ANALYSIS OF AQI FOR MAJOR CITIES IN INDIA

Group 11

Pranesh Manoharan – 50321223

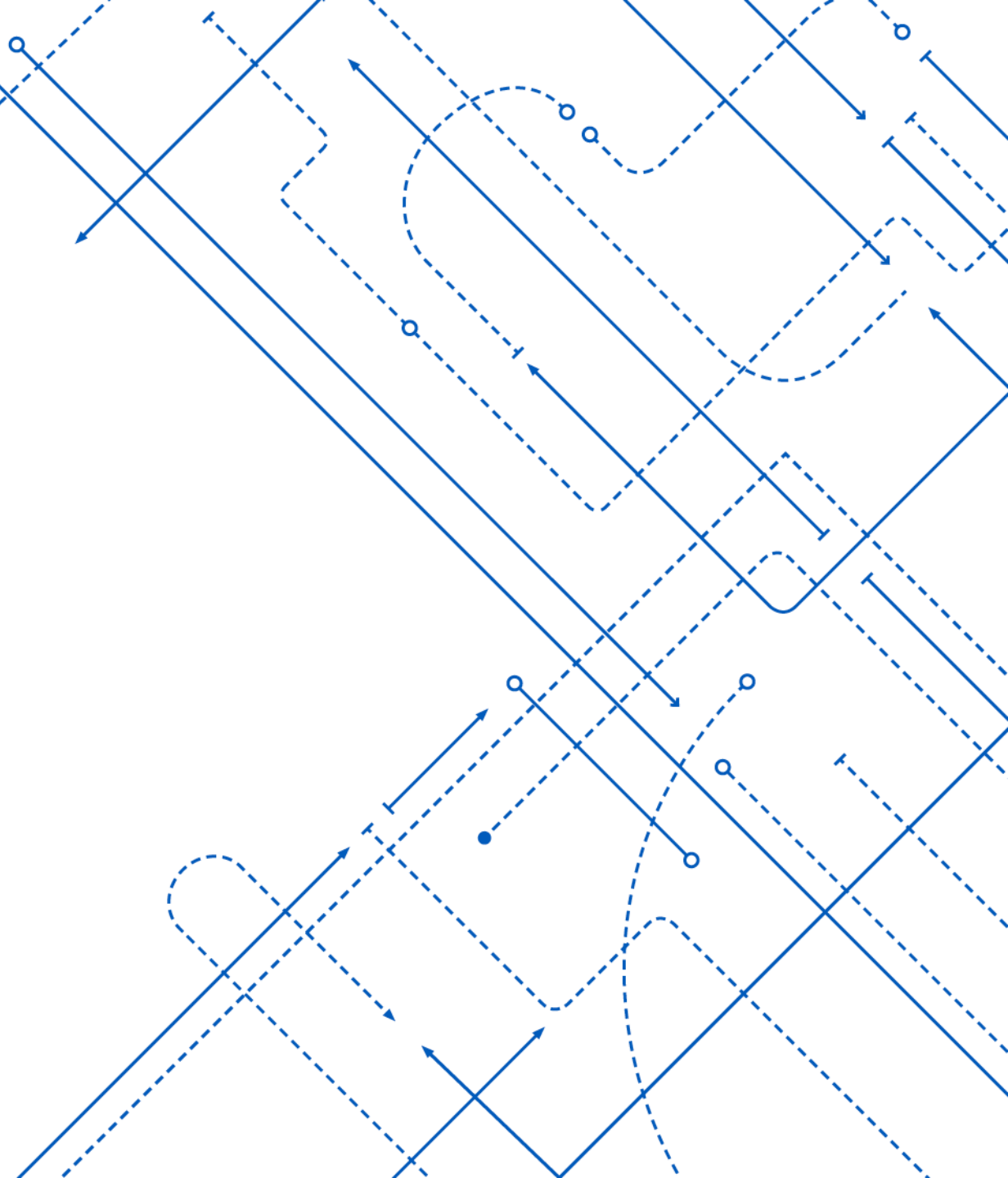
Praveen Mohan – 50321225

University at Buffalo The State University of New York

University at Buffalo The State University of New York

CONTENT

- Motivation & Problem statement
- Literature review & gaps
- Data Collection and Cleaning
- Data Visualization
- Additional Predictors
- Methodology
- Model Framework & Evaluation
- Conclusion
- Future work and hurdles
- References
- Data Source



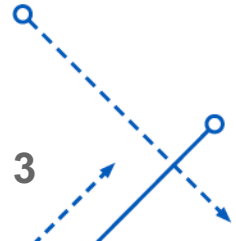
Why Air Pollution ?

Air pollution is a major problem of recent decades, which has a serious toxicological impact on human health and the environment.

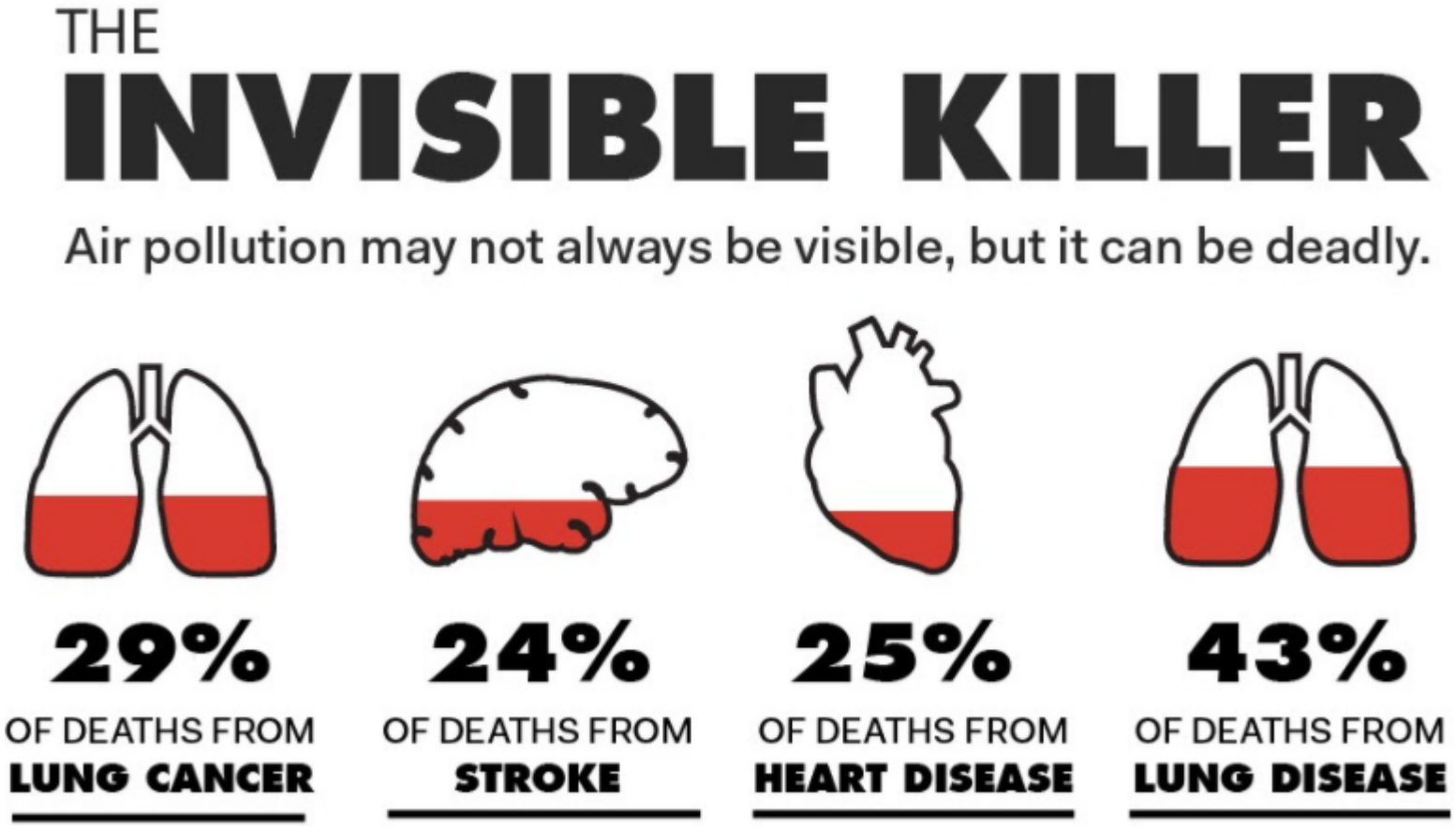
Long-term effects of air pollution on the onset of diseases such as respiratory infections and inflammations, cardiovascular dysfunctions.

Both India and China are seeing periods of rapid expansion, and much of this expansion releases toxic air pollutants that can harm people’s health and lifestyles.

It is one of the main reason for Global warming which is a global phenomenon.

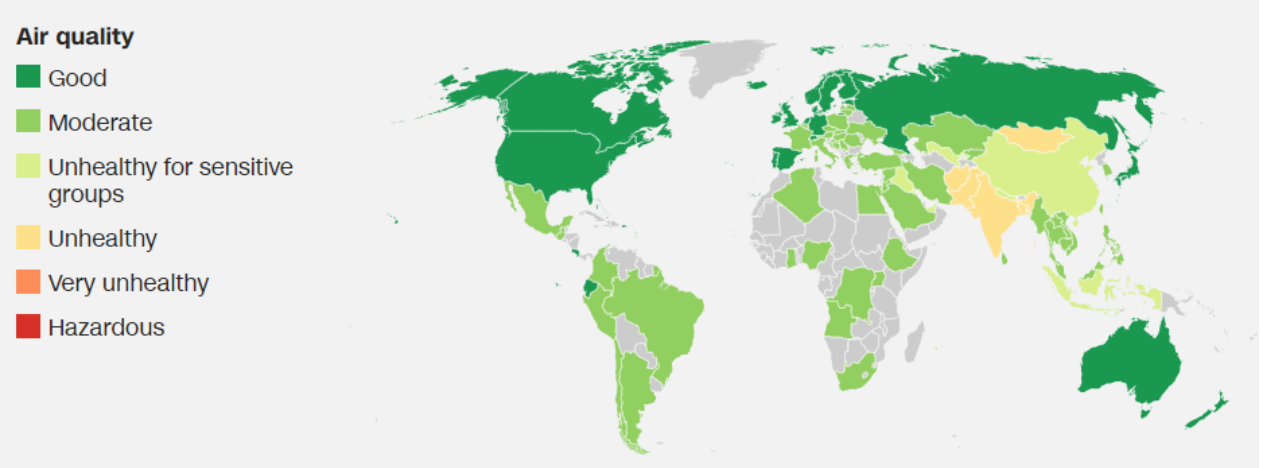


The Invisible Killer



Why India ?

