Analyzing California Wildfires: Patterns, Trends, and Insights

Yu-Ching Huang (jhuang13@usc.edu, 6413493088)

I. Research Question and Purpose

What are the temporal and geographic trends of wildfires in California between 2015 and 2024, and how do factors such as location, percent contained, and extinguished duration correlate to wildfire severity?

California wildfires have become increasingly devastating over the past decade, with far-reaching impacts on ecosystems, communities, and economies. The purpose of this project is to understand the underlying trends and patterns of wildfire incidents to gain actionable insights into their causes, behavior, and management effectiveness.

This project focuses on analyzing California wildfire data from 2015 to 2024 to identify trends and patterns in wildfire activity. The dataset includes information such as fire names, start and extinguished dates, acres burned, and location data. By visualizing and analyzing these metrics, this project aims to:

- Analyze trends, including changes in the frequency or severity of wildfires over time.
- Identify geographic regions most prone to wildfire activity.
- Evaluate the effectiveness of containment strategies and the time required for extinguishment.

II. Data Collection and Method

The California wildfire incident data is collected via API (https://www.fire.ca.gov/incidents). Since the project focuses on the past 10 years, I specified the additional API parameter ‘year’ to retrieve data from 2015 to 2024. After cleaning the raw dataset retrieved using API, I obtained 10-year California wildfire data with 10 columns (Name, Started, County, Location, AcreBurned, PercentContained, Longitude, Latitude, ExtinguishedTime), a total of 2670 data samples.

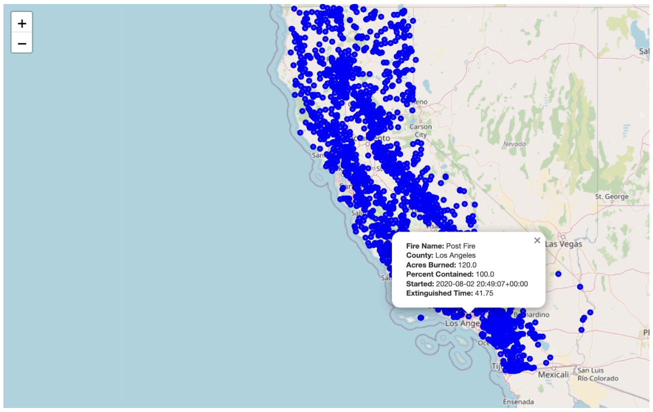
III. Data Cleaning and Preprocessing

To analyze the trend from 2015 to 2024, I added a new column ‘Year’ by extracting the year from the ‘Started’ column and calculating ‘ExtinguishedTime’ to obtain the duration of the fire, and upon reviewing the dataset, some missing values and inaccuracies have been identified:

- Year: Dropped the samples where the data is misentered (i.e. Year is not in the range 2015-2024).
- ExtinguishedTime: Filled missing values with average duration of the overall dataset[1].
- County: Removed rows with missing County values, as they were few in number. Additionally, cleaned the County column to exclude city names, retaining only the county names.
- AcresBurned: Filled missing values in the AcresBurned column with the overall average.
- PercentContained: Filled missing values in the PercentContained column with the overall average.

[1] Compared with other columns with missing values, ‘ExtinguishedDate' column has relatively more missing values, and thus resulted in missing values in the calculated column ‘ExtinguishedTime’. I decided to retain them by filling missing values with mean value as they may still provide valuable insights for analysis.

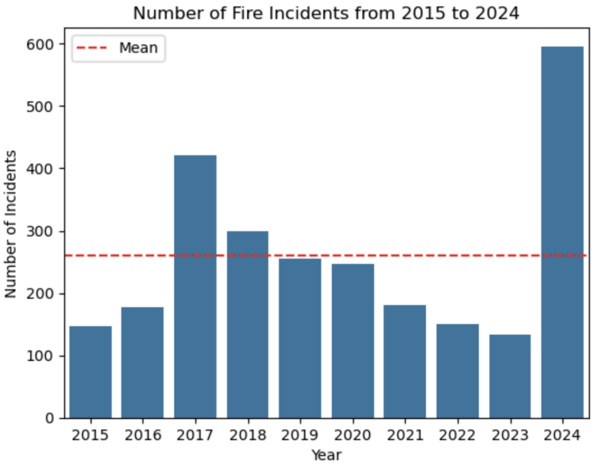
IV. Data Analysis and Visualization[2]

