An Exploratory Data Analysis of Wildfires in the United States

SI 618 Project 2

Kelley Sweitzer -- kjsweitz

Introduction **2**
Dataset **2**
Research Questions **3**
Methods **3**
Analysis & Results **4**
How does time affect the frequency of wildfires? **4**
What are the basic metrics we need to know about wildfires? 5
Have wildfires been getting more frequent in recent years? 6
What time of the year is most common for fires? 7
What factors impact the intensity of wildfires? **9**
What basic metrics about wildfire intensity should we know? 9
How do we define wildfire intensity? 10
What is the distribution of wildfire intensity? 11
What are the top causes of wildfires? **12**
What causes wildfires? 13
What percentage of wildfires are human caused? 14
Overall Challenges **15**

Introduction

Wildfires[1] are a naturally occurring ecological event that can ignite huge swaths of vegetation into uncontrollable flames. These events are incredibly destructive, harming ecosystems and wildlife, not to mention disrupting our infrastructure and cities. Wildfires cost insurers a whopping $13 billion in 2020[2] and billions more in property damage and are increasingly threatening homes year over year. The pollution in the air caused by burning can also negatively impact public health and cause respiratory problems after long-term exposure. Additionally, they impact global warming by burning off carbon stores and emitting tons of CO2 into the atmosphere which exacerbates climate change. There is no doubt that wildfires are negatively impacting our ecosystems and society.

Smokey The Bear, wildfire prevention mascot

Although they are difficult to contain, many causes of wildfires are preventable. In 2019, about 87% of wildfires were caused by humans[3]. I'm hoping this data analysis project can explore what causes fires, where fires are occurring most frequently, and what patterns we can find over time.

Dataset

The dataset I will be using for this project is the publicly available database of reports from the Fire Program Analysis (FPA) system. The data can be found on Kaggle at this link. This dataset contains 1.88 million wildfires in the US and spans from 1992-2015. It is an extremely rich dataset with 130 metrics, including georeferencing data, timestamps, and other critical information to explore these wildfires.

Research Questions

There are three primary research questions that I will be exploring. For each question, I will conduct a few different types of analysis to make a case for a concrete conclusion.

---

[1] Wildfires - https://en.wikipedia.org/wiki/Wildfire
[2] The Financial Cost of Wildfires -- https://www.reuters.com/article/us-usa-wildfires-insured-losses-trfn/
[3] Human-Caused Wildfires -- https://smokeybear.com/en/about-wildland-fire

1. **How does time affect the frequency of wildfires?**
    - What time of the year is most common for fires?
    - Have wildfires been getting more frequent in recent years?

2. **What factors impact intensity of wildfires?**
    - What is the distribution of intense wildfires vs. smaller ones?
    - Are there a lot of small wildfires or few big wildfires?
    - Is there a correlation between intensity of wildfires and some other factor?

3. **What are the top causes of wildfires?**
    - What percentage of wildfires are human caused? Has this changed over time?
    - Are there any temporal trends with statistical causes of wildfires?

Methods

In this section, I will be describing the processes I will be using to answer the research questions and arrive at conclusions.

1. **Exploratory analysis of wildfire frequency: sum, mean, median, standard deviation**
   This will simply involve grouping the wildfire data by year and getting a baseline understanding of the dataset. I'd like to find out the total number of wildfires, the average number of wildfires per year, the standard deviation, and the median number of wildfires per year. This will use the "FIRE_YEAR" attribute.

2. **Exploratory analysis of wildfire metrics: mean perimeter of fire, mean length of fire**
   Then, I will break down the data into information about each wildfire. I want to know how long they last for on average and how much acreage they cover (perimeter) over that time on average. This will be calculated for each fire and inserted in a new column called "FIRE_INTENSITY"

3. **Temporal and Seasonal analysis of wildfires**
   Using the "DISCOVERY_DATE" attribute, I will find out how many fires happen per year and see if there is any sort of correlation between time and frequency. I can also discover if there is a correlation between time and wildfire length, then time and perimeter of wildfire. These results could help me understand if wildfires have been getting worse in recent years.