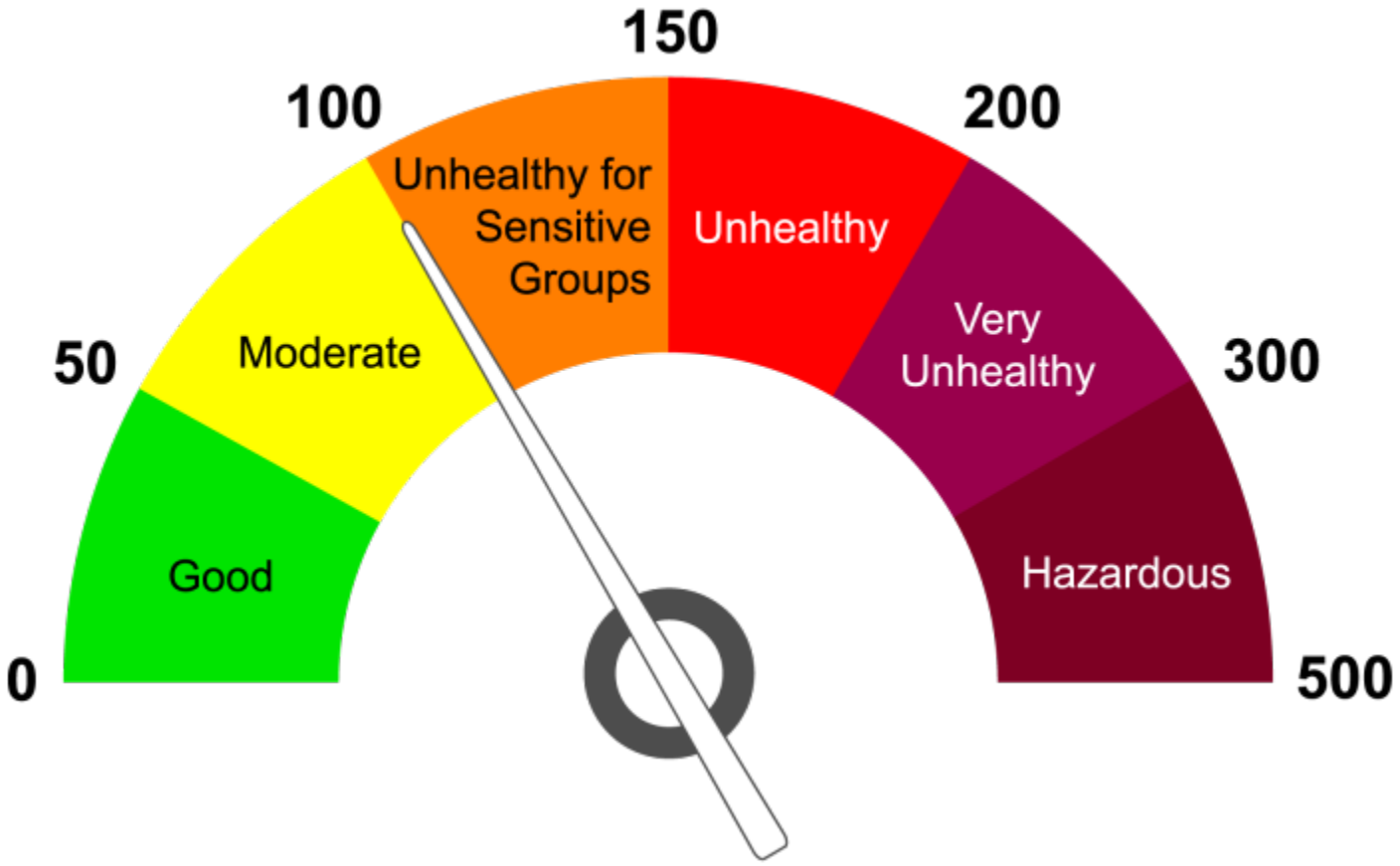
	Major city	US AQI
1	Bishkek, Kyrgyzstan	212
2	Kolkata, India	206
3	Delhi, India	199
4	Lahore, Pakistan	181
5	Karachi, Pakistan	178
6	Dhaka, Bangladesh	178
7	Kathmandu, Nepal	172
8	Kabul, Afghanistan	168
9	Shanghai, China	167
10	Mumbai, India	166
11	Wuhan, China	165
12	Hanoi, Vietnam	163
13	Hangzhou, China	156



3 out of the top 10 polluted cities belongs to India
Dated (12/10/2020)

What is an AQI ? How it is classified?

Air quality Index or AQI is a measure of how clean or polluted the air is.



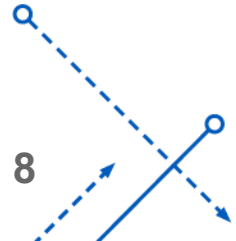
Literature review and Gaps

Many research works were made in the past to predict the AQI levels but they were missing out some of the factors. Some of the papers that were very close to our topic are listed below.

Research Work	Gap
Urban Air Quality Analysis and Prediction Using Machine Learning - IEEE	Logistic regression and Decision tree were only used other methods could be implemented
Air Quality Prediction Of Data Log By Machine Learning - IEEE	No additional predictors were used and very less models were implemented
Urban Air Quality Prediction Using Regression Analysis - IEEE	Geographical and economical factors were not included
Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization - IEEE	Complex models were used hence model interpretability was reduced

Project Scope

- Predicting Air quality index is crucial because of the increase in the toll of pollution and will be an useful investment for an individual and for a community.
- The goal is to find insights and the significance of demographic, geographical and Industrial factors that influences the Air quality index.
- This project 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.
- This analysis will also provide us useful information in identifying the safer cities with lesser pollution levels to develop residential areas for larger communities.



Data collection

AQI data for India was available on Kaggle and was used as the primary data source. Other data for some of the predictors and it source is given below.

DATASET	SOURCE
AQI dataset (2015-2020)	Kaggle
Power Consumption in India	Kaggle
Population Data	Macrotrends & Google

Data Description

- The Main AQI dataset collected from Kaggle consisted of data from year 2015 to 2020. For Prediction and Analysis, data from year 2017 to 2019 was used.
- National Thermal power generation data for each region from the year 2017 to 2019 was extracted and grouped.
- City wise demographic data was extracted for each year (2017 – 2019) and merged with the main data frame.

Data Cleaning and transformations

Initial Variables : AQI data

City	Date	PM2.5	PM10	NO
NO2	NOx	NH3	CO	SO2
O3	Benzene	Toluene	Xylene	StationName
AQI_Bucket	AQI	StationId	State	Status
Region	Month	Year	Season	Weekday_or_weekend
Regular_day_or_holiday	AQ_Acceptability			

Data Cleaning and transformations

Initial Variables : Thermal data

Date	Region	Thermal.Generation.Actual..in.MU	Thermal.Generation.Estimated..in.MU	Nuclear.Generation.Actual..in.MU
Nuclear.Generation.Estimated..in.MU	Hydro.Generation.Actual..in.MU	Hydro.Generation.Estimated..in.MU		

Initial Variables : Demographic data

City	Region	State	Year	Population

Data Cleaning and 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 together as a single particulate matter column called PM & the columns Benzene, Xylene and Toluene were combined together as a single column called as 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 to 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 variable. The geographical position of each city were added as a categorical variable with five levels.

