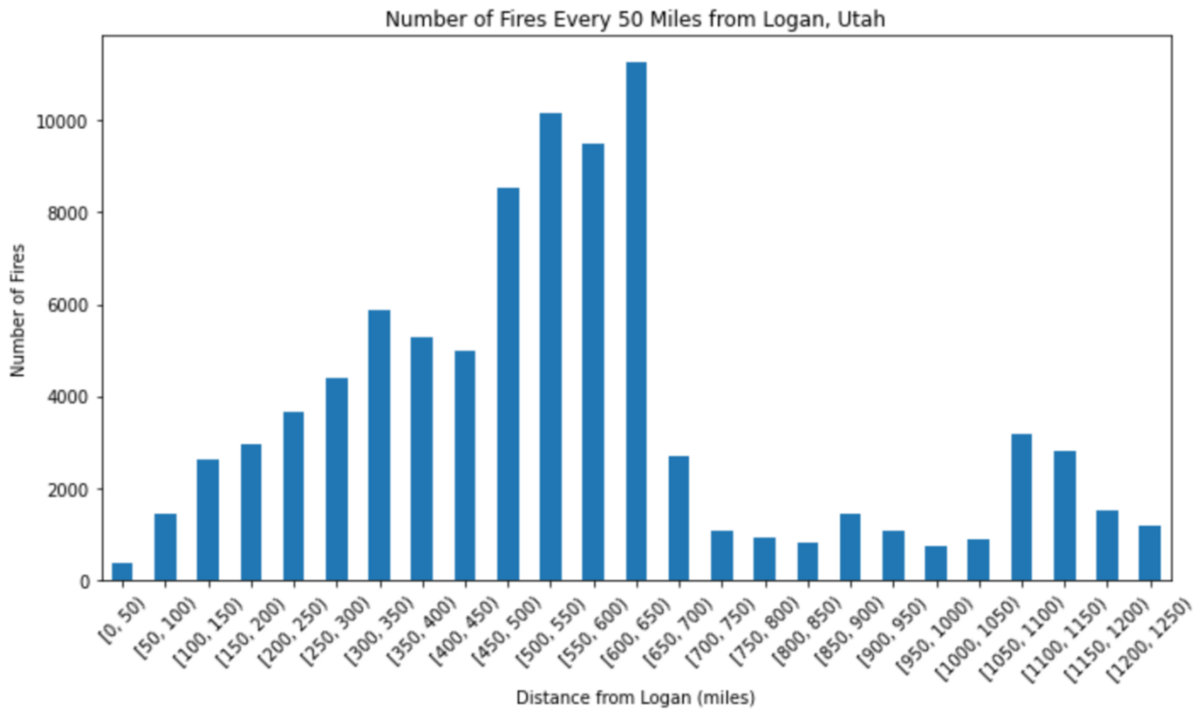
Bazham Khanatayev
Common Analysis Part 1, Visualization Explanations and Reflection

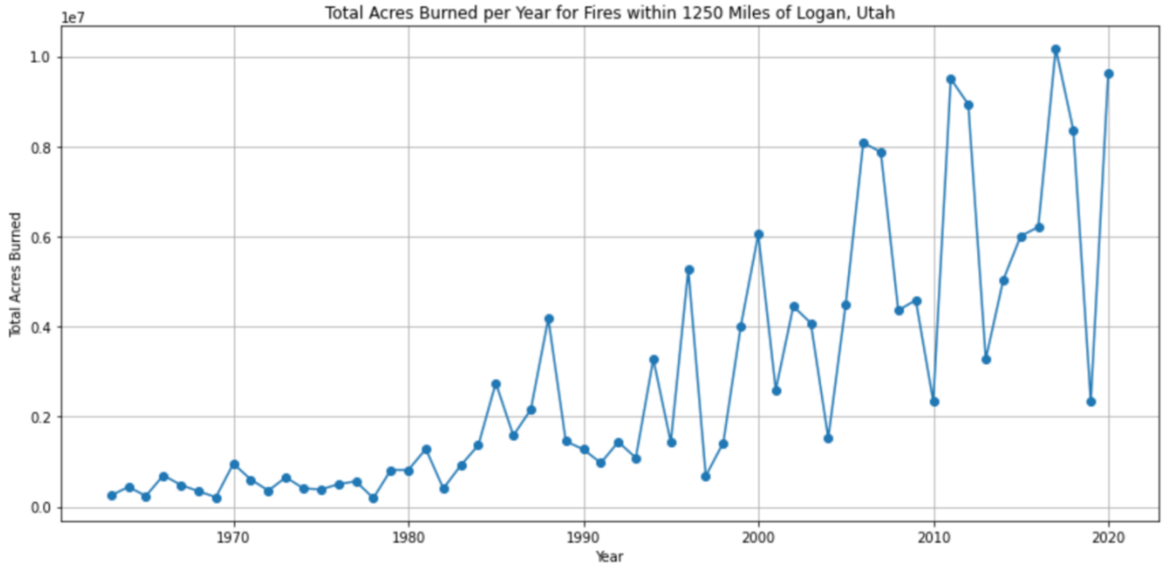
Visualization 1: Produce a histogram showing the number of fires occurring every 50 mile distance from your assigned city up to the max specified distance.