   If this yields a meaningful result, I will also use the "DISCOVERY_DOY" attribute to determine seasonal patterns in wildfire activity to see when wildfires are worse. I will graph the intensity of wildfires as a function of time to get this finding.

4. **Correlation between wildfire length and perimeter of fire**
   To be more specific, I want to know if the length of time a wildfire burns affects the spread of the fire. I can plot the wildfire time on one axis and the perimeter on the other to visualize this.

5. **Distribution of fires by statistical cause**
   The “STAT_CAUSE_CODE” and “STAT_CAUSE_DESCR” metrics in this dataset give a categorical reason for the start of the fire, for example “Smoking” or “Campfire”. If I have time, I want to really dive into this data point, but will start with a histogram of fire cause and count to determine the most frequent cause of wildfires.

Analysis & Results

After making a concrete plan, in this section I will describe the code workflow and present meaningful results and visualizations. I conducted all of this analysis in the attached Jupyter notebook, *si618-project-2-kjsweitz.ipynb*.

I. How does time affect the frequency of wildfires?

There are two temporal components I would like to explore with this dataset. The first is long-term and the second is seasonal effects. Specifically, I want to know if wildfires are happening more frequently now than 23 years ago and if so, by how much. It’s not a secret that wildfires have been appearing more in the news lately and affecting the local economy and way of life for people in certain regions in the US. I want to validate if this hypothesis is true and if it is statistically significant. For seasonal effects, I want to pinpoint at what time of year wildfires are the most frequent and graph this answer. Knowing this would also give people valuable insight into when is the best time to visit certain areas of the US and when to avoid wildfire-prone areas.

What are the basic metrics we need to know about wildfires?

After loading the dataset and taking a look at the SQLite Database through the DB Browser[4] I came up with some quick facts about wildfires. I got these numbers by grouping the wildfires by year and counting up the number of wildfires per year using sum. Here are the stats:

- Least number of fires per year (minimum): **61450**
- Highest number of fires per year (maximum): **114004**
- Average number of fires per year (mean): **78352.7083**

[4] https://sqlitebrowser.org/

- Median number of fires per year: **75615.0**
- Total recorded fires: **1880465**
- Standard Deviation of fires per year: **12759.0394**

Wow, I had no idea there were over 78,000 fires per year on average! That number really shocked me. I wonder why they do not get more coverage in the news if there are so many fires per year? I am hoping to answer that question by exploring geographic details and looking at fire intensity in the next steps of analysis.

Finding #1:

There are **78352** wildfires a year on average in the United States! There have been **1880465** total recorded fires and the highest number of fires per year was **114004,** occurring in 2015.

Have wildfires been getting more frequent in recent years?

To answer this first question, I first unpacked and cleaned the data using Pandas. Then, I simply grouped the wildfires by year and aggregated them by sum. I used these numbers to create a bar plot of the number of fires as it relates to year.

