Forest Fire EDA

Debalina Maiti, Mark Paluta, Tina Agarwal, Vivek Agarwal

5/28/2018

Introduction

In some areas, forest fires are a major environmental concern, endangering human lives and causing substantial economic damage.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 we load 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. There are no missing values in our dataset. These variables represent a mix of quantitative and categorical variables.

```
nrow(forest_fire)
```

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

```
names(forest_fire)
```

```
##  [1] "X"     "Y"     "month" "day"   "FFMC"  "DMC"   "DC"    "ISI"
##  [9] "temp"  "RH"    "wind"  "rain"  "area"
```

```
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
```

Transformations

We required several transformations to our data before we could properly explore variables and trends. Month and day of the week require sorting in chronological order rather than alphabetically. Forest fire size is skewed right but unfortunately has a large number of size zero fires, so a log transformation an be performed but has limited effectivity. We thus also mask fire size into a low, medium, and high factor. We interpret size zero fires to mean that a fire did occur but had a very small burn area, below the minimum measurable, as opposed to a data point about a fire not occuring. This is a very important assumption going forward in our analysis and it would be worth confirming with our data supplier before making recommendations to our customer.

Note: the mean fire area is 12.85. We categorize fires of size 0 as low, 0 to 12.85 as medium, and above 12.85 as high.

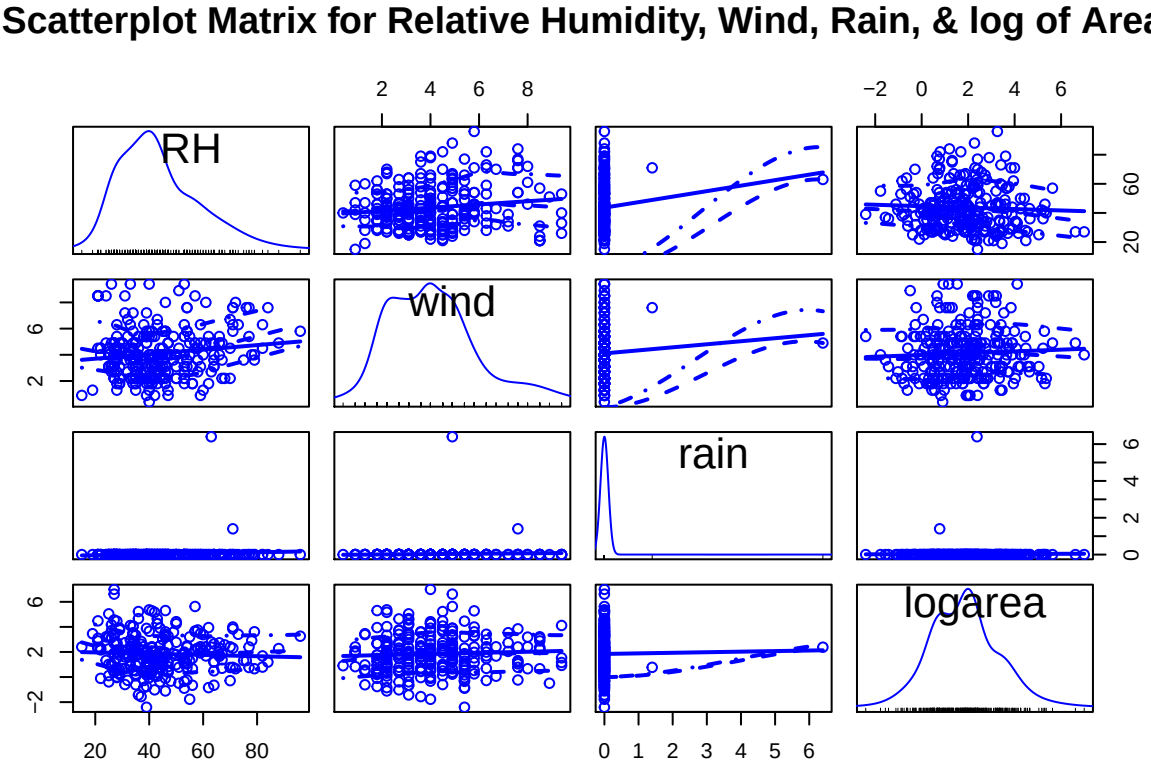
```
forest_fire$sorted_day<-factor(forest_fire$day, levels = c("mon","tue","wed","thu","fri","sat","sun" ))

forest_fire$fire_size <- cut(forest_fire$area, breaks=c(-Inf,0.01, 25, Inf),
                             labels=c('Low','Medium','High'))
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",]$dotsize<-1
Med_High_fire <- forest_fire[forest_fire$fire_size != "Low",]
forest_fire$logarea = log(forest_fire$area)
forest_fire[forest_fire == -Inf] <- NA
forest_fire$month = factor(forest_fire$month,levels(forest_fire$month)[c(5,4,8,1,9,7,6,2,12,11,10,3)])
pdfire<-forest_fire[forest_fire$fire_size=="High",] #particularly damaging fire
```

