# 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, California 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 emphasize the need for improved wildfire risk assessment and management strategies. Climate change has intensified drought conditions, reduced vegetation moisture, and increased fuel availability, making wildfire behavior more unpredictable. Understanding the factors that influence the total area burned is critical 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 primary drivers of fire spread. Key variables such as fire intensity, temperature, wind speed, humidity, precipitation, and land cover type will be evaluated for their impact on total acres burned. The objective is to develop a predictive model that estimates wildfire size based on these factors. A more accurate understanding of fire spread dynamics will provide valuable insights for firefighting agencies, land managers, and policymakers to enhance wildfire preparedness and response efforts.

Research Question:

What environmental and fire-specific factors influence wildfire size, and can we develop a predictive model to estimate burned area based on weather conditions, vegetation type, and remoteness?

Motivation & Importance

Wildfire response teams must quickly assess fire severity and 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 vegetation type, weather conditions, and remoteness, fire departments and policymakers can make more informed decisions about resource allocation and risk mitigation.

The relevance of this research is especially important given the recent large-scale wildfires in California. By understanding the key factors that influence fire size, fire agencies can prioritize high-risk areas, adjust strategies proactively, and potentially reduce the total area burned.

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

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

   NOAA gathers Integrated Surface Hourly data from thousands of global weather stations, including those managed by the National Weather Service, the Federal Aviation Administration, and international meteorological organizations. The data include hourly

observations of temperature, wind speed, humidity, and other meteorological variables, primarily collected through automated and manual weather stations.

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

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

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

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

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

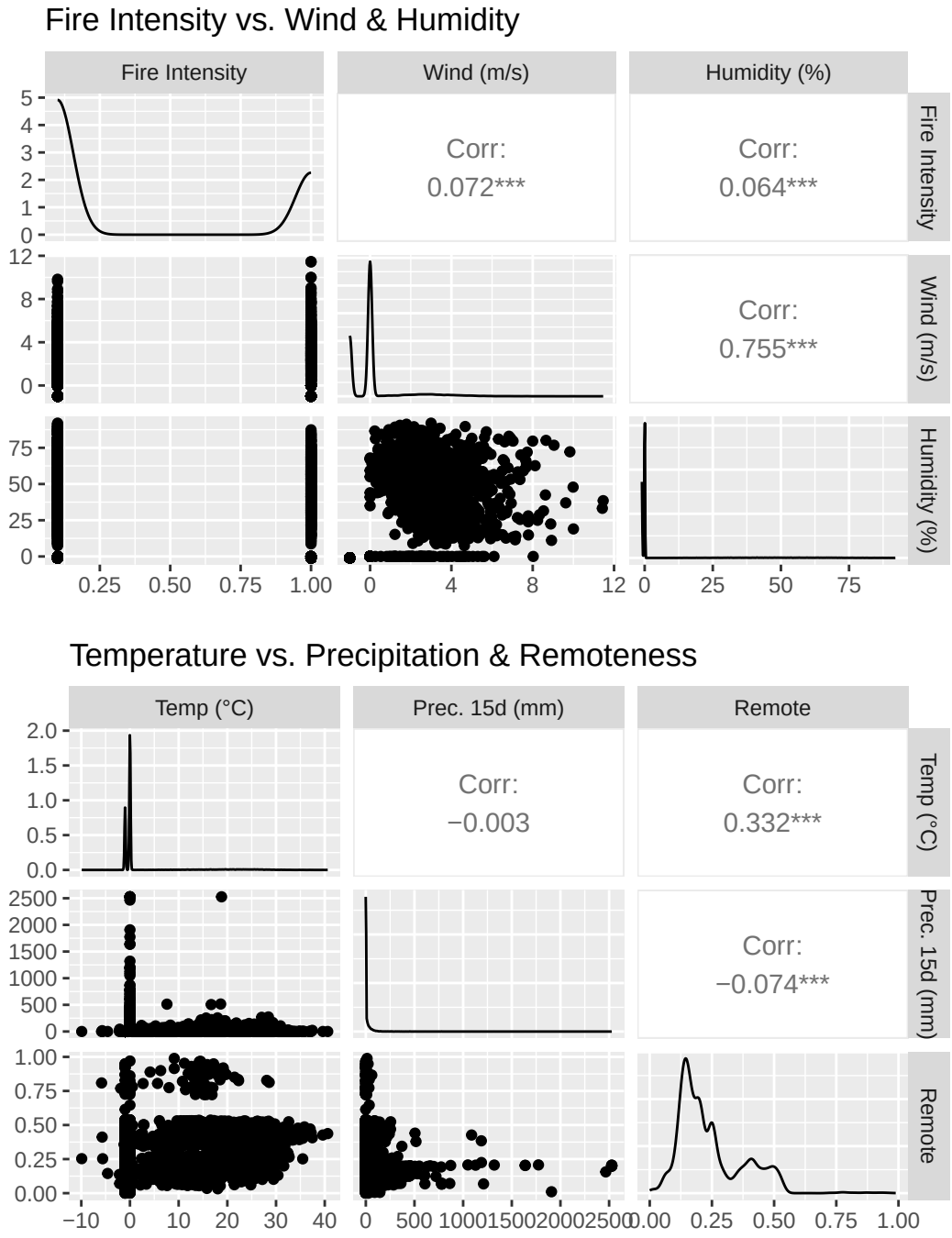
To prepare the dataset, we first removed observations with missing values in key variables (`fire_mag`, `stat_cause_descr`, `Vegetation`, `remoteness`, `Prec_pre_15`, `Hum_pre_7`, `fire_size`, `putout_time`) to ensure completeness. We then extracted numeric values from `putout_time` for consistency in 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.

