TRAIN - IT HACKATHON 2025

ImpactX: Exploratory Data Analysis and Predictive Modeling of AQI Trends in India

Presented by : Team 2023ugee059

Asmi Srivastava

Sachin Mishra

Aameya Devansh

INTRODUCTION

Why AQI Matters?

- **What is AQI? -** An air quality index (AQI) is an indicator developed by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become.

- **Why analyze AQI trends? -** It is essential to keep the public aware of impending health risks. AQI analysis will help families with children, the elderly and individuals with respiratory or cardiovascular problems prepare in time. Predictions of high AQI will help the government issue public alerts, encouraging people to stay inside.

- **Objectives of this analysis -**

1. Understanding AQI patterns across cities
2. Finding correlations between AQI, meteorological conditions (rainfall, temperature)
3. Predicting future AQI trends

Data Sources & Preprocessing

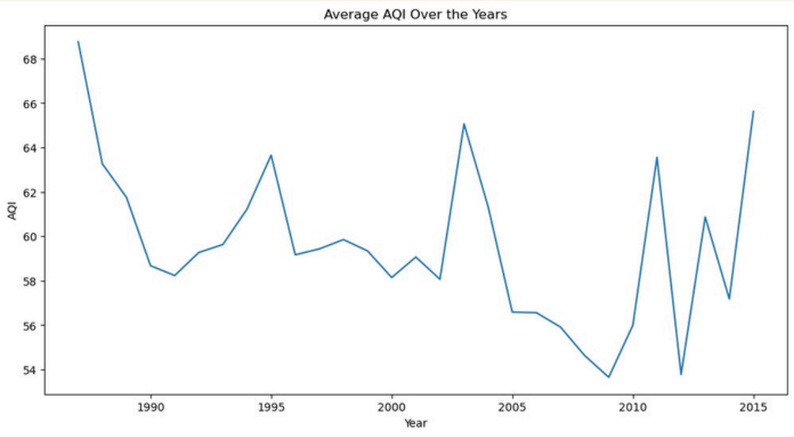
Datasets Used:

- Air Quality Data in India – Central Pollution Control Board (CPCB)
- India Rainfall Data (1901-Present) – IMD Climate Data

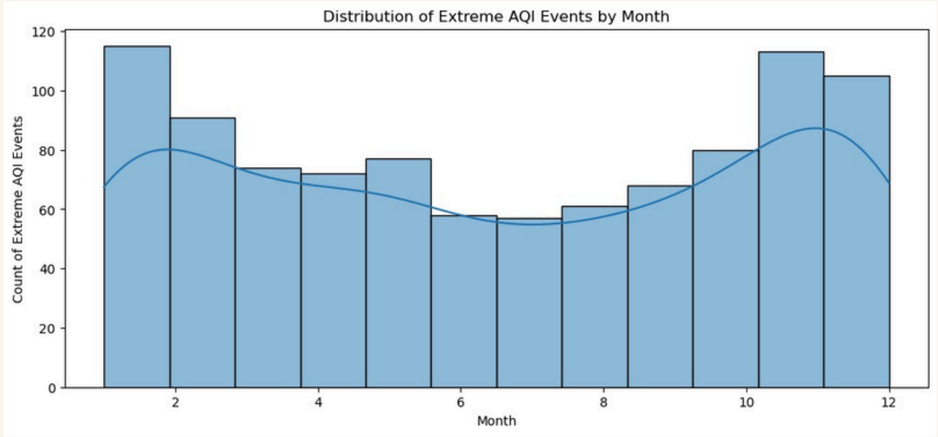
Preprocessing Steps:

- Handling Missing Values (filled using mean, median, etc.)
- Removing Duplicates
- Standardizing date-time formats
- Merging datasets for extended analysis
- Calculating AQI using CPCB formula

