Analyzing Factors Associated with Wildfire Size in the United States

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

2025-03-20

Introduction and Data

Wildfires are among the most damaging natural disasters in the U.S., with recent years bringing record-breaking destruction in terms of acreage burned, economic loss, and human displacement in places like California and Western North Carolina. Climate change has exacerbated the issue by intensifying drought, reducing vegetation moisture, and increasing fuel availability—making wildfires more frequent and less predictable. These challenges highlight the urgent need for better tools to anticipate wildfire behavior and allocate resources effectively.

This project aims to predict the total area burned by analyzing historical wildfire data alongside environmental, weather, and geographic factors. By identifying the variables most strongly associated with fire spread—such as temperature, wind, humidity, precipitation, vegetation type, and fire remoteness—we hope to develop models that can guide strategic decision-making. Improved prediction could help firefighting agencies and policymakers prioritize high-risk events, respond faster, and reduce the overall damage caused by wildfires.

Research Question:

What environmental and fire-specific factors influence wildfire size, and can we build models to estimate burned area based on those factors?

Sources:

Our compiled dataset integrates information from four sources: NOAA National Centers for Environmental Information (2021), World Cities Database, Forest Service Research Data Archive, and Frontiers of Earth Science.

Key Variables:

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.
- `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.
- `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_15` (°C): The recorded temperature at the fire location up to 15 days before the fire was contained, measured in degrees Celsius.
- `Wind_pre_15` (m/s): The wind speed at the fire location up to 15 days before containment, measured in meters per second.
- `Hum_pre_15` (%): The humidity level at the fire location up to 15 days before containment, expressed as a percentage.
- `Prec_pre_15` (mm): The total amount of precipitation recorded at the fire location up to 15 days before containment, 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

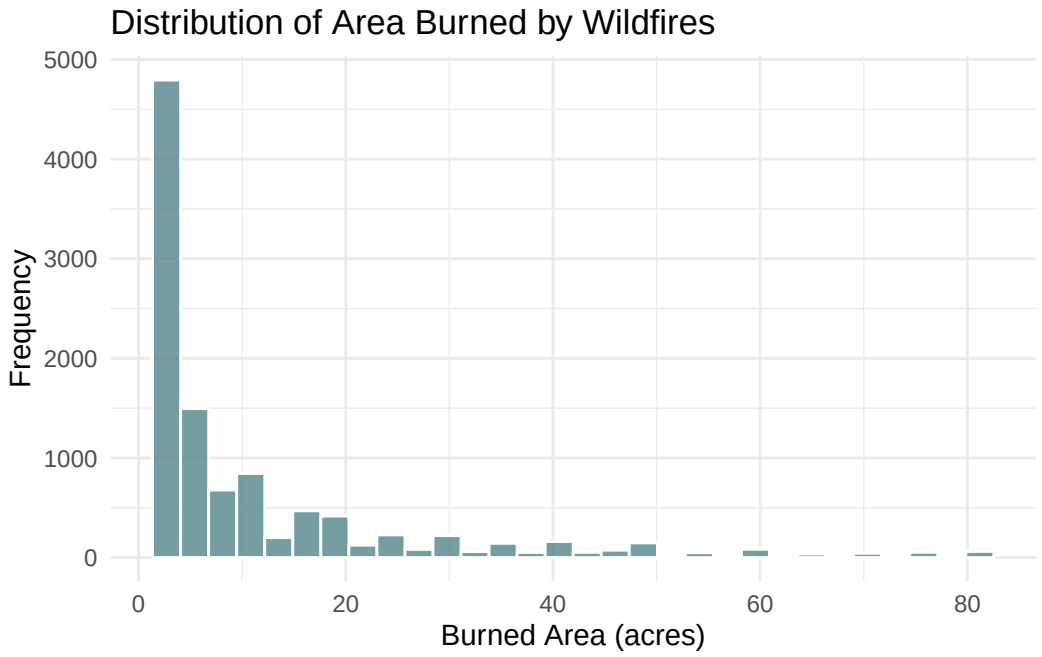
Data Cleaning

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.

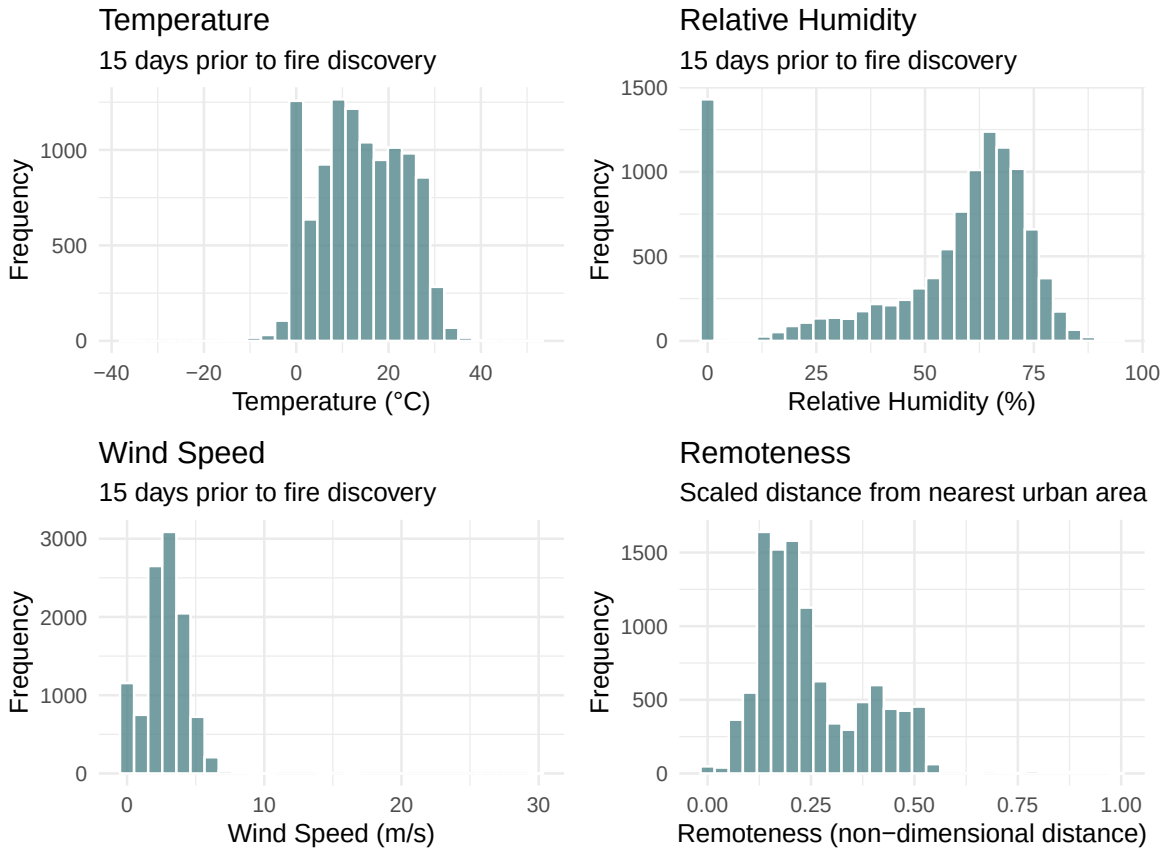
Univariate EDA

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.

