DATA 512 Part 2 - Extension Plan

Abhinav Duvvuri

Introduction

In Common Analysis - Part 1, we looked at the wildfire data for the city of Philadelphia, PA. We created an estimate of smoke using this data and used it to create a model and got predictions for the next 15 years. Simultaneously, we also looked at AQI data and compared both to see if there were any noticeable trends in both. In this second part, we extend that work by looking into the impact of smoke on health.

The overall aim of this analysis is to produce a report that can help inform the city council, city manager/mayor and city residents what they might need to plan for in the coming years. For this analysis, I will work on finding the impact of smoke on Health, specifically respiratory conditions.

In this extension plan, we first look at the problem statements and the motivation behind it. This is followed by an impact focus. Then we talk about the data being used for this analysis in detail. This is followed by a section on the unknowns and dependencies that I presume might occur. Finally, I will be providing a comprehensive timeline which I will try to follow for this analysis.

Motivation/Problem Statement:

When I was looking into the health details of the city of Philadelphia, I found a very interesting article from the Philadelphia regional center for children's environmental health website about asthma. The website had a very interesting statistic that among children in the Philadelphia region for the year 2020, there was a 21% asthma prevalence rate. This is an alarming number compared to the 5% average asthma prevalence rate in the US. Since one of the major contributors to Asthma is said to be air pollution, I wanted to understand if Asthma and other respiratory illnesses in general were linked to smoke. If we find that respiratory illnesses are indeed linked to smoke, they should help the lawmakers take proper steps towards dealing with smoke before the situation gets out of hand.

Following are some key questions I want to address:

1. How might exposure to smoke contribute to the number of respiratory disease cases in Philadelphia? Is there any trend over the years? What might the trends look like in the near future?
2. Are asthma cases in children living in Philadelphia related to smoke? Is there a trend here too? If yes, what might the future look like?

This analysis of finding respiratory diseases and smoke will indeed help everyone living in Philadelphia understand, evaluate and take action so as to make the city a better place.

Impact Focus

For this extension part, I have decided to look into the healthcare sector in Philadelphia, PA. Specifically, I want to look into respiratory diseases. Also, I will be looking into asthma cases for children as it looks like there is a high percentage of children having asthma in this city. If we find correlations, it means that the relevant authorities really need to take measures really quickly to rectify this.

Data

Dataset 1: BRFSS

The commonwealth of Pennsylvania has a Pennsylvania Asthma Surveillance System. Upon exploring the results there, they point to multiple data sources. In order to analyze asthma data for children, I found data that was published by the Centre for Disease Control(CDC) was useful. The CDC has an initiative called the Behavioural Risk Factor Surveillance System (BRFSS) which is the nation's premier system of health-related telephone surveys. Our data comes from there.

The survey was performed using a set of questions that are modified every year. There are two questions related to asthma in this questionnaire. The first question enquires about the history of asthma in the family and the second question asks if the person has asthma currently. Data is available from 1989-2022. There are two difficulties in working with this data. Firstly I will need to get each year's data separately and since every year's questionnaire is different I will have to adapt the data for each year individually. The second issue is that since the data is on a per survey basis, I will have to do a considerable amount of cleaning, grouping, and aggregations on the data. He is a link to the BRFSS Data sources. The data is published in the public domain and may be used without permission since it is from the federal agency. This can be found in point 14 of the FAQs.

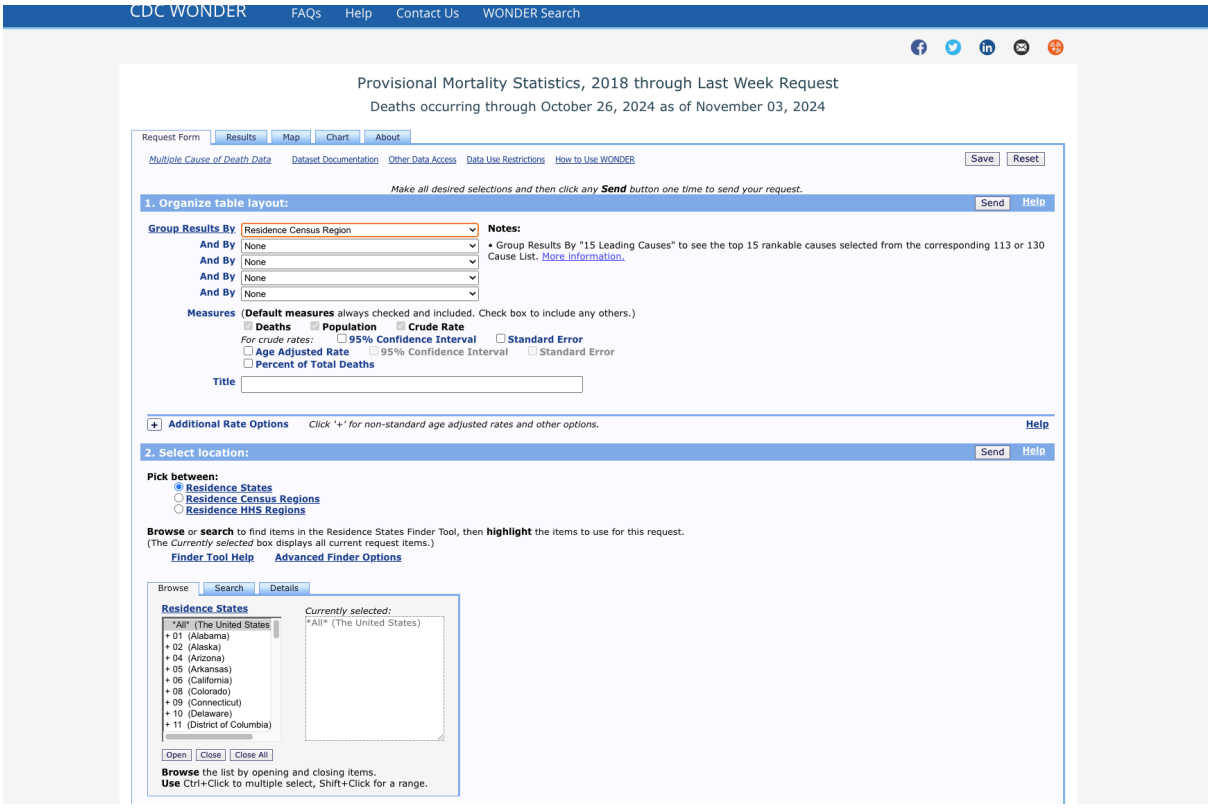
Following are the fields I will be using from here:

1. Year -> For the year in which the survey was recorded
2. County -> County where the participant resides
3. State -> State where the participant resides
4. _LTASTH1 -> Calculated variable for adults who have ever been told they have asthma
    a. **SAS Code used**: IF ASTHMA3=1 THEN _LTASTH1=2; ELSE IF ASTHMA3=2 THEN _LTASTH1=1; ELSE _LTASTH1=9;
5. ASTHMS1 -> Calculated variable for computed asthma status.
    a. **SAS Code used:** IF ASTHMA3=1 AND ASTHNOW=1 THEN _ASTHMS1=1; ELSE IF ASTHMA3=1 AND ASTHNOW=2 THEN _ASTHMS1=2; ELSE IF ASTHMA3=2 THEN _ASTHMS1=3; ELSE _ASTHMS1=9;

Dataset 2: CDC wonder

Apart from this data, I will also be using the respiratory illness mortality data from CDC wonder. Data is available from 1992-2022 here. Here also, I will need to download multiple files and consolidate. Next, I will need to clean, group and make aggregations to get the per-year statistics that are required. This data is covered by the *Public Health Service Act (42 U.S.C. 242m(d))*.

Data can be downloaded using this website. A database needs to be accessed and we need to provide the required fields and aggregations to get the data. Below screenshot shows the database request site. We can directly perform aggregations and filtering here.



Below are the fields I will be using from this dataset.

1. State -> To get Pennsylvania state
2. County -> To filter by Philadelphia county
3. Year -> To aggregate by year
4. Leading Causes of Death -> Used to get cause of death
5. Leading causes of death (infants) -> Used to get cause of death (for infants)
6. Age -> used to separate kids and adults

Following is the data use restrictions as published on their website. I will make sure to abide by this while performing my analysis.

Data obtained from the National Center for Health Statistics, including Compressed Mortality, Fetal Deaths, Linked Birth / Infant Death records, Multiple Cause of Death, Natality, Provisional Mortality, Underlying Cause of Death, and other presentations of vital records data, are also covered by the following policy:

The Public Health Service Act (42 U.S.C. 242m(d)) provides that the data collected by the National Center for Health Statistics (NCHS) may be used only for the purpose for which they were obtained; any effort to determine the identity of any reported cases, or to use the information for any purpose other than for statistical reporting and analysis, is against the law. Therefore users will:

- *Use these data for statistical reporting and analysis only.*
- *Do not present or publish statistics representing nine or fewer births or deaths, including rates based on counts of nine or fewer births or deaths, in figures, graphs, maps, tables, etc.*
- *Make no attempt to learn the identity of any person or establishment included in these data.*

- *Make no disclosure or other use of the identity of any person or establishment discovered inadvertently and advise the Director, NCHS of any such discovery*

Unknowns and dependencies

Following are some issues I foresee:

1. I will have to limit my analysis to after 1992 since the second dataset does not have data before that.
2. The data from BRFSS, while impressive, is from a survey. Hence, we need to understand more about how people were sampled. We also need to be aware of convenience sampling here. I will look into this as I do my analysis.
3. The questionnaire is separate for each year which means I will have to manually get the data, find the right fields and clean it. This is very time-taking and hence I will have to start early.

Timeline to completion

- Collect data (Target Date - Nov 11th)
    - Collect CDC wonder data from the portal. Clean it and merge into a single file.
    - Collect Asthma data from BRFSS per year. Merge by year and filter for required columns. Highly time taking.
- Clean and process data (Target Date - Nov 15th)
    - Aggregate the CDC wonder data on year after filtering for Philadelphia county and perform other processing steps.
    - Aggregate the Asthma data and other processing steps
- Build model and visualizations (Target Date - 22nd)
    - Choose a model to build. Try a few models.
    - Verify the statistical significance of the result and document.
    - Create visualizations that capture findings.
- Presentation Date Nov 27 (complete presentation by Nov 24th)
    - Create a presentation in the required format using the results.
- Documentation (Target Date - Dec 1)
    - Create documentation for repository of code
    - Start working on the final report. Reuse components from previous homeworks.
- Final touches for the repository (Dec 2,3)
    - Make sure the repository and final report adhere to all required standards and rubrics.
- Final Report and repository submission date - Dec 4

References

1. Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 1996-2022.
2. CDC wonder Data use policy: https://wonder.cdc.gov/datause.html
3. CDC wonder data: https://wonder.cdc.gov/controller/datarequest/D176
4. Asthma article on PRCC: https://prcceh.upenn.edu/focus-areas/asthma/
5. Article showing Asthma aggravated due to wildfire: https://www.inquirer.com/health/expert-opinions/air-quality-smoke-philadelphia-asthma-symptoms-20230607.html