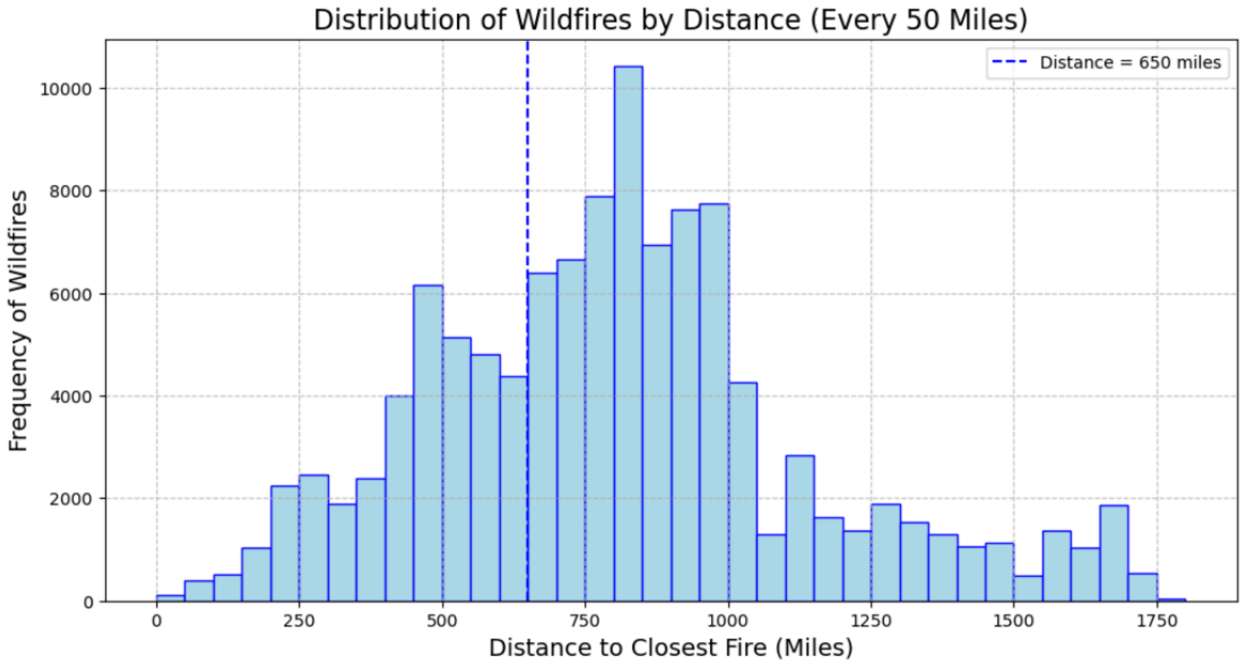
DATA 512: Part 1 - Common Analysis

In this assignment, I focused on building a predictive model to forecast wildfire smoke impacts and explored how accurate simplified smoke impact estimates could be for representing real effects on urban air quality. The questions posed in the project requirement document encouraged a deep analysis of time series modeling, data assumptions, and variable selection, and further improved my understanding of predictive modeling.

