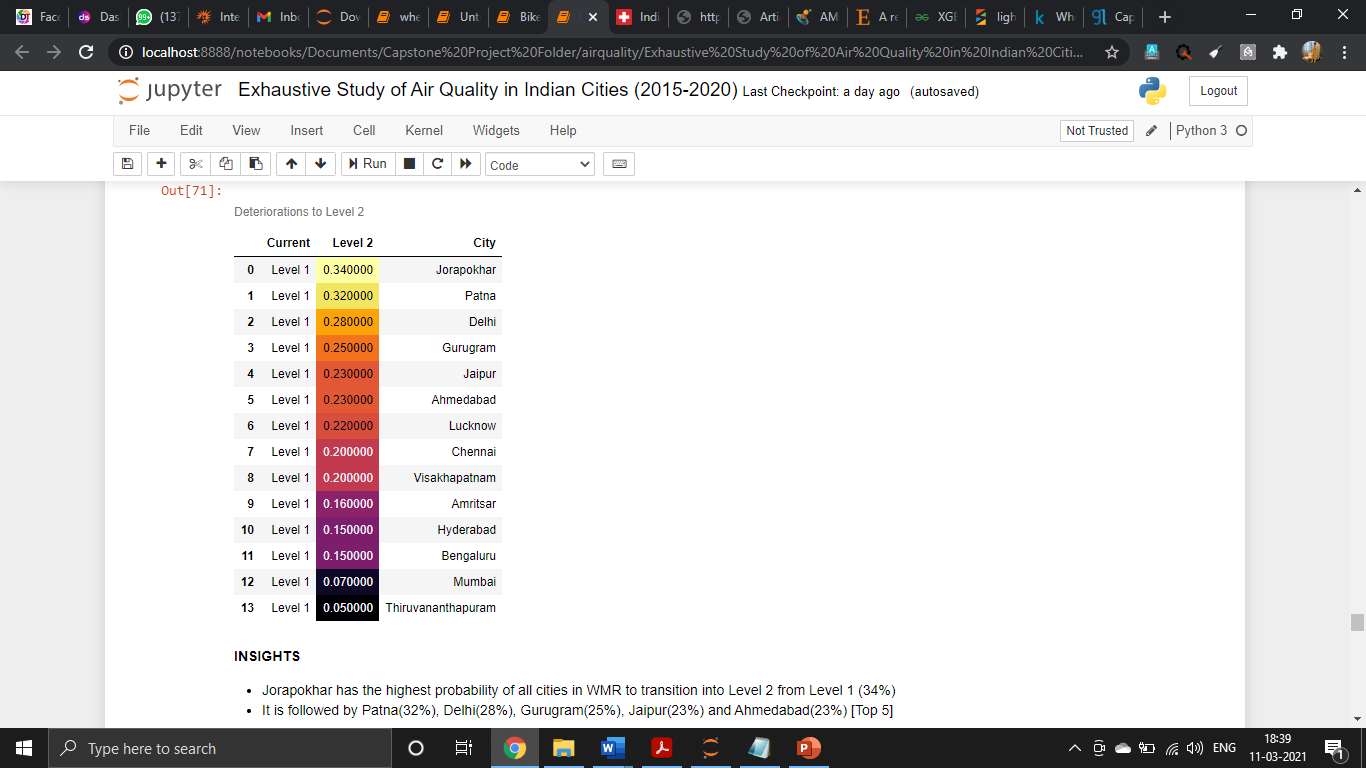
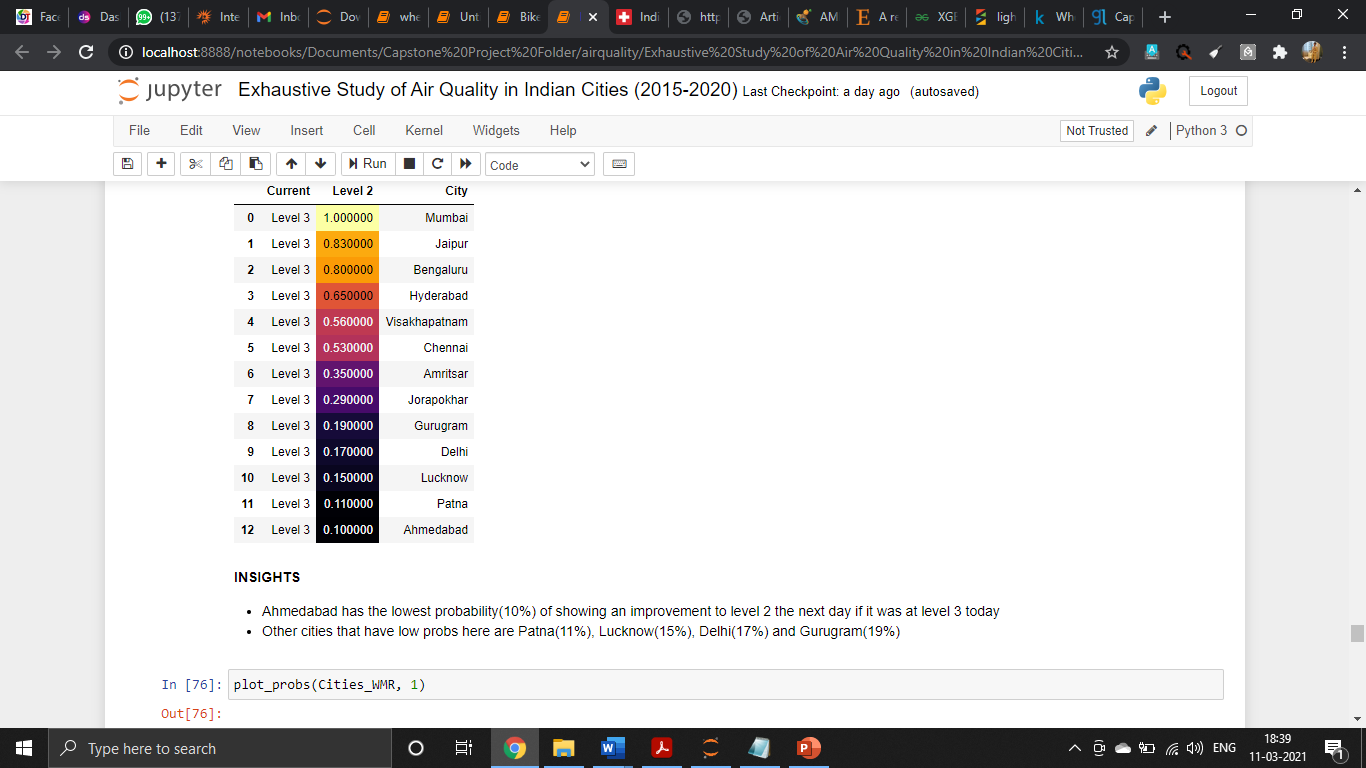