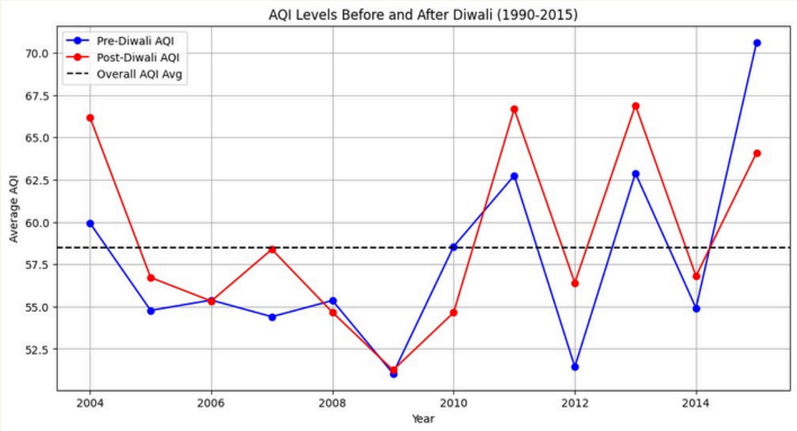
AQI Trend Analysis



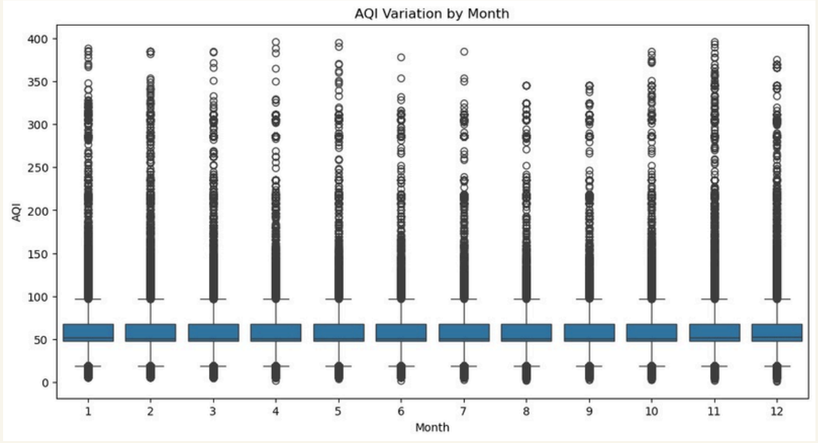
The graph displays how average AQI has changed over time



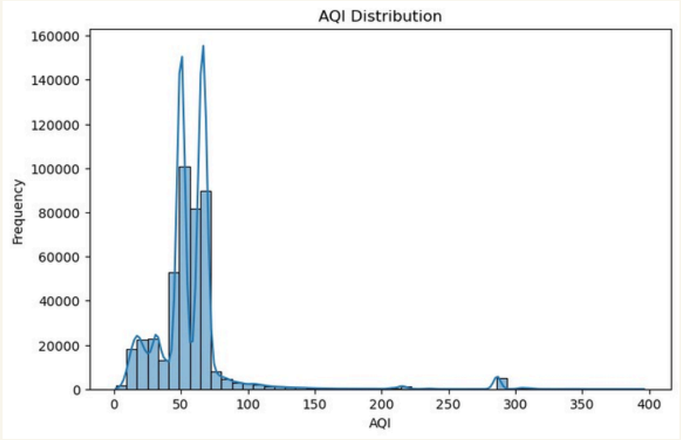
After defining extreme events (AQI > 300), this graph visualizes the spread of extreme events by month



This graph analyzes the effect of festivals like Diwali on AQI levels

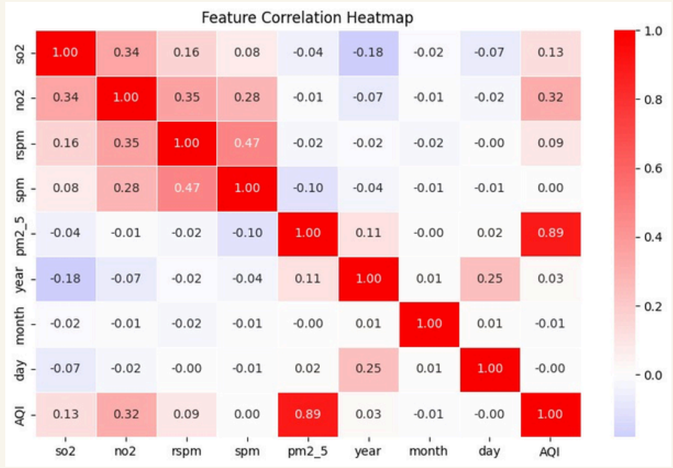


The boxplot shows the trend of median AQI across the months of the year



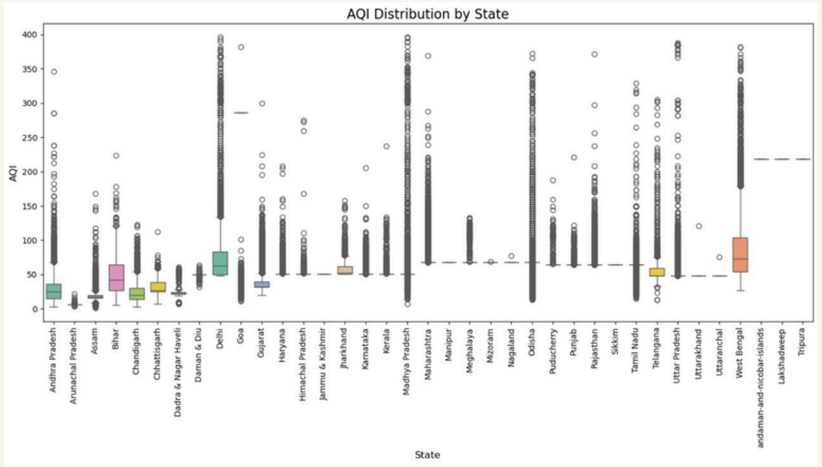
The histogram and KDE visualize the spread of frequency of AQI values