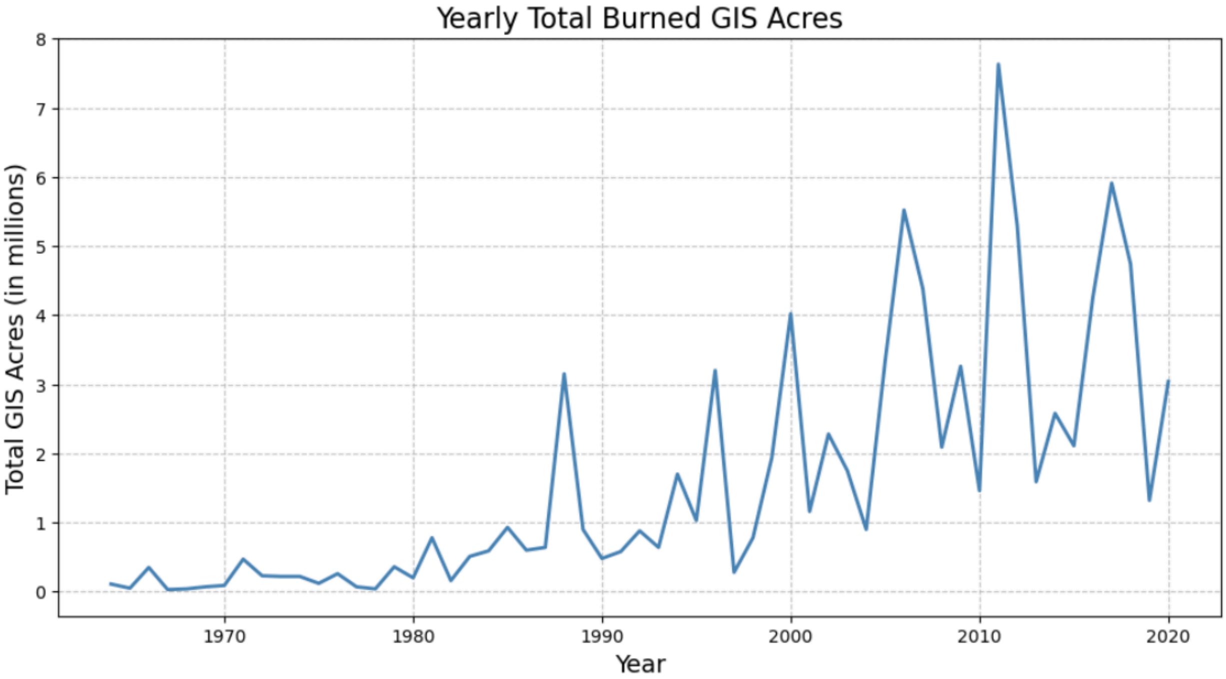
Histogram showing the number of fires occurring every 50 mile distance from your assigned city for all fires ranging up to 1800 miles away from Centennial, CO.



The count of fires that is visualized in this dataset was derived from the Combined Wildland Fire Datasets for the United States and Certain Territories, provided by the U.S. Geological Survey. This dataset contains detailed records of wildfires, including the geographical coordinates outlining the perimeter of each fire polygon. These coordinates were essential in calculating the shortest distance from each fire-affected area to Centennial, enabling a more precise understanding of proximity and potential impact on the city.

This histogram illustrates the distribution of wildfires based on their distance from a reference point, with each bar representing a 50-mile range. Here, the x-axis represents the Distance to Closest Fire (in miles) and the y-axis represents the frequency of wildfires. The distribution is unimodal and approximately symmetric, with a clear peak around 750 miles. This suggests that most wildfires tend to occur within a specific distance range, notably between 500 and 1000 miles from the reference point. While the majority of data points fall within the central range, a few wildfires occur at farther distances, extending up to around 1750 miles, though these instances are less common.

