Using Climate Variables to Predict the Magnitude of California Wildfires

Group Project – Final Report

MGT 6203 – Data Analytics for Business

Georgia Institute of Technology

Jason Ho

Andrew Snider

Vetrivel Kanakasabai

Jordan Chen

(Team 23)

Fall 2022

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

1.1 BACKGROUND INFORMATION

Wildfires threaten lives, properties, and natural resources. Nearly 7.5 million acres of forest are lost to wildfires across the United States each year. In the United States, California (CA) sees the worst wildfires compared to any other state, with 2.5 million acres burned in 2021 and 4.3 million acres burned in 2020. Many of the largest wildfires in U.S. history have happened in California, with eight wildfires reaching mega-fire status, surpassing 100,000 acres burned. In 2021, a single fire in the city of Dixie burned close to a million acres and spread across four counties. These wildfires can start due to natural causes, like lightning strikes, or by humans, like a campfire mistake or fallen powerline. Environmental factors play a large role in contributing to a fire's sheer size and intensity. Some of these contributing factors included California being in the most extreme megadrought in 1,200 years, having the hottest summer ever recorded, and receiving less than half of its normal precipitation. This same pattern was seen in the 2020 wildfire season, a record-setting season with a higher seasonal fire risk than usual due to arid periods in the months of January and February.

Weather conditions play a crucial role in a wildfire's spread. For example, high temperatures raise the flammability of dry grass, leaves, and trunks, while high wind speeds increase the pace at which wildfires spread, droughts both intensify and prolong forest fires, and lightning can trigger ignitions in dry vegetation. Current and upcoming weather conditions and their impact on forests need to be carefully considered when planning operations to prevent future wildfires. State governments and agencies can prepare and implement strategies based on current and forecasted wildfires. Increased patrols by fire service professionals with fire detection equipment and regular inspections in fire-prone areas can be carried out. This demands for wildfire prediction and prevention strategies.

Wildfires in California are dangerous natural disasters that put homes, businesses, and people at risk of loss and even death while destroying the environment and wildlife habitats in the surrounding area. By predicting and preventing wildfires, California's natural beauty and ecosystems can be preserved, reducing carbon pollution, and mitigating personal and statewide losses. The total number of acres affected in the U.S in the year 2021 was 7,125,643 and the total amount of costs incurred to suppress annual wildfires in the year 2021 was $4,389,000,000. This means the amount spent fighting wildfires for 2021 was $616 per acre and our goal with this project is to reduce this number.

1.2 Overview of project and Initial Hypothesis

The purpose of our analysis and investigation of California wildfire history is to predict the size of wildfires (in total acres burned) for future wildfires that might take place in order to allocate proper resources, mitigate risk, and make each impacted region safer from primary wildfire causes. Can the climate variables be used to predict the potential scale of wildfires by county/region in California, to help state forestry departments and other contributing agencies better prepare vital resources for the improved prevention and suppression efforts of future wildfires?

1.3 Planned Approach

Our approach to this project is to gather publicly available data on California's weather history, such as precipitation, temperature, drought, etc. on a given date, as well as data on its past wildfires (date, location, total acres burned). We will take this data, clean it, then merge it into one dataset based on California's 58 counties and date. With this dataset, we will perform exploratory data analysis, report our findings, and then move into the modeling part of our project. The primary model we plan to use is linear regression but also plan to use time series or logistic regression model as an alternative.

2 DATA

2.1 DATA SOURCES

2.1.1 Primary Datasets:

1. California fire perimeters
2. Climatological data annual summary for California

2.1.2 Supplementary Datasets:

1. Stations and counties list from here on the last pages of each year's extract, then scraped in Excel to get the lists
2. Admin units regions, helpful for google searching locations and mapping counties, admin units, and stations

2.2 DATA CLEANING

2.2.1 Initial cleaning in Excel / Google Sheets:

To join the primary data sources, each dataset needed common IDs or lookup values. The CA Counties and Admin Units list allowed us to fill in county names and admin units to their appropriate station name. Then, climatological data (precipitation and temperature data) were summarized for each station at the monthly level. Stations that appear in the precipitation and temperature data were mapped to their appropriate county. We leveraged the map in supplementary datasets and Google Maps to locate the station and assign each station to the closest admin unit and county. From the original station value, we created a new "station_clean" to better group stations together, such as names that were similar with some slight variation. Several values in the precipitation and temperature dataset had a mix of numbers and text, so a flag was added as a new column for each measure to indicate whether that data point was converted from a string to number by removing text. The actual conversion were executed in R, but the flags were added in Excel. Finally, stations, admin units, and counties were further cleaned up to prepare for vlookups into the fire perimeter's raw data. Admin unit, station_clean, and county were brought over into this dataset.

2.2.2 Subsequent cleaning in R:

Helpful libraries for further data cleaning include dplyr, xts, tidyverse, and tidyr. First, looking at the precipitation and temperature data, we replaced long dashes (–) with standard short dashes (-), and created a flag to indicate which rows were cleaned with this method. Next, we converted dates from a text format to YYYY-MM Date type. Then we added a function to extract only numeric values from a data point using a regular expression (source: StackOverflow). All measurements of precipitation and temperature were converted to numeric from string. Blanks and NAs were replaced with 0s, for no reading or insufficient data recorded.

