PSTAT 100 Final Report: California Wildfires 1969-2021

Amy Nguyen, Tess Ivinjack

Author Contributions

1. Amy worked on tidying the data, performing the initial codes, created visualizations, writing up findings, and curating final draft of the report.
2. Tess studied the CalFire documentation/background, prepared wrote the data descriptions and background information, and created visualizations for the exploratory analysis.

Abstract

As the climate increases and gets more dangerous with each day, wildfires become a more immediate threat to the people of California. In an effort to understand them and their patterns better, we have decided to analyze the CAL FIRE data from every recorded wildfire in the state of California dating back to 2009. In this project, we will discover more about what time of the year California wildfires are the most likely to occur, and when they have been the most popular in the past. In addition, we investigate the relationship between how long the fire burned for and the total acreage it burned. This project has led us to conclude that the summer months are when California is most susceptible to wildfires, as well as 2017 being the most popular year for wildfires due to wind patterns. Finally, we concluded that there is a positive relationship between the duration of the fire and how much land it burned.

Background

Wildfires have been a prevalent issue for the United States, specifically California, for hundreds of years. Wildfires can be extremely destructive, causing the state of California billions of dollars, loss of millions of acres, thousands of structures destroyed, and fatalities. They have become a really serious issue because of climate-change induced drought/unusual heat, but most of the time wildfires are human caused. The California Department of Forestry and Fire Protection (CAL FIRE) is a government agency that responds to wildland fires, structure fires, hazardous material spills, and all other kinds of medical emergencies. CAL FIRE also collects all the data from every reported incident, and our dataset we will be working with for this project is comprised of that data. It includes information about each wildfire's location, duration, how many acres it burned or structures it damaged, and much more.

We will be investigating the relationship between reported California wildfires and their time and location to further learn more about what time periods/areas are more prone to wildfires of different sizes. Going to school in California, we have been faced with the consequences of wildfires first hand, so this topic is one that we are interested in learning more about because it affects our everyday life. We hope to discover more about this current problem for our environment and provide a good summary of wildfire patterns in California.

Aims

The first question we wanted to answer with our data analysis was what months is California the most susceptible to wildfires? Along the same lines, we also wanted to know what years have been the most popular for wildfires. To answer these questions, we made visualizations displaying the number of wildfires by month and by year to give us which ones were the most frequent. We discovered that they hit their peak in June and July then start to decrease around August, and that 2017 was the most popular year for wildfires. We also did some extra analysis with the relationship between time and the severity of the wildfires.

The next topic we were curious about was the relationship between how long the fire burned for and how many acres it had burned? Using trend lines, we were able to see that there is a slight, positive correlation between a fire's duration and the total acres burned. We also explore a little deeper by researching outliers in our data, and seeing how the time of the year affects this relationship as well.

Datasets

The data are about California wildfire incidents from 1969 to 2021. It describes where and when they occurred, as well as how much support and materials went into extinguishing them. It was collected by CalFire, which is the California Department of Forestry and Fire Protection (https://www.fire.ca.gov). CalFire collects their data from the Computer Aided Dispatch (CAD) system and the National Incident Management Situation Report. It is collected on a weekly basis, but then at the end of each fire season they finalize the data in their Final Fire Season Reports, which is what our dataset consists of.

The relevant population is all incidents of wildfires in California, and the sampling frame is all reported incidents to CalFire and partnering agencies in California reporting wildfire activity between 1969 to 2021. This is administrative data and the sampling mechanism is a census because the dataset includes all reported incidents to CalFire, so consequently, there is no scope of inference.

The original dataset consisted of 23 variables and 1910 observations, where a single observation represents an incident of a California wildfire reported to CalFire or other partnering agencies in this time period. Our tidied dataset consists of 1910 observations and 6 variables. A list of these variables and their respective descriptions is summarized in **Table 1** below.

> **Table 1**: variable descriptions and units for each variable in the dataset.

Name	Variable description	Type
Incident Name	Name of Wildfire	String
County	County where fire started	String
Acres Burned	Total Acreage consumed by the fire	Numeric
Date Created	Fire start date	Datetime
Date Entinguished	Fire extinguished date	Datetime
Duration Days	Duration of fire in days	Numeric

In tidying the data, we converted `Date Created` and `Date Extinguished` in `datetime` format in order to calculate the resulting column `Duration Days`.

