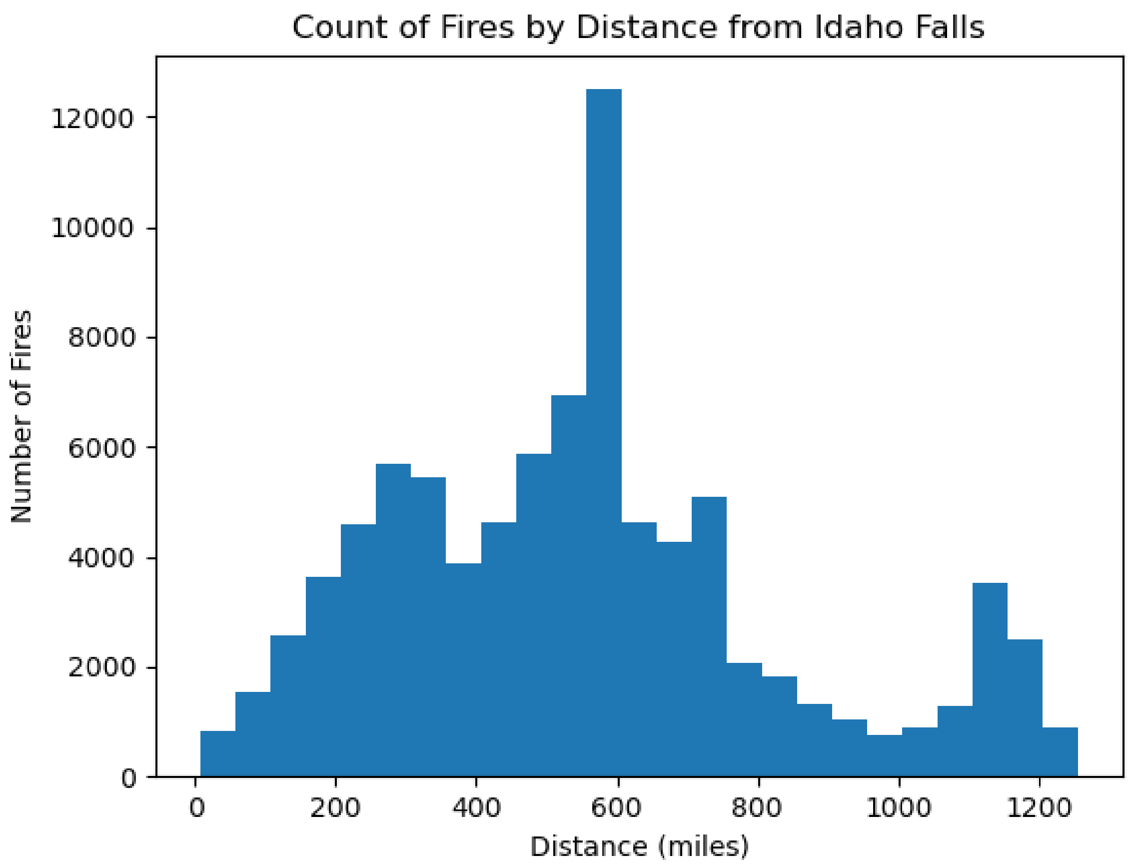
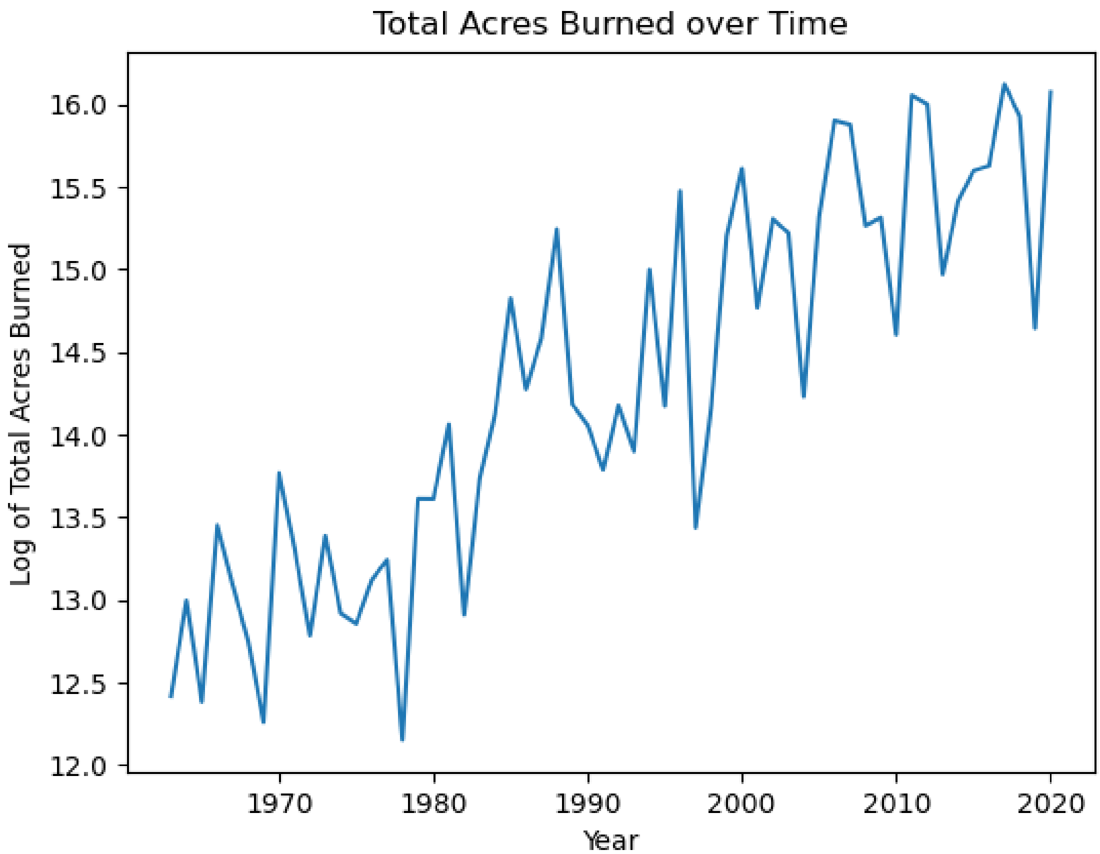
