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

Methods of Advanced Data Engineering

Utkarsh Bansal (23546954)

Introduction:

"The air quality index for Tuesday in New York City was forecast to be 101, a level considered unhealthy for sensitive groups" (Angela Weiss/Agence France-Presse). Such advisories underscore the growing concern over air pollution and its impact on public health, particularly in urban areas where industrial activities, transportation, and other human activities contribute to elevated levels of harmful pollutants. One of the most concerning pollutants is fine particulate matter (PM2.5), which is small enough to penetrate deep into the lungs and even enter the bloodstream, posing serious health risks. This study investigates the impact of PM2.5 concentrations on respiratory hospitalizations in various neighbourhoods in New York. By analyzing historical PM2.5 data alongside hospitalization records, this project aims to uncover trends, assess the magnitude of the impact.

Used Data:

1. NYC Open Data (Air-Quality Data):-

   URL:- https://data.cityofnewyork.us/Environment/Air-Quality/c3uy-2p5r/about_data

   - Description:- This dataset provides detailed measurements of PM2.5 levels, across various locations in New York and the data is in CSV format.
   - Structure and Quality:- The data is disaggregated by UHF-42, Time Period and pollutant type, enabling granular analysis of air quality trends over time. It has been verified to have consistent and clean data.
   - 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/

2. NYC Environmental Public Health Data Explorer (Hospitalisations Data):-

   - URL:-https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/health-impacts-of-air-pollution/?id=2119#display=summary
   - Description:- This dataset provides data on hospitalizations caused due to air pollution, including PM2.5 exposure, in New York City across various neighbourhoods.
   - Structure and Quality:- The data includes detailed breakdowns by UHF-42, 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/.

Analysis:

1) **Method**:

Data Cleaning and Preparation

- In the Air Quality Dataset Missing values and irrelevant columns (e.g., “messages”) were removed. Only PM2.5 data was retained, and the cleaned dataset was stored in a database.
- Hospitalization Dataset was already cleaned; directly stored in the database.

Aggregation and Joining

- The air quality data (yearly) was aggregated into multi-year time periods (e.g., 2009–2011) to match the hospitalization dataset. PM2.5 averages were calculated for each UHF-42 region.
- The datasets were joined using SQL, with **UHF-42 regions** and **time periods** as keys.

Final Dataset

- Structure: **42** UHF-42 regions, **4** time periods each (e.g., 2009–2011, 2012–2014).
- Columns: **Region**, **Time Period**, **Average PM2.5**, **Annual Hospitalizations.**
- Example: Bayside - Little Neck; 2009–2011 ; PM2.5 = 9.47 ; Hospitalizations = 7.1.

This prepared dataset enabled further analysis of PM2.5's impact on hospitalizations.

2) **Interpretation of results:**

a) The scatter plot (Figure-1) illustrates the average PM2.5 concentration in the air and the estimated annual number of hospitalizations for neighborhoods under UHF-42 in the New York city .The graph demonstrates that after each time period, PM2.5 levels decrease across all neighborhoods, leading to a reduction in the number of hospitalizations. To avoid clutter, a line graph for all 42 regions was not created. Instead, 7 random neighborhoods were selected to ensure there was no bias while maintaining clarity in the visualization.

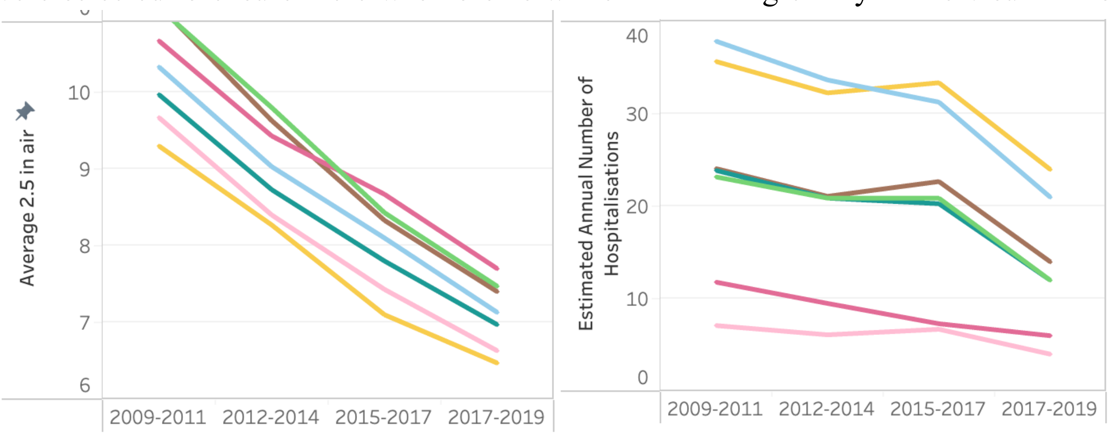

Figure 1

b)The stacked bar chart (Figure-2) shows the estimated annual hospitalizations categorized by PM2.5 levels (High, Medium, Low). Hospitalizations show a significant decline over time, primarily driven by a sharp reduction in high PM2.5 areas. Over the years, areas with initially high PM2.5 levels gradually transitioned into the medium PM2.5 category and eventually into the low PM2.5 category, reflecting the effectiveness of air quality interventions. This trend underscores the positive impact of improving air quality in reducing respiratory health risks.