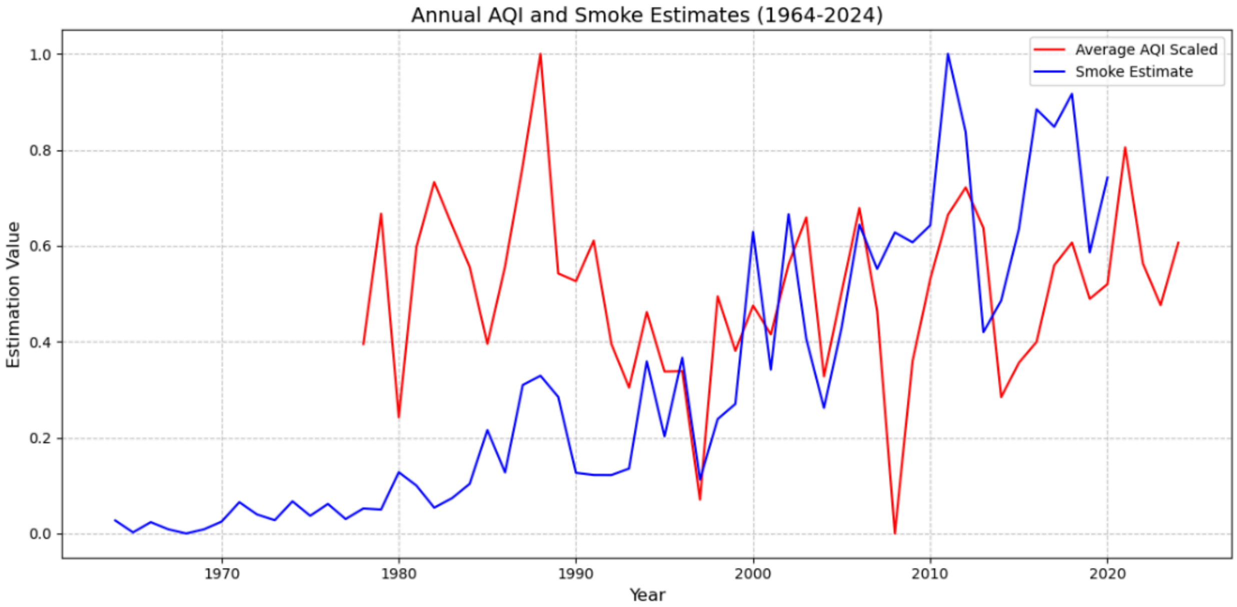
Time series graph of total acres burned per year for the fires occurring within 650 miles from Centennial, CO



The data used in the above line chart is very similar to the data used in the previous analysis. A few filters have been applied - only the wildfires within a 650 mile radius of Centennial were considered and some scaling has been done on the y-axis.

This line plot shows the yearly total burned GIS acres (in millions) from the early 1960s to the 2020s. This is a time series graph where the x-axis represents the year and the y-axis represents the total GIS acres (in millions). The trend indicates a relatively low and stable number of burned acres from the 1960s to the 1990s, with only minor fluctuations. However, starting around the year 2000, there is a noticeable increase in burned acres, with significant spikes in certain years. The most extreme peak occurs around 2010, reaching nearly 80 million acres. This increase in burned acreage over recent decades could reflect factors such as rising temperatures, drier conditions, or other environmental changes, highlighting the intensifying impact of wildfires in recent years.

Time series graph containing your fire smoke estimates and the AQI estimates for Centennial, CO



The data used to generate the line chart originates from two sources. The average AQI is derived by calculating the annual average of AQI data collected from monitoring stations in Centennial via the AQI API. The smoke estimate was developed using my own formula, incorporating wildfire data from the previous source in Centennial. This line chart shows the trends of two different variables—Average AQI (Air Quality Index) and Smoke Estimates—over time, from 1964 to 2024. Both variables are normalized on a scale from 0 to 1 to allow comparison. The x-axis represents the years from 1964 to 2020, while the y-axis shows a scaled value of the AQI and smoke estimates.

1. *Average AQI Scaled (red line)*: This line generally increases over time, showing fluctuations in air quality with significant peaks around the mid-1980s, early 2000s, and early 2010s. Periods of noticeable decline are also present, indicating temporary improvements in air quality.

2. *Smoke Estimate (blue line)*: This estimate shows a steady increase, especially after the 1980s, with notable peaks in the early 2000s and after 2010, corresponding with wildfire activity and smoke events. The line shows high variability, especially in recent years, suggesting an increase in smoke-related events.

The two lines occasionally overlap, suggesting a correlation between smoke events and higher AQI values, although not consistently. Overall, the chart indicates a general upward trend in both AQI and smoke estimates, reflecting worsening air quality and an increase in smoke events over the decades.

Reflection Statement: Learnings and Challenges

This project offered a valuable learning experience in developing a predictive model to forecast wildfire smoke impacts. Building this model required me to navigate challenges associated with time series modeling and integrating a range of variables to improve prediction accuracy. My initial approach was to create a smoke impact estimate based on historical fire and smoke data, aiming to capture potential future trends up to 2050. However, this exercise revealed the complexities and limitations of using simplified metrics and static models to predict such a dynamic phenomenon.