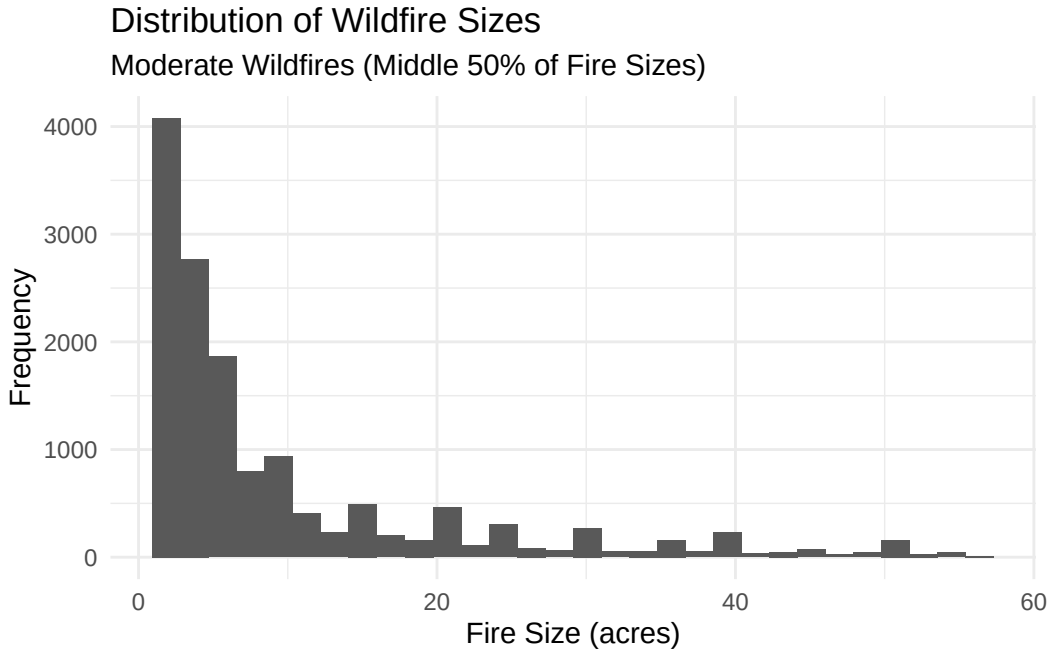



From our pairwise scatterplots, we observed that many relationships between variables do not follow a clear linear trend. One main observation is the vertical clustering of data points in several scatterplots, where points appear stacked on top of one another at specific values. This pattern is evident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

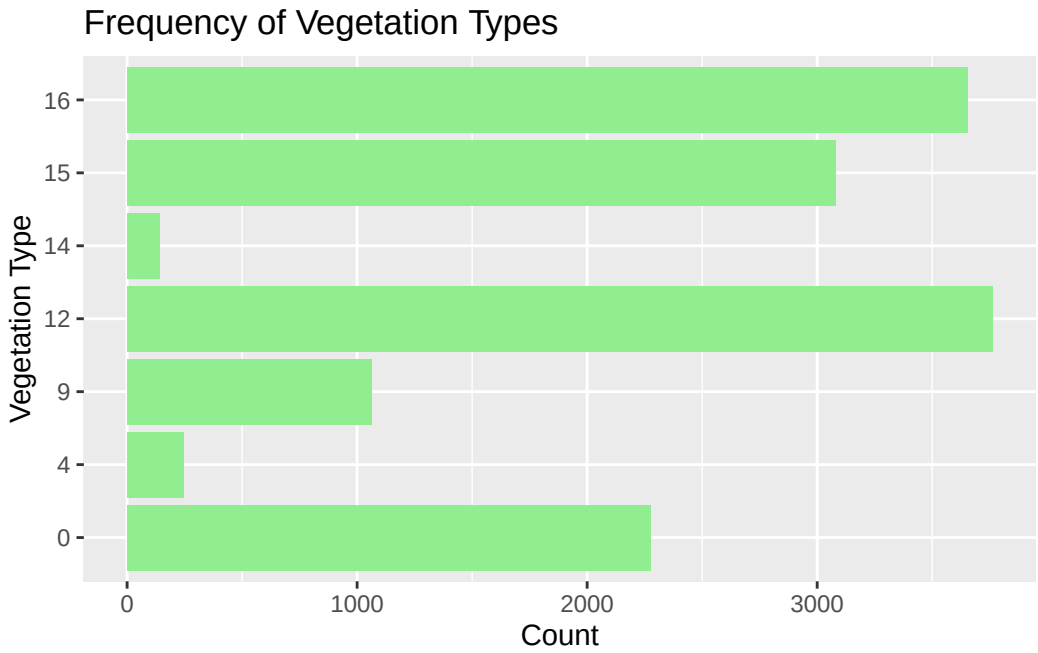
### Exploratory Data Analysis - Univariate EDA

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.

```
   0    4    9   12   14   15   16
2278  246 1065 3763  143 3081 3653
```

