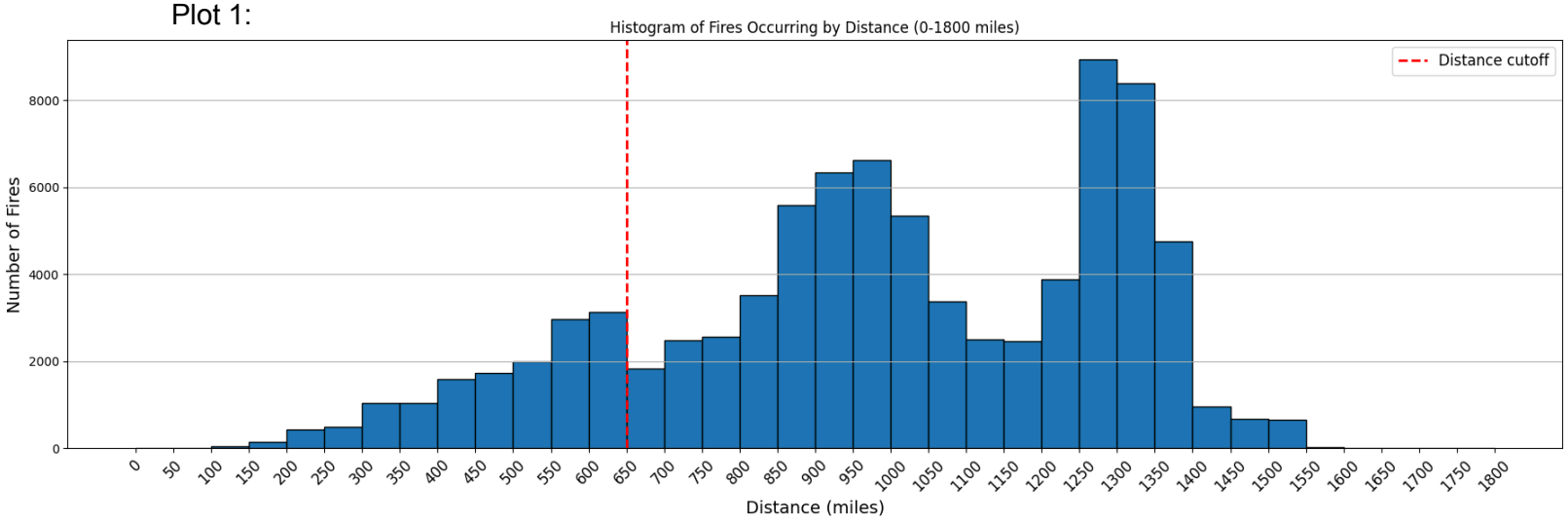
Elliott Sanger
Data 512
Final Part 1 Reflection
https://github.com/Elliott-1/Elliott-512-Final

Visualizations

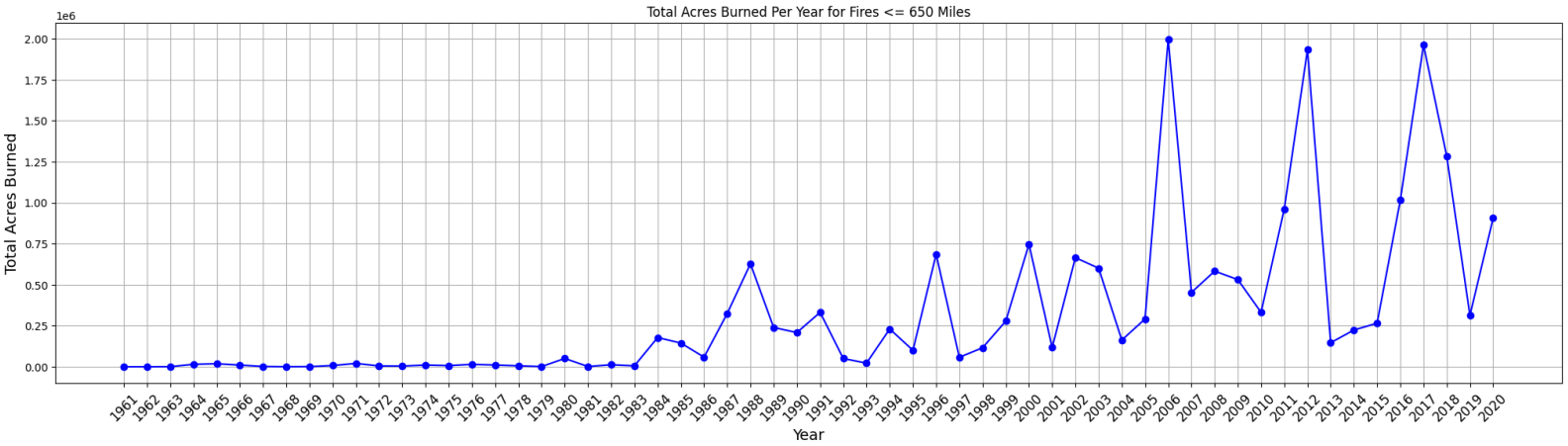
Plot 1:



The figure shows how the number of wildfires vary by distance, from Sioux Falls, my chosen reference city. The x-axis ranges from 0 to 1800 miles with each interval covering 50 miles. The y-axis is the number of fires. There's also a vertical red dashed line which indicates the 650 mile cutoff we're using for our smoke estimate. Sioux Falls is centrally located within the United States in a relatively remote area with a high wildfire count. As expected, the number of fires generally increases with distance due to the expanding area covered. This trend isn't consistent throughout, but there are notable spikes in fire counts between 850-1000 miles and 1200-1400 miles away, possibly indicating specific areas prone to wildfires or with higher record densities.

Within the 650-mile radius used for smoke estimates, the pattern reflects increasing wildfire counts as the distance from the city grows, beginning with few fires within 0-50 miles (an urban area with limited forests) and increasing steadily up to 650 miles. The accuracy of the data may be affected by the rural location and smaller population of Sioux Falls, potentially leading to underreporting in some areas.

Plot 2:

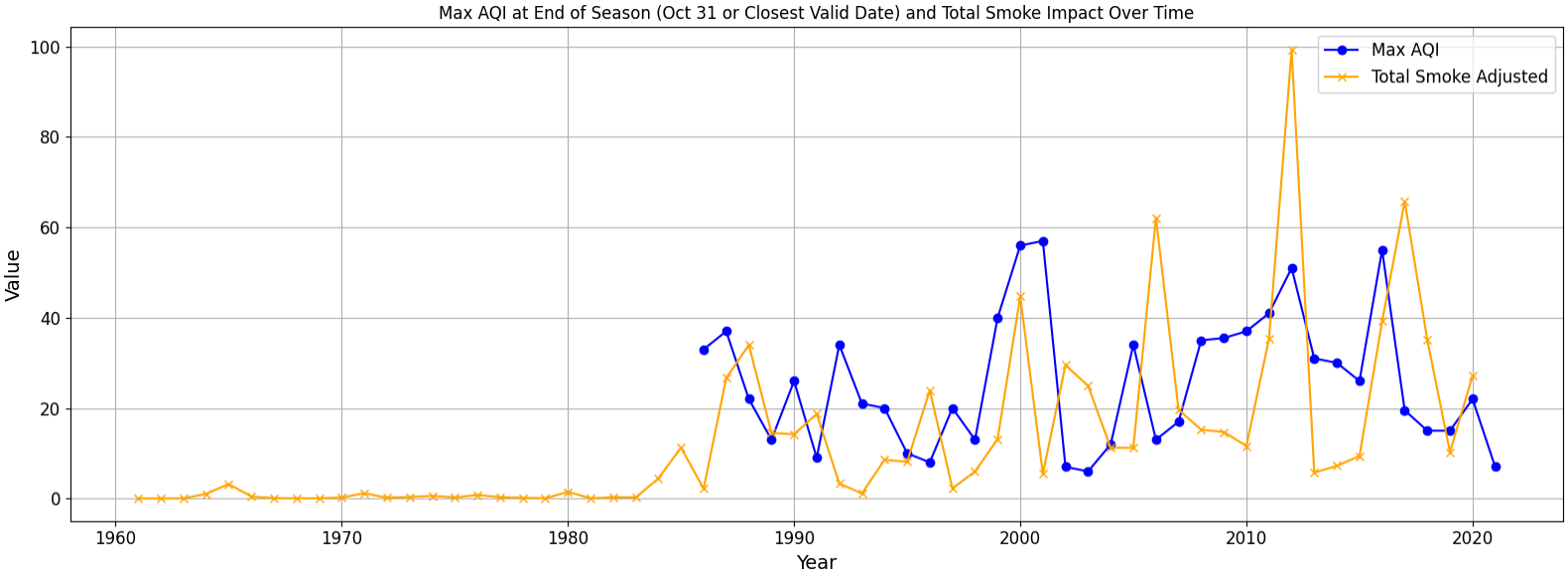


The second plot shows the total acreage burned within 650 miles of Sioux Falls from year to year. The x-axis represents the years from 1961 to 2020, and the y-axis displays the total acres burned in millions. The data reveals minimal acreage burned between 1961 and 1983, likely due to limited recording methods. Satellite technology, introduced in the 1980s, likely improved reporting accuracy.