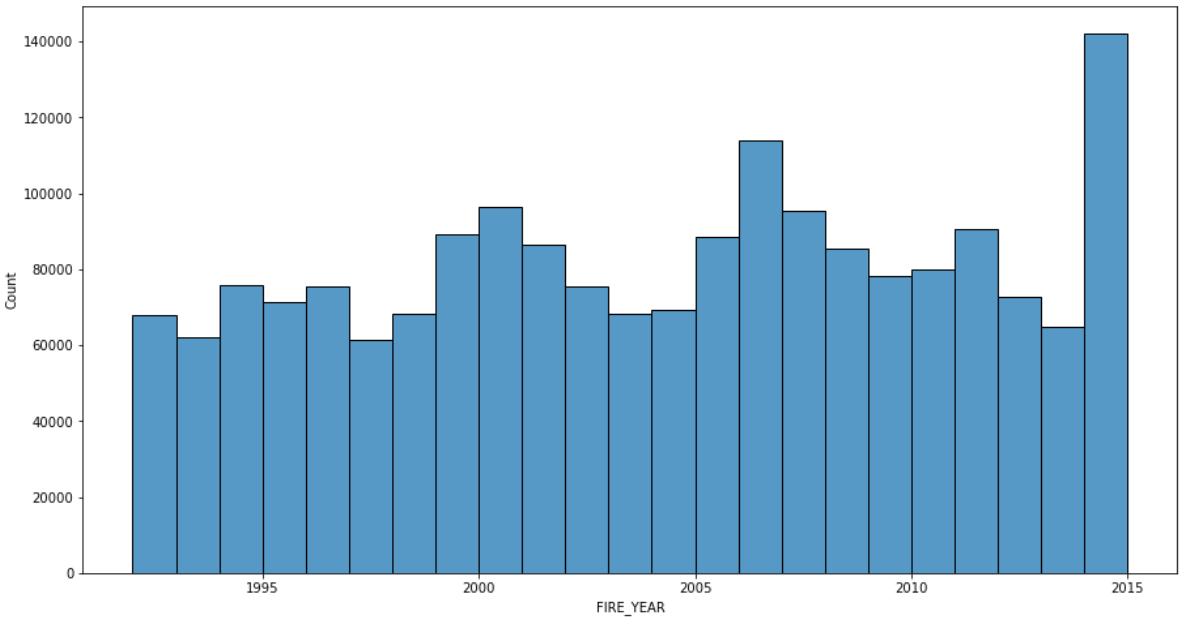


Fig. 1: Distribution of total wildfires per year

It's difficult to tell based on the graph if there is a definitive positive trend between year and frequency of wildfires. There is a sharp spike in wildfire activity towards the end of the graph, but there are also smaller peaks intermittently so it could be part of some other long-term fluctuation or seasonal pattern. I will be exploring this more in the next sections to see if there are trends relating to intensity or length of wildfires as it relates to year.

Since it was not easy to tell from the graph if there was a correlation between year and frequency, I wanted to conduct an ANOVA test to validate this. I conducted this on the fires_peryear dataframe, which contained each fire year and the total wildfire count.

ANOVA test for FIRE_YEAR ~ COUNT

	sum_sq	**df**	**F**	**PR(>F)**
COUNT	41.638914	1.0	0.826496	0.37314
Residual	1108.361086	22.0	NaN	NaN

Based on the results, the p value is 0.37 which is significantly high. It's fair to assume that there is a positive correlation between year and fire frequency.

Finding #2:

There is a positive correlation between wildfires and fire year, meaning as time goes on, fires occur more frequently.

What time of the year is most common for fires?

This time I grouped by the metric "DISCOVERY_DOY" and counted each row for each day of the year that a forest fire was discovered. The "DISCOVERY_DOY" metric tells us which day of each calendar year a fire was reported, so 1 is January 1st and 365 is December 31st.

I graphed this using a line plot and got the following visualization:

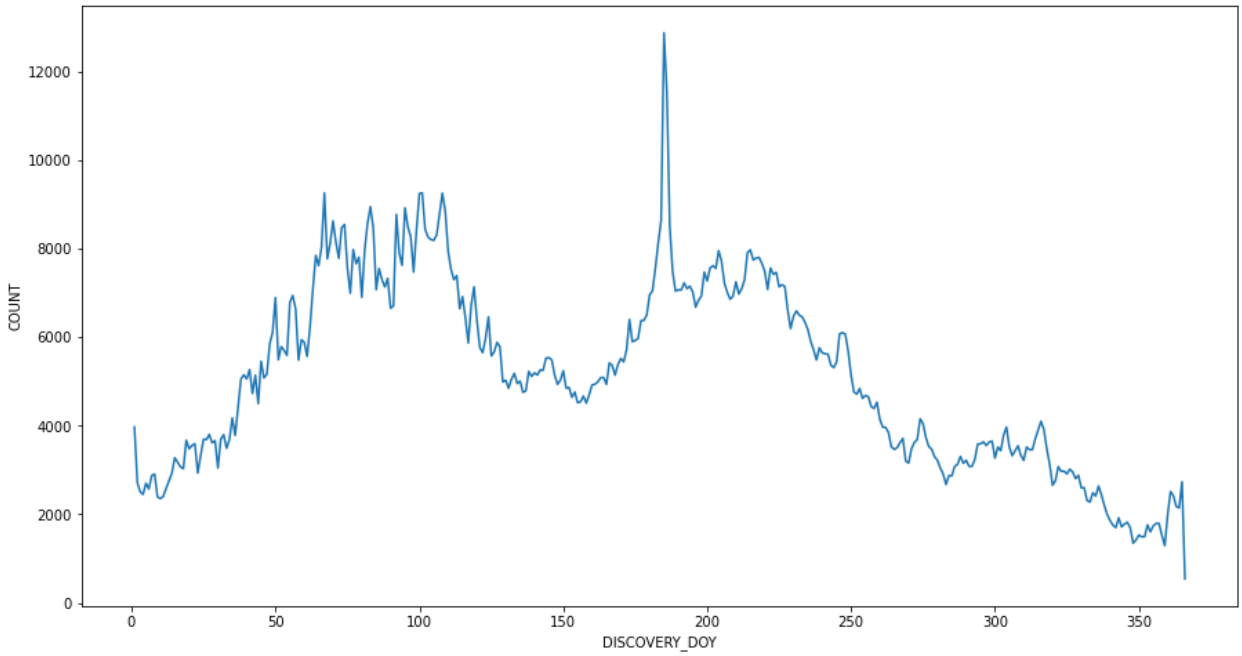


Fig 2. Wildfire Frequency as a function of day of year (January - December)

This chart clearly shows the peak seasons for wildfires in the US. Days 55-100 (Late February - Mid April) see a significant increase from the start of the year and then there is a slight drop off in the early spring. I wonder if this is due to drier climates making wooded areas more susceptible to fires, and then spring rainfall dampening potential fires. I validated this assumption by doing some research and found that wildfires are more likely to occur when the landscape is dry, meaning no rain or snow dampening the brush.[5]

The number of wildfires steadily declines into the late summer and fall, reaching a minimum during the winter months of November and December. Since seasonal rains and snow melting patterns are different in various regions of the country, I imagine that this graph would look different for different regions.

Finding #3:

Wildfires increase in frequency during dry seasons (February - April) and drop off with spring rainfalls. There is typically another peak later in the summer before frequency drops off in the winter.

[5] When is Wildfire Season? https://rainbowintl.com/frequently-asked-questions/when-is-wildfire-season

The frequency increases in June and sees a huge spike on day 185. I was curious about the significance of this date, then I realized it coincides with The Fourth of July, America's favorite explosives-based holiday...

Finding #4:

The biggest day for wildfires each year is the Fourth of July, accounting for 12,875 total wildfires, and about 0.6847% of fires in a year.

The biggest takeaway from this graph is that the Fourth of July is a really destructive holiday. I would be curious to see what proportion of the fireworks-related wildfires happen on July 4th, but I can say that these wildfires are 100% avoidable with proper fire safety protocols.

II. What factors impact the intensity of wildfires?

In part 1 of the analysis, I uncovered some interesting findings about seasonal effects on wildfire frequency. It made me curious about how to measure wildfire intensity and how that metric relates to other factors. I initially was going to plot a few separate variables, to measure these insights, but I think a pairplot with several variables would be a faster way to compare a lot of variables. Let's get into it.

