Sulaiman Mathew-Wilson
003133656
Intro to Data Science
Dr. Nerolu
3 December 2025

Final Project Part I: Summary Report

Project Title: A Comprehensive Look into the State of Air Pollution Globally

Objective:

The objective of this project is to analyze global air pollution patterns using a dataset sourced from Kaggle. The goal is to identify the most polluted regions, explore relationships among pollutant Air Quality Index (AQI) values, evaluate the distribution of air quality categories, and visually communicate the insights that I found through data visualizations. Additionally, this project will serve as a way to practice and demonstrate the skills I have learned over the course of the semester in data wrangling, visualization, and storytelling using Python.

Introduction:

Air pollution remains one of the most pressing environmental and public health issues worldwide. Rapid industrialization, urbanization, and fossil fuel reliance have dramatically increased emissions of harmful pollutants, contributing to respiratory illnesses, cardiovascular disease, and premature death. Understanding global air quality patterns and identifying the areas that are most affected by air pollution is crucial for policy development, investment allocation, and public health protection.

This project uses a dataset compiled from global air quality monitoring sources. The dataset includes AQI values and pollutant-specific AQI indicators for multiple cities across different countries. By examining pollutants such as PM2.5, CO, Ozone, and NO2, this data analysis provides a multi faceted understanding of air pollution severity and its global distribution.

In order to gain a clearer understanding of patterns, this project integrates data cleaning techniques and visualizations. Visualizations utilized include bar charts, correlation heatmaps, scatterplots, pairplots, pie charts, and a global choropleth map to highlight key insights about pollution sources, severity, and geographic disparities.

Method:

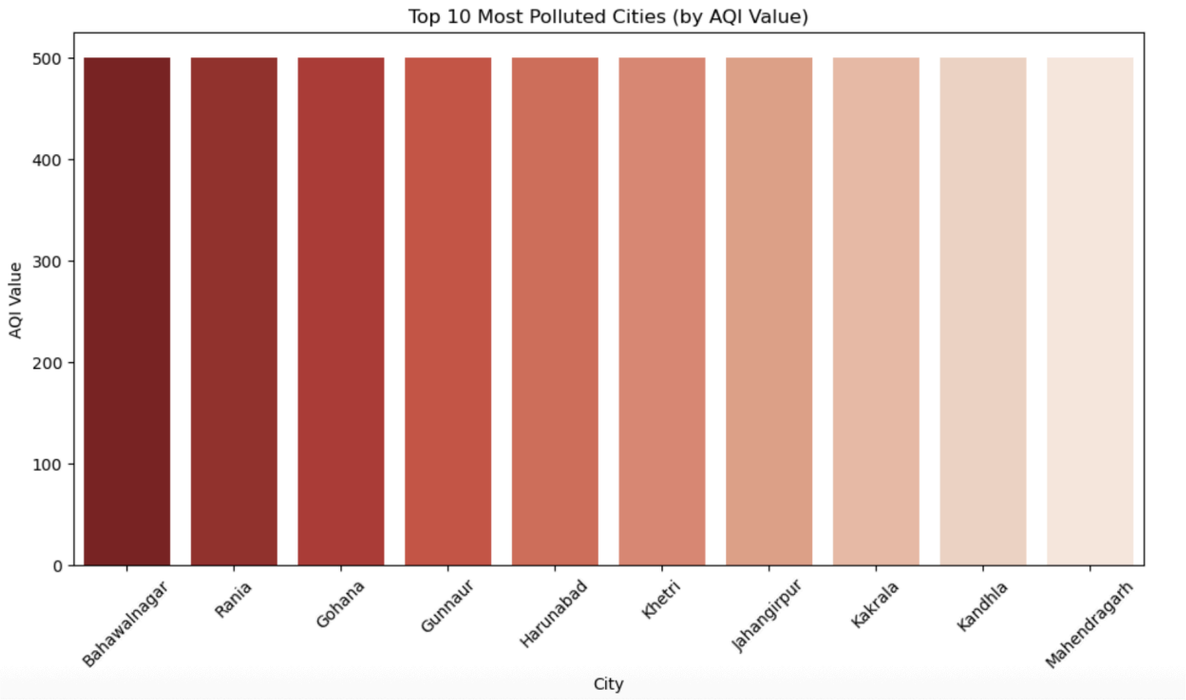
Missing values in the “Country” column were replaced with "Unknown" to preserve those rows for analysis. Rows missing City names were removed because city identifiers are essential for visualizations and cannot be reliably imputed. Pollutant AQI columns containing missing or non-numeric values were converted to numeric. These were retained when possible to avoid losing entire records, as long as the primary AQI value was present.

AQI pollutant columns were stored as strings due to the presence of non-numeric characters, but were converted to numeric values. Converting these columns was crucial for successful data visualizations, and enabled me to utilize correlation analysis, scatterplots, and descriptive statistics.