Data Cleaning and transformations

Final Variables : AQI data

City	Date	PM	BTX	NO
NO2	NOx	NH3	CO	SO2
O3	AQI	State	Region	Month
Year	Season			

Data Cleaning and transformations

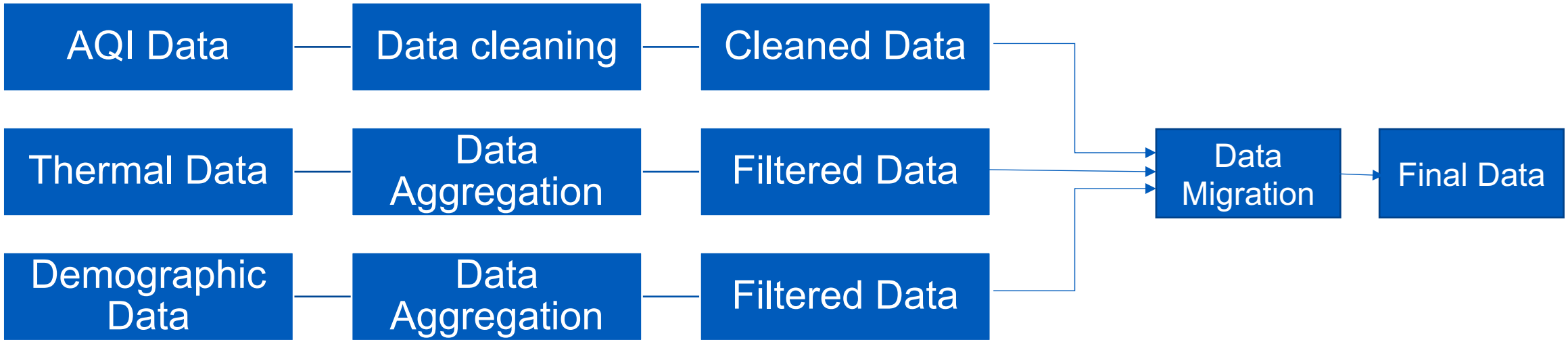
Final Variables : Thermal data

Year	Region	Thermal

Final Variables : Demographic data

City	Region	State	Year	Population

Data Cleaning and transformations



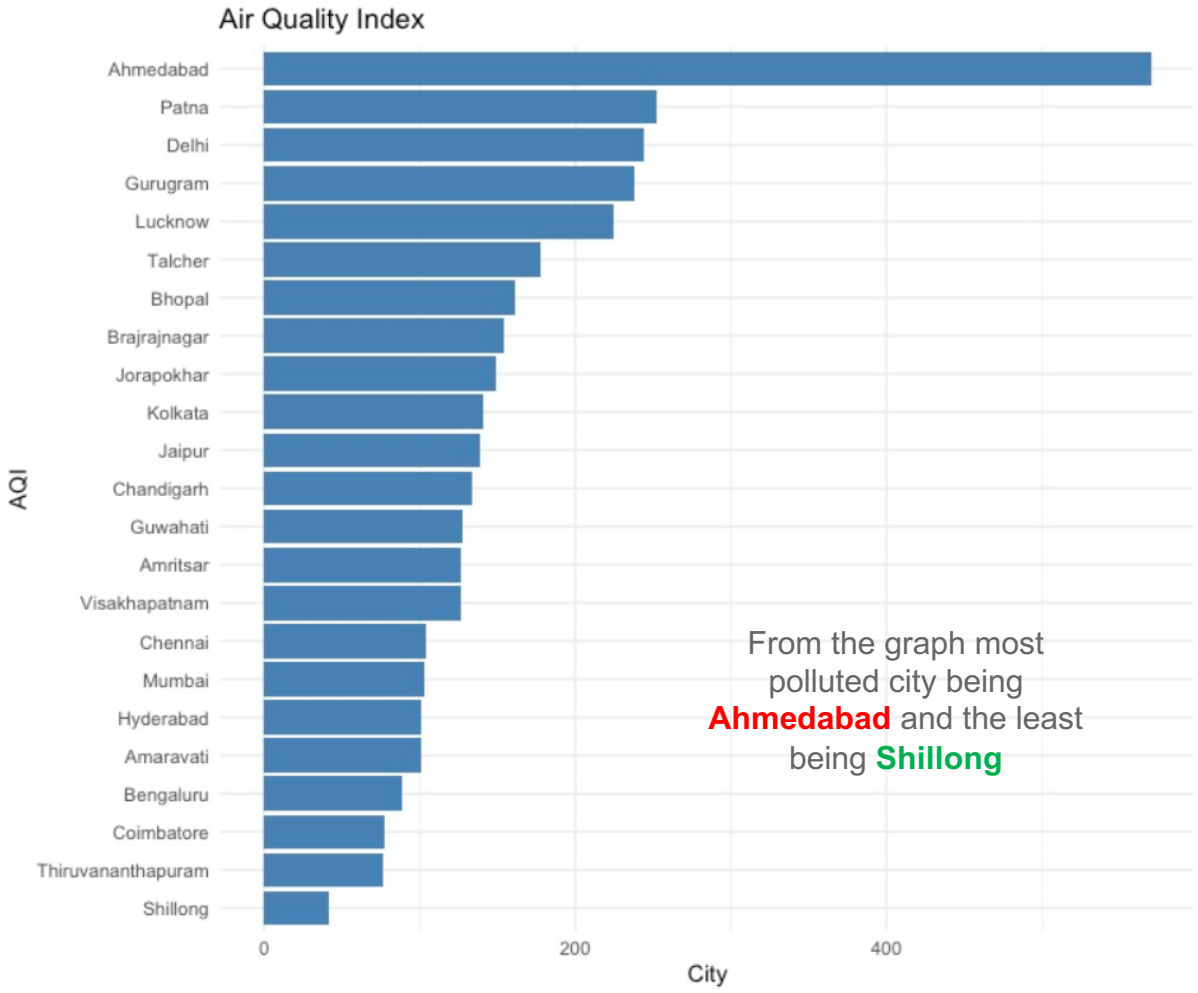
Data Cleaning and transformations

Final Data frame :

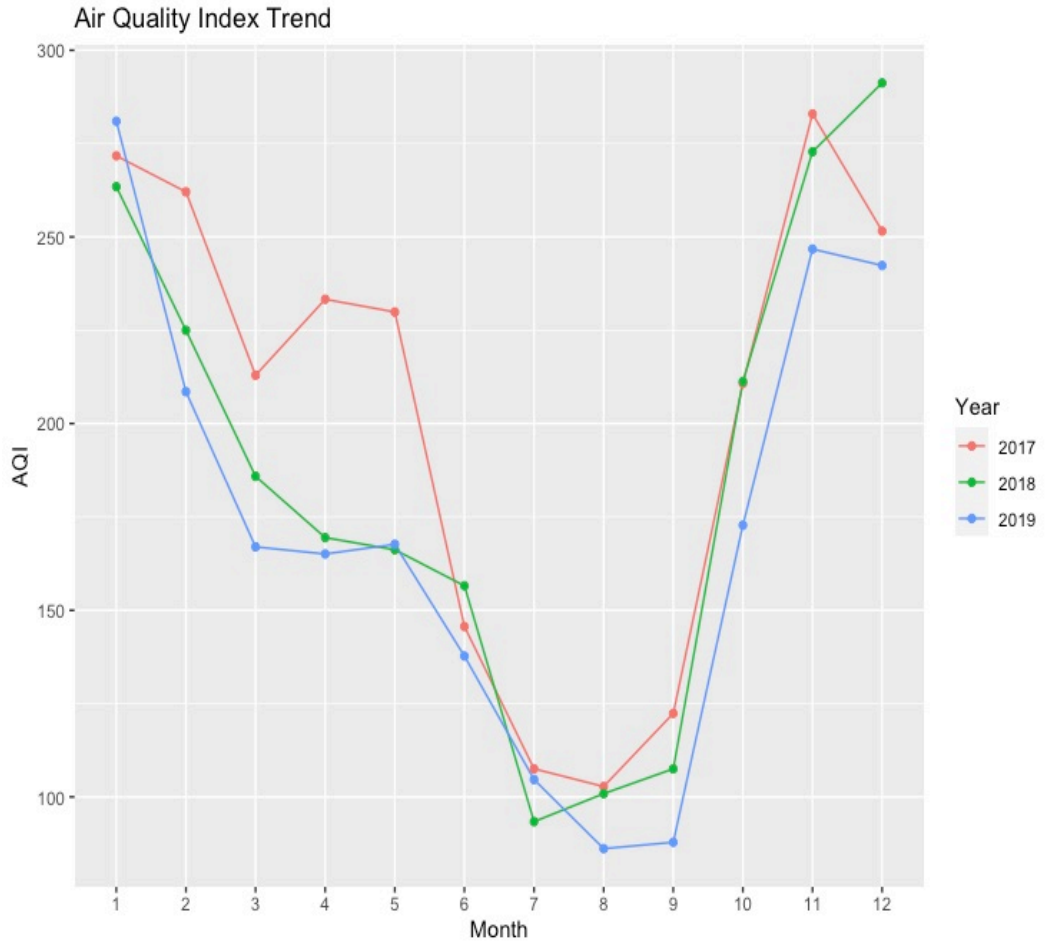
Number of rows	107308
Number of columns	18

Variable	Type	Unit
City	Chr	Category
State	Chr	Category
Region	Chr	Category
Year	Chr	Category
Month	Chr	Category
Season	Chr	Category
Population	int	Count
Thermal	num	Megaunit
PM	num	Molecule/cm^2
BTX	num	Molecule/cm^2
NO	num	Molecule/cm^2
NO2	num	Molecule/cm^2
Nox	num	Molecule/cm^2
NH3	num	Molecule/cm^2
CO	num	Molecule/cm^2
So2	num	Molecule/cm^2
O3	num	Molecule/cm^2
AQI	num	Unit measure

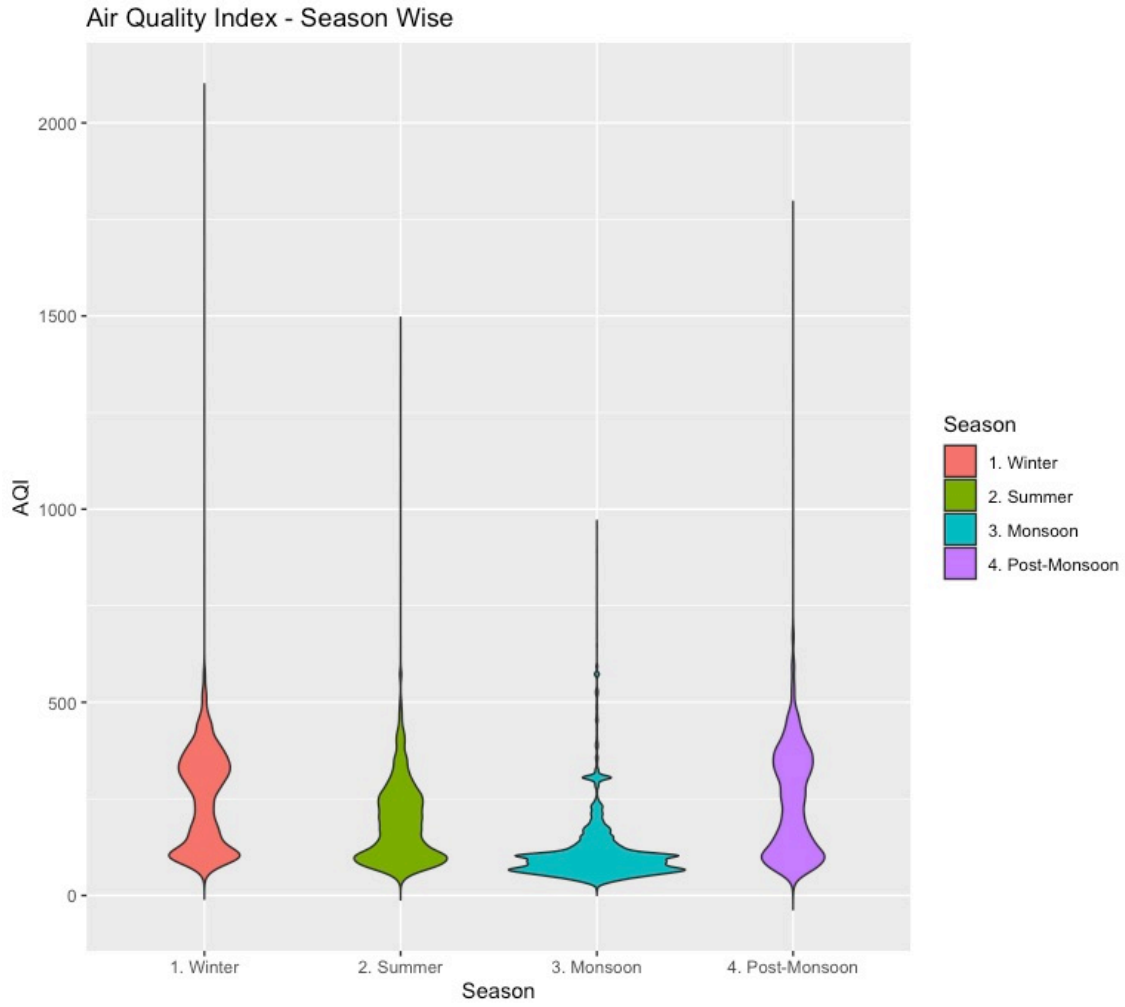
Data Visualization



From the graph most polluted city being **Ahmedabad** and the least being **Shillong**

