HW2: Wildfires in California

Stat 133, Spring 2023

Motivation

In this assignment, you are going to perform an exploratory analysis of **wildfires in California** using some of the `"tidyverse"` tools.

You will have to write a report (`Rmd+html`) containing your code, analysis and narrative. Also, you will have to record a short 3-minute video about some of the analysis carried out and their outcome.

A) Fire Perimeters Data

The data for this assignment involves the so-called **Fire Perimeters** database that is curated and maintained by the *California Department of Forestry and Fire Protection* (Cal-Fire). According to Cal-Fire, please keep in mind that:

> "The fire perimeter database represents the most complete digital record of fire perimeters in California. However it is still incomplete in many respects.

A.1) CSV Data File

The specific dataset you will be working with is in a Comma Separated Value (CSV) file: `California_Fire_Perimeters.csv`. This file is available in bCourses.

https://bcourses.berkeley.edu/courses/1521994/files/folder/hws/hw2

A.2) Data Dictionary

The associated **data dictionary** is in the text file `fire-perimeters-data-dictionary.md` also available in the above bCourses link. This document provides detailed descriptions about the fields (i.e. columns) of the CSV data file.

A.3) Data Preparation Script

In the same bCourses link, you can find an auxiliary script `hw2-data-preparation-script.html` with some starting code to import the data in R, and perform the main data-preparation steps. You can use this code in your own report (Rmd+html).

B) Univariate Exploratory Data Analysis (*Not to be reported*)

This part does not have to be included in your report, but you should carry out this exploratory analysis to get to know your data, which is something typical of almost any data analysis project.

Following the descriptions and details of the **data dictionary** document, perform an Exploratory Data Analysis (*EDA*) to inspect how data/values in the columns look like, and if things seem to make sense.

In real life, you should look at all the variables. But you can focus on just a handful of them: `YEAR`, `GIS_ACRES`, `MONTH`, `DURATION` and `CAUSE`.

About NAs: The Fire Perimeter data contains several missing values. Sometimes it's convenient to filter out such values. One way to do this is with the function `is.na()` and the logical negation `!`, for example:

```
# hypothetical example for counting values of a VARIABLE
# in a table "tbl", filtering out missing values
tbl %>% filter(!is.na(VARIABLE)) %>% count(VARIABLE)
```

In case you are curious: Although you won't be dealing directly with geospatial data (e.g. longitude-latitude coordinates, plotting maps), you may want to take a look at an interactive map of California Fire Perimeters (link below).

https://gis.data.cnra.ca.gov/datasets/CALFIRE-Forestry::california-fire-perimeters-all/explore?layer=0&location=37.425382%2C-118.986576%2C6.84

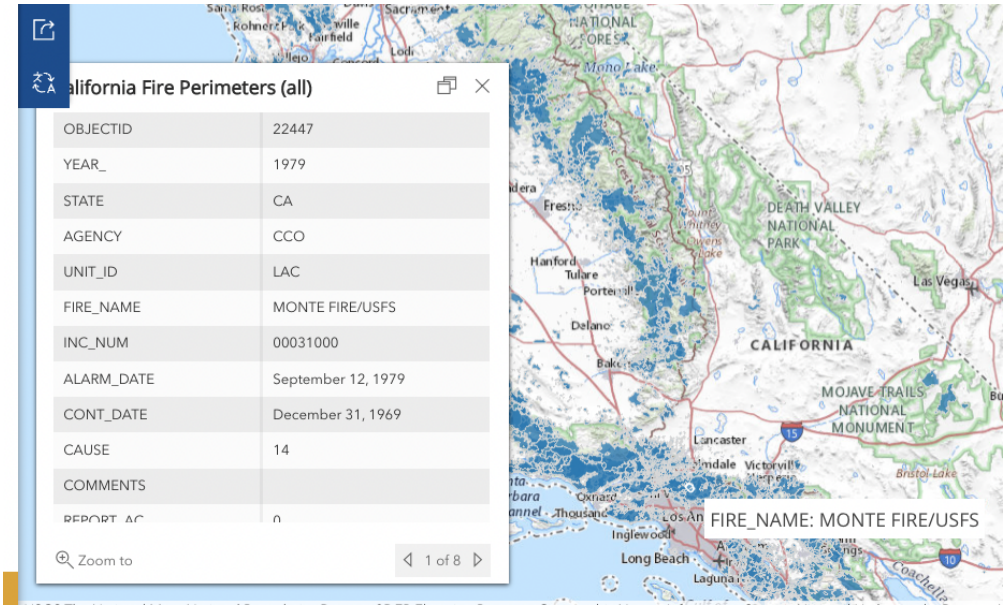


Figure 1: Screenshot of the interactive map of California Fire Perimeters.

