Project Description and Results

Methodology

Data Acquisition

Data plays a critical role in this analysis. The following steps were undertaken to acquire relevant datasets:

- **Wildfire Data:** The core dataset used in this project is the Combined Wildland Fire Datasets for the United States sourced from the US Geological Survey (USGS) [D1], which includes comprehensive historical records of wildfires. This dataset was selected for its extensive geographic coverage, merging data from over 40 sources, and its long temporal range dating back to the mid-1800s. These characteristics allow for a robust examination of trends and impacts over time.
- **Air Quality Data:** AQI data [D2] was obtained through the US Environmental Protection Agency (EPA) Air Quality Service (AQS) API. This data is essential for correlating the health impacts of wildfire smoke with quantifiable air quality measurements. The AQI is an established metric that indicates air quality and its health implications based on the concentrations of various pollutants, including particulate matter resulting from wildfires.
- **Health and Socioeconomic Data:** The analysis was further supported by datasets capturing respiratory disease mortality rates across Clark County, WA from the Institute for Health Metrics and Evaluation (IHME) [D3], alongside socioeconomic data (unemployment rates, premature deaths, and poverty rates) sourced from the Federal Reserve Bank of St. Louis (FRED) [D4][D5][D6]. This multifaceted approach to data acquisition allows for a more nuanced understanding of how smoke exposure from wildfires correlates with health outcomes and economic stability.

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:** There were not a lot of missing values, hence, I didn't really have to do any imputations. Further, all datasets were standardized to ensure consistency in column naming and types. This included unifying how wildfire incidents were described and ensuring all geographic coordinates aligned with the same reference system

(EPSG:4326). Additionally, the dataset was filtered to include only wildfires that occurred during the fire season (from May 1st to October 31st). This selection process ensured that the analysis focused on relevant wildfires that would have impacted air quality during the typical wildfire-related air contamination periods. In the next step, any duplicate records were identified and removed from the wildfire dataset to ensure the integrity of the analysis. Irrelevant entries - such as those without an identifiable fire cause or that were flagged as circular and greater than one acre - were also omitted. Finally, columns containing categorical variables, such as 'Listed_Fire_Causes' and 'Assigned_Fire_Type', were cleaned to eliminate any unnecessary characters and to create a standardized list of categorizations. For example, fire types were given numerical weights based on their expected impact on air quality.

- **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 employed to elucidate historical trends, future predictions, and the relationships between different factors relevant to wildfires and air quality. Line charts were predominantly utilized to showcase the historical and forecasted smoke estimates alongside respiratory mortality rates and poverty rates over time, providing a clear view of how these variables interact with one another, thereby identifying potential correlations and trends. Furthermore, histograms were created to display the distribution of wildfire occurrences by distance from the city, enabling a better understanding of the spatial dynamics of wildfires. Pair plots and correlation heatmaps were also employed to investigate the relationships among multiple numerical variables, revealing underlying patterns that influence smoke estimates. Each of these visual tools helped me understand the data better, highlighting key trends and interactions essential for informing public health decisions and policy implications in the context of wildfire smoke exposure. **Matplotlib** was the primary library used for creating most of the charts. **Seaborn** was further used to enhance visual appeal and clarity.

Ethical Considerations

Human-centered considerations were integral throughout the design of the study. All research practices, data sources, methodologies, findings and limitations were documented clearly to promote transparency and reproducibility.
-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 sourced from the US Geological Survey (USGS), which as mentioned before, integrated data from 40 different sources covering wildfires from the mid-1800s to the present. 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 below (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.

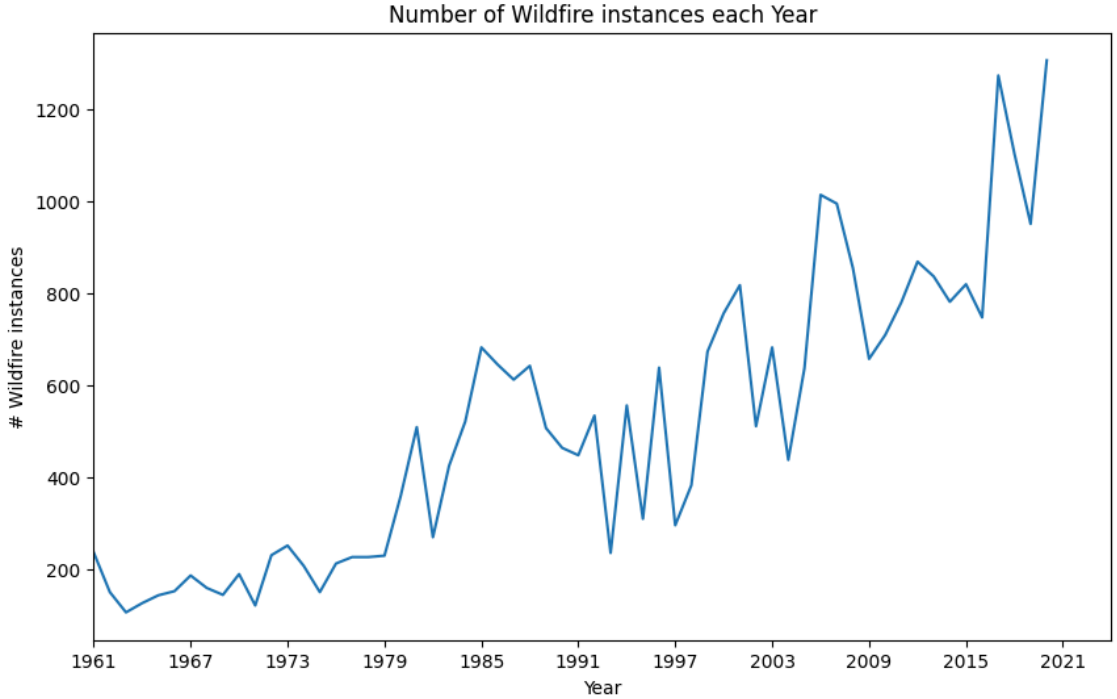

Figure 1: Number of wildfire instances each year from 1961 to 2021

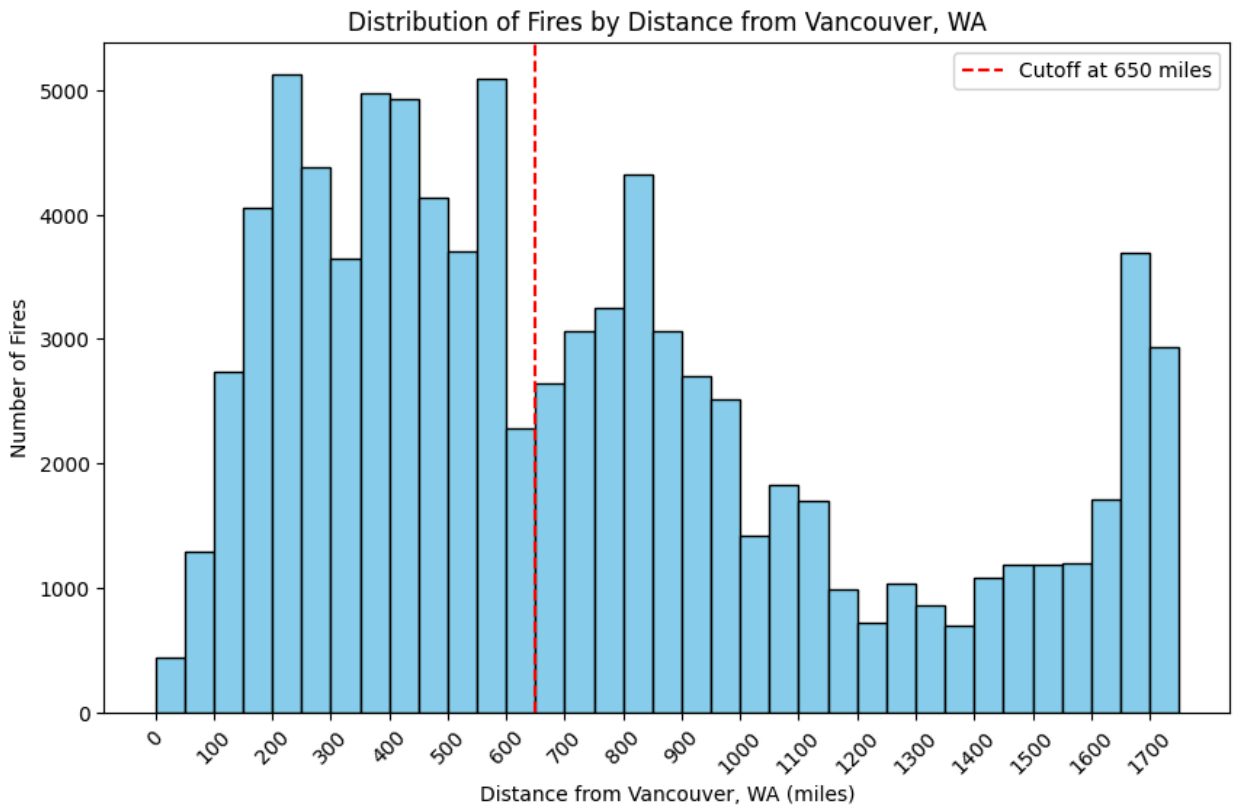

Figure 2: Distribution of Fires by Distance (cut off at 650 mile mark)

This histogram in Figure 2 visualizes the distribution of fires based on their distance from Vancouver, WA. The x-axis represents the distance from Vancouver in miles, and the y-axis represents the number of fires observed at each distance range. To interpret the figure, the viewer can examine the height of each bar. A taller bar indicates a higher number of fires at that specific distance range. For instance, the tallest bar, located around the 100-mile mark, suggests that the highest number of fires occur within 100 miles of Vancouver. The underlying data consists of the average distance from Vancouver, WA to the occurrence of fire for all dates. The distances were grouped into 50-mile intervals, and the number of fires in each interval was counted to create the histogram distribution. The highest frequency of fires occurs within 100 miles of Vancouver, suggesting that the immediate vicinity of the city is most susceptible to fires. The histogram shows a general trend of decreasing frequency with increasing distance. This indicates that the number of fires decreases as the distance from Vancouver increases, which could be attributed to factors like fuel availability, topography, and human activities. The vertical dashed line at 650 miles represents a cutoff or limit in the data collection or analysis. The data beyond this distance was excluded or not considered in the analysis. The distribution of fires by distance from Vancouver, WA, somewhat appears to be right-skewed (though I have a lot more after the 1400 miles mark). This means that there is a long tail to the right, indicating that a majority of the fires occur closer to Vancouver, with a smaller number of fires occurring at greater distances. This distribution aligns with the expected pattern of fire occurrence. Typically, wildfires are more frequent in areas with higher population density, more human activity, and

abundant fuel sources. As the distance from Vancouver increases, these factors generally decrease, leading to a lower frequency of fires. We have a lot more fires after the 1400 mark, probably referring to cities like LA in California where there are more fires reported.

