Foundations of Data Science Final Project: The Prevalence and Impact of Major U.S. Wildfires, 1992 to 2015

Kate Lampo (kel2169)

I. Background & Research Questions

The National Interagency Fire Center, which is housed in Boise, Idaho, has documented data about fires in the United States for many decades. Since 1983, there's been an average 70,000 fires recorded each year, which vary wildly in intensity. While experts acknowledge that the total number of fires each year has not increased significantly in that time, they are much more concerned with the damage caused by such fires. As global warming continues to worsen and climates become warmer and more arid, many places in the US and around the world are more susceptible to large, uncontrollable blazes.

As a Colorado native that grew up right next to the Rockies, the threat and impact of wildfires is something that I've always been acutely aware of. My junior year of high school, my cross country season was limited due to air quality issues arising from local fires. My senior year, the town over from mine experienced a devastating wildfire that displaced hundreds, burning down the apartment building that my parents lived in after college. As such, I'm interested in studying trends in fire prevalence and intensity over time, and in seeking trends that may help us mitigate damage in the future.

Therefore, in analyzing these data, I'm hoping to answer two crucial questions:

1) Can we predict the severity of a fire when it begins based on factors like location, time of year, and what caused it?; and
2) How do we expect the rate of severe fires in different states to change in the coming years?

While much prior work exists surrounding predicting the impact of a fire as they happen, and in studying broad trends and fire patterns, addressing these questions has the potential to better inform long-term infrastructure planning. If we know that certain areas (towns, cities, states) are more prone to fires, or that we statistically expect a given number of fires in a month, local authorities can be better equipped to handle these disasters, even years in advance. This analysis isn't intended to predict exactly when and where fires will occur—rather, it's focused on analyzing with as much specificity as possible when fires are *likely* to happen.

II. Data Description

To address these research questions, I will be using a Kaggle data set prepared and published by Rachel Tatman in 2019. The data set pulls from several different fire reporting sources, and serves as a comprehensive list of recorded fires in the US between 1992 and 2015. It contains over 1.88 million entries, each of which is described using a set of more than 35 descriptors. However, many of these descriptors, while crucial to identifying where the data come from (e.g. the reporting agency that recorded the fire), are less relevant to the proposed research questions. Therefore, for analysis, the data set is pared down to only the following nine attributes:

- *ID:* (Int) A unique identification number for each fire.
- *DISC_DATE:* (Float) The date that the fire was discovered, in Julian time.
- *CONT_DATE:* (Float) The date that the fire was contained, in Julian time.
- *CAUSE_CODE:* (Float) A numeric code denoting the cause of the fire.
- *CAUSE_DESCR:* (String) A string description of what caused the fire.
- *CLASS:* (String) The class of the fire, which is determined by the total number of acres burned. Classes are lettered (A: 0 to 0.25 acres; B: 0.26 to 9.9 acres; C: 10.0 to 99.9 acres; D: 100 to 299 acres; E: 300 to 999 acres; F: 1000 to 4999 acres; G: 5000+ acres
- *LATITUDE:* (Float) The latitude at which the fire originated.
- *LONGITUDE:* (Float) The longitude at which the fire originated.
- *STATE:* (String) The two-letter state code for the state in which the fire originated.

While intended primarily as a centralized database of fire occurrences, there is clearly enough information here to also extract basic information about fire impacts—particularly the amount of land area burned. While this is a simplified metric to determine the impacts of a given fire (since some large controlled blazes have little to no human impact and are crucial to the health of particular ecosystems), it is a good baseline metric. These shortcomings are discussed further in Section VI.

III. Data Cleaning

A. Importing Data

Because of the size of the original data set, it was stored in a .sqlite file format——a C database engine that allows for the storage of SQL databases in smaller archives. To begin analysis, the data were extracted using the command-line program sqlite3 and saved off into a .csv file. This process is captured in Fig. 1. From there, a SQL query was used on the .csv file to extract the fields mentioned in Section II to a Pandas data frame.

