Forest_Fire_Initial

Debalina Maiti, Mark Paluta, Tina Agarwal, Vivek Agarwal

5/24/2018

Introduction

This analysis is motivated by the following research question:

What factors lead to particularly damaging forest fires?

Setup

First, we load the car library, which gives us a convenient scatterplotMatrix function and the data set.

```
library(car)
```

```
## Loading required package: carData
```

```
forest_fire = read.table("forestfires.csv", header=TRUE, sep=",", na.string = "na")
```

Data Overview

We note that we have 517 observations and 13 variables

```
nrow(forest_fire)
```

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

```
str(forest_fire)
```

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

```
summary(forest_fire)
```

```
##        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
##
```

%% Comment: It might be a good idea to check all columns for na values up here near the top to get it over with %%

```
for (name in names(forest_fire)){
  cat("NAs in " , name , " = " , sum(is.na(forest_fire[,name])),"\n" )
  }
```

```
## NAs in  X  =  0
## NAs in  Y  =  0
## NAs in  month  =  0
## NAs in  day  =  0
## NAs in  FFMC  =  0
## NAs in  DMC  =  0
## NAs in  DC  =  0
## NAs in  ISI  =  0
## NAs in  temp  =  0
## NAs in  RH  =  0
## NAs in  wind  =  0
## NAs in  rain  =  0
## NAs in  area  =  0
```

No NAs in the dataset. Hooray!!

Transformations

```
forest_fire$sorted_day<-factor(forest_fire$day, levels = c("mon","tue","wed","thu","fri","sat","sun" ))

forest_fire$fire_size <- cut(forest_fire$area, breaks=c(-Inf,0.01, 25, Inf), labels=c('Low','Medium','H

forest_fire$dotsize<-NA
forest_fire[forest_fire$fire_size=="Medium",]$dotsize<-2
forest_fire[forest_fire$fire_size=="High",]$dotsize<-5
forest_fire[forest_fire$fire_size=="Low",]$dotsize<-1
```

```r
Med_High_fire <- forest_fire[forest_fire$fire_size != "Low",]

forest_fire$month = factor(forest_fire$month,levels(forest_fire$month)[c(5,4,8,1,9,7,6,2,12,11,10,3)])
```

Univariate Analysis of Key Variables

We have a large number of variables, we begin with a scatterplot matrix. This is helpful for getting a high-level overview of the relationships between our variables and can draw our attention to important features we want to investigate further.

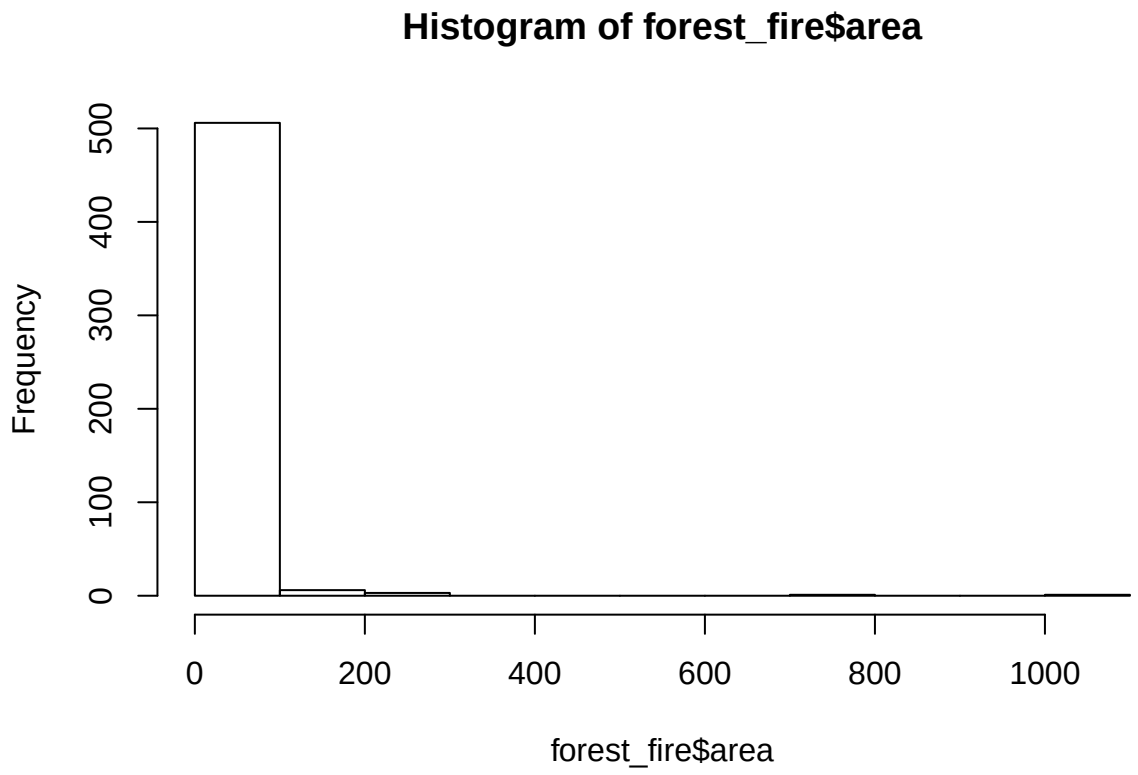
```r
#scatterplotMatrix(~ RH + wind + rain + logarea, data=forest_fire, smooth = TRUE, main = "Scatterplot M
```

Area

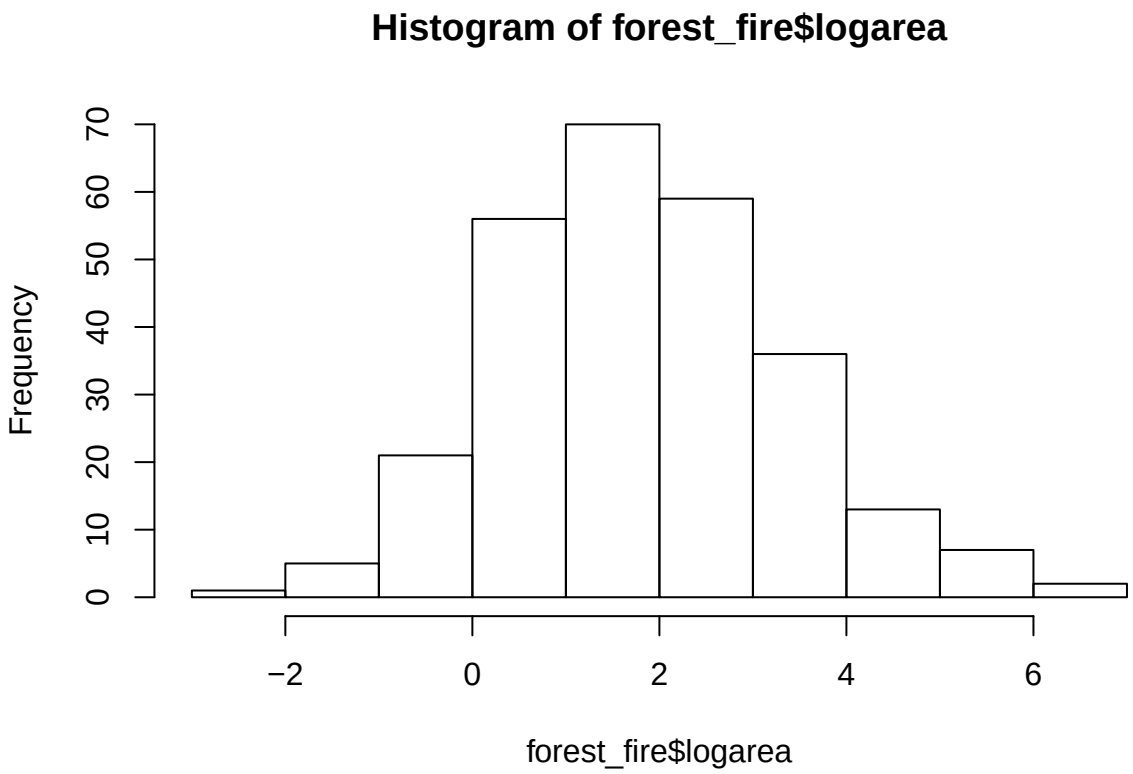
%% Comment: need to address log(0) values being dropped %%

Area seems to be right skewed. To normalize the data, we could take the log of area.

```r
forest_fire$logarea = log(forest_fire$area)
hist(forest_fire$area)
```




```r
hist(forest_fire$logarea)
```



However, as a signficant portion of our area data is zero and has an undefined logarithm, we also look at an approach of transforming area into a factor of low, medium, or high fire level.

Note: the mean fire area is 12.85.

