W203_Lab1: EDA on Forest Fire Data

Beau Kramer, Nishant Hedge, Rory Liu, Stewart Warther

January, 2018

Introduction

This analysis is motivated by the following research question:

What factors lead to particularly damaging forest fires?

We aim to answer this this question using exploratory data analysis on the Forest Fires dataset, containing measurements taken of recent fires in a Portuguese park.

First, we load the data into R and also load the car package so we can use the scatter plot matrix feature.

```r
# Load the data and setup
options(warn = -1)
library(car)
f=read.csv("forestfires.csv")
```

Data Selection and Processing

We note that we have 517 observations and 13 variables.

We proceed to run the summary command on the full dataset, to get a snapshot view of all variables.

```r
# Summarize dataset
summary(f)
```

```
##        X               Y           month      day          FFMC      
##  Min.   :1.000   Min.   :2.0   aug    :184   fri:85   Min.   :18.70  
##  1st Qu.:3.000   1st Qu.:4.0   sep    :172   mon:74   1st Qu.:90.20  
##  Median :4.000   Median :4.0   mar    : 54   sat:84   Median :91.60  
##  Mean   :4.669   Mean   :4.3   jul    : 32   sun:95   Mean   :90.64  
##  3rd Qu.:7.000   3rd Qu.:5.0   feb    : 20   thu:61   3rd Qu.:92.90  
##  Max.   :9.000   Max.   :9.0   jun    : 17   tue:64   Max.   :96.20  
##                                (Other): 38   wed:54                  
##       DMC              DC             ISI              temp      
##  Min.   :  1.1   Min.   :  7.9   Min.   : 0.000   Min.   : 2.20  
##  1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500   1st Qu.:15.50  
##  Median :108.3   Median :664.2   Median : 8.400   Median :19.30  
##  Mean   :110.9   Mean   :547.9   Mean   : 9.022   Mean   :18.89  
##  3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800   3rd Qu.:22.80  
##  Max.   :291.3   Max.   :860.6   Max.   :56.100   Max.   :33.30  
##                                                                  
##        RH              wind            rain              area        
##  Min.   : 15.00   Min.   :0.400   Min.   :0.00000   Min.   :   0.00  
##  1st Qu.: 33.00   1st Qu.:2.700   1st Qu.:0.00000   1st Qu.:   0.00  
##  Median : 42.00   Median :4.000   Median :0.00000   Median :   0.52  
##  Mean   : 44.29   Mean   :4.018   Mean   :0.02166   Mean   :  12.85  
##  3rd Qu.: 53.00   3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.:   6.57  
##  Max.   :100.00   Max.   :9.400   Max.   :6.40000   Max.   :1090.84  
## 
```

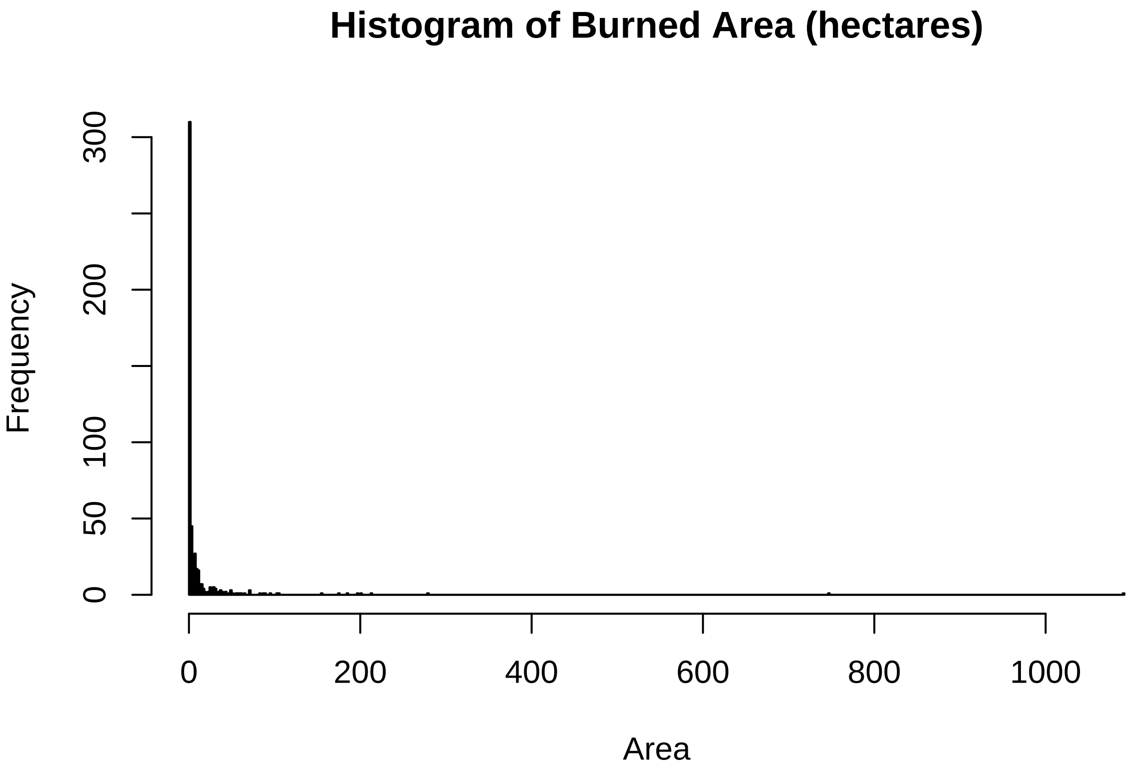
Through the summary, we see that besides the proxy for fire damage (i.e. "area"), the dataset includes some meteorological measures from the Fire Weather Index (e.g. "FFMC","DC","DMC", etc), some weather information (e.g. "rain","temp", "wind", "RH"), the month and day of week, and the spatial coordinates within the park. The dataset is rather clean. All variables are in a reasonable type and format, with no "NAs". However, there are a few things that caught our eyes in this initial summary and call for further data processing and cleaning:

1. The "month" and "day" variable were recorded as strings, which may be hard to sort and analyze. We will need to factorize these variables in our data processing.

```
# reformatting month and day
proper=function(x) paste0(toupper(substr(x, 1, 1)), tolower(substring(x, 2)))
f$month <- factor(proper(f$month),levels = month.abb)
f$day <- factor(f$day, levels= c("mon", "tue", "wed", "thu", "fri", "sat","sun"))
```

2. The variable "area", which we use as a proxy for fire damage, is largely skewed. We proceeded to look at a histogram of this variable, which revealed that almost half of the data entries actually had 0 as the recorded hectares of damage. This could either mean that there wasn't a fire on the particular day, or it could mean the fire was so small that it did cause any measurable damage. In either case, since our research question is to find factors behind "particularly damaging fires", these minimal fires may bring unncessary noise into the analysis. Therefore, it is useful to subset our data to exclude these 0-area fire entries. We could also make the assumption that these 0-area fires were controlled and put out early, that could have later turned out to be large and severe fires, however, since we do not have any data on fire containment and damage control we treat them as fires that were small or negligibly damaging.

```
# Distribution of fire damage area
hist(f$area, breaks="FD", main = "Histogram of Burned Area (hectares)", xlab="Area")
```



```
# Subsetting initial data to exclude 0-hectare fires
f_nz=f[f$area!=0,]
```

3. The “rain” variable is also highly skewed, with even the 3rd quartile being 0. We attempted to turn this into a binary rain or no-rain variable, and found that in fact, there were only 8 entries with rain above 0 (see below). Therefore, we need to exclude this variable from our relationship analysis since the result would not be statistically sound.

```
# binarize "rain"
rain_binary=cut(f$rain,breaks = c(-1,0,Inf),labels=c("No rain","Rained"))
summary(rain_binary)
```

```
## No rain  Rained
##     509       8
```