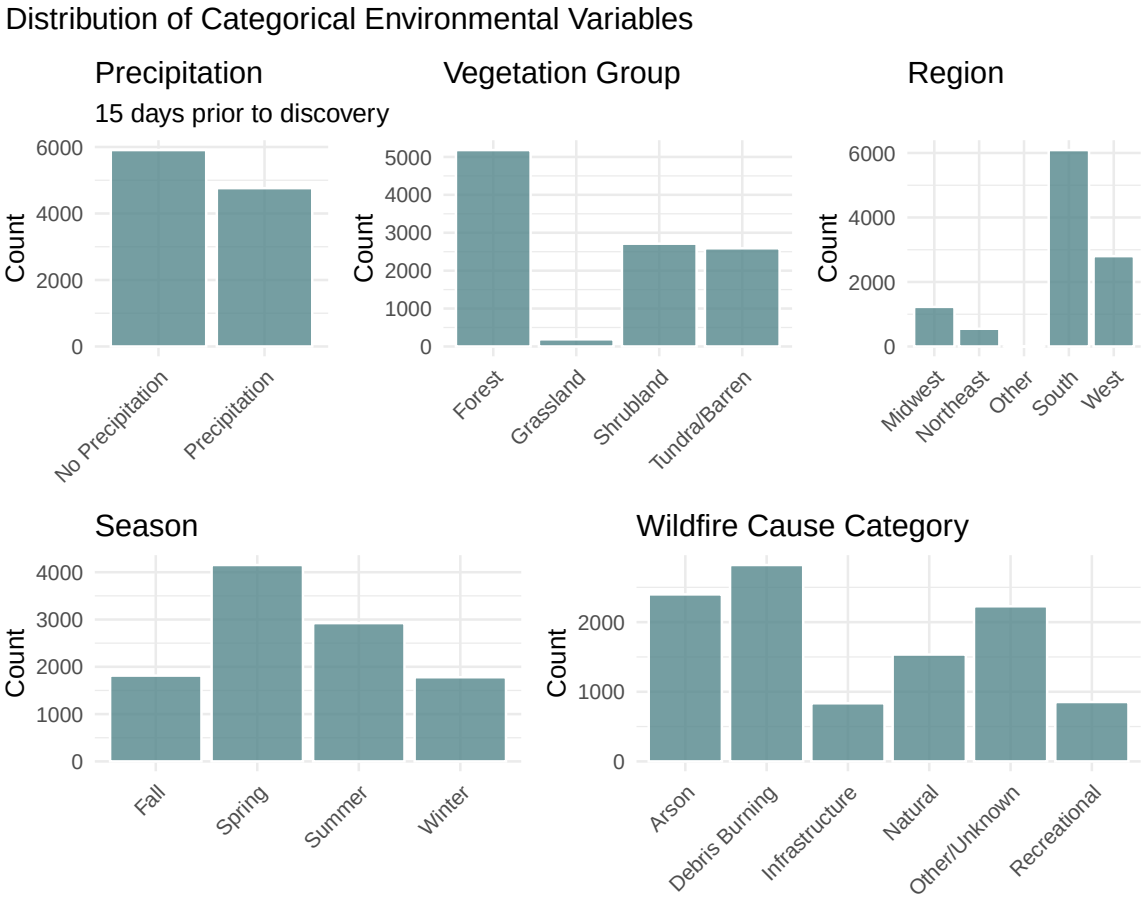


Count	Min	Q1	Median	Mean	Q3	Max	SD
10671	1.5	2.5	5	11.84	14.7	80	15.62

Distribution of Quantitative Environmental Variables



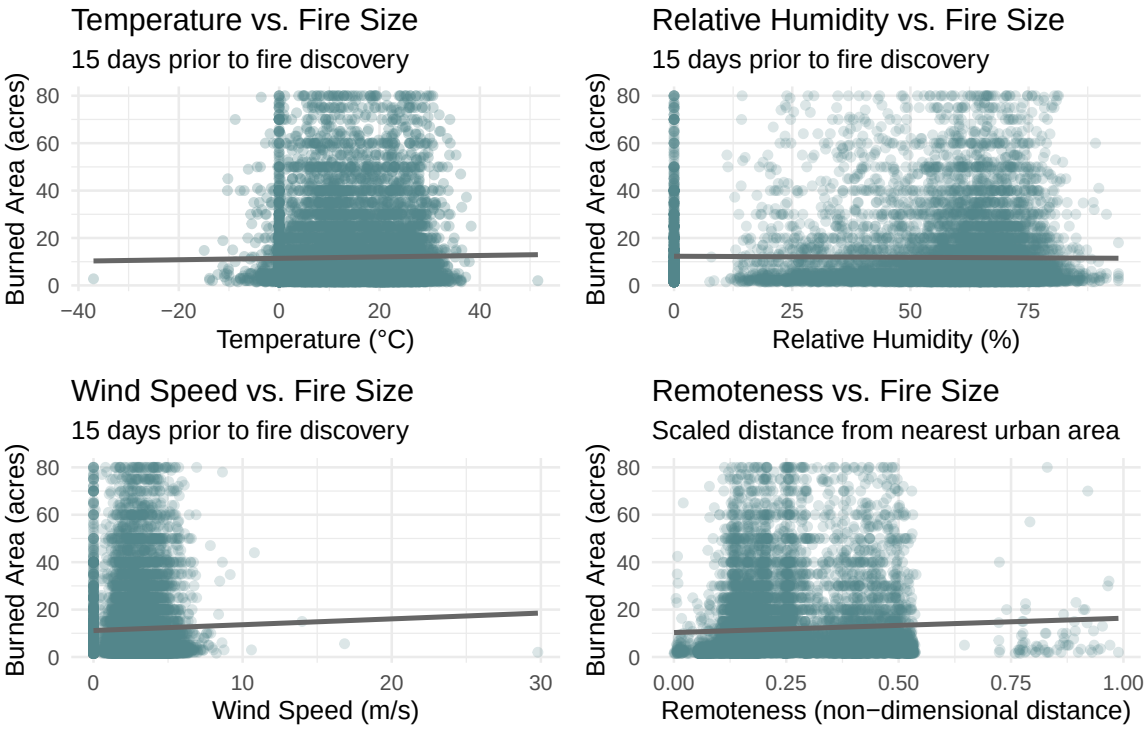
Variable	Min	Q1	Median	Mean	Q3	Max	SD
Temperature	-37	6.70	13.41	13.81	21.57	51.57	9.30
Humidity	0	43.08	61.20	52.01	68.84	94.00	24.39
Wind	0	1.91	2.81	2.77	3.74	29.80	1.51
Remoteness	0	0.15	0.21	0.25	0.35	0.99	0.13

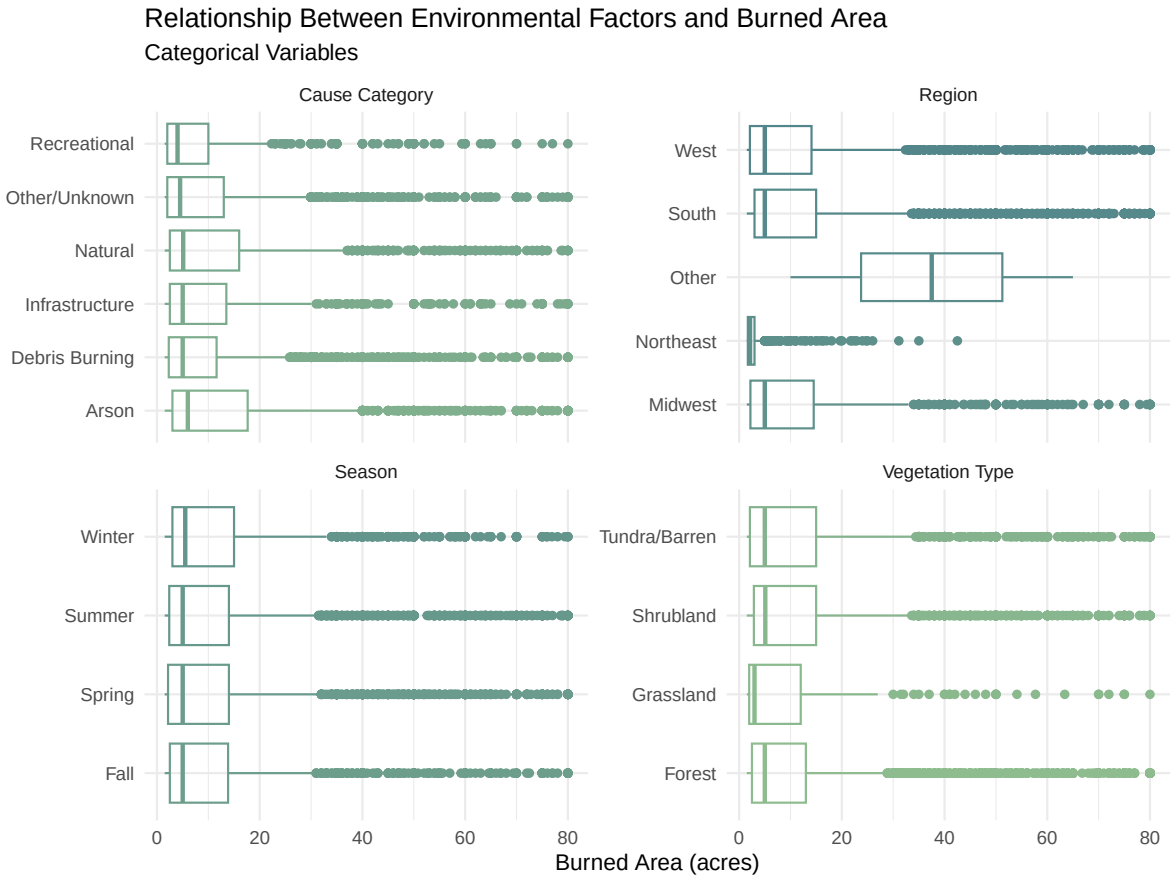


Bivariate EDA

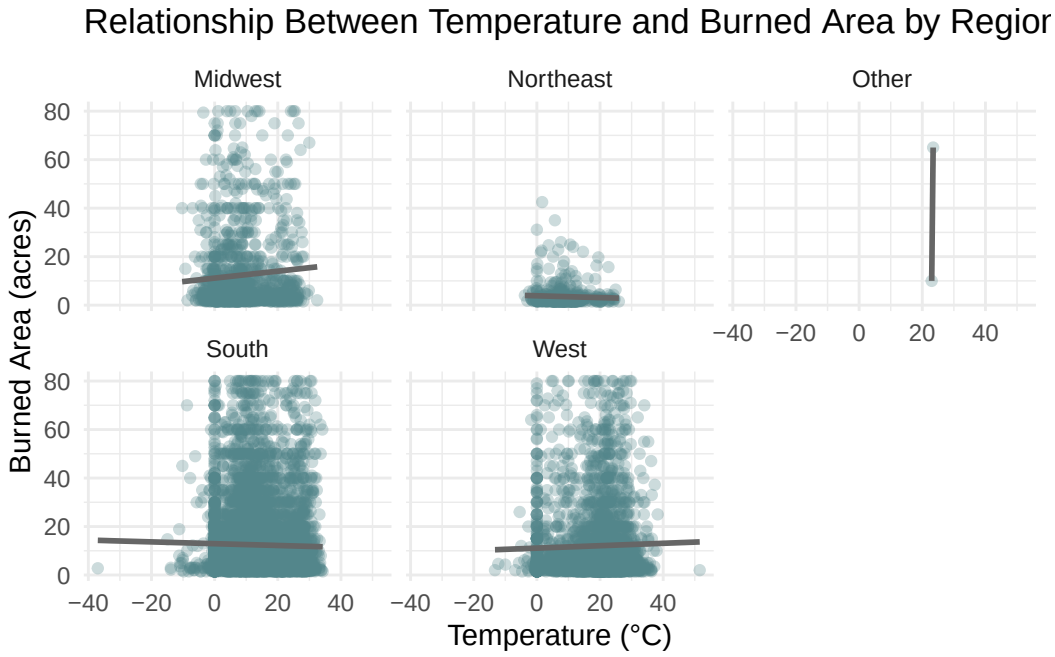
Relationship Between Environmental Factors and Burned Area

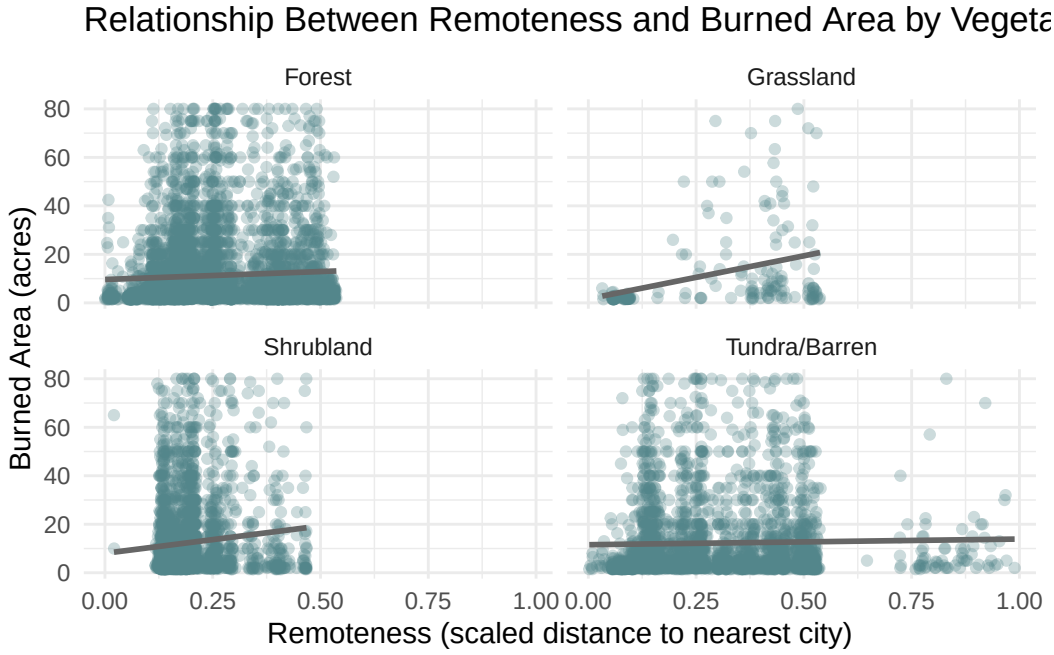
Quantitative Variables





Interaction Effects





Methodology

```
# A tibble: 1 x 12
```

```
  r.squared adj.r.squared sigma statistic  p.value    df  logLik    AIC    BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>   <dbl>  <dbl>  <dbl>
1    0.0310        0.0295  15.4      20.1 3.61e-61    17 -44304. 88646. 88784.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
              regionNortheast                    regionOther
                     1.926245                       1.006738
                  regionSouth                     regionWest
                     3.725138                       5.280966
 cause_categoryDebris Burning  cause_categoryInfrastructure
                     1.632311                       1.293260
        cause_categoryNatural   cause_categoryOther/Unknown
                     1.815560                       1.641285
   cause_categoryRecreational                    remoteness
                     1.293495                       5.184689
    Vegetation_groupGrassland     Vegetation_groupShrubland
                     1.077473                       1.339964
Vegetation_groupTundra/Barren                  Wind_pre_15
                     1.226920                       1.447867
                 Prec_pre_15                    Temp_pre_15
                     1.236883                       1.379270
                  Hum_pre_15
                     1.687038
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
1    0.0552        0.0537  1.05      36.6 9.21e-118    17 -15668. 31373. 31511.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
1    0.0561        0.0543  1.05      31.7 1.54e-117    20 -15663. 31369. 31529.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

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.