In Figure 3, I have a line chart that visualizes the total number of acres burned per year due to fires occurring within a 650-mile radius of Vancouver, WA. The x-axis represents the year, and the y-axis represents the total number of acres burned. To interpret the figure, the viewer can examine the position of the data points and the line connecting them. A higher data point indicates a larger number of acres burned in that particular year. The line connecting the points helps visualize the trend in the number of acres burned over time. The data consists of a dataset containing information about wildfires, including their distance from Vancouver, WA, size, the year they occurred, and various other factors for all dates. The data was processed to identify fires within a 650-mile radius of Vancouver, WA in the initial step. The total number of acres burned was calculated for each year, and these values were plotted on the graph. The graph shows significant fluctuations in the number of acres burned from year to year. This suggests that wildfire activity varies considerably over time, likely influenced by factors such as weather conditions, fuel availability, and human activities. While there are periods of high and low fire activity, there appears to be a general upward trend in the number of acres burned, especially in recent years. This could indicate an increase in the severity and frequency of wildfires. There are a few years with extremely high values, representing years with significantly larger wildfires. These outliers could be attributed to extreme weather events, such as droughts or heatwaves, or large-scale wildfires.

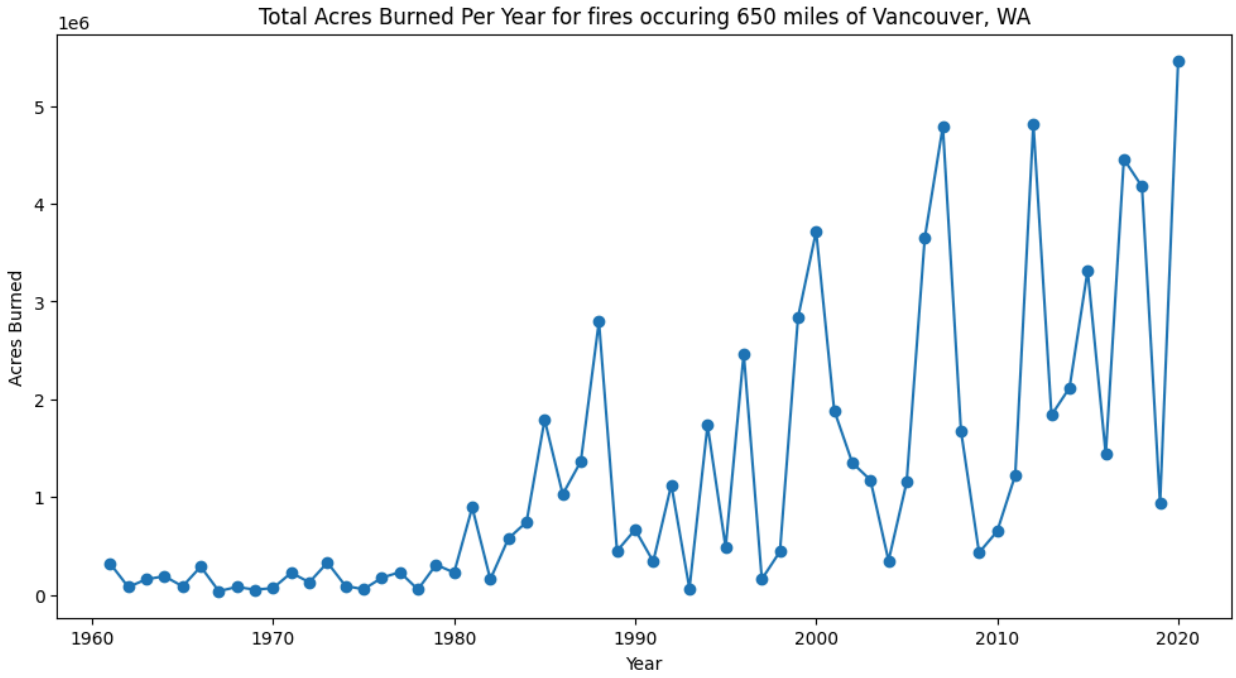

Figure 3: Total Acres Burned Per Year for fires occurring 650 miles of Vancouver, WA

Smoke Estimate Calculation

A critical objective was to calculate a smoke estimate for each wildfire incident based on the factors in the USGS dataset.

To design the formula, I first wanted to understand the relationship between the variables in the dataset. Before computing the correlation coefficient, I wanted to see if the features have a linear relationship between them. However, it seems like it doesn't. The plot below (Figure 4) shows that most of the features are not linearly related to other variables.

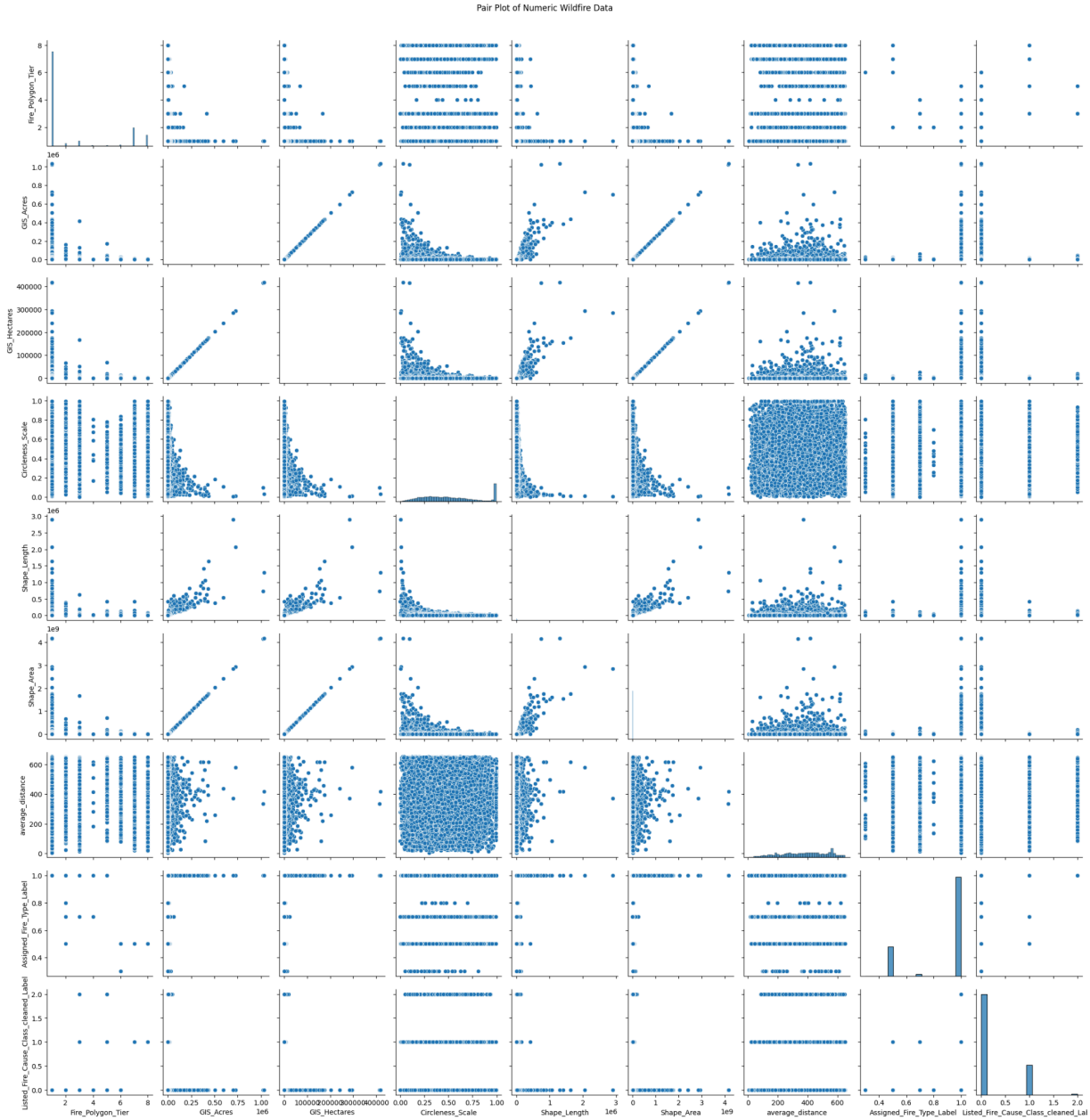

Figure 4: Pair plot showing the relationship between many variables in the USGS dataset indicating why I chose Spearman correlation over Pearson's.

To better understand how correlated the variables are, I computed the Spearman correlation.