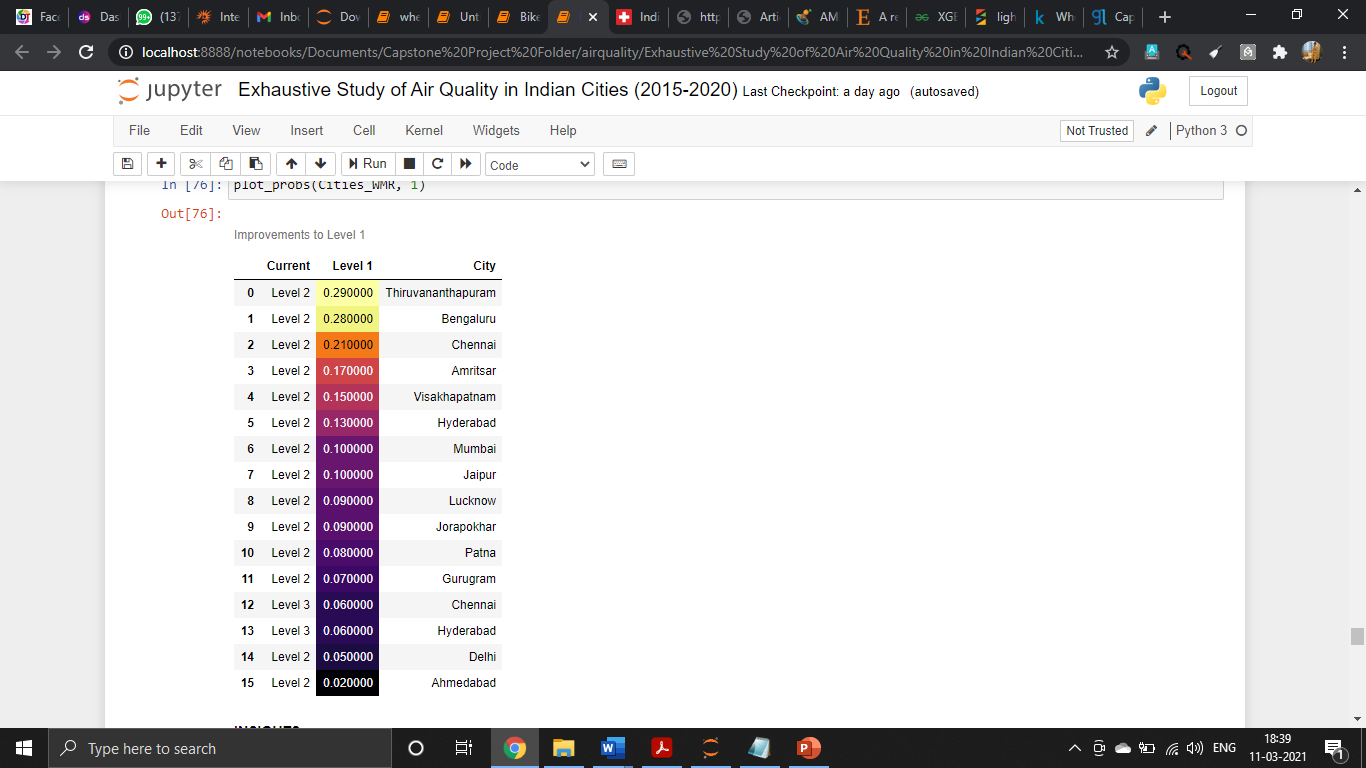
