

# Predicting Total Burned Area of Wildfires Using Environmental and Fire Characteristics

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

2025-03-20

## ! Important

Before you submit, make sure your code chunks are turned off with `echo: false` and there are no warnings or messages with `warning: false` and `message: false` in the `YAML.Introduction`

## Introduction

Wildfires are among the most destructive natural disasters, causing significant environmental, economic, and human losses. In recent years, the United States has experienced some of the largest and most costly wildfires on record. These fires have burned [millions of acres](#), forced [thousands of residents to evacuate](#), and caused [billions of dollars in damages](#). The increasing scale and frequency of wildfires highlights the need for improved wildfire prediction and management strategies. Complicating this issue, climate change has worsened drought conditions, reduced vegetation moisture, and increased fuel availability, making wildfire behavior more unpredictable. Understanding the factors that influence the total area burned is essential for developing strategies to allocate resources efficiently and minimize damage.

This project examines wildfire attributes, environmental conditions, and geographical factors to predict the total burned area of a wildfire. Historical wildfire data, meteorological variables, and vegetation characteristics will be analyzed to determine the main factors contributing to fire spread. In our analysis, variables such as fire intensity, temperature, wind speed, humidity, precipitation, and land cover type will be evaluated for their impact on total acres burned, with the end goal of developing predictive models to estimate wildfire size based on these factors. A more accurate understanding of the conditions affecting fire spread will provide valuable insights for firefighting agencies, land managers, and policymakers to improve preparedness and response efforts.

Research Question:

- What environmental and fire-specific factors influence wildfire size?
- Can we develop models to estimate burned area based on geographic factors, weather conditions, land cover, and fire-specific information?

### Motivation & Importance

Wildfire response teams must quickly evaluate fire severity in order to determine the best course of action to contain and mitigate damage. However, limited firefighting resources often require prioritization based on expected fire growth. If fire size can be accurately predicted based on geographic factors, weather conditions, land cover, and fire-specific information, fire departments and policymakers can make more informed decisions about resource allocation.

The relevance of this research is especially important given the recent large-scale wildfires in California, as well as the recent wildfires in western North Carolina. By understanding important factors that influence fire size, fire agencies can prioritize high-risk areas adjust strategies proactively, and potentially reduce the total area burned. Importantly, predicting fire size is useful for national and global wildfire management and could help with disaster preparedness and policy development efforts. By identifying the most influential factors in fire spread and burned area, we hope to provide insights to improve fire prevention/suppression planning.

### Key Variables:

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 combines 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 database 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). Some key variables from the dataset are:

- **fire\_size** (acres): The total area burned by the fire, measured in acres.
- **fire\_size\_class**: The classification of fire size based on a standard A-G scale, where each class represents a different range of fire sizes.
- **stat\_cause\_descr**: The documented cause of the fire, such as lightning, human activity, or equipment use.
- **latitude** (degrees): The geographical latitude coordinate of the fire’s point of origin, measured in decimal degrees.
- **state**: The U.S. state where the fire occurred.
- **discovery\_month**: The month in which the fire was first detected or officially reported.
- **putout\_time** (days): The total duration from the fire’s discovery to its containment, measured in days.
- **Vegetation**: The dominant type of vegetation in the fire-affected area, categorized into specific vegetation classes, such as tropical forests, grasslands, shrublands, and urban land.
- **Temp\_pre\_30** (°C): The recorded temperature at the fire location up to 30 days before the fire was contained, measured in degrees Celsius.

- **Temp\_pre\_15** (°C): The recorded temperature at the fire location up to 15 days before the fire was contained, measured in degrees Celsius.
- **Temp\_pre\_7** (°C): The recorded temperature at the fire location up to 7 days before the fire was contained, measured in degrees Celsius.
- **Temp\_cont** (°C): The temperature recorded at the fire location on the day the fire was officially contained, measured in degrees Celsius.
- **Wind\_pre\_30** (m/s): The wind speed at the fire location up to 30 days before containment, measured in meters per second.
- **Wind\_pre\_15** (m/s): The wind speed at the fire location up to 15 days before containment, measured in meters per second.
- **Wind\_pre\_7** (m/s): The wind speed at the fire location up to 7 days before containment, measured in meters per second. **Wind\_cont** (m/s): The wind speed recorded at the fire location on the day the fire was contained, measured in meters per second.
- **Hum\_pre\_30** (%): The humidity level at the fire location up to 30 days before containment, expressed as a percentage.
- **Hum\_pre\_15** (%): The humidity level at the fire location up to 15 days before containment, expressed as a percentage.
- **Hum\_pre\_7** (%): The humidity level at the fire location up to 7 days before containment, expressed as a percentage.
- **Hum\_cont** (%): The humidity level recorded at the fire location on the day the fire was contained, expressed as a percentage.
- **Prec\_pre\_30** (mm): The total amount of precipitation recorded at the fire location up to 30 days before containment, measured in millimeters.
- **Prec\_pre\_15** (mm): The total amount of precipitation recorded at the fire location up to 15 days before containment, measured in millimeters.
- **Prec\_pre\_7** (mm): The total amount of precipitation recorded at the fire location up to 7 days before containment, measured in millimeters.
- **Prec\_cont** (mm): The total amount of precipitation recorded at the fire location on the day the fire was contained, measured in millimeters.
- **remoteness** (non-dimensional): A calculated measure representing the distance of the fire's location from the nearest city or major populated area, expressed as a non-dimensional value.

## Exploratory Data Analysis

### Data Cleaning

In order to focus on wildfires that are more manageable and relevant to fire prevention strategies, we filtered our dataset to include only the middle 50% of wildfire sizes (interquartile range). We excluded the smallest fires because they often burn insignificantly small areas (<1-2 acres) and may not require extensive intervention. Similarly, we removed the largest fires since they represent extreme cases that are difficult to control and may not reflect the majority of wildfires that fire management teams can realistically contain more effectively.

To prepare the dataset, we first removed observations with missing values in key variables (`fire_mag`, `stat_cause_descr`, `Vegetation`, `remoteness`, `Prec_pre_15`, `Hum_pre_7`, `fire_size`, `putout_time`) to ensure completeness. We then extracted numeric values from `putout_time` for consistency in analysis. Finally, we filtered the dataset to retain only the middle 50% of fire sizes using the interquartile range (IQR), so that our analysis remains focused on wildfires that are more representative of common fire management challenges.

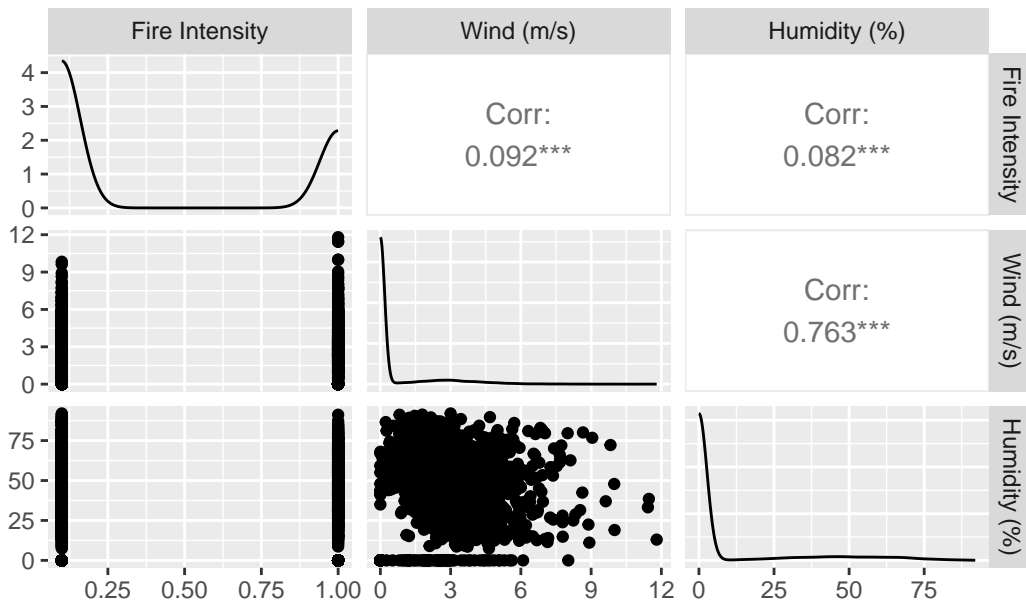
### Exploratory Data Analysis - Pairwise Plots

We began our analysis by creating pairwise plots to explore relationships between key numerical variables related to fire behavior, environmental conditions, and containment efforts. Since pairwise plots display scatterplots and correlations, we focused on continuous numerical variables that likely impact wildfire dynamics.