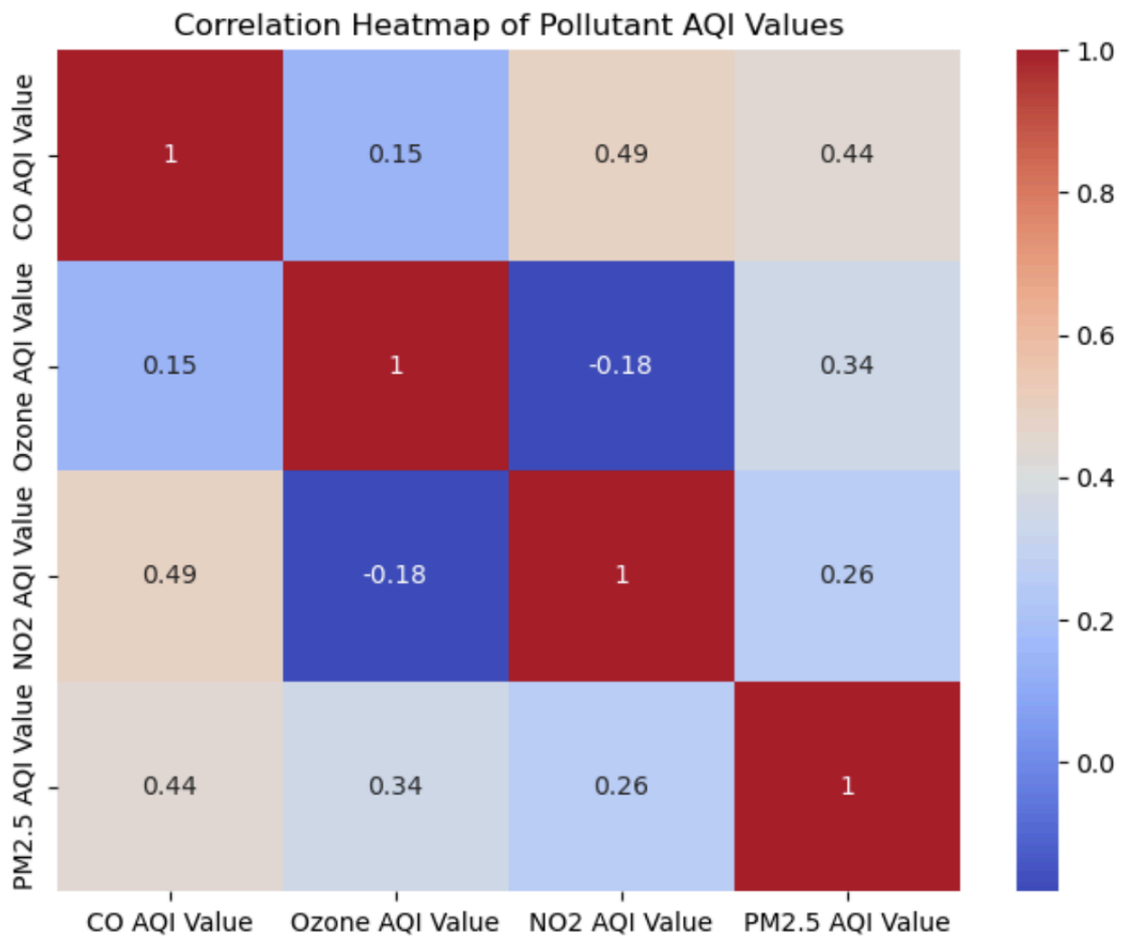
Duplicate entries were also removed to ensure accuracy in statistical analysis. Duplicate pollution readings could skew statistics, distorting the mean, maximum AQI values, and distribution categories.

Outliers were identified in the AQI and pollutant columns using statistical thresholds. Although outliers were not removed, recognizing them was important because extreme AQI events are meaningful in environmental analysis. High spikes in PM2.5 or NO_2 levels may signal hazardous pollution incidents, industrial accidents, or seasonal fire events. Keeping outliers preserved the integrity of real-world environmental variability.