Univariate Analysis of Key Variables

With transformations complete, 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. Note that to plot all the 13 variables in one matrix reduces legibility, so we will just visualize a subset of the variables here to illustrate the concept.

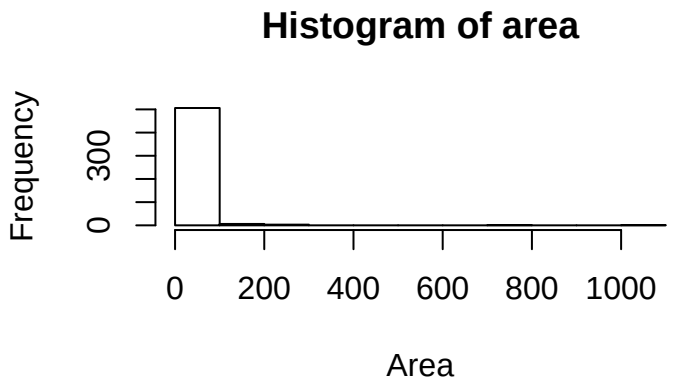
```
scatterplotMatrix(~ RH + wind + rain + logarea, data=forest_fire, smooth = TRUE,
                  main = "Scatterplot Matrix for Relative Humidity, Wind, Rain, & log of Area")
```



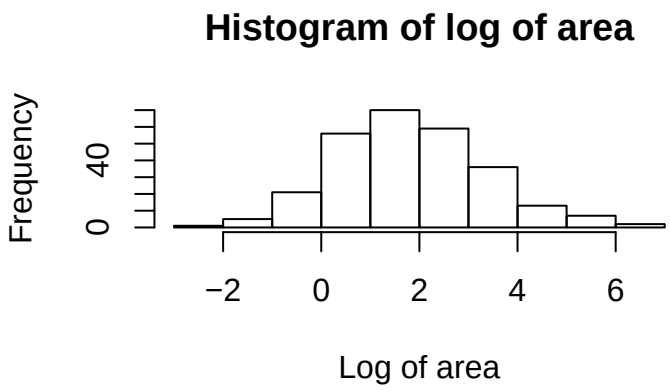
Area

Area appears to be right skewed. To normalize the data, we took the log of area above in our transformations section. Note that as addressed above, this histogram is excluding fires of size zero. Below we show histograms of the raw area, the log of area, and finally the factor of fire size as defined above.

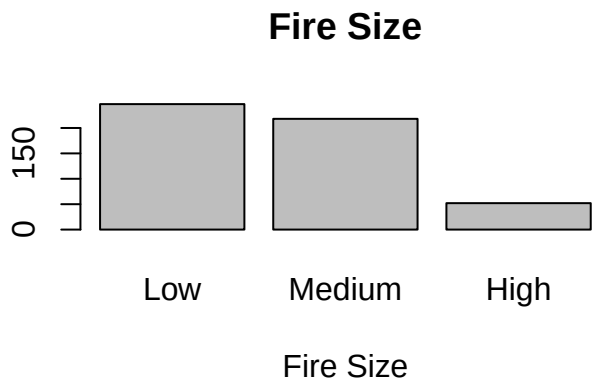
```
hist(forest_fire$area, xlab="Area", main='Histogram of area')
```




```
hist(forest_fire$logarea, xlab="Log of area",main="Histogram of log of area")
```



