# Analysis of AQI for Major cities in India

Pranesh Manoharan, Praveen Mohan

December 2020

# Contents

# 1 Abstract

This Paper presents the Analysis of Air Quality Index (AQI)[2] for Major cities in India using statistical Machine learning[7] techniques along with geographic, demographic, and industrial factors that influence the air quality index which can be used to predict the cities that require additional healthcare facilities and safe cities for residential communities. The Air Quality in India dataset by Rahul Rao was used as the primary dataset for developing models in R programming. This paper also investigates how well the predictive models are performing in predicting the Air Quality Index (AQI) given some input data, based on the pollution, population, and thermal power production in mega units per region.

# 2 Introduction

Air pollution[20] is the result of human activities such as mining, construction, industrial emissions[13], etc. But some natural calamities such as wildfires and volcanic explosions also act as a source of air pollution. The world health organization (WHO) states that Air pollution[20] kills an estimated seven million people worldwide every year. When we look at the list of highly polluted cities in the world most of the cities are from India. Since air pollution has been causing many adverse effects on our health it should be predicted using some methods so that preventive measures can be taken in advance. One of the ways to combat air pollution is to identify its source, contributing factors, and its origin. Usually, it is monitored by the respective state government's environment ministry. But these government agencies focus only on the chemical factors and tend to miss out on the demographic and other economic factors that might have an impact on the air quality levels in every city. Advancements in Machine learning[7] and statistical modeling techniques of predicting an outcome using a data set have led to some great discoveries in the past. In this project, we are going to implement some of the statistical modeling methods to predict the importance of other factors influencing the air quality index levels. This model can also answer the question of the safest and unsafe cities to live in in 2021. The data "Air Quality Data in India" by Rahul Rao which acts as the base for this project was extracted from the website Kaggle.org. The data consisted of the individual chemical levels and AQI value from the year 2015 till 2020. But for our project, we made use of the data from the year 2017 – 2019. Then the demographic data for major cities in India was obtained from the website Macrotrends.net. The "Power generation data" was extracted from Kagglewhich contained the Thermal power generation data which is considered to be one of the key contributing factors for the Air quality index level. The data was categorized based on the geographical regions. These predictors were used to build our final model.

# 3 Literature Review

Some of the previous studies done on this topic include the work done by soubhik Mahanta where they analyzed the usefulness of Regression models[4] in predicting the Air Quality index using the past data. The project did not consider any other predictors while building the model. Then the research done by Timothy used some of the advanced Machine Learning[7] techniques for the classification of Air Quality levels. This paper focuses on the alternative methods of predicting Air quality levels with the help of real-time data from various gas sensors. The work done by Kingsy Grace focuses mainly on the classification of individual pollutants[6] according to each AQI bucket. Their work used two clustering techniques and compared their accuracies. The paper by Venkat Rao Pasupuleti was also interesting. They had mainly focused on predicting the various pollutants influencing the Air Quality index[3] using meteorological information. We are going to predict the AQI similarly with the help of some other external factors which contribute significantly to the Air Quality Index. This is the research gap that we found by going through this paper. The idea to build this model was inspired by the work done by Usha Mahalingam. This paper focused on building a machine learning[7] model[5] using advanced techniques to determine the smart cities in India. Similarly, our model can be used to determine the highly polluted and Less polluted cities. This prediction[8] can be used to introduce specialized healthcare facilities into the worse areas and to determine the best cities to construct residential communities

## 3.1 Studies performed in Prediction

The work done by Soubhik Mahanta[4] involved the usage of supervised learning methods and advanced clustering techniques to predict the response. The regression plots for the various models were plotted. A comparison was made between the models with the help of RMSE values and the accuracy score to predict the best performing model. The model was also used to identify the important features[16] that influenced the response. In the research performed by Sankarganesh with the help of regression analysis techniques like MLR[15] (Multi Linear regression) and SVM (support vector machine)[5] along with gradient descent to predict[8] the AQI. The models were trained with batch gradient descent algorithms. To evaluate the best performing model various statistical measures were used. The measures used to evaluate the best performance are Mean absolute error (MAE), Mean absolute percent error (MAPE), correlation coefficient R and the root mean square error (RMSE). By using various gradient descent algorithms the model was trained in the training set. In the multiple regression model[15] (MLR) with batch gradient descent or steepest gradient descent, the model was trained with 500 epochs to determine the optimal parameters required for the model. In the second model, the multiple regression model (MLR)[15] with stochastic gradient descent as the training algorithm 50 epochs were used to determine the optimal parameters for the model. The final method that was implemented was the multiple regression model (MLR)[15] with mini-batch gradient descent as the training algorithm which used 500 epochs and resulted in a very low variance compared to the stochastic gradient method. The final regression was used as the support vector regression method. The performance evaluation was made using the statistical metrics and the Support Vector Machine model outperformed all the other models.

### 3.2 Studies Performed in Air Pollution factors

To consider the various other factors which contribute to the Air pollution levels we performed various literature reviews. From the paper published by Matthew Cole[10], we can see that there exists a relationship between the population, demographic factors, and pollution. In this research, they had also considered the urbanization rate to be one of the important factors influencing air pollution levels. They also found that SO2 emissions followed a U-shaped relationship with population - emissions elasticity rising at higher population levels. SO2 emissions were not dependent on the urbanization rate and the household size. From the paper published by S K Goyal[21], She had studied the various sources of pollution in the urban areas. This paper mainly focused on the effects of vehicles on Air pollution. This research work mainly focused on one Indian megacity which is Delhi. The contribution of transportation is more than 72 percent in Delhi. They had also considered the technological changes and government policies that are in place to reduce the Air pollution levels. They also suggested implementing various management strategies to control air pollution levels. Various factors like these have been studied through the literature review and some of the factors like population, thermal energy production have been considered for our project to build the model.

## 4 Data Cleaning and transformation

The primary AQI data by Rahul Rao was collected from Kaggle. The initial columns before preprocessing the data set can be seen in the below table. The initial dataset consisted of observations from the year 2015 to 2020. The highlighted variables are used in the analysis and prediction.

### 4.1 Initial Variables: AQI Data

| Column Names | Description |
|---|---|
| City | The major City name in India |
| Date | Date of observation |
| PM2.5 | Particulate Matter (Molecule/$cm^2$) |
| PM10 | Particulate Matter (Molecule/$cm^2$) |
| NO | Nitric Oxide (Molecule/$cm^2$) |
| $NO_2$ | Nitrogen Dioxide (Molecule/$cm^2$) |

| | |
|---|---|
| $NO_x$ | Nitrogen Oxides (Molecule/$cm^2$) |
| $NH_3$ | Ammonia (Molecule/$cm^2$) |
| CO | Carbon Monoxide (Molecule/$cm^2$) |
| $SO_2$ | Sulphur Dioxide (Molecule/$cm^2$) |
| $O_3$ | Ozone or Trioxygen (Molecule/$cm^2$) |
| Benzene | (Molecule/$cm^2$) |
| Toluene | (Molecule/$cm^2$) |
| Xylene | (Molecule/$cm^2$) |
| Station Name | Observation station name |
| AQI Bucket | Severity of AQI |
| AQI | Air Quality Index (Unit measure) |
| Station Id | Each station ID |
| State | State name |
| Status | Status of the observation |
| Region | Geographical position |
| Month | Observation month |
| Year | Observation year |
| Season | Season of the observation |
| Weekday or Weekend | Whether weekday or weekend |
| Regular day or holiday | Whether a regular day or holiday |
| AQ Acceptability | Is AQI within the acceptable range |

In the AQI data, we have performed four levels of mean imputation. Initially, the missing values are imputed with mean by grouping the observations by City, Year, and Month. The next level of imputation is by grouping the observation by City, Year, and Season. The third level of imputation is by City and Year. The final level of imputation is just grouping the observations by each city. To reduce complexity we have combined PM2.5 and PM10 columns into a single Particulate Matter (PM) column and combined Benzene, Toluene, Xylene together as a single column called BTX.