The data also shows annual variations, suggesting that the acreage burned one year may impact subsequent years. More substantial fluctuations are evident in recent years, particularly since the 2000s, with periods of extensive burning followed by much lower levels. This may reflect changes in public attitudes towards fire prevention, allowing burnable vegetation to accumulate until a single fire ignites a significant event. In the following years, reduced burned acreage may signal regrowth and forest health recovery.

Processing this data presented mixed challenges. On the positive side, acreage was consistently recorded for all wildfires, making it reliable if sourced accurately from the US Geological Survey. However, limiting the analysis to fires within the fire season (May 1 - October 31) presented difficulties in extracting exact start and end dates for wildfires. Often, only a start date was available, and when dates were missing, it was assumed the fire occurred within the fire season. As a result, the actual acreage burned could vary significantly from reported totals.

Plot 3



The third plot compares our smoke estimate with AQI data from the EPA. The x-axis represents the years from 1960 (the start of our wildfire data) to 2020, while the y-axis shows generic “values” for comparison. AQI values are measured in ppm, but I kept the smoke estimate unitless to focus on relative differences between the two metrics.

The smoke estimate was calculated by dividing the square area burned (in miles) by the squared distance from Sioux Falls to each fire, then multiplying by 5000 to bring the values closer to the AQI scale. From the plot, it appears that the smoke estimate has limited correlation with the AQI. There are, however, some multi-year periods, such as 1997-2001, where both metrics increase or decrease in tandem.

The main takeaway from this plot is that due to the sparse wildfire data, inconsistencies in AQI data, and the simplicity of the smoke estimate formula, accurately estimating smoke levels for a specific city is challenging. Additionally, since this estimate only considers smoke on October 31st each year, the level of detail is limited.

Reflection statement:

One major insight I gained from answering the research question in this assignment was the impact of each decision on the final results. Choices such as which AQI score to use and how to handle prescribed fires each carried significant weight on the outcomes. This process underscored the importance of making careful, informed choices in analysis.Additionally, this assignment reinforced the idea of “garbage in, garbage out.” My estimated smoke metric proved insufficient, as reflected in the broad 95% confidence intervals, which indicated that I couldn't reliably model smoke on a yearly basis. A more accurate model would require deeper insights into atmospheric smoke composition, local weather patterns, the types of materials burned, and the methodologies behind EPA AQI measurements. With a better understanding of these factors, I would be better equipped to estimate smoke levels accurately.

I also learned valuable lessons about working with incomplete datasets. In this project, missing data meant having to balance extracting what information was available while predicting other components. The experience emphasized how much the quality of data collection influences results: for instance, including or excluding wildfires without precise start and end dates could change the dataset size by as much as 30%. This variability shows that achieving reliable results begins at the data collection stage.

The collaborative aspect of the project added a helpful perspective. Hearing others’ approaches to computing their smoke estimates highlighted the diversity of possible methods. Initially, I planned to include a component for the duration of fires but decided against it after some of my friends warned me that there were a lot of missing values. Instead, I opted for a simpler statistical approach, aligning with others’ strategies to improve the chances of detecting patterns, even if limited.

Finally, the importance of attribution was clear throughout. I even noted in the markdown comments that some of my functions used the tutorial code and modified it to iterate multiple times. For extracting wildfire data from JSON, I diverged from the tutorial by importing the data directly into a dataframe, which I worked on with a friend to develop. This ended up saving a bunch of time when loading the data. Using dataframes also made it much easier to analyze the data compared to the JSON format, so it ended up being a great decision. Additionally, I consulted my classmates for help with a few specific challenges. They helped me create a method for converting strings into lists of lists, and assisted me in handling datetime

manipulations more intuitively. Overall, the tutorial code was invaluable, especially in guiding me through JSON processing and making API calls to the EPA, which were areas I had minimal experience with.

As a whole, this assignment provided a hands-on learning experience in data-driven decision-making, underscoring the role of data quality and the collaborative refinement of methods.