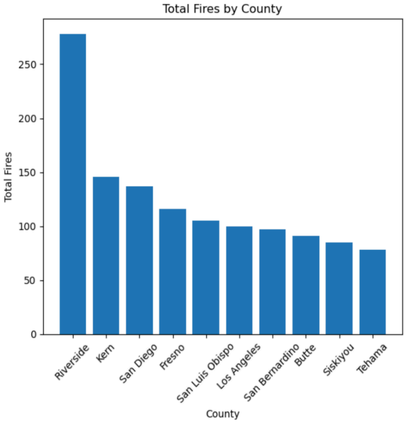


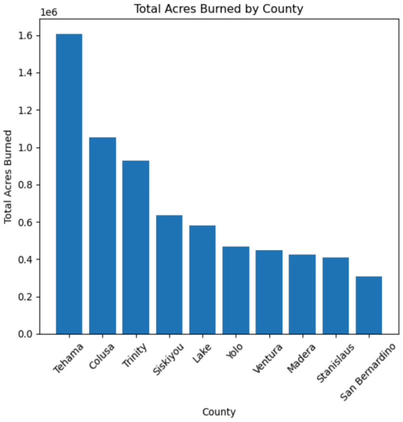
Fire Name: Post Fire
County: Los Angeles
Acres Burned: 120.0
Percent Contained: 100.0
Started: 2020-08-02 20:49:07+00:00
Extinguished Time: 41.75

The figure[3] on the left is an interactive map of the wildfire with hover-over information, which provides an overview of the locations where the wildfires took place. Click on the points on the map to show you some information about the incidents, for example: Name of the fire, county where the fire takes place, acres burned, just to name a few. The map indicates that there isn't a specific location where fires typically occur.

The bar chart on the right displays the number of fire incidents from 2015 to 2024. As shown, there is a noticeable fluctuation in the number of fire incidents across the years. A significant spike in fire incidents occurred in 2017 and 2024, with 2024 having the highest recorded number of incidents. The data suggests an irregular pattern of fire incidents, with some years experiencing extreme spikes. Further investigation could explore what factors contributed to these high or low years (e.g., weather conditions, policy changes, or mitigation efforts).







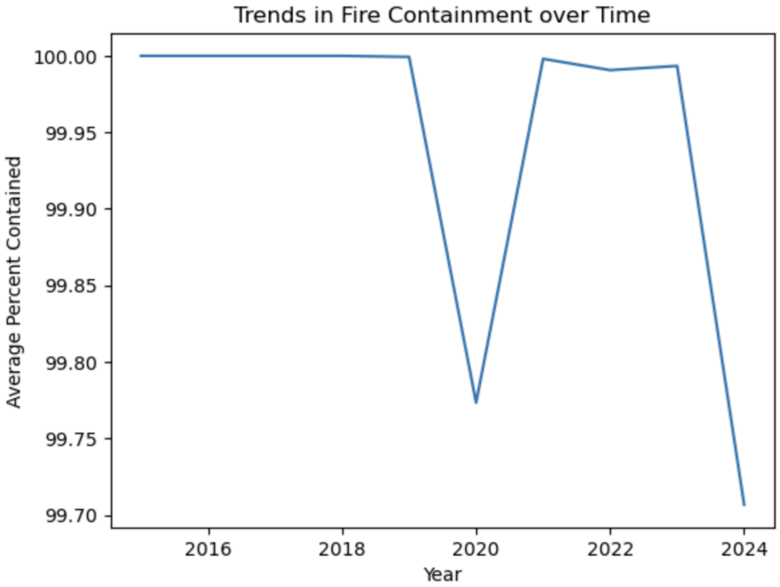
From the two graphs, we can observe that counties like Riverside, Kern, and San Diego have the highest number of fire incidents, indicating that these areas are prone to frequent wildfires. These counties are likely located in more densely populated regions or urban-adjacent areas in Southern California, where human activities and the dry climate might contribute to the elevated fire risks[4]. On the other hand, counties like Tehama, Colusa, and Trinity have the highest total acres burned, suggesting that while fires in these

