PROJECT DS303

Air Quality Prediction

Team 36 :

23B2286 Vishal Tiwari
23B0710 Uditi Raymangia
23B0067 Kashish Jain
23B2236 Hima Varsha
23B2201 Dnyaneshwari Kate

Project Description
Air Quality Prediction

Introduction

- Air pollution is a growing concern in many Indian cities.
- Accurate AQI forecasting is crucial for maintaining ambient air quality.
- Traditional methods (statistical, deterministic, physics-based) have limitations.
- Machine learning offers improved prediction accuracy.
- Abundant historical data enhances model performance.
- Deep learning techniques further boost forecast reliability.
- Using data from the last 7 days helps predict AQI effectively.

Project Description

Areas and Application

01 Environmental Science

Focuses on analyzing air pollutants and understanding their effects on health and the environment.

02 Machine Learning & Data Science

Involves building predictive models to forecast Air Quality Index (AQI) based on historical data.

03 Public Health Alerts

Provides early warnings about poor air quality to help protect vulnerable groups.

04 Government Policy Support

Assists in creating data-driven decisions for pollution control and urban planning.

Dataset Description

Link

Context

Since industrialization, environmental pollution has become a growing concern. According to the WHO, air pollution causes 7 million premature deaths annually, making it the world's largest environmental health risk. As reported by The New York Times, India’s air pollution is now deadlier than even China’s.

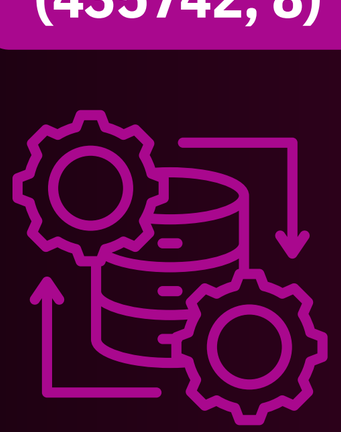
Content

The dataset used is a combined and cleaned version of historical daily ambient air quality data released by the Ministry of Environment and Forests and the Central Pollution Control Board (CPCB) under the National Data Sharing and Accessibility Policy (NDSAP). It includes records across multiple years and states in India.

Feature Description

```
Data columns (total 13 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   stn_code                      291665 non-null  object
 1   sampling_date                 435739 non-null  object
 2   state                         435742 non-null  object
 3   location                      435739 non-null  object
 4   agency                        286261 non-null  object
 5   type                          430349 non-null  object
 6   so2                           401096 non-null  float64
 7   no2                           419509 non-null  float64
 8   rspm                          395520 non-null  float64
 9   spm                           198355 non-null  float64
 10  location_monitoring_station   408251 non-null  object
 11  pm2_5                         9314 non-null    float64
 12  date                          435735 non-null  object
```

Initial Size
(435742, 13)

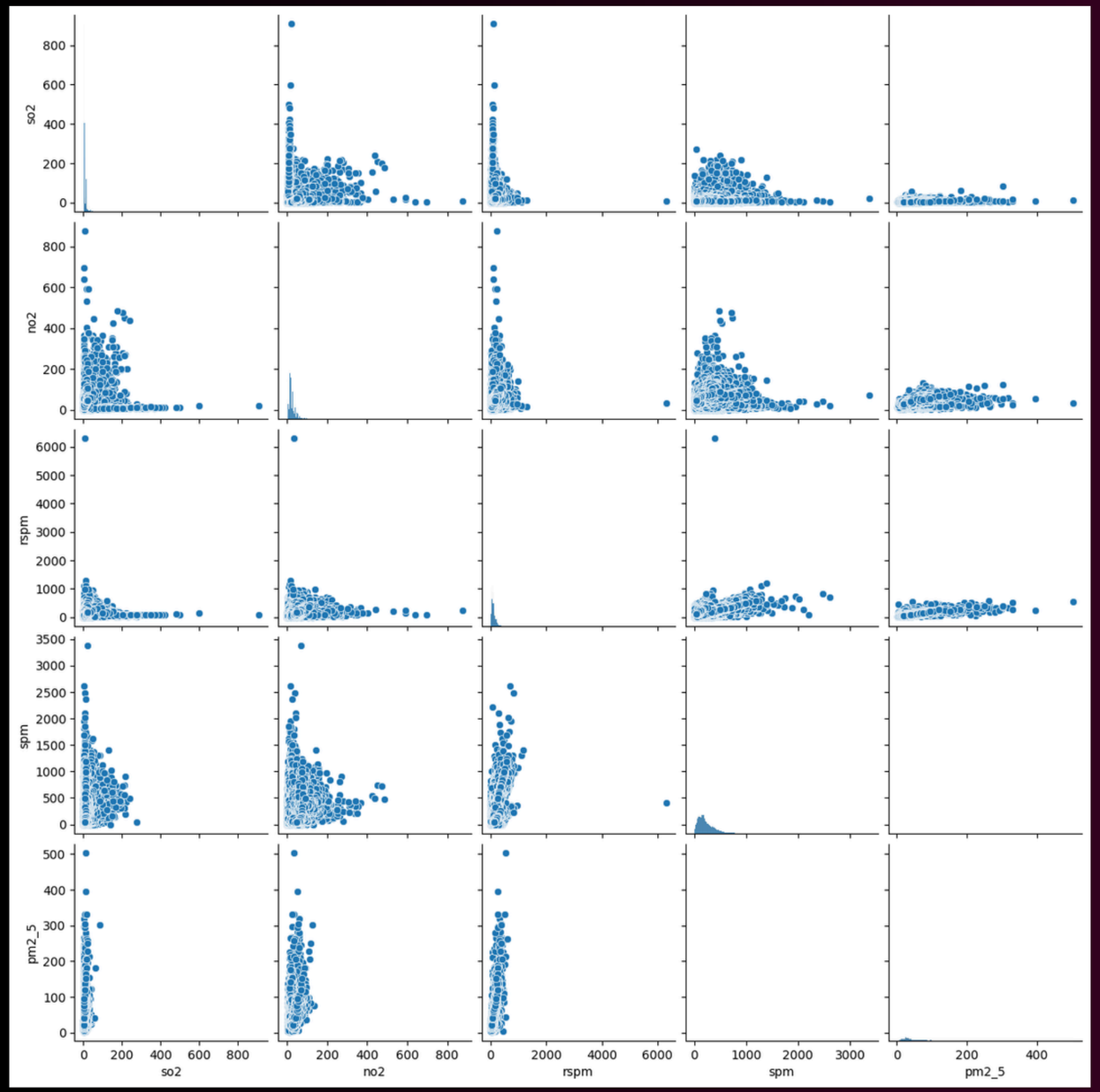


Final Size
(435742, 8)