```
Prec_pre_15 Temp_pre_15 Wind_pre_15  Hum_pre_15  remoteness
   1.200925    1.119898    1.262910    1.508329    1.195878
```

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)	1.959	0.042	46.502	0.000
remoteness	0.373	0.091	4.091	0.000
log(Prec_pre_15 + 0.001)	-0.003	0.003	-1.073	0.283
Temp_pre_15	-0.005	0.001	-3.685	0.000

term	estimate	std.error	statistic	p.value
Vegetation4	-0.285	0.082	-3.493	0.000
Vegetation9	-0.271	0.044	-6.117	0.000
Vegetation12	0.074	0.035	2.120	0.034
Vegetation14	-0.260	0.100	-2.590	0.010
Vegetation15	-0.204	0.034	-6.048	0.000
Vegetation16	0.004	0.034	0.120	0.904
stat_cause_descrCampfire	-0.290	0.063	-4.611	0.000
stat_cause_descrChildren	-0.576	0.072	-7.978	0.000
stat_cause_descrDebris Burning	-0.227	0.030	-7.658	0.000
stat_cause_descrEquipment Use	-0.026	0.050	-0.532	0.595
stat_cause_descrFireworks	-0.029	0.127	-0.232	0.816
stat_cause_descrLightning	-0.121	0.039	-3.127	0.002
stat_cause_descrMiscellaneous	-0.240	0.035	-6.893	0.000
stat_cause_descrMissing/Undefined	-0.155	0.048	-3.235	0.001
stat_cause_descrPowerline	-0.231	0.101	-2.290	0.022
stat_cause_descrRailroad	-0.234	0.103	-2.263	0.024
stat_cause_descrSmoking	-0.383	0.076	-5.029	0.000
stat_cause_descrStructure	-0.628	0.233	-2.701	0.007
Wind_cont	0.090	0.009	10.429	0.000

term	estimate	std.error	statistic	p.value
(Intercept)	2.053	0.049	42.146	0.000
remoteness	0.379	0.091	4.148	0.000
Temp_pre_15	-0.005	0.001	-3.694	0.000
log(Prec_pre_15 + 0.001)	0.021	0.008	2.833	0.005
Vegetation4	-0.507	0.101	-5.012	0.000
Vegetation9	-0.452	0.060	-7.531	0.000
Vegetation12	-0.061	0.051	-1.182	0.237
Vegetation14	-0.345	0.146	-2.372	0.018
Vegetation15	-0.300	0.049	-6.089	0.000
Vegetation16	-0.050	0.052	-0.953	0.341
stat_cause_descrCampfire	-0.295	0.063	-4.699	0.000
stat_cause_descrChildren	-0.569	0.072	-7.885	0.000
stat_cause_descrDebris Burning	-0.231	0.030	-7.784	0.000
stat_cause_descrEquipment Use	-0.029	0.050	-0.584	0.559
stat_cause_descrFireworks	-0.039	0.127	-0.305	0.760
stat_cause_descrLightning	-0.122	0.039	-3.162	0.002
stat_cause_descrMiscellaneous	-0.233	0.035	-6.708	0.000
stat_cause_descrMissing/Undefined	-0.161	0.048	-3.361	0.001

term	estimate	std.error	statistic	p.value
stat_cause_descrPowerline	-0.229	0.101	-2.279	0.023
stat_cause_descrRailroad	-0.233	0.103	-2.255	0.024
stat_cause_descrSmoking	-0.377	0.076	-4.959	0.000
stat_cause_descrStructure	-0.624	0.232	-2.685	0.007
Wind_cont	0.089	0.009	10.340	0.000
log(Prec_pre_15 + 0.001):Vegetation4	-0.083	0.025	-3.285	0.001
log(Prec_pre_15 + 0.001):Vegetation9	-0.056	0.013	-4.357	0.000
log(Prec_pre_15 + 0.001):Vegetation12	-0.035	0.010	-3.625	0.000
log(Prec_pre_15 + 0.001):Vegetation14	-0.022	0.029	-0.766	0.444
log(Prec_pre_15 + 0.001):Vegetation15	-0.025	0.010	-2.577	0.010
log(Prec_pre_15 + 0.001):Vegetation16	-0.016	0.010	-1.599	0.110

