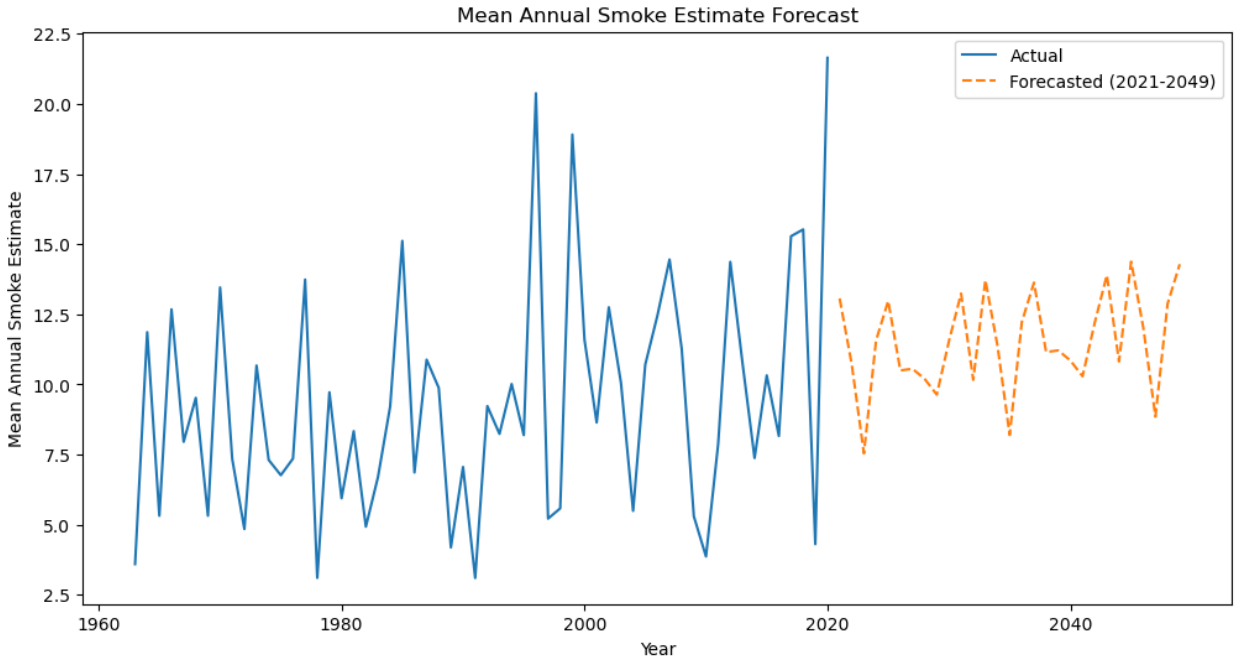
Describing the visualizations



The graph in question is a line plot that serves as a visual representation of the comparison between actual and forecasted cumulative smoke estimates. It covers a time frame from the year 2021 to 2049. The X-axis, which runs horizontally, signifies the years within this specified period, starting from 2021 and extending up to 2049. The Y-axis, the vertical dimension, represents the mean annual smoke estimate, a measurement used to gauge air pollution levels. The unit of measurement for the Y-axis is not specified but could be, for example, in micrograms per cubic meter (μg/m³), commonly used for particulate matter and other air quality metrics.

