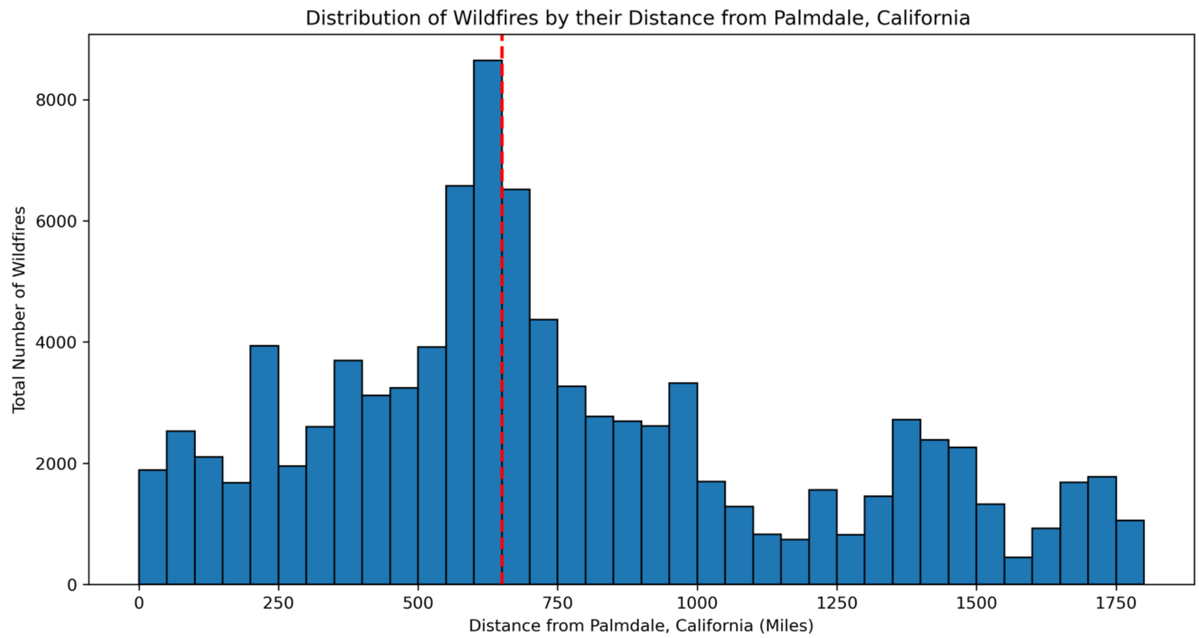
Data 512 – Project Part 1 : Common Analysis

