DA 204o: **Data Science in Practice**

Course Project Final Presentation

India's Air

Analysis and Prediction of Air Quality in Indian cities

Ashutosh Mishra, *IISC, ashutoshmish@iisc.ac.in*
Neeraj K Shete, *IISC, neerajshete@iisc.ac.in*
Tony P Joy, *IISC, tonyjoy@iisc.ac.in*
Vivek H N, *IISC, vivekhn@iisc.ac.in*

Data Science Canvas

Project: India's Air
Analysis and Prediction of Air Quality in Indian cities

Team: Ashutosh, Neeraj, Tony, Vivek

Problem Statement

Business Case & Value Added

Analyze and understand the levels of major air pollutants to calculate and predict the resultant AQI(Air Quality Index) across the selected cities in India. This analysis can be beneficial for taking precautionary and remediation measures.

Data Landscape

We need data containing the measurement of major air pollutants taken from the selected cities. This data is available from Central Pollution Control Board, India.

Model Selection

- EDA – Visualization of AQI trends across cities and across seasons.
- EDA – Correlation of major air pollutants against AQI for different cities.
- EDA – Classification of major source of air pollution for different cities.
- Regression Modelling – To be able to predict the value of AQI for a given day and city.
- Timeseries Forecasting – To be able to forecast AQI values for future dates.
- Prediction of live input data - Simulation

Model Requirements

To obtain a valid model for air quality index (AQI) prediction based on features, several requirements and considerations must be met:

- Data Quality. Consistency and accuracy
- Feature selection and transformation
- Data Preprocessing,
- Model selection

Software & Libraries

- Pandas,
- matplotlib and Seaborn for EDA
- Scikit-learn packages, implemented in Python for modelling
- Streamlit for GUI visualization
- PyTorch LSTM

Skills

Few of the essential skills required across various domains are:

- Data Handling and management
- Feature engineering
- Statistical analysis
- Machine learning and modelling
- Programming and Tools frameworks
- Domain know how

Execution & Evaluation

Model Evaluation

Key Indicators Requiring Quality Control

- RMSE, MAE, and R^2 for model accuracy.
- Correlation coefficients and feature importance for pollutants and AQI relationships.
- Trend line fit and seasonal analysis for AQI patterns.

- Real-time monitoring is essential if the system is designed for citizen alerts and real-time AQI predictions. Quality control over data feeds and model retraining will ensure the system remains reliable in real-world conditions.
- Handling of Real-time data and forecasting to be considered as stretch goal.

Data Storytelling

Requirements:

- Technical Understanding:
- Real-World Application:
- Hands-on Insights:
- Data Visualization:
- Scalability and Optimization:

Communication approach:

- Clear Breakdown of Analytical Process (Data collection, data cleaning and preparation and feature engineering)
- Rationale behind model selection
- Interactive and visual elements
- Code and practical application (Code walkthrough and live demo
- Real world impact

Data Collection & Preparation

Data Selection & Cleansing

Air pollution concentration data (PMx, NHx, COx, SOx, O3)
AQI, Location data, Temporal data, Weather data

Yes, the data needs to be cleaned up, Missing values needs to be addressed through interpolation, date and time should be in correct formats. Few columns need to be dropped

Data Collection

Using public dataset like the Central Control Room for Air Quality Management data and other government databases (like CPCB, State Pollution Control Boards) and data should have relevant pollutant data and should be consistent.
CSV merger was used to merge all city level pollutants dataset

Data Integration

Data needs to be migrated into shared google drive as we are working with Google Colab.

Explorative Data Analysis

- Study of major pollutants per city
- Study of Environmental Parameters per city
- Study of City wise AQI Values
- Impact on Environmental parameters on AQI
- Study of Vehicular and Industrial Pollution

Adopted from: https://github.com/tomalytics/datasciencecanvas

Data Science Workflow

1. Problem Definition
2. Data collection and preparation
3. Data Exploration
4. Data Modeling
5. Reflection or Inference Phase
6. Communicating & Visualizing Results

Problem Definition

Objective	Build ML models that can Predict and Forecast the AQI value based on the measured air pollutant parameters.
Deployment Model	The models can be deployed as a standalone dashboard or can be integrated with existing weather forecasting systems.
Data Science Category	This is a classical supervised regression problem. Can be used with both offline and online data input.
KPI Benchmarks (Best Model)	Validation Accuracy KPI: 86.44 % Test Accuracy KPI: 83.01 %
Stretch Goal	Time Series Forecasting Prediction for live input data.
Manual Approach	Mean of historical data for similar hour and city could be used as a prediction measure.
Assumptions	Model runtime is not a huge concern. Data taken for training is accurate. There will be authorized methods to subscribe for online data.

Data Collection and Preparation

Data Collection

Government Data – (“Central Control Room for Air Quality Management - All India”)

Link: CCR (cpcb.gov.in)

Open and Free to use.

Well Maintained along with sufficient historical data.

Data Retrieval and Storage

Shortlisted 6 cities to focus on – Bengaluru, Hyderabad, Chennai, Mumbai, Kolkata and Delhi.

Collecting yearly dataset for each of these cities for the years 2019 through 2023 (hourly data).

~8k samples for a city per year.

Dedicated Train [2019 - 22] and Test [2023] Datasets.

Combined datasets to form train and test master datasets of all cities.

Stored in shared google drive.

Data Cleaning

Set threshold of 10% and delete columns based on missing value threshold

Imputed missing values with linear interpolation as we have time series data.

Convert Timestamp column to datetime format.

Feature Engineering

Rolling average computation using sliding window

Sub Index Calculation

AQI and AQI Class calculation

Vehicular and Industrial pollution calculation.

Data Exploration

Study of major pollutants per city

- Average levels
- Change over the years
- Monthly Averages
- Seasonal Trends

Study of Environmental Parameters per city

- Monthly Averages
- Seasonal Trends.

