w203 Lab 1: Forest Fires

Dani Salah, Peter Trenkwalder

5/27/2018

Loading Packages and Data

```
library(car)
library(ggplot2)
library(cowplot)
library(pander)
ff = read.csv('forestfires.csv')
```

Introduction

Our analysis serves to explore this design question: "what factors lead to particularly damaging forest fires?

By performing exploratory analysis on this data we aim to discover commonalities in areas with sizable fire damange. The insights discovered will help inform a ditection system that provides early warnings to these regions. In order to identify these characteristics, we followed the following steps:

- examine the data and make any necessary adjustments to improve its quality and usability
- perform initial analysis on the included variables to identify key indicators
- use multivariate analysis on the material variables to converge upon some commonalities

The Data

Upon importing the data, we can see that we have 517 observations and 13 variables. These variables are mostly of the numerical type, with a few integers and two factors included as well. All of these variable types appear to be appropriate given the data.

```
str(ff)
```

```
## 'data.frame':    517 obs. of  13 variables:
##  $ X    : int  7 7 7 8 8 8 8 8 8 7 ...
##  $ Y    : int  5 4 4 6 6 6 6 6 6 5 ...
##  $ month: Factor w/ 12 levels "apr","aug","dec",..: 8 11 11 8 8 2 2 2 12 12 ...
##  $ day  : Factor w/ 7 levels "fri","mon","sat",..: 1 6 3 1 4 4 2 2 6 3 ...
##  $ FFMC : num  86.2 90.6 90.6 91.7 89.3 92.3 92.3 91.5 91 92.5 ...
##  $ DMC  : num  26.2 35.4 43.7 33.3 51.3 ...
##  $ DC   : num  94.3 669.1 686.9 77.5 102.2 ...
##  $ ISI  : num  5.1 6.7 6.7 9 9.6 14.7 8.5 10.7 7 7.1 ...
##  $ temp : num  8.2 18 14.6 8.3 11.4 22.2 24.1 8 13.1 22.8 ...
##  $ RH   : int  51 33 33 97 99 29 27 86 63 40 ...
##  $ wind : num  6.7 0.9 1.3 4 1.8 5.4 3.1 2.2 5.4 4 ...
##  $ rain : num  0 0 0 0.2 0 0 0 0 0 0 ...
##  $ area : num  0 0 0 0 0 0 0 0 0 0 ...
```

To better understand the specifics of the variables included in the dataset, we pulled details from the original data source. The definitions for each variables used in this report are included in the following table:

Variable	Name	Range	Details
X	X-axis Spatial Coordinate	1:9	
Y	Y-axis Spatial Coordinate	2:9	
month	Month of the Year		
day	Day of the Week		
FFMC	Fine Fuel Moisture Content	18.7:96.20	indicator of ease of ignition and flammability
DMC	Duff Moisture Code	1.1:291.3	indicator of fuel consumption in decomposing layers
DC	Drought Code	7.9:860.6	indicator of seasonal drought effects
ISI	Initial Spread Index	0.0:56.10	rating of expected rate of fire spread
temp	Temperature (C)	2.2:33.30	
RH	Relative Humidity (%)	15.0:100.0	
wind	Wind Speed (km/h)	0.40:9.40	
rain	Outside Rain (mm/m2)	0.0:6.4	
area	Burned Area (hectares)	0.00:1090.84	

The Cleanup

Initially in the process, we can use the summary of the dataframe to see high level details on the 13 variables. None of them appear to have unreadable or otherwise unusable values, there are no negative minimums, and all of the index variables have minimums and maximums within the appropriate ranges for each specific index.

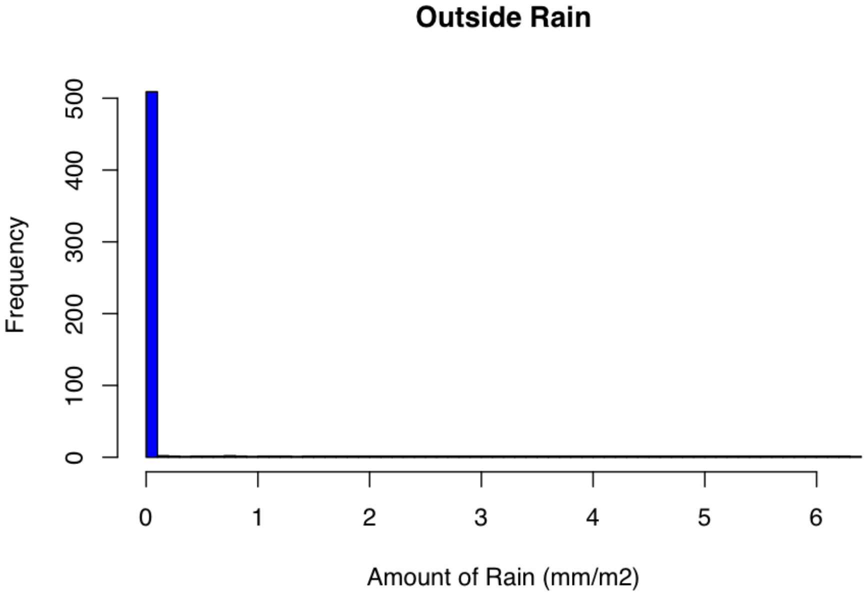
```r
summary(ff)
```

```
##        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
##
```

Rain

One variable of interest based on these statistical summaries is rain. When we look at the histogram of this variable we can start to uncover what is happening.