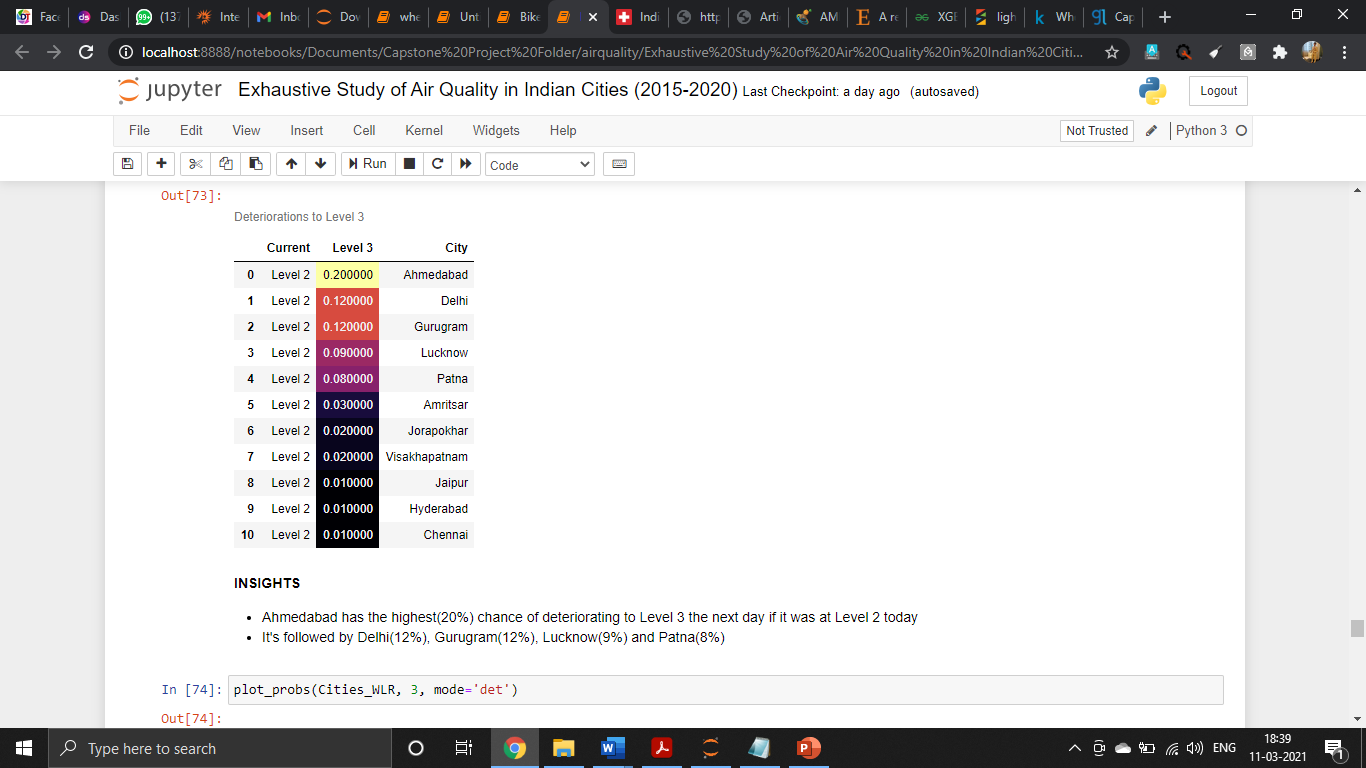
Good and Satisfactory: Level 1