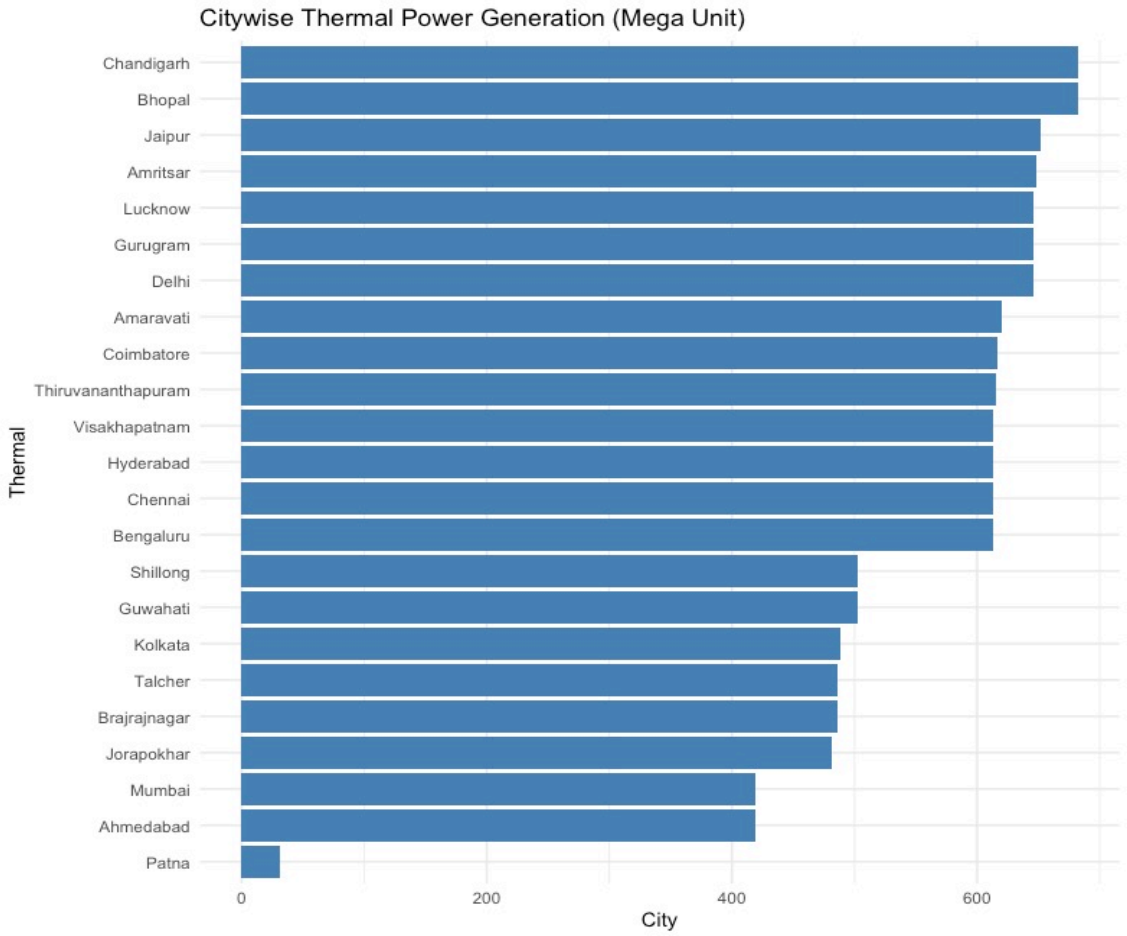


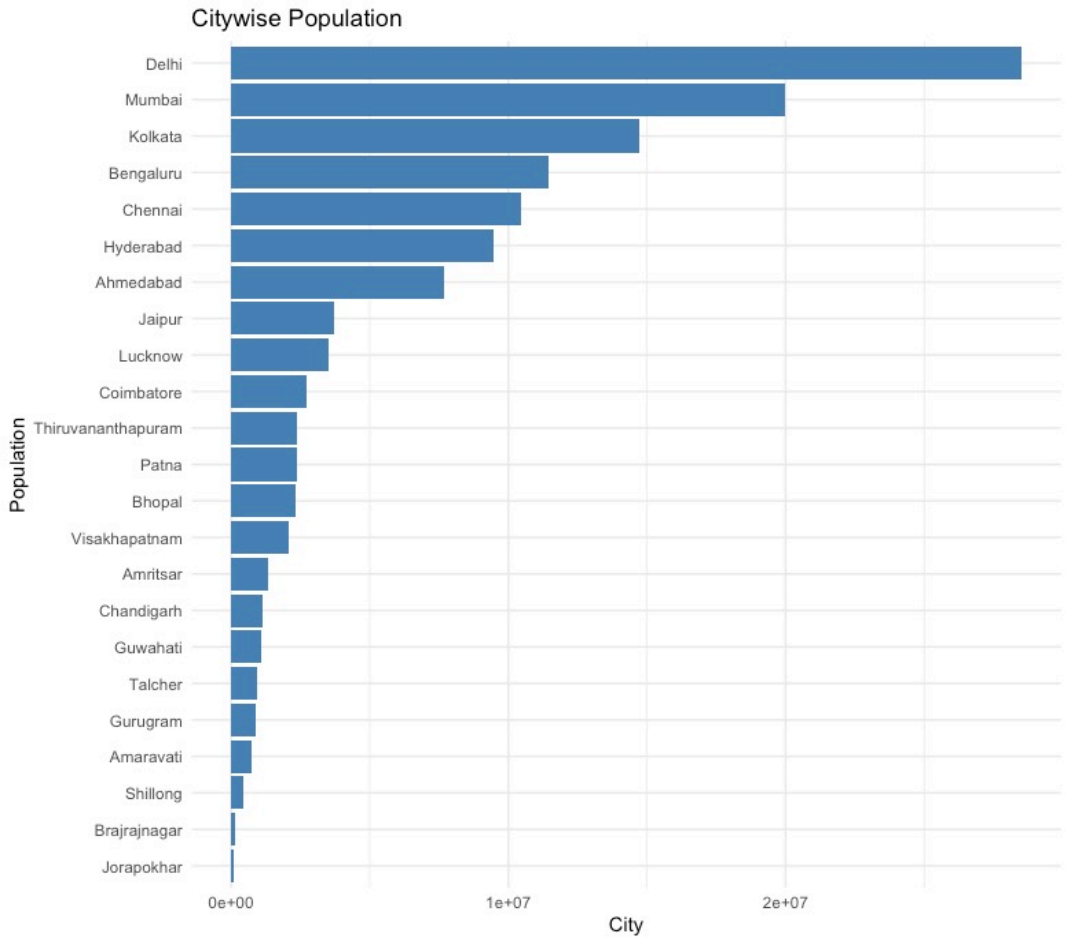
Data Visualization



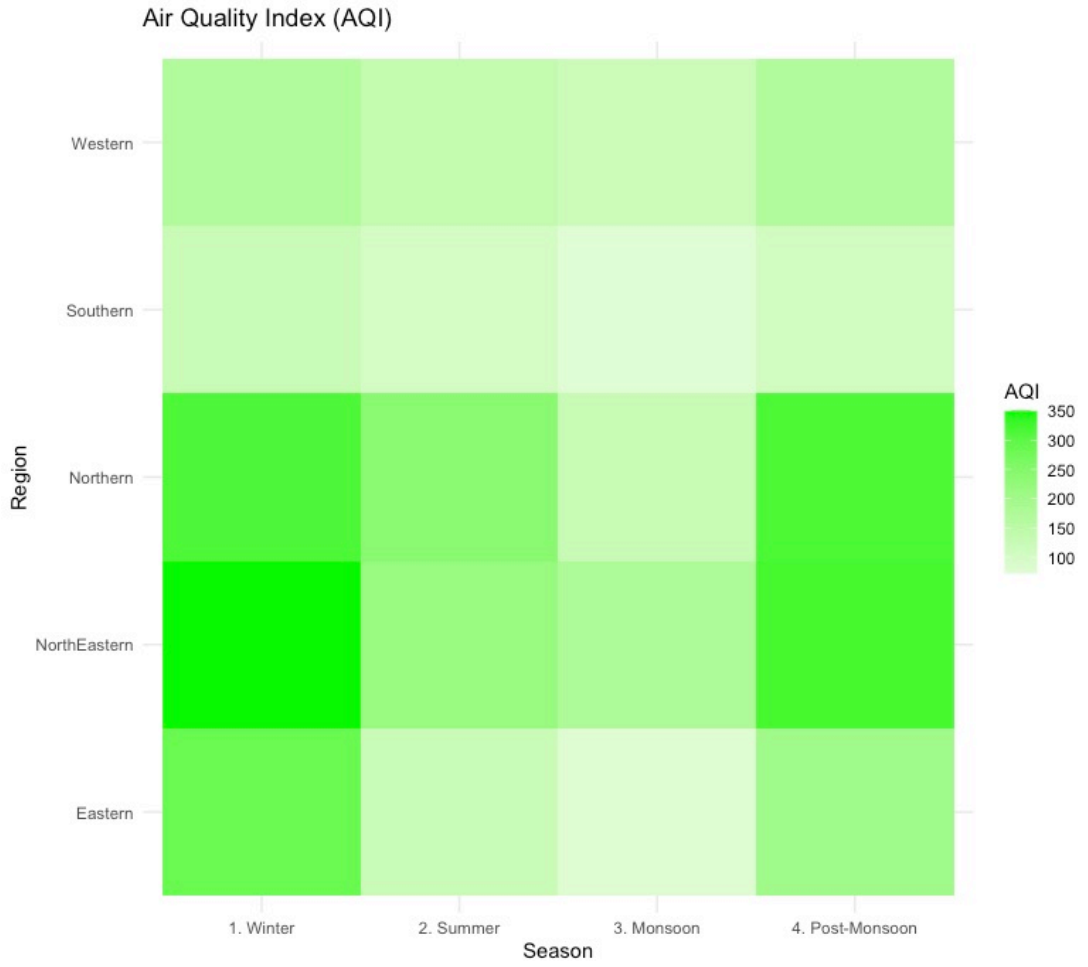
This plot helps us identify the impact of different seasons in the Air pollution levels