C) Main Analysis (*To be reported*)

You will have to analyze the Fire Perimeters Data in order to address the following sections C.1 - C.4.

- Each section, C.1 - C.4, links to a source that you must consider when following the given instructions.
- Each source is described by their **Name**, **Source**, and **URL**.
- For each source, you will find a certain piece of information: e.g. a table, a graphic, or a given claim.
- These pieces of information serve as a prompt for you to perform an analysis. You'll have to provide either numeric and/or visual output to address those prompts.

C.1) Top-20 Largest California Wildfires

- **Name**: "Top 20 Largest California Wildfires"
- **Source**: California Department of Forestry and Fire Protection (Cal-Fire)
- **URL**: https://www.fire.ca.gov/media/4jandlhh/top20_acres.pdf

Consider the table of Top 20 largest California Wildfires—screenshot with a few rows shown below—that is available in the PDF document listed in the above URL.

	FIRE NAME (CAUSE)	*DATE*	*COUNTY*	*ACRES*
1	**AUGUST COMPLEX** *(Lightning)*	August 2020	Mendocino, Humboldt, Trinity, Tehama, Glenn, Lake, & Colusa	**1,032,648**
2	**DIXIE** *(Powerlines)*	July 2021	Butte, Plumas, Lassen, Shasta & Tehama	**963,309**
3	**MENDOCINO COMPLEX** *(Human Related)*	July 2018	Colusa, Lake, Mendocino & Glenn	**459,123**
4	**SCU LIGHTNING COMPLEX** *(Lightning)*	August 2020	Stanislaus, Santa Clara, Alameda, Contra Costa, & San Joaquin	**396,625**
5	**CREEK** *(Undetermined)*	September 2020	Fresno & Madera	**379,895**

Use `"dplyr"` commands to obtain your own data table (`tibble`) of "Top 20 Largest California Wildfires". Include the following variables in your table:

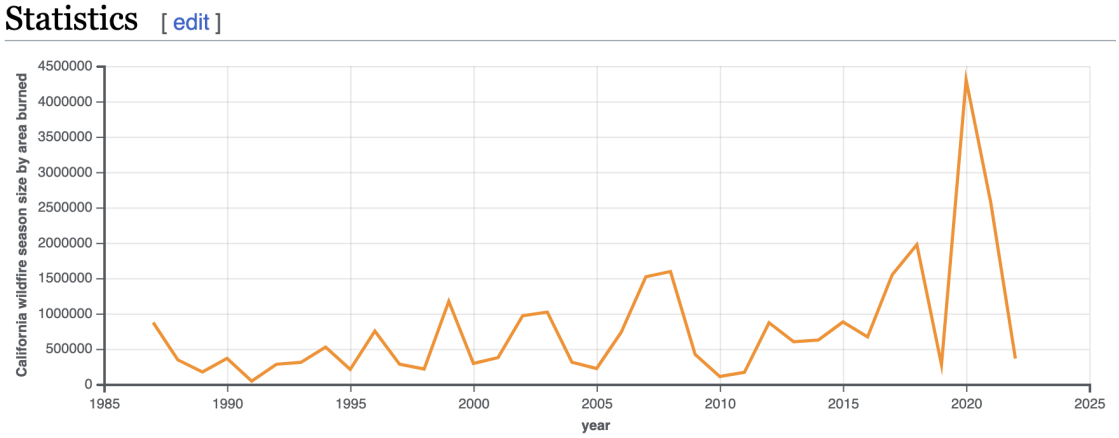
- `YEAR`
- `FIRE_NAME`
- `CAUSE`
- `ALARM_DATE`
- `GIS_ACRES`

Ignoring County, Structures, and Deaths columns, how does the content of your table compare to that reported by Cal-Fire in its `top20_acres.pdf` report? If you find any discrepancies, what seem to be the reason for such mismatches?