> **Table 2**: example rows of tidied dataset `fire`

Incident Name	County	Acres Burned	Date Created	Date Extinguished	Duration Days
Bridge Fire	Shasta	37.0	2017-10-31	2018-01-09	70.0
Pala Fire	San Diego	122.0	2009-05-24	2009-05-25	1.0
River Fire	Inyo	406.0	2013-02-24	2013-02-28	4.0
Fawnskin Fire	San Bernardino	30.0	2013-04-20	2013-04-22	2.0

Methods

The focus of exploratory analysis was targeted on understanding trends in California wildfires, both over time (between 1969-2021) and on a monthly basis. In order to further investigate, years and months for both `Date Created` and `Date Extinguished` were extracted to analyze which months and years were most frequent in wildfire incidents. Because we are also interested in the severity of these wildfires, the data was categorized into 3 different quantiles, `1–low`, `2–medium`, and `3–high` to quantify severity relative to acres burned. Using histogram visualizations, we were able to identify the month and year where there were the most reported California wildfire incidents, and subsequently, also revealed wildfire distributions both on an annual basis and in retrospect from 1969-2021. We continued our analysis by visualizing the average acres burned for each year, suggesting that frequency of annual wildfires and acres burned are independent of one another. Finally, we plotted `Acres Burned` as a function of `Duration Days` to investigate whether or not these variables exhibited a relationship with one another, and additionally if the time of year played a role in the how long or how severe the wildfire was.

Results

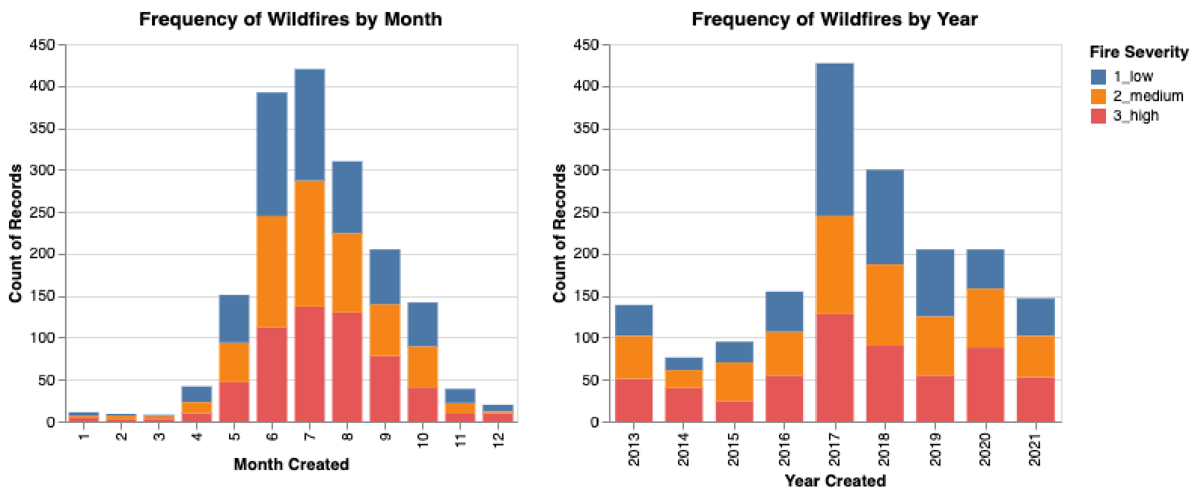
Frequency of California Wildfires

Based on our findings displayed in **Figure 1**, California Wildfires are most frequently recorded in July, and is closely followed by June, August, and September. Both the severity of the wildfire (in regards to acres burned on a low, medium, high ranking) and the frequency of the wildfires on a monthly basis appear to follow a bell-shaped normal curve with perhaps a slight left-skew. This is intuitive because these summer/fall months provide warm, dry weather, creating the perfect breeding ground for wildfires, and as the months get colder, wildfires are less frequent. Taking a look at the frequency of reported wildfires by year, 2017 exceeds all other years with just under 450 reported incidents.

An article on the 2017 fires by the *LA Times*, notes that paired with California's drought, "This summer was the hottest ever recorded in California, allowing for new vegetation to dry up."