```
[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
```

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 or 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`.

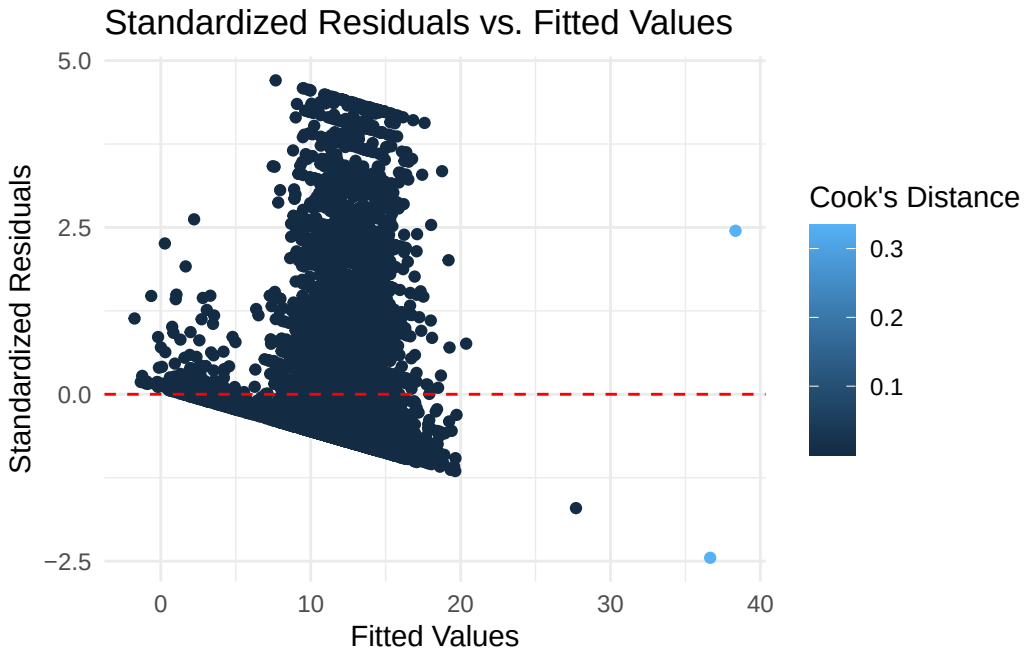
The adjusted R^2 values were low across all models. The main-effects model with continuous precipitation (`fire_main_fit`) had an adjusted R^2 of 0.037, and the interaction model (`fire_int_fit`) had an adjusted R^2 of 0.041. The models using the binary precipitation indicator produced similar results: 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., $\exp(1.06) \approx 2.88$).

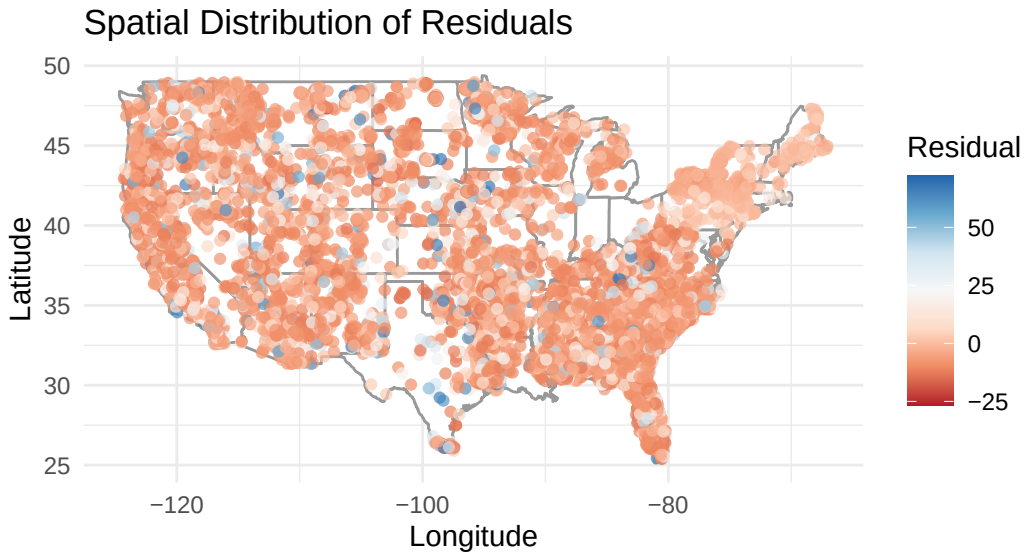
Overall, there were minimal differences in adjusted R^2 and RMSE between the models using continuous precipitation and those using a binary indicator. The addition of interaction terms with `Vegetation` slightly increased model complexity without meaningfully improving predictive performance.

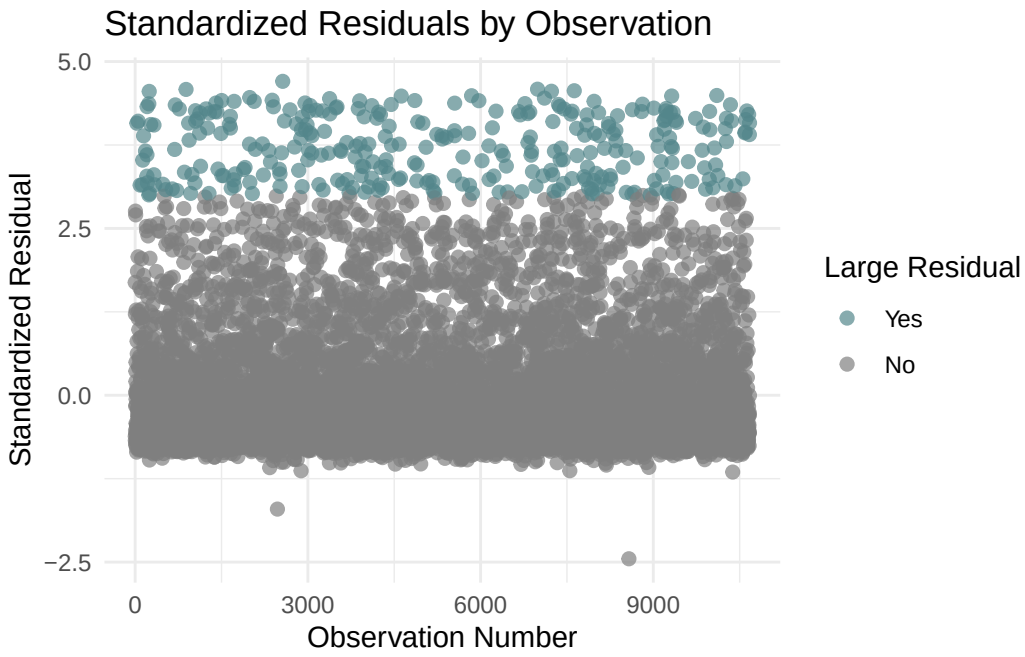
Appendix

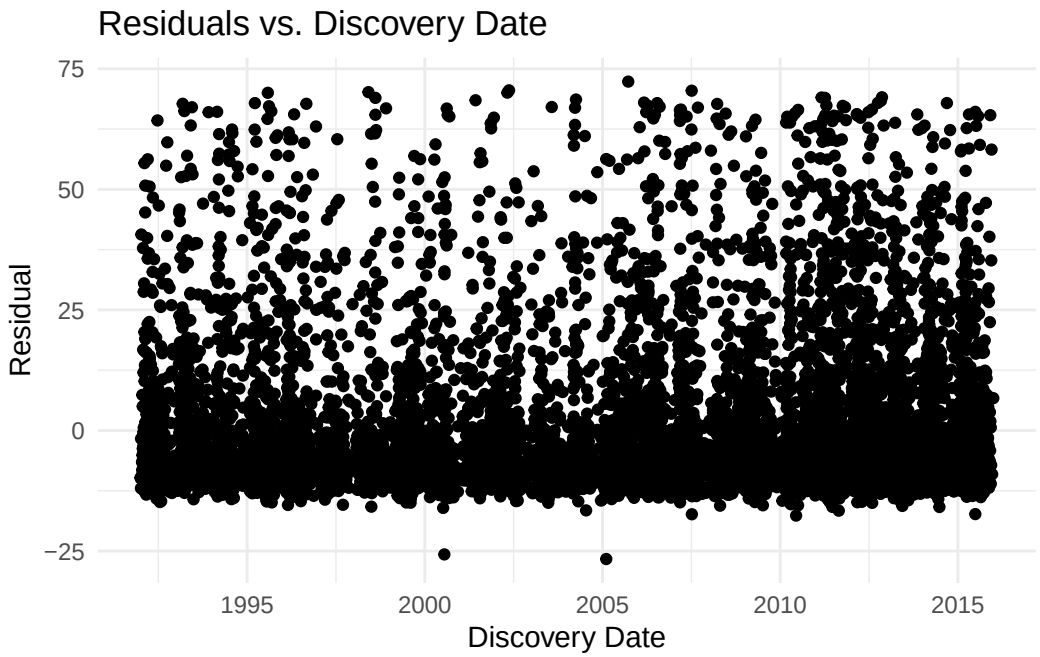
Std. residuals vs. fitted values - constant variance, linearity, cook's distance (influential points)



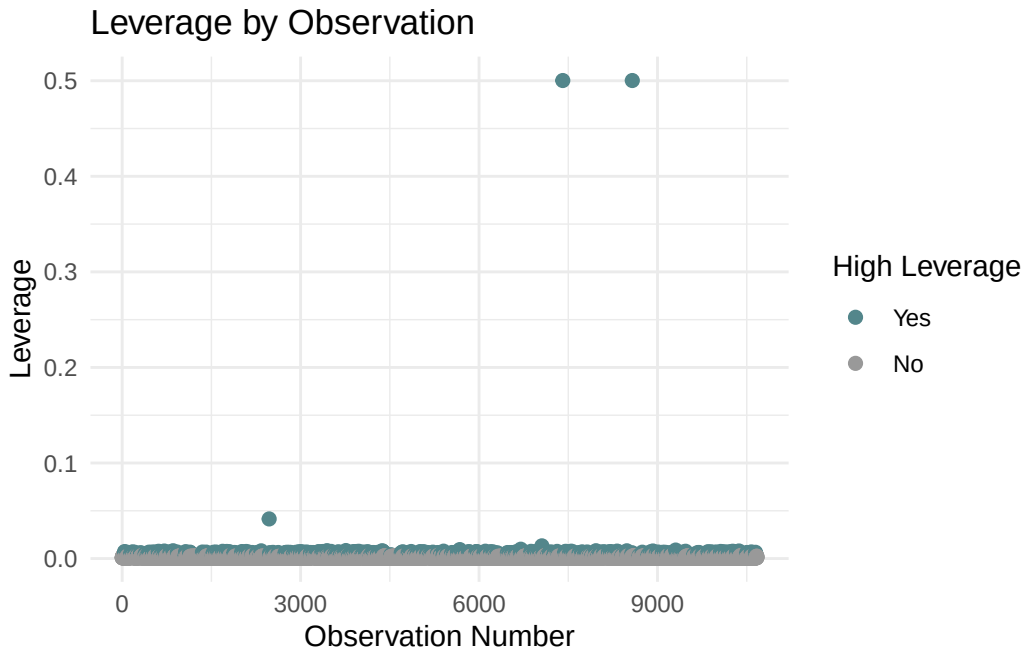
spatial distribution of residuals - independence

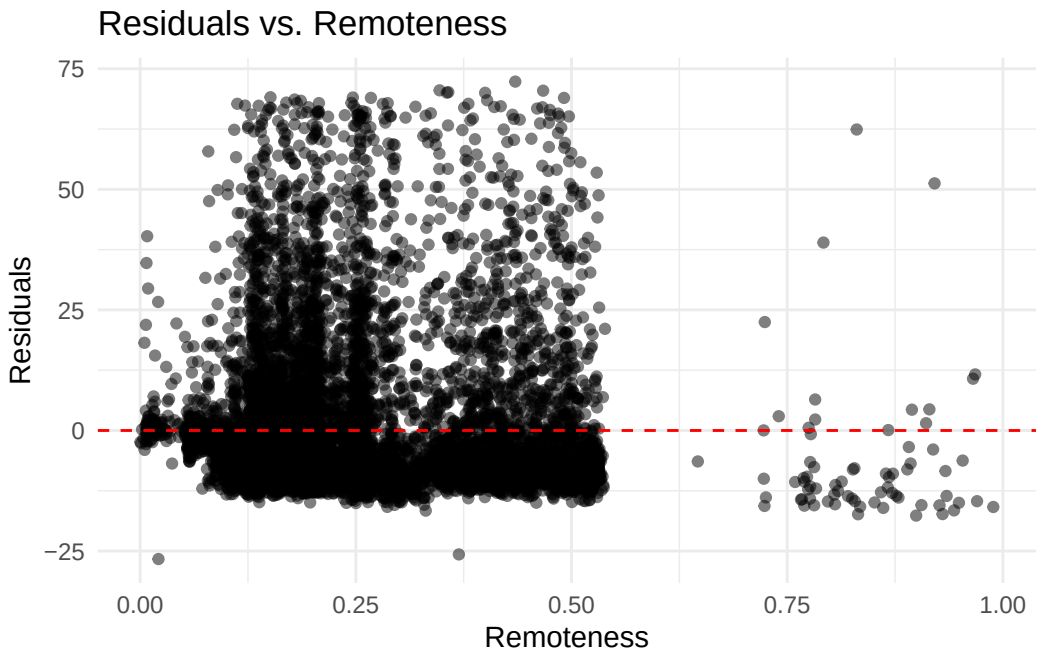

Spatial Distribution of Residuals
Latitude
Longitude
50
45
40
35
30
25
−120
−100
−80
Residual
50
25
0
−25

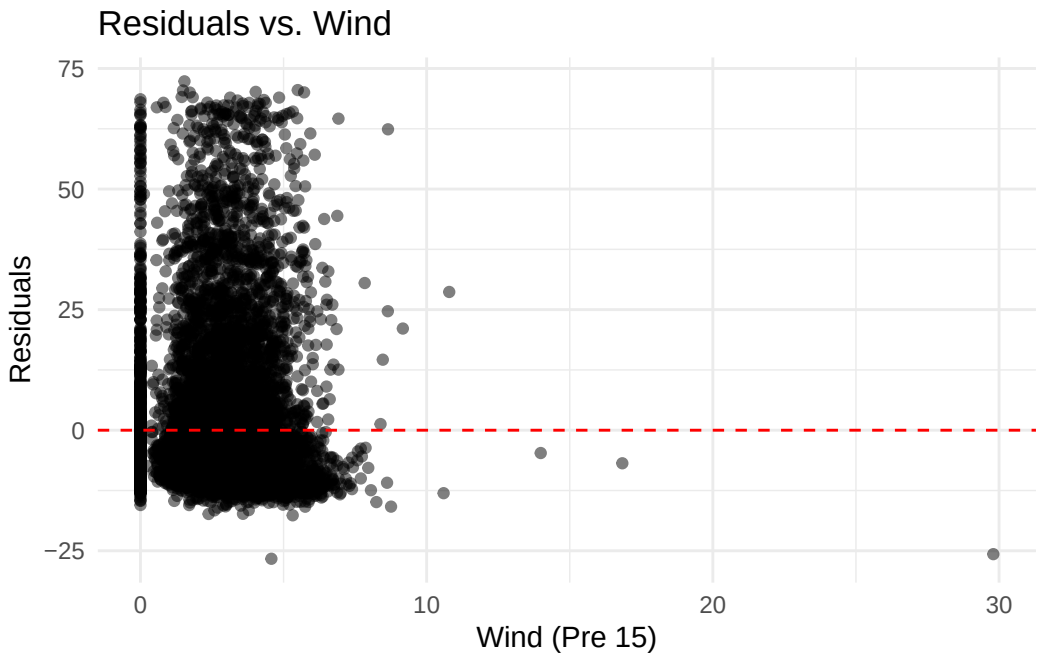