Next, looking at the fire perimeters dataset, the alarm date value is by day, to represent the beginning of when the fire was recorded. This data was converted first from string to Date, then to only year and month YYYY-mm format, to match the precipitation and temperature monthly summaries. Then, we performed a group_by on admin unit and alarm year-month, summing by acres burned for that admin unit-month period. Finally, we joined the precipitation and temperature data with the grouped fire perimeters data by admin unit-month using an outer join. The data was exported as a CSV file using write csv in R, which will be used for modeling and EDA.

3 DATA EXPLORATION

3.1 Initial Data Exploration

In the data exploration phase of our research, our group aims to better familiarize our newly cleaned data set, examining the data from several different vantage points, to extract any significant insights that may influence our decisions when approaching our data modeling. We begin by printing a simple summary of the dataset and reviewing the results for any substantial outliers in the data set.

```
> summary(data)
  ADMIN_UNIT          DATE_MONTH              YEAR          MONTH          SUM_PRECIP
 Length:6804        Min.   :2001-01-01   Min.   :2001   Min.   : 1.00   Min.   :   0.000
 Class :character   1st Qu.:2006-03-24   1st Qu.:2006   1st Qu.: 3.75   1st Qu.:   0.920
 Mode  :character   Median :2011-06-16   Median :2011   Median : 6.50   Median :   8.215
                    Mean   :2011-06-16   Mean   :2011   Mean   : 6.50   Mean   :  26.026
                    3rd Qu.:2016-09-08   3rd Qu.:2016   3rd Qu.: 9.25   3rd Qu.:  30.430
                    Max.   :2021-12-01   Max.   :2021   Max.   :12.00   Max.   :1173.060
 SUM_PRECIP_DEPART   SUM_TEMPERATURE   SUM_TEMPERATURE_DEPART   AVG_PRECIP       AVG_PRECIP_DEPART
 Min.   : -305.440   Min.   :    0.0   Min.   :-290.40        Min.   : 0.00000   Min.   :-8.0371
 1st Qu.:   -8.835   1st Qu.:  263.9   1st Qu.:  -2.50        1st Qu.: 0.05885   1st Qu.:-0.6132
 Median :   -1.615   Median :  497.4   Median :   6.90        Median : 0.58569   Median :-0.1071
 Mean   :   10.319   Mean   : 1055.7   Mean   :  20.45        Mean   : 1.78088   Mean   : 0.4237
 3rd Qu.:    3.385   3rd Qu.: 1011.1   3rd Qu.:  24.40        3rd Qu.: 2.40435   3rd Qu.: 0.2284
 Max.   :10024.000   Max.   :41034.4   Max.   :2160.40        Max.   :24.82714   Max.   :56.8750
 AVG_TEMPERATURE AVG_TEMPERATURE_DEPART SUM_GIS_ACRES      AVG_GIS_ACRES
 Min.   : 0.00   Min.   :-6.2857        Min.   :       0   Min.   :     0.0
 1st Qu.:32.00   1st Qu.:-0.2628        1st Qu.:       0   1st Qu.:     0.0
 Median :40.41   Median : 0.6750        Median :       0   Median :     0.0
 Mean   :41.38   Mean   : 0.6817        Mean   :   38679   Mean   :  1005.8
 3rd Qu.:50.17   3rd Qu.: 1.5750        3rd Qu.:     606   3rd Qu.:    32.6
 Max.   :85.50   Max.   : 8.4000        Max.   :10125587   Max.   :963405.4
```

Figure 1: Summary statistics of the data

Looking at the above summary, we can first note our factor variable, ADMIN_UNIT, used to segment the data between 27 regions across the state of California, each representing unique climate zones and each with a unique wildfire history and risk.

Next, we note our date variables, including DATE_MONTH, a mdy() date formatted column, ranging 20 years from Jan 2001 to Dec 2021. The DATE_MONTH column is used to align our data with time-series plots and segmentation. Additionally, we have extracted individual columns for YEAR and MONTH as integer values, which can be used as parameters in our subsequent regression models to retain a time-series element and aid in accurate future predictions.

Moving onto our numeric independent variables, we remark our climate variables, temperature, and precipitation. It is important to note that in our combined data set the numeric variables (precip, temperature, gis_acres) are aggregated across multiple weather stations up to individual admin units, and thus represent a many-to-one join on the data sources. We have elected to handle this potential misalignment by generating both sum and average aggregations (e.g., SUM_PRECIP and AVG_PRECIP), to provide the most comprehensive features list for identifying potential correlations and exhaustive feature selection for optimizing fit. With this note, we observe the sum total precipitation (SUM_PRECIP) ranges from 0 to 1,173.0 inches and averages (AVG_PRECIP) from 0 to 24.82 inches in total. The temperature data ranges from 0 - 41,034.4 degrees Fahrenheit in sum total aggregate (SUM_TEMPERATURE) and ranges from 0 to 85.5 degrees Fahrenheit on average (AVG_TEMPERATURE).

The variables with "_DEPART" suffix represent the recorded departure from the average. As these recorded variances can be above or below the mean, their values too can be positive or negative, and as a result, we see many stats for these variables that are below zero. These departure variables will also be explored for their potential explanatory power in forecasting wildfires.

Lastly, we examine our dependent variable, GIS_ACRES, representing the total acres burned. The sum total acres (SUM_GIS_ACRES), ranges from 0 to a total max of 10.1M acres. While this range understandably seems extreme, we can further note that the median is 0 acres, suggesting that most data points over the last 20 years and admin units saw no significant wildfires at all. Further, the mean is 38,679 acres, and while this also would seem a bit large for an average, our theory is that there is one or just a few extremely large fires in the data set that is drastically skewing the distribution of the data.