Exploratory Data Analysis

Pollutant metrics

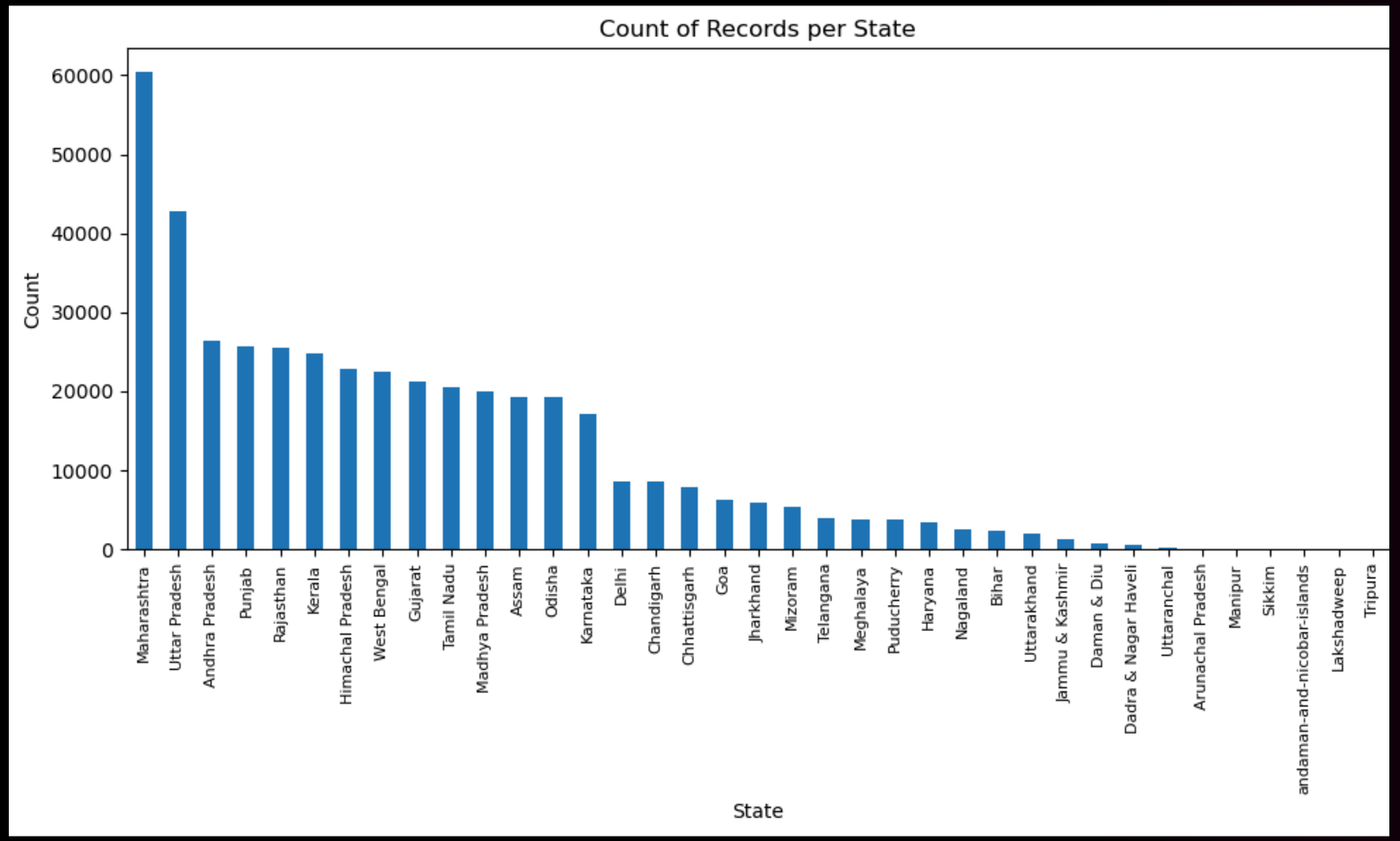
	so2	no2	rspm	spm	pm2_5
count	401096.00	419509.00	395520.00	198355.00	9314.00
mean	10.83	25.81	108.83	220.78	40.79
std	11.18	18.50	74.87	151.40	30.83
min	0.00	0.00	0.00	0.00	3.00
25%	5.00	14.00	56.00	111.00	24.00
50%	8.00	22.00	90.00	187.00	32.00
75%	13.70	32.20	142.00	296.00	46.00
max	909.00	876.00	6307.03	3380.00	504.00



PairPlot of SO_2, NO_2, rspm, spm, pm2_5

Exploratory Data Analysis

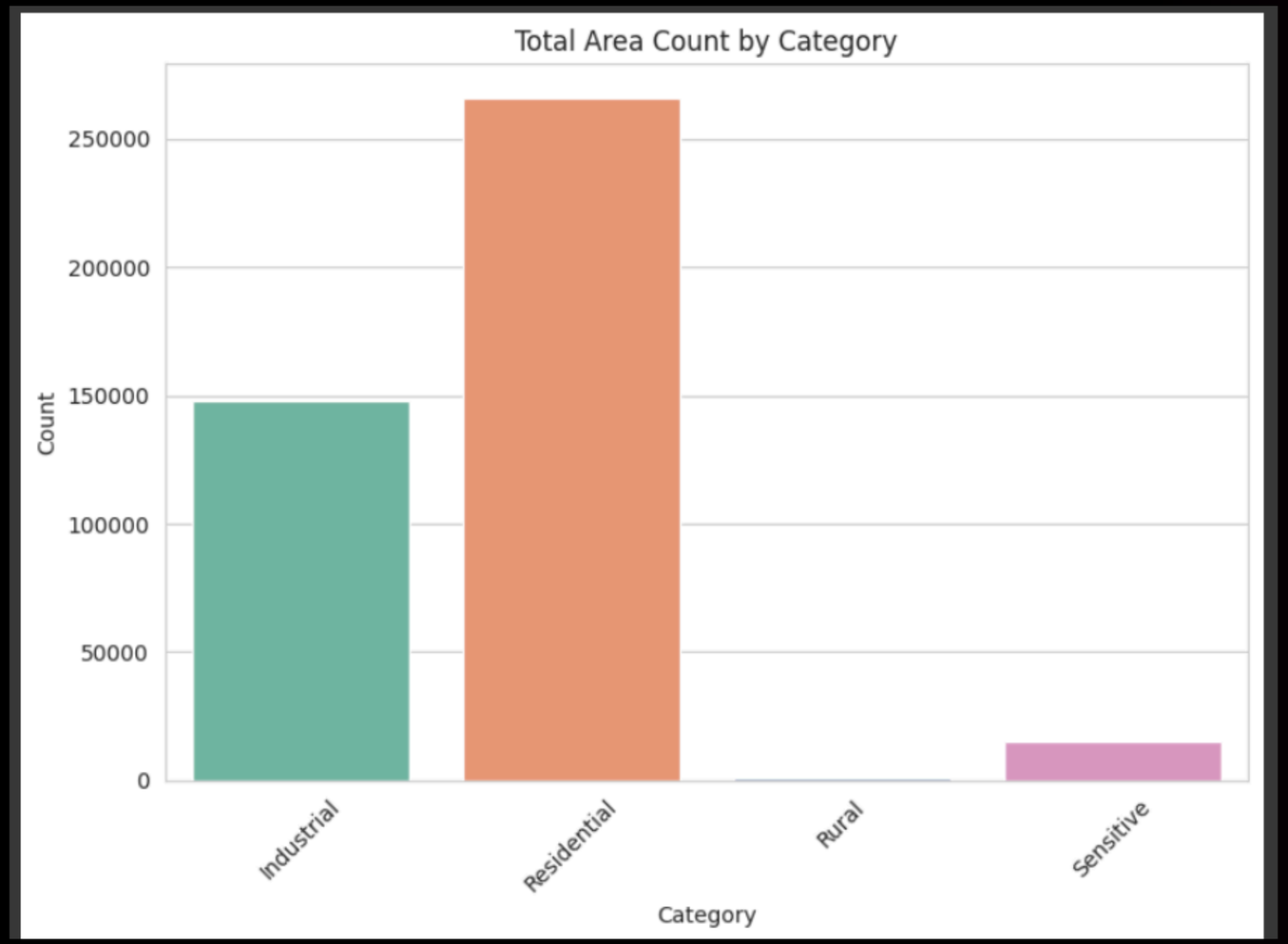
State wise data distribution



state	
Maharashtra	60384
Uttar Pradesh	42816
Andhra Pradesh	26368
Punjab	25634
Rajasthan	25589
Kerala	24728
Himachal Pradesh	22896
West Bengal	22463
Gujarat	21279
Tamil Nadu	20597
Madhya Pradesh	19920
Assam	19361
Odisha	19279
Karnataka	17119
Delhi	8551
Chandigarh	8520
Chhattisgarh	7831
Goa	6206
Jharkhand	5968
Mizoram	5338
Telangana	3978
Meghalaya	3853
Puducherry	3785
Haryana	3420
...	
Sikkim	1
andaman-and-nicobar-islands	1