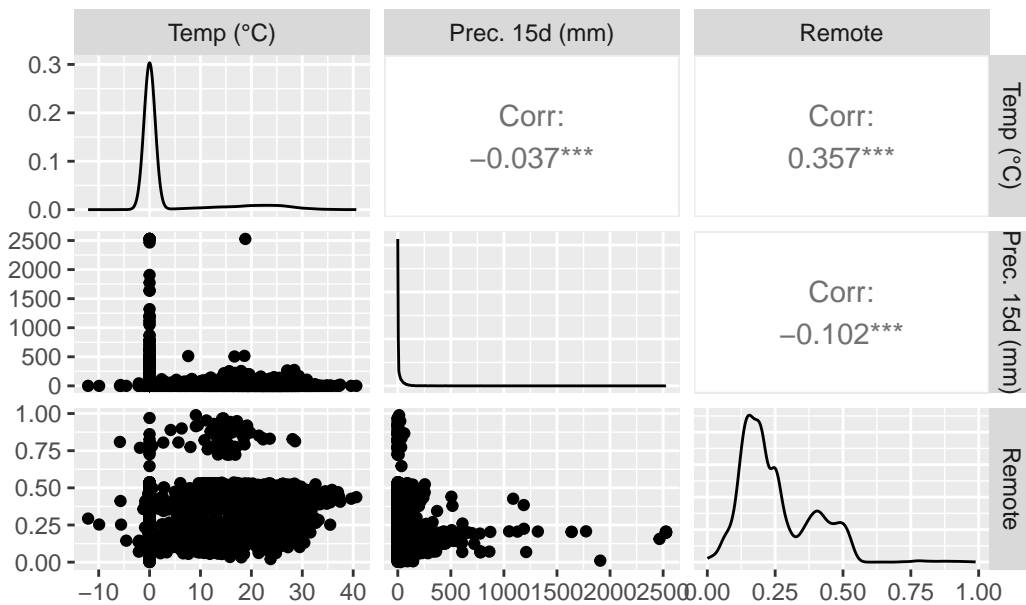
For the first pairwise plot, we examined the relationship between fire intensity, wind speed, and humidity. Fire intensity, represented by `fire_mag`, serves as a measure of how severe a wildfire is. Wind speed on the day of containment, `Wind_cont`, is an important factor because stronger winds can accelerate fire spread and make containment efforts more difficult. Humidity on the containment day, `Hum_cont`, was included since higher humidity levels can slow fire spread by increasing moisture in vegetation and the surrounding environment. Analyzing these three variables together provides insight into how atmospheric conditions influence wildfire intensity and containment efforts.

For the second pairwise plot, we selected temperature, precipitation, and remoteness to understand how fire conditions are affected by climate and location. Temperature on the day of containment, `Temp_cont`, plays a significant role because higher temperatures dry out vegetation, creating more favorable conditions for fire spread. Precipitation in the seven days prior to containment, `Prec_pre_15`, is relevant since recent rainfall can increase soil and vegetation moisture, which may reduce fire intensity. The remoteness of a fire's location, `remoteness`, influences how quickly firefighting resources can reach the site, which can affect containment time. Analysis of these variables allow us to better understand how environmental factors and accessibility impact wildfire behavior.

## Fire Intensity vs. Wind & Humidity



## Temperature vs. Precipitation & Remoteness

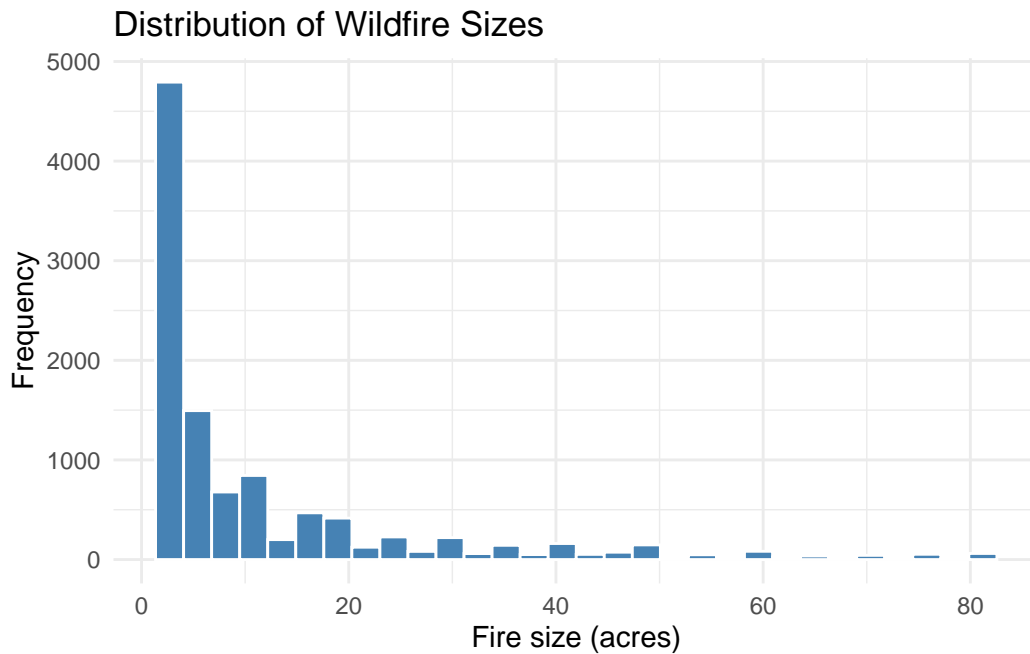


From our pairwise scatterplots, we observed that many relationships between variables do not follow a clear linear trend. One main observation is the vertical clustering of data points in several scatterplots, where points appear stacked on top of one another at specific values. This pattern is evident in fire intensity versus wind speed and fire intensity versus humidity in the

first plot, as well as temperature versus precipitation, temperature versus remoteness, and precipitation versus remoteness in the second plot. This clustering suggests that many of the measurements in our dataset are recorded in discrete increments rather than as continuous values. For example, wind speed and humidity may be rounded to the nearest whole number or recorded at set intervals, leading to apparent groupings in the data. Similarly, precipitation data may be stored as categorical or interval-based values rather than precise continuous measurements. This is an important consideration when preparing the dataset for modeling, as data transformation techniques may need to account for these discrete measurement patterns.