3.2 Histogram Plots

In the next portion of our data exploration, we proceed to review a series of histograms, to create visual representations of the distribution of values for each of our numerical variables. Beginning with our independent variables, we first examine the histogram results for precipitation data.

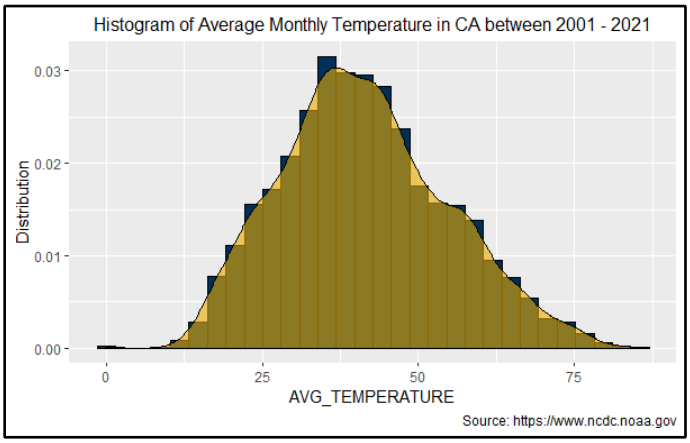


Figure 2: Histogram of AVG_TEMPERATURE

The average total precipitation data, as seen in the accompanying histogram, is a clear right-skewed distribution, exhibiting a long tail trailing off to the right. This is expected, as we saw in the previous summary section, while the range for AVG_PRECIP is extreme, the majority of periods saw very little rainfall, with the median being only 0.6 inches. A large spread such as this may make this variable a good candidate for non-linear transformation to improve fit and will be tested in our data modeling section.

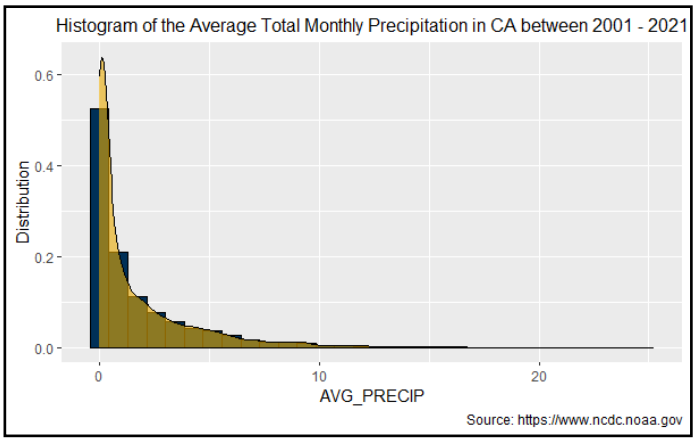


Figure 3: Histogram of AVG_PRECIP

Next, and by strong contrast, the histogram plot for temperature data (AVG_TEMPERATURE) yields a very normal-looking distribution. Further, while the plot does exhibit clear bell-curve qualities, there is possibly a slight right-skew, but this is not unexpected for California to see more warmer days above 41 degrees Fahrenheit than cooler days. With a recognizable normal distribution, we would expect this variable to be more immediately fit to a linear regression modeling.

Lastly, the histogram plot for our dependent variable, SUM_GIS_ACRES, we see that the spread of the data is so incredibly right-skewed, that the data barely registers on the chart. As mentioned previously, due to the extremely uneven distribution of the data, with just a massive few wildfires heavily skewing the distribution, when the majority of periods recorded 0 total acres and many more periods recorded significantly smaller wildfires. We can zoom closer on this effect by filtering the data set to only records where SUM_GIS_ACRES is greater than 0, then look at the lower 50th percentile, as seen in the next chart. Here we confirm the cluster of wildfire acre records centered around zero and rapidly trailing off. This dramatic distribution can provide unique challenges in modeling and forecasting.

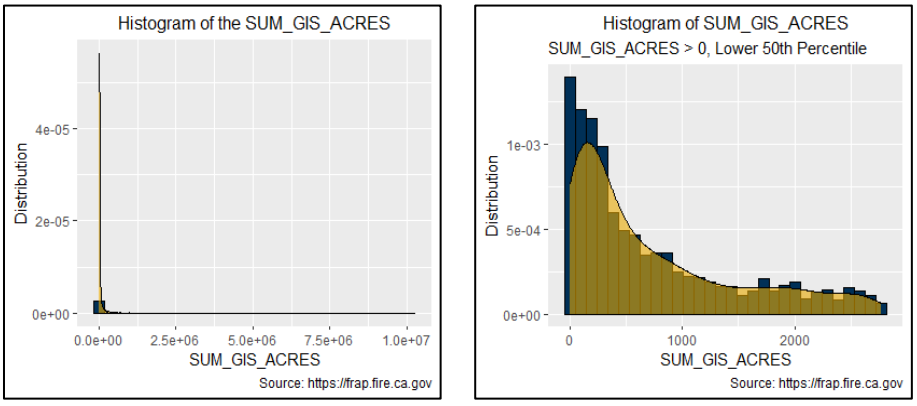


Figure 4: *Histogram of SUM_GIS_ACRES*
(Left: All data points; Right: SUM_GIS_ACRES >0, Lower 50th Percentile)

3.3 Time Series Plots

Plotted to the right is the sum total precipitation (inches) plotted by year from 2001 to 2021, which could anecdotally align with our hypothesis of climate change's influence on weather patterns. We can see a general trend of decreased cumulative total precipitation since the early 2000s, as well as observe general ebbs and flows in the total precipitation data in the years following.