Exploratory Data Analysis

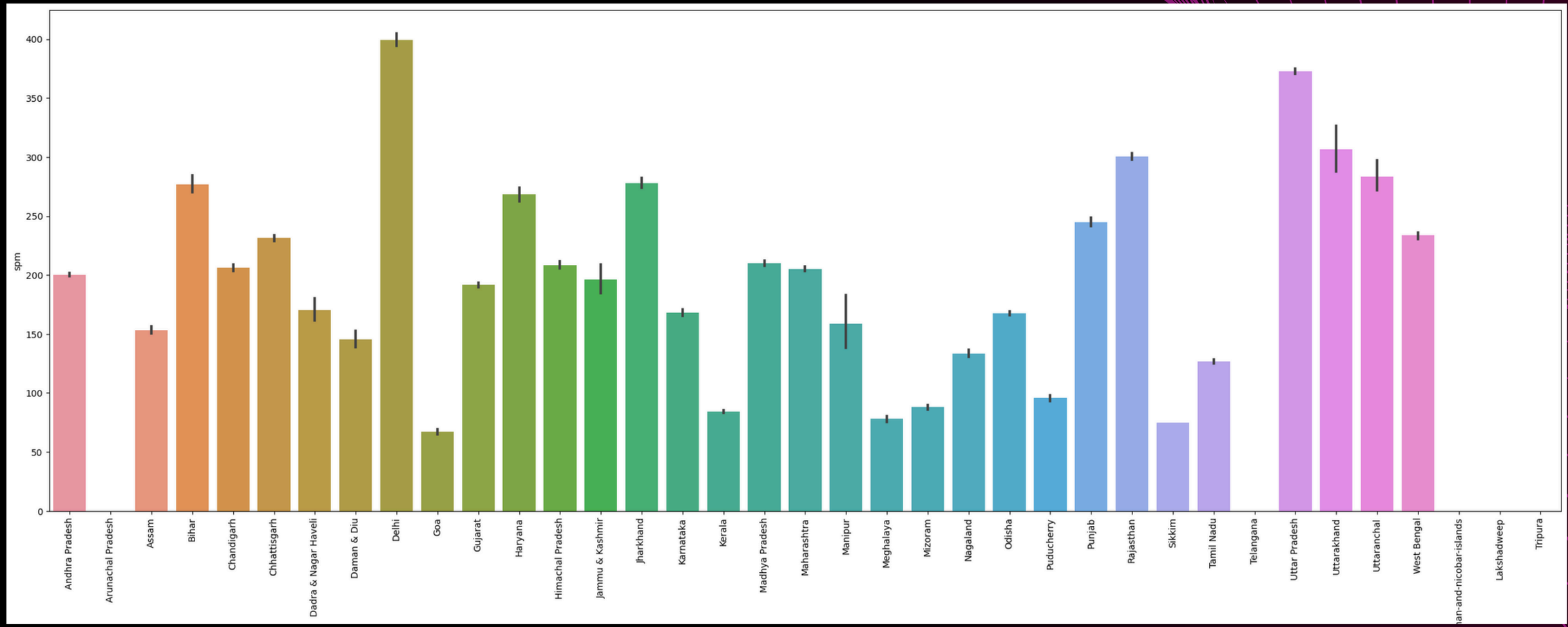
LULC (Land Use) Distribution



	Standard_Category	Count
0	Industrial	148071
1	Residential	265963
2	Rural	1304
3	Sensitive	15011