## 4.2 Initial Variables: Thermal Data

The thermal data consists of the total energy produced region wise from energy sources like thermal energy, nuclear energy, etc. The below table represents the detailed description of the thermal data and its components.

| Column Names | Description |
|---|---|
| Date | Observation date |
| Region | Geographical position |
| Thermal Generation actual in MU | Energy generated in the mega unit |
| Thermal Generation estimated in MU | Estimated value |
| Nuclear Generation actual in MU | Energy generated in the mega unit |
| Nuclear Generation estimated in MU | Estimated value |

| | |
|---|---|
| Hydro Generation actual in MU | Energy generated in the mega unit |
| Hydro Generation estimated in MIU | Estimated value |

## 4.3 Initial Variables: Demographic Data

The Demographic data collected from macro trends consists of population statistics for the years 2017-2019. The below table represents the detailed description of Demographic data from the year 2017 to 2019.

| Column Names | Description |
|---|---|
| City | Name of the City |
| Region | Geographical position |
| State | Name of the state |
| Year | Year |
| Population | Population Count |

## 4.4 Data Transformations

In the AQI dataset initially, unwanted columns were removed and four levels of mean imputations were performed by grouping various segments of the data. The columns PM2.5 and PM10 were combined as a single particulate matter column called PM & the columns Benzene, Xylene, and Toluene were combined as a single column called BTX.Converted Character columns as categorical variables and filtered data for the years 2017, 2018 & 2019. Aggregated thermal data according to each region and year. Then this data was merged with the main data frame. Aggregated demographic data according to each city and year. Then this data was combined with the main data frame. Year and Month columns were extracted from the date column and converted as Categorical variables. The geographical position of each city was added as a categorical variable with five levels.

## 4.5 Final Variables: Combined Data

After the data transformations were completed all the datasets were merged into a single final Dataframe. After cleaning the final dataset consisted of 107308 rows and 18 columns. The description of the dataset is given in the below table.

| Column Names | Description | Type |
|---|---|---|
| City | Name of City | Category |
| State | Name of State | Category |
| Region | Geographic region | Category |
| Year | Year | Category |
| Month | Month | Category |
| Season | Climatic condition | Category |
| Population | Demography | Category |
| Thermal | Power Produced | Mega Unit |
| PM | Particulate Matter | Molecule/Sq.cm |
| BTX | Benzene, Toluene,Xylene | Molecule/Sq.cm |
| NO | Nitric Oxide | Molecule/Sq.cm |
| $NO_2$ | Nitrogen DiOxide | Molecule/Sq.cm |
| $No_x$ | Oxides of Nitrogen | Molecule/Sq.cm |
| $NH_3$ | Ammonia | Molecule/Sq.cm |
| CO | Carbon Monoxide | Molecule/Sq.cm |
| $SO_2$ | Sulfur Dioxide | Molecule/Sq.cm |
| $O_3$ | Ozone | Molecule/Sq.cm |
| AQI | Air quality index | Unit Measure |

# 5 Exploratory Data Analysis (EDA)

## 5.1 Air Quality Index (AQI) Trend

Figure1 represents the bar plot of the Air Quality Index (AQI) citywide. From the plot, we can infer that the cities like Ahmedabad, Patna, Delhi, Gurugram, and Lucknow are at higher risk with an average AQI of above 200. Other cities like Talcher, Bhopal, Kolkata, Jaipur, etc. are still in the risk zone with an average AQI of above 100. Cities like Shillong, Thiruvananthapuram, Coimbatore, Bengaluru have an average AQI of below 100.

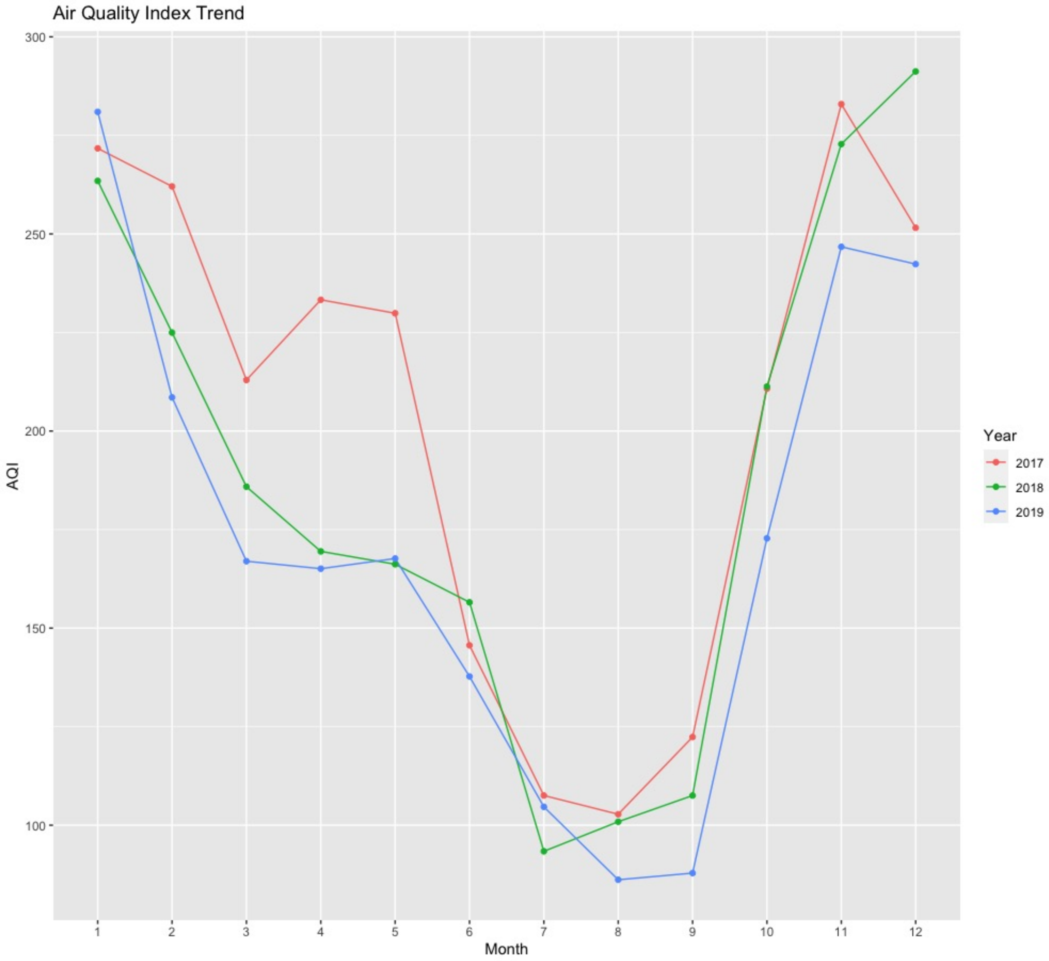


Figure 1: Air Quality Index (AQI) Trend

The below plot represents the month-wise trend of average AQI in India for the years 2017, 2018, and 2019. The trend has a unique and similar pattern for all three years. It seems like the average AQI throughout India seems to decrease rapidly after May and has a sudden increase after September. This may be due to the holiday seasons from November to January where people tend to use more public transportation and personal vehicles in traveling.