Looking at average temperature data plotted by year in California, we see a few distinct peaks in the data - most notably the high point in 2015. The above data is the monthly temperature averaged by year and it should be noted that this is the mean() temperature by year. We may also wish to add variables for the max() and min() temperature into our data, as the trends for mean temperature are curious, but max() temperature could prove more directly correlated with wildfires.

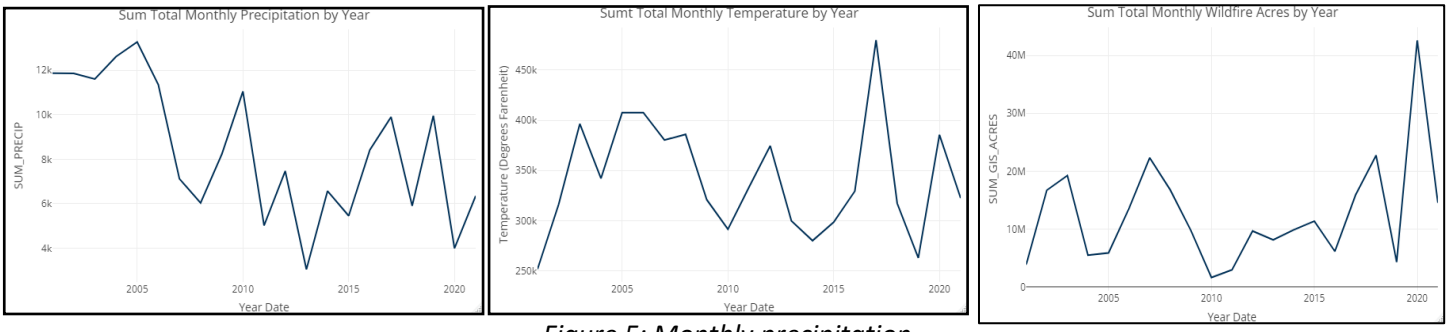


Figure 5: Monthly precipitation, SUM_TEMPERATURE and wildfire acres by year

Examining the sum total wildfire acreage by year, we note that the largest total wildfires appeared to come in waves every few years, and there is an extremely sharp uptick in total wildfire acreage in 2020. Wildfires tend to be seasonal, most often taking place in the warmer summer months. Our analysis will attempt to identify whether there is a strict data link correlated between the periods of hotter and drier weather and whether this is directly correlated with an increase in wildfires.

3.4 Correlation Matrices and Plots

An initial review of the correlation summary and plot does not immediately reveal any strong correlations between the variables, in either positive or negative directions.

Looking at the correlation values specifically for the dependent variable SUM_GIS_ACRES the most strongly correlated dependent variables look to be SUM_TEMPERATURE (0.32) with a moderate positive correlation effect and SUM_TEMPERATURE_DEPART (0.27), also with a low-moderate positive correlation. Also of note, AVG_PRECIP shows a weak, but notable negative correlation (-0.06).

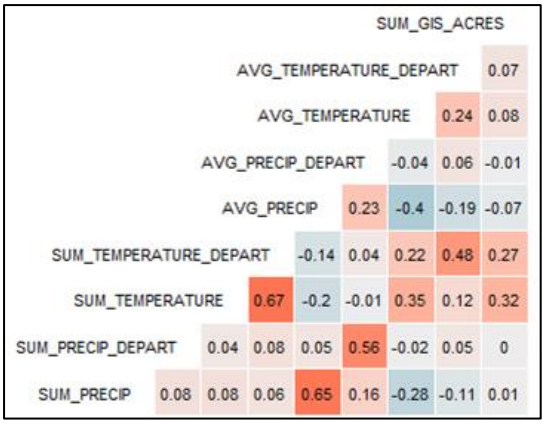


Figure 6: Correlation plot

The correlation plot results suggest that temperature and its departure variability may be the best suited to predicting SUM_GIS_ACRES. Interestingly, while rainfall does show the expected negative correlation with wildfire acreage, its correlation is substantially weaker than anticipated.

4 CHALLENGES

4.1 Data With Missing/Incomplete Values

NOAA weather data is collected by individual weather stations and months. For any given period, there can be completely blank (missing) values, or records that are marked by NOAA with "M" (insufficient or partial data), "T" (trace data; an amount too small to measure), or another symbol to indicate a potentially insufficient recording. Our team must then determine the impact of these missing and potentially invalid data points, and make a determination on how to process these data points prior to proceeding with further data exploration and modeling.

Additionally, while we had hoped to analyze the wildfire duration as a potential dependent/independent variable, the CalFire perimeters data set is missing the critical data necessary to do so. While most all recorded fires have "Alarm Dates" (the date recorded when the fire was first observed), many fire records are missing the key "Containment Date" (the date on which the fire was deemed fully contained).

4.2 Data Lacking Additional Details

We had originally hoped to analyze the dimension for the cause of each fire however, this column in the dataset appears to lack categorical structure and instead is a freeform text input field. This means that many of the responses can be similar, but different enough to make categorization very difficult. Further, many of the values, for this reason, are simply blank and therefore unusable. As a result, our team has decided to not analyze the particular reason for a fire, but instead simply attempt to model the relationship between overall wildfire magnitude by month in a given region.