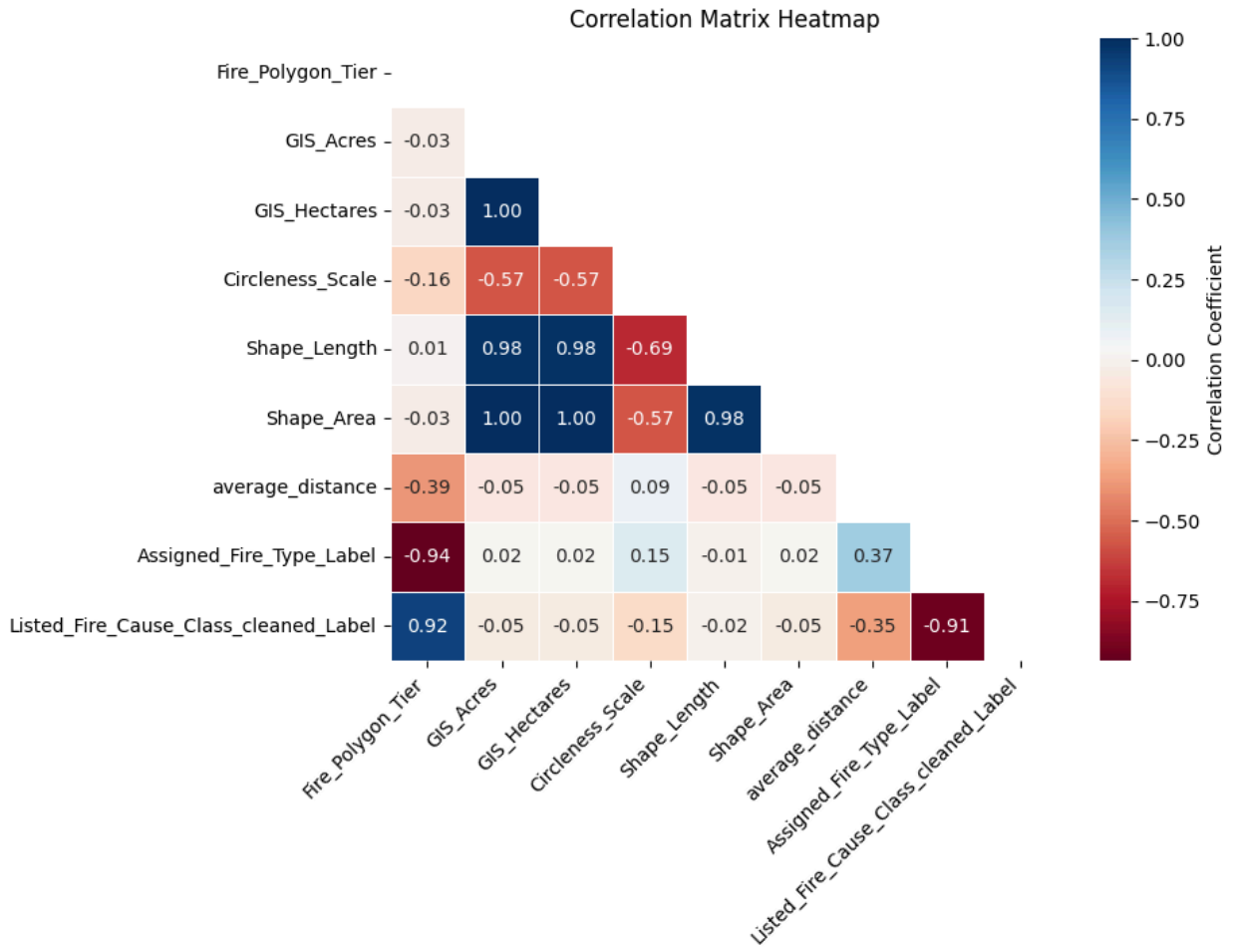

Figure 5: Spearman correlation Matrix Heatmap

From Figure 5, It was observed that,

- The Circleness_Scale was calculated by using the following equation "4*pi*(Shape_Area/(Shape_Length * Shape_Length))". Probably that is a reason why the fields are even a bit related.
- The Listed_Fire_Cause_Class_cleaned_Label seems to be highly negatively correlated with Assigned_Fire_Type_Label. Since Assigned_Fire_Type_Label can be related to smoke estimation, let us ignore Listed_Fire_Cause_Class_cleaned_Label from further analysis.
- The average_distance is also a bit negatively correlated with the Fire Type. As smoke rises and travels away from the fire, it spreads out and mixes with the clean air around it, which reduces its concentration. To better understand this effect, I decided to use average_distance in calculating my smoke estimate: divide the expected smoke impact by the average distance from the fire. This means that the closer a location is to the fire, the higher the smoke concentration will be, while areas farther away will experience less smoke.

The size of wildfires was converted from acres to square miles for a standardized measure of impact. To craft the formula, I want to understand how this converted GIS_SqMiles (in square miles) is related to the average_distance.

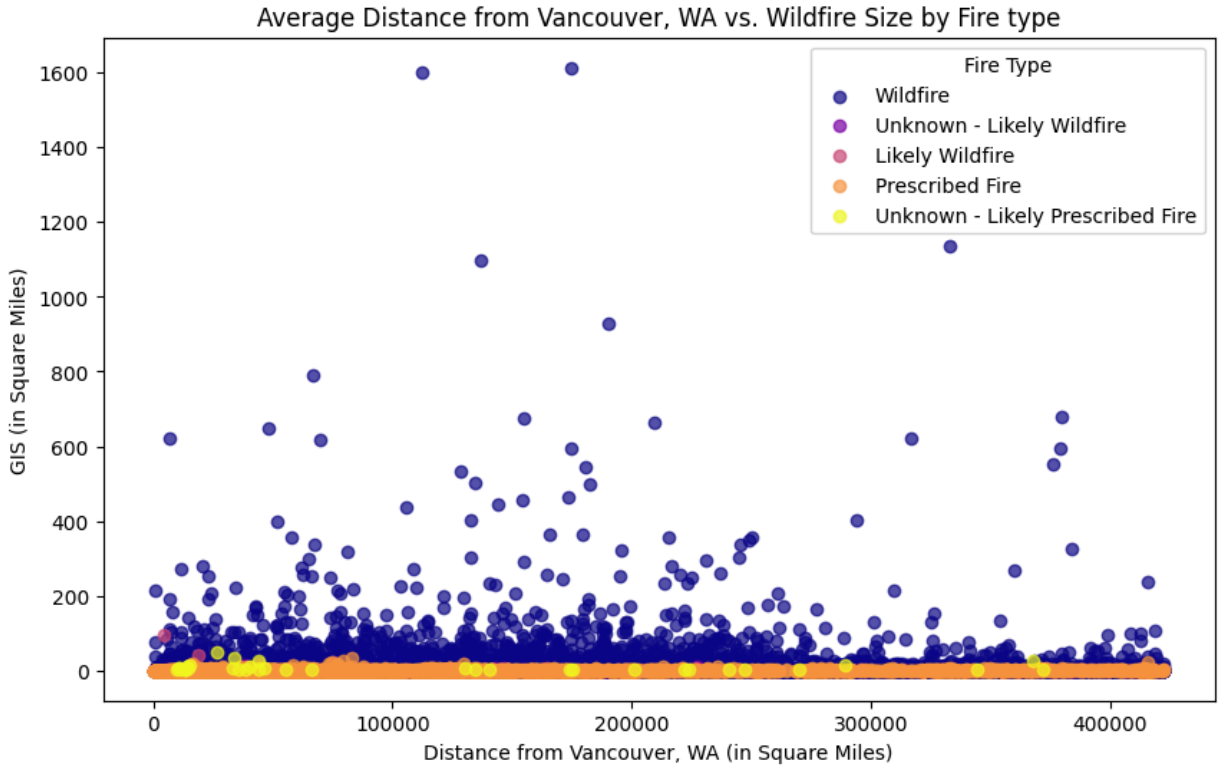

Figure 6: Average Distance from Vancouver, WA v.s Wildfire Size by Fire Type

In Figure 6, I did not see any particular trend here (Except maybe that only Fire Type = Wildfire are more in number and have larger area). They also don't seem to be very well correlated (Spearman Correlation = -0.05). It looks like most of the wildfires are small in size. They don't vary much with distance. Hence, I thought of applying a log transformation to visualize it better. By doing so, I can capture the non-linear effect. Both the area and distance are weighted now in a way that reflects their complex interactions. This also helps reduce the impact of outliers.
I also add 1 to both GIS_SqMiles and average_distance to avoid the issue of taking the logarithm of small values.

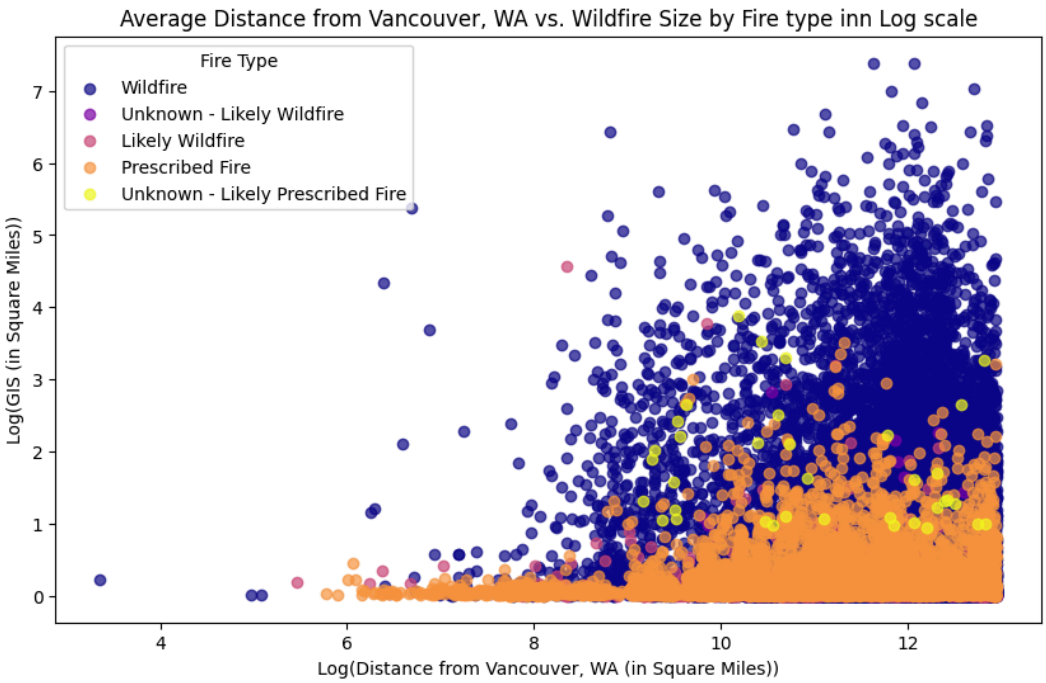

Figure 7: Average Distance from Vancouver, WA v.s Wildfire Size by Fire Type in Log scale

The chart in Figure 7 appears to be left-skewed and not normally distributed; which means that log transformation is not appropriate in this scenario. Alternative transformations might be more appropriate.

I was curious to check if it is the presence of outliers that is contributing to this skew in the log transformed dataset (i.e., I wanted to understand if all the fires "more in area" are being classified as outliers). To check that, I have a box plot below in Figure 8,

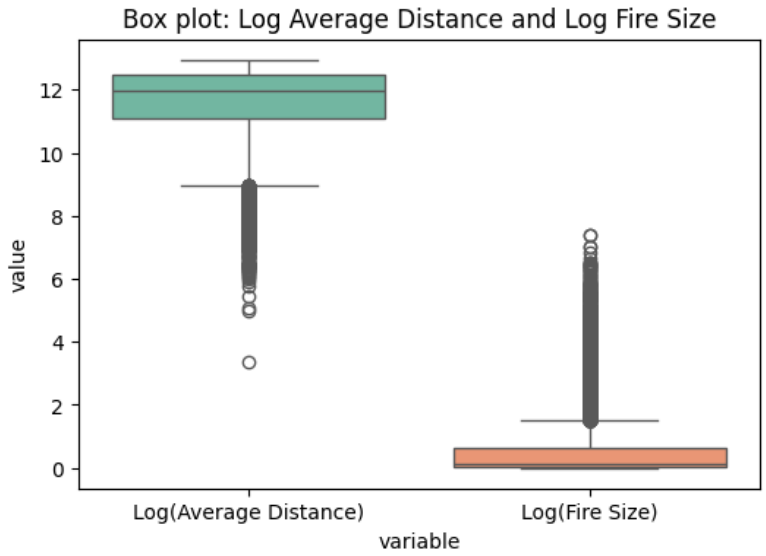

Figure 8: Box plot - log(Average Distance) and log(Fire Size)

There are not many large wildfires compared to the ones small in size; hence, mostly they fall in the outliers category. The data is tightly clustered, resulting in a flat trend in the linear scale.