For the rows with area = 0, we will populate the level value as 'low' For the rows with 0 < area < 12.85, we will populate the level value as 'medium' For the rows with area >= 12.85, we will populate the level value as 'high'

```
## Case High

case_high = subset(forest_fire, 12.85<=forest_fire$area &!is.na(forest_fire$area))
case_high["level"]<- "high"

nrow(case_high)##high count = 76
```

```
## [1] 76
```

```
## Case Medium

case_medium = subset(forest_fire, 0<forest_fire$area & forest_fire$area<12.85 &!is.na(forest_fire$area))
case_medium["level"]<- "medium"

nrow(case_medium)##medium count = 194
```

```
## [1] 194
```

```
## Case Low
```

```
case_low = subset(forest_fire, forest_fire$area==0 &!is.na(forest_fire$area))
case_low["level"]<- "low"

nrow(case_low)##low count = 247
```

```
## [1] 247
```

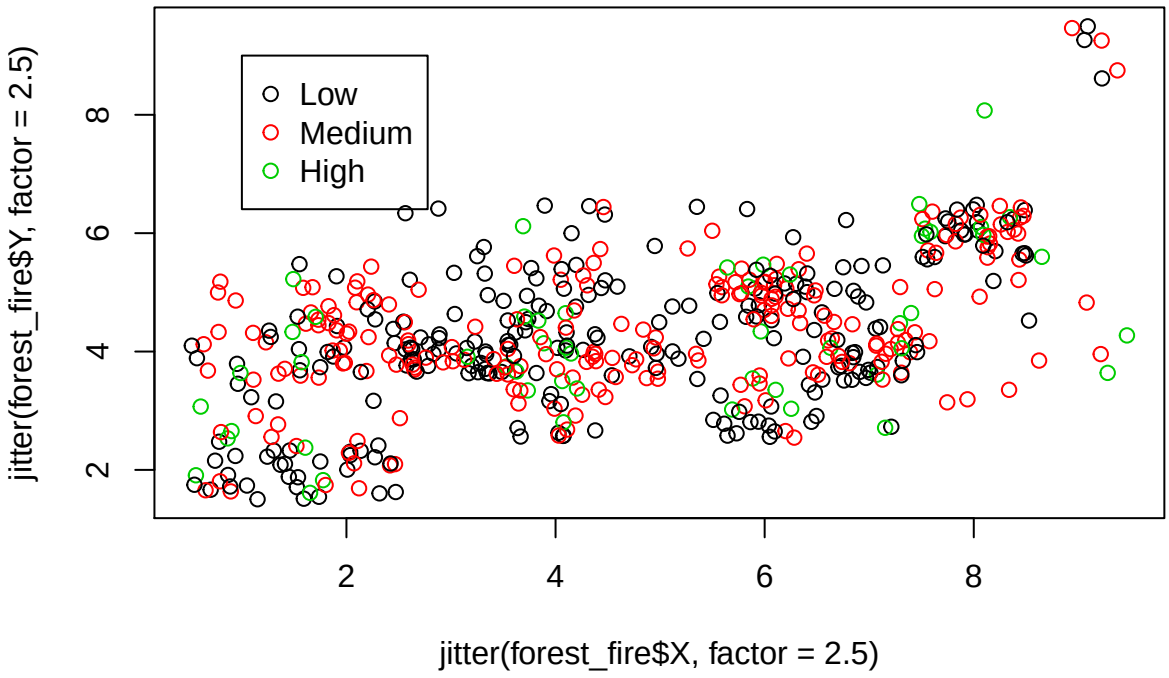
```
###Merge all three to ff_level

ff_level = merge(case_high, case_medium, all = TRUE)
ff_level = merge(case_low, ff_level, all = TRUE)

####Logarithm - looking at the data against area we see that there are huge number of data with value z
#
#ff_level$area = log(ff_level$area) - Vivek - suggest we don't overwrite an existing variable. We alrea
#str(ff_level)
```

Spatial coordinates (X & Y)

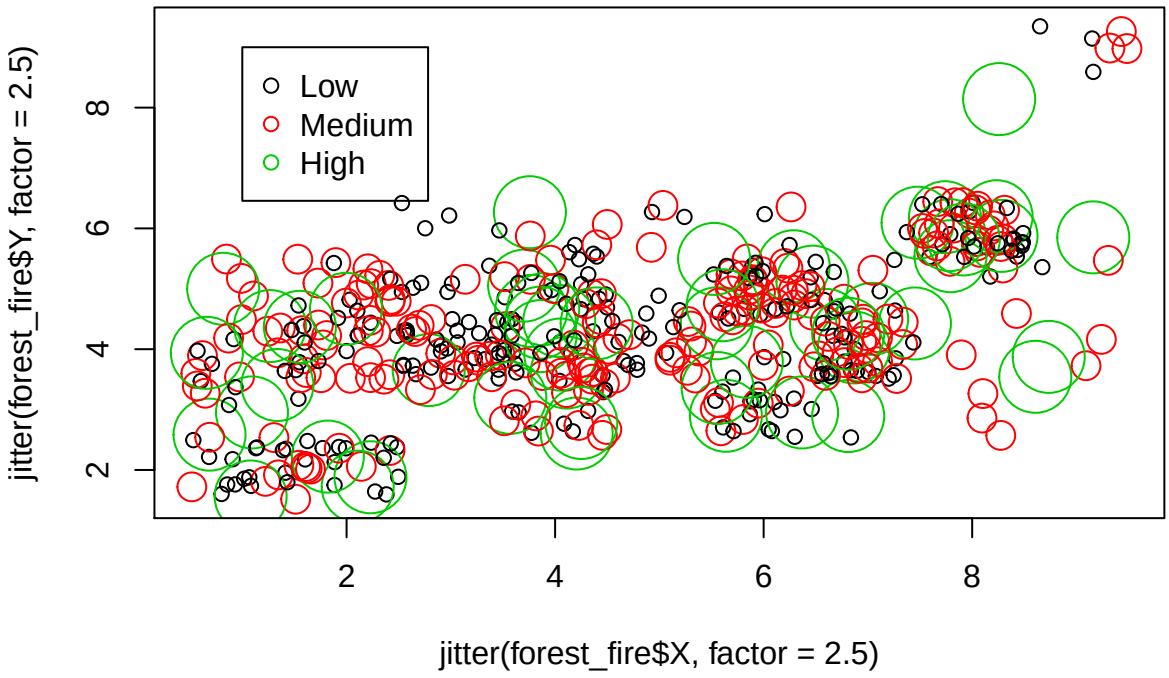
%% This might below in next section: Analysis of Key Relationships %%

Plotting the coordinates on a grid, we can see that some areas of the grid are densely populated with data and other areas are sparse. The colors represent size of the fire.

```
########## replace this filler breakdown with Debalina's high/medium/low breakdown - Done (Vivek) #####

forest_fire$fire_size <- cut(forest_fire$area, breaks=c(-0.01,0.01, 25, Inf), labels=c('Low','Medium','
plot(jitter(forest_fire$X, factor=2.5), jitter(forest_fire$Y, factor=2.5), col=forest_fire$fire_size)
legend(1,9,unique(forest_fire$fire_size),col=1:length(forest_fire$fire_size),pch=1)
```



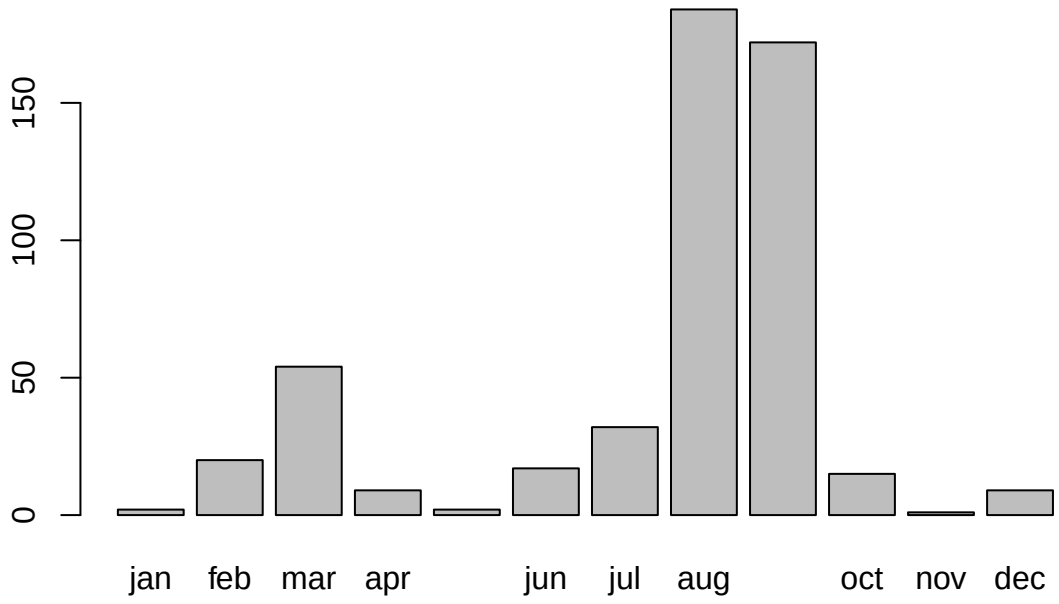

```
########## replace this filler breakdown with Debalina's high/medium/low breakdown - Done (Vivek) #####


plot(jitter(forest_fire$X, factor=2.5), jitter(forest_fire$Y, factor=2.5), col=forest_fire$fire_size, c
legend(1,9,unique(forest_fire$fire_size),col=1:length(forest_fire$fire_size),pch=1)
```