Moderate and Poor: Level 2

Very Poor and severe: Level 3

Not Available: Level 4

And now we will order the cities based on the probability for each city to transition from one level i to the next level j. We will do it in two group again with WMR and WLR.



**From Level 1 to 2:** Jorapokhar has the highest probability of all cities in WMR to transition into Level 2 from Level 1 (34%). It is followed by Patna (32%), Delhi (28%), Gurugram (25%), Jaipur (23%) and Ahmedabad (23%) [Top 5]

**From Level 2 to 3:** Ahmedabad has the highest (20%) chance of deteriorating to Level 3 the next day if it was at Level 2 today. It's followed by Delhi (12%), Gurugram (12%), Lucknow (9%) and Patna (8%).

**From Level 3 to2:** Ahmedabad has the lowest probability (10%) of showing an improvement to level 2 the next day if it was at level 3 today. Other cities that have low probs here are Patna(11%), Lucknow (15%), Delhi (17%) and Gurugram (19%).

**From Level 2 to 1:** Ahmedabad has the lowest probability (2%) of showing an improvement to level 1 the next day if it was at level 2 today. Other cities that have low probs here are Delhi (5%), Hyderabad (6%), Chennai (6%), Gurugram (7%) and Patna (8%).

The same process is repeated for WLR cities too.

With the above insights provided by the **State Transition Idea**, we now can quantify the priority of each city involved more definitively.

