Impact of Air Pollution on Respiratory Diseases

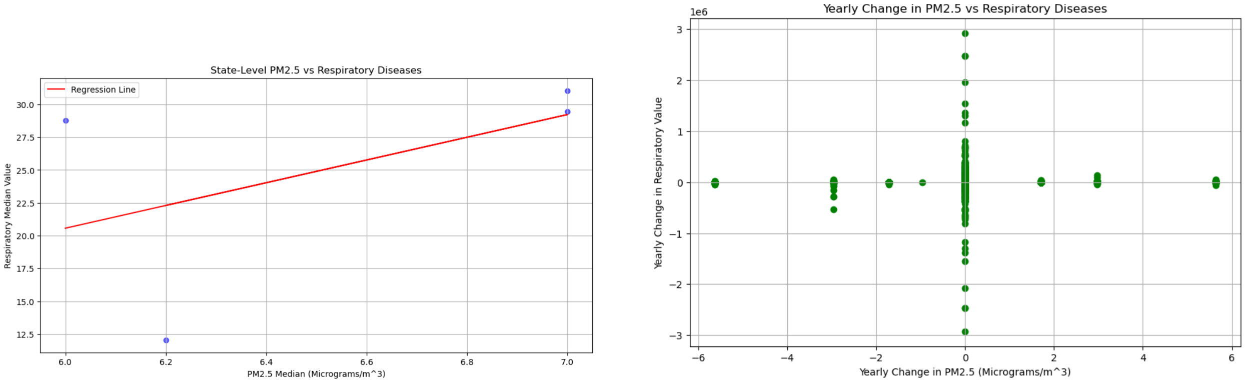
Introduction

This report examines whether air pollution, specifically PM2.5 levels, impacts respiratory diseases. Using air quality and respiratory health data, we performed correlation and regression analyses to uncover potential links.

Analysis

State-Level Analysis

State-level data shows a moderate positive Pearson correlation (0.51) between PM2.5 levels and respiratory health metrics. However, an R^2 score of 0.26 indicates PM2.5 alone explains only a small portion of the variation in respiratory outcomes. The regression trendline reflects a slight increase but lacks strong predictive power.



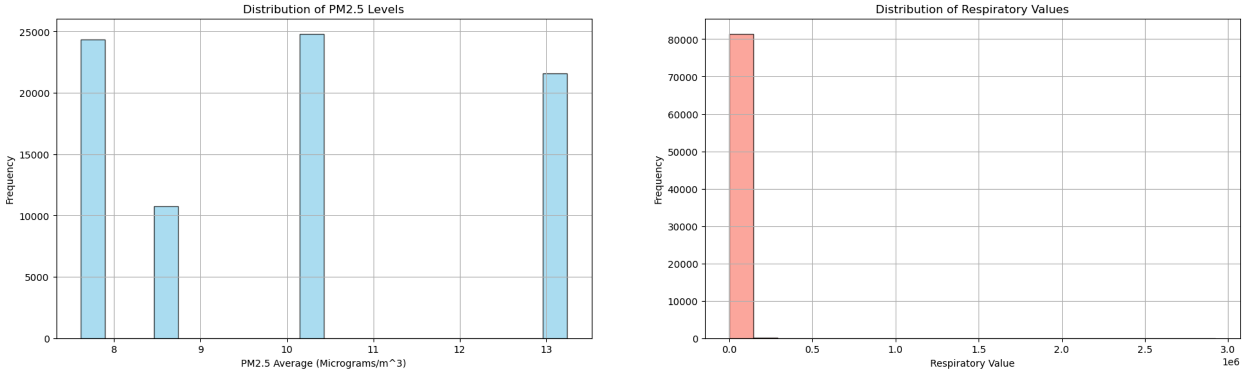
(a) State-Level PM2.5 vs Respiratory Diseases

(b) Yearly Change in PM2.5 vs Respiratory Diseases

Figure 1: State-Level and Yearly Changes Analyses

Distribution Analysis

The distributions of PM2.5 levels and respiratory metrics exhibit considerable variability. The histograms below illustrate these patterns:



