UNIVERSITY of WASHINGTON

Project Report

(DATA 512: Human-Centered Data Science)

Addressing the Public Health and Economic Impact of Wildfire Smoke in Vancouver, WA

Parvati Jayakumar

Table of Contents

Introduction

The increasing frequency and intensity of wildfires due to climate change pose a growing public health concern, particularly in urban areas like Vancouver, WA. Dense smoke from wildfires contains harmful pollutants that significantly elevate respiratory diseases [1][2][3][4], hospitalizations, and premature deaths [5]. These health challenges, coupled with economic impacts such as rising unemployment and poverty, disproportionately affect vulnerable communities. This analysis examines the interplay between wildfire smoke exposure, respiratory health, and economic stability through a human-centered data science approach, highlighting the lived experiences of impacted individuals.

This research addresses critical questions about the long-term effects of smoke exposure on public health and local economies [6]. The findings aim to guide city officials, public health practitioners, and community organizations in implementing interventions and policies to mitigate these impacts. By emphasizing community resilience and preparedness, this study contributes valuable insights into public health responses and policy initiatives needed to support affected populations.

Background

Research on the impacts of wildfire smoke on public health and socio-economic conditions has intensified, driven by the rising frequency and severity of wildfires due to climate change. Wildfire smoke exposure is strongly associated with respiratory illnesses like asthma and COPD [7], contributing to higher hospitalization rates and emergency visits, particularly among vulnerable groups such as children, the elderly, and those with pre-existing conditions [8]. These findings support hypotheses linking smoke exposure to negative health outcomes.

Wildfires also harm local economies through job losses and reduced tourism revenue caused by hazardous air quality. For example, a UCLA study estimated that wildfire smoke led to 52,500 deaths in California between 2008 and 2018, with economic losses exceeding $450 billion [9].

Advancements in analyzing wildfire impacts now combine atmospheric models with ground-level monitoring, using geospatial data to assess smoke exposure and its health effects. These techniques offer detailed insights into how fire types and sizes influence air quality, improving public health strategies [10].

This project focuses on the impacts of wildfire smoke in Vancouver, WA, using key datasets such as the **USGS Wildland Fire Combined Dataset [D1]**, **EPA Air Quality Service (AQI) data [D2], IHME respiratory disease mortality data [D3]**, and **FRED socioeconomic data [D4][D5][D6]**. By analyzing trends over the past 60 years and forecasting future risks, the study aims to provide actionable insights for city planners and public health officials.

Key hypotheses include:

1. **Differential Effects of Fire Types:** Wildfires may have more severe smoke impacts compared to controlled prescribed fires.
2. **Health Impacts of Smoke Exposure:** Higher smoke exposure correlates with increased mortality from respiratory diseases and long-term health and socio-economic challenges.
3. **Economic Instability from Wildfires:** Wildfire smoke is hypothesized to increase unemployment, poverty, and premature deaths in affected communities.

By integrating insights from existing research and employing advanced modeling techniques, this analysis aims to contribute valuable evidence for policymakers and public health officials to develop targeted interventions and resilience strategies for communities facing the growing threat of wildfires.

Methodology

This project employs a comprehensive methodology aimed at analyzing the impacts of wildfires on the city of Vancouver, WA, particularly focusing on the effects of wildfire smoke on public health and air quality. The methodological framework was structured around several pivotal stages, including data acquisition, data processing, statistical analysis, predictive modeling, and ethical considerations. Each stage was designed with a keen awareness of human-centered data science principles, emphasizing the importance of understanding and addressing the impacts of wildfires on vulnerable populations.

Data Acquisition

Data plays a critical role in this analysis. The analysis utilized several key datasets: wildfire data from the **USGS Combined Wildland Fire Datasets [D1]**, air quality data from the **EPA Air Quality Service API [D2]**, **respiratory disease mortality rates from the IHME [D3]**, and **socioeconomic indicators from the Federal Reserve Bank of St. Louis (FRED) [D4][D5][D6]**. These datasets were downloaded to support the study.

Data Processing

Several key steps employed in the data preprocessing phase are mentioned below. Python libraries, **Pandas** and **NumPy**, were primarily utilized to handle data manipulation and calculations efficiently.

- **Average Distance Estimation:** The average distance to a wildfire was computed using a defined function that averaged distances from all points on the fire's perimeter to the city coordinates. Additionally, only wildfires that were within 650 miles from Vancouver were retained. This threshold was chosen based on the requirement to ascertain the effects of wildfires specific to the city and is critical for establishing a relevant context for assessing smoke impacts.

- **Data Cleaning and Transformation:** The datasets had minimal missing values, so imputations were unnecessary. Data was standardized for consistency in column naming and types, with geographic coordinates aligned to EPSG:4326. Wildfires during the fire season (May 1–October 31) were filtered, duplicates and irrelevant entries removed, and categorical variables like fire causes and types were cleaned and numerically weighted based on their air quality impact.
- **Smoke Impact Estimation:** I developed a smoke impact estimate based on fire size and type. A formula was established to define smoke impact, integrating relevant variables such as the intensity of the fire (quantified by GIS acres) and its proximity to the city:

$$\text{Smoke Estimate} = W \times c \times \text{GIS (SqMiles)} \times \frac{1}{\text{average distance}}$$

,

Where,

- **W:** weight based on the fire type (`wildfire_df['Assigned_Fire_Type_Label']`)
- **c:** Constant of proportionality - Ideally should be determined from empirical observations as there could be factors based on specific environmental conditions that affect the smoke impact differently - For simplicity, it is set to (1) since I don't have any other information.
- **GIS (SqMiles):** Fire size in square miles, `wildfire_df['GIS_SqMiles']`
- **average distance (SqMiles):** Proximity of the fire to the city, `wildfire_df['average_distance']`

To analyze trends over time, the smoke impact was further grouped by **year.**

Statistical Analysis

Statistical methods were employed to analyze the relationships between wildfire smoke estimates and health/economic indicators systematically. The **SciPy** package was used for most of the analyses here.

- **Correlation Analysis:** Spearman correlation coefficients [11] were calculated to assess the strength and direction of relationships between smoke estimates, AQI values, and various factors in the extension plan (health-related and socioeconomic factors). This non-parametric approach was chosen to accommodate non-linear relationships and to provide a robust comparison across varied data distributions. Pearson correlation was not attempted, as I wanted to capture the non-linear associations.

- **Time Series Analysis:** Historical trends in wildfire smoke impact were visualized, and statistical tests (such as the Augmented Dickey-Fuller test [12]) were performed (using **Statsmodels**) to establish the stationarity of the time series data. This further led to the formulation of an ARIMA model to forecast future smoke estimates based on observed trends (More information about ARIMA is mentioned in the next section).

Predictive Modeling

The ARIMA (AutoRegressive Integrated Moving Average) model [13] was chosen for its strengths in time series forecasting, including its ability to capture trends, seasonality, and noise in sequential data. ARIMA's flexibility and interpretability make it an ideal choice for modeling complex temporal patterns while providing reliable forecasts, as in this case.

Due to time constraints, I focused exclusively on ARIMA and did not explore alternative models.

- **Model Selection:** A systematic grid search was conducted to identify the optimal parameters (p, d, q) for the ARIMA model, which included assessing various combinations based on minimizing Root Mean Square Error (RMSE) metrics. This methodological rigor ensured that the selected model accurately represents the underlying data patterns. Libraries utilized for this purpose include **Statsmodels** for fitting the ARIMA model and **Scikit-learn** for evaluating model performance, particularly in calculating metrics such as Root Mean Square Error (RMSE).
- **Forecasting:** The ARIMA model, once trained, was used to predict smoke impact over the next 25 years. The same model was also trained on public health and economic data. This helps authorities understand how wildfire smoke could affect community health and the economy and plan better to handle these impacts.

Data Visualization

Various data visualizations were used to highlight historical trends, future projections, and relationships between factors related to wildfires and air quality. Line charts illustrated historical and forecasted smoke estimates alongside respiratory mortality and poverty rates over time, revealing potential correlations and trends. Histograms showed the distribution of wildfire occurrences by distance from the city, offering insights into wildfire spatial dynamics. Pair plots and correlation heatmaps explored relationships among numerical variables, uncovering patterns influencing smoke estimates. These visualizations, created primarily with **Matplotlib** and enhanced with **Seaborn** for clarity and appeal, provided valuable insights to inform public health decisions and policies regarding wildfire smoke exposure.

Findings

Utilizing a range of datasets and methodologies, several key findings emerged, highlighting critical correlations and trends associated with wildfire incidents and their wider ramifications.

Wildfire Data Acquisition and Trends

The primary dataset used in this analysis was the Combined Wildland Fire Dataset [D1]. The average distance between wildfires and Vancouver was computed, with notable values indicating proximity's effect on potential smoke exposure. After filtering for wildfires that occurred within a 650-mile radius of Vancouver from 1961 to 2021 during the fire season (May 1st to October 31st), the dataset produced a total of 68,394 entries. The line chart (Figure 1)

illustrates the number of wildfire instances each year in the past 50yrs. While there are fluctuations, an overall upward trend is evident, indicating an increasing frequency of wildfires over the past six decades.

Smoke Estimate Calculation

A critical objective was to calculate a smoke estimate for each wildfire incident based on the factors, for which I primarily analyzed relationships among variables in the USGS dataset using Spearman correlation and visualizations, such as pair plots and heatmaps, to understand key factors influencing smoke dispersion. Wildfire size (converted to square miles) and average distance were log-transformed to account for non-linear effects, and fire types were weighted based on their potential smoke impact. Despite limited information, I intuitively modeled smoke estimation.

