DATA 512 - Human Centered Data Science

Project Part 1: Common Analysis

Name: Tanushree Yandra
Date: 11/08/2023

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

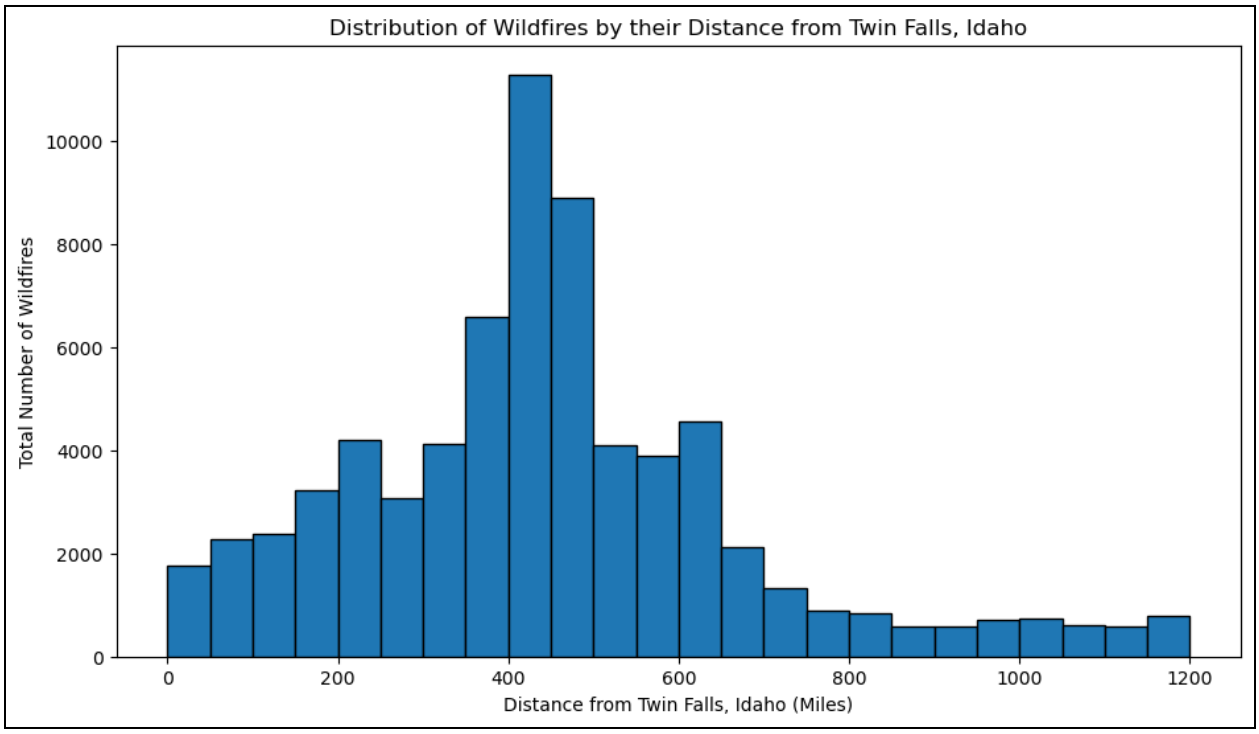

Fig. 1. Histogram of Distribution of Wildfires by their Distance from Twin Falls, Idaho

The above plot shows the distribution of the wildfires as a histogram for the total wildfires occurring every 50 miles distance. In essence, each bar covers the total wildfires within 50 miles buckets. To help read the figure, consider the tallest bar in the histogram which corresponds to the bucket of 400-450 miles. It can be inferred that the most number of wildfires are present in the 400-450 miles radius of Twin Falls, Idaho. The x-axis represents the Distance from Twin Falls, Idaho. Note that for this analysis, only the wildfires within the 1250 miles radius of Twin Falls, Idaho were considered. The y-axis represents the Total Number of Wildfires for a particular bucket of distance.