Standardized Residuals by Observation
Standardized Residual
Observation Number
5.0
2.5
0.0
−2.5
0
3000
6000
9000
Large Residual
Yes
No

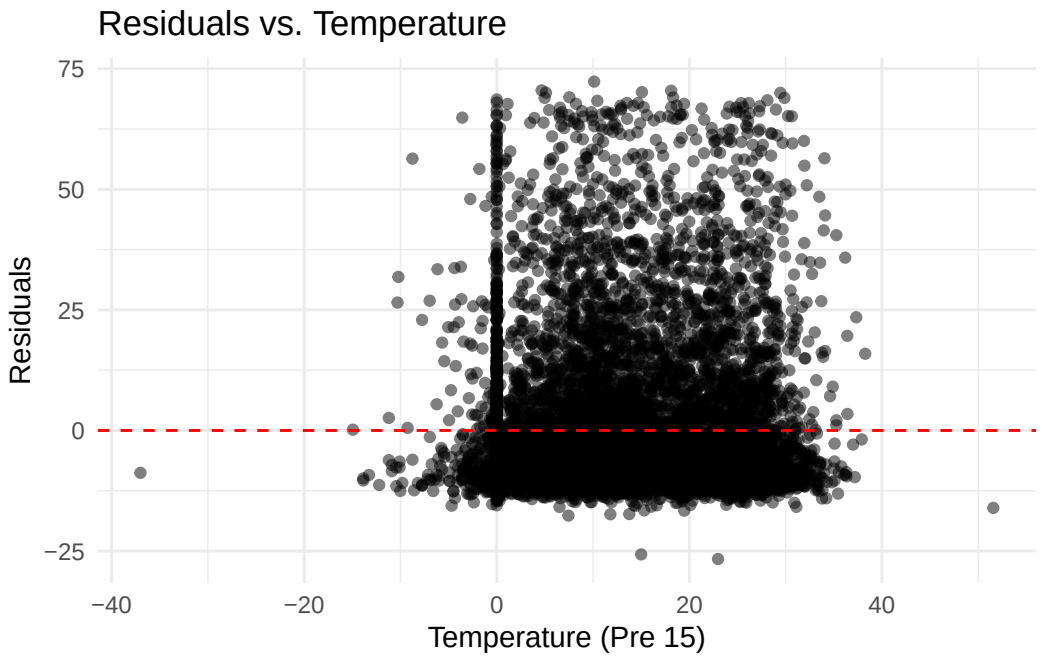


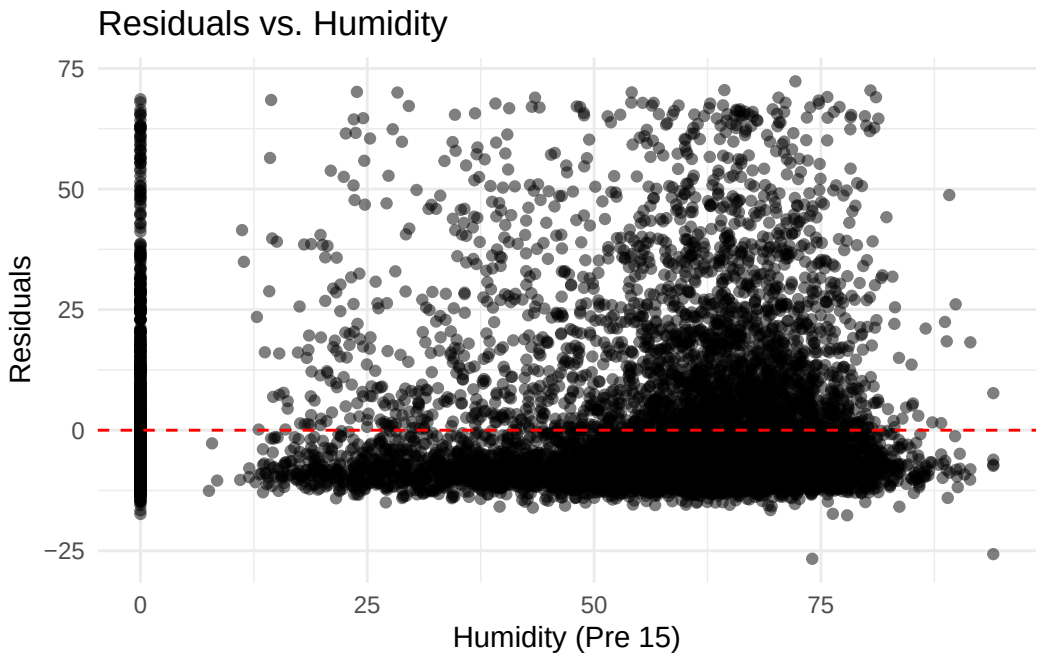
large leverage

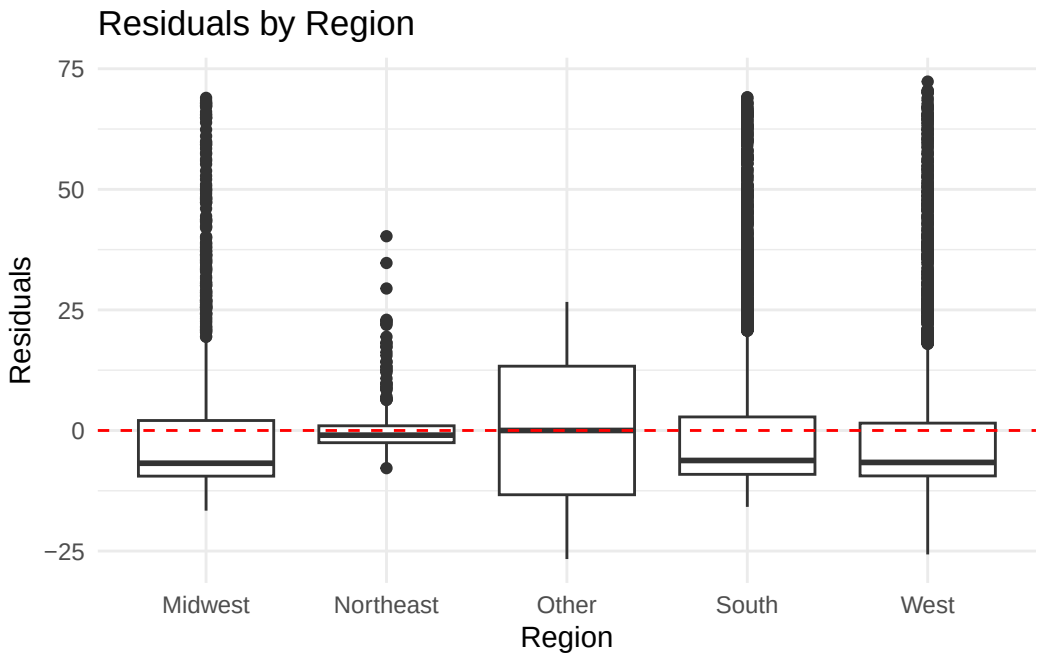


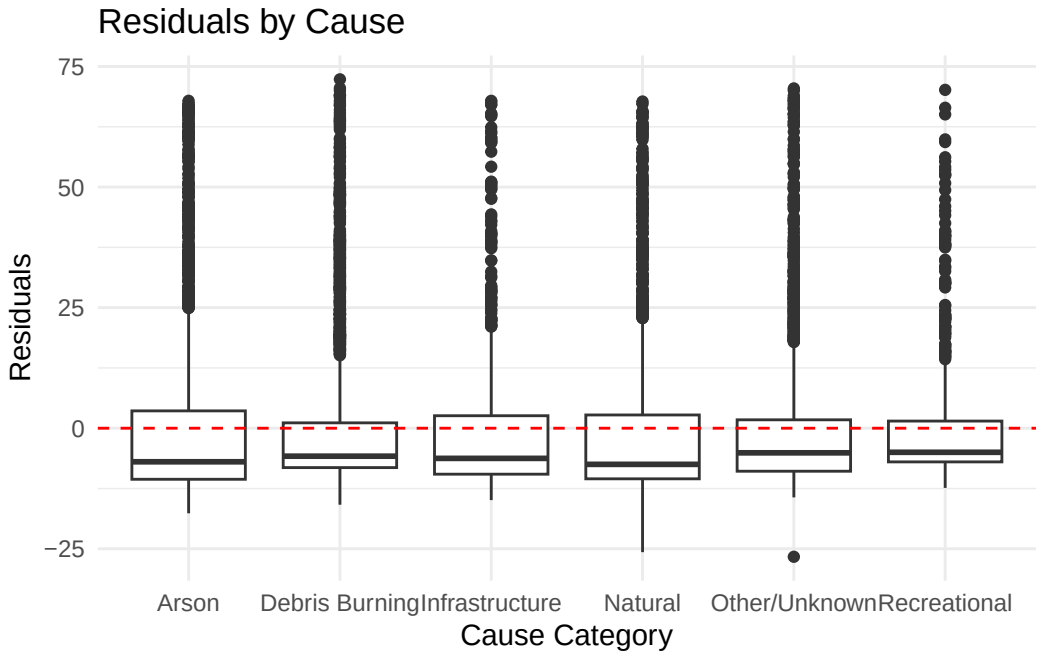

Residuals vs. Remoteness
Residuals
75
50
25
0
−25
0.00
0.25
0.50
0.75
1.00
Remoteness

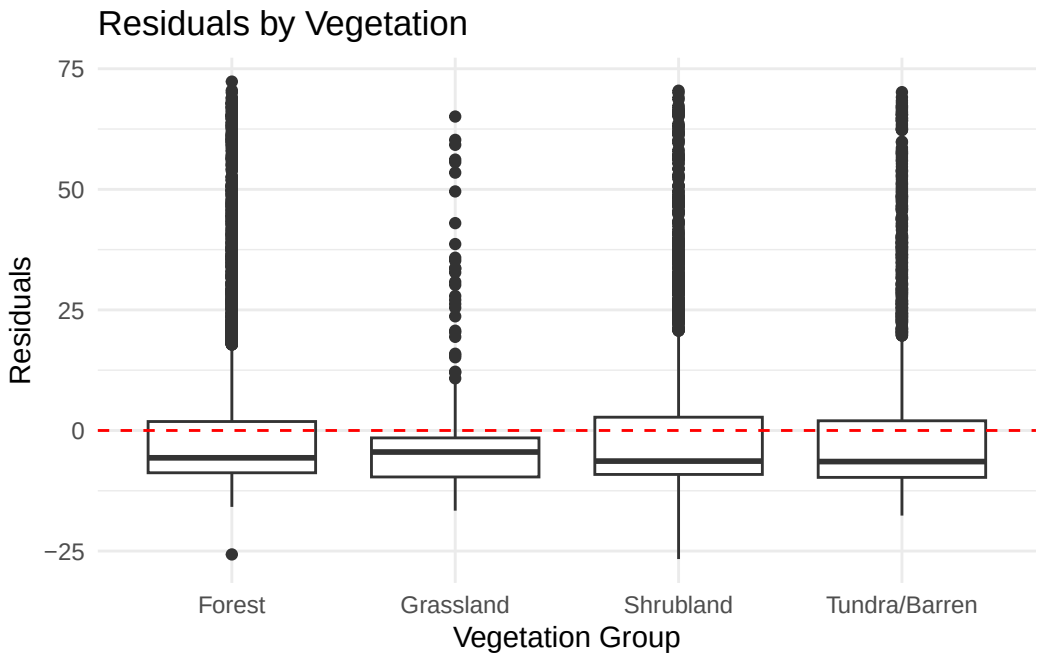

Residuals vs. Wind
Residuals
75
50
25
0
−25
0
10
20
30
Wind (Pre 15)

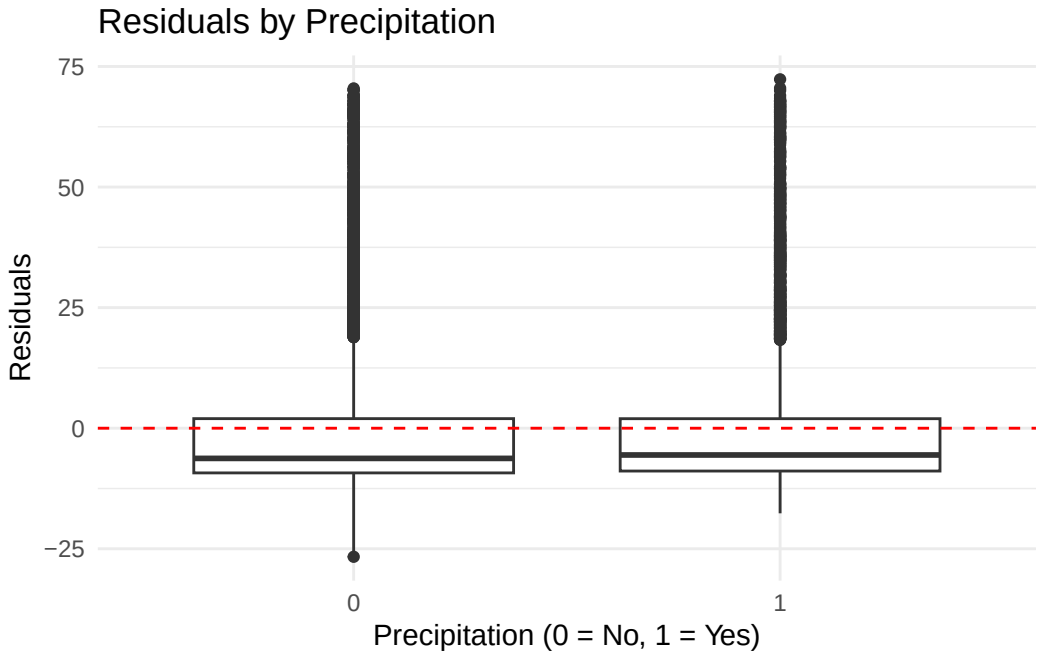

Residuals vs. Temperature
Residuals
75
50
25
0
−25
−40
−20
0
20
40
Temperature (Pre 15)


Residuals vs. Humidity
Residuals
75
50
25
0
−25
0
25
50
75
Humidity (Pre 15)

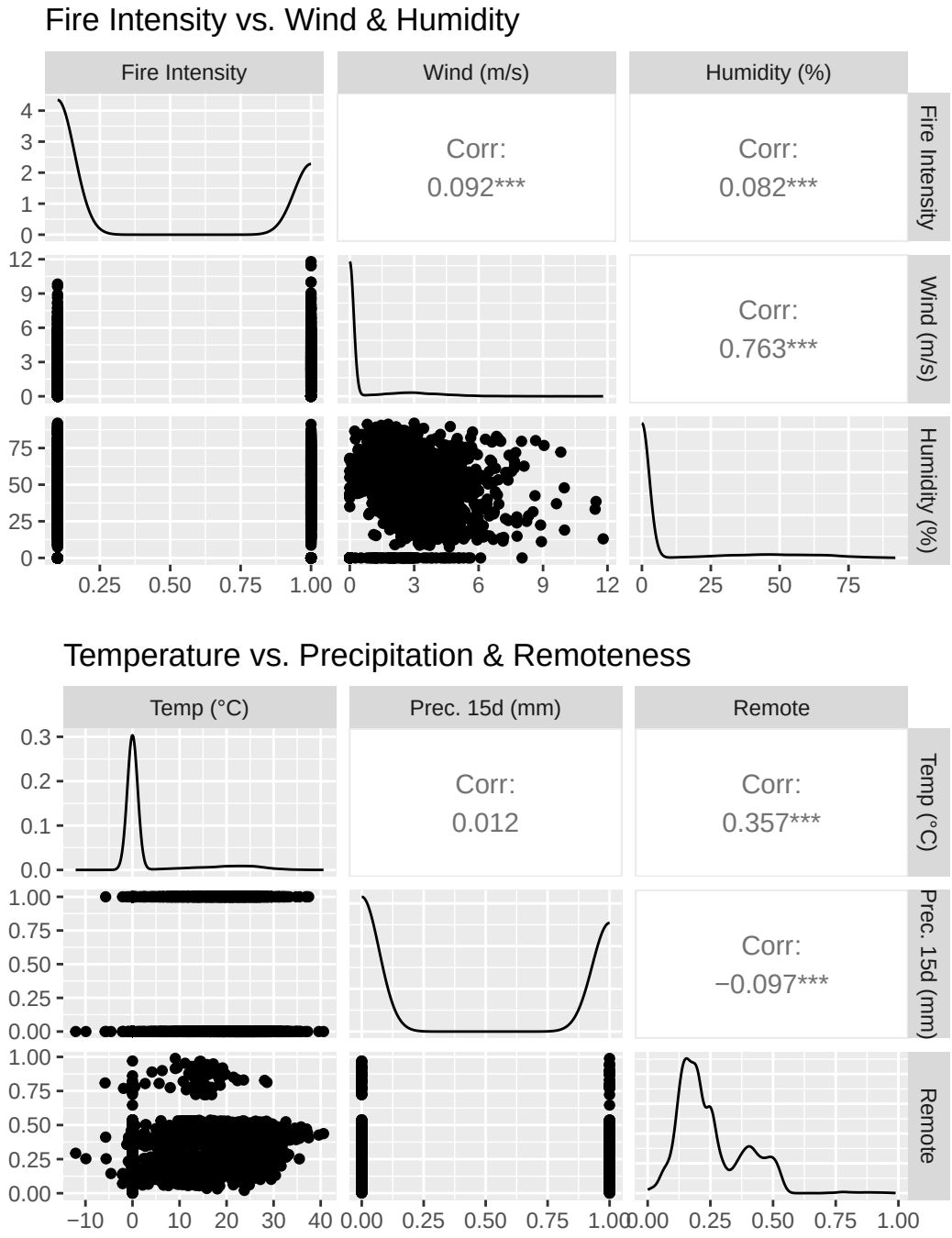

Residuals by Region
75
50
25
0
−25
Residuals
Midwest
Northeast
Other
South
West
Region


Residuals by Cause
75
50
25
0
−25
Residuals
Arson
Debris Burning
Infrastructure
Natural
Other/Unknown
Recreational
Cause Category





Pairwise plots (if they add value)

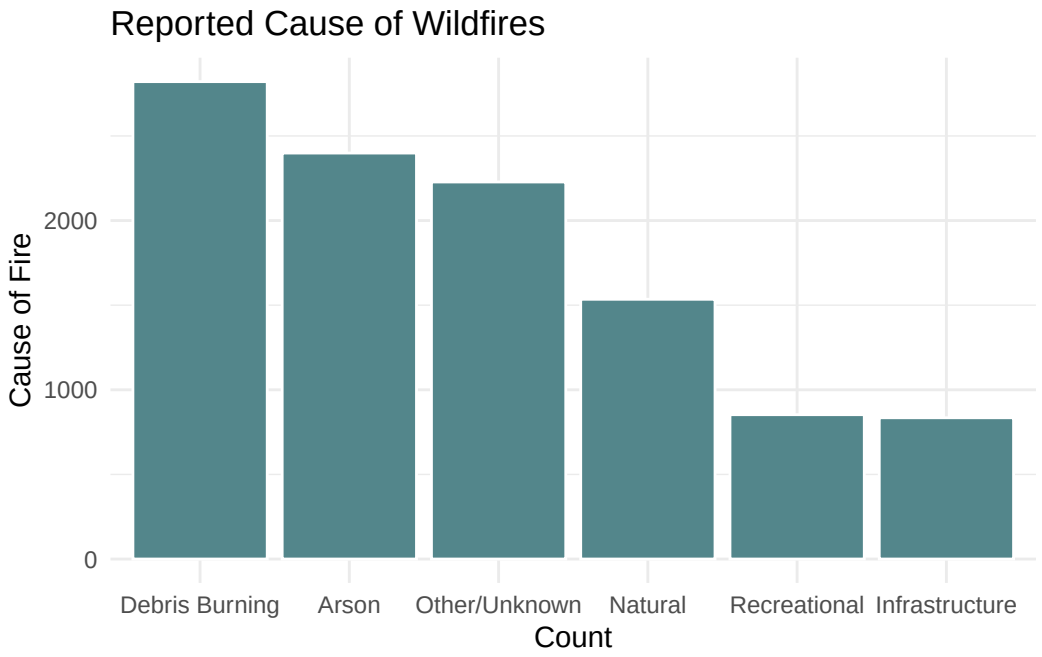


We began our analysis by creating pairwise plots to explore relationships between key numerical variables related to fire behavior, environmental conditions, and containment efforts. Since pairwise plots display scatterplots and correlations, we focused on continuous numerical variables that likely impact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