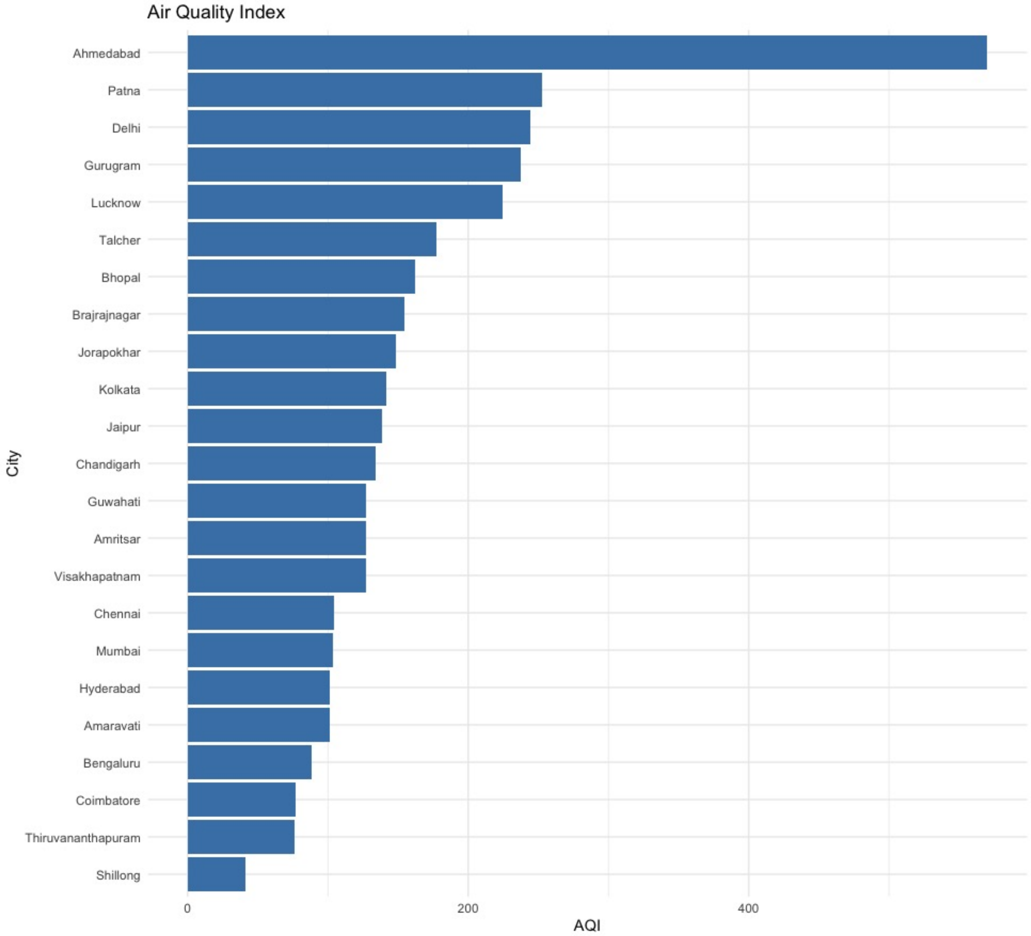


Figure 2: Citywise AQI

The violin plot represents the season-wise distribution of AQI throughout India. It seems like the AQI is in a lower and safer region during monsoon and is in a higher risk region during winter and summer. This explains that the power consumption in households and industries will be higher due to the usage of heaters and air conditioners which results in higher demand for electricity. And the majority of the electricity in India is generated from thermal power plants which is a major cause of air pollution.

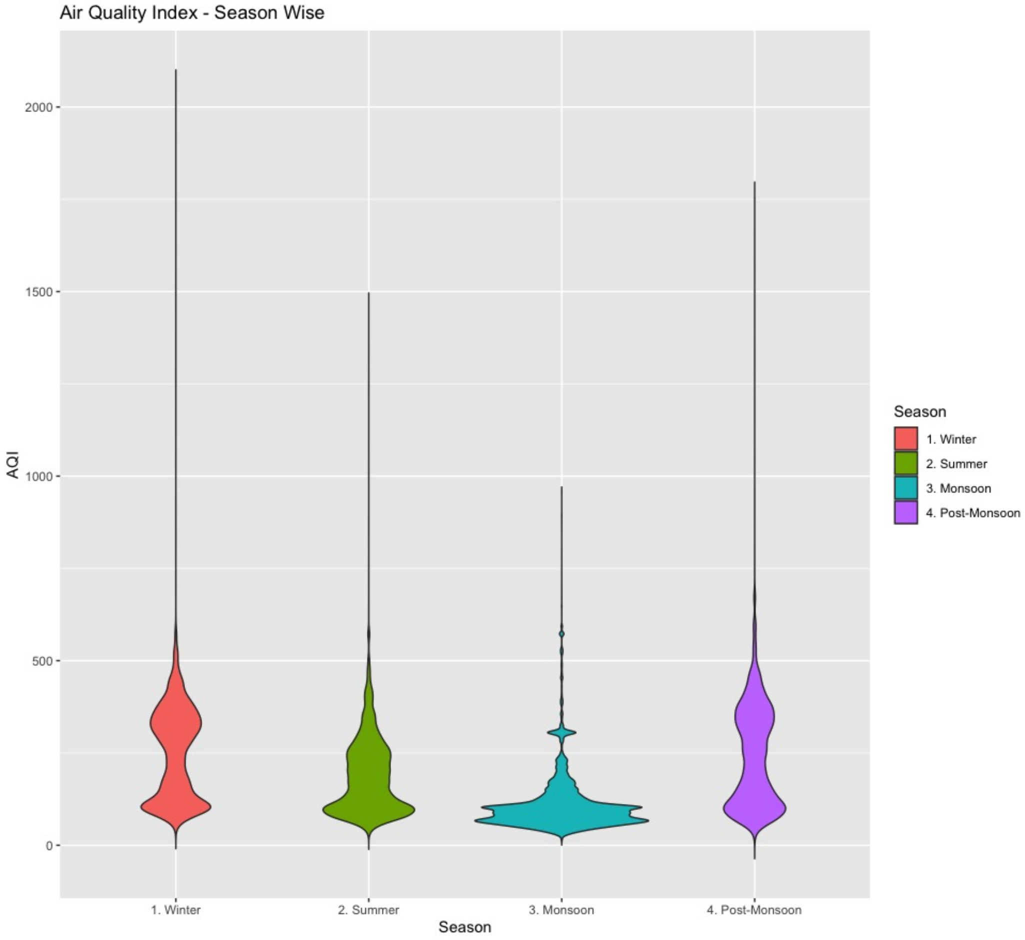


Figure 3: Violin Plot

Figure 4 represents the citywide average Thermal Power Generation in mega units. Almost all the cities are producing electricity through thermal power plants on an average of more than 400 mega units. The plot represents the region-wise and season wise heat map of AQI. The AQI is in a higher risk region in northern, northeastern, and eastern areas during winter and post-monsoon.