C.2) Area Burned Over Time

- **Name**: "List of California Wildfires"
- **Source**: Wikipedia
- **URL**: https://en.wikipedia.org/wiki/List_of_California_wildfires

Refer to the wikipedia entry "List of California Wildfires", and consider the timeline graph—see screenshot below—of California wildfire season size by area burned (1985 - present)



Carry out an analysis and use `"ggplot2"` functions to replicate—as much as possible—this timeline. In addition to your own replica, create another timeline plot but this time consider data for more years (e.g. 1950-present or some other period of time much longer than 1985-present). For this second plot, feel free to use a `geom_...()` function different from the one used in your replica.

C.3) California Fire Season (part 1)

- **Name**: "California Fire Season: In-Depth Guide"
- **Source**: Western Fire Chief Association
- **URL**: https://wfca.com/articles/california-fire-season-in-depth-guide/

In the "California Fire Season: In-Depth Guide" (see above URL) it is claimed that:

> *"It is a common misconception that the most dangerous time for fires in California is during July and August. While there may be fewer fires in September and October, the fires that do occur . . .* ***burn through many more acres****. This explosive effect is due to a combination of dry vegetation from hot summer weather, and intense dry winds that blow through the state during fall.*

Carry out an analysis and obtain output(s) that allow you to answer the following questions:

- Is it true that July and August are the months where fires tend to start more frequently?

- Do fires that start either in September or in October burn through many more acres than those starting in July or August?

C.4) California Fire Season (part 2)

- **Name**: "California Fire Season: In-Depth Guide"
- **Source**: Western Fire Chief Association
- **URL**: https://wfca.com/articles/california-fire-season-in-depth-guide/

In the "California Fire Season: In-Depth Guide" (see above URL) it is claimed that:

> *"The length of the fire season in any given year in California depends on summer temperatures, rainfall, and wind, with the most fires historically occurring between May and October. However, recent data show that, due to rising temperatures and decreased rainfall,* ***the season is beginning earlier and ending later each year, approaching a year-round fire season****"*

Carry out an analysis and obtain output(s) that allow you to answer the following questions:

- Is it true that, historically, most wildfires in California occur between May and October?
- Is it true that the fire season is beginning earlier and ending later each year, approaching a year-round fire season?

D) Submission

You will have to submit your source `Rmd` file, the generated `html` document, and the public link of your 3-min video.

D.1) Report (`Rmd` + `html`)

Your report should include:

1) all your code with results, outputs, and graphics,
2) your narrative: descriptions, explanations, interpretations, etc.

Above all, do not simply include code chunks, with minimal descriptions, and a boring list of conclusions for each research question. We want to see content that reflects your "thinking process" and how you organize your workflow, all of this by reading and looking at your report, without you there to explain it to us. Therefore, your submission must "speak for itself".

- Do not be afraid to use as many code chunks as necessary. Your code should be easy to read and understand, using descriptive names, well commented, and well organized.

- For visualizations, describe motivations behind the particular ones built and what they illuminate. Make sure any visualizations are functionally labeled (e.g. title, possibly a subtitle, axis labels, units of measurement when applicable).
- You must communicate your key findings and insights clearly.
- Keep the content length of your report to no more than 2500 words. To give you a rough idea: single spaced, 2500 words yields about 5 pages (excluding images). Obviously there are no pages in an html document, but use this guideline to keep track of your code and narrative.

D.2) Video

In addition to the report (Rmd and html files), you will also have to record a video (3-minute max length) in which you use your results to address either section B.3 or B.4 (just one section).

Ideally we want to see your face (it's really you and not someone else) and also a capture of your screen, explaining to us your analysis (e.g. numerical and/or graphical evidence), your results, interpretations, and conclusions.

We recommend recording the video using your Zoom Berkeley account, and then sharing with us the **public link**. Optionally, you can also record a video using other tools, upload it to your Berkeley Box, or your google drive account, or youtube, or vimeo, or a similar video-hosting platform.

Regardless of where you decide to host the file of your video, you must share the **public link** (don't make it private) in bCourses: use the "comments section" of the submission to add this link as a comment. To be clear: do NOT upload the video to bCourses, just its link.

Make sure that the video does not exceed 3-minutes, that its resolution is okay, without too much background noise, avoid very low volume or inaudible audio, and also ensure that your image is stable (don't make us watch a shaky video).