4.3 Joining Data

Our next challenge was in joining our two separate data sets: NOAA weather data, with the geographic identifier of the weather station name, and the CalFire perimeter data set, with the geographic identifier of the administration unit name. In order to complete this data join, we decided to map both datasets geo-key's back to California Counties in order to bridge the data. However, there were not readily available tables that we could locate to easily map weather stations and admin units back to counties, so this task had to be done largely manually, and some compromises had to be made where there were not clear exact matches in the map. These compromises could lead to added deficiencies in our combined data set and have unintended impacts on our resulting data analysis and models.

4.4 Data Correlation

As noted in the previous section, there was not an immediate, strong correlation between our chosen independent and dependent variables. This unexpected lack of a distinct correlated trend between weather variables, precipitation and temperature, and the recorded wildfires in California, goes in direct contrast to our initial hypothesis and challenges the general perception of the trends of climate change and its impact on the perceived rise in destructive natural disasters around the globe. Additionally, the lack of a distinct strong feature correlation may lead to further challenges in developing future effective data models later.

5 DATA MODELING

5.1 Modeling Considerations

After completing our exploratory data analysis and processing potential modeling challenges, we turned our attention to identifying the optimal data models to test. Considering the range of variables in our data set, their unique characteristics, and interlocking relationships, we looked for models that would be well-suited to our data set and simultaneously most useful in our original goal of predicting the magnitude of California wildfires using weather data.

Ultimately, we decided on examining a combination of regression models, including (1) classic multiple linear regression, (2) non-linear regression models, utilizing log transformations, and (3) logistic regression models. With this

selection of regression models, we aim to comprehensively test the varying relationships of our independent variables to find the optimal model for predicting wildfire magnitude. Lastly, we will also look to explore a logistic regression model, in hopes of predicting the probability of wildfires, and intended to work as either an independent model or a complement to the linear model.

5.2 Linear Regression

We decided to apply a linear regression model for the reason that given a set of variables, we can predict the sum of acres burned by wildfire for an admin unit region of California. The full model to predict the sum of acres on all variables yielded very poor results, having R2 and adjusted R2 coefficients of 0.1211 and 0.1164, respectively. Using precipitation and temperature variables as trailing X months variables provided unexpectedly poor results as well. The date and temperature variables tended to be more consistently significant.

```
Call:
lm(formula = SUM_GIS_ACRES ~ ., data = data_san_bern)

Residuals:
     Min       1Q   Median       3Q      Max
-898065  -85100  -29014   21921 3929838

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.767e+07  8.156e+06   2.167 0.031251 *
YEAR                   -8.798e+03  4.059e+03  -2.167 0.031197 *
MONTH                   5.049e+03  6.511e+03   0.775 0.438895
SUM_PRECIP             -1.128e+03  1.360e+03  -0.830 0.407478
SUM_PRECIP_DEPART      -1.308e+02  2.162e+02  -0.605 0.545639
SUM_TEMPERATURE         4.125e+00  1.515e+01   0.272 0.785705
SUM_TEMPERATURE_DEPART  1.255e+03  3.197e+02   3.926 0.000113 ***
AVG_PRECIP              9.555e+03  5.221e+04   0.183 0.854930
AVG_PRECIP_DEPART       1.241e+03  1.654e+04   0.075 0.940280
AVG_TEMPERATURE         1.017e+03  2.457e+03   0.414 0.679296
AVG_TEMPERATURE_DEPART -3.219e+04  2.002e+04  -1.608 0.109152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 341500 on 241 degrees of freedom
Multiple R-squared:  0.2981,    Adjusted R-squared:  0.269
F-statistic: 10.24 on 10 and 241 DF,  p-value: 2.43e-14
```

Figure 7: San Bernardino model summary

The decision to narrow the dataset by seven admin units was made to reduce clutter and noise while focusing on a few admin units was made under the following criteria and logic. Seven admin units were explored with individual linear regression models based on their total acres burned from 2001 through 2021, having at least 20 million acres burned. Then, linear models with all precipitation and temperature variables were run on each of the seven admin units. Upon gathering results in the form of RMSE and R^2 for the fitted models, San Bernardino and Siskiyou admin units performed within the top three in both RMSE and R^2. We also reached this decision as the two admin units are spread towards the north and south of California, with Siskiyou and San Bernardino units covering those areas. For future models, exploration, and testing, we can advise using the San Bernardino and Siskiyou models created for new data to test whether that period or next will have a wildfire, and if so, how large. Model selection will depend on which part of California we are testing. The below table captures RMSE and R^2 by admin unit, with the total acres burned for context.