The underlying data used for this graph is the ‘Wildlife_Data_Processed.csv’ which contains the cleaned wildfires dataset. This dataset was processed by removing irrelevant columns, filtering out overlapping wildfires, and ignoring circular fires of size greater than 1 acre. The final dataset had 72608 rows and 9 columns. One of these 9 features was the Distance which has been plotted above.

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

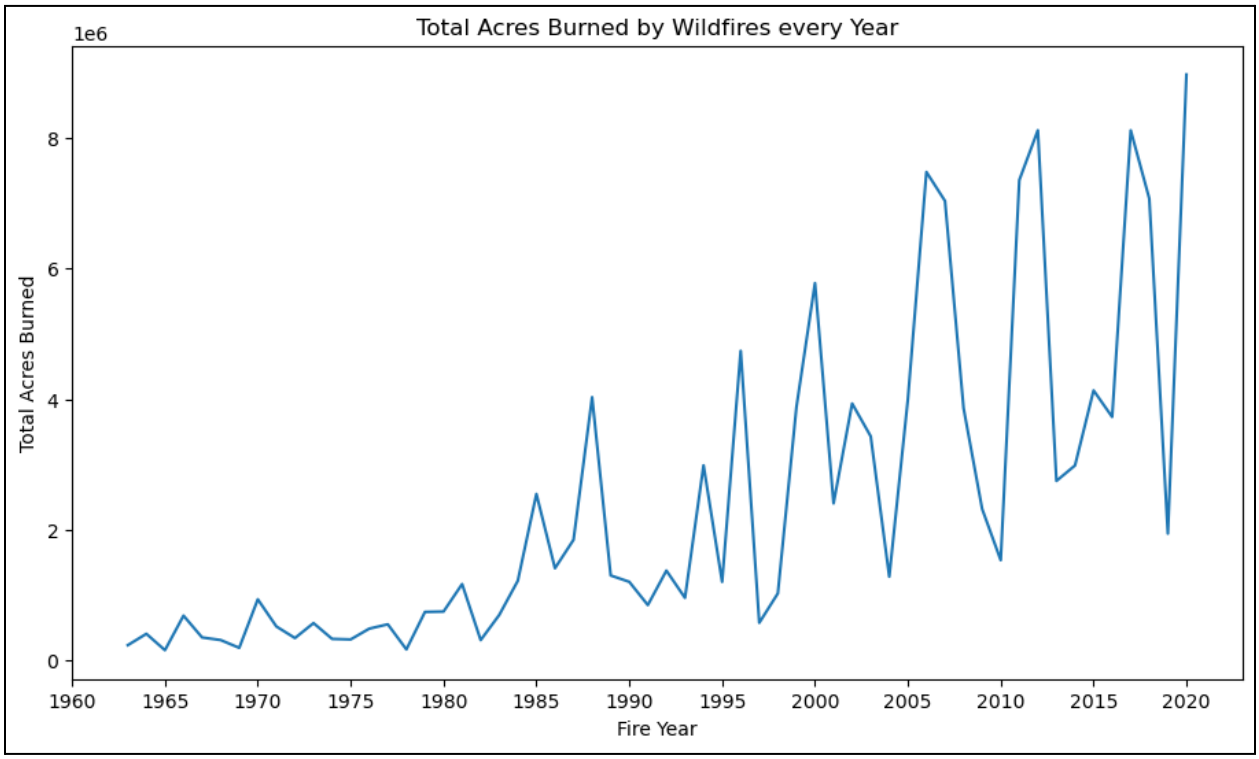

Fig. 2. Total Acres Burned by Wildfires every Year

This figure shows the time series graph of the total acres burned by wildfires per year for the fires occurring within 1250 miles of Twin Falls, Idaho from 1963 to 2020. To read the figure, one can pick any point on the line chart. The corresponding value on the x-axis represents the year and the corresponding value on the y-axis represents the total acres that were burned. Thus, the x-axis represents the year in which the wildfires occurred, and the y-axis represents the total acres that were burned by those wildfires. Based on this information, it can be seen that the total acres burned go up and down periodically. However, the peaks and the dips steadily increase every time indicating that although the total burned acres do go down at times, the average value of the total burned acres over time has been increasing.