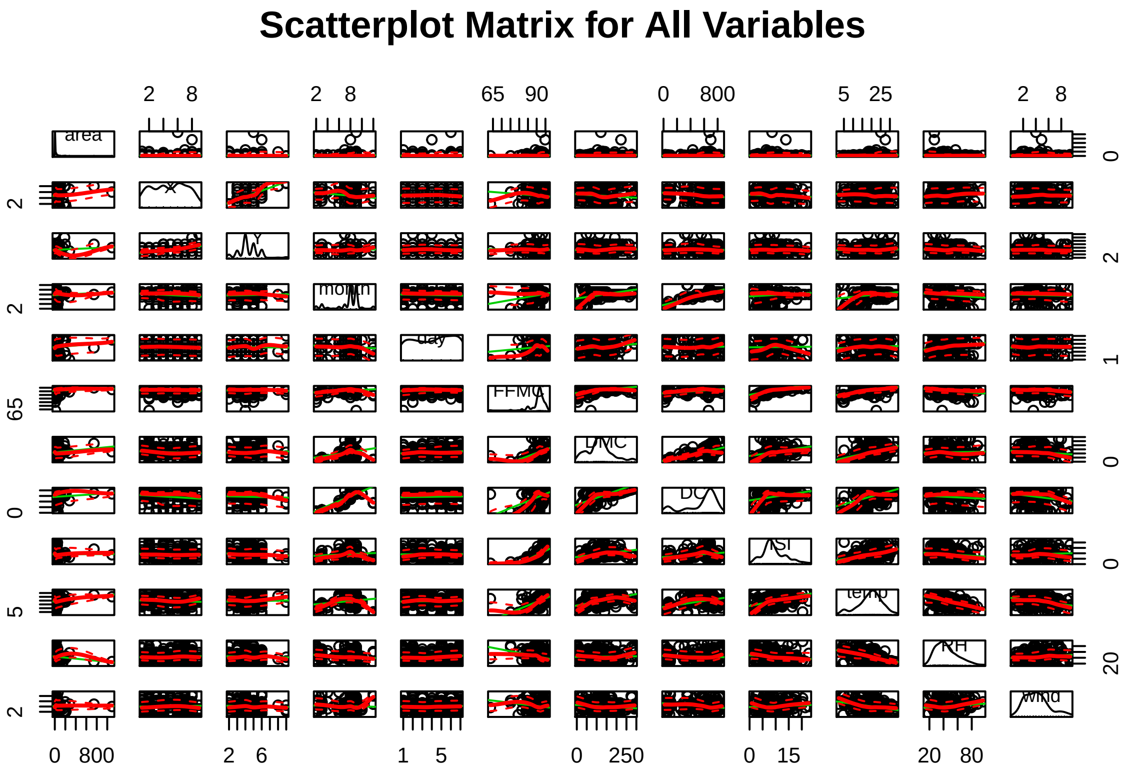
Data Summary

Now that we have completed our initial data processing and cleaning, we will look at the data summary again on our relevant data subset and create a scatterplot matrix. This, together with a correlation matrix will help us get an overview of relationships between variables.

```
# Subsetting initial data to exclude 0-hectare fires
summary(f_nz)
```

```
##        X               Y              month      day          FFMC      
##  Min.   :1.000   Min.   :2.000   Aug    :99   mon:39   Min.   :63.50  
##  1st Qu.:3.000   1st Qu.:4.000   Sep    :97   tue:36   1st Qu.:90.33  
##  Median :5.000   Median :4.000   Mar    :19   wed:32   Median :91.70  
##  Mean   :4.807   Mean   :4.367   Jul    :18   thu:31   Mean   :91.03  
##  3rd Qu.:7.000   3rd Qu.:5.000   Feb    :10   fri:43   3rd Qu.:92.97  
##  Max.   :9.000   Max.   :9.000   Dec    : 9   sat:42   Max.   :96.20  
##                                  (Other):18   sun:47                  
##       DMC              DC              ISI              temp      
##  Min.   :  3.2   Min.   : 15.3   Min.   : 0.800   Min.   : 2.20  
##  1st Qu.: 82.9   1st Qu.:486.5   1st Qu.: 6.800   1st Qu.:16.12  
##  Median :111.7   Median :665.6   Median : 8.400   Median :20.10  
##  Mean   :114.7   Mean   :570.9   Mean   : 9.177   Mean   :19.31  
##  3rd Qu.:141.3   3rd Qu.:721.3   3rd Qu.:11.375   3rd Qu.:23.40  
##  Max.   :291.3   Max.   :860.6   Max.   :22.700   Max.   :33.30  
##                                                                  
##        RH             wind            rain               area        
##  Min.   :15.00   Min.   :0.400   Min.   :0.00000   Min.   :   0.09  
##  1st Qu.:33.00   1st Qu.:2.700   1st Qu.:0.00000   1st Qu.:   2.14  
##  Median :41.00   Median :4.000   Median :0.00000   Median :   6.37  
##  Mean   :43.73   Mean   :4.113   Mean   :0.02889   Mean   :  24.60  
##  3rd Qu.:53.00   3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.:  15.42  
##  Max.   :96.00   Max.   :9.400   Max.   :6.40000   Max.   :1090.84  
## 
```

```
scatterplotMatrix(~ area + X + Y + month + day + FFMC + DMC + DC + ISI + temp + RH + wind, data=f_nz, ma
```



```
cor(f_nz[ ,c("area", "RH", "wind", "temp", "ISI", "DC", "DMC", "FFMC")])
```

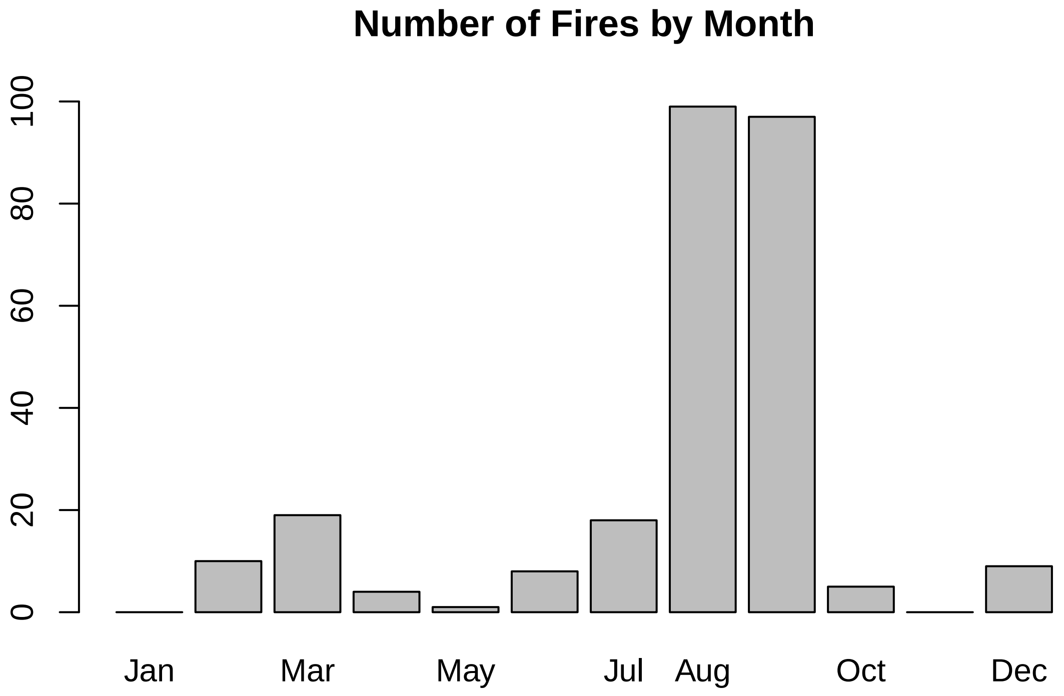
```
##              area          RH         wind       temp         ISI
## area  1.000000000 -0.10484626  0.002085841  0.1102929  0.002121065
## RH   -0.104846261  1.00000000  0.138489957 -0.4975479 -0.148803620
## wind  0.002085841  0.13848996  1.000000000 -0.3205632  0.072651791
## temp  0.110292941 -0.49754787 -0.320563247  1.0000000  0.466027115
## ISI   0.002121065 -0.14880362  0.072651791  0.4660271  1.000000000
## DC    0.046734570 -0.08221711 -0.237593410  0.4957027  0.256826065
## DMC   0.089088179  0.02786140 -0.137897089  0.5016432  0.329656141
## FFMC  0.054323454 -0.28599043 -0.161384446  0.5622557  0.704170493
##              DC         DMC        FFMC
## area  0.04673457  0.08908818  0.05432345
## RH   -0.08221711  0.02786140 -0.28599043
## wind -0.23759341 -0.13789709 -0.16138445
## temp  0.49570270  0.50164319  0.56225571
## ISI   0.25682607  0.32965614  0.70417049
## DC    1.00000000  0.66892587  0.40763825
## DMC   0.66892587  1.00000000  0.48025000
## FFMC  0.40763825  0.48025000  1.00000000
```

From the scatterplot matrix, we derive the following preliminary findings:

1. The X and Y location showed strong correlation between X and Y coordinates. This suggests that there is possibly a diagnoal “fire prone” zone that is in this park.

2. The data shows seasonality. Most of the fires happened during the summer months. The months of August and September had the highest frequency of fires and also had a higher number of large fires. The weather conditions during these months might be conducive to severe fires.

```
counts <- table(f_nz$month)
barplot(counts, main = "Number of Fires by Month", ylim=c(0,100),cex.lab=0.8)
```



3. No weekly effect was seen in the data.
4. The correlations between fire damage and the various weather indices are as strong as we have hoped. The damage area seems to be positively correlated with all the indices (FFMC, DMC, DC, ISI) but not to a very significant degree. In our relationship analysis we will dive in and see if these impact become more apparent after we group the fires into various severity levels.
5. We do see some correlation among the various weather indices and the weather variables. This is expected because many weather indices are based on common sets of raw weather data. We need to be careful when interpreting our results in our relationship analysis stage.

Univariate Analysis of Key Variables

From the preliminary conclusions drawn from the scatterplot matrix, we narrow down the list of variables for univariate analysis to the following: 1) area, 2) drought code, 3) temperature, and 4) relative humidity.