Correlation Analysis – AQI vs Pollutants

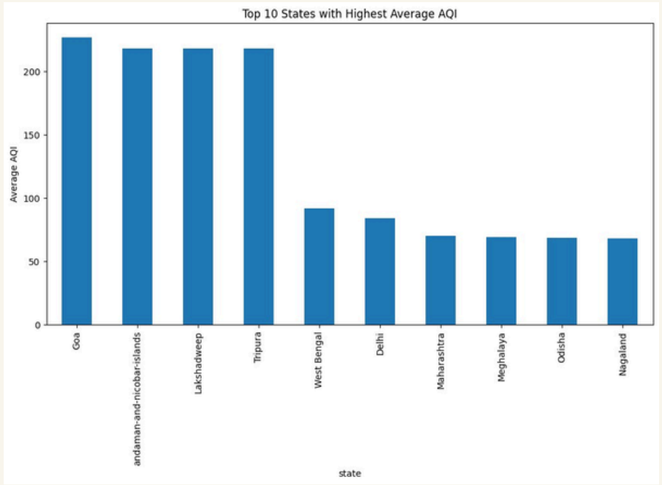


The heatmap pictures the correlation among the various features

Geographical AQI Distribution



This boxplot displays AQI variation across the different states of India



This graph analyzes the states with the 10 highest average AQI

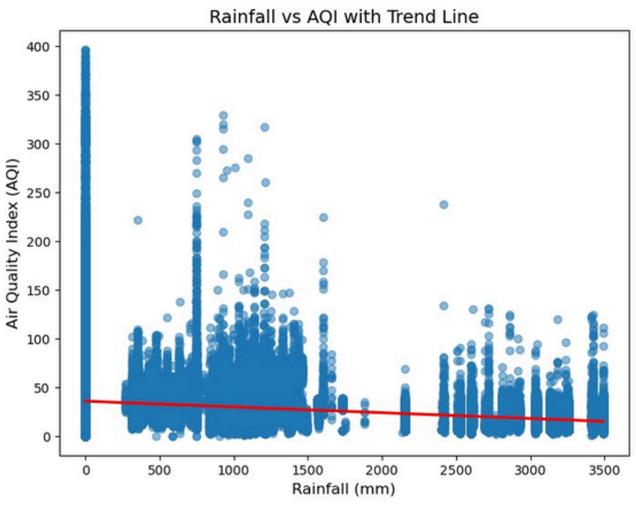
```python
# --- 11. Most Polluted Cities ---
city_avg_aqi = df.groupby('location')['AQI'].mean().sort_values(ascending=False)
print("Most Polluted Cities:")
print(city_avg_aqi.head(10))
```

```
Most Polluted Cities:
location
Panaji      286.344828
Panjim      286.344828
Vasco       286.344828
Mormugao    286.344828
MALDAH      218.068966
SILIGURI    218.068966
ULUBERIA    218.068966
DANKUNI     214.655172
Kalyani     214.655172
HALDIA      214.655172
Name: AQI, dtype: float64
```

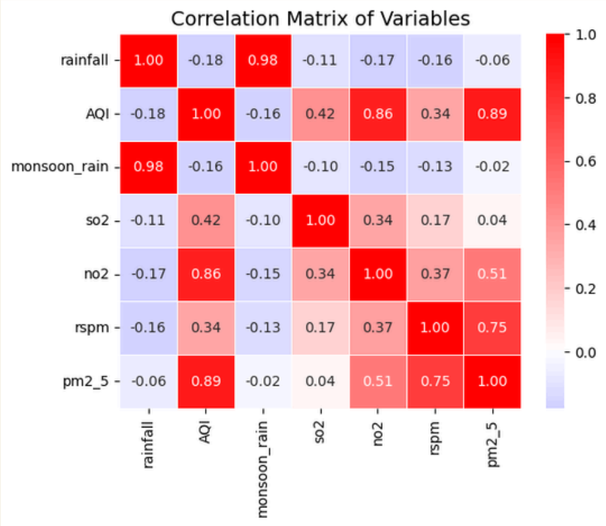
Here is a list of the top 10 most polluted cities with their corresponding AQI values

