Writeup DATA 512 Project Part -1

- Abhinav Duvvuri

Visualization 1

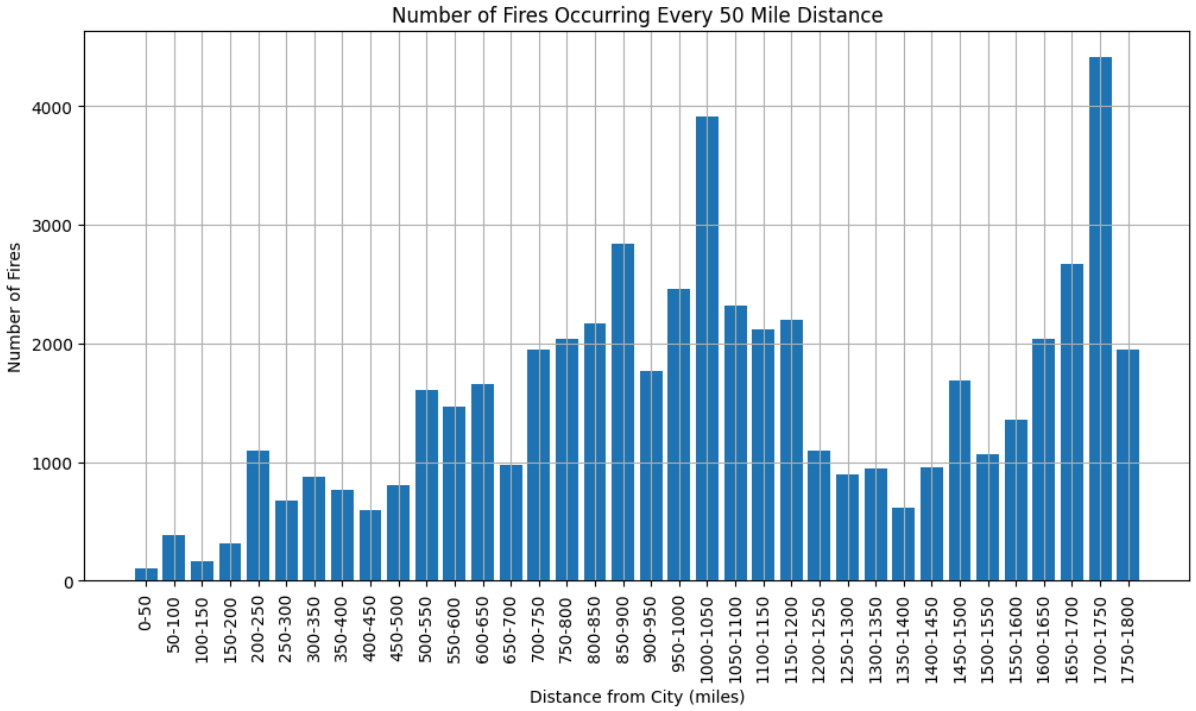

Figure 1: Number of fires occurring at every 50 mile distance

The figure shows the number of fires that were there for every 50 mile distance for the city of Philadelphia in the state of Pennsylvania. The histogram shows the number of fires that occurred at a distance of 50 miles, 100 miles, 150 miles and so on from the city of Philadelphia. As mentioned in the assignment, I have used a distance cut off of 1800 miles. This means that all the fires that are more than 1800 miles away from the city of Philadelphia have been removed from the data. The axis shows the distance from the city in miles and is grouped for every 50 miles. The Y axis shows the number of fires that have occurred over the years for each bin. The data for this graph comes from the wildlife data that was taken from the government website and retrieved using the API provided and code can be found in the first notebook. The shortest distance was computed using the pyproj module in python. Next filtering was done to filter out the distance above 1800 miles the data was grouped for every 50 miles and the count of the fires was taken bin wise. Using Python's matplotlib library I was able to plot this histogram.