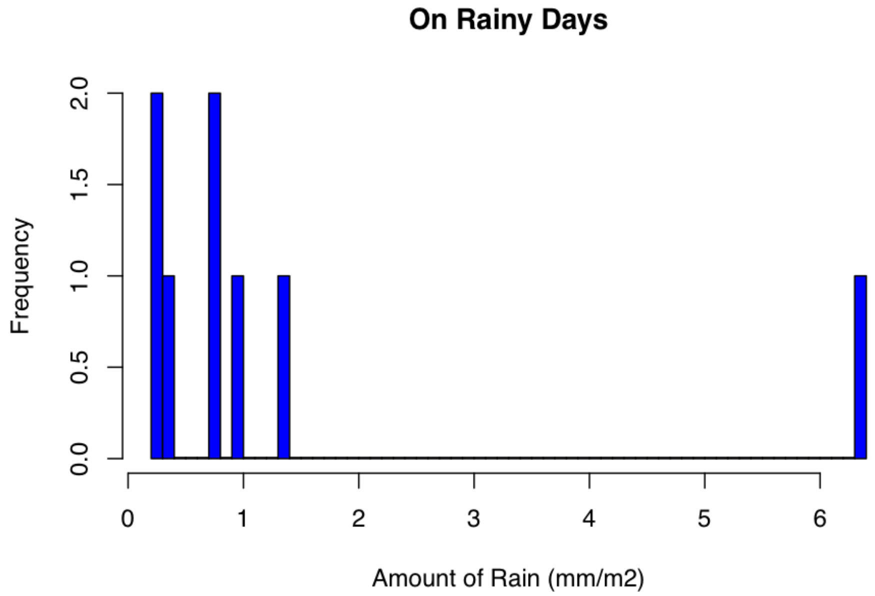
```
hist(ff$rain, breaks = 80, col = "blue",
     xlab = "Amount of Rain (mm/m2)",
     main = "Outside Rain")
```



We immediately notice that there is a signficiant number of observations with no rain at all. While this detail may be important to the analysis in the future, it initially entirely obscures our observations of what the data looks like. If we graph a histogram of the data removing any 0 values, we can

```
rainNonZero <- subset(ff$rain, ff$rain > 0.00)

hist(rainNonZero, breaks = 80, col = "blue",
     xlab = "Amount of Rain (mm/m2)",
     main = "On Rainy Days")
```



From here, a few things become clear about the variable:

- There are only 8 observations on days with any rain.
- All but one of these observations falls below 1.5 mm/m2.
- This single observation is actually quite an outlier at 6.4 mm/m2.

```r
rainyDay <- subset(ff, ff$rain > 6)
rainyDay
```

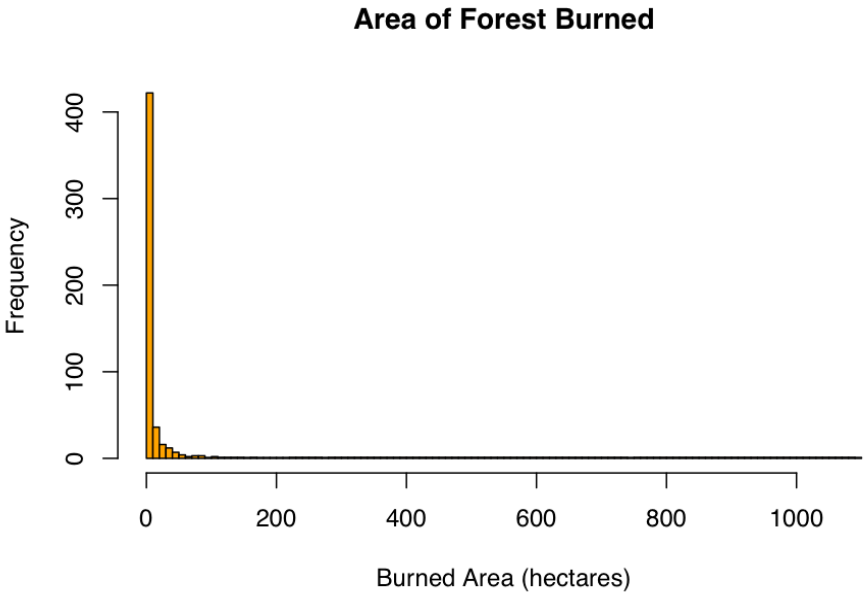
```
##     X Y month day FFMC   DMC    DC  ISI temp RH wind rain  area
## 500 7 5   aug tue 96.1 181.1 671.2 14.3 27.3 63  4.9  6.4 10.82
```

When we isolate the data for this observation, we can see that it occurred in aug which is a reasonable time of the year for a very rainy day. Interestingly this observation also includes 10.82 hectares of burned forest, which is close to the mean of 12.8472921 this variable.

Area

Similar to rain, we notice an interesting distribution in the area variable.

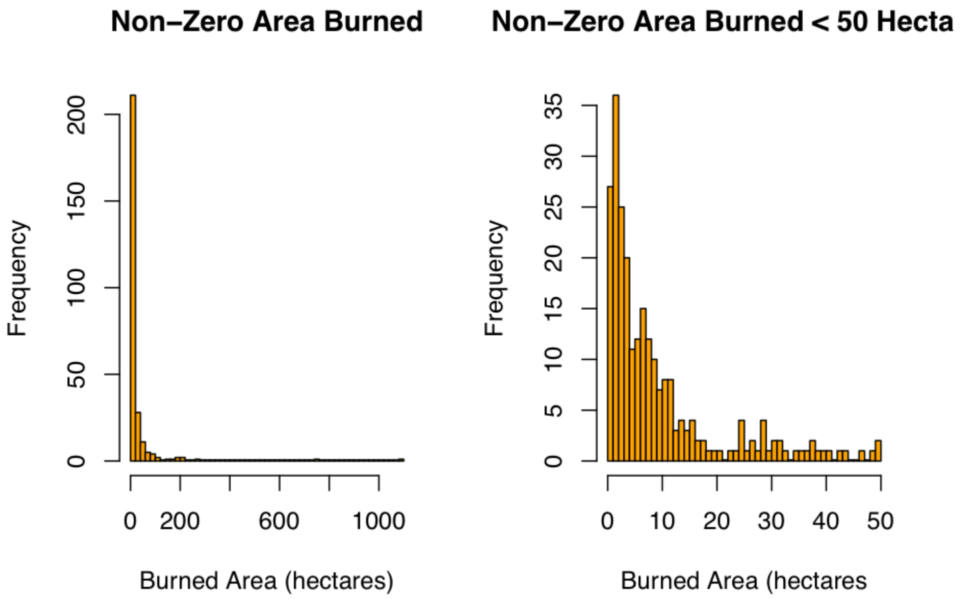
```r
hist(ff$area, breaks = 80, col = "orange",
     xlab = "Burned Area (hectares)",
     main = "Area of Forest Burned")
```



It appears that a significant portion of the observations have zero or near-zero hectares burnt. In fact, while the maximum value for this variable is 1090.84 hectares, the mean is only 12.8472921 hectares, greatly skewed by the zero values. A total of 247 observations, or 47.78% have 0 hectares burned.

