Semester Project

Report

Submitted By:

Asif Hussain 2023-CS-646

Subject:

Introduction To Data Science

Submitted To:

Mam Alina Munir

Dated:

28th December 2025

Department of Computer Science

University of Engineering and Technology Lahore, New Campus

Title:

Global Air Quality Analysis

1. Introduction

Air pollution is a major global issue that directly affects human health, climate, and ecosystems. Rapid urbanization, industrialization, and increased vehicular emissions have significantly degraded air quality in many cities worldwide. This project analyzes the *Global Air Quality Dataset* consisting of 10,000 records collected from major cities across different countries.

The analysis focuses on understanding air pollution patterns, calculating Air Quality Index (AQI), exploring relationships between pollutants and meteorological factors, and building predictive machine learning models. The work is aligned with the course *CSC380 – Introduction to Data Science* and fulfills the CCP requirements.

2. Objectives of the Project

The main objectives of this project are:

- To analyze global air quality data
- To calculate and categorize AQI
- To study pollutant distributions and trends
- To visualize relationships between pollutants and weather
- To prepare data for machine learning analysis

3. Dataset Description

The dataset contains 10,000 records collected from different cities and countries around the world.

3.1 Features Used

- PM2.5: Fine particulate matter (very harmful)
- PM10: Coarse particulate matter
- NO2: Nitrogen dioxide
- SO2: Sulfur dioxide

- CO: Carbon monoxide
- O3: Ozone
- Temperature
- Humidity
- Wind Speed
- City, Country, Date

4. Tools and Technologies

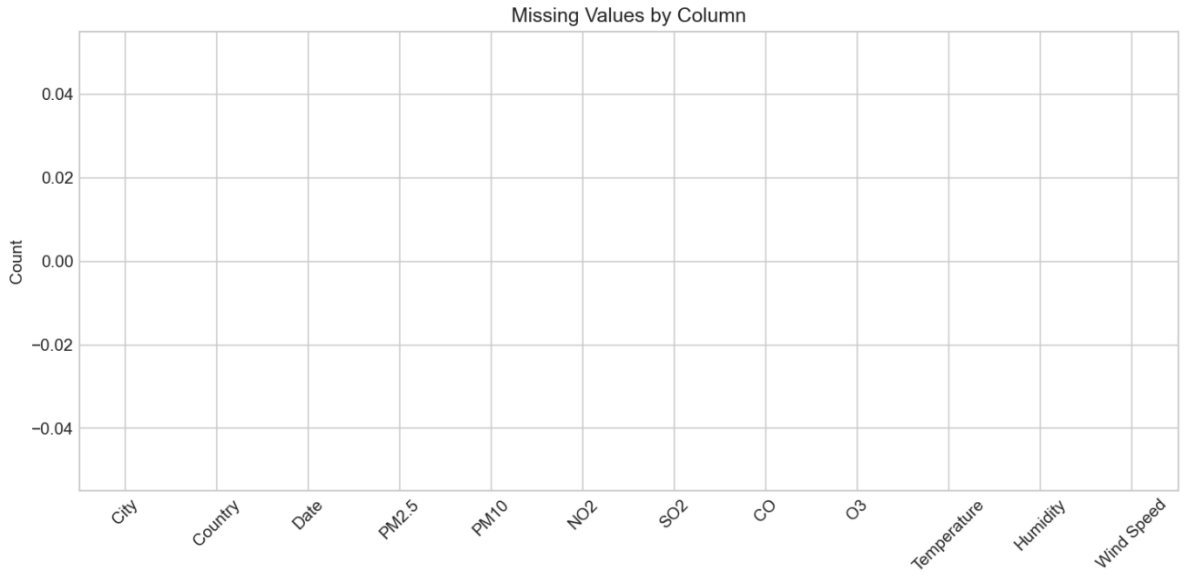
The following tools and libraries are used:

- Python programming language
- Pandas for data handling
- NumPy for numerical operations
- Matplotlib and Seaborn for data visualization
- Scikit-learn for preprocessing and machine learning

5. Data Preprocessing

5.1 Missing Value Analysis

The dataset was checked for missing values and no missing data was found.