The figure seems to suggest that many of the fires that were very near the city were actually less in comparison to the fires that were farther away from the city. One reason could be that because the cities are urban areas and have less forest in general and places outside the cities might have more for us and that might explain more fires there. Also using the same

hypothesis, I suspect that I don't 50 to 1500 miles away from the city. There are fewer fires likely because there are other cities in the distance. However, this is just a hypothesis and I would need more data to confirm this. Another interesting observation is that most of the fires have happened farther than 650 miles from the city; however since we have only considered less than 650 miles of fires, we might miss out on some amount of pollution that has traveled from these fires into the city.

Visualization 2

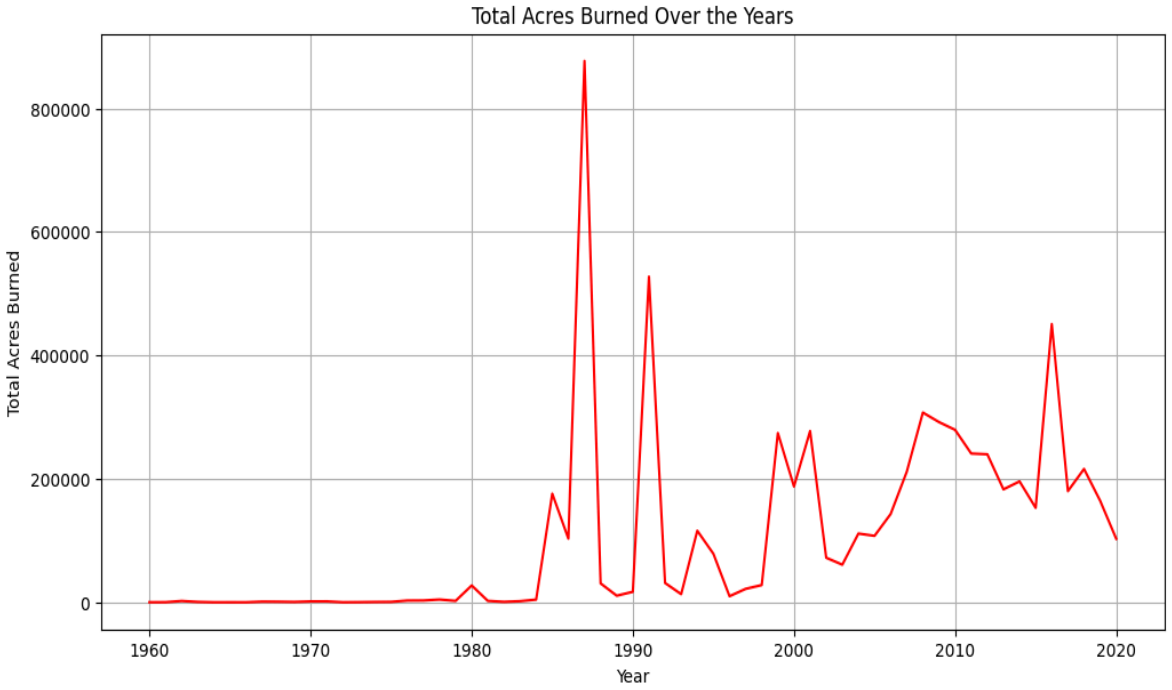

Figure 2: Total acres of area burnt over the years

Figure 2 depicts the total area burnt over the years. The X-axis shows us the year from the year 1960 to the year 2020. The Y-axis shows the total number of acres burnt. The data of this graph comes from the wildfire data that is taken from the government website. For a given year, the graph can be read by looking at the Y-axis and seeing the total number of acres that were burnt in that year. The data for this graph too was fetched, using an API and the code for this can be found in the first notebook.

The same data, cleaning steps written above, were used for this as well. The grouping and aggregations, however, were different for this case. First, I filtered out the fires that were over 650 miles away from the city. Next I grouped the data by year and aggregated the number of acres column, which is required to show the Y-axis by the sum of the number of acres. Finally I sorted the data by the ear and filtered out the years that were before 1960.