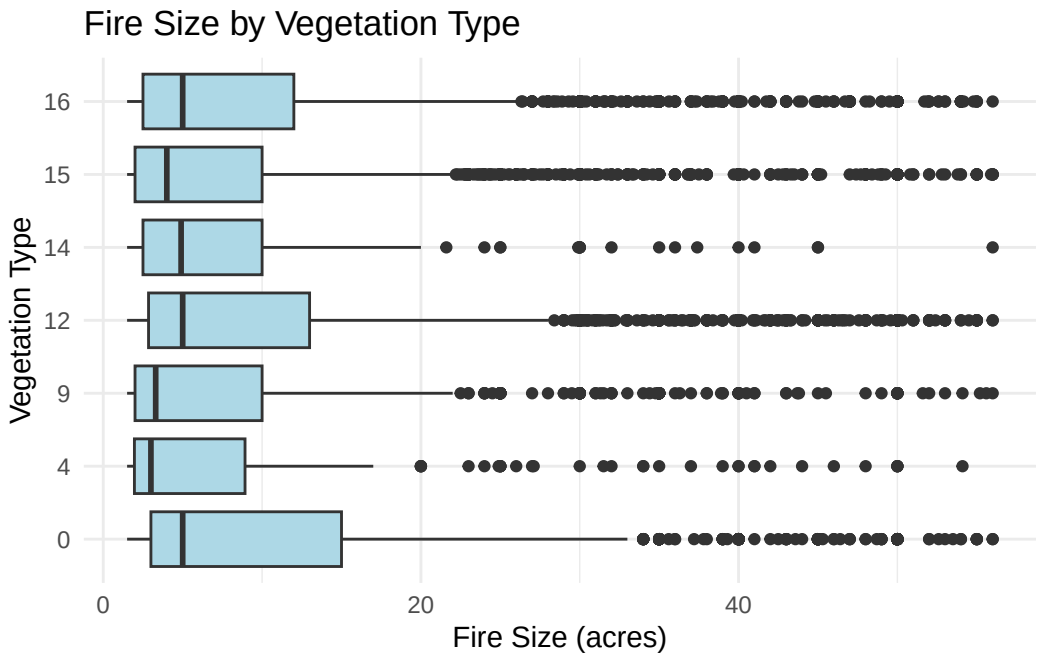



```
        0         4         9        12        14        15        16
16.009558  1.728864  7.484714 26.445991  1.004990 21.652962 25.672921
```

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.



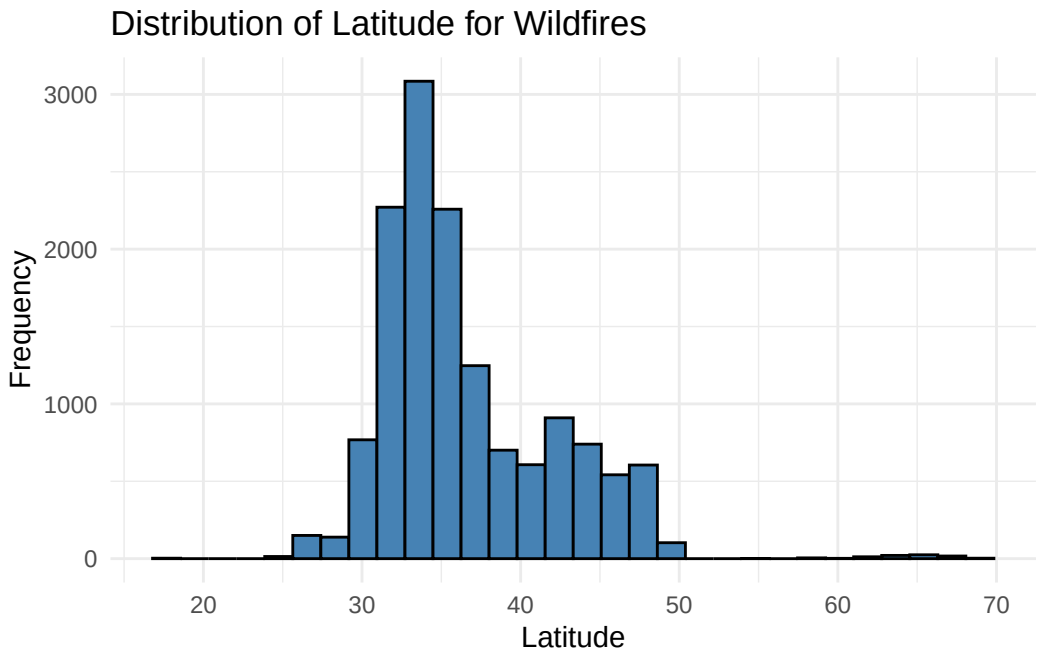

```
# A tibble: 7 x 4
  Vegetation count mean_fire_size median_fire_size
  <fct>      <int>          <dbl>            <dbl>
1 0           2278          11.0                5
2 12          3763          10.2                5
3 16          3653           9.67               5
4 14           143           8.98               4.9
5 15          3081           8.91               4
6 4            246           8.39               3
7 9           1065           8.33               3.3
```

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.