From Figure 2, I can see that the resultant smoke estimates indicated significant variations across the dataset with an overall upward trend is evident. This suggests that the average amount of smoke in the atmosphere has been increasing over time.

Analyzing the trends of AQI with smoke estimates helps us understand the relationship between particulate matter from smoke and overall air quality. This information can be useful for public health officials, policymakers, and residents to take appropriate measures to protect public health during periods of poor air quality and better prepare for socioeconomic issues.

Figure 3 compares the time series of Smoke Estimate and AQI Value for Vancouver, WA (for the period when AQI data was available). Both Smoke Estimate and AQI Value exhibit significant fluctuations over time, indicating variability in air quality conditions. There also seems to be a positive correlation between the two metrics, suggesting that higher smoke estimates are often associated with higher AQI values. This is expected, as increased smoke in the air can contribute to poor air quality and higher AQI readings. Periods of high smoke estimates, often associated with wildfire events, coincide with spikes in AQI values, highlighting the significant impact of wildfires on air quality.

The Spearman correlation coefficient between the Smoke Estimate and the Air Quality Index (AQI) is ~0.23. This indicates a weak positive correlation, suggesting that as the Smoke Estimate increases, the AQI tends to increase slightly, but the relationship is not strong. The p-value of ~0.22 suggests that this correlation is not statistically significant, indicating that we cannot confidently reject the null hypothesis that there is no correlation between the two measures.

Next, I wanted to forecast the trends in the smoke estimate for the next 25yrs. By doing so, stakeholders can gain insights into the expected variations in air quality due to wildfires and other factors. I used the ARIMA model for this purpose. While ARIMA models can deal with non-stationarity up to a point, they cannot effectively account for time-varying variance. Here I used the Augmented Dickey-Fuller test to tell us if our data has a constant mean and variance.

The ADF Statistic observed was 0.0038. The value is very close to zero, and it generally implies a lack of evidence to support the presence of stationarity within the series. For a series to be stationary, we typically expect the ADF statistic to be significantly negative. Even the p-value was 0.9588. This high p-value (much greater than common significance levels such as 0.05) indicates that we cannot reject the null hypothesis that the series is non-stationary. A p-value of this magnitude suggests that there is substantial evidence indicating that the series may contain trends or other time-dependent properties. There are many methods to transform the data into a stationary series, for example, Differentiating and Performing transformations. I tried differentiating and then re-ran the ADF test. Now I had the ADF statistic as -7.42. This value is significantly negative, far below zero. Even the p-value is exceedingly low now (9.43e-11), much lower than even the significance level of 0.05. Hence, in this case, we can say that the differencing process successfully transformed the "smoke_estimate_scaled" series into a stationary series.

To find the best p, d, and q values, I systematically performed hyperparameter tuning [14][15][16][17]. The optimal ARIMA configuration was identified to be (6, 1, 3). I further forecasted the smoke estimate for the next 25yrs. It is visualized in Figure 4.

The straight blue line shows significant fluctuations in smoke estimates from the 1960s to the early 2020s in Vancouver, WA. This suggests that smoke levels have varied considerably over the past decades. While there are ups and downs, there seems to be a general upward trend in smoke estimates, especially towards the end of the historical period. This could indicate that smoke levels have been increasing over time. The dotted blue line on the other hand represents the forecasted smoke estimates from 2025 to 2050. It shows a continued upward trend, suggesting that smoke levels are expected to increase further in the future. The reasons for this trend could be various factors such as climate change, increased wildfires, or changes in human activities. The shaded area represents the 95% prediction interval. This means that there is a 95% chance that the actual smoke estimates will fall within this range. The widening of the interval towards the end of the forecast period indicates increasing uncertainty in the predictions.

Analyze Respiratory variation with smoke impact

The dataset provides annual mortality rates for various respiratory diseases, including Chronic Respiratory Diseases (CRDs) such as Chronic Obstructive Pulmonary Disease (COPD), which are linked to wildfire smoke. Exposure to smoke accelerates disease progression, increases hospitalizations, and raises mortality risk in individuals with CRDs. Studies show significant spikes in respiratory-related deaths during wildfire events, highlighting the need for public health measures.

From Figure 5, we can observe that Chronic Respiratory Diseases and COPD have average mortality rates of 48.6% and 56.8%, respectively, with males generally being more affected, except in the case of asthma.

I filtered the dataset for CRDs to focus on this cause, including both sexes in the analysis.

From Figure 6, both the smoke estimate and mortality rate show upward trends, indicating a possible link between rising air pollution and worsening respiratory health. Mortality rates fluctuate more than smoke estimates, with a noticeable rise in the early 2000s followed by a gradual decline after mid-decade.

I calculated the Spearman correlation between smoke estimates and mortality rates, which yielded a coefficient of 0.318 (p-value = 0.063). This suggests a weak to moderate positive monotonic relationship. While not statistically significant, the result indicates a potential trend that could be more evident with more data.

To explore the potential long-term impact, I used a time series modeling approach to project the trajectory of smoke estimates and their influence on respiratory mortality over the next 25 years, as shown in Figure 7. The historical data (straight red line) reveals fluctuations in mortality rates, likely influenced by air quality, socioeconomic conditions, and healthcare access. The forecast (dotted red line) predicts continued increases, with a widening 95% prediction interval (shaded red area) indicating greater uncertainty over time. This suggests that future mortality rates could be influenced by various unpredictable factors.

Improving forecast accuracy requires incorporating more detailed data, advanced modeling techniques, and ongoing evaluation of public health interventions to better address the respiratory health risks associated with smoke exposure.

Analyze socio-economic Unemployment Rate with Smoke Impact

Fluctuations in smoke estimates as seen in Figure 8 appear to align with changes in the unemployment rate, though some instances show a delayed effect. The Spearman correlation coefficient of 0.045 (p-value = 0.809) indicates a very weak correlation, suggesting any observed patterns are likely coincidental.

Given the weak correlation, it seems unlikely that smoke exposure directly impacts unemployment rates, though external factors like economic disruptions or health policies might play a role.

Unlike the previous section, where we used time series modeling to forecast long-term trends, this section does not involve such projections, as the relationship between smoke estimates and unemployment rates appears weak and not statistically significant.

Analyze socio-economic Poverty Rate with Smoke Impact

Figure 9 shows that both variables exhibit significant fluctuations over time, but the relationship between them is complex. While there isn't a clear, consistent pattern, some periods show a positive correlation, where both smoke estimates and poverty rates rise or fall together. This could be influenced by economic downturns or severe wildfire seasons that impact both health

and economic conditions. However, in other periods, increases in poverty rates seem to follow high smoke estimates, suggesting that the effects of air pollution and wildfires might have a delayed impact on economic conditions.

The Spearman correlation coefficient of -0.411 (p-value = 0.033) suggests a moderate negative relationship, with higher poverty rates associated with lower smoke estimates. However, this could reflect delays in both physical recovery from health impacts and economic recovery from wildfire disruptions.

Following this, I used a time series modeling approach to project the future trajectory of smoke estimates and their potential impact on poverty over the next 25 years, as shown in Figure 10. The historical data (straight green line) reveals fluctuations in poverty levels, likely influenced by various factors, including economic cycles and social policies. The forecast (dotted green line) predicts continued fluctuations, with periods of increase and decrease, reflecting the complexity of addressing poverty. The wide prediction interval (shaded green area) shows the uncertainty of future poverty rates, influenced by the many unpredictable factors at play.

Although the forecast suggests that poverty will remain a persistent issue, effective policies and interventions could help mitigate its long-term impact.

Analyze socio-economic Premature Deaths with Smoke Impact

As seen in Figure 11, premature death rates fluctuate with changing smoke estimates, with a slight upward trend in both variables over time. While the correlation between smoke estimates and premature deaths is weak (Spearman coefficient = 0.193, p-value = 0.391), there may still be a slight association. The lack of statistical significance suggests that this relationship could be due to chance, and further analysis would be needed to confirm any meaningful link.

Further analysis, potentially with additional variables or more specific data, would be needed to draw definitive conclusions. In this case, I did not proceed with a time series modeling approach to project the trajectory of smoke estimates and their influence on premature deaths over the next 25 years.

Combined Forecast Display

A comprehensive plot combining the historical smoke estimates, mortality rates, and poverty rates while including forecasted estimates is plotted in Figure 12. The plot illustrates a complex interplay between smoke estimates, mortality rates, and poverty rates over time. While there are fluctuations, a general upward trend is evident in all three variables, suggesting a potential link between increased smoke exposure and adverse health and socioeconomic outcomes.

The historical data reveals a clear upward trend in smoke estimates, suggesting an increase in wildfire activity or other factors contributing to air pollution. This rise in smoke exposure seems to correlate with the increasing mortality rates, likely due to the adverse health effects of prolonged air pollution, including respiratory problems and cardiovascular diseases. Additionally, poverty rates have shown fluctuations over time, with a general upward trend in recent years.

The economic impacts of wildfires, such as job losses, business closures, and heightened healthcare costs, could exacerbate poverty, contributing to greater socio-economic inequality. These observations suggest a potential relationship between smoke exposure, mortality, and economic hardship, underscoring the broader societal impacts of wildfire-induced air pollution.

Discussion/Implications

The findings from this project highlights the significant public health and economic challenges posed by increasing wildfire smoke exposure in Vancouver, WA. As in Figure 12, the data reveal a concerning upward trend in both smoke estimates and associated health impacts, particularly respiratory mortality rates.