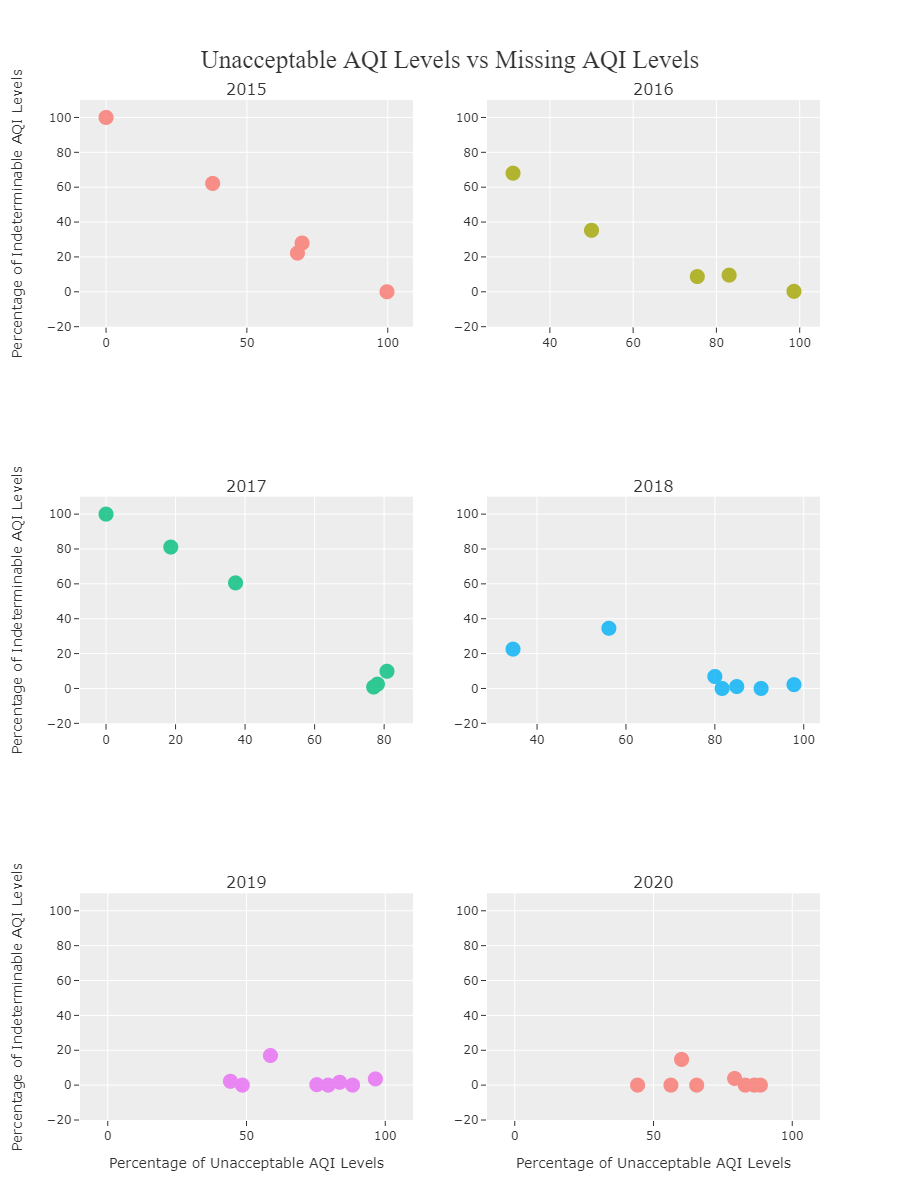
**Which cities are under the radar now?**

As seen in all the analyses of the previous sections, **Ahmedabad** is a **high-priority city** when it comes to AQI. It has the highest probabilities to deteriorate from one level to the other (5th highest for level 1 >> level2 and highest for level2 >> level3). It also has the lowest probabilities to improve from one level to another. Therefore, Ahmedabad is most likely the city that requires the initial monetary funding to improve its pollution.

Other cities that are still under consideration for the second and third spots are: Patna, Delhi, Gurugram, Lucknow, Kolkata, Talcher and Guwahati.

**6.4 Methodology 4: Unacceptable AQI Levels and Indeterminable AQI Levels - New metrics?**

We know that, how we calculated UALP and IALP in the methodology section. We have plotted the graph for Unacceptable AQI Levels vs Missing AQI Levels for Year 2015 – 2020.

****

**Observations:** As the years progress, a general trend sees the cities move towards the bottom right. The bottom right is a region with high UALP and low IALP. This indicates that data collection has in general become better for these cities over the given time period. The year 2017 is an outlier in the sense that both Ahmedabad and Patna have shown a higher IALP measure than what they did have in 2016. Especially, Patna with a rise of about 50% from 2016 in the IALP. This however is restored to 1% in 2018. This does cast reasonable doubt. The following states are removed from this list of 8 states for the following reasons:

Talcher: It has a population of around 40,000 only which is way lesser than the other cities. So, a fair comparison is not possible