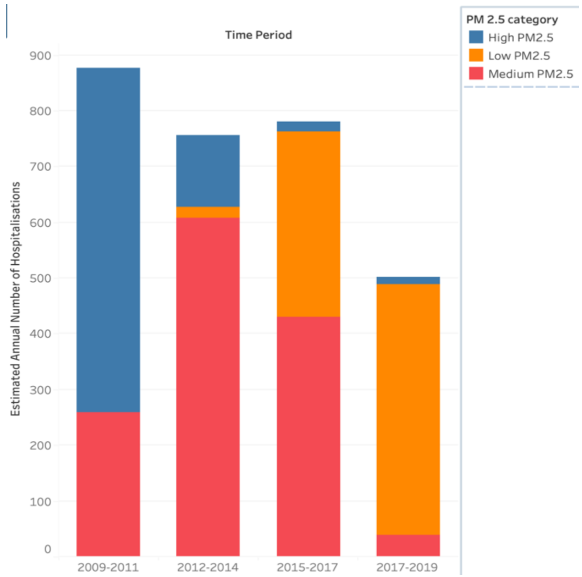

Figure-2

Correlation between Average 2.5 concentration and Hospitalisations

c) The scatter plot (Figure-3) shows a meaningful link between PM2.5 levels and respiratory hospitalisations. The **p-value** of **0.022** indicates a strong relationship between PM 2.5 concentration in atmosphere and the number of hospitalisations. However, the $\boldsymbol{R^2}$ value of **0.29** means that PM2.5 explains only **29%** of the variation in hospitalisations, suggesting that other factors also play a big role. This highlights the need to look at additional influences, like demographics or healthcare access, to fully understand what drives respiratory health issues

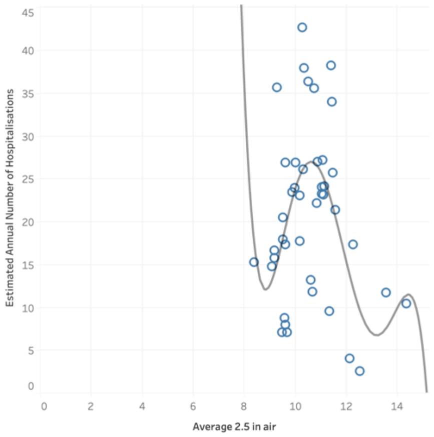

Figure - 3

Identifying Disparities in Air Pollution and Health Outcomes

d) The scatter plot (figure-4) groups all neighborhoods across all time periods into three distinct clusters. Cluster-1(**red**) represents neighborhoods with low PM2.5 concentrations and low hospitalization numbers, indicating areas with minimal air pollution and its associated health impacts. Cluster 2 (**blue**) includes neighborhoods with high PM2.5 concentrations and high hospitalization numbers, highlighting regions where air pollution is directly linked to increased health risks. Cluster 3 (**orange**) identifies neighborhoods with high PM2.5 concentrations but surprisingly low hospitalization numbers. This disparity suggests the influence of external factors, such as inadequate access to medical facilities or underreporting of cases. The graph underscores the need to prioritize regions in Cluster 3 by improving healthcare infrastructure and accessibility to address potential hidden health impacts.

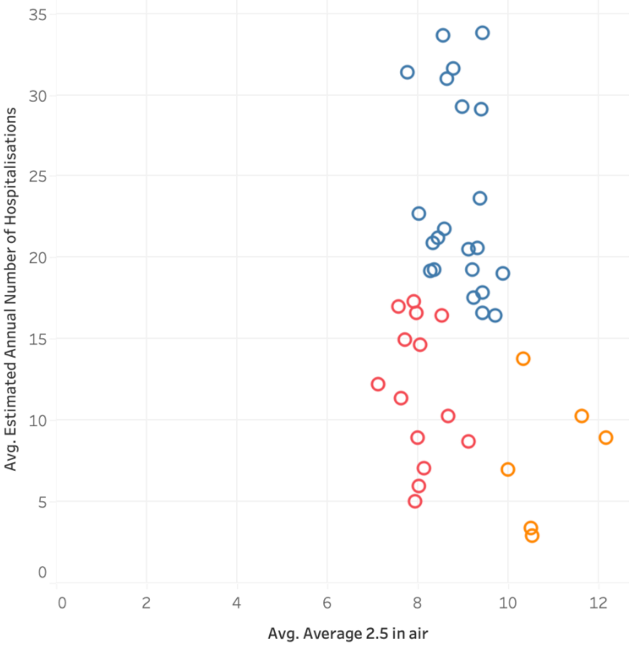

Figure-4

Conclusion:-

The analysis demonstrates a clear link between PM2.5 concentrations and respiratory hospitalizations in New York City neighborhoods. While reductions in PM2.5 levels over time have led to decreased hospitalizations, the relationship is only partially explained by air pollution ($R^2 = 0.29$). Clustering analysis highlights the need to prioritize neighborhoods in Cluster 3, where high pollution levels and low hospitalizations indicate potential gaps in healthcare access. These findings emphasize the importance of addressing both environmental and systemic factors to improve respiratory health outcomes citywide.