The observed increase in smoke estimates correlated with higher respiratory mortality rates suggests a direct link between wildfire activity and public health. This finding is particularly critical considering that vulnerable populations, including children, the elderly, and individuals with pre-existing health conditions, are disproportionately affected by wildfire smoke. The implications extend beyond health; economic stability is jeopardized as communities grapple with healthcare costs, increased absenteeism, and declining productivity due to poor air quality.

Given these insights, it's imperative for the city council, city manager, and public health officials to take proactive measures to mitigate these impacts. The time to act is now — as wildfire seasons become more prolonged and severe due to climate change, decisions made today will shape the health and economic resilience of the community in the coming years.

To address the impact of wildfire smoke, both immediate and long-term actions are needed. In the short term, real-time air quality monitoring, public health advisories, clean air shelters, and enhanced healthcare services for respiratory illnesses are essential. Long-term strategies include improving forest management to reduce wildfire frequency, supporting climate change mitigation, and implementing community plans to reduce smoke exposure. Economic recovery programs should aid affected communities, while investing in research and innovation can improve monitoring, fire management, and health interventions, strengthening overall public health responses.

Reflection on Human-Centered Data Science Principles

Throughout this project, human-centered data science principles were pivotal in shaping the methodological framework and guiding decision-making processes. By prioritizing the needs of the community, this analysis went beyond merely documenting numerical trends to acknowledge and connect to the lived experiences of residents. For instance, understanding the direct health implications of smoke exposure informed targeted recommendations, ensuring vulnerable populations; such as children, the elderly, and those with pre-existing health conditions received appropriate protective measures. During the development of the smoke estimate, the analysis explicitly considered the geographical proximity of wildfires to vulnerable communities, acknowledging that those closest would experience the most significant effects. By utilizing

localized data, it was clear that certain neighborhoods in Vancouver, WA, faced higher risks based on their proximity to wildfire events, which directly informed the creation of recommendations regarding clean air shelters in these high-risk areas.

In selecting data, considerable attention was paid to include socioeconomic indicators alongside health metrics. By incorporating data on unemployment rates, poverty levels, and respiratory disease mortality, the analysis embraced a broader perspective that reflects the interconnectedness of health and socioeconomic conditions. It highlighted the importance of considering economic vulnerability when recommending public health interventions, thus aiming to support equity in health outcomes.

Additionally, the project adhered to best practices for open scientific research [19], ensuring transparency and reproducibility throughout its processes. By documenting data sources and methodologies clearly, stakeholders can verify and replicate the analysis, fostering trust within the community regarding the findings and recommendations. This commitment to transparency reflects an understanding that truly human-centered data science is not merely about collecting and analyzing data; it's about empowering communities through accessible and actionable insights.

Limitations

This analysis, while providing valuable insights into the potential impacts of wildfire smoke in Vancouver, WA, is not without limitations. Several factors could potentially influence the accuracy and generalizability of the findings:

Data Quality and Availability:

- **Incomplete Wildfire Data:** The USGS Wildland Fire Combined Dataset [D1], while comprehensive, may not capture all relevant wildfires due to reporting inconsistencies or gaps in historical records, particularly for smaller fires or those occurring in remote areas. This could result in an underestimation of the true smoke impact.
- **Limited Air Quality Data:** The AQS API provides historical air quality data [D2], but it is limited in its spatial and temporal coverage, and the availability of data in Clark County. Vancouver is part of this county, but it may not accurately reflect the exact smoke conditions specific to Vancouver. This may make it challenging to draw direct correlations between smoke exposure and air quality indices.
- **Data Gaps and Missing Values:** The dataset for Respiratory Disease Mortality from IHME (1980-2014) [D3] does not capture mortality trends beyond this period, and it is based on Clark County, which may not be representative of Vancouver. This limits the ability to analyze long-term trends and impacts across a wider timeframe. The FRED dataset for Premature Deaths [D5] only has data from 1999 - 2024, making it difficult to see the long-term trends. We also don't have enough data for the whole period for [D4] and [D6].

Assumptions and Simplifications:

- **Smoke Estimation:** The smoke estimate developed in this analysis relies on simplified assumptions regarding fire size, type, and distance from the city. This estimation does not account for numerous factors that influence smoke distribution, such as wind direction, atmospheric conditions, and fire duration, potentially leading to inaccuracies.
- **Fire Type Weights:** The assigned weights for different fire types are based on general assumptions about their impact on smoke generation. These weights may not accurately represent the actual smoke output of different fire types, especially in specific environmental conditions.
- **Correlation Analysis:** While Spearman correlation was used to account for non-linear relationships between variables, the strength of these relationships is still relatively weak. This suggests that other factors not captured by the analysis may also be influencing the observed trends.

Modeling Assumptions and Limitations:

- **Model Selection:** Due to time constraints, this analysis focused solely on ARIMA for time series modeling. Exploring other models such as SARIMA, Exponential Smoothing, or machine learning techniques could potentially yield more accurate results and provide additional insights.
- **ARIMA Model:** The ARIMA model used for forecasting assumes stationarity in the time series data, and though differencing was applied to achieve stationarity, this process can potentially lose inherent trends within the data, especially for long-term predictions. This limitation could impact the accuracy of the forecasts.

Temporal and Spatial Considerations:

- **Seasonal Variations:** The analysis did not account for seasonal variations in fire occurrence and weather patterns. Smoke dispersion and its effects can vary significantly based on factors like wind direction, humidity, and temperature, which are influenced by seasons.
- **Spatial Heterogeneity:** The analysis focused on Vancouver, WA. However, it is essential to consider that wildfires occurring in distant regions can also impact air quality, even if to a lesser degree. The model does not account for the cumulative effects of multiple wildfires occurring beyond the specified cutoff distance of 650 miles.

Conclusion

This analysis set out to explore the potential impacts of wildfire smoke on Vancouver, WA, focusing on its implications for public health and economic stability. The core research question was: How does wildfire smoke exposure affect respiratory health outcomes, hospitalization rates, and the economic stability of local industries? To answer this question, I explored several hypotheses:

1. **Differential Effects Based on Fire Type:** I hypothesized that different types of fires, like wildfires and prescribed fires, would have varying smoke impacts. This analysis found that larger wildfires, particularly those closer to Vancouver, had a greater impact on air quality, as measured by the smoke estimate.
2. **Increased Smoke Exposure Correlates with Negative Health Outcomes:** I hypothesized that increased smoke exposure correlates with higher mortality rates related to chronic respiratory diseases. The analysis revealed a weak to moderate positive correlation between the smoke estimate and respiratory mortality rates, although the correlation was not statistically significant at the 5% level. This suggests a potential link, but further research is needed to establish a definitive causal relationship.
3. **Wildfire Smoke Affects Economic Stability:** I hypothesized that wildfire smoke negatively impacts economic indicators, leading to higher unemployment rates, increased poverty levels, and premature deaths. Our findings showed a weak negative correlation between the smoke estimate and poverty rate, suggesting that areas with higher poverty rates tend to have lower smoke estimates. This finding may be due to a delay in the observed increase in poverty after high smoke estimates, as the economic and health impacts of wildfires can take time to manifest. The correlation between smoke estimate and unemployment rate was very weak and not statistically significant, suggesting that these factors might not be directly related. I did not observe a significant correlation between smoke estimate and premature death rate.

These findings highlight a complex interplay between wildfire smoke exposure and public health, as well as economic outcomes in Vancouver, WA. While the correlations observed were not always statistically significant, they suggest potential connections that warrant further investigation.

This study also demonstrates key aspects of human-centered data science by using data analysis to quantify the impacts of environmental events, like wildfires, on communities. It emphasizes a holistic perspective by integrating health, economic, and environmental data, highlighting the need for a multidisciplinary approach to address complex societal challenges. The study also brings attention to the ethical considerations of using data to understand public health and economic impacts, stressing the importance of responsible data collection, analysis, and communication.

The limitations discussed in the previous section highlight the need for further research using more extensive data, improved modeling techniques, and a deeper understanding of the interconnectedness of these factors. By continuing to explore these connections, we can create more effective strategies for mitigating the impacts of wildfire smoke and building resilient communities.