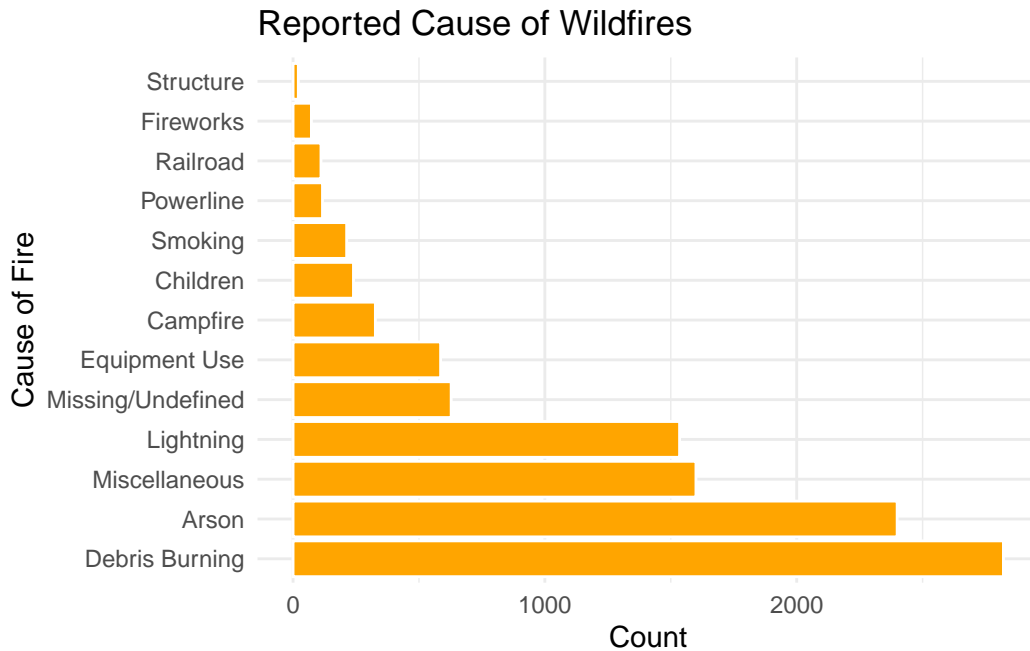
## Exploratory Data Analysis - Univariate EDA

As part of our univariate EDA, we first analyzed our response variable, `fire_size`, to understand its distribution and variability. Based on the histogram below, the distribution of fire size in our dataset appears to be right-skewed and unimodal. (As we can see in the histogram, the distribution has a long right tail, which indicates that while most wildfires in the dataset are relatively small, a smaller number of larger fires extend the range and pull the mean fire size to the right.) Only the middle 50% (IQR) of wildfires are shown in this histogram to focus on moderate-sized fires that are more relevant for management and containment strategies. Even within this subset, there is a clear peak around a frequency of ~4000 for wildfires that burn less than 1-2 acres of land.



Next, to better understand the primary causes of wildfires in our dataset, we examined the `stat_cause_descr` variable, which provides the reported cause for each fire. The most fre-

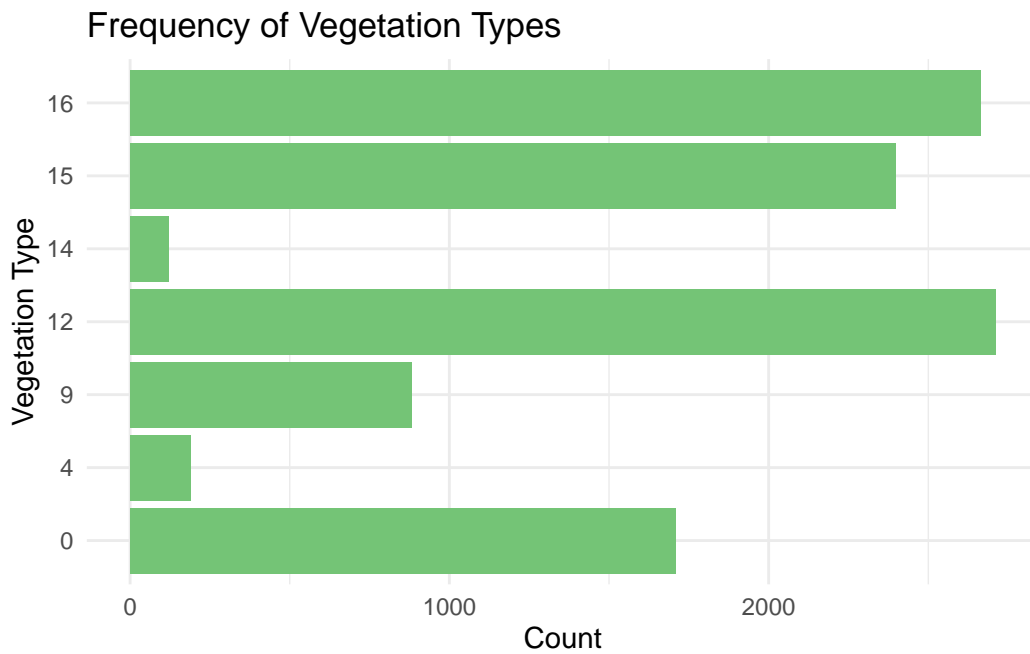
quently recorded cause was debris burning (2,821), followed by arson (2,399), miscellaneous causes (1,600), and lightning (1,535). Other causes included equipment use (586), campfires (327), children (240), smoking (213), and powerline-related fires (117). A total of 628 fires were labeled as missing or undefined. Interestingly, we found that many of the top causes (e.g., debris burning, arson, and equipment use) are related to human activity.



```
# A tibble: 13 x 2
  stat_cause_descr      n
  <chr>              <int>
1 Debris Burning      2821
2 Arson                2399
3 Miscellaneous        1600
4 Lightning            1535
5 Missing/Undefined    628
6 Equipment Use        586
7 Campfire             327
8 Children             240
9 Smoking              213
10 Powerline           117
11 Railroad             111
12 Fireworks            73
13 Structure            21
```



0	4	9	12	14	15	16
1708	189	882	2711	120	2397	2664

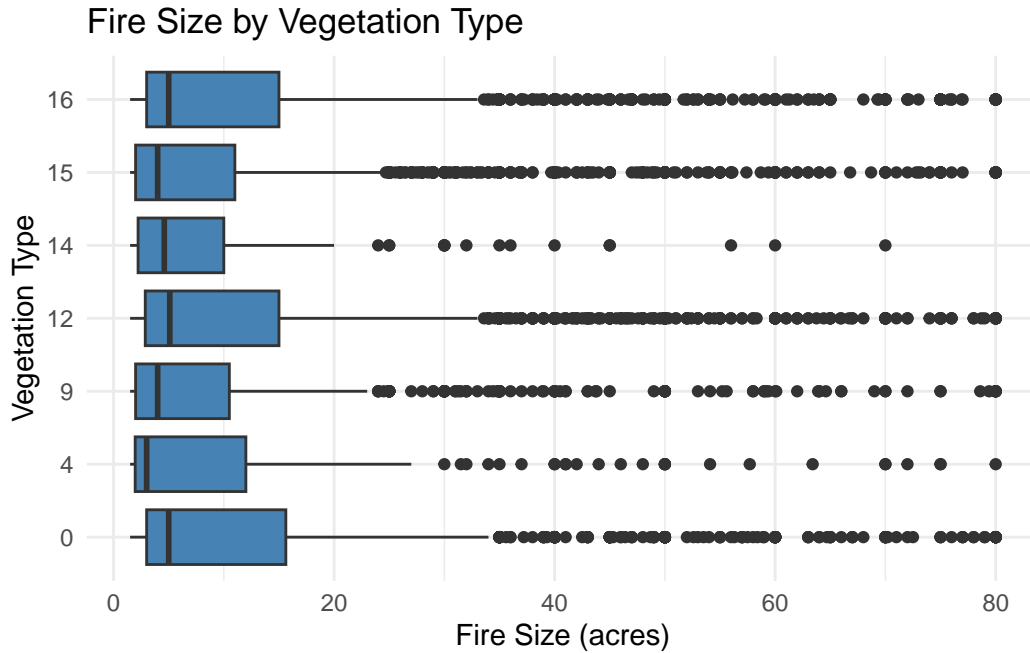


0	4	9	12	14	15	16
16.005998	1.771155	8.265392	25.405304	1.124543	22.462750	24.964858

The dataset includes 28 distinct vegetation types, each classified by a numerical code. Among these, a few vegetation types dominate the data. The most frequently occurring types are:

- Open Shrubland (code 12) with 3,763 observations (about 26% of the dataset)
- Secondary Tropical Evergreen Broadleaf Forest (code 16) with 3,653 observations (about 26%)
- Polar Desert/Rock/Ice (code 15) with 3,081 observations (about 22%)

Less common vegetation types represented in the data include desert and temperate evergreen needleleaf forests.