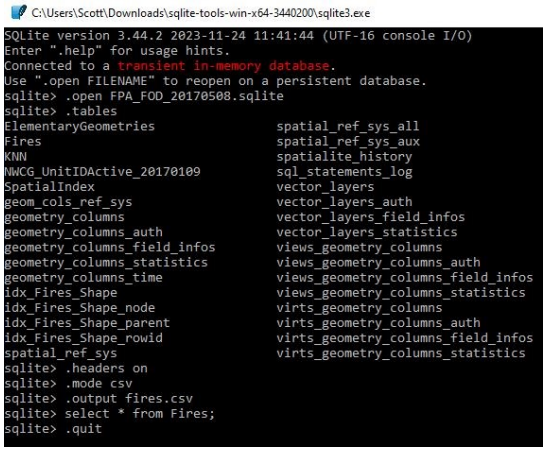

Fig. 1. The CMD line process used to convert the original .sqlite data to a workable .csv file.

B. Cleaning

To begin the cleaning process, the *CAUSE_DESCR* and *CLASS* columns were modified to remove inconsistencies in the set. Grouping by *CAUSE_DESCR* revealed that some entries were erroneously populated with an integer cause code instead of the string description (eg. "1.0" instead of "Arson"). To remedy this, a *desc_code* data frame was created to store the relations between the numerical codes and their corresponding descriptions. By merging this data frame with the original *fire_data* data frame, erroneous numeric values were replaced by their appropriate descriptions.

Similarly, there were several *CLASS* values that were populated with a numeric value (asumed to be the number of acres burned) instead of the associated letter code. This was remedied using slicing and a boolean mask for each category. At this point, fires of class A, B, or C were also dropped from the data set, since the focus of this project is large, damaging fires, which are defined here to be 100+ acres in size.

After cleaning the *CLASS* attribute, there remained a few irreconcilable values in the *STATE* column (eg. "14.0"), which were ill-defined with no clear fix. These rows were dropped, but other rows with missing values were kept, as the data that they contain was still useful for some analyses.

	ID	CLASS	LATITUDE	LONGITUDE	STATE	CAUSE_DESCR	DISC_DATE	CONT_DATE	DURATION
0	17	G	38.523335	-120.211670	CA	Equipment Use	2004-10-06	2004-10-21	15 days
1	18	G	38.779999	-120.260002	CA	Equipment Use	2004-10-13	2004-10-17	4 days
2	40	D	36.001667	-81.589996	NC	Debris Burning	2005-02-12	2005-02-13	1 days
3	119	D	43.899166	-102.954720	SD	Lightning	2005-07-16	2005-07-17	1 days
4	120	D	43.892776	-102.948059	SD	Lightning	2005-07-16	2005-07-16	0 days
...	...	...	...	...	...	...	...	...	...
54088	300345398	D	40.463516	-124.386818	CA	Missing/Undefined	2009-01-15	2009-01-16	1 days
54089	300345499	D	37.072712	-119.694153	CA	Debris Burning	2015-06-21	NaT	NaT
54090	300346248	F	40.956982	-121.321236	CA	Lightning	2008-06-22	NaT	NaT
54091	300347461	E	37.600388	-120.191170	CA	Equipment Use	2014-06-06	2014-06-08	2 days
54092	300347643	D	37.222500	-121.808334	CA	Children	2014-06-30	2014-06-30	0 days

Fig. 2. The cleaned data frame, *fire_data_cleaned.*

Finally, the discovery and containment dates were converted to datetime values. Using these new values, the duration of each fire was calculated by creating a timedelta attribute representing the difference of the two. This column was added to the *fire_data* dataframe in a series called *DURATION*. During the conversion, many timestamps were found to be illogical in the Julian system (eg. "2014.0", which would correspond to 4708 B.C.). Because it was impossible to reconstruct a date from those values (and time is important to this analysis), they were dropped from the set.

The final data frame, renamed *fire_data_cleaned*, is given in Fig. 2. It contains just over 54,000 rows, representing the top 3% of the most destructive fires recorded in the data set's 24-year span. Of those, 52% are class D, 26% are class E, 14% are class F, and 7% are class G.

IV. Exploratory Data Analysis

Data exploration begin with identifying the most prevalent causes of fires, which are given in Table I. Evidently, many of the causes are undefined or miscellaneous, suggesting that the data on the causes of fires are far from complete. This makes it more difficult to use fire source as a predictor of fire intensity, duration, or likelihood. Additionally, among the sources that are defined, there's no clear delimitation between natural and man-made sources, both of which seem to be equally at fault. All other fire sources (those not included in the table) each accounted for less than 1,012 of the 54,000 large fires in the cleaned data set.

TABLE I

Table giving the most common sources of large wildfires.