In exploring the ARIMAX model, I realized the importance of incorporating variables like wind speed, temperature, and humidity, as these added essential context that improved the model's accuracy. Initially, I hesitated to include multiple variables, concerned about overfitting or unnecessary complexity. Yet, testing showed that these exogenous factors provided significant predictive value. This insight helped me understand that adding relevant data, even if it makes the model more complex, can make predictions more meaningful, especially for phenomena as multifaceted as wildfire impacts.

However, despite refining the model, I became increasingly skeptical about its ability to forecast long-term trends reliably. The model's predictions suggested a simple upward trend without capturing seasonal fluctuations or variations, raising questions about its applicability over a 25-year horizon. The historical data spanned roughly 60 years, but forecasting further out introduces significant uncertainty. I began to see that wildfire impacts could be influenced by evolving climate and environmental factors, which a single, static model may struggle to account for. This highlighted the inherent limitations of long-term forecasting in such contexts and the importance of quantifying and openly acknowledging prediction uncertainties.

Furthermore, my approach to creating a simplified smoke impact estimate raised critical questions about whether such metrics truly reflect the complex realities of wildfire smoke effects. While this composite estimate provided a workable foundation for prediction, it lacked the nuances that come from more granular data. A more comprehensive analysis would ideally consider factors like terrain, local wind patterns, rainfall, and specific demographic vulnerabilities to smoke exposure. I came to recognize that while simplified estimates are useful for initial modeling, they may fall short of accurately representing the diverse and serious impacts that wildfire smoke can have on communities.

Overall, this project underscored both the potential and the limitations of predictive modeling in capturing real-world complexities. While data-driven forecasts can provide valuable insights, simplifying assumptions and static models may not fully capture the evolving dynamics of environmental impacts, especially in the face of climate change. This experience highlighted the need for continual refinement and caution when interpreting long-term forecasts, as well as the importance of integrating detailed, contextual data wherever possible.

Reflection Statement: Contributions of Collaboration to My Learning and Thinking

Collaboration played an important role in shaping my approach and overcoming challenges throughout this project. I took Manasa Shivappa's suggestion to address missing AQI data through imputation rather than simple omission. Initially, I was inclined to discard incomplete data, as I believed it would simplify the dataset and avoid potential distortions in model accuracy. However, Manasa pointed out that roughly 30% of the AQI data was missing—a substantial gap that, if left unaddressed, would compromise both the dataset's coverage and the model's reliability. Her suggestion to impute these values encouraged me to reconsider my approach, as I began to understand that a more complete dataset, even if imputed, could provide a more accurate foundation for analysis.

Similarly, working with Sushma on the daily AQI calculations helped deepen my understanding of the impacts that data choices have on model outcomes. I initially leaned towards using the average daily AQI, assuming it would provide a balanced representation of air quality trends without overemphasizing extremes. However, Sushma Vankayala and I engaged in discussions on the pros and cons of this approach, and through our research, we realized that capturing maximum daily AQI values might better represent the intensity of smoke events. High AQI spikes are often more significant for public health and could be key indicators of severe smoke impact periods that an average might dilute. By collaboratively brainstorming the implications of each approach, we concluded that maximum values could serve our research question more effectively. This conversation illustrated the value of looking beyond initial assumptions and reinforced how data preprocessing choices can profoundly influence the final model's relevance and accuracy, especially in health-related studies.

Additionally, I had a discussion with Abhinav Duvvuri which helped me make critical decisions around model selection. At one point, I was unsure whether to pursue a time series model, like ARIMAX, or to use a more conventional multiple regression approach. We both shared our insights on the strengths and limitations of each approach, helping me appreciate how temporal dependencies could be leveraged in a time series model to improve predictive accuracy. We discussed the importance of capturing seasonal trends and interdependencies, which ultimately led me to choose a time series approach.

Working with peers allowed me to test ideas, receive constructive feedback, and apply alternative approaches I hadn't considered on my own. My code incorporates insights from discussions with my peers as appropriate; however, it is entirely my own original work.