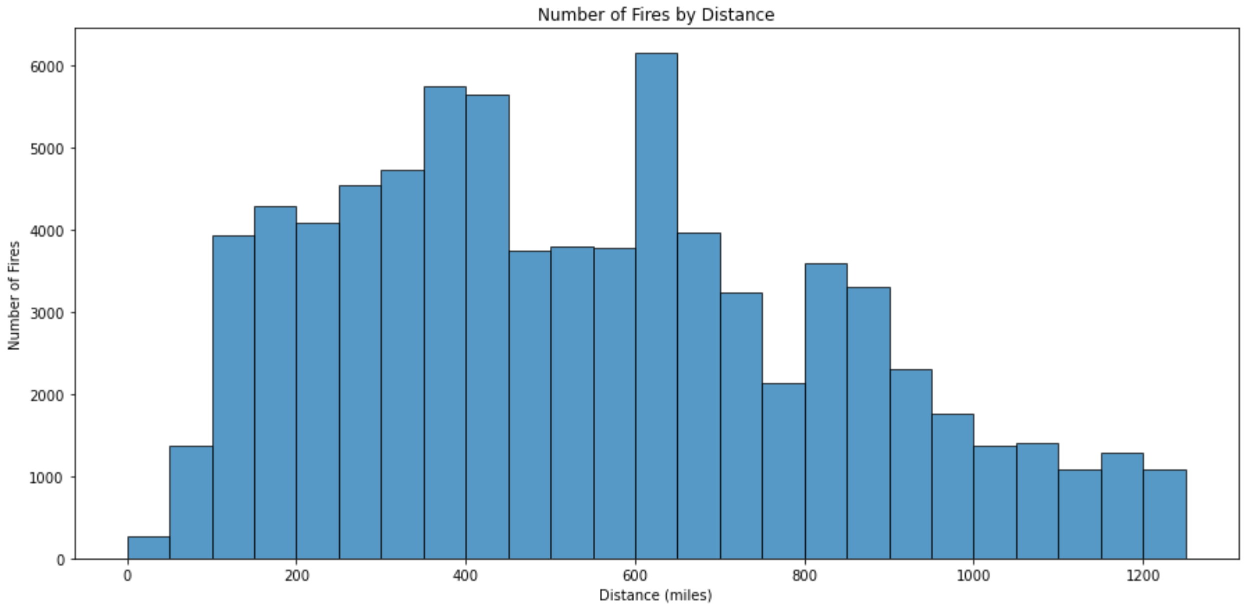
Write and reflect: Common Analysis

Graph 1

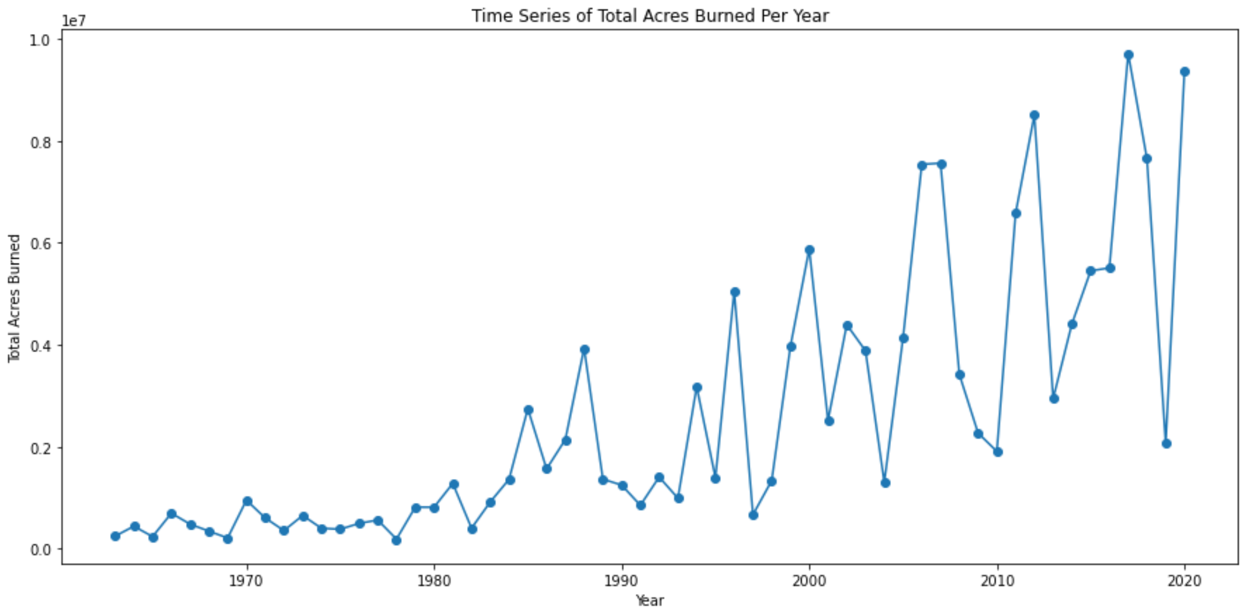


The graph above represents the number of fires at various distances from the city of interest. In this project, I was assigned the city of Lewiston, Idaho. The x-axis represents the distance from the source, measured in miles. This data was obtained through a function defined to calculate the average distance from the fire's boundary to the source. The y-axis represents the count of fires that occurred within this distance radius from the source. The study covers the time period from 1963 to 2023.

The underlying data for this graph is sourced from the USGS dataset on combined wildfires in the USA, which contains information about wildfires across the country. For the purpose of this assignment, the data was filtered to include only wildfires that occurred within a 1250-mile radius of the source city and within the years 1963 to 2023. However, upon applying this filter, it became apparent that there was no data available for the years 2021, 2022, and 2023.