Study of City wise AQI Values

- Monthly AQI Distribution
- City Wise AQI Distribution

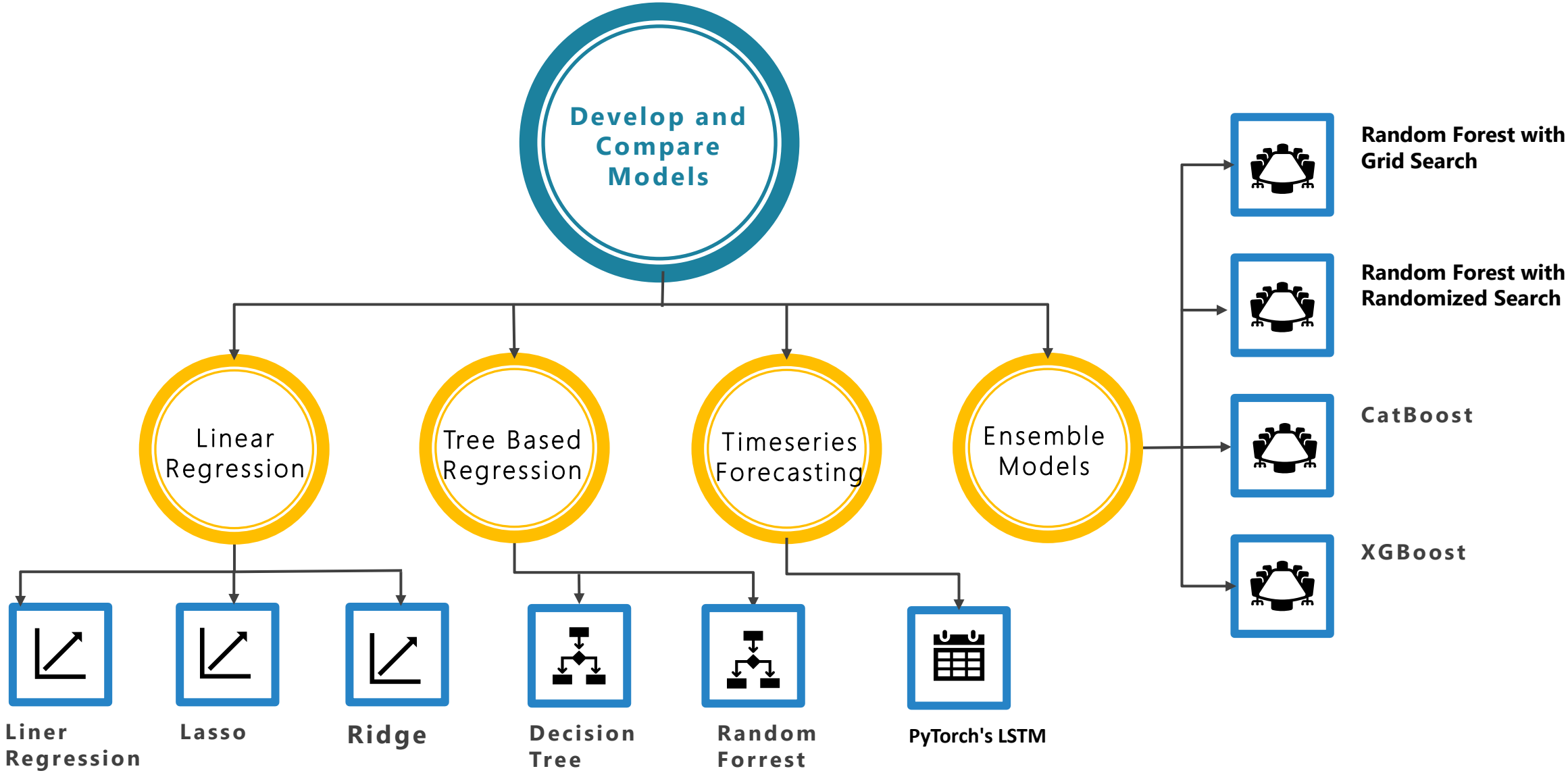
Impact on Environmental parameters on AQI

- Correlation Analysis
- AQI Class Distribution

Study of Vehicular and Industrial Pollution

- City wise distribution
- Monthly and Yearly Average Trends

Model Development



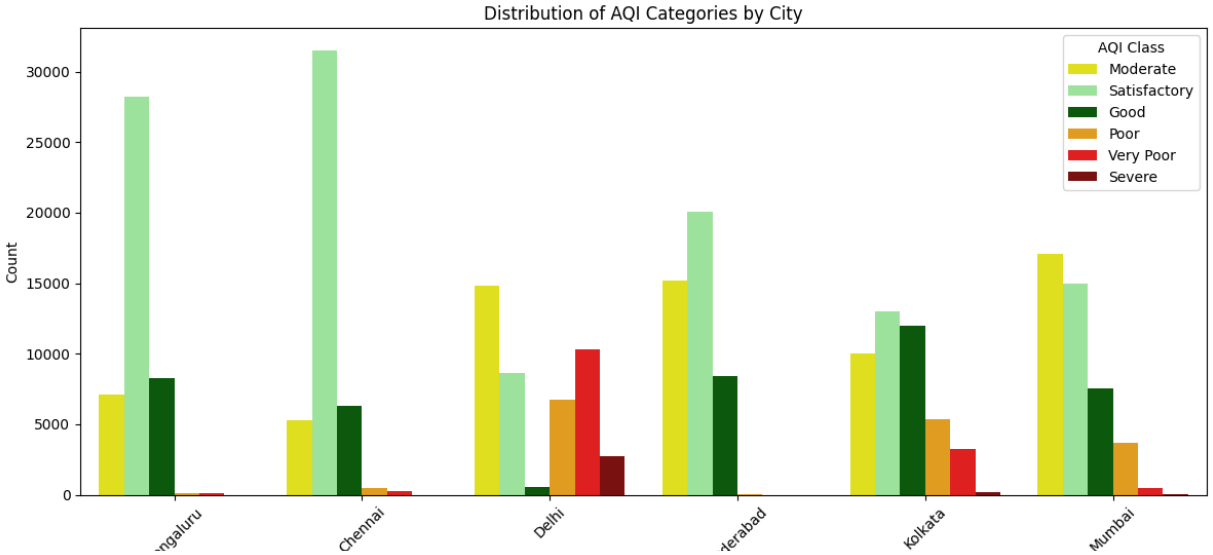
Reflection and Inference : Data Preparation and EDA