A correlation of ~0 does not necessarily mean that the variables are independent. They might still have a nonlinear relationship that is not captured by the correlation coefficient. However, for simplicity, I assumed that **GIS_SqMiles and average_distance are independent of each other.**

Further, I wanted to understand the relationship between Fire Type and Average distance (Figure 9)

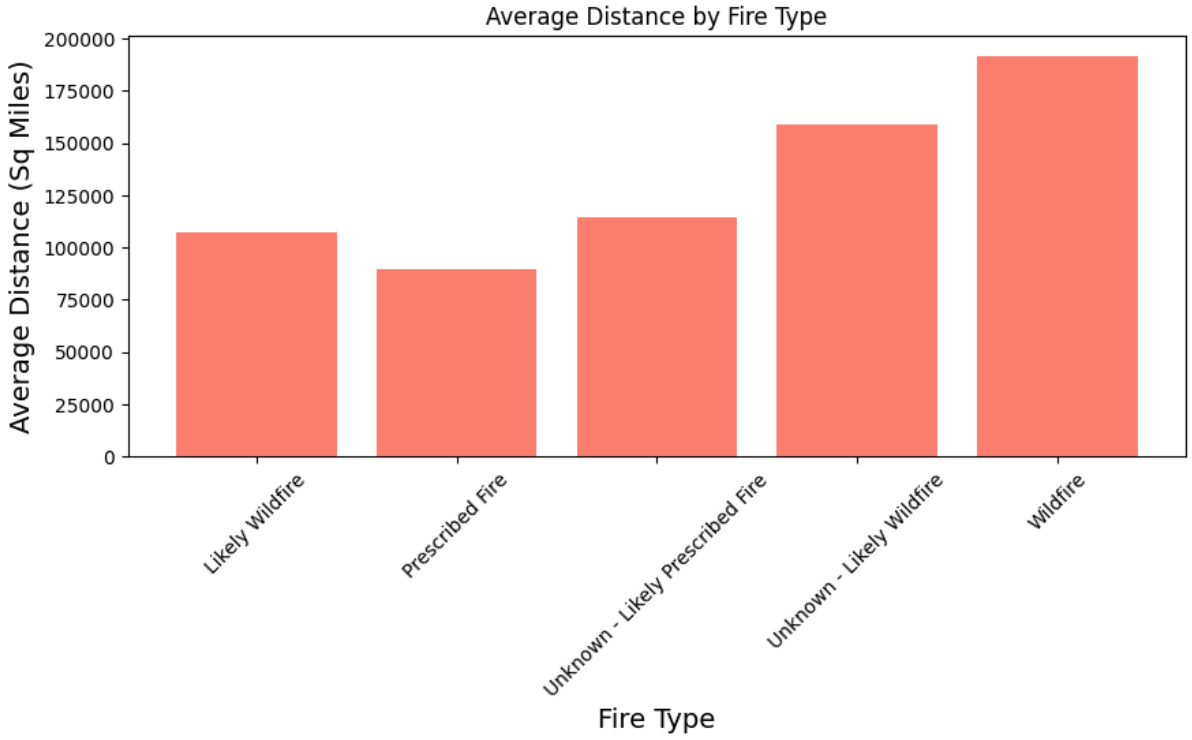

Figure 9: Average Distance by Fire Type

There seems to be a clear distinction between the types and the average distance. Intuitively, it made sense to give more importance to the wildfire and less importance to the Prescribed fire. I thus used this mapping to scale the smoke estimation impact: wildfires have a higher weight of 1.0, while prescribed fires, which are controlled, have a lower weight of 0.5. Other types, like “Likely Wildfire” and “Unknown - Likely Wildfire,” are given weights of 0.7 and 0.8, respectively, based on their potential intensity. "Unknown - Likely Prescribed Fire" is given an even smaller weight of 0.3.

To model the smoke estimate accurately, I need to explore other factors influencing the wildfire size / average distance. However, since I don't have any in our dataset, I am intuitively modeling in this case.

The formulated equation for smoke impact is:

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

From Figure 1, 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.

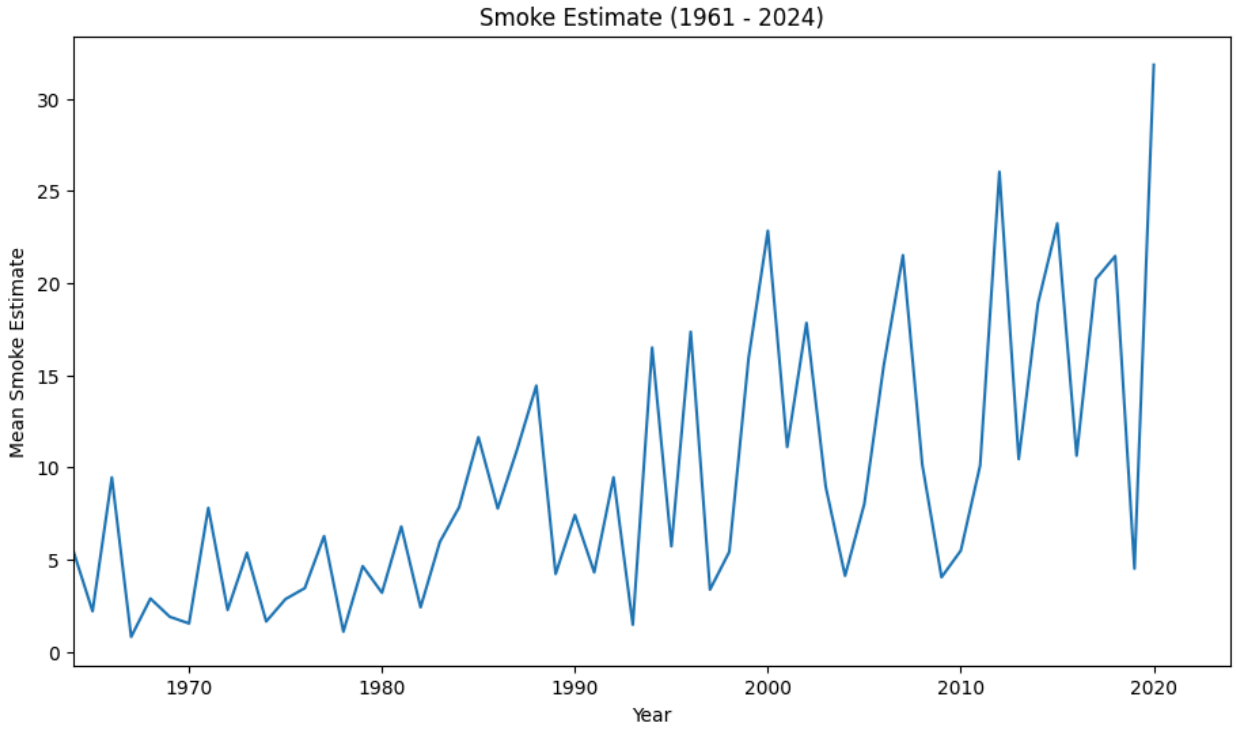

Figure 10: Mean smoke estimate from 1961 to 2021

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 11 compares the time series of Smoke Estimate and AQI Value for Vancouver, WA. The goal of this comparison is to understand the relationship between these two metrics and how they might influence each other or be influenced by external factors. The x-axis represents the year (ranging from 1990 - 2020; I only have particulate and gaseous data from 1990), and the y-axis represents the value of both smoke estimates and AQI. To interpret the figure, the viewer can examine the position of the data points and the lines connecting them. A higher data point indicates a higher value for either smoke estimate or AQI. 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.

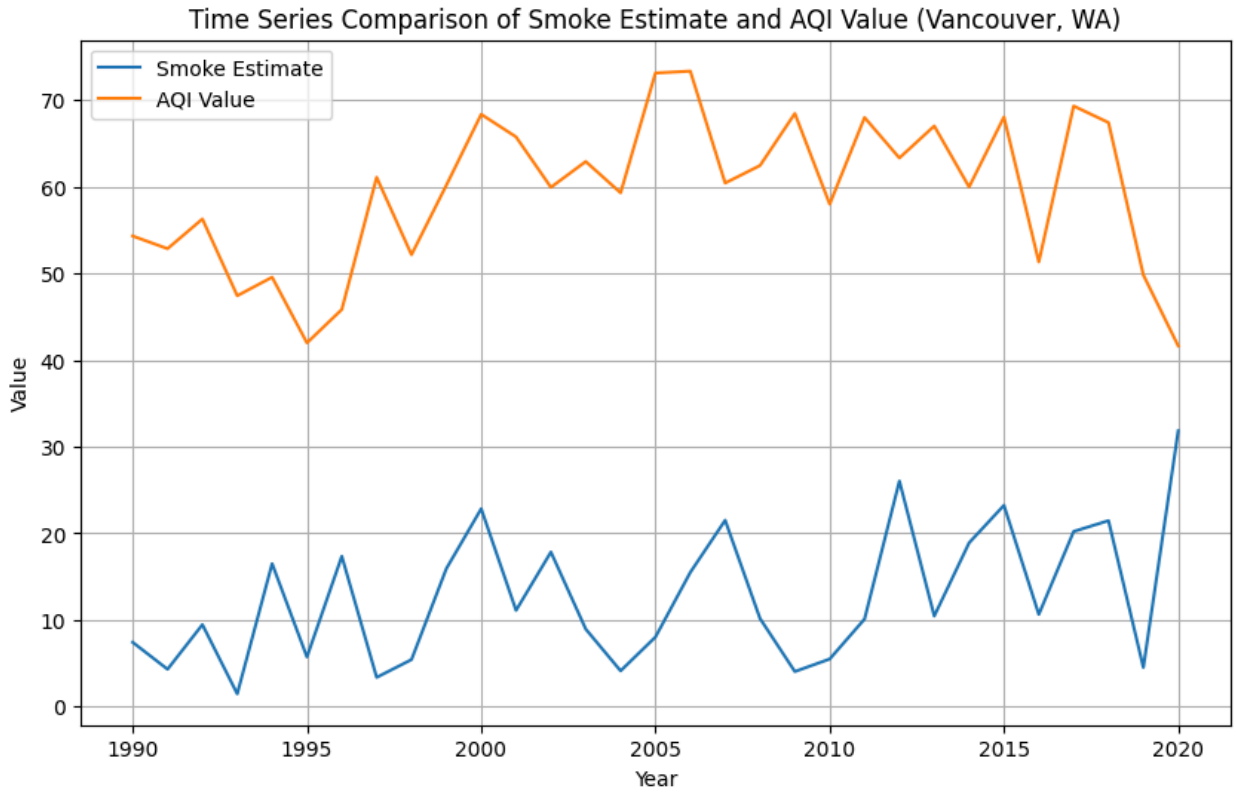

Figure 11: Time Series Comparison of Smoke Estimate and AQI value

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.

Time forecasting models are statistical and machine learning techniques used to predict future values based on historical time-series data. Statistical models like, ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal ARIMA), Exponential Smoothing (ETS), Machine Learning Models like, Linear/Polynomial Regression, Decision Trees, Boosting Models, SVMs, RNNs, LSTMs, Transformers, and Hybrid models such as ARIMA-LSTM are some example.

ARIMA (AutoRegressive Integrated Moving Average) is a robust model specifically designed for analyzing and forecasting time series data. It is capable of modeling trends (e.g., increasing smoke estimates over the years), particularly through its differencing parameter (d) which helps in achieving stationarity. They incorporate both autoregressive (past values) and moving average (past forecast errors) components, allowing it to capture complex relationships in the data that simpler models may overlook. ARIMA is grounded in solid statistical theory, providing a credible basis for forecasts. The ability to produce confidence intervals for predictions is a crucial feature of ARIMA. They can be fine-tuned by adjusting the parameters (p, d, q) to optimize performance according to the specific characteristics of our data and with readily available libraries (e.g., Statsmodels in Python), implementing ARIMA models is straightforward. In this project, I did not explore any other models in depth. I will be using ARIMA for wildfire smoke estimate prediction.