The figure shows the number of acres burnt has been increasing over the years. Particularly till the year 1985 it looks like very few acres burnt were recorded. This is consistent with the

fact that the data might not have been recorded properly in the early years. This also means that before 1985 the smoke estimates that were calculated might be very off from the true value because of lack of data. Moving on, the data shows a very high peak for the total number of acres burned around 1985 to 1995. It will be interesting to find out what caused the peak during this time. I will analyse this later.

Visualization 3

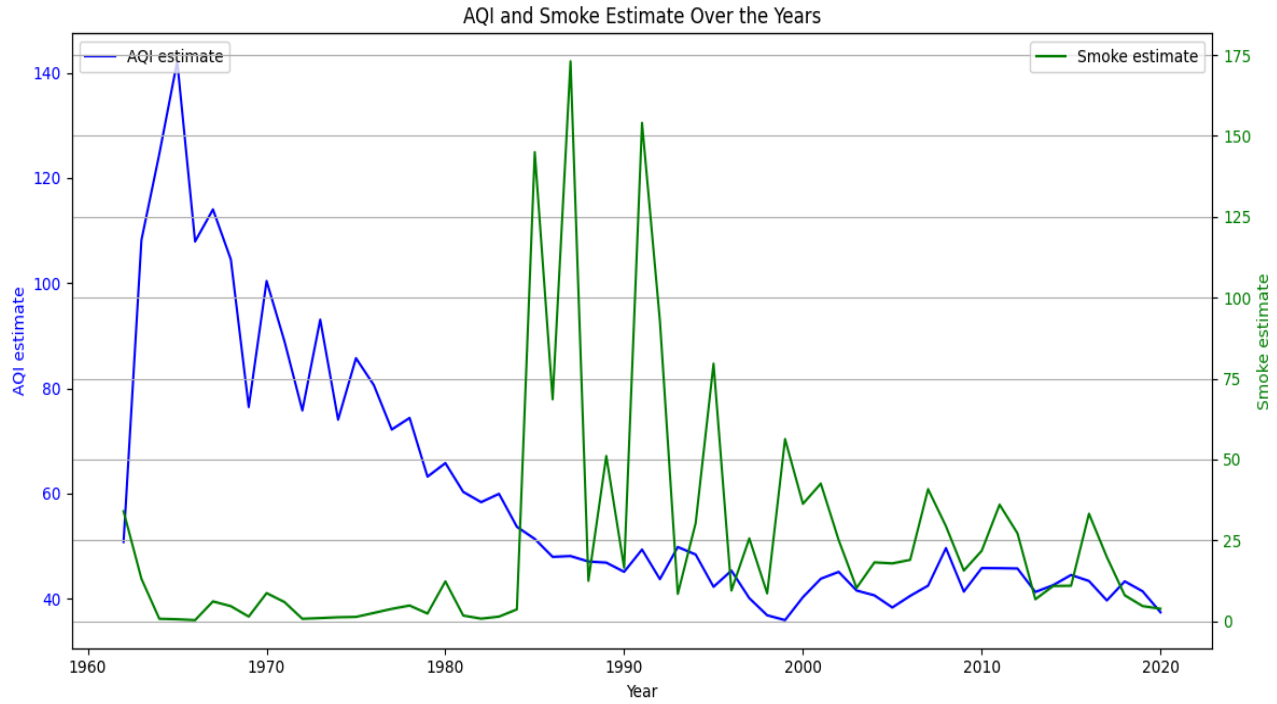

Figure 3: Comparing AQI and Smoke estimates over the years

The figure shows a dual axis chart of the AQI and the smoke estimates for the city of Philadelphia. A dual axis chart is a chart which has 2 Y-axes. The x-axis shows the year from 1960 to 2020. AQI here is the average AQI for all the sites in the city of Philadelphia. The smoke estimate is the estimate of the amount of smoke that was there in the city because of the wildfires. The data for this graph comes from the AQI data and the wildfire data that was taken from the government website. The data was cleaned and aggregated as mentioned above under figure 1.