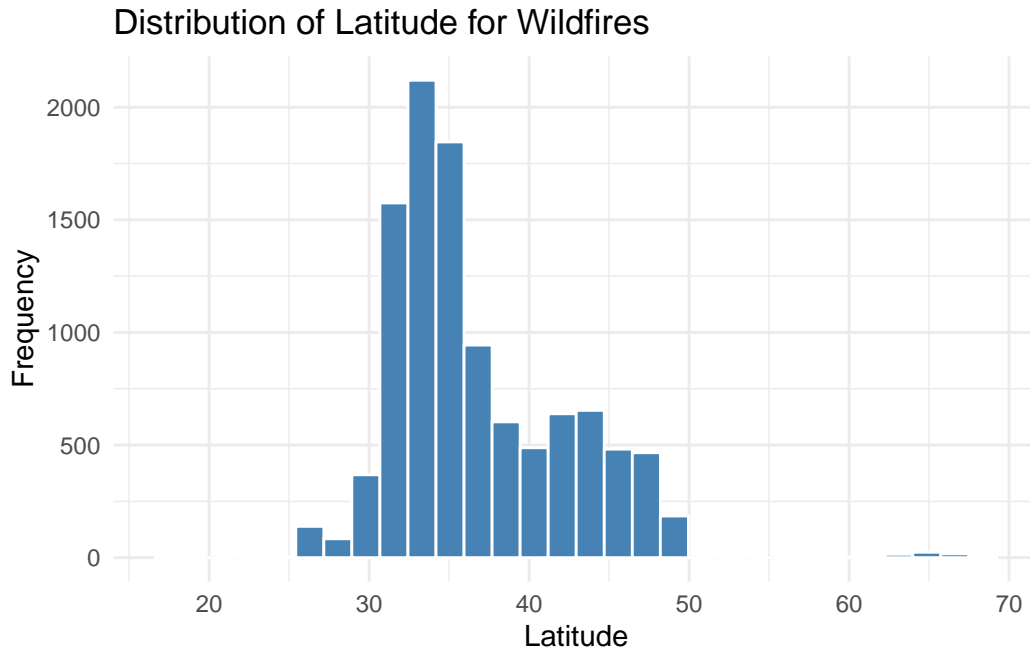
```
# A tibble: 7 x 4
  Vegetation count mean_fire_size median_fire_size
  <fct>      <int>      <dbl>         <dbl>
1 0          1708      13.2           5
2 12         2711      12.4          5.1
3 16         2664      12.0           5
4 4           189      11.7           3
5 15         2397      10.6           4
6 9           882      10.4           4
7 14           120       9.73          4.6
```

We explored the relationship between vegetation type and fire size using a boxplot and summary statistics. The boxplot shows variation in fire size distributions across vegetation types. Notably, vegetation type 0 has the largest spread and the highest median fire size at 5.0 acres. This type also has the highest mean fire size at 11.02 acres, suggesting a tendency for larger fires in this vegetation category.

Vegetation types 12 and 16 also exhibit relatively high median fire sizes (both around 5.0 acres) with mean fire sizes of 10.15 and 9.67 acres, respectively. Conversely, vegetation types 9 and 4 have lower median fire sizes (3.3 and 3.0 acres), along with the lowest mean fire sizes, 8.33 and 8.39 acres, respectively.

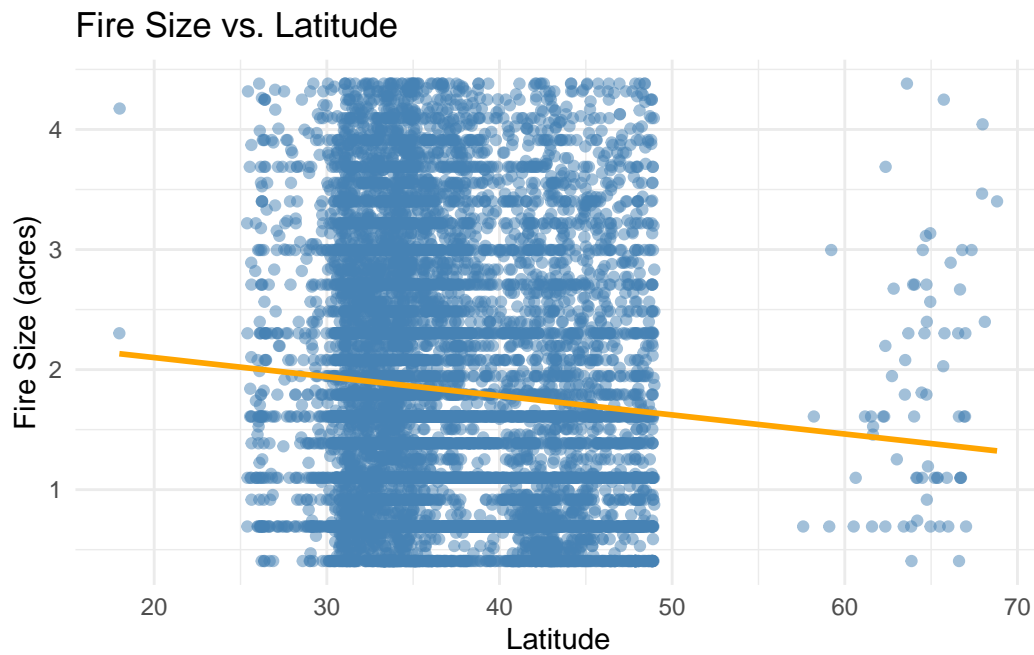
Overall, there is noticeable variation in fire size depending on vegetation type. Some vegetation types appear to be more prone to larger fires, which could be due to factors like fuel availability or vegetation density.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17.98	32.96	34.98	36.84	40.75	68.82



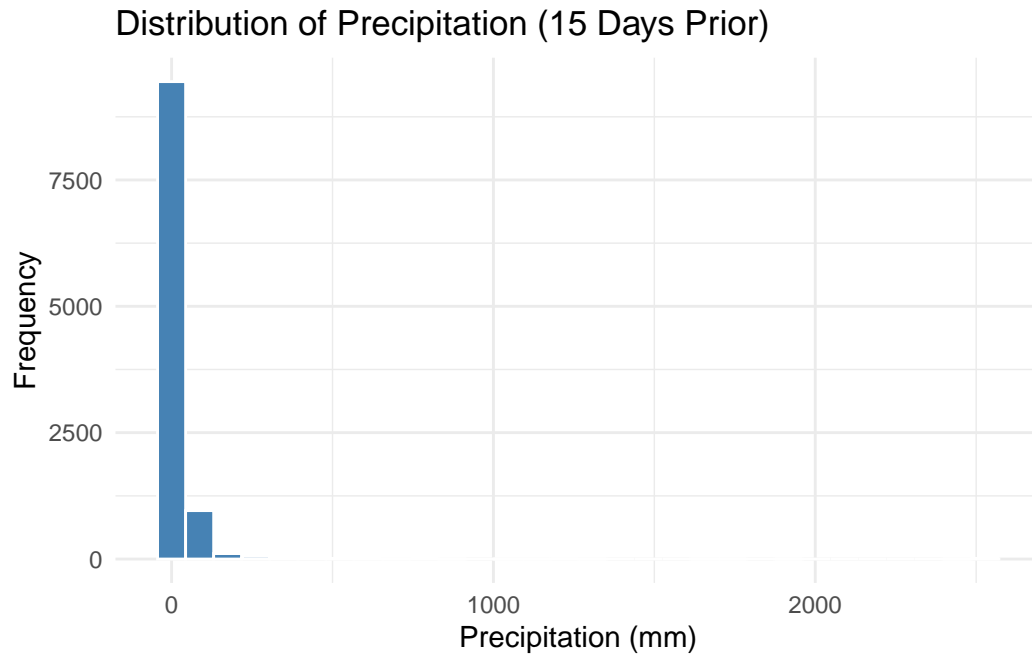
The latitude of wildfires in the dataset ranges from 17.98° to 69.26°, with a median of 34.84° and a mean of 36.62°. The middle 50% of the data falls between 32.85° and 39.93°, suggesting that most wildfires occur in mid-latitude regions of the United States. This range corresponds to areas that commonly experience wildfires, such as parts of California and other western states.

The distribution of latitudes appears to be centered around the mid-30s to upper-30s, which may reflect the concentration of fire-prone areas in those geographic zones.