In time series forecasting, stationarity means that the statistical properties of the data, like mean and variance, do not change over time. Stationary data is often easier to model accurately because it's predictable and consistent in its patterns, while non-stationary data can be much harder to predict due to its trends or varying volatility.

Keeping things stationary makes the modeling task a lot easier, helps to improve our model accuracy and in return provides us with more reliable predictions. 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. Here I subtracted the previous data point from the current data point. The first difference can often eliminate trends and stabilize the mean of the series over time. I tried differentiating and then re-run the ADF test. Now I have the ADF statistic as -7.42. This value is significantly negative, far below zero. In general, more negative ADF statistics indicate stronger evidence against the null hypothesis of non-stationarity. A value this low suggests that the time series is likely stationary. Even the p-value is exceedingly low now (9.43e-11), much lower than even the significance level of 0.05. A p-value this small provides strong evidence to reject the null hypothesis of non-stationarity, indicating that the series does not exhibit a unit root and likely shows stationarity.

Hence, in this case, we can say that the differencing process successfully transformed the "smoke_estimate_scaled" series into a stationary series.

When we build an ARIMA model, we have to consider the p, d, and q terms that go into our ARIMA model.

- The first parameter, p, is the number of lagged observations. By considering p, we effectively determine how far back in time we go when trying to predict the current observation. I do this by looking at the autocorrelations of our time series, which are the correlations in our series at previous time lags.
- The second parameter, d, refers to the order of differencing. Differencing simply means finding the differences between consecutive timesteps. It is a way to make our data stationary, which means removing the trends and seasonality. d indicates differencing at which order you get a process stationary.
- The third parameter q refers to the order of the moving average (MA) part of the model. It represents the number of lagged forecast errors included in the model. Unlike a simple moving average, which smooths data, the moving average in ARIMA captures the relationship between an observation and the residual errors from a moving average model applied to lagged observations.

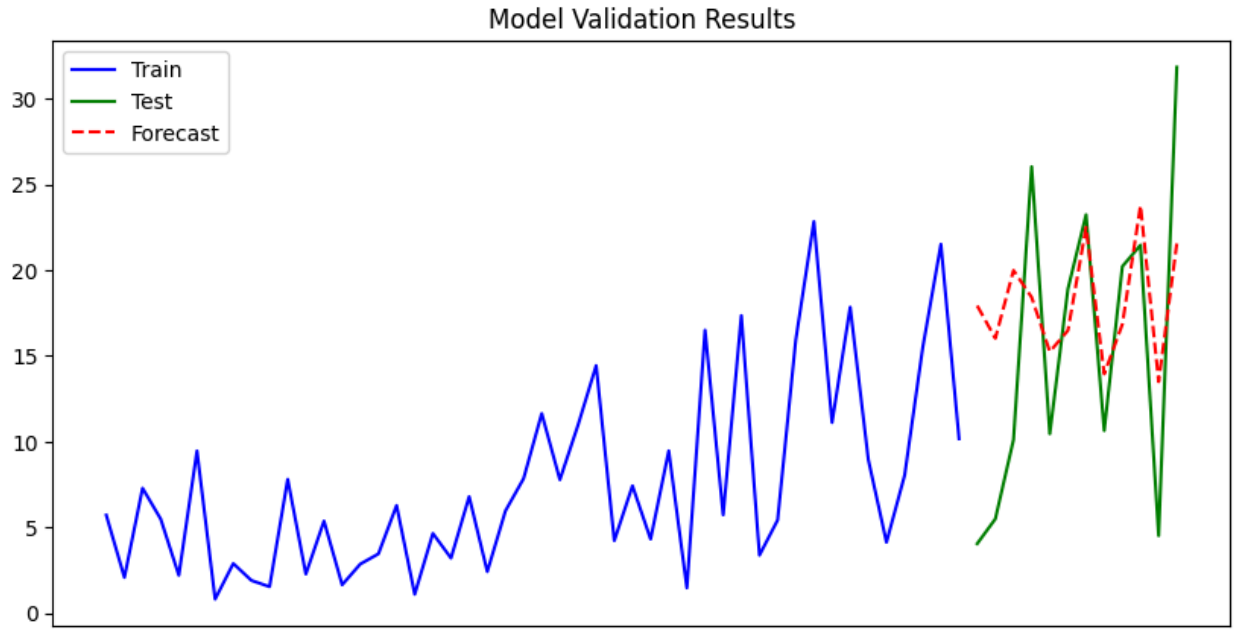

Figure 12: ARIMA model fit - Validation results

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):

- p = 6: Model includes six lagged observations of the dependent variable in predicting future values. The autoregressive component (AR) captures the relationship between the current observation and its previous four values, allowing the model to use more information from the recent past.

- d = 1: Model applies first differencing to the time series data to achieve stationarity. Differencing one time removes trends and seasonality from the data, effectively making the series stationary for better model performance.
- q = 3: Model incorporates three lagged forecast errors in its predictions. The moving average component (MA) helps to account for the influence of past error terms on the current observation.

To forecast using an ARIMA model, I started by using the fitted model to predict future values based on the data. Once predictions were made I visualized them by plotting the predicted values alongside the actual values (Figure 12). Doing this helped me see how well the model performs on unseen data.

The lowest average RMSE (Root Mean Square Error) obtained using this best order was ~7.46. When interpreted on this scale (0-30), the model's prediction errors are relatively low, given the narrower range of possible values. This indicates that while some errors occur, they remain manageable within this defined scale

Now that we have the model, I forecasted the smoke estimate for the next 25yrs. It is visualized in Figure 13 below.

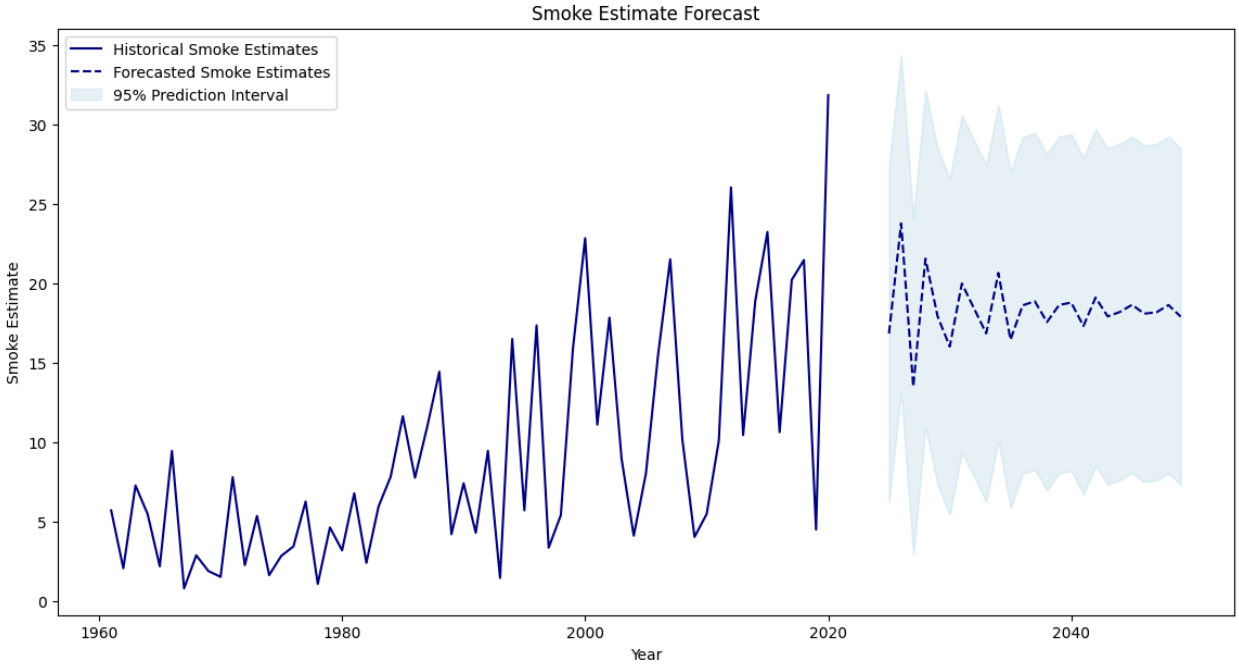

Figure 13: Smoke Estimate Forecast for the next 25 years (2025 - 2050)

The straight blue line above 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.

As discussed in the Introduction and Background sections, dense smoke generated from these wildfires carries harmful pollutants that pose significant risks to respiratory health, manifesting in increasing respiratory elevated rates of respiratory diseases, and premature deaths. The economic repercussions; such as rising unemployment and poverty rates - further exacerbate the challenges faced by vulnerable communities. Hence, I thought it would be interesting to see how public health and economic outcomes vary with smoke estimates.

Analyze Respiratory variation with smoke impact

For each year, we have data on multiple respiratory diseases. The data includes annual mortality rates for males, females, and the combined population, covering various respiratory diseases.

Figure 14 shows how the mortality rate is distributed across various diseases.

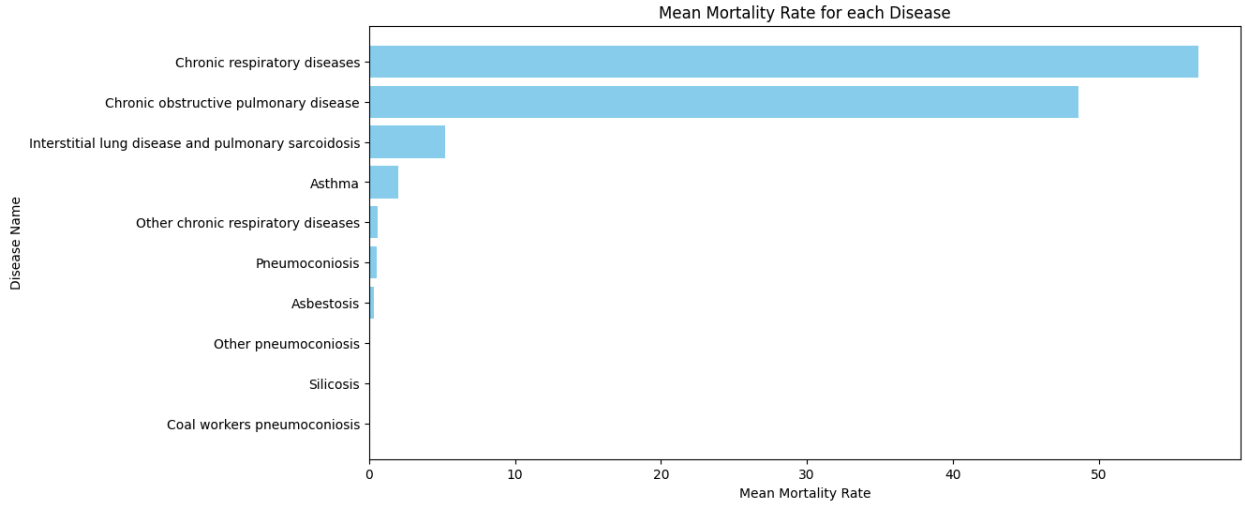

Figure 14: Mean Mortality Rate for each disease

It can be observed that the Chronic Respiratory diseases and Chronic Obstructive Pulmonary diseases on an average have 48.6% and 56.8% mortality rate.

