Project Plan

Impact of PM2.5 Concentration in atmosphere on Respiratory Hospitalizations in New York

Utkarsh Bansal

November 21, 2024

Main Question

How does the increase in PM2.5 levels affect the number of respiratory-related hospitalizations in New York?

Data Sources

Selected Data Source: NYC Open Data

- **Data URL:** `https://data.cityofnewyork.us/Environment/Air-Quality/c3uy-2p5r/about_data`
- **Data Format:** CSV, Excel

Description:
The Air Quality dataset from NYC Open Data provides detailed measurements of air pollutant concentrations, including PM2.5 levels, across various locations in New York City. The data is disaggregated by borough, date, and pollutant type, enabling granular analysis of air quality trends over time

Licensing:
The dataset is publicly accessible for research and analysis, with attribution to the NYC Department of Health and Mental Hygiene required for use.
Details at: `https://opendata.cityofnewyork.us/overview/`.

Integration into the Project:
The Air Quality dataset forms a core component of the analysis, providing essential environmental data that can be correlated with health outcomes to assess the impact of PM2.5 on respiratory-related hospitalizations.

Selected Data Source: NYC Environmental Public Health Data Explorer

- **Data URL:** `https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/health-impacts-of-air-pollution/?id=2119#display=summary`

- **Data Format:** CSV, Excel

Description:
The Health Impacts of Air Pollution dataset from the NYC Environmental Public Health Data Explorer provides data on hospitalizations related to air pollution, including PM2.5 exposure, in New York City. The data includes detailed breakdowns by borough, demographic groups, and time periods, supporting an in-depth examination of air pollution's effects on public health.

Licensing:
Apache License Version 2.0, January 2004.
Details at: `https://github.com/nychealth/EH-dataportal?tab=Apache-2.0-1-ov-file/`.

Integration into the Project:
This dataset provides critical health outcome data to complement air quality measurements, enabling a robust analysis of the association between PM2.5 levels and respiratory-related hospitalizations in New York City.

Data Pipeline

Technology:
The data pipeline was implemented using Python, utilizing libraries such as `pandas` for data processing and `matplotlib` for visualization.

Transformation and Cleaning:

- Removal of irrelevant columns.
- Normalization of date formats.
- Handling of missing values through imputation or removal.
- Aggregation of data by year and UHF42 column.

Challenges and Solutions:

- *Problem:* Presence of irrelevant columns.
  Solution: Remove unnecessary columns to simplify the dataset and focus on key variables like PM2.5 levels and locations.
- *Problem:* Missing data for a few specific boroughs or demographics.
  Solution: Exclude the incomplete rows from the analysis and focus on the remaining data.

Meta-Quality Measures:

- Implementation of logging to track errors during data processing.
- Validation of data integrity after each transformation.
- Automatic notifications upon detection of anomalies or missing data.

Results and Limitations

Output Data:
The final cleaned and transformed dataset is stored in a CSV file that combines PM2.5 levels from the Air Quality dataset with respiratory-related hospitalization data from the Health Impacts of Air Pollution dataset. The datasets are joined on the year and location (UHF42) columns, resulting in a comprehensive dataset

Data Structure and Quality:

- **Columns:** Year , Location (Borough) , PM2.5 Levels (from the Air Quality dataset) , Respiratory-Related Hospitalizations (from the Health Impacts dataset).
- **Quality:** High data quality after cleaning and merging. Some gaps may persist in the original datasets, such as missing PM2.5 measurements for specific time periods or incomplete hospitalization records for certain boroughs or years.

Data Format:
CSV was chosen due to its widespread use and compatibility with most analysis tools.
Critical Reflection:

- Potential Biases: The analysis may be biased due to missing data points in either datase.
- Data Collection Differences: Variations in how air quality and hospitalization data were collected could lead to inconsistencies across boroughs or time periods.
- Future Work: Incorporating additional datasets, such as demographic or weather data, could improve the validity and depth of the analysis.