	Admin_Unit	Total_Acres_Burned	RMSE	R_Squared_fit	R_Squared_lwr	R_Squared_upr
5	San Bernardino Unit	23193722	708074	0.70560490	0.42645860	0.76854170
7	Siskiyou Unit	20730362	531590	0.61318010	0.53709220	0.62605770
4	San Diego Unit	24344547	1470014	0.27460380	0.18352940	0.33423010
6	Lassen-Modoc Unit	23098284	1066009	0.19333470	0.17633850	0.20644730
3	Shasta-Trinity Unit	24425666	614594	0.19214540	0.18745780	0.19646690
1	Sonoma-Lake-Napa Unit	38800459	1711409	0.18522880	0.07043268	0.29578440
2	Los Angeles County	25642648	944464	0.09712883	0.09193059	0.08933765

Figure 8: RMSE and R2 for each fitted Admin Unit model

5.3 Nonlinear Transformations (Linear-Log & Log-Linear)

After testing linear regression models, we explored the options of linear-log and linear-log transformations with the goal to increase model fit and improve R^2 values as well. We learned above that the full dataset to predict sum acres performed poorly so we started the nonlinear transformation testing by splitting the full dataset into two subsets of sum data and average data. Many models were tested with individual logged factors being tested. R^2 and adjusted R^2 coefficients were compared for each test to see what logged factors would improve and hurt model fit. The best-

performing model was the sum data subset using a log-linear transformation. As seen in the table, the average data subset performed worse consistently and should not be considered in future model creation. When testing linear-log models, the log precipitation yielded better results than the log temperature. A challenge we ran into when running these nonlinear transformations was missing data. Missing numerical data was filled in as zero and the log of zero is undefined. We were unable to run the regression model with undefined data so missing data was removed from the dataset. Overall, the full dataset nonlinear transformations still performed poorly with the highest R^2 value and adjusted R^2 both being 0.23.

Log Model	R^2	ADJ R^2
SUM LOG-LNIEAR	0.2324788	0.2312075
SUM LINEAR-LOG (LOG SUM PRECIP DEPART)	0.1996714	0.19823
SUM LINEAR-LOG (LOG SUM PRECIP)	0.1164213	0.1158557
SUM LINEAR-LOG (LOG SUM TEMP DEPART)	0.1063905	0.1056219
SUM LINEAR-LOG (LOG SUM TEMP)	0.08435022	0.08381113
AVG LOG-LNIEAR	0.04245445	0.04086845
AVG LINEAR-LOG (LOG AVG PRECIP DEPART)	0.01085305	0.009072408
AVG LINEAR-LOG (LOG AVG PRECIP)	0.009247634	0.008613349
AVG LINEAR-LOG (LOG AVG TEMPERATURE)	0.00257538	0.001988142
AVG LINEAR-LOG (LOG AVG TEMPERATURE DEP	0.001606647	0.0007472598

Figure 9: Nonlinear model R2

The next step for our nonlinear transformation exploration was to test and narrow our scope to admin units and we chose two admin units, Siskiyou and San Bernardino, to use. Running the same tests as above, the best model for Siskiyou was a linear-log with log precipitation transformation and the best model for San Bernardino was the log-linear model. The Siskiyou model performed better than the models discussed previously with an R^2 and adjusted R^2 was .43 and .31 but the same cannot be said about the San Bernardino unit as itsR^2 and was around the same as the full dataset models. Compared to the regular linear regression models above, the nonlinear transformations seemed to perform worse and shouldn't be used as the final model.

```
Call:
lm(formula = SUM_GIS_ACRES ~ YEAR + MONTH + LOG_SUM_PRECIP +
    SUM_TEMPERATURE, data = trimmed_siskiyou_data)

Residuals:
    Min      1Q  Median      3Q     Max
-890758 -134072  -38140   87688 1683589

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.558e+07  1.210e+07  -2.940  0.00400 **
YEAR             1.767e+04  5.997e+03   2.947  0.00392 **
MONTH            1.107e+04  1.554e+04   0.712  0.47797
LOG_SUM_PRECIP  -4.592e+04  2.450e+04  -1.874  0.06357 .
SUM_TEMPERATURE  1.675e+02  1.961e+01   8.540 8.43e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 377300 on 110 degrees of freedom
Multiple R-squared:  0.4346,    Adjusted R-squared:  0.4141
F-statistic: 21.14 on 4 and 110 DF,  p-value: 5.962e-13
```

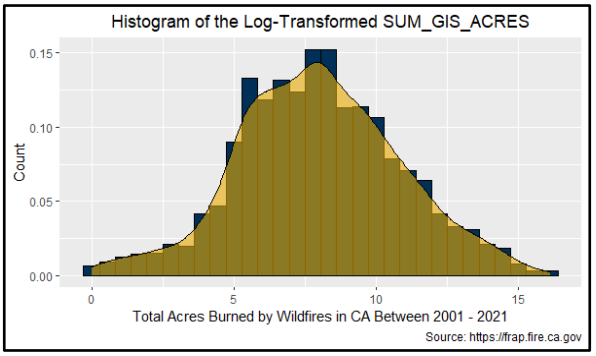
Figure 11: Siskiyou Linear-Log Final Model

```
Call:
lm(formula = LOG_SUM_GIS_ACRES ~ YEAR + MONTH + SUM_PRECIP +
    SUM_TEMPERATURE, data = trimmed_san_bernardino_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7494 -1.5279 -0.0498  1.5741  5.0302

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.241e+02  6.617e+01   1.875   0.0631 .
YEAR            -5.734e-02  3.289e-02  -1.743   0.0837 .
MONTH            6.757e-03  7.263e-02   0.093   0.9260
SUM_PRECIP      -3.401e-03  5.425e-03  -0.627   0.5318
SUM_TEMPERATURE  2.234e-04  4.176e-05   5.350 3.96e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.123 on 127 degrees of freedom
Multiple R-squared:  0.2194,    Adjusted R-squared:  0.1948
F-statistic: 8.924 on 4 and 127 DF,  p-value: 2.203e-06
```

Figure 10: San Bernardino Log-Linear Final Model

5.4 Nonlinear Transformations (Log-Log)

Next, we tested a full log-log model, in an attempt to further increase the overall model fit by transforming all variables, paying particular attention to precipitation and gis acres, which as noted previously, do not naturally normally distributed.