```
# A tibble: 13 x 2
   stat_cause_descr      n
   <fct>             <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
```

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 frequently 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

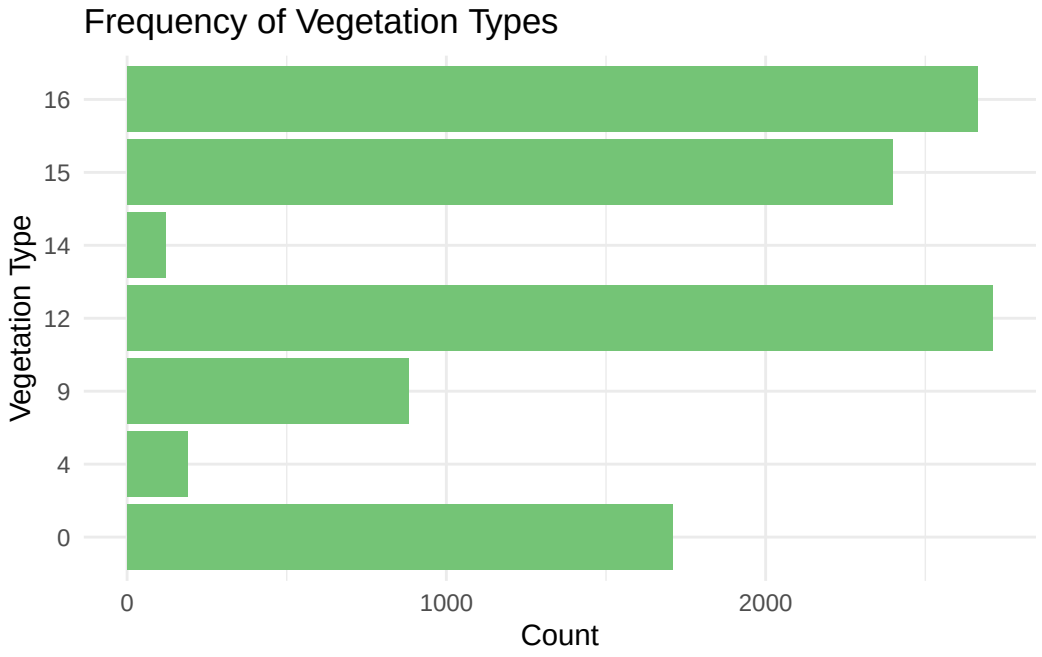
were labeled as missing or undefined. Interestingly, we found that many of the top causes (e.g., debris burning, arson, and equipment use) are related to human activity.

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%)

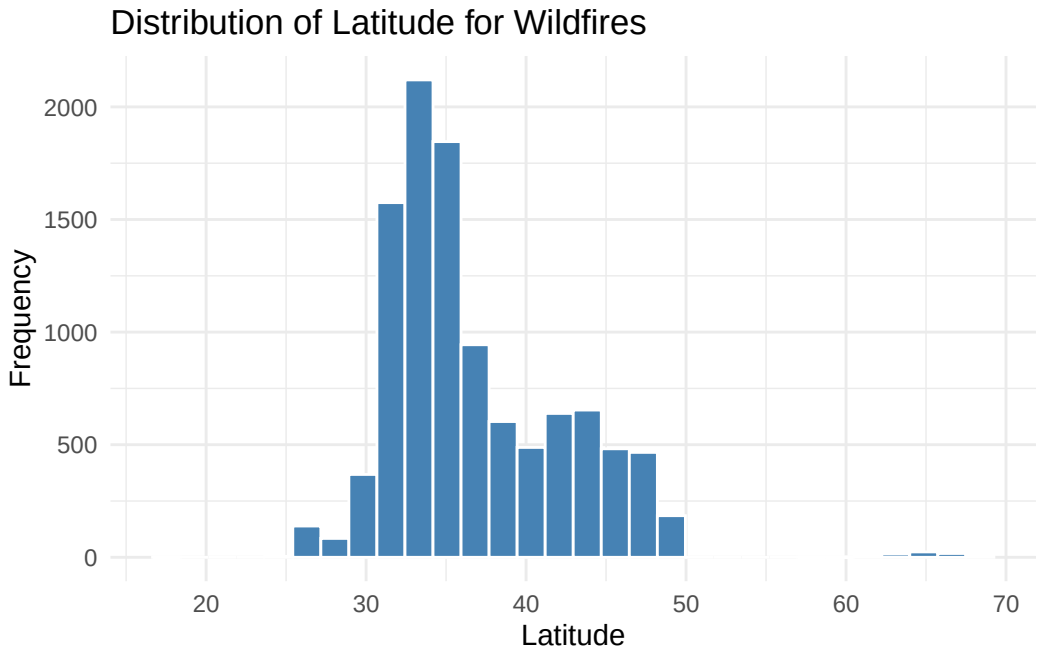
Less common vegetation types represented in the data include desert and temperate evergreen needleleaf forests.

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



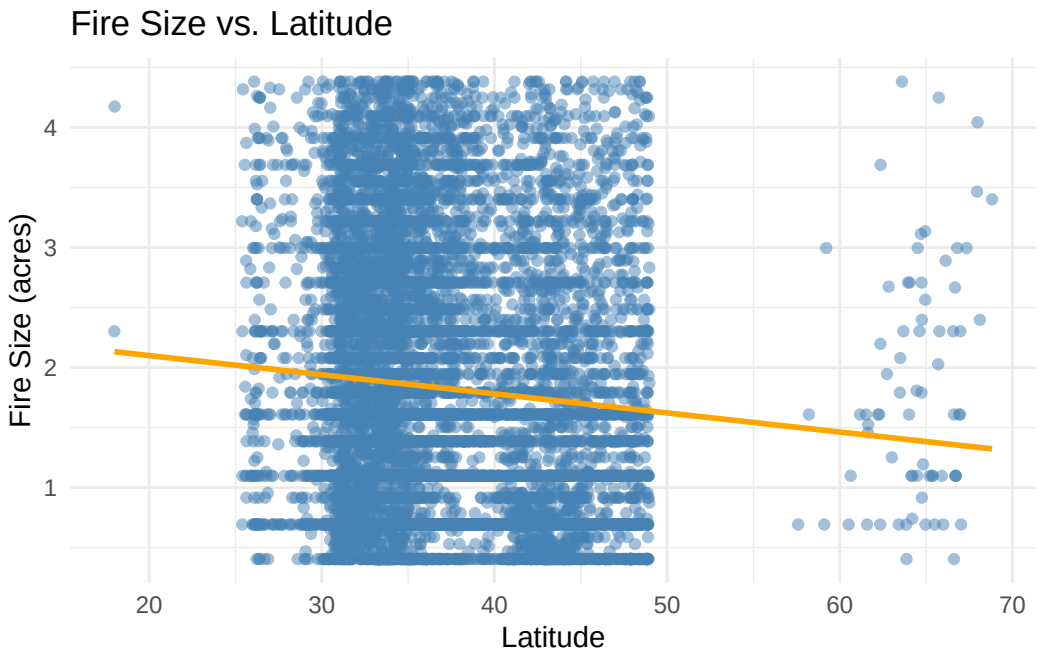