The x-axis of the bar chart represents the distance from Logan in miles, ranging from 50 to 1250 in increments of 50 miles. The y-axis represents the number of wildfires that occurred within each distance range. The best way to read this figure is to look at the x axis to get an idea of the ranges and then look at the y axis to see how many fires fit into each of those buckets. Each row in my dataset represents a fire and each fire has a specific distance from Logan, Utah. This distance was calculated between the coordinates of the city and the coordinate point within the fire polygon which is closest to the city. Only the fires with a distance less than or equal to 1250 miles were included in the final dataset. It is also important to note that before the 1990's, there was significantly less data available on wildfires and diligent documentation was not as prevalent. There could be some bias in the data if the nature of the landscape was different before 1990. This bias would skew the distribution of fires we see in the histogram.

We can see clearly that the bulk of wildfires that are within 1250 miles are within 650 miles of the city. There is also a range of wildfires that occur between 650 and 950 miles from Logan, with a slight decrease in frequency beyond 950 miles. It picks up again past 1000 miles and drops after 1100. This suggests that the proximity to forests, grasslands, and other fire-prone areas plays a significant role in the occurrence of wildfires near Logan.

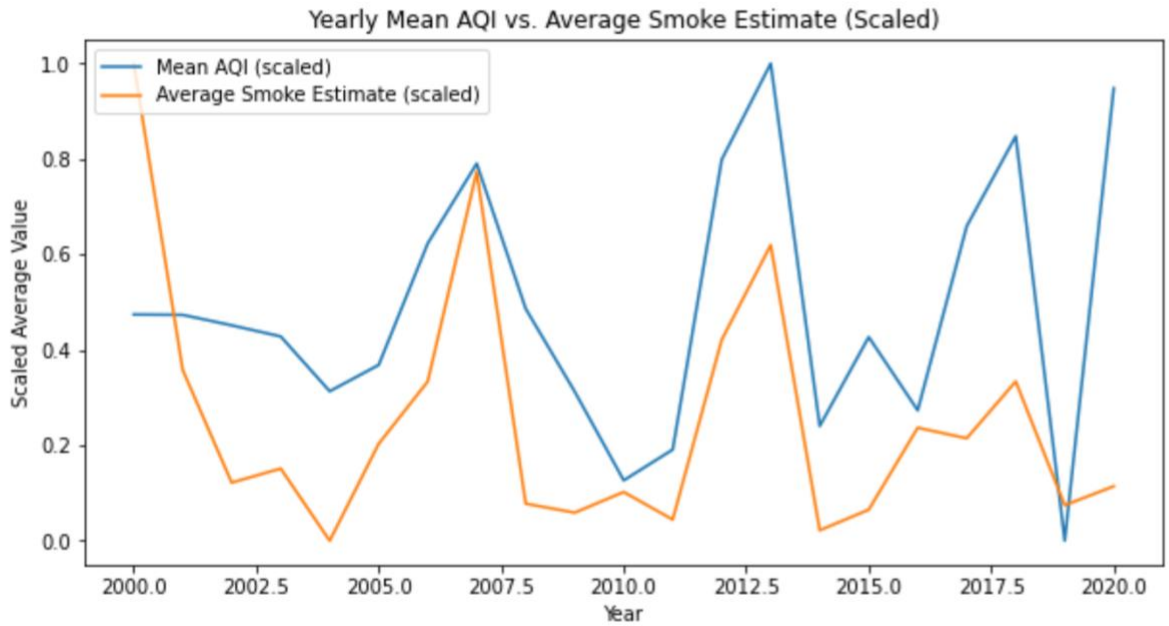
Visualization 2: Produce a time series graph of total acres burned per year for the fires occurring in the specified distance from your city.