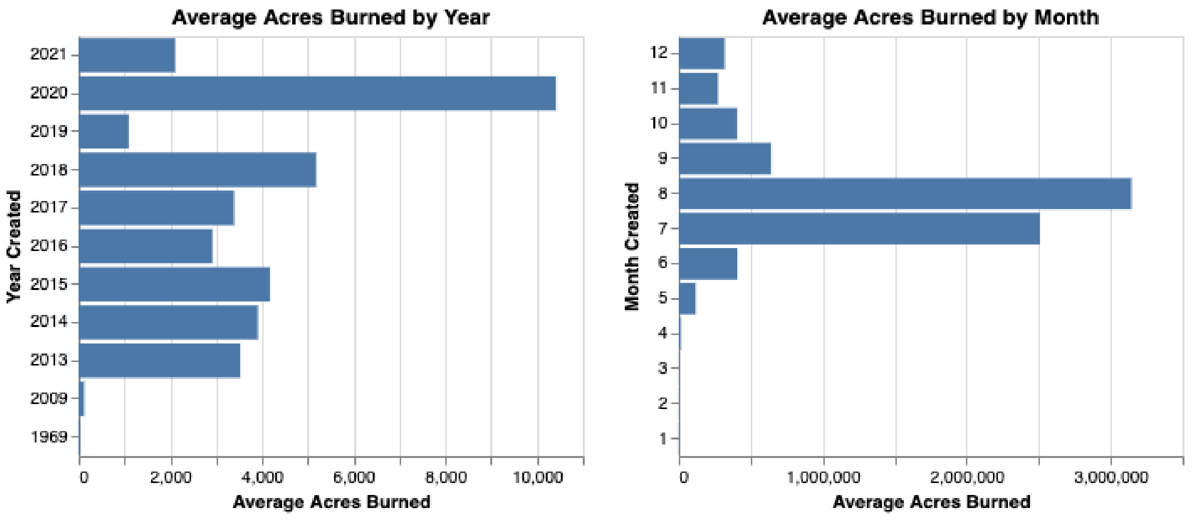
> “Why the 2017 Fire Season Has Been One of California's Worst.” Los Angeles Times, *Los Angeles Times*, https://www.latimes.com/projects/la-me-california-fire-seasons/.

> **Figure 1**: Frequency of reported California Wildfire incidents by year and by month, where severity in terms of acres burned is indicated by color. Wildfire incidents follows a bell-shaped curve and on a monthly basis is slightly left-skewed. Historically, the most reported CA wildfires occur in July, and the year with the most reported wildfires is in 2017.



What's interesting though is, when observing the average acres burned by each month and year shown in **Figure 2** below, 2020 very clearly surpasses all of the other recorded years, including 2017, with on average nearly 6,000 more acres burned than 2017. Additionally, on average August burns more acres than July, so while there may not be as many occurences of wildfires recorded, when there are fires, they are incredibly destructive. These findings are critical to our analysis because it showcases that count of incidents and acres burned are not entirely dependent on one another. Even though 2017 had the most occurences of CA wildfires that year, 2020 had on average far more burned acres and what we would consider to be a more destructive year since the amount of land destoyed exceeded all other years.

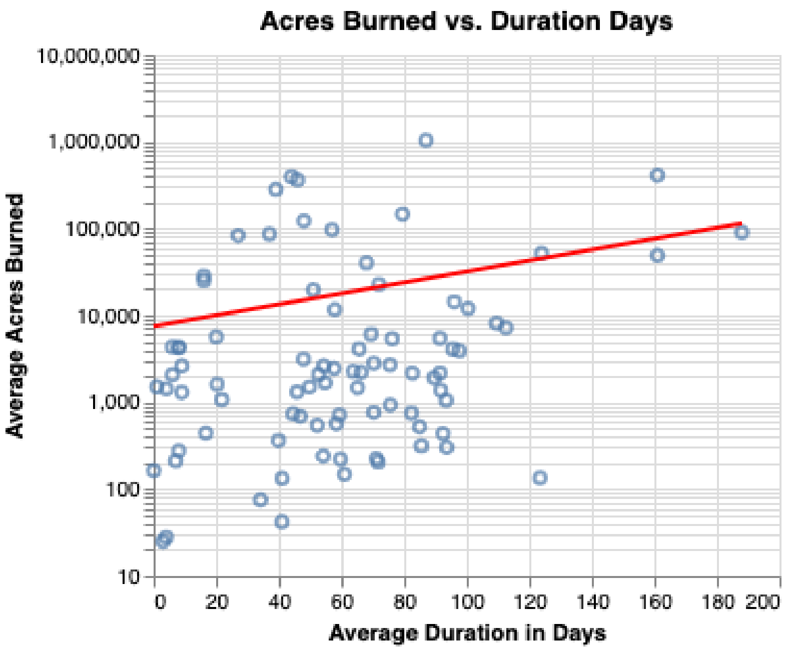
> **Figure 2**: Displays the average acres burned for each recorded year and each month. 2020 surpassed the amount of acres burned for all other years by a landslide, and August historically has the most destructive wildfires.