References

[1] Why Wildfire Smoke is a Health Concern | US EPA. (2024). US EPA. https://www.epa.gov/wildfire-smoke-course/why-wildfire-smoke-health-concern

[2] Human Exposures, Health Impacts, and Mitigation - The Chemistry of Fires at the Wildland-Urban Interface - NCBI Bookshelf. (2024). nih.gov. https://www.ncbi.nlm.nih.gov/books/NBK588649/

[3] Cascio, W. E. (2018). Wildland fire smoke and human health. Science of The Total Environment. https://doi.org/10.1016/j.scitotenv.2017.12.086

[4] Portland-Vancouver-Salem, OR-WA. (2024). lung.org. https://www.lung.org/research/sota/city-rankings/msas/portland-vancouver-salem-or-wa

[5] rate-trend-comparison. nih.gov. https://hdpulse.nimhd.nih.gov/data-portal/home

[6] em_climateresilience_economic_impact_jan22.pdf. (2022). wa.gov. https://www.dnr.wa.gov/publications/em_climateresilience_economic_impact_jan22.pdf

[7] Reid, C. E et.al., (2016). Critical review of health impacts of wildfire smoke exposure. Environmental Health Perspectives, 124(9), 1334–1343. https://doi.org/10.1289/ehp.1409277

[8] Cecilia E. Rouse. (2024). Health Consequences of Wildfire Smoke. NBER. https://www.nber.org/bh/20243/health-consequences-wildfire-smoke

[9] angela wu. (2024). UCLA study of wildfire smoke's long-term health effects finds upwards of 50,000 deaths in 11 years | UCLA Luskin Center for Innovation. UCLA Luskin Center for Innovation. https://innovation.luskin.ucla.edu/2024/06/18/ucla-study-of-wildfire-smokes-long-term-health-effects-finds-upwards-of-50000-deaths-in-11-years/

[10] Community Wildfire Mitigation Best Practices Toolbox – Coalitions & Collaboratives. (2024). co-co.org. https://co-co.org/community-wildfire-mitigation-best-practices-toolbox/

[11] Spearman's rank correlation coefficient - Wikipedia. https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

[12] Augmented Dickey–Fuller test - Wikipedia. https://en.wikipedia.org/wiki/Augmented_Dickey–Fuller_test

[13] ARIMA for Time Series Forecasting: A Complete Guide. DataCamp. https://www.datacamp.com/tutorial/arima

[14] What are the values p, d, q, in ARIMA? - StackExchange. https://stats.stackexchange.com/questions/44992/what-are-the-values-p-d-q-in-arima

[15] Şeyma Aysu Demir, A Guide to Parameter Tuning in auto_arima() Function for Time Series Forecasting - Medium, https://medium.com/@aysuudemiir/a-guide-to-parameter-tuning-in-auto-arima-function-for-time-series-forecasting-aec50fb1523a

[16] Jason Brownlee, How to Grid Search ARIMA Model Hyperparameters with Python. machinelearningmastery.com, December 10, 2020, https://machinelearningmastery.com/grid-search-arima-hyperparameters-with-python/

[17] Using k-fold cross-validation for time-series model selection - StackExchange. https://stats.stackexchange.com/questions/14099/using-k-fold-cross-validation-for-time-series-model-selection

[18] Which Populations Experience Greater Risks of Adverse Health Effects Resulting from Wildfire Smoke Exposure? | US EPA. (2024). US EPA. https://www.epa.gov/wildfire-smoke-course/which-populations-experience-greater-risks-adverse-health-effects-resulting

[19] The practice of reproducible research by Justin Kitzes, Daniel Turek, Fatma Deniz - paper. University of California Press. (n.d.). https://www.ucpress.edu/books/the-practice-of-reproducible-research/paper

Appendix

The images referenced in the report are included below:

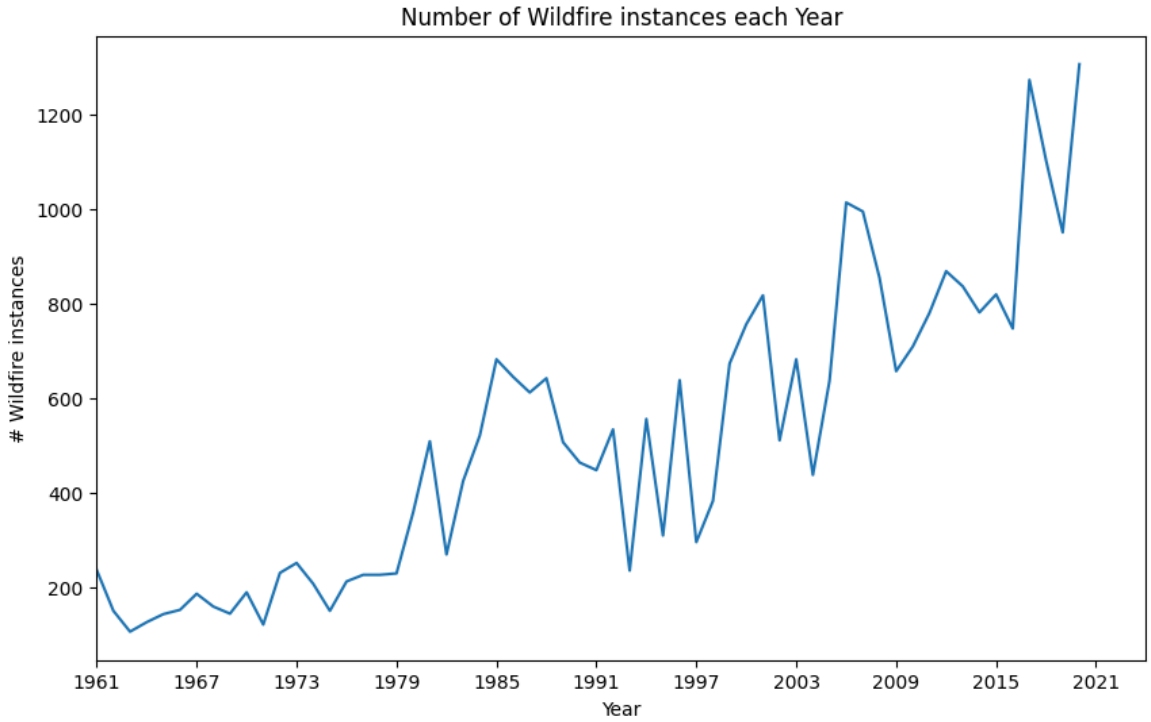

Figure 1: Number of wildfire instances each year from 1961 to 2021

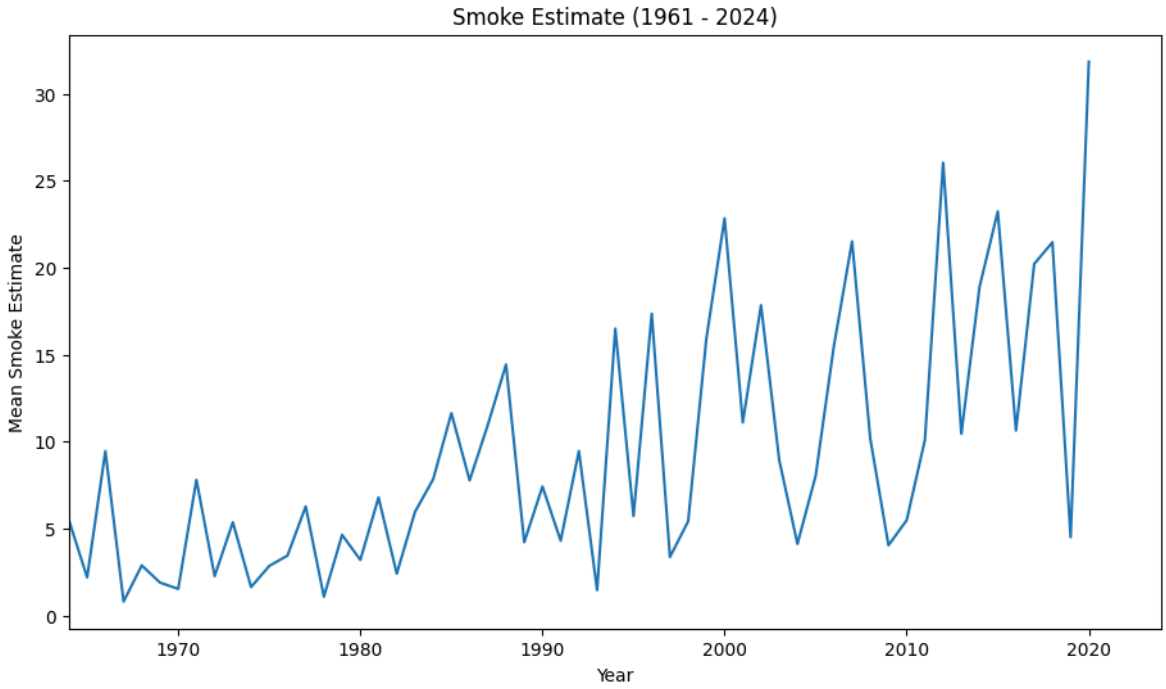

Figure 2: Mean smoke estimate from 1961 to 2021

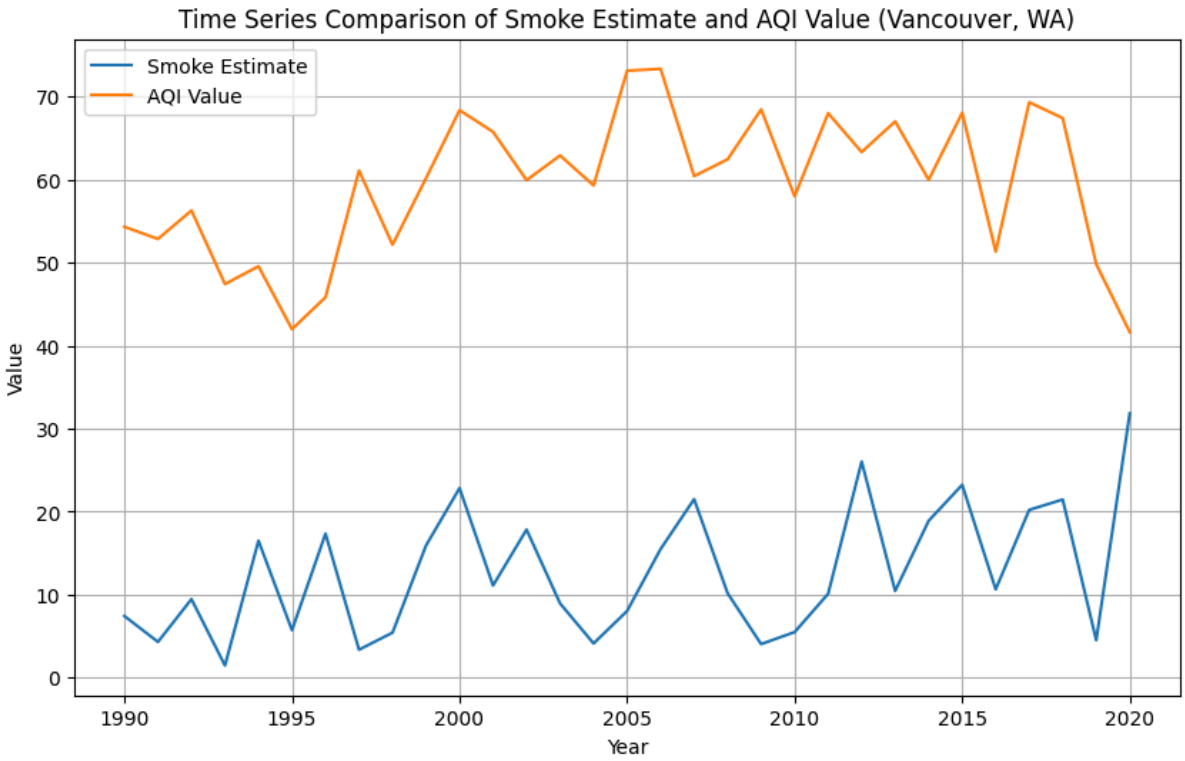

Figure 3: Time Series Comparison of Smoke Estimate and AQI value

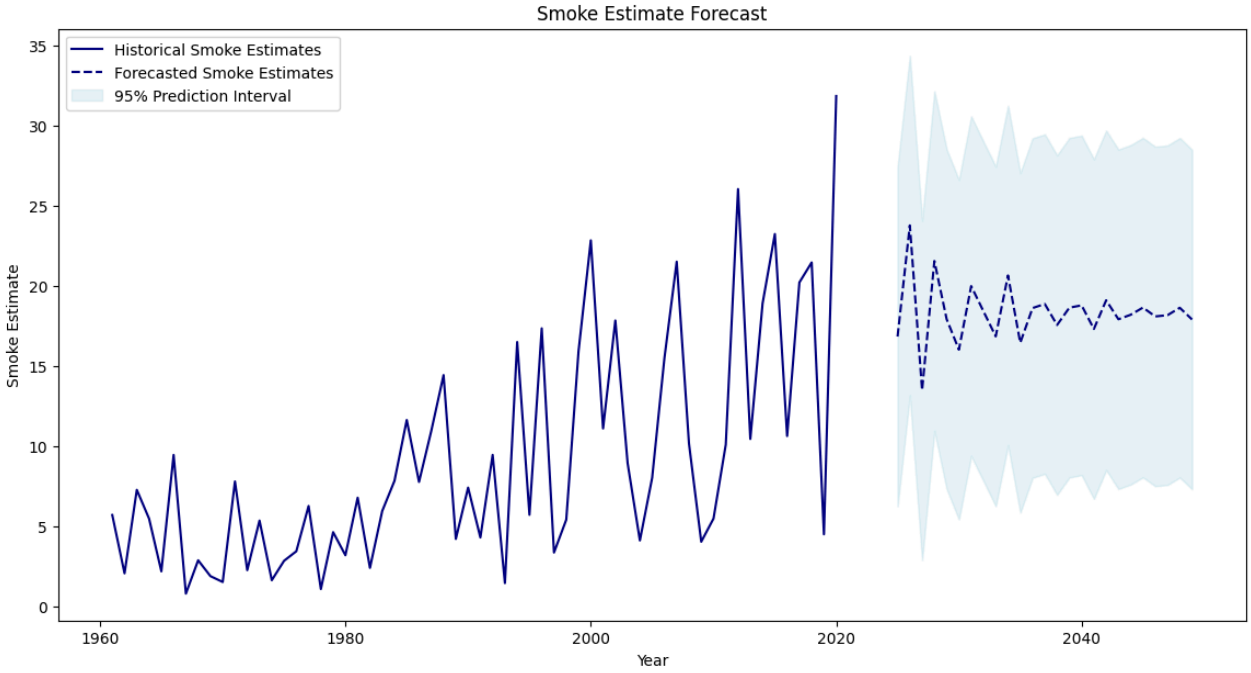

Figure 4: Smoke Estimate Forecast for the next 25 years (2025 - 2050)

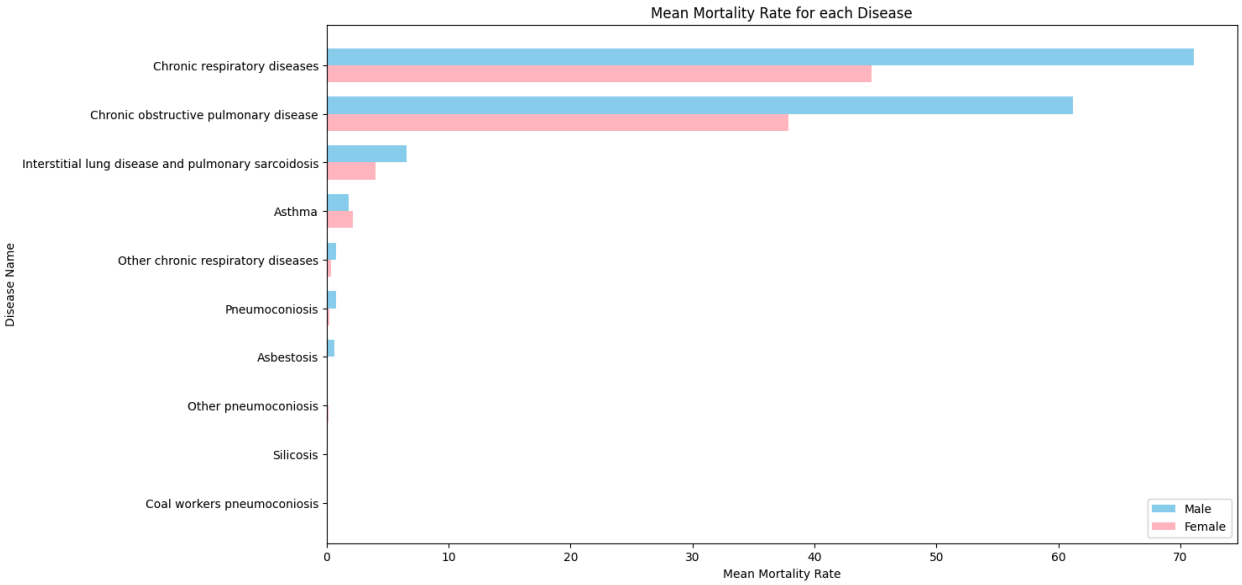

Figure 5: Mean Mortality Rate for each Disease

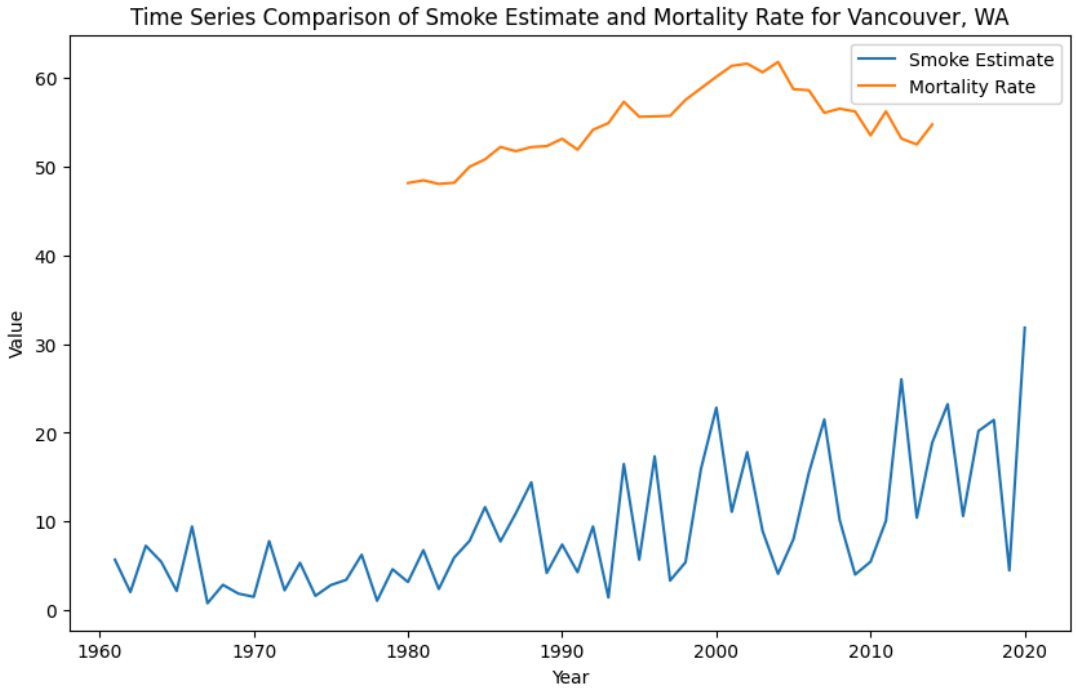

Figure 6: Time Series Comparison of Smoke Estimate and Mortality Rate for Vancouver, WA