Cause Description	Quantity
Lightning	15944
Arson	9441
Miscellaneous	8082
Debris Burning	6247
Missing/Undefined	5805
Equipment Use	4519

Next, to better examine the distribution of fires across the US, the number of fires in each state over the same time period were calculated. The top seven states by number of "G" class fires is given in Table II.

TABLE II

The top seven states with the most class G wildfires.

State	Class D	Class E	Class F	Class G	Total
AK	349	378	413	650	1790
CA	2122	1184	756	394	4456
ID	1062	810	693	394	2959
NV	483	416	407	295	1601
OR	638	417	353	256	1664
TX	3782	1781	798	238	6599
NM	927	710	493	236	2366

Alaska, the largest state in the US, is home to the most G class fires–more than double California, which is the state with the second most. However, the state with the most fires overall is Texas, which had 6599 fires in the period of study. Overall, the fires are concentrated in the western US, and are larger in states with larger land mass, which is to be expected. It's also important to note that much of Alaska's wilderness is uninhabited, meaning that the fires with the most severe

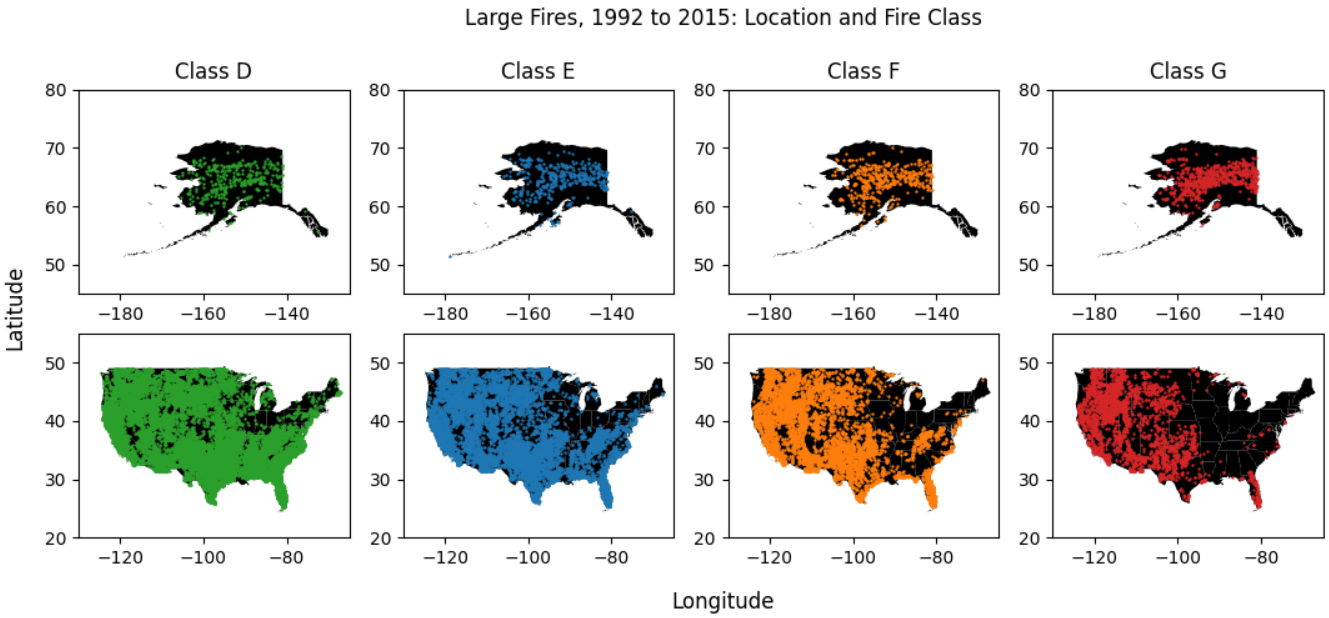

Fig. 3. Map illustrating the frequency of fires in different locations, 1992-2015. Each dot represents one occurrence of a fire. Hawaii is omitted for having only 210 total fires in that period.

human impact are likely to be found in states like California and Texas, which are far more population dense.

Adding further to this analysis, the number of large wildfires by type is visualized in Fig. 3. While smaller fires occur across the contiguous US, more severe ones are clearly more prevalent in the West and in Alaska.