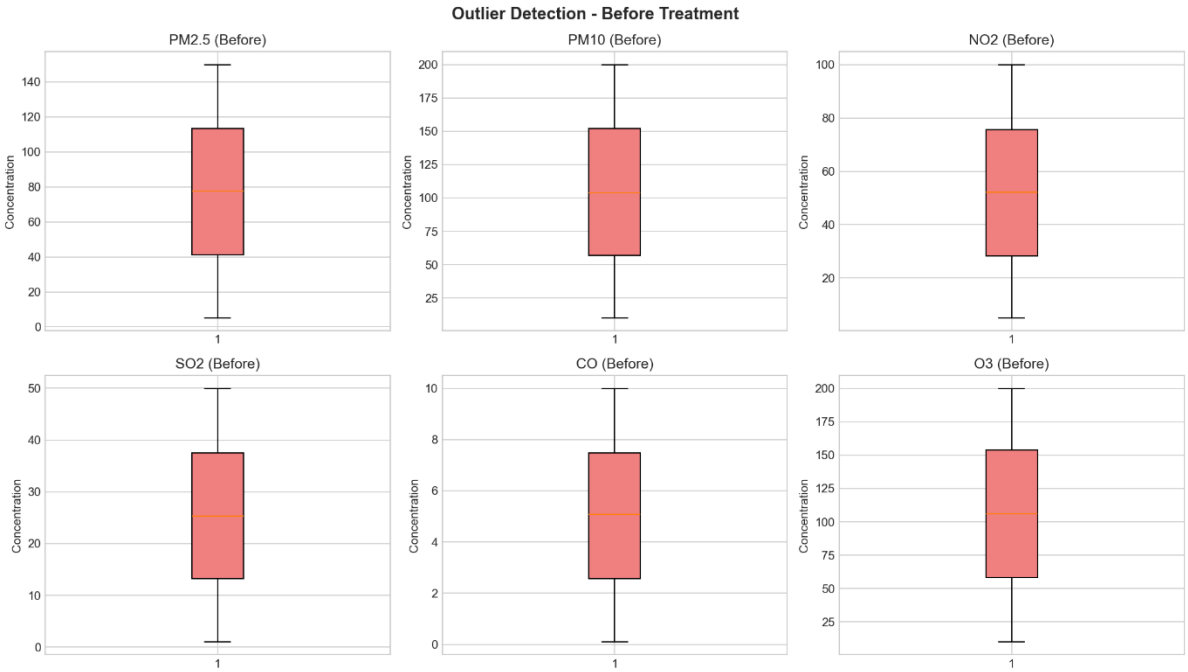


5.2 Outlier Detection and Treatment

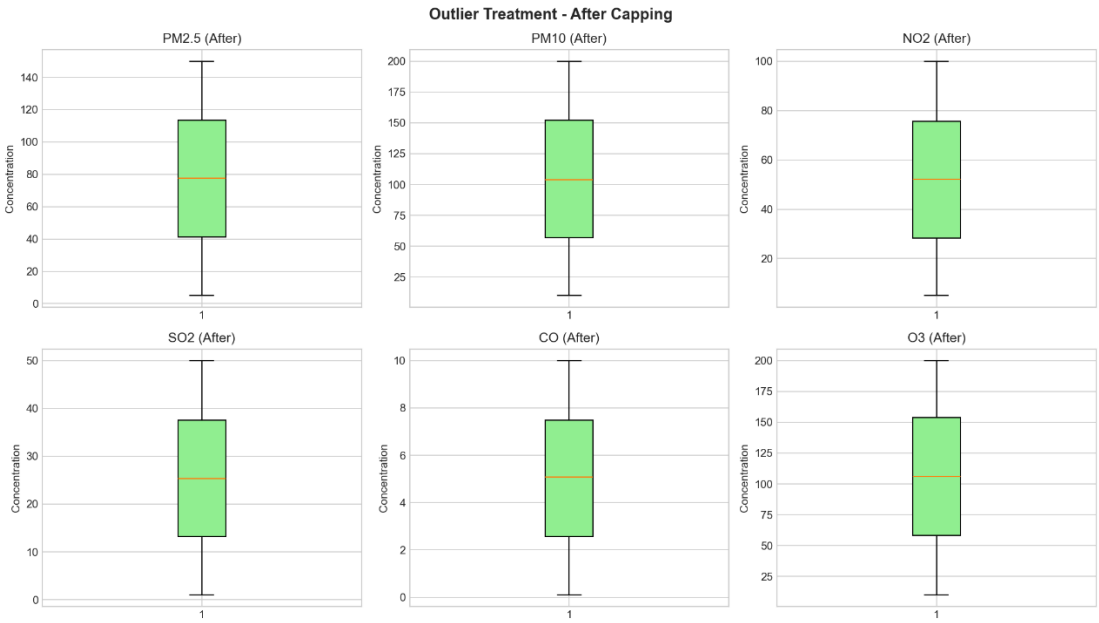
Outliers were detected using Box Plots, IQR method, and Z-score method. Since extreme values can affect analysis, capping was applied instead of removing data.

Graphs Used:

- Box Plot (Before Outlier Treatment)



- Box Plot (After Outlier Treatment)



6. Air Quality Index (AQI)

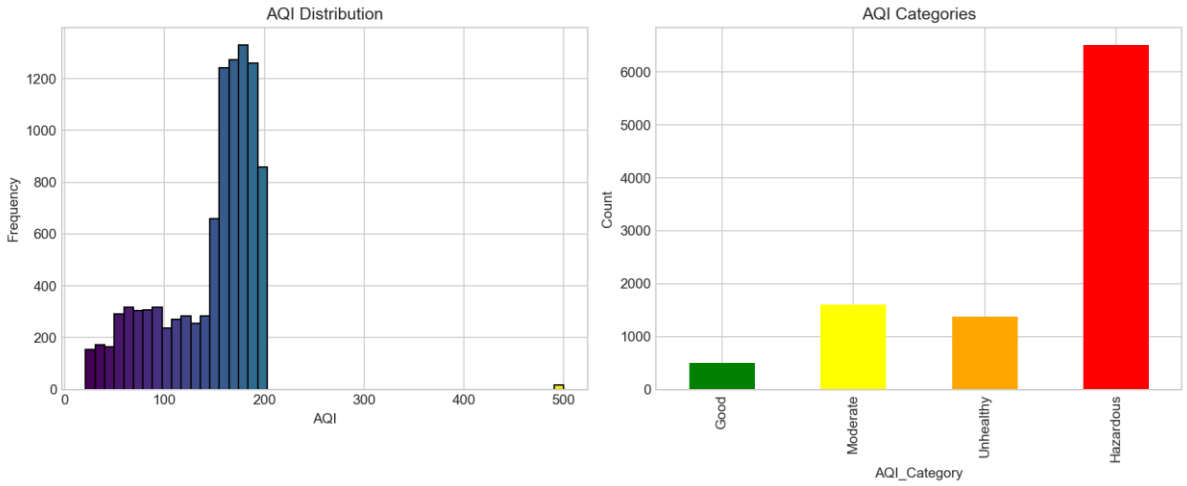
AQI is a numerical scale used to represent overall air pollution levels. In this project, AQI is calculated mainly using PM2.5 values according to standard breakpoints.

6.1 AQI Categories

- Good
- Moderate
- Unhealthy
- Very Unhealthy
- Hazardous

Graph Used:

- AQI Distribution Histogram
- AQI Category Bar Chart



7. Exploratory Data Analysis (EDA)

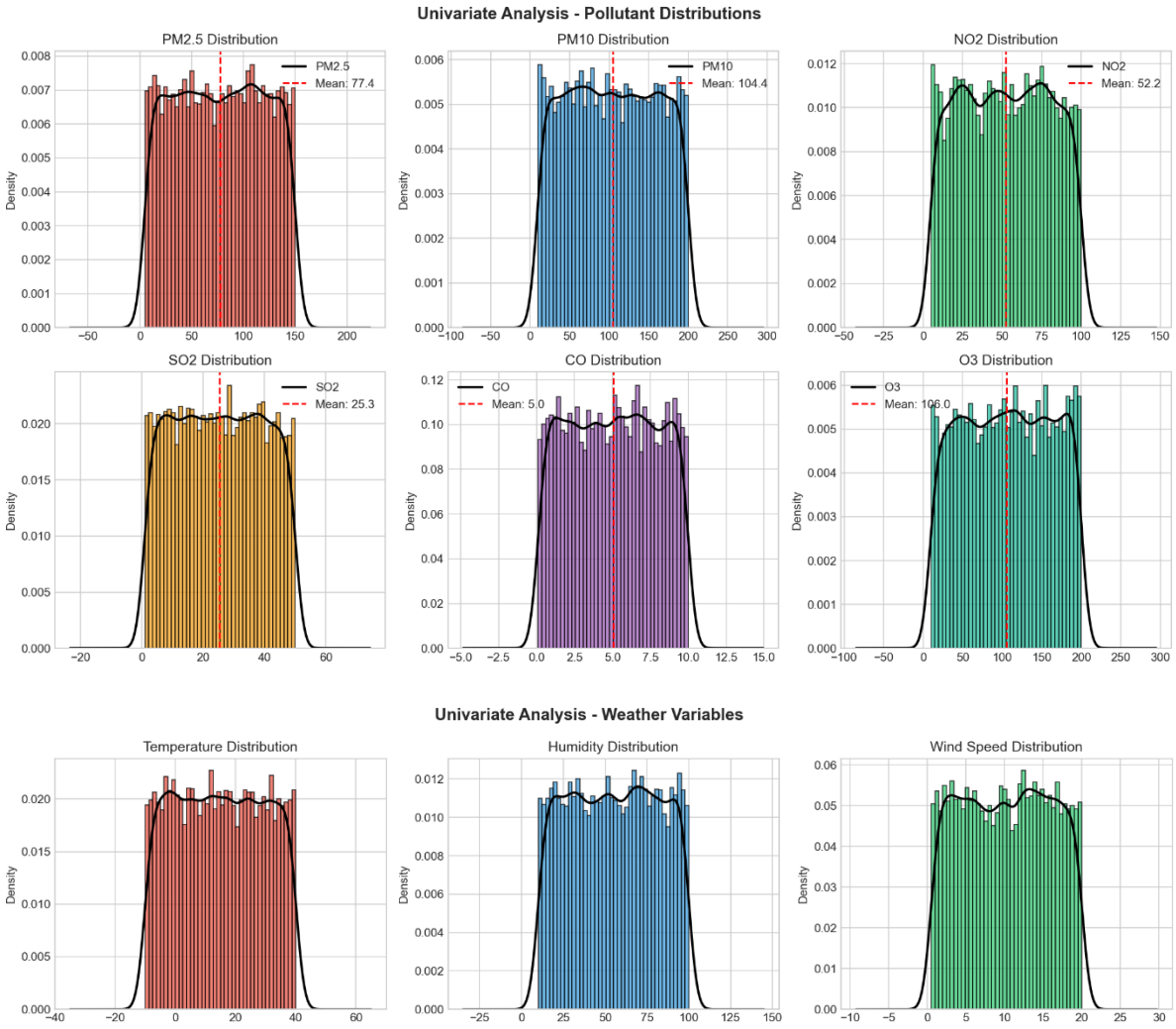
EDA is performed to understand data distribution, patterns, and relationships.

7.1 Univariate Analysis

Each pollutant is analyzed individually to study its distribution.

Graphs Used:

- Histogram with KDE Plot for PM2.5
- Histogram with KDE Plot for PM10
- Histogram with KDE Plot for NO2, SO2, CO, and O3

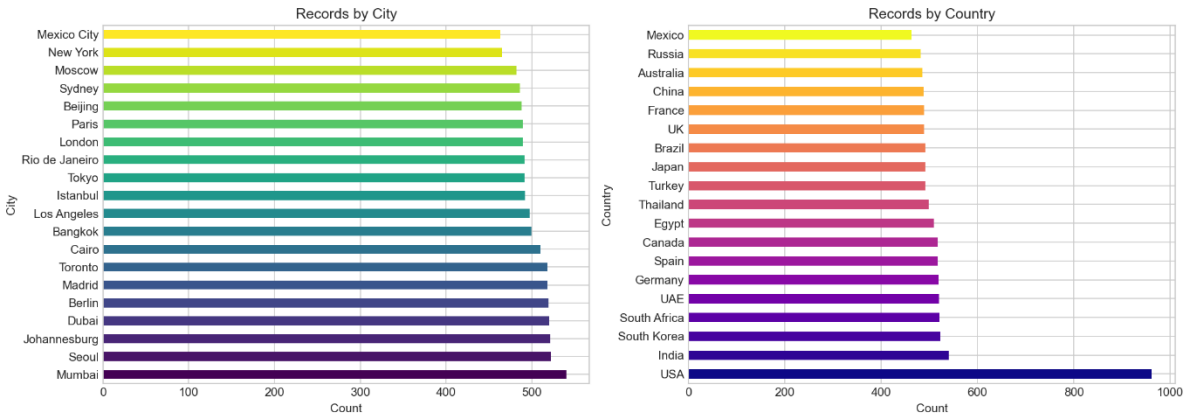


7.2 Categorical Analysis

City-wise and country-wise analysis is performed to study data distribution.

Graphs Used:

- City-wise Bar Chart
- Country-wise Bar Chart



8. Feature Scaling

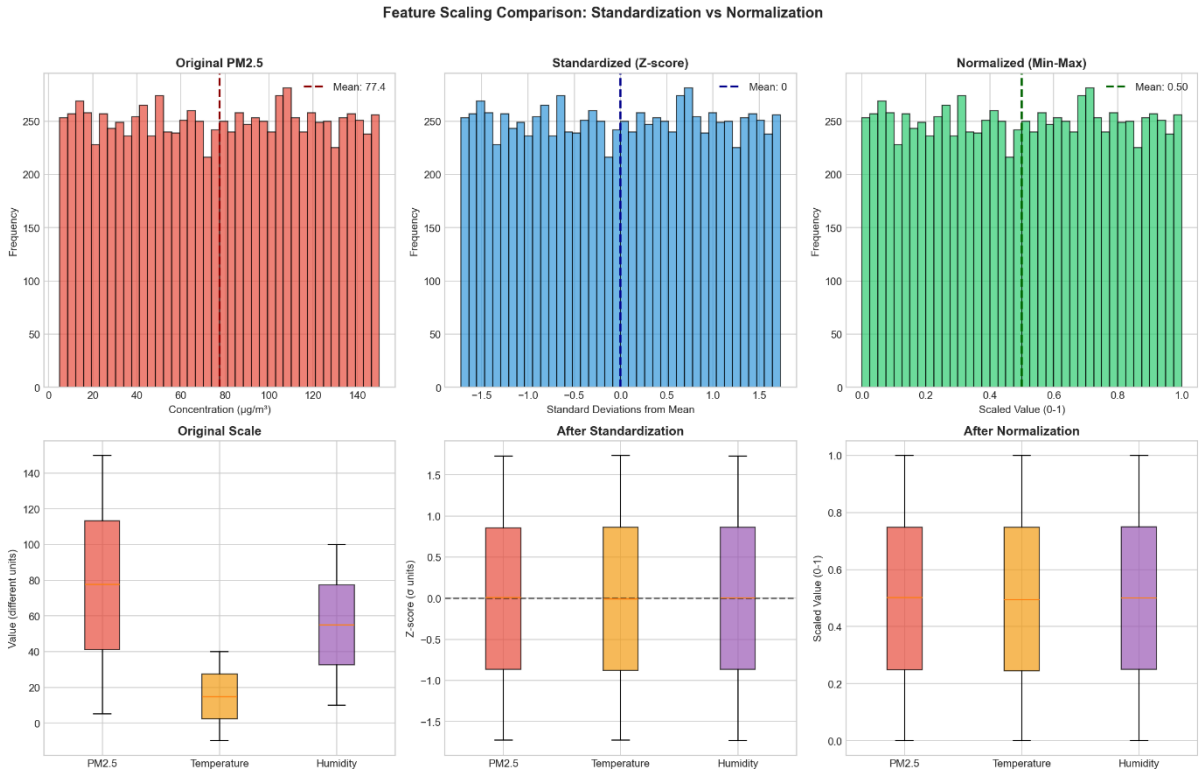
Feature scaling is applied to ensure all variables contribute equally to machine learning models.

Methods Used:

- StandardScaler
- MinMaxScaler

Graphs Used:

- Original Data Distribution Histogram
- Standardized Data Histogram
- Normalized Data Histogram



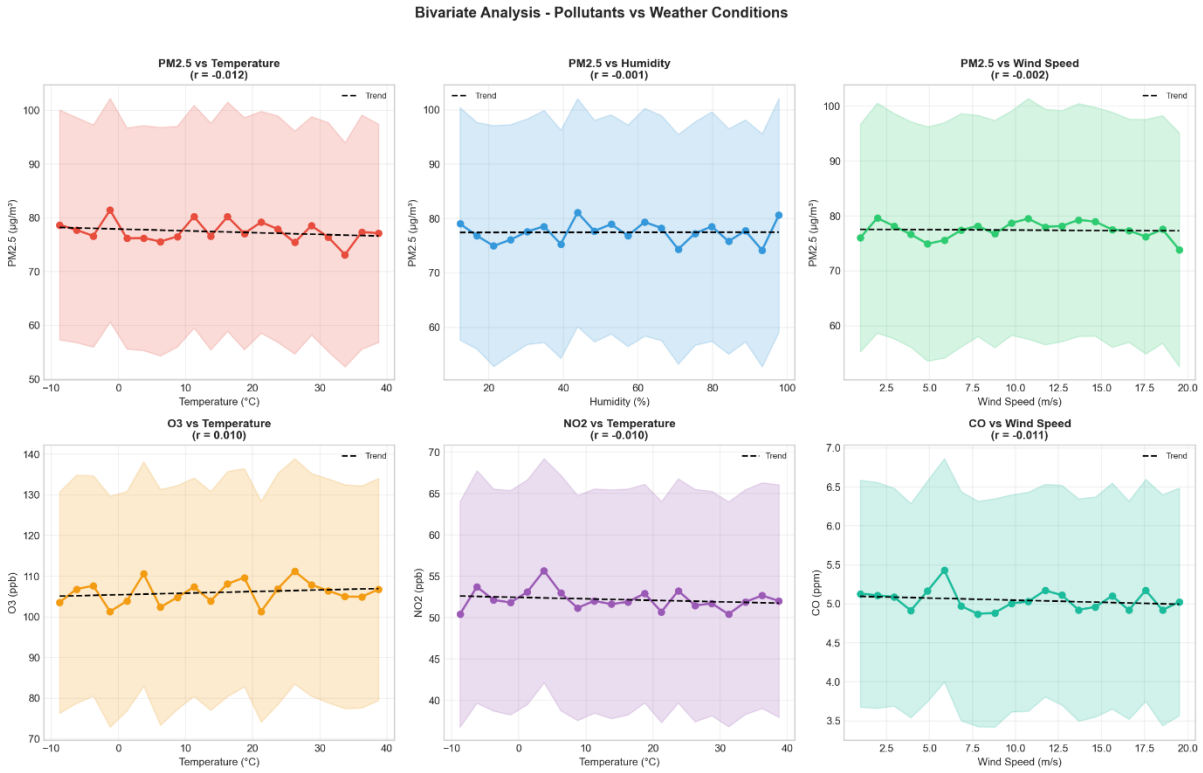
9. Bivariate Analysis

Bivariate analysis is used to study relationships between two variables.

Graphs Used:

- Line Plot (PM2.5 vs Temperature)

- Line Plot (PM2.5 vs Wind Speed)
- Line Plot (O3 vs Temperature)

