

DSCI 510: PRINCIPLES OF PROGRAMMING FOR DATA SCIENCE

FINAL PROJECT PROPOSAL

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]

Problem Statement:

Major cities around the world differ greatly in terms of air quality, climate, and population density - all of which are important aspects of urban planning and public health. The data is dispersed across several sources like - Wikipedia, weather APIs, environmental services.

This project addresses this by creating a unified, programmatically generated dataset that combines:

- Weather data
- Air quality and pollution indicators
- City population statistics

A clean, analysis-ready environment dataset for significant comparisons between major global cities will be produced by integrating web-scraped data with open APIs.

Collection of Data:

Two primary categories of data sources will be used to construct the dataset:

- A. Metadata and City Population (Static Web Scraping):

The city name, nation, and population will be scraped from a static page like Wikipedia using Python, BeautifulSoup, and requests

- B. Environmental Information (Public APIs):

Two free REST APIs will be used to gather environmental indicators:

- Weather (Open-Meteo API): Obtains temperature, wind direction, and wind speed.
- Air Quality (WAQI API): Calculates pollution levels using coordinates, and indicates the AQI, PM2.5, PM10, NO₂, and O₃ levels.

Analysis and Visualization:

1. Analysis:

The unified dataset (CSV) will be used to rank cities according to environmental stress indicators (AQI, PM2.5), find correlations, and identify geographic patterns.

2. Visualizations:

- Scatter Plots (i.e., AQI vs. Temperature, Population vs. PM2.5).
- Bar Charts (i.e., Top 10 most polluted cities, country comparisons).
- Heatmap for the correlation matrix between all variables.