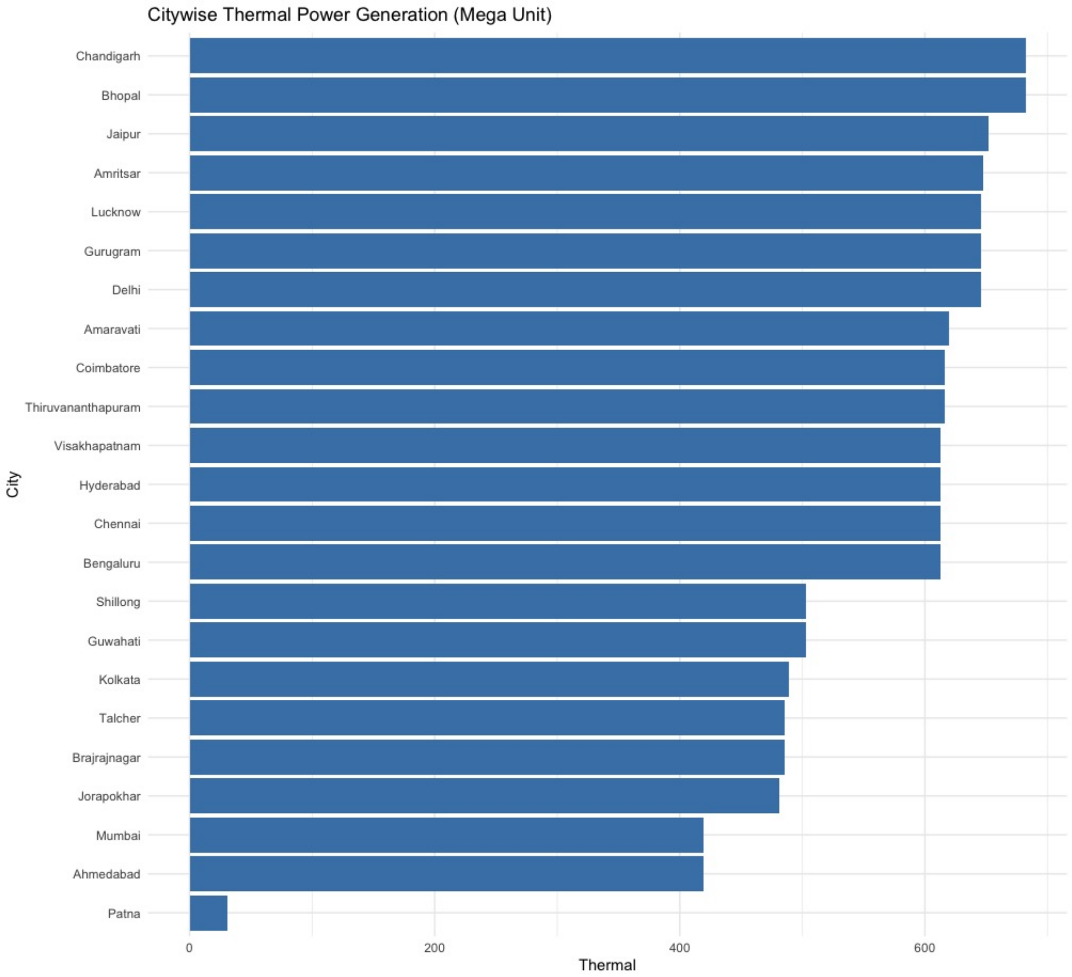


Figure 4: Thermal Power Generation

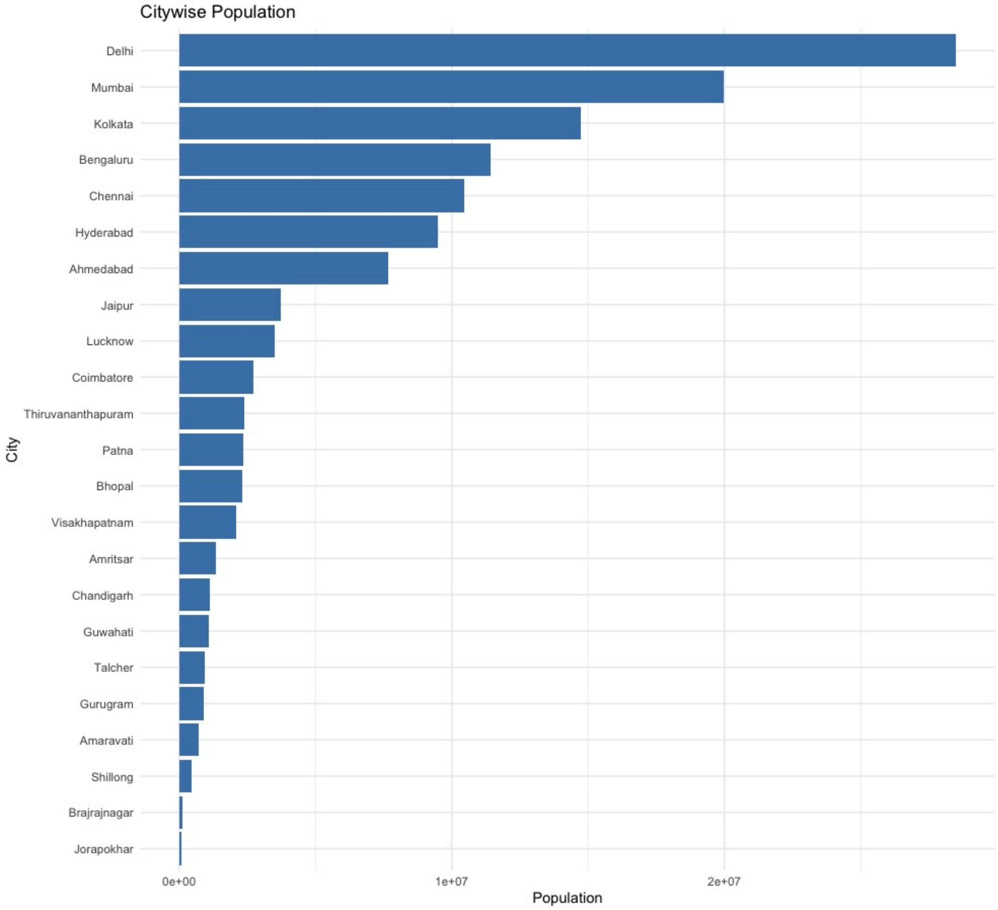


Figure 5: Citywise Population

And the overall AQI is lower during the monsoon season. The following plots are year wise and month wise AQI, Particulate Matter (PM), Carbon Monoxide (CO), BTX (Benzene, Toluene, Xylene) heat map. By comparing the four heatmaps we can understand how each pollutant has a significantly similar pattern throughout the year. This helps us to identify the relationship between each pollutant.

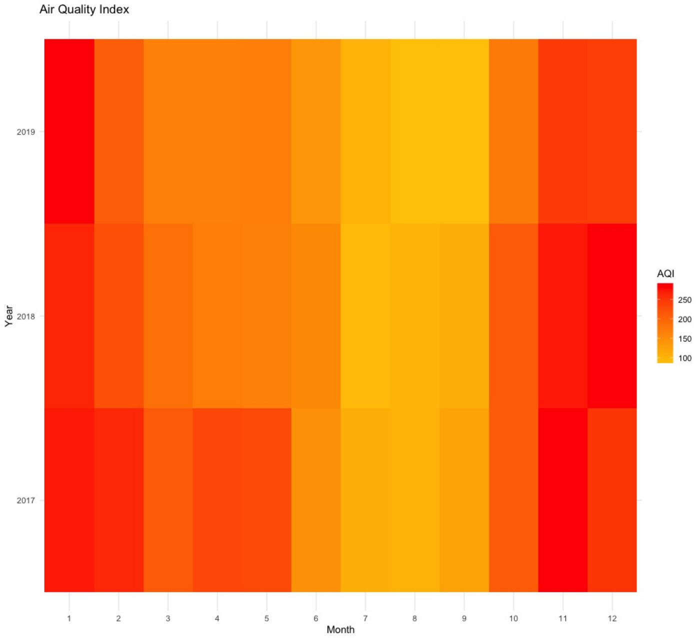


Figure 6: AQI Heatmap

The plot represents the distribution of thermal power generation for various states in India in mega units. This plot represents the distribution of thermal power generation based upon the geographical position of each city in India. The northern region is producing more electricity (MU) from thermal power plants when compared to other regions. This plot represents the year-wise distribution of the thermal power generation in mega units. We can see a gradual increase in the amount of electricity generated from the year 2017 to 2019.

The figure represents the correlation between the numerical explanatory variables in the data. We can identify that the highly correlated variables with the AQI are Particulate Matter (PM), Nitrogen Dioxide (NO2), BTX (Benzene, Toluene, Xylene), Population, etc. We can also infer some interesting facts by looking at the interactions between predictor variables. Thermal power generation and Ammonia (NH3) are highly correlated with each other, this may be because of the harmful emissions from the thermal power plants. Similarly, there is a significant correlation between population and other pollutants like