After performing log transformations on the dependent variable, we observed a much more normal-looking bell curve for SUM_GIS_ACRES. Further, by performing a log transformation on the independent variables, and running an initial model with all coefficients, we saw noticeable improvements to residuals vs fitted and Q-Q plots, indicating a better fit for modeling the data.

After developing an initial log-log model and getting baseline results with all coefficients, we tested several different combinations of variables, and through the process of elimination based on p-values and R^2 scores, we arrived at a simple model utilizing ln_SUM_PRECIP and ln_SUM_TEMPERATURE. This model had a nearly identical Adj-R2 value (0.4414), but also has the added benefit of decreased potential for multicollinearity and increased interpretability - qualities which are especially important when working with already complex log transformations in data modeling.

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -85.831798  12.757891  -6.728 1.86e-11 ***
YEAR                0.036886   0.006334   5.824 6.01e-09 ***
MONTH               0.079689   0.011172   7.133 1.08e-12 ***
ln_SUM_PRECIP      -0.591832   0.024568 -24.090  < 2e-16 ***
ln_SUM_TEMPERATURE  2.423859   0.035659  67.973  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.136 on 6799 degrees of freedom
Multiple R-squared:  0.4417,    Adjusted R-squared:  0.4414
F-statistic:  1345 on 4 and 6799 DF,  p-value: < 2.2e-16
```

Figure 12: Summary of log-log model

5.5 Logistic Regression

We created a logistic regression model to identify whether a fire will happent. Downloaded fire data has a variable for the amount of acreage burnt because of wildfire. The factor variable for the logistic regression model was created based on the variable amount of acreage burnt. When the amount of acreage burnt is more than zero, the factor variable is specified as 1 indicating there was fire, and otherwise specified as zero. Year, month, monthly average precipitation, monthly average temperature, sum of precipitation and sum of temperature were included in the model. The model with variables year, month, sum of precipitation, and sum of temperature resulted in better accuracy. Individual logistic regression models were created for all the admin units, San Bernardino (southernmost part of California) and Siskiyou admin unit (northernmost part of California). Within each group, three different probability thresholds were tested. The probability thresholds were compared based on the confusion matrix metrics. In the case of forest fires, the cost of false negatives is higher compared to other metrics in the confusion matrix. A 0.25 threshold yielded the least false negatives and was picked as the final model.

```
Call:
glm(formula = fire ~ MONTH + SUM_PRECIP + SUM_TEMPERATURE, family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3907  -0.7040  -0.5047   0.3658   3.3566

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.349e+00  8.557e-02 -27.454  < 2e-16 ***
MONTH           6.588e-02  9.377e-03   7.026 2.13e-12 ***
SUM_PRECIP     -1.544e-02  1.058e-03 -14.591  < 2e-16 ***
SUM_TEMPERATURE 2.111e-03  7.022e-05  30.060  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8857.3  on 6803  degrees of freedom
Residual deviance: 6108.1  on 6800  degrees of freedom
AIC: 6116.1

Number of Fisher Scoring iterations: 7
```

```
Call:
glm(formula = fire ~ YEAR + MONTH + SUM_PRECIP + SUM_TEMPERATURE,
    family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.92748  -0.70314   0.00003   0.32182   2.04992

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    39.4450161 68.6588104   0.575   0.5656
YEAR           -0.0213268  0.0340254  -0.627   0.5308
MONTH          -0.0130841  0.0487792  -0.268   0.7885
SUM_PRECIP     -0.0120277  0.0069473  -1.731   0.0834 .
SUM_TEMPERATURE  0.0024212  0.0004609   5.253 1.49e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 348.06  on 251  degrees of freedom
Residual deviance: 194.04  on 247  degrees of freedom
AIC: 204.04

Number of Fisher Scoring iterations: 8
```

```
Call:
glm(formula = fire ~ YEAR + MONTH + SUM_PRECIP + SUM_TEMPERATURE,
    family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0470  -0.6495  -0.3445   0.1436   2.3166

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.177e+02  6.826e+01  -3.189  0.00143 **
YEAR            1.064e-01  3.380e-02   3.147  0.00165 **
MONTH          -7.241e-03  5.251e-02  -0.138  0.89032
SUM_PRECIP     -4.463e-03  6.357e-03  -0.702  0.48266
SUM_TEMPERATURE  9.237e-03  1.656e-03   5.578 2.43e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 348.33  on 251  degrees of freedom
Residual deviance: 184.80  on 247  degrees of freedom
AIC: 194.8