[2] I believe that visualization is an integral part of the analysis process, so I combined the two Python scripts (run_analysis and visualize_resutls) into one single file: analyze_visualize_results.py

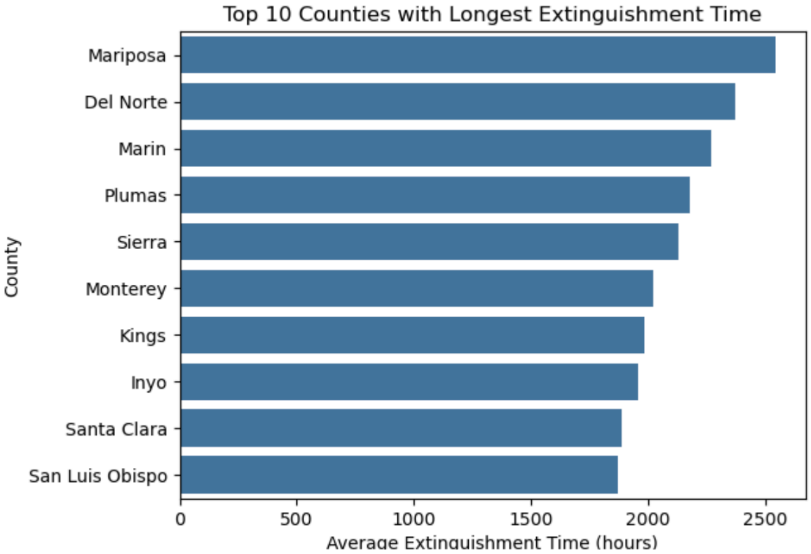
[3] For the complete version of the visualizations, go to results > final_project_viz.ipynb.
For visualization files only, go to results > visualizations.

[4] https://wildfiretaskforce.org/southern-california-regional-profile/
https://www.kqed.org/science/1994311/southern-california-wildfires-are-so-intense-theyre-creating-their-own-weather

regions may be less frequent, they tend to be more catastrophic in scale. These counties are predominantly located in Northern or Central California, characterized by rural or forested landscapes where fires can spread extensively due to the availability of continuous vegetation and challenging terrain[5]. Overall, Southern California appears to experience more frequent but smaller fires, while Northern and Central California suffer from fewer but larger-scale wildfires.