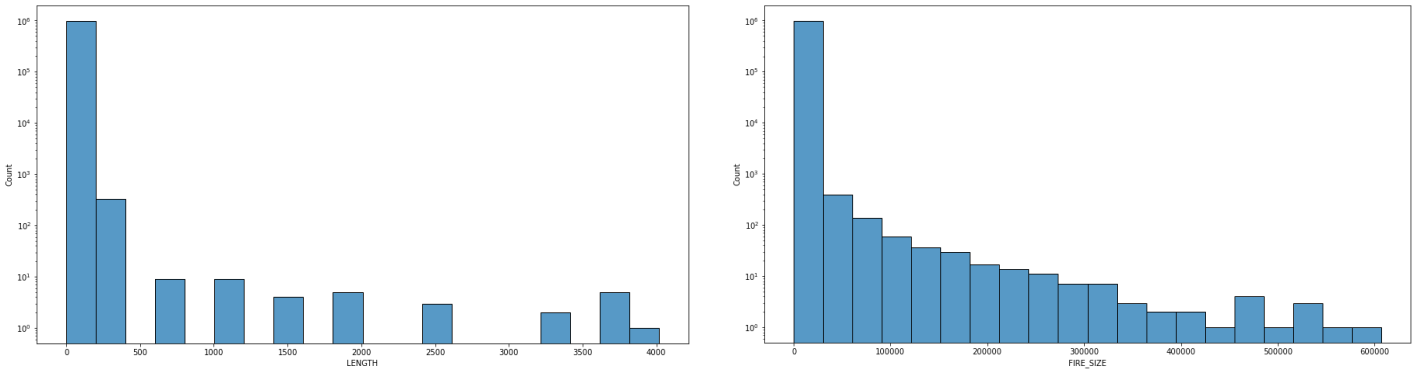
What basic metrics about wildfire intensity should we know?

There are two columns in the database that constitute "intensity" -- FIRE_SIZE and the difference between DISCOVERY_DATE and CONT_DATE. The size is measured in acres and is an estimate of the final perimeter of the fire. There is also a categorical variable FIRE_SIZE_CLASS which indicates if a fire is small or large. I'm going to calculate the length of the fire by subtracting the CONT_DATE, the date the fire was declared "contained", and the DISCOVERY_DATE, the date the fire was discovered and add it to a column called FIRE_LENGTH. I ran a few pandas functions on the intensity metrics and got the following results:

	Fire Size	Length of Fire
Maximum	606945.0 acres	4018.0 days
Minimum	0.0001 acres	0.0 days

Mean	119.268 acres	1.23 days
Median	0.5 acres	0.0 days
Standard Deviation	3374.63	15.618

As you can see, there is a lot of variability for both of these metrics. Fire size ranges from barely a few square feet to the Inowak Fire that burned 606945 acres -- close to the size of Yosemite National Park! The average size is 119 acres, which roughly ¾ the size of Disneyland[6]. Most fires last less than a day, which is surprising to me. The conclusion I'd draw is that fires spread rapidly and can go from a small flame to a few square miles in a few hours!
Both of these distributions are skewed pretty far to the left, indicating most fires are not that severe. It's possible that the fire reporting protocol requires every fire to be recorded, no matter how small.



Histograms of Fire Size (left) and Fire Length (right), log scaled on y-axis

Finding #5:

Most wildfires really aren't that intense: the average wildfire lasts for less than a day and covers ¾ the size of Disneyland -- 119 acres!

How do we define wildfire intensity?

From the variables described above, I'm going to define intensity as the following equation:

Intensity = FIRE_SIZE * (FIRE_LENGTH / 3) + 1

[6]All estimations done through The Measure of Things

For example, a fire that burns 10 acres for 20 days would have an intensity score of 100. A fire that burns for 100 days on 250 acres would be 12,500.

The reason I decided to divide FIRE_LENGTH by 3 is because I believe fire size has a bigger impact on the destructiveness of a fire than length. A fire that doesn't spread very far, but lasts a long time would be far less destructive than one that burns a lot of area very quickly. I still think it's important to include fire length in the equation because if two fires burn the same sized area, but one takes longer to put out, the longer one would be more intense. I arbitrarily said by a factor of 3, but I may adjust this constant and see what results appear.

