SUPER
SAIYANS

JUST ‘Q’
IN AI

OUR TEAM

Joseph Malombe

Mary Mwangi

Anthony Thuita

Rwenji Murengaitta

Maureen Oketch

PROJECT OVERVIEW

Monitoring air quality is crucial for protecting public health, the environment, and ensuring regulatory compliance. By measuring contaminants in the atmosphere, societies can promote sustainable development and take proactive measures to reduce pollution and create cleaner and healthier living environments.

PROBLEM STATEMENT

In our busy world, where each breath holds significance, hidden perils linger in the air essential for life. From urban streets to remote locales, air pollution is pervasive, affecting millions yearly. Yet, amidst this challenge lies opportunity. Envision a realm where technology defends against pollution, with real-time air quality monitoring offering hope. Each data point moves us towards a future where clean air is a universal entitlement. Our system empowers individuals and communities to safeguard their well-being.

OBJECTIVES

Main objective is to create a system that can predict values with simulated figures fed to it.

- To enable easy management of air quality : Enhance the efficiency of air quality management by providing accurate predictions and timely alerts
- To Integrate time series forecasting techniques like ARIMA, SARIMA, or LSTM to predict future values of AQI and alert users about potential changes in air quality
- To utilize regression models to analyze the relationship betweeN AQI and pollutants
- To Create a value proposition system for health care:

AIR QUALITY INDEX

AQI Category	Associated Health Impact
Good (0 to 50)	Minimal impact
Satisfactory (51 to 100)	May cause minor breathing discomfort to sensitive people
Moderately Polluted (101 to 200)	May cause breathing discomfort to the people with lung disease such as asthma and discomfort to people with heart disease, children and older adults
Poor (201 to 300)	May cause breathing discomfort to people on prolonged exposure and discomfort to people with heart disease
Very Poor (301 to 400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases
Severe (401 to 500)	May cause respiratory effects even on healthy people and serious health impacts on people with lung/heart diseases. The health impacts may be experienced even during light physical activity

BUSINESS UNDERSTANDING

According to various studies and reports, Of the 30 most polluted cities in the world, 21 were in India in 2019.

Air pollution causes 2 million premature deaths in India annually, from vehicle and biomass emissions.

Industrial pollution accounts for 51%, vehicles for 27%, crop burning for 17%, and other sources for 5% of the pollution.

Importance of clean air is crucial as concerns about pollution and its impact on health grow, demanding innovative and effective solutions.

Metric Of Success

- Evaluate how well the regression model fits the observed data.
- Measure the average magnitude of errors in the predictions.
- Measure the square root of the average of the squared differences between the predicted and actual values

R-squared (R^2)

Mean Absolute Error (MAE):

Root Mean Squared Error (RMSE)

Data Description

Introduce the project. Provide a quick background and rationale. Briefly share its overall scope as well as expected outcomes.

01 The data represents information from 2015- 2020

02 The dataset has 707,876 rows and 16 columns

03 29,810 Null Values

04 NO Duplicates

Data Description

Columns information

City

The city where the monitoring station is located.

Date time

The date and time of the air quality measurement.

PM2.5

Particulate Matter with a diameter of 2.5 micrometers or less.

PM10

Particulate Matter with a diameter of 10 micrometers or less.

Pollutants

NO, NO2, Nox, NH3, CO, SO2, O3, Benzene, Toluene, Xylene .

AQI

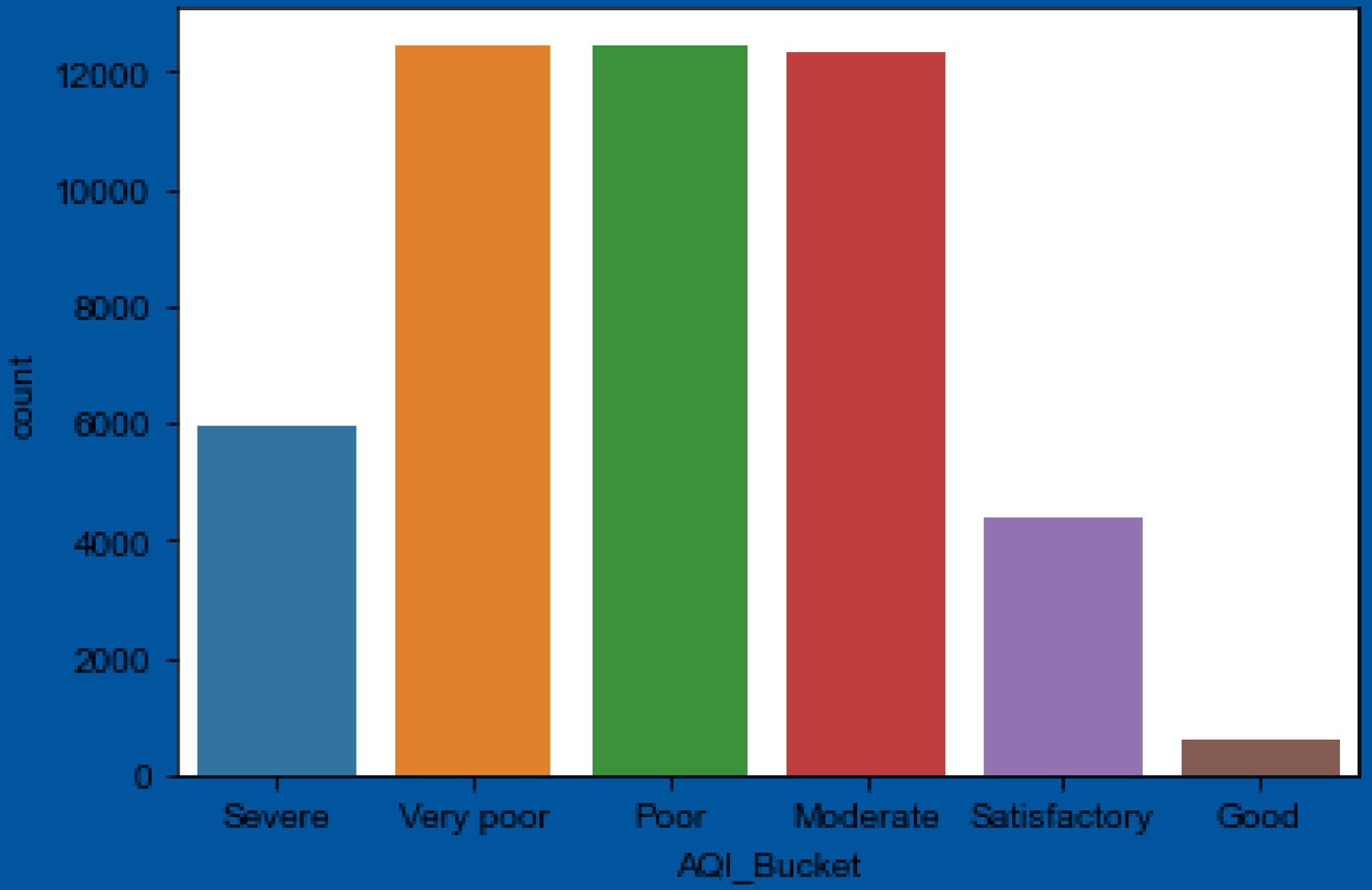
An air quality index, derived from pollutant concentrations, typically using standardized formulas from environmental agencies.

AQI_Bucket

Lorem Ipsum is simply a dummy text of the typesetting industry.

