# Global Air Analysis Report

## Trends, Insights & Machine Learning for AQI Prediction

**Prepared by: Iman Fasasi**

**Date: March 21, 2025**

# Table of Contents

## 1. Introduction

Air pollution is one of the most pressing environmental challenges of our time, contributing to millions of premature deaths each year and causing widespread damage to ecosystems and urban infrastructure. Understanding the dynamics of air quality—how it varies across regions, which pollutants drive it, and how it can be predicted—is essential for creating effective policies and protecting public health.

This project presents a data-driven analysis of global air quality, using 2024 AQI data from over 170 countries and 300+ cities. It focuses on four major air pollutants:

- **PM2.5** (Fine Particulate Matter)
- **Carbon Monoxide (CO)**
- **Ozone ($O_3$)**
- **Nitrogen Dioxide ($NO_2$)**

By analyzing city-level pollutant concentrations and their corresponding AQI values, this study aims to:

- Uncover regional air quality patterns
- Identify the most and least polluted locations
- Explore correlations between pollutants and overall AQI
- Build machine learning models to predict AQI from pollutant data

Through this approach, the project not only provides insights into global pollution trends but also showcases how data science can be leveraged to support environmental monitoring, public health decision-making, and sustainable urban development.

## 2. Dataset Overview

The dataset used in this project provides a detailed snapshot of global air quality for the year 2024. It includes data from over **300 cities across 170+ countries**, offering a diverse and representative view of air pollution levels around the world.

Each record in the dataset represents a city and contains both the **overall Air Quality Index (AQI)** and the **individual AQI values for key pollutants**.

## Columns & Descriptions

| Column Name | Description |
|---|---|
| country_name | Name of the country |
| city_name | Name of the city |
| aqi_value | Overall AQI score for the city |
| aqi_category | AQI category (e.g., Good, Moderate, Unhealthy) based on overall AQI |
| co_aqi_value | AQI value for Carbon Monoxide (CO) |
| co_aqi_category | AQI category based on CO levels |
| ozone_aqi_value | AQI value for Ozone ($O_3$) |
| ozone_aqi_category | AQI category based on Ozone levels |
| no2_aqi_value | AQI value for Nitrogen Dioxide ($NO_2$) |
| no2_aqi_category | AQI category based on $NO_2$ levels |
| pm2.5_aqi_value | AQI value for PM2.5 (fine particulate matter, ≤2.5µm diameter) |
| pm2.5_aqi_category | AQI category based on PM2.5 levels |

- Total rows (after cleaning): **23,035**
- Total columns: **12**
- Focuses on pollutants with significant public health impact
- Useful for both **exploratory data analysis** and **predictive modeling**

## 3. Data Cleaning & Preprocessing

Before analysis and modeling, the dataset underwent several important cleaning and preprocessing steps to ensure consistency, accuracy, and readiness for machine learning.

### 1. Column Name Fixes

- One of the columns had a formatting issue:
  - co_aqi_value\t (included a hidden tab character)
    Renamed to co_aqi_value for consistency.

### 2. Missing Values

- country_name: 427 missing entries
- city_name: 1 missing entry
  All rows with missing country_name or city_name were **dropped**, as they are critical identifiers.

### 3. Duplicate Rows

- Checked for duplicate records across all columns
  All duplicate rows were **removed** to avoid skewing statistics and model training.

### 4. Column Standardization

- Ensured all pollutant columns were in consistent numeric format
- Confirmed that AQI categories were properly labeled

### 5. Final Dataset Shape

- **Rows:** 23,035
- **Columns:** 12
  The cleaned dataset is now ready for **exploratory data analysis** and **machine learning modeling**.

## 4. Exploratory Data Analysis

The Exploratory Data Analysis phase focused on understanding the global distribution of air quality, identifying pollution patterns, and investigating relationships between pollutants and overall AQI values.

### 1. Most and Least Polluted Cities

- **Most Polluted:** Several cities in **India** reported the maximum AQI value of **500 (Hazardous)**.
- **Least Polluted:** Cities like **El Torno (Bolivia)**, **Macas (Ecuador)**, and **Tari (Papua New Guinea)** recorded AQI values as low as **6 (Good)**.