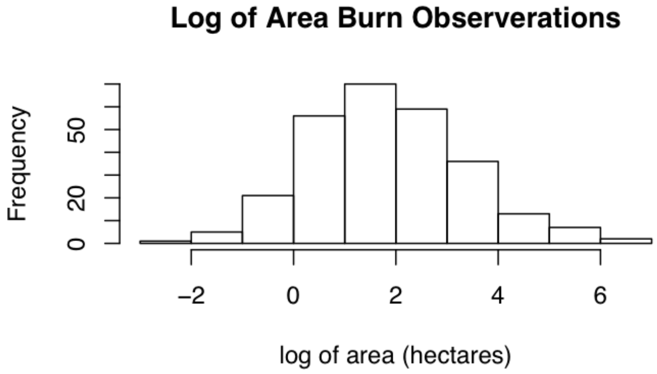
```r
areaNonZero <- subset(ff$area, ff$area > 0.00)

par(mfrow = c(1,2))
hist(areaNonZero, breaks = 40, col = "orange",
     xlab = "Burned Area (hectares)",
     main = "Non-Zero Area Burned")
hist(subset(ff$area, ff$area > 0 & ff$area <= 50), breaks = 40, col = "orange",
     xlab = "Burned Area (hectares",
     main = "Non-Zero Area Burned < 50 Hectares")
```



Removing non-zero values for area doesn't do much to alter the view, as there are still a large number of observations in the lowest bucket of the histogram. When we alter the scale to look only at non-zero values less than 50 hectares, we see that there is no obvious cutoff point at which we can call a fire "significant." However, we can transform the area observations on a log basis and see that the log transformation of the area approaches a normal distribution.

```
hist(log(ff$area),main = "Log of Area Burn Observerations",xlab = "log of area (hectares)")
```

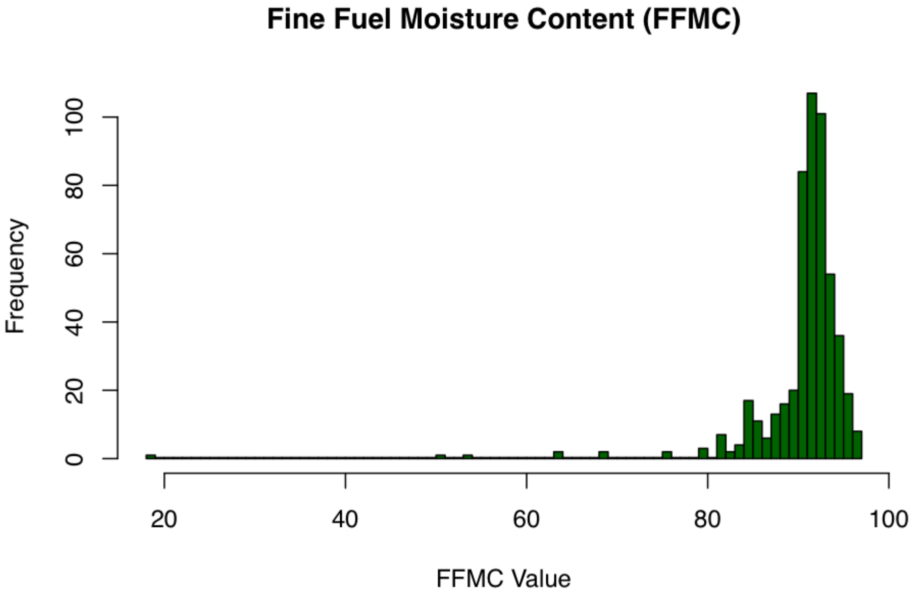


Because area will be our dependent variable in this analysis, we will compare other variables against the log of the area burned. Given the number of 0 area observations and the fact that the log(0) is undefined, we will add 1 to each input in the area column so that the minimum of our dependent variable will be 0 on the log scale.

Fine Fuel Moisture Content (FFMC)

The third variable that has a noticeable skew is the FFMC, which is values on the index that measures the moisture content of fine fuel sources such as grass, pine needles, tree moss and figs, for example. Unlike the two previously examined variables, the distribution of the FFMC skews left. While there are outlier observations on the low end of the range, all the values appear to be reasonable and should therefore be kept for analysis.

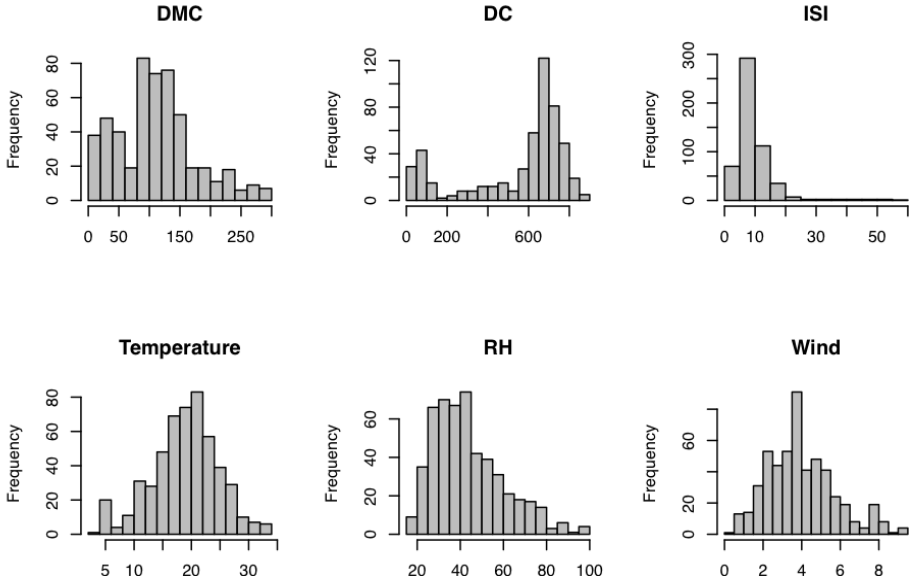
```r
hist(ff$FFMC, breaks = 80, col = "dark green",
     xlab = "FFMC Value",
     main = "Fine Fuel Moisture Content (FFMC)")
```