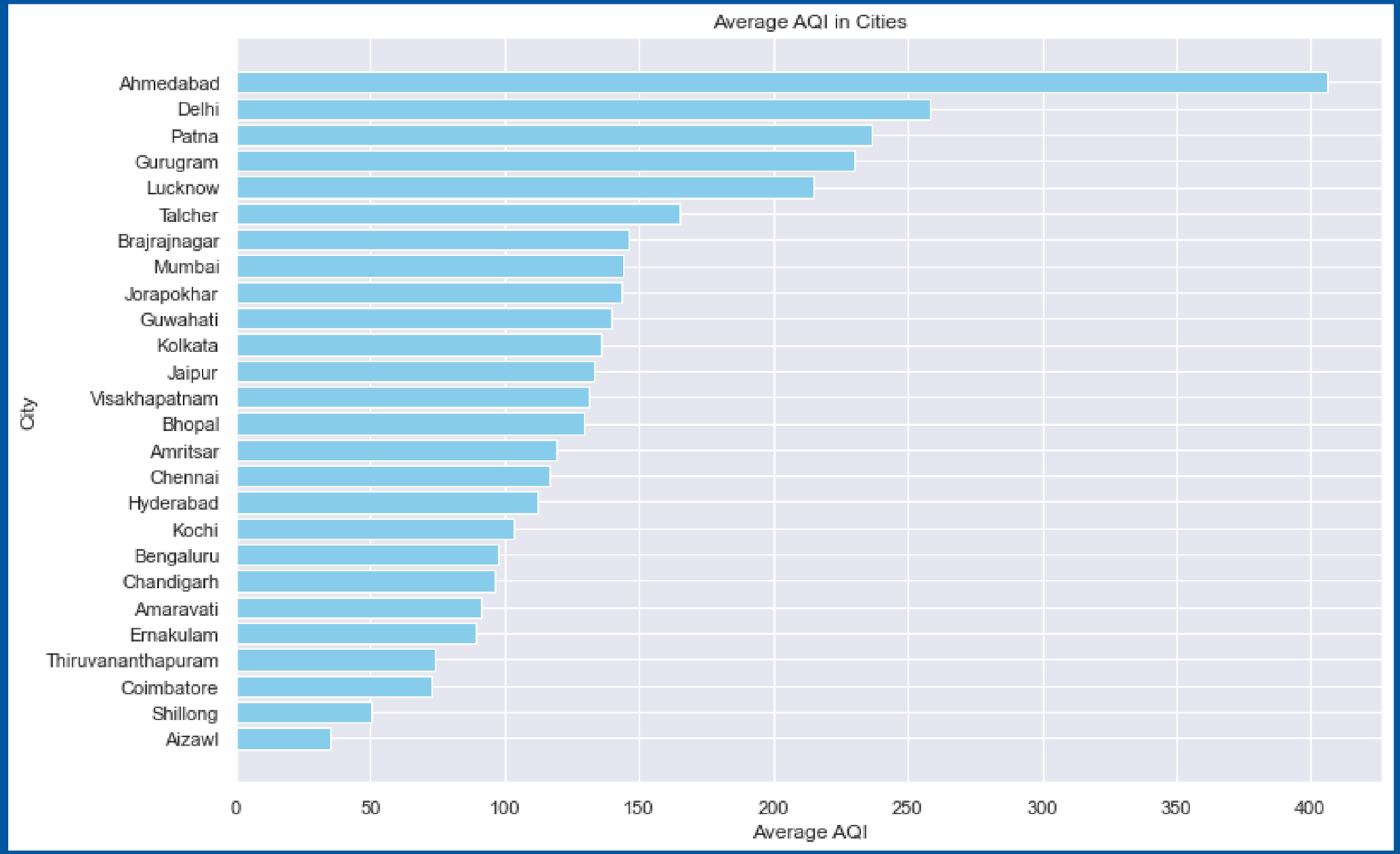
EXPLORATORY DATA ANALYSIS

DISTRIBUTION OF THE AQI_BUCKET



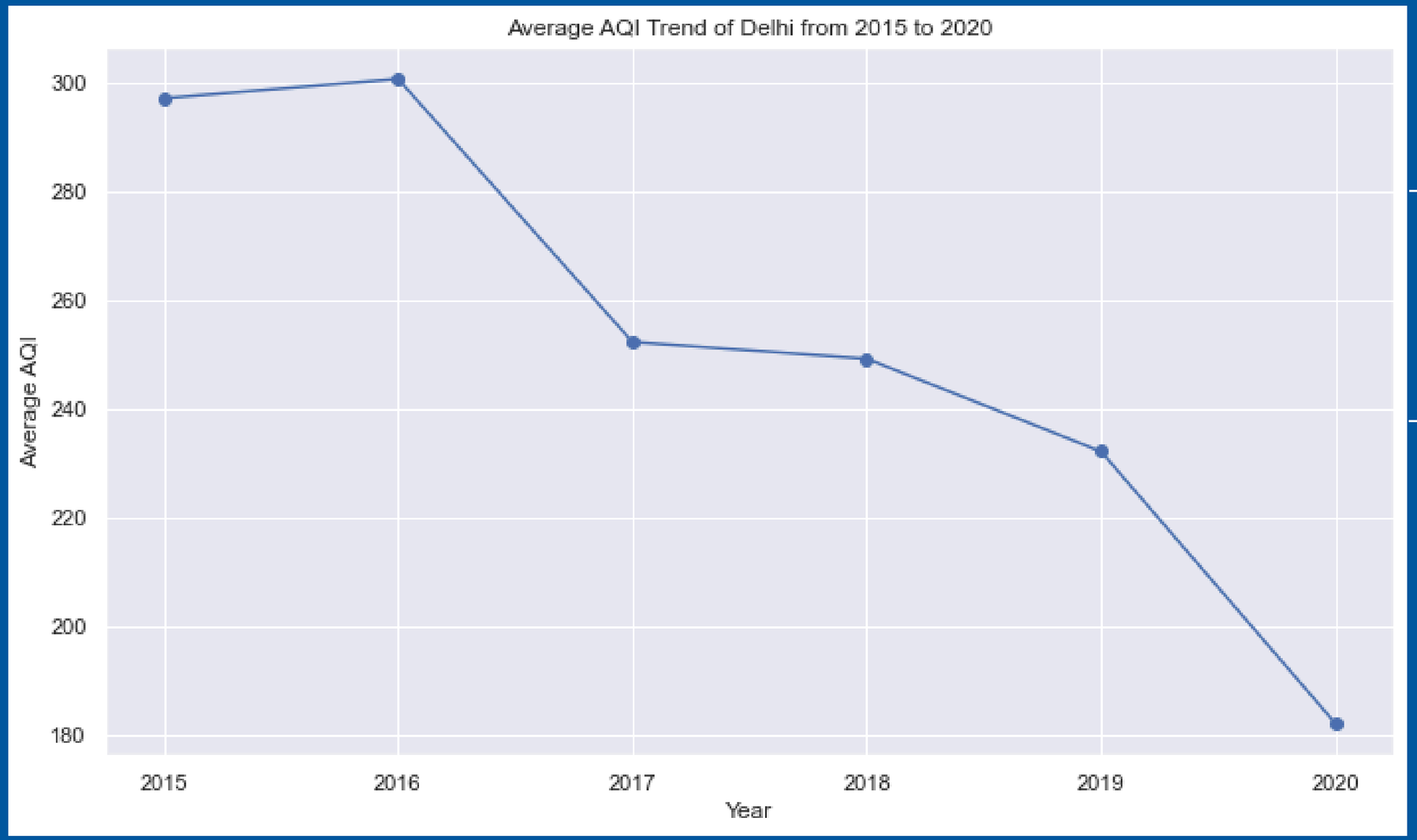
Moderate, Satisfactory, and Adequate represent the majority classes, embodying a significant portion of the spectrum, whereas Severe, Good, Poor, and Very Poor stand as the minority classes, delineating the edges of the continuum.