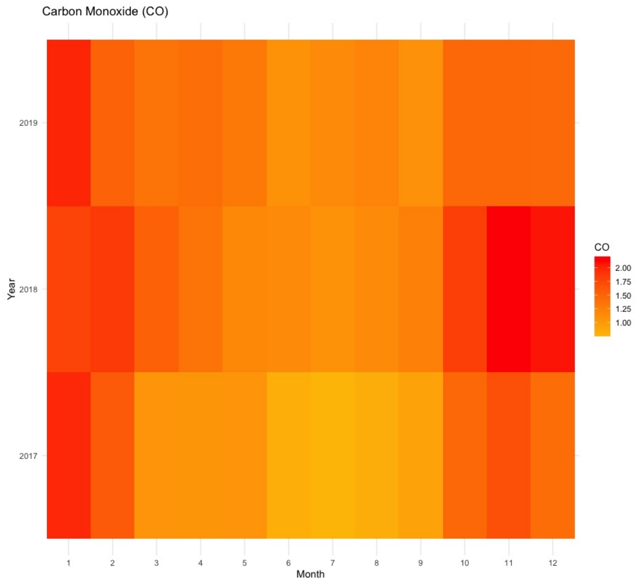


Figure 7: CO Heatmap

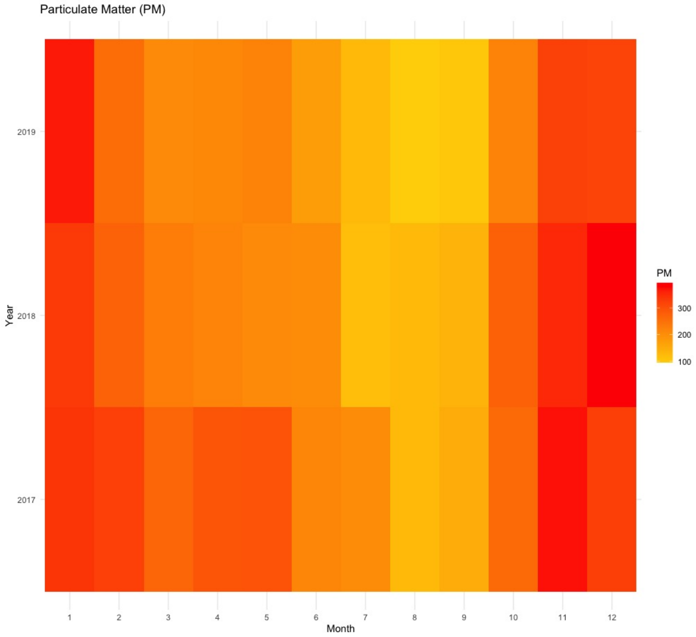


Figure 8: PM Heatmap

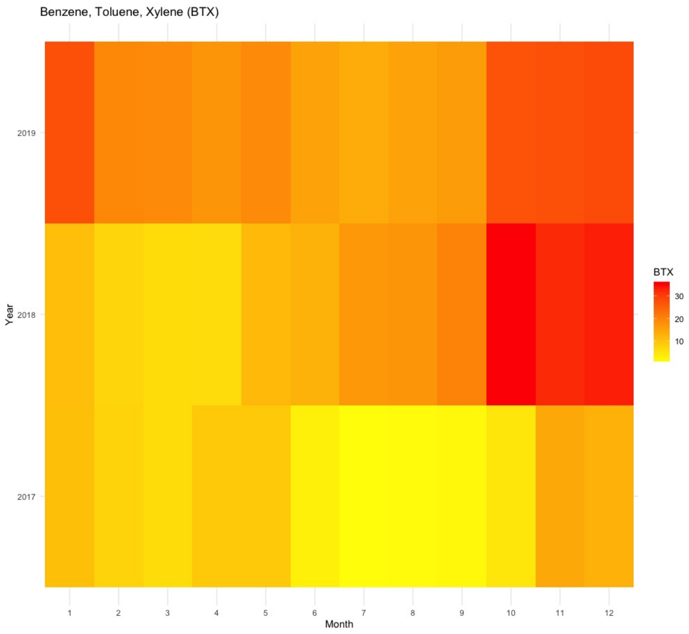


Figure 9: BTX Heatmap

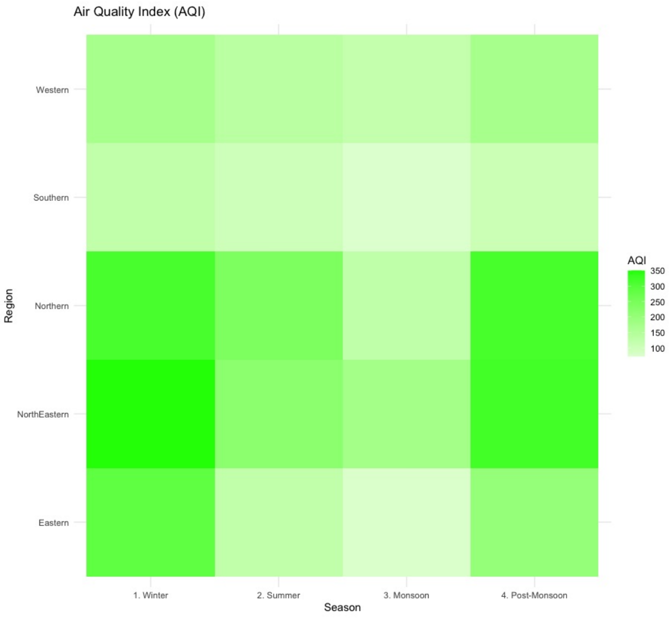


Figure 10: Region wise AQI heatmap

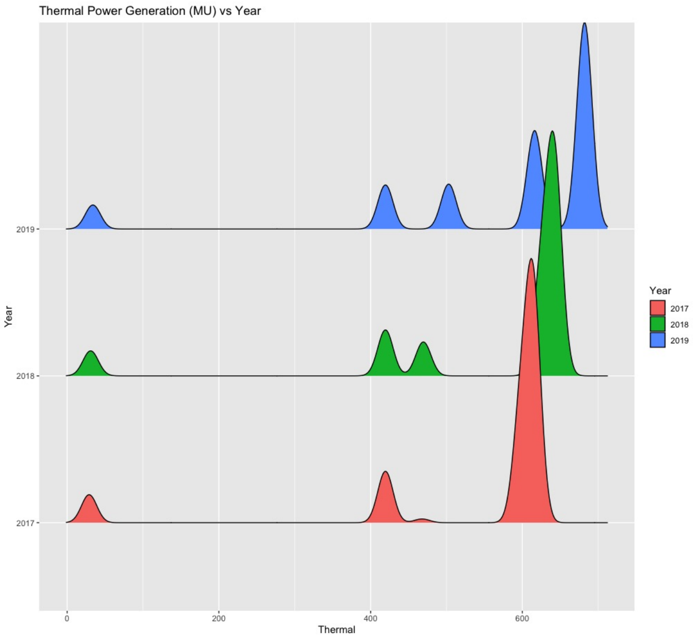


Figure 11: Thermal Power Generation - Year

Nitric Oxide(NO), Nitrogen Dioxide (NO2), etc.