Other Variables

The other variables included in the dataset appear to have reasonable distributions without any necessary adjustments.

```r
par(mfrow=c(2,3))
hist(ff$DMC, breaks = 20, col = "grey", xlab = NULL, main = "DMC")
hist(ff$DC, breaks = 20, col = "grey", xlab = NULL, main = "DC")
hist(ff$ISI, breaks = 20, col = "grey", xlab = NULL, main = "ISI")
hist(ff$temp, breaks = 20, col = "grey", xlab = NULL, main = "Temperature")
hist(ff$RH, breaks = 20, col = "grey", xlab = NULL, main = "RH")
hist(ff$wind, breaks = 20, col = "grey", xlab = NULL, main = "Wind")
```



The Time

To aid in later analysis, we created an additional variable called "season" that is calculated based on the month of the observation. We did this to allow for testing of the hypothesis that the factors contributing to fire risk differed at different times of the year.

```
ff$season[ff$month == 'jun' | ff$month == 'jul' | ff$month == 'aug' ] <- "Summer"
ff$season[ff$month == 'sep' | ff$month == 'oct' | ff$month == 'nov' ] <- "Autumn"
ff$season[ff$month == 'dec' | ff$month == 'jan' | ff$month == 'feb' ] <- "Winter"
ff$season[ff$month == 'mar' | ff$month == 'apr' | ff$month == 'may' ] <- "Spring"
```

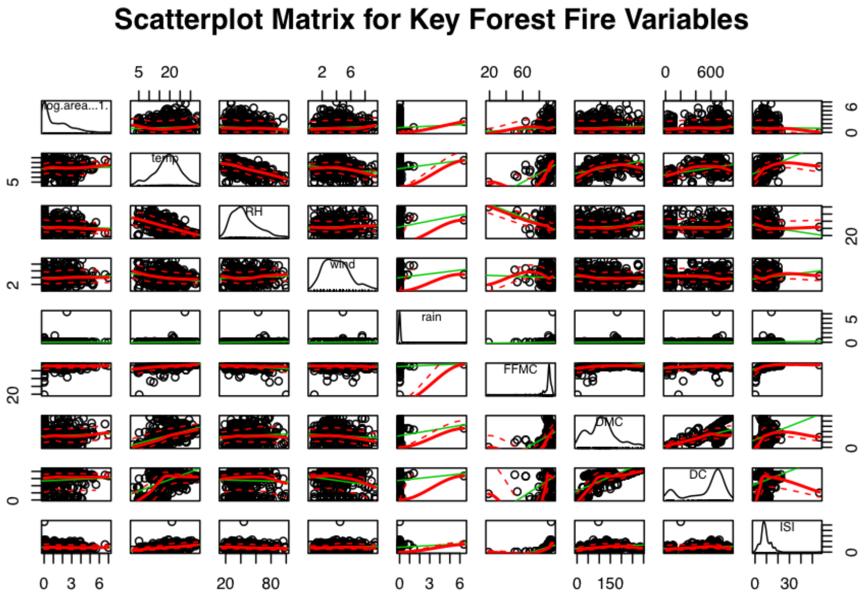
We can now observe that there are 233 observations in the summer, 188 observations in the autumn, 31 observations in the winter, and 65 observations in the spring.

Analysis of Key Relationships

Investigating Relationships between Quantitative Variables and Area

Given the large number of variables in our dataset, we'll initially use a scatterplot matrix to help direct our focus on which variables might serve as stronger signals in leading to particularly damaging forest fires. Note that the scatterplot matrix will only give us meaningful linear trends with quantitative variables. So we'll investigate categorical variables like location, day, and month separately against our dependent variable, area.

```
scatterplotMatrix( ~ log(area + 1) + temp + RH + wind + rain + FFMC + DMC + DC + ISI, data = ff,
                   main = "Scatterplot Matrix for Key Forest Fire Variables")
```



From our scatterplot matrix, we note a few noticeable sightings between our dependent variable and the others:

1. Temperature shows an interesting parabolic looking relationship which is unexpected because we intuitively think that lower temperatures would be associated with less fire spread
2. There appears to be a slight positive relationship between wind and the log of the area burned which seems to change slope around 5-6 km/h. This may indicate some sort of wind effect beyond a particular rate.
3. There's a semblance of an exponential relationship between FFMC (fine fuel moisture code) and the log(area). Perhaps beyond a certain FFMC rating, the impact on area burned increases incrementally with FFMC.
4. Lower relative humidities (RH) appear to be somewhat assoicated with larger areas burned
5. There is clustering around ranges of DC (duff code) and burn area

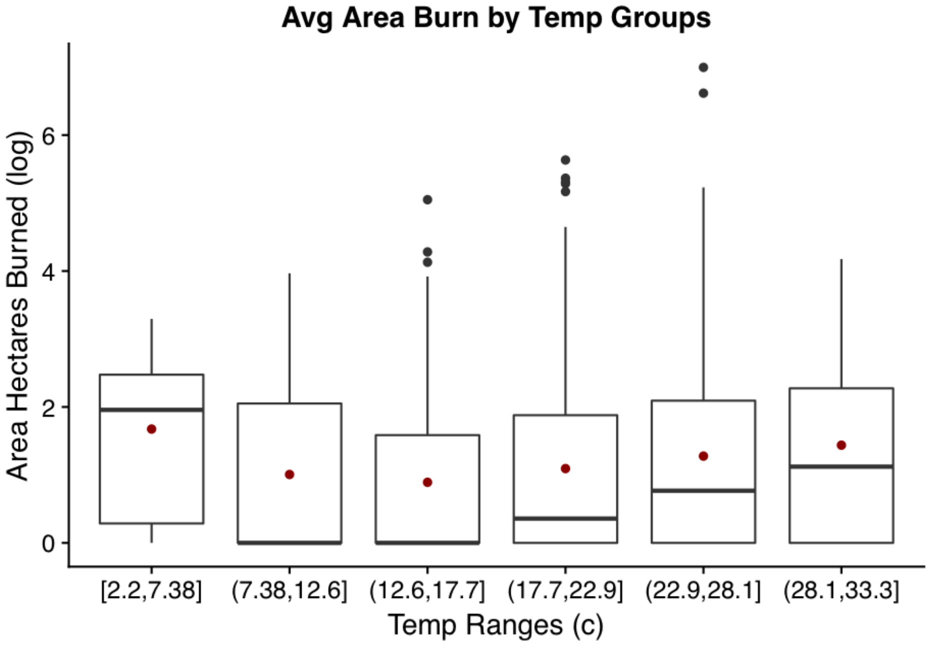
Collinearity between Quantitative Variables:

We note what appears to be collinearity between: temperature and RH, temperature and DMC, temperature and DC, and DMC and DC. We know from outside research that relative humidity is influenced by temperature. Because warm air can hold more water vapor than cool air, relative humidity falls when the temperature rises if no moisture is added to the air. The lack of rain helps explain this trend. We can probably consider relative humidity as a secondary signal and leave it out of this exploratory analysis. The same applies to DC and DMC where the Duff layer (DC) is the deepest layer of the moisture codes sitting below the layer DMC measures.