Month

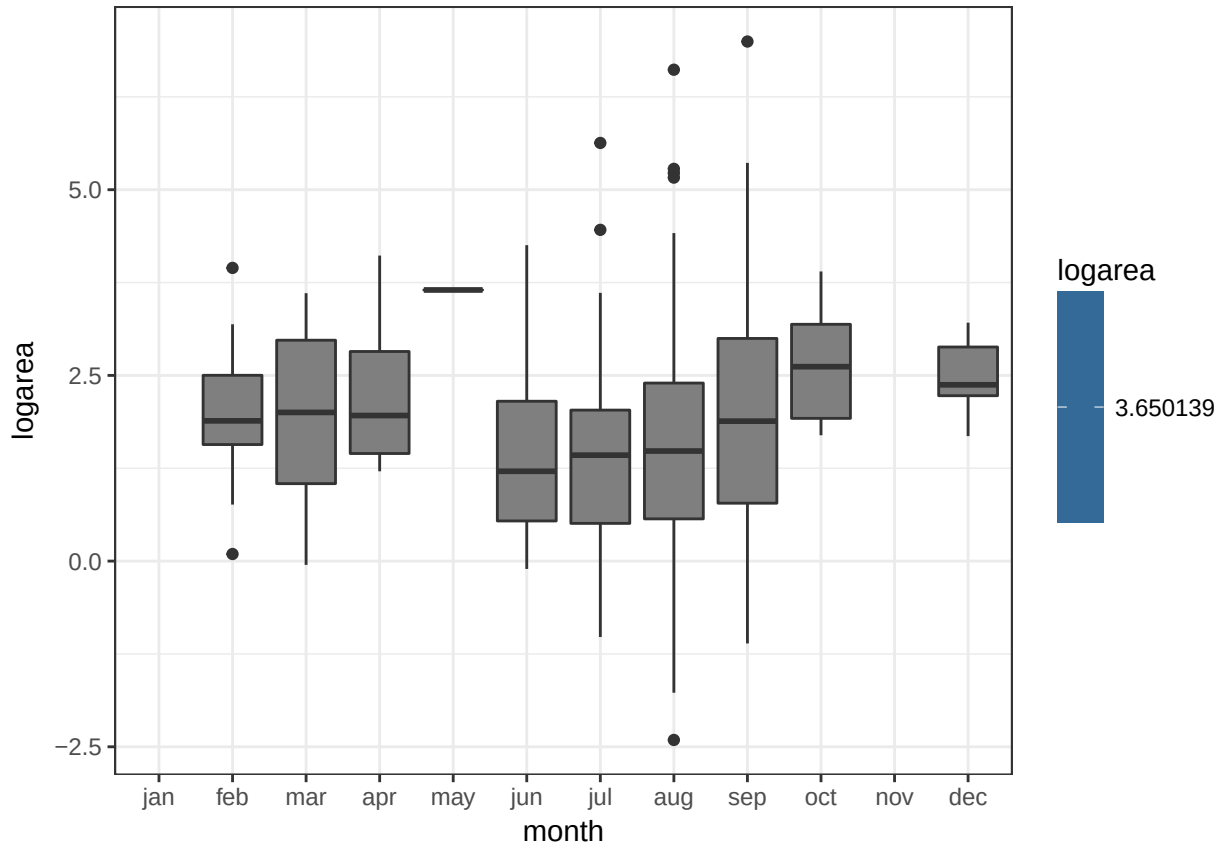
A histogram shows that most of the data comes from spring and summer months. We can use box plots of month and log of area to see the distributions of burned area by month.

```
barplot(table(forest_fire$month))
library(ggplot2)
```




```
ggplot(data = forest_fire, aes(x = month, y = logarea)) + geom_boxplot(aes(fill = logarea), width = 0.8)
```

```
## Warning: Removed 247 rows containing non-finite values (stat_boxplot).
```

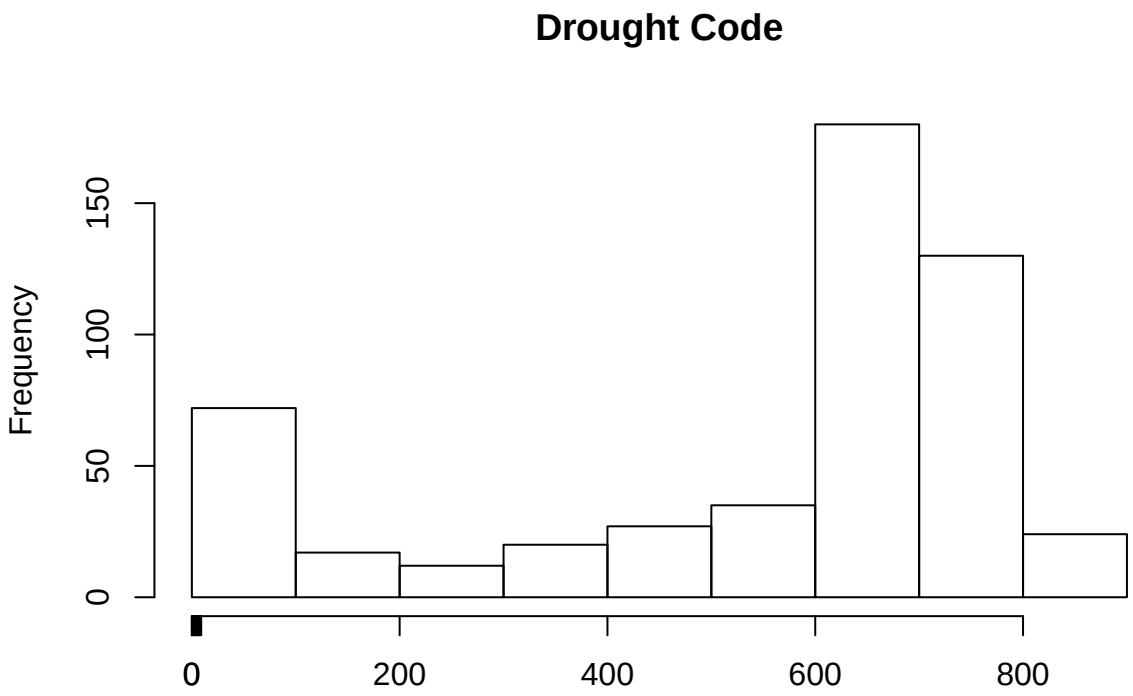


DC

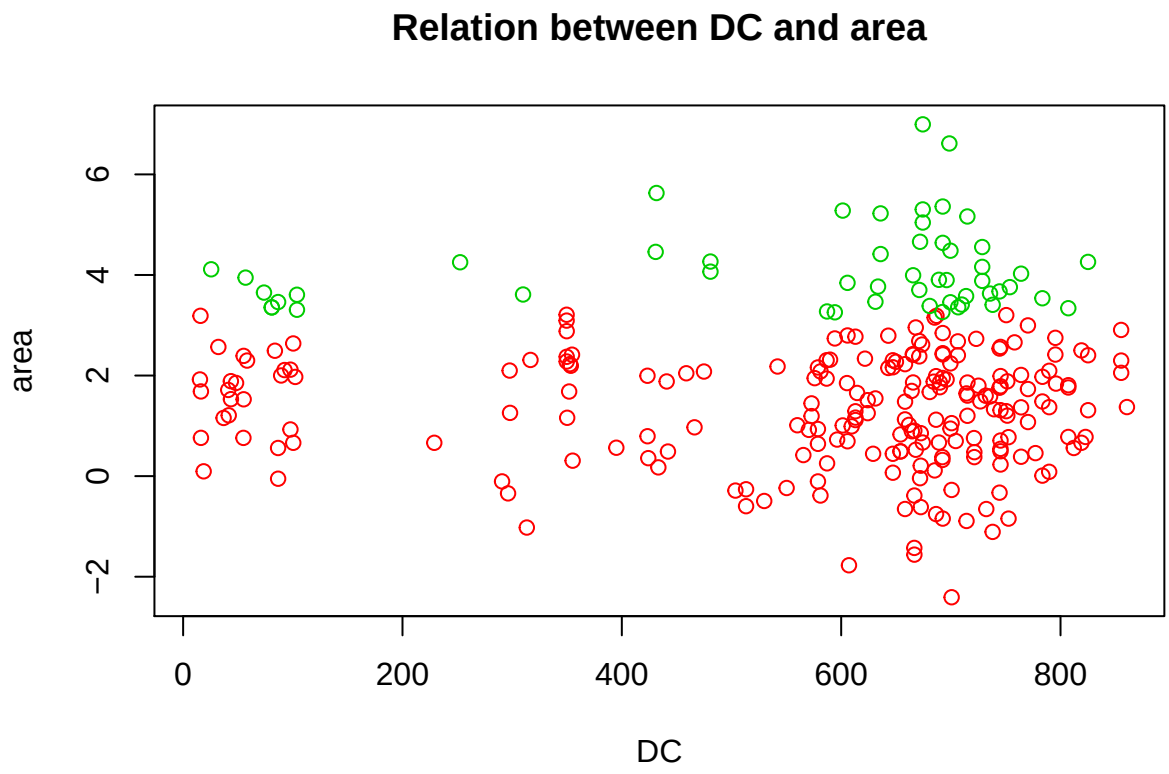
```
# Let's take a look at DC index
summary(ff_level$DC)
```

