Part 2: Extension Plan

Motivation/problem statement:

With wildfires becoming more frequent and intense, especially across the western United States, smoke exposure has become a growing concern for communities, even far away from the actual fires. This smoke doesn't just harm the environment, it also impacts the health of people who live in areas where the smoke travels to. The particles in wildfire smoke are so small that they can easily reach deep into people's lungs, which can worsen breathing problems and even affect overall health (Yale Medicine, 2023).

In Part 1 of this project, I created an initial model to estimate the smoke impact on Murfreesboro, Tennessee. This first model used smoke impact scores based on the size and distance of fires within a 650-mile radius of the city. However, I want to take this analysis further by including health data to see how the smoke may be affecting the people in Murfreesboro, TN over time. Specifically, I'm looking to add data on deaths from chronic respiratory illnesses, which could be related to smoke exposure. The goal of this extension is to model how smoke exposure may connect to death rates from respiratory illnesses and to use this model to predict how health impacts from smoke could look in the future, up to 2050. I hope this analysis can provide helpful insights for city officials, healthcare providers, and policymakers who are responsible for keeping people safe. It could also be valuable for Murfreesboro residents, helping them understand the health risks that come with wildfire smoke and encouraging them to take steps to protect their health, like using air filters or avoiding outdoor activities on smoky days.

Impact focus:

The main focus of this project is on healthcare, specifically looking at how wildfire smoke exposure could affect deaths due to chronic respiratory illnesses. There's growing concern about the health impacts of wildfires, and more and more studies are coming out that show how wildfire smoke can worsen existing health problems. According to Dr. Carrie Redlich, a doctor at Yale Medicine who specializes in occupational and environmental health and asthma, wildfire smoke particles are small enough to get deep into the lungs and even enter the bloodstream, which can make certain respiratory and heart conditions worse (Yale Medicine, 2023). By expanding my smoke impact model to include health data, I hope to give a clearer picture of how smoke exposure from wildfires may be impacting people's health over time. This project could be really useful for healthcare providers and city officials who need to make decisions about public health, like planning for hospital resources or setting up warnings for high-smoke days. Plus, it can help inform residents in Murfreesboro about why wildfire smoke is a serious health concern and how they can protect themselves.

Data to be used:

I am planning to use a dataset from the Institute for Health Metrics and Evaluation (IHME) which is an independent population health research organization based at the University of Washington. One of the IMHE's largest studies is the Global Burden of Disease study which is a comprehensive study of diseases over time and country for various metrics (Institute for Health Metrics and Evaluation, n.d.). I am choosing to use a subset of the data from this study which I can download from their visualization tool. I am focusing on the number of deaths for all ages, men and women, caused by chronic respiratory diseases which includes: asthma, COPD, interstitial lung disease and pulmonary sarcoidosis, pneumoconiosis (including silicosis, asbestosis, coal-worker pneumoconiosis, and other pneumoconiosis), and other chronic respiratory diseases (VizHub- GBD Compare, n.d.).

The most granular location breakdown that the GBD study gives is by state, so I am using the state of Tennessee for my analysis, since I could not find a dataset specific to Rutherford county or the city of Murfreesboro. Additionally, the data is only available for the years 1990 to 2021. The link to access this data is: https://vizhub.healthdata.org/gbd-compare/**.** You must use the visualization tool to filter for the state of Tennessee, all age groups, both genders, the death metric, and for the disease. When you download the dataset, it downloads as a csv with the following columns: "Location", "Year", "Age", "Sex", "Cause of death or injury", "Measure", "Value", "Lower bound", "Upper bound". The value I am going to use, which is the number of deaths for that year, is in the "Value". The "Lower" and "Upper" columns provide an uncertainty interval, but for the purpose of this analysis, I will just be using the "Value" column. This data is available to use for non-commercial use in accordance with IHME Free-of-Charge Non-commercial User Agreement which can be found here: https://www.healthdata.org/Data-tools-practices/data-practices/ihme-free-charge-non-commercial-user-agreement.