Area

Beginning with area, we plot a histogram (zeros removed) to see that there is a long tail to this distribution with the majority of fires being categorized as small.To better assess key relationships later, we bin the area variable based on the quartiles of data after removing the area values equal to zero. We bin these zero values into their own "zero" bin as they consitute roughly half the dataset.

Before splitting out the zero fires the mean area is 12.847 with a standard deviation of 63.656. After removing the zero fires, the mean area is 24.6 with a standard deviation of 86.502.

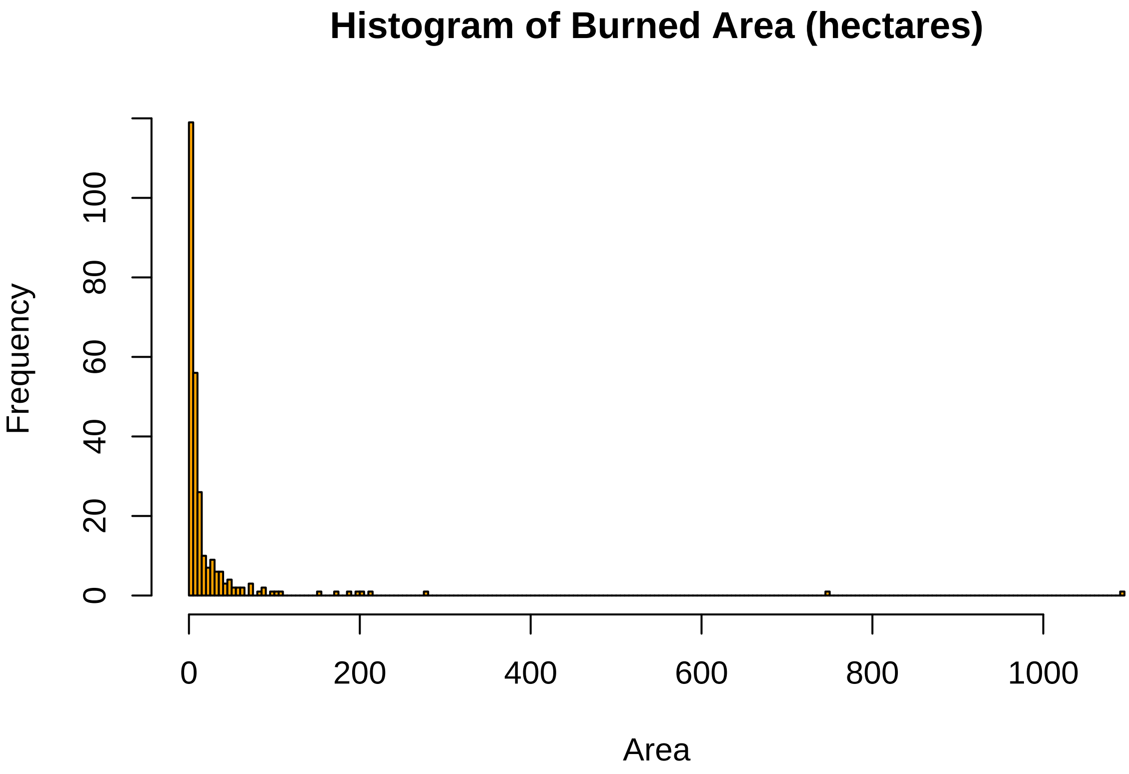
```
## Analysis of Area
summary(f_nz$area)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.09    2.14    6.37   24.60   15.42 1090.84
```

```
quantile(f_nz$area,c(0,0.1,0.25,0.3,0.4,0.5,0.6,0.75,0.8,0.9,1))
```

```
##        0%       10%       25%       30%       40%       50%       60% 
##    0.0900    1.0050    2.1400    2.6190    4.1300    6.3700    8.2680 
##       75%       80%       90%      100% 
##   15.4225   24.3100   46.8850 1090.8400
```

```
hist(f_nz$area, breaks="FD", col="orange", xlab="Area", main = "Histogram of Burned Area (hectares)")
```



```
area_bin=cut(f$area,breaks = c(-1,0,15.4,46.9,84.8,Inf),
             labels=c("zero","small","medium","large","very large"))

summary(area_bin)
```

```
##        zero       small      medium       large very large 
##         247         202          41          13          14
```

Drought Code

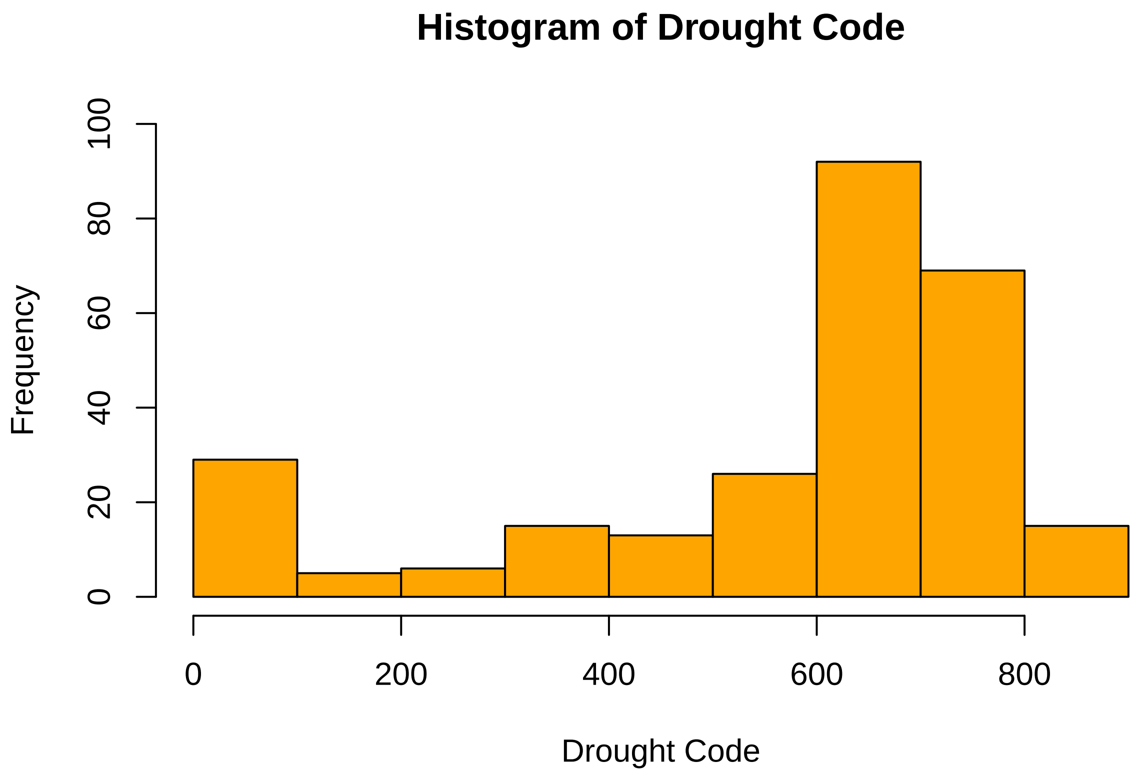
Next, we look at the drought code (DC) variable using a histogram. We note the following key observations:

1. DC data exhibits a slight negative skew with a heavy left tail indicating periods with high moisture content.
2. The shape of the histogram gives the impression of an almost a bimodal distribution. Frequency spikes between 0 and 100 then tail off until reaching 300 then rising again until spiking between 600 and 800. We will examine this further and the potential relationship with area burned in the Key Relationships section.

```
## Analysis of Drought Code (DC)
summary(f_nz$DC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    15.3   486.5   665.6   570.9   721.3   860.6
```

```
hist(f_nz$DC, breaks="FD", col="orange", xlab="Drought Code", main="Histogram of Drought Code", ylim=c(
```



Temperature

We observe the Temperature variable and notice the following: 1. Visually the histrogram appears to have minimal skew.

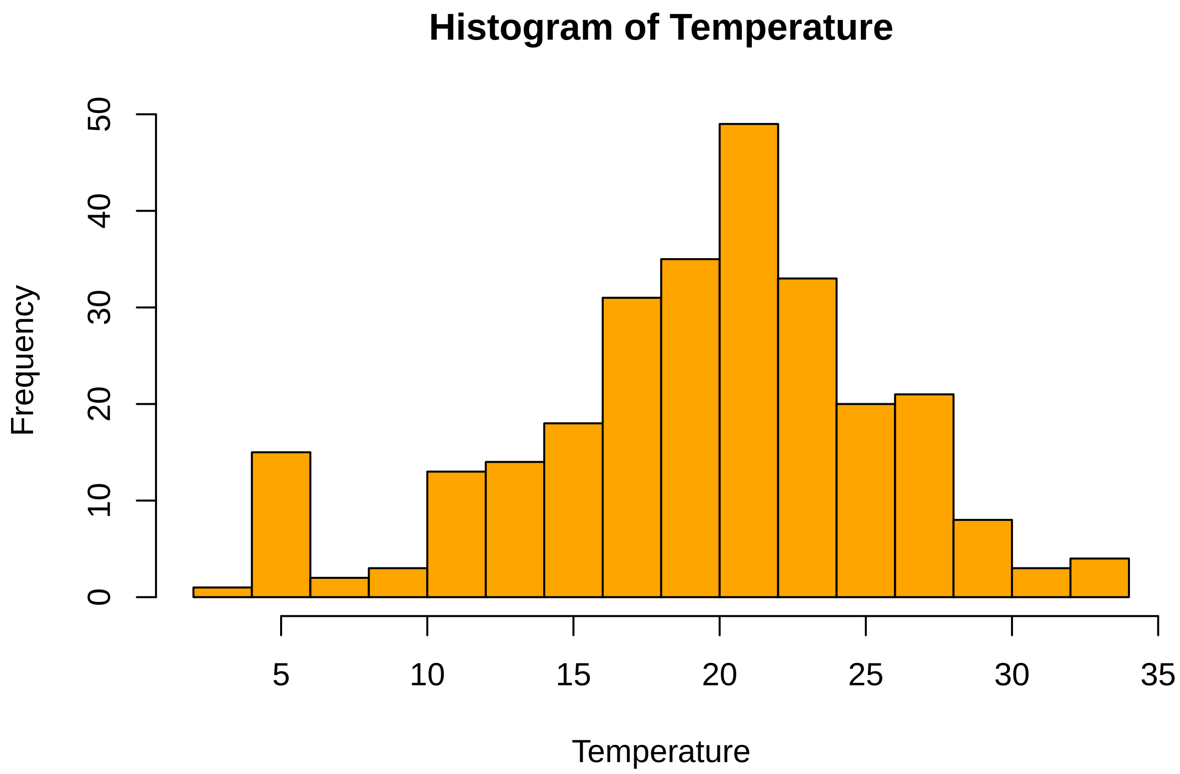
2. The data itself has an interesting range as this is park has a climate that at times is near freezing and other times is sweltering.

3. There is an interesting concentration of temperatures around the 5 degree level. This could indicate certain weather conditions that cause fires at very low temperatures. Low temps (temp < 10) seem to be negatively correlated with wind with a correlation coefficient of -0.1649. High winds at low temps might ceate conditions leading to fires.

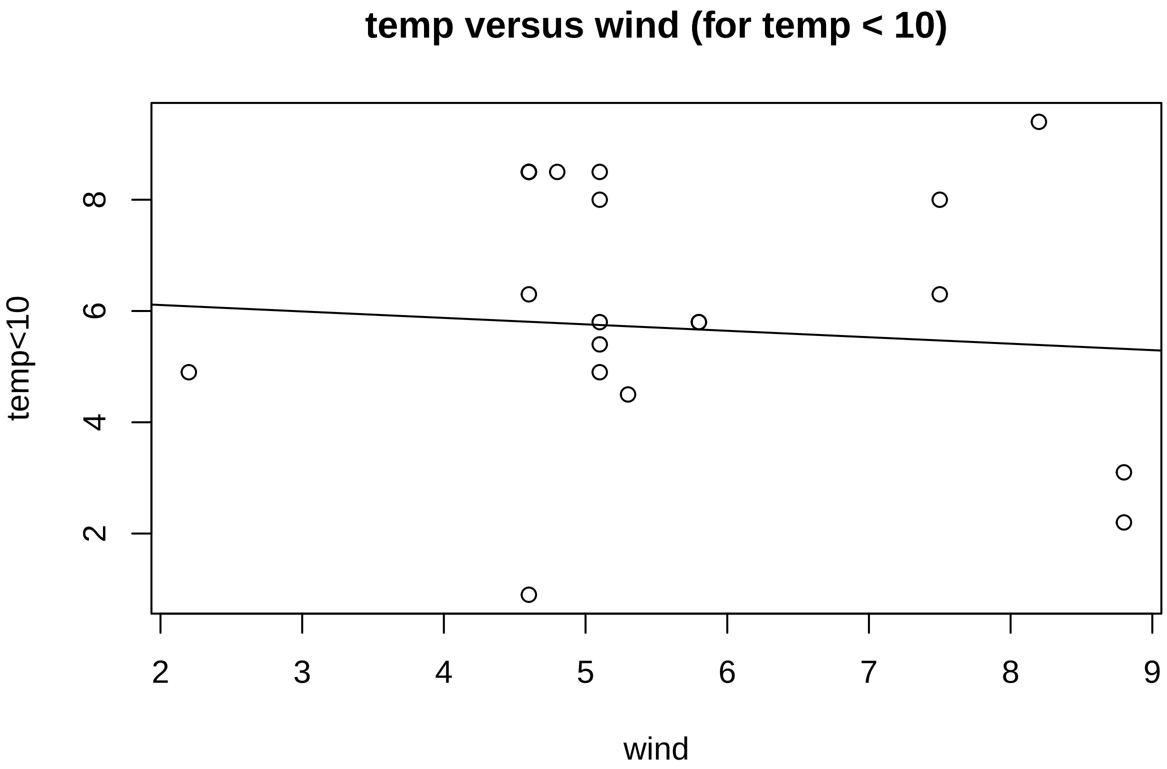
```
## Analysis on Temperature
summary(f_nz$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.20   16.12   20.10   19.31   23.40   33.30
```

```
hist(f_nz$temp, breaks="FD", col="orange", xlab="Temperature", main ="Histogram of Temperature")
```



```
## Low Temperature vs Wind
plot(f_nz[f_nz$temp<10,]$temp, f_nz[f_nz$temp<10,]$wind, xlab = "wind", ylab = "temp<10",
     main = "temp versus wind (for temp < 10)")
abline(lm(f_nz[f_nz$temp<10,]$temp ~ f_nz[f_nz$temp<10,]$wind))
```



Relative Humidity

Finally, we observe the Relative Humidity variable and note the following:

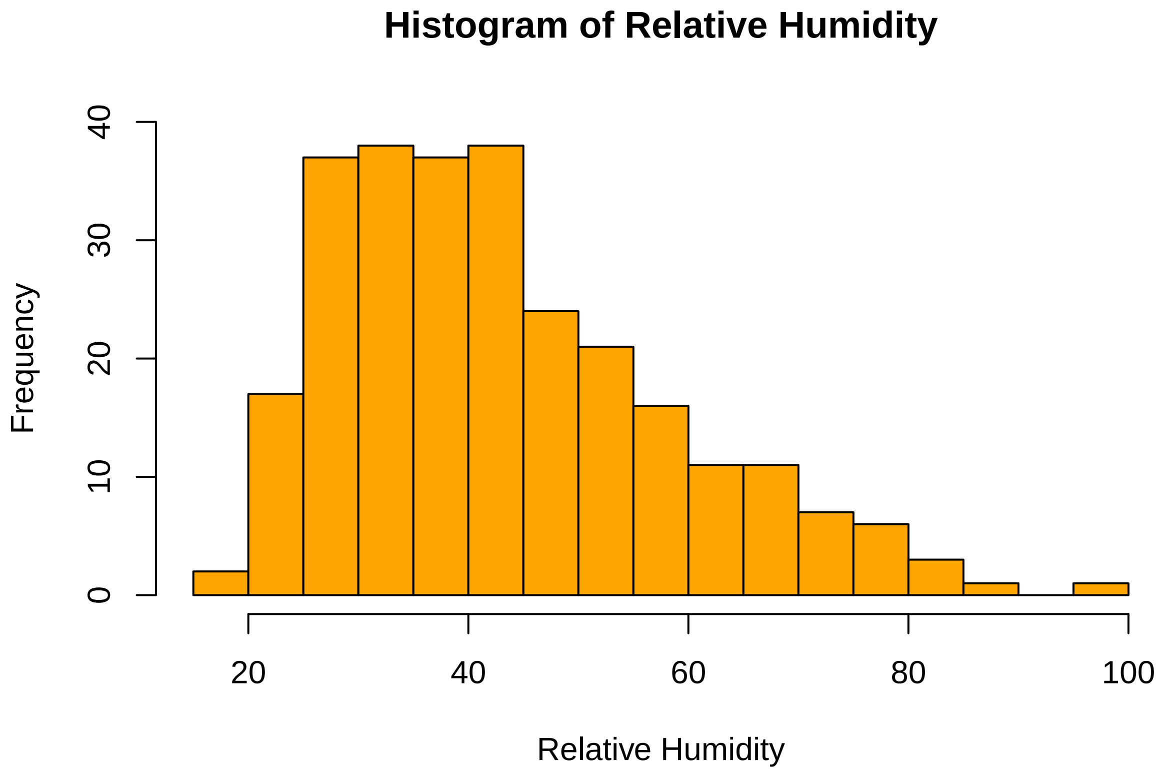
1) Visually there appears to be a positive skew to the with most of the measurements clustered between 25 and 45 then falling off toward 100. This could be due to the absence of the zero fire data. Fires may be more likely at lower levels of humidity. This could explain the skew.

2) The histogram has a high peak between 25 and 45 then sharply drops off. This could suggest a threshold that greatly increases the likelihood or severity of fire below a certain level of humidity.

```
## Analysis of Relative Humidity (RH)
summary(f_nz$RH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   33.00   41.00   43.73   53.00   96.00
```

```
hist(f_nz$RH, breaks="FD", col="orange", xlab="Relative Humidity", main="Histogram of Relative Humidity
```



Analysis of Key Relationships

Next we will examine how these variables relate to the area burned variable which is the focus of this EDA.

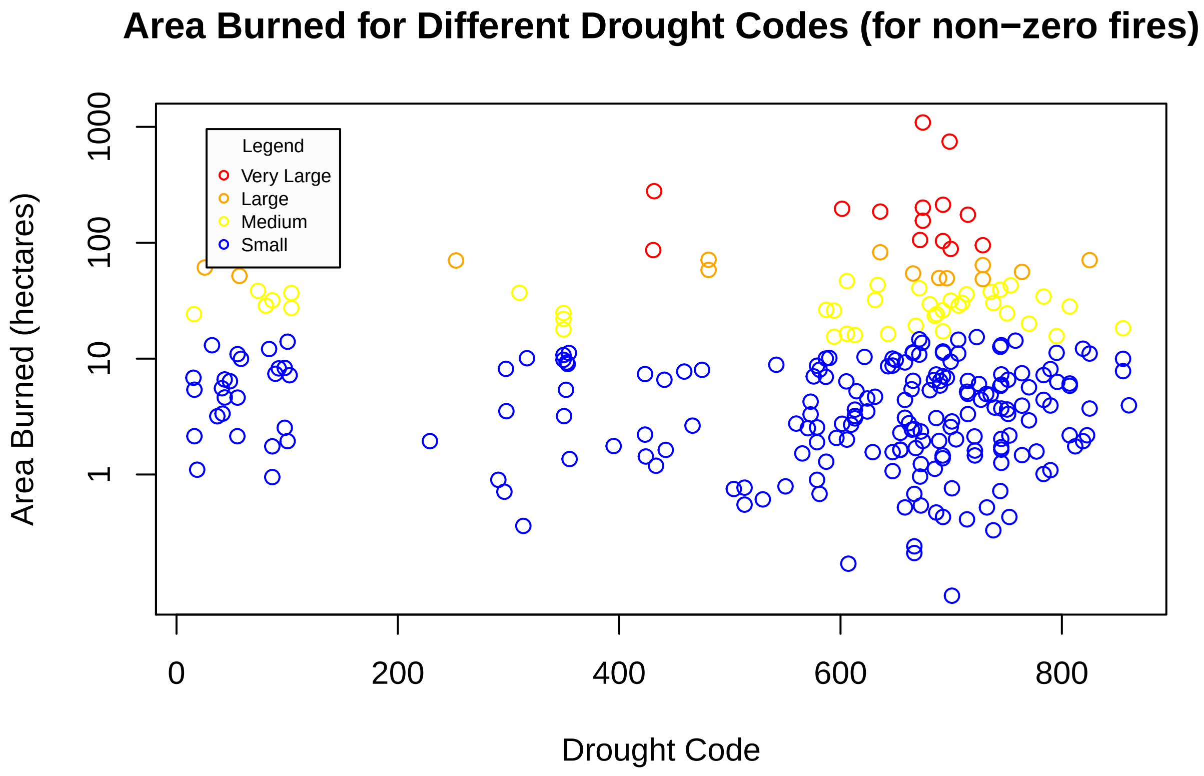
Drought Code

First we will examine the Drought Code (DC). The Drought code captures the moisture content of the deepest organic layers in the ground. We first plot a scatterplot. We segment the data by fire size and color code this. We also use a logarithmic scale on the y axis for a better view.

```r
# Area Burned vs Drought Code (color)
plot(f_nz$DC, f_nz$area,
    xlab = "Drought Code",
    ylab = "Area Burned (hectares)",
    main = "Area Burned for Different Drought Codes (for non-zero fires)",
    col = ifelse(f_nz$area > quantile(f_nz$area, 0.95), "red",
                 ifelse(f_nz$area > quantile(f_nz$area, 0.9),"orange",
                        ifelse(f_nz$area > quantile(f_nz$area, 0.75),"yellow","blue"))),
    log="y",
    yaxt="n")

axis(side = 2,
     at = c(0, 1, 10, 100,1000),
     labels = c(0, 1, 10, 100,1000))
```

```
legend("topleft",
       inset=.05,
       cex = 0.6,
       title="Legend",
       c("Very Large","Large","Medium","Small"),
       horiz=FALSE,
       pch = c(1,1,1,1),
       col=c("red","orange","yellow", "blue"),
       bg="grey99")
```



Observations:

1) Clustering of fires between about 600 and 800.

This may suggest that the existence of a fire is more likely in these ranges. This would be a potential point to examine in a statisical analysis.

2) Bigger fires cluster between 400 and 800

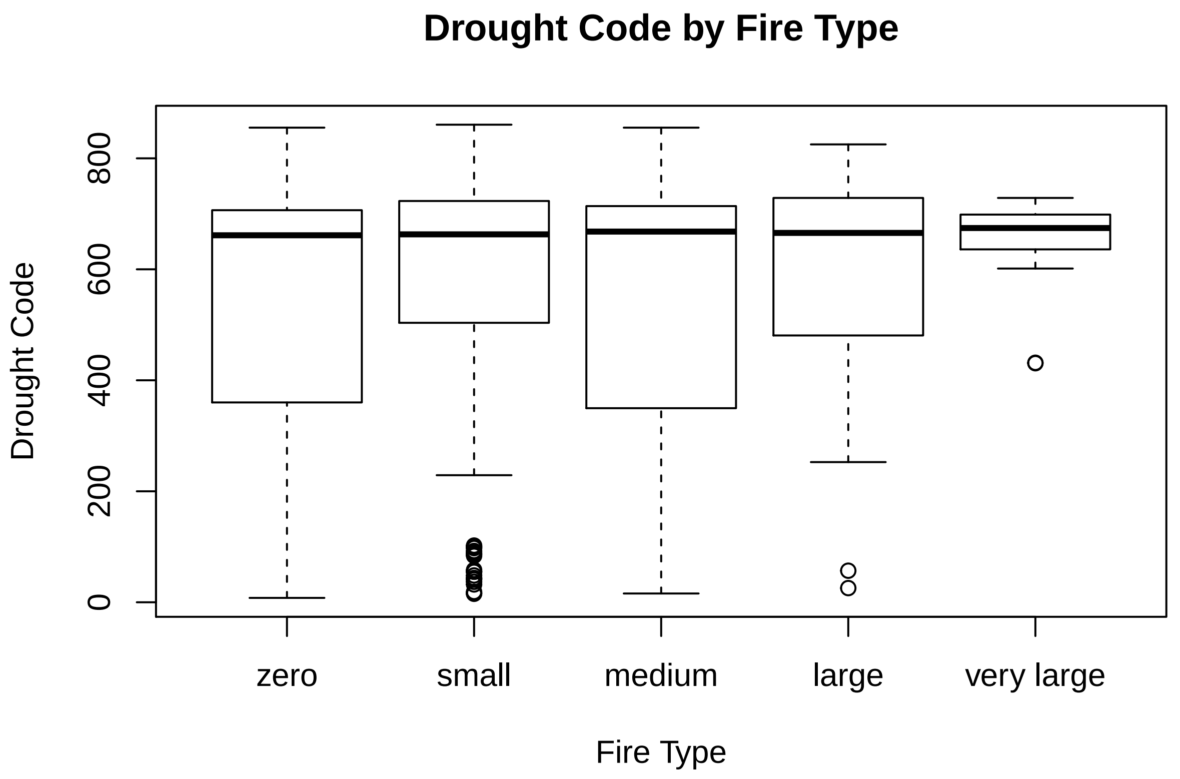
With a few exceptions, medium, large, and very large fires all occur between a Drought Code of about 400 and 800. The higher the Drought Code the dryer the deep layers of oragnic material so although this is not a causal analysis this is an expected logical relationship.

3) Fires at low Drought Codes

Fires of medium and large size do occur with low drought codes at left end of plot. These are puzzling observations. It could suggest that fires of these size can occur regardless of Drought Code. There may be other factors driving these observations. For example, the dryness of shallower layers: the Duff Moisture Code (DMC) or Fine Fuel Moisture Code (FFMC). However, we do note that no very large fires exist with these low Drought Code observations. This could suggest that the Drought Code acts as a sort of limit on the area burned. Medium or large fires could become very large ones were it not that the deep organic layers (the Drought Code) so moist.

To gain a more detailed view, we will use a boxplot.

```r
boxplot(DC ~ area_bin, data = f,
        main = "Drought Code by Fire Type",
        xlab = "Fire Type", ylab = "Drought Code")
```



Observations:

1) Many small and medium fires at low drought codes.

We note the presence of many small and medium fires at very low Drought Codes. As noted above, these smaller fires may be related to dryness at shallower layers of soil. This is interesting, especially when compared to the absence of larger fires at these low drought codes.

2) Higher and tighter ranges for large and very large fires.

It is interesting to note the wide range of Drought Codes for small and medium fires. The range for very large fires is comparatively tight and high. This could possibly be a function of the number of very large fires (14). Perhaps more data would show them having a similar range to other fires. This does, however, reaffirm our earlier observations in the scatterplot. Although medium and smaller sized fires may occur roughly regardless of Drought Code, very large fires seem to be limited to a narrower range of higher Drought Codes.

The Drought Code, measuring the moisture of a thick, deep layer of organic material, represents magnitudes more fuel than in shallower layers. The lack of existence of very large fires with low Drought Codes makes sense because if the deep layers of organic material are moist then potential fires have much less fuel to burn. So low Drought Codes could make it unlikely for a small fire to escalate into a very large one. Simply, it does not have as much fuel to burn. Convesrely, it may be that high drought codes create the *potential* for very large fires. For a group concerned about particularly damaging fires, the Drought Code may be an important indicator of the potential for very large fires. However, the relationship here does not seems clearly linear and may require a more complicated model to capture.

Next we will examine how temperature relates to area burned.

Temperature

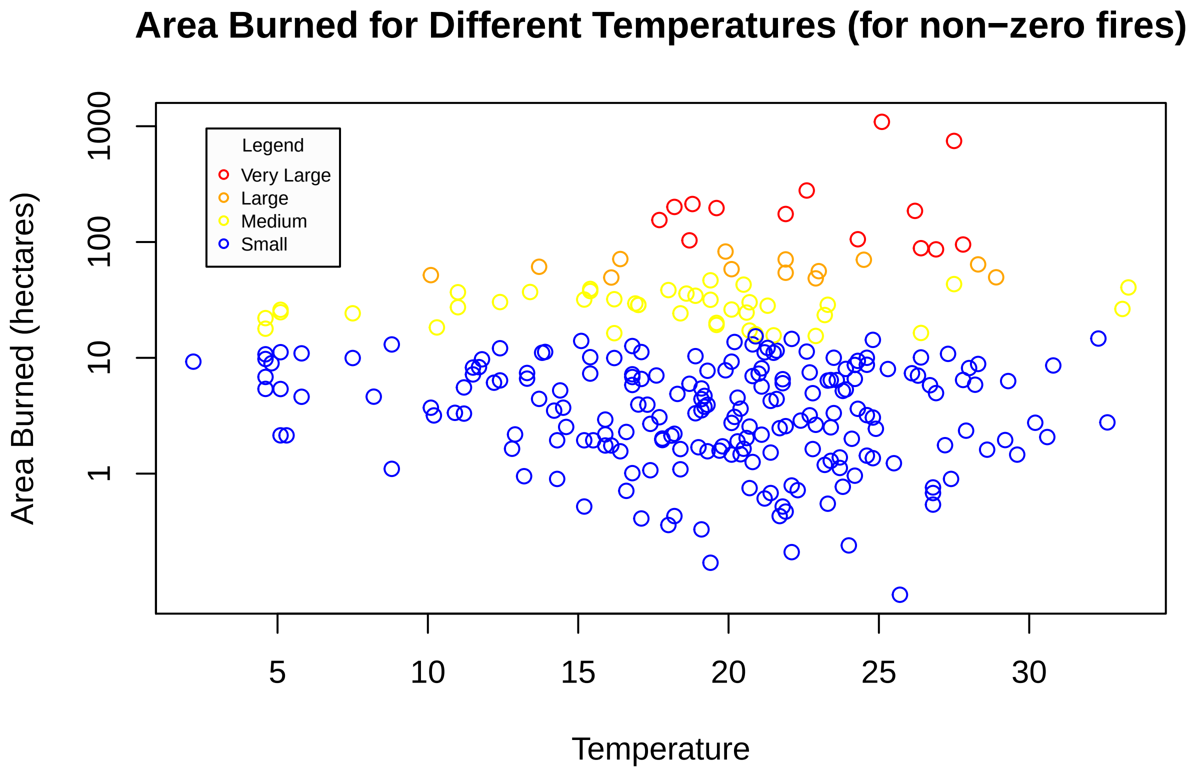
We first plot a scatterplot. Again, we segment the data by fire size and color code this. We also use a logarithmic scale on the y axis for a better view.

```
# Area vs Temperature

plot(f_nz$temp, f_nz$area,
    xlab = "Temperature",
    ylab = "Area Burned (hectares)",
    main = "Area Burned for Different Temperatures (for non-zero fires)",
       col = ifelse(f_nz$area > quantile(f_nz$area, 0.95), "red",
                 ifelse(f_nz$area > quantile(f_nz$area, 0.9),"orange",
                        ifelse(f_nz$area > quantile(f_nz$area, 0.75),"yellow","blue"))),
    log="y",
    yaxt="n")

axis(side = 2,
     at = c(0, 1, 10, 100,1000),
     labels = c(0, 1, 10, 100,1000))

legend("topleft",
       inset=.05,
       cex = 0.6,
       title="Legend",
       c("Very Large","Large","Medium","Small"),
       horiz=FALSE,
       pch = c(1,1,1,1),
       col=c("red","orange","yellow", "blue"),
       bg="grey99")
```



Observations:

1) Fires occur across wide range of temperatures.

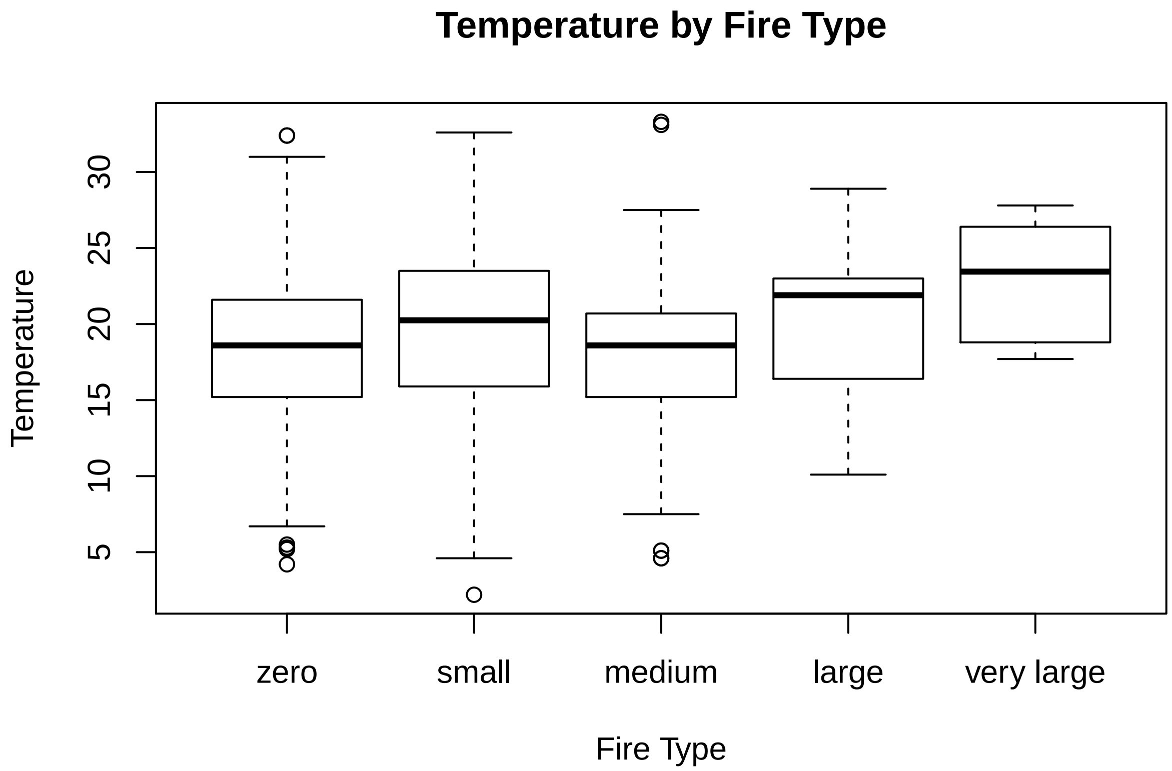
It is somewhat surprising to see the existence of fires at relatively cold temperatures. Although the main body of fires is grouped toward higher temperatures, it is worth noting that some fires do occur at cold temperatures like 5 degrees and below.

2) Very large fires cluster between 17 and 30.

Similar to the Drought Code relationship, we notice that very large fires are clustered within a range of higher temperatures. Here between about 17 and 30 degrees.

We will also examine this data in a boxplot to try to identify any other insights.

```
boxplot(temp ~ area_bin, data = f,
        main = "Temperature by Fire Type",
        xlab = "Fire Type", ylab = "Temperature")
```



Observations:

1) Tightening stairstep as severity increases.