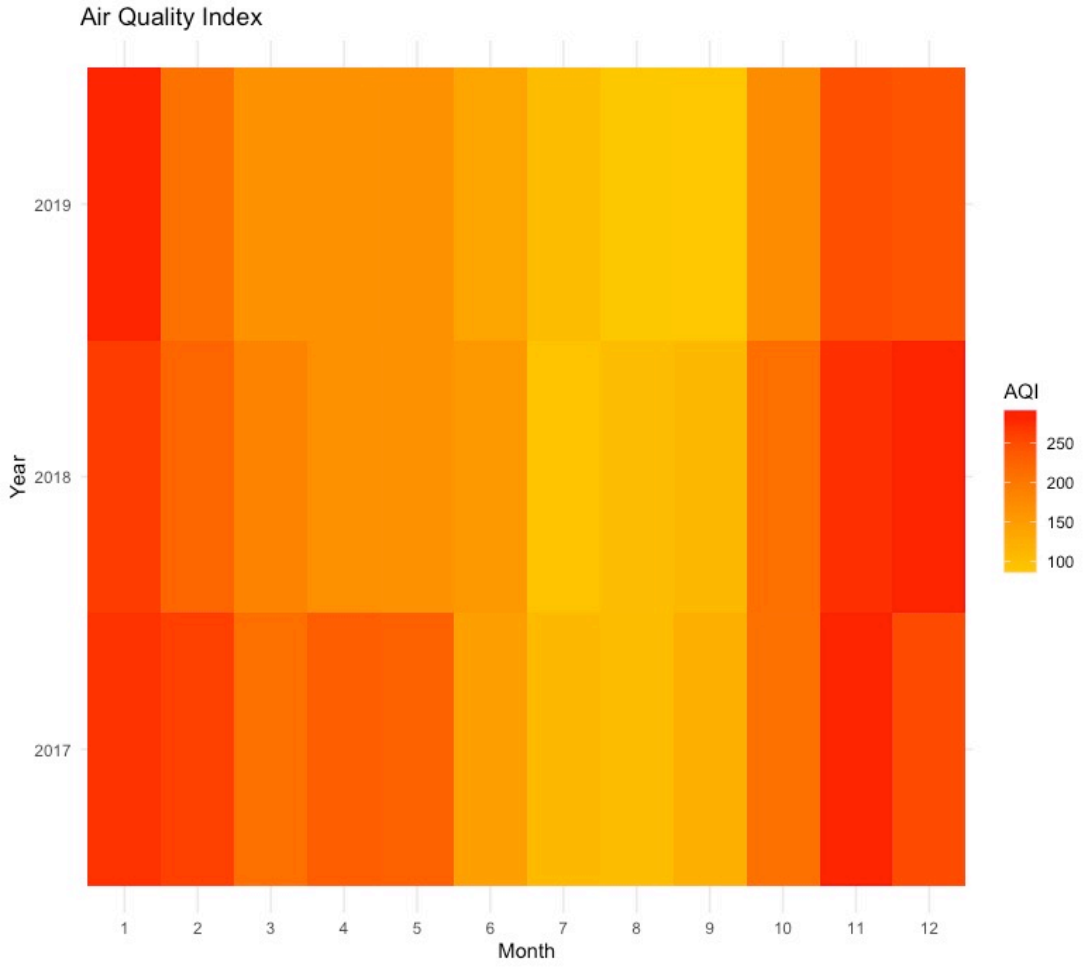
Data Visualization



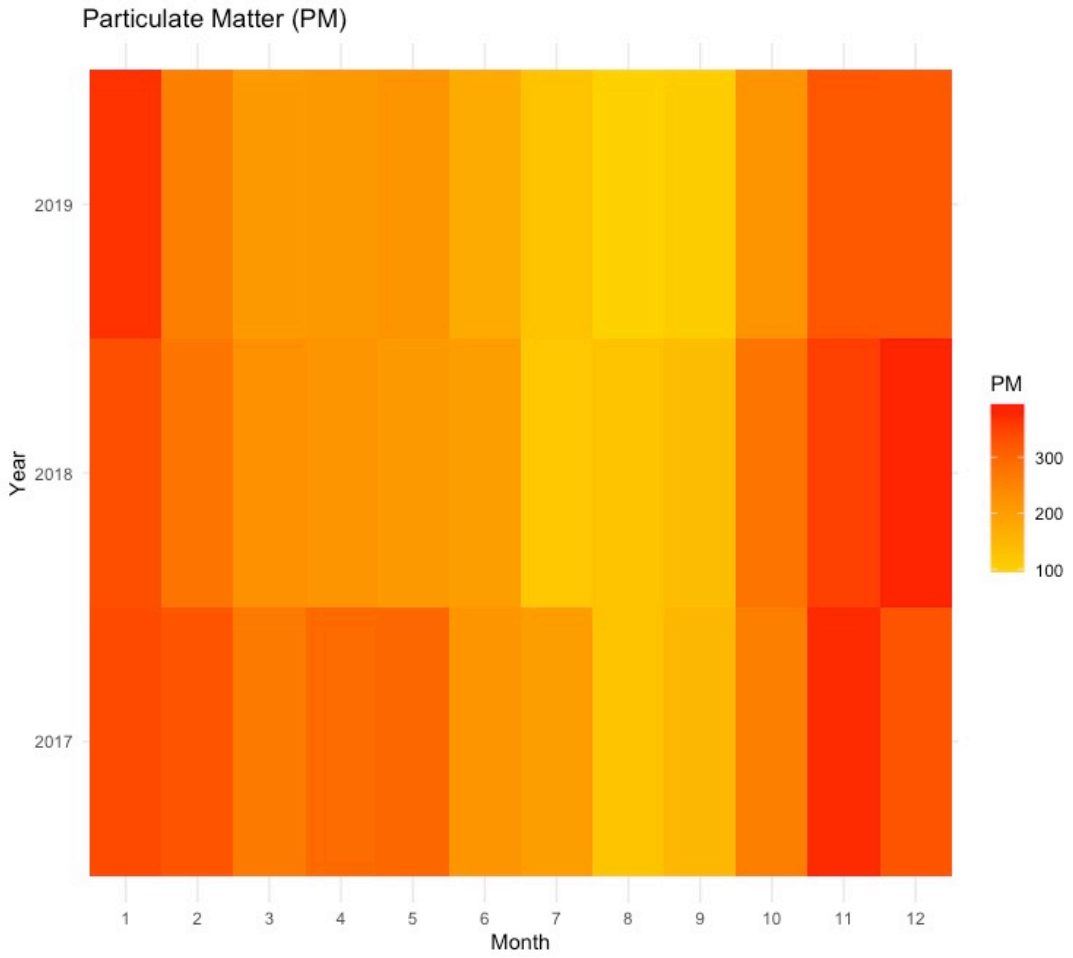


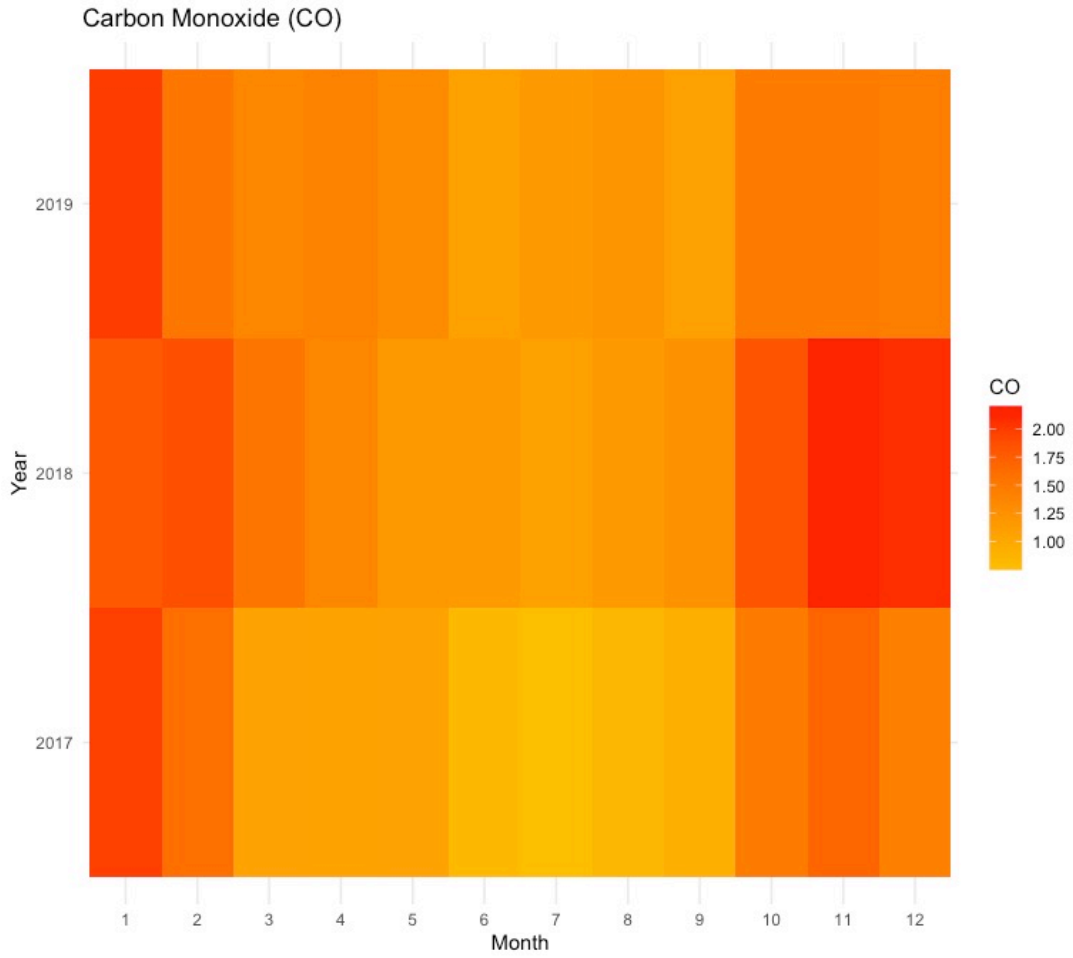
Data Visualization



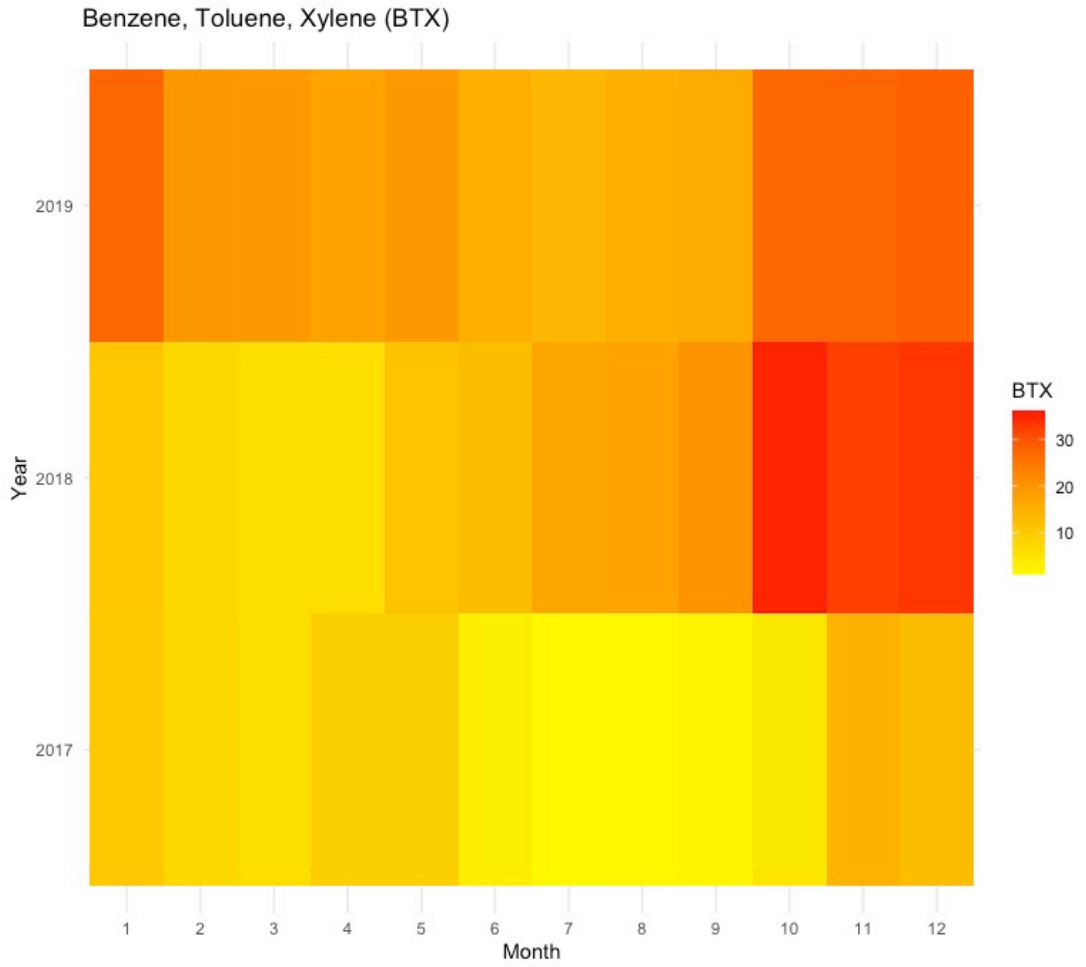


Data Visualization

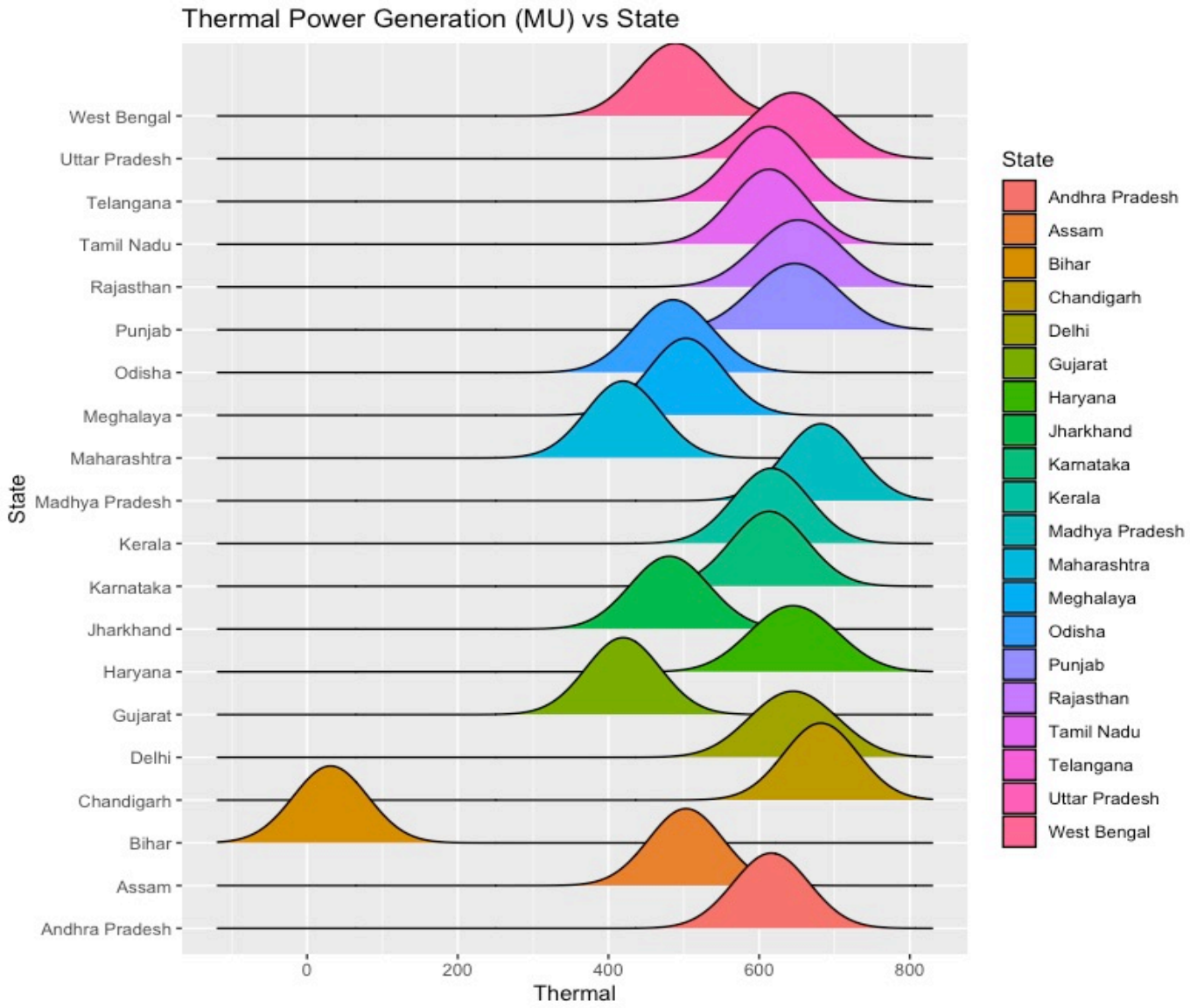




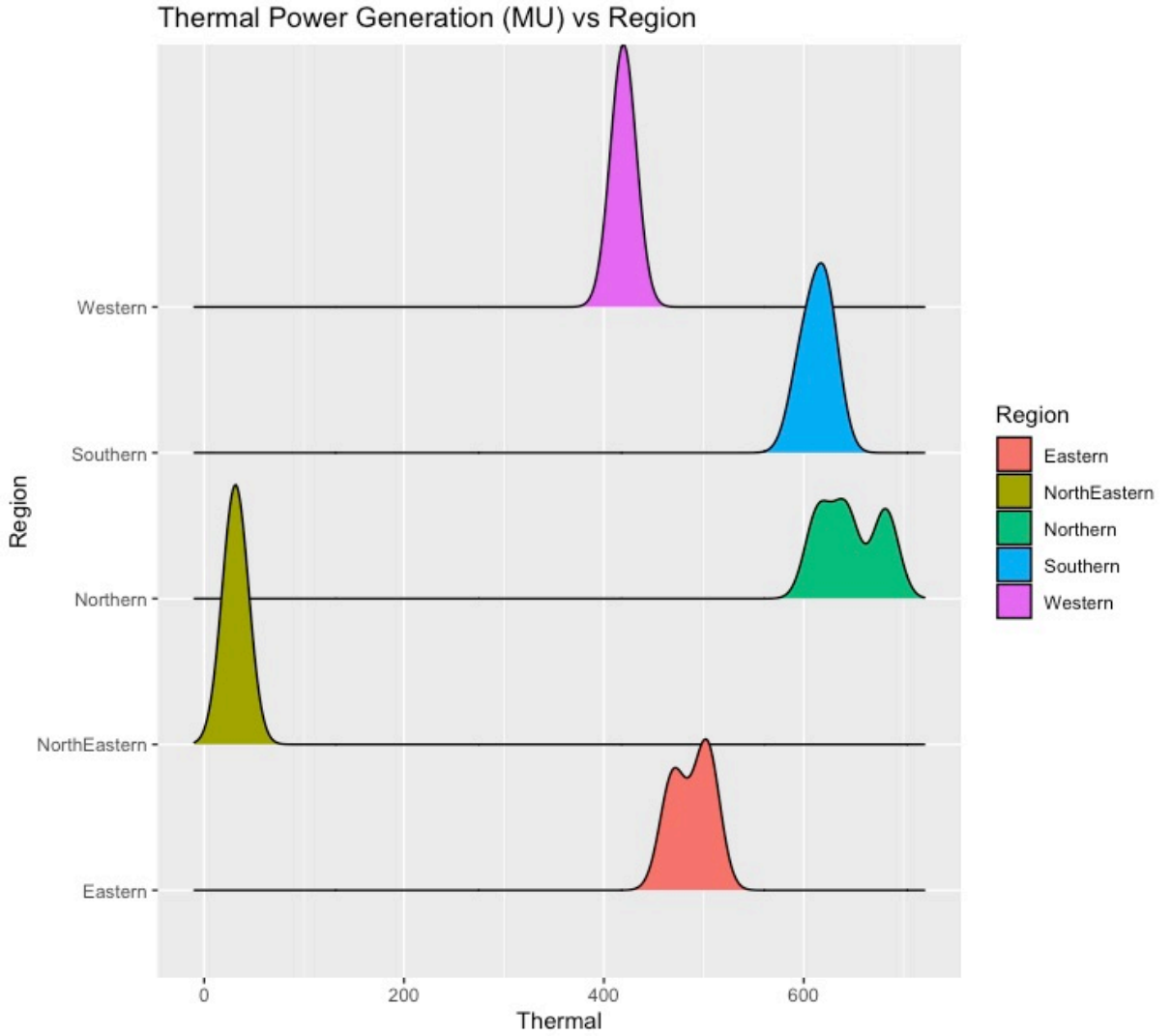
Data Visualization

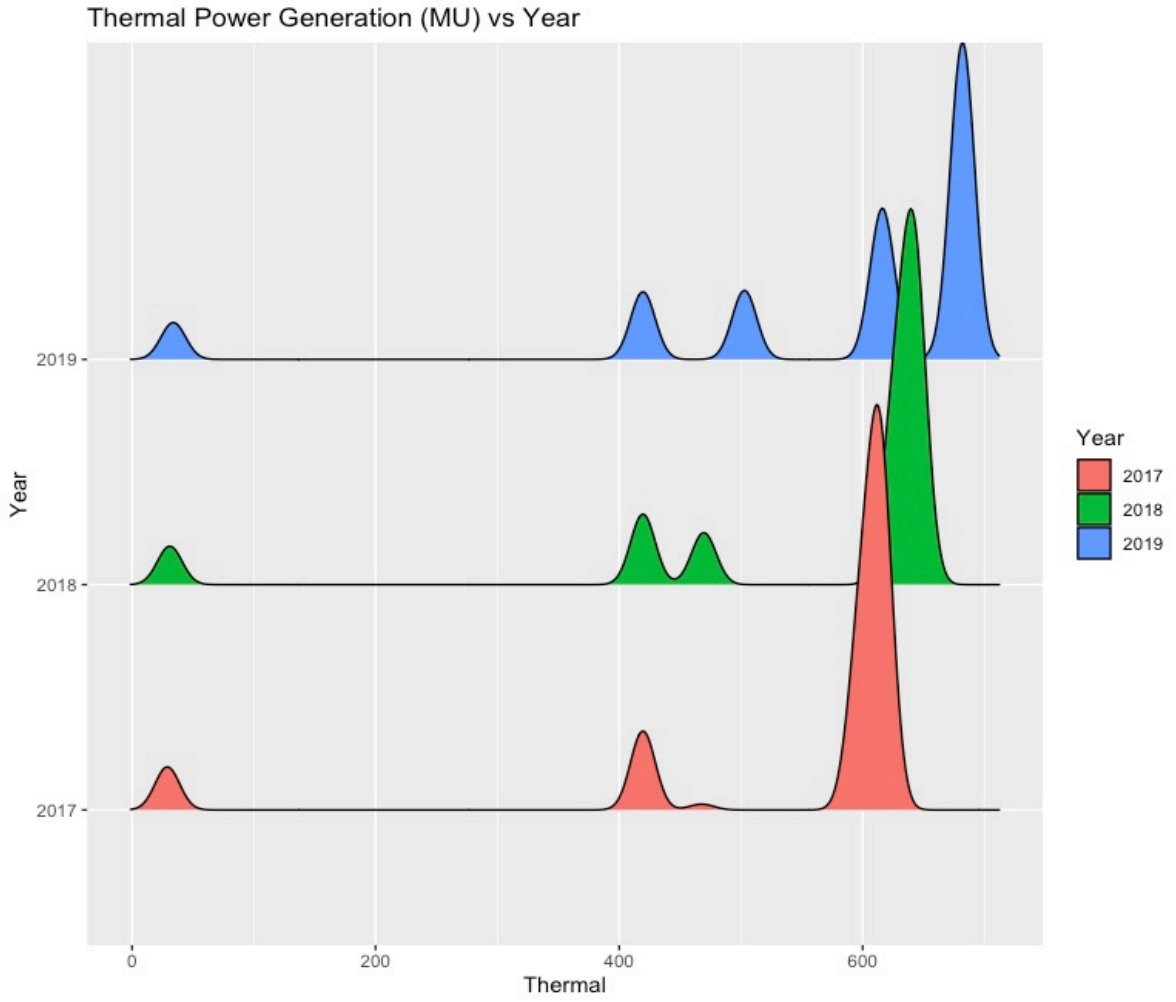


Data Visualization

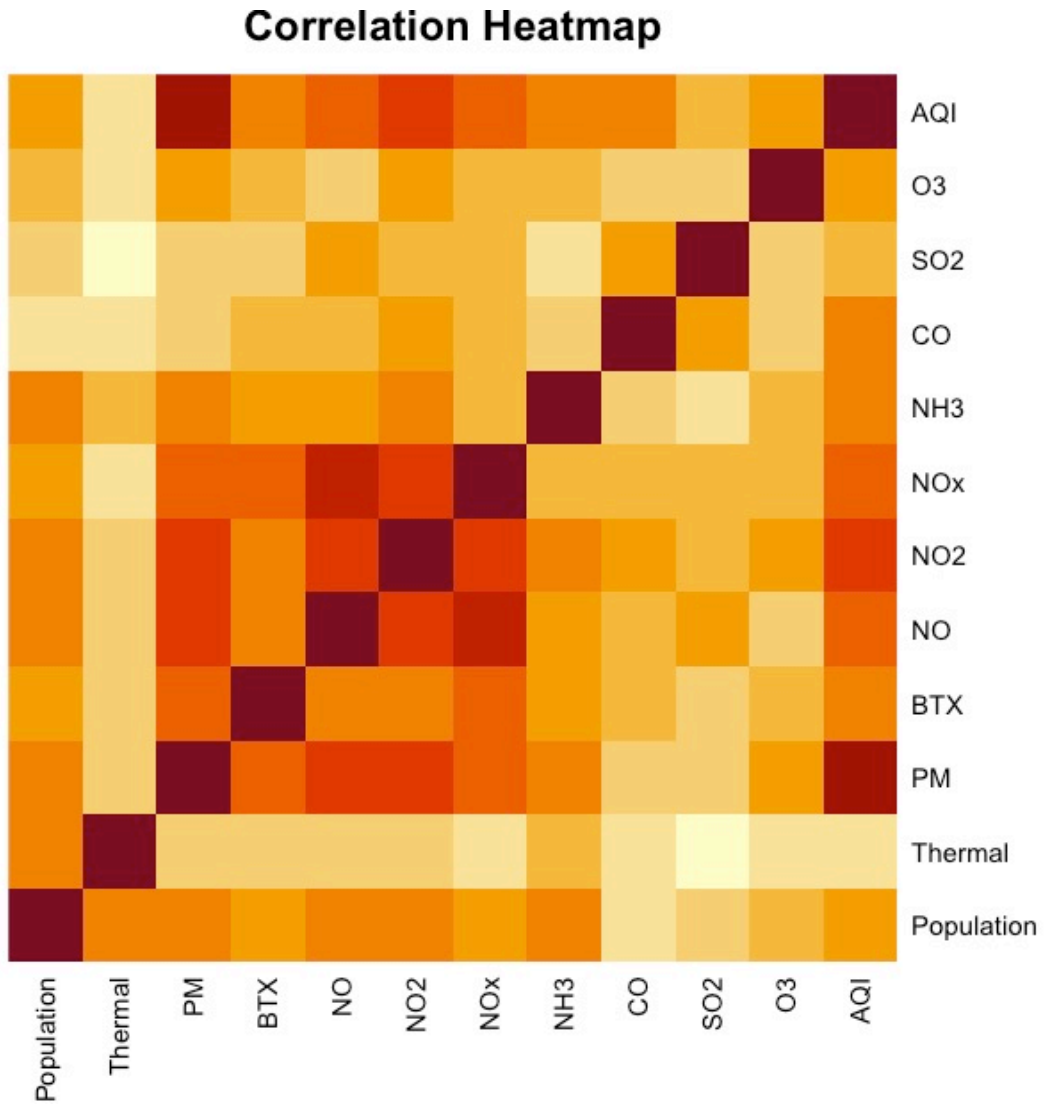


Data Visualization