Guwahati: Data available only for 2019 and 2020

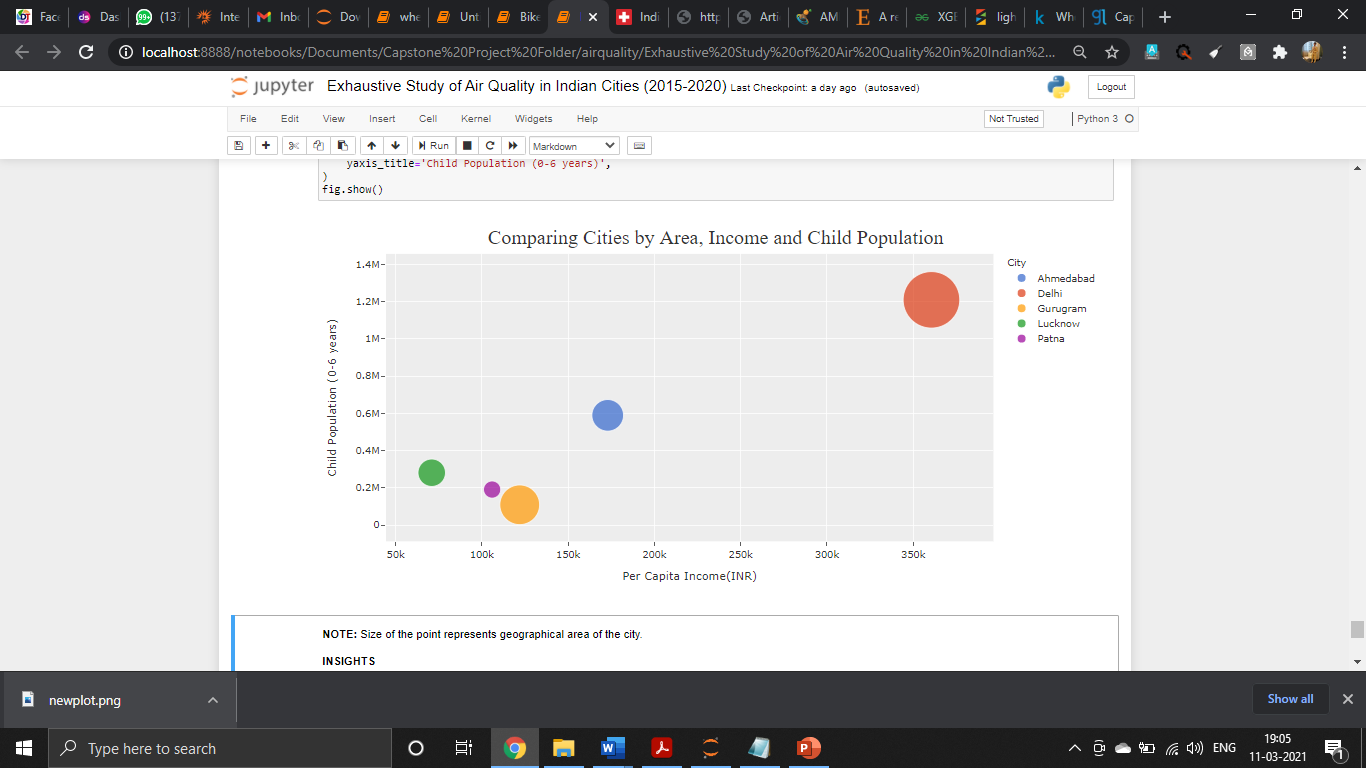
Kolkata: Data available for only 2018,2019 and 2020

**NOTE:** In an earlier statement, I had cast doubt on Kolkata's administration because of the lower number of records in spite of having more stations in place. This could be because Kolkata has only begun registering records since 2018(3 years lesser than the time for most other cities in this data)

The above visualization does provide insights. But a more improve way of seeing a pattern would be using a bubble chart with only the 5 top cities (Ahmedabad, Delhi, Patna, Gurugram and Lucknow) in consideration.

**6.5 Methodology 5: Comparing our top 5 cities on the basis of socio-economic factors**

Now we have our top 5 cities and this will be our last metric to filter out top 3 cities from these. The Area in sq.km, child population from age 0-6 and per-capita income of that city in INR is compared in this single graph.



