W203 Lab 1: Forest Fire EDA

Smith_Vincent_He_Sered

9/24/2017

1. Introduction

1.1 Research Purpose

The purpose of this analysis is to examine what factors are associated with particularly damaging forest fires. For this analysis, area burned will serve as a proxy for damage.

```
library(car)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(grid)
library(corrplot)
```

```
setwd("C:/Users/Asher/Desktop/MIDS/w203/MIDS/Lab1/Forest_Fire_EDA/")
df_n = read.csv("forestfires.csv", stringsAsFactors = FALSE)
df <- df_n
```

The examined data set consists of 517 observations on 13 variables from a sinlge park in Portugal. The observed numerical variables are both discrete and continuous. In addition, there are two categorical values indicating the month and the day of the week.

```
str(df_n)
```

```
## '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: chr  "mar" "oct" "oct" "mar" ...
##  $ day  : chr  "fri" "tue" "sat" "fri" ...
##  $ 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 ...
```

```
head(df_n, 5)
```

```
##   X Y month day FFMC  DMC    DC ISI temp RH wind rain area
## 1 7 5   mar fri 86.2 26.2  94.3 5.1  8.2 51  6.7  0.0    0
## 2 7 4   oct tue 90.6 35.4 669.1 6.7 18.0 33  0.9  0.0    0
## 3 7 4   oct sat 90.6 43.7 686.9 6.7 14.6 33  1.3  0.0    0
## 4 8 6   mar fri 91.7 33.3  77.5 9.0  8.3 97  4.0  0.2    0
## 5 8 6   mar sun 89.3 51.3 102.2 9.6 11.4 99  1.8  0.0    0
```

1.2 Data Description and Quality

As an initial observation, the data set did not contain any missing values. For the time oriented categorical variables, values are present for each day of the week and each day of the month. However, the temporal range for the data set is not known. As an example, the data set would not reveal whether two observations occurring in September occurred in the same year, or whether they occurred five years apart. This temporal aspect carries to the numerical values. As examples, it is not clear from the data whether the wind, temperature, and rain variables represent the total rainfall for the given month/day combination, whether they represent the average wind and total rain up to the time of the fire, or some other configuration like a trailing sum/average over a specified time window. In addition, the dataset does not provide contextual information that may be helpful such as historical averages for the variables, which could be compared to the observed time of the fire. The set does not provide other potentially helpful information like available fuel, terrain, duration of fire, or cause of fire. Finally, it is not known whether the observations list all recorded fires in the region, fires from a certain period, or a sampling of recorded fires. As a result, analysis related to counts and correlations on the categorical variables must be caveated, since they may be the result of the underlying sampling methods. This makes evaluating and developing context difficult. Finally, the unit of measure for the key Area variable is measured in hectares to two decimal places. Because a hectare is relatively large unit of area, measuring small fires with precision may be difficult. It additionally may be responsible for generating 0 values for the Area variable and obscuring relationships that might be better revealed by a more precise unit of measure.

```
summary(df_n)
```

```
##        X               Y          month              day
##  Min.   :1.000   Min.   :2.0   Length:517         Length:517
##  1st Qu.:3.000   1st Qu.:4.0   Class :character   Class :character
##  Median :4.000   Median :4.0   Mode  :character   Mode  :character
##  Mean   :4.669   Mean   :4.3
##  3rd Qu.:7.000   3rd Qu.:5.0
##  Max.   :9.000   Max.   :9.0
##       FFMC            DMC              DC             ISI
##  Min.   :18.70   Min.   :  1.1   Min.   :  7.9   Min.   : 0.000
##  1st Qu.:90.20   1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500
##  Median :91.60   Median :108.3   Median :664.2   Median : 8.400
##  Mean   :90.64   Mean   :110.9   Mean   :547.9   Mean   : 9.022
##  3rd Qu.:92.90   3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800
##  Max.   :96.20   Max.   :291.3   Max.   :860.6   Max.   :56.100
##       temp             RH              wind            rain
##  Min.   : 2.20   Min.   : 15.00   Min.   :0.400   Min.   :0.00000
##  1st Qu.:15.50   1st Qu.: 33.00   1st Qu.:2.700   1st Qu.:0.00000
##  Median :19.30   Median : 42.00   Median :4.000   Median :0.00000
##  Mean   :18.89   Mean   : 44.29   Mean   :4.018   Mean   :0.02166
##  3rd Qu.:22.80   3rd Qu.: 53.00   3rd Qu.:4.900   3rd Qu.:0.00000
##  Max.   :33.30   Max.   :100.00   Max.   :9.400   Max.   :6.40000
```

```
##       area
##  Min.   :   0.00
##  1st Qu.:   0.00
##  Median :   0.52
##  Mean   :  12.85
##  3rd Qu.:   6.57
##  Max.   :1090.84
```

```
table(df_n$day)
```

```
##
## fri mon sat sun thu tue wed
##  85  74  84  95  61  64  54
```

```
table(df_n$month)
```

```
##
## apr aug dec feb jan jul jun mar may nov oct sep
##   9 184   9  20   2  32  17  54   2   1  15 172
```

1.3 Data Processing and Preparation

To facilitate the evaluation of correlations with categorical variables, dummy variables were created for each level in a parallel dataframe. The dummy variable takes on a value of 1 for instances of the category and 0 for instances where another level is present. In addition, categorical variables along with their associated dummy variables were created for the twelve months, the four seasons, and for weekdays and weekends. Weekdays are considered Monday through Friday for purposes of the analysis. Summer is considered June through August, fall is considered September through November, winter is December through February, and Spring is March through May. An additional variable was created to concatenate the XY coordinate. Because of the presence of 0 valued observations in the Area variable, a variable adding 1 to each observation was created to facilitate data transforms. And, finally, a variable for relative fire severity based on Area was created with the criteria of 0 hectare as "Min", 0.09 to 10 as "Small", >10 to 20 as "Medium", >20 to 100 as "Large", and >100 as "Severe". The designations are based on the distribution of non-zero values relative to the sample, although it may be useful to consult an expert for additional analysis to confirm any assumptions.

It is worth noting that because of the range of the numerical scales in the variables it would be useful to place each variable on a common scale if a modeling exercise were undertaken.

```
# to facilitate matching
df_n <- mutate(df_n,
       sun = ifelse(df_n$day=="sun", 1, 0),
       mon = ifelse(df_n$day=="mon", 1, 0),
       tue = ifelse(df_n$day=="tue", 1, 0),
       wed = ifelse(df_n$day=="wed", 1, 0),
       thu = ifelse(df_n$day=="thu", 1, 0),
       fri = ifelse(df_n$day=="fri", 1, 0),
       sat = ifelse(df_n$day=="sat", 1, 0),
       time.of.week = ifelse(df_n$day == "sat" | df_n$day == "sun", "Weekend", "Weekday"),
       weekend_n = ifelse(df_n$day == "sat" | df_n$day == "sun", 1, 0),
       season = ifelse(df_n$month == "mar" | df_n$month == "apr" | df_n$month == "may", "spring",
                ifelse(df_n$month == "jun" | df_n$month == "jul" | df_n$month == "aug", "summer",
                ifelse(df_n$month == "sep" | df_n$month == "oct" | df_n$month == "nov", "fall",
                ifelse(df_n$month == "dec" | df_n$month == "jan" | df_n$month == "feb", "winter", "NA")))
       spring_n = ifelse(df_n$month == "mar" | df_n$month == "apr" | df_n$month == "may", 1, 0),
       summer_n = ifelse(df_n$month == "jun" | df_n$month == "jul" | df_n$month == "aug", 1, 0 ),
       fall_n = ifelse(df_n$month == "sep" | df_n$month == "oct" | df_n$month == "nov", 1, 0  ),
```

```
      winter_n = ifelse(df_n$month == "dec" | df_n$month == "jan" | df_n$month == "feb", 1, 0),
      XY_Coord = paste(X, Y),
      size_cat = ifelse(df_n$area  == 0, "Min",
                        ifelse(df_n$area >= 0.09 & df_n$area <= 10, "Small",
                        ifelse(df_n$area > 10 & df_n$area <= 20, "Medium",
                        ifelse(df_n$area > 20 &df_n$area <= 100, "Large",
                        ifelse(df_n$area >100, "Severe", "Check"))))),
      area_p1 = df_n$area + 1,
      jan = ifelse(df_n$month=="jan", 1, 0),
      feb = ifelse(df_n$month=="feb", 1, 0),
      mar = ifelse(df_n$month=="mar", 1, 0),
      apr = ifelse(df_n$month=="apr", 1, 0),
      may = ifelse(df_n$month=="may", 1, 0),
      jun = ifelse(df_n$month=="jun", 1, 0),
      jul = ifelse(df_n$month=="jul", 1, 0),
      aug = ifelse(df_n$month=="aug", 1, 0),
      sep = ifelse(df_n$month=="sep", 1, 0),
      oct = ifelse(df_n$month=="oct", 1, 0),
      nov = ifelse(df_n$month=="nov", 1, 0),
      dec = ifelse(df_n$month=="dec", 1, 0)
)

# factor for graphics
levels(df_n$month)
df_n$month <- factor(df_n$month, levels = c("jan", "feb", "mar", "apr", "may", "jun", "jul",
                                            "aug", "sep", "oct", "nov", "dec"))
levels(df_n$day)
df_n$day <- factor(df_n$day, levels = c("mon", "tue", "wed", "thu", "fri", "sat", "sun"))
```

```
df_n_no0 <- df_n[df_n$area != 0,]
summary(df_n_no0)
```

```
##        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
##
##       sun              mon              tue              wed
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1741   Mean   :0.1444   Mean   :0.1333   Mean   :0.1185
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##       thu              fri              sat         time.of.week
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Length:270
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :0.0000   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :0.1148   Mean   :0.1593   Mean   :0.1556
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##    weekend_n         season            spring_n          summer_n
##  Min.   :0.0000   Length:270         Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.0000   Class :character   1st Qu.:0.00000   1st Qu.:0.000
##  Median :0.0000   Mode  :character   Median :0.00000   Median :0.000
##  Mean   :0.3296                      Mean   :0.08889   Mean   :0.463
##  3rd Qu.:1.0000                      3rd Qu.:0.00000   3rd Qu.:1.000
##  Max.   :1.0000                      Max.   :1.00000   Max.   :1.000
##
##      fall_n          winter_n         XY_Coord           size_cat
##  Min.   :0.0000   Min.   :0.00000   Length:270         Length:270
##  1st Qu.:0.0000   1st Qu.:0.00000   Class :character   Class :character
##  Median :0.0000   Median :0.00000   Mode  :character   Mode  :character
##  Mean   :0.3778   Mean   :0.07037
##  3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000
##
##     area_p1              jan         feb               mar
##  Min.   :   1.09   Min.   :0   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:   3.14   1st Qu.:0   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :   7.37   Median :0   Median :0.00000   Median :0.00000
##  Mean   :  25.60   Mean   :0   Mean   :0.03704   Mean   :0.07037
##  3rd Qu.:  16.42   3rd Qu.:0   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1091.84   Max.   :0   Max.   :1.00000   Max.   :1.00000
##
##       apr               may               jun               jul
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.000000   Median :0.00000   Median :0.00000
##  Mean   :0.01481   Mean   :0.003704   Mean   :0.02963   Mean   :0.06667
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.00000   Max.   :1.00000
##
##       aug              sep               oct               nov
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0
```

```
##  Mean   :0.3667   Mean   :0.3593   Mean   :0.01852   Mean   :0
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :0
##
##       dec
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.03333
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

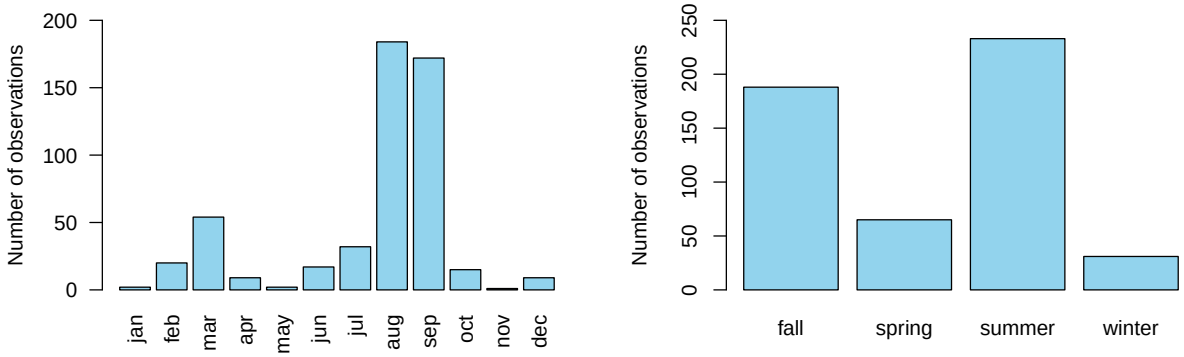
Our parallel data frame contains 42 variables in aggregate.

2. Univariate Analysis of Key Variables

2.1 Time and Location

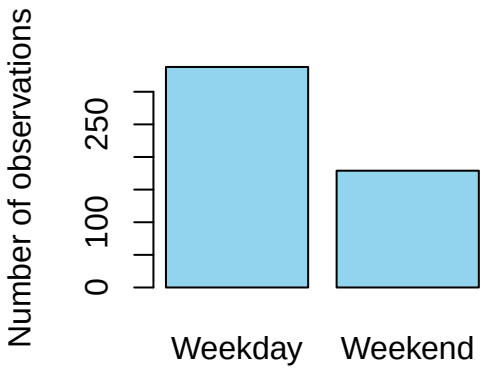
We begin by exploring the distribution of the times and locations of the measurements in the dataset. A month by month comparison shows that the majority of the observations are in August and September with the third most in March. Looking in terms of seasonality, we can see that the majority of observations came in the fall and summer with very few in the winter.

```
layout(matrix(c(1,2), 1, 2))
barplot(table(df_n$month), ylab = "Number of observations", col = "#92d3ed", ylim=c(0,200), las=2)
barplot(table(df_n$season), ylab = "Number of observations", col = "#92d3ed", ylim=c(0,250))
```



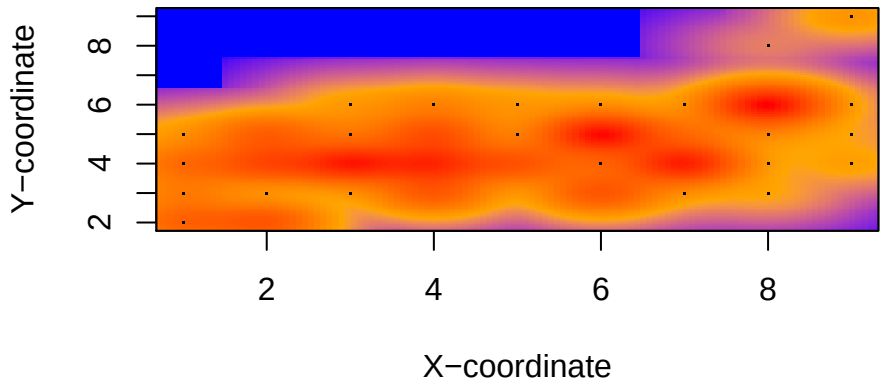
As far as when the observations were taken, we examine whether observations were more likely to come during the week or on weekends. Measurements appear to have been fairly well spread out.

```
barplot(table(df_n$time.of.week), ylab = "Number of observations", col = "#92d3ed")
```



In addition to investigating how spread out the observations are over time, we also look at how they are distributed over the area of the park. The observations come with an (x,y) created by partitioning the park into a 9X9 grid. The following visualization shows a heatmap of the density of observations from given locations within the park.

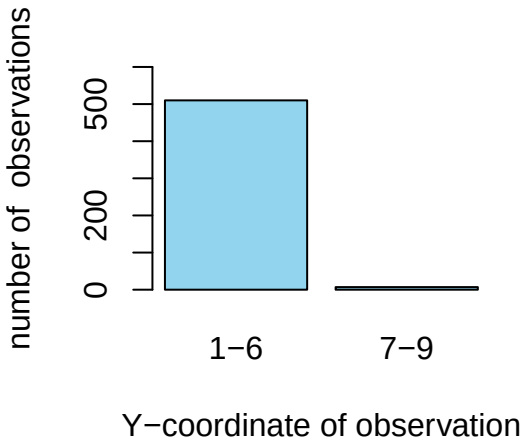
```
Lab.palette <- colorRampPalette(c("blue", "orange", "red"), space = "Lab")
smoothScatter(df_n$X, df_n$Y, colramp = Lab.palette, bandwidth = .45,
              xlab="X-coordinate", ylab="Y-coordinate")
```



We can see that the observations are generally spread out in the x direction, but seem to exclude the top third or so of the park. Indeed, we can see the ratio of observations from the top third of the park to the bottom two thirds.

```
barplot(c(length(df_n$Y[df_n$Y<7]),length(df_n$Y[df_n$Y>6])), names.arg = c("1-6", "7-9"),
        ylab = "number of  observations", xlab = "Y-coordinate of observation",
```

```
        col = "#92d3ed", ylim=c(0,600))
```



To summarize, the dataset shows that a large majority of measurements take place in a narrow span of the year during the months of August and September, and that the data also lacks observarions from the top part of the park. The fact that the observations are so unevenly distributed in physical and temporal merits caution and making conclusions about the park as a whole based on the data provided.

2.2 FWI Indices

Turning our attention from examining the distribution of the location of the observations, we change our focus to the measured values themselves. The dataset includes four separate indices from the Fire Weather Index, a Canadian system for fire danger. Below are some summary statistics for those four indices. We note that there are no missing or NA values.

```
cat("Summary of FFMC index:\n")
```

```
## Summary of FFMC index:
```

```
summary(df_n$FFMC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.70   90.20   91.60   90.64   92.90   96.20
```

```
cat("\nSummary of DMC index:\n")
```

```
##
## Summary of DMC index:
```

```
summary(df_n$DMC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.1    68.6   108.3   110.9   142.4   291.3
```

```
cat("\nSummary of DC index:\n")
```

```
##
## Summary of DC index:
```

```
summary(df_n$DC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.9   437.7   664.2   547.9   713.9   860.6
```

```
cat("\nSummary of ISI index:\n")
```

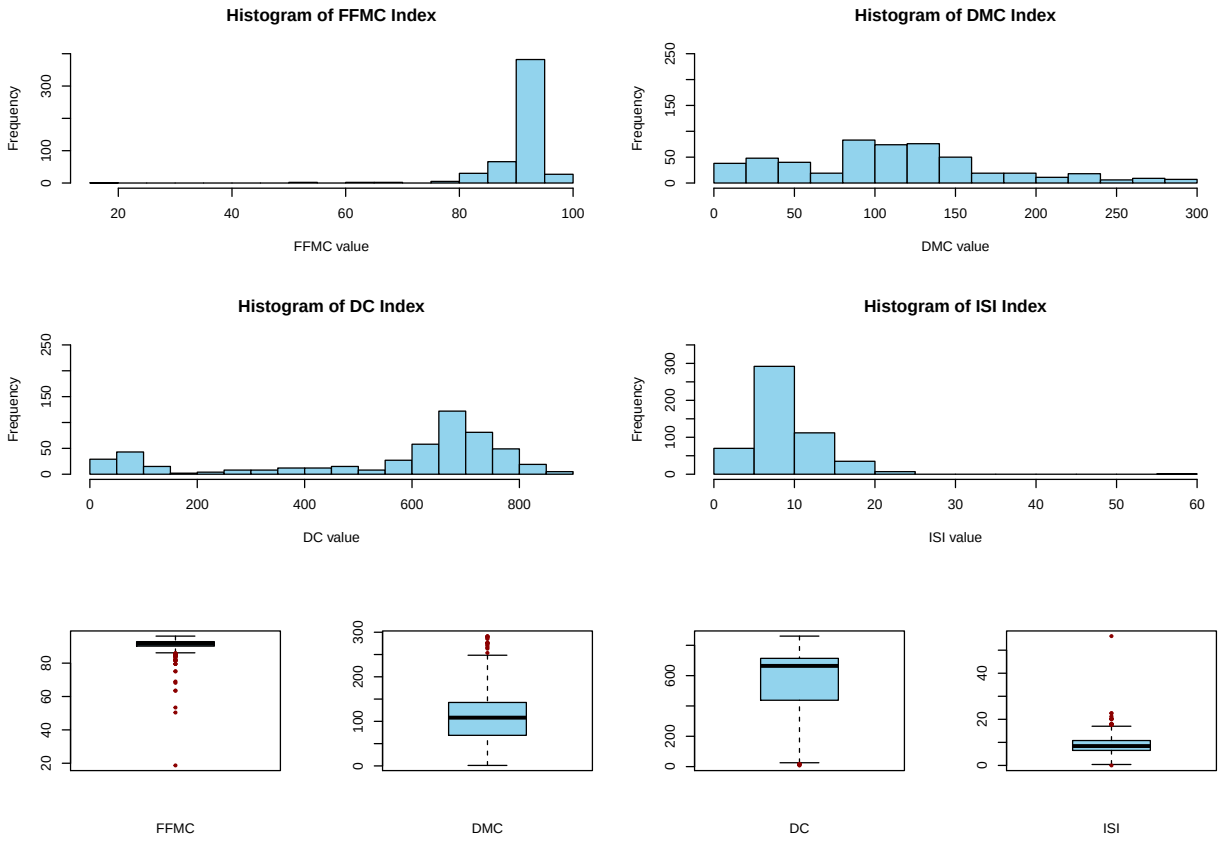
```
##
## Summary of ISI index:
```

```
summary(df_n$ISI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.500   8.400   9.022  10.800  56.100
```

Comparing the minx/max values with the quartile distributions, some of these variables seem like they might have a handful of outlier values. We create histograms and boxplots below to assess the distributions.

```
layout(matrix(c(1,1,2,2,3,3,4,4,5,6,7,8), 3, 4, byrow=TRUE))
hist(df_n$FFMC, breaks=20, xlab="FFMC value", main = "Histogram of FFMC Index",
     col = "#92d3ed", ylim=c(0,400))
hist(df_n$DMC, breaks=20, xlab="DMC value", main = "Histogram of DMC Index",
     col = "#92d3ed", ylim=c(0,250))
hist(df_n$DC, breaks=20, xlab="DC value", main = "Histogram of DC Index",
     col = "#92d3ed", ylim=c(0,250))
hist(df_n$ISI, breaks=20, xlab="ISI value", main = "Histogram of ISI Index",
     col = "#92d3ed", ylim=c(0,350))
boxplot(df$FFMC, xlab="FFMC", col = "#92d3ed", outcol = "darkred", outpch = 20)
boxplot(df$DMC, xlab="DMC", col = "#92d3ed", outcol = "darkred", outpch = 20)
boxplot(df$DC, xlab="DC", col = "#92d3ed", outcol = "darkred", outpch = 20)
boxplot(df$ISI, xlab="ISI", col = "#92d3ed", outcol = "darkred", outpch = 20)
```



From these variables, it appears we have outliers on the min side for the FFMC index and on the max side for the ISI index. These outliers are not necessarily an indication of data integrity issues, but we should be careful to ensure that they are accounted for when assessing means or correlations.

2.3 Temperature and Weather

We perform a similar analysis on temp, humidity, rain and wind to see if there are data integrity issues on those measurements.

```
cat("Summary of temp:\n")
```

```
## Summary of temp:
```

```
summary(df_n$FFMC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.70   90.20   91.60   90.64   92.90   96.20
```

```
cat("\nSummary of relative humidity:\n")
```

```
##
## Summary of relative humidity:
```

```
summary(df_n$DMC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.1    68.6   108.3   110.9   142.4   291.3
```

```
cat("\nSummary of wind:\n")
```

```
##
## Summary of wind:
```

```
summary(df_n$DC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.9   437.7   664.2   547.9   713.9   860.6
```

```
cat("\nSummary of rain:\n")
```

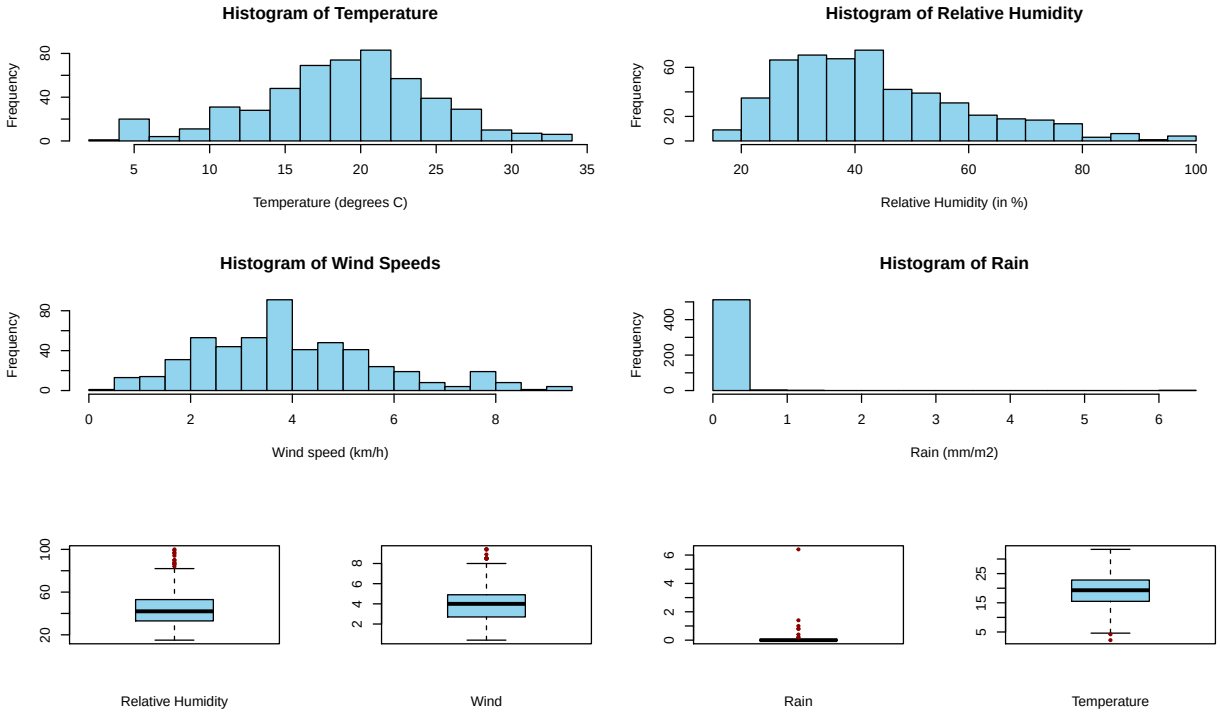
```
##
## Summary of rain:
```

```
summary(df_n$ISI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.500   8.400   9.022  10.800  56.100
```

Once again, we see that there are no NA values. We plot histograms to see if there are outliers or other data integrity issues to look out for.

```
layout(matrix(c(1,1,2,2,3,3,4,4,5,6,7,8), 3, 4, byrow=TRUE))
hist(df_n$temp, breaks=20, xlab="Temperature (degrees C)",
     main = "Histogram of Temperature", col = "#92d3ed")
hist(df_n$RH, breaks=20, xlab="Relative Humidity (in %)",
     main = "Histogram of Relative Humidity", col = "#92d3ed")
hist(df_n$wind, breaks=20, xlab="Wind speed (km/h)",
     main = "Histogram of Wind Speeds", col = "#92d3ed")
hist(df_n$rain, breaks=20, xlab="Rain (mm/m2)",
     main = "Histogram of Rain", col = "#92d3ed")
boxplot(df$RH, xlab="Relative Humidity", col = "#92d3ed", outcol = "darkred", outpch = 20)
boxplot(df$wind, xlab="Wind", col = "#92d3ed", outcol = "darkred", outpch = 20)
boxplot(df$rain, xlab="Rain", col = "#92d3ed", outcol = "darkred", outpch = 20)
boxplot(df$temp, xlab="Temperature", col = "#92d3ed", outcol = "darkred", outpch = 20)
```

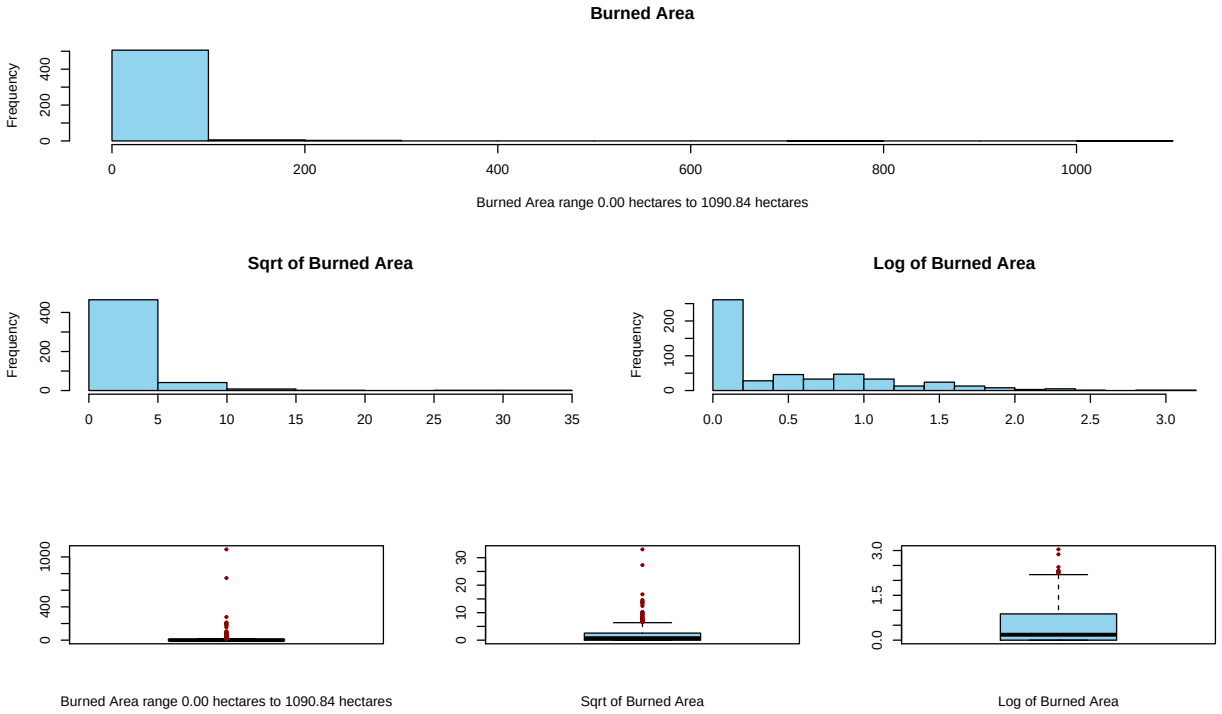


For the most part, there is nothing that stands out too strongly in regards to the distributions of the weather variables. Most of the variables appear normally distributed with some outliers occuring with relative humidity, wind, and rain. Relative humidity also exhibits some positive skew. We also note that rain is a variable that is often quite small, but has a small number of very large values. This is consistent with major storms, but may also be the result of how the measurement was taken. All in all, we have no reason to believe there are siginificant data integrity issues with the weather variables, but we should be cautious in using these variables to draw conclusions based on the issues discussed.

2.4 Area

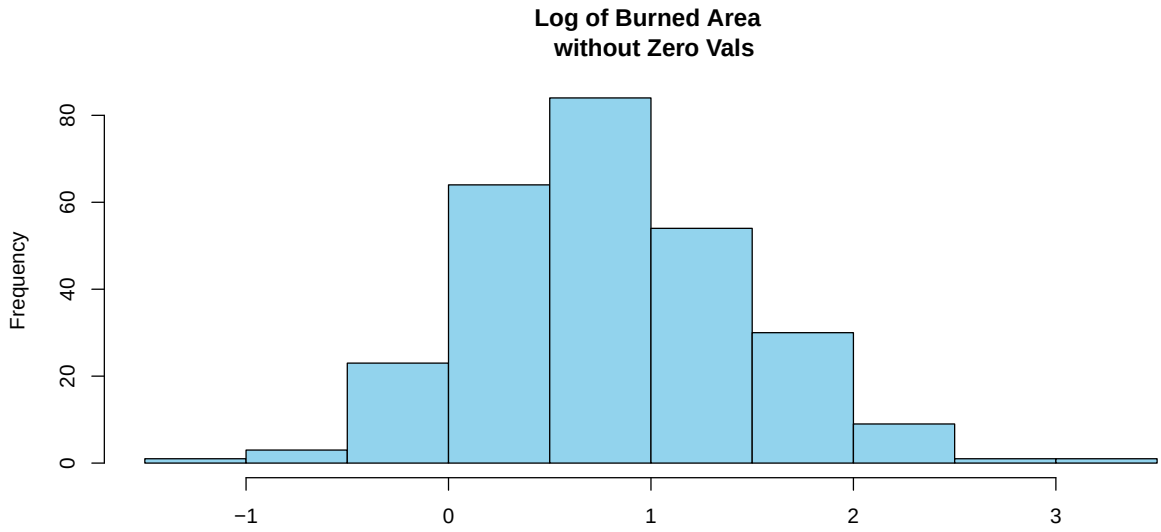
Finally, we examine our key Area variable using histograms and boxplots.

```
layout(matrix(c(1,1,1,1,1,1,2,2,2,3,3,3,4,4,5,5,6,6), 3, 6, byrow=TRUE))
hist(df_n$area, xlab="Burned Area range 0.00 hectares to 1090.84 hectares",
     main="Burned Area", col = "#92d3ed")
hist(sqrt(df_n$area), xlab="",
     main="Sqrt of Burned Area",col = "#92d3ed")
hist(log(df_n$area_p1, base=10), xlab="",
     main="Log of Burned Area", col = "#92d3ed")
boxplot(df_n$area, xlab="Burned Area range 0.00 hectares to 1090.84 hectares",
     xlab="Burned Area", col = "#92d3ed", outcol = "darkred", outpch = 20)
boxplot(sqrt(df_n$area),
     xlab="Sqrt of Burned Area",col = "#92d3ed", outcol = "darkred", outpch = 20)
boxplot(log(df_n$area_p1, base=10),
     xlab="Log of Burned Area", col = "#92d3ed", outcol = "darkred", outpch = 20)
```



Evaluating the base line histogram we see that the distribution does not conform to a normal distribution and exhibits significant positive skew. This is reinforced by the supporting boxplot. We first attempted to normalize the distribution with a square root transform, but it did not have any significant normalizing effect. Next, we attempted a transform. Because of the prevalence of 0 values in the set, we added 1 to each Area observation before performing the transform. The shape of the distribution was relatively constant, although it is noted that if the 0 values were not present the log transform may have a normalizing effect. It is likely the data set could benefit from a more precise measure of area to eliminate these 0 values if they do in fact represent the occurrence of a fire.

```
hist(log(df_n_no0$area, base=10), xlab="",
     main="Log of Burned Area \n without Zero Vals", col = "#92d3ed")
```

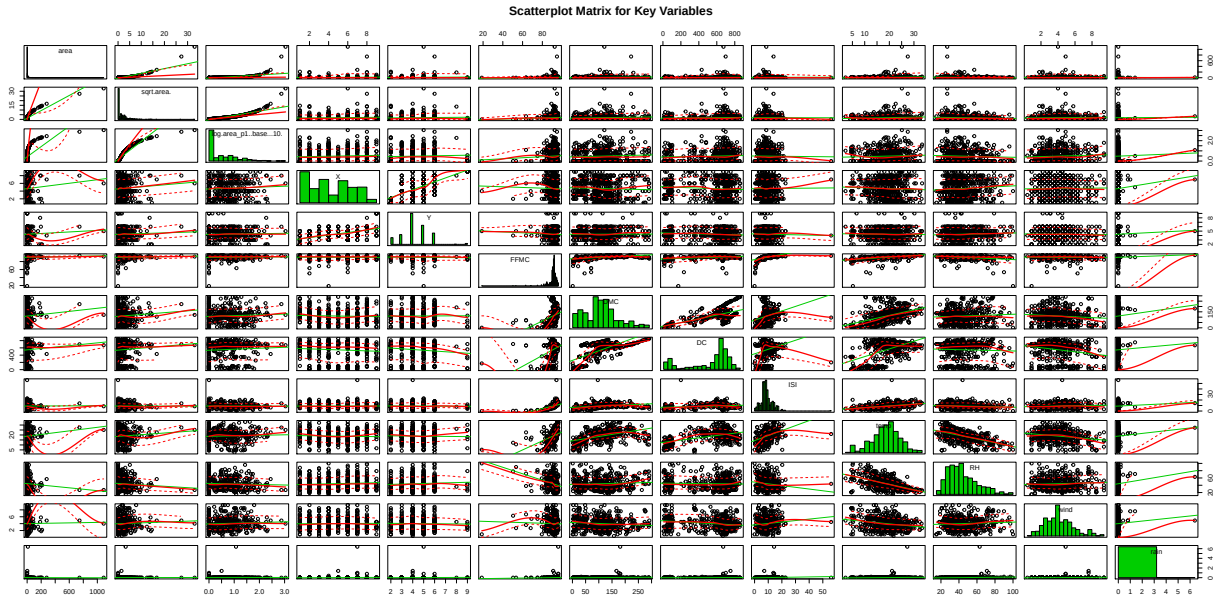


3. Analysis of Key Relationships

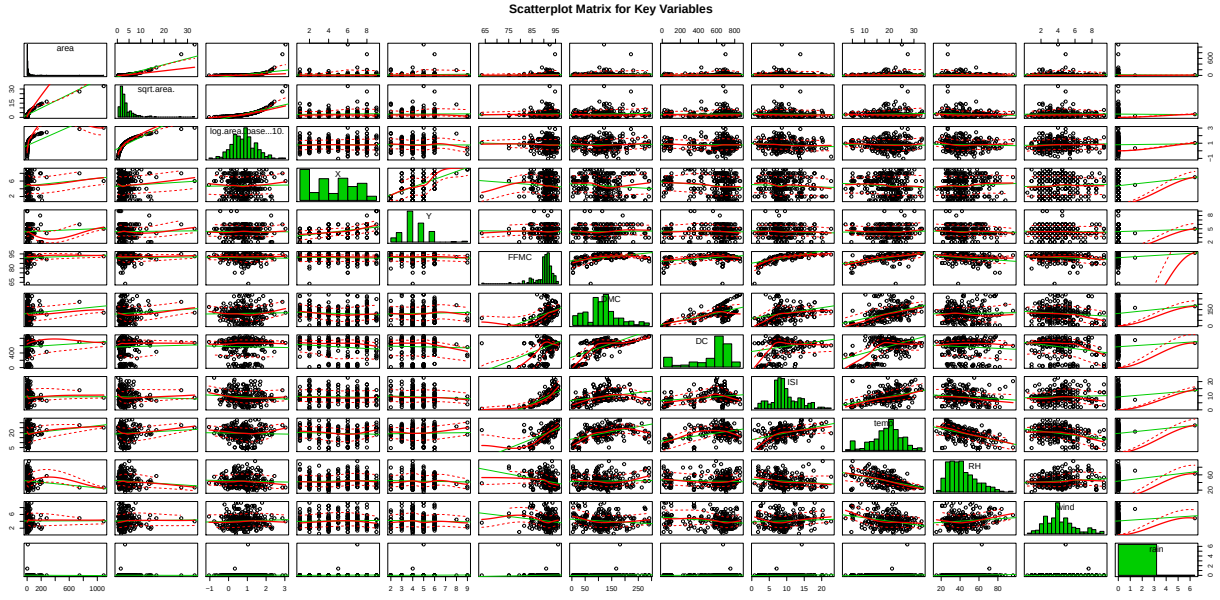
3.1 Linear Bivariate Relationships

We first want to create a scatterplot matrix as a high-level overview of the relationships between our variables. A scatterplot matrix reveals that some of the variables exhibit linear and potentially other non-linear relationships that warrant further exploration. As we noticed in the univariate analysis, Area is positively skewed with majority of the values close to 0. Hence, in order for us to better understand bivariate relationships that Area has with other variables, we repeat the same analysis on both the original dataset (dataset 1) as well as a dataset with all 0 Area excluded (dataset2). Moreover, we also created a logarithm and a square root transformation for the Area variable in both datasets to explore how these transformations, that decrease the positive skewness, affect the bivariate relationship. Again, since log(0) is undefined, we decided to add 1 to all Area values in dataset 1 when calculating their logarithms.

```
## 1. Dataset 1: all data
scatterplotMatrix(~ area + sqrt(area) + log(area_p1, base = 10) +
                    X + Y + FFMC + DMC + DC + ISI +
                    temp + RH + wind + rain, data = df_n,
                    main = "Scatterplot Matrix for Key Variables",
                    diagonal = "histogram")
```




```
## 2. Dataset 2: excluding 0 area
scatterplotMatrix(~ area + sqrt(area) + log(area, base = 10) +
                    X + Y + FFMC + DMC + DC + ISI +
                    temp + RH + wind + rain, data = df_n_no0,
                    main = "Scatterplot Matrix for Key Variables",
                    diagonal = "histogram")
```



Both dataset 1 and dataset 2 demonstrate similar relationships, according to the least-square regression lines. However, looking at the smoothing lines for Area-related scatterplots, we can see that dataset 1 has more noisy data values.

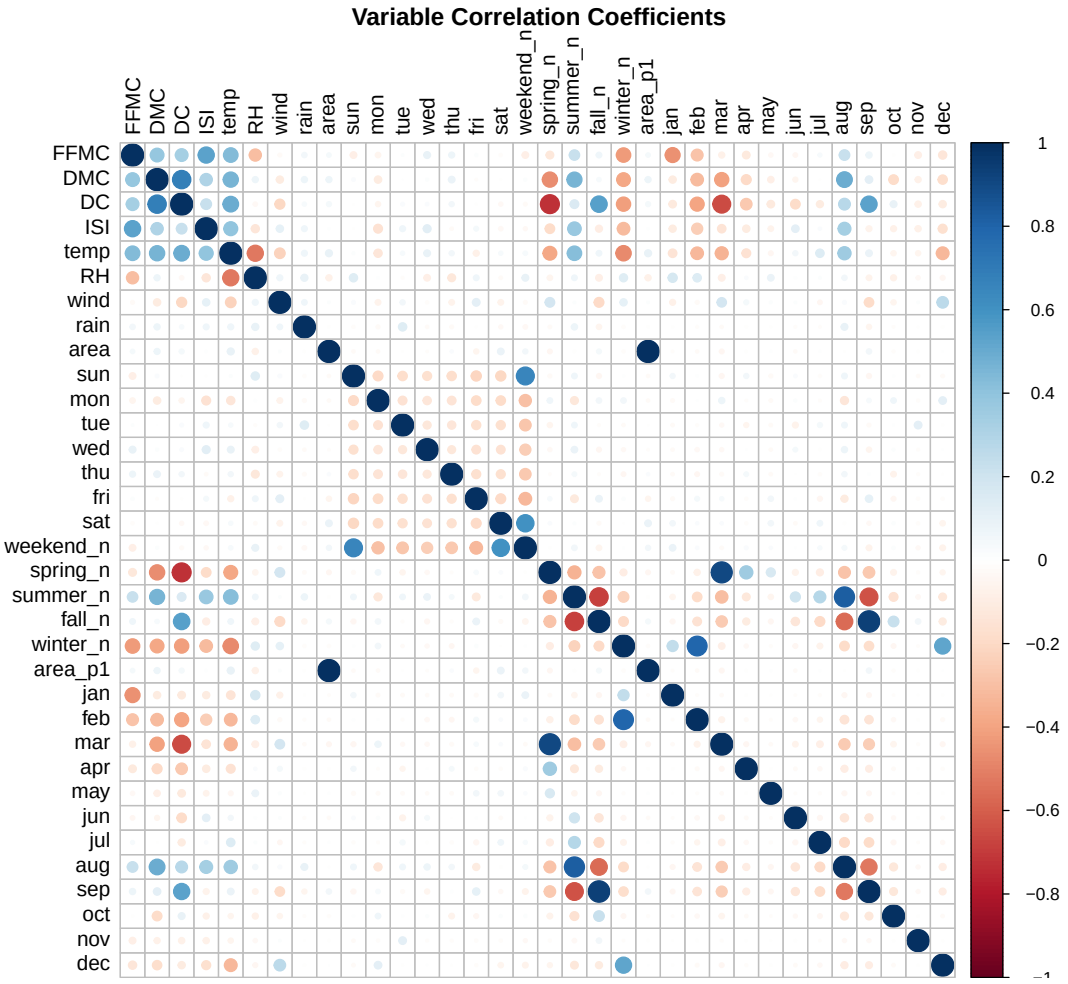
As shown in the graphs above, we can see the following: 1) Neither the square root transformation or logarithm transformation significantly improve any correlation strength. 2) Area has a noticeable positive correlation with X, Y, DMC, DC, and Temp, of which the correlations with DMC and Temp seem to be the

strongest. 3) Area and RH are negatively correlated. 4) Area doesn't show much correlation with FFMC, ISI, Wind or Rain. We note that Rain only has 8 non-zero data points. 5) DMC, DC, and Temp are positively correlated. 6) FFMC, DMC, DC, ISI, and Temp are positively correlated. 7) FFMC and RH show a relatively strong negative correlation.

By confirming that dataset 2 demonstrates very similar and less noisy relationships, we can likely treat dataset 2 as a useful subset in this EDA. However, since the underlying meaning of the 0 Area values is unclear, we want to include analyses on both dataset 1 and 2 to make sure that we are not skewing our results.

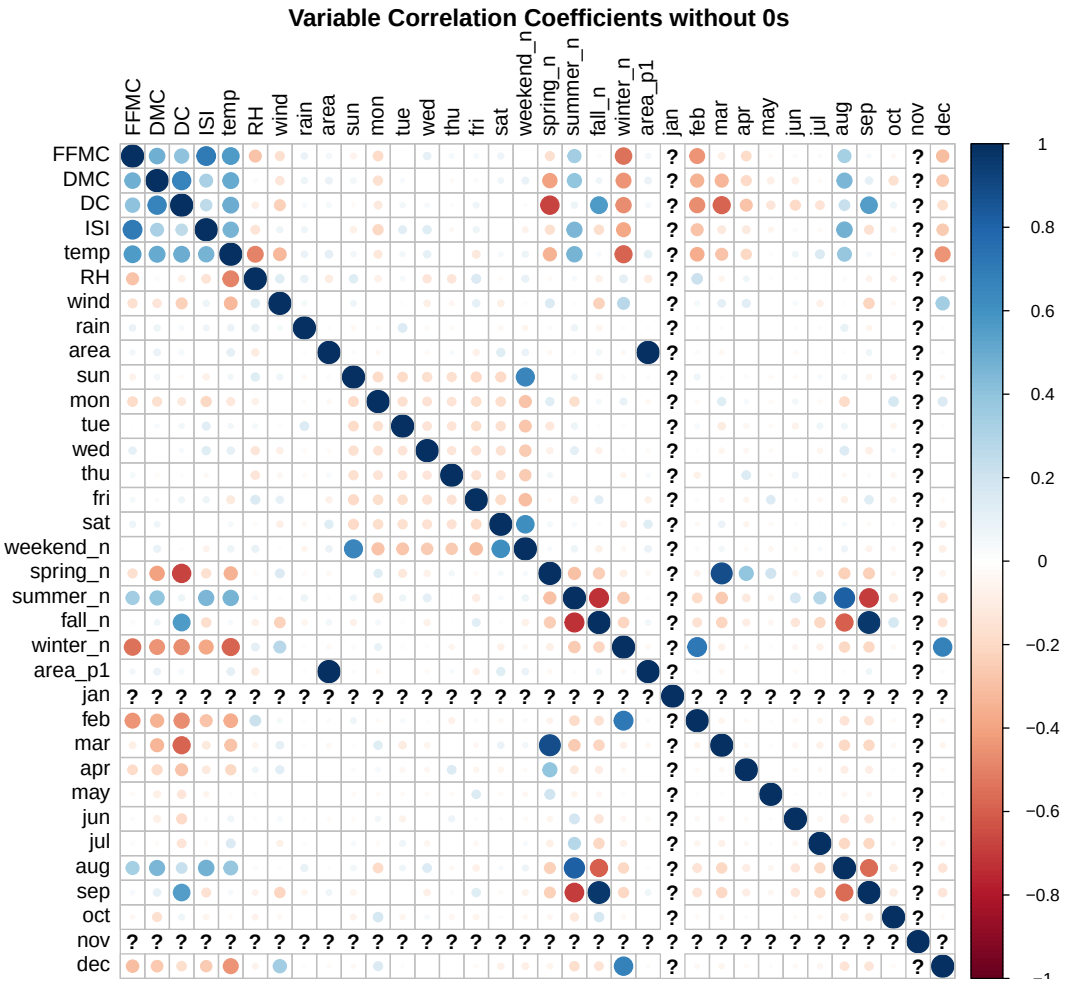
Next, we want to quantify the strengths of the correlations shown in the scatterplot matrices above for both datasets.

```
corrplot(cor(df_n[,c(5:20, 22, 24:27,30:42)]),
         title = "Variable Correlation Coefficients", tl.col = "black", mar=c(1,0,1,0))
```




```
corrplot(cor(df_n_no0[,c(5:20, 22, 24:27,30:42)]),
         title = "Variable Correlation Coefficients without 0s", tl.col = "black", mar=c(1,0,1,0))
```

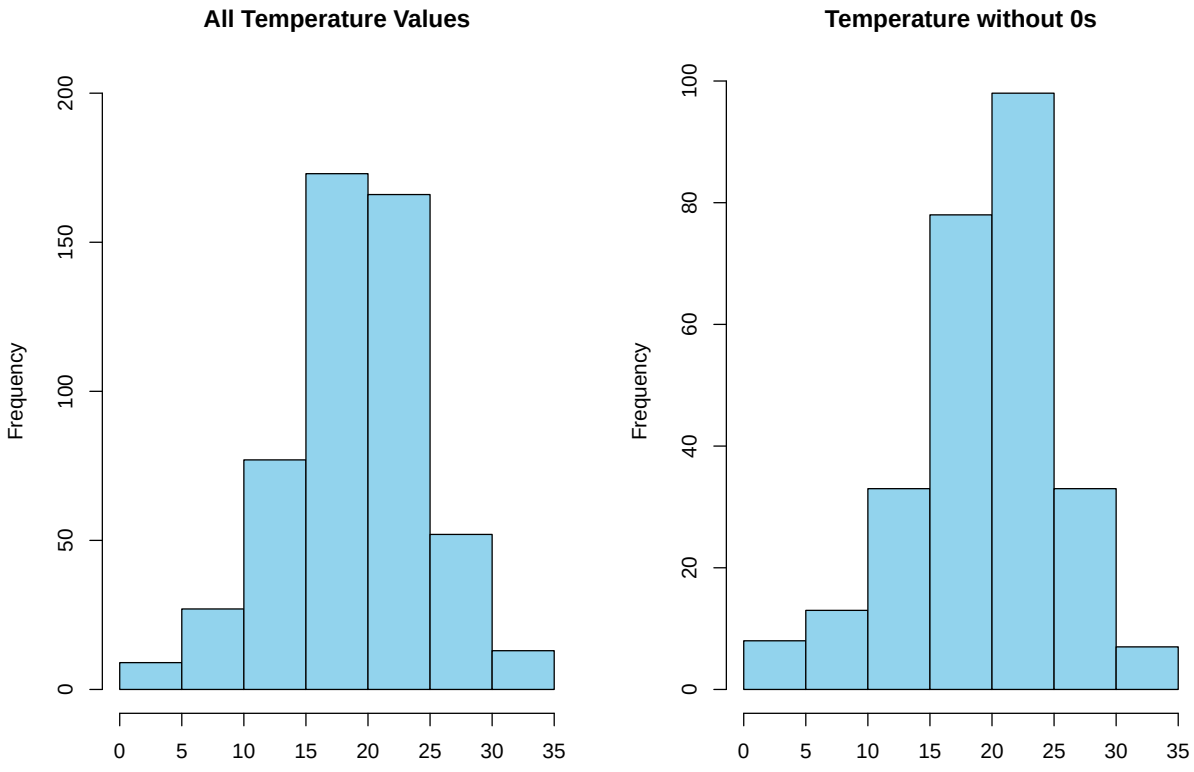
```
## Warning in cor(df_n_no0[, c(5:20, 22, 24:27, 30:42)]): the standard
## deviation is zero
```



The correlation coefficients above validate the relationships we noted from the scatterplots. As we can see, the strongest correlation that Area has, besides its own transformations, is with Temperature. However, in both datasets, the correlation coefficients between Area and Temperature are merely around 0.10. It's possible that we don't have enough data points to demonstrate stronger correlations related to Area, but the strong correlations among Temp and FFMC, DMC, DC, ISI, RH, and wind mean that these variables could also affect the bivariate relationships we see.

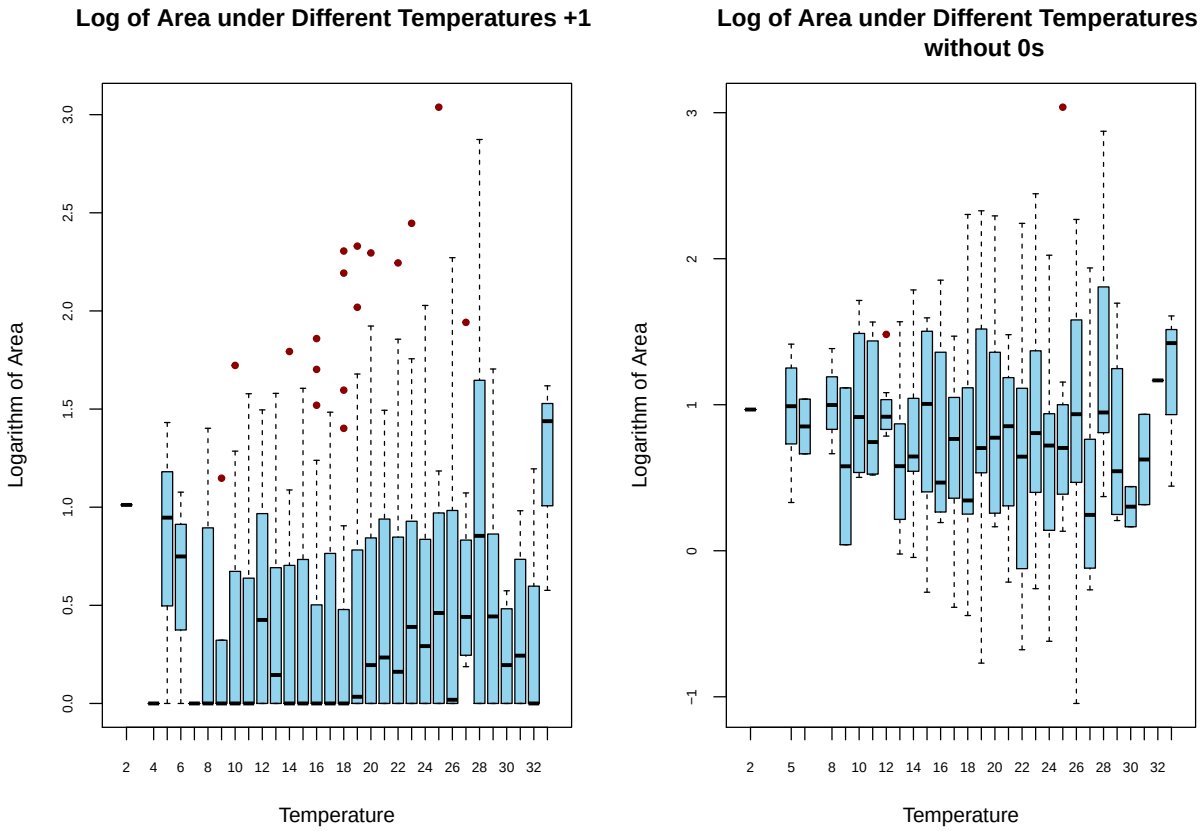
Let's look closely at the relationship between Area and Temp, the strongest correlated variable according to the analysis above.

```
layout(matrix(c(1,2), 1, 2))
hist(df_n$temp, xlab="", main="All Temperature Values", col = "#92d3ed", ylim=c(0,200))
hist(df_n_no0$temp, xlab="", main="Temperature without 0s", col = "#92d3ed")
```



As we have seen in the univariate analysis, we don't have many data points for Temp lower than 10 degrees or higher than 30 degrees. We need to keep that in mind when looking at the bivariate relationship. In order to have a cleaner box plot graph, we round temperature readings to their closest integer.

```
layout(matrix(c(1,2), 1, 2))
boxplot(log(area_p1, base= 10) ~ round(temp, 0), data = df_n, # log transformation on area, quotient fo
at = sort(unique(round(df_n$temp,0))), cex.axis = .7,
xlab = "Temperature", ylab = "Logarithm of Area",
main = "Log of Area under Different Temperatures +1 \n",
col = "#92d3ed", outcol = "darkred", outpch = 20)
boxplot(log(area, base= 10) ~ round(temp,0), data = df_n_no0, # log transformation on area, quotient fo
at = sort(unique(round(df_n_no0$temp,0))), cex.axis = .7,
xlab = "Temperature", ylab = "Logarithm of Area",
main = "Log of Area under Different Temperatures \n       without 0s",
col = "#92d3ed", outcol = "darkred", outpch = 20)
```



In order to accommodate a few large outliers in the visualizations, we used the log transformation of Area (log(area+1) for dataset 1, log(area) for dataset 2). If we focus on Area values in Temperature range of [20 degrees, 30 degrees], we see a slight positive correlation, confirming what we noted in the previous analyses.

3.2 Nonlinear Bivariate Relationships

So far, we have only looked at the linear bivariate relationships that Area has with other variables. However, according to both our understanding of the variables and the scatterplots we see above, there could be some nonlinear bivariate relationships that we haven't yet explored.

3.2.1 Month

We now dive deeper into the relationship between Month and Area.

```
#Dataset 1
summary(df_n$month)
```

```
## jan feb mar apr may jun jul aug sep oct nov dec
##   2  20  54   9   2  17  32 184 172  15   1   9
```

```
#Dataset 2
summary(df_n_no0$month)
```

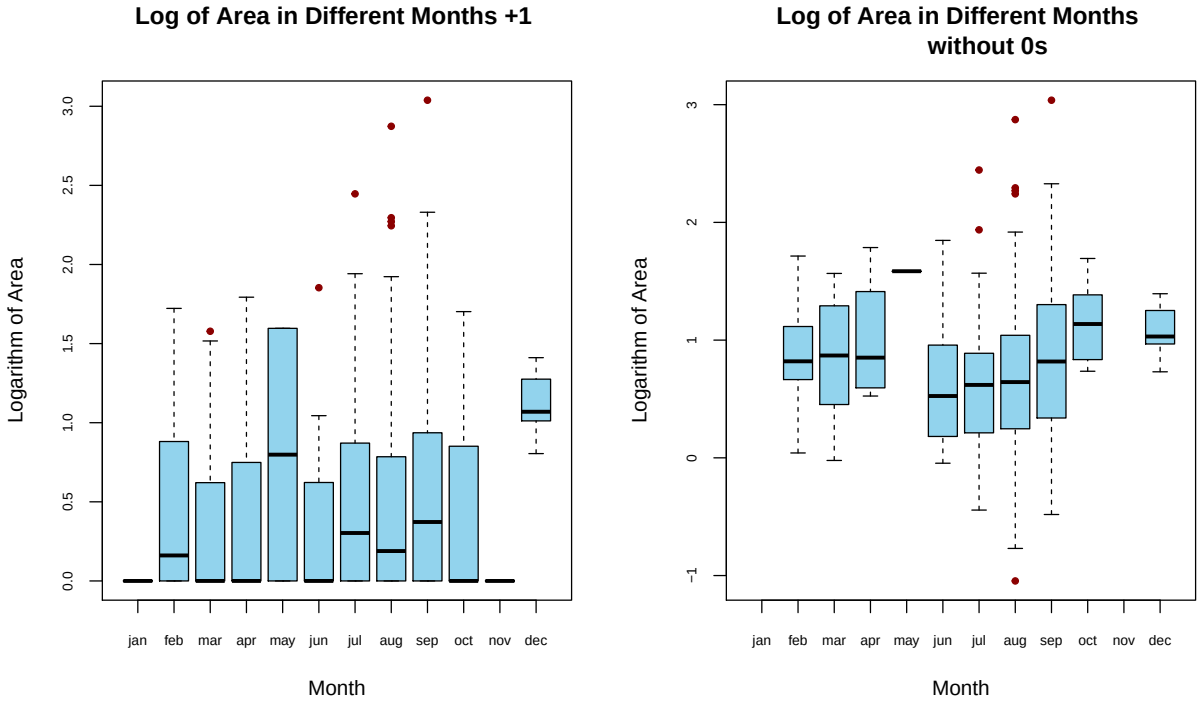
```
## jan feb mar apr may jun jul aug sep oct nov dec
##   0  10  19   4   1   8  18  99  97   5   0   9
```

The Month data points aren't distributed evenly. Note that there are less than 10 observations for Jan, Apr, May, Nov, and Dec in the original sample (dataset 1) and after excluding the 0 Area values in dataset 2, we

don't have any observations in Jan or Nov.

If this sample is taken in a time frame that's a multiple of one year and captures all fires in the region, then we can see fires that cause Area to be greater than 0 happen most frequently in Aug and Sept, frequently in Mar and Jul and less frequently in other months. We should keep this in mind when interpreting the relationship between Month and Area.

```
layout(matrix(c(1,2), 1, 2))
# Dataset 1
boxplot(log(area_p1, base = 10) ~ month, data = df_n, # log transformation on area,
cex.axis = .7, xlab = "Month", ylab = "Logarithm of Area",
main = "Log of Area in Different Months +1 \n",
col = "#92d3ed", outcol = "darkred", outpch = 20)
# Dataset 2
boxplot(log(area, base = 10) ~ month, data = df_n_no0, # log transformation on area,
cex.axis = .7, xlab = "Month", ylab = "Logarithm of Area",
main = "Log of Area in Different Months \n        without 0s",
col = "#92d3ed", outcol = "darkred", outpch = 20)
```



May shows the highest median, but since it only has 2 observations in dataset 1, and 1 in dataset 2, it is unlikely to be a significant finding. With so few observations in certain months, the comparison of the medians across different months doesn't reveal much, but we can see that all fire with burned Area >100 hectares happened in Jul, Aug, and Sep.

3.2.2 Day

Similar to Month, Day might have nonlinear relationship with Area.

```
#Dataset 1
summary(df_n$day)
```

```
## mon tue wed thu fri sat sun
```

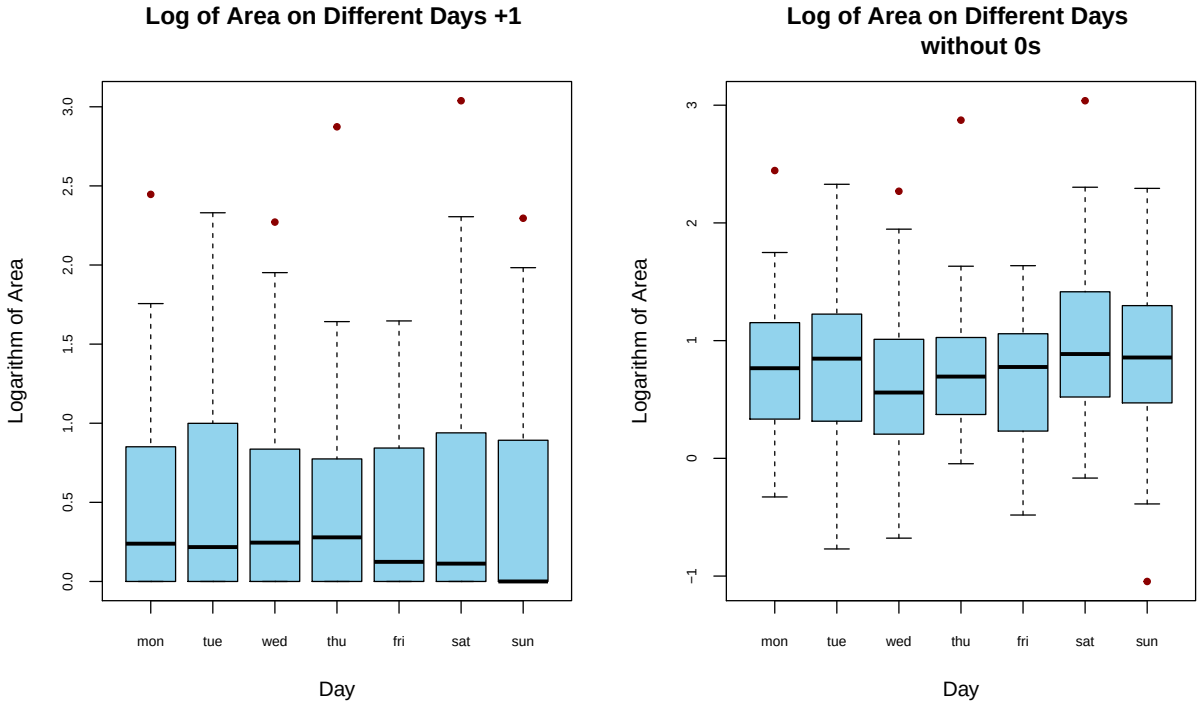
```
##  74  64  54  61  85  84  95
```

```
#Dataset 2
summary(df_n_no0$day)
```

```
## mon tue wed thu fri sat sun
##  39  36  32  31  43  42  47
```

Unlike Month, the data points for Day are distributed more evenly. This can be helpful if we try to compare the medians of the burned area on different days.

```
layout(matrix(c(1,2), 1, 2))
# Dataset 1
boxplot(log(area_p1, base = 10) ~ day, data = df_n, # log transformation on area,
cex.axis = .7, xlab = "Day", ylab = "Logarithm of Area",
main = "Log of Area on Different Days +1 \n",
col = "#92d3ed", outcol = "darkred", outpch = 20)
# Dataset 2
boxplot(log(area, base = 10) ~ day, data = df_n_no0, # log transformation on area,
cex.axis = .7, xlab = "Day", ylab = "Logarithm of Area",
main = "Log of Area on Different Days \n      without 0s",
col = "#92d3ed", outcol = "darkred", outpch = 20)
```



We used similar log transformations here for the visualizations. From the box plots, we can't see a strong bivariate relationship between Area and Day as the medians for different days appear to be pretty close.

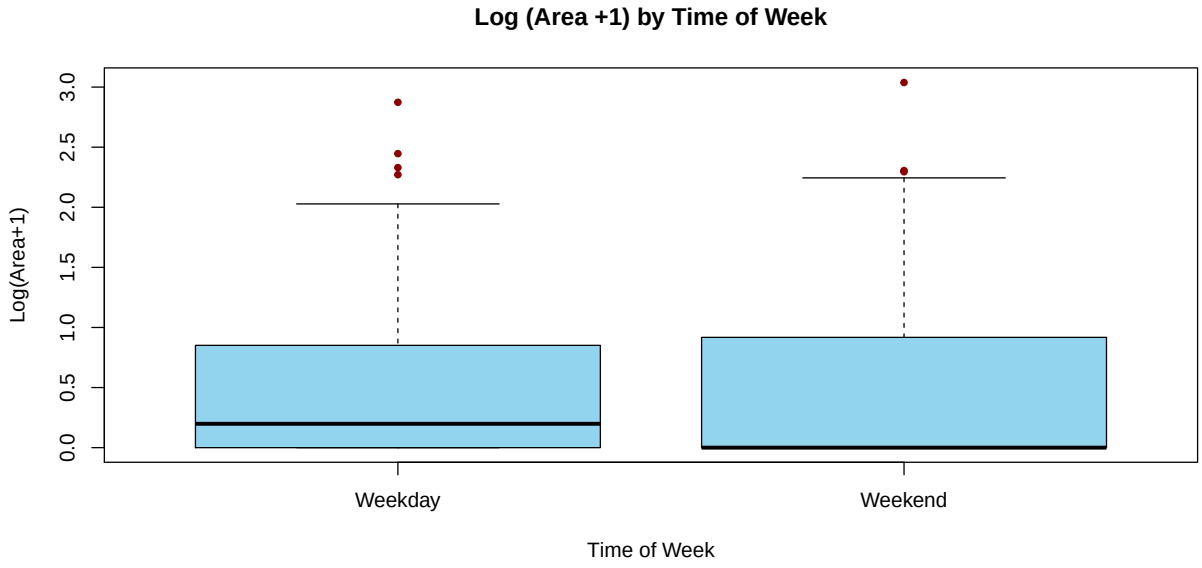
4. Analysis of Secondary Effects

4.1 Weekday v. Weekend Effects

When examining secondary effects, we first looked at our weekday data through a slightly modified lens. Using a dummy variable to assign observations to either a weekday or weekend category, we plotted these

categories against log(area+1). Looking at the plot below, we see that weekday observations show a higher median of log(area+1) than weekend observations. We could speculate that this may be due to differences in human activity in the park, which may be higher during the week versus weekend (or vice versa). However, since we don't have information regarding trends in park visits, we cannot say for certain what may be driving this relationship. It is worth noting that there are nearly double the amount of weekday observations as weekend, which is generally in line with expectations if sampling from the week randomly.

```
boxplot(log(area_p1, base = 10) ~ time.of.week, data = df_n,
        main = "Log (Area +1) by Time of Week", xlab = "Time of Week", ylab = "Log(Area+1)",
        col = "#92d3ed", outcol = "darkred", outpch = 20)
```




```
table(df_n$time.of.week)
```

```
##
## Weekday Weekend
##     338     179
```

4.2 Seasonal Effects

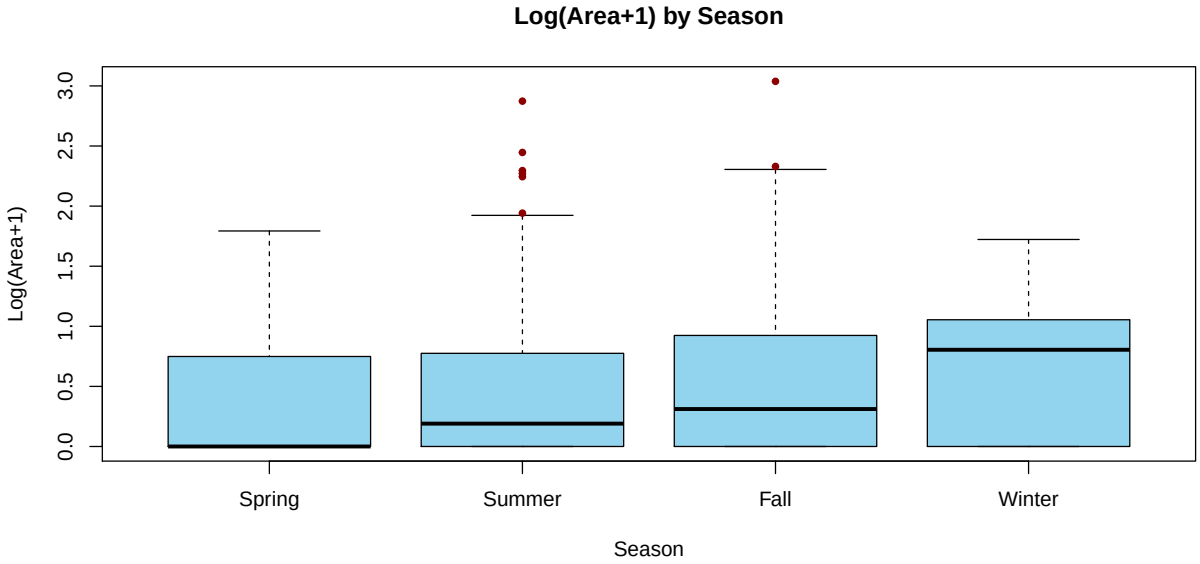
Next, we took a similar approach and assigned each month value to seasonal categories. As a reminder:

March, April, May = Spring, June, July, August = Summer, September, October, November = Fall, December, January, and February = Winter.

Plotting these categories against log(area+1) produced the chart below. Looking at the plot we can see some apparent positive trends in log(area+1) median with season throughout the year. We could postulate that this might be due to things like seasonal weather patterns or seasonal human activity. However, it is also important to note that Winter only has 31 total observations, so the sample size of that category may not be meaningful. Additionally, the majority of oberservations within Summer occur in August, and the large majority of observations in Fall occur in September. If we had defined our seasons differently, these trends could differ.

```
df_n$season <- factor(df_n$season, levels = c("spring", "summer", "fall", "winter"),
                      labels = c("Spring", "Summer", "Fall", "Winter"))
boxplot(log(area_p1, base=10) ~ season, data=df_n,
```

```
        main = "Log(Area+1) by Season", xlab = "Season", ylab = "Log(Area+1)",
        col = "#92d3ed", outcol = "darkred", outpch = 20)
```




```
table(df_n$season)
```

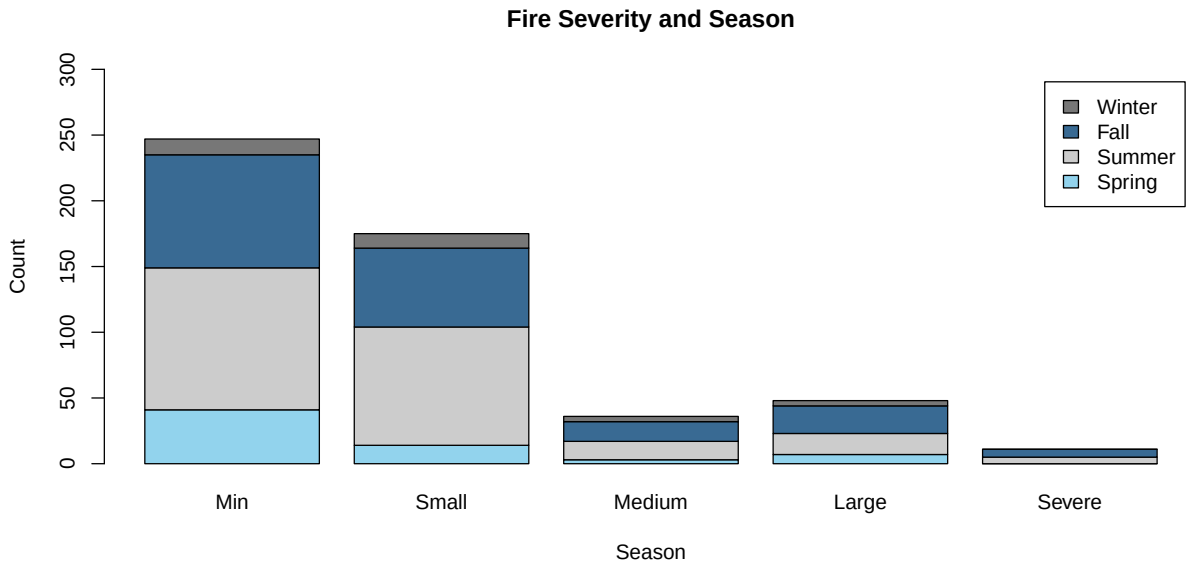
```
##
## Spring Summer   Fall Winter
##     65    233    188     31
```

```
table(df_n$month)
```

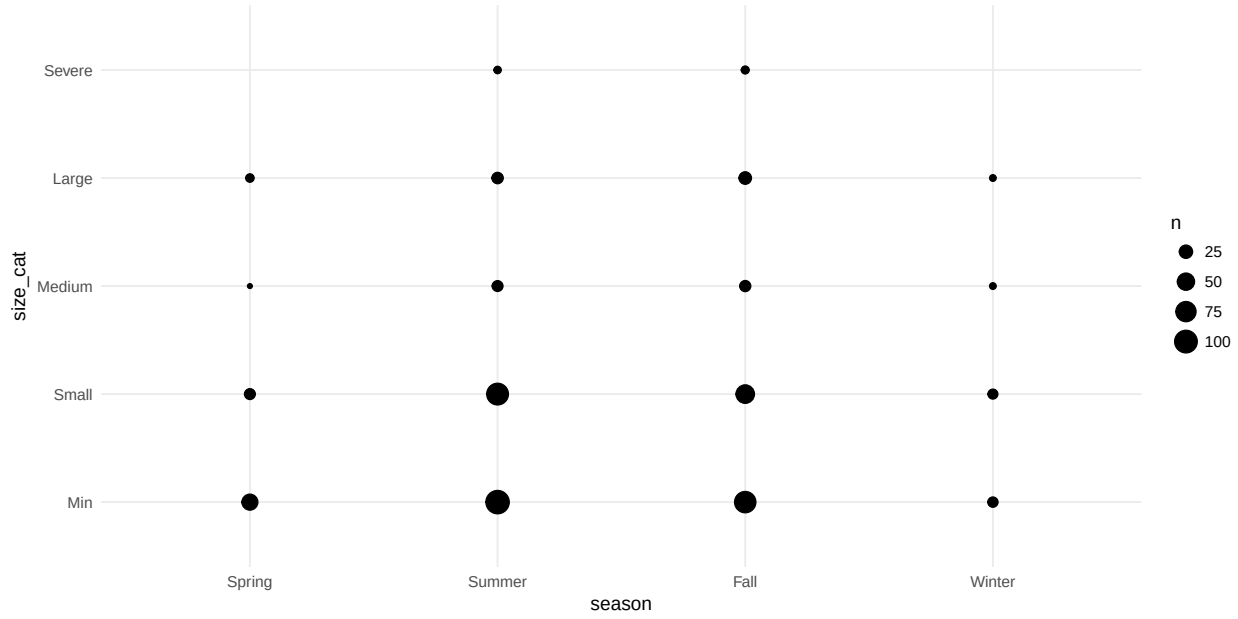
```
##
## jan feb mar apr may jun jul aug sep oct nov dec
##   2  20  54   9   2  17  32 184 172  15   1   9
```

To examine this trend further, we also produced the following graphs, where we can more clearly see the most large and severe fires occurring in the Summer and Fall.

```
df_n$size_cat <- factor(df_n$size_cat, levels = c("Min", "Small", "Medium", "Large", "Severe"))
seasoncat<-table(df_n$season, df_n$size_cat)
barplot(as.matrix(seasoncat),
        main = "Fire Severity and Season", xlab = "Season", ylab = "Count", ylim = c(0, 300),
        col = c("#92d3ed", "#cccccc", "#396a93", "#777777"), legend = rownames(seasoncat))
```




```
ggplot(df_n) + geom_count(aes(season, size_cat)) + theme_minimal()
```

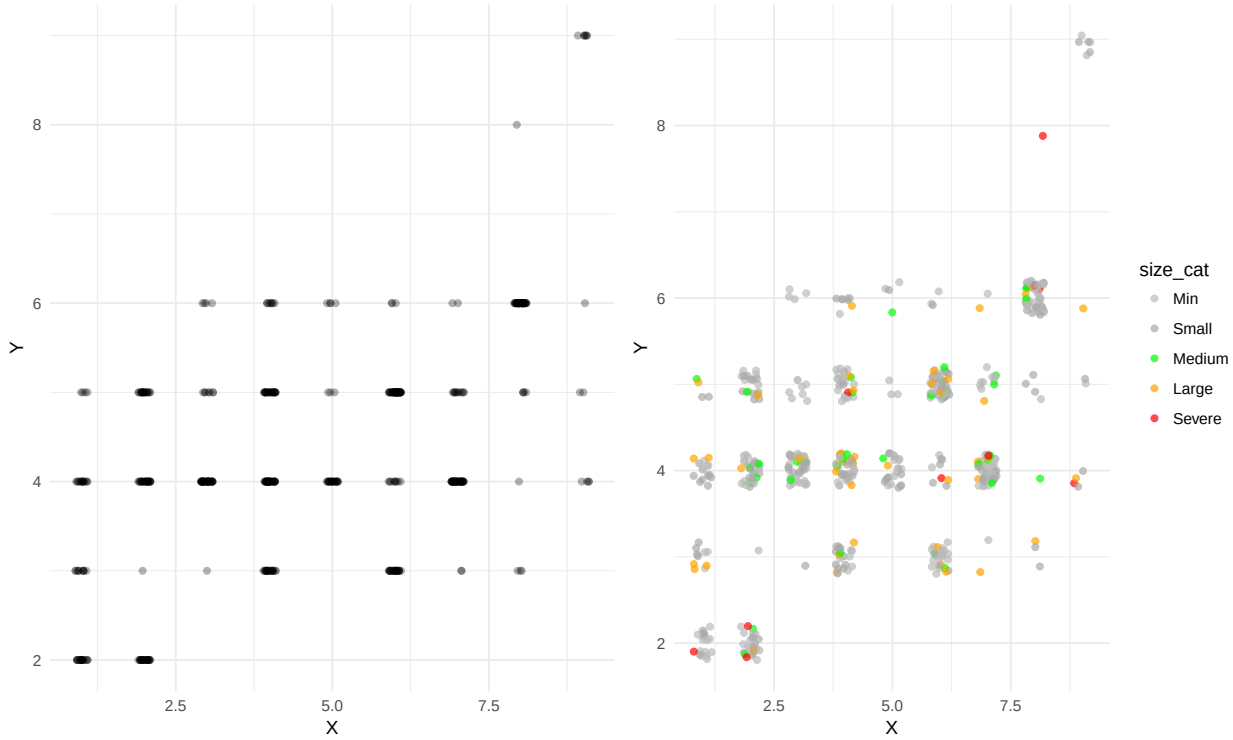


4.3 Regional Effects

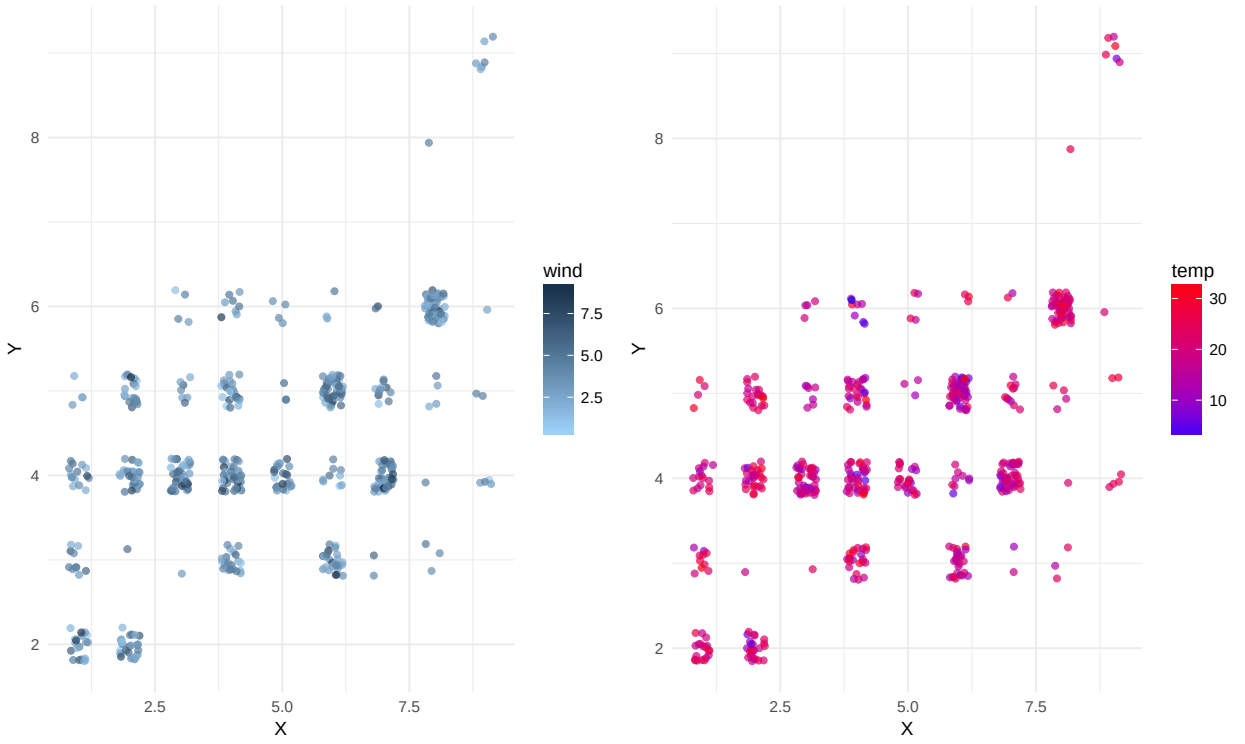
Lastly, we looked at the association of X,Y coordinates together with fire occurance to see if there are any regions within the park that might trend with severity. Looking at the plots below, we do see that there are pockets within the park with greater concentrations of fire counts. Using the size_cat variable, it appears that large and severe fires occur equally within these regions. This would lead us to hypothesize that perhaps these regions are more forested or share a similar density of trees, or share common characteristics (e.g. Wind and Temp), but without knowing anything about the layout of the park we can't be certain.

```
p1 <- ggplot(df_n) + geom_point(aes(X, Y), alpha=0.3, position = position_jitter(.1,0)) +
      theme_minimal()
p2 <- ggplot(df_n) +
```

```
      geom_point(aes(X, Y, color=size_cat), alpha=0.7, position = position_jitter(.2,.2)) +
      scale_color_manual(values=c("#bbbbbb", "#aaaaaa", "green", "orange", "red")) +
      theme_minimal()
p3 <- ggplot(df_n) +
      geom_point(aes(X, Y, color=wind), alpha=0.7, position = position_jitter(.2,.2)) +
      scale_color_gradient(low = "#9ed2fa", high = "#132B43",
      space = "Lab", na.value = "grey50", guide = "colourbar") + theme_minimal()
p4 <- ggplot(df_n) +
      geom_point(aes(X, Y, color=temp), alpha=0.7, position = position_jitter(.2,.2)) +
      scale_color_gradient(low = "blue", high = "#ff0000",
      space = "Lab", na.value = "grey50", guide = "colourbar") + theme_minimal()
grid.newpage()
pushViewport(viewport(layout = grid.layout(1, 2)))
print(p1, vp = viewport(layout.pos.row = 1,layout.pos.col = 1))
print(p2, vp = viewport(layout.pos.row = 1,layout.pos.col = 2))
```




```
grid.newpage()
pushViewport(viewport(layout = grid.layout(1, 2)))
print(p3, vp = viewport(layout.pos.row = 1,layout.pos.col = 1))
print(p4, vp = viewport(layout.pos.row = 1,layout.pos.col = 2))
```



5. Conclusion

Our analysis didn't reveal any strong correlations within the sample set related to our key variable. Whether the lack of correlation is due to the sample set or the measurement methods used with the variables is not made clear by the data set. Moreover, the data set often shows counter-intuitive results. As an example, it could be expected that temperature and wind play a role in fire severity. However, examining temperature and wind more closely shows larger fires are not especially correlated to these factors.

We suspect that secondary factors and the methods of measure may be playing a sigificant role. Specifically, we believe access to information like the cause of fire, other environmental factors, the proximity of fire suppression services, and levels of human activity may play a role. In addition, information related to the sampling method and methods of measure would help to elucidate the relationships within the data. As an example, the observations are more heavily represented in certain days of the week or in certain seasons. Although it can't be ascertained from the data whether this is due to the sampling methods used, it is plausible that the increased average count seen during weekends could be associated to increased human activity. Similarly, seasonality could be associated to drier weather or increased fuel availability, which are components of fire risk.

Cumulatively, it is difficult to assess whether fires in the region are well-represented by the data set. Accordingly, it would be difficult to predict the most severe fires. Notably, the inverse problem may be fairly easy to predict given that most of the fires are very small.