Analyzing Air Quality: Insights into the Impact of Weather and Traffic in Big Cities

Understanding the Dataset

Atreyo Das
das.at@northeastern.edu
Omer Seyfeddin Koc
koc.o@northeastern.edu
Shruti Suhas Kute
kute.s@northeastern.edu

Instructor: Dr. Fatema Nafa

Northeastern University, Boston MA
Master of Science in Data Science

Github Link

1 Dataset Overview and Project Motivation

This project analyzes air quality by integrating publicly available traffic and weather data from New York, Chicago, and Los Angeles. These cities were selected due to their high population densities, significant traffic congestion, and availability of structured datasets, making them ideal for studying the relationship between traffic, weather, and air pollution.

1.1 Dataset Overview

The dataset was created by combining historical records from the following sources:

- **Air Quality Data:** Air Quality Data
- **Weather Data:** Weather Data
- **New York Traffic Data:** NYC Automated Traffic Volume Counts

The dataset consists of three main categories:

- **Traffic Data:** Vehicle counts recorded at specific locations.
- **Pollutant Data:** Measurements of PM2.5, PM10, and NO2 concentrations.
- **Weather Data:** Parameters such as temperature, wind speed, and humidity.

All data is stored in `.csv` or Excel formats, and no APIs were used for acquisition. The integration of these datasets enables a comprehensive analysis of urban air quality trends.

1.2 Project Motivation

This project is driven by the need to understand and mitigate urban air pollution, which poses significant public health challenges. Key motivations include:

- **Analyzing Environmental Challenges:** Examining how traffic and weather contribute to pollution levels in major cities.
- **Enhancing Prediction Models:** Improving air quality predictions by integrating traffic data and using advanced machine learning models.
- **Informing Public Policy:** Providing actionable insights to help policymakers reduce pollution and protect urban populations.

The insights from this study are expected to contribute to better urban planning and scalable solutions for managing air quality in metropolitan areas.

1.3 Alignment with Goals

The dataset has been carefully selected to align with the project's objective of analyzing air quality by integrating traffic and weather data. It includes key variables such as pollutant concentrations (PM2.5, PM10, NO2), vehicle counts, and meteorological parameters (temperature, wind speed, and humidity), making it well-suited for understanding the relationship between urban pollution, traffic congestion, and weather patterns.

2 Data Preprocessing and Transformation

The dataset underwent several preprocessing steps to ensure its quality, consistency, and readiness for analysis:

- **Handling Missing Data:** Interpolation techniques were used to address missing values, ensuring no gaps that could compromise the integrity of the models.
- **Removing Duplicate Records:** Duplicate entries were identified and removed to eliminate redundancy and maintain the reliability of the dataset.
- **Data Normalization:** All features were scaled to a consistent range through normalization, enabling fair comparisons across variables with different units and magnitudes.
- **Dataset Integration:** Data from multiple sources, including pollutant measurements, meteorological variables, and traffic data, were combined to create a unified and comprehensive dataset.

These steps provided a robust foundation for accurate air quality prediction and analysis.

3 Data Limitations

While the dataset is quite robust, it does have some drawbacks like:

- **Missing Data:** The dataset seems to be missing values for the pollutants of certain regions. The values of that part had to be supplemented from other sources to avoid major gaps in the data. Methods like imputation can also be used to fill in some missing information.
- **Large Size of Dataset:** The Air Quality dataset provided by EPA is divided based on their pollutants and the year they were recorded on. Each of these datasets have millions rows of data. While having this kind of data can be very useful, it makes it much harder to handle such datasets because things like outliers or repeating data becomes very hard to track.
- **Geographical Bias:** The dataset and the project only focuses on a limited amount of counties for the air quality data which reduces its generalizability to other regions with different environmental or socioeconomic conditions. This geographic limitation could introduce sampling bias and restrict the applicability of findings to a broader population.