USC University of Southern California

DSCI 510: PRINCIPLES OF PROGRAMMING FOR DATA SCIENCE

FINAL REPORT

How are population statistics and weather variables correlated with air pollution levels across major global cities?

Team Members:

1. Dubi Sao [3931757486]
2. Manasa Vijayaraghavan [2685584788]

How are population statistics and weather variables correlated with air pollution levels across major global cities?

This project examines global urban air quality by collecting environmental data from hundreds of cities worldwide. $PM_{2.5}$ concentrations are used to calculate Air Quality Index (AQI) values in accordance with U.S. Environmental Protection Agency (EPA) standards. The study analyzes variations in air quality across cities, population sizes, temperature conditions, and pollutant types, utilizing statistical analysis and visualizations.

Data Collection:

City-level environmental and demographic data were collected for large global cities from. The final cleaned dataset includes **787 cities**, reduced from the original 822 samples following data cleaning and deduplication.
Each city record includes:

- City name and country
- Population
- Latitude and longitude
- Current temperature, wind speed, and wind direction
- Air pollutant concentrations: $PM_{2.5}$, PM_{10}, NO_2, and O_3
- Derived $PM_{2.5}$-based AQI and categorical air quality labels

Data Sources and Approach

The dataset was created using the following sources:

1. **City and Population Data**
   - Source: World Population Review
   - URL: https://worldpopulationreview.com/cities
   - Approach: Web scraping using requests and BeautifulSoup to extract city names, countries, and population values from an HTML table.
2. **Geographic Coordinates**
   - Source: Nominatim (OpenStreetMap)
   - URL: https://nominatim.openstreetmap.org/search
   - Approach: API used to retrieve latitude and longitude for each city using city–country pairs.

3. **Weather Data**
    - Source: Open-Meteo Weather API
    - URL: https://api.open-meteo.com/v1/forecast?
    - Approach: API used to retrieve temperature, winds peed, wind direction
4. **Air Quality Data**
    - Source: Open-Meteo Air Quality API
    - URL: https://air-quality-api.open-meteo.com/v1/air-quality
    - Approach: API used to retrieve $PM_{2.5}$, PM_{10}, NO_2, O_3

Raw data was saved to CSV files, then cleaned and standardized before analysis.

Changes from initial proposal:

The initial methodology involved retrieving AQI values directly from the WAQI API. However, access restrictions and reliability concerns necessitated a change in approach. Consequently, AQI values were calculated programmatically using standard EPA formulas, while pollutant concentrations were sourced from the Open-Meteo Air Quality API.

Data Cleaning and Processing:

1) Numeric variables were standardized by converting all relevant fields to numeric types using the **pandas** library.
2) Missing values were removed using the **dropna()** function in pandas.
3) Duplicate entries were removed using the **drop_duplicates()** function in pandas.
4) The resulting dataset is stored and processed as a **pandas DataFrame.**

Analysis and Visualization:

Analytical Methods Employed

The study utilized the following analytical techniques:

- EPA-based AQI calculation using $PM_{2.5}$concentration thresholds
- Categorical classification of AQI severity levels
- Grouped statistical summaries (mean AQI by population group and temperature category)
- Correlation analysis between AQI, pollutants, and weather variables
- City-level ranking to identify most and least polluted cities

- Exploratory visualization using scatter plots, box plots, and correlation heatmaps

$PM_{2.5}$ -Based AQI Computation

AQI values were determined using official U.S. EPA breakpoint formulas, which facilitated direct comparison of pollution severity across cities.

$$AQI = \frac{I_{hi} - I_{lo}}{C_{hi} - C_{lo}} \left(C - C_{lo}\right) + I_{lo}$$

Here,

C- represents the $PM_{2.5}$ concentration

C_{lo}, C_{hi}, I_{lo}, I_{hi} **- are the corresponding concentration and AQI breakpoints defined by U.S. EPA**

Key Findings:

Megacities demonstrate the highest average Air Quality Index (AQI) values. Temperature and wind speed display only weak associations with AQI. The most polluted cities are concentrated in South Asia, East Asia, and parts of Africa, whereas the cleanest cities are primarily found in North America and Northern Europe. These findings underscore pronounced regional disparities in global urban air quality.

Description of Visualizations:

Several targeted visualizations were developed to facilitate the analysis:

- A correlation heatmap summarizes the relationships among AQI, pollutants, and weather variables, emphasizing the significant influence of $PM_{2.5}$.
- A scatter plot depicting AQI versus temperature for the upper quartile of values illustrates the relationship between extreme pollution levels and temperature, offering preliminary validation of the AQI calculations.
- Boxplots of AQI distribution by population group demonstrate differences in pollution distributions across city sizes, with higher medians and greater variability observed in mega cities.
- AQI distributions across temperature categories enable comparison of pollution levels in cold, moderate, and hot climates.
- A ranked bar chart of the ten most polluted cities highlights geographic concentrations of extreme air pollution.

Each figure is constructed to address a specific research question.

Observations and Conclusions

Air quality varies significantly across cities of different sizes and climatic conditions. Larger urban areas typically exhibit higher and more variable Air Quality Index (AQI) levels. Additionally, cities in moderate-temperature regions tend to have higher average AQI values than those in extremely cold or hot climates. Although weather factors such as temperature and wind speed affect air quality, they do not constitute the primary determinants. Instead, urban and regional characteristics may serve as more influential drivers of pollution.

Impact of Findings:

The results of this project yield several significant implications:

- **Public health awareness:** Identification of high-risk cities enables prioritization of targeted mitigation efforts.
- **Urban planning:** The analysis underscores the environmental costs associated with rapid urbanization.
- **Policy support:** The findings inform data-driven decision-making for air quality regulation.

Future Scope:

With additional time, the project could be enhanced through the following approaches:

- Incorporating time-series air quality index (AQI) data rather than relying solely on single-point observations
- Conducting regional or country-level comparative analyses
- Expanding the analysis to include additional pollutants, such as carbon monoxide (CO) and sulfur dioxide (SO_2)
- Developing an interactive dashboard to facilitate exploratory data analysis
- Examining both seasonal and long-term trends in air pollution