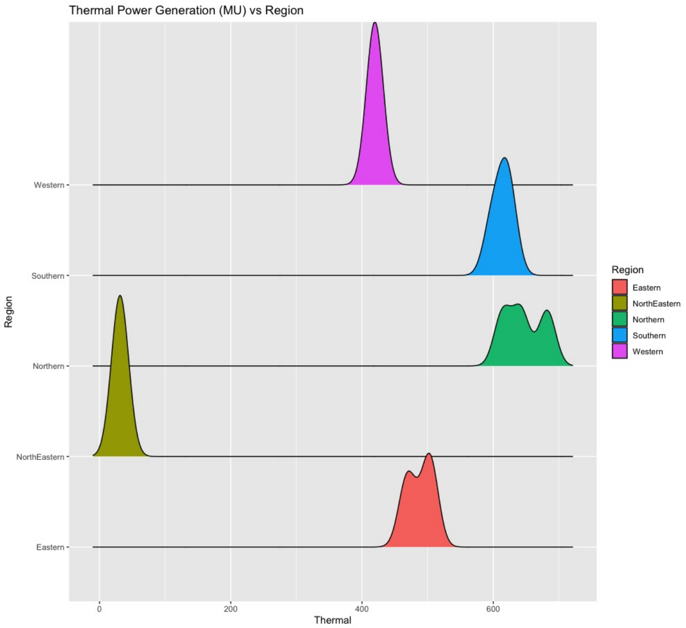


Figure 12: Thermal Power Generation - Region

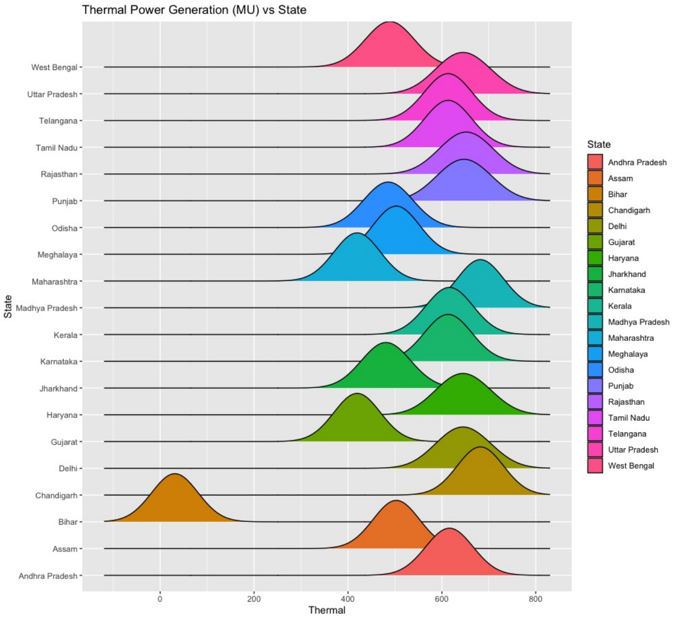


Figure 13: Thermal Power Generation - State

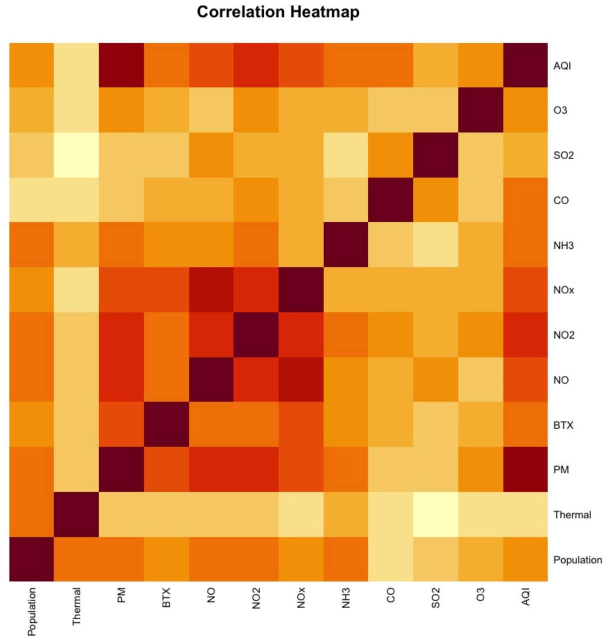


Figure 14: Correlation

# 6 Modeling Framework

The modeling framework helps us define the skeleton or the major structure of this project which is a flow in which the data is resampled and trained to various models and then tested with suitable evaluation metrics to estimate the performance of each model. The below diagram is a brief representation of the framework we have used in this project.

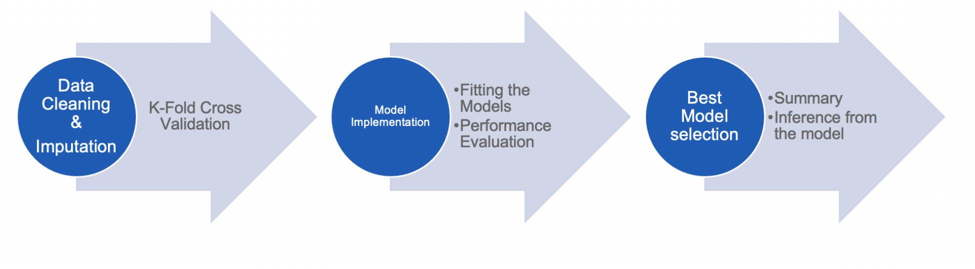


Figure 15: Model Framework

## 6.1 K-Fold Cross-Validation

Resampling methods are simple mathematics that is used while sampling from a sample or a population to estimate the accuracy or precision of a statistical model. These methods help us to use the available data more efficiently and economically making sure that almost all the anomalies present in the data are fed to the model while estimating its performance. Various resampling methods are available to use and some of the most common methods are given below.

1. Leave p-out cross-validation
2. Leave one out cross-validation
3. Holdout cross-validation
4. Repeated random subsampling validation
5. k-fold cross-validation
6. Stratified k-fold cross-validation
7. Time Series cross-validation
8. Nested cross-validation

In this project we have utilized the standard 10 - Fold Cross Validation to estimate our model performance. The below figure is a pictorial representation of how k-fold cross-validation works. The data is initially split into K equal subsets. And K different sets of training and test data are fed to the model for K number of times until all the K subsets have fed to the model as test data at most once. For each of the K iterations, the model performance is measured and finally, the performance results of all the folds are averaged and final metrics are generated. In this method, almost every observation is fed to the model as training data which helps us to estimate the performance of the model in various subsets of training and testing data.

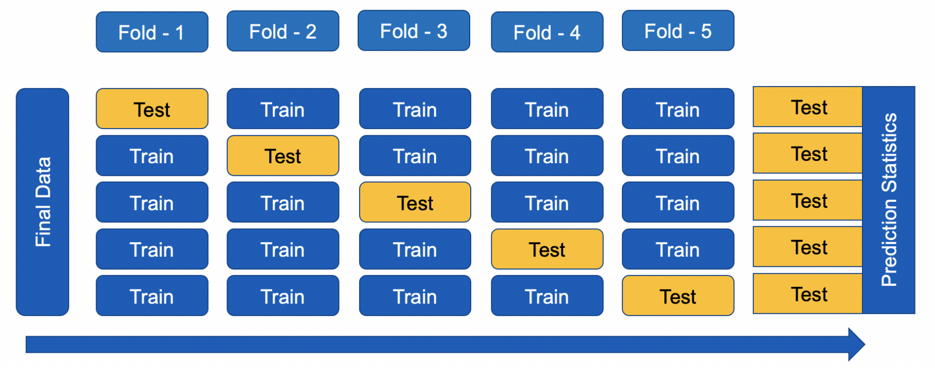

Fold - 1
Fold - 2
Fold - 3
Fold - 4
Fold - 5
Final Data
Test Train Train Train Train
Train Test Train Train Train
Train Train Test Train Train
Train Train Train Test Train
Train Train Train Train Test
Test
Test
Test
Test
Test
Prediction Statistics

Figure 16: K-fold cross validation

## 6.2 Modeling Methods

The response variable which we are trying to predict is the Air Quality Index (AQI) which is a continuous variable and hence we are using Regression Models.

### 6.2.1 Linear Regression

Linear regression is a statistical model that predicts the response variable by a linear combination of the corresponding predictor variables. It predicts the response by fitting a straight line in case of two-dimensional data or a plane in case of three-dimensional data to the response variable by the Ordinary Least Square (OLS) method. The goal is to fit a line or a plane that has the least sum of squared error concerning the response variable. Below is a basic formulation of the Linear Regression model.

$$Y_t = \beta_0 + \beta X_t + \epsilon_t$$

### 6.2.2 Stepwise Regression

Stepwise regression is an extension of the Linear Regression model in which the predictors are chosen automatically based upon their significance using various tests like F-test, T-test, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), etc. The model automatically decides whether to include an explanatory variable based upon the above test. This model is based on three approaches. The first approach is called Forward Selection in which initially no variable is included in the model and starts testing the model by adding each predictor using a chosen model fit criterion. It includes predictors that have the most significant improvement in the fit. This process is repeated until no other predictors have a significant impact on improving the model fit. Another approach is called Backward Elimination in which all the predictors are included in the model and eliminating each predictor step by step which is insignificant toward the model fit. This step is repeated until no other predictors can be eliminated without any insignificant loss of the model fit. The third approach is called Bidirectional Elimination which is a combination of both Forward Selection and Backward Elimination.

### 6.2.3 Decision Tree

Decision Tree [17] is one of the popular models in which a tree-shaped structure is generated from top to bottom by training the model. The tree consists of a Root node, Internal Nodes, and Leaf nodes. These nodes are a representation of each predictor variable in the data. This method uses metrics like Gini Index, Gini Impurity, and Gain in creating the tree. Initially, the root node is selected among the predictor variables which has the lowest Gini Impurity value which means that the particular node has the best capability of segmenting the response variable. Then the next internal node is selected by choosing the predictors which have the highest Gain value which means that the particular node has the most information gain from the previous node. This step is repeated until there are no more predictors left to split and leaf nodes are generated.

$$Gini\,Index = 1 - (Probability\,of\,Yes)^2 - (Probability\,of\,No)^2$$

*Gini Impurity = (Weight of left node * Gini Index of the left node) + (Weight of right node * Gini Index of the right node)*

### 6.2.4 LASSO & RIDGE Regression

LASSO (Least Absolute Shrinkage and Selection Operator)[15] and RIDGE are a form of regression model which automatically performs the best variable selection and regularization technique to increase the model accuracy and interpretability. Regularization is a method to avoid overfitting which results in improved model performance by reducing its complexity. LASSO and RIDGE are often referred to as L1 and L2 Regularization respectively.

