SAGNIK GHOSAL

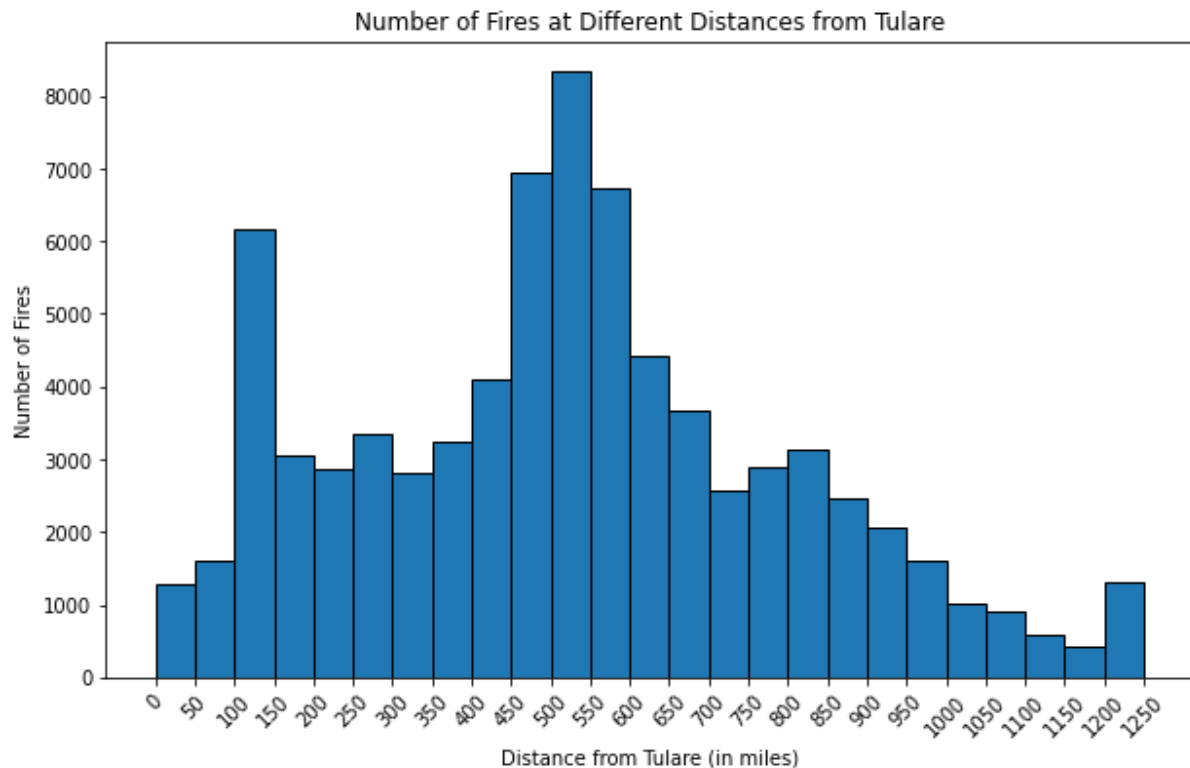## DATA 512: Course Project - Common Analysis

Visualization Explanations:

1.

The histogram visualization represents the number of fires occurring at various distances from the assigned city, Tulare, spanning a range of up to 1250 miles. This histogram provides valuable insights into the historical distribution of fires in proximity to Tulare between the years 1963 and 2023.

On the y-axis, the viewer can observe the count of fires, while on the x-axis, distances are measured in miles from Tulare. The data used for this histogram is derived from the original fire dataset. The dataset was filtered to include fires within a range of up to 1250 miles from Tulare and between the years 1963 to 2023, with the "Distance" column serving as the basis for constructing the histogram. The histogram bins are set in 50-mile increments, ranging from 0 to 1300 miles.

Upon examining the plot, several key patterns emerge. The trend illustrates that the number of fires is lower within the first 100 miles from Tulare. As the distance from the city increases, there is a notable rise in the number of fires, particularly between the 100 to 450-mile range, where the count ranges between 3000 and 4000 fires. The most pronounced peak occurs between 450 and 600 miles from Tulare, with the number of fires spiking to as high as 7000 to 8000 incidents. Subsequently, the number of fires steadily decreases as the distance from Tulare extends beyond 600 miles, with fewer than 1000 fires recorded at distances of 1000 miles or more from Tulare.