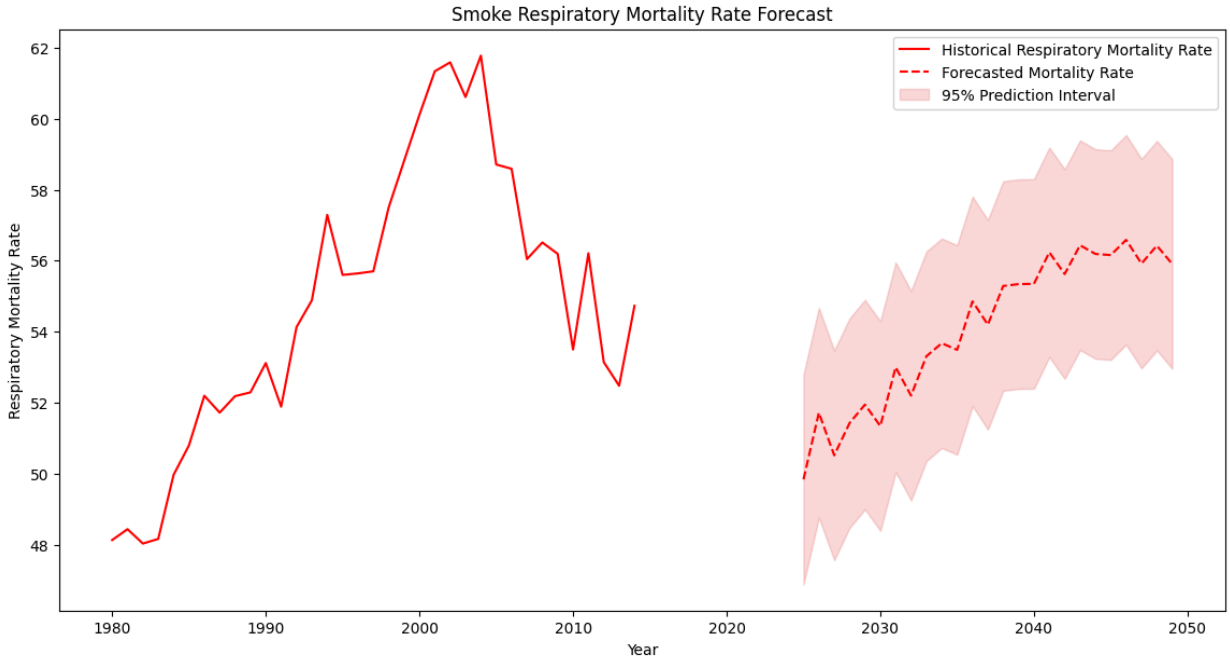

Figure 7: Respiratory Mortality Rate Forecast

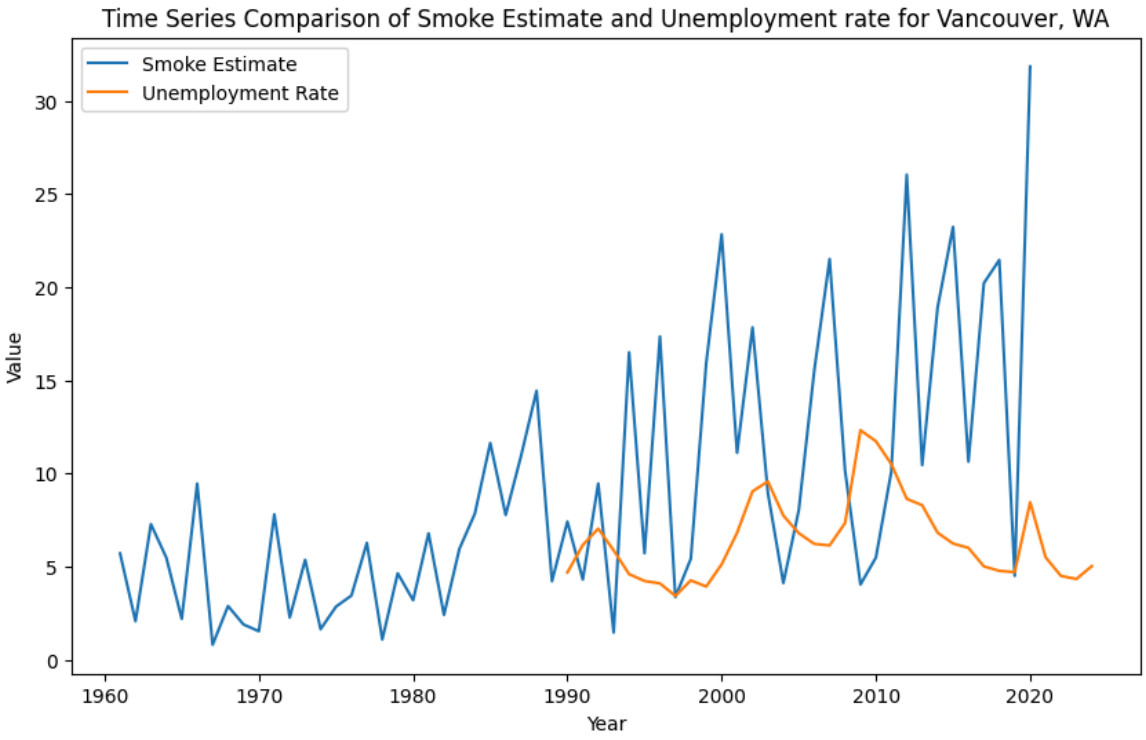

Figure 8: Time Series Comparison of Smoke Estimate and Unemployment rate for Vancouver

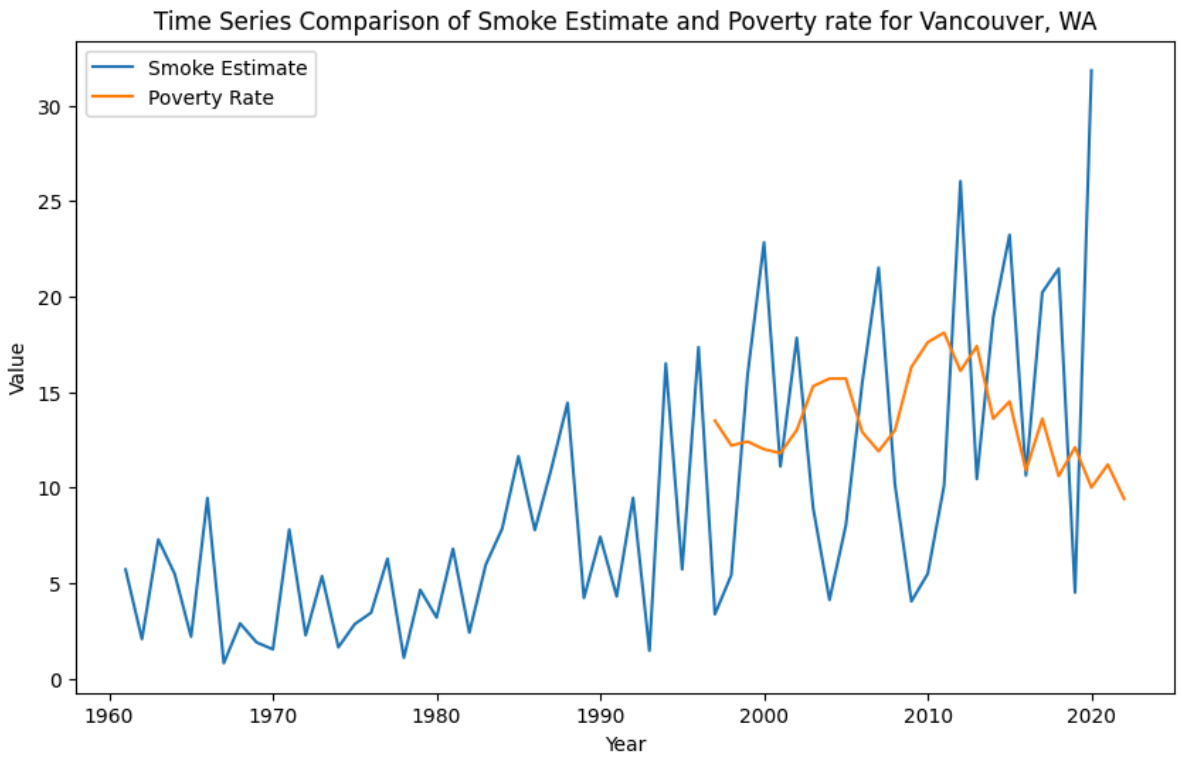

Figure 9: Time Series Comparison of Smoke Estimate and Poverty rate for Vancouver, WA

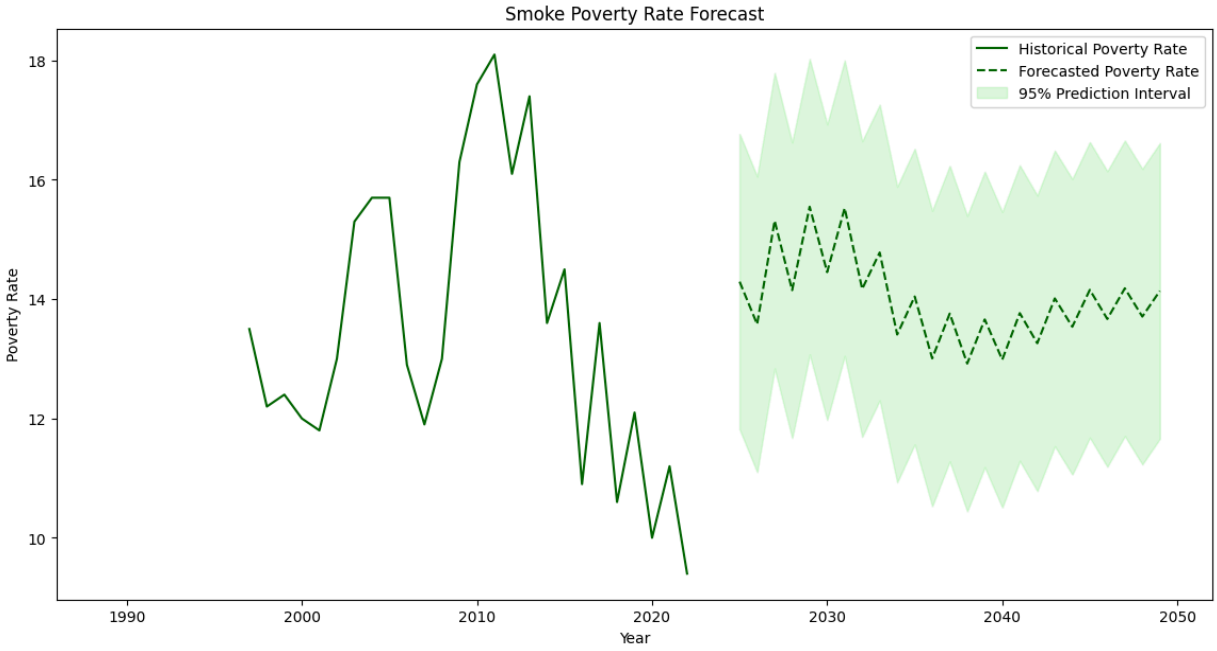

Figure 10: Poverty Rate Forecast

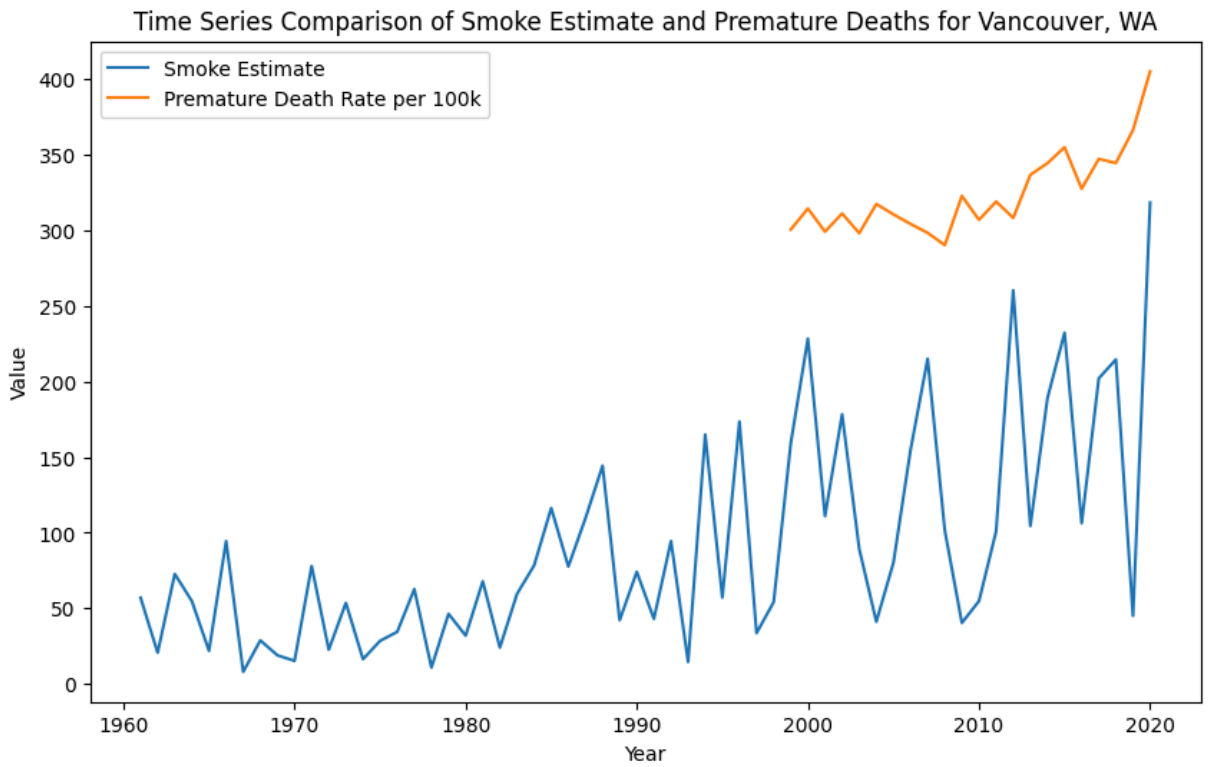

Figure 11: Time Series Comparison of Smoke Estimate and Premature Deaths for Vancouver

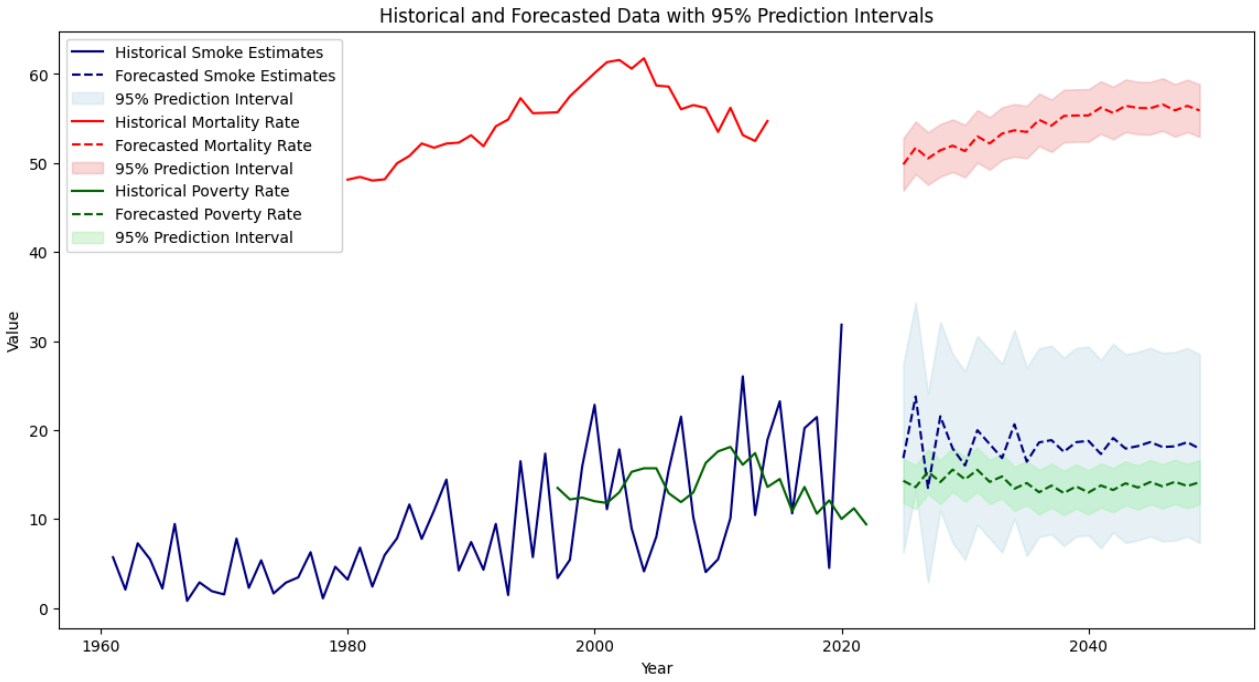

Figure 12: Historical and Forecasted Data with 95% Prediction Intervals for Smoke Estimate, Mortality, and Poverty Rate.

Data Sources

[D1] USGS Wildland Fire Combined Dataset

This dataset serves as the primary source of data for the project and contains comprehensive wildfire data.

a. **Description:** The dataset includes detailed information on wildfires and prescribed fires across the United States from the mid-1800s to the present. It merges data from 40 different sources, providing a rich and comprehensive view of wildfire activity over time. The dataset can be accessed from Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons).

b. **Columns:** The dataset contains a variety of attributes, including:
   - **Fire_Year:** The calendar year when the fire occurred as determined by the dataset creators.
   - **Fire_Polygon_Tier:** The tier from which the fire polygon was generated; one or more polygons within the tier could be combined to create the fire polygon.
   - **Fire_Attribute_Tiers:** Lists all fire tiers that contributed attributes to the fire feature, indicating where a polygon intersects the fire perimeter in space and time.
   - **GIS_Acres:** The GIS-calculated area of the fire polygon in acres, derived using ArcGIS Pro's Calculate Geometry tool.
   - **GIS_Hectares:** The GIS-calculated area of the fire polygon in hectares, derived using ArcGIS Pro's Calculate Geometry tool.
   - **Source_Datasets:** Lists all original source datasets contributing to either the polygon or its attributes, with the number of polygons contributed indicated in parentheses.
   - **Listed_Fire_Types:** Each fire type from the merged dataset that intersects the polygon in space and time, with the number of features contributing the fire type listed in parentheses.
   - **Listed_Fire_Names:** Each fire name from the merged dataset that intersects the polygon in space and time, with the number of features contributing the fire name listed in parentheses.
   - **Listed_Fire_Codes:** Codes for each fire from the merged dataset that intersect the polygon; number of contributing features is indicated in parentheses.
   - **Listed_Fire_IDs:** Each fire ID from the merged dataset that intersects the polygon in space and time, with the number of contributing features indicated in parentheses.
   - **Listed_Fire_IRWIN_IDs:** Each IRWIN ID from the merged dataset intersecting the polygon, with the number of contributing features indicated in parentheses.
   - **Listed_Fire_Dates:** Dates associated with each fire in the merged dataset that intersect the polygon; the number of contributed features is indicated in parentheses.
   - **Listed_Fire_Causes:** Causes for each fire in the merged dataset that intersect the polygon; the number of contributing features is indicated in parentheses.
   - **Listed_Fire_Cause_Class:** Classifications for the causes of fires in the merged dataset that intersect the polygon; the number of contributing features is indicated in parentheses.