Strange Behvaior in Temperature and Area Burned

We can get more insight into trends in temperature and the effect on burn area by grouping the temperature observations into ranges. We do so using the cut function to break the observations into 6 range groups.

```
temp_cut = cut_interval(ff$temp,6)
ggplot(data=ff, aes(temp_cut,log(area + 1))) + geom_boxplot() +
  stat_summary(fun.y=mean, colour="darkred", geom="point") +
  labs(title ="Avg Area Burn by Temp Groups",
       x = "Temp Ranges (c)", y = "Area Hectares Burned (log)")
```

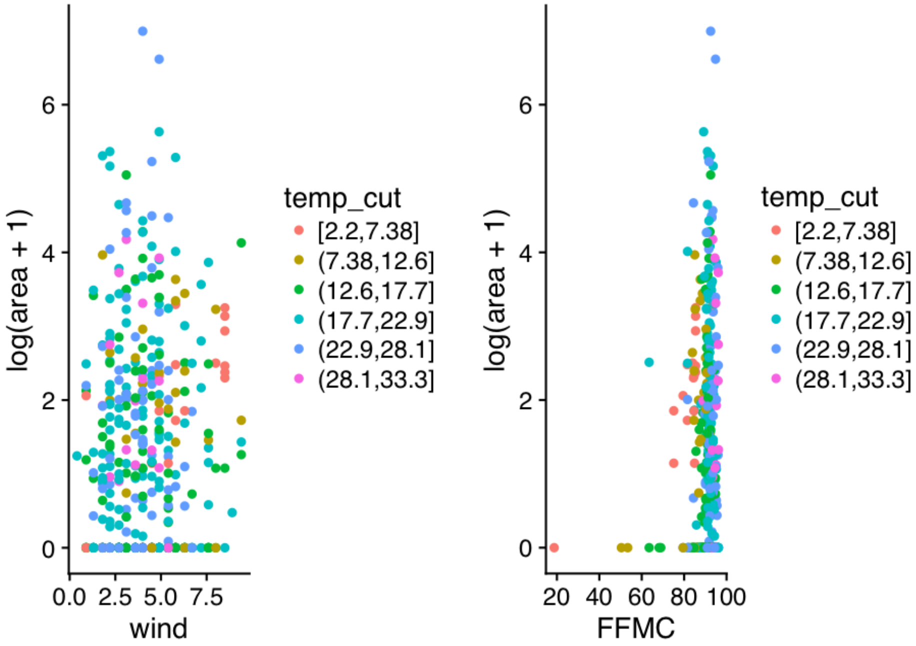


After binning temperature observations into groups, we see a more noticeable trend between the range of temperatures and the average area burned barring the first temperature group.It appears as temperatures exceed 17 degrees celcius, both the median and average area burned tend to increase.

What's particularly interesting is that both median and average area burned are at their highest when temperatures are at their lowest. Perhaps something else was happening during these low temperature observations that was driving up the fire damage.

To help inform us, we'll investigate where these temperature groups reside when our other suspect variables, wind and FFMC, are plotted against area burn. We do so by color coding the points in the scatterplot by their temperature range.

```
gwind = ggplot(ff,aes(wind,log(area + 1),colour = temp_cut)) +   geom_point()
gffmc = ggplot(ff,aes(FFMC,log(area + 1),colour = temp_cut)) + geom_point()
plot_grid(gwind,gffmc)
```

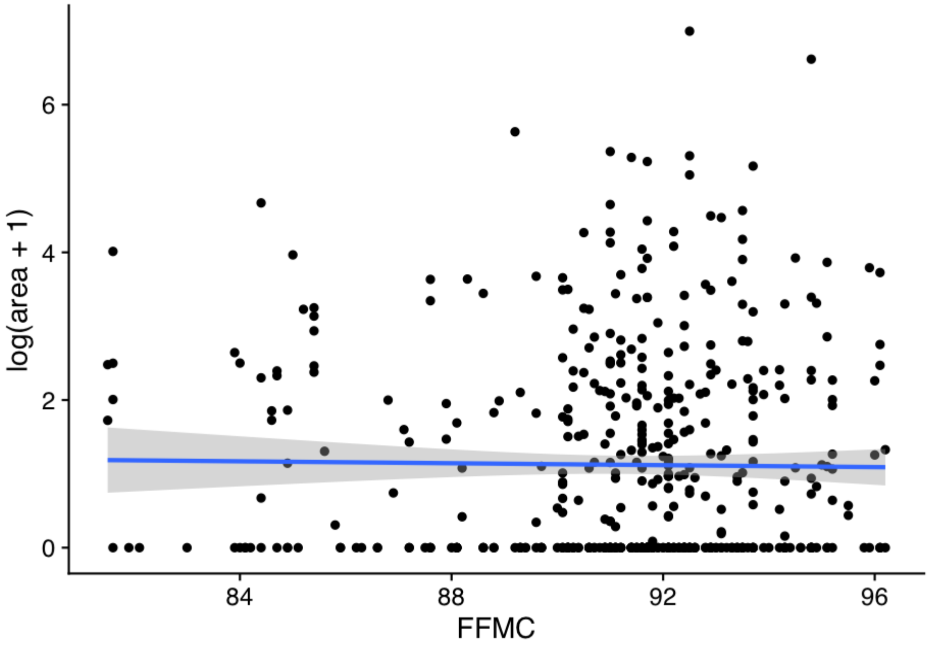


Now we can see that most of the low range temperatures associated with larger area burns are concentrated in peak wind periods. In fact, those low temperature observations have the highest count of peak-wind occurrences.This makes sense intuitively as we'd expect fires to burn across greater areas when winds are faster (all else equal). What this also helps demonstrate is that variables are likely acting together to influence area burn and so a multi-variate analysis seems more appropriate.