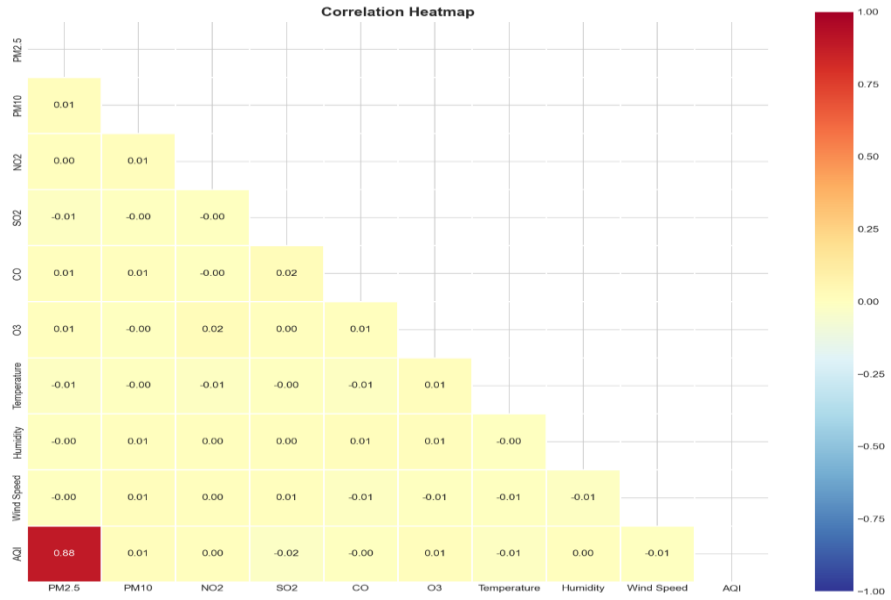


10. Correlation Analysis

Correlation analysis helps identify relationships between pollutants and AQI.

Graph Used:

- Correlation Heatmap



11. Model Building & Prediction

To fulfill the CCP requirement, five different machine learning models were implemented to predict AQI categories. Using multiple models helps in performance comparison and ensures robustness of results.

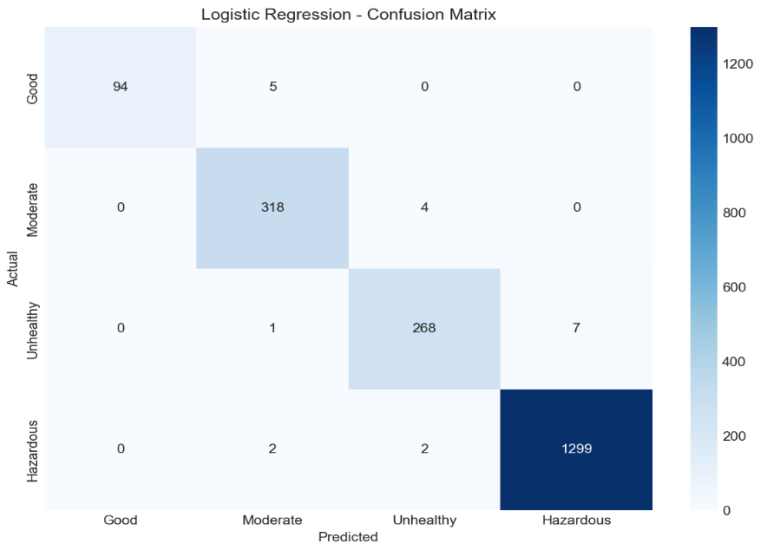
11.1 Logistic Regression

Logistic Regression is a linear classification algorithm used for categorical prediction problems. It estimates the probability of a class using a logistic (sigmoid) function. In this project, it serves as a baseline model for AQI classification.

Why used: Simple, fast, and interpretable.

Graph Used:

- Confusion Matrix for Logistic Regression



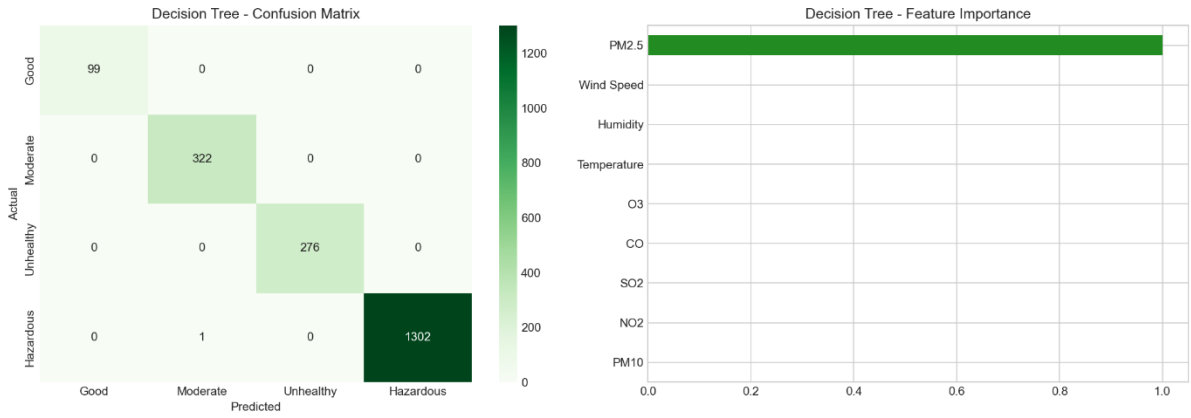
11.2 Decision Tree Classifier

Decision Tree is a non-linear model that splits data into branches based on feature values. It can capture complex relationships between pollutants and AQI.

Why used: Easy to understand and handles non-linearity well.

Graph Used:

- Confusion Matrix for Decision Tree



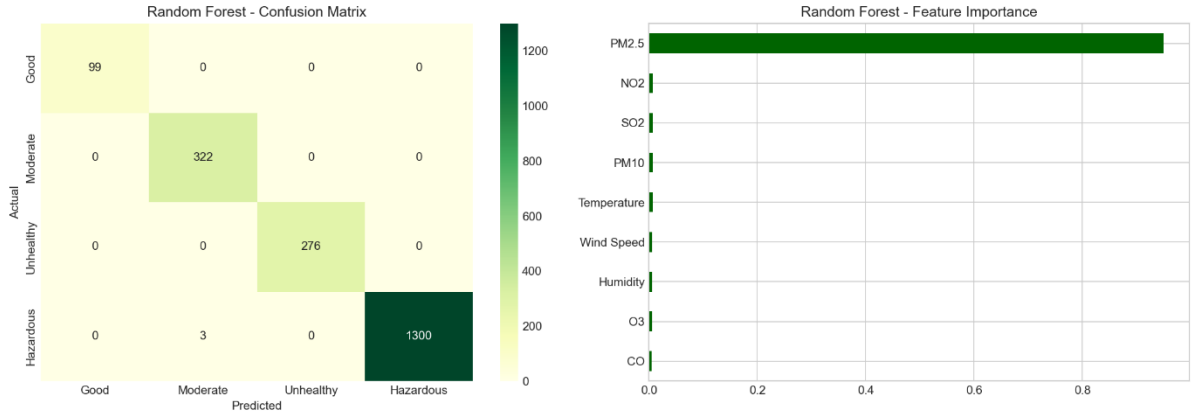
11.3 Random Forest Classifier

Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting.