Relationship between acres burned and duration of fire

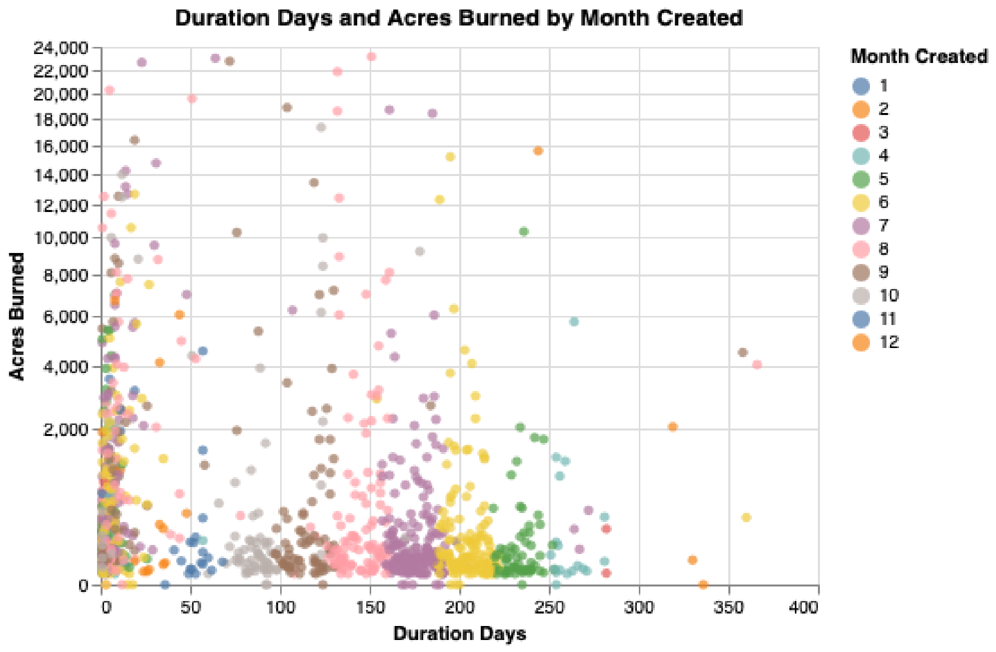
After removing outliers and plotting `Average Acres Burned` vs. `Average Duration in Days`, we see in **Figure 3** that while the relationship is not entirely obvious, there appears to be a slightly positive linear relationship. This indicates that longer wildfires in turn corresponds with more acres burned, which is intuitive enough and not entirely surprising, however, we further dissect this relationship by seeing how this relationship varies with month.

> **Figure 3**: Plot of Average Acres Burned as a function of Average Duration in Days.



