Data 604 - Project Proposal

Health outcomes in New York City

Jason Wong
Boru Sun
Angela Li
Viktoriia Shcherbachuk

Group: L01 - 12
Instructor: Dr. Leanne Wu
November 7, 2022

Introduction

For our project, we will be examining the relationships between air quality and neighbourhood poverty, and seeing what effect this has on cancer rates and incidents of asthma in New York City. In our research for this project, we discovered that lung cancer is strongly correlated to air pollution, and chose to investigate that along with 3 other types of cancer that we think will also be affected and that we found interesting. The cancers that we will be investigating are: Lung cancer, Larynx cancer, Esophageal cancer, and Non-Hodgkins Lymphoma.

According to the World Health Organization, 1 in 10 people who are exposed to poor air quality live in extreme poverty, and 9% of all lung cancer deaths are caused by air pollution. In the 2009 scientific article “Healthy neighbourhoods: Walkability and air pollution. Environmental Health Perspectives”, Marshall et.al describe the poorest neighbourhoods as having the worst air quality and are closest to sources of severe air pollution such as landfills, factories, and power plants, while the most affluent neighbourhoods have the best air quality and are the furthest away from these severe sources of air pollution. Although cigarette smoking accounts for more than 80% of lung cancers, substantial numbers of lung cancer cases are observed among people who have never smoked. This is likely caused by air pollution, especially Particulate Matter (PM) , in which there is clear evidence linking it to lung cancer incidence and mortality world-wide (Turner et al. 2020).

Understanding the relationships between these variables will allow us to address issues regarding environmental injustice as well as be the key drivers for positive societal change.

Data Sets

We will be looking at four individual datasets, one for each group member. We obtained these datasets through different databases that were collected and published by the American Community Survey, which are publicly available on the New York City Department of Health Environment and Health Data Portal from the New York City government website. For each database, we pulled certain sets of data that were relevant to what we are studying.

For Air Quality, we will be looking at several forms of pollution: Black Carbon, Fine Particles (PM2.5), Nitric Oxide (NOx), Ozone (O3), Sulfur Dioxide (SO2). We also added a Carbon Monoxide dataset as well, which was not included in the Air Quality database, but we decided that this would be worth looking in to. The datasets are tabular in format, with 9 columns named: Time, GeoType, GeoID, GeoRank, Geography, Measurement, Annual Average, General Population Action Days, and Sensitive Population Action Days. Angela Li will be taking on this dataset.

For Asthma, we pulled data pertaining to incidence rate for the past 12 months for adults. The data is in tabular format, with 8 columns named: Time, GeoType, GeoID, GeoRank, Geography, Age-Adjusted rate per 10,000, Estimated annual rate per 10,000, and Number. Viktoriia Shcherbachuk will be responsible for this dataset.

The Neighbourhood Poverty dataset is also structured in a tabular format, with 9 columns named: Time, GeoType, GeoID, GeoRank, Geography, Number, Percent and Rank. The GeoType has six subsections as Citywide, Borough, UHF42, Sub-borough, CD and NTA. There are 5 boroughs, 42 adjoining zip code areas in UHF (United Hospital Fund), 45 sub-boroughs, 27 congressional districts (CD) and 195 Neighbourhood tabulation areas (NTA). Boru Sun will be responsible for this dataset.

The 4 Cancer datasets are all structured in the same way; tabular format with 11 columns named Time, GeoType, GeoID, GeoRank, Geography, Age-adjusted rate (Females) per 100,000, Age-adjusted rate (Males) per 100,000, Age-adjusted rate per 100,000, Average annual number, Average annual number (Females), Average annual number (Males). The GeoType has three subsections as Citywide, Borough and Sub-borough. There are 5 boroughs and 54 sub-boroughs. Jason Wong will be working with this dataset.

Methodologies

Our methodologies include: establishing a database in MariaDB or MongoDB to store our datasets, clean the data by removing irrelevant columns, populating empty values with “N/A”, and dropping rows with too many “N/A” values, and also establishing an ERD (Entity-Relationship Diagram) of our relational database. We will use our guiding questions to graph any correlations, and create helpful visualizations that will appropriately communicate what our investigations. This will allow us to draw conclusions based on our data, even if any hypothesized correlations don’t actually exist.