Chronic Respiratory Diseases (CRDs), including Chronic Obstructive Pulmonary Disease (COPD), are closely linked to wildfire smoke due to its harmful pollutants like fine particulate matter (PM2.5), carbon monoxide, and volatile organic compounds. These substances

exacerbate symptoms of CRDs by triggering airway inflammation, reducing lung function, and causing acute respiratory distress. Prolonged exposure to wildfire smoke accelerates disease progression, increases hospitalizations, and raises mortality risk in individuals with CRDs. There are several studies (an example is [18]) showing significant spikes in emergency visits and respiratory-related deaths during wildfire events, highlighting the need for public health measures to protect vulnerable populations.

Figure 15 below how the diseases are distributed across sexes.

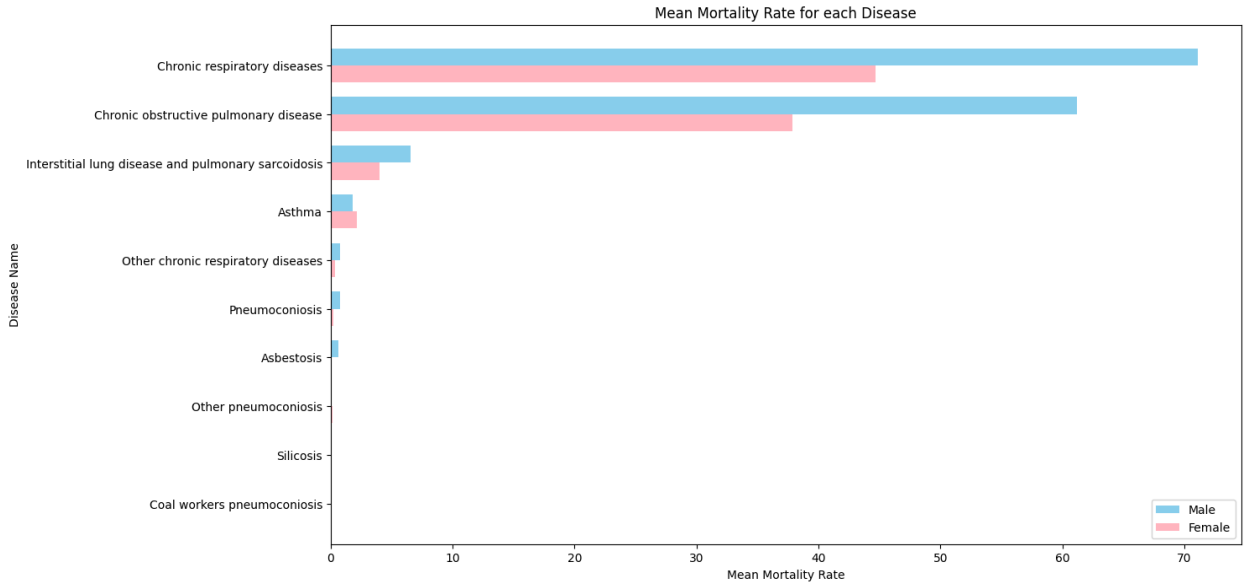

Figure 15: Mean Mortality Rate for each Disease

Except in case it is Asthma, we can infer that Males are mostly affected.

Because Chronic Respiratory Diseases are closely linked to wildfire smokes, I filtered the dataset for this cause. I also filtered to include both the sexes in the analysis. Figure 16 shows the relationship between Smoke estimate impact and Mortality Rate.

From the figure below, it can be observed that both the smoke estimate and mortality rate seem to exhibit increasing trends over the years. This suggests a potential link between increased air pollution and worsening respiratory health. The mortality rate exhibits more fluctuations compared to the smoke estimate - I do see the mortality rate rising in the early 2000s and gradually decreasing after the mid 2000s.

Increased exposure to particulate matter and other pollutants in smoke can directly irritate the respiratory system, leading to chronic respiratory diseases like asthma, chronic obstructive pulmonary disease (COPD), and lung cancer. Smoke can also exacerbate existing respiratory conditions, leading to increased hospitalizations and mortality.

To strengthen the causal link between smoke exposure and chronic respiratory disease mortality, I computed the spearman correlation. It came out to be 0.318, suggesting a weak to

moderate positive monotonic relationship. This means that as smoke estimates increase, there is a tendency for respiratory mortality rates to increase as well, but the relationship is not very strong.

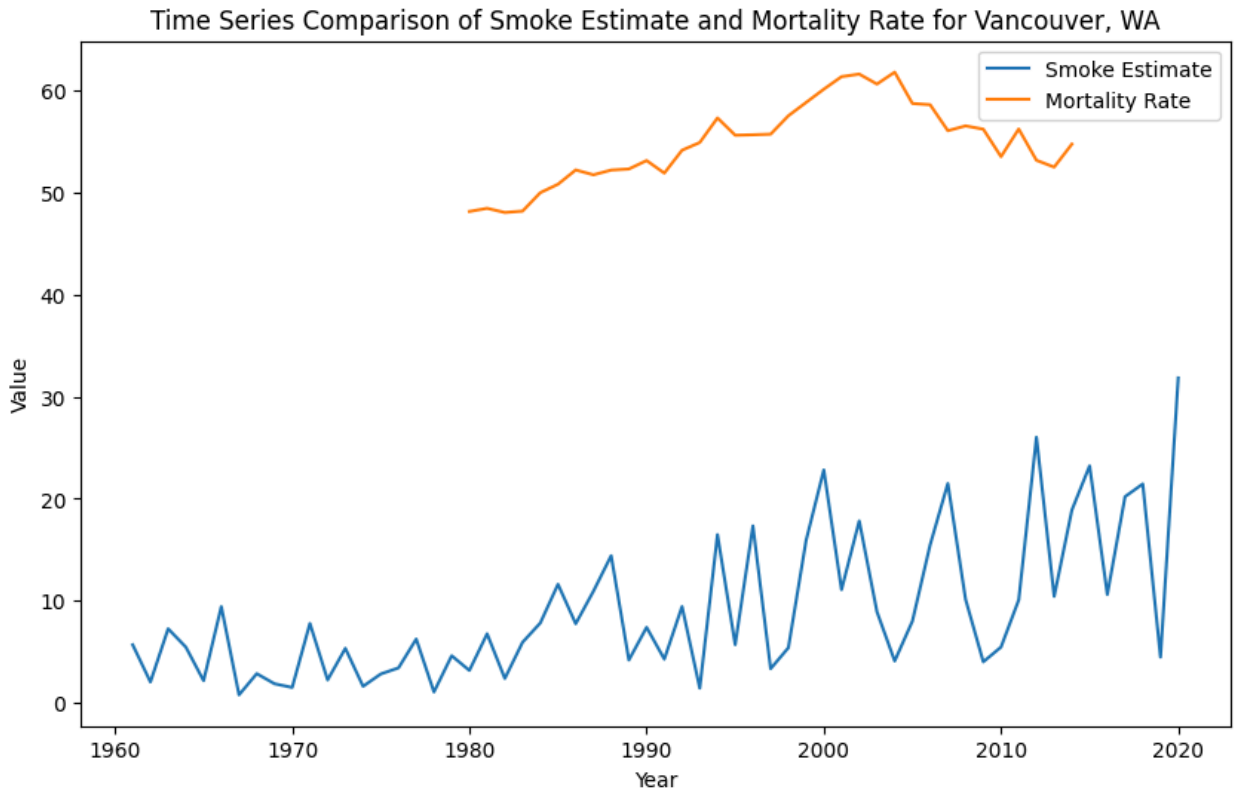

Figure 16: Time Series Comparison of Smoke Estimate and Mortality Rate for Vancouver, WA

The p-value of 0.063 is slightly above the typical significance threshold of 0.05, indicating that this correlation is not statistically significant at the 5% level. However, it is close enough to suggest a potential trend, and with more data, the relationship might reach statistical significance.

To better understand its potential long-term impact, I utilized the same time series modeling approach to project the trajectory of smoke estimates and their influence on respiratory mortality over the next 25 years in the next step (shown in Figure 17).

There's a general upward trend in respiratory mortality rates, suggesting a potential increase in respiratory health issues over time. The historical data (straight red line) also exhibits fluctuations, indicating variations in mortality rates from year to year, likely influenced by factors such as air quality, socioeconomic conditions, and healthcare access. The forecast (dotted red line) predicts a continued upward trend in respiratory mortality rates, with a widening prediction interval towards the end of the forecast period. The wider 95% prediction interval (shaded red area) indicates increased uncertainty in the future projections, suggesting that multiple factors

could influence the actual trajectory of respiratory mortality rates.trend in respiratory mortality rates, with a widening prediction interval towards the end of the forecast period.

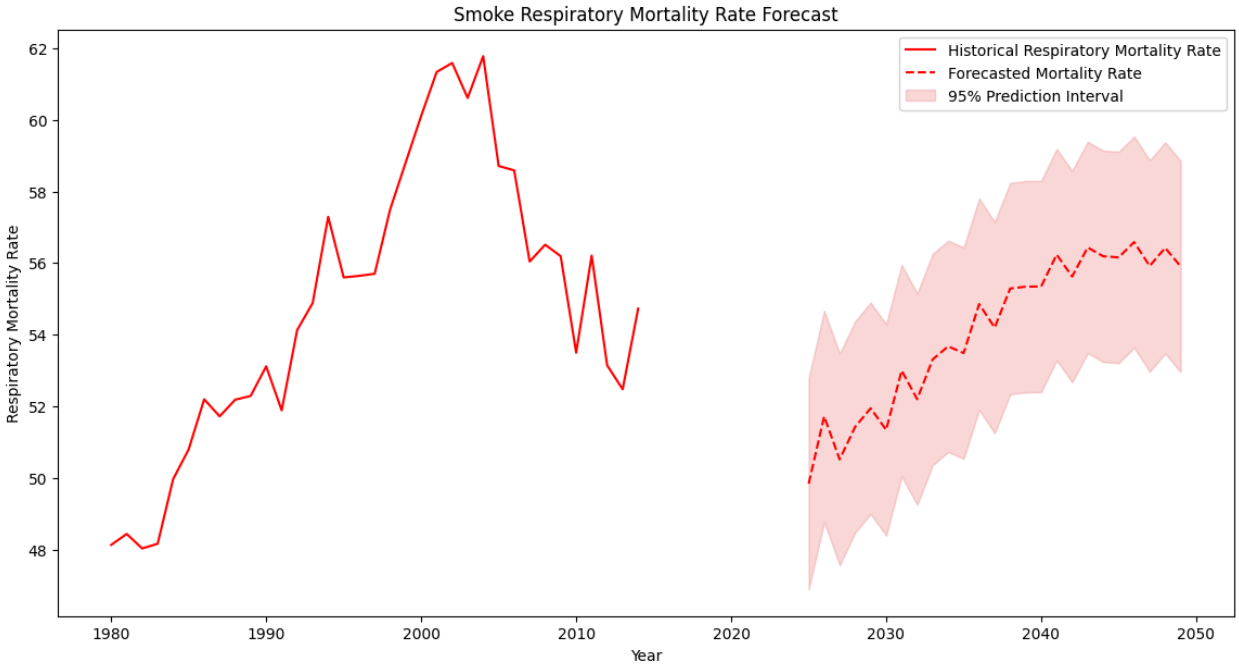

Figure 17: Respiratory Mortality Rate Forecast

To improve the accuracy of the forecast, it's crucial to consider other factors and incorporate more detailed data and advanced modeling techniques. Additionally, ongoing monitoring and evaluation of public health interventions are essential to mitigate the impact of air pollution and other factors on respiratory health.