The underlying data used for this graph is the ‘Wildlife_Data_Processed.csv’ which contains the cleaned wildfires dataset. This dataset was processed by removing irrelevant columns, filtering out overlapping wildfires, and ignoring circular fires of size greater than 1 acre. The final dataset had 72608 rows and 9 columns. Two of these 9 features were the ‘Fire_Year’ and ‘GIS_Acres’. The dataframe was then grouped by the ‘Fire_Year’ and the ‘GIS_Acres’ variable was aggregated to get the total burned acres per year.

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

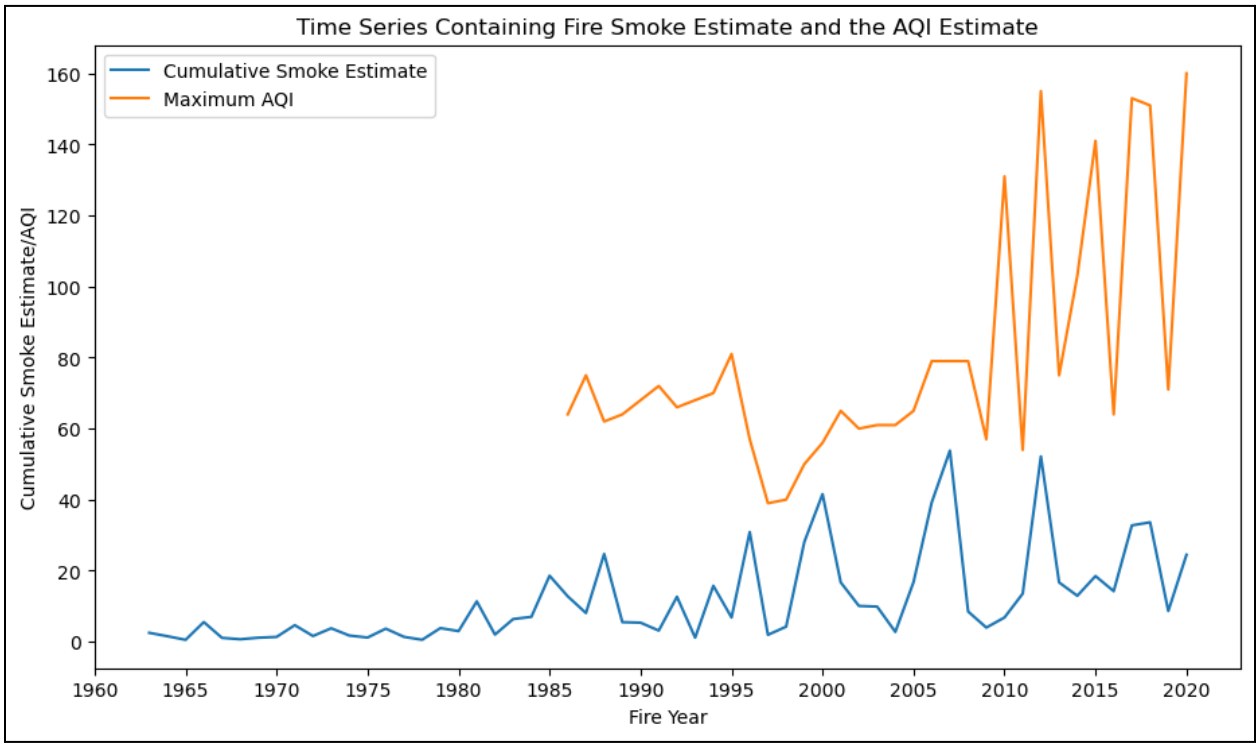

Fig. 3. Time Series Containing Fire Smoke Estimate and the AQI Estimate

The above plot shows the time series graph containing the yearly cumulative smoke estimate and the yearly maximum AQI estimate for Twin Falls, Idaho. The cumulative smoke estimate is calculated using the area burned by a wildfire and its distance from Twin Falls, Idaho. Larger and closer wildfires have a higher smoke estimate when compared to smaller and farther wildfires. The AQI estimate is calculated for the 'fire season' every year which lasts from May 1st to October 31st. For every fire season, the maximum AQI was chosen as the AQI estimate for that year. Since this analysis is concerned about the potential extreme impacts of wildfires on air quality and how high pollution events correlate with smoke estimates, considering the maximum AQI for each year felt suitable. To read the figure, one can pick any year from the x-axis and go vertically up until they hit one of the points on the two line graphs. Those points correspond to the Cumulative Smoke Estimate and the AQI Estimate respectively. The x-axis of the plot represents the year of the wildfires' occurrence and the y-axis represents the two estimates - smoke and AQI.

The underlying data used for this graph are the 'Wildlife_Data_Processed.csv' which contains the cleaned wildfires dataset, and the 'Yearly_AQI_Data.csv' which contains the maximum AQI values for every year. The first dataset was processed by removing irrelevant columns, filtering out overlapping wildfires, and ignoring circular fires of size greater than 1 acre. The final dataset had 72608 rows and 9 columns. Two of these 9 features - 'Distance' and 'GIS_Acres' were transformed (distance was converted to square of distance, and acres were converted to square

miles) to obtain the smoke estimate. The second dataset was obtained directly from the API. The only processing that had to be done was to convert the daily AQI data to annual AQI data. This was achieved by taking the maximum AQI value for the fire season of every year.

Reflection:

This assignment was very insightful and I really liked the 'individual' aspect of the project where everyone is assigned a different city and thus, no two individuals will have the same findings. Keeping it individual, yet opening the possibility of collaboration was a very nice way of letting students help each other out.