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

```
# Here is the histogram of DC index

hist(ff_level$DC, main = "Drought Code",
     xlab = NULL)
axis(1, at = 0:9)
```




```
# Plot DC and Burned Area

plot(jitter(forest_fire$DC, factor=2),jitter(forest_fire$logarea, factor=2 ),
     col=factor(forest_fire$fire_size),
     xlab = "DC", ylab = "area",
     main = "Relation between DC and area")
```



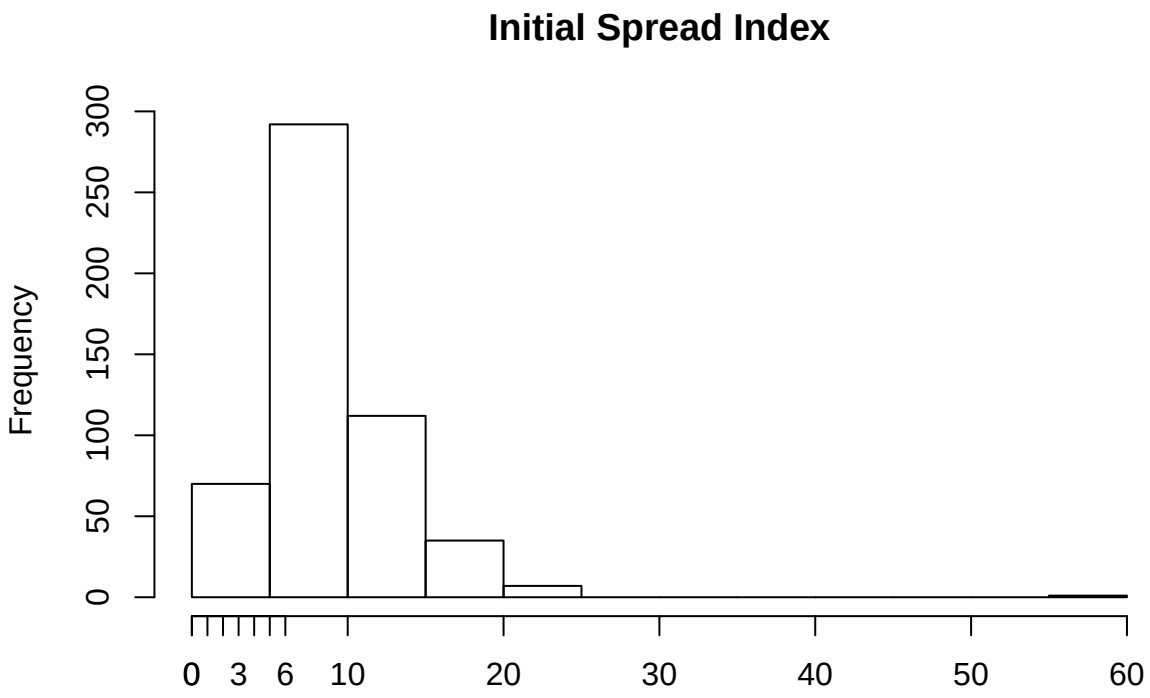

```
legend=levels(ff_level$fire_size)

#Observation: Could not find very distinctive feature. But for 600 <= Drought Code <=800, fire tendency
```

ISI

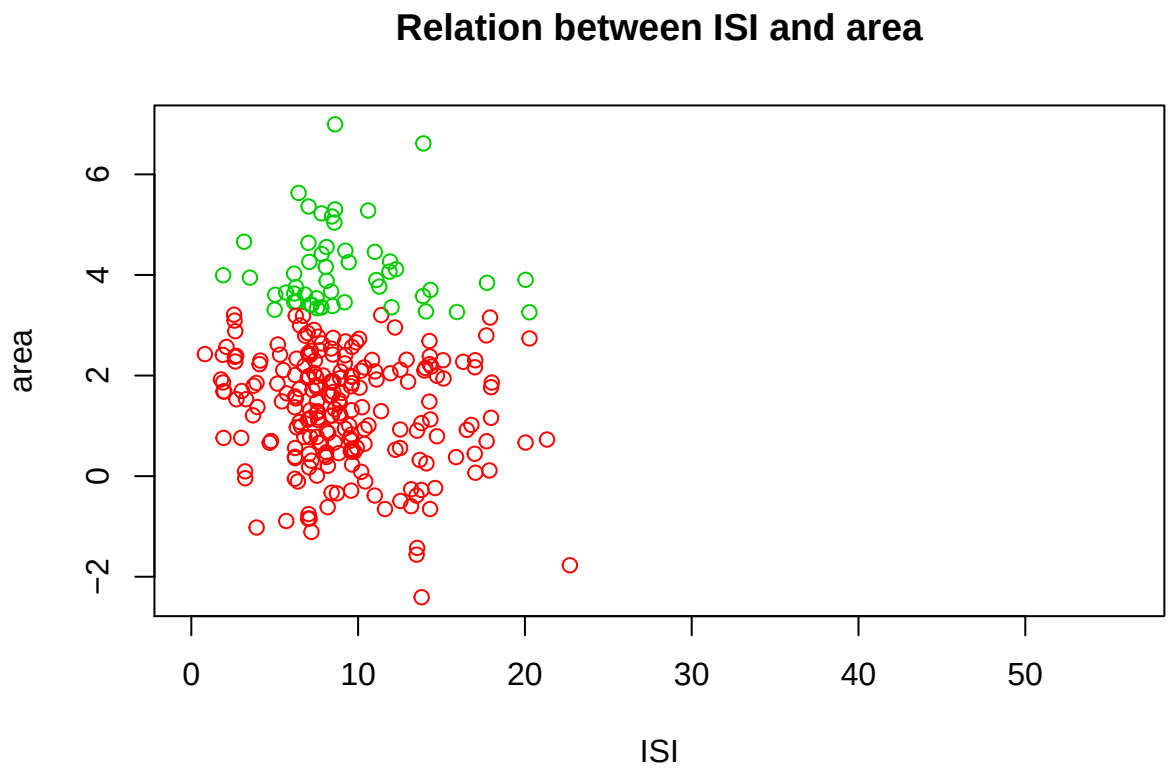
```
# Now we will check ISI index

summary(forest_fire$ISI)
```

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

```
# Here is the histogram of ISI index

hist(forest_fire$ISI, main = "Initial Spread Index",
     xlab = NULL)
axis(1, at = 0:6)
```




```
# Plot ISI and Burned Area

plot(jitter(forest_fire$ISI, factor=2),jitter(forest_fire$logarea, factor=2 ),
     col=factor(forest_fire$fire_size),
     xlab = "ISI", ylab = "area",
     main = "Relation between ISI and area")
```