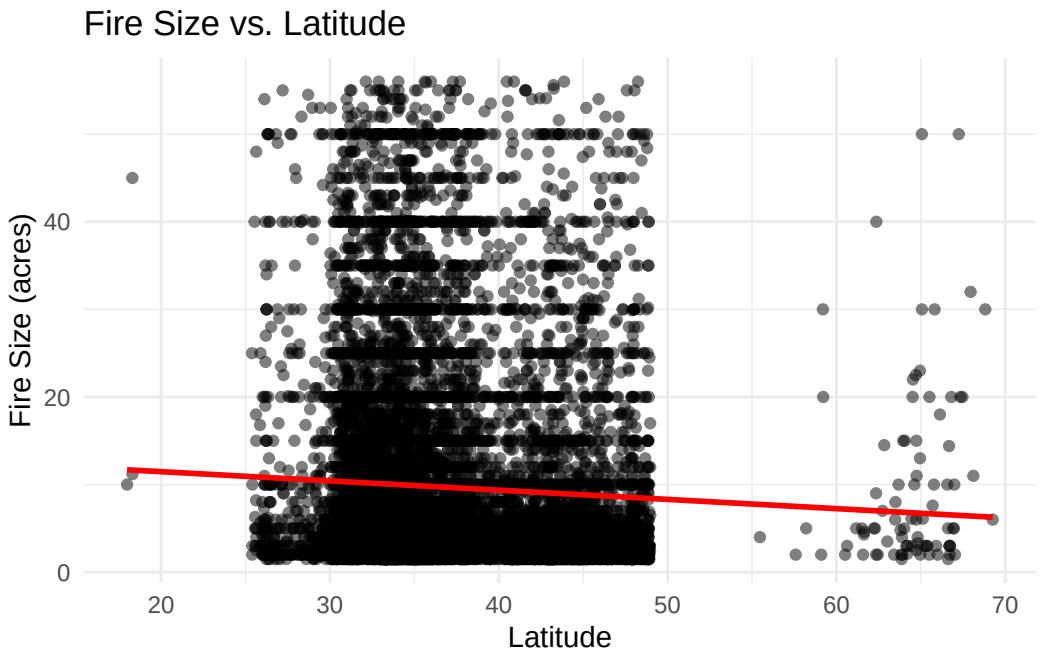
Overall, there is noticeable variation in fire size depending on vegetation type. Some vegetation types appear to be more prone to larger fires, which could be due to factors like fuel availability or vegetation density.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.98   32.85   34.84   36.62   39.93   69.26
```




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.



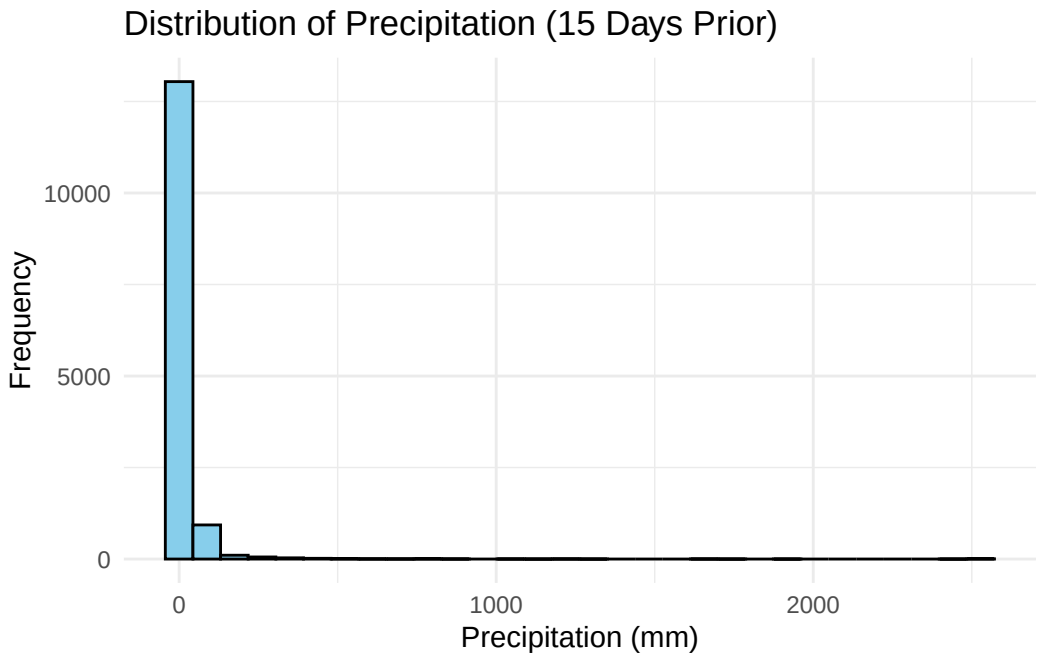

```
[1] -0.05139251
```

Not really any relationship between latitude and fire_size