(a) Distribution of PM2.5 Levels

(b) Distribution of Respiratory Values

Figure 2: Distributions of PM2.5 Levels and Respiratory Values

Trends Over Time

State-level respiratory disease trends show year-to-year variation. The heatmap below captures these dynamics:

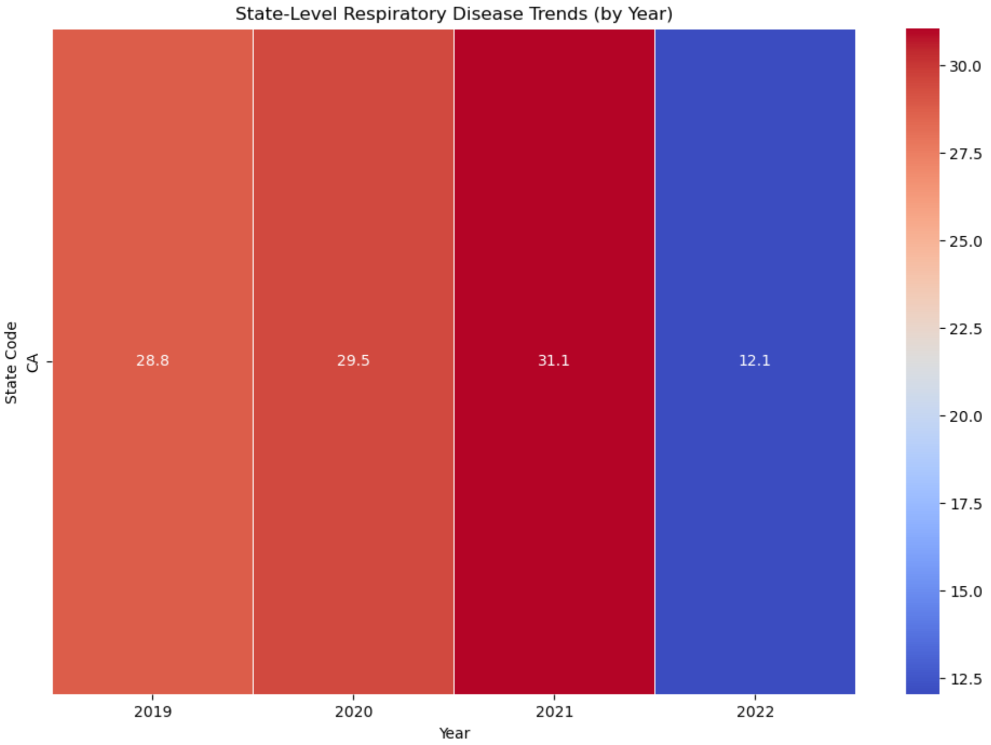


Figure 3: State-Level Respiratory Disease Trends (by Year)

Pollutant Correlations

Additional analysis of pollutants and respiratory health indicates weak correlations, as shown in the chart below:

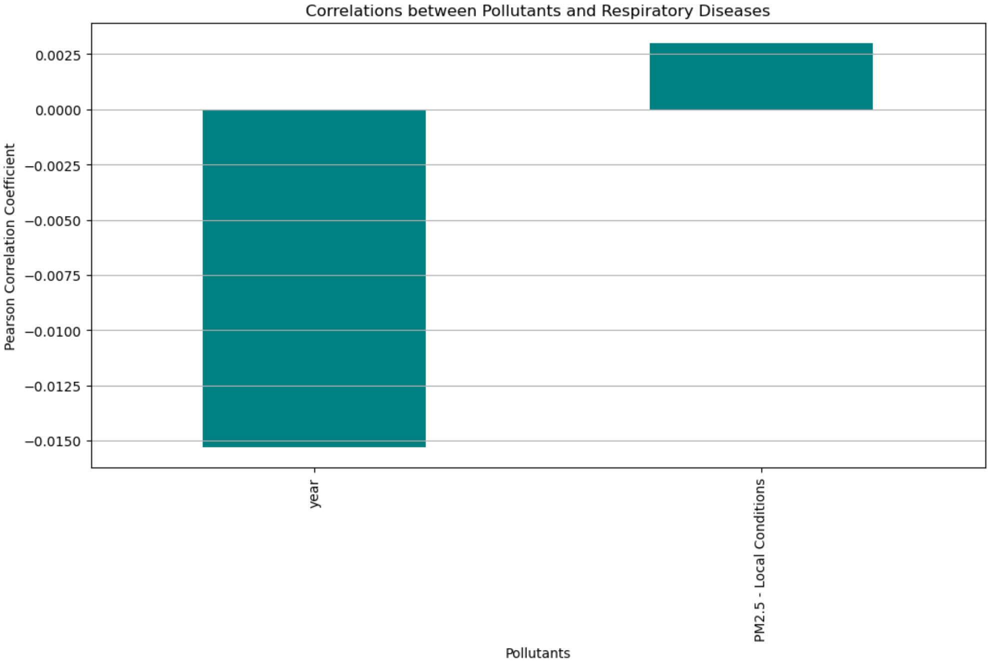


Figure 4: Correlations between Pollutants and Respiratory Diseases

Conclusion

The findings suggest a moderate association between PM2.5 levels and respiratory diseases at the state level, though with limited explanatory power (R^2). Yearly changes and county-level data provide no substantial evidence of direct impact. Improved datasets are essential for a more definitive understanding of air pollution's effects on respiratory health.

Data Sources and Pipeline

Air Quality Data

Source: EPA Air Quality System (AQS) **Reason for Selection:** Provides detailed PM2.5 pollution measurements across California, critical for understanding environmental health impacts. **Content:** Includes pollutant levels, sampling locations, dates, and measurement units. **Structure and Quality:**

- Structured tabular data with attributes such as `state_code`, `date_local`, and `sample_measurement`.
- Missing values handled during preprocessing.
- High reliability as data is maintained by the EPA.

License and Obligations: Licensed under open access by the EPA. Obligations include attribution, addressed in the project documentation. License Details

Respiratory Health Data

Source: CDC Chronic Disease Indicators Dataset **Reason for Selection:** Provides health metrics for respiratory conditions (e.g., asthma) at state and county levels. **Content:** Includes respiratory health indicators, demographic stratifications, and annual statistics. **Structure and Quality:**

- Structured tabular data with fields such as `topic`, `state_code`, `respiratory_value`, and `year`.
- Data contains noise, requiring filtering for relevant indicators.

License and Obligations: Open Data license with attribution requirements, fulfilled via documentation. License Details

Pipeline Overview

The analysis pipeline is summarized below:

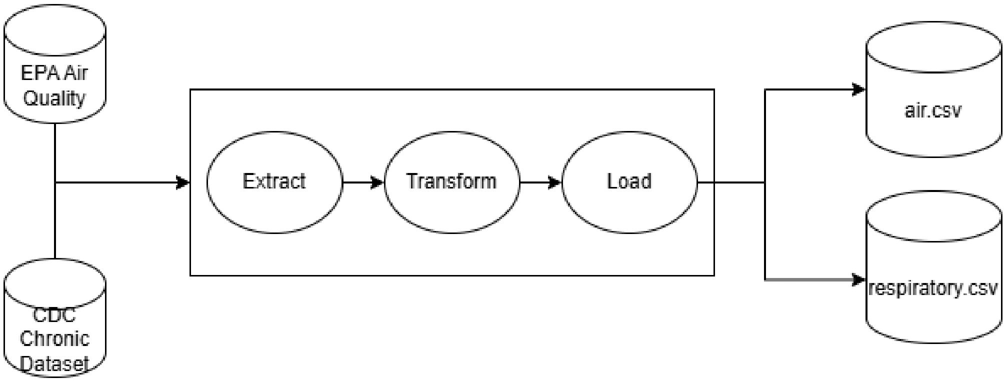


Figure 5: Data Processing Pipeline

Implementation Highlights:

- **Multithreading:** Data extraction, transformation, and loading were performed in parallel to reduce runtime.
- **Modular Design:** The project codebase is organized into modules, ensuring clear segregation of responsibilities:
  - `extract.py`: Extracts raw data from sources.
  - `transform.py`: Cleans and processes the extracted data.
  - `load.py`: Saves the transformed data for analysis.
- **Reproducibility:** The modular approach ensures each step is well-documented, tested, and easily reproducible.