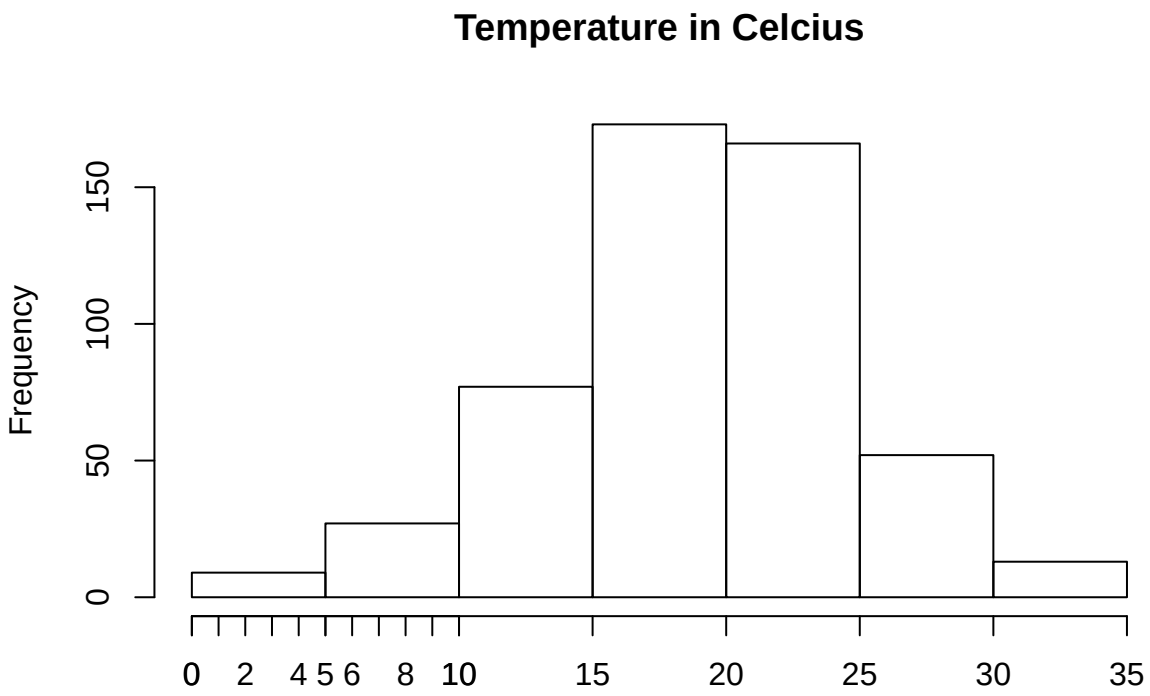

```
legend=levels(ff_level$fire_size)

#Observation: Fire tendency is highest when ISI is from 5 to 15
```

Temperature

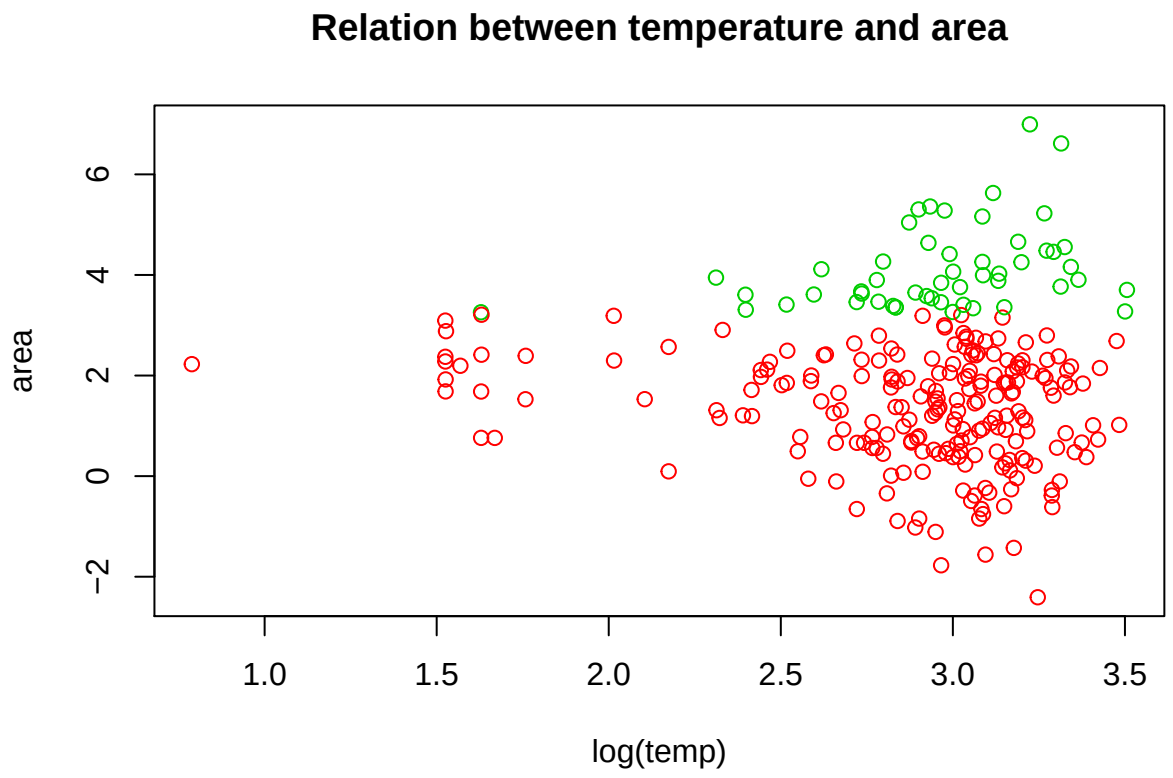
```
#How temperature data looks like. First looking at the summary:

summary(forest_fire$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.20   15.50   19.30   18.89   22.80   33.30
```

```
# Here is the histogram of Temperature in Celcius

hist(forest_fire$temp, main = "Temperature in Celcius",
     xlab = NULL)
axis(1, at = 0:10)
```




```
# Let's plot temperature and burned area

plot(jitter(log(forest_fire$temp), factor=2),jitter(forest_fire$logarea, factor=2 ),
     col=factor(forest_fire$fire_size),
     xlab = "log(temp)", ylab = "area",
     main = "Relation between temperature and area")
```




```
legend=levels(forest_fire$fire_size)

#Observation: Maximum fire incident occurs when temperature is in between 15 - 25
```

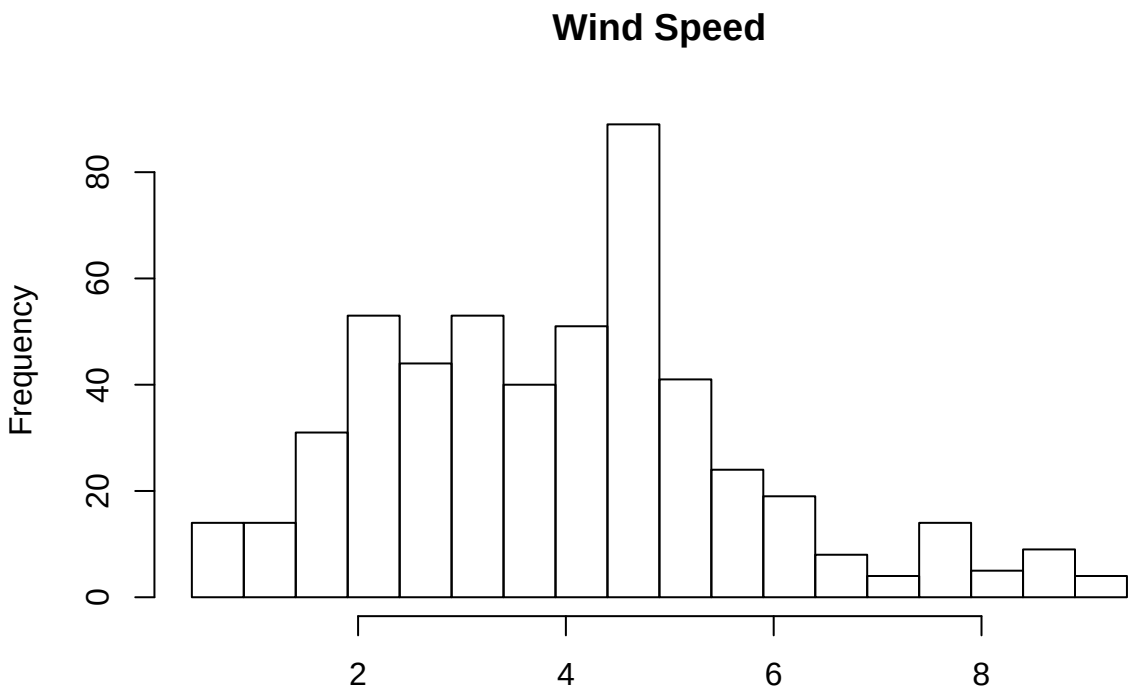
Wind

Below is a histogram of wind.

```
summary(forest_fire$wind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.400   2.700   4.000   4.018   4.900   9.400
```

```
hist(forest_fire$wind,breaks=seq(0.4,9.4,0.5),main = "Wind Speed",xlab= NULL)
```



Visually, the histogram seems to have a positive skew (right skew). This means that there are observations stretching further to the right of the bulk of the data. Note from summary that the mean is greater than the median, which is typically what we see for positively skewed variables.