Next, looking for temporal trends, we can look at the number of fires in the dataset that occur in each month of the year. The histogram in Fig. 4 gives the total number of fires in the data set that fall in each month of the year (where 1:January and 12:December).

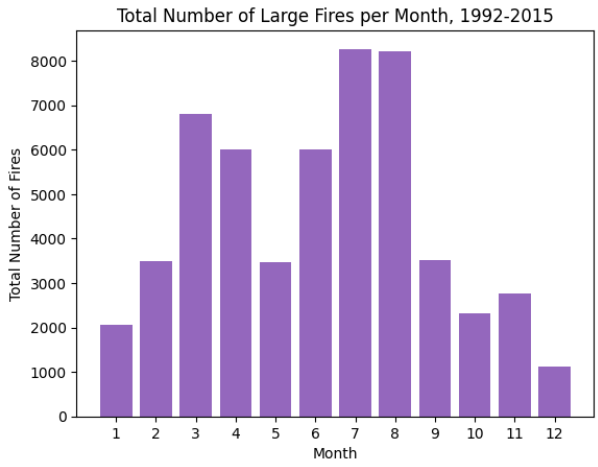

Fig. 4. Total number of fires starting in each month.

Evidently, there are more fires in the summer months, which is to be expected given the elevated heat during those times. However, we also see a spike in fires in March and April, which is slightly less intuitive. This may be due to the high winds that are characteristic of late spring in some areas, which, when coupled with an ever-drying climate, may contribute to higher rates of fires.

It's also interesting to compare the duration of a given fire to the time of year at which it started, which is visualized in Fig. 5.

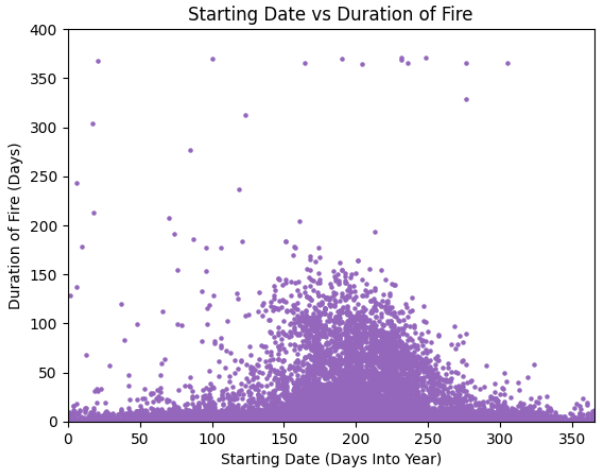

Fig. 5. Scatterplot of fire duration vs time of year, illustrating that the longest fires occur in the Summer months.

Here, we see that, aside from there being more fires in the summer months, those fires also tend to last longer (save a few outliers—extremely long fires that start at any given time of year, which are primarily Alaskan).

By calculating the average duration of each class of fire (Table III), we can also see that there is a direct correlation between duration and area burned, indicating that higher class fires are more likely to occur in those same Summer months.

Finally, we can observe trends over time in the period accounted for by the data set. Beginning with the total

TABLE III

THE AVERAGE DURATION OF EACH CLASS OF LARGE FIRE.

Class	Average Duration (days)
D	4.12
E	7.41
F	14.32
G	32.40

number of fires per year over time, we see a gradual increase as illustrated in Fig. 6.

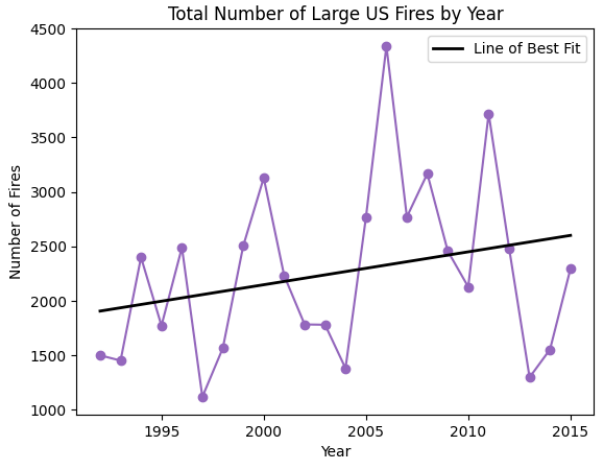

Fig. 6. The total number of fires per year over time.