Data Visualization – Correlation Matrix



Adding other Predictors to the data

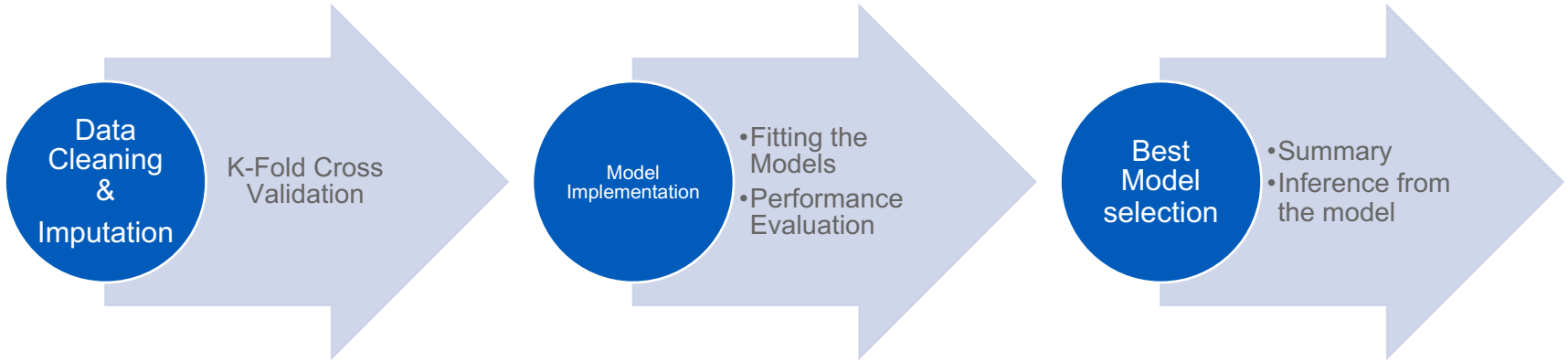
Thermal Power Generation Data: The conventional Thermal Power plants that use coal for producing electricity. Hence we added the column that contains Power output at that year. The data was collected from Wikipedia and Govt websites.

Region wise category : The geographical position of the city might play an important role in