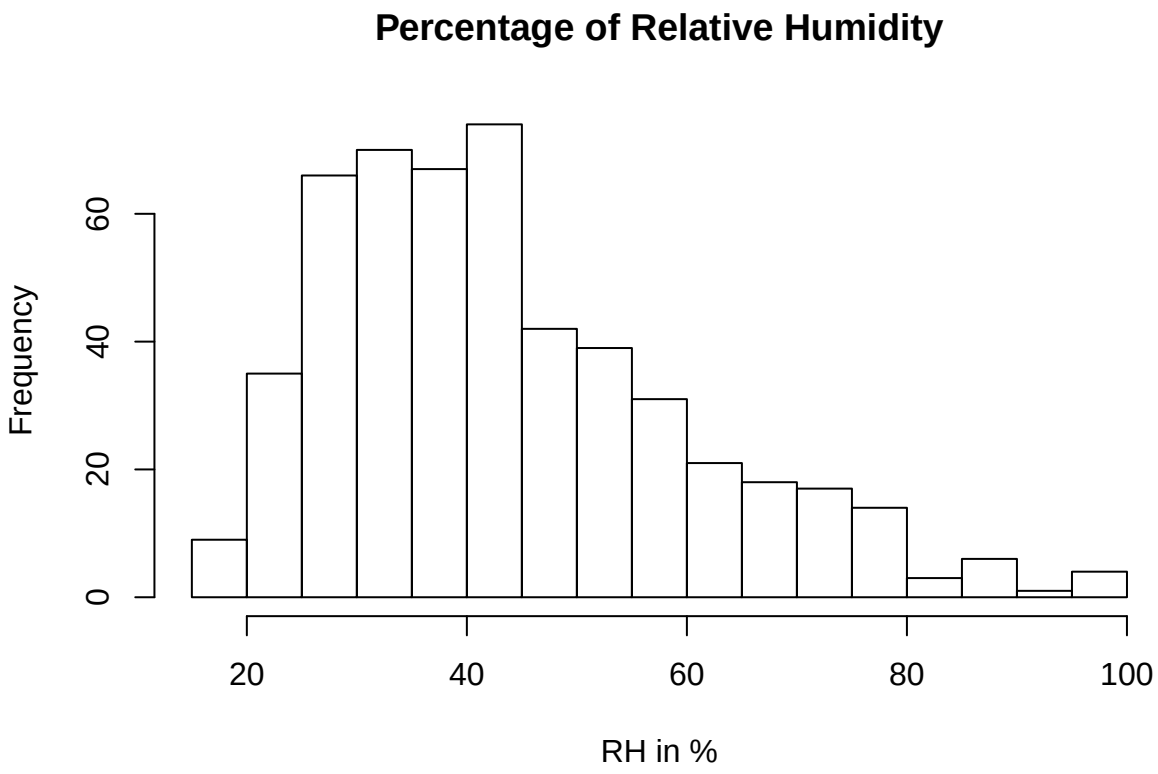
Relative Humidity

Next we look at a histogram of Relative humidity

```
summary(forest_fire$RH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   33.00   42.00   44.29   53.00  100.00
```

```
hist(forest_fire$RH,breaks=seq(15,100,5),main = "Percentage of Relative Humidity",xlab= "RH in %")
```



Again, we see the same evidence of right skew as in wind speed.

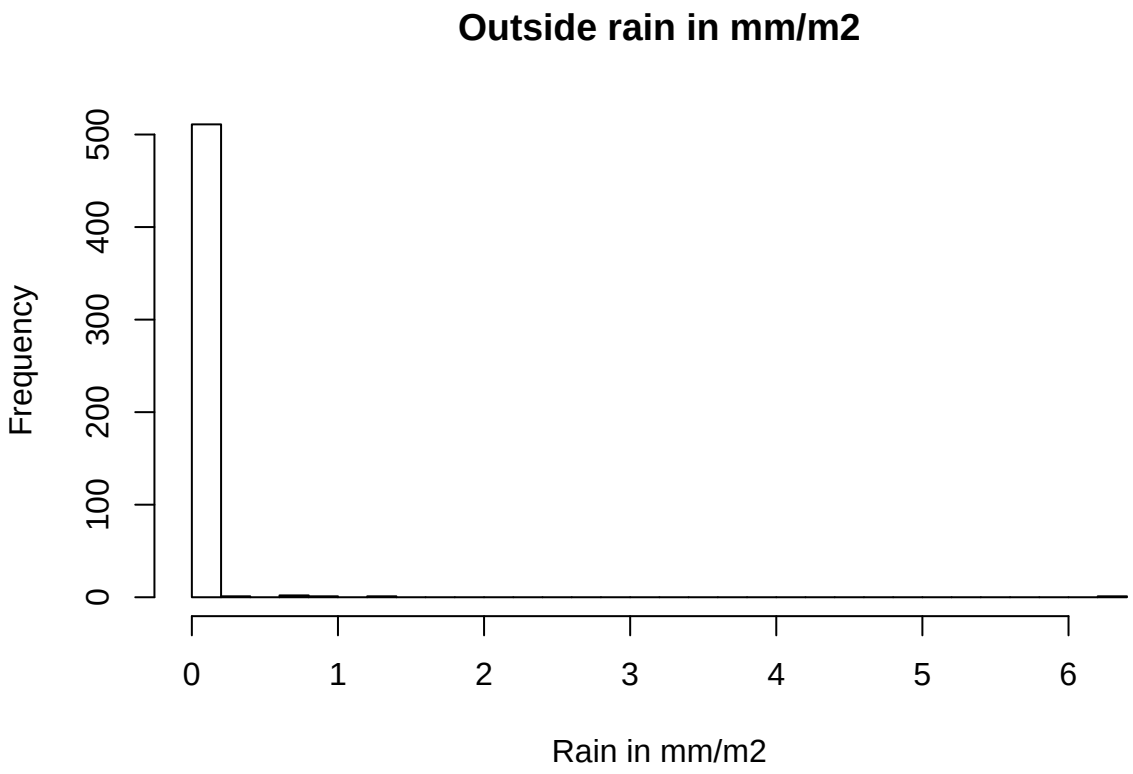
Rain

We next examine our Rain variable. We notice that it is mostly 0's.

```
summary(forest_fire$rain)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.02166 0.00000 6.40000
```

```
hist(forest_fire$rain, breaks=seq(0.0,6.4,0.2), main = "Outside rain in mm/m2", xlab = "Rain in mm/m2" )
```



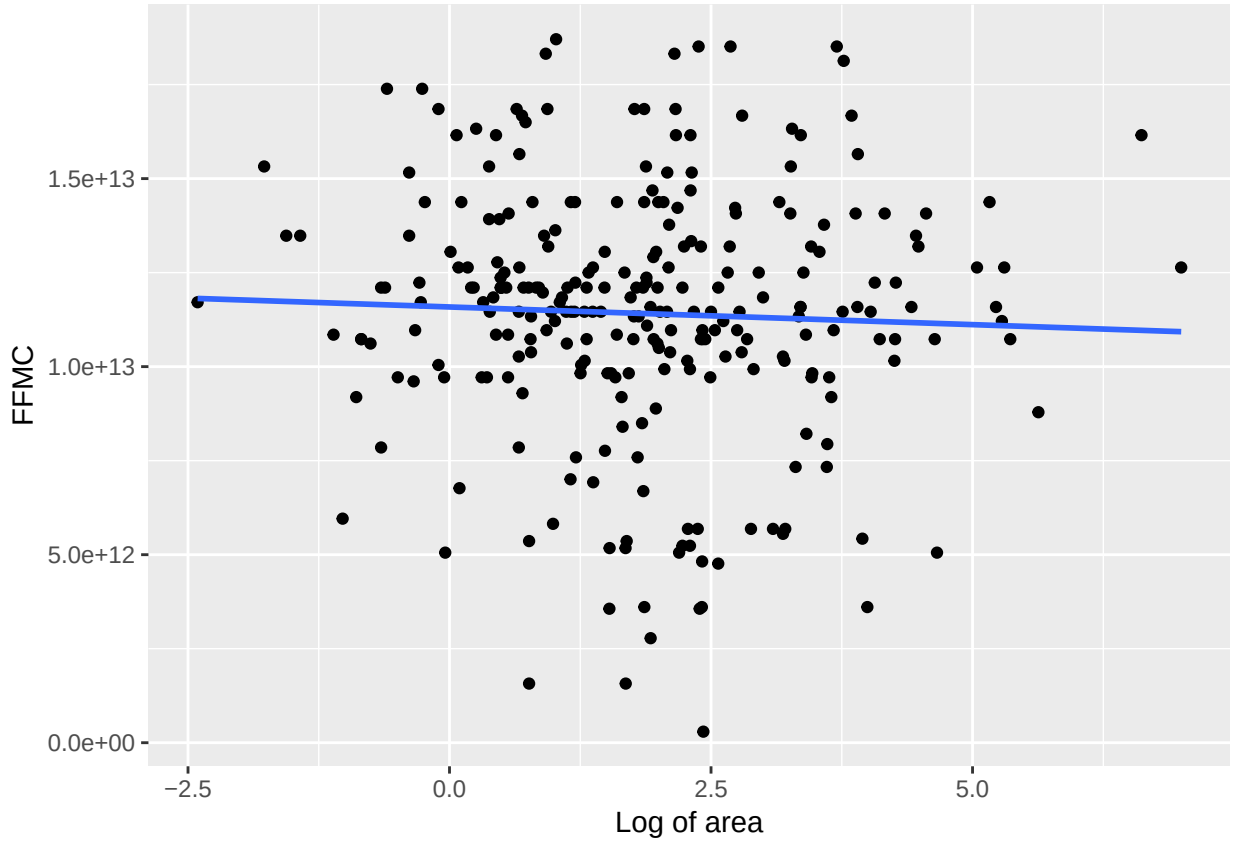
The distribution seems to spike around zero, which means that there was no rain at all.