```
        0         4         9        12        14        15        16
16.005998  1.771155  8.265392 25.405304  1.124543 22.462750 24.964858
```

```
   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.

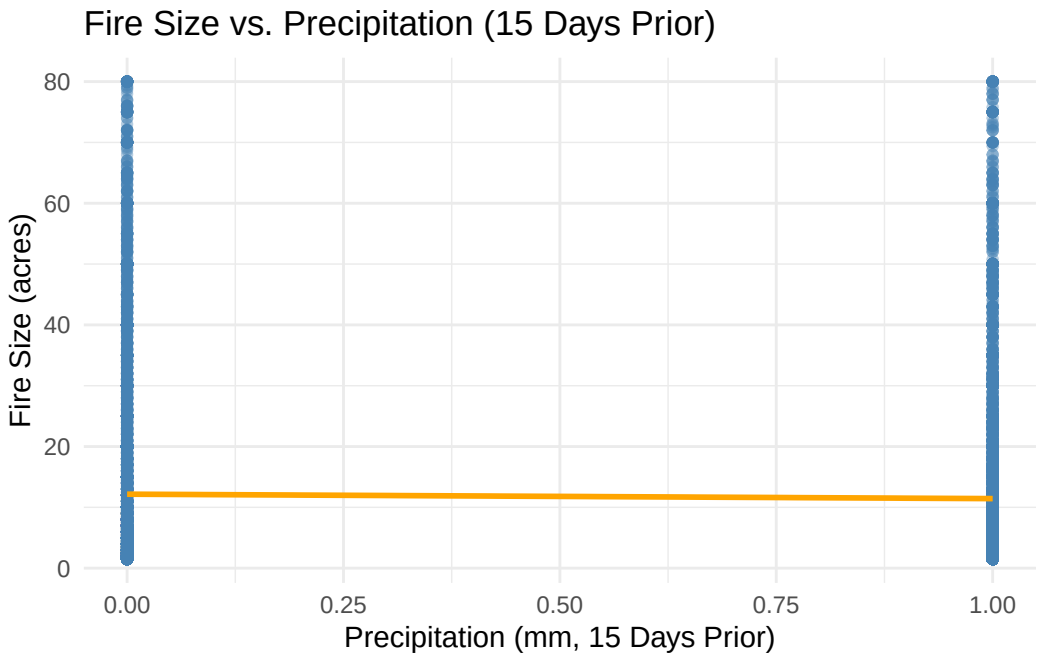
The distribution of latitudes appears to be centered around the mid-30s to upper-30s, which may reflect the concentration of fire-prone areas in those geographic zones.

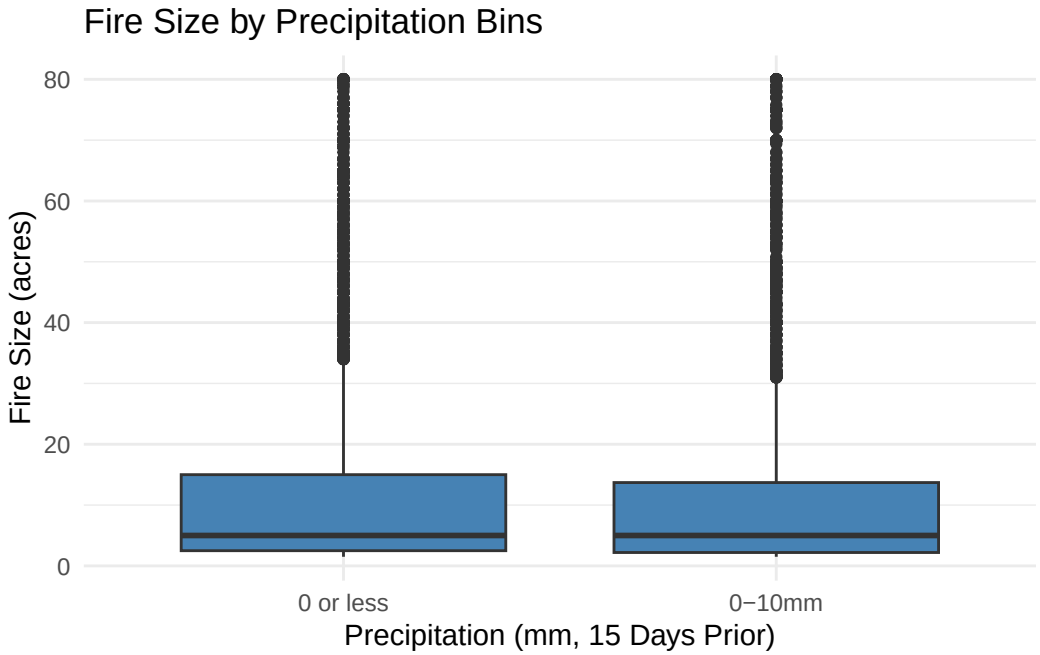



```
[1] -0.04720852
```

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.