Unknowns and dependencies:

There are several unknowns and dependencies that could impact how reliable this analysis will be. One major challenge is that time series forecasting, especially over a long period, like 25 years, is difficult to get right. Predicting this far into the future assumes that trends from the past will continue, but in reality, things could change due to climate change, new wildfire management strategies, or health advancements. This could lead to more error over time, especially since wildfire activity and smoke impact can vary so much from year to year. Also, the smoke impact scores I'm using are estimates based on fire acreage and distance, so they aren't exact measurements of the smoke Murfreesboro impact.

Another unknown is the mortality data itself. Since I'm using an external regressor to add the mortality rates to my existing model, I need historical and future values in order to

predict the impact estimates. I'll need to make assumptions for future mortality rates, which could introduce more uncertainty, especially if healthcare or policies change significantly over the next few decades. I'll have to either assume they stay constant or base them on historical trends and use a model to predict them. The mortality data I have only covers the years from 1990 to 2021, while the smoke impact data goes back to 1960. This mismatch means I'll mostly be using data from the years that overlap which is a smaller subset of historical trends to base the future predictions on.

There's also a limitation in the way the mortality data is recorded. The dataset tracks all deaths from chronic respiratory illnesses, not just those that may be related to wildfire smoke exposure. Chronic respiratory diseases can be caused by lots of factors, like genetics, smoking, or other environmental pollutants, so it's hard to pinpoint how much wildfire smoke specifically contributes to these deaths. This means that any conclusions drawn from this analysis should be taken with caution since other causes could be at play. Additionally, the mortality data is for the entire state of Tennessee, but my analysis is focused on Murfreesboro. Statewide data may not fully capture what's happening in Murfreesboro specifically, as there could be differences in population health, healthcare access, or air quality. However, since we are looking at such a wide range of wildfires to include in the data, 650 mile radius, I feel that it is okay to assume mortality data for the whole state, but it does have to be disclosed and considered when the estimated predictions are being analyzed.

Overall, these unknowns and dependencies mean that while this model can give us useful insights, we should treat the predictions as general trends rather than precise outcomes. There are many factors outside of my control that could influence the results, so it's important to remember that the model's predictions are just one way to understand the potential impact of wildfire smoke on Murfreesboro's health.

Timeline to completion:

The following table outlines tasks that need to be completed in order to complete this analysis by the date it is due which is December 4, 2024. The table has a column for tasks that need to be completed and a column, "Completion Date", which is the date the tasks need to be completed by. Some of these tasks are when to have certain steps done and some are tasks for starting a step of the project.

Task	Completion Date
Collect data - Download and clean	11/7/2024
Add data to existing smoke impact model	11/12/2024

and test model - Make sure that adding External Regressor to ARIMA is appropriate - Find the best hyperparameter - Find accuracy	
Make predictions	11/15/2024
Visualize results and document process - Make visualizations for presentation and report - Document in repo and take notes	11/18/2024
Begin final report	11/18/2024
Create presentation and have draft of report - Have a good draft of final report ready to base presentation on	11/23/2024
Turn in presentation	11/27/2024
Finalize report and repo	12/2/2024
Present project	12/2/2024
Turn in final report and repo	12/4/2024

References:

Institute for Health Metrics and Evaluation. (n.d.). *Global burden of disease (GBD)*. IHME. Retrieved November 4, 2024, from https://www.healthdata.org/research-analysis/gbd

VizHub- GBD Compare, n.d. *GBD compare | Viz hub*. IHME. Retrieved November 4, 2024, from https://vizhub.healthdata.org/gbd-compare/

Yale Medicine. (2023). *How bad is wildfire smoke for your health?* Yale Medicine. Retrieved from https://www.yalemedicine.org/news/how-bad-is-wildfire-smoke-for-your-health