Analyzing Global Disparities in Air Quality

Project Draft

Hendrick Octavius
Data Science
Wentworth Institute of Technology
Boston, MA, United States
octaviush@wit.edu

ABSTRACT

The project aims to analyze and visualize annual ambient air pollution across cities using the dataset tiled, “WHO Ambient Air Quality Database 2023.” It compiles mean concentrations of key pollutants in over 100 countries. The analysis focuses on identifying regions with the best and worst air quality, determining how many cities meet World Health Organization safety standards, and exploring the relationship between national wealth and pollution levels. Additionally, the study incorporates a quantitative modeling component utilizing clustering algorithms to categorize countries based on pollutant profiles, thereby identifying the main sources of air pollution across different urban environments.

KEYWORDS

World Health Organization (WHO), Air Quality, Pollutants, Clustering, Disparities.

1 Introduction

This project, entitled "Analyzing Global Disparities in Air Quality," focuses on analyzing and visualizing annual ambient air pollution levels across cities around the world. It utilizes data gathered by the World Health Organization (WHO) to understand the public health implications of air quality and explores the relationship between national wealth and pollution levels. Additionally, this analysis will classify cities based on pollutant profiles through a quantitative clustering method.

2 Data

The dataset for this project was obtained from Kaggle and was last updated in 2023 and is entitled, “WHO Ambient Air Quality Database”. This dataset contains annual mean concentrations of key air pollutants from monitoring stations in over 100 countries. By tracking these specific metrics across different cities, the data enables the quantitative assessment of air quality disparities and the identification of regions falling short of the standards established by the WHO.

2.1 Source of dataset

The dataset was created with data from he World Health Organization and made available on Kaggle by the user called Nataly Reguerin.

2.2 Characters of the datasets

The dataset is displayed in tabular form in an excel file of less than 4MB. The columns relevant to our analysis are

- who_region – Geographical areas recognized by the WHO
- country_name – Name of the countries in which cities are located
- city – Name of the cities that will be considered for our analysis

- pm10_concentration – polluting particulate matter with a diameter of 10 micrometers
- pm25_concentration – polluting particulate mater with a diameter of 2.5 micrometers
- no2_concentraion – Nitrous Oxide level in the area
- population – Number of residents in a city

3 Methodology

K-Means Clustering algorithm will be implemented as our quantitative clustering model. This unsupervised learning method is chosen to categorize cities and countries based on their distinct pollutant profiles. The analysis will be conducted using Python within a Jupyter Notebook. The following libraries will be utilized: Pandas Matplotlib, Seaborn, and Scikit-learn.

3.1 Heading Level 2

3.2 Heading Level 2

4 Results

4.1 Heading Level 2

5 Discussion

6 Conclusion

ACKNOWLEDGMENTS

REFERENCES