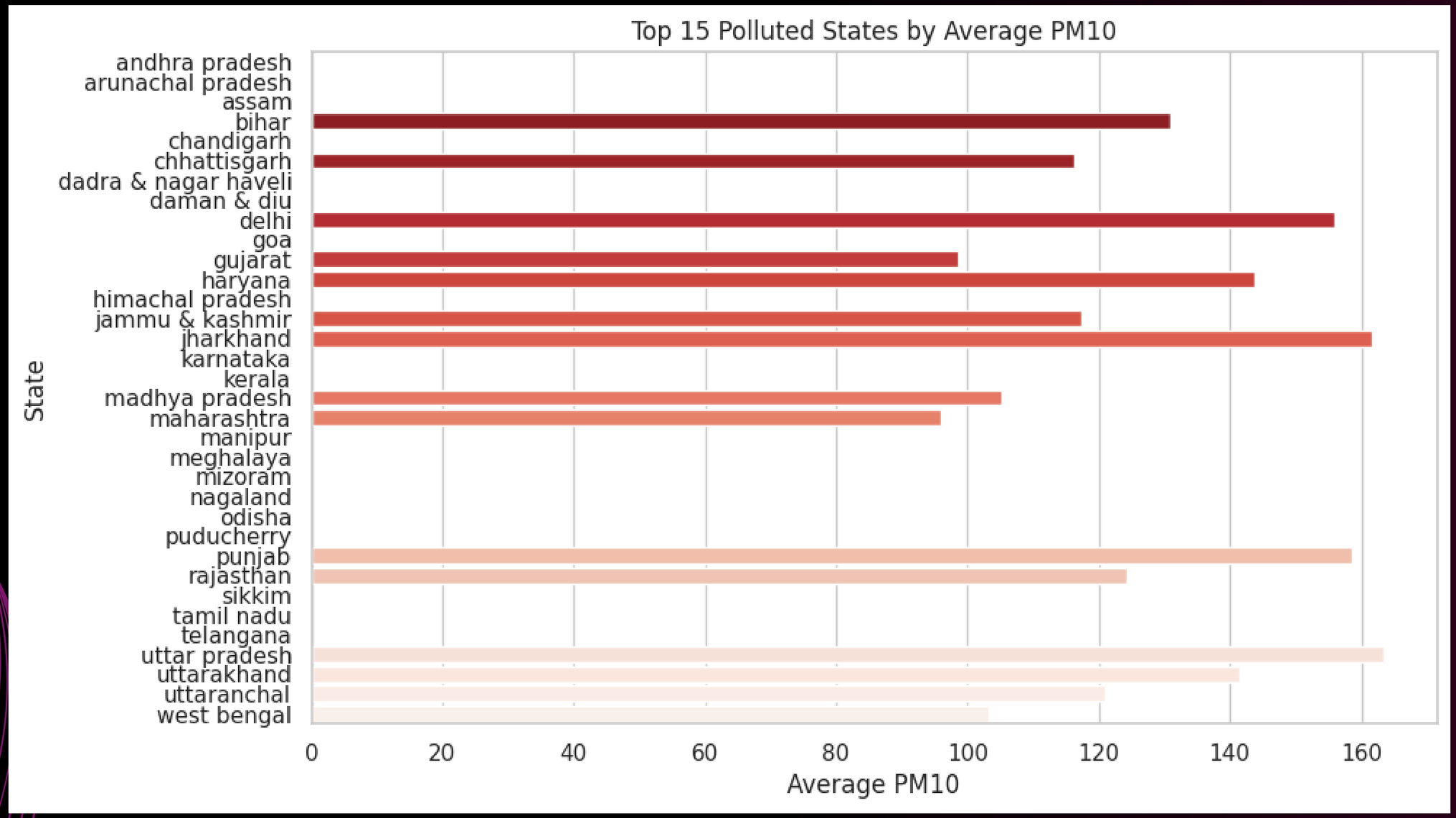
Exploratory Data Analysis

SPM pollutant state distribution



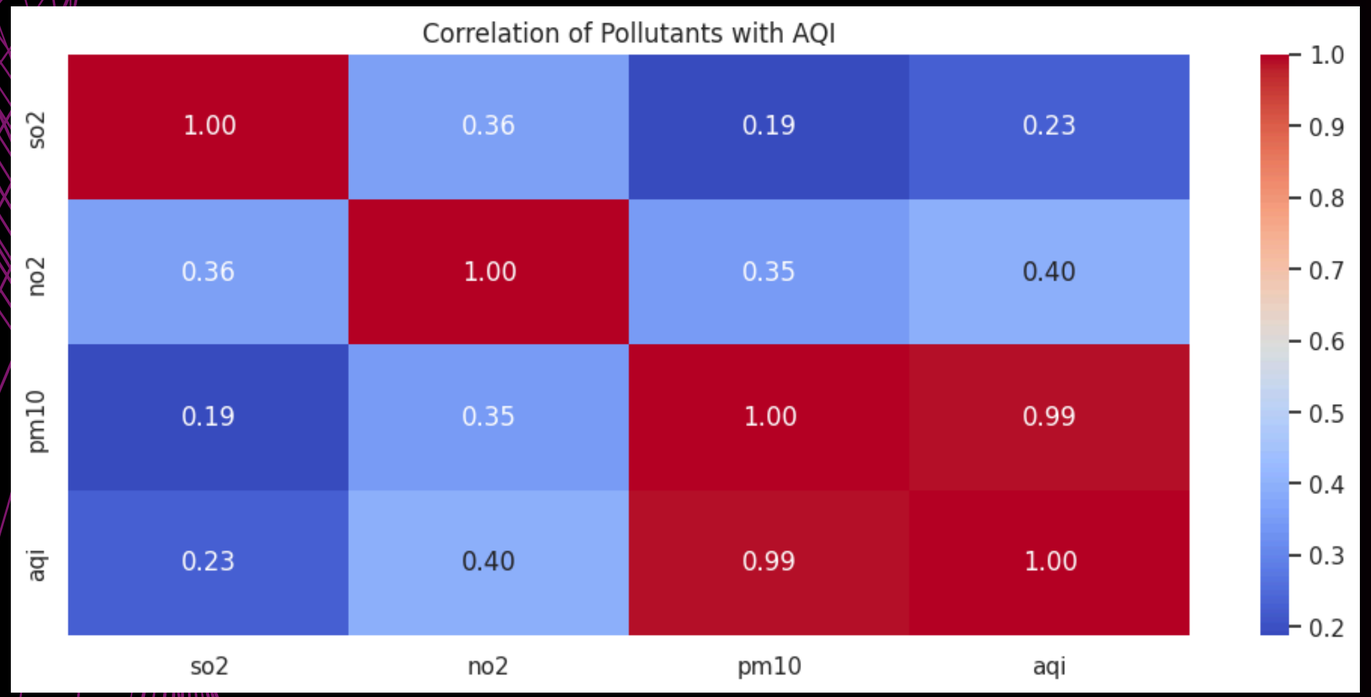
Exploratory Data Analysis

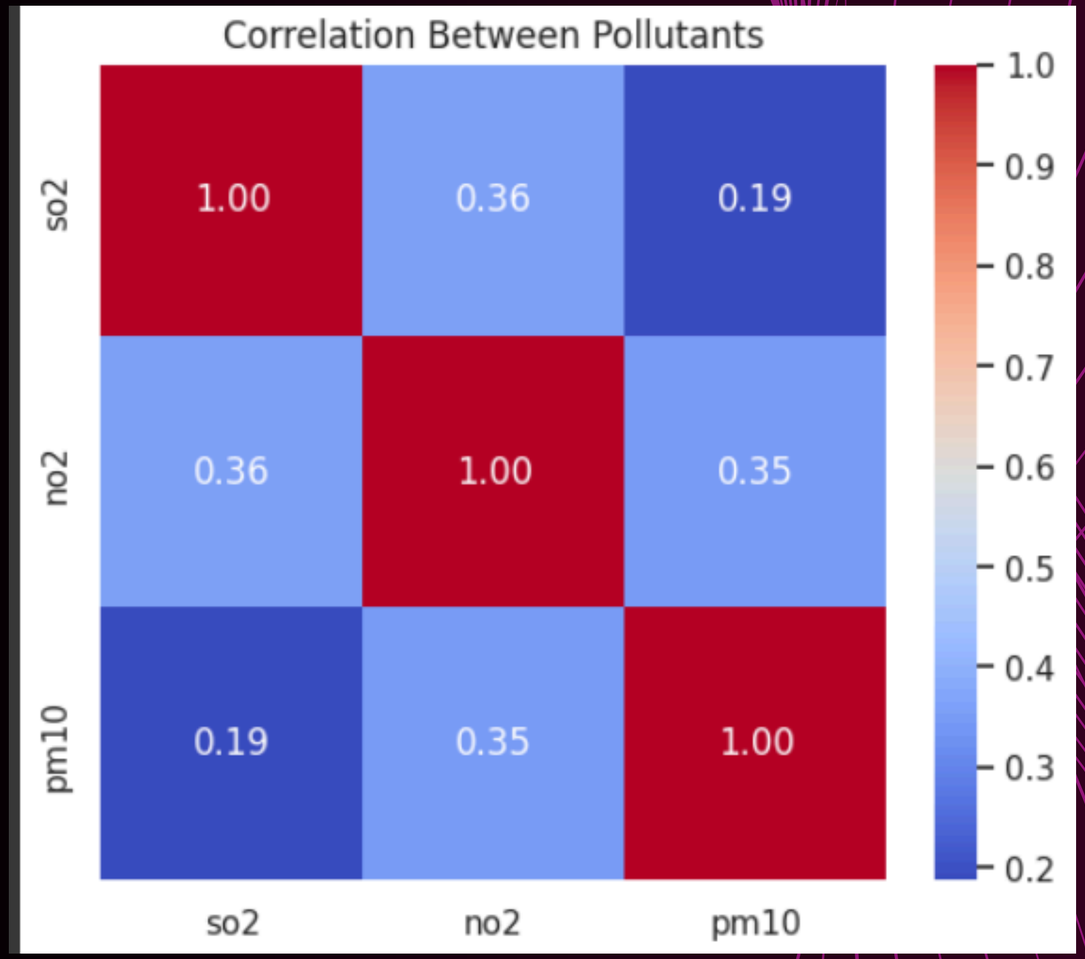
Most Polluted States (by average PM10)