[1] -0.04720852

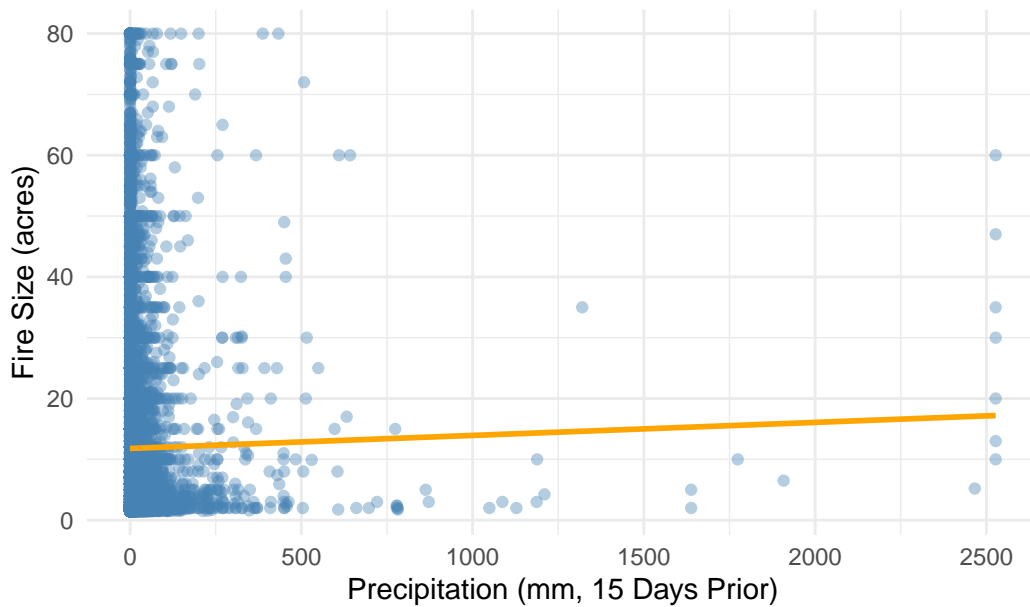
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	0.0	0.0	20.8	13.8	2527.0



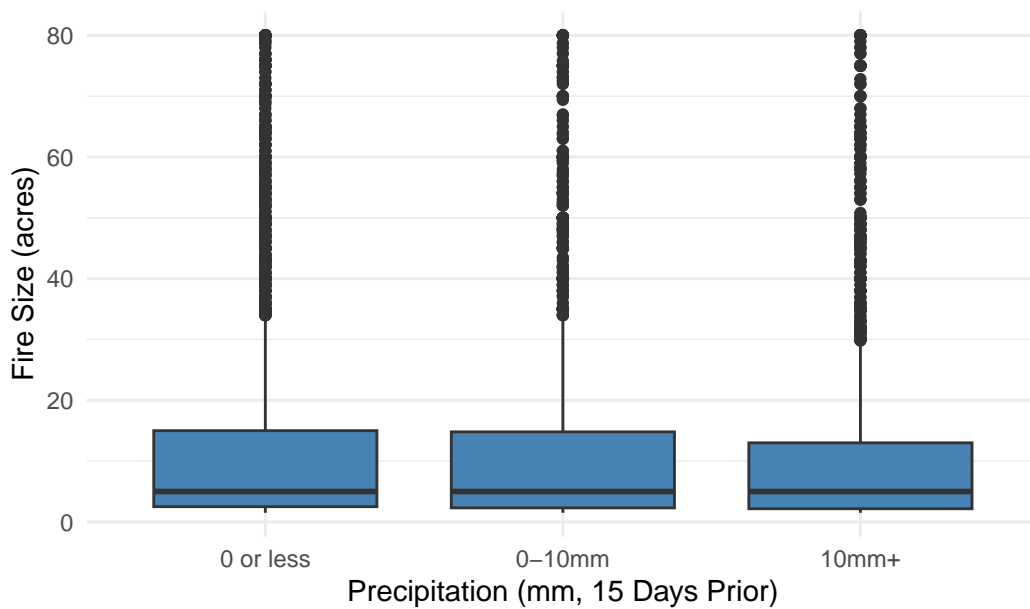
Most values are 0 mm or -1 mm, likely indicating missing data. The median is 0 mm, while the mean is 14.63 mm, skewed by extreme outliers (up to 2,527 mm).

The distribution is highly right-skewed, with most fires occurring after little to no precipitation, consistent with dry conditions increasing fire risk.

Fire Size vs. Precipitation (15 Days Prior)



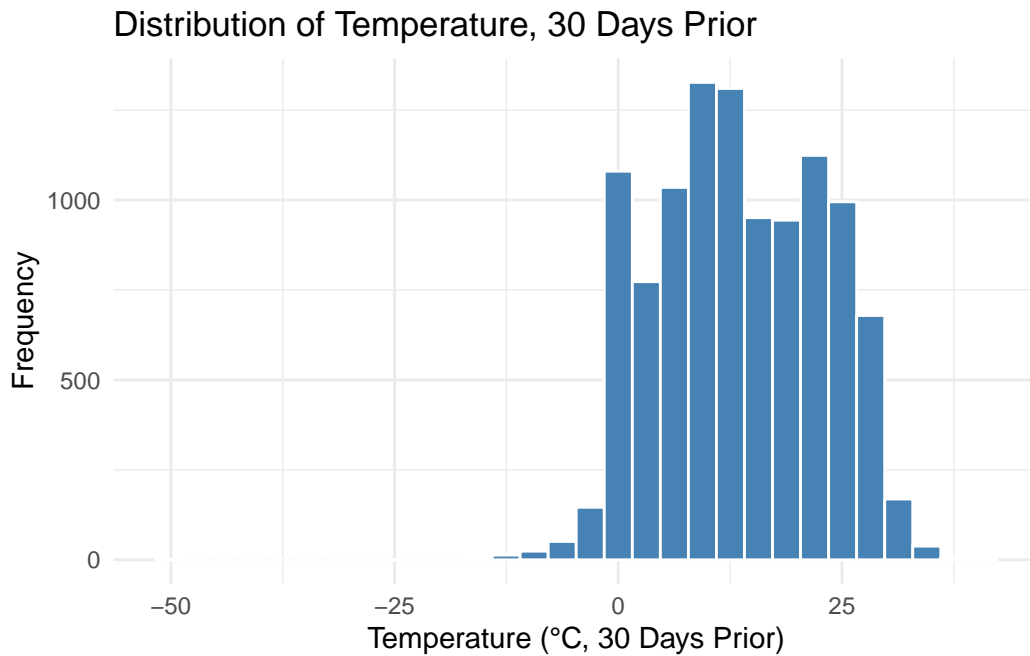
Fire Size by Precipitation Bins



The scatterplot of `Prec_pre_15` and `fire_size` shows no strong relationship between prior precipitation and fire size. Most fires occurred with little to no precipitation, and there is substantial variability in fire size regardless of precipitation level. A few extreme precipitation values do not appear to have a significant effect on fire size.