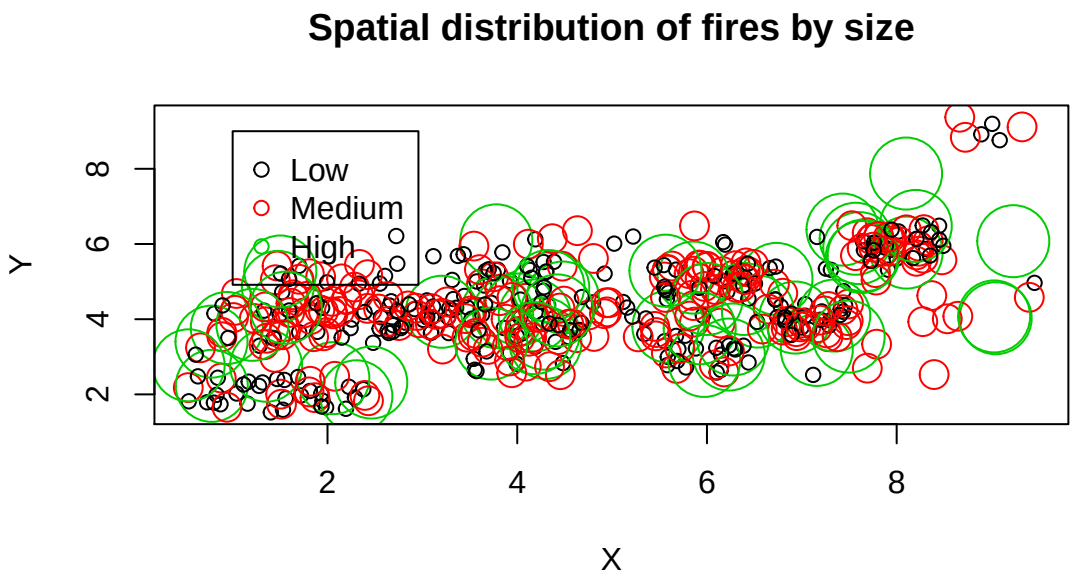

```
barplot(table(forest_fire$fire_size), xlab="Fire Size",main="Fire Size")
```



Spatial coordinates (X & Y)

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. We also exaggerate the size of the circles so that it is easier to notice the difference. Fires appear to be contained in a strip ranging from lower left corner to the upper right corner, so we suspect this represents the footprint of the park. Also note that the upper left corner is especially bare of fires, so it may not be within the park or may not have many trees. Beyond that, we don't see a trend of fire size with respect to geography within the affected areas.

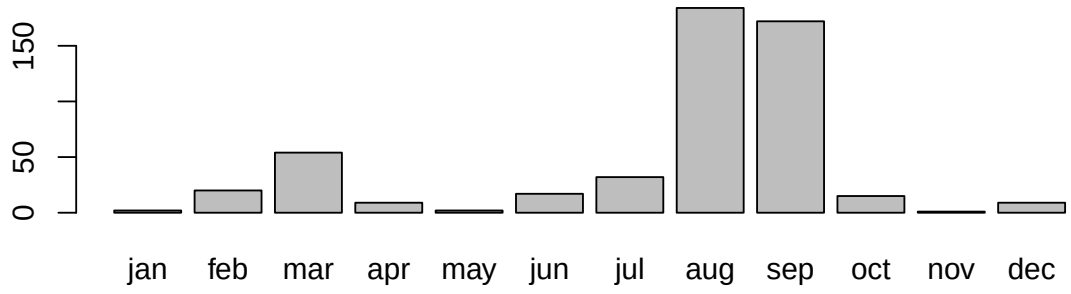
```
plot(jitter(forest_fire$X, factor=2.5), jitter(forest_fire$Y, factor=2.5),
     col=forest_fire$fire_size, cex=forest_fire$dotsize, xlab="X", ylab="Y", main = "Spatial distributi
legend(1,9,unique(forest_fire$fire_size),col=1:length(forest_fire$fire_size),pch=1)
```



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. Our interpretation is that summer results in the most fires with a smaller bimodal effect in the spring.