```
[1] 14.62986
```

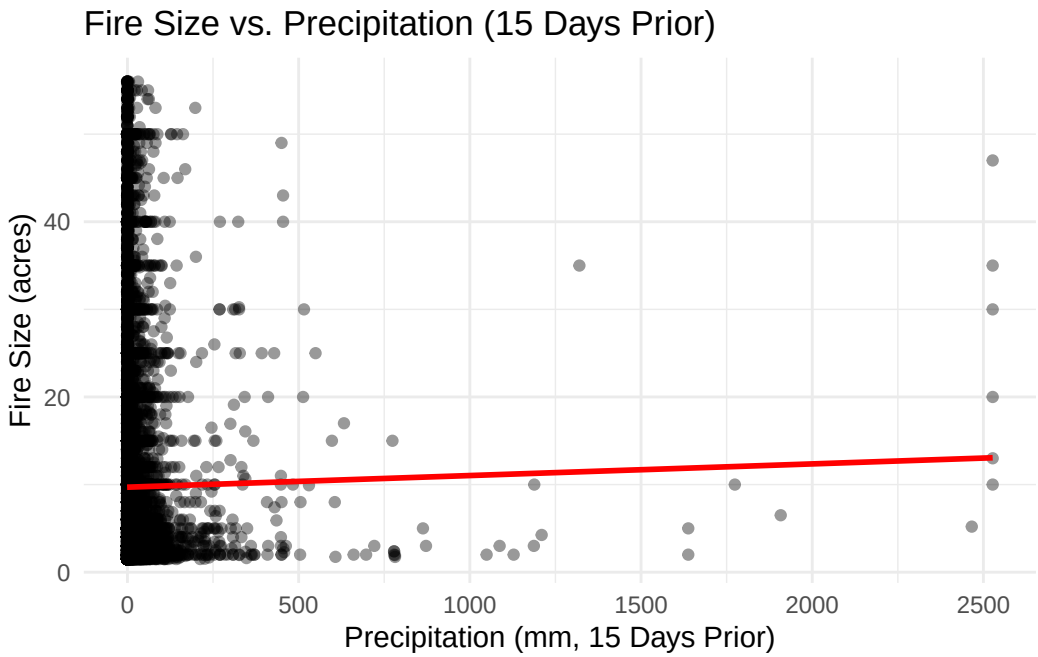
```
[1] 0
```

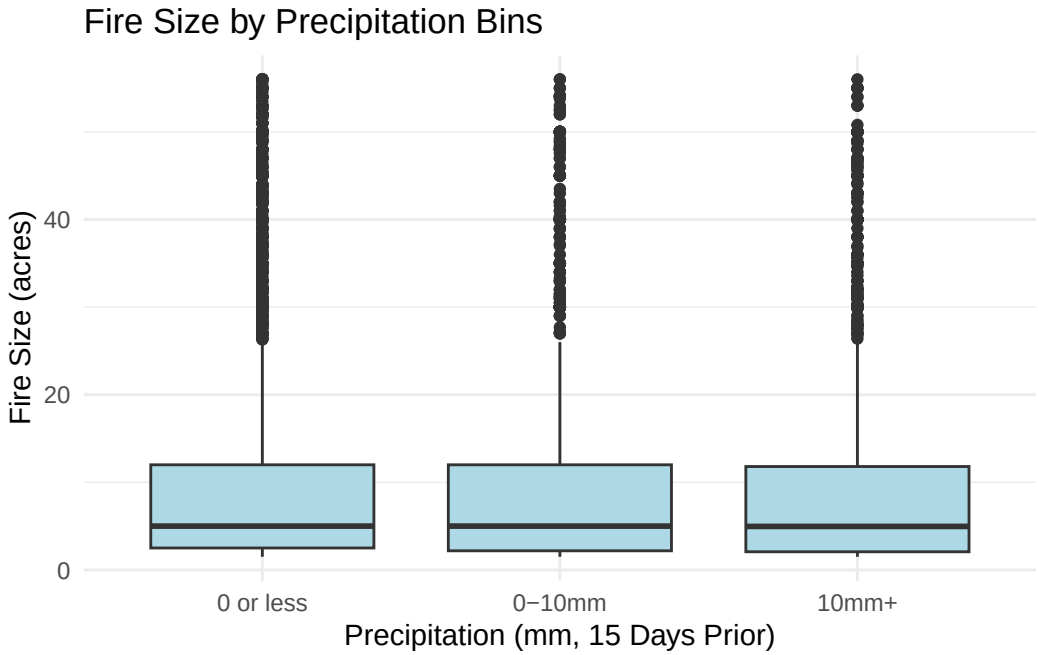
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.0000 -1.0000  0.0000 14.6299  4.3000 2527.0000
```




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.



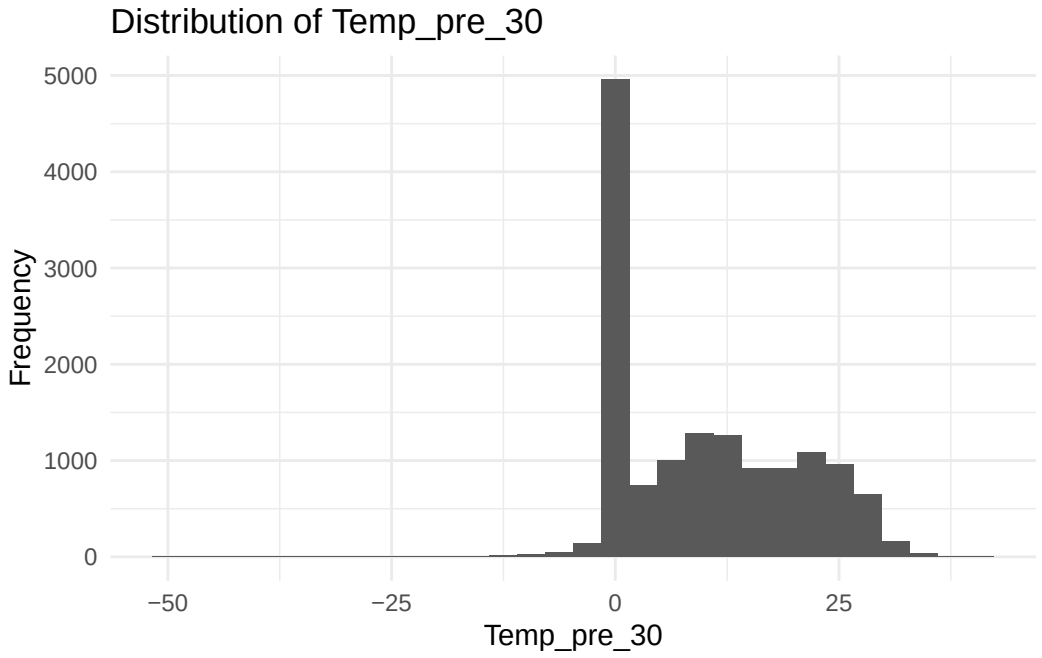




The scatterplot of `Prec_pre_15` and `fire_size` shows no strong relationship between prior precipitation and fire size. Most fires occurred with little to no precipitation, and there is substantial variability in fire size regardless of precipitation level. A few extreme precipitation values do not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

---

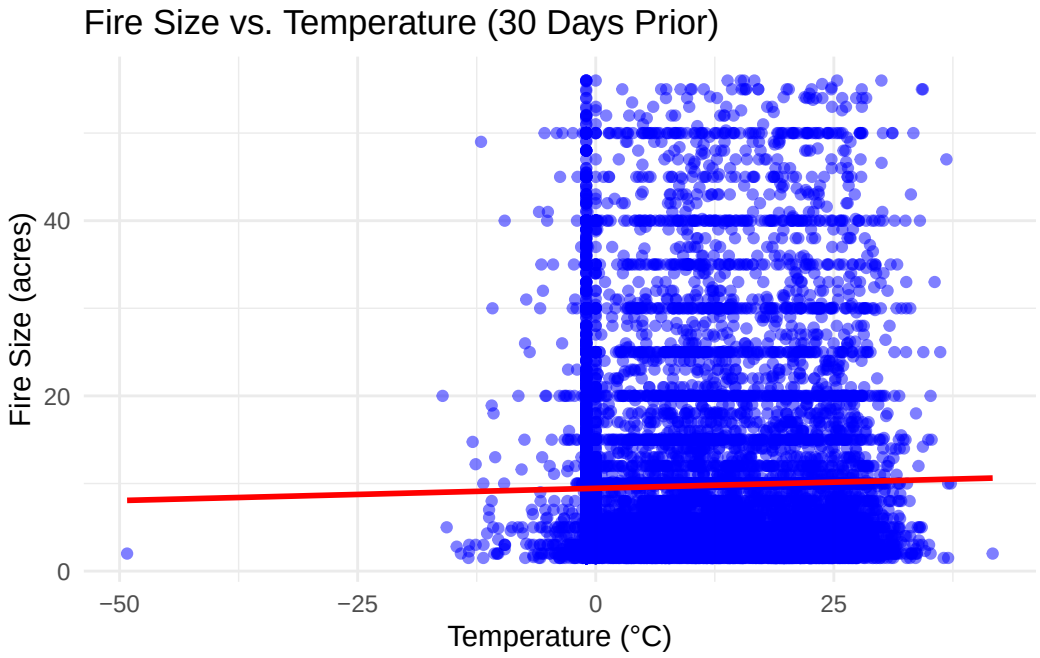
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-49.211  -1.000   8.310   9.562  18.365  41.678
```




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 could 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.

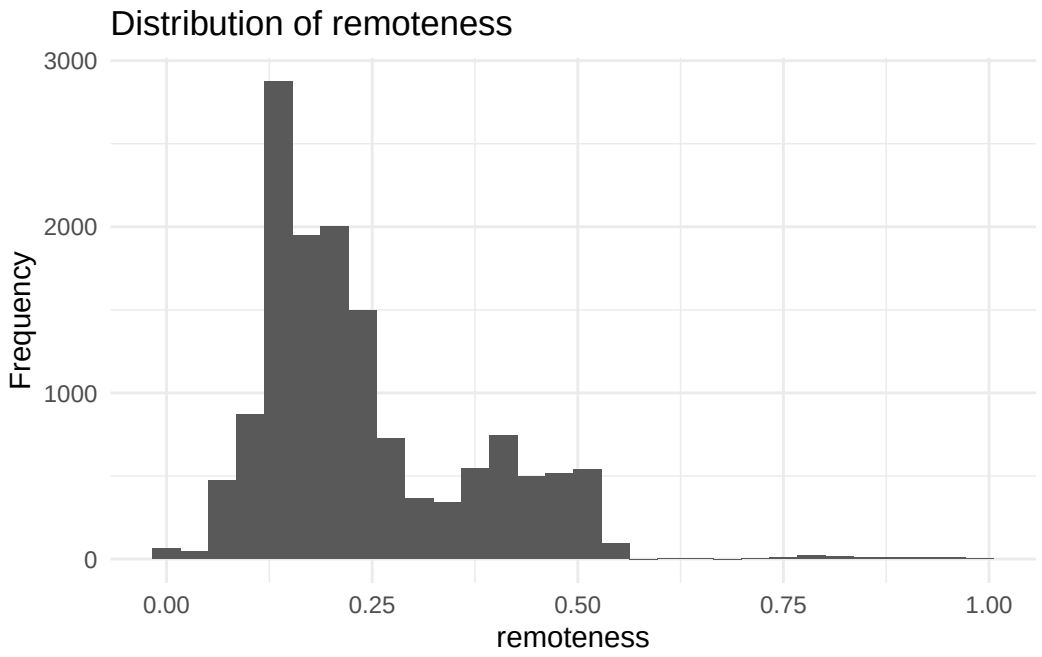
Looking at the histogram, the data appears to be slgihtly right-skewed, with most temperatures falling between 0°C and 25°C, while some extend into negative values as low as -50°C. The presence of extreme negative temperatures raises concerns about outliers or data recording errors, as fires typically 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.




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. 
 0.0000  0.1450  0.2002  0.2403  0.3018  0.9889 
```