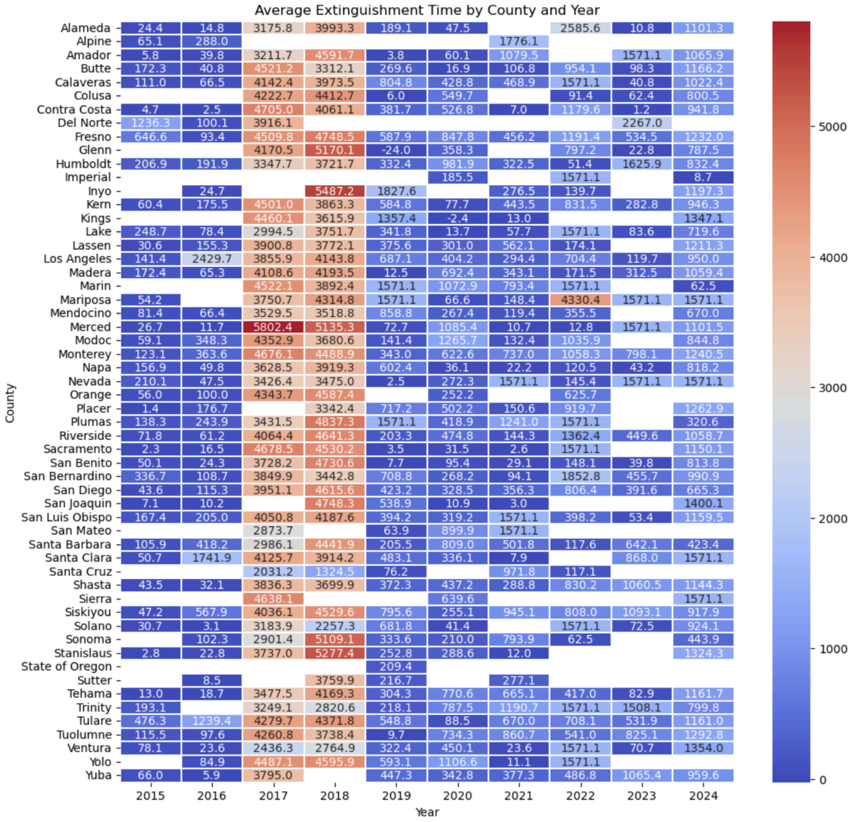


The graph shows that fire containment rates are consistently high, averaging close to 100% across most years, indicating effective fire management overall. However, there are sharp drops in 2020 and 2024, where average containment dips to around 99.7%. These anomalies might indicate challenges such as larger or more uncontrollable fires, unusual weather conditions, or resource constraints during these years. Alternatively, they could reflect missing or incomplete data. Further investigation into fire records, regional contributions, and data quality for these years is recommended to identify the underlying causes.



The bar chart on the right highlights the top 10 counties with the longest average extinguishment times for fires, with Mariposa County having the highest average. However, the results are affected by missing data, which was replaced with the overall average, potentially distorting the true distribution. Counties with fewer actual data points may have artificially leveled averages. The high extinguishment times in some counties could indicate larger or more complex fires, challenging terrains, or resource limitations.

[5] https://www.nationalgeographic.com/environment/article/climate-change-california-wildfire



The heatmap reveals significant disparities in average fire extinguishment time across counties and years. Counties like Mariposa, Del Norte, and Plumas consistently exhibit longer extinguishment times, suggesting challenges like larger fires, difficult terrain, or resource limitations. Temporal spikes in extinguishment times, such as in Fresno (2020) and Inyo (2024), may indicate exceptional fire events or operational constraints. Conversely, counties like Alameda and Contra Costa show shorter times, possibly due to smaller fires or better resources. However, the uniform values in some cells (i.e., 1571.1) reflect missing data replaced with averages, which may affect reliability. Further investigation into outliers and improved data imputation are recommended to enhance insights and guide fire management strategies.

This project analyzed California wildfires from 2015 to 2024 to uncover patterns in fire activity, containment, and extinguishment times. Southern California counties, like Riverside and San Diego, experienced frequent but smaller fires, while Northern counties, such as Mariposa and Plumas, faced fewer but larger fires due to dense vegetation and challenging terrain. Containment efforts were highly effective, averaging nearly 100%, though anomalies in 2020 and 2024 highlighted challenges with larger or more complex fires. Extinguishment times varied significantly, with Mariposa and Plumas showing prolonged durations, likely due to terrain and fire scale, while counties like Alameda demonstrated greater efficiency. Despite valuable insights, missing data required imputation, which may have affected some results. These findings underscore the need for improved data collection, targeted resource allocation, and localized strategies to enhance wildfire management.

IV. Future Work and Improvements[6]

There are several directions to expand and improve this project. One potential direction is to train predictive models to forecast key wildfire metrics such as severity, acres burned, and extinguishment times. For instance, regression models could predict the size of fires or the time required for containment, while classification models could assess fire risk in specific regions. These models could incorporate features like weather conditions, historical fire data, and proximity to urban areas. Time series forecasting could also be applied to anticipate future trends in fire incidents.

Another significant improvement involves integrating additional data sources to enrich the analysis. Including environmental data such as temperature, humidity, wind speed, and drought indices could provide valuable insights into fire behavior. Topographical information like elevation and vegetation type could help explain terrain challenges, while socioeconomic data, including population density and human activity, could reveal patterns related to fire causation. Satellite data, such as vegetation health and burn areas, would add a spatial perspective to the analysis.

Future work could also focus on quantifying the economic and environmental impacts of wildfires, such as property damage, firefighting costs, carbon emissions, and biodiversity loss. Additionally, regional resource optimization models could recommend the allocation of firefighting resources based on fire risk and historical patterns, ensuring efficient preparedness and response.

[6] I skipped the **‘Changes from Original Proposal’** section because the project has no changes from the original proposal.