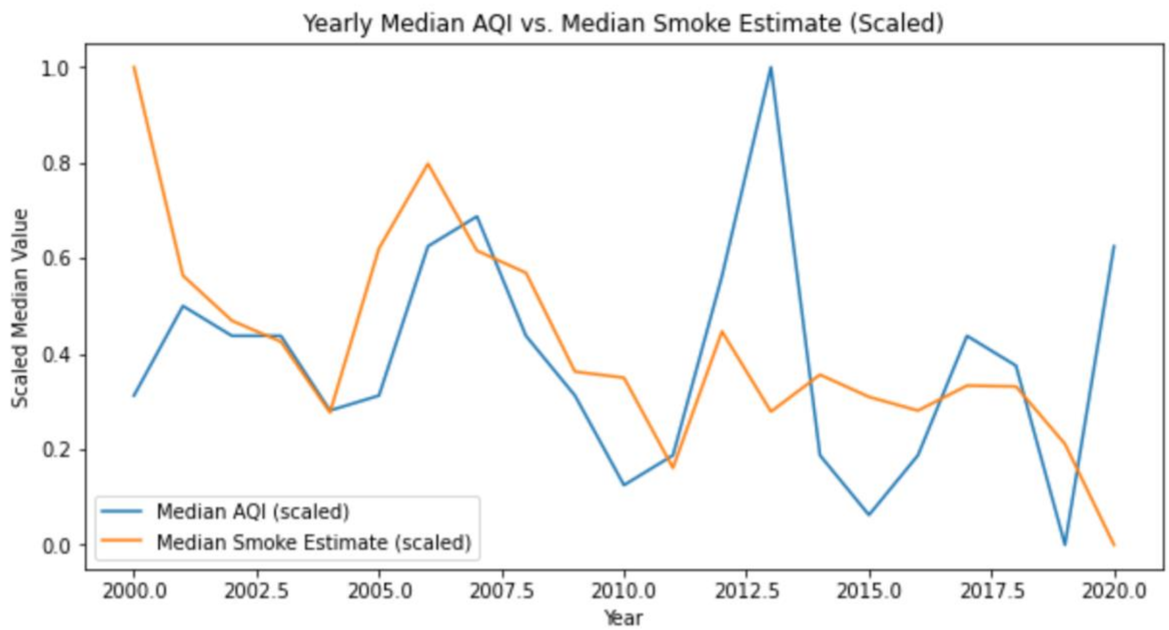


My graph depicts the total acres burned due to wildfires within 1250 miles of Logan, Utah, from 1970 to 2020. The visualization employs a line graph to showcase the trend of total acres burned over time. The x-axis represents the year, ranging from 1963 to 2020, while the y-axis represents the total acres burned in each year, with a maximum value of approximately 10,000 acres. The calculations behind this visualization were straightforward, our given dataset (the USGS Combined wildland fire dataset) contained a feature called GIS Acres that provided us the acreage of the fire; we only included fires that were within 1250 miles of Logan, Utah. This distance was calculated between the coordinates of the city and the coordinate point within the fire polygon which is closest to the city.

The line graph clearly demonstrates a significant increase in the total acres burned due to wildfires over the years. In 1970, only around 2000 acres were burned. However, the trend steadily increased throughout the 1980s and 1990s, reaching a peak of nearly 10,000,000 acres in 2020. This suggests that the frequency and intensity of wildfires in the region surrounding Logan have been on the rise.

Visualization 3: Plot 3: Produce a time series graph containing your fire smoke estimate for your city and the AQI estimate for your city.





My graphs showcase the relationship between the average AQI (Air Quality Index) and the average smoke estimate for the years 2000 to 2020. The visualization employs a line graph to display the trends for both metrics over time. The reason that is starts at 2000 is because that is when the estimates became available from the EPA for my specific city.

We also have to make that we are clear on the units. The Air Quality Index (AQI) is a dimensionless index created by the Environmental Protection Agency (EPA) to communicate how polluted the air currently is or how polluted it is forecast to become. It's a scale that runs from 0 to 500, where the higher the AQI value, the greater the level of air pollution and the greater the health concern. An AQI value of 50 or below represents good air quality, while an AQI value over 300 represents hazardous air quality.

For PM2.5 (particulate matter less than 2.5 micrometers in diameter), which is often a major component of smoke from wildfires, the AQI is based on the concentration of PM2.5 particles in the air over a 24-hour period and is typically measured in micrograms per cubic meter ($\mu g/m^3$).