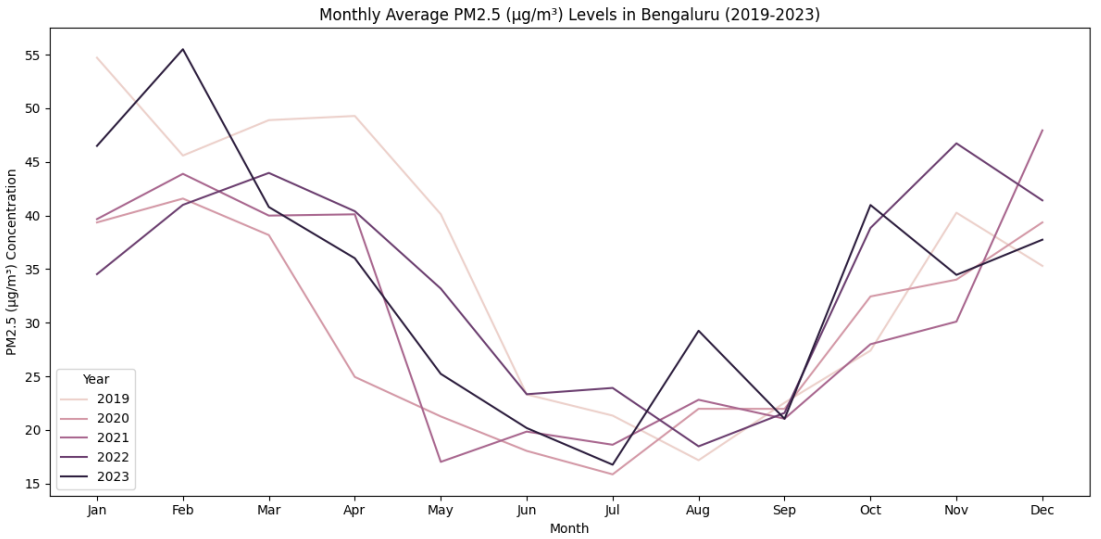
Data Preparation and EDA metrics	Success Criteria	Results Shown
Data Cleaning and Preprocessing	Achieve 90%+ clean, usable data across selected cities and pollutants	Successfully cleaned the dataset with 0% missing values.
Distribution of Pollutants Across Cities	Understand the distribution of major pollutants for comparison across cities.	Observed flatter distributions in Delhi for PM pollutants; lower values in Bengaluru and Chennai.
Average Value of Major Pollutants for Each City	Visualize and compare average pollutant values across cities.	Found airborne PM as the dominant pollutant; Delhi and Mumbai had the highest averages for most pollutants.
Pollutant Trends (2019-2024)	Compare pollutant trends over the years and identify long-term patterns.	Chennai showed sustained control; Mumbai exhibited upward trends in pollutants; Delhi remained consistently high.
Impact of Environmental Parameters on AQI	Identify relationships between AQI and environmental parameters like temperature, humidity, and wind.	Poor AQI linked to lower temperatures and higher humidity; higher wind speeds correlated with reduced pollution.
Monthly Pollutant Levels (Seasonal Trends)	Identify seasonal changes and peak months for pollutants.	PM pollutants peaked in winter; NH3 and SO2 levels were highest during summer; monsoon reduced pollution.
AQI Trends Across Cities	Assess monthly AQI trends and identify best and worst months for air quality.	Delhi worsened in winter; June to September showed healthy AQI in most cities except Delhi.
Industrial vs. Vehicular Pollution	Determine dominant pollution sources for each city and their trends.	Industrial pollution exceeded vehicular in all cities; Delhi showed consistently high vehicular pollution levels.
City-Specific Insights	Highlight unique trends for each city.	Chennai and Bengaluru maintained better air quality; Delhi and Mumbai had higher pollutant levels and poor AQI

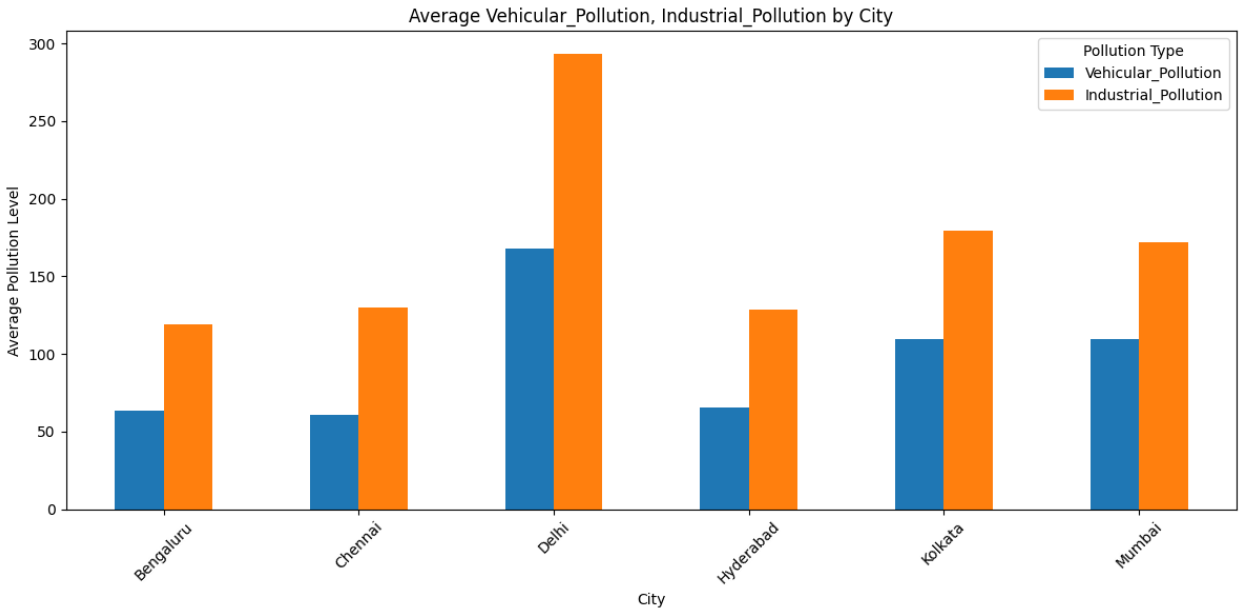
Reflection and Inference : Data Modelling

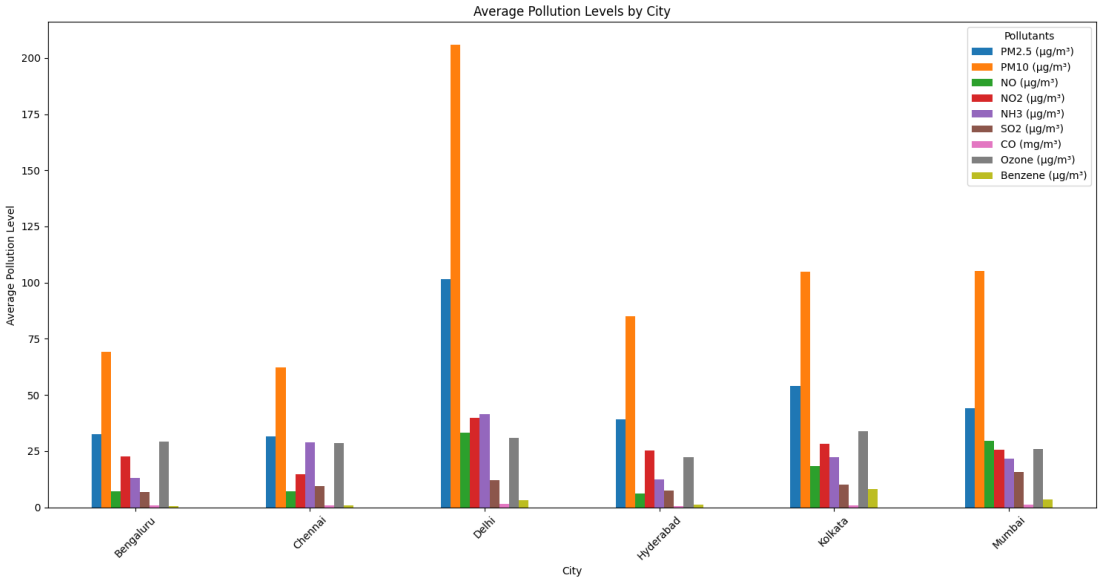
Models	RMSE		Training Accuracy		Testing Accuracy		Validation Accuracy	
	Achieved	Target	Achieved(%)	Target(%)	Achieved(%)	Target(%)	Achieved(%)	Target(%)
Linear Regression	39.07	35	79.81	80	80.53	75	79.51	75
Lasso Regression	39.08	35	79.80	80	80.58	75	79.50	75
Ridge Regression	39.06	35	79.81	90	80.53	75	79.51	75
Decision Tree Regressor	44.94	35	100	90	67.25	70	72.63	70
Random Forest with Grid Search	31.48	35	98.15	90	83.01	80	86.44	85
Random Forest with Randomised search	31.48	35	98.15	90	83.01	80	86.44	85
Catboost	32.1941	35	87.90	90	83.58	80	86.19	85
XGboost	31.79	35	93.02	90	83.01	80	86.53	85

Visualizing Results: EDA

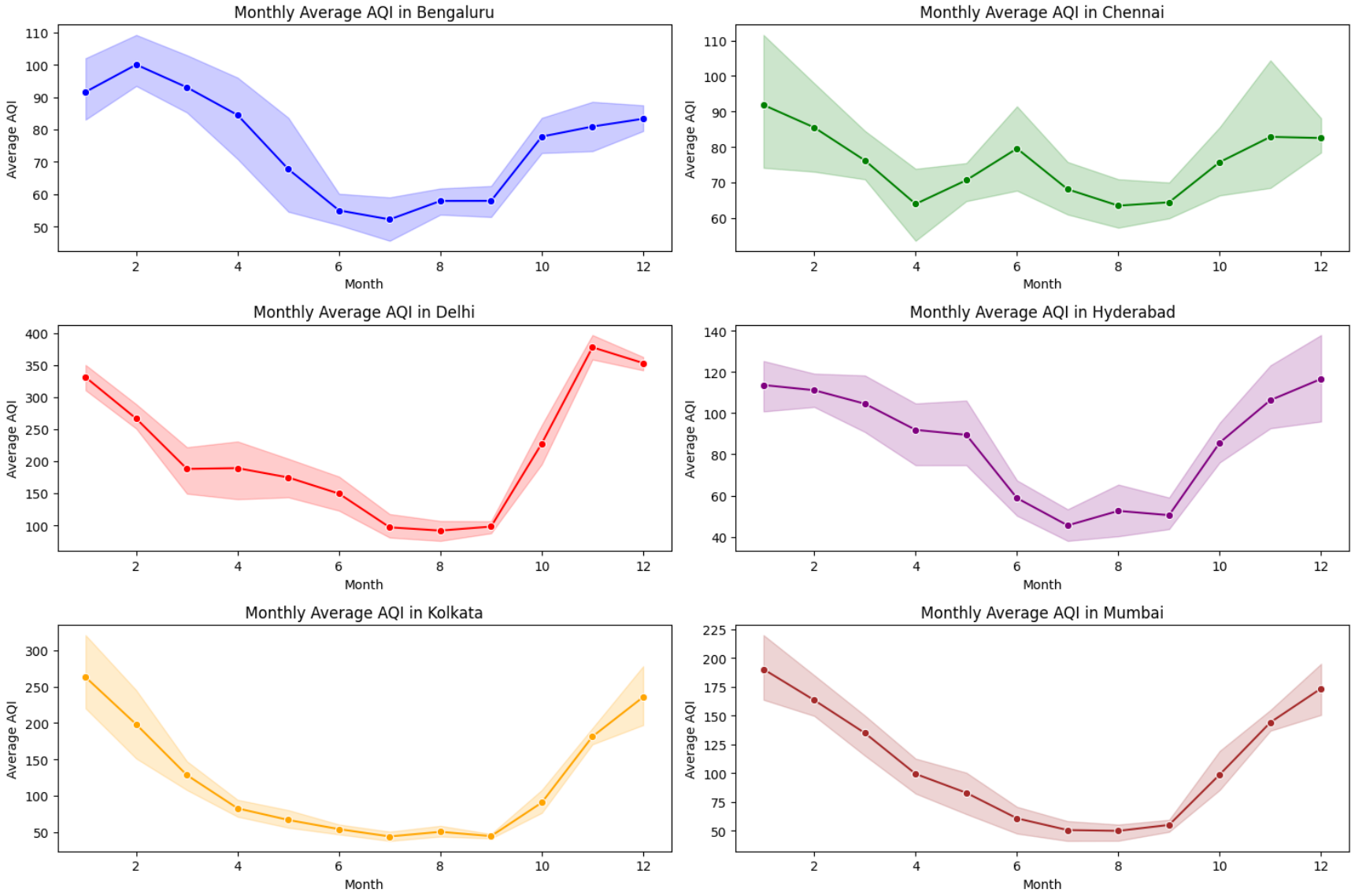




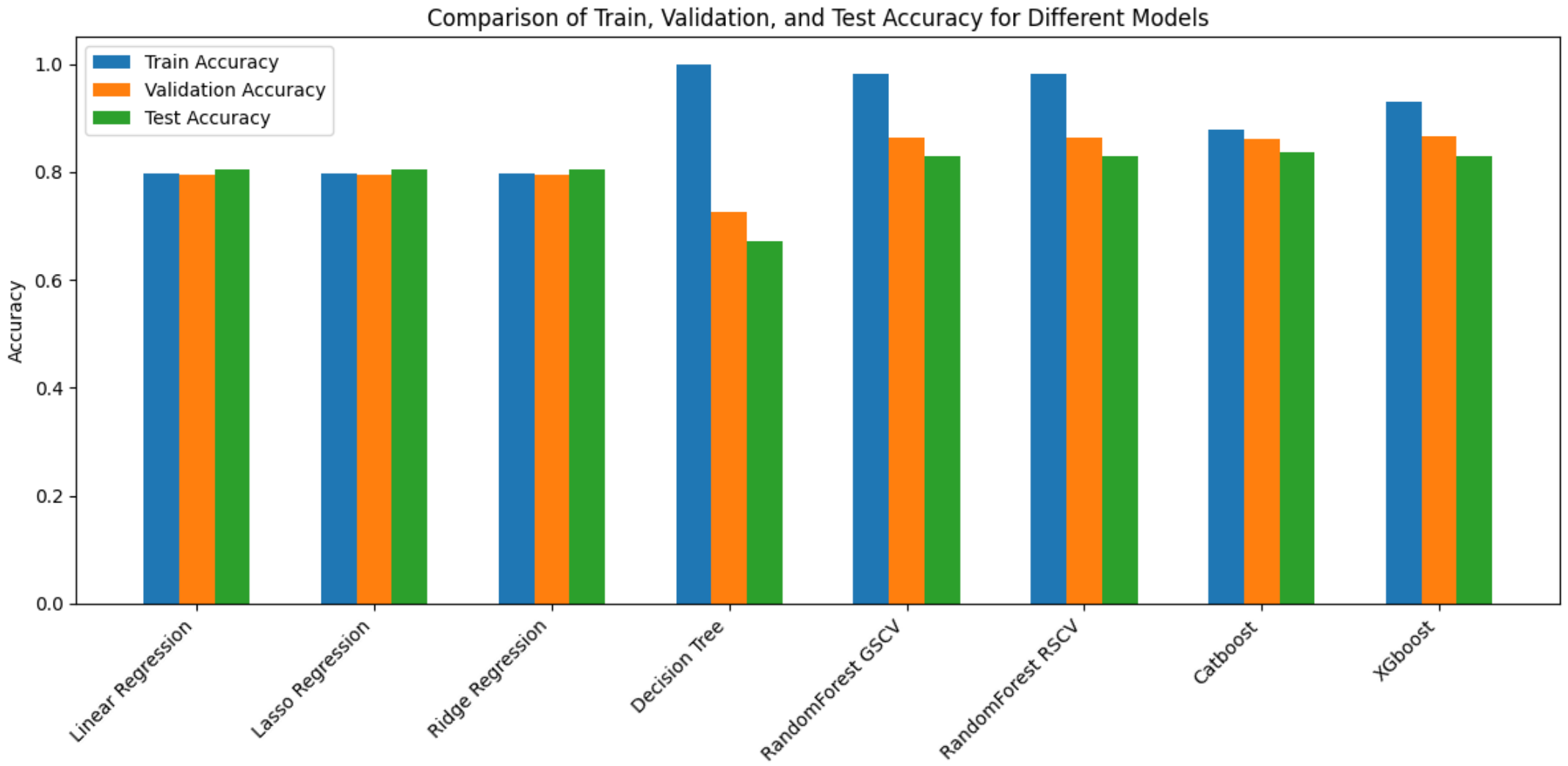




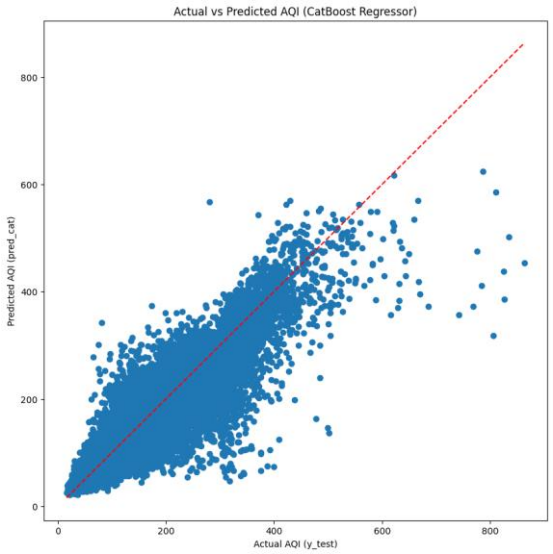
Visualizing Results: EDA

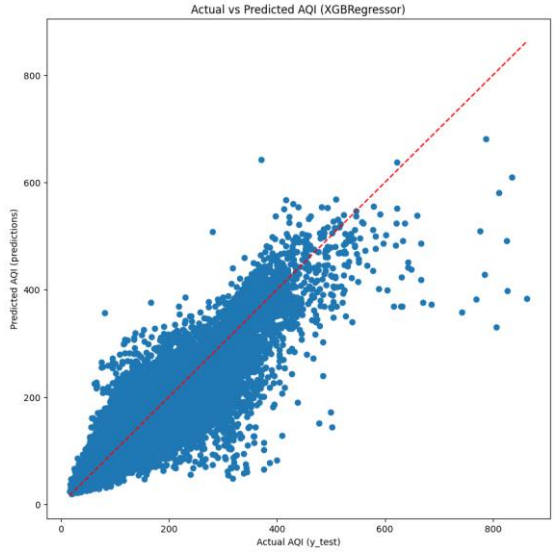


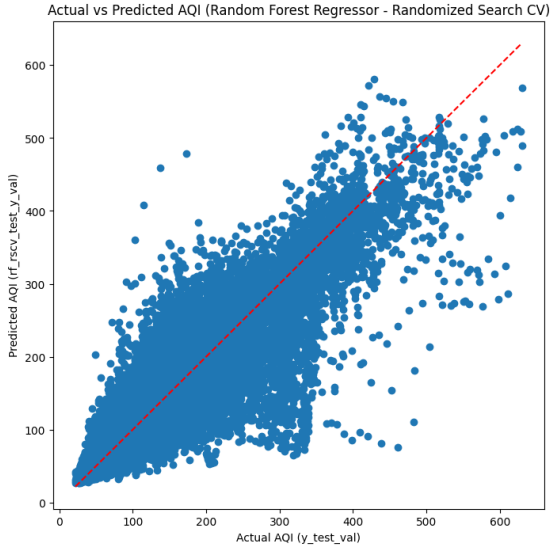