Focusing in on FFMC Influence on Area Burn

From our original scatterplot matrix we saw that there were very few observations of FFMC less than 80 which could be a byproduct of the park's natural environment. We'll zoom in on the relationship by excluding FFMC observations less than 80 and investigate any new patterns

```r
ggplot(ff[ff$FFMC >= 80,],aes(FFMC,log(area+1))) + geom_point() + geom_smooth(method = "lm")
```

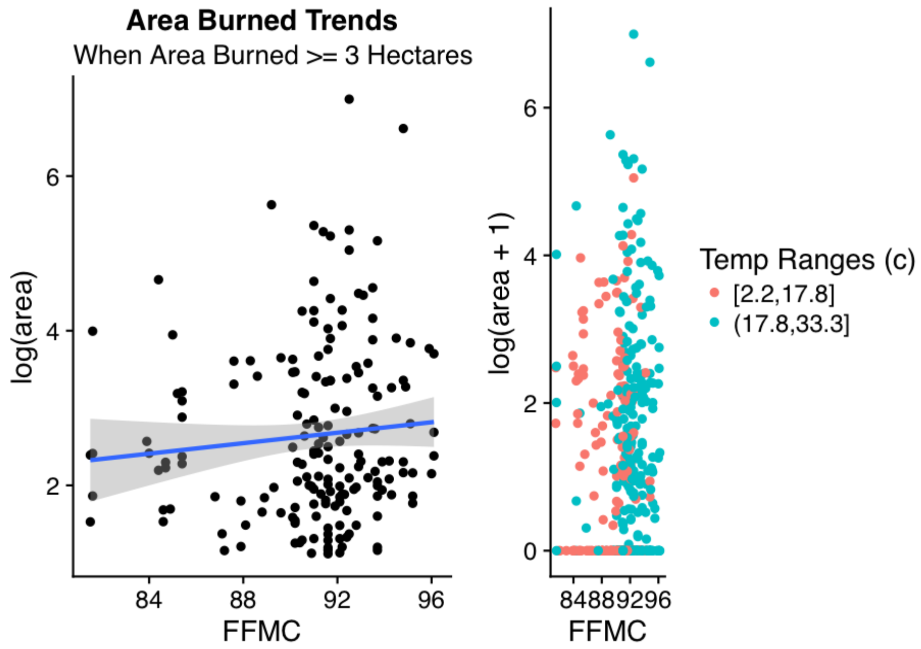


After focusing in on the observations there doesn't appear to be any linearity between FFMC and Area. However, we note that the relationship may change depending on how we define damaging fires and how other variables might interact with FFMC to influence area burn. We could refine our study by defining potentially damaging fires as those that are greater than or equal to 3 hectares (~7.5 acres). The thinking is that larger fires are harder to contain and risk spreading further, threatining greater damage. We can also look to see how temperature interacts with FFMC to influence area burn.

```
g_ffmc_3_hectares = ggplot(ff[ff$FFMC >= 80 & ff$area >= 3,],aes(FFMC,log(area))) + geom_point() +
  geom_smooth(method='lm') + labs(title ="Area Burned Trends", subtitle = "When Area Burned >= 3 Hectar

g_ffmc_temp = ggplot(ff[ff$FFMC >= 80,],aes(FFMC,log(area+1),colour = cut_interval(temp,2))) + geom_poi

plot_grid(g_ffmc_3_hectares,g_ffmc_temp)
```



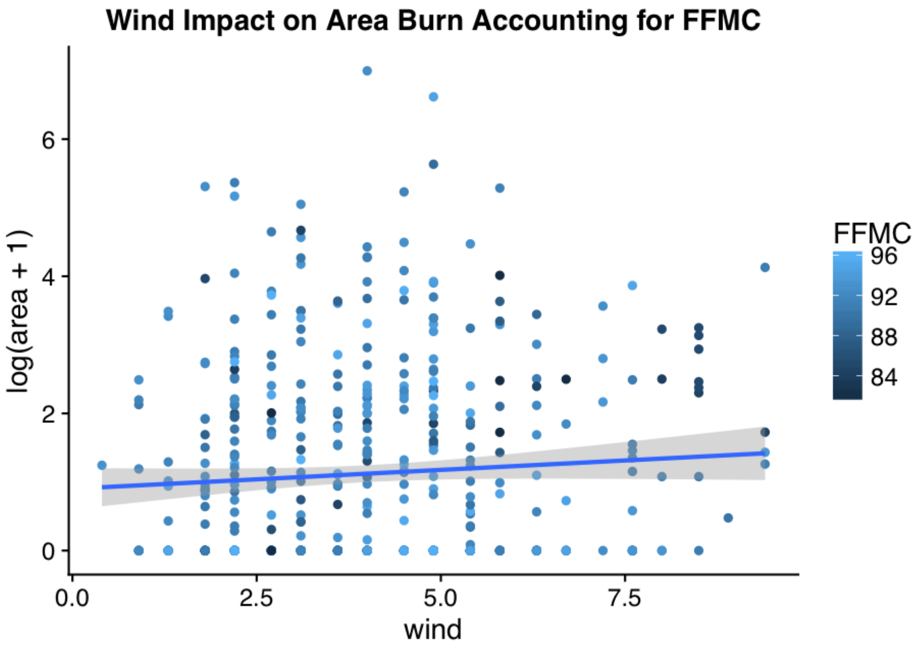
When viewing the data this way, we see different patterns. Now there appears to be an obvious linear relationship between FFMC and area burns greater than 3 hectares. What's also interesting is that we see most of the largest area burn observations occur at higher temperatures and FFMC. However, there seems to be a clear divide in FFMC observations and area temperature groupings. It looks like temperatures above 17 celcius are directly proportional to FFMC ratings above 92. FFMC might be dependent on temperature.

All said, we really need a better understanding of the park's ecosystem before we classify damaging based on area burned. Small areas of the park may contain a richer habitat, economy, or ecology than other largers areas of the park. We won't be able to make these distinctions until we have access to the park's information. For these reasons, we'll consider all area burns as potentially damaging during our exploratory analysis.