Eventually we would like to establish a database on AzureDB to store our datasets, as this is more relevant to industry. We would also like to write stored procedures that will provide a process to automatically store and clean our data, as well as have some data pipelines that will continuously pull raw JSON files that will aid us in our investigation/analysis. Finally, we would like to have an opportunity to alter our database to have tables that accurately reflect our ERD and follow relational schema.

Ultimately, we are aiming to expand our knowledge of data storage and processing, improve data wrangling skills, and gain exposure to cloud-based data management systems in order to effectively do an in-depth exploratory analysis of our chosen topic.

References

Literature:

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature, 525(7569), 367–371. https://doi.org/10.1038/nature15371

DeMarini, D. M. (2015). How does air pollution cause cancer? Cancer Prevention Research, 8(10_Supplement). https://doi.org/10.1158/1940-6215.prev-14-cn08-03

Turner, M. C., Andersen, Z. J., Baccarelli, A., Diver, W. R., Gapstur, S. M., Pope, C. A., Prada, D., Samet, J., Thurston, G., & Cohen, A. (2020). Outdoor Air Pollution and cancer: An overview of the current evidence and public health recommendations. CA: A Cancer Journal for Clinicians, 70(6), 460–479. https://doi.org/10.3322/caac.21632

Turner, M. C., Krewski, D., Diver, W. R., Pope, C. A., Burnett, R. T., Jerrett, M., Marshall, J. D., & Gapstur, S. M. (2017). Ambient air pollution and cancer mortality in the cancer prevention study II. Environmental Health Perspectives, 125(8), 087013. https://doi.org/10.1289/ehp1249

Marshall, J. D., Brauer, M., & Frank, L. D. (2009). Healthy neighbourhoods: Walkability and air pollution. Environmental Health Perspectives, 117(11), 1752–1759. https://doi.org/10.1289/ehp.0900595

Stewart, J. A., Mitchell, M. A., Edgerton, V. S., & VanCott, R. (2015). Environmental justice and health effects of Urban Air Pollution. Journal of the National Medical Association, 107(1), 50–58. https://doi.org/10.1016/s0027-9684(15)30009-2

Nitrogen Dioxide. American Lung Association. (n.d.). Retrieved November 6, 2022, from https://www.lung.org/clean-air/outdoors/what-makes-air-unhealthy/nitrogen-dioxide#:~:text=Increased%20inflammation%20of%20the%20airways,Increased%20asthma%20attacks%3B%20and

Dataset sources:

New York City Department of Health, Environment & Health Data Portal. "Neighborhood Air Quality" data. Black carbon. Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/air-quality/?id=2024#display=summary" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Neighborhood Air Quality" data. Fine particles (PM 2.5). Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/air-quality/?id=2023#display=summary" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Neighborhood Air Quality" data. Nitric oxide (NOx). Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/air-quality/?id=2028#display=summary" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Neighborhood Air Quality" data. Nitrogen dioxide (NO2). Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/air-quality/?id=2025#display=summary" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Neighborhood Air Quality" data. Ozone (O3). Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/air-quality/?id=2027#display=summary" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Neighborhood Air Quality" data. Sulfur dioxide (SO2). Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/air-quality/?id=2026#display=summary" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Adults asthma prevalence" data. Adults with asthma (past 12 months). Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/asthma/?id=18#display=summary" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Economic conditions" data. Neighborhood poverty. Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/economic-conditions/" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Carbon monoxide" data. Carbon monoxide incidents. Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/carbon-monoxide-incidents/" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Cancer" data. Lung and bronchus cancer. Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/cancer/" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Cancer" data. Non-Hodgkin's lymphomas. Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/cancer/" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Cancer" data. Esophageal cancer. Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/cancer/" on 10/31/2022.

New York City Department of Health, Environment & Health Data Portal. "Cancer" data. Larynx cancer. Accessed at "https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/cancer/" on 10/31/2022.