The x-axis represents the year, ranging from 2000 to 2020. The y-axes represent the average AQI in either median or mean. Both axis have been scaled because their original scales were vastly different and it would be difficult to make a comparison on the same plot. The effect of this scaling is to translate each feature individually such that it is in the given range on the training set, which is typically between zero and one. This has the benefit of bringing different features into a common scale while maintaining relationships in the data.

Let us discuss how we calculated the values. Starting with smoke estimates. The smoke impact score that I calculated is based on the size of the fire and its distance from a specific location, following the inverse square law of dispersion. This law suggests that the intensity of an effect such as smoke diminishes with the square of the distance from the source.

Here is the formula for the smoke impact score:

$$\text{Smoke Impact Score} = \frac{\text{Size of Fire}}{(\text{Distance from the City})^2}$$

In this formula:

- "Size of Fire" could be represented by the number of acres burned.
- "Distance from the City" is the distance between the fire's location and the city in question (in this case, Logan, Utah).

By applying this formula to each fire event, you can estimate the smoke impact on the city. The smoke impact score is higher for larger fires and those closer to the city, and it decreases as the distance increases. This score helps in understanding and estimating the potential smoke exposure to the city from various fires.

Every fire had one of these scores and I just simply took the mean and median of all of the fires for a given year.

Now let’s discuss the AQI score. When calling the API, this is the logic I used when deciding on which sensors to focus on:

The code is designed to retrieve daily summary data for particulate matter with a diameter of 2.5 micrometers or smaller, commonly referred to as PM2.5. The 'param' parameter in the API call is set to '88101', which is the code for PM2.5 in the EPA AQS database. PM2.5 is a significant pollutant often used as an indicator of air quality because these fine particles can penetrate deep into the lungs and even into the bloodstream, causing health issues. The concentration of PM2.5 is particularly relevant when estimating smoke impacts from wildfires, as these particles are a major component of wildfire smoke.

When looking at the two graphs, we can see how they follow each other pretty well until they reach 2012 and then it diverges more. This suggests that my smoke estimate and the AQI score are moderately similar in measuring the smoke impacts from wildfire.

Part 2 Reflection:

Working on this project felt like navigating a solo journey within a team framework. The assignment underscored the value of collaboration, yet it turned out that I spent a good chunk of time working through it on my own. This isn't to say teamwork wasn't important, but rather that the independent route I took had its own set of perks. I ended up really diving into the data and understanding it very well. Also, I believe that being a part time student makes a bit more difficult to work with my peers and establish relationships. I have done so and made progress, but there is an inherit difference between being part time and being full time. I was more collaborative in my undergraduate studies and I believe the time spent on campus and in classes contributed to that.

Delving into the complex data surrounding wildfires, I found that working independently helped me really get to grips with the intricacies involved. However, I did not completely by myself the entire time. Firstly, I asked a question in slack after initially calculating and filtering out the fires within 1250 miles of my city. I only got about 15 fires and after using geopandas, I only got about 1500 fires. I mentioned the number in slack and my classmates responded with the numbers of fires that they got after filtering. There numbers were all significantly higher than mine and it made me realize that there must be some error in my code. I reviewed the example notebook very carefully and was able to start the distance calculation successfully. However, I kept getting an index error after around half an hour of running the code. This was a pain point for me and I ended up asking one of my classmates. She suggested that I take a look at the features and filter out missing values. Later that night, I took her suggestion and made my function more robust by skipping the samples with missing values. I am sure that eventually I would have figured out the issue myself. However, asking a classmate made it so much faster.

Since we all have different backgrounds and different ways of thinking, some types of problems may be easier to solve for certain people. In this case, no code was shared. Even though I did not share any code this assignment, I believe that it is valuable to see how other people implement solutions. I have seen some very clever things this way. The first main takeaway for me was seeing first-hand how a small tip from a colleague can make solving a problem so much faster.

The second takeaway I got from this assignment was how much bias there could be in even very straightforward calculations. Let us take distance from Logan, Utah. Since I calculated this distance using the coordinate polygon data, I had to decide on whether I should measure from the center of the polygon or measure from the closest coordinate edge. I am also aware in some bias in the data. For example, not much fire data was collected before 1980 and it could skew the results. I discussed with some of my classmates, and it seemed like most people went with the shortest distance path. Knowing that my classmates are having to make similar decisions does put me at ease to some extent. I know that I can at least get a bit of feedback or direction from my peers, and I know how valuable it can be.

It is also important to mention that the provided code was very useful. I used the reader class that was provided as well as the general outline for the API. The code simply worked, and I knew what inputs and outputs I needed. The code looked fine and it did not make sense for me to not use it.