Analyze socio-economic Unemployment Rate with Smoke Impact

In the chart below (Figure 18), we can see if the unemployment rate varies with changing smoke estimates.

It can be seen that both the variables (smoke estimate and unemployment) exhibit significant fluctuations over time, reflecting dynamic changes in environmental and economic conditions. There are periods where increases or decreases in smoke estimates appear to align with similar trends in the unemployment rate, suggesting a potential correlation. However, in some instances, higher smoke estimates seem to precede increases in unemployment, hinting at a possible delayed impact.

To better understand the significance of the relationship, I computed the Spearman correlation. The coefficient of 0.045 and the p-value of 0.809 suggest that there is a very weak positive relationship between the Smoke Estimate and the Unemployment Rate in Vancouver, WA, but this relationship is not statistically significant. The high p-value indicates that any observed correlation is likely due to random chance rather than a meaningful connection. This finding

suggests that fluctuations in smoke estimates and unemployment rates may not be directly related.

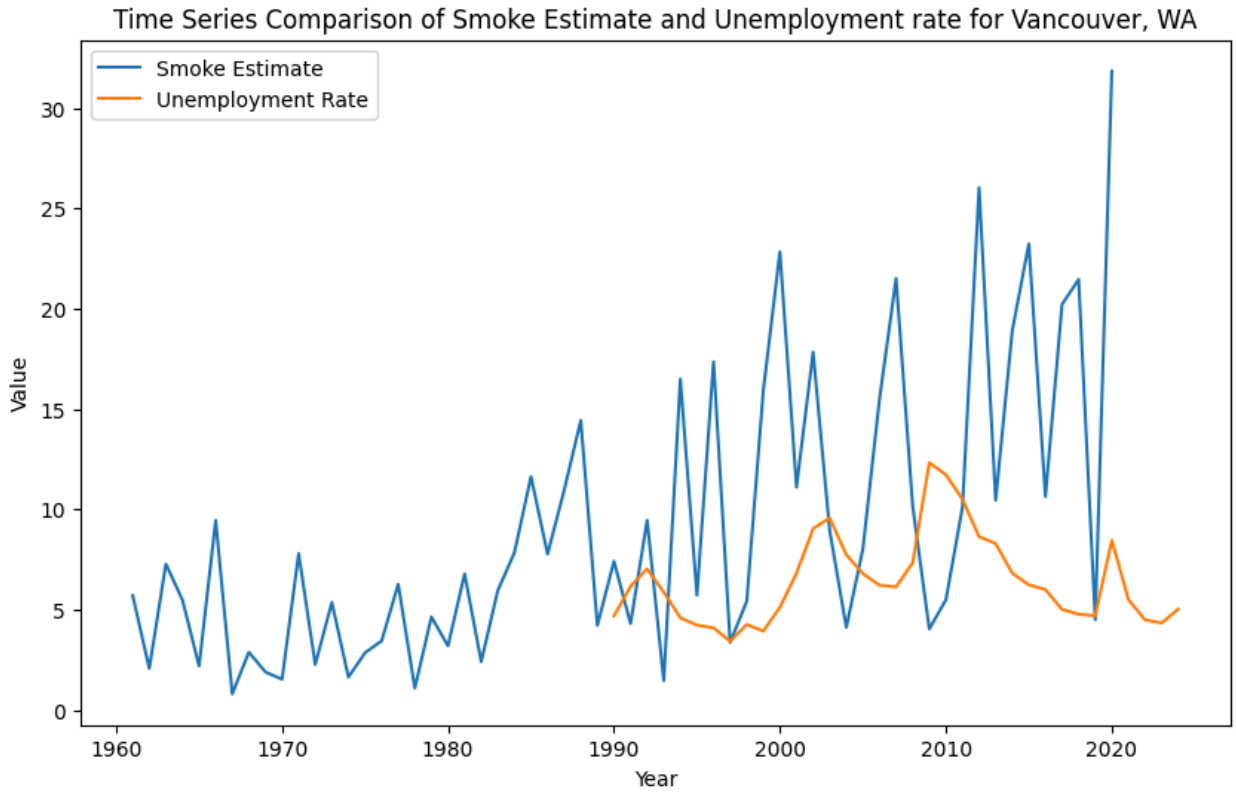

Figure 18: Time Series Comparison of Smoke Estimate and Unemployment rate for Vancouver

Further analysis could explore whether external factors, such as economic policies or health-related disruptions, might influence both variables indirectly. In this case, I did not proceed with a time series modeling approach to project the trajectory of smoke estimates and their influence on unemployment over the next 25 years.

Analyze socio-economic Poverty Rate with Smoke Impact

In the chart below (Figure 19), we can see how the poverty rate varies with changing smoke estimates.

It is interesting to observe that the plot shows a complex relationship between smoke estimates and poverty rates in Vancouver, WA. While there isn't a clear, consistent correlation, we can observe some interesting patterns! Both smoke estimates and poverty rates exhibit significant fluctuations over time. In some periods, I see a positive correlation, where both smoke estimates and poverty rates increase or decrease together. This could be due to factors like economic downturns or severe wildfire seasons that impact both economic conditions and smoke impact. However, in all other instances, increases in poverty rates seem to follow periods of high smoke estimates. This might suggest that the economic and health impacts of wildfires and air pollution can have delayed effects on poverty levels.

To gain a deeper understanding of the relationship between smoke and poverty, I computed the Spearman coefficient.

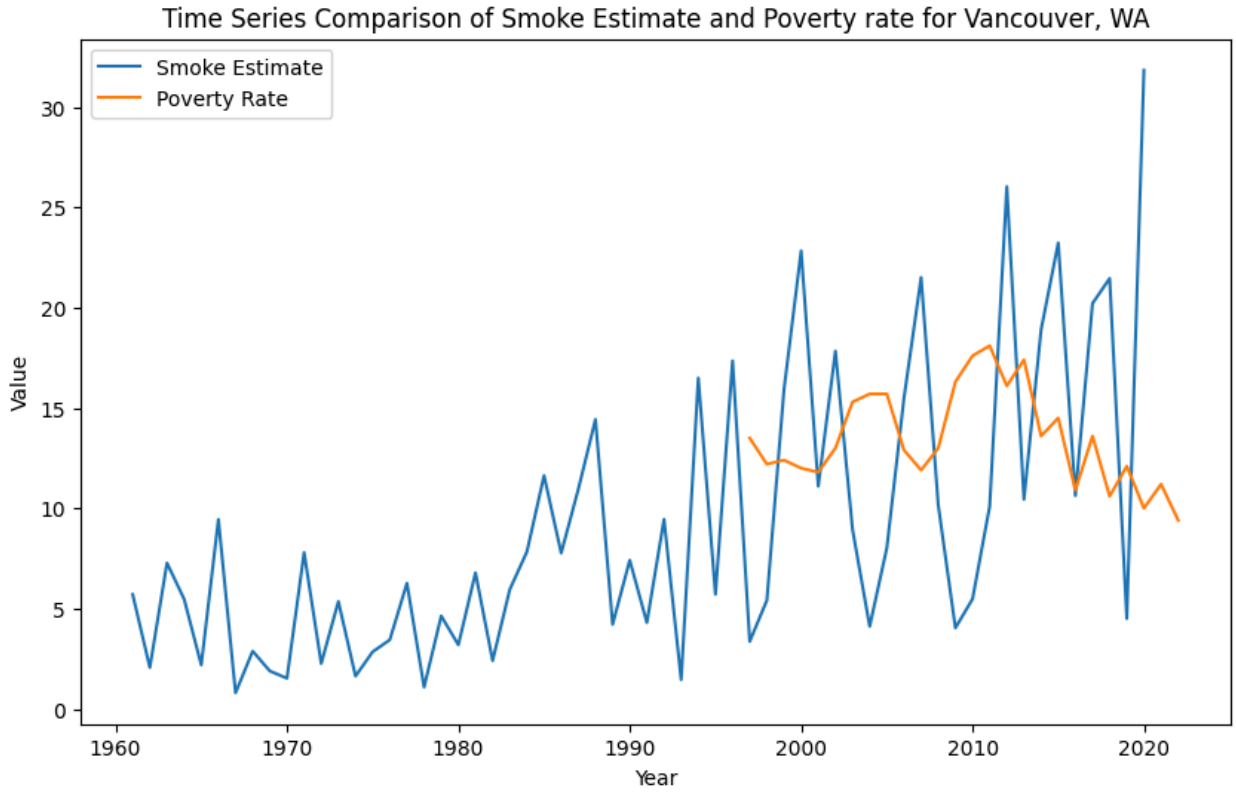

Figure 19: Time Series Comparison of Smoke Estimate and Poverty rate for Vancouver, WA

The Spearman correlation of -0.411 and the p-value of 0.033 suggest a moderate negative relationship between the Smoke Estimate and Poverty Rate, which means that areas with higher poverty rates tend to have lower smoke estimates. This finding might seem counterintuitive at first, however, it can be observed that increases in poverty rates seem to follow periods of high smoke estimates.

The delay in the observed increase in poverty after high smoke estimates could be due to a lag in both physical and economic recovery. The immediate health impacts might not show up right away, and the economic disruptions caused by wildfires (such as job losses and business closures) may take some time to manifest fully. Once these effects accumulate, they can lead to long-term increases in poverty, especially for communities already vulnerable due to their low-income status.

To better understand its potential long-term impact, I utilized a time series modeling approach to project the trajectory of smoke estimates and their influence on poverty over the next 25 years in the next step. It can be visualized in Figure 20.