Linearity between Wind and Area Burn

Wind seemed to be the only variable that showed a consistent linear relationship with the log of area burned. We note that there's a correlation of 0.0467986 between those two variables. As we've seen with the temperature analysis, it's likely that wind interacts with other variables to influence area burn. We get a hint of that in that there are a relatively high number of low FFMC ratings during high wind periods which could be dragging down the slope of the area line.

```
ggplot(ff[ff$FFMC>80,],aes(wind,log(area + 1),colour = FFMC)) + geom_point() +
  geom_smooth(method = 'lm') + labs(title = "Wind Impact on Area Burn Accounting for FFMC")
```

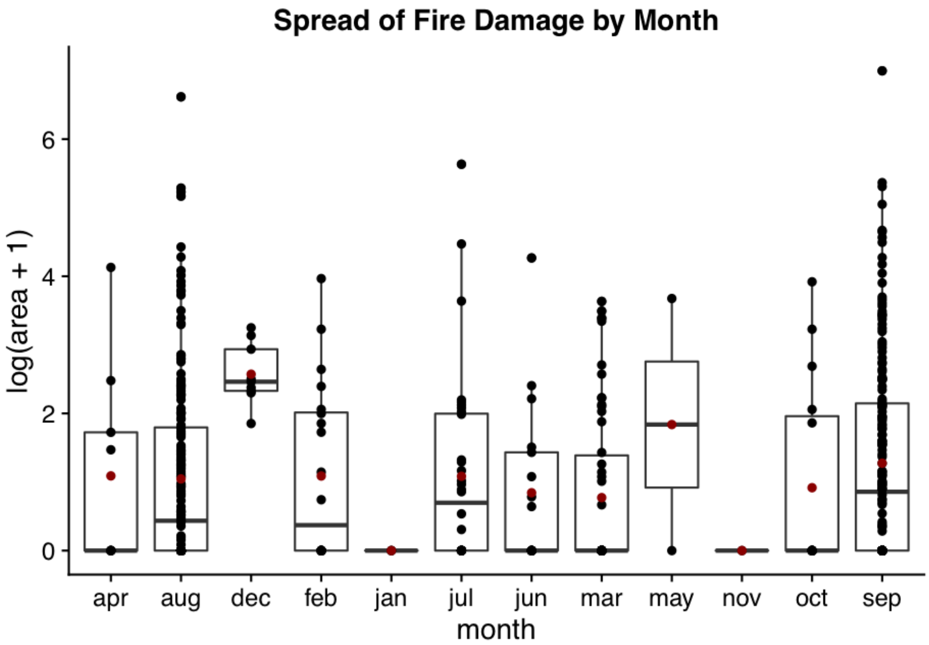


Investigating Relationships between Categorical Variables and Area

Since the scatterplot matrix isn't effective in highlighting categorical trends with area, we will evaluate each category separately.

Exploring Area Burn by Months

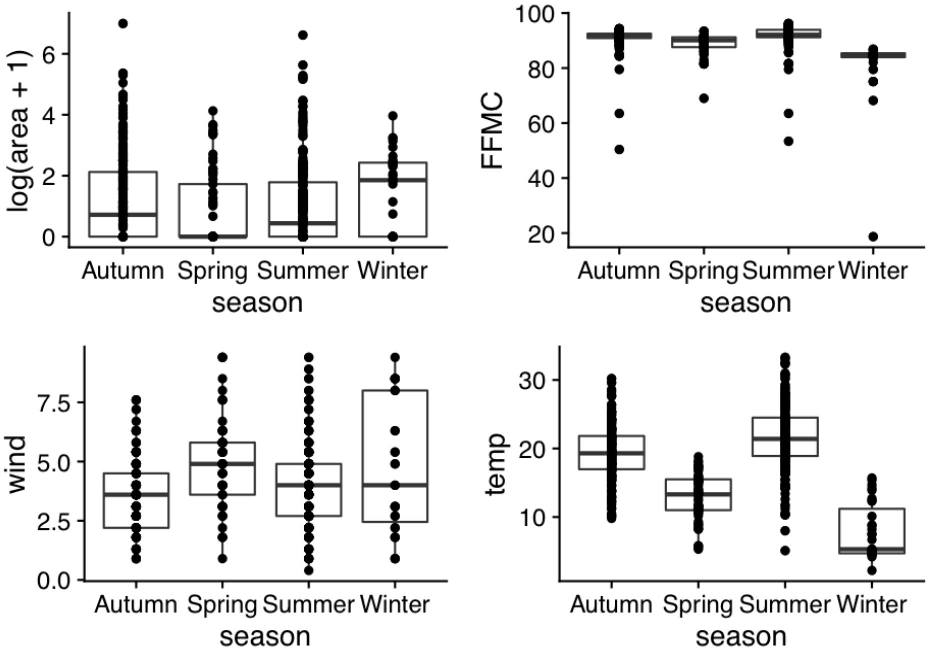
```
ggplot(ff, aes(month,log(area + 1))) + geom_boxplot() + geom_point() + stat_summary(fun.y=mean, colour="
                 labs(title = "Spread of Fire Damage by Month",ylab = "Log of Burned Area (in hectares)"
```



Looking at the distribution of fire damage by month we note December has the higest average (red dot) and median burn area. This is a little unexpected given Portugal's seasonality (dec is part of their winter). Both the preceding month (nov) and following month (jan) have 0 oberservations of area burn. So its interesting that december would have 9 instances of high area burn.