Baisakhi Sarkar – MSDS 2023-25, UW, Seattle

Explanation of Visualizations

1. Distribution of Wildfires by Distance



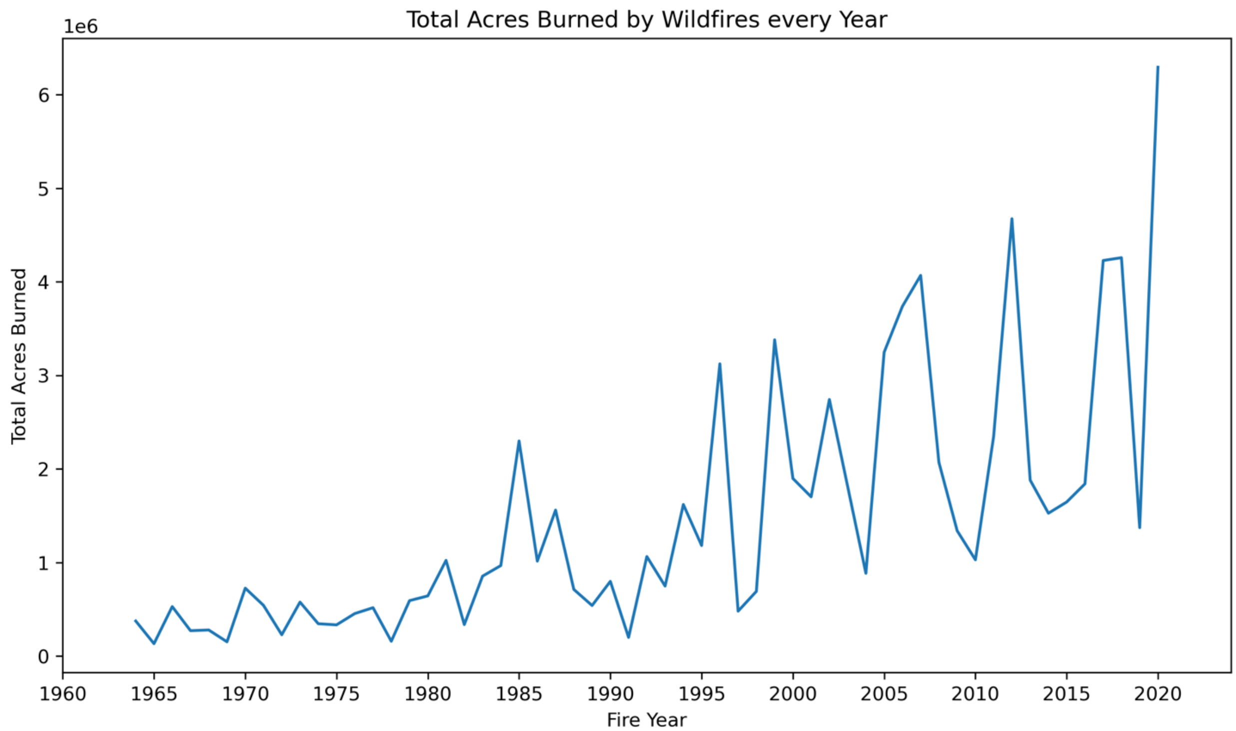
What the figure shows:
This histogram displays the **distribution of wildfires** occurring at various distances from **Palmdale, California**, grouped into **50-mile intervals** up to **1,800 miles**. It provides insight into how wildfire frequency varies with increasing distance from the city. The **red vertical line at 650 miles** marks the cutoff used to filter the wildfires that were included in the analysis to estimate **smoke impacts**.

How to read the figure:
The **x-axis** represents the **distance from Palmdale, California** in miles, divided into 50-mile intervals. Each bar corresponds to the total number of wildfires that occurred within that distance range. For example, if the tallest bar is in the **600-650 mile range**, it indicates that the **largest number of wildfires** occurred within this interval. The **y-axis** represents the **total number of wildfires** in each interval. The **red vertical line at 650 miles** indicates the distance beyond which wildfires were excluded from the smoke impact analysis, focusing on fires that could have a more direct influence on air quality.

Why this figure is relevant:
This visualization is critical to **understanding the geographical distribution** of wildfire activity near Palmdale, CA, which helps determine which fires are likely to impact **local air quality**. The concentration of wildfires within specific distance ranges can inform **fire management strategies** and **mitigation efforts**.

Underlying data and processing:
The data for this plot is sourced from the **Combined Wildland Fire Datasets for the United States and certain territories, 1800s-Present**. This dataset was **filtered to include only wildfires** within **1,800 miles of Palmdale**. However, for smoke impact analysis, only **wildfires within 650 miles** were selected to focus on **relevant fires** that are more likely to impact air quality in Palmdale. The histogram helps to identify where most wildfires occur within this range, informing future predictions and environmental planning.