	FIRE_SIZE	DISCOVERY_DATE	CONT_DATE	INTENSITY
0	0.10	2453403.5	2453403.5	1.000000
1	0.25	2453137.5	2453137.5	1.000000
2	0.10	2453156.5	2453156.5	1.000000
3	0.10	2453184.5	2453189.5	1.166667
4	0.10	2453184.5	2453189.5	1.166667
5	0.10	2453186.5	2453187.5	1.033333
6	0.10	2453187.5	2453188.5	1.033333
7	0.80	2453437.5	2453437.5	1.000000
8	1.00	2453444.5	2453444.5	1.000000

I completed the calculations for each row using Pandas .apply() function and got a new column called INTENSITY. I realized that there were many fires that started and ended on the same day, so the value came out to be 0. I solved this problem by adding 1 to all of the values as a smoothing factor.

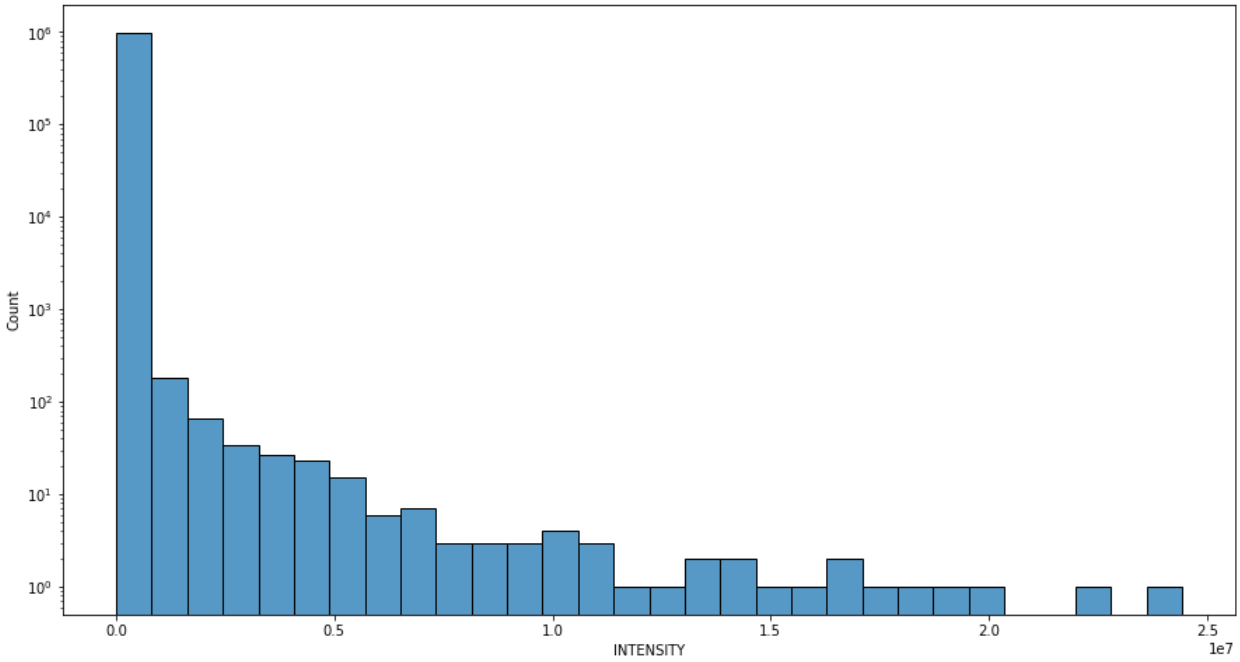
Also, note that the datetime format is the Julian calendar[7], meaning it's the number of days that have elapsed since the start of the calendar. This shouldn't affect our results -- the length attribute will remain accurate in the unit days since we're calculating the difference.

What is the distribution of wildfire intensity?

Now that we have defined an objective measure for wildfire intensity, it is time to make some visualizations! I made a few graphs to illustrate my findings. Here's the code I used to make the first visualization:

[7] Julian Calendar -- https://en.wikipedia.org/wiki/Julian_calendar

```
f, ax = plt.subplots(figsize=(15, 8))
ax.set(yscale="log")
sns.histplot(data=intensity,
                  x=intensity.INTENSITY, bins=30,
                  ax=ax)
```



The first graph is a histogram of the wildfire intensity scores. There are a few things to note about this graph -- firstly that it is log scaled on the y-axis. There are a significantly higher number of low-intensity fires than high-intensity ones. In fact, the median fire size is only 0.5 acres, which is about a third of a football field.[8] This aligns with the finding we had earlier about wildfire size and length.