Visualizing Results: Model

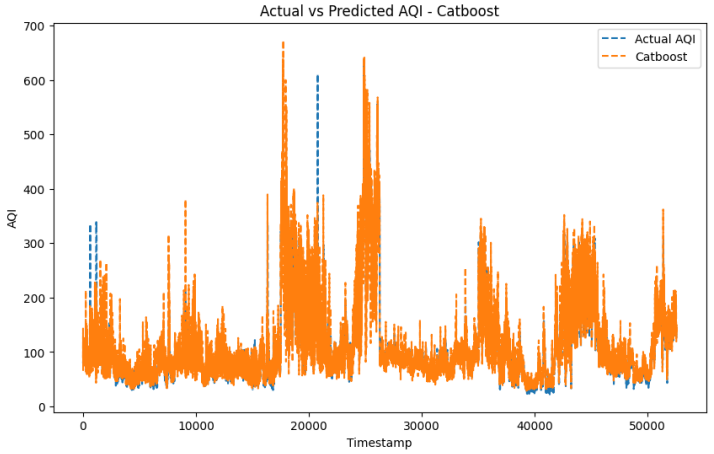


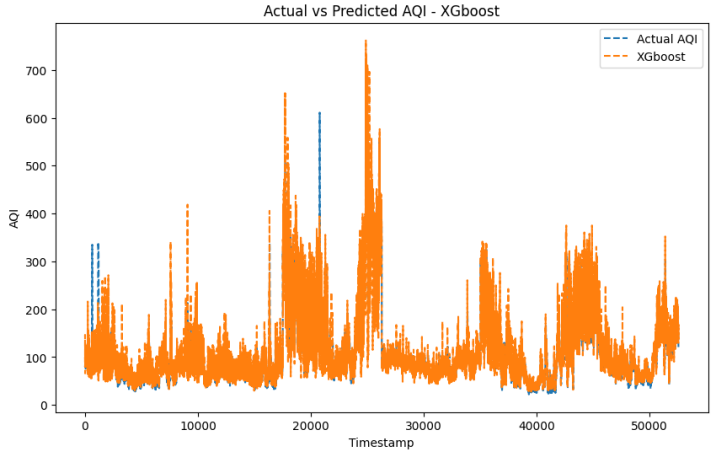
Visualizing Results: Model

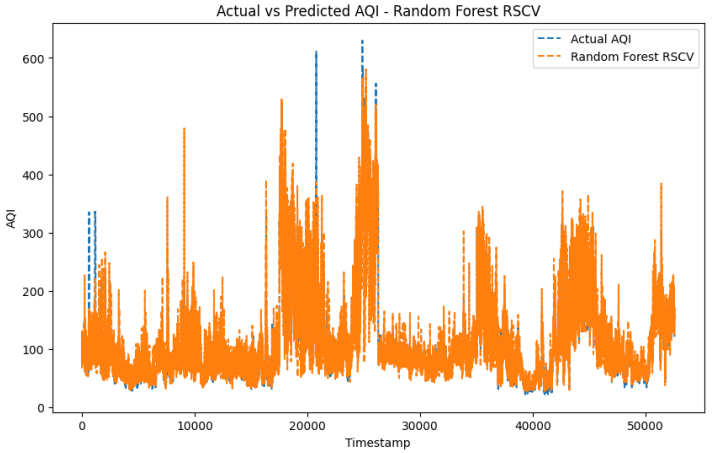




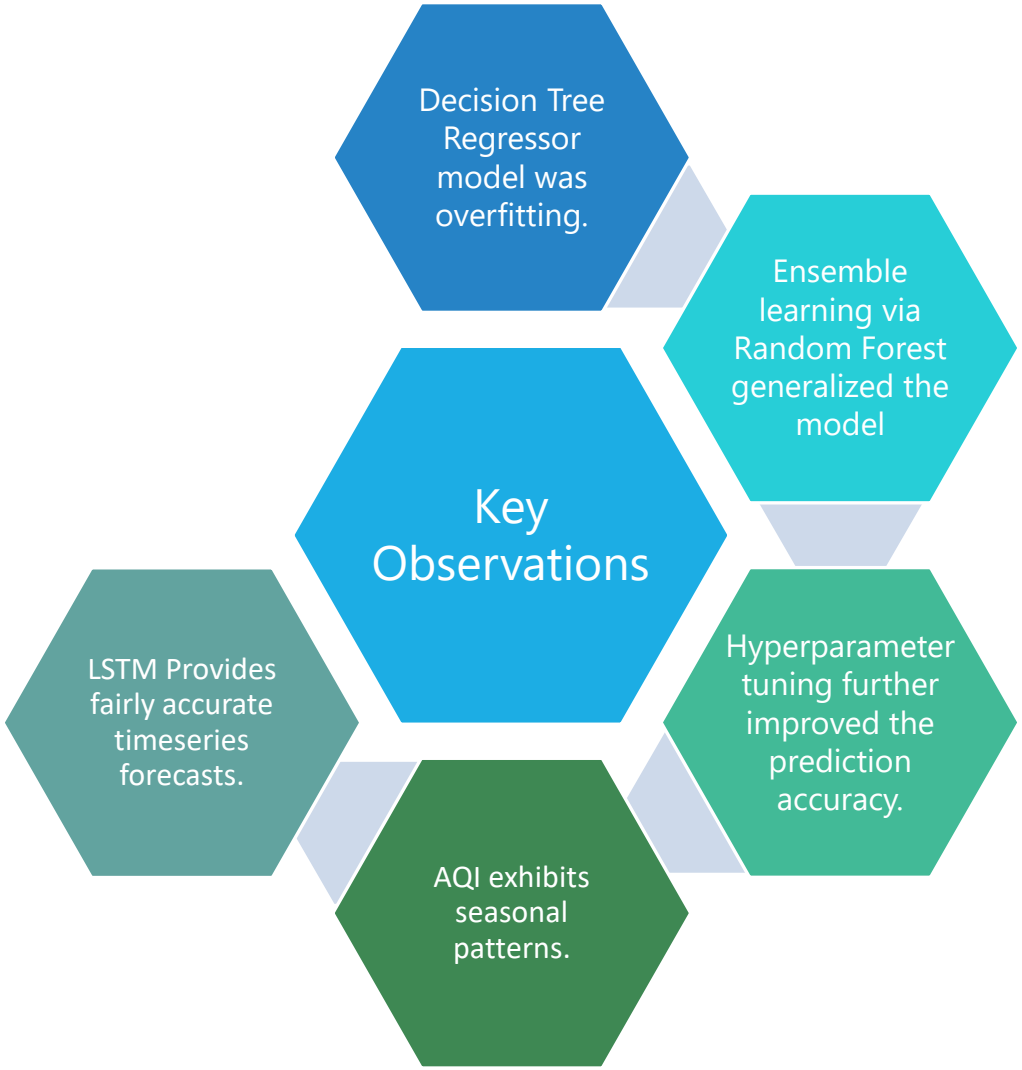




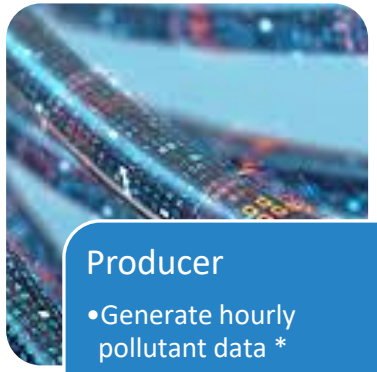




Key observations

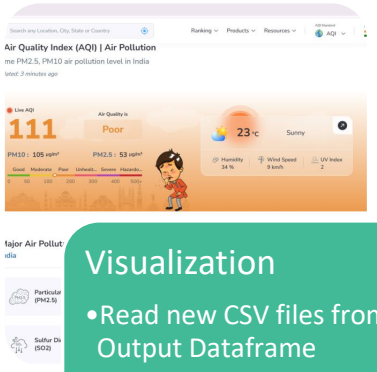


Deployment for Demo



Shared Input Store

Output Dataframe



Producer

- Generate hourly pollutant data *
- Generate consolidated CSV for all cities.
- Write consolidated hourly CSV to Input Store

