# Portfolio assessment Part 1:
# *Wildfires in California between 2013 and 2020*

ADSE3200 – Visualization

March 23rd, 2023 *Candidate No. 148* Word count: *3,823*

## Table of contents

# 1 Task 1. The utility value and the audience:

## 1.1 Background and relevance to the theme

For this assignment, I have chosen to focus on the United Nations' (UN's) Sustainable Development Goal (SDG) 13, *«Climate Action»*, which is aimed at combatting climate change and its negative impacts (United Nations General Assembly, 2015). The dataset I have chosen to work with is *«California WildFires (2013-2020)»* (ARES, 2020), released into the public domain under the Creative Commons CC0 1.0 licence, and accessed through Kaggle on February 24th, 2023. The data was originally scraped from the official website of the California Department of Forestry and Fire Protection (CAL FIRE), and includes detailed information on all reported wildfire incidents in California within the relevant time frame.

While I have chosen to apporoach the dataset from the angle of climate change, the dataset could also be relevant to UN's SDG 15, *«Life on Land»*, and to a lesser SDG 3, *«Good Health and Well-Being»* (Delfino et al., 2009; Kunzli et al., 2006).

California is well known for her fire season, which typically begins in early summer or late spring and runs until late fall. Californian wildfires often rank among the deadliest and most destructive in the world, and in recent years, much attention has been drawn to the potential effects of global climate change on the destructiveness and deadliness of wildfires in California and elsewhere (Parks & Abatzoglou, 2020).

In addition to the increase in temperature caused by global warming, it is speculated that droughts are becoming more frequent as a result of changes in weather patterns brought about by climate change (Diffenbaugh et al., 2015). This facilitates an increase in both the frequency and intensity of wildfires, and makes wildfire suppressing operations more difficult as well as more dangerous for the firefighters and rescue crews involved. This development is cause for grave concern, both for its environmental impacts and for the potential for harm to people and property.

The stated aim of SDG 13 is to *«Take urgent action to combat climate change and its impacts»*. This entails not only efforts to stop or reverse global warming, which is but one of the effects of climate change, but also involves taking concrete measures to limit or prevent damage caused by natural disasters and extreme weather events globally. In order to implement effective measures towards this goal, we must first understand the relevant phenomena and their effects.

## 1.2 Who and why – audience and utility

The dataset contains information about wildfire events in California over a span of seven years (January 2013 - December 2019). From this data we can extrapolate information about current trends in wildfire incidents as they relate to global climate change. Since climate change is very much a hot topic, this infomation is of general public interest. The dataset is therefore well suited to popular dissemination and education about the topic, i.e. climate change and its impacts.

More specifically, the information can be used to predict where and when wildfires are likely to occur within the state of California. This information may inform decisions made by Californian policy makers concerning wildfire preparedness, e.g. in which administrative units are additonal funds, equipment, and personnel needed; and development, e.g. which areas are particularly susceptible or vulnerable to wildfires, and what measures can be implemented to combat this?

For this assignment, due to some of the techniques that I will utilise, the target audience includes mainly people who are already somewhat familiar with the reading of scientific graphs and with basic statistical concepts such as *trend lines*. During the user testing, I will therefore select people who are either currently pursuing or have already completed their higher education.

## 1.3 What are my goals with the assignment?

The information I wish to visualise from the dataset concerns mainly the *spatiotemporal* distribution of wildfires in California, i.e. where and when do wildfires most frequently occur?

Specifically, the questions I want to explore are:

1) Where in California do major wildfires most frequently occur?
2) Where in California do wilfires cause the most damage?
3) When does the Californian fire season start, peak and end?
4) Are wildfires becoming more frequent, and how is this reflected in the onset and duration of the Californian fire season?

It would be very interesting to combine these data with meteorological data, including temperature and precipitation data for the focal period, geographic data describing e.g. landscape features and vegetation types which could influence the spread and severity of wildfires, or demographic data including population density, which is likely to be a primary factor when looking at damage. However, since this would require using data from several additional sources, these analyses are beyond the scope of this assignment.

## 2 Task 2. Dataset insights

My chosen dataset contains data on Californian wildfires spanning the seven years from Janurary 2013 to December 2019. The object in the context of this dataset is the individual wildfire incident. Each row represents a wildfire incident, i.e. an *observation*, and the various columns represent *attributes* or *variables* pertaining to that particular wildfire. There are 40 columns of attributes in the dataset. For descriptions, see Table 1.

The dataset is formatted as a *comma-separated values* (.csv) text file, which can be natively imported as a spreadsheet and manipulated by most data handling software, e.g. Microsoft Excel, Tableau or R.

Table 1: Summary of all the attributes in the dataset including textual descriptions of each attribute. The attributes are listed in the same order in which they appear in the dataset. I was not able to provide a description of the attributes *«Featured»* or *«Public»* as I was unable to identify their meaning from the available documentation.