### 2. AQI Distribution by Category

- AQI values ranged widely from **6 to 500**, with a median near **55**.
- A boxplot of AQI by category showed:
  - **"Hazardous"** and **"Unhealthy"** categories had significant spread and outliers.
  - **"Good"** and **"Moderate"** categories were more concentrated with lower AQI values.

### 3. Dominant Pollutants by City

- PM2.5 was the **most common dominant pollutant** across cities.
- Fewer cities were dominated by **Ozone**, **$NO_2$**, or **CO**.
- This indicates PM2.5 plays a major role in determining overall AQI globally.

### 4. Top Polluted Countries (by Average AQI)

- Countries with the highest average AQI include:
  - India
  - Bangladesh
  - Pakistan
- These countries consistently show high concentrations of particulate matter and ozone pollution.

### 5. Correlation Analysis

- A heatmap of numeric features revealed:
    - **PM2.5 AQI** had the **strongest positive correlation** with overall AQI.
    - **Ozone** and **$NO_2$** showed moderate correlation.
    - **CO** showed weak correlation, suggesting it's less impactful globally.

### Key Takeaways from EDA

- **PM2.5** is the primary pollutant driving poor air quality.
- There are stark differences in air quality between regions.
- Some cities experience dangerously high AQI levels, which require immediate attention from policymakers.
- These findings helped shape the modeling strategy by prioritizing **PM2.5 as a key feature**.

## 5. Predictive Modeling

The goal of this section was to build and evaluate machine learning models that predict the **overall AQI value** (aqi_value) using the AQI levels of individual pollutants:

- co_aqi_value
- ozone_aqi_value
- no2_aqi_value
- pm2.5_aqi_value

### Modeling Approach

Two regression models were trained and evaluated:

1. **Linear Regression**
    - Assumes a linear relationship between pollutant AQI levels and overall AQI.

    - Easy to interpret, useful as a baseline.

2. **Decision Tree Regressor**
    - A non-linear, rule-based model.
    - Captures interactions between features more flexibly.

**Model Performance**

| **Model** | **R² Score** | **MAE** | **RMSE** |
|---|---|---|---|
| **Linear Regression** | 0.9744 | 4.91 | 9.23 |
| **Decision Tree Regressor** | 0.9955 | 0.28 | 3.88 |

**R² Score** measures how well the model explains variance (closer to 1 is better).

- **MAE** (Mean Absolute Error) and **RMSE** (Root Mean Squared Error) show the average and penalized prediction errors.

**Predictions vs. Actual**

- Both models performed well, but the **Decision Tree Regressor** had near-perfect alignment with actual AQI values.
- Visual scatterplots confirmed this:
    - Linear Regression showed minor deviations at extreme AQI levels.
    - Decision Tree predictions closely followed the ideal diagonal line.

**Why Decision Tree Performed Better**

- AQI data may have **non-linear relationships** and **threshold effects** (e.g., sudden jumps in category).
- Decision Trees are **non-parametric**, making them ideal for capturing such patterns.

**Summary**

- **AQI can be accurately predicted** using pollutant AQI values, especially PM2.5.
- The **Decision Tree Regressor** is a powerful model for this task, achieving **$R^2$ of 0.9955**.
- These models can be used in **real-time air quality monitoring** or early-warning systems.

## 6. Conclusion & Recommendations

This project presented a detailed analysis of global air quality data across more than 170 countries and 300+ cities. Through data cleaning, exploratory analysis, and machine learning, the project uncovered key insights about pollution dynamics and demonstrated how AI can support environmental monitoring.

Key findings include:

- **PM2.5 is the most critical pollutant**, strongly correlated with overall AQI in most regions.
- Cities in countries like **India** consistently showed the highest AQI values (up to 500).
- **Cleanest air** was observed in parts of **South America and Southeast Asia**.
- A **Decision Tree Regressor** accurately predicted AQI with an $R^2$ of 0.9955, MAE of 0.28, and RMSE of 3.88, outperforming a baseline linear model.

This project illustrates how data science can play a significant role in **tracking environmental hazards**, **guiding policy**, and **supporting public health efforts**.

## Recommendations

Based on the analysis and predictive insights:

1. **Target PM2.5 Reduction**
    - Enforce stricter regulations around industrial emissions, burning, and transportation sources.
    - Promote cleaner technologies and alternative fuels.

2. **Improve Monitoring Infrastructure**
    - Invest in real-time AQI sensors in high-risk cities.
    - Share AQI updates publicly to improve awareness and decision-making.
3. **Use Predictive Models in Policy Planning**
    - Leverage machine learning models to anticipate AQI spikes.
    - Use predictions to issue health alerts and prepare urban response strategies.
4. **Expand Public Access to Air Quality Data**
    - Create user-friendly dashboards for citizens, researchers, and policy-makers.
    - Encourage community-driven reporting and feedback loops.

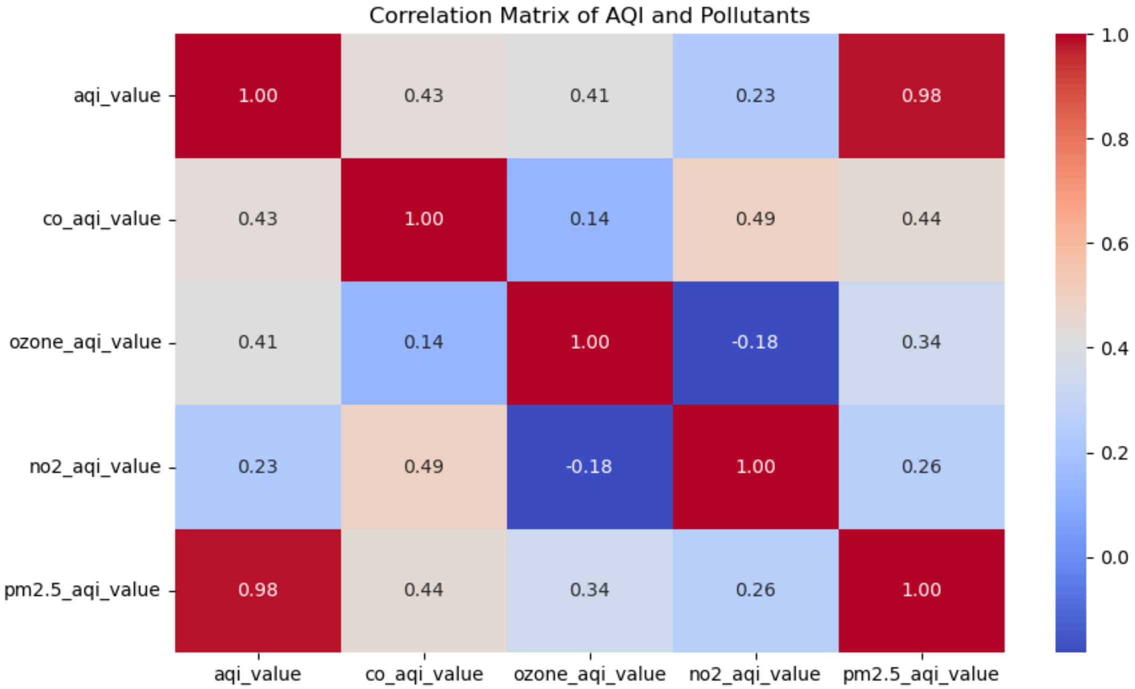
## 7. Appendix

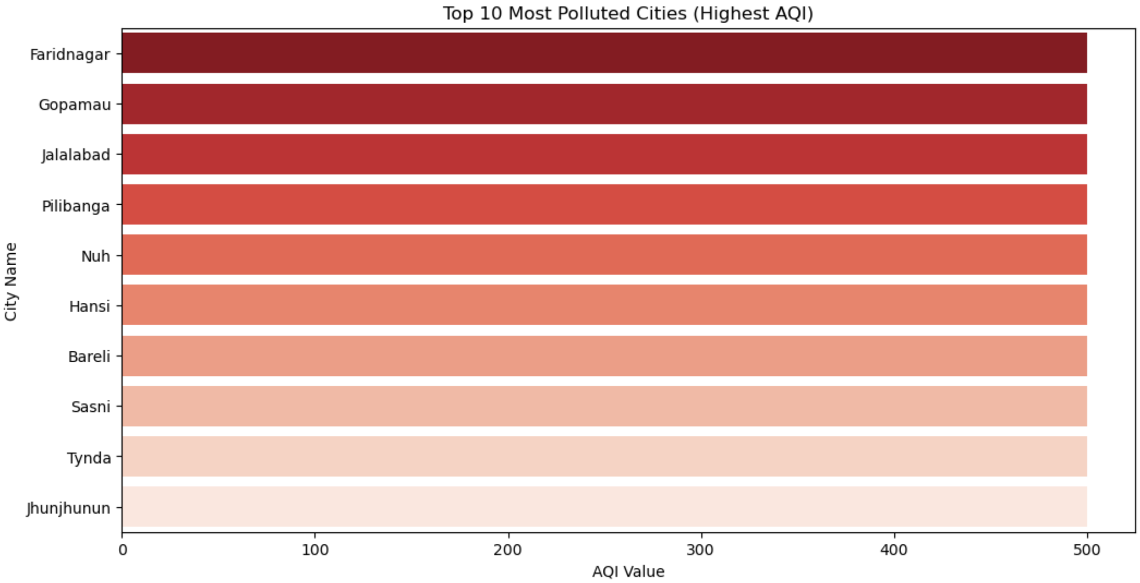
This section includes supplemental content that supports the analysis, modeling, and insights presented in the main report.

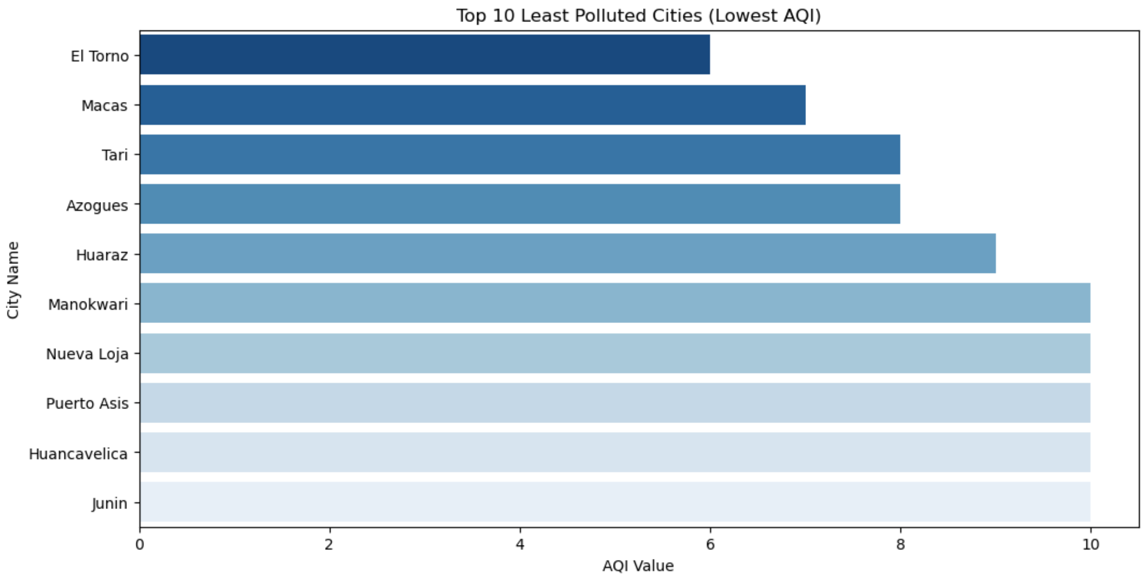
### A. Dataset Snapshot

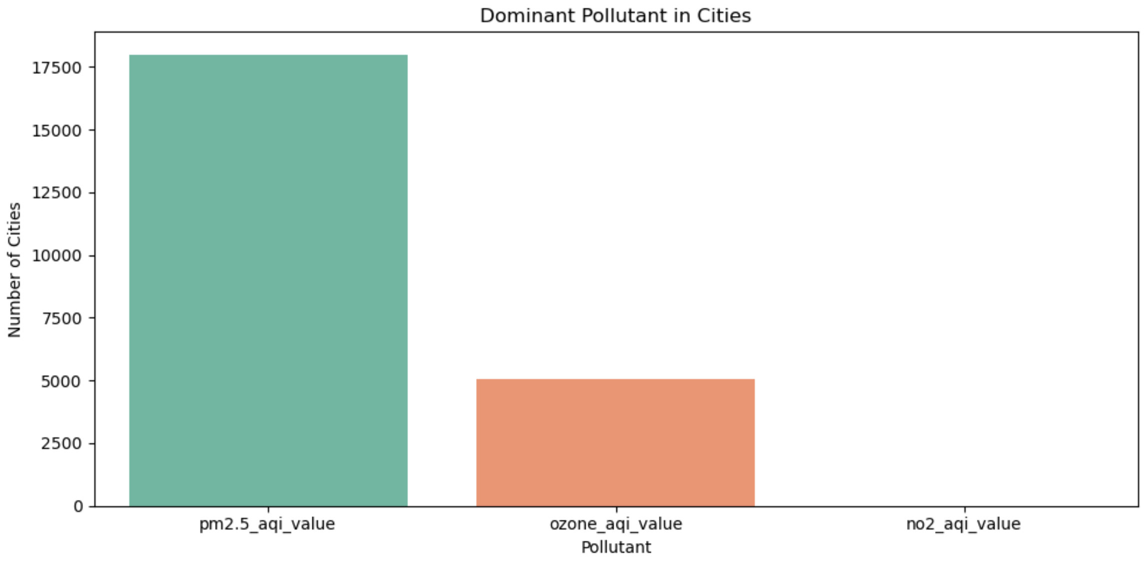
- Source: Global air quality dataset (2024)
- Total observations after cleaning: **23,035**
- Features: 12 columns including AQI and pollutant-specific values and categories
- Sample columns:
    - country_name, city_name
    - aqi_value, aqi_category
    - co_aqi_value, ozone_aqi_value, no2_aqi_value, pm2.5_aqi_value

### B. Visuals & Plots


Correlation Matrix of AQI and Pollutants
aqi_value
co_aqi_value
ozone_aqi_value
no2_aqi_value
pm2.5_aqi_value
1.00
0.43
0.41
0.23
0.98
0.43
1.00
0.14
0.49
0.44
0.41
0.14
1.00
-0.18
0.34
0.23
0.49
-0.18
1.00
0.26
0.98
0.44
0.34
0.26
1.00
1.0
0.8
0.6
0.4
0.2
0.0


Top 10 Most Polluted Cities (Highest AQI)
City Name
Faridnagar
Gopamau
Jalalabad
Pilibanga
Nuh
Hansi
Bareli
Sasni
Tynda
Jhunjhunun
0
100
200
300
400
500
AQI Value





## C. Model Evaluation Code Snippet

```
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error

import numpy as np

# Evaluation function

def evaluate_model(y_true, y_pred, model_name):

    print(f"{model_name} Performance:")
    print("R² Score:", r2_score(y_true, y_pred))

    print("MAE:", mean_absolute_error(y_true, y_pred))

    print("RMSE:", np.sqrt(mean_squared_error(y_true, y_pred)))
```

**D. AQI Categories Reference**

| **AQI Range** | **Category** |
|---|---|
| 0–50 | Good |
| 51–100 | Moderate |
| 101–150 | Unhealthy for Sensitive Groups |
| 151–200 | Unhealthy |
| 201–300 | Very Unhealthy |
| 301–500 | Hazardous |

**E. Tools & Libraries Used**

- Python 3.10+
- Jupyter Notebook
- pandas, numpy
- seaborn, matplotlib
- scikit-learn
- docx (for report generation)