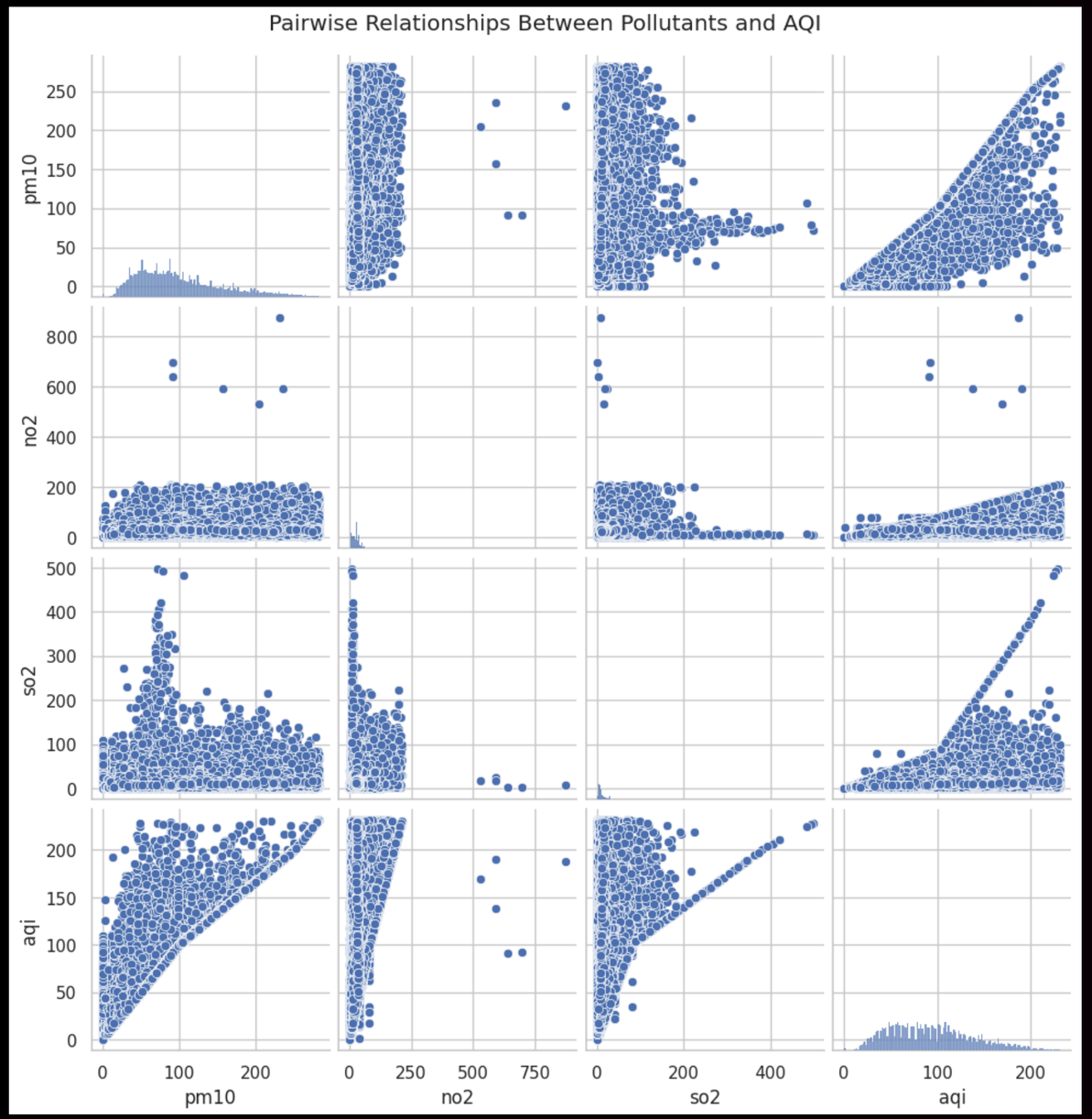
Exploratory Data Analysis

Correlation HeatMap

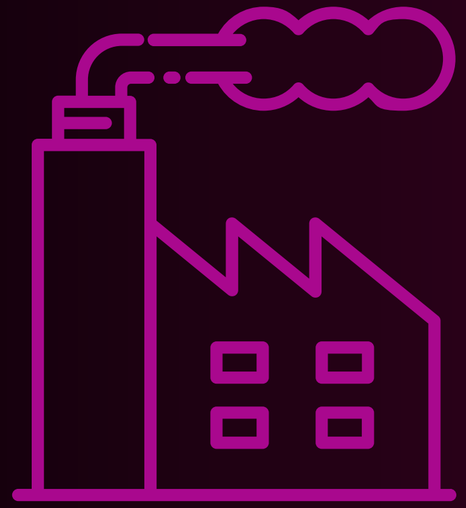




Exploratory Data Analysis



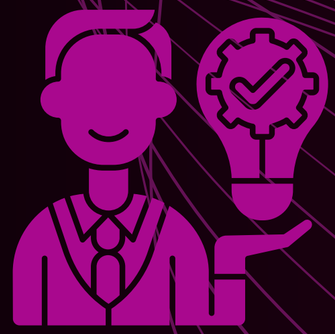
Pairwise plot based on AQI



Related Works: Source of Data and Model Inspiration

WHAT THEY DID

- Cleaned and processed data from OpenAQ and CPCB sources.
- Calculated AQI using pollutant concentration data and CPCB's official formula.
- Trained multiple classifiers on environmental features.
- Evaluated how well each model predicts air quality categories.



Link to Source

Related Works: Source of Data and Model Inspiration

WHAT THEY LACKED

- No actual AQI prediction was performed—AQI was computed deterministically using predefined pollutant-to-AQI conversion formulas.
- Lacked any machine learning or statistical modeling to anticipate future AQI values.
- Did not explore temporal patterns or feature-driven learning.

Link to Source

Related Works: Source of Data and Model Inspiration

WHAT DID WE DO

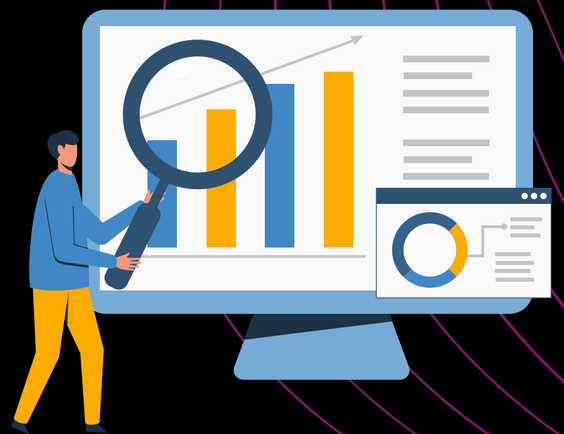
- Instead of relying on the AQI formula for each day, we predict future AQI using past AQI trends and environmental metadata.
- Modeled AQI as a time series regression task.
- Introduced lag features, temporal encodings, and location encoding, which are entirely absent in the source project.
- Our work brings predictive insight beyond visualization or static computation.

Link to Source

Our Approach

Data Preprocessing

Cleaned missing values, filled categories, removed outliers.



Handling Missing Data

Filled missing categorical values with 'Unknown' and dropped rows with all pollutants missing.

AQI Calculation

A shortage of skilled AI professionals can hinder the development and deployment of AI solutions.

Outlier Removal

Applied the IQR method to eliminate extreme values affecting model accuracy.

Data Integrity Check

Ensured consistency and completeness before feeding data into the model.

Our Approach
Feature Engineering

Engineered temporal, lag-based, and location features to capture seasonality, recent trends, and spatial variation in AQI.

Temporal Features

Extracted month and dayofweek from dates to capture seasonal and weekly AQI patterns

Lag Features

Created aqi_lag_1 to aqi_lag_7 to incorporate recent AQI trends, enabling time-aware prediction.

Location Encoding

Transformed categorical location into numerical form using LabelEncoder to help models learn spatial patterns.