All in all, I don't think the wildfire intensity score we defined made a big difference in the analysis. Fire size and fire length are two separate, yet meaningful, variables. After doing the intensity score calculations, I realized it made more sense to analyze the two variables on their own rather than use a combined variable. I imagine that for more comparative analyses with other factors, I could use the intensity score, but it abstracts important data and overall isn't that effective.

[8] How big is an acre? https://measuringstuff.com/7-examples-of-how-big-an-acre-is-with-visuals/

III. What are the top causes of wildfires?

The last goal I had for this project was to understand what causes wildfires. There is a lot we can learn about preventing wildfires from knowing why wildfires start in the first place. I have a hypothesis that many of these reasons are human-caused. I'd like to find out how many wildfires are human-related and the top reasons causing them..

What causes wildfires?

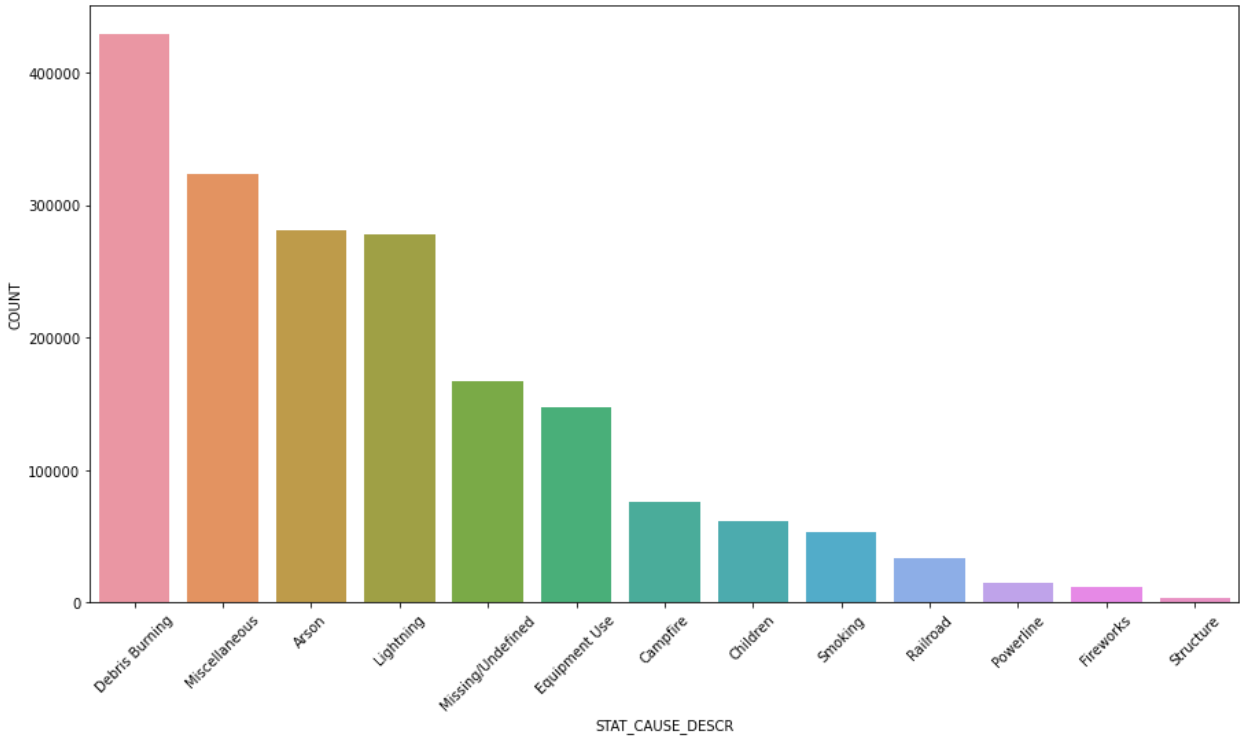
To answer this question, I referred to the STAT_CAUSE_DESCR column. It contains all of the official causes of what started a specific wildfire. I grouped the wildfires by STAT_CAUSE_DESCR using Pandas and aggregated them by count. I sorted the results by value and got the following DataFrame:

STAT_CAUSE_DESCR	COUNT
Debris Burning	429028
Miscellaneous	323805
Arson	281455
Lightning	278468
Missing/Undefined	166723
Equipment Use	147612
Campfire	76139
Children	61167
Smoking	52869
Railroad	33455
Powerline	14448
Fireworks	11500
Structure	3796

I used the NWCG Glossary [9] to define some of the causes. The #1 cause for wildfires in the United States is debris burning. In fire suppression terminology, this means "a fire spreading from any fire originally ignited to clear land or burn rubbish, garbage, crop stubble, or meadows (excluding incendiary fires)." It seems that in some rural areas in the United States, fires are started by an individual trying to burn land or trash. Arson was the #3 cause of forest fires, which is intentionally burning your own or someone else's property.

[9] National Wildlife Coordinating Group Glossary - https://www.nwcg.gov/glossary/a-z

A few notable causes are “Children” causing a total of 61,167 recorded fires and “Campfire” causing 76,139.



This bar graph shows the distribution of fire causes. There are no extreme outliers, but each cause is a fairly significant percentage of total fires.

What percentage of wildfires are human caused?

I wanted to compare the number of human-caused fires to the total number to get an idea of human impact on ecosystems. I defined a list of human-caused fires by selecting a subset of the fire causes above and adding the counts together using Pandas. The events that I considered to be “human caused” as defined by the NWCG Glossary are Arson, Campfire, Children, Debris Burning, Equipment Use, Fireworks, Powerline, Railroad, Smoking, and Structure.

Total recorded wildfires:	1111469
Human Caused Recorded Wildfires:	1880465
Natural Recorded Wildfires:	768996
Percentage Human Caused:	59%

The conclusion we can draw is that 59% of wildfires are human-caused! That is a bit lower of a percentage than I expected, but since “Miscellaneous” and “Missing/Undefined” take up a large chunk of fire causes, it’s possible that this percentage is different.

Finding #6:

About 59% of all recorded wildfires are caused by human-related events. The top human wildfire causes are debris burning, arson, and equipment use.

Overall Challenges

I am excited about the findings I was able to uncover from this rich dataset, but not everything went as planned. I want to quickly touch on the main challenges I faced in both the programming, analysis, and design of this project.

- **SQLite Database:** The database I worked with came in the form of a SQLite database instead of a CSV or TXT, which is different from what we used in class for the most part. This challenge wasn't a huge issue once the data was loaded, but it was a little difficult to manipulate the whole dataset using just Python.
- **Large Dataset**: There were almost 2 million entries in the database. My computer struggled at times to process all of that, especially when working with all 130 columns in the database! I mitigated this problem by selecting just the columns I needed in their own dataframes instead of manipulating a 2 million row x 130 column set.
- **Intensity Factor:** I originally thought it would be a good idea to create a new metric called "intensity" to measure the combined fire area burned and fire length, but I later realized it wasn't as useful as I thought. I still think it was worthwhile to explore and learn from, but it ultimately did not produce any meaningful findings.
- **TMI:** The dataset had 130 columns, which meant there were a lot of factors to look into in regards to wildfires. I did my best to pull out the ones most relevant to my goals, but there were many that went unused. I think in the future, I could definitely revisit this dataset and create a different type of analysis altogether.
- **Overambitious:** I think the proposal I submitted was a little ambitious in terms of length and didn't include any high-level data analysis techniques. I think there could have been areas to incorporate clustering or other types of machine learning, but I instead opted for many smaller and easier analyses. I also wanted to do more pairplots and correlation

analysis, but was unable to due to the size of the dataset. If I had more time, I would have figured out how to work around that.