Number of Fisher Scoring iterations: 9
```

Figure 13: Logistic Regression Results: All admin units, San Bernardino and Siskiyou Unit

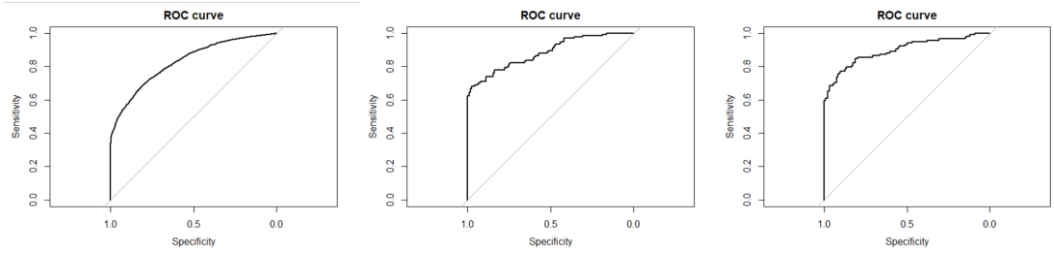


Figure 14:ROC Curve: All admin units, San Bernardino and Siskiyou Unit

Admin Unit	Threshold	True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
All units	0.25	1871	3030	1354	549	0.72	0.77	0.69
San Bernardino Unit	0.25	126	53	64	9	0.71	0.93	0.45
Siskiyou Unit	0.25	102	92	42	16	0.77	0.86	0.69

Figure 15:Comparison of key metrics for all, San Bernardino and Siskiyou Units

6 CONCLUSION

6.1 Model Comparison & Selection

After thoroughly developing and analyzing the above models, we reflect on the overall final performance of each. To accurately compare the effectiveness of each model relative to one another, we looked at adjusted-R2 values.

Data Model	Adj R^2
Linear	0.7056
Log-Linear	0.1948
Linear-Log	0.4141
Log-Log	0.544

Ultimately, to provide the most impactful forecasts to governing agencies in the effort to preempt future wildfires, we decided to adopt a combination of models. Based on superior accuracy and adjusted-R2 values, we recommend the use of the logistic regression model to predict the probability of a wildfire occurring, and the multiple linear regression model to predict the magnitude of a wildfire in a given month and administrative unit.

6.2 Future Prediction

The probability of fire happening for the San Bernardino and Siskiyou unit were computed for the month of November 2022 from the logistic regression model.

```
> head(new_data)
  YEAR MONTH SUM_PRECIP SUM_TEMPERATURE
1 2022    11      21.13         1238.94
> predict(lr, new_data, type="response") # San Bernardino Unit
         1
0.2542607
```

```
> head(new_data)
  YEAR MONTH SUM_PRECIP SUM_TEMPERATURE
1 2022    11      32.83          245.32
> predict(lr, new_data, type="response") # Siskiyou Unit
         1
0.3568327
```

Probability of fire for San Bernardino and Siskiyou Unit

Predicted sum of burned acres, if a fire is expected to be present in the San Bernardino and Siskiyou regions, is 55,579 and 28,551 acres burned, respectively.

6.3 Suppression Cost

Previous year fire acreage burnt and suppression cost [7] was used to create a linear regression model which was used to calculate the fire suppression cost for the future.

```
Call:
lm(formula = Total ~ Year + Acres, data = data_cost_no_outlier)

Residuals:
       Min         1Q     Median         3Q        Max
-611809461 -156155511   69743099  163270771  484919382

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.348e+10  1.197e+10  -6.976 7.91e-08 ***
Year         4.202e+07  6.013e+06   6.988 7.66e-08 ***
Acres        7.682e+01  2.226e+01   3.452  0.00163 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.82e+08 on 31 degrees of freedom
Multiple R-squared:  0.8173,    Adjusted R-squared:  0.8055
F-statistic: 69.32 on 2 and 31 DF,  p-value: 3.616e-12
```

Figure 16: Linear Regression Model for Fire Suppression Cost

Logistic and linear regression models were combined to predict the future possibility of fire. The logistic regression model was used to find out the probability of fire and the linear regression model was used to find out how many acres can be expected to burn. 2021 average for dollars spent suppressing burned acres: ~$616 per acre. 2021 average for dollars spent per fire: ~$74,409. Fire suppression cost for the predicted acres burned was quantified based on the suppression cost linear regression model above. Based on the cost prediction, appropriate budgets and resources can be allocated to the admin units accordingly.

7 REFERENCES

[1] California Department of Forestry and Fire Protection (CAL FIRE). (n.d.). Stats and events. Cal Fire Department of Forestry and Fire Protection. Retrieved October 30, 2022, from https://www.fire.ca.gov/stats-events/

[2] Cragcrest. (2018, November 16). Why California's wildfires are so destructive, in 5 charts. FiveThirtyEight. Retrieved October 30, 2022, from https://fivethirtyeight.com/features/why-Californias-wildfires-are-so-destructive-in-5-charts/

[3] Fleck, A., & Richter, F. (2022, July 13). Infographic: The growing danger of Californian wildfires. Statista Infographics. Retrieved October 30, 2022, from https://www.statista.com/chart/14462/California-wildfire-deadly/

[4] Marisa Iati, D. M. (2021, September 17). Anatomy of a wildfire: How the dixie fire became the largest blaze of a devastating summer. The Washington Post. Retrieved October 30, 2022, from https://www.washingtonpost.com/climate-environment/interactive/2021/dixie-fire/

[5] Sleight, M. (n.d.). U.S. wildfire statistics. Bankrate. Retrieved October 30, 2022, from https://www.bankrate.com/insurance/homeowners-insurance/wildfire-statistics/#the-worst-wildfires-in-us-history

[6] Wildfire prevention: How to prevent & control forest fires. EARTH OBSERVING SYSTEM. (2022, July 15). Retrieved October 30, 2022, from https://eos.com/blog/wildfire-prevention/

[7] https://www.nifc.gov/fire-information/statistics/suppression-costs