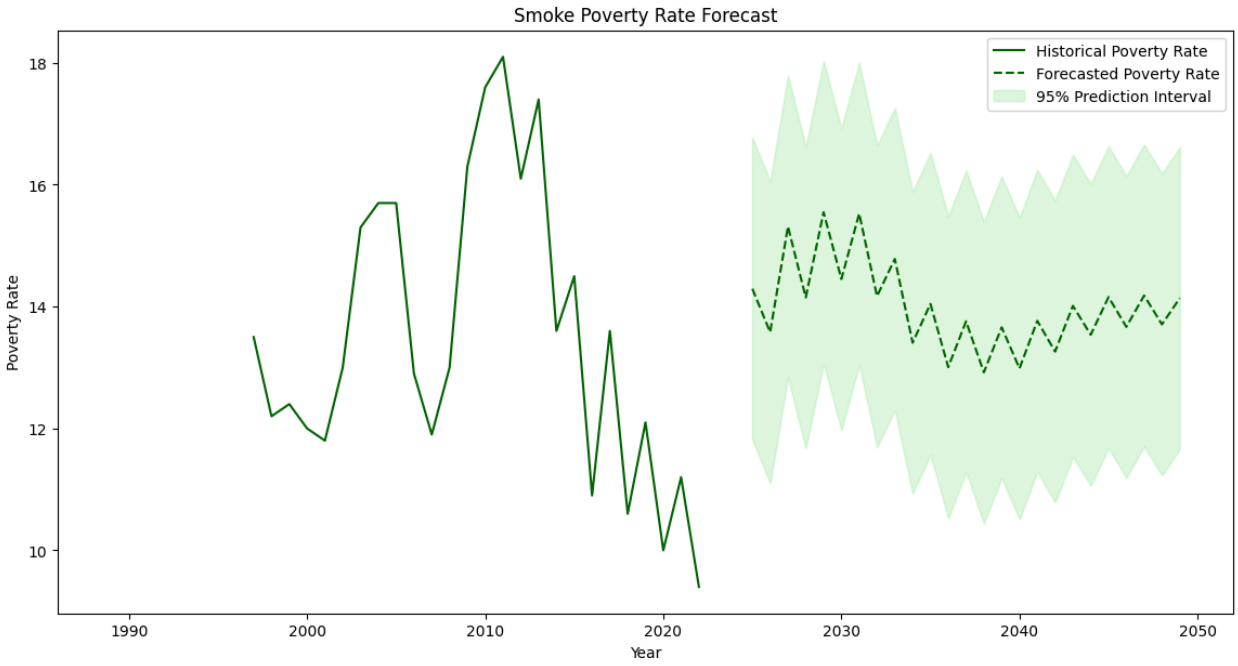

Figure 20: Poverty Rate Forecast

The historical poverty rate (straight green line) exhibits significant fluctuations, suggesting that various factors influence poverty levels. These factors could include economic cycles, social policies, and demographic changes. The forecast (dotted green line) suggests a continued fluctuating trend in poverty rates, with periods of increase and decrease. This indicates that poverty is likely to remain a complex issue with no simple solution. The wide prediction interval (shaded green area) highlights the uncertainty associated with future poverty rates (95% prediction interval). This uncertainty is likely due to the numerous factors that can influence poverty levels and the difficulty in accurately predicting future economic and social conditions.

While the forecast suggests that poverty will remain a persistent issue, it's important to note that effective policies and interventions can help mitigate its impact.

Analyze socio-economic Premature Deaths with Smoke Impact

Figure 21 shows how premature deaths vary with changing smoke estimates.

It can be observed that, over the years, there is an overall increasing trend in smoke estimates, which could be linked to factors such as heightened wildfire activity, evolving air quality monitoring techniques, or population growth.
In contrast, the premature death rate per 100,000 people fluctuates, influenced by a variety of factors including demographic shifts, advancements in healthcare, and socioeconomic conditions. While there may not be a strong, direct correlation between the two time series, it is

plausible that periods of higher smoke estimates could align with increases in premature death rates, especially for vulnerable populations.

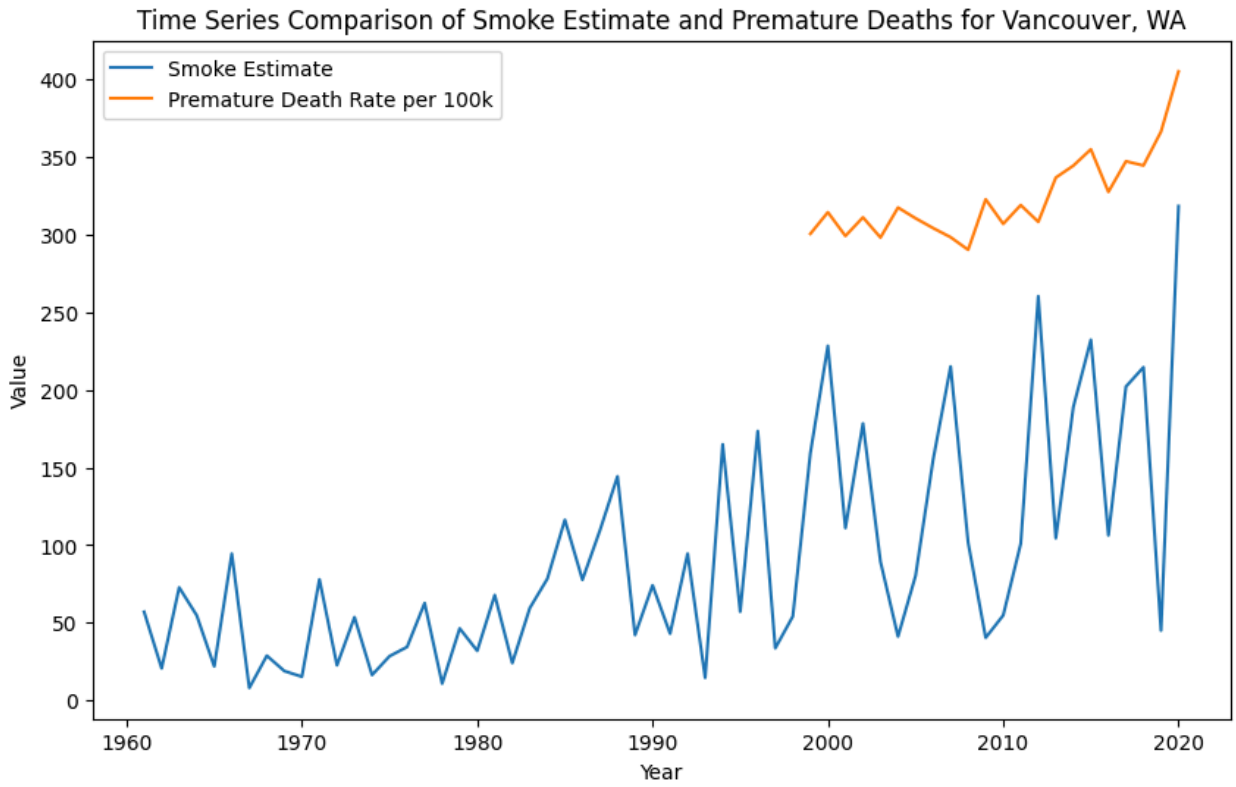

Figure 21: Time Series Comparison of Smoke Estimate and Premature Deaths for Vancouver

As in the other sections, here as well I computed the Spearman Correlation Coefficient. The Spearman correlation of 0.193 between the smoke estimate and premature deaths per 100k suggests a weak positive relationship, meaning that as smoke estimates increase, there is a slight increase in premature deaths, but the relationship is not strong. The p-value of 0.391 is much higher than the typical significance threshold of 0.05, indicating that this correlation is not statistically significant. This suggests that while there may be some association between the two variables, it is likely due to chance, and there is no clear evidence to support a strong or meaningful link between higher smoke estimates and an increase in premature death rates in this case.

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 below in Figure 22.

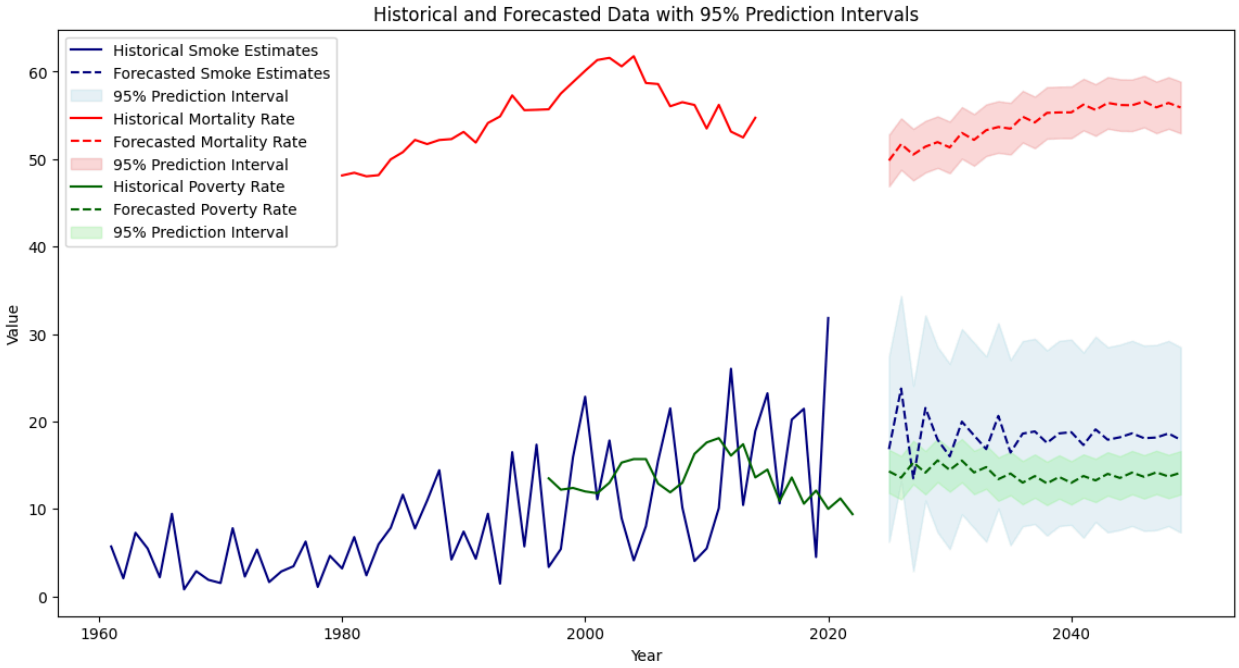

Figure 22: Historical and Forecasted Data with 95% Prediction Intervals for Smoke Estimate, Mortality, and Poverty Rate.

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.

Key Observations:

- **Increasing Smoke Estimates:** The historical smoke estimates show a clear upward trend, indicating a potential increase in wildfire activity or other factors contributing to air pollution.
- **Rising Mortality Rates:** The historical mortality rates also exhibit an upward trend, which could be partially attributed to the increasing exposure to air pollution.
- **Fluctuating Poverty Rates:** The poverty rate shows fluctuations over time, but there is a general upward trend, particularly in recent years.

Potential Relationships:

- **Smoke Impact and Mortality:** Increased exposure to air pollution, particularly from wildfires, can lead to respiratory problems, cardiovascular diseases, and other health issues, which can contribute to increased mortality rates.
- **Smoke Impact and Poverty:** Wildfires can have significant economic impacts, such as job losses, business closures, and increased healthcare costs. These economic disruptions can exacerbate poverty and inequality.