THE AVERAGE AQI TRENDS IN THE CITY



Ahmedabad has a poorer air quality in this city while Aizawl has better air quality in this city.

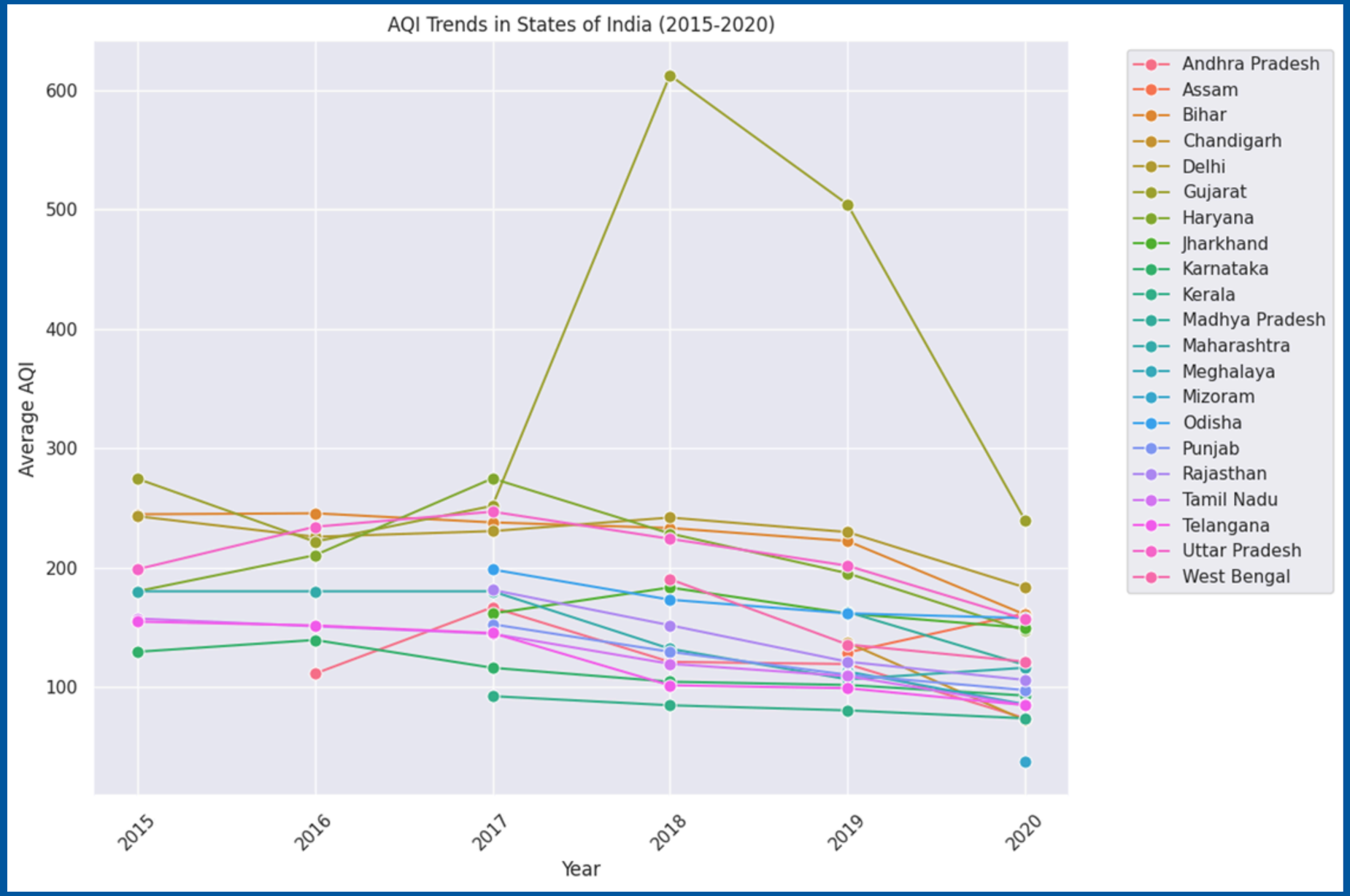
Visualize the AQI trend of Delhi from 2015-2020



-There seems to be a decreasing trend in the average AQI over the years. The AQI was highest in 2016(300) and gradually decreased over the following years, reaching its lowest point in 2020 (182)

-The decreasing trend in AQI suggests an improvement in air quality over time probably due to awareness creation around this challenge