| **Field** | **Description** |
|---:|:---|
| AcresBurned | The number of acres consumed by the wildfire incident. |
| Active | Is the wildfire incident currently ongoing? |
| AdminUnit | The administrative unit where the fire started. Administrative unit borders follow California county borders fairly closely, but are not identical. |
| AirTankers | The number of air tankers deployed to fight the fire. |
| ArchiveYear | The year in which the data was archived. This also corresponds to the year in which the wildfire took place. |
| CalFireIncident | Was the wildfire treated as a CAL FIRE incident, i.e. did it occur within the jurisdiction of the department? |
| CanonicalUrl | The fire.ca.gov web page containing information about the incident. |
| ConditionStatement | Textual observations and notes about the wildfire incident. |
| ControlStatement | Public statments issued concerning the fire. |
| Counties | Name of the county where the wildfire started. |
| CountyIds | ID of the county where the wildfire started. |
| CrewsInvolved | The numbers of firefighters and other crews involved. |
| Dozers | The number of bulldozers employed. |

| Field | Description |
|---:|---|
| Engines | The number of fire engines (i.e. fire trucks) assigned |
| Extinguished | The date when the wildfire was extinguished. |
| Fatalities | The number of casualties in the fire. |
| Featured | N/A |
| Final | Is this the final update on the fire in CAL FIRE's systems? |
| FuelType | The type of material which burned. |
| Helicopters | The number of helicopters assigned. |
| Injuries | The number of injured personnel. |
| Longitude | The longitudinal coordinate of where the wildfire started. |
| Location | Textual description of the location. |
| Longitude | The latitudinal coordinate of where the wildfire started. |
| MajorIncident | Was the wildfire considered a major incident? |
| Name | Name of the wildfire. |
| PercentContained | Containment percentage, i.e. 100% for a finalised incident. |
| PersonnelInvolved | The number of CAL FIRE personnel involved. |
| Public | N/A |
| SearchDescription | Textual description of the wildfire incident. |
| SearchKeywords | List of keywords descriping the incident. |
| Started | The date when the wildfire started. |
| Status | Fire status, i.e. either ongoing or finalised. |
| StructuresDamaged | The number of structures damaged in the fire. |
| StructuresDestroyed | The number of structures destroyed in the fire. |
| StructuresEvacuated | The number of structures that were evacuated. |
| StructuresThreatened | The number of structures threatened by the fire. |
| IniqueId | Unique alphanumerical ID. |
| Updated | The date of the last update on the fire. |
| WaterTenders | The number of water tenders (i.e. water transport trucks) assigned. |

The dataset contains several different types of data. Much of the data is *qualitative* rather than *quantitative*. Examples include the attributes *«AdminUnit»*, which specifies in which administrative unit the wildfire incident took place, *«ConditionalStatement»*, which contains textual desciptions of the wildfire incident, *«FuelType»*, which describes the type of material that burned, and so on. The attributes in this category can be further subdivided into several categories.

The attributes *«Active»*, *«CalFireIncident»*, *«MajorIncident»*, *«Status»*, etc. are examples of *boolean* or *binary* data, i.e. data whose value can take on only one of two states, typically *«true»*/*«false»* or 0/1. Attributes such as *«AdminUnit»* and *«Counties»* are examples of *nominal* data.

Inversely, the quantitative attributes in the dataset include among other things the attributes *«AcresBurned»*, *«Injuries»*, and *«PercentContained»*. Like the qualitative data types, these data can be further subdivided into several categories. The attributes *«AcresBurned»*, *«ArchiveYear»*, *«Extinguished»*, and *«Injuries»* are examples of *interval* data, while *«PercentContained»* belongs to the category *ratio* data.

Additionally, the dataset in its entirety can be classified as *geospatial* data, since each observation is directly associated with a geographic location as specified by a set of coordinates, i.e. *«Latitude»* and *«Longitude»*.

For this assignment and for my purposes, the quantitative attributes are the most relevant (with a few exceptions). However, there are many possible options for visualising the qualitative attributes. For example, a word cloud could be created from the keywords listed in *«SearchKeywords»*.

# 3 Task 3: Visualisation and analysis

## 3.1 Design

All visualisations and analyses were produced in the R programming environment version 4.2.3 (R Core Team, 2023) (see Appendix A), through the integrated development environment Rstudio (Posit team, 2022).

## 3.2 Figures 1. & 2. Wildfire frequency and total area burned

Figure 1 and Figure 2 are meant to be presented together and are by design very similar. Both figures were created using map visualisation methods. The data points were overlaid on top of a map graphic using the coordinate values from the *«Latitude»* and *«Longitude»* columns of the dataset, in this instance onto a polygonal shapefile of California. The shapefile was downloaded from the California Open Data Portal (CAL FIRE, 2022).

Since the dataset consists of spatial geodata, another option for Figure 1 would be to use the coordinate values to plot each wildfire event as an individual point or create a heatmap overlaid on top of the map. Had I only wanted to create Figure 1, I would have gone for this approach, since it much more accurately represents where wildfire events occured geographically.

However, in reality a wildfire isn't constricted to a single point in space. While the approach mentioned above would be well suited to visualise where wildfires started, it is poorly suited to convey the impact or scale of wildfire events, as I wanted to do with Figure 2. To make sure that Figure 1 and Figure 2 were easily comparable, I decided that it would be unwise to use very different visualisation techniques for the two visualisation. By sacrificing a level of detail in Figure 1, I make sure that the two visualisations work well together.

### 3.3 Figure 3. Monthly frequency of wildfires

I used a line chart to visualise the monthly frequency of wildfires between January 2013 and December 2019 (Figure 3). Each data point in the plot represents the wildfire frequency for a given month, and lines are drawn between data points in order to demonstrate continuity in the data and a natural order in the arrangement of values. In the final visualisation, wildfire frequency is represented on the $y$ axis, while the individual months are represented on a discrete scale along the $x$ axis.

The visualisation also features a trend line which showcases the increasing trend in wildfire frequency. The trend line was calculated using a simple *linear regression model* (lm) where the attributes $n$ (number of wildfires) and *month* are used as the *response* and *explanatory* variables respectively.

