# Project Proposal

Info Innovators - Kevin Mao, Arnav Meduri, Ben Trokenheim, Ricardo Urena

```r
library(tidyverse)
library(tidymodels)
library(knitr)
# add other packages as needed
```

```r
wildfire <- read_csv("data/FW_Veg_Rem_Combined.csv")
```

## Introduction

Wildfires are among the most destructive natural disasters, causing significant environmental, economic, and human losses. Just this past year, California has experienced some of the most severe and costly wildfires in history, with fires destroying millions of acres, displacing thousands of residents, and costing billions in damages. The increasing frequency and intensity of these fires demonstrates the urgent need for improved wildfire response strategies. As climate change affects drought conditions and fuel availability, accurately predicting fire containment time is becoming more important for firefighting agencies and policymakers.

In this project, we will analyze wildfire characteristics, environmental conditions, and geographical factors to predict the time required to contain a fire. By using historical wildfire data, weather patterns, and vegetation information, we hope to identify key determinants of fire containment time and develop a predictive model to aid decision-making in fire management.

Research Question:

What factors influence the time it takes to put out a wildfire, and can we develop a predictive model to estimate wildfire containment time based on environmental and fire-specific variables?

Motivation & Importance

Wildfire response teams must quickly assess fire severity and determine the best course of action to contain and extinguish fires efficiently. However, limited firefighting resources often require prioritization based on containment difficulty. If containment time can be accurately predicted based on fire size, vegetation type, weather conditions, and remoteness, fire departments and policymakers can make more informed decisions about resource allocation/risk mitigation.

The relevance of this research is especially important given the recent wildfires California. By understanding the key factors that impact containment time, fire agencies can allocate resources more effectively and reduce response times, which could potentially reduce the destruction caused by wildfires.

Beyond California, predicting containment time is useful for national and global wildfire management and could help with disaster preparedness and and policy development efforts. By identifying the most influential factors in fire containment, we hope to provide insights to improve containment planning.

Hypothesis

We hypothesize the following relationships between key variables and wildfire containment time:

- Larger fires will take longer to contain
- Weather conditions significantly impact containment time. Higher temperatures, stronger winds, and lower humidity levels will be associated with longer containment times, while increased precipitation will reduce containment time
- Vegetation type influences containment difficulty. Fires in densely forested areas may take longer to contain than those in grasslands or shrublands (due to fuel availability).
- Remoteness increases containment time. Fires located farther from cities and firefighting infrastructure will have longer putout times due to delayed response efforts.

## Data description

The dataset we will be using in this project is a subset of U.S. wildfire data, supplemented with additional attributes related to weather, vegetation, and fire remoteness. The data is compiled from four sources:

Sources:

1. Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

   This dataset provides information on 1.88 million recorded wildfires in the United States, including information fire size, cause, location, and containment time. A random sample of 50,000 fires was selected for this dataset. The dataset integrates records from the Fire

Program Analysis (FPA) Fire-Occurrence Database (FOD), which consolidates information from agencies such as the U.S. Forest Service, Bureau of Land Management, and National Park Service.

2. NOAA National Centers for Environmental Information (2001): Integrated Surface Hourly (1992-2015) - https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database

   NOAA gathers Integrated Surface Hourly data from thousands of global weather stations, including those managed by the National Weather Service, the Federal Aviation Administration, and international meteorological organizations. The data include hourly observations of temperature, wind speed, humidity, and other meteorological variables, primarily collected through automated and manual weather stations.

3. Meiyappan, Prasanth, and Atul K. Jain. "Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years." Frontiers of Earth Science 6.2 (2012) - https://bg.copernicus.org/preprints/11/C2254/2014/bgd-11-C2254-2014-supplement.pdf

   This dataset provides information on the dominant vegetation types in the areas where fires occurred (which can impact fire behavior). Records in this datase span from 1765 to 2010.

4. "World Cities Database." Simplemaps, https://simplemaps.com/data/world-cities

   Simplemaps compiles city location data from various authoritative sources, including government databases, geographic surveys, and other publicly available records. The dataset includes information such as city names, coordinates, population estimates, and administrative divisions.

The compiled dataset (which we will be using in our analysis) consists of 43 variables, including attributes such as fire name, size, class, cause, location (latitude/longitude, state), discovery month, containment time, and environmental conditions before and during the fire event. Weather-related variables include temperature, wind speed, humidity, and precipitation recorded at multiple time points (30, 15, and 7 days before containment, as well as on the day the fire was contained).

## Data Processing

All of our data is contained in the `wildfire` data set, so we will not need to join any data sets. We will need to handle missing data values, especially in our response variable column (`putout_time`) and columns that we intend to use as predictors. We will need to look at each column on a case by case basis to determine if we want to drop NA values or if we are able to interpolate values. Finally, there are some potential outliers in the data set, which we will need to investigate further.

One data processing technique we might use is to potentially convert `cont_date_final` and `disc_date_final` into DateTime variables. These two columns not only contain the date, but also a recorded time. Currently, `putout_time` is only measured in integer values of days, which leads to a generalization of putout time that especially impacts fires with shorter durations. By converting the aformentioned columns into DateTime variables, we could find the fire duration in terms of hours as opposed to whole days. Other data processing techniques we might use are converting data types (character to numeric, etc.), extracting values, and potentially re-scaling variables.

Before investigating our response variable, we need to do some data processing and cleaning. To begin, we drop NA values from the `putout_time` column. Next, we need to extract the number of days it took to put out the fire, and we need to convert this value to a numeric. The website where we found the code to do this is: https://www.statology.org/r-extract-number-from-string/. We also include a check to ensure that we properly converted the values to the numeric type.

```r
wildfire_copy <- wildfire

wildfire_copy <- wildfire_copy |>
  filter(!is.na(putout_time))

wildfire_copy <- wildfire_copy |>
  mutate(
    putout_time_num = parse_number(putout_time)
  )

is.numeric(wildfire_copy$putout_time_num)
```

```
[1] TRUE
```

```r
wildfire_copy |>
  summarize(
    mean = mean(putout_time_num),
    median = median(putout_time_num),
    min = min(putout_time_num),
    max = max(putout_time_num),
    sd = sd(putout_time_num),
    Q1 = quantile(putout_time_num, 0.25),
    Q3 = quantile(putout_time_num, 0.75),
    IQR = IQR(putout_time_num)
  ) |>
  kable()
```

| mean | median | min | max | sd | Q1 | Q3 | IQR |
|---|---|---|---|---|---|---|---|
| 6.033592 | 0 | 0 | 3287 | 27.80376 | 0 | 1 | 1 |

```
wildfire_copy |>
  count(putout_time_num) |>
  arrange(desc(n)) |>
  slice_head(n = 10)
```

```
# A tibble: 10 x 2
   putout_time_num     n
             <dbl> <int>
 1               0 18680
 2               1  2598
 3               2   940
 4               3   616
 5               4   457
 6               5   349
 7               6   300
 8               7   249
 9               8   183
10               9   154
```

This output shows the fire put out durations (in number of days) with the highest frequency. This shows that the majority of our data is clustered in low fire put out times, which we will also show using visualizations.

```
(18680)/nrow(wildfire_copy)
```

```
[1] 0.6798413
```

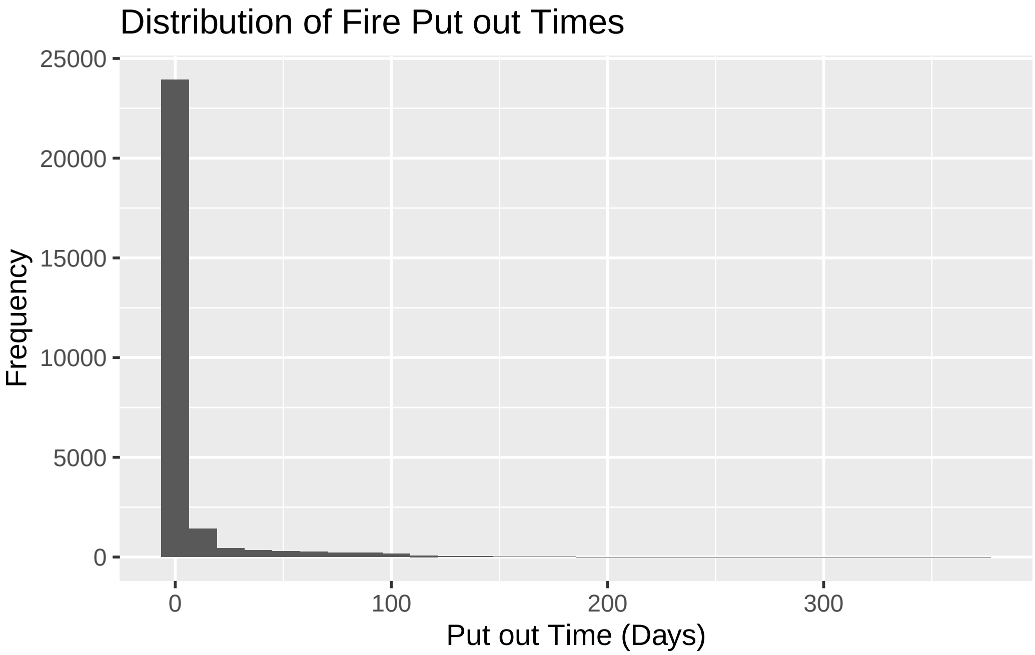
Fire put out times of 0 days account for more than half of the data in our data set. This also suggests that using DateTime variables to get more specific put out times in terms of hours might be beneficial.

```
wildfire_copy |>
  arrange(desc(putout_time_num)) |>
  select(putout_time_num)
```

```
# A tibble: 27,477 x 1
   putout_time_num
             <dbl>
 1            3287
 2             371
 3             312
 4             261
 5             255
 6             237
 7             222
 8             213
 9             211
10             204
# i 27,467 more rows
```

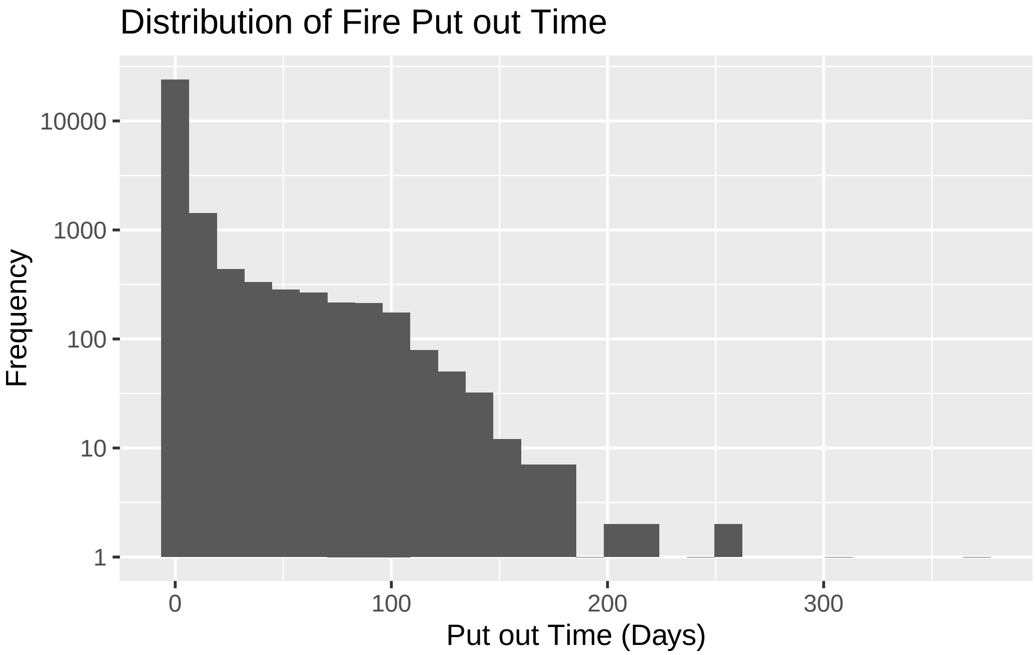
Looking at the output, the longest fire put out time of 3287 days is far longer than the second longest fire put out time of 371 days. For readability of the histogram, we will exclude the data point with a putout time of 3287 days, as it will stretch the horizontal axis too much.

```r
wildfire_copy |>
  filter(putout_time_num <= 400) |>
  ggplot(aes(x = putout_time_num)) +
    geom_histogram() +
    labs(title = "Distribution of Fire Put out Times",
         x = "Put out Time (Days)",
         y = "Frequency")
```



Because the frequency of fire putout times of 0 days and 1 day is so high, it makes it hard to read the histogram. Thus, we can re-scale the frequencies on the log scale. The website where we found the information to do this is: https://www.r-bloggers.com/2021/08/beginning-a-ggplot2-series-logarithmize-your-scales/.

```
wildfire_copy |>
  filter(putout_time_num <= 400) |>
  ggplot(aes(x = putout_time_num)) +
    geom_histogram() +
    scale_y_log10() +
    labs(title = "Distribution of Fire Put out Time",
         x = "Put out Time (Days)",
         y = "Frequency")
```



The distribution of fire put out times is unimodal and highly right skewed, with most of the values being low. This is shown by the visualizations and the fact that the mean (6.03 days) is far higher than the median (0 days). It has an IQR of 1 day, with the 25th percentile at 0 days and the 75th percentile at 1 day. The maximum value put out time of 3287 days and put out times greater than or equal to 250 days appear to be potential outliers. ADD DISCUSSION OF STANDARD DEVIATION?

### Analysis approach

Since the goal of our project is to understand how different fire characteristics and weather, geographic, and location-based factors influence fire containment time, `putout_time` will be the response variable in our analysis. So far, we have identified the following variables as candidate predictors that may help explain the variability in `putout_time`:

Numerical variables:

- Fire Size (`fire_size`): The total area burned by the fire (acres). We think this would be a strong predictor because larger fires generally take more time to contain (Fires with a greater burned area require more resources and extended efforts from firefighting teams.)
- Fire Intensity (`fire_mag`): A scaled measure of fire intensity based on fire size. We think this would be a strong predictor because more intense fires may be harder to control, and higher fire intensity might be correlated with faster fire spread and greater difficulty in containment efforts.

- Temperature on Containment Day (`Temp_cont`): The temperature in degrees Celsius at the fire location on the day the fire was contained. We think this would be a strong predictor because higher temperatures can dry out vegetation and increase fire spread, making containment more difficult. Lower temperatures might help slow fire progression.
- Wind Speed on Containment Day (`Wind_cont`): The wind speed in meters per second at the fire location on the day the fire was contained. We think this would be a strong predictor because stronger winds can spread fires rapidly and therefore make containment more challenging. Fires in areas with low wind speeds may be easier to manage.
- Precipitation 7 Days Before Fire Containment (`Prec_pre_7`): The amount of precipitation in millimeters at the fire location in the seven days leading up to containment. We think this would be a strong predictor because recent precipitation can increase soil and vegetation moisture, whic could reducing fire intensity and help with containment efforts.

Categorical variables:

- Cause of Fire (`stat_cause_descr`): The reported cause of the fire (e.g., lightning, human activity, or equipment use). We think this would be a strong predictor because fires caused by natural events like lightning might have different containment challenges compared to those caused by human activities (i.e., some causes may be associated with fires that spread more aggressively).
- Dominant Vegetation Type (`Vegetation`): The main form of vegetation present in the fire-affected area (categorized into forest, shrubland, grassland, or urban land). We think this would be a strong predictor because different vegetation types burn at different rates; for example, denser forests/vegetation may sustain fires for longer periods of time, while grasslands might allow for quicker containment.
- Remoteness (`remoteness`): A non-dimensional measure of the distance to the closest city.
  We think this would be a strong predictor because fires in remote areas may take longer to contain becuse of limited access to firefighting resources and difficulty in transporting equipment (delayed response times).

To understand the relationship between each of our candidate predictors and fire containment time, we will first conduct EDA. For numerical predictors, we will use correlation coefficients to assess their strength of association with fire containment time. For categorical predictors, we will mainly rely on visualizations to determine whether there are significant differences in fire containment time across different levels of categorical variables (e.g., side-by-side boxplots to compare the distribution of fire containment time across different categories).

Based on the strongest predictors identified in our exploratory analysis, we will fit multiple linear regression models to predict fire containment time using a combination of the selected variables. In this step, we may need to sample from our data rather than using the entire dataset, since since our dataset consists of time-series attributes, which, as we learned in class, may violate the independence criterion for regression (There may be correlations between fire incidents occurring close in time or similar environmental conditions.) Multiple linear regression is the most appropriate modeling approach for this project because our response variable (fire containment time) is a continuous, numerical variable; since we are not predicting a categorical outcome, logistic regression would not be applicable. As part of our modeling methodology, we will fit both main effects models and interaction effects models (Interactions will be included based on factors that may have a different relationship with fire containment time depending on another category based on EDA – for example, fire size may have a different effect on fire containment time depending on vegetation type).

We will fit models with 95% confidence intervals for coefficients to estimate the range in which the true slope of the relationship between each predictor and fire containment time is likely to fall. After fitting our models, we will use model assessment metrics to determine which models provide the best fit. We will evaluate models based on adjusted $R^2$ (which accounts for the number of predictors and helps identify models that explain the most variability in fire containment time) and RMSE (average difference between observed and predicted values of the response variable). Another technique that might be useful (which we covered in class) is standardizing numerical variables in our model so we can compare each of the numerical variables and determine which are the most "impactful" predictors of fire containment time.

Based on the performance of each of our models, we will be able to draw conclusions about which factors most strongly influence fire containment time and how different environmental and fire-related conditions contribute to variability in putout time.

## Data dictionary

The data dictionary can be found here.