To create this graph, the distance measure was pre-processed into bins of 50, 100, 150, and so on, up to 1250 miles from the source city. This was accomplished using a simple list comprehension in Python. To generate the histogram, the matplotlib library in Python was utilized. This library provides a specific function, 'hist,' designed for creating histograms. It accepts parameters defining the number of bins, the values to be binned (in this case, distances), and various styling options such as transparency and edge color.

Graph 2

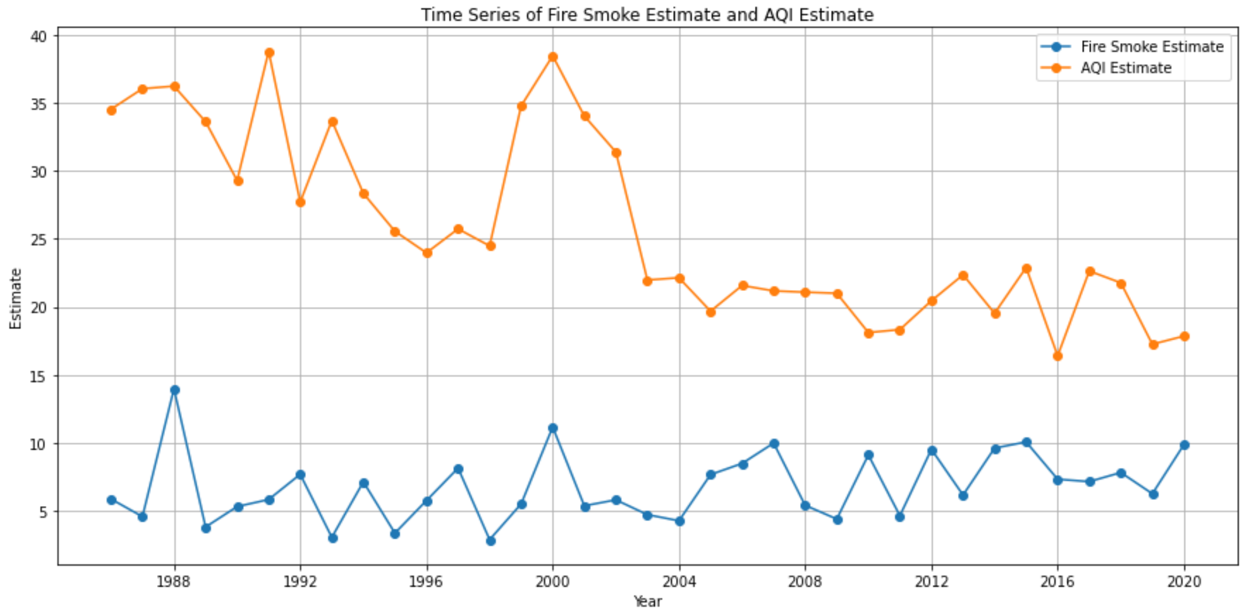


The graph above represents the total acres burned per year within a 1250-mile radius of the source city. The x-axis represents the years from 1963 to 2020. The original intent was to include data for the years 1963 through 2023; however, I discovered that there was no data available for the years 2021, 2022, and 2023.