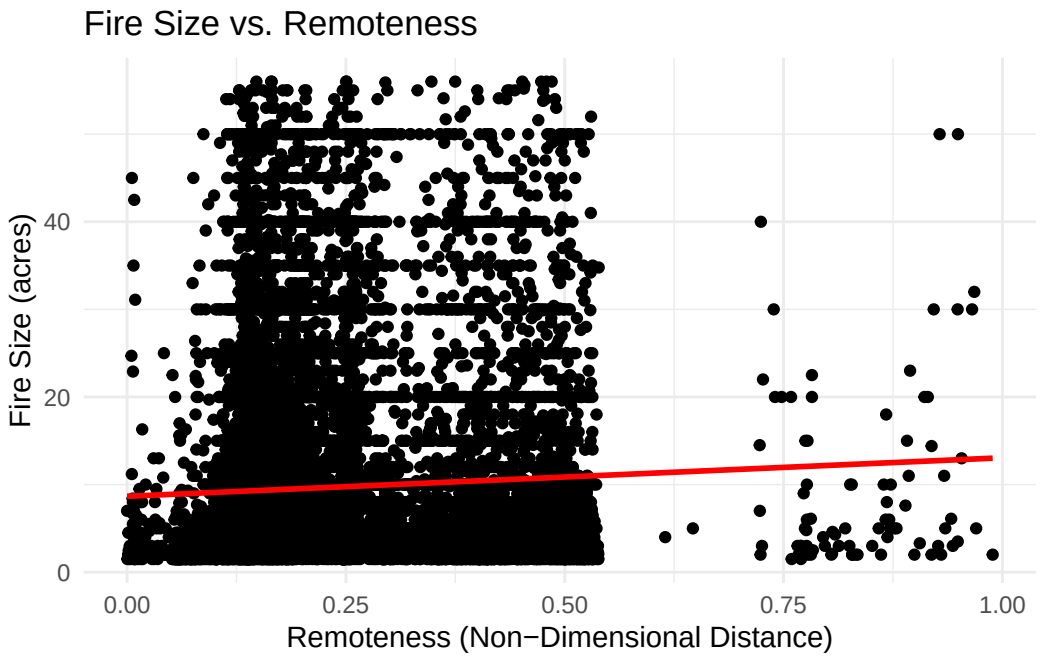



The variable remoteness, which represents the non-dimensional distance to the closest city, has a distribution ranging from 0.0000 to 0.98899. The minimum value of 0 suggests that some fires occurred basically within cities, while the maximum value of 1 indicating the farthest fire away. However, the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

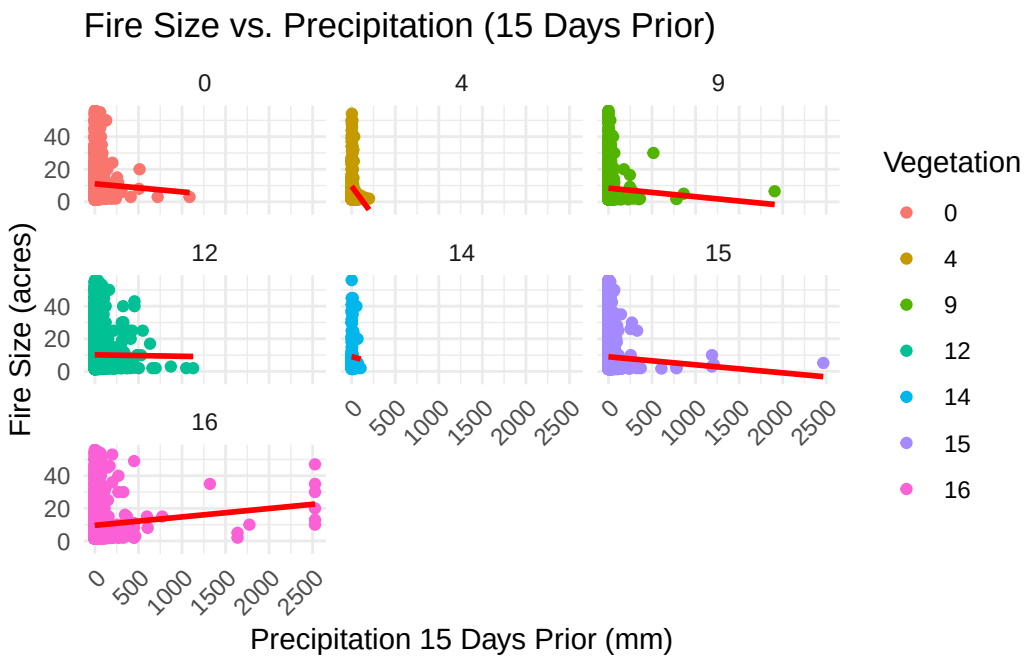
This distribution suggests that most fires tend to occur closer to urban areas rather than in extremely remote locations, but a minority of cases involve fires in highly remote regions.