The boxplot comparing precipitation bins (0 or less, 0-10mm, 10mm+) reveals similar distributions of fire size across all groups, with no meaningful differences in medians or spread. This suggests that precipitation up to 15 days prior to a fire may have limited impact on the size of the fire in this dataset.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-49.211	6.531	12.972	13.578	21.356	41.678



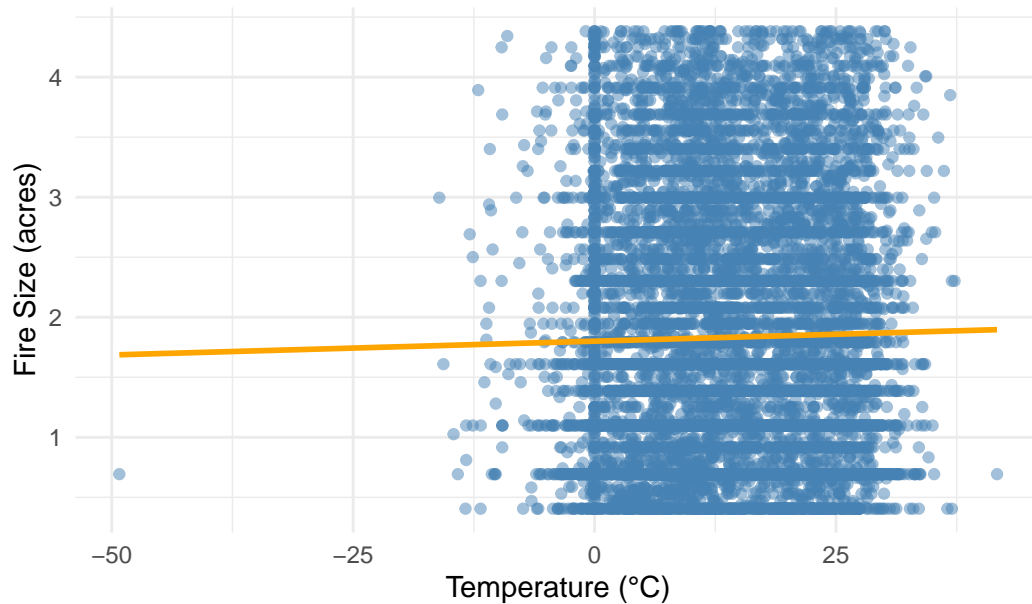
The distribution of Temp\_pre\_30, which represents the temperature at the location of a fire up to 30 days prior, shows a wide range of values, spanning from -49.211°C to 41.678°C. The minimum recorded temperature of -49.211°C is exceptionally low and may indicate an outlier or a potential data entry error, especially considering that there are few locations that can get that cold and have a fire. Additionally, the first quartile of -1.000°C suggests that at least 25% of recorded fires happened in temperatures at or below freezing, which could indicate that fires occurred in cold regions or during winter months.

The median temperature of 8.310°C suggests that half of the recorded fires occurred in moderate conditions, while the mean of 9.562°C, which is slightly higher than the median, indicates that the distribution is right-skewed. This means that some extremely high-temperature values may be pulling the average up. The maximum recorded temperature of 41.678°C suggests that some fires took place in extremely hot environments, which does make a lot of sense.

Looking at the histogram, the data appears to be slightly right-skewed, with most temperatures falling between 0°C and 25°C, while some extend into negative values as low as -50°C. The

presence of extreme negative temperatures raises concerns about outliers or data recording errors, as fires typically occur in warmer conditions. As there is a huge spike in temperatures of 0°C, we will probably need to look more into whether or not these are errors or actual measurements.

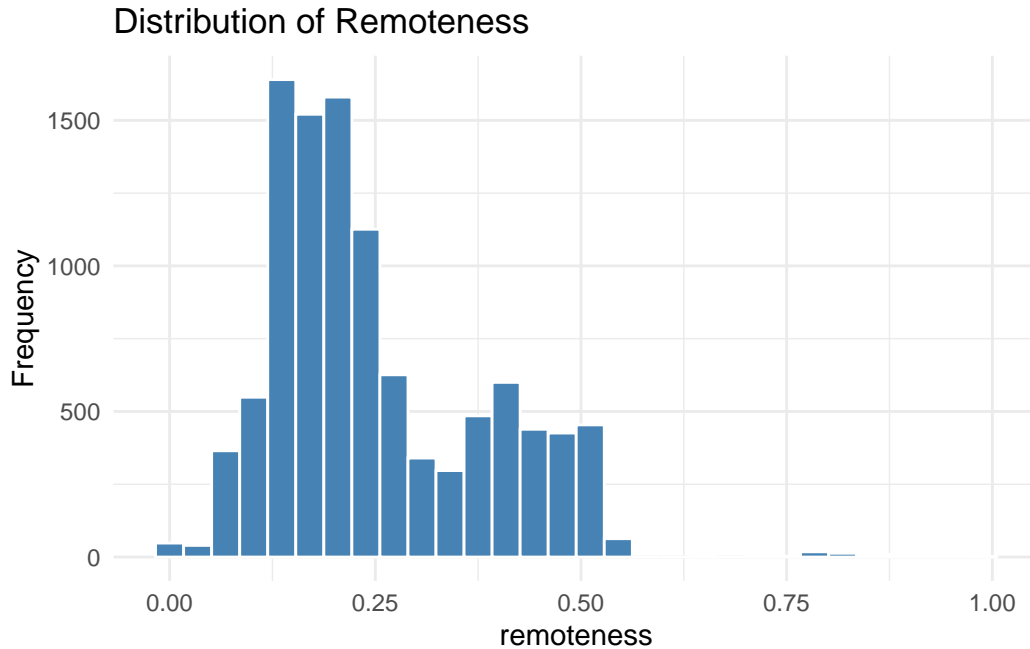
Fire Size vs. Temperature (30 Days Prior)



The scatter plot shows the relationship between fire size and temperature 30 days prior, with little evidence of a strong correlation. The nearly flat regression line suggests that temperature alone is not a key predictor of fire size. Most fires occurred between 0°C and 25°C, with fire sizes clustered at lower values. There also does seem to be a large amount of fires that are recorded at 20, 30, 40, and 50 acres for fire size indicating that many of these may have been rounded.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0003903	0.1541690	0.2063234	0.2514805	0.3469825	0.9888925



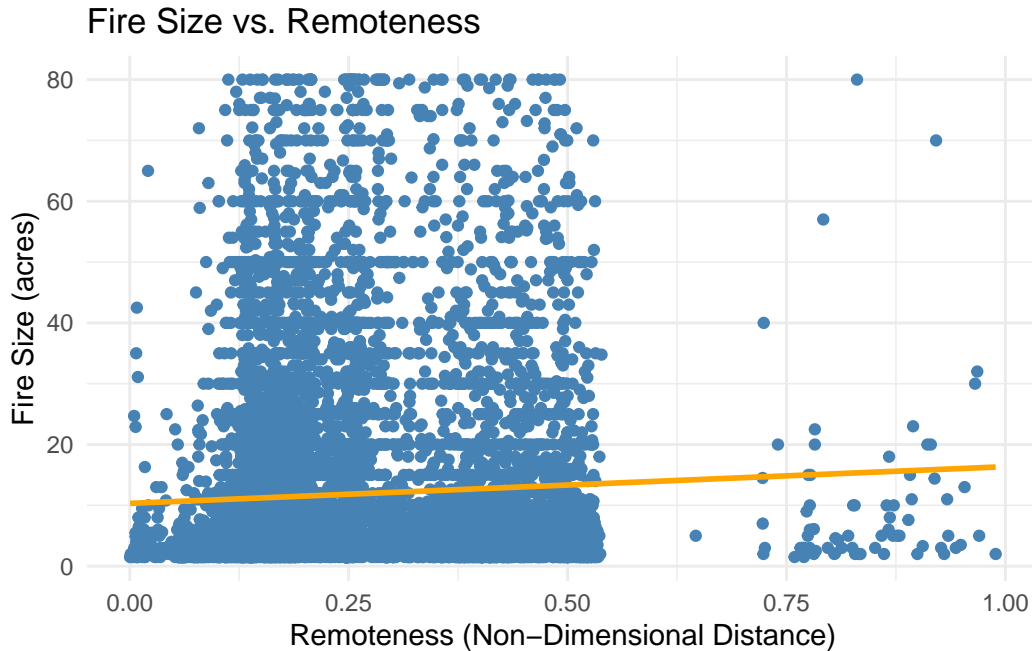


The variable remoteness, which represents the non-dimensional distance to the closest city, has a distribution ranging from 0.0000 to 0.98899. The minimum value of 0 suggests that some fires occurred basically within cities, while the maximum value of 1 indicating the farthest fire away. However, the fire with value of 1 in remoteness may have been removed when we cleaned the data.

The first quartile of 0.1450 means that 25% of the fires occurred in areas where remoteness was relatively low, suggesting proximity to cities or towns. The median value of 0.2002 indicates that half of the fires took place in areas with remoteness below this threshold, meaning that most fires are moderately close to urban areas rather than in extremely remote locations. However, the mean value of 0.2403 is slightly higher than the median, which suggests that the distribution is right-skewed, meaning that a small number of fires occurred in highly remote areas, pulling the average upward. This skewness is further confirmed by the third quartile (Q3) of 0.3018, showing that 75% of fires occurred in areas with remoteness below this level, while the remaining 25% took place in much more remote regions.

