Lab 1 - Forest Fires EDA

W203 Statistics for Data Science

Group 4 - Daniel Alvarez, Anup Jha, Mumin Khan, Peter Wang

September 25th, 2018

Contents

Introduction

This analysis is motivated by the following main research question:

> What factors or combination of factors lead to forest fires with the largest area burned?

Other corollary questions from this exploration are:

- Are there any salient observations we can surmise from the data?
- Are there data issues uncovered and how do we deal with them?
- Are there confounding effects on the relationships identified?

We will address this question using exploratory data analysis techniques.

Setup

Loading our data into R:

```r
fires <- read.csv("forestfires.csv")
# Note how many columns and rows in the dataset
dim(fires)
```

```
## [1] 517  13
```

```r
nrow(fires)
```

```
## [1] 517
```

```r
#show all variables
str(fires)
```

```
## '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 ...
```

```r
# Check if there are any null values
fires_naomit = na.omit(fires)
dim(fires_naomit)
```

```
## [1] 517  13
```

There are 517 observations with no missing values in the dataset. It is described in the table below.

Variable	Type	Description
X	categorical	x-axis spatial coordinate within the Montesinho park map: 1 to 9
Y	categorical	y-axis spatial coordinate within the Montesinho park map: 2 to 9
month	categorical	month of the year: January to December
day	categorical	day of the week: Monday to Sunday
FFMC	metric	Fine Fuel Moisture Code (FFMC) index from the FWI system: 18.7 to 96.20
DMC	metric	Duff Moisture Code (DMC) index from the FWI system: 1.1 to 291.3
DC	metric	Drought Code (DC) index from the FWI system: 7.9 to 860.6
ISI	metric	Initial Spread Index (ISI) from the FWI system: 0.0 to 56.10
temp	metric	Temperature in Celsius degrees: 2.2 to 33.30
RH	metric	Relative humidity in %: 15.0 to 100
wind	metric	Wind speed in km/h: 0.40 to 9.40
rain	metric	Outside rain in mm/m2 : 0.0 to 6.4
area	metric	The burned area of the forest (in hectares): 0.00 to 1090.84

```r
# Get a summary of the variables
summary(fires)
```

```
##        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  
## 
```

The summary statistics reveal a few interesting observations:

1. The rain variable mostly contains zeros.
2. The area variable contains many zero values.
3. Most of the variables appear to be positively skewed with the mean greater than the median (except for the FFMC and DC variable).

Restricting the Data Set

Given our research objective about understanding factors that lead to particularly damaging forest fires, we decided to restrict our dataset to just those observations with non-zero values of the area variable. We find that there are 270 observations of the area variable greater than zero of the total 517 observations in the data set. We can assume that the zero values of the area variable were either inaccurately reported by the data collectors or are not meaningful measures of interest for our research objective. We proceed with the analysis with the dataset with observations where there are non-zero values of the area variable.

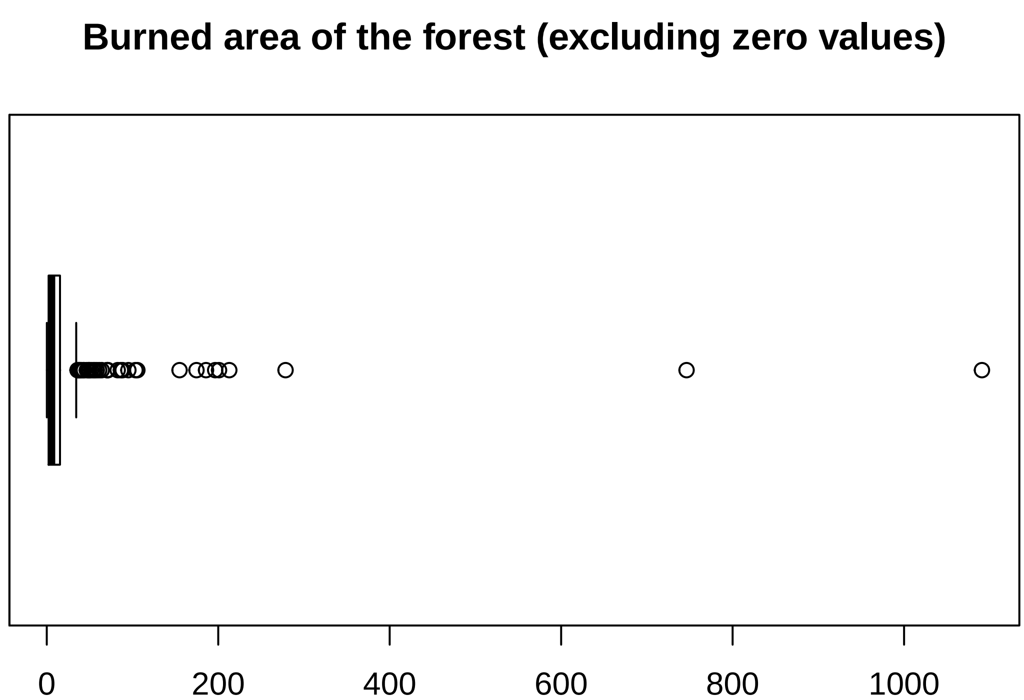
```
#Create dataset with nonzero values of area
fires_nonzeroarea <- fires[fires$area>0,]
```

To further examine the distribution of each of the variables, we produced plots of each variable and describe findings below.

Univariate Analysis of Key Variables

Univariate analysis of the Area Variable

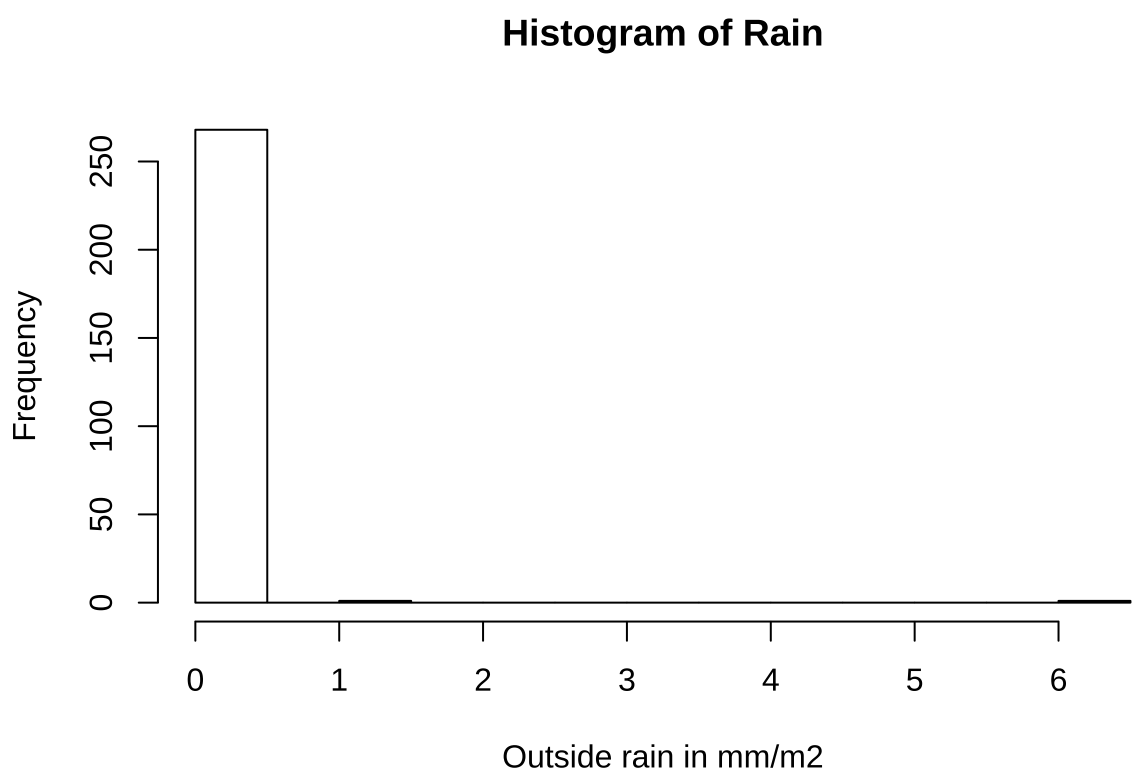
```
# area variable (excluding zero values)
boxplot(fires$area[fires$area>0], main = "Burned area of the forest (excluding zero values)",
        main = "Area in hectares", horizontal = TRUE)
axis(side=2, at=seq(0,max(fires$area[fires$area>0]), 100),
     labels=seq(0,max(fires$area[fires$area>0]),100))
```



We examined the distribution of the area variable before and after removing the non-zero values using the boxplot illustrated above. In both datasets, we find evidence of a highly right-skewed (positive-skewed) distribution with the mode and bulk of the variables clustered around zero. We observe outliers in the extreme right of the distribution of the Area variable, indicating that fires might spread exponentially. We can conclude that the area variable is a candidate for a log transformation.