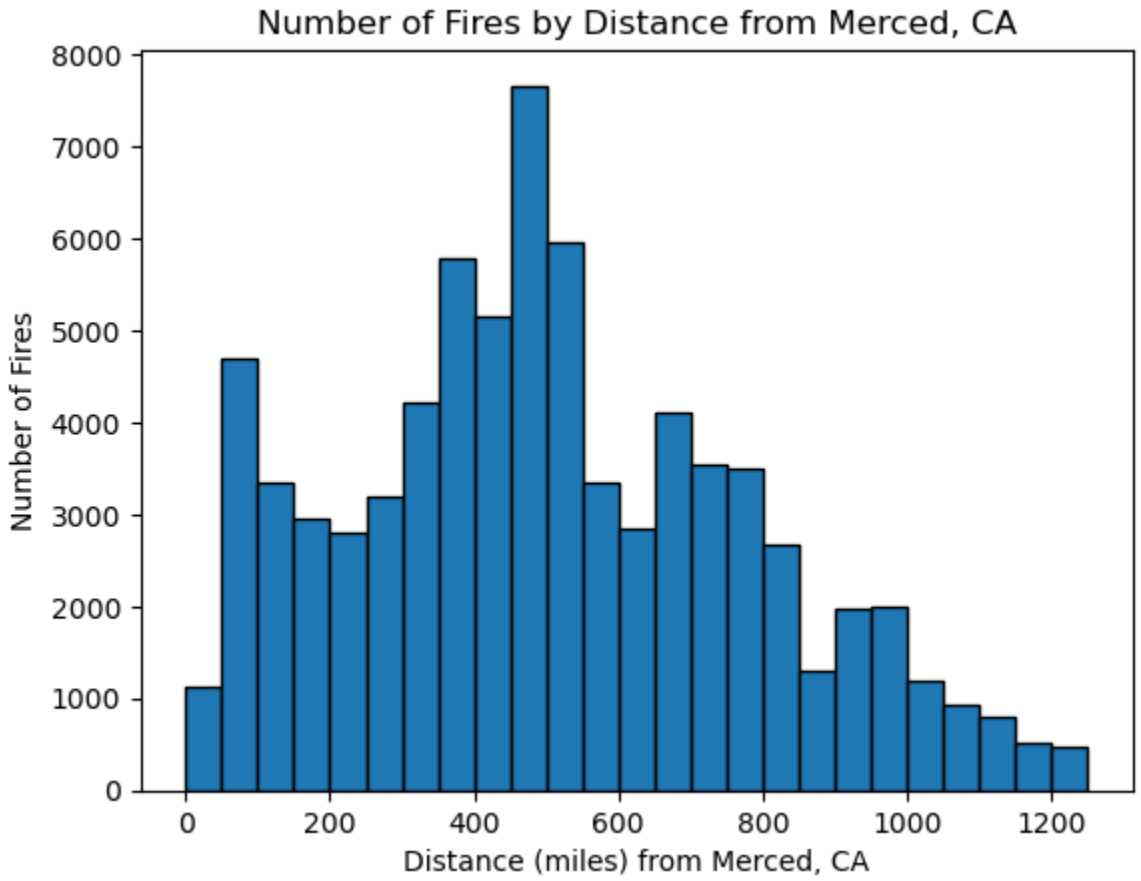
Within the graph, the blue solid line illustrates the actual historical data for smoke estimates. These values are derived from historical records and represent observed levels of smoke or air pollution for each year in the dataset. This historical data is crucial as it serves as a baseline for evaluating the accuracy and reliability of the forecasted estimates.

The orange dashed line, on the other hand, represents the forecasted smoke estimates for the years 2021 to 2049. These estimates are generated by an Exponential Smoothing model, taking into account historical data and the model's understanding of trends and seasonality in the dataset. The model's forecasts are depicted as a line that extends into the future, providing an outlook on how smoke estimates are expected to change over the specified years.

To clarify the contents of the graph, a legend is provided in the upper-right corner, indicating the colors and labels used in the graph. It distinguishes between the "Actual" data, represented by the blue line, and the "Forecasted (2021-2049)" data, indicated by the orange dashed line. Additionally, the graph is given a descriptive title, "Mean Annual Smoke Estimate Forecast," which offers context and conveys the purpose of the visual representation.

In summary, this graph is designed to provide a clear visual comparison between historical and forecasted smoke estimates for the years 2021 to 2049. It offers insights into the model's performance and its ability to predict future trends and fluctuations in smoke levels based on past data, ultimately aiding in the assessment of the model's accuracy and reliability for future forecasting.

Graph 1



This is a histogram used to visualize the distribution of wildfires by distance from Merced, CA, within the years 1963 to 2020.

The code begins by filtering the dataset (`df_fires`) to focus on data related to wildfires occurring between the years 1963 and 2020. This filtering ensures that only relevant data within this timeframe is considered for the histogram.

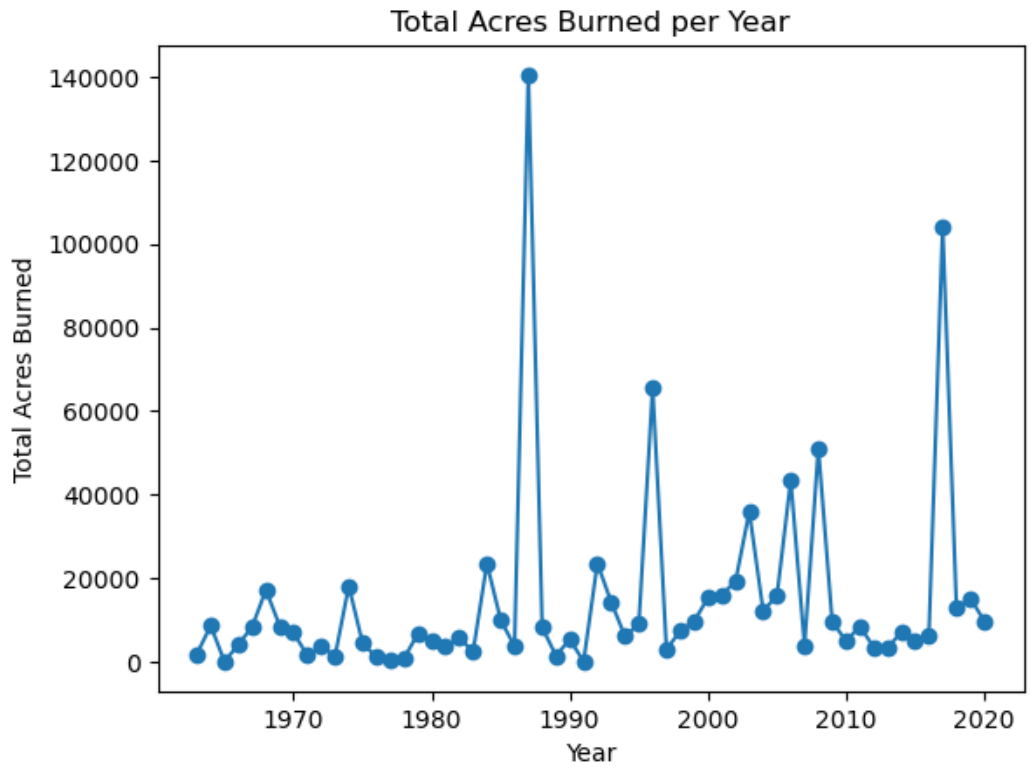
The histogram construction starts with defining the range and bin width for the distance data. The maximum distance from Merced, CA, is calculated from the filtered data, rounded to the nearest whole number. A bin width of 50 miles is chosen to create intervals for grouping the data.