the pollution levels. Ozone levels might differ from region to region hence adding region as a predictor could help us study the effect of geographical location in determining the AQI levels.

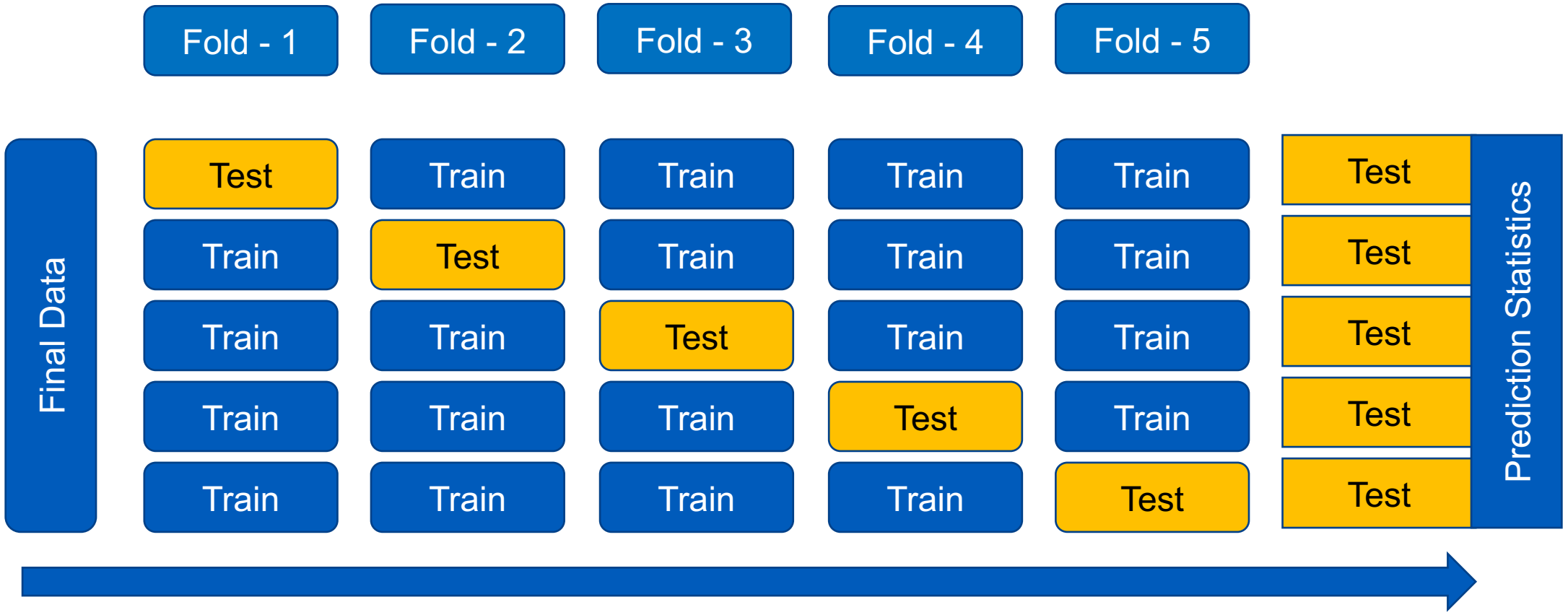
Geographical Position
Northern
Northeastern
Southern
Eastern
Western

Model creation and testing Framework

The Data was cleaned and new predictor values were added to the data frame. The testing framework is given below.