boxplot

```
#boxplot(rain ~ RH ~ logarea, data = forest_fire)
```

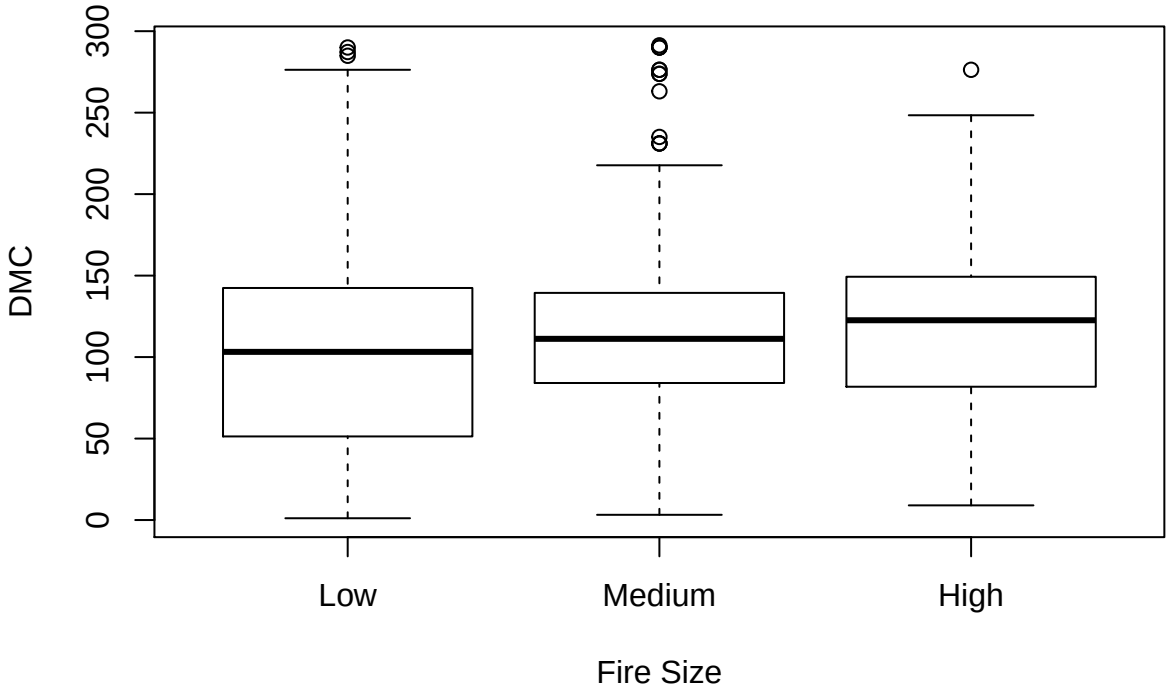
FFMD, DMC and day

```
#plot(forest_fire$fire_size, forest_fire$FFMC^10/factorial(10), xlab= 'Fire Size', ylab = 'Transformed
ggplot(Med_High_fire, aes(x=log(area), y=FFMC^10/factorial(10)) ) +geom_point()  + labs(x="Log of area"
```



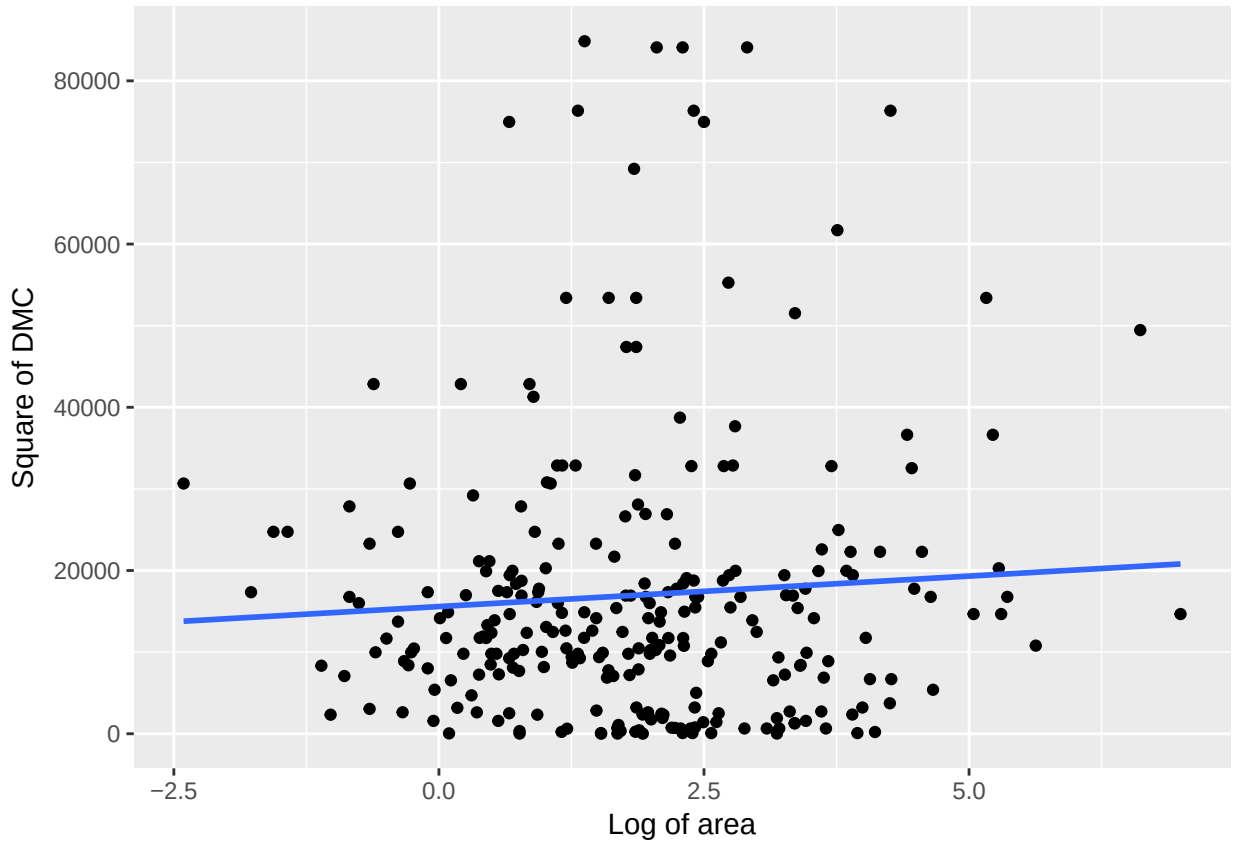
FFMC seems to have no bearing on the size of the fire.

```
plot(forest_fire$fire_size, forest_fire$DMC, xlab= 'Fire Size', ylab = 'DMC')
```