NOTE: Size of the point represents geographical area of the city.

**Observations:** Delhi has the highest child population, but also the largest per capita income. In fact, Delhi is an outlier in this 5-city group!Ahmedabad has already been identified as a highly polluted city w.r.t the **State/Level Transition Diagram** analysis before**.** In the above bubble plot, it can also be noticed that Ahmedabad has a high child population value. The other 3 cities are of specific interest**. Lucknow** is a city with a higher child population and lower per capita income than Patna and Gurugram. This makes it a *region of urgency* (relative to these 5 cities)**. Patna** is between Lucknow and Gurugram in terms of the two factors considered here**. Gurugram** has been noticed to have a sharp decline in AQI w.r.t to the other cities in previous analysis. However, it has a very low count of child population and a higher per capita income than the Patna and Lucknow.

Based on the analysis so far, **Ahmedabad** is the city that should receive the initial funding for 3 years. This is because of the findings in the Level Transition Diagram-based approach that indicates Ahmedabad to have constantly deteriorating air quality and long periods of bad air quality and very little subsequent natural improvement. **Lucknow and Patna** are to be the 2 cities that should be funded if the funding to Ahmedabad is successful. This is because of the relative importance to these regions due to low income and high child population.

**Why not Gurugram?**

Gurugram has shown steep decline, no doubt. However, it's a city with a larger area. Moreover, being one of the fastest growing areas in the country the amount required to perform any kind of improvement or initiative to curb air pollution will most likely be significantly higher. Since there is such a doubt cast, Gurugram has been excluded from this top 3 list.

**Why not Delhi?**

Same reasons as that for Gurugram. But, in addition to that Delhi's air pollution problem is internationally discussed. Everybody is talking about it already. Therefore, it will probably receive funding from other sources. It is the other cities that are only placed in "lists of most polluted cities" and see little media visibility that require funding to improve the lives of its citizens.

**6.6 Rough Plan to Use the Investment**

****

**6.6.a Invest in Clean Technology**

Clean technology is any process, product or service that reduces negative environmental impacts through significant energy efficiency improvements, the sustainable use of resources, or environmental protection activities. (<https://en.wikipedia.org/wiki/Clean_technology>). A couple of ways the uncle could invest his money would be to

1. Set up a Money-lending firm or bank

- Provide low-interest loans for people to convert their vehicles into CNG-run from petrol-run or diesel-run

- Provide incentives to those who are willing to give away their old fuel-driven vehicles and switch to electric-vehicles. The extra money could be used by the family for a purpose like long-term deposits for children

2. Launch an Electric Carpool Start-up ask

- The traditional carpool system with only electric cars

- Provide free rides and offers to attract people

- This can generate livelihood for cab-drivers without causing pollution to the environment

- Autorickshaw drivers can be targeted and brought into the revolution as (<https://bengaluru.citizenmatters.in/e-rickshaws-air-pollution-bengaluru-policy-transport-28136#:~:text=According%20to%20it%2C%20in%20a,sector%20is%200.44%20million%20tonne> )

- Under this start-up, more solar-powered battery charge stations for any e-vehicle can be setup. This is important, sans this everything else fails

**6.6.b Invest in Community-driven Change**

Initiatives fail because people are not ready to accept it. People are not ready to accept because most initiatives and ideas are brought into effect without a public study or social research. Therefore, investment has to be put into creating an organization that would ask people questions and act on their responses. Basically, bring about effective change in existing behaviour of citizens, by involving them as key stakeholders in any relevant policy decision.

- Ask the Right Questions

- Use social surveys to collect data

- Make data open, but maintain ethical standards

- Understand the \*why\* behind behaviours

- Why do people not change their polluting vehicles?

- Why do people not adhere to government regulations?

- Why do people not want to carpool?

- Is their cultural relevance to how people's lifestyles are?

- Make People Responsible

- Involve local communities like for example, an apartment locality

- Make people take initiative, rather than wait for govt. policy

**6.6.c Invest in Making Connections**

Alone, it will be difficult to make a change. So, it’s important to form the right partnerships. Develop plans with corporations to help out as part of their CSR. Create accountability by keeping all transactions open and public.