Why used: High accuracy and robustness.

Graph Used:

- Confusion Matrix for Random Forest



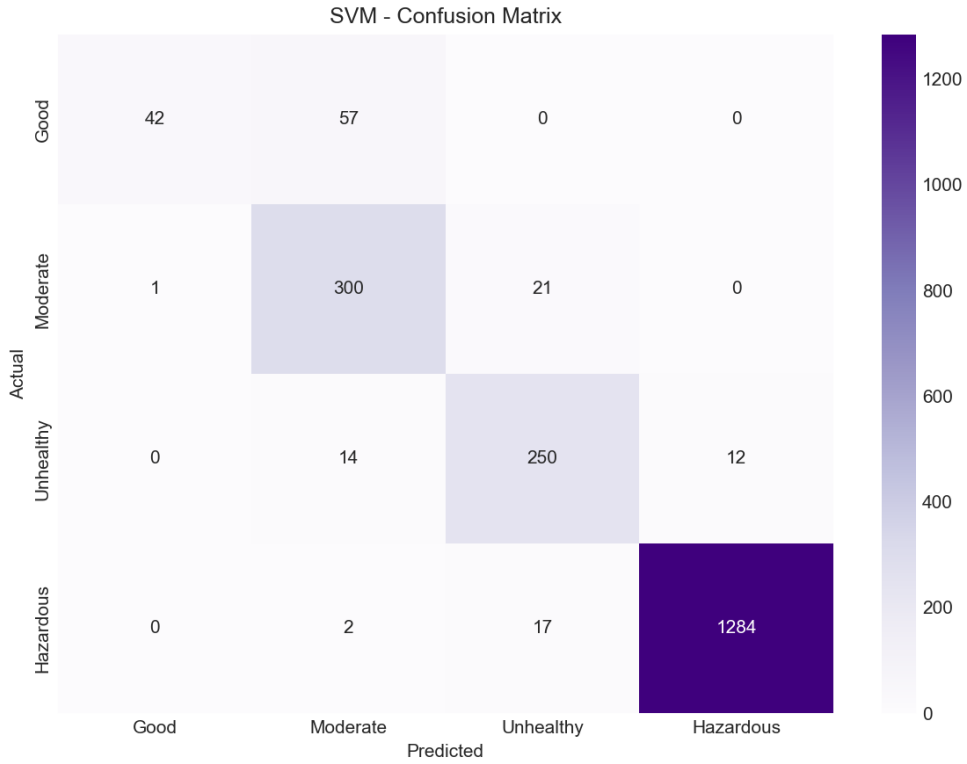
11.4 Support Vector Machine (SVM)

SVM finds an optimal hyperplane that separates classes with maximum margin. It performs well in high-dimensional spaces.

Why used: Effective for complex decision boundaries.

Graph Used:

- Confusion Matrix for SVM



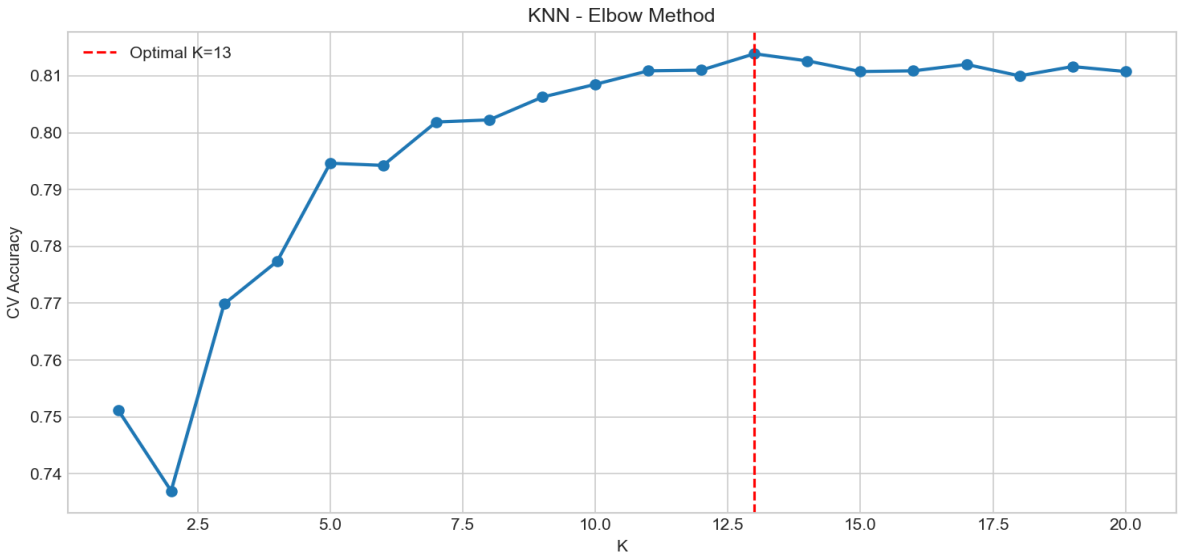
11.5 K-Nearest Neighbors (KNN)

KNN is an instance-based learning algorithm that classifies data based on the majority class of nearest neighbors.