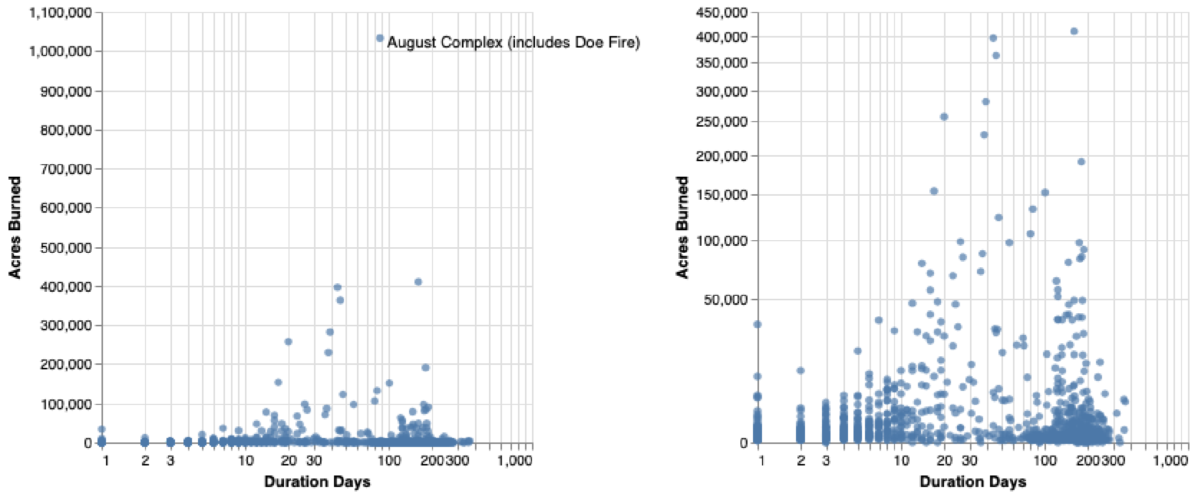
As you can see in **Figure 4**, the months represented by color are almost aggregated into their own cohorts and follows a pattern. April and May are most distuishable grouped months with the longest duration. Then in descending duration, the months are grouped consecutively such as June, July, August, etc. As the months regress and go earlier back in the year, the fire's duration gets longer, which is an interesting observation to make through this visual. Additionally, this scatterplot confirms our earlier discoveries that July (purple) and August (pink) have the most acres burned.

> **Figure 4**: Scatterplot shows the relationship between the duration of the fire and the total acres it burned by month.



It is also important to note that one outlier, the August Complex Fire in 2020, burned 1,032,648 acres and significantly skewed the data. The average duration of wildfires was 73.78 days and while the August Complex Fire significantly exceeded all other firesin terms of burned acres, it only lasted for 87 days. While it exceeded average duration, the max duration was 17901 days. **Figure 5** shows that removing this outlier gives a clearer depiction of the relationship between acres burned and duration.

> **Figure 5**: The two visualizations show the impact that one outlier has made on our data. The left chart includes the massive August Complex Fire of 2020, which burned 1,032,648 acres. When you remove that one outlier, the spread of the chart is much easier to read, and now we can clearly visualize the relationship between the acres burned and duration days.



Discussion

The purpose of this project was explore CalFire's dataset of all reported incidents of California Wildfires from 1969-2021, and more specifically analyze when specifically are these wildfires at their worst. Upon visualizing the counts of incidents by month and year, we found that July was the most common month for California wildfires and 2017 was the year with the most records of fires. The resulting histograms also revealed that shape of the distribution of wildfires follows a left-skewed bell curve meaning that the peak season for fires are summer/fall, and occurences of fires depreciates further from these months as the seasons get colder. This bell-shaped/normal distribution also coincided with severity of the fires with less severe fires and highest concentrations of low severity fires.

Perhaps most shocking in our analysis was ouor findings in **Figure 2**. Prior to our analysis, we assumed that because 2017 had the most wildfires relative to the other years, it would have also been the year with the most destroyed land. However, this was not the case as 2020 had the most average acres burned by far, followed by 2018, and *then* 2017. This is most notably attributed to the 2020 August Complex Fire in Northern CA, our outlier which burned over 1 million acres! The same goes for months- even though July was historically the most susceptible month for fires, August had the most acres burned. These findings are crucial to the understanding of the severity of California Wildfires and implies that even in future cases, even if there are many occurring fires throughout the year, they may be small, or not as destructive, as they *could* be in August per se. These trends are also contingent with California's droughts and rising summer temperatures. Quoted by George Morris of Cal Fire, *"The 2020 fire year 'is on a scale that has not been experienced in California in at least 100 years.'"*

> Cart, Julie. “California's 2020 Fire Siege: Wildfires by the Numbers.” *CalMatters*, 29 July 2021, https://calmatters.org/environment/2021/07/california-fires-2020/.

This outlier although significant in California's historical record, skewed our data and removing it dramatically changed the relationship between acres burned and duration. Additionally, we did notice certain imperfections/typos in the CalFire dataset, which we manually had to resolve, although there's no sure answer that our corrections skewed the data too. Our findings are both informative, but also daunting. California's wildfires are a reoccurring topic of discussion annually, and with rising temperatures and drier months, we should anticipate that fires as dustructive as the 2017 or 2020 fires are in the near future. Perhaps future explorations could include seeing how these wildfire trends correlate with California precipitation or summer temperatures, as well as seeing what regions of California these wildfires are most concentrated in.

In [ ]: