Analyzing Factors Associated with Area Burned by Wildfires 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

Before conducting our analysis, we applied many data cleaning steps to prepare our dataset for modeling and interpretation. One of the major decisions we made as part of our data cleaning process was to filter our response variable, acres burned, since the majority of observations in our dataset (over 13,000) recorded fires that burned one acre or less of land (i.e., small-scale wildfires). In our analysis, we decided to focus only on the interquartile range (middle 50%) of wildfires by acres burned because a) the goal of our analysis was to focus on wildfires that can reasonably be addressed during early containment efforts (rather than fires that had already expanded beyond an early intervention phase) and b) to narrow our practical range for analysis (i.e., wildfires representative of "typical" wildfire events).

Our data cleaning process also involved transforming many of our variables of interest to make downstream analysis and interpretation more manageable. For example, we observed

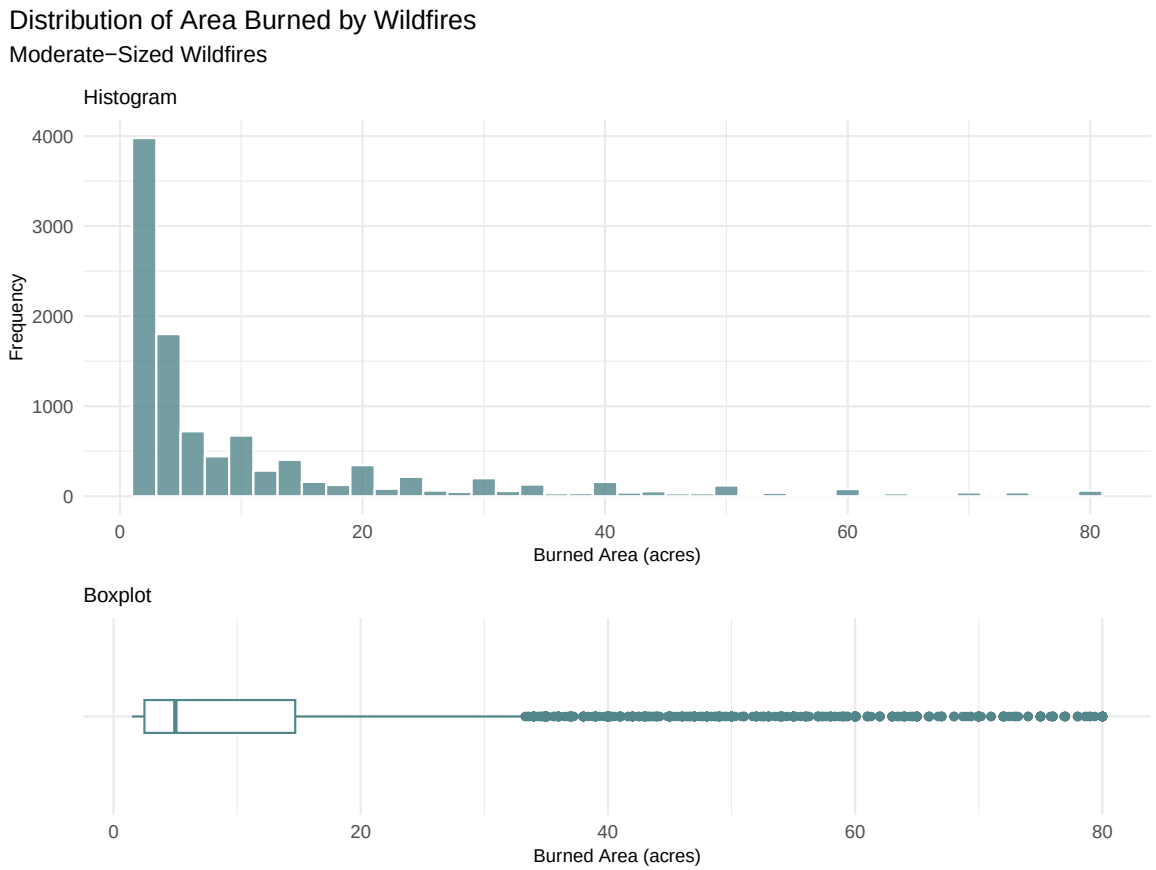
that most precipitation-related variables in our dataset were heavily left-skewed (i.e., the majority of observations corresponded to wildfire events with no recorded precipitation during a given time period before discovery), so we transformed all precipitation variables into binary indicators (0 = no precipitation, 1 = precipitation greater than 0). Additionally, we "collapsed" many categorical variables with a large number of levels in our dataset, mainly to help with model interpretability and to reduce model complexity. Specifically, we grouped states into four broader regions based on U.S. Census classifications (Northeast, Midwest, South, and West), vegetation types into broader environmental categories (e.g., Forest, Shrubland, Grassland), and cause descriptions into broader cause categories (e.g., Natural, Recreational, Infrastructure).

Following standard data cleaning practices, we dropped all observations with missing values for any variables of interest in our dataset. Finally, we converted all categorical variables of interest (e.g., vegetation type, fire cause, geographic region) into factor variables to support appropriate handling during modeling and analysis.

Univariate EDA

As part of our analysis, we first conducted univariate EDA to better understand the distribution of variables in our dataset and to identify patterns (e.g., skewness, outliers) and potential data quality issues that could affect subsequent modeling. As mentioned previously, we focused on the middle 50% of wildfires by acres burned because these represent moderate-sized fires that are more likely to be responsive to early containment efforts.

Based on the histogram of burned area (top panel), we can see that the distribution of moderate-sized wildfires is right-skewed and unimodal. Additionally, we can see that there is a clear peak at 3,978 observations corresponding to wildfires that burned less than 2 acres of land. According to the summary statistics, the mean burned area within this subset is 11.84 acres, and the typical (median) wildfire size is 5 acres. The middle 50% of burned area values in our dataset falls between 2.5 acres (first quartile) and 14.7 acres (third quartile), and the minimum and maximum values are 1.5 and 80 acres, respectively; the standard deviation is 15.62 acres, which tells us there is substantial variability in fire size within this range. Additionally, based on the boxplot visualization (bottom panel), we can see that there are many wildfires with burned areas above 30 acres that extend beyond the typical range of values. These wildfires are moderate outliers within this subset.



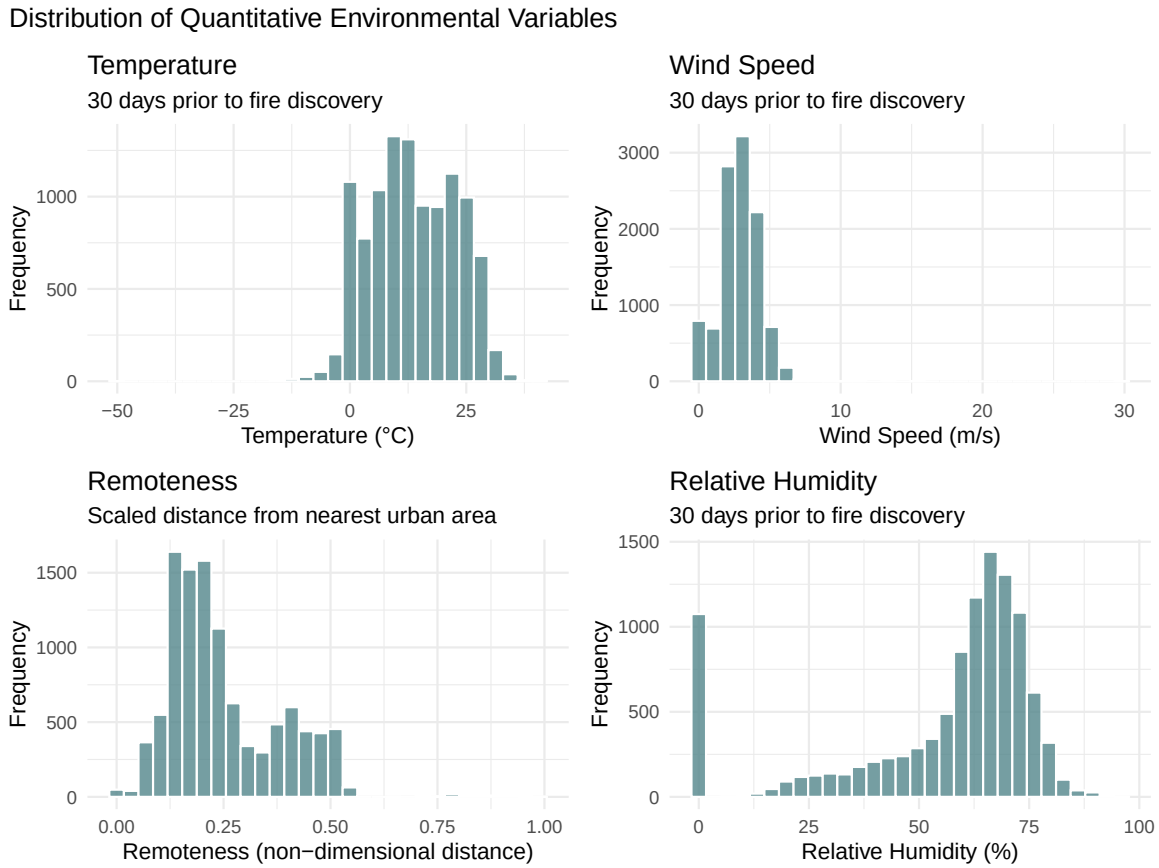
Count	Min	Q1	Median	Mean	Q3	Max	SD
10671	1.5	2.5	5	11.84	14.7	80	15.62

Next, we visualized the distribution of the quantitative variables of interest in our dataset (temperature, relative humidity, and wind speed from 30 days prior to fire discovery, and remoteness), along with summary statistics to better understand their scale and variability. Among these predictors, we first examined the distribution of temperature a month before discovery (top left panel), which we observed to be approximately normal and unimodal. (This is consistent with what we would generally expect, since temperature tends to change gradually across different areas rather than clustering at specific values.) The median temperature for wildfire observations in our dataset is 12.97°C and the mean is 13.58°C, with most observations between 6.53°C and 21.36°C, and a range from -49.21°C to 41.68°C.

We then examined the distributions of wind speed thirty days before discovery and remoteness (top right and bottom left panels, respectively). Wind speed appears to be very slightly right-skewed and unimodal, while remoteness appears more strongly right-skewed and unimodal.

This indicates that most fires occurred relatively close to populated areas (lower remoteness) and that most wind speeds were low to moderate. The median remoteness for observations in our dataset is 0.21 and the mean is 0.25, with most observations between 0.15 and 0.35, and values ranging from 0 to 0.99. The median wind speed for observations in our dataset is 2.89 m/s and the mean is 2.87 m/s, with most observations between 2.06 m/s and 3.76 m/s, and a range from 0 m/s to 29.80 m/s.

Lastly, we looked at the distribution of relative humidity thirty days before discovery (bottom right panel), which appears to be left-skewed and unimodal. This tells us that lower humidity conditions were more common among wildfires in our data. The median relative humidity for observations in our dataset is 63.53% and the mean is 55.20%, with most observations between 48.80% and 69.97%, and values from 0% to 96%. However, we noticed that a large number of observations in our data recorded a relative humidity of exactly 0%. Given that 0% relative humidity is unlikely under normal atmospheric conditions (since even very dry air typically contains some moisture), we decided to exclude these observations from our analysis.

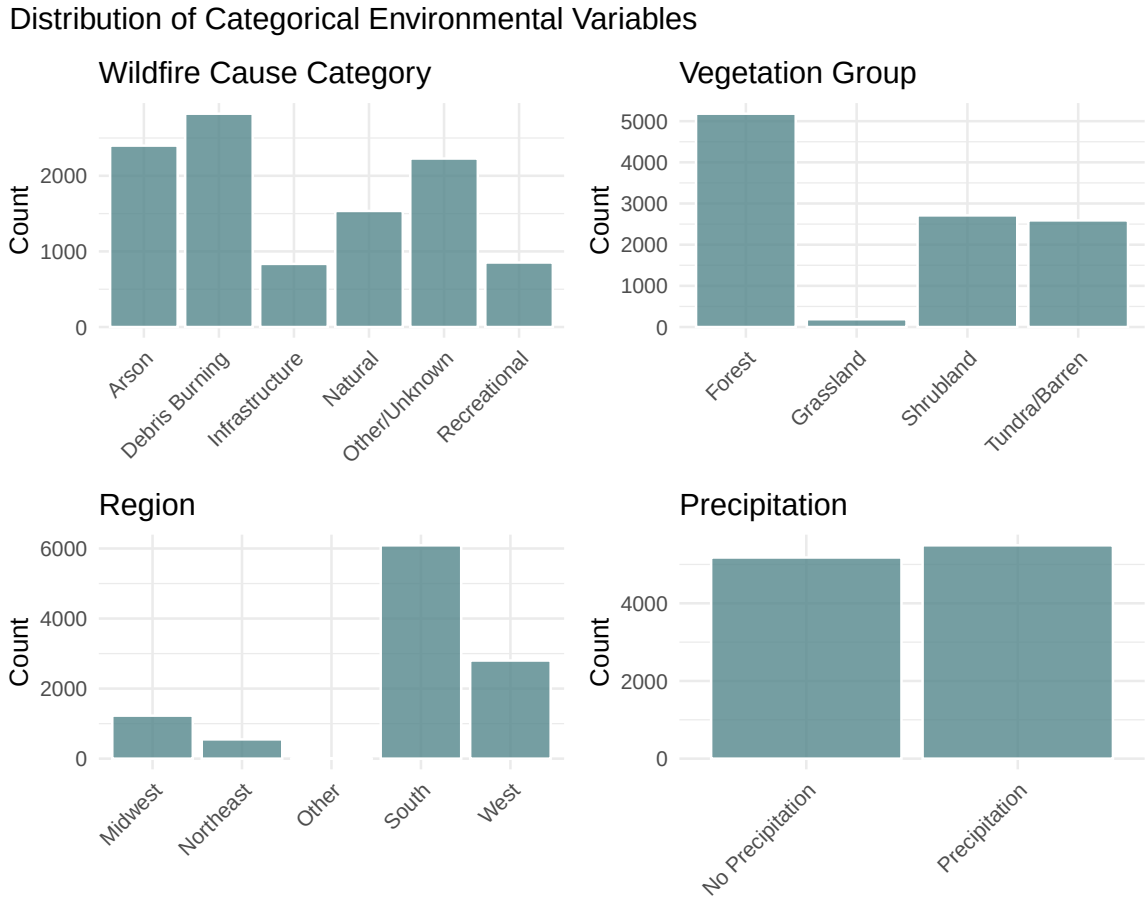


Variable	Min	Q1	Median	Mean	Q3	Max	SD
Temperature	-49.21	6.53	12.97	13.58	21.36	41.68	9.21
Humidity	0.00	48.80	63.53	55.20	69.97	96.00	22.76
Wind	0.00	2.06	2.89	2.87	3.76	29.80	1.40
Remoteness	0.00	0.15	0.21	0.25	0.35	0.99	0.13

In addition to examining the distributions of quantitative predictors, we also visualized the distribution of categorical variables of interest in our dataset (i.e., wildfire cause category, vegetation group, region, and precipitation 30 days prior to wildfire discovery). Starting with precipitation 30 days before discovery, we observed that the majority of wildfire observations (5492 observations or 51.5%) corresponded to fires with no precipitation during that period, while the remaining observations (5179 observations or 48.5%) corresponded to fires where precipitation occurred.

Looking more closely at vegetation group, we can see that most fires occurred in forested areas (5181 observations or 48.6%), followed by shrubland (2711 observations or 25.4%) and tundra or barren areas (2590 observations or 24.3%), with relatively few fires in grassland environments (189 observations or 1.77%). Examining wildfire cause categories, we found that a large portion of fires were attributed to human-related causes rather than natural causes. Debris burning accounted for the highest proportion (2821 observations or 26.4%), followed by arson (2399 observations or 22.5%) and other or unknown causes (2228 observations or 20.9%). Natural causes made up a smaller proportion (1535 observations or 14.4%), followed by recreational causes (853 observations or 8.0%) and infrastructure-related causes (835 observations or 7.82%).

Finally, we examined the regional distribution of fires. Most fires in our dataset occurred in the South (6090 observations or 57.1%), followed by the West (2801 observations or 26.2%), Midwest (1228 observations or 11.5%), and Northeast (550 observations or 5.15%). Only two fires (2 observations or less than 0.01%) were categorized as occurring in the "Other" region classification.



Bivariate EDA

As the next step in our analysis, we examined the relationship between each of our quantitative predictors and fire size to better understand their associations. As we can see in the scatter plot of burned area versus temperature (top left panel), the relationship between these two variables appears to be very weak and slightly positive (i.e., suggesting that larger fires may be slightly more likely at higher temperatures). The relationship also appears to be nonlinear, since most fires in our dataset seem to be pretty small regardless of temperature, and the largest fires (greater than 60 acres in size) most commonly occur between 5°C and 20°C. This tells us that while moderate temperatures may allow for larger fires, temperature alone does not strongly control fire size.

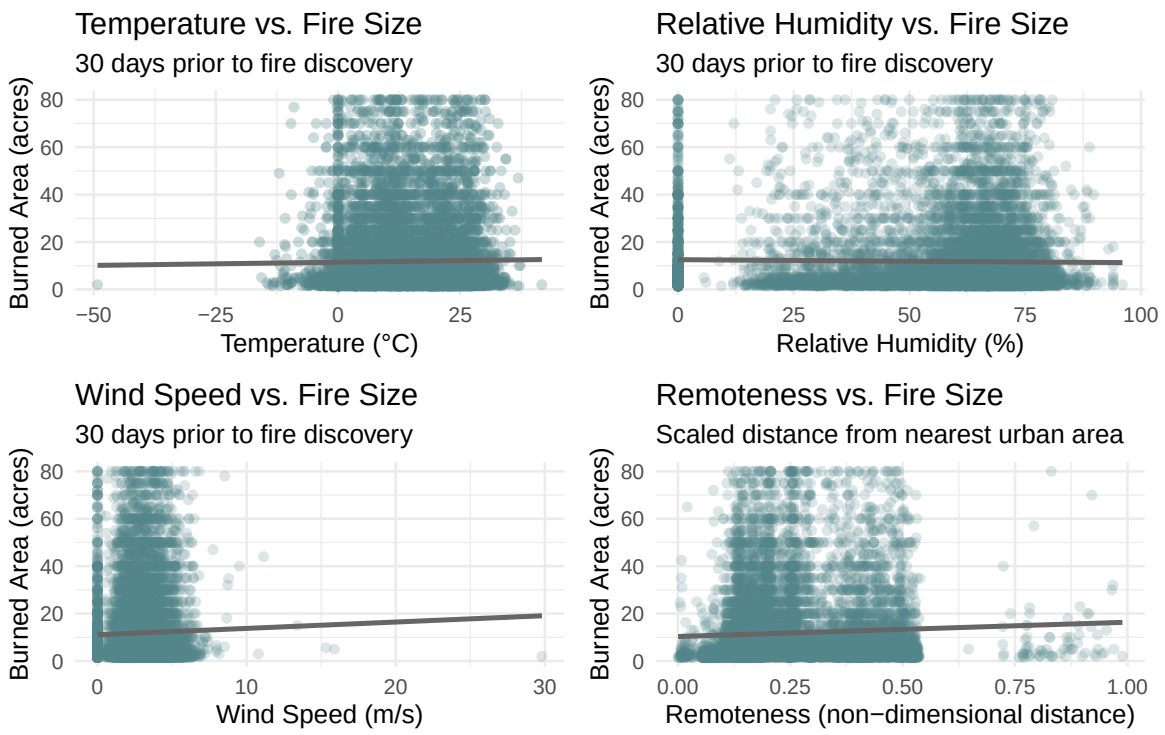
Similarly, the relationship between relative humidity and burned area (top right panel) appears to be very weak and highly nonlinear (with fires of all sizes occurring across almost all humidity levels and the largest fires tending to occur at mid-range humidity values at approximately 25% to 75%), but slightly negative (i.e., suggesting that burned area slightly decreases as

humidity increases). As with temperature, the wide vertical spread of data points around the line of best fit tells us that relative humidity may not be a strong predictor of burned area on its own.

Additionally, the relationship between wind speed and burned area (bottom left panel) appears to be slightly positive (i.e., indicating a slight tendency for larger fires to occur at higher wind speeds), very weak, and highly nonlinear. As we can see in the plot, most data points in the plot are concentrated at lower wind speeds, and most larger fires (fires greater than 40 acres in size) occur below 5 m/s, which indicates that higher wind speeds are not consistently associated with larger burned areas.

Lastly, looking at the scatter plot of remoteness versus burned area (bottom right panel), we see a weak, slightly positive relationship between these two variables (i.e., suggesting that fires originating further away from cities may end up burning more acres of land, on average). This relationship also appears to be highly nonlinear, since the largest fires occur at moderate remoteness values between 0.25 and 0.50, and relatively few large fires at very high remoteness values greater than 0.75 (which tells us that burned area does not consistently increase or decrease across the range of remoteness values).

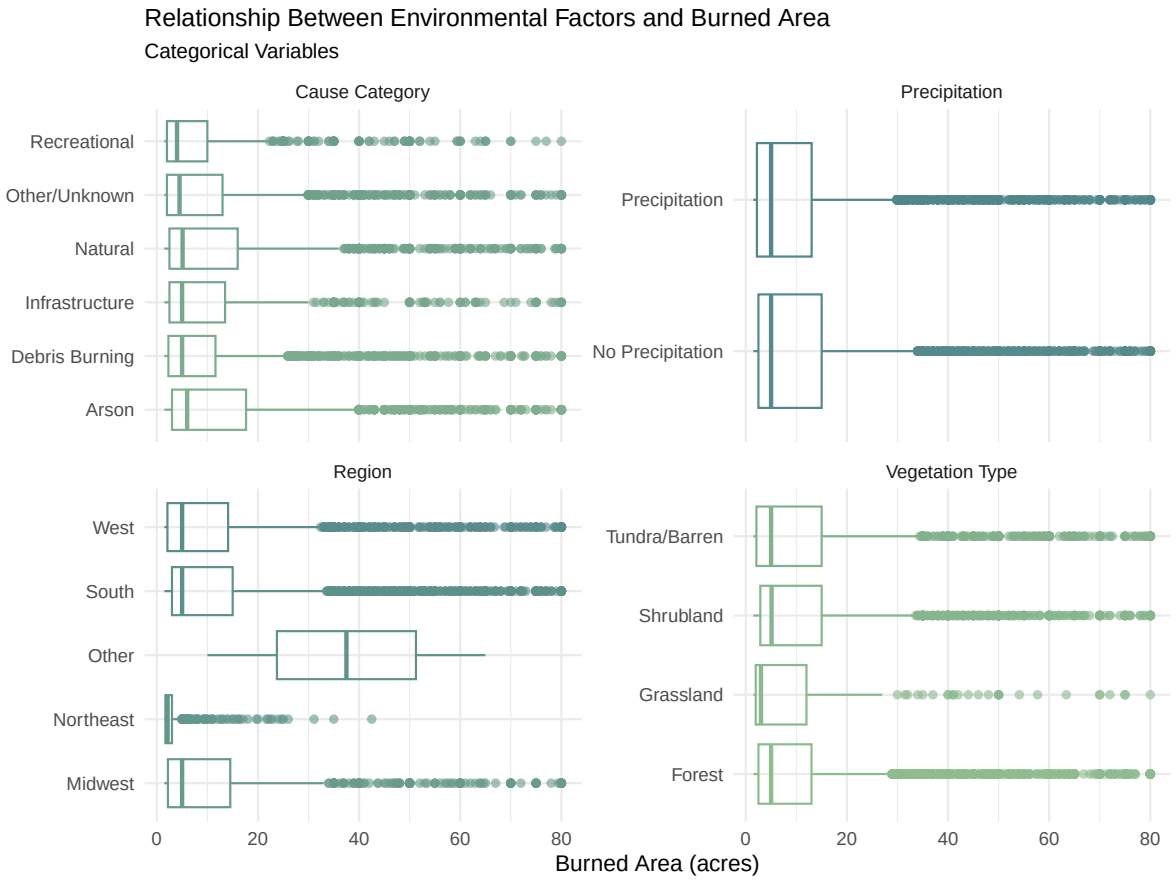
In general, there is a wide vertical spread (i.e., high variability) in the data points around the lines of best fit for all predictors, which suggests that the quantitative predictors in our dataset may have relatively limited predictive power for burned area when considered individually. That being said, this is somewhat expected given that wildfire size can be influenced by many factors beyond the limited set of quantitative predictors available in our dataset.



As part of our bivariate EDA, we also analyzed the relationship between each of our categorical variables of interest and burned area to better understand differences in fire size across groups. We can see from the visualization (top left panel) that the distribution of acres burned is relatively similar for all wildfire cause categories, with medians between 4.00 and 6.00 acres and means between 9.02 and 13.74 acres, and outliers beyond the 30-acre mark across all cause types. This indicates that cause category on its own does not explain much variability in fire size. We observed a similar pattern in terms of the distributions between observations corresponding to no precipitation 30 days before discovery and observations corresponding to precipitation (top right panel). The distributions of fire size appear to be very similar, with medians of 5.00 acres in both groups and means of 12.32 acres for no precipitation and 11.38 acres for precipitation, although there is slightly more spread in the middle 50% of fires with no precipitation compared to fires with precipitation.

We observed more apparent differences in the distribution of acres burned across different regions and vegetation groups (bottom left and bottom right panels, respectively). In particular, the Northeast had a lower median (3.00 acres) and mean (6.04 acres) compared to other regions, with much less variability in the middle 50% of fires and most outliers limited to about the 40-acre mark. In contrast, distributions for the West, South, and Midwest were more similar, with medians between 5.00 and 6.00 acres, means between 12.00 and 13.00 acres,

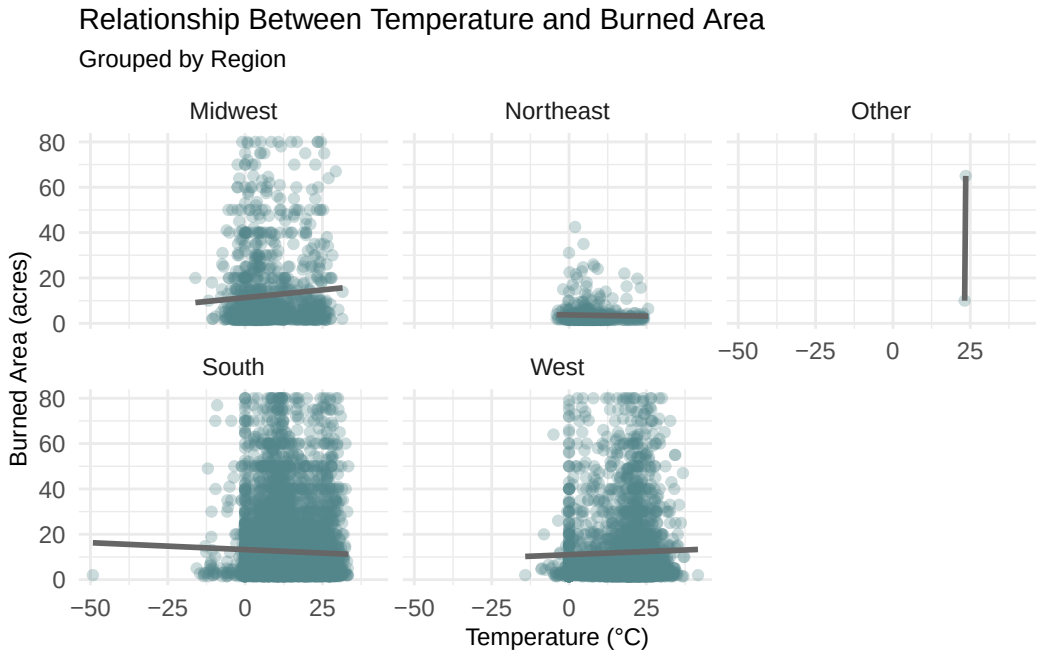
and a large number of outliers beyond the 30-acre mark. For vegetation type, the distribution of fire size was fairly similar across categories, although grasslands had a slightly lower median (4.00 acres) and mean (8.43 acres) compared to other vegetation groups. That being said, the middle 50% of fire sizes and the overall range of outliers were generally similar across vegetation types, with grasslands showing somewhat fewer extreme outliers. Overall, region and vegetation group seem to explain more of the variability in fire size on their own compared to cause category or precipitation, and could potentially be useful predictors in modeling fire size.



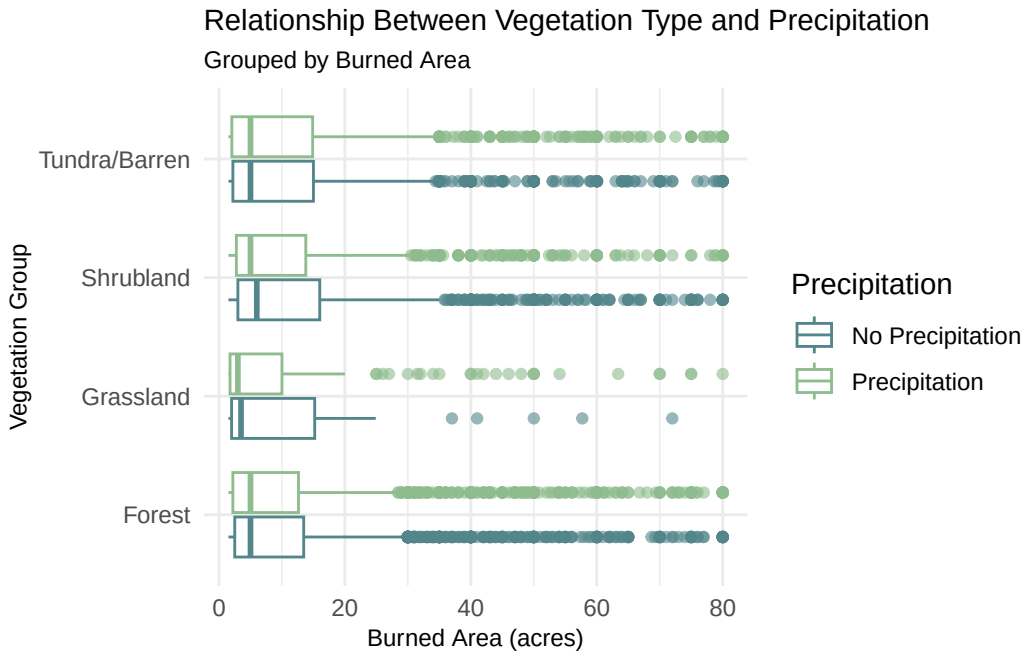
Potential Interaction Effects

Since many of our variables of interest relate to environmental conditions and geographic location, there are many possible interaction effects among our predictors that could help explain variability in acres burned. As part of our analysis, we looked at two potential interaction effects that we thought might be the most meaningful. First, we analyzed whether there was a potential interaction between temperature and region in explaining variability in burned area, since temperature may influence wildfire size differently across regions (e.g., regions like

the Northeast generally experience cooler temperatures, which could lead to different fire behavior compared to hotter, drier regions like the West). As we can see in the visualization below, there does appear to be an interaction effect, since the magnitude and direction of the relationship between temperature and burned area vary across regions. (Specifically, the relationship between temperature and burned area appears to be slightly positive for regions like the Midwest and West, while the trend appears slightly negative for the South and Northeast.) A possible explanation for this could be that, in cooler regions, increases in temperature may support fire spread, while in already hot and humid regions, temperature may not have the same effect due to other limiting factors like vegetation or moisture levels. This interaction would be interesting to look into more formally when modeling fire size to determine whether it significantly improves model performance.



We were also interested in a potential interaction between vegetation type and precipitation, since vegetation and fuel moisture could influence burned area together. However, based on the visualization below, the distributions of burned area appear relatively similar between precipitation and no precipitation conditions for each vegetation group. In forest, the median burned area is 5.0 acres for both no precipitation (mean 11.48 acres) and precipitation (mean 11.14 acres), while in grassland, the median is 3.5 acres for no precipitation (mean 12.26 acres) and 3.0 acres for precipitation (mean 11.54 acres). Additionally, in shrubland, the median is 6.0 acres for no precipitation (mean 13.52 acres) and 5.0 acres for precipitation (mean 11.31 acres), while in tundra/barren, the median is 5.0 acres for both no precipitation (mean 12.81 acres) and precipitation (mean 11.86 acres). Even though there does not seem to be a very strong interaction effect, and we will further examine this during modeling.




```
# A tibble: 8 x 7
# Groups:   Vegetation_group [4]
  Vegetation_group Precipitation    count mean_fire_size median_fire_size
  <fct>            <fct>            <int>          <dbl>            <dbl>
1 Forest           No Precipitation  2642           11.5                5
2 Forest           Precipitation     2539           11.1                5
3 Grassland        No Precipitation    40           12.3                3.5
4 Grassland        Precipitation      149           11.5                3
5 Shrubland        No Precipitation  1397           13.5                6
6 Shrubland        Precipitation     1314           11.3                5
7 Tundra/Barren    No Precipitation  1100           12.8                5
8 Tundra/Barren    Precipitation     1490           11.9                5
# i 2 more variables: min_fire_size <dbl>, max_fire_size <dbl>
```

Methodology

forward seleciton apprahc

```
# 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    0.0316        0.0300  15.4      20.4 2.43e-62    17 -44301. 88640. 88778.
```

```
# 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    0.0360        0.0338  15.4      16.6 7.09e-68    24 -44277. 88606. 88795.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>


              regionNortheast                    regionOther
                     1.927904                       1.006995
                  regionSouth                     regionWest
                     3.773257                       5.362973
 cause_categoryDebris Burning  cause_categoryInfrastructure
                     1.632644                       1.294878
        cause_categoryNatural   cause_categoryOther/Unknown
                     1.815988                       1.642033
   cause_categoryRecreational                    remoteness
                     1.293178                       5.206690
    Vegetation_groupGrassland     Vegetation_groupShrubland
                     1.078625                       1.347203
Vegetation_groupTundra/Barren                   Wind_pre_30
                     1.227884                       1.353596
                 Prec_pre_30                    Temp_pre_30
                     1.203873                       1.335423
                  Hum_pre_30
                     1.590041


# 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
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

term	estimate	std.error	statistic	p.value
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
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
```

```
# 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    0.0391        0.0371  1.06      19.7 2.83e-76    22 -15758. 31564. 31739.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

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

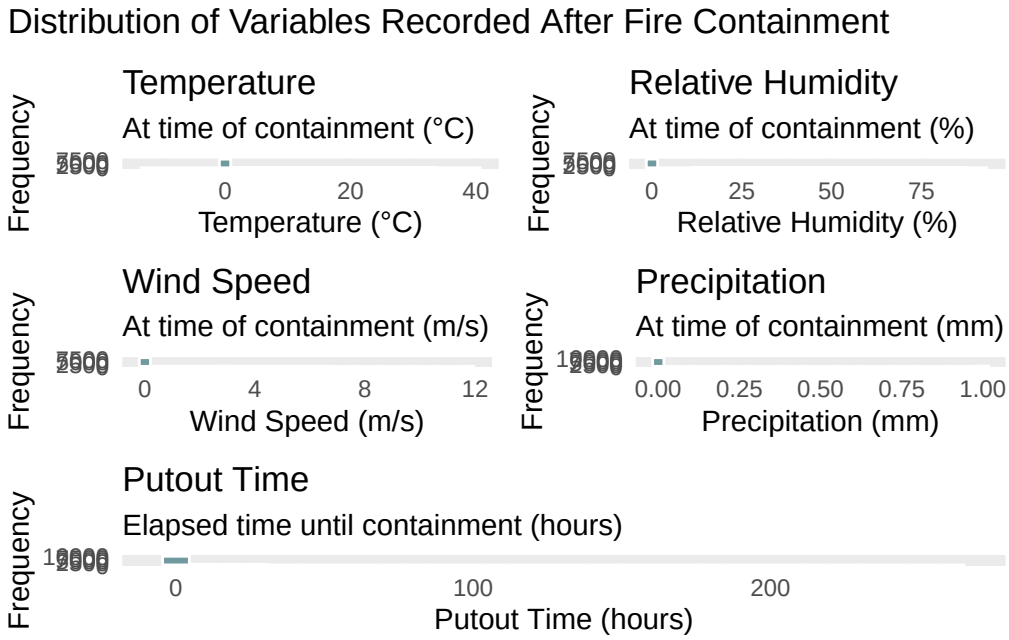
(`fire_int_fit`) had an adjusted R^2 of 0.041. The models using the binary precipitation indicator produced similar results: 0.037 for the main-effects model (`fire_main_bin`) and 0.039 for the interaction model (`fire_int_bin`).

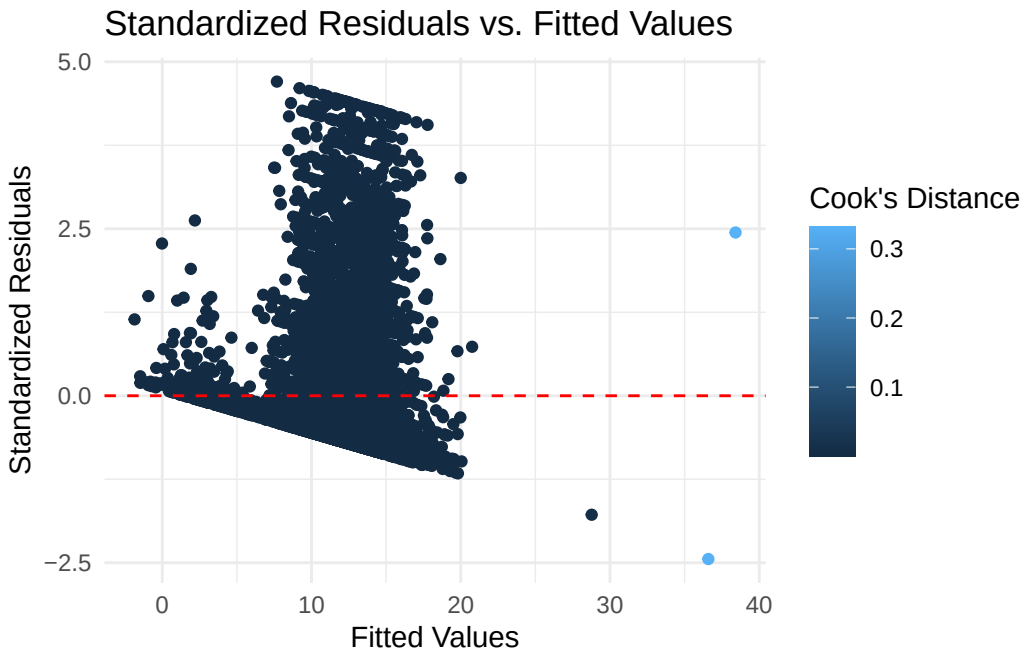
We evaluated prediction error using root mean squared error (RMSE) on the log scale. The RMSE was 1.059 for both `fire_main_fit` and `fire_main_bin`, and slightly lower for the interaction models: 1.057 for `fire_int_fit` and 1.058 for `fire_int_bin`. These values indicate that the average prediction 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.

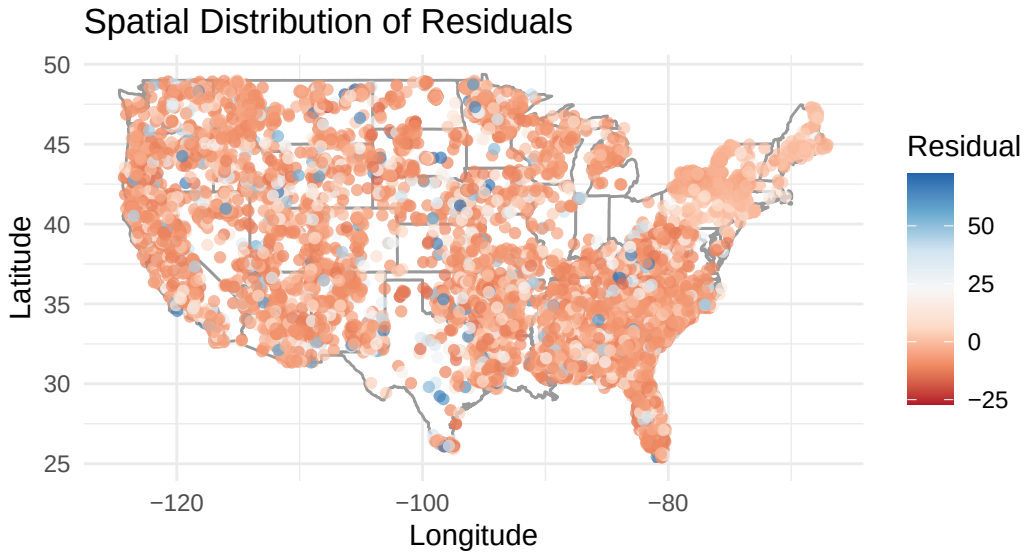
Appendix

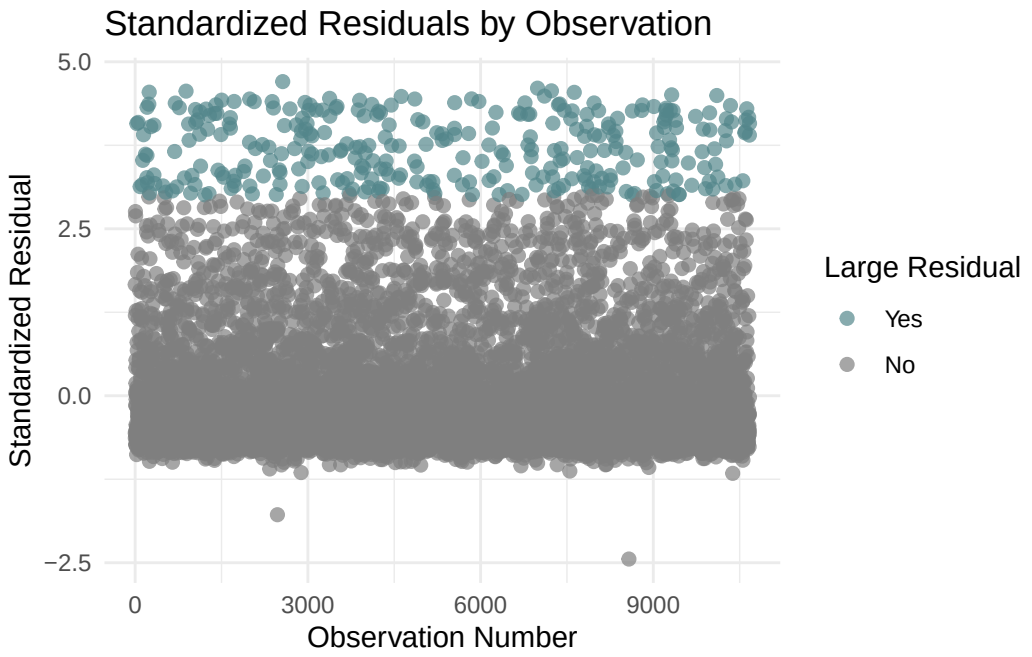
Std. residuals vs. fitted values - constant variance, linearity, cook's distance (influential points)

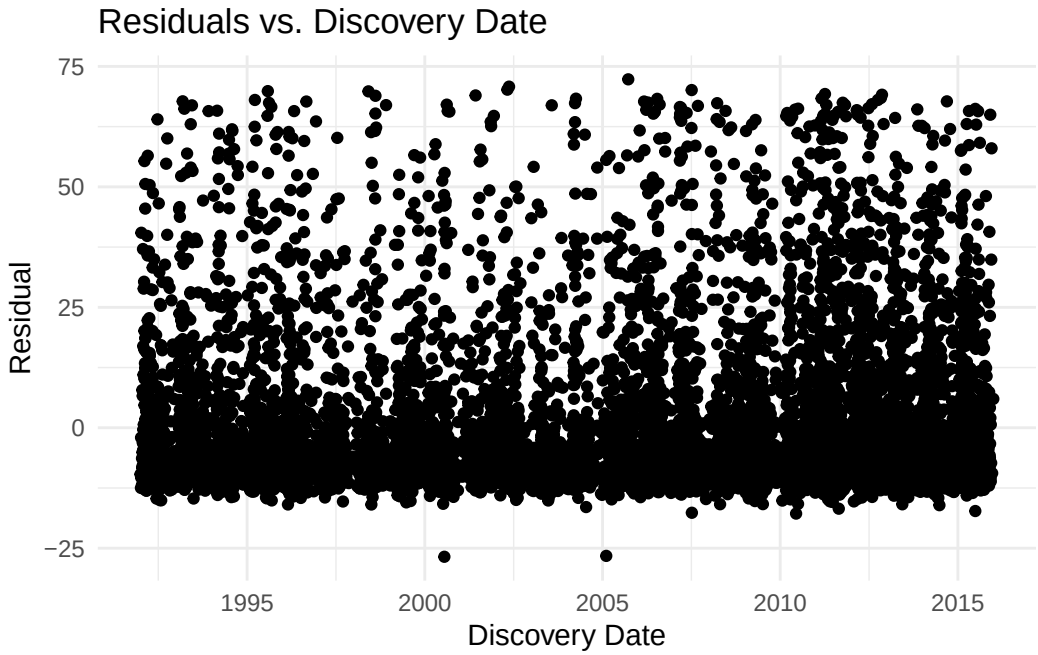




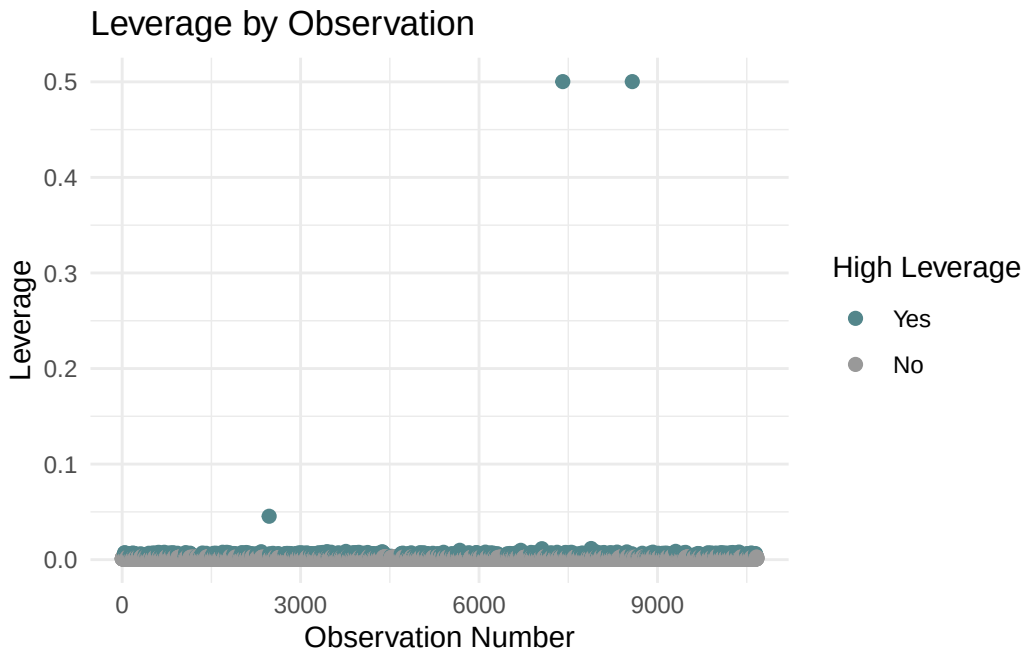
spatial distribution of residuals - independence

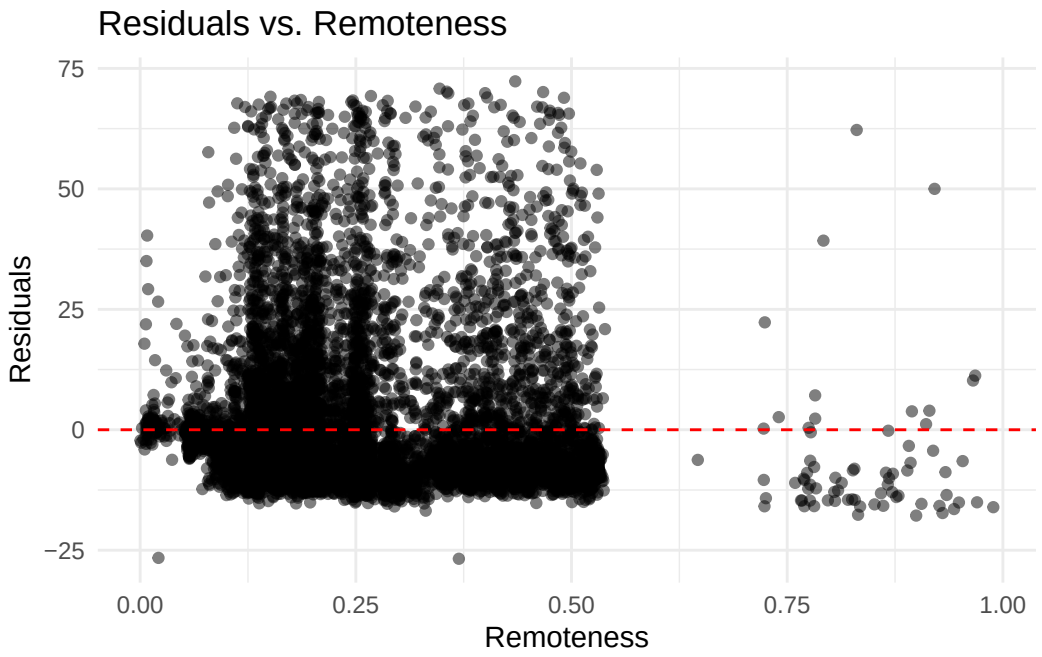


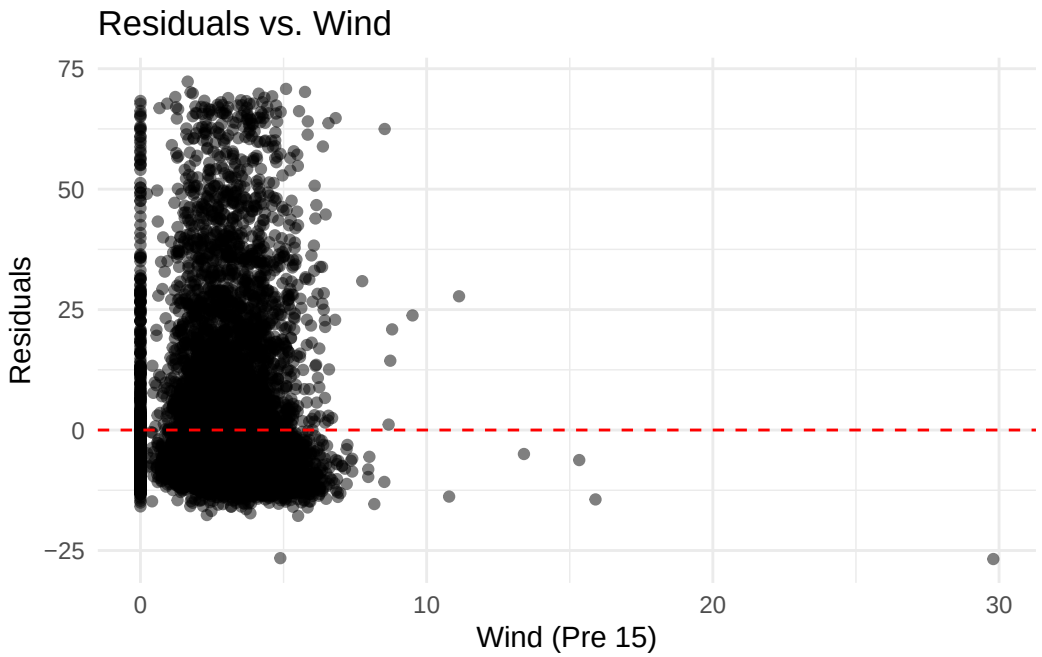


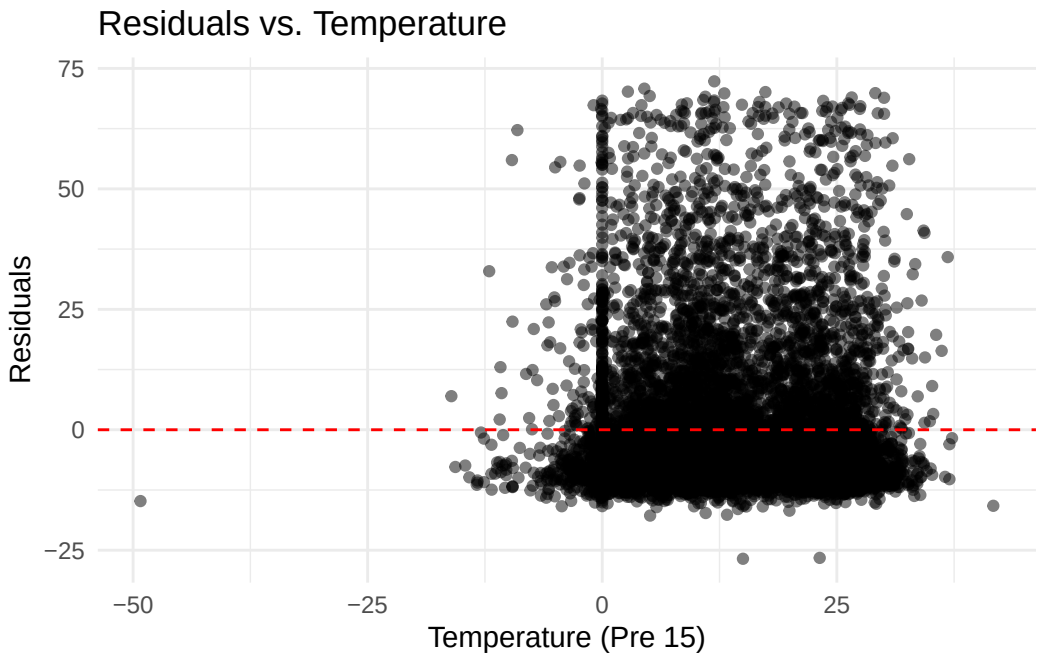


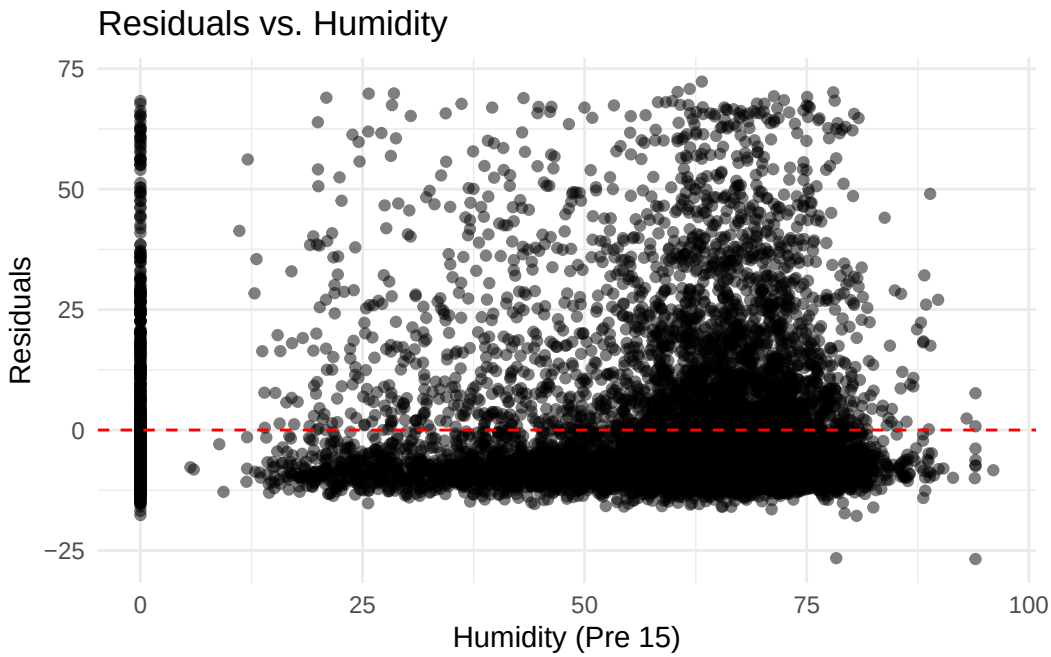
large leverage

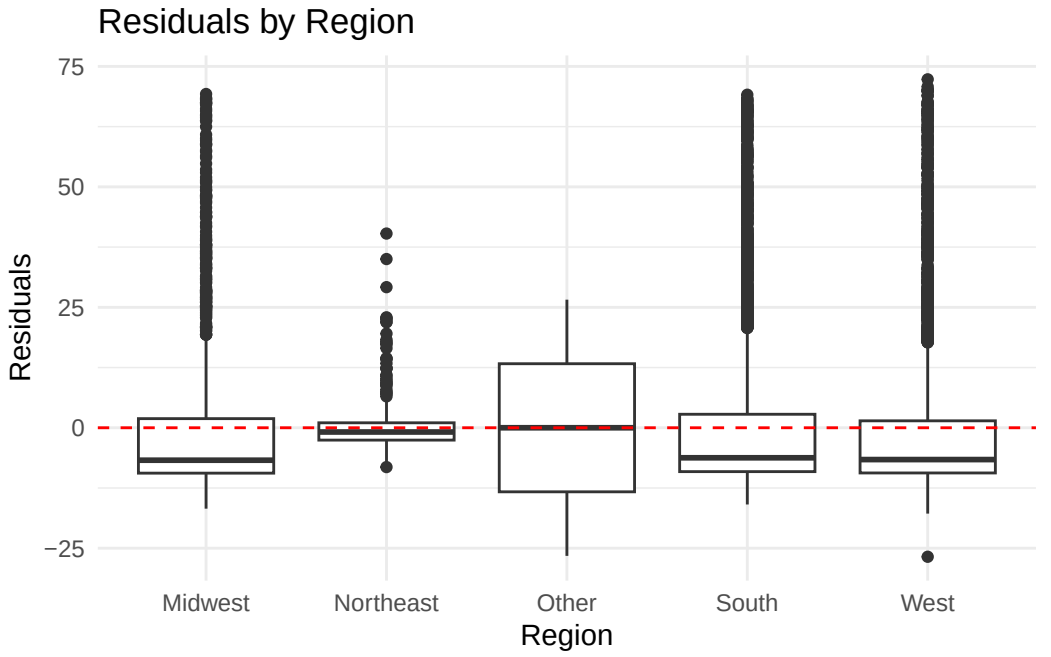

Leverage by Observation
Leverage
0.5
0.4
0.3
0.2
0.1
0.0
0
3000
6000
9000
Observation Number
High Leverage
Yes
No

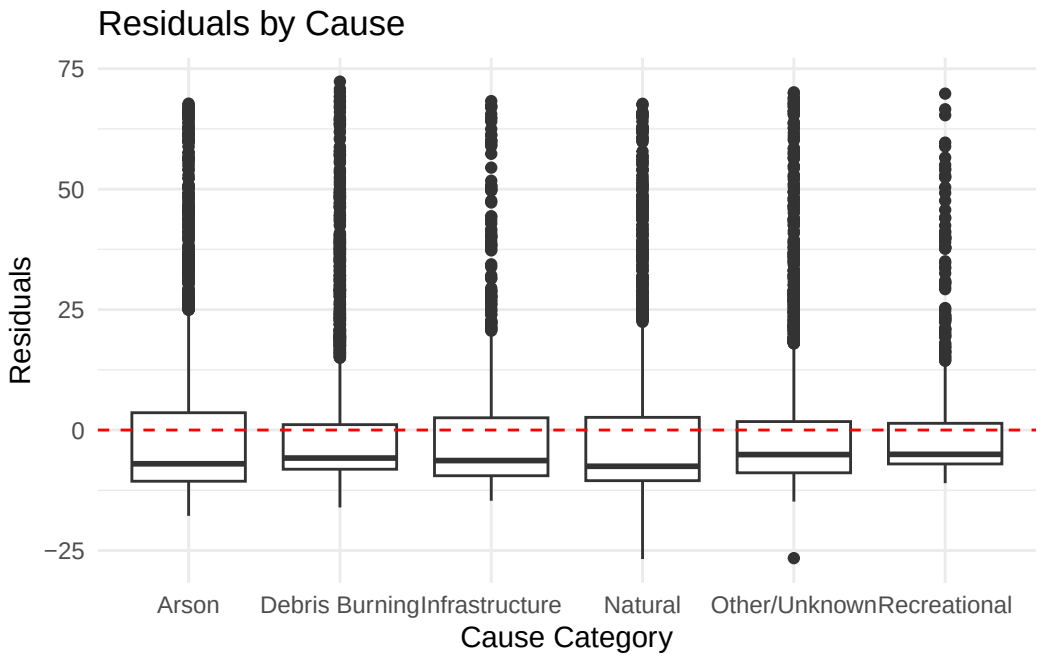

Residuals vs. Remoteness
Residuals
75
50
25
0
−25
0.00
0.25
0.50
0.75
1.00
Remoteness

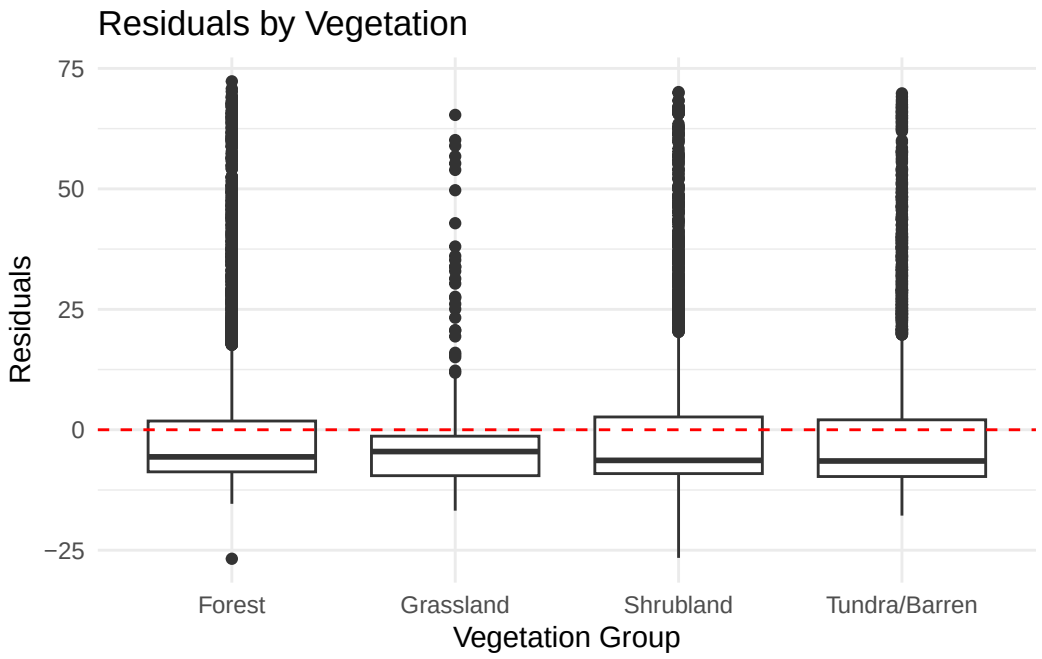

Residuals vs. Wind
Residuals
75
50
25
0
−25
0
10
20
30
Wind (Pre 15)

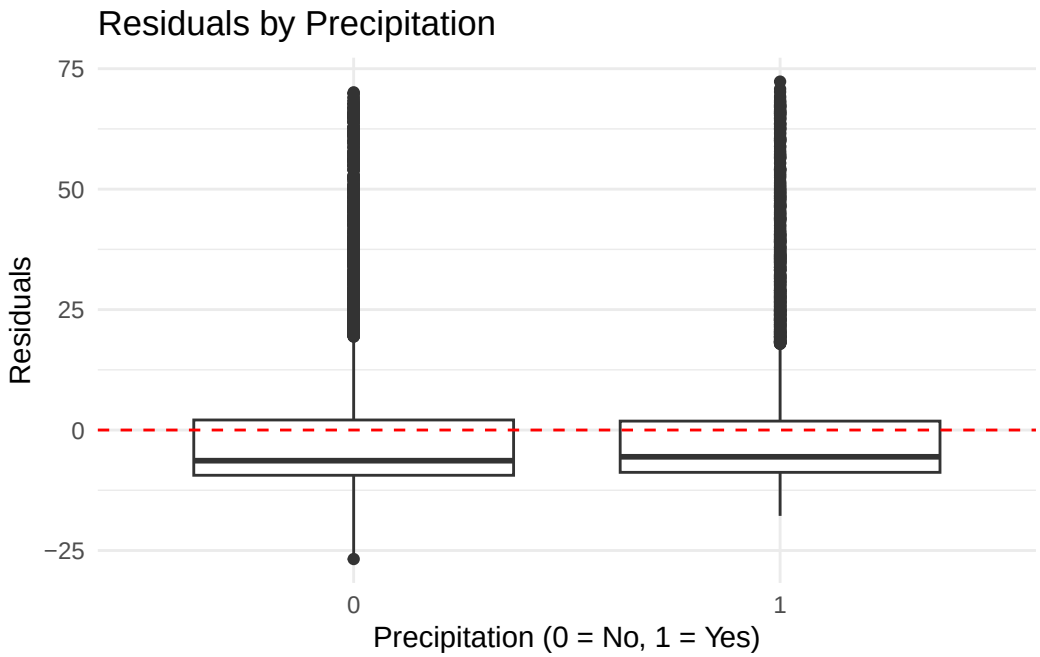

Residuals vs. Temperature
Residuals
75
50
25
0
−25
−50
−25
0
25
Temperature (Pre 15)


Residuals vs. Humidity
Residuals
75
50
25
0
−25
0
25
50
75
100
Humidity (Pre 15)

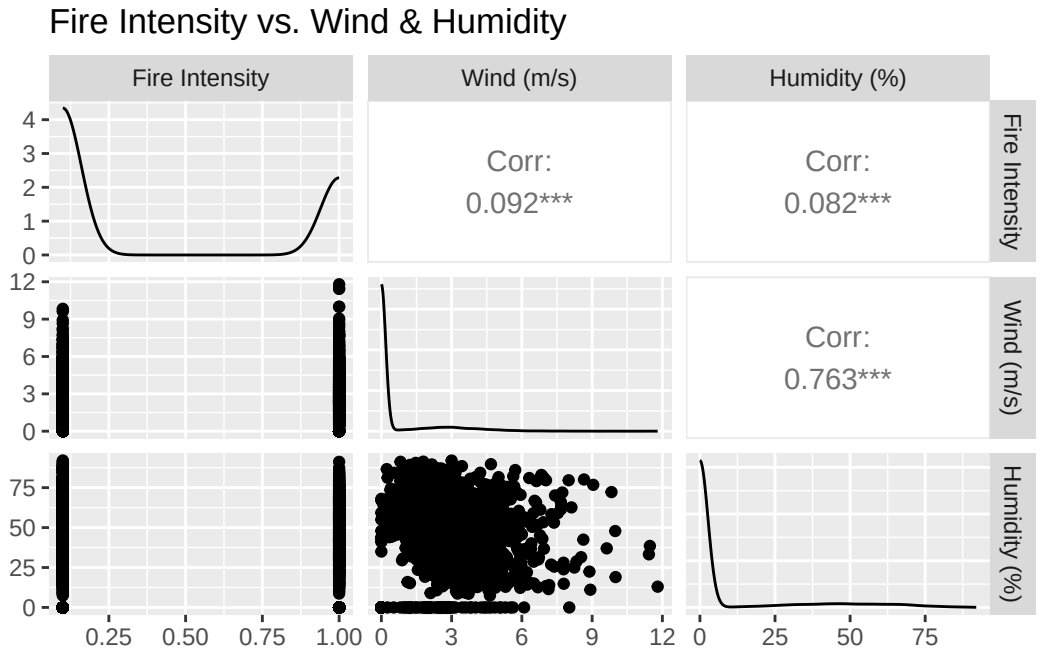

Residuals by Region
Residuals
75
50
25
0
−25
Midwest
Northeast
Other
South
West
Region

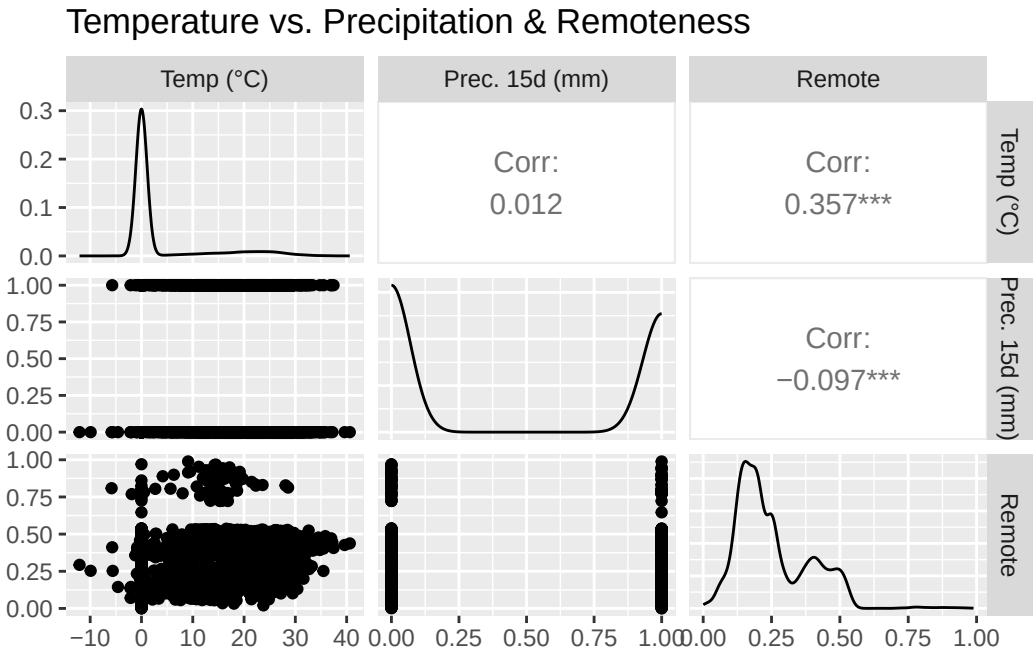

Residuals by Cause
75
50
25
0
−25
Residuals
Arson
Debris Burning
Infrastructure
Natural
Other/Unknown
Recreational
Cause Category


Residuals by Vegetation
75
50
25
0
−25
Residuals
Forest
Grassland
Shrubland
Tundra/Barren
Vegetation Group



Pairwise plots (if they add value)



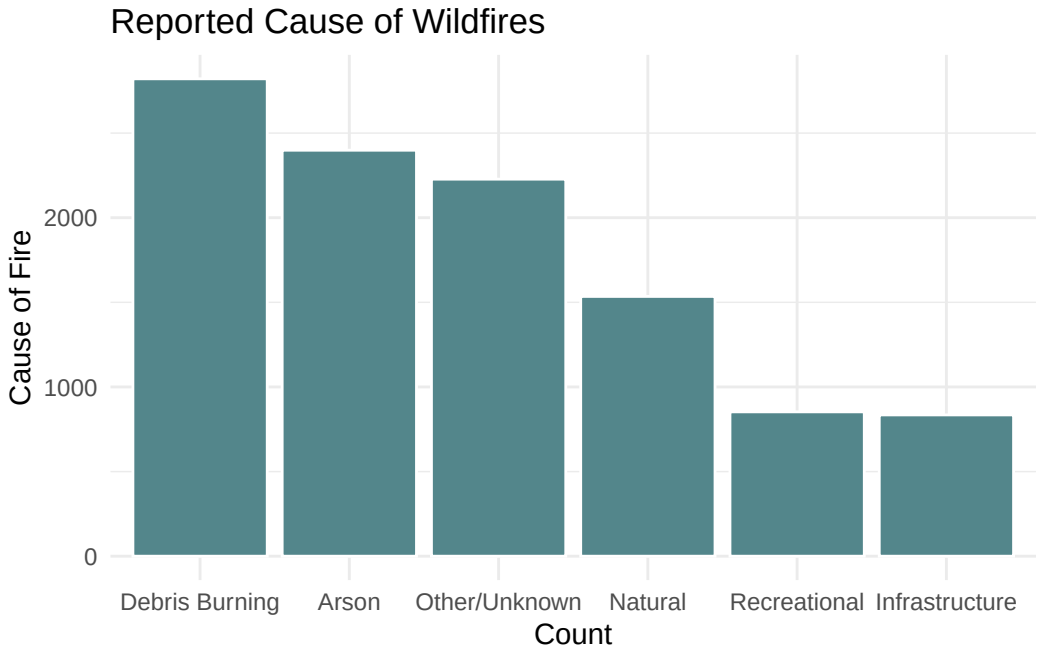


We began our analysis by creating pairwise plots to explore relationships between key numerical variables related to fire behavior, environmental conditions, and containment efforts. Since pairwise plots display scatterplots and correlations, we focused on continuous numerical variables that likely impact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

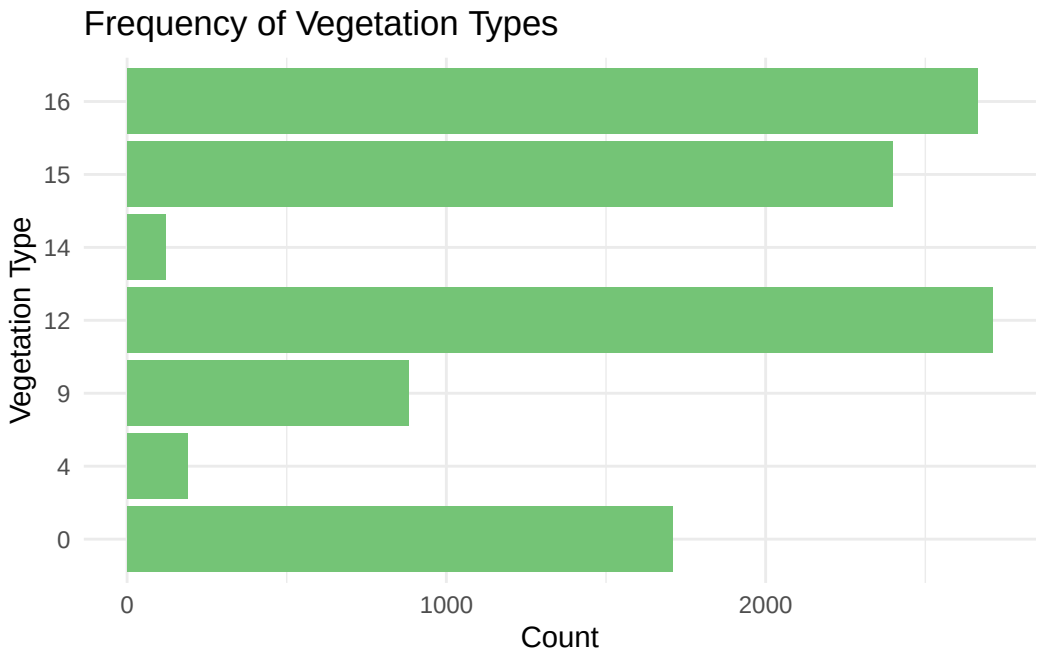
Next, to better understand the primary causes of wildfires in our dataset, we examined the stat_cause_descr variable, which provides the reported cause for each fire. The most frequently recorded cause was debris burning (2,821), followed by arson (2,399), miscellaneous causes (1,600), and lightning (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%)

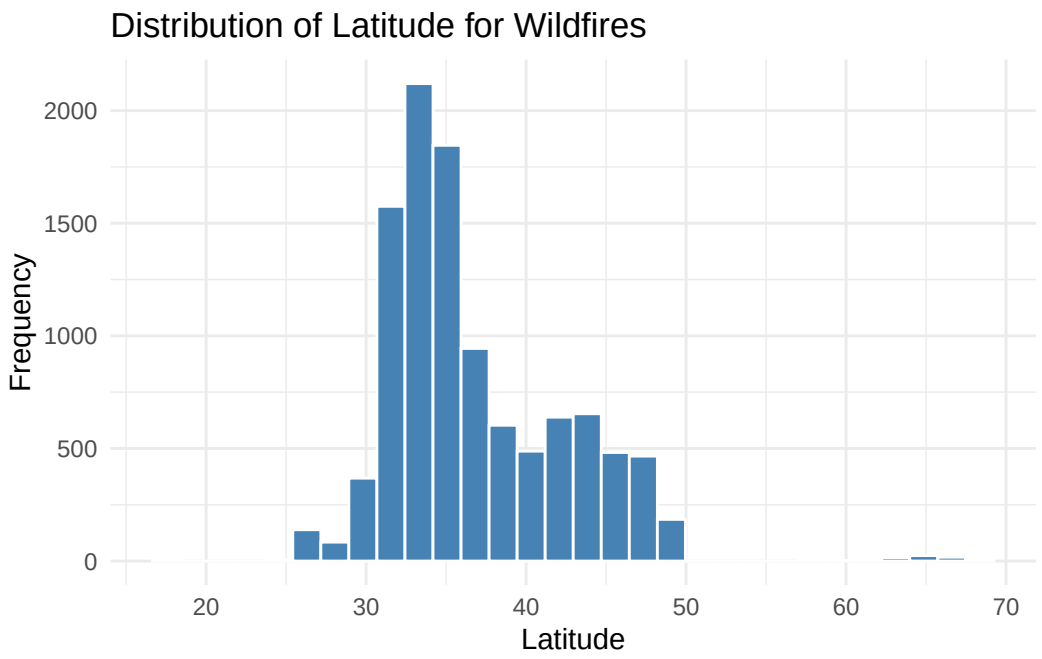
Less common vegetation types represented in the data include desert and temperate evergreen needleleaf forests.

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



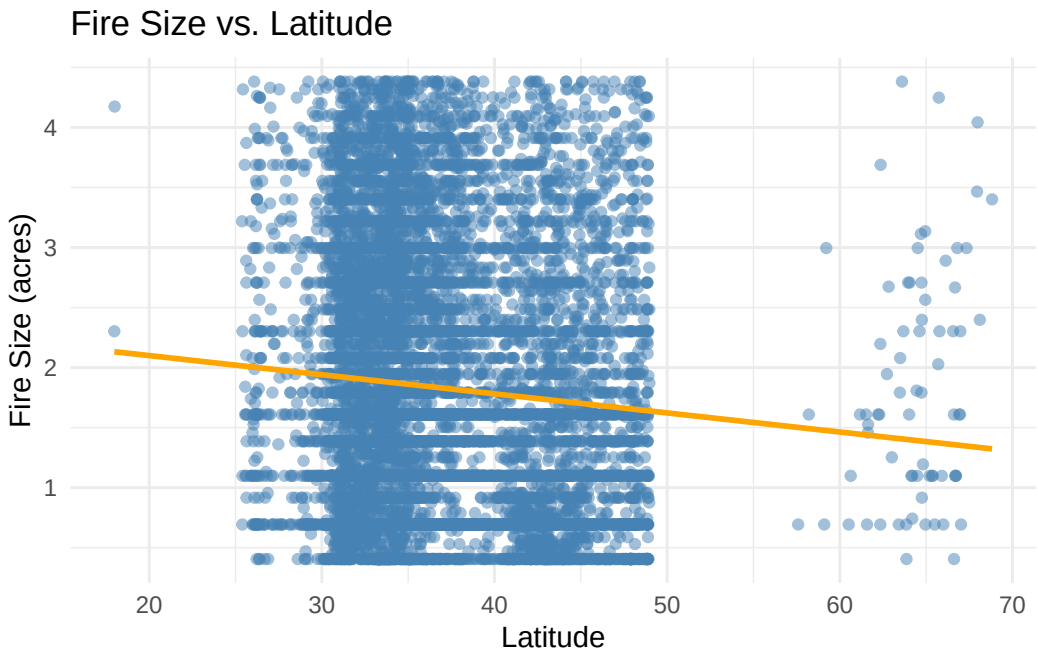

```
       0         4         9        12        14        15        16
16.005998  1.771155  8.265392 25.405304  1.124543 22.462750 24.964858
```

```
   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.

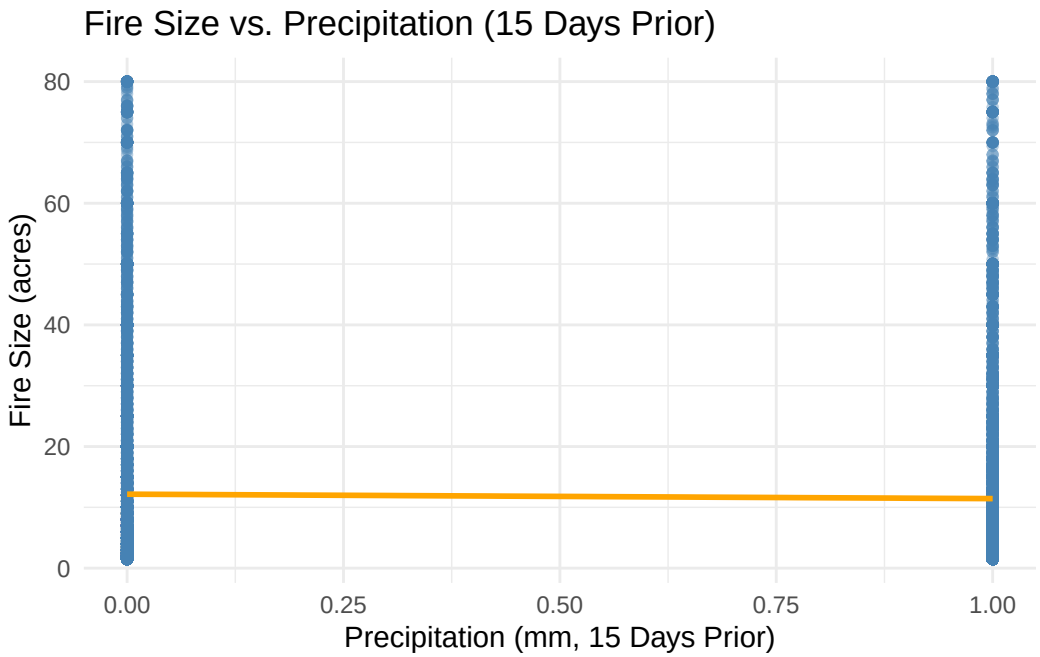
The distribution of latitudes appears to be centered around the mid-30s to upper-30s, which may reflect the concentration of fire-prone areas in those geographic zones.

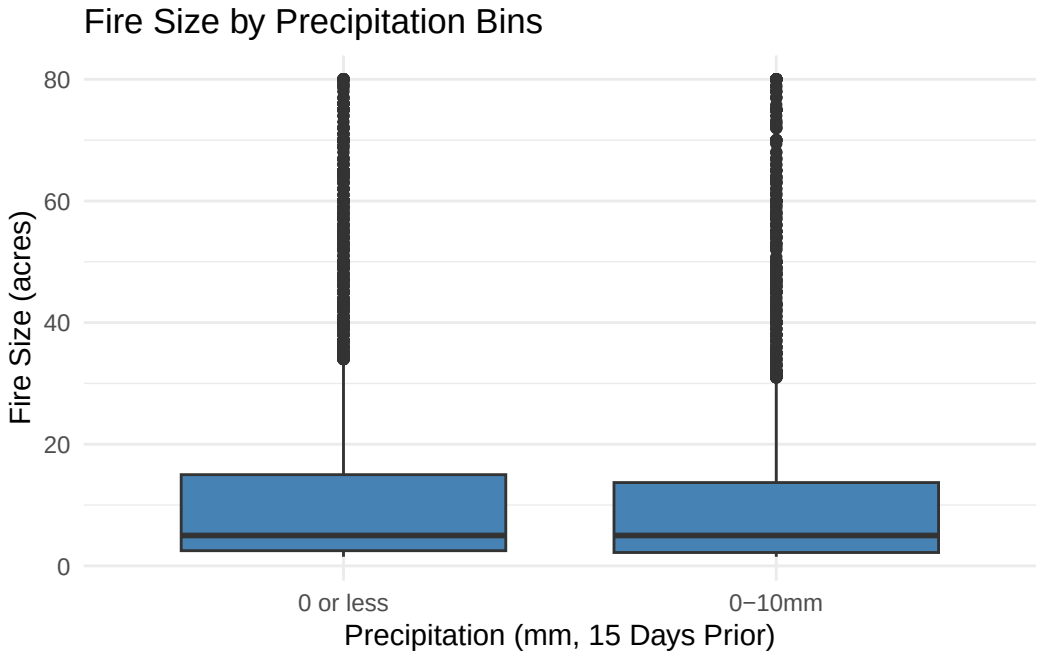



```
[1] -0.04720852
```

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.

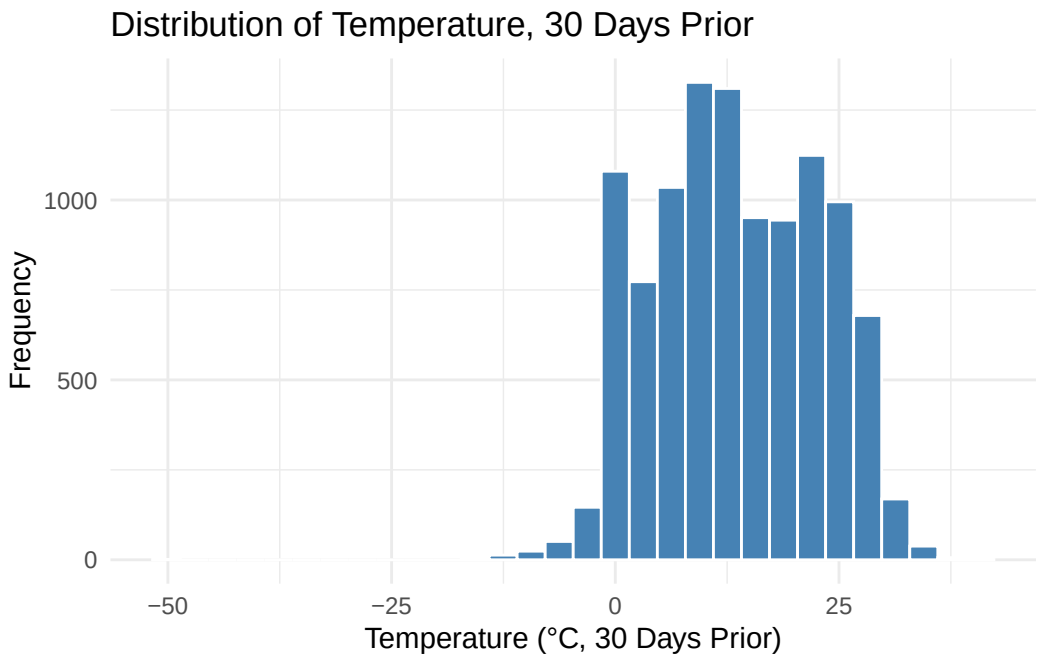




The scatterplot of `Prec_pre_15` and `fire_size` shows no strong relationship between prior precipitation and fire size. Most fires occurred with little to no precipitation, and there is substantial variability in fire size regardless of precipitation level. A few extreme precipitation values do not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

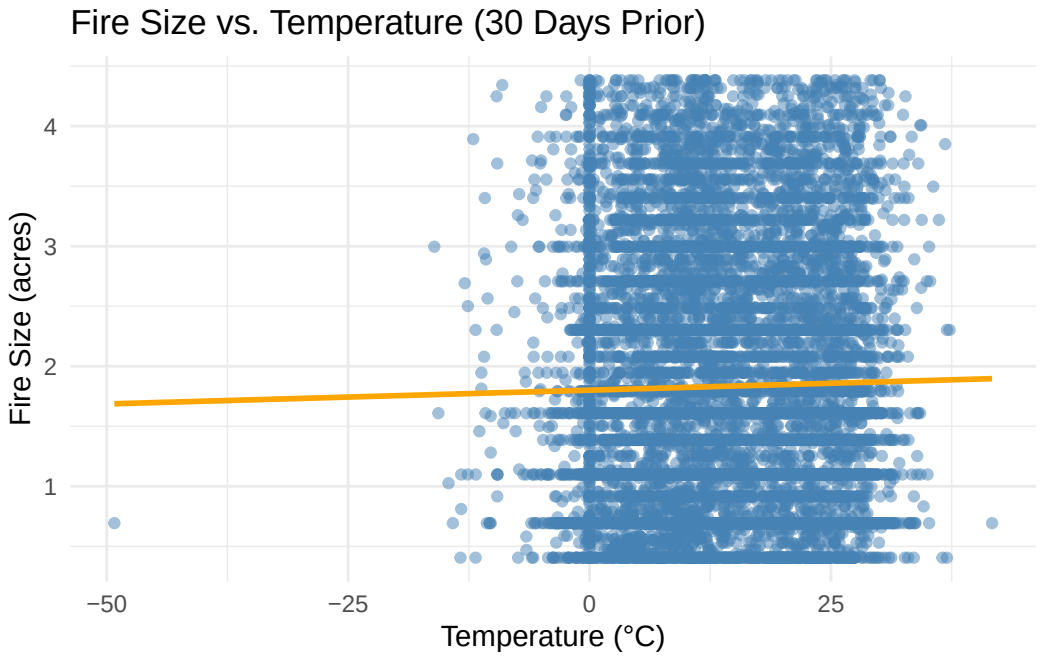


The distribution of Temp_pre_30, which represents the temperature at the location of a fire up to 30 days prior, shows a wide range of values, spanning from -49.211°C to 41.678°C. The minimum recorded temperature of -49.211°C is exceptionally low and may indicate an outlier or a potential data entry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

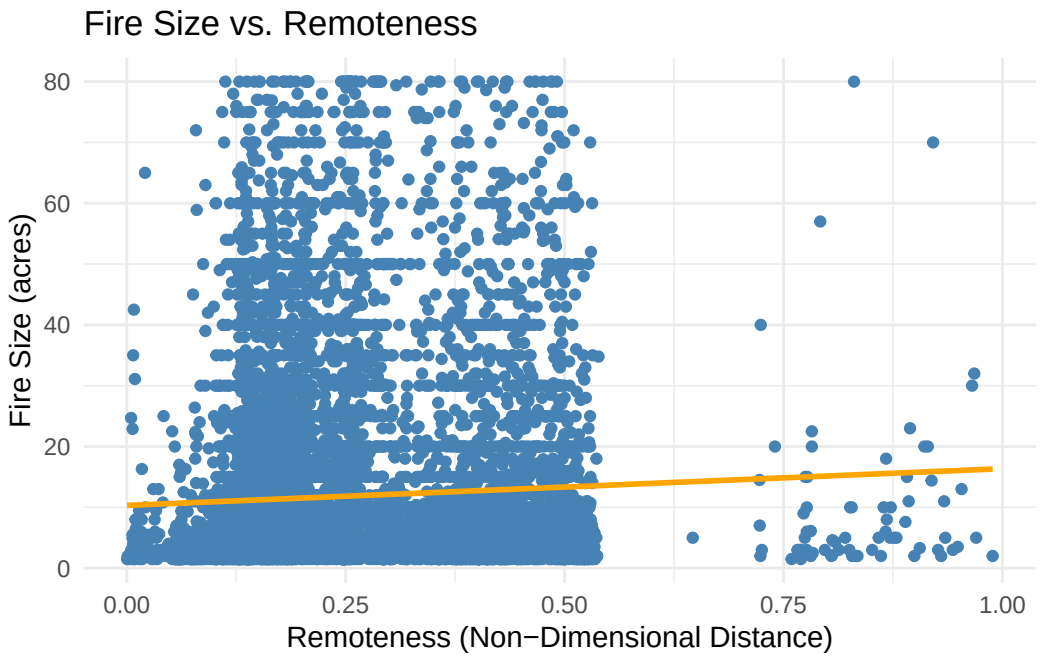
The median temperature of 8.310°C suggests that half of the recorded fires occurred in moderate conditions, while the mean of 9.562°C, which is slightly higher than the median, indicates that the distribution is right-skewed. This means that some extremely high-temperature values may be pulling the average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



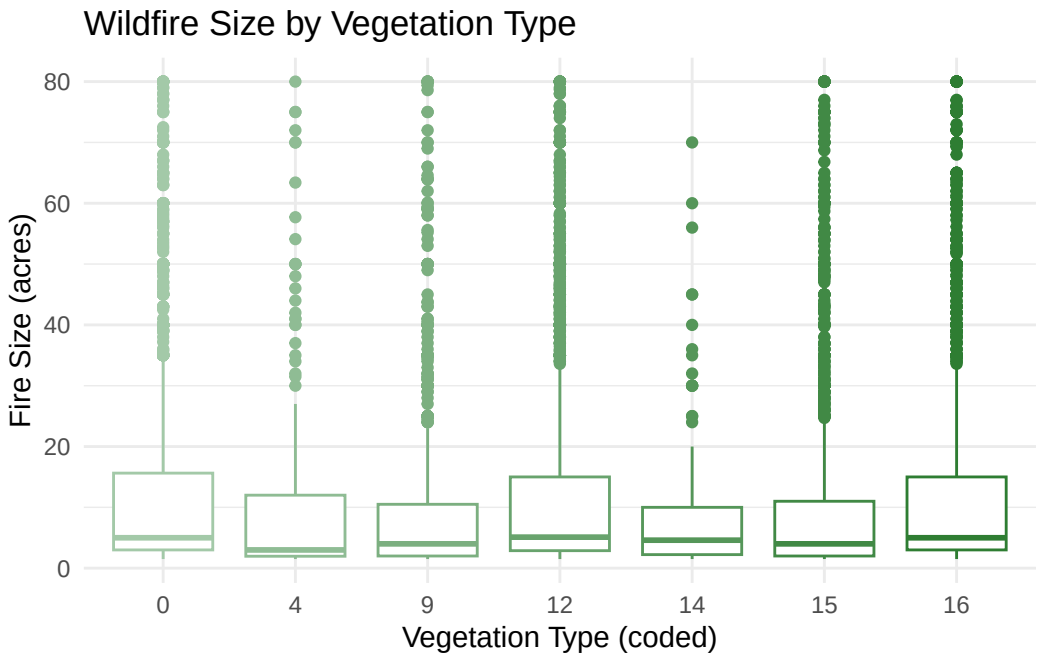
The scatter plot shows the relationship between fire size and temperature 30 days prior, with little evidence of a strong correlation. The nearly flat regression line suggests that temperature alone is not a key predictor of fire size. Most fires occurred between 0°C and 25°C, with fire sizes clustered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