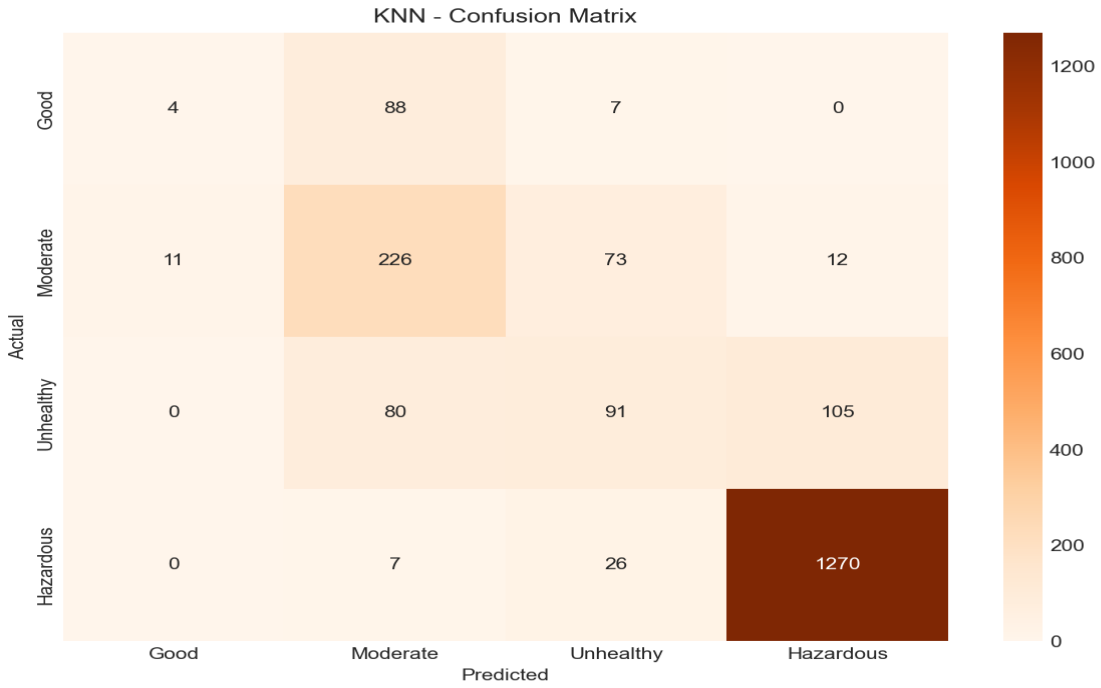
Why used: Simple and effective for pattern recognition.

Graph Used:

- K-Nearest Neighbors (KNN) with Elbow Method



- Confusion Matrix for KNN

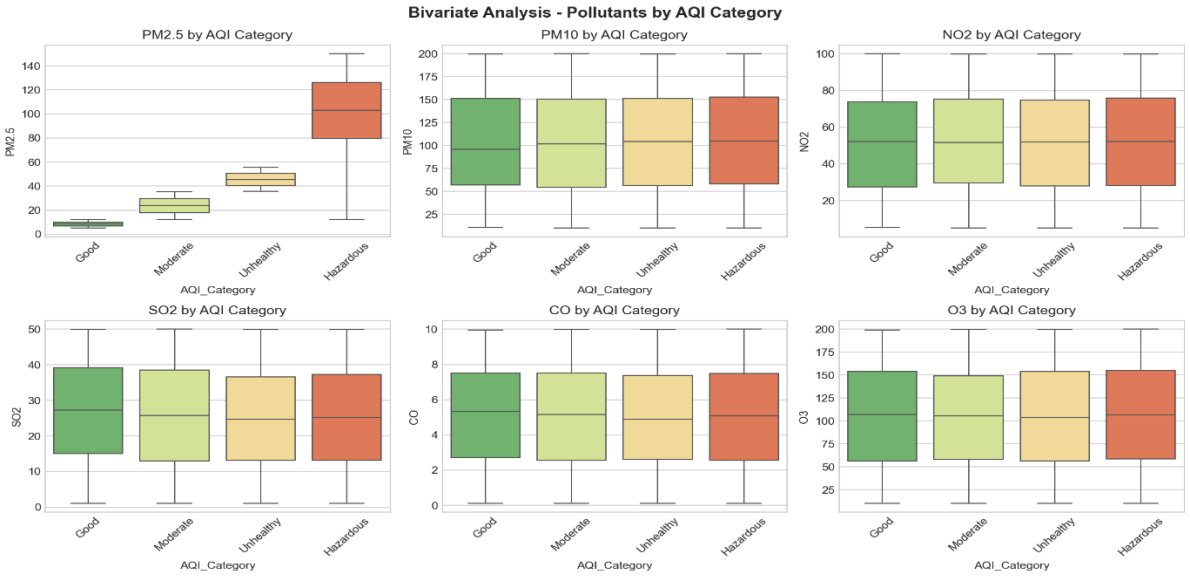


12. AQI Category-Based Analysis

Pollutant levels are compared across different AQI categories.

Graph Used:

- Box Plot of Pollutants by AQI Category

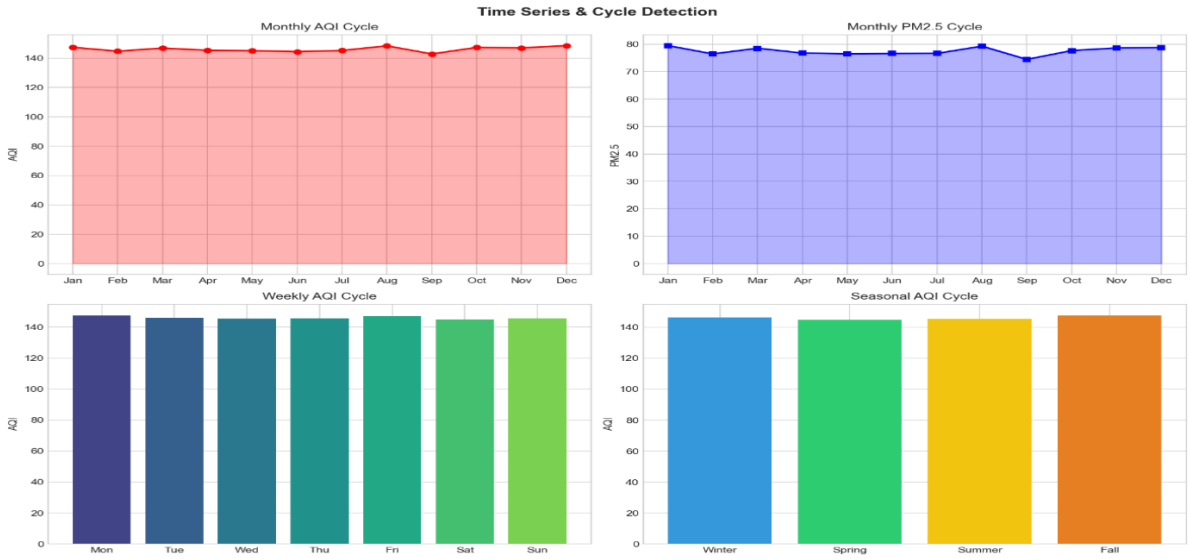


12. Time Series Analysis

Time-based analysis is performed to identify seasonal trends in air pollution.