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.

### Interaction Effects




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.

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 11.526 | 0.369 | 31.195 | 0.000 |
| remoteness | 3.949 | 0.844 | 4.680 | 0.000 |
| Prec_pre_15 | 0.001 | 0.001 | 1.163 | 0.245 |
| Temp_pre_15 | -0.035 | 0.011 | -3.221 | 0.001 |
| Vegetation4 | -2.409 | 0.760 | -3.170 | 0.002 |
| Vegetation9 | -2.616 | 0.422 | -6.205 | 0.000 |
| Vegetation12 | -0.169 | 0.313 | -0.539 | 0.590 |
| Vegetation14 | -2.378 | 0.976 | -2.437 | 0.015 |
| Vegetation15 | -2.087 | 0.313 | -6.671 | 0.000 |
| Vegetation16 | -0.912 | 0.311 | -2.931 | 0.003 |
| stat_cause_descrCampfire | -2.457 | 0.592 | -4.149 | 0.000 |
| stat_cause_descrChildren | -4.311 | 0.672 | -6.415 | 0.000 |
| stat_cause_descrDebris Burning | -2.274 | 0.272 | -8.354 | 0.000 |
| stat_cause_descrEquipment Use | -0.847 | 0.471 | -1.801 | 0.072 |
| stat_cause_descrFireworks | 0.986 | 1.167 | 0.846 | 0.398 |
| stat_cause_descrLightning | -1.168 | 0.356 | -3.284 | 0.001 |

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| stat_cause_descrMiscellaneous | -2.092 | 0.328 | -6.386 | 0.000 |
| stat_cause_descrMissing/Undefined | -1.512 | 0.381 | -3.969 | 0.000 |
| stat_cause_descrPowerline | -1.821 | 0.994 | -1.833 | 0.067 |
| stat_cause_descrRailroad | -3.754 | 0.982 | -3.824 | 0.000 |
| stat_cause_descrSmoking | -3.141 | 0.688 | -4.565 | 0.000 |
| stat_cause_descrStructure | -4.356 | 2.141 | -2.034 | 0.042 |
| Wind_cont | 0.759 | 0.086 | 8.794 | 0.000 |

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 11.663 | 0.377 | 30.949 | 0.000 |
| remoteness | 3.632 | 0.848 | 4.285 | 0.000 |
| Temp_pre_15 | -0.033 | 0.011 | -3.039 | 0.002 |
| Prec_pre_15 | -0.005 | 0.006 | -0.908 | 0.364 |
| Vegetation4 | -1.515 | 0.837 | -1.810 | 0.070 |
| Vegetation9 | -2.599 | 0.432 | -6.014 | 0.000 |
| Vegetation12 | -0.230 | 0.324 | -0.710 | 0.478 |
| Vegetation14 | -2.441 | 1.062 | -2.299 | 0.022 |
| Vegetation15 | -2.098 | 0.321 | -6.527 | 0.000 |
| Vegetation16 | -1.062 | 0.319 | -3.326 | 0.001 |
| stat_cause_descrCampfire | -2.428 | 0.592 | -4.100 | 0.000 |
| stat_cause_descrChildren | -4.310 | 0.672 | -6.415 | 0.000 |
| stat_cause_descrDebris Burning | -2.260 | 0.272 | -8.307 | 0.000 |
| stat_cause_descrEquipment Use | -0.858 | 0.470 | -1.824 | 0.068 |
| stat_cause_descrFireworks | 0.950 | 1.166 | 0.815 | 0.415 |
| stat_cause_descrLightning | -1.174 | 0.356 | -3.302 | 0.001 |
| stat_cause_descrMiscellaneous | -2.054 | 0.328 | -6.267 | 0.000 |
| stat_cause_descrMissing/Undefined | -1.512 | 0.381 | -3.969 | 0.000 |
| stat_cause_descrPowerline | -1.820 | 0.993 | -1.832 | 0.067 |
| stat_cause_descrRailroad | -3.779 | 0.981 | -3.850 | 0.000 |
| stat_cause_descrSmoking | -3.161 | 0.688 | -4.595 | 0.000 |
| stat_cause_descrStructure | -4.385 | 2.141 | -2.049 | 0.041 |
| Wind_cont | 0.765 | 0.086 | 8.860 | 0.000 |
| Prec_pre_15:Vegetation4 | -0.059 | 0.024 | -2.463 | 0.014 |
| Prec_pre_15:Vegetation9 | 0.001 | 0.007 | 0.145 | 0.884 |
| Prec_pre_15:Vegetation12 | 0.004 | 0.006 | 0.583 | 0.560 |
| Prec_pre_15:Vegetation14 | 0.008 | 0.057 | 0.135 | 0.892 |
| Prec_pre_15:Vegetation15 | 0.001 | 0.006 | 0.213 | 0.832 |
| Prec_pre_15:Vegetation16 | 0.010 | 0.006 | 1.681 | 0.093 |

```
[1] 0.02168202
```

```
[1] 0.02259321
```