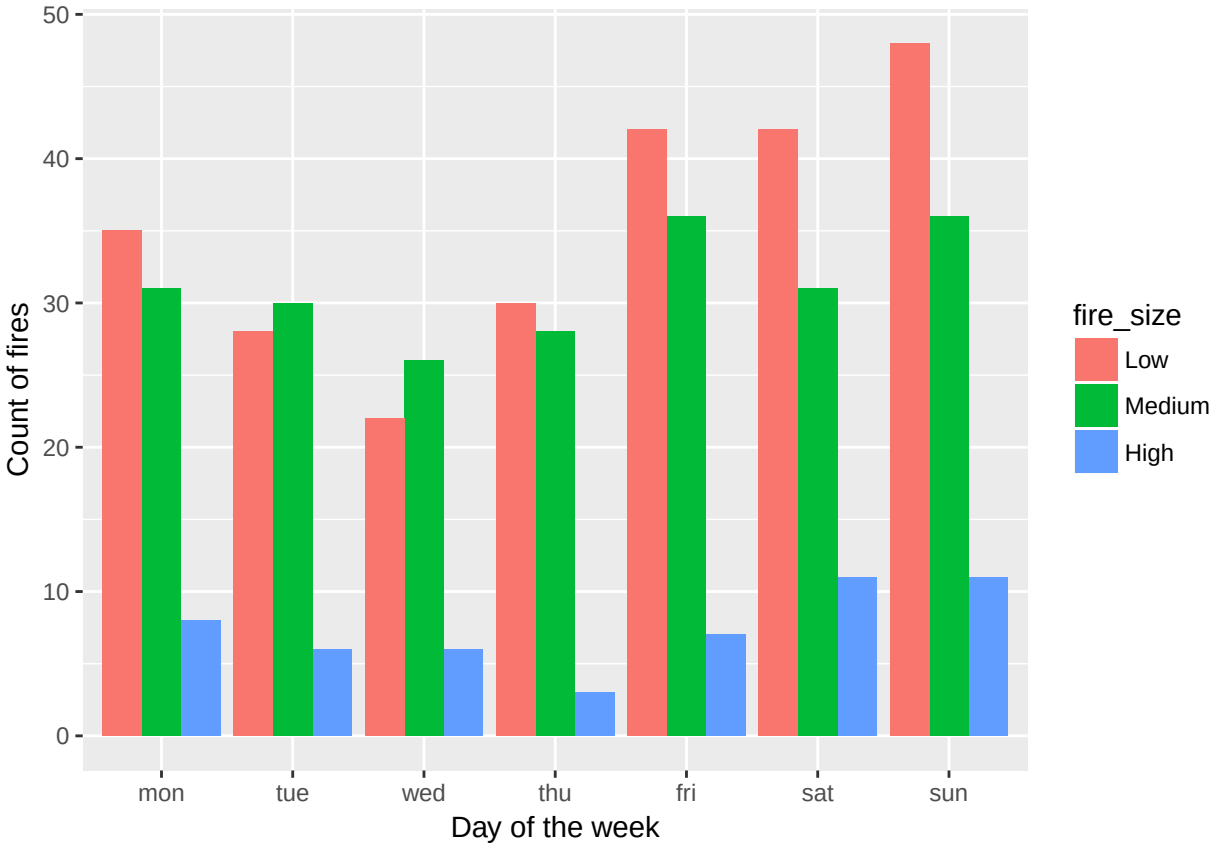
DMC seems to have no relationship to the size of the fire.

```
#plot(log(Med_High_fire$area), Med_High_fire$DMC, xlab= 'Fire Size', ylab = 'DMC')
ggplot(Med_High_fire, aes(x=log(area), y=DMC^2) ) +geom_point()  + labs(x="Log of area", y="Square of DM
```



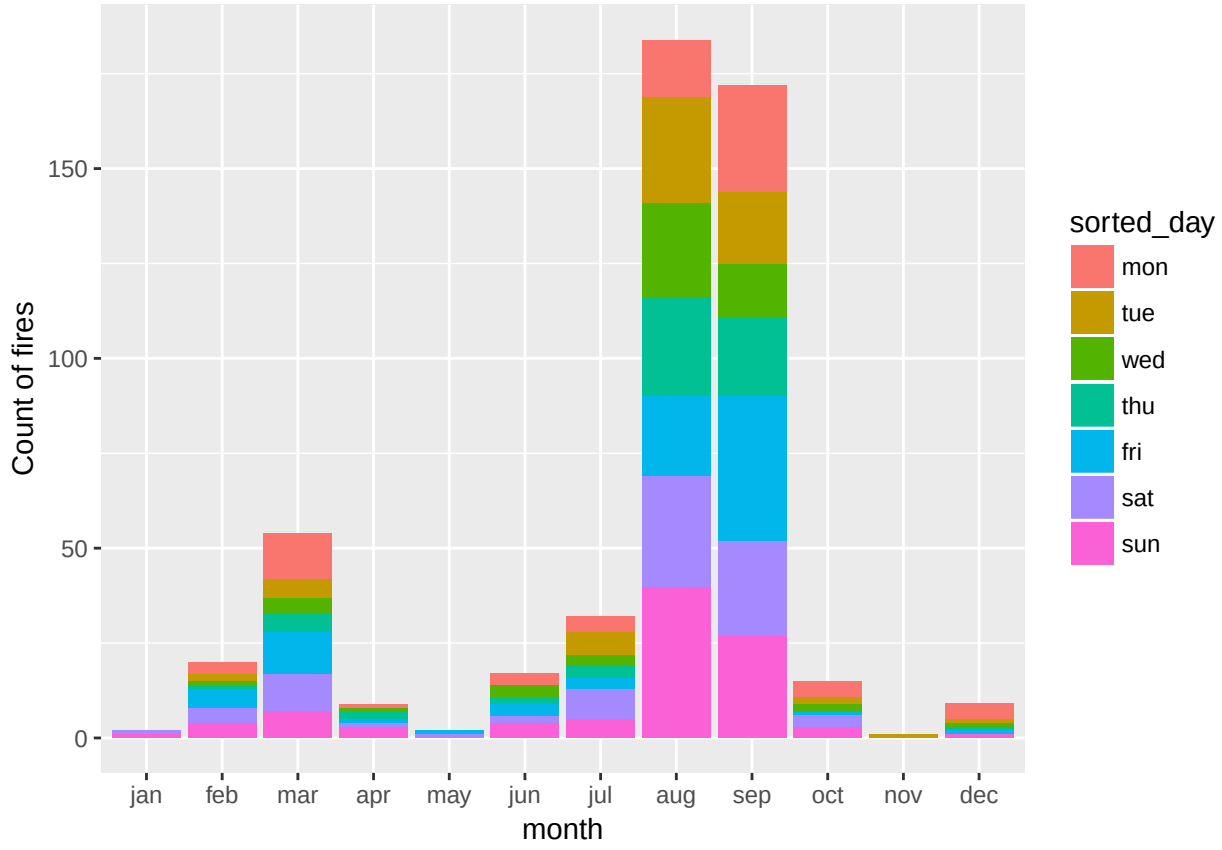
Even after scaling the variables, slope is barely perceptible

```
ggplot(forest_fire, aes(sorted_day, ..count..))+geom_bar(aes(fill=fire_size), position = "dodge") + labs
```



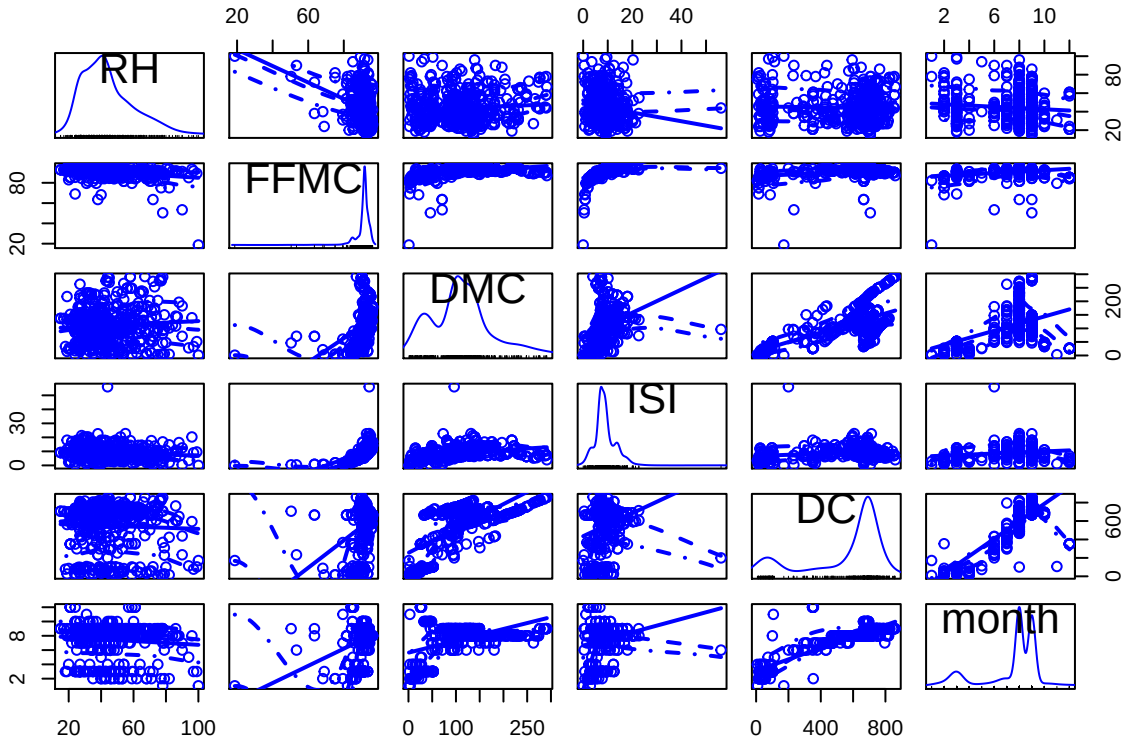
Friday, Saturday and Sunday seem to have seen more fires overall. Saturday and Sunday have seen more large fires

```
ggplot(forest_fire, aes(month, ..count..))+geom_bar(aes(fill=sorted_day)) + labs(x="month") + labs( y="
```




```
# + facet_grid (fire_size~.)
```

```
scatterplotMatrix(~RH+FFMC+DMC+ISI+DC+month, data=forest_fire)
```



Analysis of Key Relationships

Analysis of Secondary Effects

Conclusion