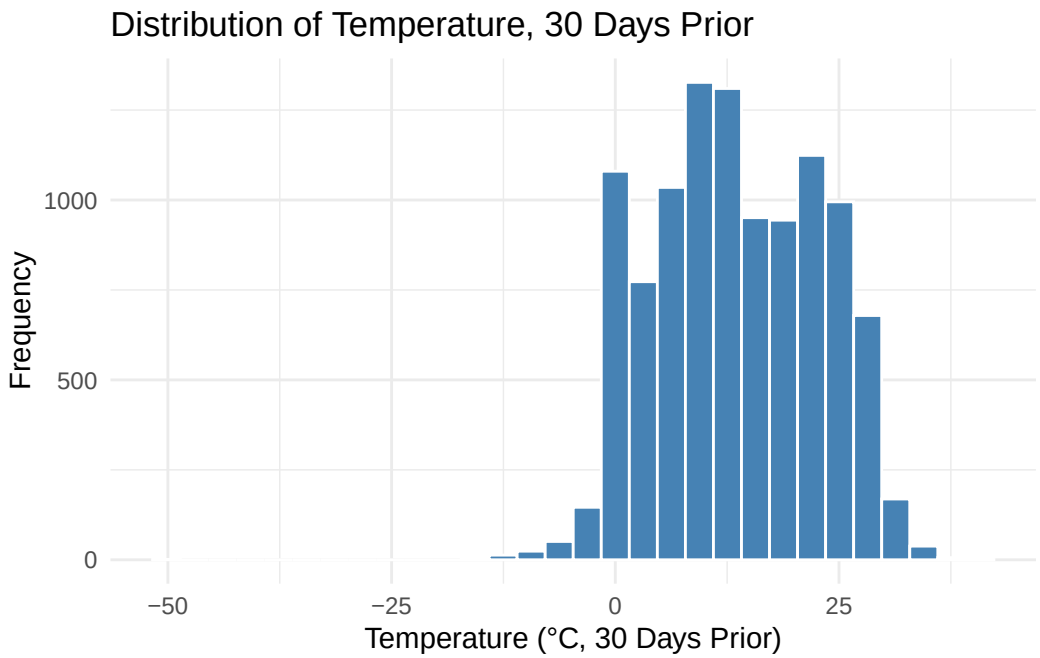




The scatterplot of `Prec_pre_15` and `fire_size` shows no strong relationship between prior precipitation and fire size. Most fires occurred with little to no precipitation, and there is substantial variability in fire size regardless of precipitation level. A few extreme precipitation values do not 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   6.531  12.972  13.578  21.356  41.678
```

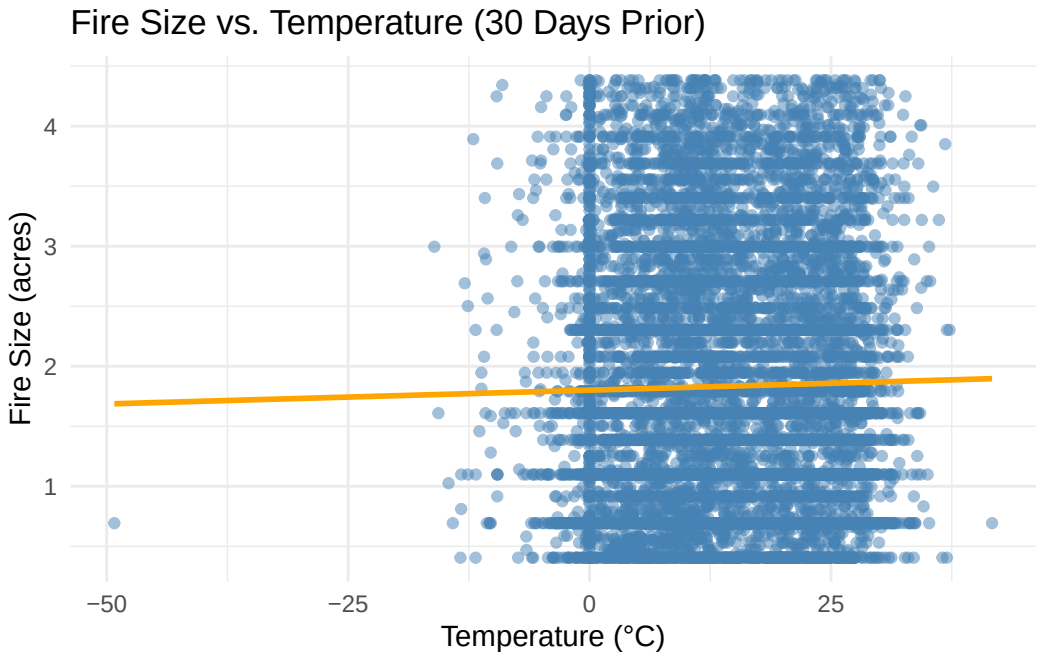


The distribution of Temp_pre_30, which represents the temperature at the location of a fire up to 30 days prior, shows a wide range of values, spanning from -49.211°C to 41.678°C. The minimum recorded temperature of -49.211°C is exceptionally low and may indicate an outlier or a potential data entry error, especially considering that there are few locations that can get that 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.

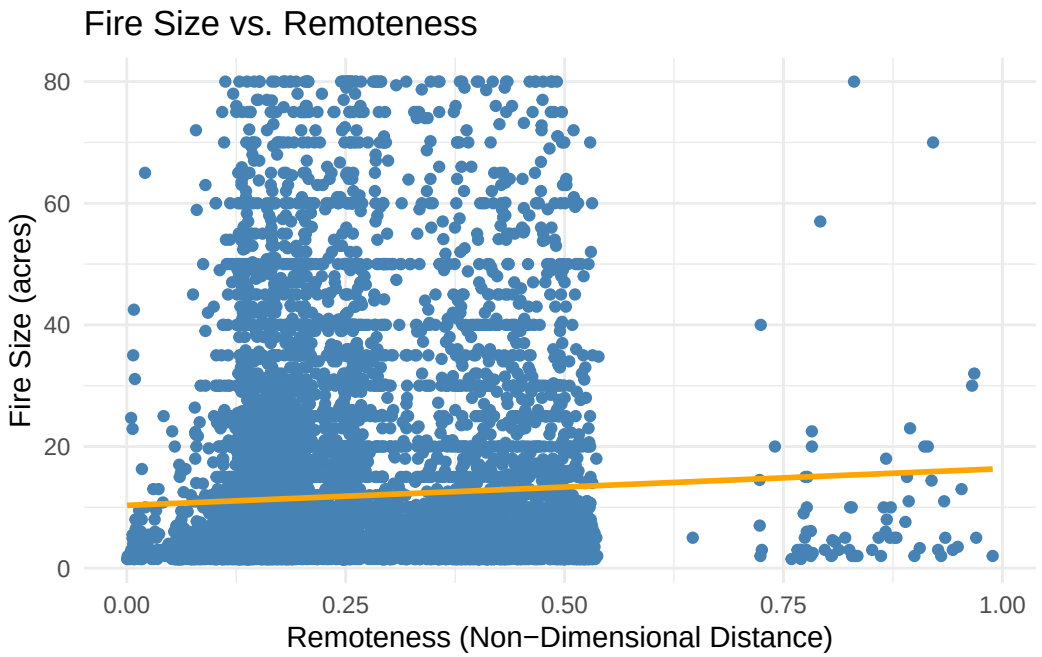
The median temperature of 8.310°C suggests that half of the recorded fires occurred in moderate conditions, while the mean of 9.562°C, which is slightly higher than the median, indicates that the distribution is right-skewed. This means that some extremely high-temperature values may be pulling the average 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.



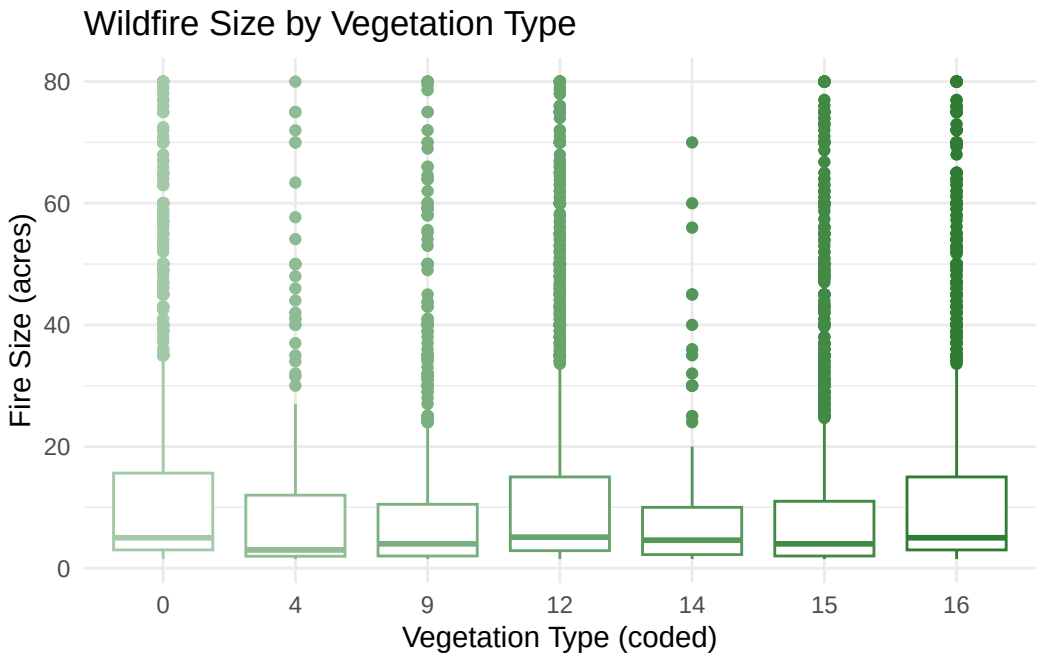
The scatter plot shows the relationship between fire size and temperature 30 days prior, with little evidence of a strong correlation. The nearly flat regression line suggests that temperature alone is not a key predictor of fire size. Most fires occurred between 0°C and 25°C, with fire sizes clustered 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.



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.

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.

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.