* Hourly Samples from

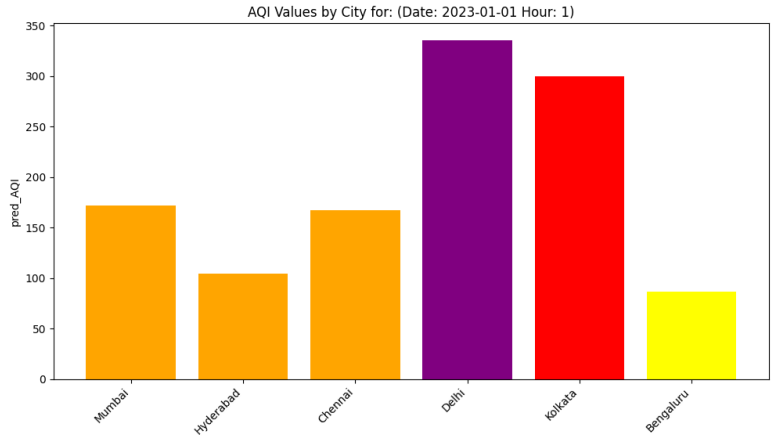
Live Prediction

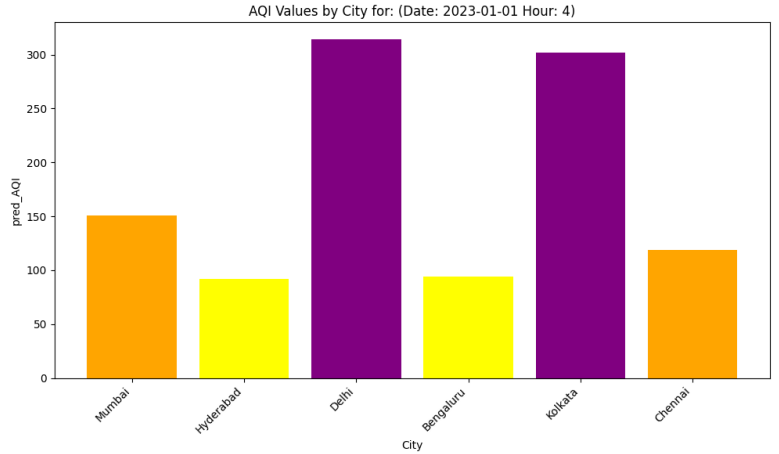
- Read new CSV files from Input Store
- Convert to Dataframe.
- Predict AQI with saved model.
- Generate Output Dataframe.

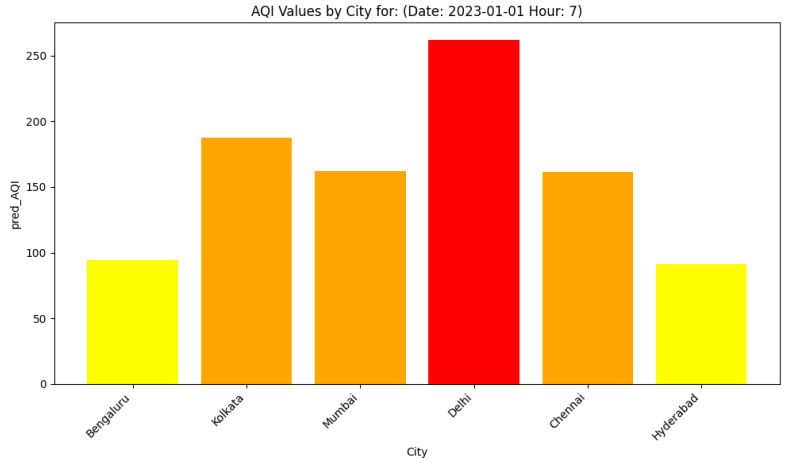
Visualization

- Read new CSV files from Output Dataframe
- Visualize the Predicted AQI for each city *

* New bar plots



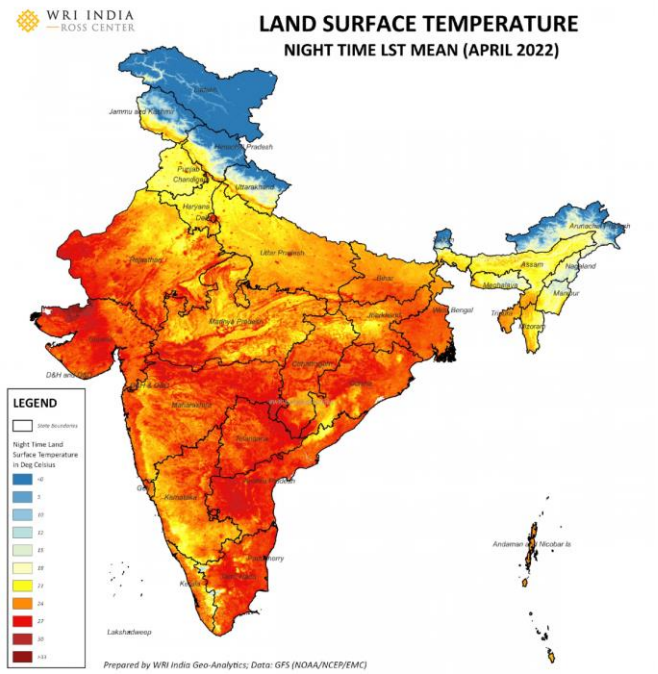


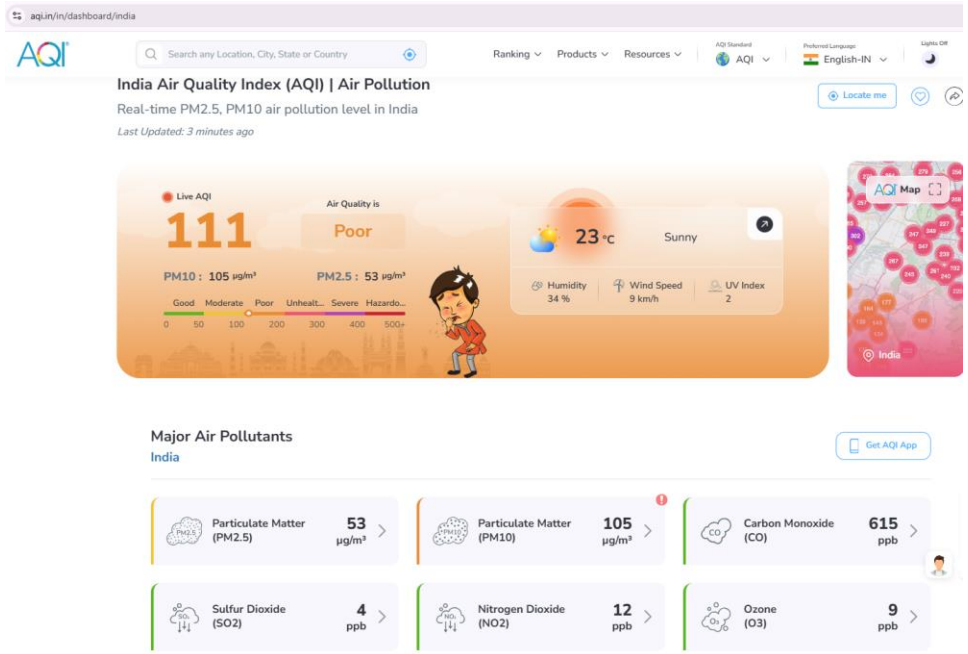


Enhancements/Future work

City wise MODELS

Continuous Training and Integration





LIVE DASHBOARD

Forecasts, Prediction and Live Measurements

Thank You !!