**7. LIMITATIONS**

The overall project with 3 objectives will result good. There still few regions where we can work with more domain knowledge. The clustering and modelling part is gone well. The data will not perform the best with new data due to R2 scores and RMSE we have got.

The last part of identifying 3 cities from 26 can be done very smoothly with more knowledge about air quality and some measures to quantify it. Then the part to suggest where to invest was all about domain.

There are so much more we can do with this data. As per as our level understanding and access to official resources we used our overall potential to use and present it in the project.

**8. CLOSING REFLECTIONS**

The one line which comes in mind is, ‘Everything is Data and Data is Everything`. The takeaways from this project are many:

1. Finding the data for project only with domain name was so much exploratory. It gave us the way to think how many pillars can be there on which this particular domain standing.
2. Writing and creating problem statement by just studying the data was too good. By this, we were able to think about the approach before even framing the problem statement.
3. In this particular project the main challenge was that this data was not about any specific company organization. It was about a country, where we have no direction to think upon or put our finger upon. Thinking from all the possible aspects and achieving the ultimate aim was a great journey.

References:

1. Srinivas, J., and Purushotham, A.V., 2013, Determination of Air Quality Index Status in Industrial areas of Visakhapatnam, India, Research Journal of Engineering Sciences, Vol. 2(6), 13-24.
2. Choking cities. 04 March 2016 Available from <http://www.scroll.in/article/801365/not-just-delhi-these-indian-citieshave-an-air-pollution-problem-worse-than-beijing>
3. Dales R, Burnett RT, Smith-Doiron M, Stieb DM, Brook JR (2004) Air pollution and sudden infant death syndrome. PEDIATRICS 113: e628–e631. <https://doi.org/10.1542/peds.113.6.e628>
4. Darryl DM (2017) <https://indiaclimatedialogue.net/2017/11/24/indiabiggest-emitter-so2/>
5. Limaye S, Salvi S (2010) Ambient air pollution and the lungs: what do clinicians need to know? Breathe 6:234–244
6. Mukesh Sharma, 2003, Interpretation of air quality data using an air quality index for the city of Kanpur, India, Journal of Environmental Engineering and Science, Vol 2: 453–462
7. Balakrishnan K, Cohen A, Smith KR (2014) Addressing the burden of disease attributable to air pollution in India: the need to integrate across household and ambient air pollution exposures. Environ Health Perspect 122:A6–A7. <https://doi.org/10.1289/ehp.1307822>
8. Pant P, Lal RM, Guttikunda SK, Russell AG, Nagpure AS, Ramaswami A, Peltier RE (2018) Monitoring particulate matter in India: recent trends and future outlook. Air Qual Atmos Health 12:45–58. <https://doi.org/10.1007/s11869-018-0629-6>
9. CPCB (2016) Annual Report 2015–16. <http://cpcb.nic.in/annual-report.php>
10. Kim K-H, Kabir E, Kabir S (2015) A review on the human health impact of airborne particulate matter. Environ Int 74:136–143. <https://doi.org/10.1016/j.envint.2014.10.005>
11. <https://www.kaggle.com/romandovega/chaieda-air-quality-in-india-eda-using-tableau>
12. <https://www.kaggle.com/parulpandey/breathe-india-covid-19-effect-on-pollution>
13. <https://www.kaggle.com/rohanrao/calculating-aqi-air-quality-index-tutorial>
14. Air Pollutants <https://en.wikipedia.org/wiki/Air_pollution#Pollutants>
15. UNEP Monitoring Air Quality <https://www.unep.org/explore-topics/air/what-we-do/monitoring-air-quality>
16. UNEP's Air Programme <https://www.unep.org/explore-topics/air>
17. UNEP's 2016 report on "Actions on Air Quality" <https://www.unenvironment.org/resources/assessment/actions-air-quality?_ga=2.209423066.934997296.1598950157-760778986.1598950157>
18. UNEP's Air Programme <https://www.unep.org/explore-topics/air>
19. Ministry of Environment, Forestry and Climate Change - Govt. of India <http://moef.gov.in/environment/pollution/>