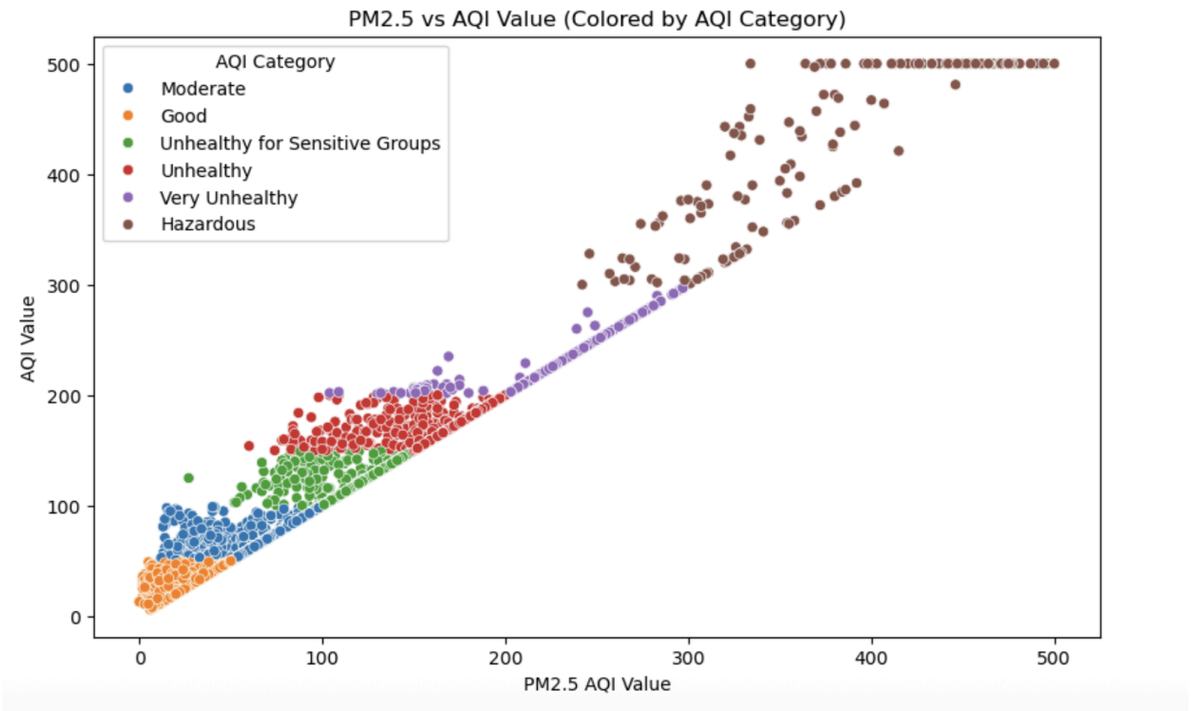
Storytelling:



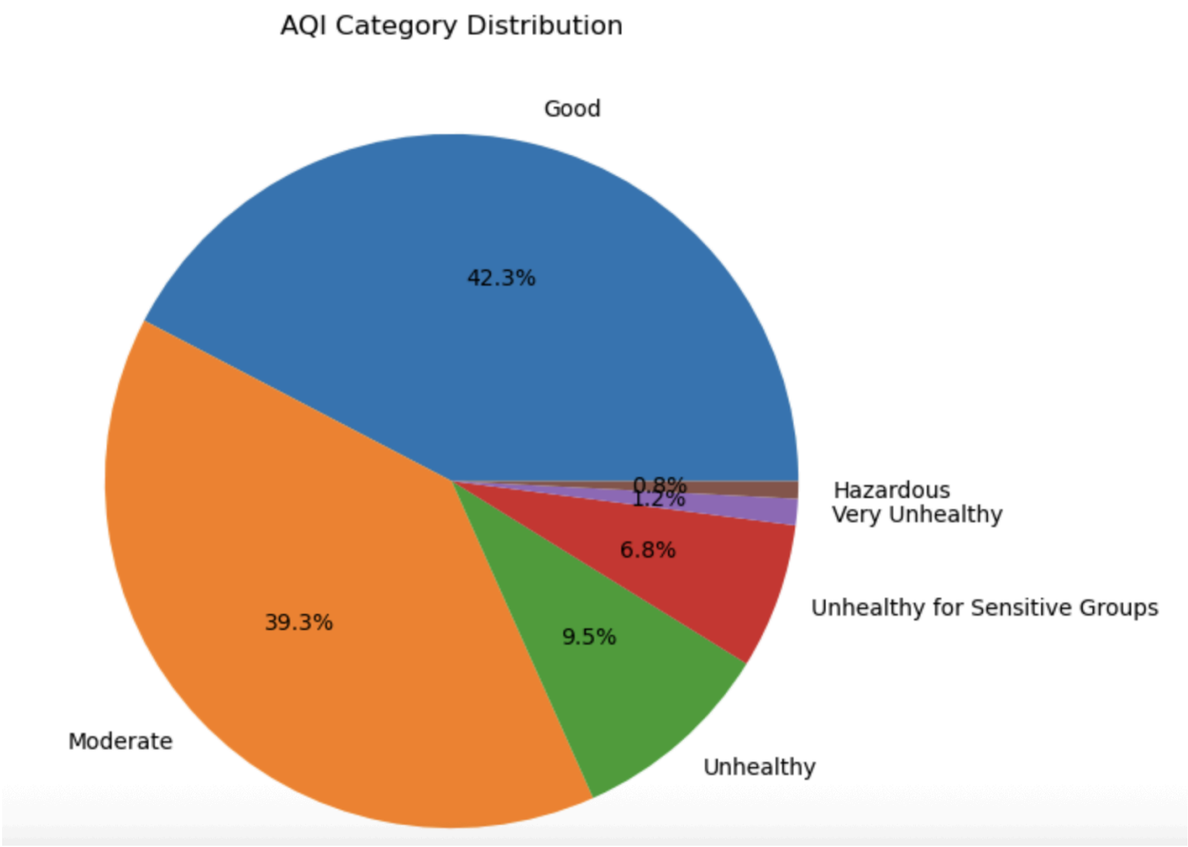
Top 10 Most Polluted Cities (Bar Chart): A bar chart ranking the top 10 cities by overall AQI value revealed which urban areas experience the most severe air pollution. Cities with AQI values approaching or exceeding the "Hazardous" threshold stood out. This visualization immediately highlighted global hotspots where residents face the highest exposure to harmful pollutants.