### 6.2.5 MARS (Multivariate Adaptive Regression Splines)

The MARS model[17] is a non-parametric regression which is an extension of a linear regression that creates a piecewise linear model thus providing an efficient and convenient approach in capturing the nonlinearity and interactions present in the data. The MARS model is developed in two phases, the forward pass and the backward pass. In a forward pass, the algorithm initially starts with a model consisting of just the intercept term and then repeatedly includes the basic function in pairs to the model. During each iteration, the pair of basis functions which produces the maximum reduction in the residual error is chosen. After each iteration, each of the basis functions is multiplied by a hinge function which is defined by a variable and a knot. A knot is an observation in each of the predictors.

$$\widehat{f}(x) = \sum_{i=1}^{k} c_i B_i(x).$$

### 6.2.6 GAM (Generalized Additive Model)

The GAM model is a generalized linear model in which the relation between the response variable and the predictors variable may be linear, non-linear, or a combination of linear and various smooth functions of the predictors. The GAM model is a semi-parametric model that identifies the components for which the smoothing functions like smoothing splines and regression smoother needs to be incorporated by an algorithm called Backfitting.

## 6.3 Evaluation Metrics

Evaluation metrics play a vital role in evaluating the performance, accuracy, and fit of a statistical model. Since our research involves regression problems, we used metrics like RMSE (Root Mean Squared Error), R2, and MAE (Mean Absolute Error).

### 6.3.1 Root Mean Squared Error (RMSE)

The RMSE is a measure of variability between the predicted response by a statistical model and the actual observation. RMSE is a non-negative measure in which a value closer to zero represents a good fit to the data. Higher RMSE represents that there is a higher variability between the predicted response and the actual value which means the model has a poor fit.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

### 6.3.2 $R^2$ (R - Squared)

The $R^2$ also referred to as the coefficient of determination, is the amount of variability explained by a statistical model on the response variable. It is the ratio between the explained variability and the total variability.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

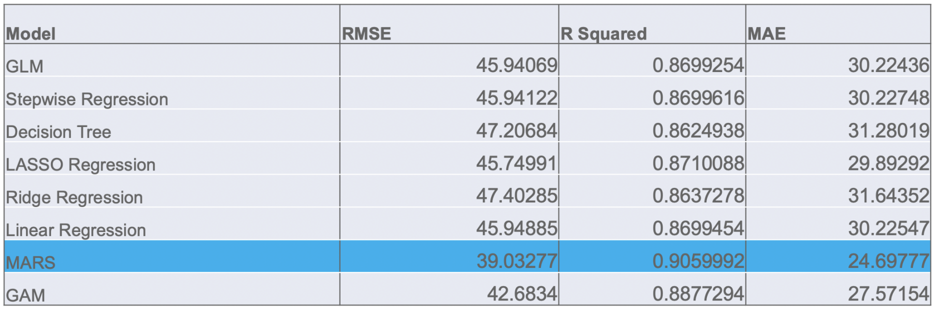
### 6.3.3 Mean Absolute Error (MAE)

The MAE is the measure of error between the predicted response and the actual observation.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

## 6.4 Model Results

The results after performing 10-fold cross-validation for the test data in each model are shown in the below table.



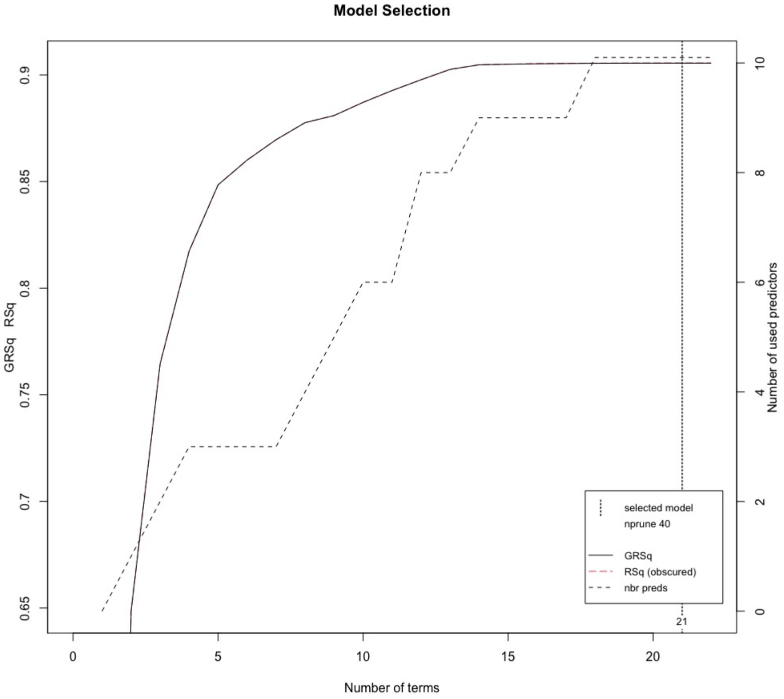
| Model | RMSE | R Squared | MAE |
|---|---|---|---|
| GLM | 45.94069 | 0.8699254 | 30.22436 |
| Stepwise Regression | 45.94122 | 0.8699616 | 30.22748 |
| Decision Tree | 47.20684 | 0.8624938 | 31.28019 |
| LASSO Regression | 45.74991 | 0.8710088 | 29.89292 |
| Ridge Regression | 47.40285 | 0.8637278 | 31.64352 |
| Linear Regression | 45.94885 | 0.8699454 | 30.22547 |
| MARS | 39.03277 | 0.9059992 | 24.69777 |
| GAM | 42.6834 | 0.8877294 | 27.57154 |

Figure 17: Severity Count

## 6.5 MARS (Multivariate Adaptive Regression Splines)

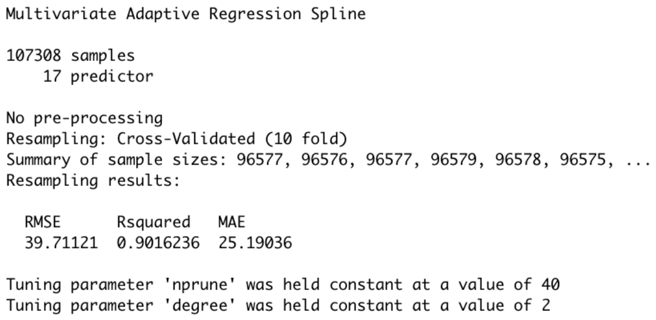
The model selection plot tells us at which the number of terms the final model was selected based upon the GR-Squared and R-Squared values on the y-axis. As the number of predictors increases the variability explained by the model increases. The parameter and coefficients table helps us identify which predictors and the interactions between them were used in the final model and also their corresponding coefficient value.

After performing 10-fold cross-validation the following results were generated with an RMSE of 39.08, an R-squared of 90.6%, and an MAE of 24.56. The below figure represents the final model from the cross-validation. The variable importance plot was obtained from the final MARS model from which we can infer the predictor variables which had a significant impact in predicting the AQI. The Generalized Cross-Validation (GCV) plot and the Residual Sum of Squares (RSS) plot will mostly produce similar variable importance plots. From these plots, it is evident that the extracted features[16] like Population, Thermal Power Generation (MU) had a significant impact in predicting the response. And also other predictors like Particulate Matter (PM), Carbon Monoxide (CO), Season, City, etc. had a significant impact as well.



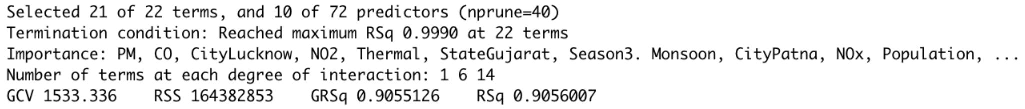
```
|parameter                                   |  coef  |
|:-------------------------------------------|:------:|
|(Intercept)                                 | 367.02 |
|h(PM-467.03)                                |  0.41  |
|h(467.03-PM)                                | -0.67  |
|h(CO-8.72)                                  | 12.56  |
|h(8.72-CO)                                  | -4.16  |
|CityLucknow                                 | 193.73 |
|CityLucknow*h(PM-145.805)                   |  0.05  |
|CityLucknow*h(145.805-PM)                   | -1.48  |
|h(NO2-95.73)*h(CO-8.72)                     | -0.01  |
|h(95.73-NO2)*h(CO-8.72)                     | -0.16  |
|h(Thermal-419.55)*h(8.72-CO)                |  0.00  |
|h(419.55-Thermal)*h(8.72-CO)                |  0.05  |
|h(419.55-Thermal)*h(467.03-PM)              |  0.10  |
|CityPatna*h(467.03-PM)                      | -37.70 |
|Season3. Monsoon*h(8.72-CO)                 | -9.38  |
|Season3. Monsoon*h(467.03-PM)               |  0.19  |
|StateGujarat                                | 136.31 |
|h(NOx-34.16)*h(8.72-CO)                     | -0.01  |
|h(34.16-NOx)*h(8.72-CO)                     | -0.08  |
|h(Population-1.1883e+07)*h(467.03-PM)       |  0.00  |
|h(1.1883e+07-Population)*h(467.03-PM)       |  0.00  |
```