We notice that the boxplots tighten and move upward as the severity of the fire increases. This tight stairstep pattern seems to confirm that there is potentially a linear relationship between tempearture and area burned.

2) Wide range of temperatures for small fires.

Small fires seem to occur regardless of temperature. There is even a fire occuring near 0 degrees! The wide range of small fires suggests that although the presence of a fire may not depend much on the temperature, its severity does.

The relationship between Temperature and Area seems more linear. Intuitively this makes sense. Higher temperatures dry out fuel at various depths creating conditions conducive to larger fires. However, the direct causal link may not be as clear. There could be a number of factors confounding this relationship, some not even captured in the dataset. Perhaps higher temperatures draw more visitors to the park who may cause fires by accident. We can examine this and other potential confouding effects in our analysis of secondary effects.

Relative Humidity

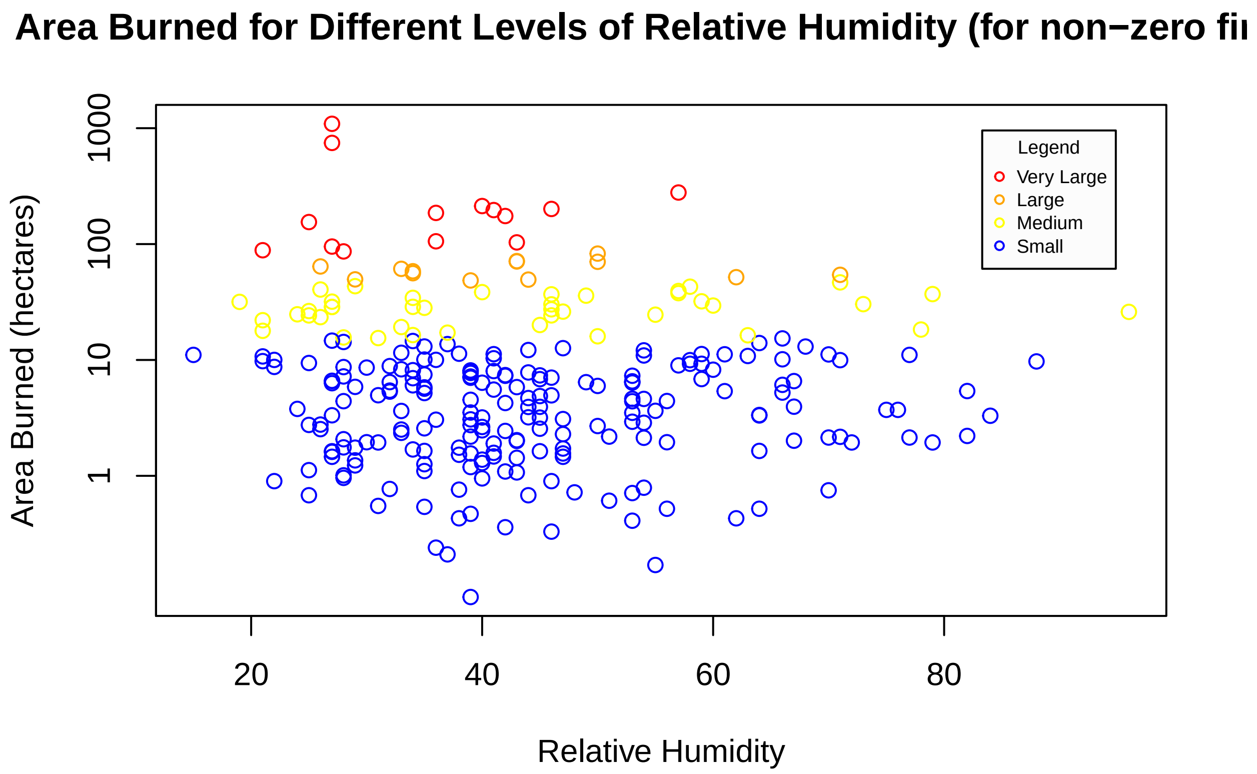
Finally, we examine relative humidity. We start with a scatterplot color coded by fire size and log scaled.

```
# Area vs Relative Humidity
```

```r
plot(f_nz$RH, f_nz$area,
    xlab = "Relative Humidity",
    ylab = "Area Burned (hectares)",
    main = "Area Burned for Different Levels of Relative Humidity (for non-zero fires)",
       col = ifelse(f_nz$area > quantile(f_nz$area, 0.95), "red",
                 ifelse(f_nz$area > quantile(f_nz$area, 0.9),"orange",
                        ifelse(f_nz$area > quantile(f_nz$area, 0.75),"yellow","blue"))),
    log="y",
    yaxt="n")

axis(side = 2,
     at = c(0, 1, 10, 100,1000),
     labels = c(0, 1, 10, 100,1000))

legend("topright",
       inset=.05,
       cex = 0.6,
       title="Legend",
       c("Very Large","Large","Medium","Small"),
       horiz=FALSE,
       pch = c(1,1,1,1),
       col=c("red","orange","yellow", "blue"),
       bg="grey99")
```



Observations:

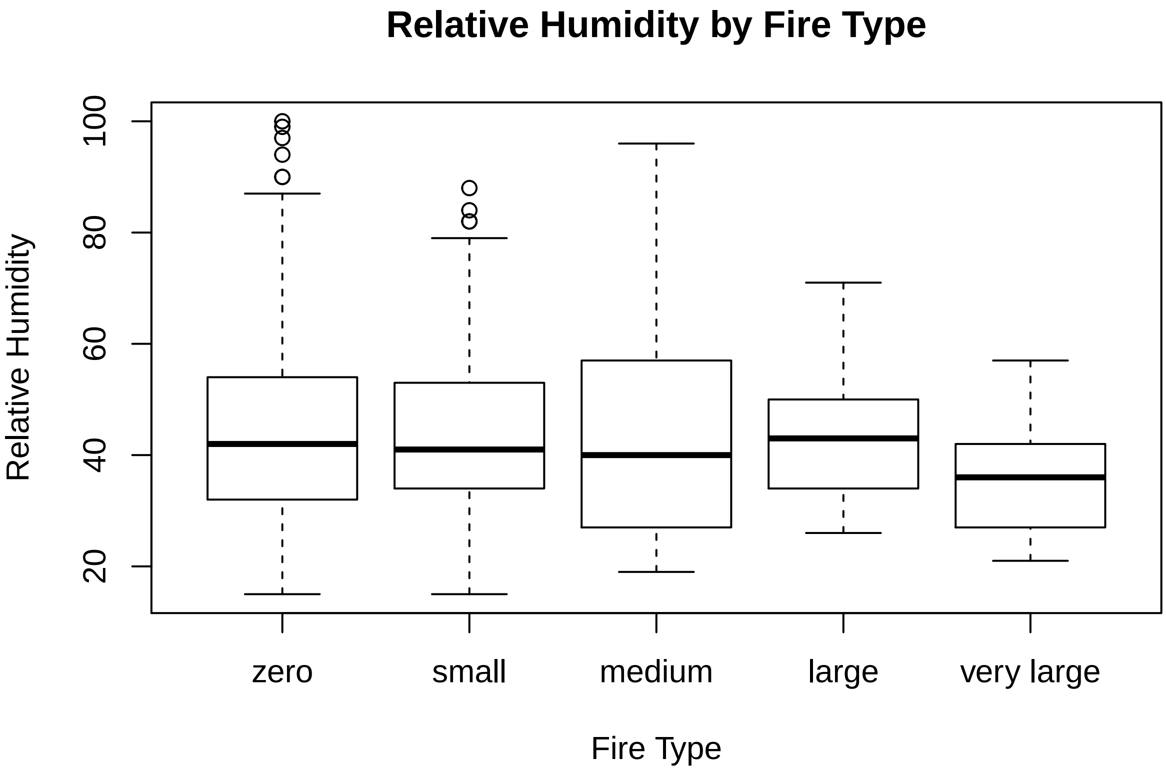
1) Grouping of fires below 60.

We note that most of the data exists below a relative humidity level of 60. This may suggest some key threshold for an increased likelihood of a fire starting or it could be a typical humidity level for this region.

2) Severe fires group toward lower levels of humidity.

We observe that large and very large fires tend to group in lower levels of relative humidity with a few exceptions. This seems like a potential a relationship between relative humidity and area burned. The boxplot will help confirm this.

```r
boxplot(RH ~ area_bin, data = f,
        main = "Relative Humidity by Fire Type",
        xlab = "Fire Type", ylab = "Relative Humidity")
```



Observations:

1) Tightening stairstep as relative humidity decreases.

Similar to temperature, we notice that the boxplots tighten and fall as relative humidity declines. This pattern suggests that a relationship may exist between relative humidity and area burned.

The relationship between relative humidity and area burned seems somewhat linear. Again, this makes some intuitive sense. Low humidity means less water vapor in the air potentially capturing periods where the ground is dry. These dry conditions could lead to very large fires or increase the likelihood that small fires escalate. So while the causal link remains unclear, relative humidity does show a noticeable relationship to area burned.