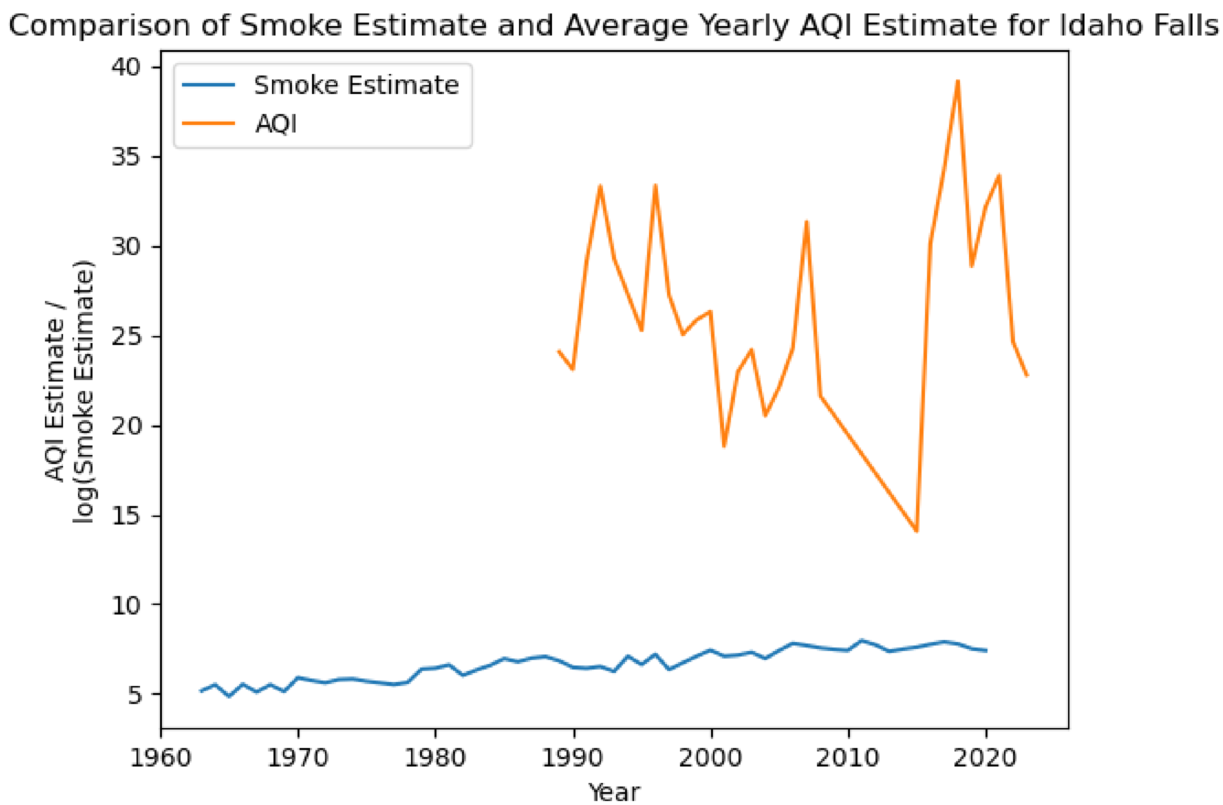
Figure Descriptions