The line of best fit on the graph can be characterized by the function $y = 30.21x - 58281$ where y is the number of fires in a given year and x is the year. This indicates an overall increase of just 30 fires per year, which is an increase of only about 1%. However, the relative proportion of each class of fire over time yields a more interesting trend. Fig. 7 gives the normalized proportion of each class of fire over time. To produce the plot, the proportion of total fires that each class accounted for in 1992 (58.44% D, 24.55% E, 13.14% F, and 3.87% G) was first computed. The same was done for each subsequent year, and all values were then normalized to the original 1992 proportion.

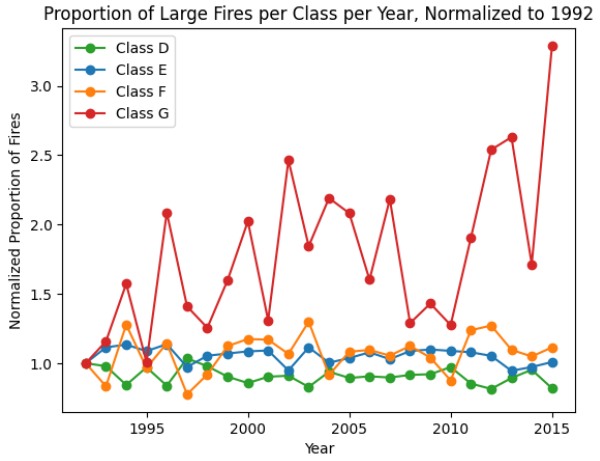

Fig. 7. The proportion of fires each year belonging to each class, normalized to 1992.

Plotting these trends reveals that, although the proportion of Class D, E, and F fires has remained relatively the same over time, the proportion of Class G fires increased by nearly three times between 1992 and 2015. This indicates that, although there aren't necessarily many more fires each year, the ones that do occur are more likely to be highly destructive.

V. WILDFIRE PREDICTION & ANALYSIS

Now that general trends in the data have been established, we can move on to prediction. Both prediction analyses discussed here focus on projecting the trends in these data into future years to better understand the risks that wildfires will pose. The first uses a classification tree to predict the fire class based on other factors, and the second addresses the number of fires per month in different states over time, directly addressing both of the research questions posed in Section I.

A. Predicting Severity

The first analysis predicts the severity of a fire based on its location, cause, the day into year it began, and the year it began. In the analysis, three different models are used: location based on latitude and longitude, location based on state, and location based on both. Because latitude and longitude are continuous variables, it's hard to use them in this type of prediction because of the dispersion of fires as indicated in Fig. 3. For example, a decrease in longitude indicates that the fire is further west, where there are generally more (and more intense) fires. However, at a longitude of -130, for example, there are no data points (since there are no US states there—it's between the continental US and Alaska). But, once longitude is low enough to be in Alaksa, fires return again. Therefore, there is clearly no perfect direct relation between longitude and severity as a continuous variable would suggest.

Using state (a categorical variable) as the predictor instead therefore may yield better results, but comes at the cost of losing detail throughout the state. For example, again referencing Fig. 3, we see that for states in the middle of the country, the western edge of a state is likely more prone to more damaging fires. Using state means that this specificity is lost, even though it mitigates the continuous variable issue.

To conduct these analyses, state and cause values were converted into binary columns in the data frame for use in a decision tree analysis. Then, using an 80/20 training and testing split, the three models were run, yielding accuracy scores of 0.5283 (state only), 0.5176 (lat/long only), and 0.5185 (both).

While none of the models are particularly accurate, they do offer a significant advantage over randomly guessing (which we would expect to have an accuracy of 0.25). We also see that the state only model is slightly more accurate than the ones using latitude and longitude, indicating that using continuous variables in a non-continuous context is problematic as anticipated. Moreover, for all three models, a maximum depth of 10 was found to produce the most accurate results, which is demonstrative of the fact that all

types of fires can stem from all types of causes in all places, so too many splits are detrimental to the model.

Finally, as an example of the application of the model, the state-only version was retrained on the full data set, then a sample point was used as a demonstration of its shortcomings. The 2021 wildfire mentioned in Section I, known as the Marshall fire, took place in December 2021, and burned over 6,000 acres, making it Class G. It occured in Boulder, Colorado, and was started by a power line. Feeding this information into the model yields a prediction of class D, which is grossly untrue. However, the Marshall fire is unique in that it followed an extremely dry season in Colorado, and was damaging largely thanks to high winds during that time. Therefore, this test point illustrates nuance that the model cannot capture, even though it provides some level of accuracy in prediction.