Analysis of Secondary Effects

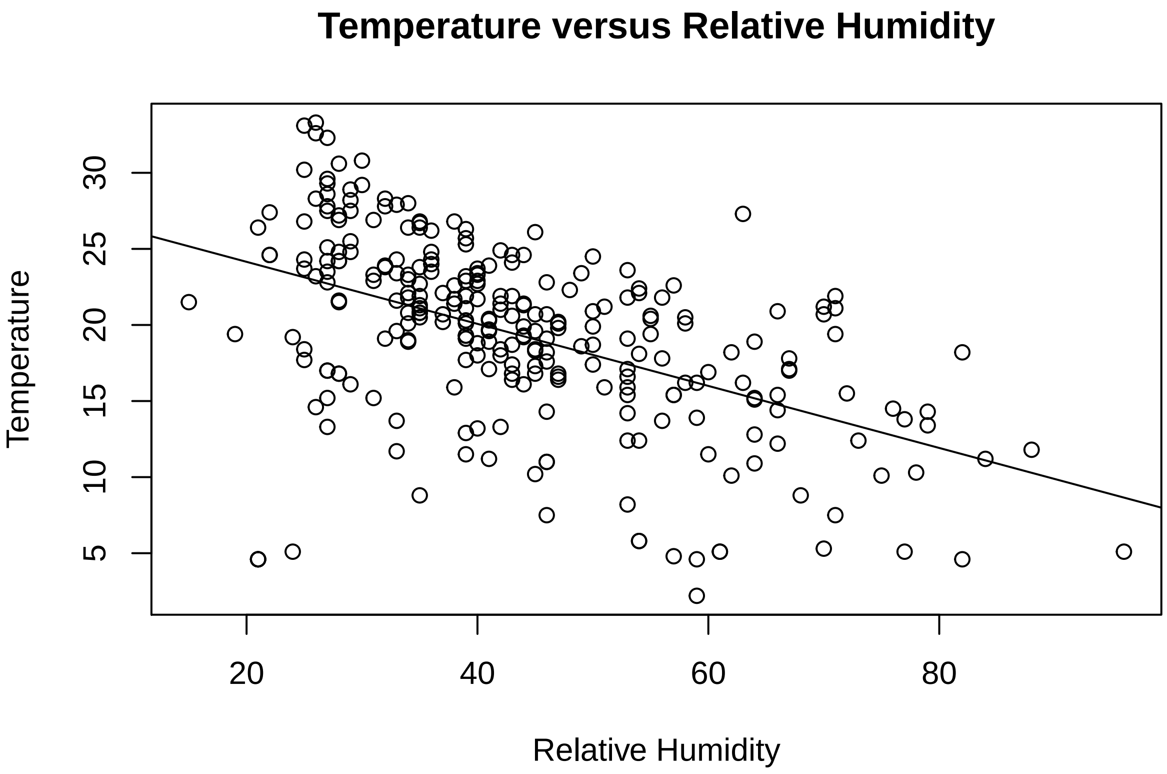
Based on the motivation behind the research question and the variables shortlisted in the previous sections, we can think of area as our output variable. Drought Code, Temperature, and Relative Humidity can be thought of as our input variables. We are not going to be doing any statistical modelling, however, we feel that our analysis needs to be evaluated in this context to assess whether there are secondary effects. In this section we will discuss if any other variables, included in our dataset, might be confounding our analysis.

1. We notice that Relative Humidity (RH) is negatively correlated with Temperature. This could mean that the higher temperatures are related to lower relative humidity. If we base a model that predicts the likelihood of a large fire based solely on temperature, we might be simply capturing dry periods with low relative humidity. This could be driving the relationship and not temperature itself.

```
cor(f_nz$RH, f_nz$temp)
```

```
## [1] -0.4975479
```

```
plot(f_nz$RH, f_nz$temp, xlab = "Relative Humidity", ylab = "Temperature",
     main = "Temperature versus Relative Humidity")
abline(lm(f_nz$temp ~ f_nz$RH))
```

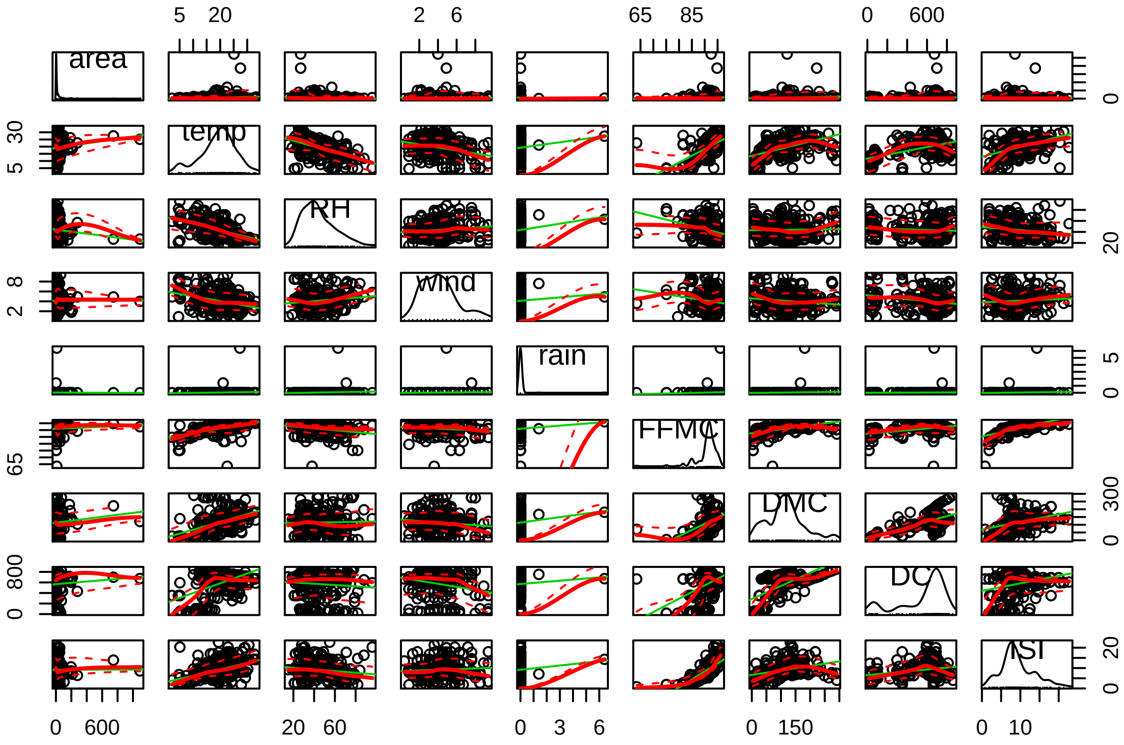


2. We notice that all of our dataset has very little rainfall. This could mean that this particular area recevies very low rainfall throughout the year. However, it is worth noting that the data contains measurements of recent fires and no information was given on time frame used for the data collection. If the data was collected from only one year, and that year happened to have an unusually low amount of rainfall, we might be biasing our conclusions drawn from this analysis.

3. The four Fire Weather Index variables (FFMC, DMC, DC and ISI) are positively correlated with each other and also show a positive relationship with area burned. This shows that all four variables use

different ways of rating fire danger and they use the same measurement direction in the rating i.e. a higher measurement denotes higher danger. They are also positively correlated with temperature, and do not have much correlation with other weather conditions. This shows that temperature is probably a factor in the mesurement calculation, but wind and RH do not seem to be playing an obvious part. If we had to create a model that predicted fire danger we would include only one of these 4 varaiables, so that there is independence between the predictors.

```
options(warn=-1)
scatterplotMatrix( ~ area + temp + RH + wind + rain + FFMC + DMC + DC + ISI, data = f_nz)
```



4. Relative Humidity seems to be the least affected by secondary effects from the FWI index variables. If we had to create a statsitical model to predict damaging fires we would consider this variable as a candidate for a predictor.

Conclusions

We believe that a combination of low relative humidity and high temperatures can lead to conditions that have a high risk for severe fires. To summarize our results, we think that Temperature, Drought Code, and Relative Humidity are variables of interest and our analysis of the sample shows them being correlated with damaging fires. Even though the correlation between Duff Mositure Code (DMC) and Area was a little higher than the correlation between Drought Code (DC) and Area, Drought Code was chosen since the relationship looked less prone to secondary effects.

Based on data collected in this sample, it seems that the months of August and Sepetember have weather conditions most likely to cause severe fires. High temperatures in these summer months, together with low relative humidity, can create conditions conducive to large fires. The diagonal “fire prone” region (based on the X and Y spatial coordinates) in the park denotes that there might be certain conditions in the soil or with the vegetation in that region that might allow fires to spread quickly.

As previously stated, there are a couple of items we would like to further investigate to make claims on what factors might be leading to severe fires.

1. Data on containment and damage control could inform us on how to treat fires with small or zero area i.e. should they be treated as fires that did not burn or are they fires that would have caused severe damage had they not been contained early.

2. Low rainfall in our data might be indicative of a selection bias with our sample and we suggest obtaining the time frame for when this was collected.