Univariate analysis of the Rain Variable

```
# Naive Histogram of rain variable
hist(fires_nonzeroarea$rain, xaxt='n', main = "Histogram of Rain", xlab="Outside rain in mm/m2")
axis(1, at=0:max(fires_nonzeroarea$rain))
```



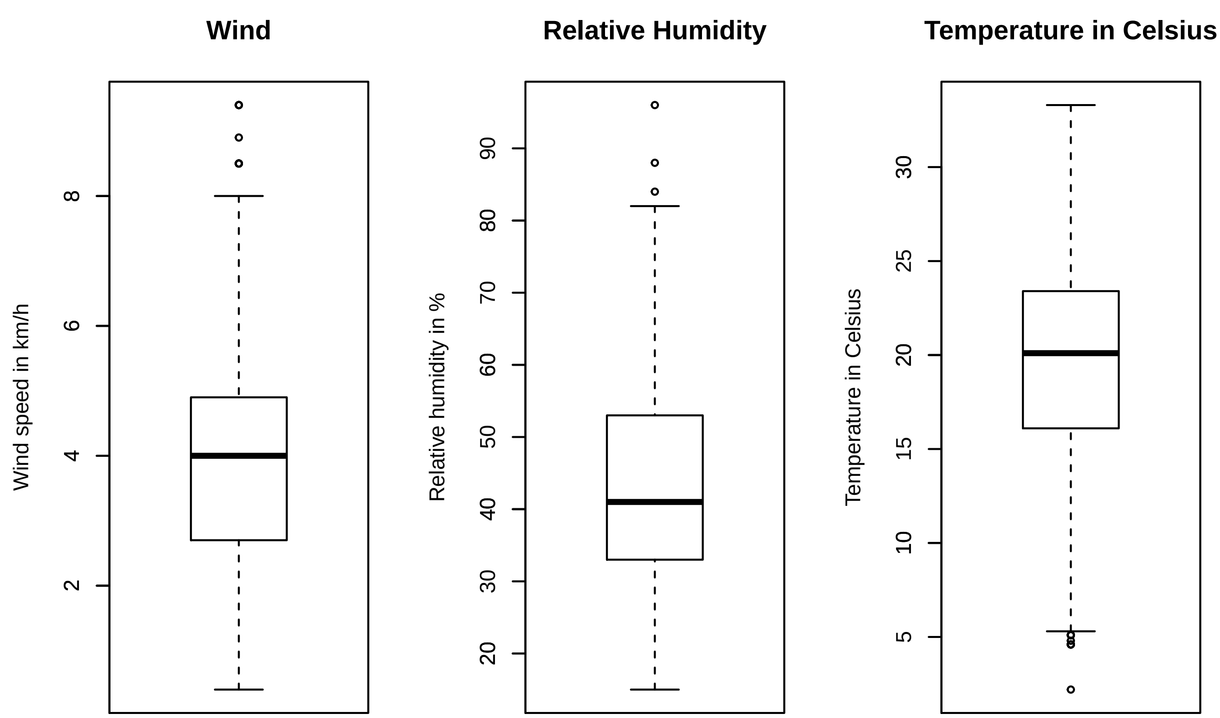
The histogram of the Rain variable provides evidence of a highly right-skewed (positive-skewed) distribution with the mode and bulk of the variables clustered at zero. We observe a few outliers in the extreme right of the distribution of rain. Moreover, we find that there are only 2 observations of the Rain variable greater than zero (following the restriction of the dataset to non-zero Area values). This might imply that values of the Rain variable were not entered by the data collectors on a consistent basis. We can consider removing the rain variable from the analysis as there are very few meaningful observations.

Univariate Analysis of the Wind, Relative Humidity, and Temperature Variables

```
# 3 boxplot figures for Wind, RH and Temp variables arranged in one row
par("mar")
```

```
## [1] 5.1 4.1 4.1 2.1
```

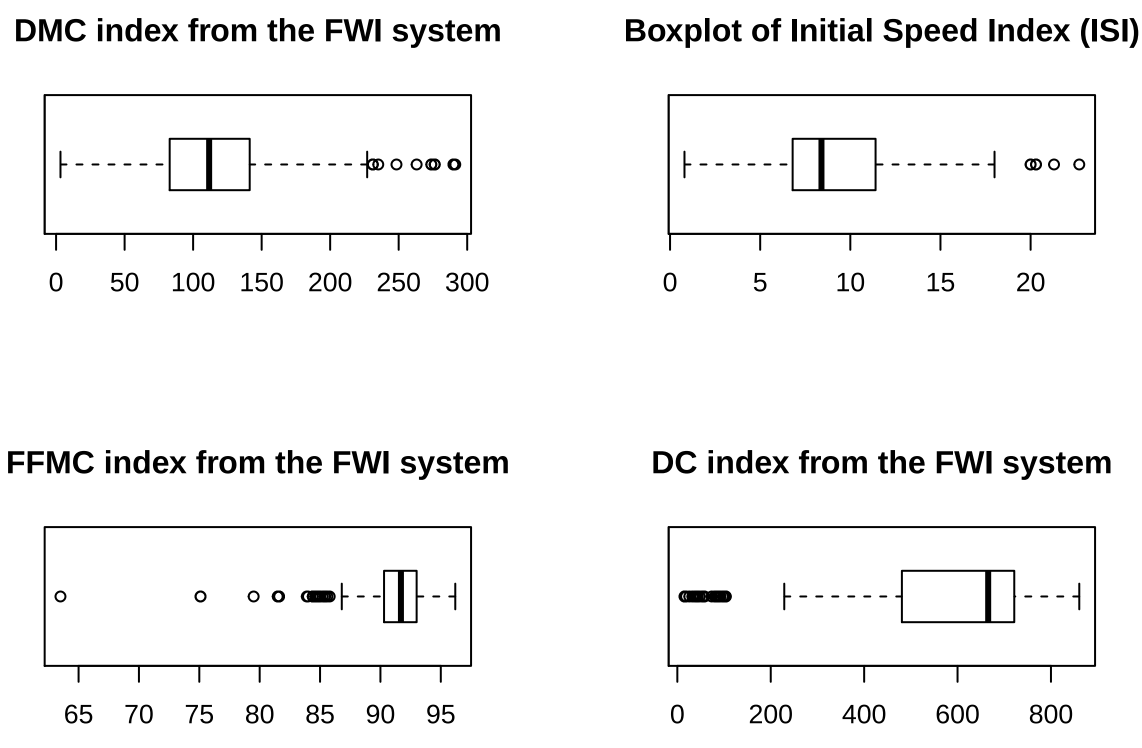
```r
par(mfrow=c(1,3))

# First row - Wind variable
### Boxplot
boxplot(fires_nonzeroarea$wind, main = "Wind", ylab="Wind speed in km/h")
axis(side=2, at=seq(0,max(fires_nonzeroarea$wind), 2), labels=seq(0,max(fires_nonzeroarea$wind),2))

# Second row - RH variable
### Boxplot
boxplot(fires_nonzeroarea$RH, main = "Relative Humidity", ylab="Relative humidity in %")
axis(side=2, at=seq(0,max(fires_nonzeroarea$RH), 10), labels=seq(0,max(fires_nonzeroarea$RH),10))

# Third row - Temp variable
### Boxplot
boxplot(fires_nonzeroarea$temp, main = "Temperature in Celsius", ylab="Temperature in Celsius")
axis(side=2, at=seq(0,max(fires_nonzeroarea$temp), 10), labels=seq(0,max(fires_nonzeroarea$temp),10))
```



As shown in the boxplots above, the wind and relative humidity variables are approximately normally distributed with a few outliers in the extreme right of the distibution. Conversely, the temperature variable has outliers to the left of the distibution. All three variables contain only non-zero, positive values. There is little concern with non-reporting of the variable and no compelling need to transform any of these variables.

Univariate Analysis of the DMC, DC, FFMC, ISI Variables

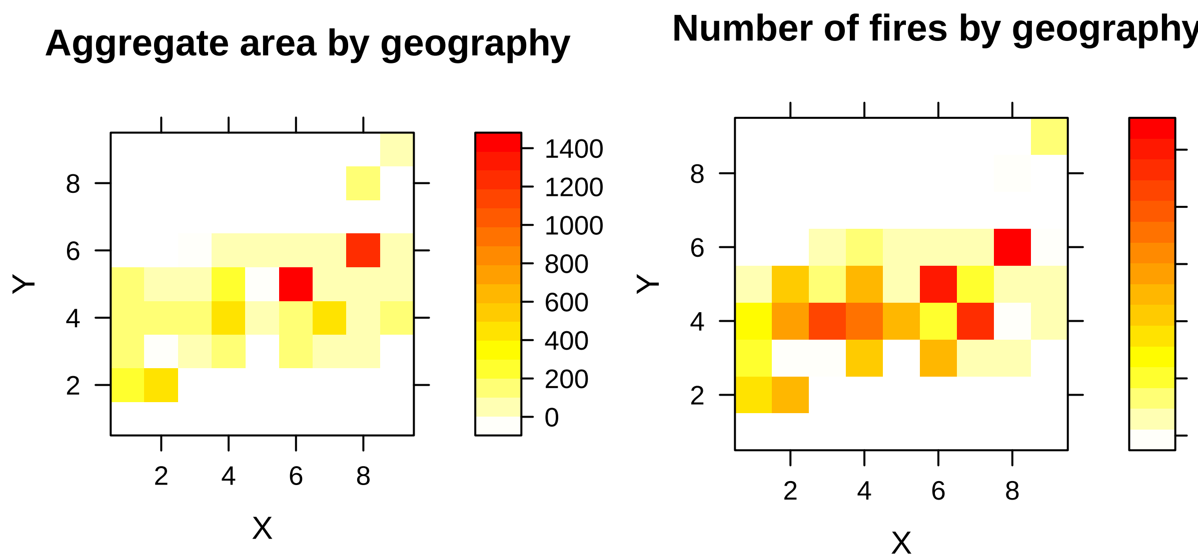
```r
par(mfrow=c(2,2))

# DMC variable
boxplot(fires_nonzeroarea$DMC, main = "DMC index from the FWI system", horizontal = TRUE)
axis(side=2, at=seq(0,max(fires_nonzeroarea$DMC), 100), labels=seq(0,max(fires_nonzeroarea$DMC),100))

# Boxplot of ISI variable
boxplot(fires_nonzeroarea$ISI, main = "Boxplot of Initial Speed Index (ISI)",horizontal = TRUE)
axis(side=2, at=seq(0,max(fires_nonzeroarea$ISI), 10), labels=seq(0,max(fires_nonzeroarea$ISI),10))

# FFMC variable
boxplot(fires_nonzeroarea$FFMC, main = "FFMC index from the FWI system", horizontal = TRUE)
axis(side=2, at=seq(0,max(fires_nonzeroarea$FFMC), 100), labels=seq(0,max(fires_nonzeroarea$FFMC),100))

# DC variable
boxplot(fires_nonzeroarea$DC, main = "DC index from the FWI system", horizontal = TRUE)
axis(side=2, at=seq(0,max(fires_nonzeroarea$DC), 100), labels=seq(0,max(fires_nonzeroarea$DC),100))
```



The boxplots of the Dulf Moisture Code (DMC) and Initial Speed Index (ISI) variables provide evidence of outliers in the extreme right of the distribution for both variables. The boxplots of the Drought Code (DC) and Fine Fuel Moisture Code (FFMC) variables show outliers to the left of the distribution. There are only non-zero, positive values for all four variables. Therefore, there is little concern with non-reporting of the variables and no compelling need to transform the variables.

Analysis of Key Relationships

Analysis of Spatial Coordinates

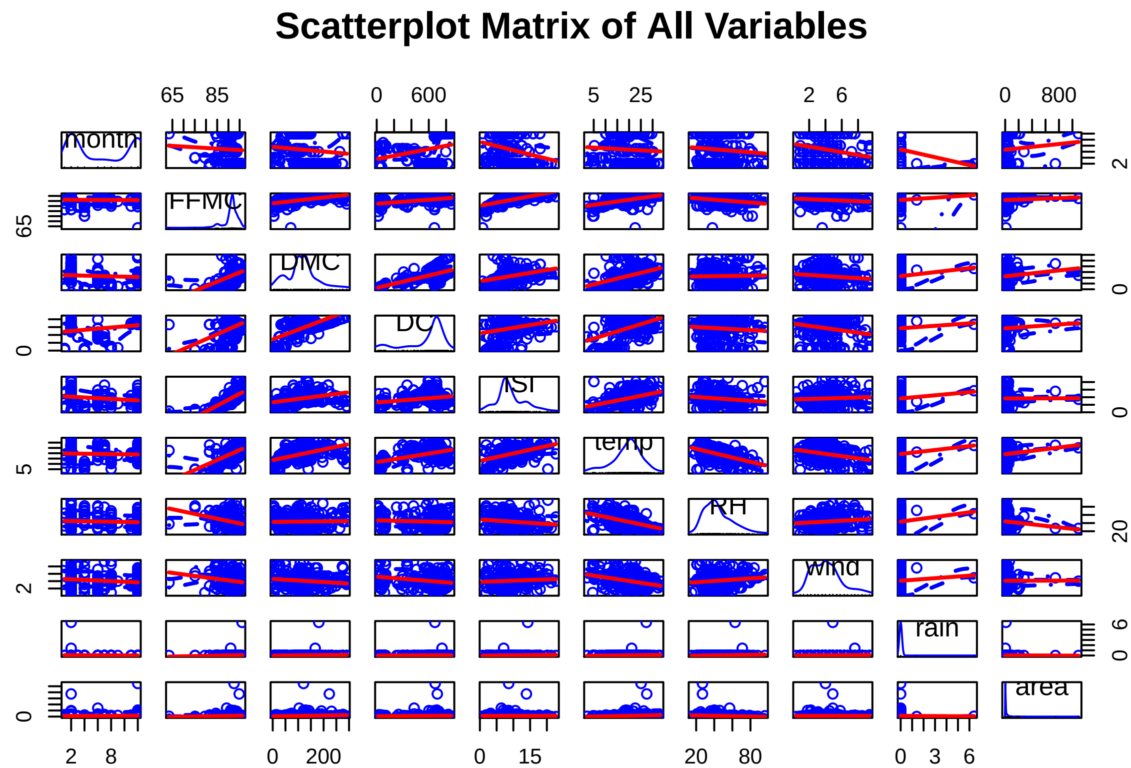


We see that there are only 36 combination of the spatial coordinates, X and Y, which are available in the sample. While there could have been up to 81 combinations from 9 unique X and 9 unique Y values.

We see from the above levelplot that the regions corresponding to the spatial coordinates X,Y = 8,6 and 6,5 are where the largest forest fires measured by burned area occurred and also where the highest number of incidences of forest fires occurred. There must be some geographical reason unique to these regions of the park which make them particularly susceptible to forest fires.

Notably, the region with spatial coordinates X,Y = 5,5 appears to stand out for being resistant to forest fires, despite being surrounded by forest fire affected regions. We suspect that this region might be deforested or may have a water body or some other land form resistant to fire, although we cannot confirm with the data.

Scatterplot matrix

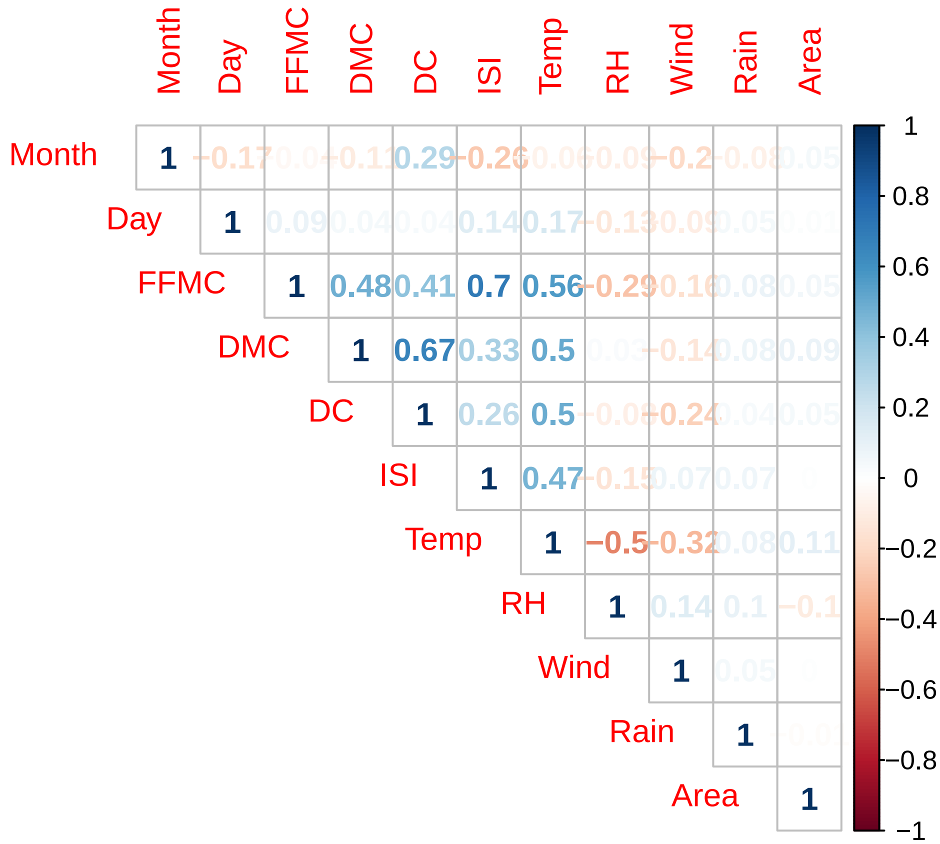


The Scatterplot Matrix shows the relationship of all variables to each other. From the matrix, we observe that there is:

1. Positive relationship between FFMC, DC, DMC, ISI and temp variables.
2. Negative relationship between temp, RH and wind variables.
3. There is very little relationship between rain and other variables.
4. There is very little relationship between area and other variables.
5. There is a strong, yet spurious positive relationship between month and DC.

Correlation matrix

To further examine the relationships between the variables, we ran a correlation matrix. The relationships observed in the scatterplot matrix holds. Of particular interest, we observe negligible correlations between the area variable and other variables.



Analysis of Area

In this section, we explore the Area variable, the outcome variable of interest which measures the magnitude of burned area of the forest fire. In the previous sections, we restricted our analysis to non-zero values of Area. Considering our interest in the most particularly damaging forest fires, we further restrict the dataset to values of Area greater than the arithmetic mean of non-zero Area values.

```
fires_highArea <- fires_nonzeroarea[fires_nonzeroarea$area >= mean(fires_nonzeroarea$area),]
summary(fires_highArea)
```

```
##        X               Y              month      day          FFMC
##  Min.   :1.000   Min.   :2.000   sep    :22   fri: 7   Min.   :81.60
##  1st Qu.:4.000   1st Qu.:3.000   aug    :17   mon: 8   1st Qu.:90.50
##  Median :6.000   Median :4.000   mar    : 5   sat:11   Median :91.70
##  Mean   :4.981   Mean   :4.283   jul    : 3   sun:11   Mean   :91.44
##  3rd Qu.:7.000   3rd Qu.:5.000   apr    : 1   thu: 3   3rd Qu.:93.30
##  Max.   :9.000   Max.   :8.000   dec    : 1   tue: 7   Max.   :96.10
##                                  (Other): 4   wed: 6
##       DMC              DC              ISI              temp
##  Min.   :  9.0   Min.   : 25.6   Min.   : 1.900   Min.   : 5.1
```

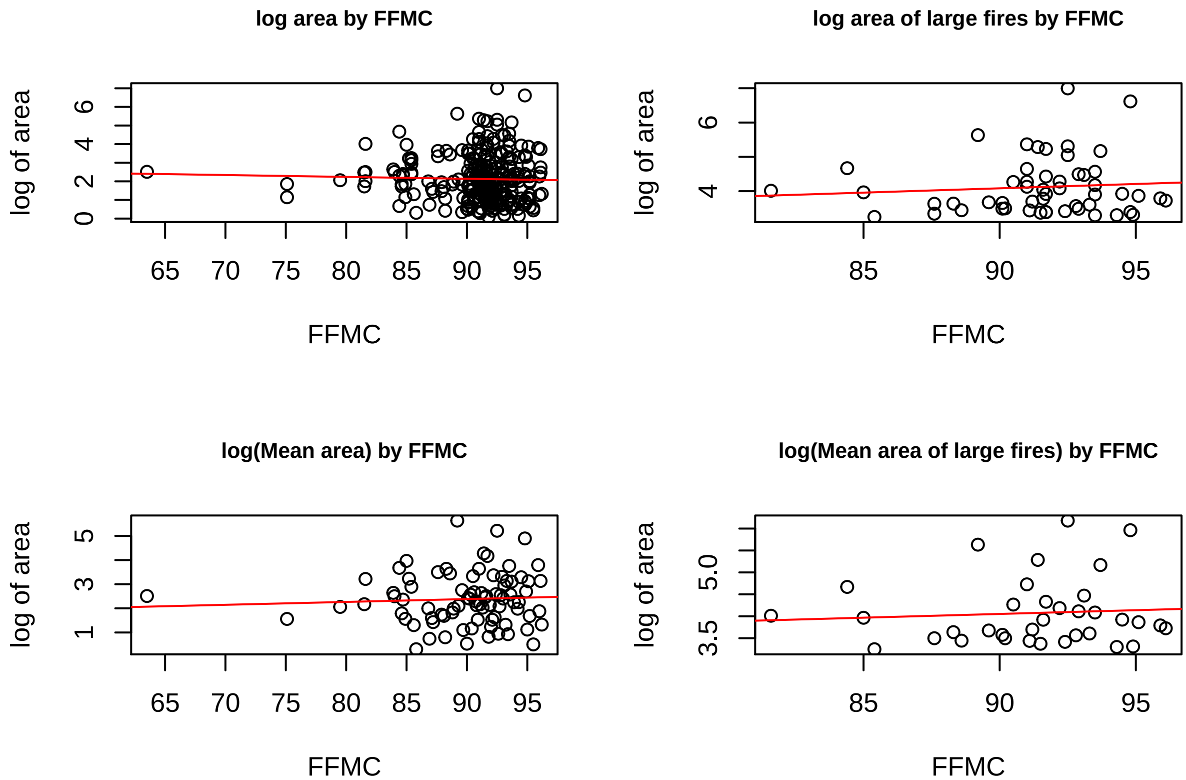
```
##  1st Qu.: 81.8   1st Qu.:480.8   1st Qu.: 6.800   1st Qu.:16.4
##  Median :121.1   Median :674.4   Median : 8.100   Median :19.6
##  Mean   :118.0   Mean   :560.9   Mean   : 9.025   Mean   :19.9
##  3rd Qu.:149.3   3rd Qu.:709.9   3rd Qu.:11.100   3rd Qu.:23.3
##  Max.   :276.3   Max.   :825.1   Max.   :20.300   Max.   :33.3
##
##        RH             wind            rain        area
##  Min.   :19.00   Min.   :1.300   Min.   :0   Min.   :  24.77
##  1st Qu.:28.00   1st Qu.:3.100   1st Qu.:0   1st Qu.:  31.86
##  Median :40.00   Median :4.000   Median :0   Median :  48.55
##  Mean   :42.36   Mean   :4.168   Mean   :0   Mean   : 101.36
##  3rd Qu.:50.00   3rd Qu.:4.900   3rd Qu.:0   3rd Qu.:  86.45
##  Max.   :96.00   Max.   :9.400   Max.   :0   Max.   :1090.84
##
```

We see that the range of the Area variable is still relatively large spanning from 24.77 to 1090.84 hectares. We also observe that by restricting our analysis to the most particularly devastating forest fires we now have a sample of just 53 observations. This represents a substantial reduction from the sample of 270 observations from the dataset with just non-zero Area values.

We would need transformation of Area variable to visualize any relation with other key variables. We decided apply a logarithmic transformation to the Area variable as intuitively we think that burned area increases exponentially in forest fires.

As an alternative transformation, will also took the logarithm of the mean Area + 1, as this will remove negative log values for the values of Area which are less than 1. Mean area with respect to different variables would be calculated using 'by' function.

Let us look at relation between the log-transformed Area variable and FFMC and log-transformation of the mean of the Area variable and the FFMC index.




```r
cor(fires_nonzeroarea$FFMC, log(fires_nonzeroarea$area+1))
```

```
## [1] -0.02969747
```

```r
cor(fires_highArea$FFMC, log(fires_highArea$area+1))
```

```
## [1] 0.08762157
```

```r
cor(sort(unique(fires_nonzeroarea$FFMC)),log(area_mean_by_FFMC+1))
```

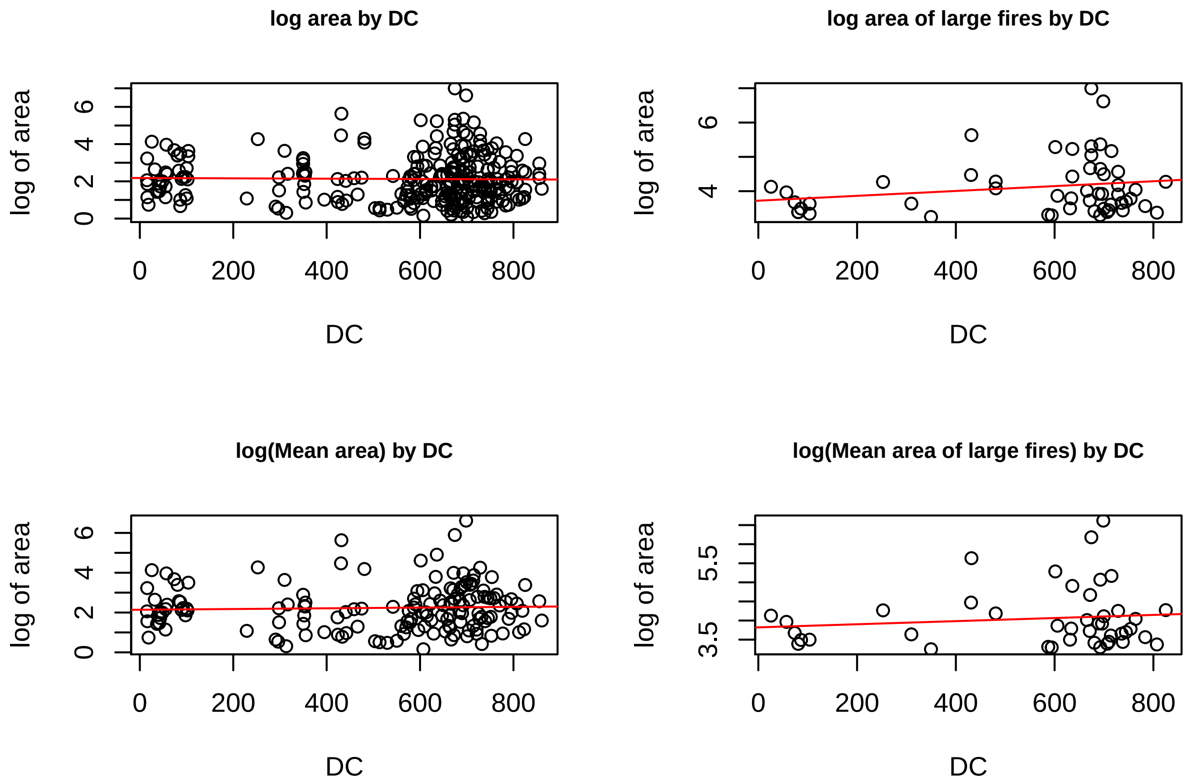
```
## [1] 0.0557785
```

```r
cor(sort(unique(fires_highArea$FFMC)), log(large_area_mean_by_FFMC+1))
```

```
## [1] 0.0751442
```

The scatterplots above indicate that there is a very weak correlation between the log-transformations of the Area and FFMC variable (and mean Area and FFMC). Furthermore, the correlation coefficients are provide evidence of a weak positive relationship. However, we notice that the outliers of the log Area values appear at larger values of FFMC. The correlation is only slightly stronger when the relationship is restricted to fires with higher Area values.

Next, we analyze log-transformed Area variable and DC index and log-transformation of the mean of the Area variable and the DC index.




```
cor(fires_nonzeroarea$DC, log(fires_nonzeroarea$area+1))
```

```
## [1] -0.01659247
```

```
cor(fires_highArea$DC, log(fires_highArea$area+1))
```

```
## [1] 0.2019966
```

```
cor(sort(unique(fires_nonzeroarea$DC)), log(area_mean_by_DC+1))
```

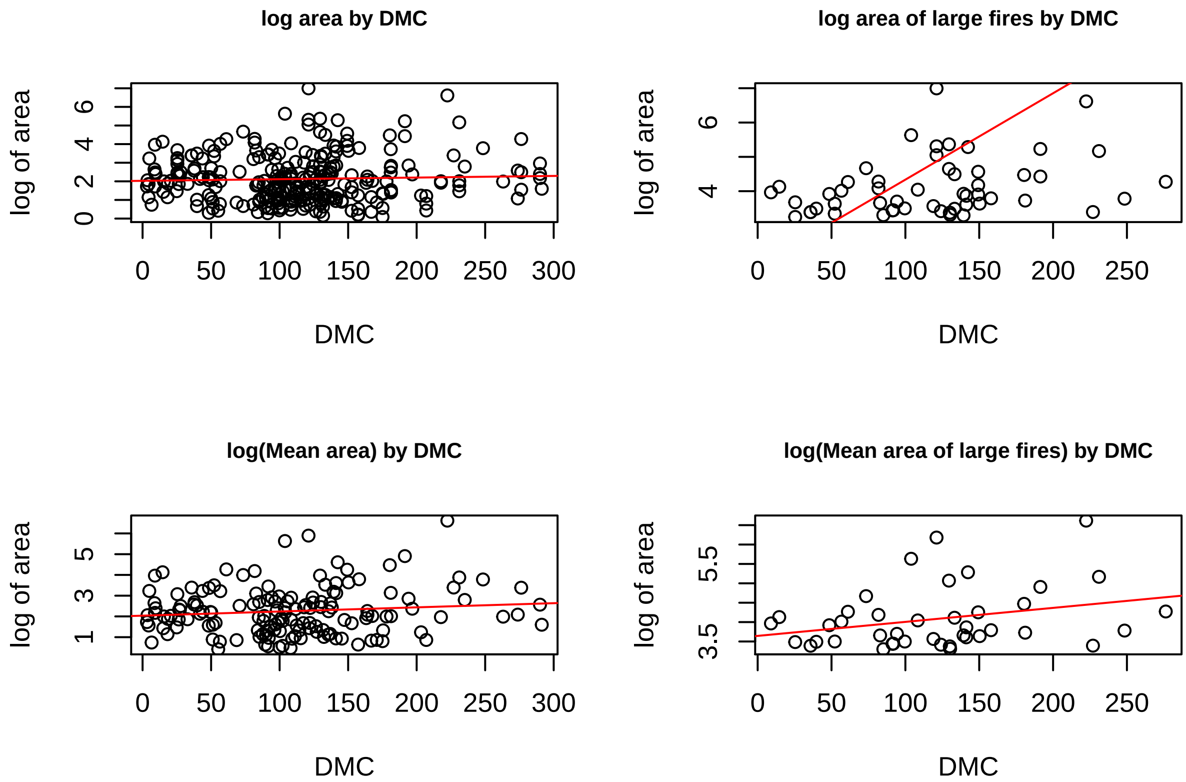
```
## [1] 0.03930777
```

```
cor(sort(unique(fires_highArea$DC)), log(large_area_mean_by_DC+1))
```

```
## [1] 0.1224547
```

The scatterplots and correlation coefficients above indicate that there is a very weak correlation between the log-transformations of the Area and DC variable (and mean Area and DC). As with the analysis of the relationship with the FFMC variable, we notice that the outliers of the log Area values appear at larger values of DC. The correlation is only slightly stronger when the relationship is restricted to fires with higher Area values.

Next, we analyze log-transformed Area variable and DMC index and log-transformation of the mean of the Area variable and the DMC index.




```r
cor(fires_nonzeroarea$DMC, log(fires_nonzeroarea$area+1))
```

```
## [1] 0.04308562
```

```r
cor(fires_highArea$DMC, log(fires_highArea$area+1))
```

```
## [1] 0.3067803
```

```r
cor(sort(unique(fires_nonzeroarea$DMC)), log(area_mean_by_DMC+1))
```

```
## [1] 0.114564
```

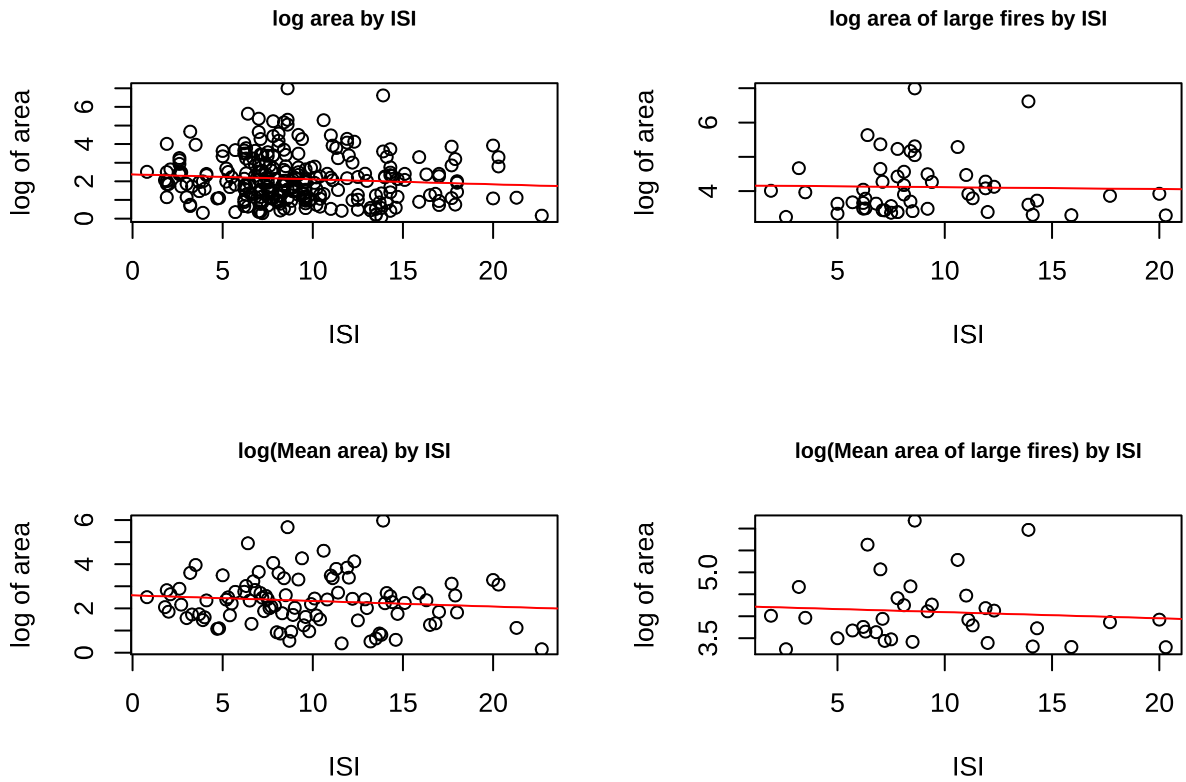
```r
cor(sort(unique(fires_highArea$DMC)), log(large_area_mean_by_DMC+1))
```

```
## [1] 0.2936497
```

The scatterplots and correlation coefficients above indicate that there is a positive and substantive correlation between the log-transformations of the Area and DMC variable (and mean Area and DMC). As with the analysis of the relationship with the FFMC and DC variables, we notice that the outliers of the log Area values appear at larger values of DMC. However, in this case, the correlation is much stronger when the relationship is restricted to fires with higher Area values.

This might support there being a positive association between forest fire damage and larger values of DMC. Larger DMC values might indicate a more damaging forest fire.

Let us now focus on relation between area and ISI




```r
cor(fires_nonzeroarea$ISI, log(fires_nonzeroarea$area+1))
```

```
## [1] -0.08788882
```

```r
cor(fires_highArea$ISI, log(fires_highArea$area+1))
```

```
## [1] -0.02547997
```

```r
cor(sort(unique(fires_nonzeroarea$ISI)), log(area_mean_by_ISI+1))
```

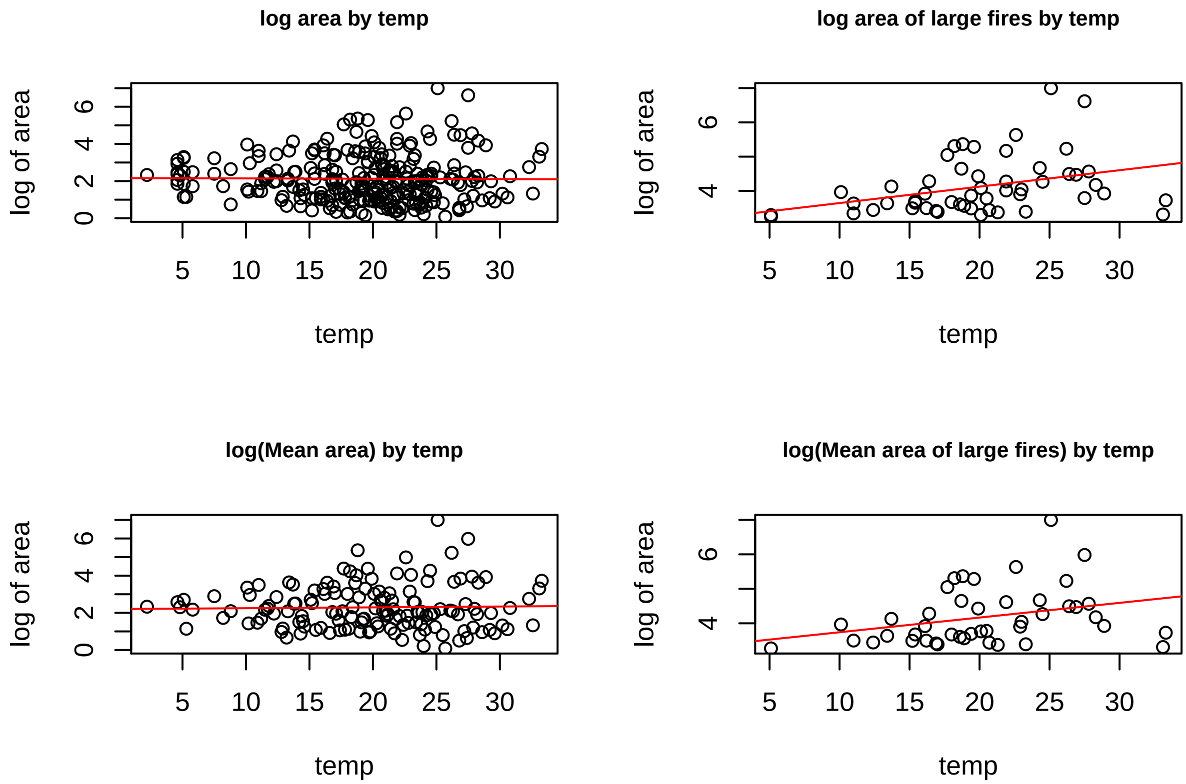
```
## [1] -0.1096299
```

```r
cor(sort(unique(fires_highArea$ISI)), log(large_area_mean_by_ISI+1))
```

```
## [1] -0.08493546
```

The scatterplots and correlation coefficients above indicate that there is a very weak, negative correlation between the log-transformations of the Area and ISI variable (and mean Area and ISI). The correlation is slightly weaker (and negative) when the relationship is restricted to fires with higher Area values.

Now, we will focus on relation of the Area variable with the temperature variable.




```
cor(fires_nonzeroarea$temp, log(fires_nonzeroarea$area+1))
```

```
## [1] -0.00873371
```

```
cor(fires_highArea$temp, log(fires_highArea$area+1))
```

```
## [1] 0.3495278
```

```
cor(sort(unique(fires_nonzeroarea$temp)), log(area_mean_by_temp+1))
```

```
## [1] 0.02502732
```

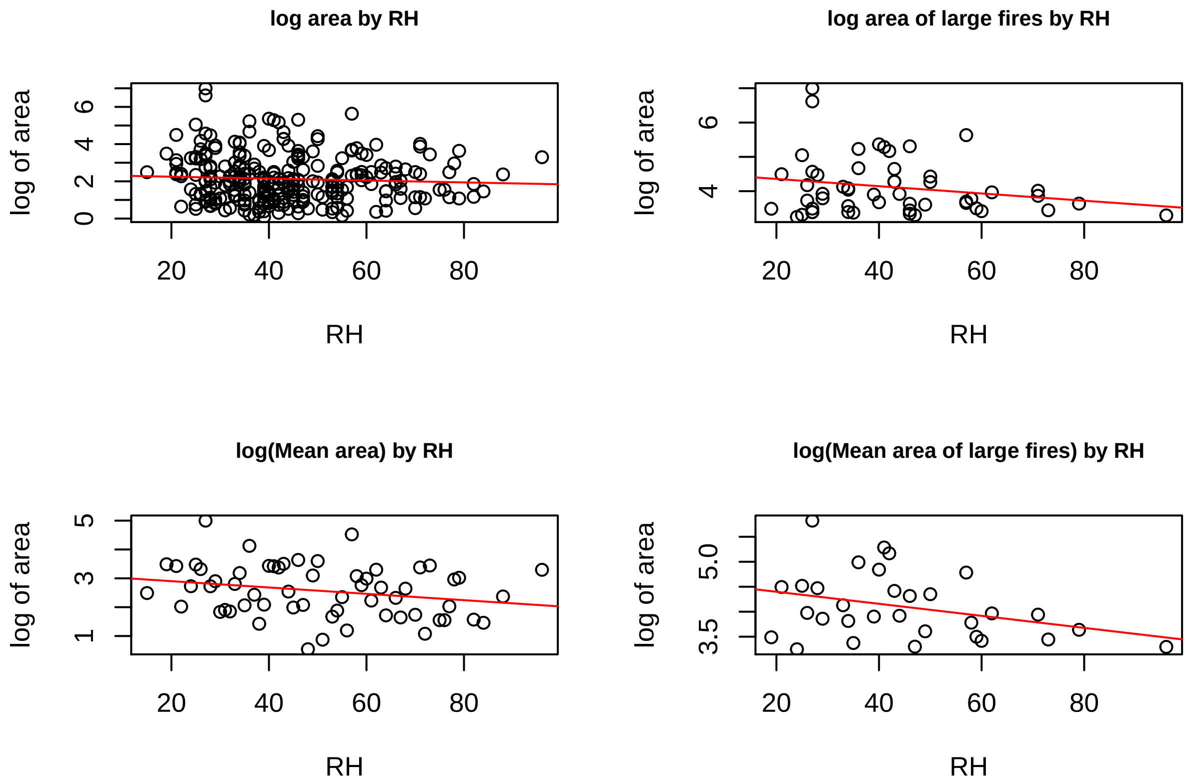
```
cor(sort(unique(fires_highArea$temp)), log(large_area_mean_by_temp+1))
```

```
## [1] 0.3060864
```

Here we see that for the fires with larger area of devastation have a stronger, positive relation with temperature. So it seems temperature has some relation with larger fires. This is confirmed by higher, positive correlation coefficients observed when the data restricted to the particularly damaging forest fires.

While temperature does not have a strong correlation with forest fires in the larger dataset, larger forest fires are more strongly associated the higher temperature.

We will now examine the relationship between area and relative humidity.




```r
cor(fires_nonzeroarea$RH, log(fires_nonzeroarea$area+1))
```

```
## [1] -0.0613912
```

```r
cor(fires_highArea$RH, log(fires_highArea$area+1))
```

```
## [1] -0.208099
```

```r
cor(sort(unique(fires_nonzeroarea$RH)), log(area_mean_by_RH+1))
```

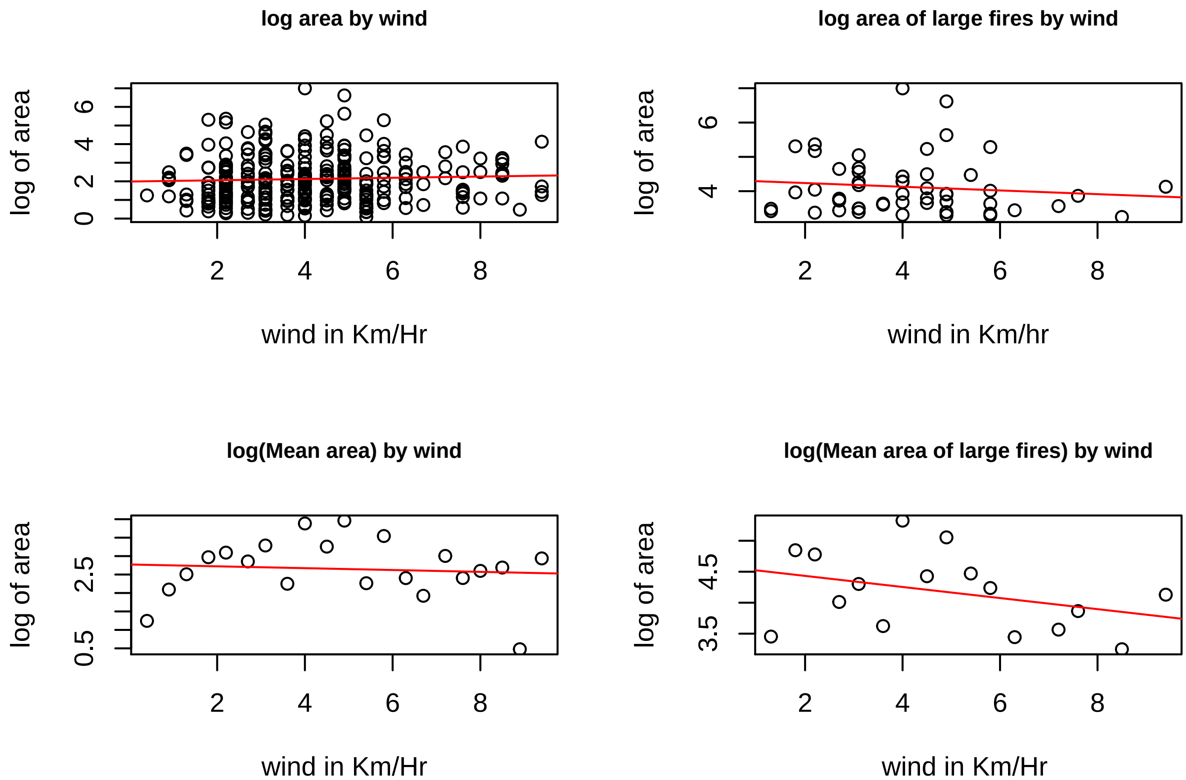
```
## [1] -0.2406195
```

```r
cor(sort(unique(fires_highArea$RH)), log(large_area_mean_by_RH+1))
```

```
## [1] -0.3365854
```

We see that the RH variable has the strongest negative relation with the log-transformed Area variable for both non-zero values of Area and particularly high values of Area. Moreover, we see that the correlation coefficient increases for higher values of the Area variable. The can make the association between lower relative humidity and larger forest fires.

Now we will examine wind with respect to area.



We see that there is not much of a relationship between all fires with some area to wind but when we focus on larger area fires the relation is strongly negative. The correlation coefficients increases between the log-transformed Area variable restricted to the particularly higher damage forest fires. This is actually counter intuitive as larger windspeed should mean more spread of fire but it does not seem so in the sample data. We suspect that there might be something lurking explaining this negative association.

```
cor(fires_nonzeroarea$wind, log(fires_nonzeroarea$area+1))
```

```
## [1] 0.04990456
```

```
cor(fires_highArea$wind, log(fires_highArea$area+1))
```

```
## [1] -0.1113709
```

```
cor(sort(unique(fires_nonzeroarea$wind)), log(area_mean_by_wind+1))
```

```
## [1] -0.08492572
```

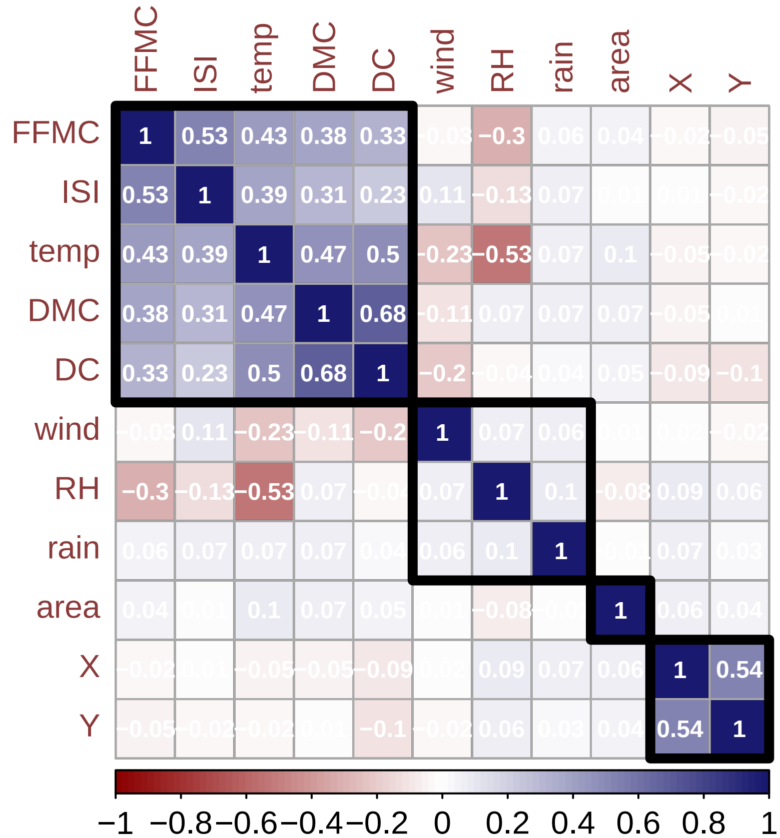
```
cor(sort(unique(fires_highArea$wind)), log(large_area_mean_by_wind+1))
```

```
## [1] -0.3514072
```

Analysis of Secondary Effects

Given the examinations above, in relation to area we find that temperature and relative humidity show intuitive correlations with larger area forest fires and wind shows a counter intuitive negative correlation with larger area forest fires. Temperature and relative humidity variables may impact larger forest fires in both directions as higher temperature and lower relative humidity lead to larger forest fires or more damaging forest fires may lead to high temperature and lower RH. On the other hand, the relation we observe with wind appears to be counter intuitive as larger area forest fires show less severe wind levels. This is a confounding effect in the previous area data set.

Next we examine the potential confounding effects in other secondary variables.



We first take the numeric variables of the data and examine the correlation among the variables other than the area variable which we extensively studied above. We find that there are a few interesting correlation from the map above (we define as threshold correlation of 0.45 or above).

Given temperature, RH and wind are relevant variables to larger forest fire area, we move on to examine secondary variables that have strong relationship to these three variables.

```
cor (fires$temp, fires$FFMC)
```

```
## [1] 0.4315323
```

```
cor (fires$temp, fires$DMC)
```

```
## [1] 0.4695938
```

```
cor (fires$temp, fires$DC)
```

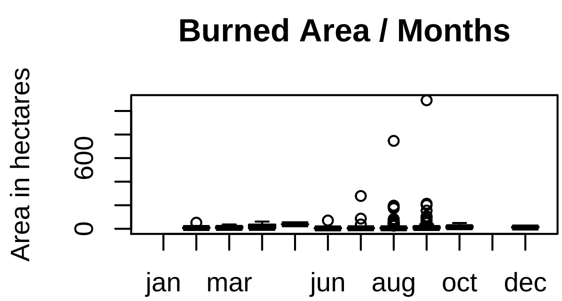
```
## [1] 0.4962081
```

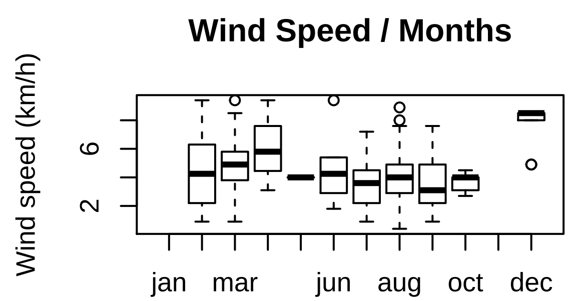
Temperature shows a relatively strong (0.43) correlation with FFMC and (0.46) with DMC and (0.5) with DC as temperature is likely an input in the FFMC index and DMC index and the DC index.

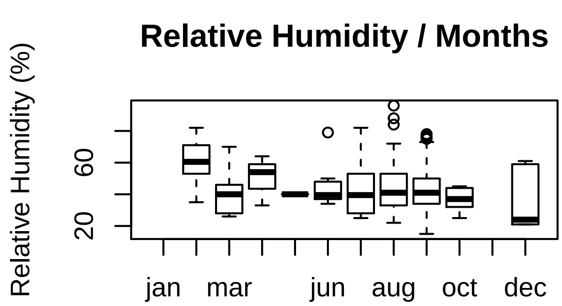
Seasonality Across the Variables

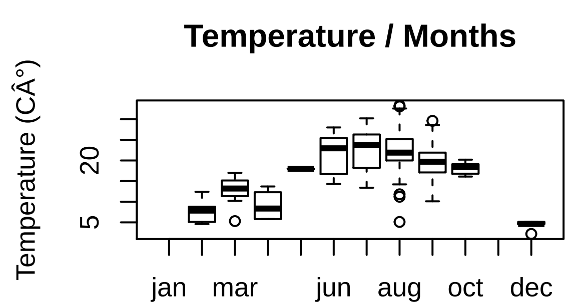
We also consider the association of variables over the months of the year to understand whether theres is seasonality in the data. The following boxplots show the distribution of the variables over the months of the year. We observe seasonality on the variation of the variable values, particularly with respect to:

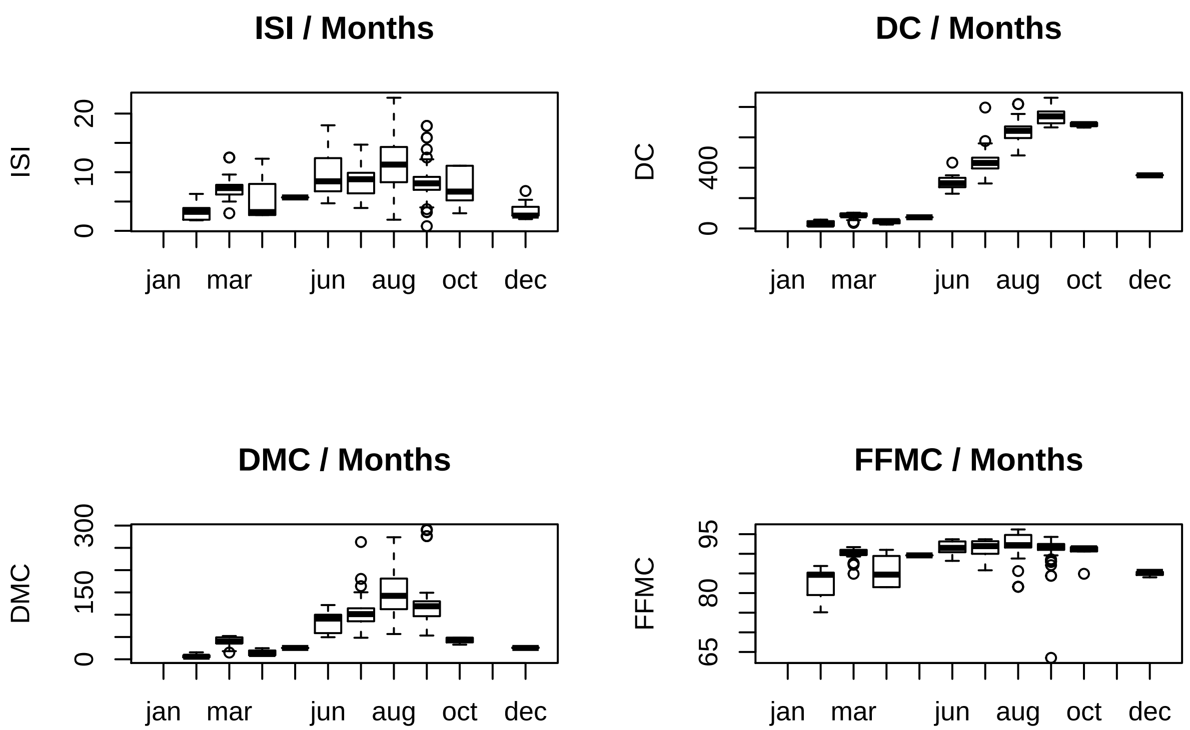
1. Extreme area values observed in summer months (July, August and September).
2. Extreme rain values observed in August.
3. Extreme outliers in wind speed values observed in March, June and August, although there are high average monthly wind speeds observed in April and December.
4. Extreme outliers in relative humidity values observed in June, August and September. There are higher average monthly relative humidity measures observed in February, March and April.
5. Higher monthly temperature values observed in summer months (June, July, August and September). Outliers are observed in August.
6. Higher monthly average ISI values observed in summer months (June, July, August and September). Largest dispersion observed in ISI values observed in August. Outlier values observed in September.
7. Higher monthly average DC values observed in summer months (June, July, August and September) and October. DC values are relatively much lower in February, March and April.
8. Higher monthly average DMC values observed in summer months (June, July, August and September). Extreme right outliers are observed in July and September.
9. Higher monthly average FFMC values observed in summer months (June, August and September) and October. Extreme left outliers are observed in August, September and October.











Given the seasonality findings above, we suspect that there is a larger concentration of forest fires in the summer and early fall months. Moreover, we hypothesize that the larger concentration of forest fires in the summer and early fall may help to explain some of the counter-intuitive relationships we observed.

Indeed the following frequency output shows that the vast majority of all non-zero area forest fires occur in August and September (196 of the 270). This indicates that there is a large concentration of forest fires within a short period of the year and while most are relatively small, a small fraction in this short time period are the most damaging.

```r
# First, re-order factors in month variable
fires_nonzeroarea$month = factor(fires_nonzeroarea$month,
                                 levels(fires$month)[c(5,4,8,1,9,7,6,2,12,11,10,3)])
count(fires_nonzeroarea$month)
```

```
##      x freq
## 1  feb   10
## 2  mar   19
## 3  apr    4
## 4  may    1
## 5  jun    8
## 6  jul   18
## 7  aug   99
## 8  sep   97
## 9  oct    5
## 10 dec    9
```

We also extend this analysis to days of the week to show that the majority of forest fires occur of weekend days (Fridays, Saturdays and Sundays) when we suspect there are larger number of visitors to the forest park.

```r
# First, re-order factors in days variable
fires_nonzeroarea$day = factor(fires_nonzeroarea$day,
                               levels(fires_nonzeroarea$day)[c(2,6,7,5,1,3,4)])
count(fires_nonzeroarea$day)
```

```
##     x freq
## 1 mon   39
## 2 tue   36
## 3 wed   32
## 4 thu   31
## 5 fri   43
## 6 sat   42
## 7 sun   47
```

Relative humidity relative to other Secondary Variables

RH shows a strong negative correlation with temperature and appears that these two variables are inverse related. The higher the relative humidity, the lower the temperature. The lower the relative humidity, the higher the temperature.

```r
cor (fires$RH, fires$temp)
```

```
## [1] -0.5273903
```

Wind does not seem to show any strong relationship with any of the secondary variables.

Other than the three variables above, we also observe strong (>0.5) correlation relationships between DMC and DC, and ISI and FFMC.

```r
cor (fires$ISI, fires$FFMC)
```

```
## [1] 0.5318049
```

```r
cor (fires$DMC, fires$DC)
```

```
## [1] 0.6821916
```

FFMC shows a strong relationship with ISI and likely to be an input into the ISI index. DMC and DC index also strong very strong correlation and possibly include the same inputs as each other.

Conclusion

Summary

There are a few salient observations we can make following this study:

- Transformation of the outcome variable of interest, Area is necessary to make analytical associations with other variables. We found that a log transformation plus 1 improved the analytical interpretation of the Area variable and subsequent associative analysis.
- There are large seasonal effects in the data with large majority of forest fires occurring in the summer and early fall months. These summer months are associated with higher average temperatures and lower average relative humidity.
- We further observe an association with particularly larger area forest fires and higher temperature and lower relative humidity. This reinforces our observation that larger forest fires are seasonal in nature with a tendency to occur in the summer and early fall months.
- We observed a counter-intuitive relationship between larger area forest fires and lower wind speed. We suspect that this might be a spurious result as higher winds should be associated with larger fires. It is also apparent that average wind speeds are somewhat lower in summer months, so there might be notable size effect in the data.
- We observe a higher incidence of forest fires on weekend days (Fridays, Saturdays and Sundays) leading us to suspect that forest fires are likely associated with larger number of human visitors.
- We observe there are spatial coordinate regions of the forest park that are particularly susceptible to forest fires. We suspect these area might be exposed to greater human activity.

What are we missing?

Lastly, it is important to note what we are not able to observe from the data. We are missing indicators of human activity, which we opine can have a large impact on forest fire incidence. The higher incidence of forest fires on weekend days can lead us to believe that measures of human activity such as number of visitors, concentration of visitors in a given region of the forest, types of activity or even length of stays can help to explain forest fire incidence and damage. Furthermore, we suspect that for large forest fires arsonist activity might be a lurking culprit. However, surveillance of human activity might be particularly costly and infeasible, particularly if entry to this forest is not well monitored.