AQI TRENDS IN STATES IN INDIA

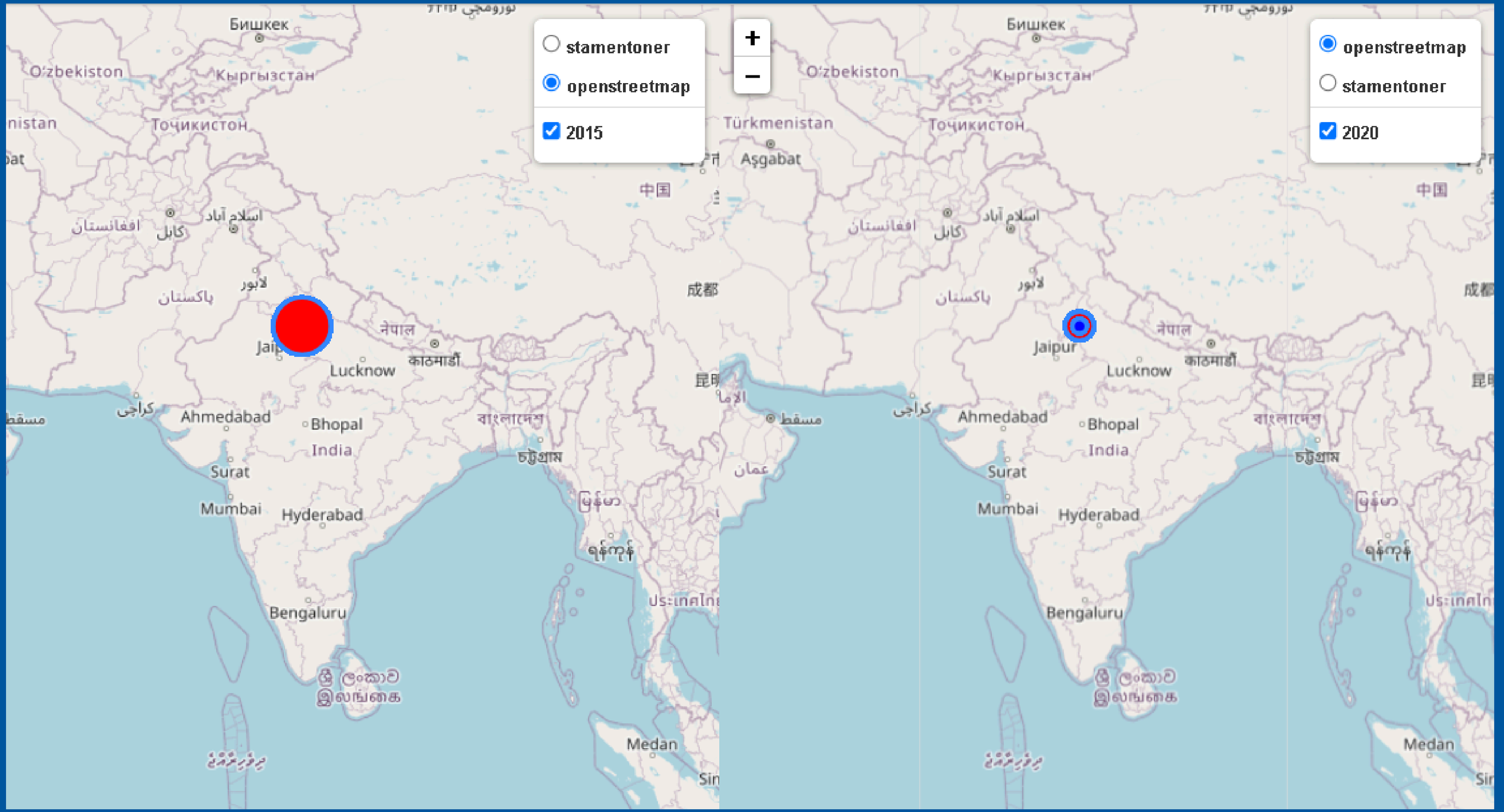


Gujarat consistently faces high AQI values due to urbanization and industrialization, while Delhi's AQI has fluctuated slightly since 2017, remaining relatively high compared to other states.

GEOGRAPHIC ANALYSIS

Geographic analysis is important in understanding the spatial patterns of air quality for 2015 and 2020.We created a dual map with two layers: one showing the average AQI data for the year 2015 and the other showing the average AQI data for the year 2020.

By comparing AQI data between 2015 and 2020, we can identify trends in air quality over time. This analysis helps us understand whether air quality has improved, deteriorated, or remained stable in different regions over the specified period.


stamentoner
openstreetmap
2015
+
−
Бишкек
O'zbekiston
Кыргызстан
Точикистон
Türkmenistan
Aşgabat
中国
افغانستان
کابل
اسلام آباد
پاکستان
لاہور
成都
नेपाल
काठमाडौं
Jaipur
Lucknow
昆明
الأما
مسقط
کراچی
Ahmedabad
Bhopal
India
বাংলাদেশ
চট্টগ্রাম
Surat
Mumbai
Hyderabad
عمان
ရန်ကုန်
Bengaluru
ශ්‍රී ලංකාව
இலங்கை
Medan
openstreetmap
stamentoner
2020

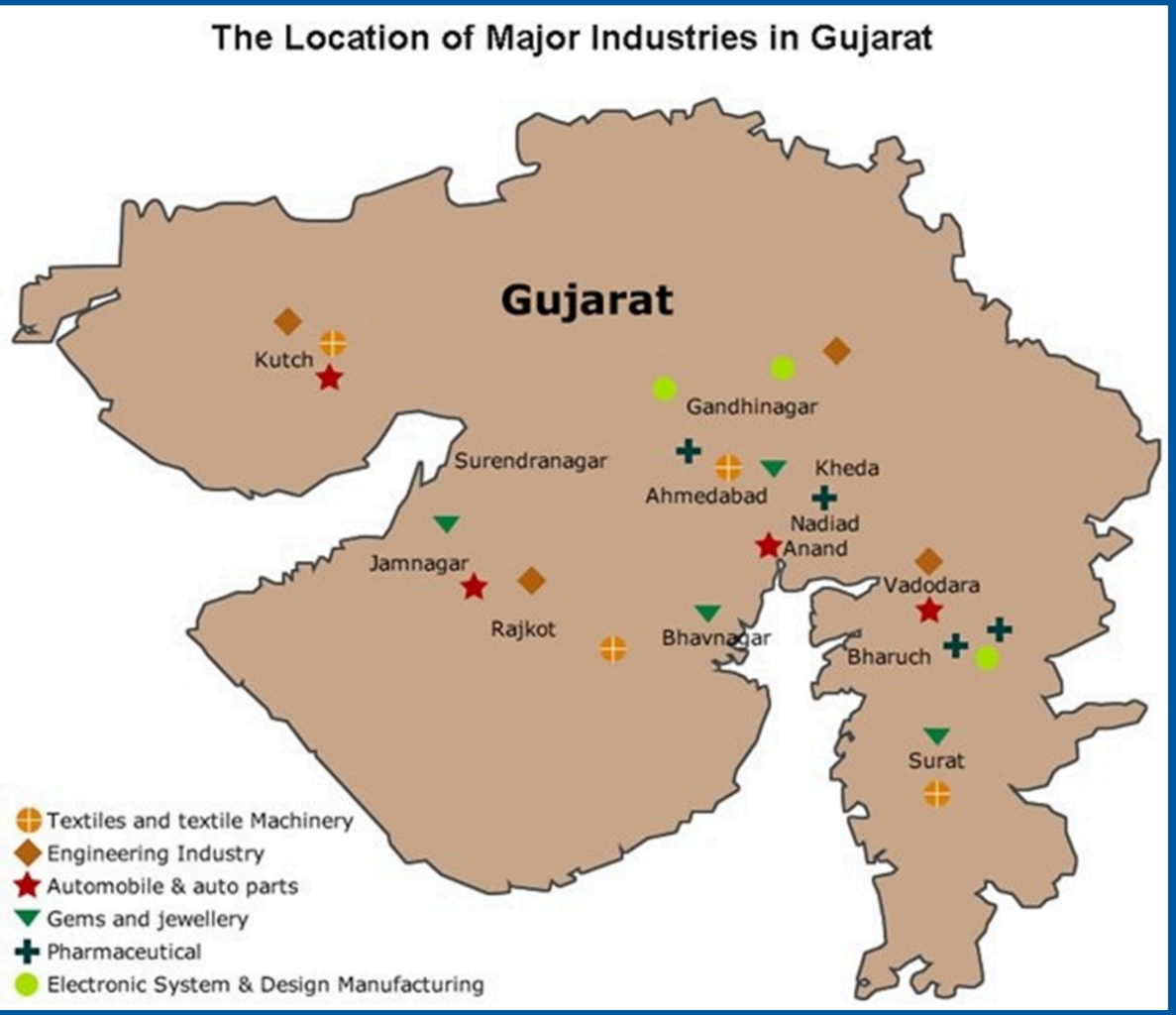
In the analysis:

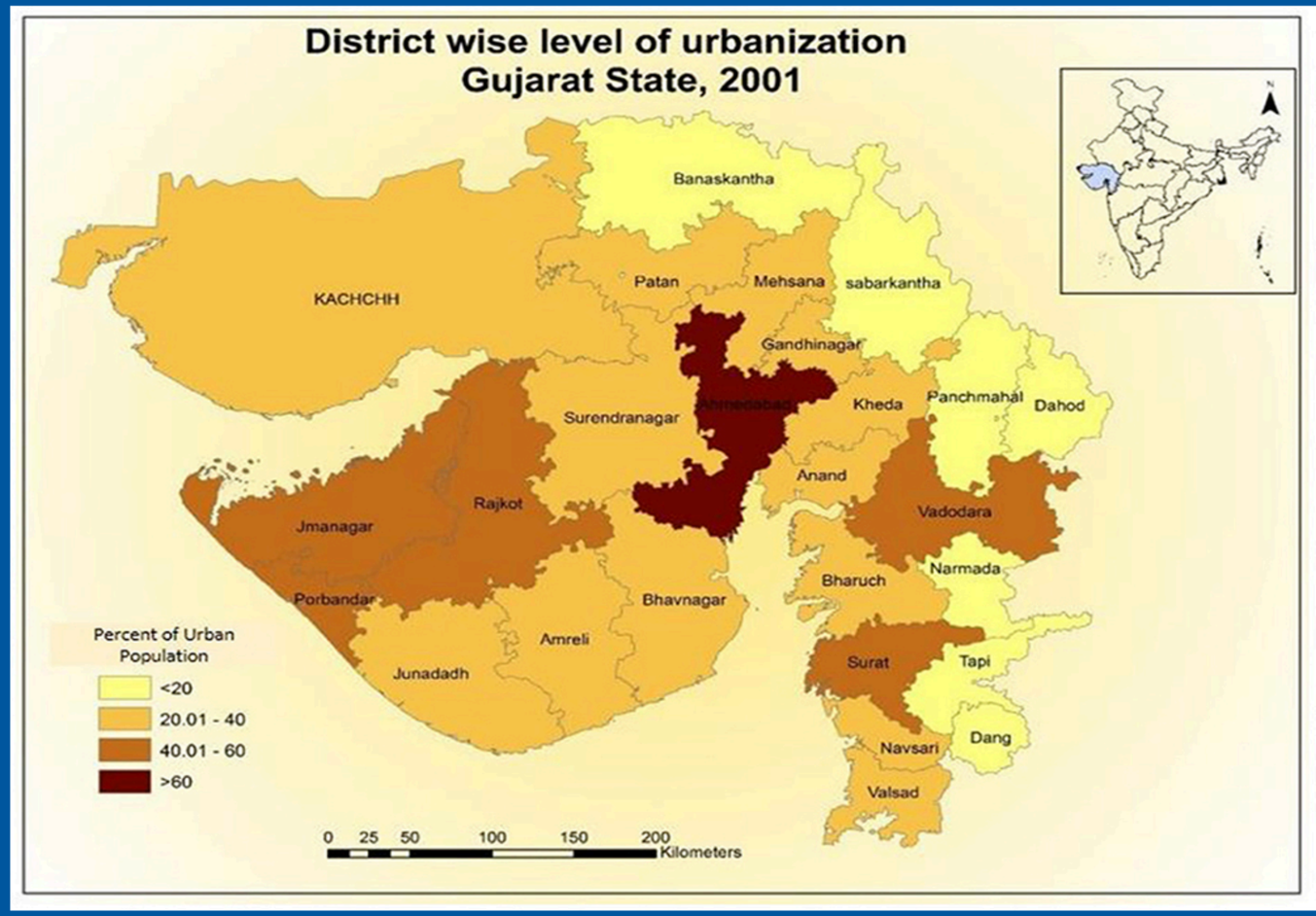
- Red Circle (larger radius): Indicates cities with a high average AQI value (> 230) for the respective year. The larger the circle, the higher the AQI value, representing poorer air quality.

- Blue Circle (smaller radius): Indicates cities with a moderate to low average AQI value (⩽ 230) for the respective year. The smaller the circle, the lower the AQI value, representing relatively better air quality.

- The average AQI value decreased from 2015 to 2020, indicating improved air quality over time likely due to increased awareness about this issue.

THE AVERAGE AQI TRENDS IN CITY IN INDIA DISTRIBUTION OF THE AQI_BUCKET



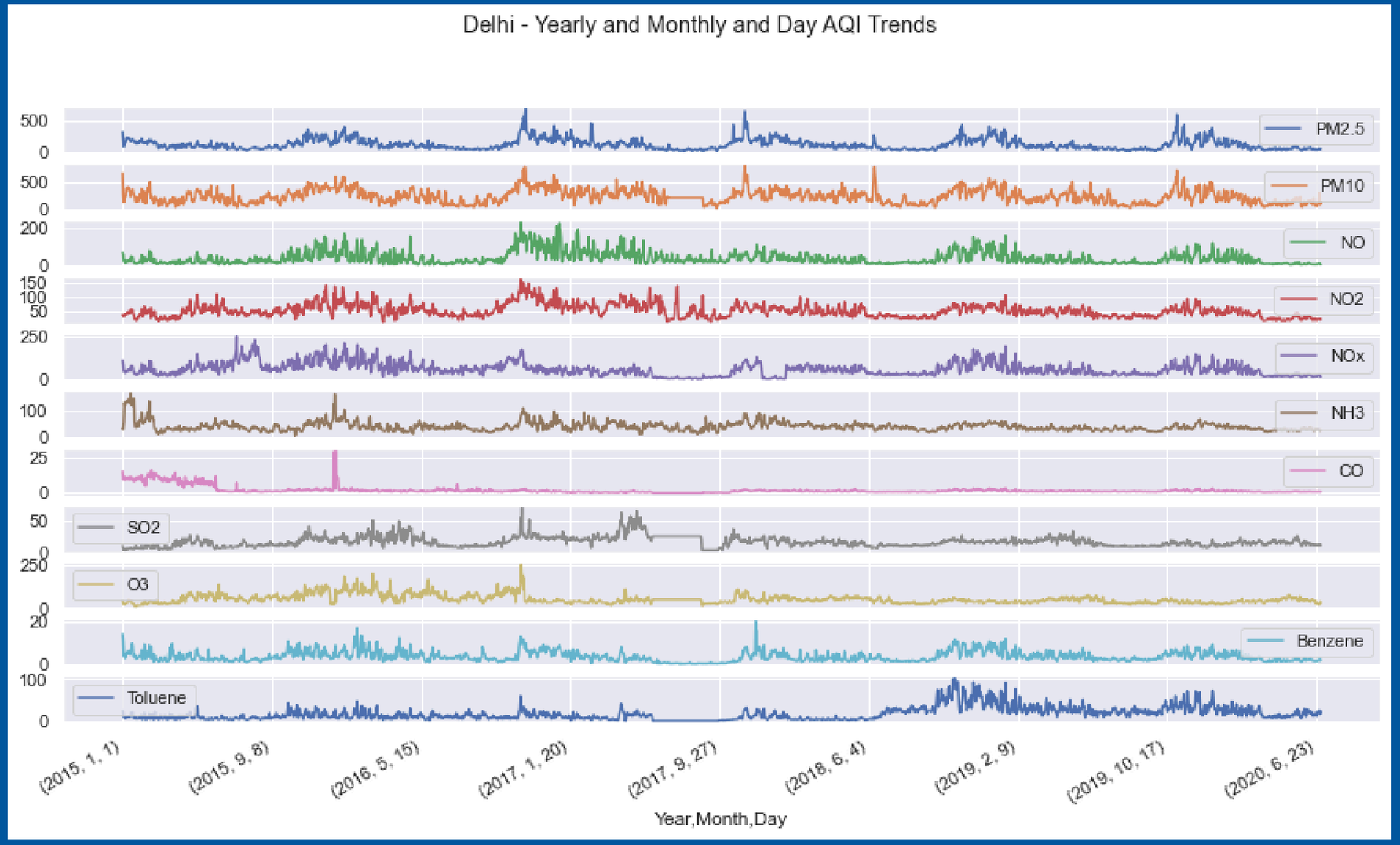


TIME SERIES ANALYSIS

ANALYSIS

Time series analysis is for identification of seasonality, trends and patterns in the data over time. It allows in predicting air quality levels based on past observations