2. Total Acres Burned per Year (1964-2020)



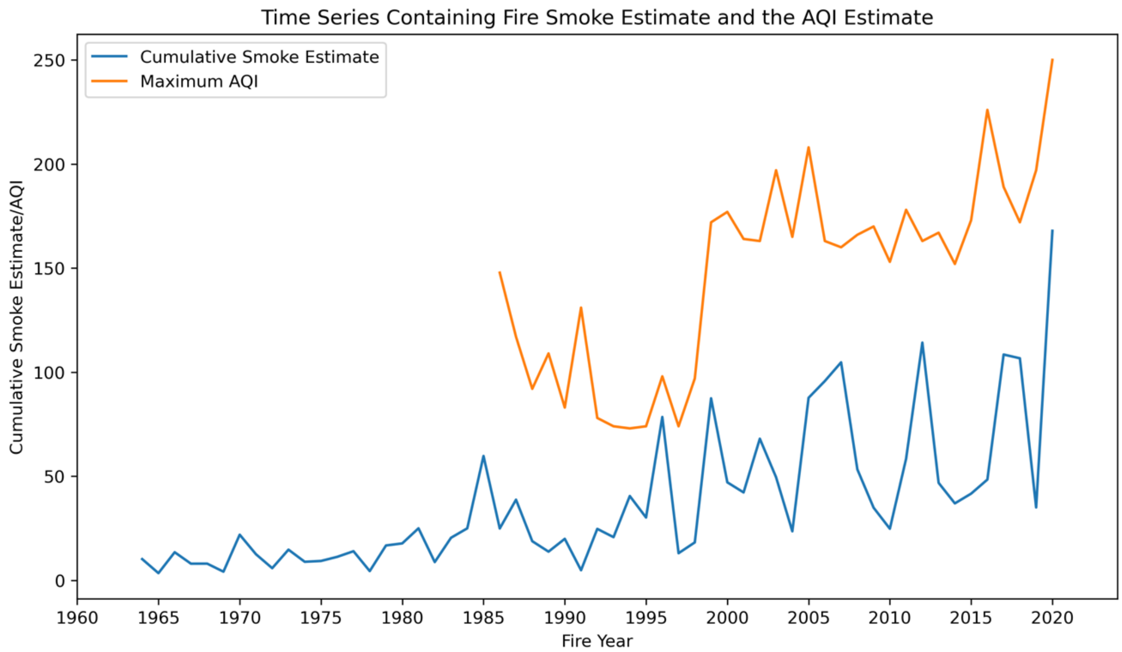
What the figure shows:
This **time series plot** illustrates the **total acres burned each year** by wildfires within **650 miles of Palmdale, California**, covering the period from **1964 to 2020**. The plot highlights fluctuations in wildfire activity, with peaks indicating years of significant fire events and dips corresponding to relatively quiet fire seasons. It provides insight into how wildfire activity has evolved over the past six decades and reflects the **increasing severity** and **frequency of fires** over time.

How to read the figure:
The **x-axis** represents the **years**, while the **y-axis** shows the **total acres burned** in those years. Each point on the plot corresponds to the **total area burned in that particular year**. **Peaks** in the plot indicate years with **high wildfire activity** and larger burned areas, while **dips** suggest **calmer fire seasons** with fewer or smaller wildfires. For instance, a sharp peak might indicate a year with multiple or large fires, while flat or declining periods suggest seasons with limited fire activity. This plot helps identify **patterns and trends** over time, such as whether fire activity is increasing, decreasing, or fluctuating unpredictably.

Why this figure is relevant:
This time series is essential for understanding the **temporal trends** in wildfire activity. Identifying years with **extensive wildfire damage** can help in **evaluating environmental changes** and **policy interventions** that might have impacted fire behavior over time. For example, clusters of high fire activity years could prompt **further analysis** of contributing factors, such as **climate conditions, forestry management practices**, or **urban development**. Moreover, by observing these trends, policymakers and environmental managers can **predict future fire risks** and **allocate resources** accordingly to mitigate fire damage and protect public health.

Underlying data and processing:
The data for this plot was sourced from the **Combined Wildland Fire Datasets for the United States and certain territories, 1964-Present**. It was filtered to include only **wildfires within 650 miles** of Palmdale to focus on those most likely to impact the region.

3. Smoke Estimate vs. AQI Comparison (1964-2020)



What the figure shows:
This **line plot** compares the **calculated smoke estimate** with the **maximum AQI values** for each year between **1964 and 2020**. The plot visually explores the relationship between **wildfire activity and air quality** by placing both metrics side by side. The **smoke estimate** reflects the impact of wildfires on air quality, while the **AQI values** capture the highest pollution levels recorded during the **fire season** (May 1st to October 31st). This comparison highlights **years with severe wildfire activity** and their corresponding **impact on air quality**.

How to read the figure:
The **x-axis** represents the **years** (1964-2020), and the **y-axis** displays the **smoke estimate** and **maximum AQI values**. Each year has two corresponding data points on the plot—one for the **smoke estimate** and one for the **AQI**. By following the trends over time, we can identify **correlations** between **high smoke estimates** (indicating extensive wildfire activity) and **elevated AQI values** (indicating poor air quality). A **year with both high smoke estimates and high AQI** suggests that wildfire activity was a major driver of poor air quality during that year. For example, if both metrics peak in the same year, it signals a **direct impact of wildfire smoke on air pollution**.