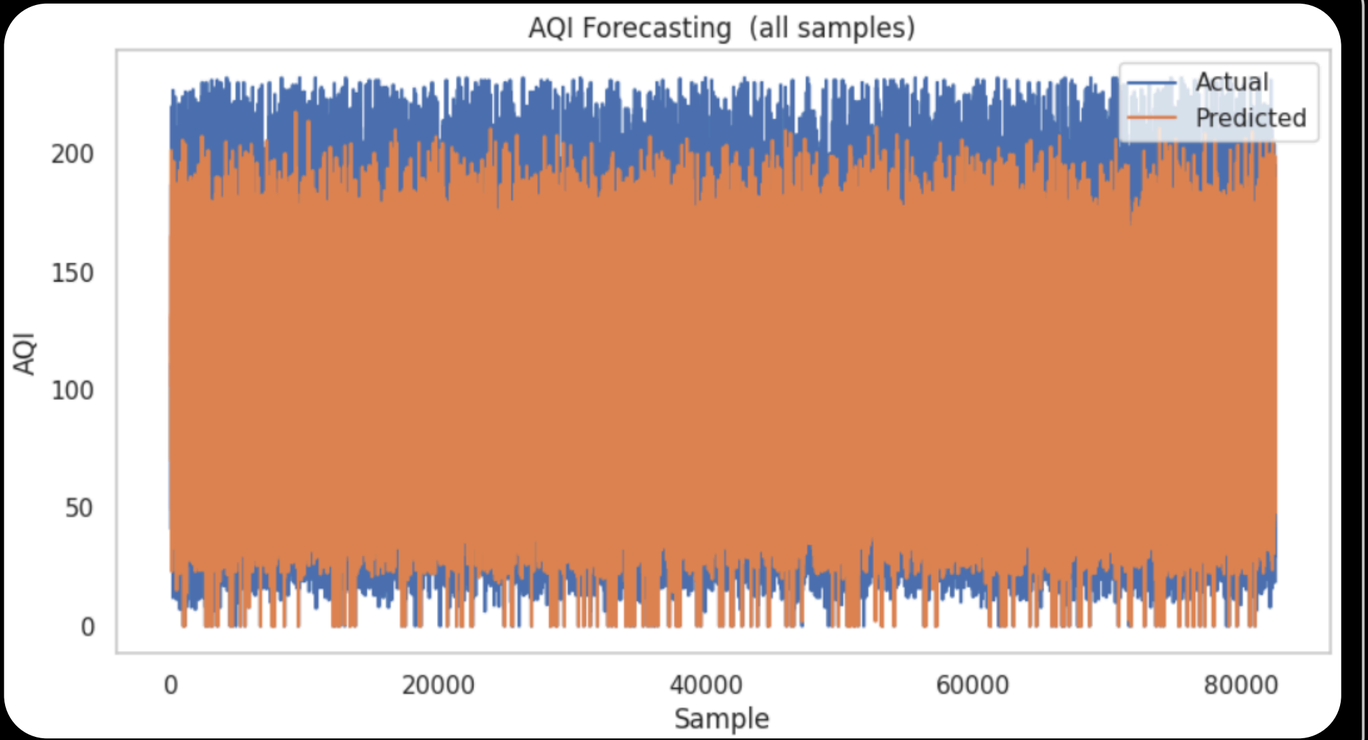
Model Selection

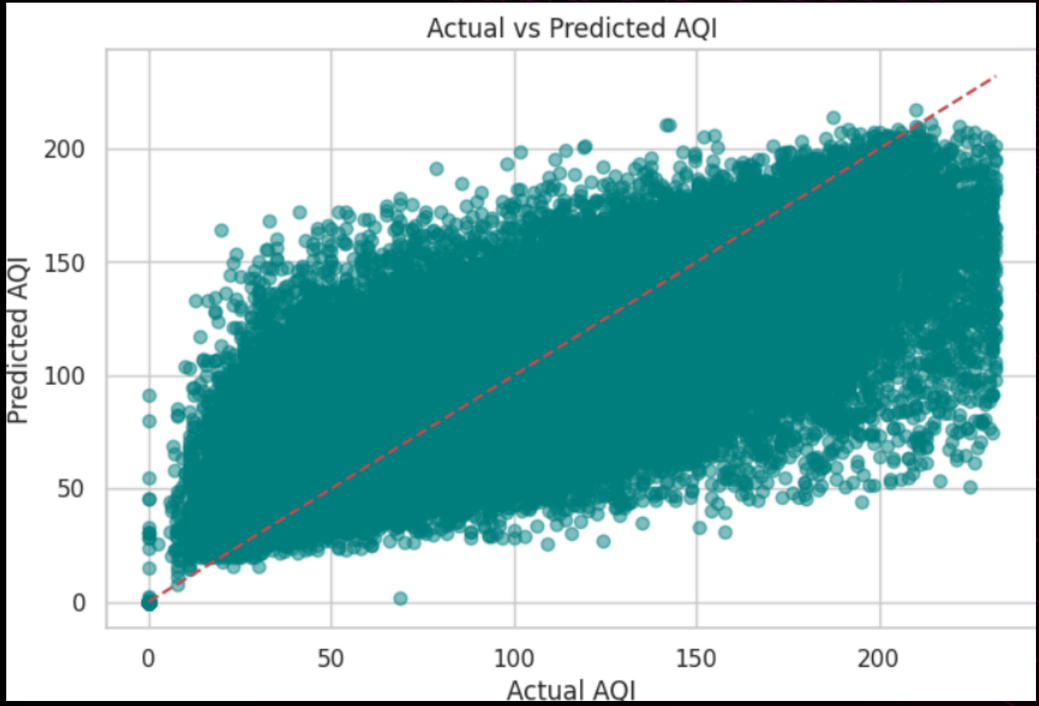
Random Forest Regressor

Why it's used: Random Forest is an ensemble of Decision Trees that reduces overfitting and increases predictive performance. It’s typically more robust and accurate than a single tree.

Results

MAE	19.941
RMSE	28.584
R^2 score	0.6144





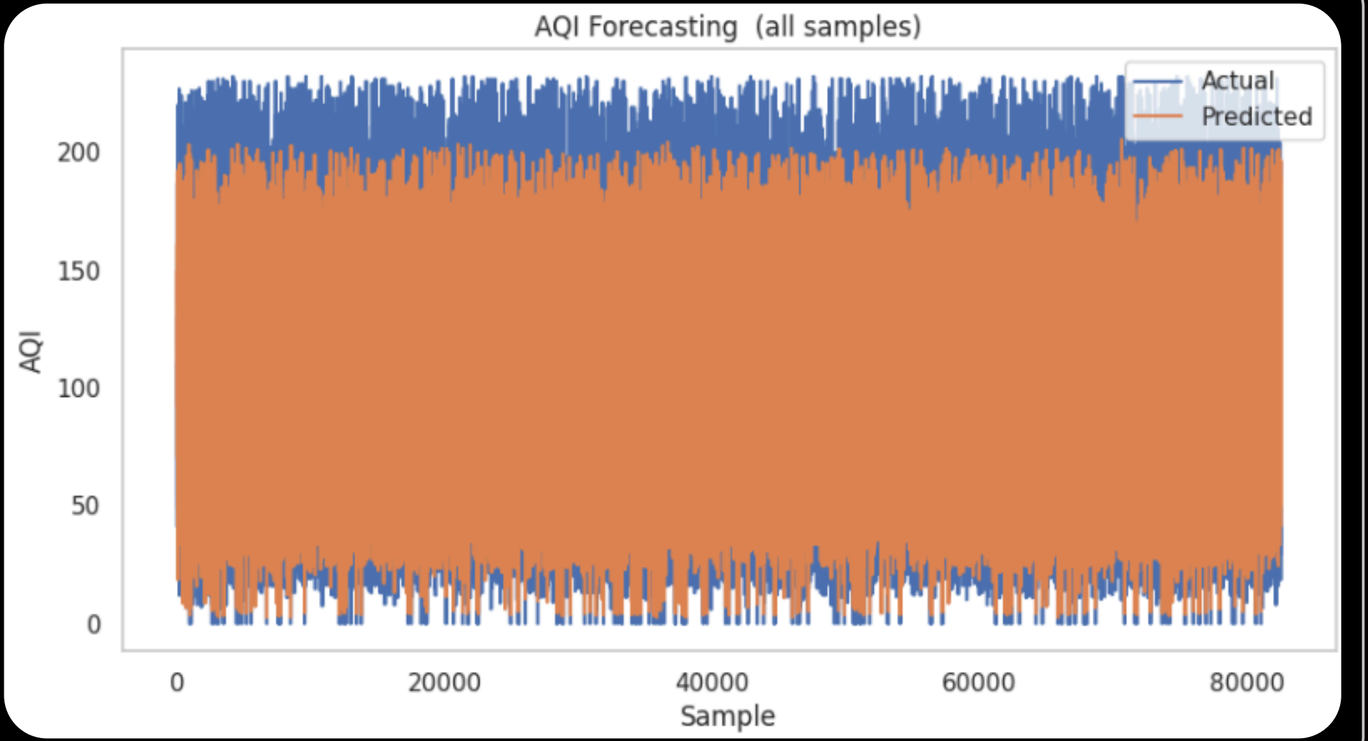
Model Selection

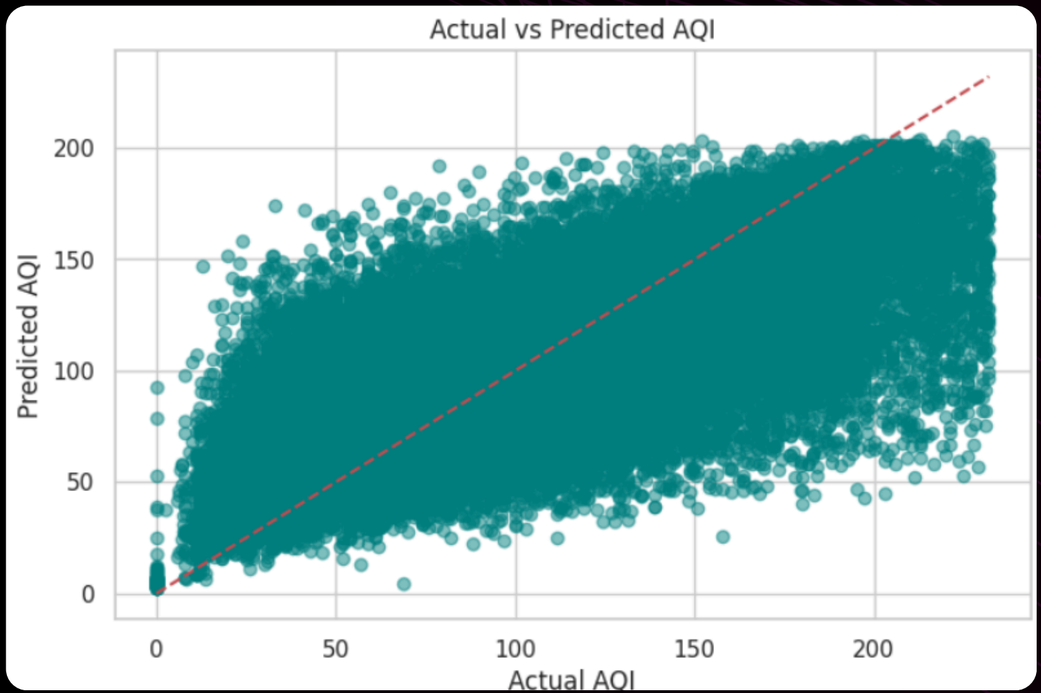
XGBRegressor

XGBoost is a powerful and efficient gradient boosting algorithm known for its speed and accuracy. It handles missing data well and used structured data regression like AQI prediction.