These trends may be attributed to various factors. The lower number of fires within the initial 100 miles could be influenced by the presence of an urban wildland interface (UWI) zone, which typically experiences reduced fire activity due to increased human presence, improved fire prevention measures, and swift responses to fire incidents. Conversely, the peak in fires occurring between 450 and 600 miles may be shaped by ecological factors, including

vegetation types, weather patterns, and topography. Certain ecosystems within this range may be more susceptible to natural or lightning-caused fires, resulting in a heightened frequency of fire incidents.

2.
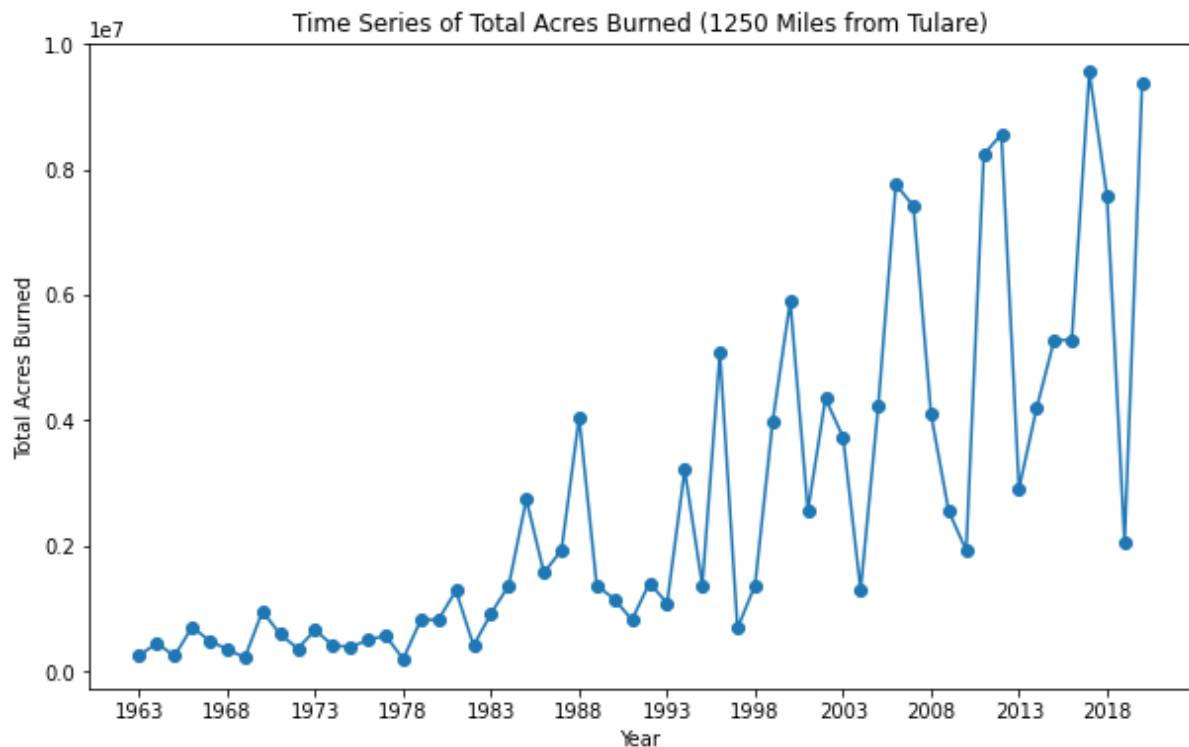
The visualization above is a time series graph depicting the total acres burned per year for fires occurring within a 1250-mile radius from Tulare. This graph serves as a powerful tool to comprehend the dynamics of wildfires in the region over a specific time frame.

The x-axis represents the years, while the y-axis signifies the total acres burned for each respective year. This straightforward arrangement allows viewers to easily grasp the relationship between time and the extent of wildfires.

The underlying data for this time series plot originates from the original fire dataset. To create this visualization, the dataset was carefully filtered to encompass only those fires that fell within a distance of up to 1250 miles from Tulare and occurred within the time frame of 1963 to 2023. It's worth noting that there's an absence of data post-2020 meeting both criteria. The filtered data was then grouped by year, and the total acres burned for each year were calculated, culminating in the construction of this time series plot. To enhance readability, the x-axis was configured to display years at 5-year intervals.