### 3.4 Figure 4. Year-by-year comparison

For Figure 4, I utilised a technique called *facet wrapping*. In essence, this means that a plot is divided into several subplots which each include a subset of the data, usually aligned along the same axes in order to make comparisons visually intuitive. Where Figure 3 makes it easy to compare the peak of one fire season with that of the next, i.e. by comparing the height of the data points along the $y$ axis, the goal of Figure 4 was to make it easy to compare the timing of the fire season between years, i.e. when does the fire season start, peak, and end?

By aligning the subplots along the same $x$ axis, i.e. months (January - December), a line can be drawn (in this case literally) between a data point in one subplot and the equivalent data point in another. The subplots were given different colours in part to provide visual interest and in part to make it easier for the reader to navigate between the subplots.

To compare the total number of wildfire incidents between years, one can look at the area under the curve. A greater coloured area indicates a higher total number of wildfire incidents.

## 3.5 User testing

I performed in all three user tests. The participants were shown the visualisations one at a time and asked to explain what they were looking at. For each visualisation, the initial question asked was *«What does this figure tell you?»*. The participants were allowed to explain without interruption, and follow-up questions related to different aspects of the visualisation were asked to fill in holes. Examples of follow-up questions include *«What does the colour of the county's polygon tells you?»*, *«What do you think this line means?»*, *«Do you see an interesting pattern?»*, and so on.

**Participant 1:** The first participant understood Figures 1 and 2 well. He understood that the various Californian counties were represented by individual polygons and that their colour indicated the frequency of wildfires in that county, and in the case of Figure 2, that a grey colouration indicated that data about fire frequency is missing for that county.

The participant had some trouble understanding the trend line in Figure 3, and speculated that the line might show the accumulative number of wildfires for the time period. This was rectified with the inclusion of an explanation in the figure text. The participant had no trouble reading Figure 4.

**Participant 2:** The second participant was seemingly familiar with most of the visualisation techniques that I used for the assignment and had no troube reading any of the plots. Considering her background in statistics this was not very surprising. She recognised that wildfires were most common during the summer months and noted that the wildfire frequency seems to have increased after 2017 (Figure 3; Figure 4).

**Participant 3:** The third participant commented that he did not know whether the white colouration of some counties in Figure 1 idicated a missing value or a value of zero before he saw the legend for Figure 2.

Presented with Figure 3, the participant speculated that the trend line could show the mean frequency of wildfires. However, he corrected himself after reading the relevant part of the figure text.

The participant had no trouble reading Figure 4 and noted without being promted that it looks like the fire season ran longer into autumn during the later years of the range (participants 1 and 2 had also recognised this pattern, but did not mention it in their initial explanation of the figure).

The partcipant expressed concern about whether all of the visualisations conformed to the *Web Content Accessibility Guidelines* (WCAG) requirements concerning colour contrasts. I did not change the colour values for any of the visualisations following this feedback, but I did add an alternative text to each of the visualisations (I am unsure as to whether this effort survived the transition to PDF).

## 3.6 Finalised visualisations

The finalised visualisations can be viewed below.