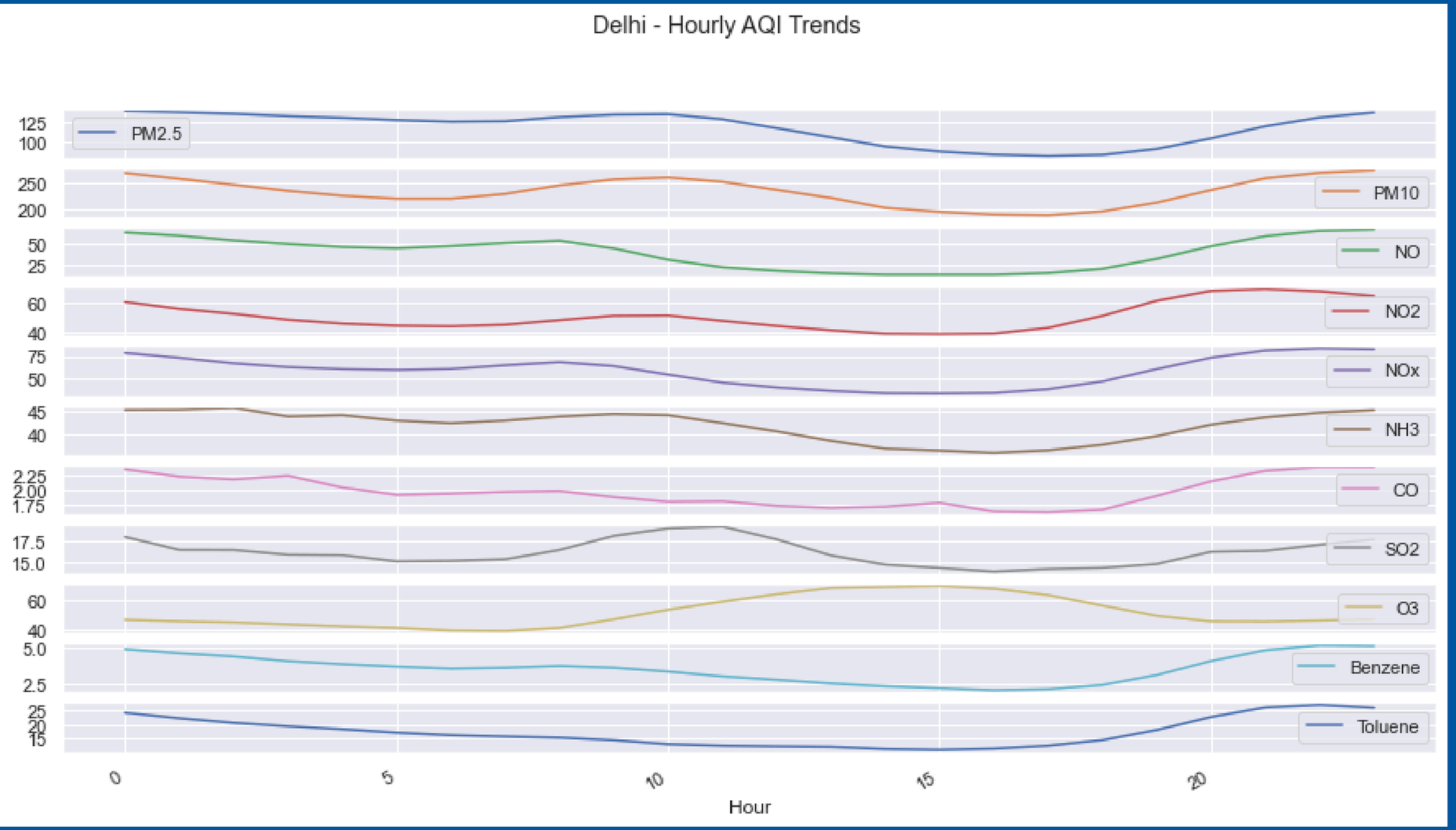
YEARLY AND MONTHLY VISUALIZATION



PM2.5, PM10, NO, NO2, NOx, NH3, CO, Benzene show a decreasing trend over time

SO2, O3, Toluene show an increasing trend over time

HOURLY VISUALIZATION



The average AQI for PM2.5, NH3, and SO2 decreases over time, suggesting a reduction in emissions or improved air quality management. Conversely, PM10, NO, NO2, NOx, O3, Benzene, and Toluene AQI show an increasing trend, likely due to factors like vehicular emissions, industrial activities, or combustion processes.

ARIMA MODEL SUMMARY FOR DELHI

he ARIMA (AutoRegressive Integrated Moving Average) model is a popular statistical method used for analyzing and forecasting time series data. It combines three components: AutoRegressive (AR), Integrated (I), and Moving Average (MA)

-RMSE is approximately 1.15

-MAE is approximately 0.78.

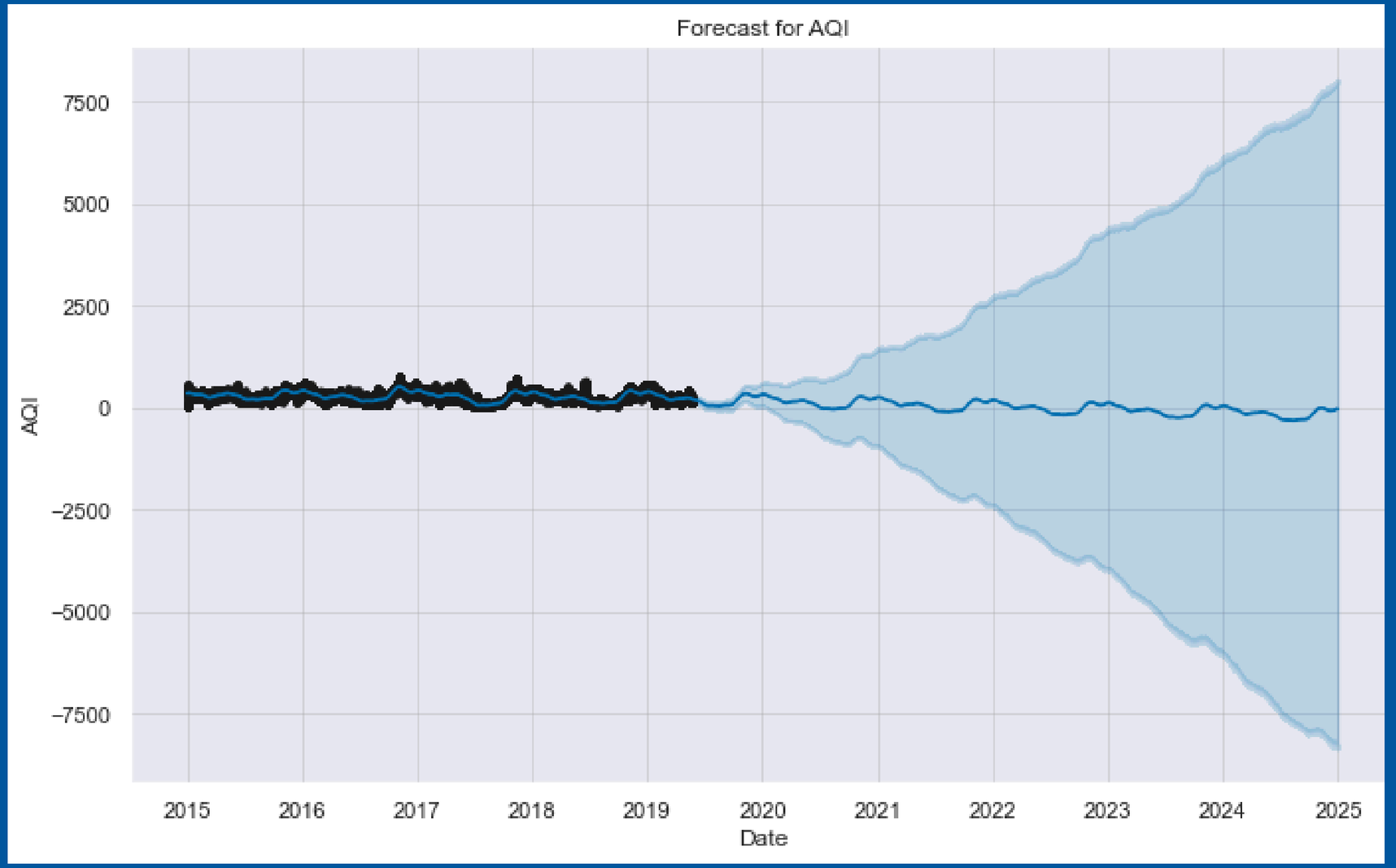
-R-squared is approximately 0.00012, indicating poor model fit