Analyzing the graph, a distinct trend emerges. There has been a substantial increase in the total acres burned as the years progress. In fact, the total acres burned in 2020 is nearly 10^7 times greater than that in 1963. While there are occasional fluctuations and temporary declines in between, the overarching pattern is prominent.

Several factors contribute to this alarming trend. Rising temperatures and prolonged droughts create conditions favorable to wildfires, increasing their likelihood and intensity. The introduction of non-native, invasive plant species can alter ecosystems, making them more

susceptible to fires. Additionally, population growth and urbanization can lead to more accidental ignitions and fires in the wildland-urban interface.

3.
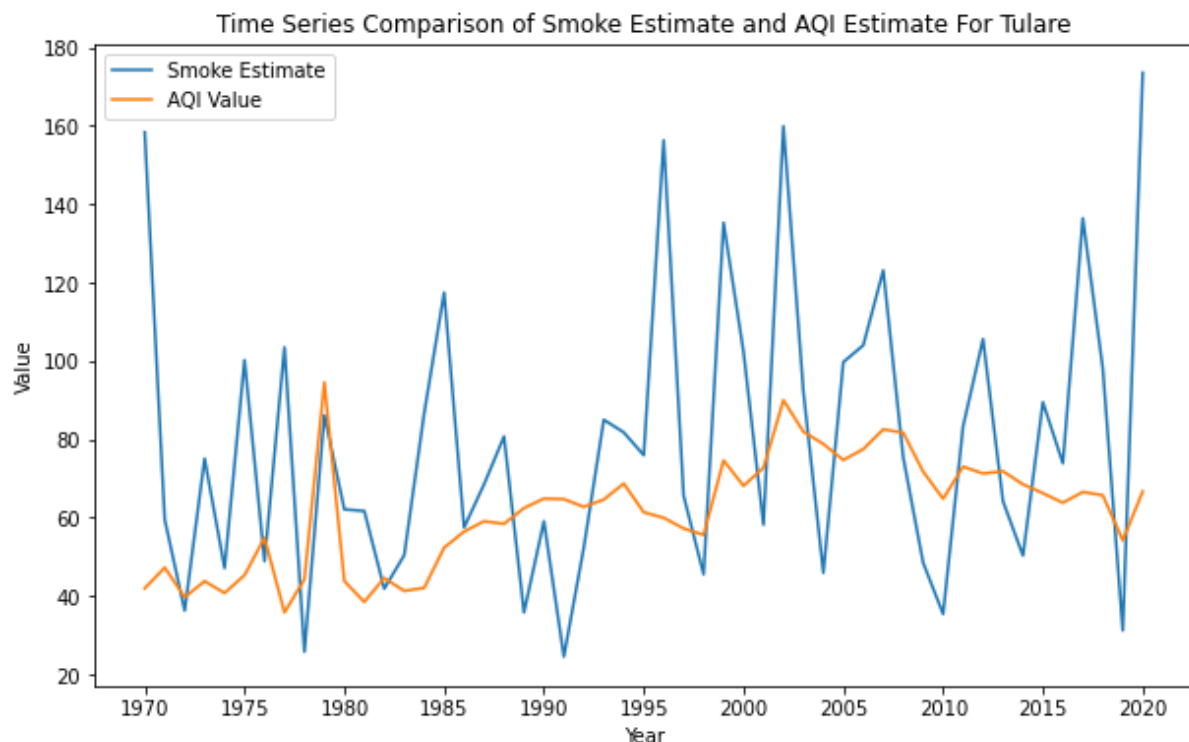
The plot presented represents a time series graph illustrating the fire smoke estimate and the Air Quality Index (AQI) estimate for the assigned city, Tulare. On the y-axis, or the vertical axis, we have the fire smoke estimate and the AQI value, while the x-axis, or the horizontal axis, represents the years.

In this visualization, two trend lines are depicted. The blue trend line represents the fire smoke estimate, which was generated through a data analysis process. On the other hand, the orange trend line represents the AQI value over the years, which is provided by the US Environmental Protection Agency (EPA).

To create this visualization, two sets of data were used. The first dataset involved the fire smoke estimate, which was calculated by filtering fires that occurred within 1250 miles from Tulare between the years 1963 and 1970. The fire estimate formula considered various factors, such as fire types, fire size, and the distance of the nearest point of the fire from Tulare. The estimates for all qualifying fires were summarized, and the mean of these estimates was considered as the fire smoke estimate for each year. The second dataset involved the AQI values for gaseous and particulate pollutants in Tulare County, obtained through an API. These values were stored in separate CSV files and subsequently combined into a single dataset. The greater AQI value, whether from gaseous or particulate pollutants, was considered as the AQI for a particular day. The mean of all daily AQI values was calculated to obtain the AQI for each year.