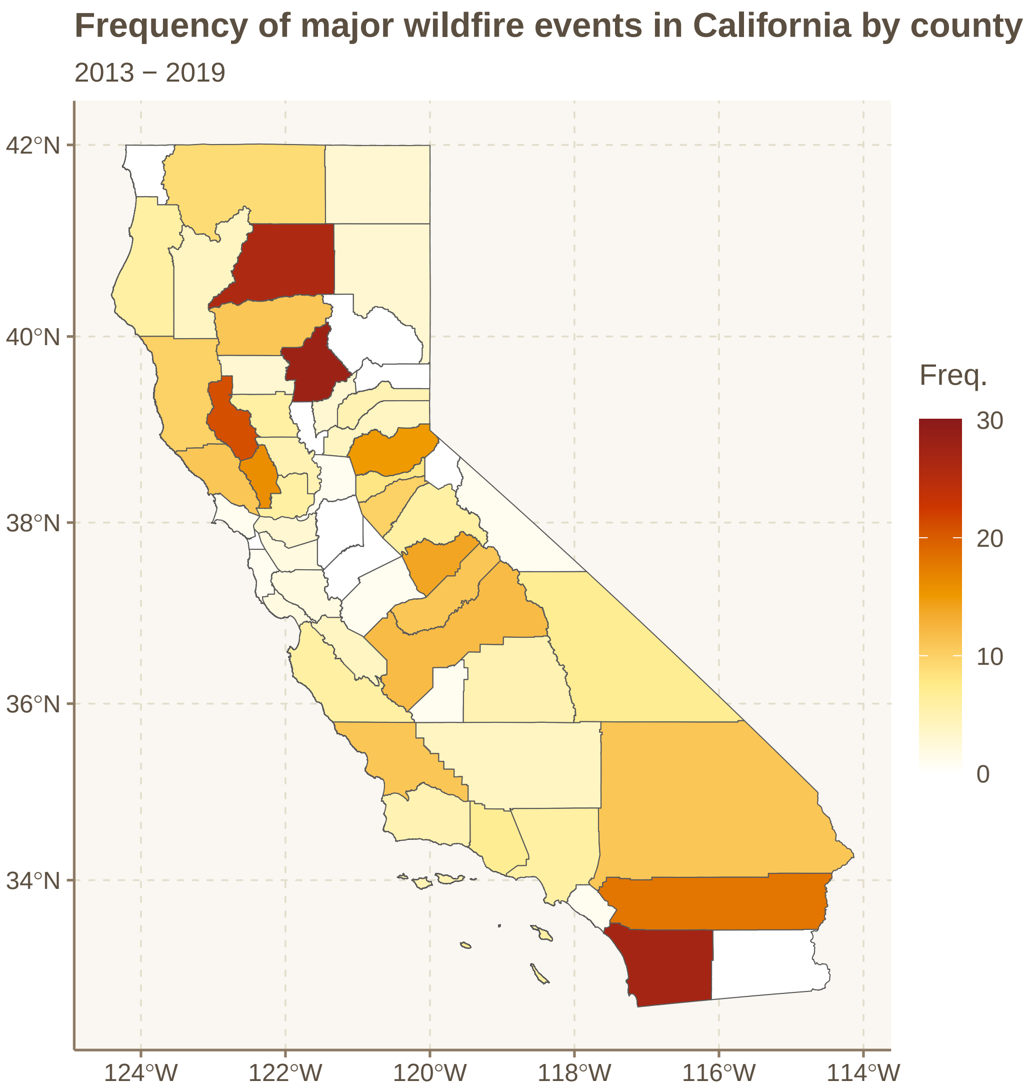

Figure 1: *Frequency of recorded major wildfire events in California by county (2013 - 2019). Each polygon represents a Californian county, and its colour represents the number of wildfires which started in the county within the relevant time frame – the darker the colour the higher the number. Each wildfire event is included in the statistic only once, meaning that fires which started in one country and later spread into another are only counted in their county of origin.*

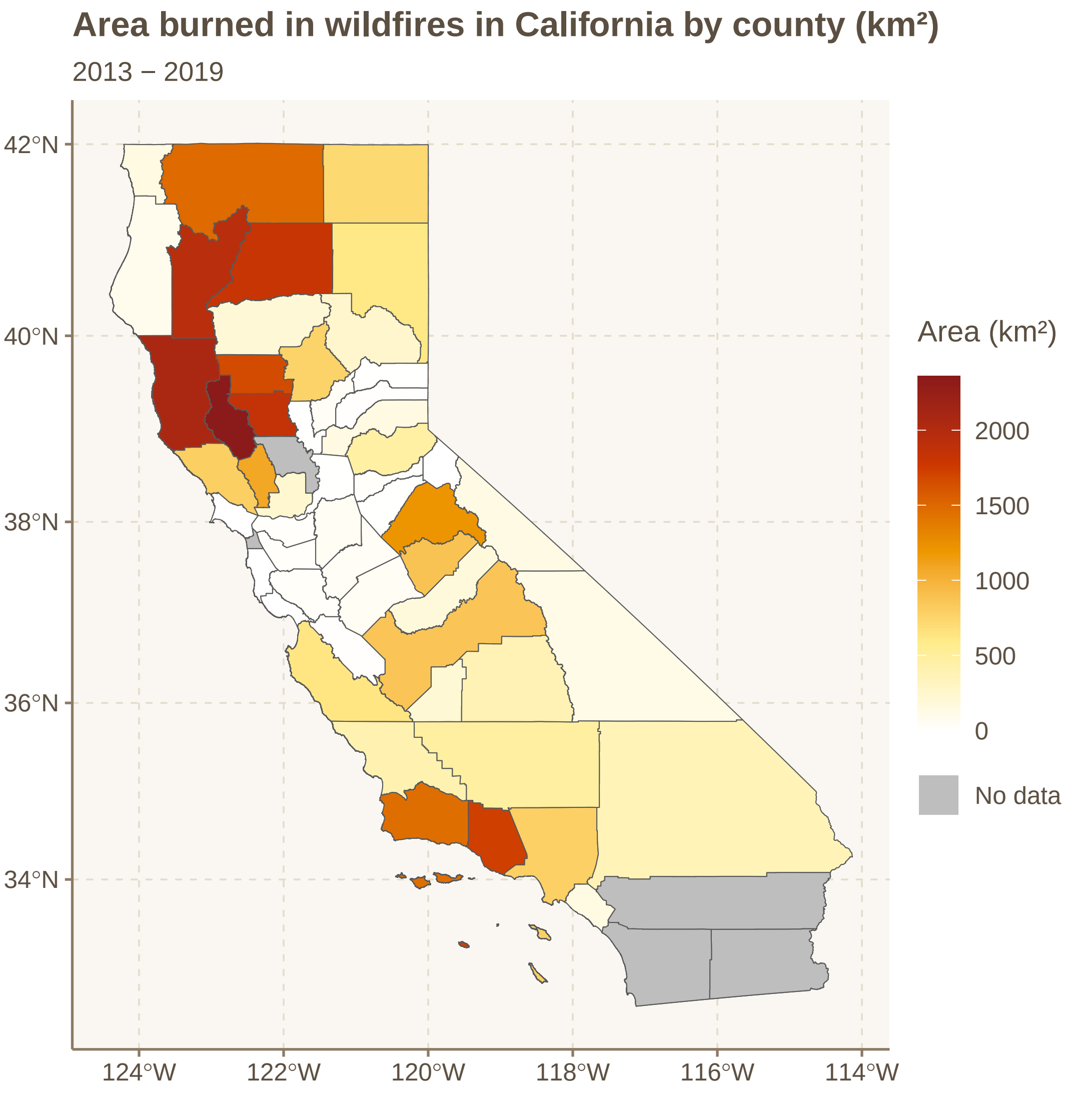

Figure 2: *Total land area burned in California by county (2013 - 2019), measured in km². Each polygon represents a Californian county, and its colour represents the total area consumed by wildfires within the relevant time frame – the darker the colour the larger the area.*