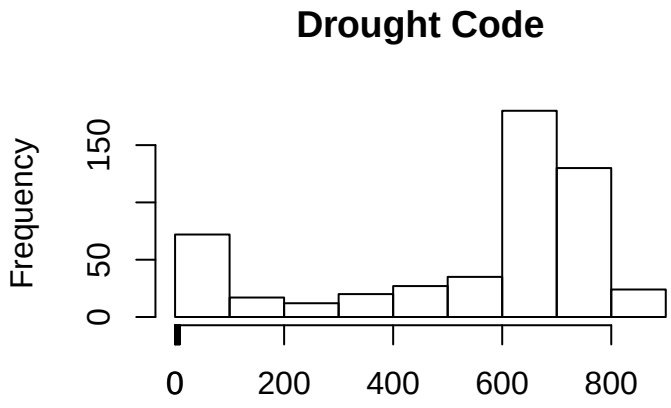
```
barplot(table(forest_fire$month))
library(ggplot2)
```



DC

Looking into Drought Code, we see a left skew with its primary mode around 700 and a second smaller bump near zero.

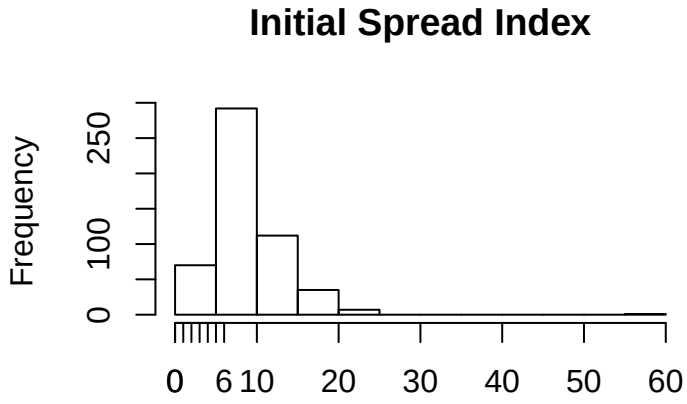
```r
hist(forest_fire$DC, main = "Drought Code",
     xlab = NULL)
axis(1, at = 0:9)
```



ISI

Looking at the Initial Spread Index, we see a skewed right distribution with mode between 5 and 10.

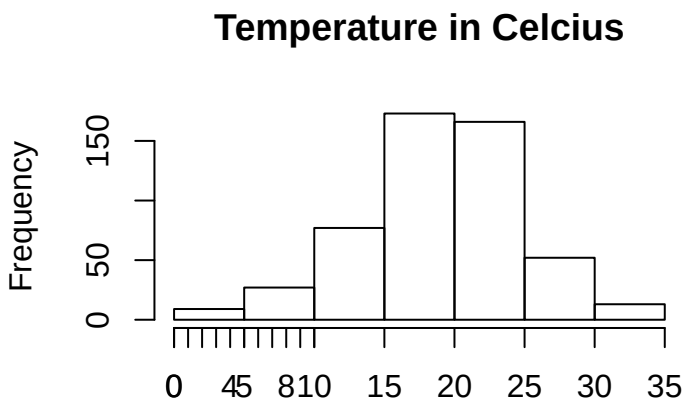
```r
hist(forest_fire$ISI, main = "Initial Spread Index",
     xlab = NULL)
axis(1, at = 0:6)
```



Temperature

Temperature appears to be approximately normally distributed and spans from approximately 0 to 35 degrees Celcius.

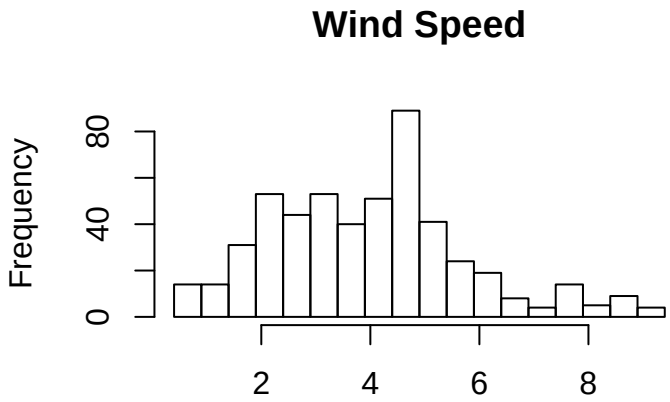
```
hist(forest_fire$temp, main = "Temperature in Celcius",
     xlab = NULL)
axis(1, at = 0:10)
```



Wind

Below is a histogram of wind. The histogram shows a slight positive skew.

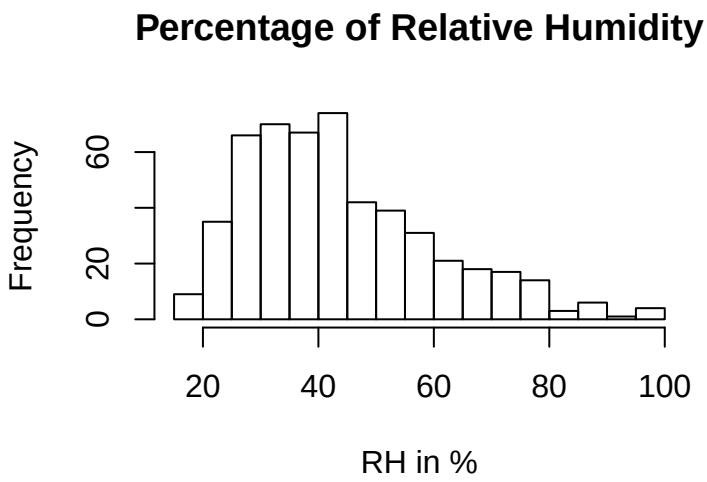
```
hist(forest_fire$wind,breaks=seq(0.4,9.4,0.5),main = "Wind Speed",xlab= NULL)
```



Relative Humidity

Next we look at a histogram of Relative humidity, which exhibits a noticeable right skew.

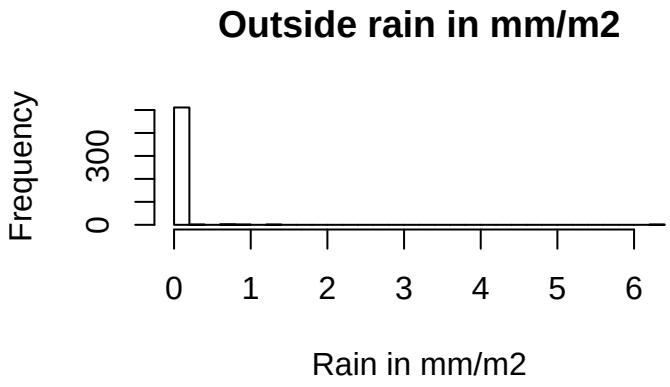
```
hist(forest_fire$RH,breaks=seq(15,100,5),main = "Percentage of Relative Humidity",xlab= "RH in %")
```



Rain

We next examine our Rain variable. We notice that it is mostly zeroes, meaning no rain or no measureable rain. This variable will likely not be very effective in predicting fire size as it has such little variation overall. We could convert this variable to "rain" versus "no rain", but with almost no non-zero data, it would be difficult to capture any significant relationship to area.

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

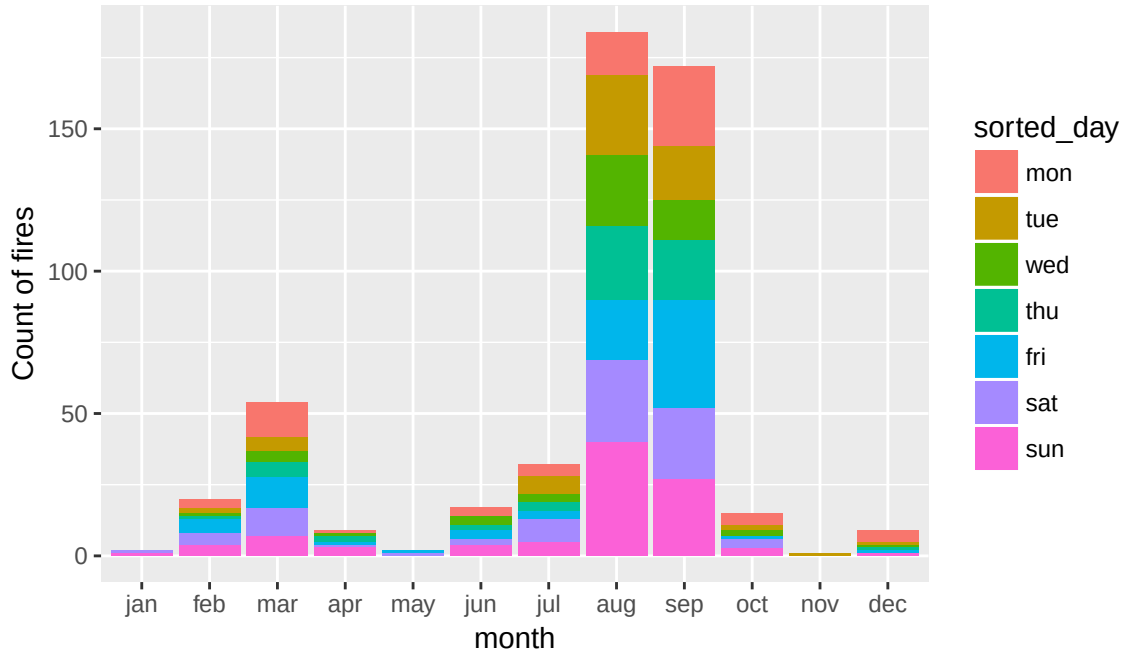


From the above Univariate Analysis of Key Variables the major observations are as follows. Note that these observations are thus far only related to quantity of fires, and we will look into fire size and particularly damaging fires in the next section.

1. Most fire incidents took place during the months of August and September
2. The number of fire incidents is higher on Friday, Saturday, and Sunday than any other day of the week.
3. In most of the fire events FFMC was maximum: >90
4. Most fires occured when DMC was in the range of 80-140,
5. Fire tendency is highest for 600 <= Drought Code <=800,
6. Fire tendency is highest when ISI is from 5 to 15
7. Temperature in the range from 15 - 25 degree celcius is most prone to fires
8. When RH <= 50%, fire tendency is high
9. Most fire incidents occurred when wind speed was 4-5 km/ hr.

Bivariate Relationship between Month, Day of the Week and Count of Fire

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



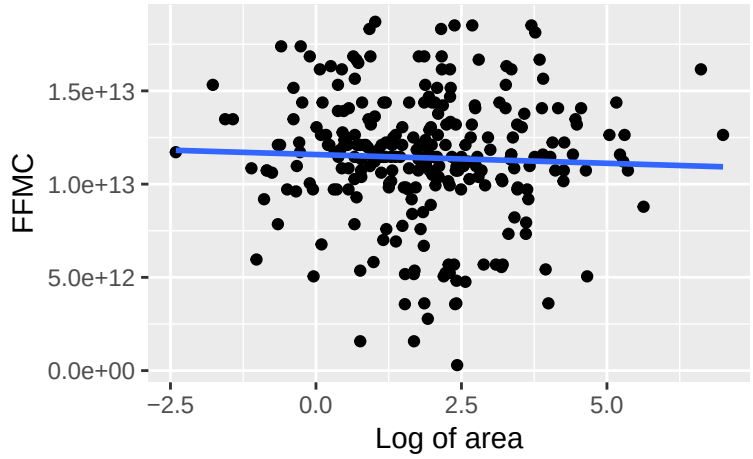
Observation: From the above plot we clearly see that fire tendency is much more higher during the weekend-s (Fri, Sat, Sun) of August and September than the other time of the year.

So while we will be analyzing other key relationships, we will check the conditions of other matrics during August and September.

Bivariate Relationship between FFMC and Fire

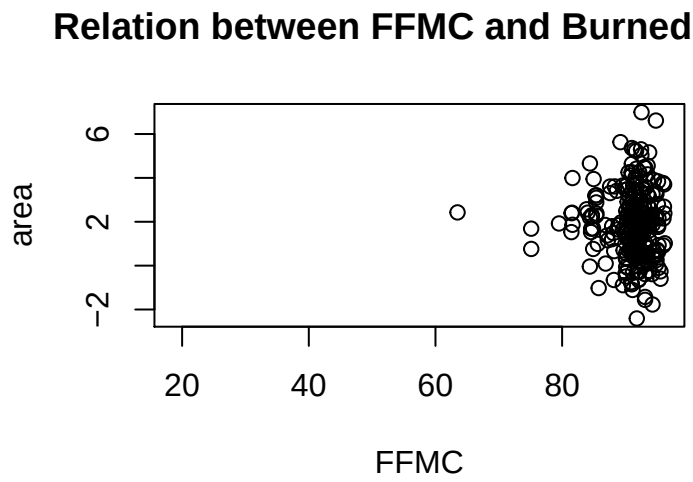
Let's first plot FFMC against fire size:

```
ggplot(Med_High_fire, aes(x=log(area),
        y=FFMC^10/factorial(10)) ) +geom_point()+ labs(x="Log of area",
        y="FFMC")+geom_smooth(method="lm", se=F)
```