The smoke estimate was calculated using the formula mentioned in the notebook. The data was then grouped by year and the AQI and the smoke estimates were aggregated by the mean. The data was then sorted by the year and the years before 1960 were filtered out. The figure shows, very interesting trend first for the years from 1960 to 1985. There seems to be no correlation whatsoever between the AQA estimate and the smoke estimate. I feel this is consistent with the fact that the data for the wildfires for these Years is very poorly maintained leading to incorrect estimate values. Also, the AQI values for early years were not vetted thoroughly as stated by the professor in the class.

Next, looking at the trend between 1985 and 1995 it looks like there is a small correlation between the peaks in both estimates. However, the smoke estimate is quite high as compared to the estimate and I don't think they are correlated. I tried to do some initial Analysis to see why this might be the case however, I have not found anything as yet.

Finally, the trend from 1995 to 2020 seems to have good correlation and the peaks seem to match in both cases. Check the graphs created in the notebook where I am showing only 20 years of data. One interesting fact to note is that for some years When the smoke estimate seems to go up, the AQI estimate seems to also rise in the subsequent year. One of the reasons this might happen in my opinion is that since we're taking data between May and October, the smoke readings might only be recorded in the subsequent year. Also, we saw in graph 1 that many of the fires were actually further away from the city and hence the estimate of 650 miles might not be very accurate as many large fires farther away from the city would contribute to smoke based on the winds.

Reflection

One of the issues that I faced in this week's assignment was coming up with a smoke estimate. The file where the metadata was stored for this data set was actually stored in a XML format, which I am not very used to. Further, the elaborate descriptions while being useful were also too lengthy and the formatting itself made it more difficult to understand what each field meant. However, after thinking for some time, I was able to come up with what I think is a proper estimate for the smoke. I am actually happy with the estimate that I created. I also faced some difficulty in understanding how the AQI API worked and with understanding modeling for time series. Since this was a collaborative assignment, I leveraged the opportunity to ask a few classmates and take help.

I think being able to collaborate on this assignment proved to be a very good opportunity for learning new things as well as understanding how other people think. One major Takeaway I had was that the weather people deal with the problem statement is very different from how I think about it. Initially, we understood the problems our own way and then we brainstormed how we could solve this assignment. Specifically for getting the AQI data, I was facing a lot of issues with understanding how the API worked and how to format the data effectively and efficiently. This is where I took help from Raagul Nagendran. I took his help in understanding how to call the API and took a function from him, which helps format the data received from the API into a JSON file. He actually gave me a lot of tips and showed me logical and easier ways of manipulating data to the required formats. Now I have a better understanding of how to work with this type of data.

Another issue that I faced was working with time series modeling. This is because I have never worked on time data before and hence I did not understand what kind of model to use. This is where I took help from Navya Eedula who was very kind in explaining how time series is different from the normal regression models and what are the various alternatives that can be used. She has prior experience working with this kind of data. It is upon talking to her that I understood that ARIMA and ARIMAX are the models that I should consider. I have

also used a function that she provided to model ARIMAX for my time series data. Although I have used her code, I have modified it to make it suitable for my use case and I took the time to understand what it means and what it does. We also talked about plot diagnostics and I got to know about the various diagnostics required for understanding if a given time series model can be used or not. This was actually way better than having to spend hours trying to learn the basics and implement everything from scratch given the tight deadlines of this assignment.

I think such constructive collaboration on assignments while helping get perspective from others also helps with sharing knowledge.

I have attributed all the code in the notebooks itself above where I have used them in the notebooks. I have used a total of 2 functions/blocks of code as mentioned above.

One thing I learnt about the data in general after trying to answer the research question is that smoke might not be the only factor for the air quality in a region. Though I was able to see significant trends for the later years of 2000 to 2020 which suggest that the smoke estimate calculated using the wildfire data is actually correlated with AQI. However there are many years where the trend is not followed. I think, along with looking at wildfires, impacts on other things like the environment, harming activities, such as deforestation, vehicle pollution, etc. might also contribute to the quality of the air which have not been considered in the study. If the sole research question was to answer if wind fire smoke impacted air quality then we would need more data to actually model air quality degradation over the years. That said, I was actually impressed to see that there was a good correlation between AQI and the wildfire smoke estimate.