- **Listed_Rx_Reported_Acres:** Reported acres of prescribed fires in the merged dataset that intersect the polygon; number of contributing features is indicated in parentheses.
- **Listed_Map_Digitize_Methods:** The methods used to digitize the polygons in the merged dataset that intersect the polygon; the number of contributing features is indicated in parentheses.
- **Listed_Notes:** Notes related to each fire in the merged dataset that intersects the polygon; the number of contributing features is indicated in parentheses.
- **Processing_Notes:** Indicates any attribute changes made during processing, along with a justification for the changes.
- **Wildfire_Notice:** A notice indicating the quality of wildfire data in the dataset.
- **Prescribed_Burn_Notice:** A notice indicating the quality of prescribed burn data in the dataset.
- **Wildfire_and_Rx_Flag:** A flag indicating whether the fire was classified as both a wildfire and a prescribed fire; may indicate errors in assignment.
- **Overlap_Within_1_or_2_Flag:** Flags whether a fire overlaps with previous fires, indicating the extent of overlap and related attributes.
- **Circleness_Scale:** A measure of how closely the polygon resembles a true circle, calculated using the shape area and perimeter.
- **Circle_Flag:** Flags whether a polygon's circle-ness value is indicative of being circular, potentially highlighting incorrect classifications.
- **Exclude_From_Summary_Rasters:** Indicates whether a fire was excluded from raster summary calculations due to being classified as circular.
- **Shape_Length:** The calculated perimeter length of the polygon in meters.
- **Shape_Area:** The calculated area of the polygon in square meters.

c. **Access:** The dataset is available in a zip format ('GeoJSON Files.zip' file) and is cited by Welty, J.L., and Jeffries, M.I. (2021). Combined wildland fire datasets for the United States and certain territories, 1800s-Present: U.S. Geological Survey data release. https://doi.org/10.5066/P9ZXGFY3. USGS-authored or produced data and information are considered to be in the U.S. Public Domain. This dataset is not copyrighted material, hence, when using information from USGS information products, publications, or websites, we need to provide proper credit. Credit can be provided by including a citation. For the ones not in the domain (this one is in the domain), USGS policy on the use of copyrighted material. These materials are generally marked as being copyrighted. To use these copyrighted materials, we must obtain permission from the copyright holder under copyright law. For more information, refer to usgs.gov: copyrights-and-credits

[D2] Air Quality Index Data

Air Quality Data was needed to evaluate the performance of the smoke estimate created.

a. Description: This data was requested from the US Environmental Protection Agency (EPA) Air Quality Service (AQS) API. This is a historical API and does not provide real-time air quality data. The documentation for the API provides definitions of the different call parameters and examples of the various calls that can be made to the API. The US EPA was created in the early 1970's. The EPA reports that they only started

broad based monitoring with standardized quality assurance procedures in the 1980's. Many counties will have data starting somewhere between 1983 and 1988. However, some counties still do not have any air quality monitoring stations. The API helps resolve this by providing calls to search for monitoring stations and data using either station ids, or a county designation or a geographic bounding box. Some additional information on the Air Quality System can be found in the EPA FAQ on the system.

b. **Columns:**
    - **date:** Date of the observed air quality measurement.
    - **site_id:** Unique identifier for the air quality monitoring site.
    - **state:** The state where the monitoring site is located.
    - **county:** The county designation for the air quality monitoring site.
    - **pollutant_type:** Type of pollutant being measured (e.g., PM2.5, Ozone).
    - **sample_duration:** Length of time the sample was taken (e.g., 24 hours).
    - **arithmetic_mean:** Average concentration of the pollutant for the specified sample duration.
    - **AQI:** Air Quality Index value calculated based on the measured pollutant concentration, indicating the overall air quality level.
    - **value_represented:** Description of what is captured by the parameter being measured.
    - **observation_count:** Total number of observations used to calculate the arithmetic mean for that day.

c. **Access:** The AQS API is publicly accessible and provides access to ambient air sample data collected by various pollution control agencies. Users are expected to be familiar with the terms of service that govern the use of the API. These terms outline acceptable usage practices and any restrictions on data access and storage

[D3] Respiratory Disease Mortality (IHME)

This dataset provides information on mortality rates due to various respiratory diseases in Clark County, which includes Vancouver.

a. **Description:** Contains data related to various respiratory conditions, categorized by cause and sex for years: 1980 to 2014. The dataset can be accessed from here: https://ghdx.healthdata.org/record/ihme-data/united-states-chronic-respiratory-disease-mortality-rates-county-1980-2014 and can be further filtered to obtain Clark County, WA.

b. **Columns:**
    - **date:** Date of the observed measurement.
    - **measure_id:** Unique identifier for the health measure being reported.
    - **measure_name:** Name of the health measure.
    - **location_id:** Unique identifier for the geographic location.
    - **location_name:** Name of the location where the data was collected (e.g., "Clark County").
    - **FIPS:** Federal Information Processing Standards code, a unique identifier for geographic areas.
    - **cause_id:** Unique identifier for specific causes of respiratory disease.
    - **cause_name:** Name of the cause of death (e.g., "Chronic respiratory diseases").

- **sex_id:** Unique identifier for sex categorization (e.g., Male, Female).
- **sex:** The sex of individuals represented in the data (e.g., "Both", "Male", "Female").
- **age_id:** Unique identifier for age groups in the dataset.
- **age_name:** Description of the age group (e.g., "All Ages").
- **year_id:** Year corresponding to the data entry (e.g., "1980").
- **metric:** Specifies the type of metric (e.g., "Rate" or "Count").
- **mx:** Value representing the mortality rate for the specified cause (e.g., deaths per 100,000).
- **lower:** Lower confidence interval for the mortality rate estimate.
- **upper:** Upper confidence interval for the mortality rate estimate.

c. **Access:** The dataset is available in a CSV format. The terms and conditions for using the data from the Institute for Health Metrics and Evaluation (IHME) are governed by the IHME FREE-OF-CHARGE NON-COMMERCIAL USER AGREEMENT. Users are permitted to use, share, modify, or build upon the data for non-commercial purposes. For inquiries related to commercial use, it is advised to refer to the IHME Terms and Conditions. For detailed information, it is recommended to consult the IHME website directly.

[D4] Unemployment rate (FRED)

Includes the data from the Federal Reserve Bank of St. Louis Economic Data (FRED) that covers unemployment rates and economic activity in Clark County.

a. **Description:** This dataset includes time series data on unemployment rates (years: 1990 - 2024), which can be correlated with smoke events and economic downturns. The dataset can be accessed from here: https://fred.stlouisfed.org/series/WACLAR1URN

b. **Columns:**
   - **DATE:** Date when the measurement was recorded.
   - **WACLAR1URN:** Unemployment Rate (%)

c. **Access:** The dataset is available in a CSV format and can be used freely. The data is available for public access without any subscription or payment required. When using the data, proper citation of the source is encouraged. It is important to credit FRED as the source of the data. Users are mainly encouraged to use the data for non-commercial purposes. FRED data is typically released under a Creative Commons license which allows for use, sharing, modification, and adaptation as long as appropriate credit is given. For detailed terms, refer to the FRED website.

[D5] Poverty rate (FRED)

Includes the data from the Federal Reserve Bank of St. Louis Economic Data (FRED) that covers poverty rates (age 0-17) and economic activity in Clark County.

a. **Description:** This dataset includes time series data on poverty rates (years: 1989-2022), which can be correlated with smoke events and economic downturns. The dataset can be accessed from here: https://fred.stlouisfed.org/series/PPU18WA53011A156NCEN

b. **Columns:**
   - **DATE:** Date when the measurement was recorded.

    - **PPU18WA53011A156NCEN:** Poverty Rate (%)
c. **Access:** The dataset is available in a CSV format and can be used freely. The data is available for public access without any subscription or payment required. When using the data, proper citation of the source is encouraged. It is important to credit FRED as the source of the data. Users are mainly encouraged to use the data for non-commercial purposes. FRED data is typically released under a Creative Commons license which allows for use, sharing, modification, and adaptation as long as appropriate credit is given. For detailed terms, refer to the FRED website.

[D6] Premature Death Rate (FRED)

Includes data from the Federal Reserve Bank of St. Louis Economic Data (FRED) that covers premature death rates in Clark County.

a. **Description:** This dataset includes time series data on premature death rates, which can be correlated with smoke events and economic downturns. The dataset can be accessed from here: https://fred.stlouisfed.org/series/CDC20N2UAA053011
b. **Columns:**
    i. **DATE:** Date when the measurement was recorded.
    ii. **CDC20N2U053011:** Premature Death (per 100k)
c. **Access:** The dataset is available in a CSV format and can be used freely. The data is available for public access without any subscription or payment required. When using the data, proper citation of the source is encouraged. It is important to credit FRED as the source of the data. Users are mainly encouraged to use the data for non-commercial purposes. FRED data is typically released under a Creative Commons license which allows for use, sharing, modification, and adaptation as long as appropriate credit is given. For detailed terms, refer to the FRED website.