This figure shows a histogram depicting the number of fires that have occurred within 50-mile increments away from Idaho Falls. The higher bars represent a higher number of fire occurrences within that 50-mile distance segment. The x-axis represents the distance away from Idaho Falls. The further right on that axis means the further that the fire occurred from Idaho Falls. The y-axis represents the numerical counts of each instance of fires. This visualization was created from a dataset that included all fires within a 1250 mile radius of Idaho Falls, occurring on or after 1963. Each record in the dataset corresponded to one fire event, so this histogram was just a count of recorded distances.



This figure shows a time series graph of the total number of acres burned by all wildfires within a 1250 mile radius of Idaho Falls, each year. The higher the line, the more total land area that has been burned in that year. However, there is a caveat with this graph. Since the total number of acres is a measurement that ranges between three different orders of magnitude, I decided to log-transform this measurement, to make the visualization much easier to read. Thus, the y-axis shows the log-transformed total acres burned, while the x-axis shows the individual year. This visualization also originated from the wildfire dataset that I mentioned previously, but I aggregated the data by year and summed the total number of acres burned. Then, I took the log of these values for visualization purposes.



This figure shows two lines representing two different estimates over time. The blue line represents the smoke estimate that I calculated to measure the impact of smoke on Idaho Falls. The orange line represents the Average AQI estimate for the given year, where the average is taken from all daily AQI readings, which itself is an average of all sensors within Bonneville County for that day. The x-axis is simply the year, but the y-axis represents two different kinds of estimates. In order to visualize both lines on the same graph, I needed to log transform my smoke estimate so that both lines could be readable. In terms of data origins, the smoke estimate comes from the wildfire data, where the smoke estimate for an individual event was calculated as the log of the acres burned divided by the log of the distance in miles. This number was then summed over an entire year to get the total impact of smoke for that year. In this graph, those year estimates are once again log transformed for readability purposes. The AQI data originated from repeated API calls to the US EPA Air Quality System. This data was then aggregated by date to get an average daily AQI count. The data was then aggregated by year to get annual averages.

Reflection Statement

This assignment has helped me understand the value of collaborating with peers to get insight on how to solve different problems. For instance, Zachary mentioned on Slack about using an ARMA model to predict smoke estimates, which gave me the inspiration to try out an ARIMA model on my own data. This revelation helped me create my own, personalized solution to the problem. When I initially read the instructions for the problem, I had considered a time series model, but I was not sure how to implement it for this data. I had also considered linear regression, although this wouldn't be the perfect use case for linear regression. Answering the research question also made me realize that there is no one-size-fits-all method for defining parameters. For instance, when creating my smoke estimate, I took the log of the acres divided by the log of the distance. I am positive that other people calculated it differently. In a way, as data scientists, we have to define our own measures, and we have to validate them through model building. Furthermore, we have to think critically about why we are using certain calculations, what weights we are using to calculate different measures, and whether the assumptions we are using are valid. Since we are analyzing data, we have to be prepared to justify any calculations and models we decide to use.

When working on this assignment, I also realized how much code I reused from Professor McDonald, who thankfully created the functions for creating API calls to the AQI service, as well as creating a function to calculate distances between two geographic coordinate points. Just creating those functions can be time consuming, so I am glad that I did not have to worry about creating those functions myself. In fact, most of the setup in my notebook is code that was adapted from Dr. McDonald's examples, and I attribute each specific function and code chunk that I used in my notebook. When creating the ARIMA model, I had to do several Google searches on ARIMA models in Python to understand what libraries to import, as well as the functions that I would need for my data. I even visited this link to learn more about the ARIMA parameters to make sure that I was choosing the right model.