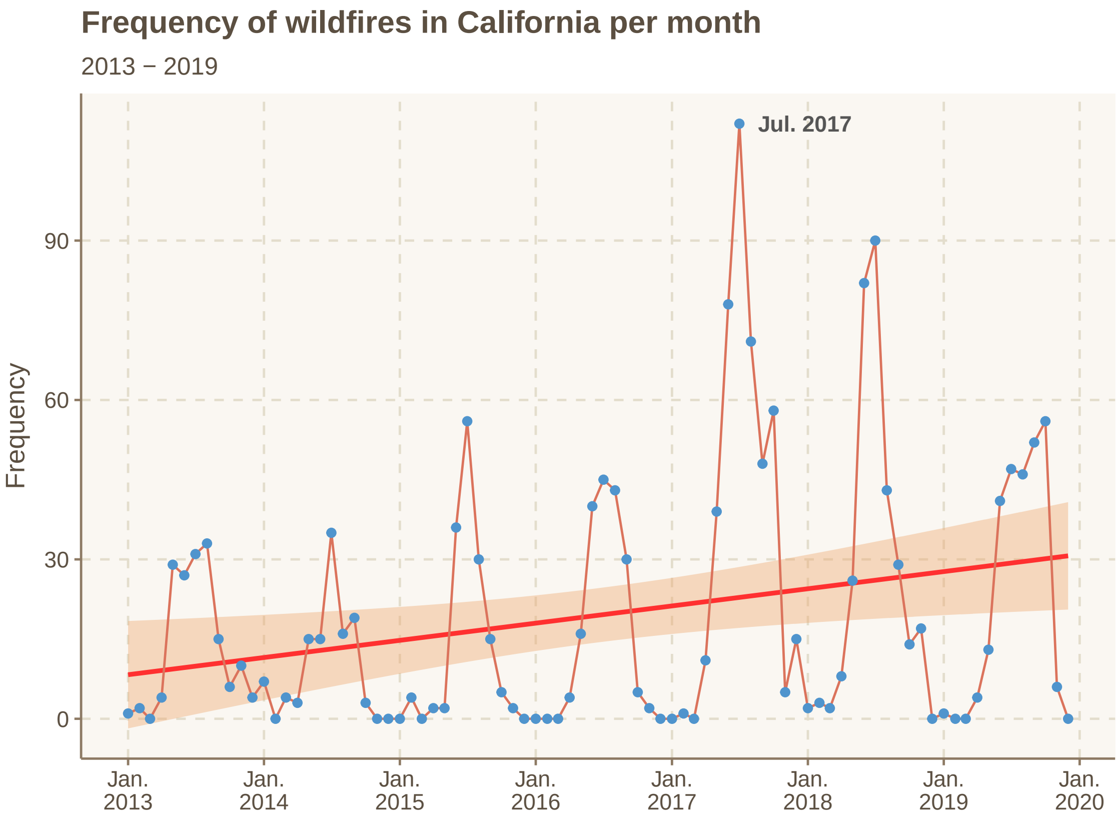

Figure 3: *Monthly frequency of wildfires in California (2013 - 2019). The line chart shows the number of recorded wildfire incidents in California per month between Jan. 2013 and Dec. 2019, with each point representing the number of fires in a given month. The fire season of 2017 saw an especially high number of wildfires, peaking in Jul. 2017 (highlighted in the plot). The trend line shows that there is an increasing trend in the number of wildfires which occur during the Californian fire season.*

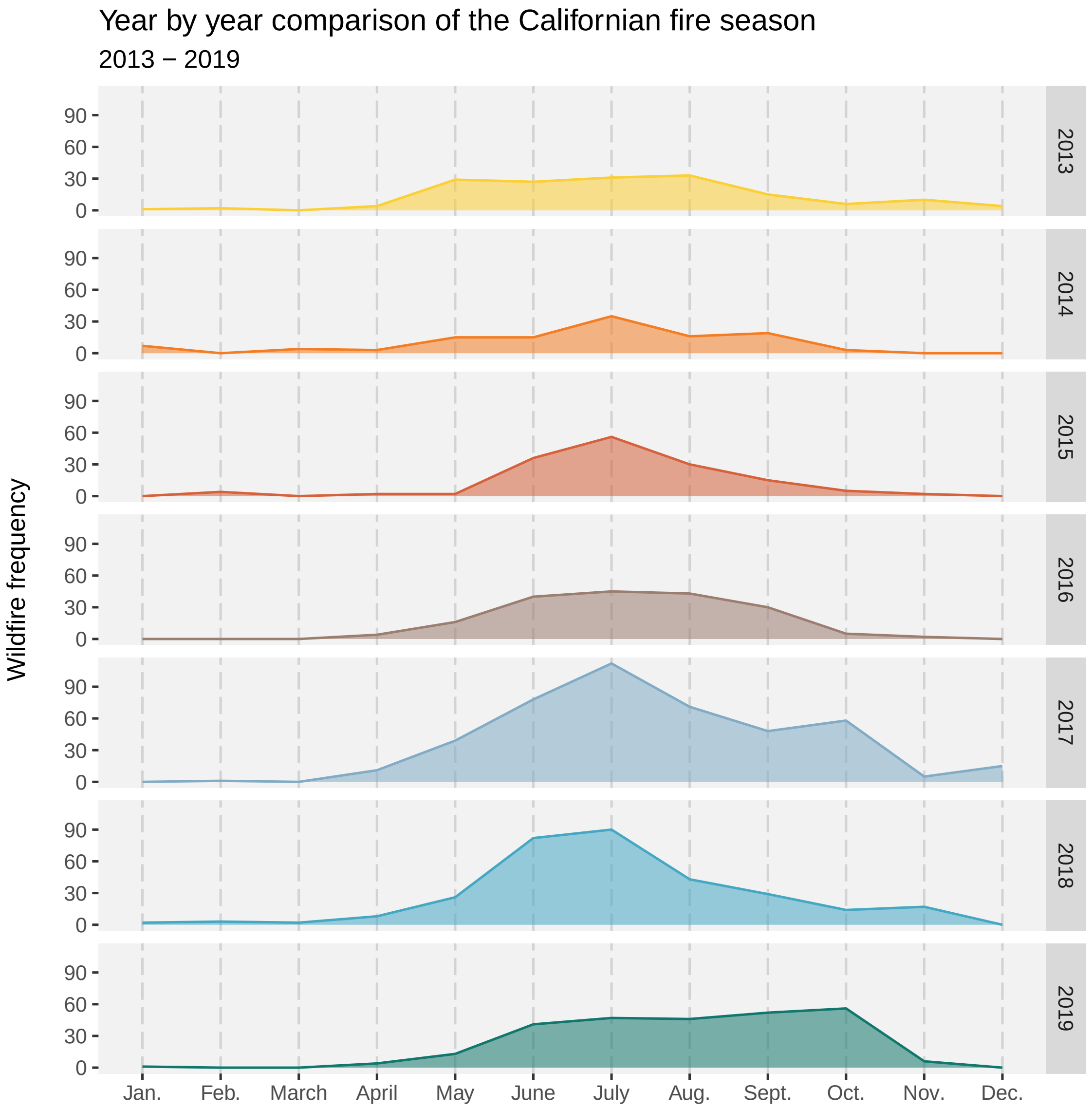

Figure 4: *Year by year comparison of the Californian fire season. Each subplot represents a year in the range 2013 - 2019 and shows the monthly frequency of wildfires for that year. There is no obvious trend in when the fire season starts. In addition to increase in wildfire frequency, the the fire season seems to run longer into autumn during the later years in the range.*

# 4 Task 4. Discussion and conclusion:

## 4.1 Figures 1 & 2

For Figures 1 and 2 I wanted to highlight the disparity between which areas experienced the highest number of wildfires and those areas in which the largest amount of land area was consumed in wildfire events.

Wildfires were by far the most frequent in southern California. Riverside county in southern California especially stood out, with almost 150 recorded wildfires between 2013 and 2019. However, most of these fires were small and inconsequential. This area of California is densely populated, and a large portion of the fires in Riverside occured in urban environments. Most of these fires caused little to no damage and were extingusished relatively quickly.

Naturally, a wildfire is more likely to spread in an area of natural vegetation than in an urban environment where where the amount of fuel is limited and fire sites are generally easily accessible for extinguishing work, but they are also more likely to be started due to human activities. It is possible that the observed differences in the frequency of wildfires between counties is caused in part by differing recording practices and guidelines between administrative units. This would introduce considerable bias towards those counties where smaller wildfire incidents are more diligently registered into the CAL FIRE system; As a point of comparison, no wildfire events were recorded in San Fransisco county in the same time period, despite its much higher population density.

Since I found this information to be both much more interesting and more relevant to the theme of the assignment, I decided to only include those wildfire incidents which were recorded in CAL FIRE's database as *«major events»*. This brought Riverside county more in line with the rest of the dataset, with 18 recorded major wildfire incidents. This is still the fifth highest number out of all the counties.

Most people living in Norway, including myself, are not very familiar with Californian geography. I contemplated including text labels to identify individual counties. However, I found that this resulted in a cluttered look and severely limited readability.

## 4.2 Figure 3

Figure 3 does a good job of showcasing the cyclical nature of the Californian fire season. Predictably, the majority of wildfires incidents occur during the summer months, while only a few occur in the winter months. The visualisation also does a good job at showcasing the general increase in wildfire frequency over the years. There is some variation, e.g. the number of wildfire incidents in 2017 is higher than the number in each of the following years, i.e. 438 wildfires in 2017 vs. 316 and 266 wildfires in 2018 and 2019 respectively, but each year from 2017 forwards saw a higher frequency of wildfire incidents than any year before 2017.

## 4.3 Figure 4

Figure 4 was created with the primary goal of comparing the timing of the fire season between years. I hypothesised that the fire season would start sooner in the spring and run later into autumn during the later years in the time frame.

Contrary to my expectation, there did not seem to be a trend towards an earlier onset of the fire season. There was however a visible trend towards an increase in the duration of the fire season. In each year after (and including) 2017, the fire season lasted well into October. In each year before 2017, the fire season ended in September (with occasional wildfires occuring in later autumn). Keep in mind that the visualisation (though useful) does not convey any information about statistical significance.

Furthermore, a sample size of only seven years is very limited, especially when looking at long-term trends. In the case of global climate change, a real trend may be masked by natural oscillations in temperatures and weather patterns, potentially for decades (Ghil & Vautard, 1991). If the time frame is of insufficient length to pick up on these overarching patterns, our conclusions may be biased in one direction or another simply depending on where we are in the cycle. Additionally, when the sample size is small, stochastic variation between observations can have a much great impact on estimates and conclusions drawn from the data, e.g. if one observation is extreme. In the context of this dataset, a much greater number of wildfire incidents occured in 2017 than in each of the years both before and after, which could be caused by random variation.

## 4.4 What have I learned?

Going into this assignment, I spent a significant amount of time searching for a dataset which I found appealing. Since I have a background in natural science and academia, I approached the assignment from a very specific angle. During this process, I focused on finding a dataset which was relevant to my field rather than look for one with which I could tell an appealing story through visualisation. As a result, I spent a lot of time looking before I could get started with any actual work. While I am satisfied with the dataset that I ended up choosing, I realise in retrospect that it would have been wiser to be less critical upfront and perhaps allow myself to experiment a bit with the data before rejecting it.

I found the process of creating the visualisations very enjoyable. Overall, I think this has been a very useful excercise in viewing data from different angles and thinking critically about what information can be extracted from it. I think I could done a more thorough job with the user testing, but lacking previous experience I learned that during the process.

I would also have liked to explore other aspects of the dataset, especially relating to injuries and physical damage to buildings and other structures, but I felt that this would be best done in combination with data from external sources.

## 4.5 Conclusion

For this assignment I wanted to answer the question about when and where wildfires occur in California during the fire season. I was able to visualise information about the spatial distribution of wildfires by using map visualisation techniques and show in which Californian counties wildfires most frequently occur, and in which counties the largest amount of land area was impacted in the relevant time frame. I was also able to highlight an increase in the frequency of wildfire incidents between 2013 and 2019, as well as an increase in the duration of the fire season during the same period of time.

Visualisation can be a tedious process. With the right tools, it can be easy to create a nice-looking graph, but that does not necessarily mean that the visualisation conveys any meaningful information, or that it conveys the information in a meaninful manner, i.e. one which makes the information easy to parse for the reader. To avoid these pitfalls, it is important to get familiar with the data that you're working with and to identify the objects and attributes which are most relevant to the story you wish to tell.

However, once you have found a suitable *angle of attack*, a good visualisation can make information which would otherwise require a detailed understanding of the underlying data accessible to anyone at a glance. Visualisation can even open up entirely new ways to explore data, especially if the dataset consists of a very large number of observations or if the data is highly multidimensionsional.

# REFERENCES

ARES. (2020, February). *California WildFires (2013-2020)*. https://www.kaggle.com/datasets/ananthu017/california-wildfire-incidents-20132020 (Accessed on February 24, 2023)

CAL FIRE. (2022, October). *California county boundaries*. California Open Data Portal. https://data.ca.gov/dataset/california-county-boundaries2 (Accessed on March 14, 2023)

Delfino, R. J., Brummel, S., Wu, J., Stern, H., Ostro, B., Lipsett, M., Winer, A., Street, D. H., Zhang, L., Tjoa, T., et al. (2009). The relationship of respiratory and cardiovascular hospital admissions to the southern california wildfires of 2003. *Occupational and Environmental Medicine*, *66*(3), 189–197.

Diffenbaugh, N. S., Swain, D. L., & Touma, D. (2015). Anthropogenic warming has increased drought risk in california. *Proceedings of the National Academy of Sciences*, *112*(13), 3931–3936.

Ghil, M., & Vautard, R. (1991). Interdecadal oscillations and the warming trend in global temperature time series. *Nature*, *350*(6316), 324–327.

Kunzli, N., Avol, E., Wu, J., Gauderman, W. J., Rappaport, E., Millstein, J., Bennion, J., McConnell, R., Gilliland, F. D., Berhane, K., et al. (2006). Health effects of the 2003 southern california wildfires on children. *American Journal of Respiratory and Critical Care Medicine*, *174*(11), 1221–1228.

Parks, S. A., & Abatzoglou, J. T. (2020). Warmer and drier fire seasons contribute to increases in area burned at high severity in western US forests from 1985 to 2017. *Geophysical Research Letters*, *47*(22), e2020GL089858.

Posit team. (2022). *RStudio: Integrated development environment for R*. Posit Software, PBC. http://www.posit.co/

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

United Nations General Assembly. (2015). *Transforming our world: The 2030 agenda for sustainable development. Draft resolution referred to the United Nations summit for the adoption of the post-2015 development agenda by the General Assembly at its sixty-ninth session*. United Nations. UN Doc. A/70/L.1 of 18 September 2015.

# Appendices

## Appendix A: The raw R code

```r
### Import packages
library(dplyr)
library(ggplot2)
library(lubridate)
library(scales)
library(sf)

### Loading the data
# Import the dataset:
fires <- readr::read_csv("data/California_Fire_Incidents.csv")
fires[1020,32] <- as.POSIXct("2017-05-01")   # Manually correct dates which were
fires[1262,32] <- as.POSIXct("2018-08-01")   # incorrectly recorded.

# Import shapefiles for mapping geodata:
# california <- st_as_sf(st_read("data/California_State_Border/california.dbf"))
county_map <- st_as_sf(st_read("data/California_County_Boundaries/cnty19_1.dbf",
                               stringsAsFactors = FALSE))

# Add column "Frequency" (of fires) to 'county_map' dataframe:
county_map <- left_join(county_map,                             # Join this...
                        fires$Counties %>%                      # ...to this.
                          table() %>%
                          as.data.frame(stringsAsFactors = FALSE) %>%
                          select(COUNTY_NAM = 1,   # Match colname in 'county_map'
                                 Frequency  = 2))  # New column to be inserted

# Add column "MajorFreq" (i.e. the same as above, but only major events):
county_map <- left_join(county_map,
                        fires[which(fires$MajorIncident == TRUE), "Counties"] %>%
                          table() %>%
                          as.data.frame(stringsAsFactors = FALSE) %>%
                          select(COUNTY_NAM = 1,
                                 Major_freq = 2))

# Add column "Acres_burned" to 'county_map':
county_map <- left_join(county_map,
                        fires[,c("Counties", "AcresBurned")] %>%
                          group_by(Counties) %>%
                          summarise(Acres_burned = sum(AcresBurned)) %>%
                          select(COUNTY_NAM = 1,
                                 Acres_burned = 2))

# Add column "Acres_major" (i.e. the same as above, but only major events):
county_map <- left_join(county_map,
                        fires[which(fires$MajorIncident == TRUE),c("Counties",
                                                                   "AcresBurned")] %>%
                          group_by(Counties) %>%
                          summarise(Acres_major = sum(AcresBurned)) %>%
                          select(COUNTY_NAM = 1,
                                 Acres_major = 2))

# Create list of months from January 2013 - December 2019:
months <- seq(from = ym('2013-01'),
                to = ym('2019-12'),
                by = 'month') %>%
  as.data.frame() ; colnames(months)[1] <- "month"

### Figure 1 - Fires by county
# Determine which colours are to be used in the gradient:
cols <- c("white", "lightgoldenrod1", "orange2", "orangered3", "firebrick4")
# Specify theme:
dust_theme <- ggthemr::ggthemr("dust", set_theme = FALSE)

# Create visualisation:
county_map %>%
  ggplot() +
  geom_sf(aes(fill = Major_freq)) +
  scale_fill_gradientn(name = "Freq.",
                       colors = cols,
                       breaks = c(0, 10, 20, 30),
                       labels = c(0, 10, 20, 30),
                       limits = c(0, 30),
                       na.value = "white",
                       guide = guide_colorbar(barheight = 10,
                                              draw.ulim = F,
                                              draw.llim = F,
                                              title.vjust = 3)) +
  labs(title = "Frequency of major wildfire events in California by county",
       subtitle = "2013 - 2019") +
  dust_theme$theme

# The plot above includes only the frequency of major wildfire events, as recorded
# by CAL FIRE. To include all wildfire events, incl. minor events, use the following
# code instead:

# fires_by_county <- ggplot(data = county_map) +
#   geom_sf(aes(fill = Frequency)) +
#   scale_fill_gradientn(name = "Freq.",
#                        colors = cols,
#                        breaks = c(0, 50, 100, 150),
#                        labels = c(0, 50, 100, 150),
#                        limits = c(0, 150),
#                        na.value = "white",
#                        guide = guide_colorbar(barheight = 10,
```

```r
#                                               draw.ulim = FALSE,
#                                               draw.llim = FALSE,
#                                               title.vjust = 3)) +
#   labs(title = "Frequency of wildfires in California by county",
#        subtitle = "2013 - 2019") +
#   dust_theme$theme

# Create visualisation:
county_map %>%
  ggplot() +
  geom_sf(aes(fill = Acres_burned * 0.004047)) +   # Convert to km²
  scale_fill_gradientn(name = "Area (km²)",
                       colors = cols,
                       na.value = "grey",
                       guide = guide_colorbar(barheight = 10,
                                              draw.ulim = FALSE,
                                              draw.llim = FALSE,
                                              title.vjust = 3,
                                              order = 2)) +
  ggnewscale::new_scale_fill() +
  geom_sf(data = subset(county_map, is.na(Acres_burned)),   # Create 'No data' legend
          aes(fill = "grey")) +
  scale_fill_manual(name = NULL,
                    labels = "No data",
                    values = "grey",
                    guide = guide_legend(override.aes = list(linetype = 0))) +
  labs(title = "Area consumed in wildfires in California by county (km²)",
       subtitle = "2013 - 2019") +
  dust_theme$theme

### Figure 4 - Yearly comparison of fire season:
cols <- yarrr::piratepal(palette = "nemo")   # Set new colour palette

# Create new data frame:
fire_by_month <- fires %>%
  group_by(month = floor_date(fires$Started,
                              unit='month')) %>%
  summarise(n = n()) %>%
  right_join(., months) %>%
  mutate(year = year(month),
         month = month(month)) %>%
  data.table::setnafill(fill = 0)

# Create visualisation:
fires_by_month %>%
  ggplot(aes(x = month,
             y = n)) +
  geom_line(aes(col  = year)) +
  geom_area(aes(fill = year,
                group = year,
                alpha = 0.5)) +
  scale_colour_gradientn(colours = cols) +
  scale_fill_gradientn(colours = cols) +
  facet_wrap(vars(year),
             ncol = 1,
             strip.position = "right") +
  theme(legend.position = "none",
        panel.grid.minor = element_blank(),
        panel.grid.major.y = element_blank(),
        panel.grid.major.x = element_line(colour = "lightgrey",
                                          linetype = "longdash"),
        panel.background = element_rect(fill = "gray95"),
        axis.title.x = element_blank()) +
  labs(title = "Year by year comparison of the Californian fire season",
       subtitle = "2013 - 2019",
       y = "Wildfire frequency\n") +
  scale_x_continuous(breaks = 1:12,
                     labels = c("Jan.","Feb.", "March", "April",
                                "May", "June", "July", "Aug.",
                                "Sept.", "Oct.", "Nov.", "Dec."))

### Figure 3 - Line chart
# Run this code after Fig. 4 to avoid issues caused by ggthemr::theme_reset()).

# Create new data frame:
fire_points <- fires %>%
  group_by(month = floor_date(fires$Started,
                              unit='month')) %>%
  summarise(n = n()) %>%
  right_join(., months) %>%
  data.table::setnafill(fill = 0)

ggthemr::ggthemr("dust")   # Set theme.

# Create visualisation:
fire_points %>%
  ggplot(aes(x = month,
             y = n)) +
  geom_smooth(method = lm,            # Calculates a smooth line using the
              col = "firebrick1") +   # formula 'month ~ n'
  geom_line() +
  geom_point(col="steelblue3") +
  scale_x_datetime(date_labels = "%b\n%Y",
                   breaks = 'year') +
  labs(title = "Frequency of wildfires in California per month",
       subtitle = "2013 - 2019",
       y = "Frequency") +
  theme(axis.title.x = element_blank()) +
  geom_text(data = fire_points[which(fire_points$n == max(fire_points$n)),],
            mapping = aes(label = "Jul. 2017",
                          fontface = 2),
            hjust = -0.2,
            size = 3.5)
```