Results

MAE	**19.818**
RMSE	28.201
R^2 score	0.6246





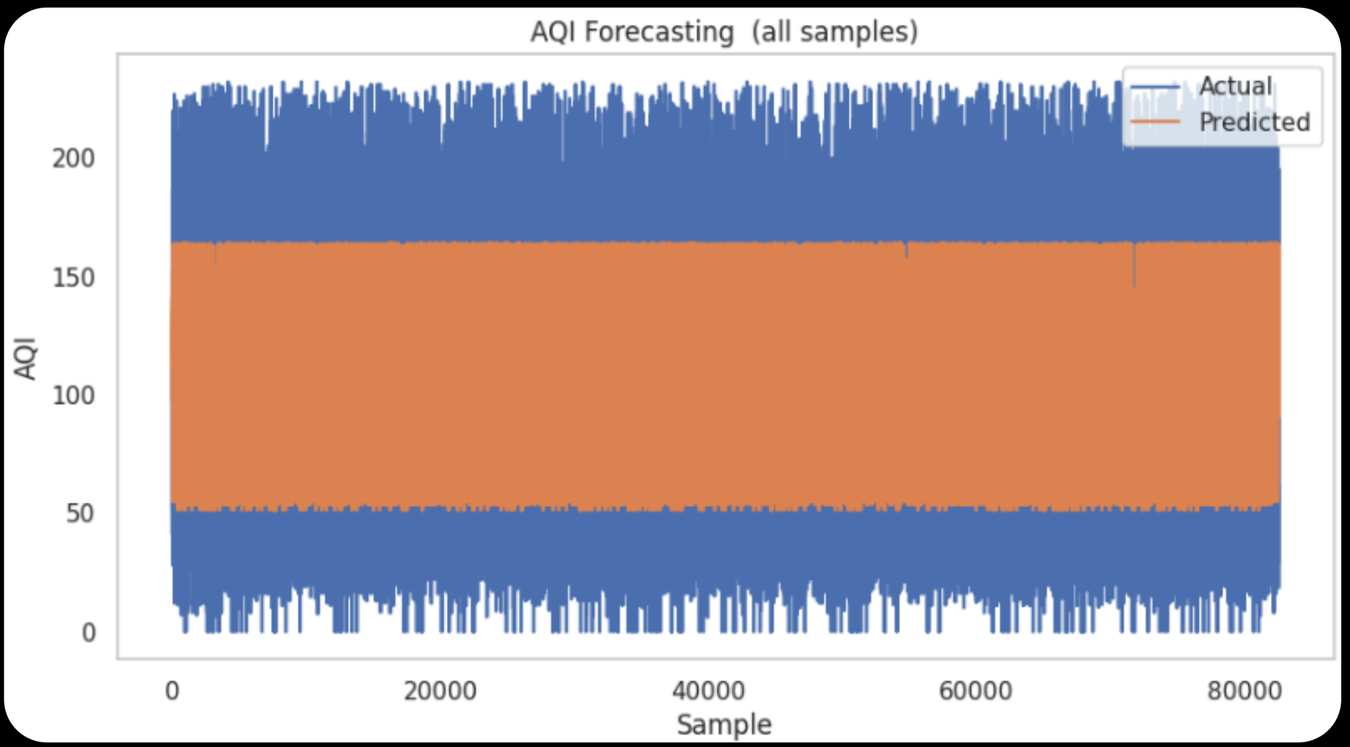
Model Selection

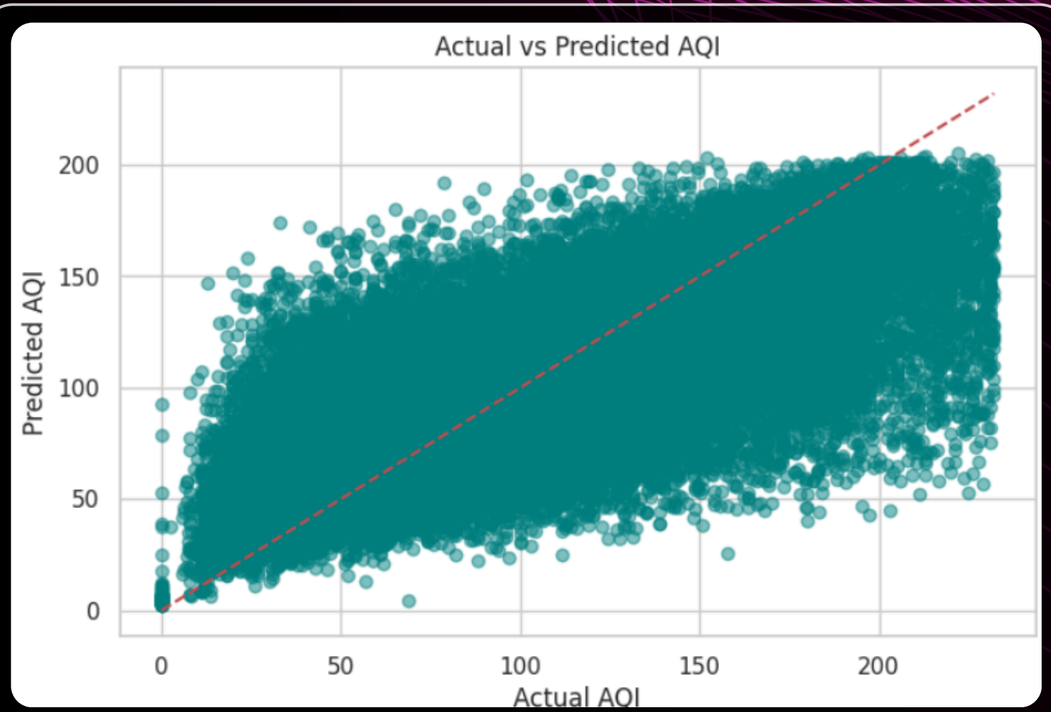
AdaBoostRegressor

AdaBoost combines multiple weak learners (typically decision trees) to improve prediction accuracy. It focuses more on difficult-to-predict samples, making it useful for handling noise.

Results

MAE	29.824
RMSE	36.668
R^2 score	0.3654





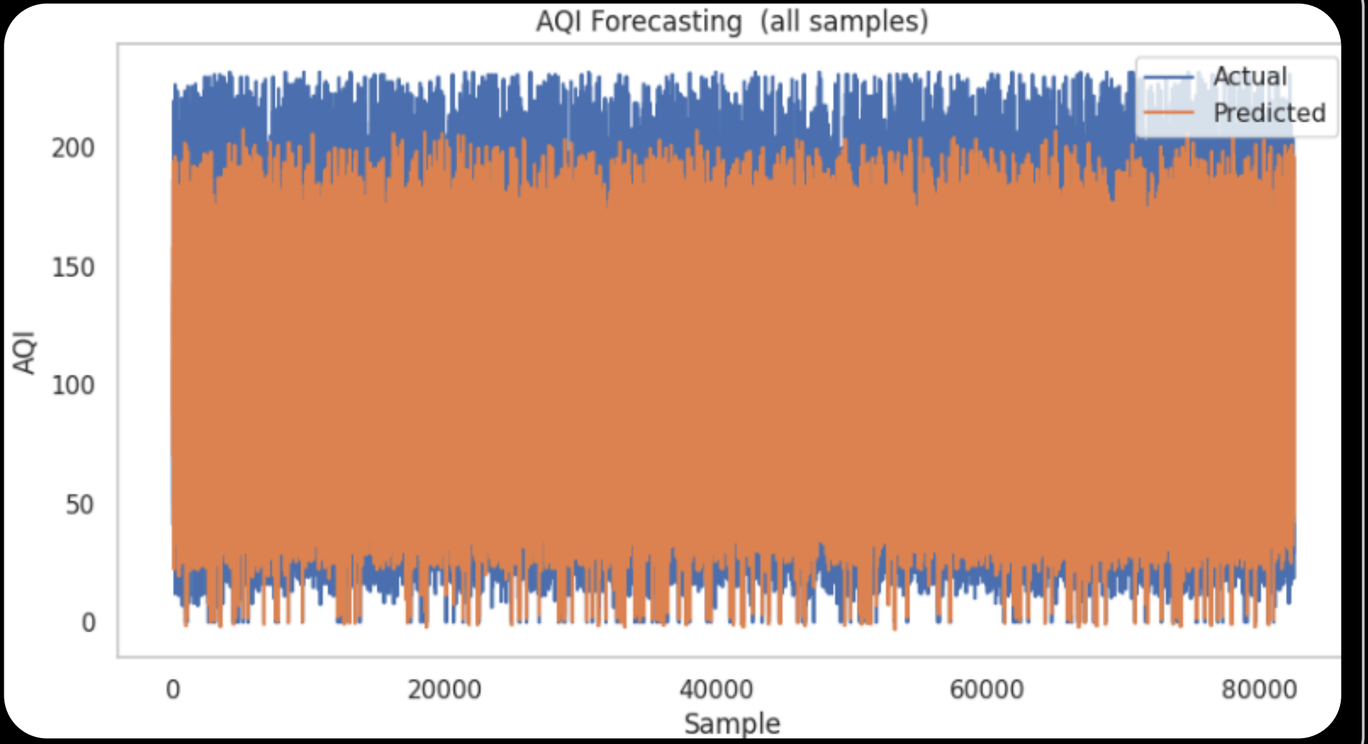
Model Selection

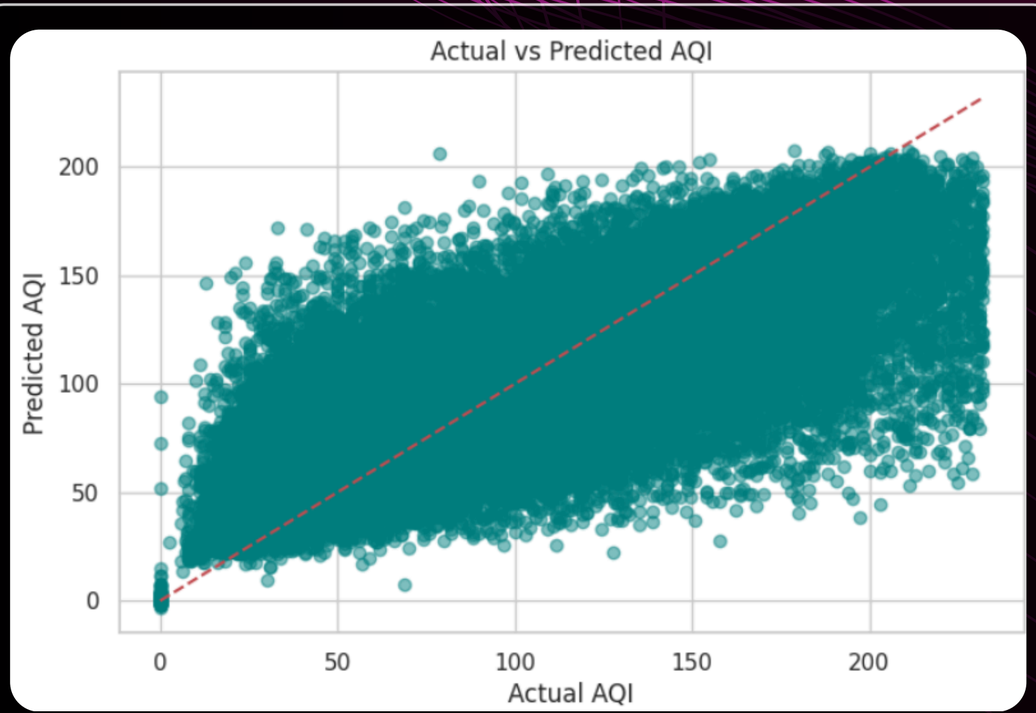
CatBoost Regressor

CatBoost is a gradient boosting model that handles complex, nonlinear relationships well and performs efficiently with minimal preprocessing. It often provides high accuracy in structured data

Results

MAE	19.545
RMSE	27.862
R^2 score	0.6336





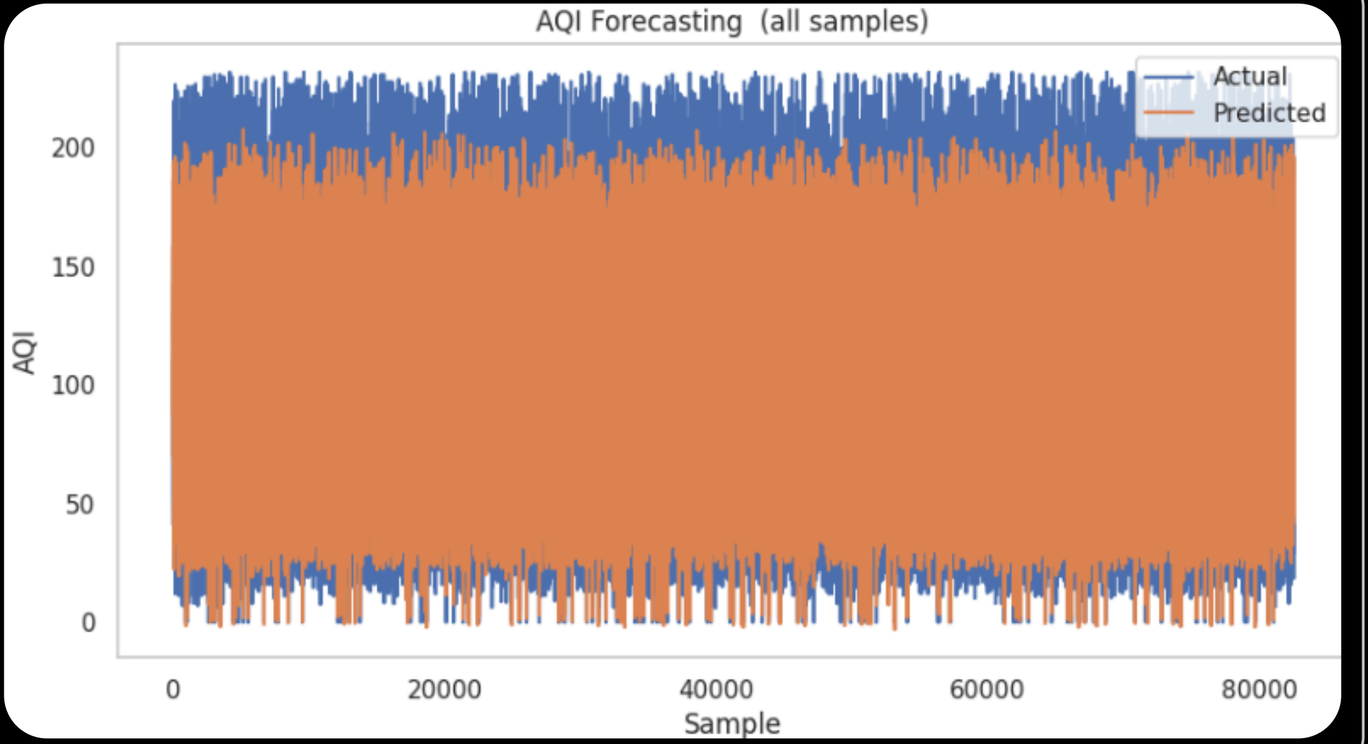
Model Selection

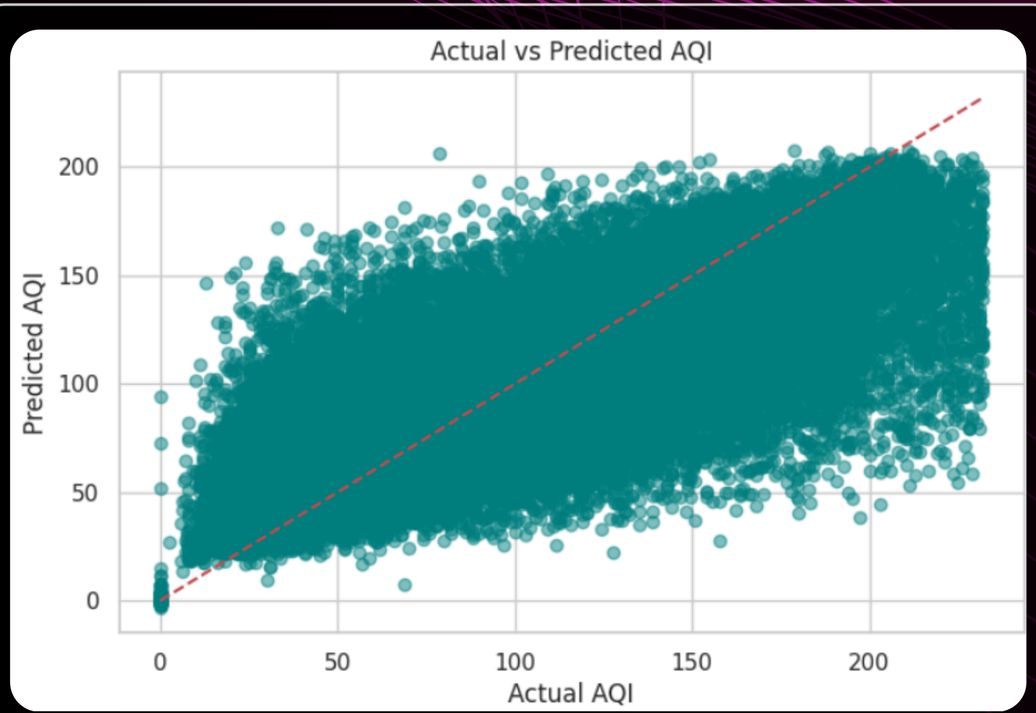
CatBoost Regressor

CatBoost is a gradient boosting model that handles complex, nonlinear relationships well and performs efficiently with minimal preprocessing. It often provides high accuracy in structured data

Results

MAE	19.545
RMSE	27.862
R^2 score	0.6336






Thank
You!!!