Rainfall & AQI

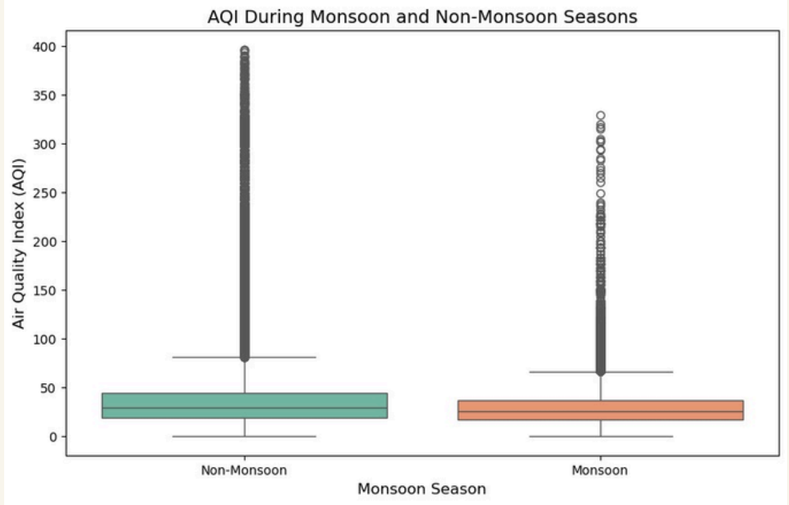
Merged Dataset



This plot shows the AQI values corresponding to the amount of rainfall and a trend line depicting a slight inverse relation



This heatmap shows the scale of impact that rainfall and other factors have on AQI value

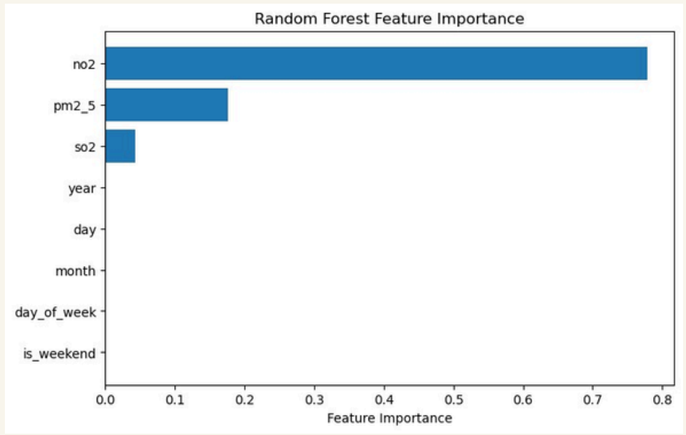


This boxplot compares AQI values during Monsoon (Jun-Sep) and Non-monsoon (Oct-May) seasons

AQI Prediction

Model & Evaluation

- **Data Splitting :** To evaluate the model effectively, we split the dataset into training and testing sets. 80% was used as training data, and the remaining 20% as testing data.
- **Hyperparameter Tuning :** To optimize our Random Forest Regressor, we use RandomizedSearchCV to find the best hyperparameters and train the model.
- **Model Training :** After tuning, we train the Random Forest Regressor with the best-found hyperparameters to predict AQI.
- **Evaluation :** After training the model, we analyze its performance using three key metrics: MAE (=0.0993901487639963 5), RMSE (=1.141327362522169), R^2 Score (=0.997841989121896 3)



	Feature	Importance
6	no2	0.778480
7	pm2_5	0.176468
5	so2	0.043530
0	year	0.000810
2	day	0.000338
1	month	0.000218
3	day_of_week	0.000143
4	is_weekend	0.000013

The graph and the corresponding table show how important each feature (pollutant levels, date-related info) is in predicting AQI.

Key Insights:

- AQI worsens in winter & festival seasons
- PM2.5 and NO_2 are the strongest AQI determinants
- Rainfall does help reduce pollution, but not always (Pearson coefficient = -0.177, indicating slight inverse relationship)
- Certain cities (Delhi, Panaji) remain highly polluted despite weather changes
- Extreme AQI events are becoming more frequent in recent years

Challenges Faced:

- Missing data issues
- Difficulty integrating rainfall dataset
- Model overfitting in some cases

Conclusion

Future Scope:

- Real-time AQI monitoring models using IoT
- Deep Learning models (LSTM) for time-series forecasting
- Government policy recommendations based on predictions

Final Thoughts:

In current times, there is a lot of uncertainty about the state of our environment and its impact. This project aims to use machine learning and data analytics to help government agencies and general public be better prepared for hazardous situations by accurately predicting AQI levels beforehand.