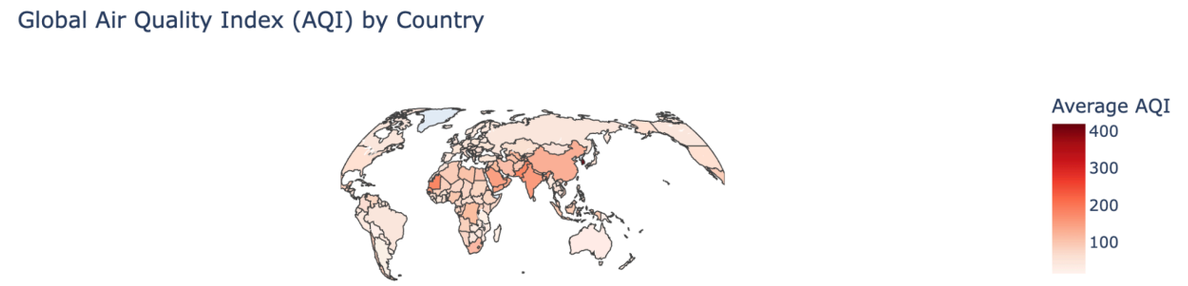
Correlation Heatmap of Pollutant AQI Values: A correlation heatmap was used to explore the relationships among pollutant AQI metrics: PM2.5, NO2, Ozone, and CO. Strong correlations were found between certain pollutants—particularly combustion-related emissions such as CO and NO2—suggesting common sources like vehicle emissions or industrial activity. PM2.5 showed very strong positive correlations with several pollutants, underscoring its role as a key indicator of poor air quality.



PM2.5 vs. Overall AQI Scatter Plot: A scatter plot comparing PM2.5 AQI values to overall AQI demonstrated a clear upward relationship. Cities with higher PM2.5 levels tended to have significantly higher AQI values. Coloring by AQI category provided additional insight into how particulate pollution drives dangerous air quality classifications such as “Unhealthy,” “Very Unhealthy,” and “Hazardous.”



AQI Category Distribution: Using country-level mapping, a choropleth visually communicated global differences in air quality. Regions such as South and East Asia displayed noticeably higher average AQI values, while parts of Europe and North America generally appeared cleaner. The map provided strong geographic context, allowing viewers to identify continental patterns and regional disparities.



Global Choropleth Map of AQI Values:

A chart showing the distribution of AQI categories highlighted how frequently cities fall into "Good," "Moderate," "Unhealthy," or more hazardous classifications. This analysis showed that a substantial proportion of global air quality readings fall into categories above the "Moderate" level, raising concerns about long-term exposure and public health risks.

Conclusion:

This project provides a detailed exploration of global air pollution using a large, multi-variable dataset from Kaggle. By applying systematic data cleaning procedures, including handling missing values, converting data types, removing duplicates, and standardizing column names, the dataset was transformed into a reliable foundation for analysis.

Visualizations revealed important environmental insights:

- Several cities experience dangerously high AQI levels.
- PM2.5 plays a major role in determining overall air quality.
- Many pollutants are closely correlated, reflecting shared emission sources.
- Air pollution is unevenly distributed across the globe, with certain regions experiencing significantly poorer air quality.

Taken together, these findings underscore the importance of environmental monitoring, emission control strategies, and policy interventions aimed at improving global air quality. This analysis not only demonstrates the power of data science tools in uncovering environmental insights but also highlights the urgency of addressing air pollution as a global public health challenge.

References:

Kaggle. *Global Air Pollution Dataset.* Retrieved from: https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset/code

U.S. Environmental Protection Agency (EPA). *Air Quality Index (AQI) Basics.*

World Health Organization (WHO). *Global Air Quality Guidelines.*

Acknowledgements:

I would like to acknowledge my instructor, Dr. Nerolu, for her guidance, instruction, and feedback throughout the course.