Perhaps there are other variables in our dataset that can help explain what might be going on. We'll look at those variables that seem to be the strongest signals we have to see how they relate to the months. Before doing so, we'll recategorize the months into seasons to make the plotting a little more digestible.

```
season = Recode(ff$month,"c('jun', 'jul','aug')='Summer'; c('sep','oct','nov')='Autumn';
                c('dec','jan','feb')='Winter'; c('mar','apr','may')='Spring'")
season_area = ggplot(ff, aes(season,log(area + 1))) + geom_boxplot() + geom_point()
season_ffmc = ggplot(ff, aes(season,FFMC)) + geom_boxplot() + geom_point()
season_wind = ggplot(ff, aes(season,wind)) + geom_boxplot() + geom_point()
season_temp = ggplot(ff, aes(season,temp)) + geom_boxplot() + geom_point()
plot_grid(season_area,season_ffmc,season_wind,season_temp)
```

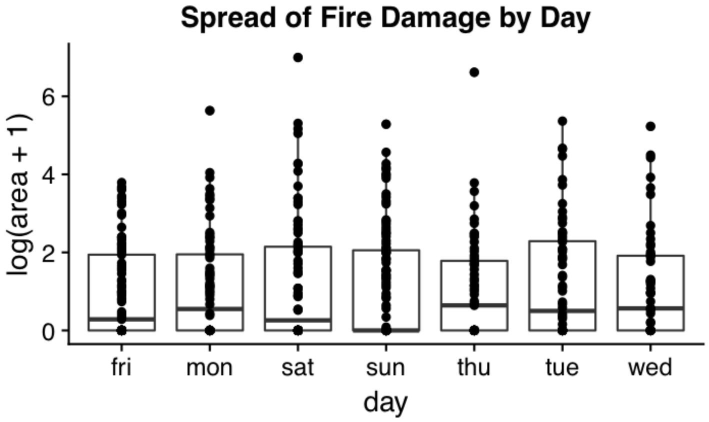


What's odd is that despite having the highest average/median burn area, winter has the lowest FFMC ratings, the lowest temperatures, and no noticeable difference in wind observations than the other seasons. This could suggest that there are other variables at play that aren't in our dataset. For instance, there may be controlled fires intentionally set that we are unaware of.

Exploring Area Burn by Day of Week

We can also see that the day-of-week is a pretty weak signal when looking at its relationship with the area burned.

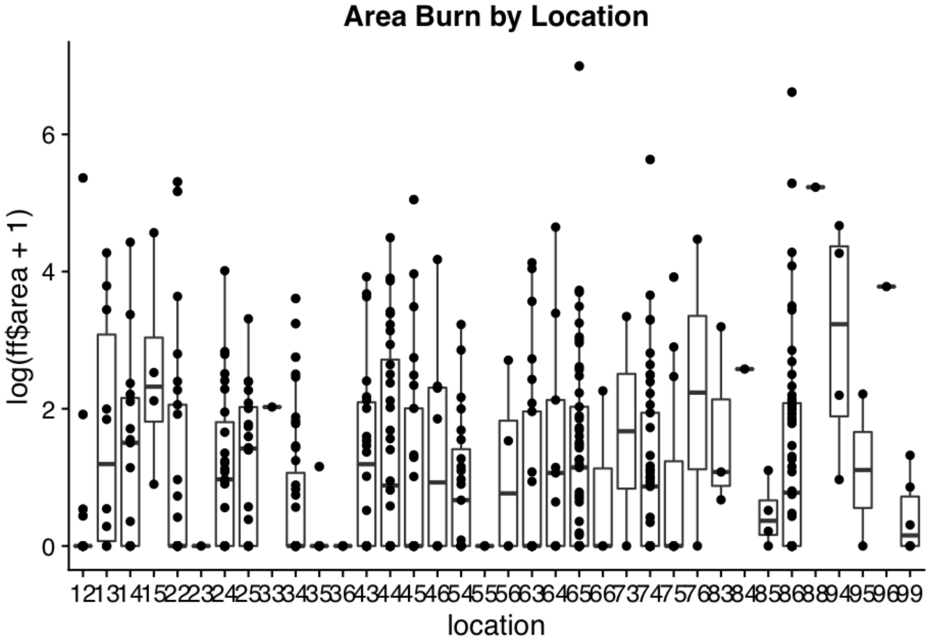
```r
ggplot(ff, aes(day,log(area + 1))) + geom_boxplot() + geom_point() +
  labs(title = "Spread of Fire Damage by Day",ylab = "Log of Burned Area (in hectares)")
```



Exploring Area Burn by Location

We are provided with separate x,y coordinates of the park's map. We combine those coordinates to created particular location points. We'll check to see if the magnitude of fire damage is concentrated in particular areas of the park. In the below graph, the numbers can be interpreted as (x,y) coordinate points.

```r
location = paste0(ff$X,ff$Y)
ggplot(ff,aes(location,log(ff$area + 1))) + geom_boxplot() + geom_point() +
  labs(title = 'Area Burn by Location')
```



The coordinates 44,65,74, and 86 stand out as having relatively higher distributions of fire damage than other locations. This could be an area of interest to pursue, but we'll likely need more information as these distributions could be a result of confounding variables such as camping zones.

Analysis of Secondary Effects

During our preliminary analysis we noted the potential of outside influences that may be disambiguating our interpretation of relationships. We saw that the month of December generally had more area burn events than other months despite having relatively lower recordings of other variables that seemed to be signals for area burn like temperature and FFMC. It is not uncommon for public parks to conduct controlled fires as a means for fire spread prevention and, so, it is possible that excercises in controlled burning are confounding our findings. Additionally, we highlighted higher distributions of area burn at particular park locations. Again these locations did not show a higher presence of other signal variables. With that said, we know nothing about camping or social activity within the park. Humans have a history of being the cause behind many forest fires and we are not able to control for that variable with the given dataset. Social behaviors within the park could also be confounding our analysis.

Conclusion

Identifying particularly damaging forest fires first requires a definition of damaging. We used area as a proxy for damaging, but admit that area may not necessarily equate to damage. Smaller areas can have much richer ecosystems and economies than larger areas, and if impacted, result in greater biological and or financial damage. Given we did not have this data, we conducted our analysis against the variable area alone. In doing so, we found what appear to be good candidates for core signals in potential area burn from park fires.

Generally higher temperatures tend to be associated with higher average (log) area burns in the park. However, we observed that wind and temperature can interact to exacerbate fire spread. For instance, low temperatures can still be associated with higher area burns if wind speeds are also relatively high. We note the same relationship with wind as there is a slight positive linear relationship with wind and area burn. While FFMC appeared to be a strong canidate as a signal for area burn, it seems to be dependent on temperature and possibily other variables. We saw this type of collinearity between DMC and DC as well as temperature and RH. Given the aforementioned interactions and collinearities between the independent variables noted, we've concluded that our analysis needs to extend beyond a bi-variate approach into multi-variate analyses. We need to hold other variables constant to get a better understanding of true influence. Moreover, we need to account for probable secondary variables like social activities within the park and any controlled fire events. Without more information from the park and a more rigourous statistical model, we can only speculate that higher winds and higher temperatures tend to lead to higher area burn.