One crucial point to note is the difference in data availability. The fire smoke estimate data covers the period from 1963 to 2020, while the AQI data provided by the US EPA only extends from 1970 to 2020. Consequently, the plot displays data starting from 1970 for both the fire smoke estimate and the AQI estimate.

Upon examining the plot, it is evident that there is no distinct trend observable for the fire smoke estimate. However, the AQI estimate demonstrates a slight upward trend. This suggests that the AQI value has been gradually increasing over the years. According to AQI standards, a range of 0-50 is considered "good" air quality, while values exceeding 50 are categorized as "moderate". Furthermore, a correlation value of 0.27 is reported for the obtained AQI values and the fire estimate that I created.

Reflection Statement:

In the realm of Data Science and human-centered design, collaboration emerges as a pivotal asset, significantly enhancing our ability to fine-tune our approaches. In this particular task, marked by its intricate nature and the limited knowledge available on the subject, collaboration played a pivotal role. Engaging in a collective cohort setting, we exchanged a wealth of ideas, designs, approaches, and algorithmic considerations, leading us towards a more robust and insightful solution. The iterative process of trying multiple algorithms and metrics allowed us to conduct a detailed analysis and ultimately arrive at the final visualization.

One of the most critical takeaways from this assignment is the significance of thorough data cleaning. Collaboration exposed the potential pitfalls of data gaps and outliers, prompting us to address these issues effectively. For instance, upon close inspection of the fire data, we discovered a small number of curveRings instances, accounting for just 35 out of 135,000 data points. Given their limited impact on the results, we opted to exclude them, highlighting the importance of carefully navigating data nuances. As discussed in class, the process of datafication can inadvertently overlook certain subtleties due to scope and data collection considerations. Given that our data is in the form of a time series, we incorporated data smoothing techniques by utilizing moving averages to address outliers effectively.

Addressing the research question, several specific insights regarding wildfires and their impact on Tulare surfaced. One standout finding was the alarming upward trend in total acres burned within a 1250-mile radius of Tulare. Over the past six decades, we observed a substantial increase, with the total acres burned in 2020 nearly 10^7 times greater than in 1963. This trend sheds light on the influence of climate change, prolonged droughts, invasive species, and urbanization, collectively intensifying wildfires in the region.

Another noteworthy observation emerged from the analysis of the fire smoke estimate and the Air Quality Index (AQI) for Tulare. The fire smoke estimate exhibited no clear trend, whereas the AQI estimate displayed a gradual upward trajectory over the years. This suggests that the AQI values have been slowly transitioning from "good" to "moderate" air quality in Tulare, underscoring the long-term impact of wildfires on the local air quality. This finding underscores the urgency of implementing measures to mitigate the effects of smoke on public health and the environment.

Furthermore, the exploration of the spatial distribution of wildfires, as depicted in the histogram visualization, unveiled significant insights. Notably, there were fewer fires within the initial 100 miles from Tulare, likely attributable to urban wildland interface (UWI) zones with enhanced fire prevention measures. In contrast, a notable spike in the number of fires occurred between 450 and 600 miles from Tulare, potentially driven by ecological factors like vegetation types and weather patterns. This finding underscores the need for region-specific fire management strategies, taking into account the varying fire frequencies across different distances from the city.

I had initially explored traditional approaches to understanding the patterns of wildfires, but through collaboration with Prerit, I became familiar with the concept of geospatial data analysis, which opened up a new dimension in our research. Prerit's expertise in geospatial data analytics enabled us to explore the spatial distribution of fires and their impact on Tulare more comprehensively. Simultaneously, my collaboration with Ishank Vasania proved to be instrumental in refining our understanding of air quality and its relationship with wildfires. Initially, I had been considering the use of linear regression for prediction. However, it was through discussions with Ishank that the idea of employing an ARIMA model emerged. This shift from linear regression to the ARIMA model marked a significant turning point. It allowed me to embrace the temporal dynamics of wildfire and air quality data, enabling more accurate predictions for future years. The decision to adopt this time series model was a direct result of collaborative brainstorming and the insights gathered from my cohort, ultimately enhancing the quality and precision of the research outcomes.