This distribution suggests that most fires tend to occur closer to urban areas rather than in extremely remote locations, but a minority of cases involve fires in highly remote regions.

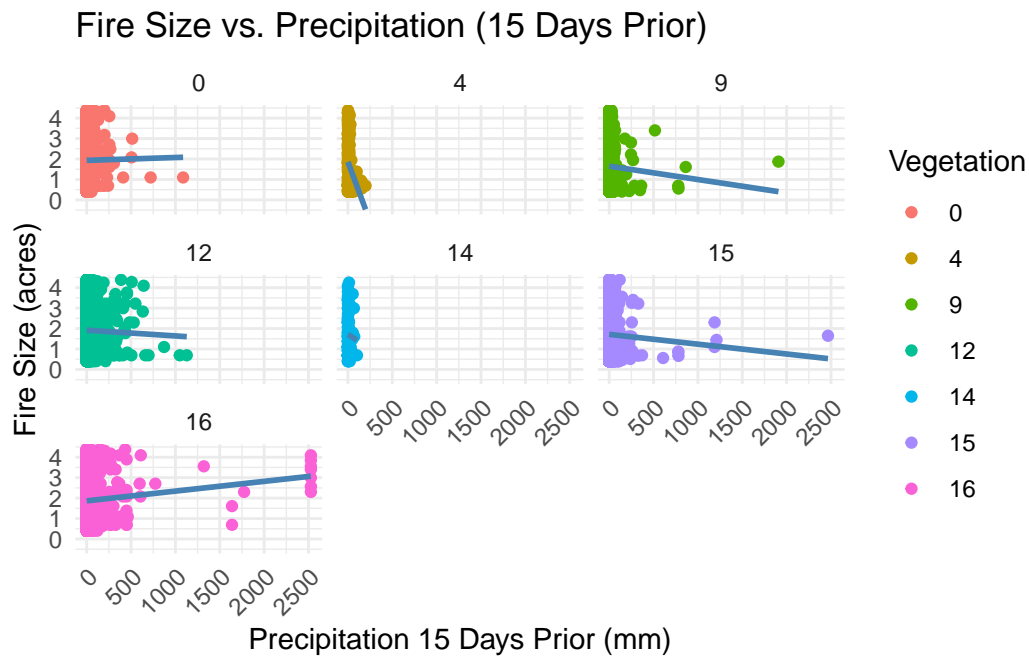
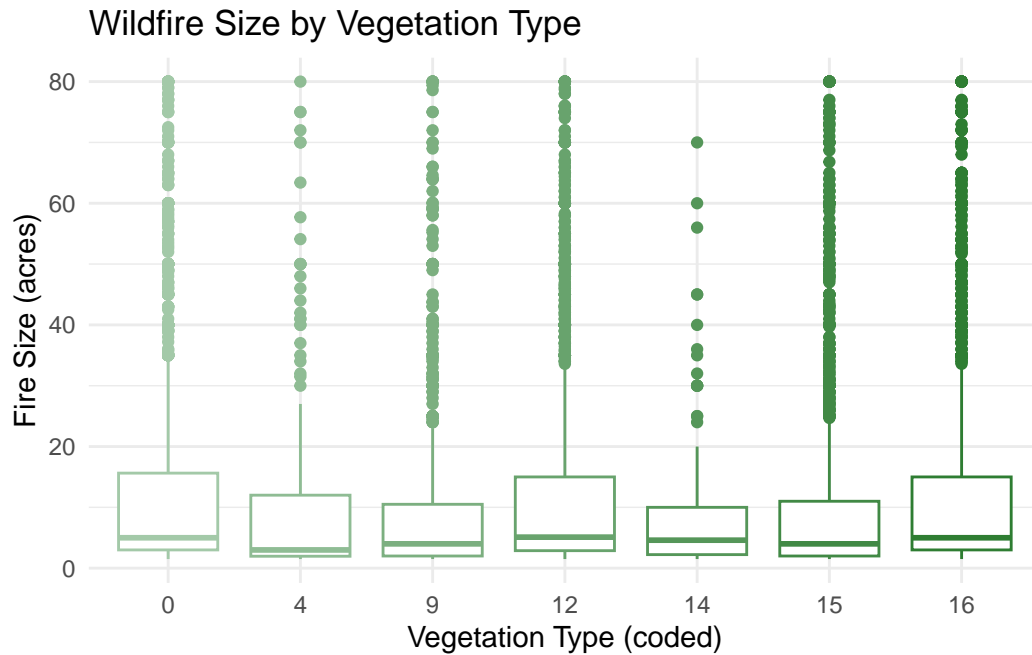
Looking at the histogram, the distribution is right-skewed, with most observations between 0.1 and 0.3. There is a sharp peak around 0.15, however, indicating that this is about the area most fires occur. As remoteness increases beyond 0.5, the frequency of observations declines significantly, meaning very few fires occurring in extremely remote areas.



The scatter plot illustrates the relationship between fire size and remoteness. The red regression line shows a slight positive trend, suggesting that more remote fires tend to be slightly larger. Most fires occur in areas closer to cities, as indicated by the dense clustering on the left side of the plot. While some larger fires occur in highly remote areas, the overall pattern does not strongly indicate that remoteness is a key driver of fire size.

### Exploratory Data Analysis - Bivariate EDA

This plot shows the distribution of wildfire sizes across different vegetation types, which represent the dominant land cover where each fire occurred. The median fire size is relatively consistent across most vegetation types, typically falling between 5 and 10 acres. However, certain vegetation categories show greater variability in fire size, particularly type 0 (i.e., Other), type 12 (i.e., Open Shrubland), and type 16 (i.e., Secondary Tropical Evergreen Broadleaf Forest). Additionally, nearly all vegetation types include significant outliers (i.e., fires extending beyond 40 acres). This suggests that while typical fire behavior is similar across land cover types, variability still exists in more extreme cases.



This figure suggests that the relationship between precipitation (15 days prior) and fire size varies across different vegetation types. The slopes of the regression lines differ by vegetation category. This could indicate that precipitation might have differing effects on fire size depending on the vegetation type. Thus, an interaction between precipitation and vegetation

type could add to a model predicting fire size.

## Methodology

First, as part of our modeling process and to determine which variables to include as predictors of wildfire size, we evaluated multicollinearity among the numeric predictors in our dataset. Specifically, we considered environmental variables measured 15 days prior to fire discovery, including precipitation (`Prec_pre_15`), temperature (`Temp_pre_15`), wind speed (`Wind_pre_15`), and humidity (`Hum_pre_15`), as well as `remoteness`, a scaled measure of distance to the nearest city. These variables were selected because they are directly tied to environmental conditions and are available for most observations. To assess multicollinearity, we used the Variance Inflation Factor (VIF), which quantifies how much the variance of each predictor is inflated due to linear dependence with the others. VIF values greater than 5 were used as a threshold to flag potential multicollinearity concerns. As shown in the output, all five predictors had VIF values well below the threshold of 5: precipitation (1.03), temperature (1.12), wind speed (1.24), humidity (1.39), and remoteness (1.20). These results suggest that there is no problematic multicollinearity among the selected predictors, and all variables were retained for use in the model.

<code>Prec_pre_15</code>	<code>Temp_pre_15</code>	<code>Wind_pre_15</code>	<code>Hum_pre_15</code>	<code>remoteness</code>
1.028289	1.120186	1.242966	1.387552	1.201539

After evaluating multicollinearity, we fit two linear regression models to predict wildfire size using a log-transformed version of the response variable (`log(fire_size)`). We applied the log transformation to reduce skewness in the fire size distribution and to stabilize variance. A small constant (+ 0.001) was added to `Prec_pre_15` before log transformation to avoid taking the logarithm of zero.