B. Predicting the Rate of Severe Fires

Moving on to the next guiding question, the second model aims to predict the frequency of each class of severe fire over time in different states. This is accomplished by modifying the data set to count the number of each type of fire in each state, stratified by month and year. From there, a linear model is constructed for each fire class, using month, year, and state as inputs and the number of fires as the output. Training the set on 80% of the data (stratified by year to ensure a full time scale is represented) and testing it on the other 20% yields mean squared error (MSE) values as given in Table IV. The table also contains information about the standard deviation σ of the residuals. For context, the mean number of fires per month of each type is given as well.

TABLE IV
SUMMARY OF THE LINEAR MODEL FOR EACH CLASS OF FIRE.

Class	MSE	σ Residuals	Mean # Fires
D	55.57	7.45	1.97
E	27.95	5.29	0.98
F	24.00	4.90	0.54
G	19.17	4.38	0.26

Overall, the model has residuals that are more than 300% the mean values, indicating significant error in prediction. However, further investigating the statistically significant (with 95% confidence) variables in the models can uncover some interesting trends about fire patterns over time. In this case, the categorical variable of state has a baseline value associated with the number of fires in Alaska, which has average numbers of Class D, E, and F fires, but an extremely large number of Class G variables. For Class D fires, 11 states (AL, CA, FL, ID, KY, MN, MO, MS, OK, TX, WV) have statistically significant coefficients, all of which are positive. This indicates that these states are more prone to Class D fires. More than half of the months in the year also have statistically significant coefficients (in comparison to April, which has an above-average total number of fires as seen in Fig. 4). Of these, March and July have positive coefficients, indicating that Class D fires are more likely to occur then, while December, January, May, November, October, and September have negative coefficients.

For class E and F fires, we see a similar trend: about half the months of the year boast statistically significant coefficients, with March, June, July, and August having a positive influence. For class E fires, 5 states (CA, ID, KY, OK, TX) have statistically significant positive coefficients, and 8 states (CO, GA, HI, LA, NC, NE, NJ, SC) have statistically significant negative coefficients. For Class F fires, half of the states have statistically significant coefficients, with a mix between negative and positive. This trend reinforces the idea that, while class D fires tend to occur everywhere, with little variation between states, the higher class fires tend to be concentrated in certain parts of the country, meaning state has a greater influence on those models.

Finally, for Class G fires, 27 of the states have statistically significant coefficients, while only two months (June and August) have them, both of which are positive. Additionally, unlike Class D, E, and F fires, the "year" attribute is also statistically significant for Class G fires. The coefficient is positive, which corroborates the trend given in Fig. 7: while the numbers of D, E, and F fires are not increasing over time, the numbers of destructive G fires certainly are[1].

VI. LIMITATIONS & CONCLUSION

As evidenced by the performance of the models in the previous section, there are obvious limitations to fire prediction using this data set. One of the major limitations of the models arises from the information available, which doesn't encompass many of the factors that impact fire development. Weather conditions, including wind, heat, and humidity, are major drivers of fire impact, as are factors like distance from cities are and how equipped local firefighters are. Additionally, this analysis looks at trends on the national level, but trends may vary from state to state—perhaps Colorado is prone to fires in December, for example, which is a nuance that cannot be represented here.

It's also interesting to consider that not all fires are necessarily bad, and using size as the only judgement of impact is a simplified metric. Fires are actually vital to the maintenance of healthy ecosystems, so a better, more informed metric may consider impact to animal populations, structure damage, or human displacement. Tracking, fighting, and predicting wildfires is complex, nuanced, and difficult topic, and while this analysis does a good job of capturing the large trends, it lacks the information necessary to predict and mange fires on a local level.

In the end, this data set and the summaries and analyses provided in this report speak to the urgency of addressing the rise in severe fires across the US. Data illustrating the year-round prevalence of fires that impact nearly every state in the country should demonstrate the need for climate mitigation strategies and better tools for fire fighting. However, to fully address the nuances of fires, more research on a smaller scale needs to be conducted—research where every state and community addresses the myriad of factors that contribute to anticipating and mitigating fires close to home.

[1]While unruly to include here because of the categorical variables, full lists of coefficients can be found in the associated Google Colab notebook.