FACEBOOK PROPHET

Since both the ARIMA and Prophet models were exhibiting poor performance in forecasting the AQI data, there are several steps you can take to potentially improve the models, by Consider more sophisticated time series models that can capture complex patterns and dependencies in the data. Models like SARIMA (Seasonal ARIMA), VAR (Vector Autoregression), or LSTM (Long Short-Term Memory) neural networks. But due to time constraits we didn't get to address this issue

RMSE: 1.148983991805148
MAE: 0.7781811727784964
R-squared: 0

Predicting future Air Quality Trends for the years



We used Prophet to forecast until 2024, calculating the time from the last observation to the end of the year.

REGRESSION ANALYSIS

Linear Regression
(BASELINE MODEL)

Linear Regression Evaluation Metrics with Scaling:
RMSE: 69.41208480868359
MAE: 53.67148894662943
R-squared: 0.6766272364831947

These model suggest moderate performance. While the R-squared value indicates that approximately 67% of the variability in AQI values is explained by the model, the RMSE and MAE values could be further reduced for better accuracy.

Linear Regression
(BASELINE MODEL)

The predicted AQI values are generally higher than the actual AQI values. This suggests that the model tends to overestimate the AQI.

-There is some variation in the predictions, as seen from the spread of the points around the line of perfect prediction (the red dashed line).

-Some points fall close to the line, indicating accurate predictions, while others are farther away, indicating larger prediction errors

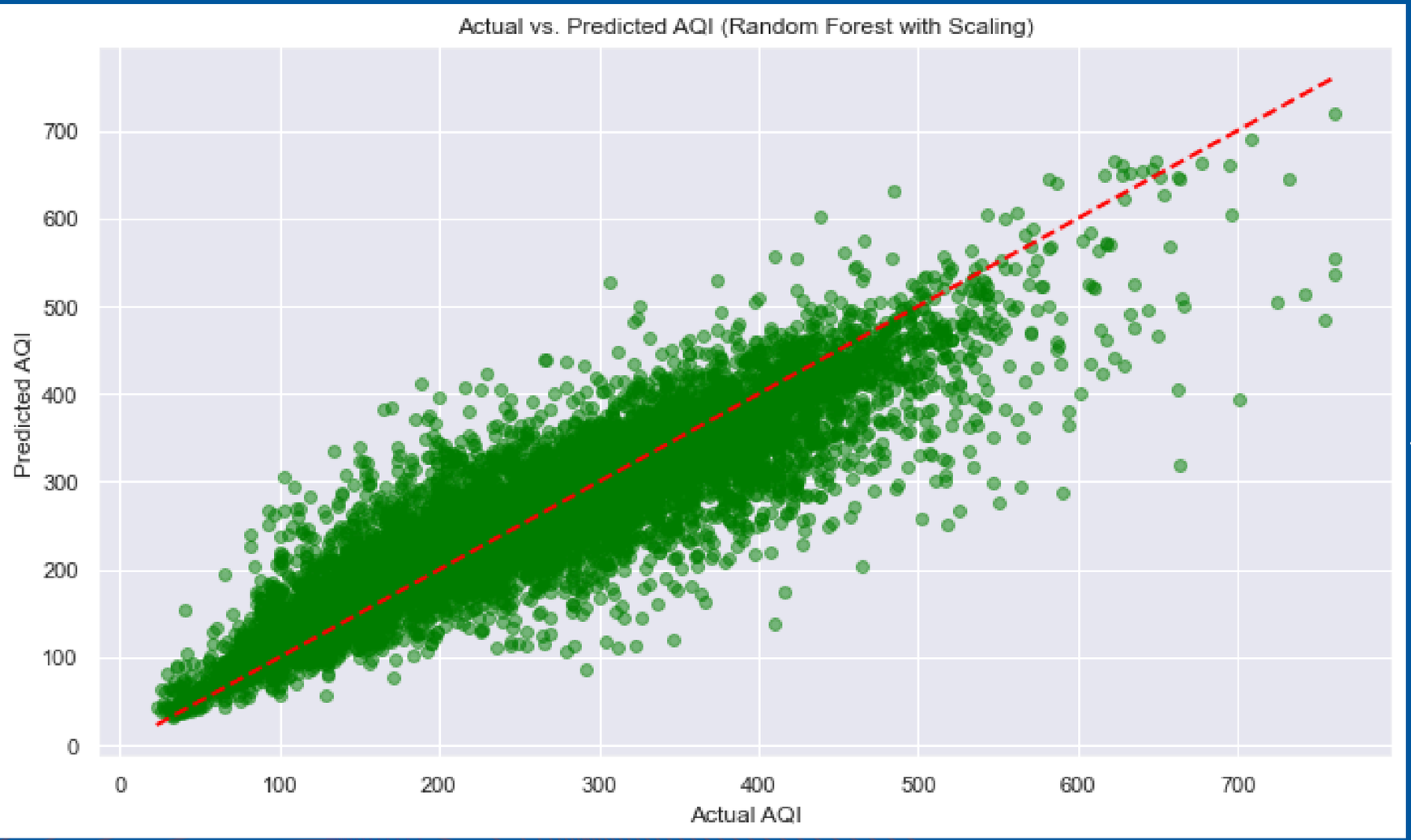
Random Forest Regressor

Random Forest Regression Evaluation Metrics with Scaling:
RMSE: 51.19426820187462
MAE: 36.791710357994575
R-squared: 0.8240959092610834

-These metrics suggest good performance. The higher R-squared value(82%) indicates a better fit of the model to the data.

-The lower RMSE and MAE values indicate less average deviation of predicted AQI values from the actual values.

Random Forest Regressor



-The predicted AQI values are generally closer and some are similar to the actual AQI values compared to Linear Regression. This suggests that Random Forest performs better in predicting AQI.

-There is less variation in the predictions compared to Linear Regression, as seen from the tighter clustering of points around the line of perfect prediction.

-Overall, the points are closer to the line, indicating smaller prediction errors.