The first model (`fire_main_fit`) included the main effects of `remoteness`, log-transformed `Prec_pre_15`, `Temp_pre_15`, `Vegetation`, `stat_cause_descr`, and `Wind_cont`. The second model (`fire_int_fit`) added an interaction term between `log(Prec_pre_15 + 0.001)` and `Vegetation` to explore whether the effect of precipitation on fire size varies by vegetation type.

The adjusted  $R^2$  values for the models were low: 0.037 for the main-effects model and 0.041 for the interaction model. Although the interaction model showed a slightly higher adjusted  $R^2$ , the difference in explained variance was minimal.

To evaluate predictive accuracy, we computed the root mean squared error (RMSE) for both models using the log-transformed fire size. RMSE represents the average difference between the observed and predicted values on the log scale. The main-effects model had an RMSE of approximately 1.06, and the interaction model had a nearly identical RMSE of 1.059. These values indicate that, on average, the predicted log fire sizes differed from the observed values

by about 1.057 log-units. Because the models are fit on a logarithmic scale, an error of this magnitude corresponds to a multiplicative factor of roughly 2.7 on the original fire size scale (in acres).

term	estimate	std.error	statistic	p.value
(Intercept)	1.967	0.041	47.815	0.000
remoteness	0.366	0.092	3.986	0.000
log(Prec_pre_15 + 0.001)	-0.003	0.002	-1.221	0.222
Temp_pre_15	-0.005	0.001	-3.680	0.000
Vegetation4	-0.285	0.082	-3.491	0.000
Vegetation9	-0.271	0.044	-6.125	0.000
Vegetation12	0.074	0.035	2.101	0.036
Vegetation14	-0.260	0.100	-2.594	0.010
Vegetation15	-0.204	0.034	-6.063	0.000
Vegetation16	0.003	0.034	0.101	0.919
stat_cause_descrCampfire	-0.290	0.063	-4.612	0.000
stat_cause_descrChildren	-0.576	0.072	-7.983	0.000
stat_cause_descrDebris Burning	-0.227	0.030	-7.656	0.000
stat_cause_descrEquipment Use	-0.027	0.050	-0.539	0.590
stat_cause_descrFireworks	-0.030	0.127	-0.237	0.812
stat_cause_descrLightning	-0.121	0.039	-3.127	0.002
stat_cause_descrMiscellaneous	-0.240	0.035	-6.896	0.000
stat_cause_descrMissing/Undefined	-0.155	0.048	-3.237	0.001
stat_cause_descrPowerline	-0.231	0.101	-2.293	0.022
stat_cause_descrRailroad	-0.234	0.103	-2.266	0.023
stat_cause_descrSmoking	-0.383	0.076	-5.028	0.000
stat_cause_descrStructure	-0.628	0.233	-2.703	0.007
Wind_cont	0.090	0.009	10.429	0.000

term	estimate	std.error	statistic	p.value
(Intercept)	2.018	0.043	47.478	0.000
remoteness	0.362	0.092	3.921	0.000
Temp_pre_15	-0.005	0.001	-3.777	0.000
log(Prec_pre_15 + 0.001)	0.016	0.005	2.956	0.003
Vegetation4	-0.362	0.083	-4.368	0.000
Vegetation9	-0.357	0.047	-7.544	0.000
Vegetation12	0.006	0.039	0.162	0.871
Vegetation14	-0.301	0.115	-2.615	0.009
Vegetation15	-0.257	0.038	-6.807	0.000
Vegetation16	-0.012	0.039	-0.303	0.762

term	estimate	std.error	statistic	p.value
stat_cause_descrCampfire	-0.297	0.063	-4.729	0.000
stat_cause_descrChildren	-0.569	0.072	-7.894	0.000
stat_cause_descrDebris Burning	-0.231	0.030	-7.807	0.000
stat_cause_descrEquipment Use	-0.029	0.050	-0.591	0.555
stat_cause_descrFireworks	-0.041	0.127	-0.325	0.745
stat_cause_descrLightning	-0.121	0.039	-3.136	0.002
stat_cause_descrMiscellaneous	-0.229	0.035	-6.594	0.000
stat_cause_descrMissing/Undefined	-0.165	0.048	-3.451	0.001
stat_cause_descrPowerline	-0.230	0.101	-2.285	0.022
stat_cause_descrRailroad	-0.237	0.103	-2.296	0.022
stat_cause_descrSmoking	-0.375	0.076	-4.935	0.000
stat_cause_descrStructure	-0.622	0.232	-2.681	0.007
Wind_cont	0.089	0.009	10.328	0.000
log(Prec_pre_15 + 0.001):Vegetation4	-0.077	0.018	-4.336	0.000
log(Prec_pre_15 + 0.001):Vegetation9	-0.047	0.009	-5.015	0.000
log(Prec_pre_15 + 0.001):Vegetation12	-0.025	0.007	-3.779	0.000
log(Prec_pre_15 + 0.001):Vegetation14	-0.016	0.022	-0.740	0.459
log(Prec_pre_15 + 0.001):Vegetation15	-0.021	0.007	-2.949	0.003
log(Prec_pre_15 + 0.001):Vegetation16	-0.009	0.007	-1.287	0.198

[1] 0.03713495

[1] 0.04071697

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>         <dbl>
1 rmse    standard         1.06
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>         <dbl>
1 rmse    standard         1.06
```

After evaluating our interaction model, we decided to explore a simpler modeling strategy by converting precipitation into an indicator variable. Rather than using continuous log-transformed precipitation, we created a binary variable to indicate whether any measurable precipitation was recorded 15 days prior to the fire. This choice was motivated by the idea that the presence or absence of precipitation may be more informative for fire behavior than small

variations in precipitation amounts, particularly given the high frequency of zero or near-zero values in the dataset.

To do this, we created a new variable called `precip_indicator`, which takes the value 1 if `Prec_pre_15` is greater than 0 and 0 otherwise. We then fit two new models: a main-effects model and an interaction model including the new binary indicator and its interaction with vegetation type. We evaluated model performance using adjusted  $R^2$  and RMSE on the log-transformed fire size, consistent with earlier analyses.

The adjusted  $R^2$  of the main-effects model was 0.037. The interaction model had an adjusted  $R^2$  of 0.039. RMSE values were 1.059 for the main-effects model and 1.058 for the interaction model. These values indicate that the average prediction error on the log scale was approximately 1.06. This corresponds to a multiplicative error of about 2.88 on the original scale of fire size. The differences between these models and the earlier continuous-precipitation models were minimal in both adjusted  $R^2$  and RMSE.

```
[1] 0.03710429
```

```
[1] 0.03934171
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>         <dbl>
1 rmse    standard         1.06
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>         <dbl>
1 rmse    standard         1.06
```

## Results

We fit four linear regression models to predict `log(fire_size)` using combinations of environmental and fire-related predictors. Two models included precipitation as a continuous predictor (`log(Prec_pre_15 + 0.001)`), and two models used a binary indicator variable representing the presence of precipitation. Each modeling approach included a main-effects model and an interaction model with `Vegetation`.

The adjusted  $R^2$  values were low across all models. The main-effects model with continuous precipitation (`fire_main_fit`) had an adjusted  $R^2$  of 0.037, and the interaction model

(`fire_int_fit`) had an adjusted  $R^2$  of 0.041. The models using the binary precipitation indicator produced similar results: 0.037 for the main-effects model (`fire_main_bin`) and 0.039 for the interaction model (`fire_int_bin`).

We evaluated prediction error using root mean squared error (RMSE) on the log scale. The RMSE was 1.059 for both `fire_main_fit` and `fire_main_bin`, and slightly lower for the interaction models: 1.057 for `fire_int_fit` and 1.058 for `fire_int_bin`. These values indicate that the average prediction error on the log-transformed scale was around 1.06. On the original scale, this corresponds to a multiplicative prediction error of approximately 2.88 (i.e.,  $\exp(1.06) \approx 2.88$ ).

Overall, there were minimal differences in adjusted  $R^2$  and RMSE between the models using continuous precipitation and those using a binary indicator. The addition of interaction terms with `Vegetation` slightly increased model complexity without meaningfully improving predictive performance.