Figure 18: Parameters & Co-efficient



```
Multivariate Adaptive Regression Spline

107308 samples
    17 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 96577, 96576, 96577, 96579, 96578, 96575, ...
Resampling results:

  RMSE      Rsquared   MAE
  39.71121  0.9016236  25.19036

Tuning parameter 'nprune' was held constant at a value of 40
Tuning parameter 'degree' was held constant at a value of 2
```

Figure 19: K-Fold Results



```
Selected 21 of 22 terms, and 10 of 72 predictors (nprune=40)
Termination condition: Reached maximum RSq 0.9990 at 22 terms
Importance: PM, CO, CityLucknow, NO2, Thermal, StateGujarat, Season3. Monsoon, CityPatna, NOx, Population, ...
Number of terms at each degree of interaction: 1 6 14
GCV 1533.336    RSS 164382853    GRSq 0.9055126    RSq 0.9056007
```

Figure 20: Final Model

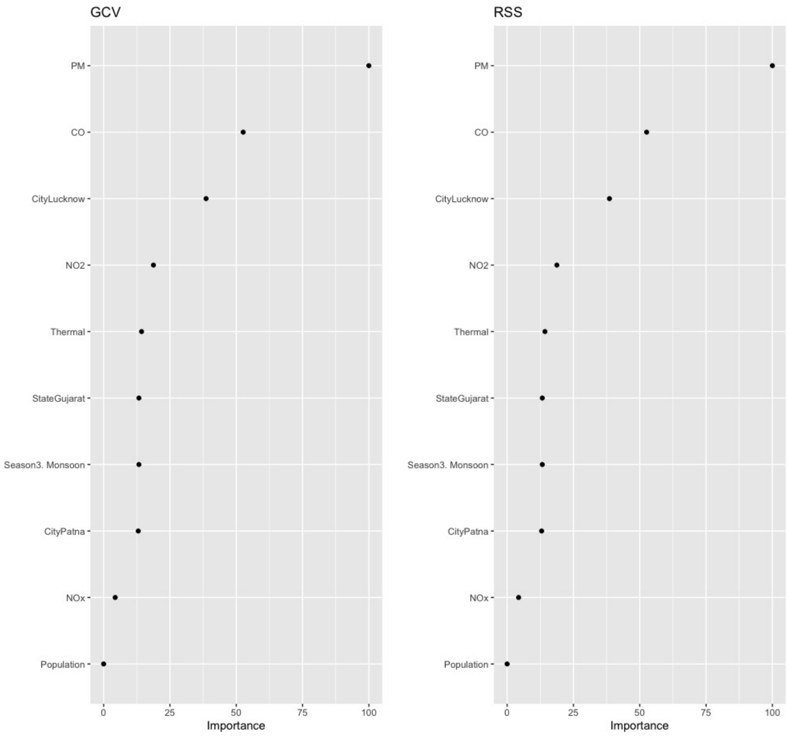


Figure 21: Variable Importance Plot

# 7 Conclusion

The goal of this research is to predict the Air Quality Index (AQI) of major cities in India based upon the Demographic, Geographical, Economical, and Weather characteristics of each city. After building eight models with 10-fold cross-validation, the MARS model obtained better results based on the evaluation metrics. The variable importance plot from the final model helped us how significantly the extracted features[16] had an impact in predicting the AQI. Also from the exploratory analysis, we have identified some interesting facts about how each predictor was influencing one another, and unique patterns in the data were also captured. With 107308 observations and 17 predictors and for 10-fold cross-validation, the MARS model obtained a test error of 39.03, R-Squared of 90.6%, and MAE of 24.6. This extensive analysis will help us make informed decisions in developing healthcare facilities focusing mainly on respiratory diseases based on the severity of the air pollution levels. Based on the predictions from our model we can take preventive measures in the future to combat the increasing emission levels from each city.

# 8 Future Works

The model was trained along with the available factors like population, Thermal production. We would like to include some of the features in the future to make a better prediction. The factors like Coal consumption data, Industrial emissions[13] data, Electric vehicle sales data, Fuel consumption data, Vehicle sales data, the Mortality rate due to respiratory diseases. These factors help us in a better prediction of AQI. The coal consumption data can be collected from the import data in major cities. There should be a proper repository of the industrial emissions[13] in India so that it can be used for building our model. When conventional vehicles are replaced by electric vehicles it has a significant impact on the Air Quality levels. By implementing these factors in our model it gives us a better prediction of the Air Quality levels.

# 9 References

1. *K. Nandini and G. Fathima, "Urban Air Quality Analysis and Prediction Using Machine Learning," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2019*

2. *P. Pullan, C. Gautam, and V. Niranjan, "Air Quality Management System," 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2020*

3. V. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth, and H. K. Reddy, "Air Quality Prediction Of Data Log By Machine Learning," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020

4. *S. Mahanta, T. Ramakrishnudu, R. R. Jha and N. Tailor, "Urban Air Quality Prediction Using Regression Analysis," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019*

5. T. M. Amado and J. C. Dela Cruz, "Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization," TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018

6. *R. K. Grace, K. Aishvarya S., B. Monisha, and A. Kaarthik, "Analysis and Visualization of Air Quality Using Real-Time Pollutant Data," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 34-38*

7. *U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha and G. Kedam, "A Machine Learning Model for Air Quality Prediction for Smart Cities," 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India, 2019*

8. *S. S. Ganesh, S. H. Modali, S. R. Palreddy, and P. Arulmozhivarman, "Forecasting air quality index using regression models: A case study on Delhi and Houston," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017*

9. *W. Zhenghua and T. Zhihui, "Prediction of Air Quality Index Based on Improved Neural Network," 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC), Dalian, 2017*

10. *Cole, M.A., Neumayer, E. Examining the Impact of Demographic Factors on Air Pollution. Population and Environment 26, 5–21 (2004)*

11. *M. Korunoski, B. R. Stojkoska, and K. Trivodaliev, "Internet of Things Solution for Intelligent Air Pollution Prediction and Visualization," IEEE EUROCON 2019 -18th International Conference on Smart Technologies, Novi Sad, Serbia, 2019*

12. *H. ALTINÇÖP and A. B. OKTAY, "Air Pollution Forecasting with Random Forest Time Series Analysis," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018*

13. *C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2019*

14. *Meng-i Liao and Hwong-wen Ma, "CO2 emission reduction through crucial industrial by-product exchange in an industrial area," 2011 Second International Conference on Mechanic Automation and Control Engineering, Hohhot, 2011*

15. *H. Hirose, Y. Soejima and K. Hirose, "NNRMLR: A Combined Method of Nearest Neighbor Regression and Multiple Linear Regression," 2012 IIAI International Conference on Advanced Applied Informatics, Fukuoka, 2012*

16. *R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, 2016*

17. *W. Zhou, T. Wang, J. Shi, B. Peng, R. Zhao, and Y. Yu, "Remotely Sensed Clear-Sky Surface Longwave Downward Radiation by Using Multivariate Adaptive Regression Splines Method," IGARSS 2018 - 2018*

18. *S. Patil and U. Kulkarni, "Accuracy Prediction for Distributed Decision Tree using Machine Learning approach," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019*

19. *M. Etkind, "Air pollution-an overview," IEE Colloquium on Pollution of Land, Sea, and Air: An Overview for Engineers, London, UK, 1995*

20. *P. Gupta, R. Kumar, S. P. Singh, and A. Jangid, "A study on monitoring of air quality and modeling of pollution control," 2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Agra, 2016*

21. *Goyal SK, Ghatge SV, Nema P, M Tamhane S. Understanding urban vehicular pollution problem vis-a-vis ambient air quality case study of a megacity (Delhi, India). Environ Monit Assess. 2006 Aug*