K – Fold Cross Validation



Evaluation Metrics

1. Root Mean Squared Error

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

2. R - Squared

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

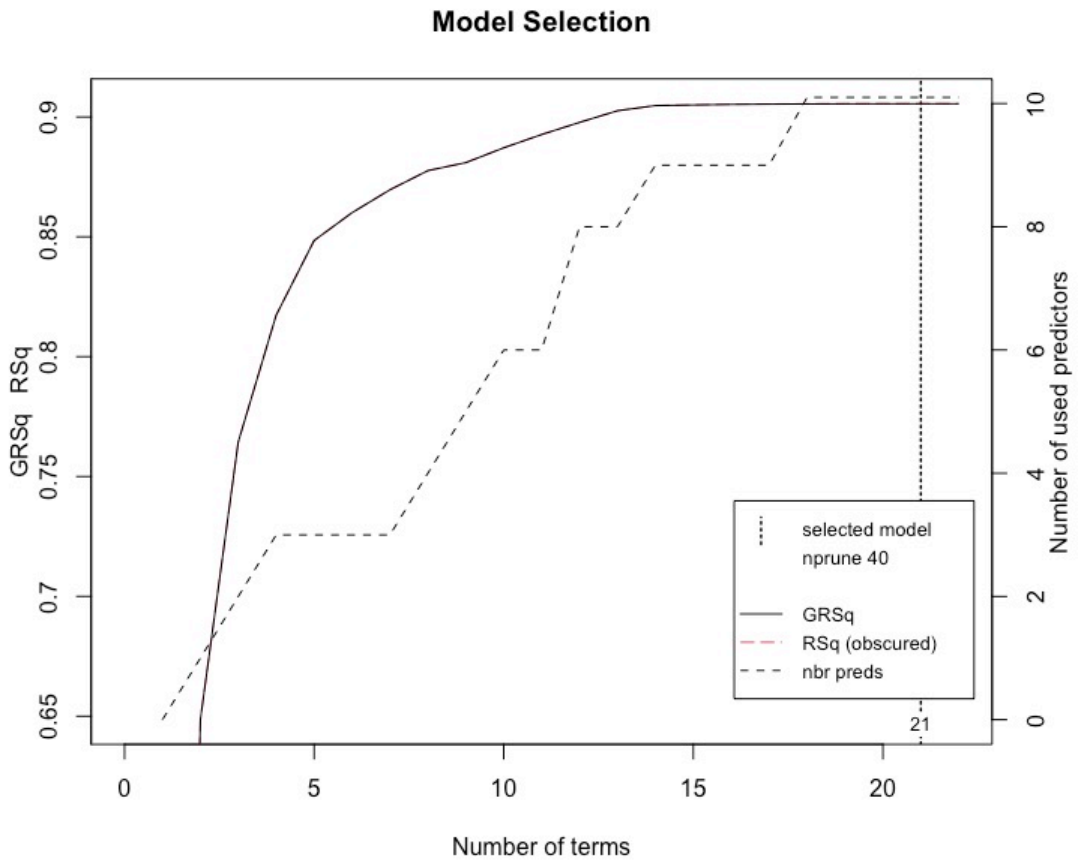
3. Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j|$$

Model Performance Evaluation – 10 fold Cross Validation - Test Data Results

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

Multivariate Regression Splines – Results



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

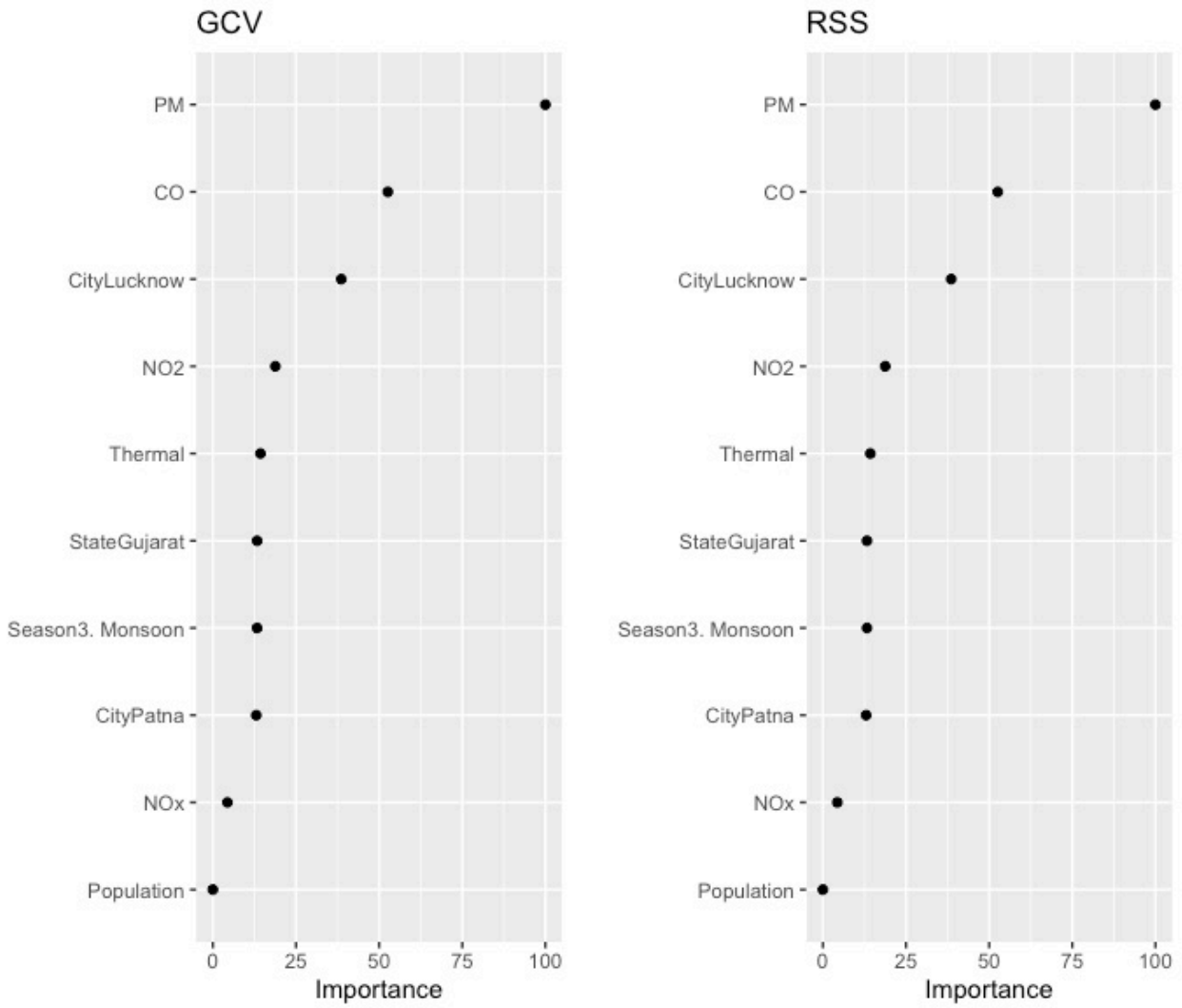
Multivariate Regression Splines – Results

```
Multivariate Adaptive Regression Spline

107308 samples
    17 predictor

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

  RMSE      Rsquared   MAE
  39.08002  0.9055027  24.56274

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

Final Model

```
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, ...
Number of terms at each degree of interaction: 1 6 14
GCV 1533.336    RSS 164382853    GRSq 0.9055126    RSq 0.9056007
```

Multivariate Regression Splines – Variable Importance



Variables with higher and lower importance can be seen from the Plot after training and testing the model

Conclusion

- With the help of k – Fold Cross Validation we have obtained the evaluation metrics for the implemented model in test data and chose MARS as the best model.
- MARS model has the best results when compared to other implemented models and its easy to interpret as well.
- From the parameters and coefficients table we can identify how each predictors and the interactions between them influenced in fitting the spline to the data.
- From the variable importance plot we can evidently state that the extracted features like thermal power generation and population had a significant impact in predicting the Air Quality Index.
- Also other factors like Particulate Matter, Season, Carbon Monoxide, etc. had a significant impact as well.
- With 107308 observations and 17 predictors and for a 10-fold cross validation, MARS obtained a test error of 39.08, R-Squared of 90.6% and MAE of 24.6

Future Work

We would like to improvise the research by adding additional predictors in the future. Some of the factors we would like to include are as follows.

- Coal consumption per city (in tons)
- Industrial emissions data
- Electric vehicle sales data
- Conventional vehicle usage data (Petroleum & Diesel)
- Mortality rate due to respiratory diseases for each cities.

We would also like to test our model with the data which will be obtained for the year 2020.

Hurdles in future research

- CENSUS data will be available only after 2021
- No standard repository for Imports and exports data
- Different emission norms for different type of Industries

References

- 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
- 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
- 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.
- 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
- 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*

Data Source

- Urban Air Quality Analysis and Prediction Using Machine Learning. Source: https://ieeexplore.ieee.org/document/9063845
- Air Quality Prediction Of Data Log By Machine Learning. Source: https://ieeexplore.ieee.org/document/9074431
- Urban Air Quality Prediction Using Regression Analysis. Source: https://ieeexplore.ieee.org/document/8929517
- Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization. Source: https://ieeexplore.ieee.org/document/8650518
- Analysis and Visualization of Air Quality Using Real Time Pollutant Data. Source: https://ieeexplore.ieee.org/document/9074283