Gradient Boosting Regressor

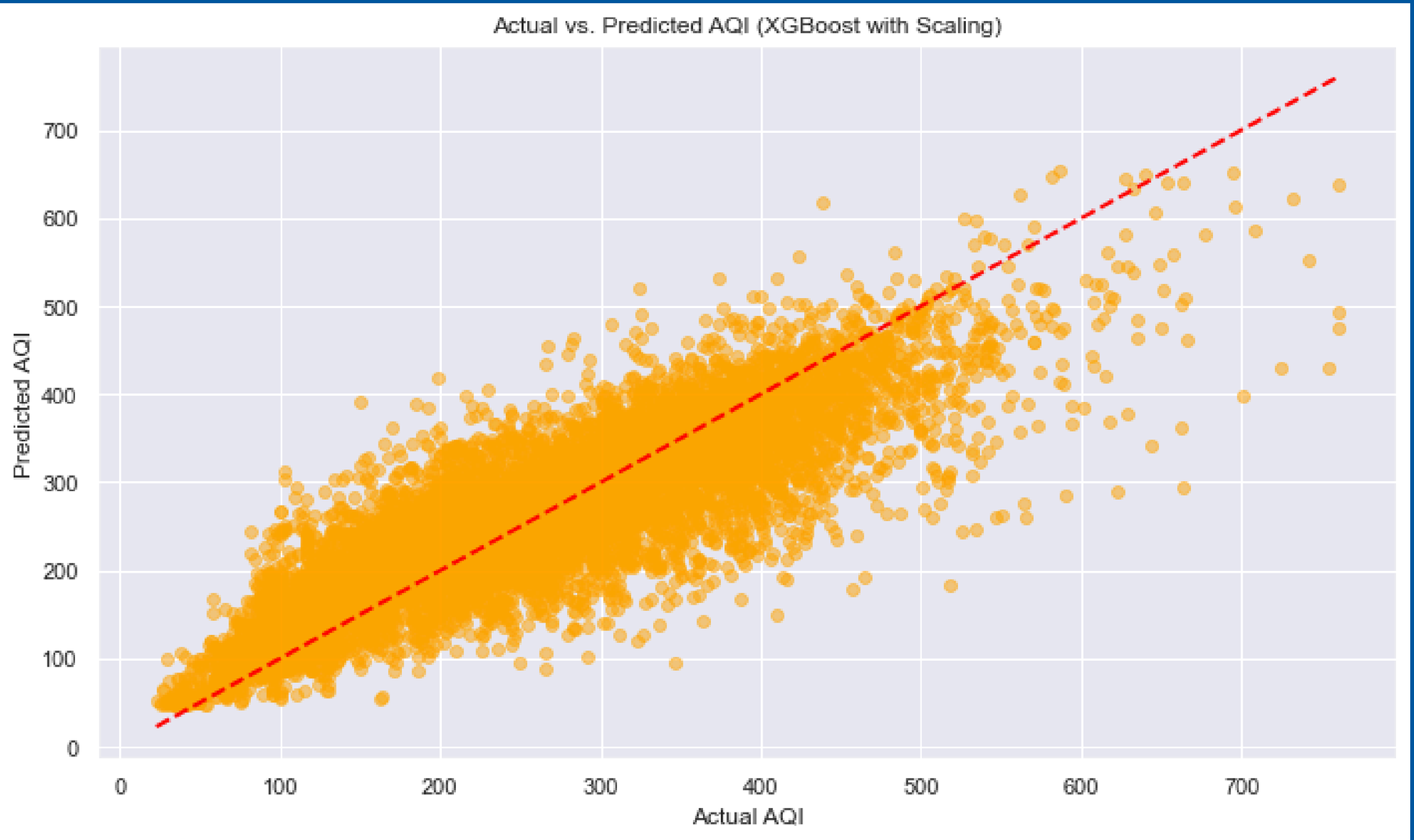
Gradient Boosting Regression Evaluation Metrics with Scaling:
RMSE: 60.3692639951799
MAE: 45.23178411615964
R-squared: 0.755395173159795 8

These metrics also suggest good performance, although slightly less than Random Forest.

The R-squared value is still relatively high(75%) but low compared to random forest

The RMSE and MAE values are higher compared to Random Forest, indicating slightly higher average deviation, but indicating a good fit.

Gradient Boosting Regressor

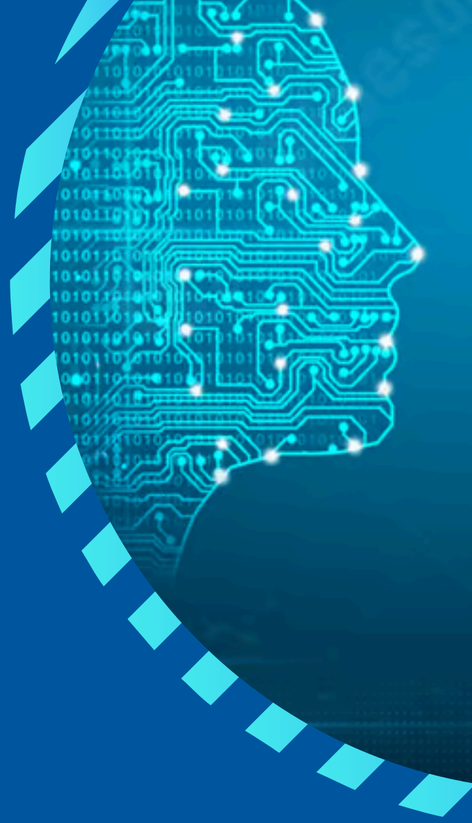


-The predicted AQI values are also generally closer to the actual AQI values compared to Linear Regression, but slightly less so compared to Random Forest.

-Similar to Random Forest, there is less variation in the predictions compared to Linear Regression, with the points clustering more tightly around the line of perfect prediction.

-Overall, the points are closer to the line, indicating smaller prediction errors, but there may be slightly more spread compared to Random Forest.

CONCLUSIONS



Pollutant Impact on AQI:

The regression analysis revealed the impact of various pollutants on the Air Quality Index (AQI). It was observed that pollutants such as CO, PM2.5, PM10, NO, NO2, and SO2 have significant positive correlations with AQI, indicating that higher concentrations of these pollutants contribute to poorer air quality.

On Time Series Analysis

Neither model effectively captured the underlying patterns and dynamics of AQI data, suggesting limitations in forecasting accuracy.

Despite the limitations, more advanced time series models like SARIMA, VAR, or LSTM could be explored to improve forecasting accuracy.

Due to time constraints, further model refinement was not conducted.

The Prophet model was used to forecast AQI trends until the end of 2024, providing insights into potential air quality trajectories beyond the analyzed period.

On Regression Analysis

Error Metrics:

Random Forest Regression shows the lowest RMSE and MAE, at 72.66 and 28.59 respectively. This implies superior prediction accuracy compared to other models. Gradient Boosting Regression follows with RMSE of 87.56 and MAE of 42.60, still better than Linear Regression which has RMSE of 103.71 and MAE of 56.92

01

Model Fit:

Random Forest Regression has the highest R-squared value of 0.79, explaining the largest proportion of the variance in the target variable. Gradient Boosting Regression follows with R-squared of 0.69, better than Linear Regression's R-squared of 0.57.

02

Recommendations

Given these results, Random Forest Regression is recommended for predicting the target variable due to its superior performance in error metrics and model fit. Monitoring its performance and exploring enhancements are advised for further optimization.

03

RECOMMENDATIONS

- Advanced forecasting techniques can significantly improve air quality management efficiency by providing accurate predictions, leading to timely interventions. Prophet's capabilities generate alerts for proactive measures, simplifying stakeholder decision-making.

- Further enhancing predictive performance is possible through model tuning, especially for Random Forest Regression and Gradient Boosting Regression. Techniques such as GridSearchCV or RandomizedSearchCV optimize hyperparameters, refining models for better results.

- Policymakers should prioritize air quality management and enact policies that address the root causes of air pollution effectively. This may involve incentivizing the adoption of clean technologies, imposing penalties for non-compliance with emission standards, and fostering international cooperation to tackle transboundary air pollution issues.

- Regression analysis aids policymakers in identifying key pollutants driving air quality decline, enabling targeted interventions. Accurate AQI prediction combined with regression insights helps healthcare providers anticipate respiratory issues, supporting proactive healthcare management.

- Feature importance analysis reveals pollutants with the most significant AQI impact, informing effective pollution mitigation strategies for policymakers and urban planners.

SUPER SAIYAN

Thank's For Watching

Connect with us.