Why this figure is relevant:
This visualization provides a **critical comparison** between two key metrics: the **smoke estimate** and **air quality index (AQI)**. Understanding how **wildfire smoke affects air quality** is essential for **public health and policy decisions**. If years with high smoke estimates consistently align with **high AQI values**, it reinforces the need for **mitigation strategies** to minimize health risks during fire seasons. This comparison also helps in evaluating the **accuracy** of the smoke estimate model by validating it against real-world AQI data.

Underlying data and processing:
The **smoke estimates** were calculated using wildfire data, with a focus on **fire proximity, size, and duration**. Fires that were **closer and larger** contributed more to the smoke estimate than smaller, more distant fires. The wildfire data was sourced from the **Combined Wildland Fire Datasets for the United States and certain territories.** This data was filtered to include only **wildfires within 650 miles** of Palmdale, CA, to capture fires most likely to affect local air quality.

The **AQI data** was obtained through the **EPA's AQS API**, which provides historiscal air quality measurements. For each fire season (May 1st to October 31st), the **highest AQI value** was selected to represent that year's air quality. This approach focuses on capturing **extreme pollution events**, which are often driven by wildfire smoke.

Reflection Statements

This assignment provided me with valuable insights not only into the **wildfire analysis and modeling process** but also into the **benefits of collaboration**. One of the most significant things I learned was the importance of **shared discussions** when working with **complex datasets** and **unfamiliar tools**, such as **GeoJSON files** and **geospatial analysis techniques**. The collaborative environment encouraged me to **explore new methods**, refine my approach, and troubleshoot issues more efficiently.

At the beginning of the assignment, I faced several challenges, such as **understanding GeoJSON structures** and calculating **geodetic distances** to filter fires based on proximity to my assigned city, **Palmdale, California**. Through discussions with my peers on platforms like MS Teams or in person discussions, I gained a better understanding of how others were handling similar challenges. For instance, a conversation around **handling curveRings** in the fire polygons dataset helped me adjust my code to **read and process GeoJSON files** correctly. This exchange saved me significant time and gave me a clearer understanding of **geospatial data processing**.

Given my **interest in sustainability**, I found the project **fascinating and rewarding**. It felt like **detective work**—uncovering patterns in wildfire behavior and understanding how they influence air quality. The process of **building the smoke estimate from scratch**, **experimenting with different features**, and **optimizing the model** was both **tedious and exciting**.

Another key takeaway was learning to **experiment with different models** for time series forecasting. The collaborative discussions helped me realize that multiple approaches could be applied to the same problem. While I initially focused on **LSTM models**, conversations with my peers introduced me to the **SARIMAX model**, which turned out to be more effective in capturing the peaks and dips in the smoke estimate data. This experience reinforced the importance of **keeping an open mind** and being willing to pivot to new strategies when necessary.

Although I did not directly reuse any code from my peers, the **guidance provided by Dr. David W. McDonald** was instrumental in shaping my approach. I used his **'wildfire' module** for reading wildfire data iteratively and employed his **geodetic distance calculation code** to filter the dataset based on proximity. Additionally, I followed his **sample code for API requests** to retrieve AQI data from the **EPA AQS API** and aggregate it into annual summaries.

One of the most helpful aspects of this collaborative environment was the **unintentional collaboration** that emerged through **discussion threads**. Even when I didn't actively participate in a discussion, reading through the responses and solutions posted by others gave me new ideas for **solving my own challenges**.

Overall, this assignment demonstrated the value of **collaboration in independent projects**. While each person worked on their assigned city individually, the collaborative environment fostered a **sense of community and shared learning**. It allowed me to **learn from others' experiences** and **improve my project through informed experimentation**. This assignment also taught me the importance of **attribution** and **academic integrity**, as I ensured that all borrowed code, methods, and techniques were properly credited.

In conclusion, this assignment has significantly improved my skills in **geospatial analysis, time series forecasting, and collaboration**. It has shown me that **even individual projects benefit from peer support**, and that sharing ideas can lead to **better problem-solving** and **more effective outcomes**.