The y-axis represents the total acres burned during a wildfire. The data underlying this graph is sourced from the USGS dataset on combined wildfires in the USA, which contains information about wildfires that have occurred in the country. For the purpose of this assignment, the data was filtered to include only wildfires that occurred from 1963 to 2023 and were within a 1250-mile radius of the source city. To calculate the number of acres burned by a wildfire, an attribute called "GIS_Hectares" was converted to acres and utilized. This conversion is documented in the USGS data collected for each wildfire.

To create this graph, the Python 'groupby' function was employed to calculate the sum of acres burned for each year. To create the actual plot, the matplotlib library in Python was used. The 'plot' function was employed, accepting parameters for plotting on the x and y-axes, as well as other styling options.

Graph 3



The graph above compares the calculated fire smoke estimates with the AQI estimate from the US Environmental Protection Agency (EPA) Air Quality Service (AQS) API. The fire smoke estimates were obtained for the years 1963 to 2020 because no data was available for the years 2021, 2022, and 2023. These estimates were calculated using USGS data on combined wildfires in the USA, which contains information about wildfires that have occurred in the country. The formula for calculating the smoke estimates is as follows: smoke_estimate = (GIS_Hectares/distance) * Assigned_Fire_Type_Code.

In this formula, GIS_Hectares represents the area of the wildfire, and distance represents the distance of the wildfire from the source. The distance measurement is converted to miles. The Assigned_Fire_Type_Code is a column representing the type of fire where the fire types can be categorized as follows:

- Likely Wildfire: The fire was likely a wildfire, but there is no way to confirm this.

- Wildfire: The fire is a confirmed wildfire based on available attributes.

- Prescribed Wildfire: The fire is a confirmed prescribed fire based on available attributes.

- Unknown - Likely Wildfire: This polygon came from MTBS and was labeled "Unknown," but research by Karen Short indicates that the fire is likely a wildfire.

- Unknown - Likely Prescribed Wildfire: This polygon came from MTBS and was labeled "Unknown," but research by Karen Short indicates that the fire is likely a prescribed fire.

Based on these values, mapping was done to numeric values based on the assurance of the fire occurrence and type (prescribed vs. wildfire).

The AQI data was obtained from the API exposed by the EPA. Although there were no stations at the source city (Lewiston), data was obtained from stations within a 50-mile radius of the city. Two stations were used for different time periods, but the data obtained from these stations only covers the years 1986 to 2023. Therefore, the graph above includes information from 1986 to 2020.

To create this graph, the Python Matplotlib library was used, and the 'plot' function within the library was employed. Various styling parameters were used to assign different colors to the lines.

Reflection Statement

This assignment involved understanding the data related to wildfires in the USA and its impact on air quality. Although the problem statement seems straightforward, it required collating data from different sources and filtering it to align with the scope of this assignment. Since we were each assigned a city to work with and study wildfires in its vicinity, it was crucial to comprehend the concepts of geodetic distance computation techniques and coordinate systems, as we were dealing with geospatial data. This was one of the specific aspects that I learned during this assignment. Understanding the notation of attributes and geometry within a GeoJSON file was interesting, and working with different projections and identifying the correct one for distance calculation was also a significant part of this assignment and a valuable learning experience for me. Although it took some time to become familiar with the data, I believe that it was extremely beneficial, as I am now acquainted with a completely different type of data and could potentially work with similar datasets in the future.

Another valuable lesson I gained from addressing the research questions posed in this assignment was understanding the various factors related to wildfires that might contribute to smoke estimates and, consequently, air quality. By performing an exploratory data analysis (EDA) and comprehending the attributes used in the data provided by the USGS, using the metadata they shared, I was able to intuitively formulate a combination of factors that could potentially serve as an estimate for the amount of smoke generated by a wildfire.

Lastly, this assignment involved collaboration with my peers to grasp the problem statement. Since we were dealing with real-world data, it was essential to ensure that the correct data source from the USGS website was downloaded, and it was also necessary to confirm this with my peers. I enjoyed collaborating and discussing with my peers about creating a smoke estimate. Everyone had different ideas on how factors could be combined to generate a smoke estimate, and it was intriguing to hear everyone's perspectives on how it could be accomplished. This encouraged me to rethink and refine my initial ideas.

I want to acknowledge that I utilized code from Prof. David McDonald's Python notebook for both reading the wildfire data from the USGS and calling the AQI API exposed by the EPA. However, the remainder of the preprocessing and coding was my original work, and I discussed ideas for the smoke estimate and techniques like processing the API response with my peers.