The `plt.hist()` function is then utilized to create the histogram. The 'distance' data from the filtered dataset is used as the main data source for the histogram. The `bins` parameter is set to the predefined bins, which represent distance intervals. The `edgecolor` parameter is set to 'k' (black) to outline the bars in the histogram.

The X-axis is labeled 'Distance (miles) from Merced, CA,' denoting the distance intervals, while the Y-axis is labeled 'Number of Fires,' representing the count of wildfires in each distance interval. The title of the graph, 'Number of Fires by Distance from Merced, CA,' provides clear context for what the histogram illustrates.

Overall, this histogram visually presents how the number of wildfires is distributed across different distances from Merced, CA, within the specified timeframe (1963 to 2020). The graph provides a clear understanding of the spatial distribution of wildfires concerning the city, which can be valuable for assessing wildfire risk and management strategies in the region.

Graph 2



This code segment that generated this graph is designed to filter data within a specified distance range from your city and then create a time series graph to visualize the total acres burned per year within that distance range.

First, the code filters the dataset (`df_fires`) to focus on data within the specified distance range, which, in this example, is defined as 0 to 50 miles from your city. This filtering ensures that only wildfires falling within this distance interval are considered for further analysis.

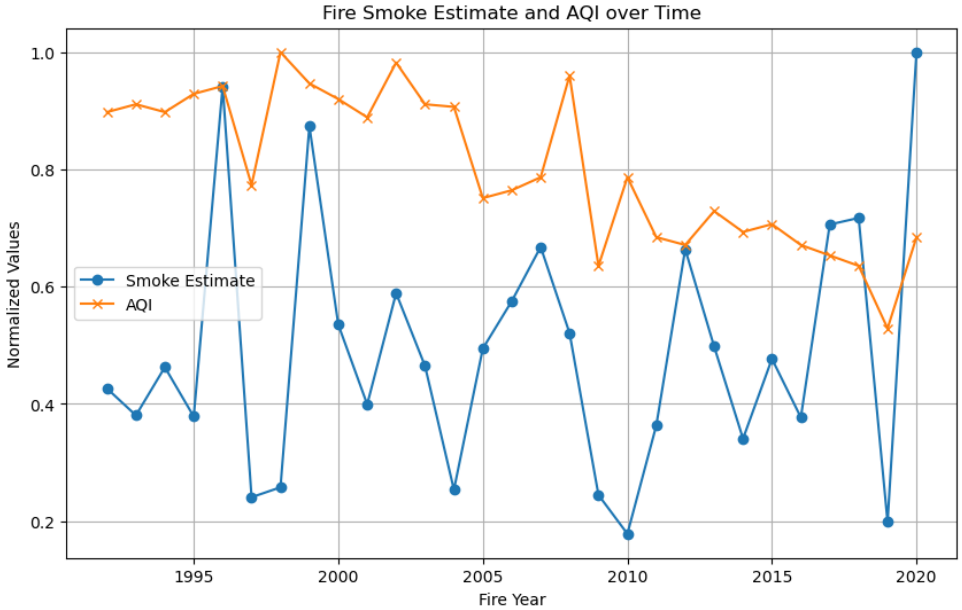
The next step involves grouping the filtered data by year and calculating the total acres burned for each year. This is achieved by using the `.groupby()` function to group the data by 'Fire_Year' and then summing up the 'GIS_Acres' (acres burned) for each year. This results in a series that contains the total acres burned per year for the specified distance range.