A few highlights of specific things I learned while answering the research questions posed in the assignment are:

1. This assignment was an eye-opener on the number of wildfires that have taken place in the US in the last 60 years. It also taught me to deal with a new type of data files - GeoJSON.
2. Secondly, having a keen interest in sustainability, this project was really fun and insightful. It felt like detective work where one probes further and further into the data to understand why a certain phenomenon is taking place. The whole aspect of creating a smoke estimate from scratch, trying out different combinations of features, finding the best combination, evaluating it, and then finally modeling was tedious yet very exciting to work on. The modeling process especially was very difficult as I tried several models before arriving at the moving average model. This assignment thus taught me about the various models that can be used to model a data with a lot of peaks and dips when one has little to no predictors.

As mentioned earlier, the possibility of collaboration while working individually was really useful. While I did not take any code snippet or a particular method/technique from a peer, unintentional collaboration especially in the form of slack was really helpful. The various doubts that were discussed on slack were some that I already had and was struggling with. Thus, when the professor, the TAs and other peers actively took part in the conversation to resolve such doubts, and give some tips, those points really helped shape my project. For instance, the situation where 'curveRings' were popping up while reading the data, threw me off-guard as I did not expect such a variable. Going through the slack conversation, I realized how different students were dealing with that variable. Thus, collaboration played an important role for me to go about my minor bugs during the course of the project.

Attributions:

During the data retrieval process, I used the 'wildfire' user module provided by Dr. David W. McDonald. This module was useful in reading the wildfire features iteratively from the wildfires dataset. I also used the sample code for computing geodetic distances which was again provided by Dr. David W. McDonald. When retrieving the AQI data from the US EPA API, I once again used the

sample code provided by Dr. David W. McDonald for generating the API request and creating daily summary of the AQI data.

Some of my data preprocessing work involved relying on the doubts and solutions provided by my fellow peers on the Slack channel. The doubt asked by Jenny Wong and the solution provided by Dr. David W. McDonald helped me navigate during this process.

For the modeling aspect, I tried out several models ranging from defining my own functions to trying out regression models. Zachary Bowyer's doubt on slack channel gave me an idea to try moving average models. His idea inspired me to choose the Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX) model.