Graphs Used:

- Monthly AQI Trend Line Graph
- Monthly PM2.5 Trend Line Graph

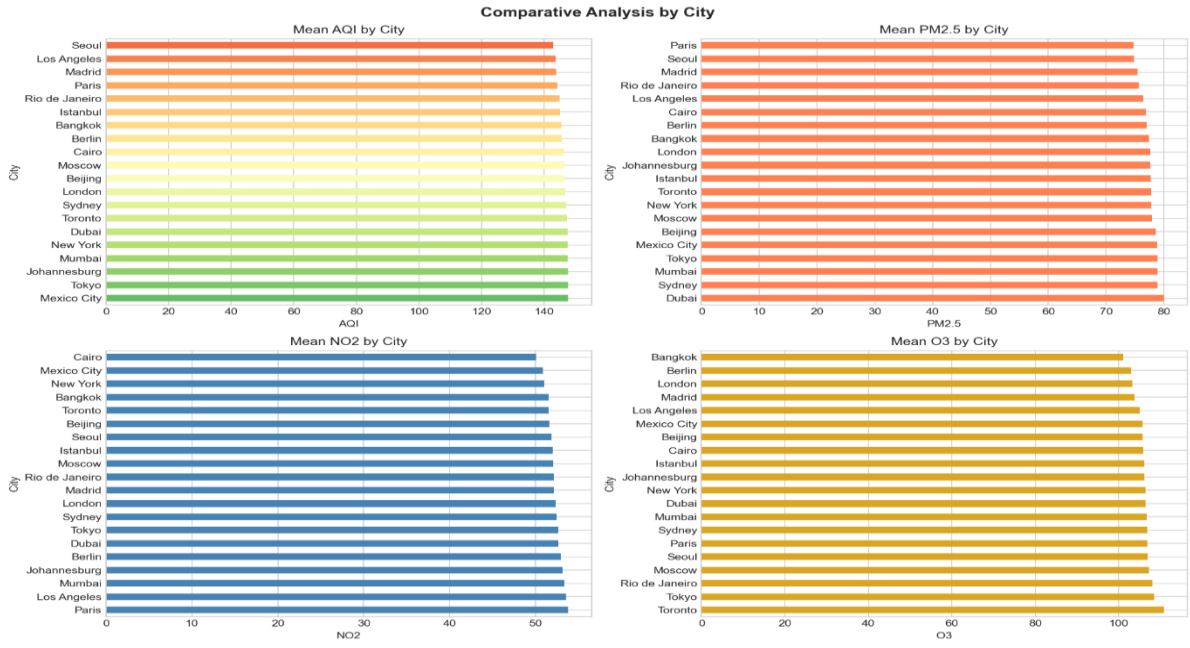


13. City-Wise Comparison

Average AQI values are calculated for each city to identify highly polluted cities.

Graph Used:

- City-wise Mean AQI Bar Chart



14. Model Interpretation and Recommendations

14.1 Model Interpretation

The machine learning models developed in this project were trained to classify air quality levels based on pollutant concentrations and meteorological factors. Among all implemented models, ensemble-based approaches such as **Random Forest** demonstrated superior performance due to their ability to capture non-linear relationships and interactions between multiple pollutants.

The analysis revealed that **PM2.5 is the most influential pollutant** in determining AQI categories, followed by **PM10, NO2, and O3**. These pollutants showed strong positive correlations with poor air quality levels. Meteorological variables such as **wind speed** showed a negative relationship with pollution levels, indicating that higher wind speeds help disperse pollutants.

Confusion matrix analysis confirmed that the models performed well in distinguishing between **Good**, **Moderate**, and **Unhealthy** air quality categories, with slightly lower performance for extreme classes due to class imbalance.

14.2 Impact of Pollutants on Human Health

The findings of this study align with established environmental health research:

- **PM2.5 and PM10** penetrate deep into the lungs, causing respiratory infections, asthma, chronic bronchitis, and cardiovascular diseases.
- **Nitrogen Dioxide (NO2)** increases the risk of lung inflammation and reduces immunity against respiratory infections.
- **Ozone (O3)** exposure can cause chest pain, coughing, throat irritation, and aggravate asthma.
- **Sulfur Dioxide (SO2)** contributes to breathing difficulties, especially in children and elderly individuals.
- **Carbon Monoxide (CO)** reduces oxygen delivery in the bloodstream, leading to dizziness and, in extreme cases, fatal poisoning.

High AQI categories indicate severe health risks, particularly for vulnerable populations such as children, elderly individuals, and patients with pre-existing respiratory or heart conditions.

14.3 Environmental Improvement Recommendations

Based on the analytical and predictive results of this project, the following strategies are recommended:

- **Strengthen vehicle emission regulations** and promote electric or hybrid vehicles.
- **Improve public transportation systems** to reduce traffic congestion and pollution.

- **Enforce industrial emission controls** through strict environmental policies.
- **Increase urban green spaces**, such as trees and parks, to naturally absorb pollutants.
- **Promote renewable energy sources** like solar and wind power.
- **Deploy real-time air quality monitoring systems** to inform the public and policymakers.

These measures can significantly reduce pollution levels and improve public health outcomes.

15. Conclusion

This project successfully analyzed global air quality data using statistical methods, data visualization techniques, and machine learning models. The calculation and categorization of Air Quality Index (AQI) enabled a clear understanding of pollution severity across different regions. Exploratory data analysis revealed meaningful patterns and correlations between pollutants and meteorological factors.

Multiple machine learning models were implemented and evaluated, with ensemble models showing strong predictive performance. The results demonstrate the effectiveness of data-driven approaches in environmental monitoring and decision-making. Overall, this study highlights the critical role of air quality analysis in protecting human health and supporting sustainable environmental policies.

16. Future Scope

The scope of this project can be expanded in the following ways:

- Integrate **real-time air quality data** from IoT sensors and government APIs.
- Apply **advanced machine learning and deep learning models** for improved prediction accuracy.
- Incorporate **health data** to directly study pollution-related disease impacts.
- Develop an **interactive web-based dashboard** for real-time visualization and public awareness.
- Extend the analysis to include **seasonal and long-term climate trends**.