The code then proceeds to plot a time series graph using `plt.plot()`. The x-axis is labeled 'Year,' signifying the chronological years from the dataset, while the y-axis is labeled 'Total Acres Burned,' representing the cumulative acres burned for each year. Each data point in the time series is marked with a circular marker ('o') for clarity.

The title of the graph, 'Total Acres Burned per Year,' provides a clear context for what the graph illustrates and what the viewer can infer from it.

In summary, this time series graph visually displays the annual variation in total acres burned by wildfires that occurred within the specified distance range from your city (0 to 50 miles). It offers insights into the historical trends of wildfire severity and can be valuable for understanding the local impact of wildfires in the region and for making informed decisions regarding fire management and prevention measures.

Graph 3



The code that creates this graph is responsible for creating a time series graph that compares two sets of normalized values—'Smoke Estimate' and 'AQI' (Air Quality Index)—over time, represented by the 'Fire Year.'

The code begins by creating a DataFrame, `df_aqi_data`, from a dictionary or data object, with two columns: 'Fire_Year' and 'AQI,' representing the year and the corresponding normalized AQI values. 'Fire_Year' is cast as an integer data type, and 'AQI' values are normalized by dividing each value by the maximum AQI value in the dataset.

Similarly, a second DataFrame, `df_smoke_estimate`, is created from the 'df_fires' dataset, containing 'Fire_Year' and 'smoke_estimate' columns. The 'smoke_estimate' values are then averaged by year and filtered for the years within the range 1992 to 2020. Like the AQI values, the 'smoke_estimate' values are also normalized.

The two DataFrames, 'df_aqi_data' and 'df_smoke_estimate,' are merged based on the 'Fire_Year' column to create a combined DataFrame, 'merged_df,' which holds both normalized AQI and smoke estimate values for the same years.

Finally, a time series graph is plotted using `plt.plot()`. The x-axis is labeled 'Fire Year,' representing the chronological years in the dataset. The y-axis represents 'Normalized Values.' Two lines are plotted, one for the 'Smoke Estimate' data and another for the 'AQI' data. Data points on the graph are marked with different markers (circles for 'Smoke Estimate' and 'x' for 'AQI). The legend displays labels for each line. The title of the graph, 'Fire Smoke Estimate and AQI over Time,' provides context for the graph's content, and the 'grid(True)' function adds a grid to improve data readability.

In summary, the resulting time series graph visually compares the normalized values of 'Smoke Estimate' and 'AQI' over time, allowing for the assessment of trends and potential correlations between air quality and smoke estimates from 1992 to 2020. This graph is useful for understanding how these two parameters have evolved in relation to each other over the specified period.

Reflection Statement

This assignment involved delving into the intricacies of wildfire-related data in the USA and its impact on air quality. While the task seemed straightforward on the surface, it necessitated the consolidation of data from diverse sources and meticulous filtering to align with the specific objectives of the assignment. Given that each of us was assigned a particular city to investigate and analyze wildfires in its vicinity, it was imperative to grasp the concepts of geodetic distance calculations and coordinate systems, given the geospatial nature of the data. This aspect stood out as one of the key takeaways from the assignment. Understanding the attributes and geometry notations within a GeoJSON file proved to be a fascinating learning experience. Furthermore, working with various projections and determining the most appropriate one for accurate distance calculations emerged as a significant aspect of the assignment, enhancing my knowledge in this area. While it did require some time to become acclimated to the data, I consider this learning process highly beneficial, as it has introduced me to a distinct data type and the potential to handle similar datasets in the future.

Another valuable insight gained from addressing the research questions in this assignment was developing a comprehensive understanding of the multifaceted factors associated with wildfires, which could contribute to smoke estimates and, consequently, air quality. By conducting an exploratory data analysis (EDA) and diligently interpreting the attributes provided in the data from the USGS, coupled with a keen utilization of the metadata shared, I was able to intuitively devise a composite of factors that might serve as an estimate for quantifying the volume of smoke generated by a wildfire.

Lastly, collaboration with peers played a pivotal role in shaping my comprehension of the assignment's objectives. Since we were handling real-world data, it was imperative to ensure that we were working with the correct data source from the USGS website, and this required mutual confirmation among peers. Engaging in discussions and collaborative problem-solving with my peers regarding the creation of a smoke estimate was particularly rewarding. Each team member brought distinct perspectives and innovative ideas to the table, allowing me to reevaluate and refine my initial concepts. Additionally, I leveraged code snippets from Professor David McDonald's Python notebook for tasks such as reading wildfire data from the USGS and making API calls to the EPA for AQI data. However, the bulk of the data preprocessing and coding represented original work, and I engaged in fruitful discussions with peers to explore concepts related to smoke estimation and techniques for processing the API responses.