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

```
plot(jitter(forest_fire$FFMC, factor=2),jitter(forest_fire$logarea, factor=2 ),
     xlab = "FFMC", ylab = "area", main = "Relation between FFMC and Burned Area")
```

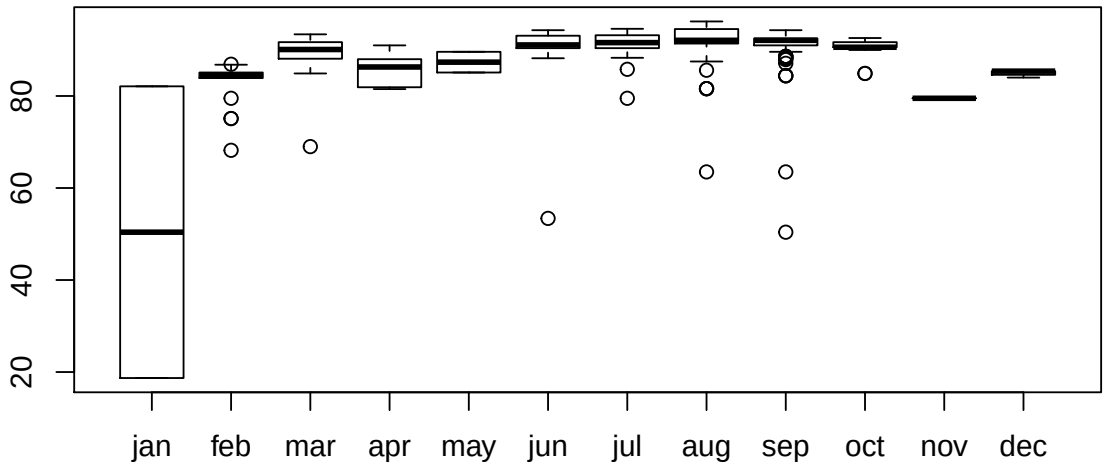



```
legend=levels(forest_fire$fire_size)
```

Observation: We don't see a specific linear relationship here; but plot clearly indicates fire occurs when FFMC>80.

Now let's see what is the approximate value of FFMC during August and September:

```
plot(forest_fire$month, forest_fire$FFMC)
```

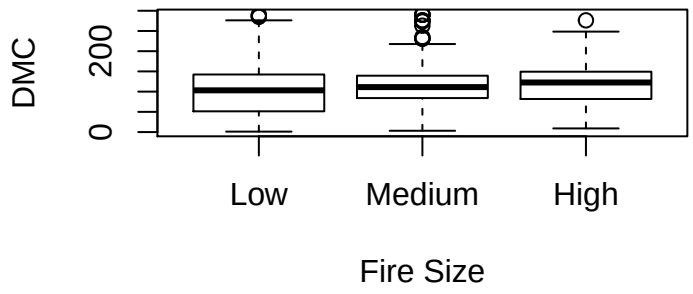


And the plot shows that FFMC > 80 during this time, which is one of the stimulating condition of forest fire.

Bivariate Relationship between DMC and Burned Area

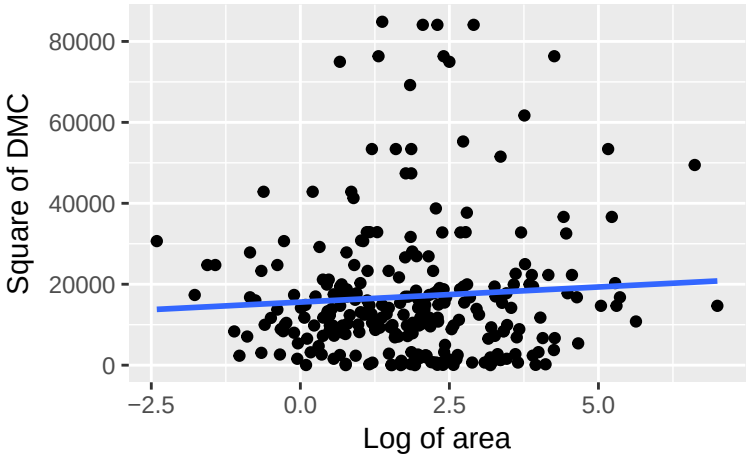
Let's first plot DMC against fire size:

```
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.Let's scale the data little bit and try to analyse relationship

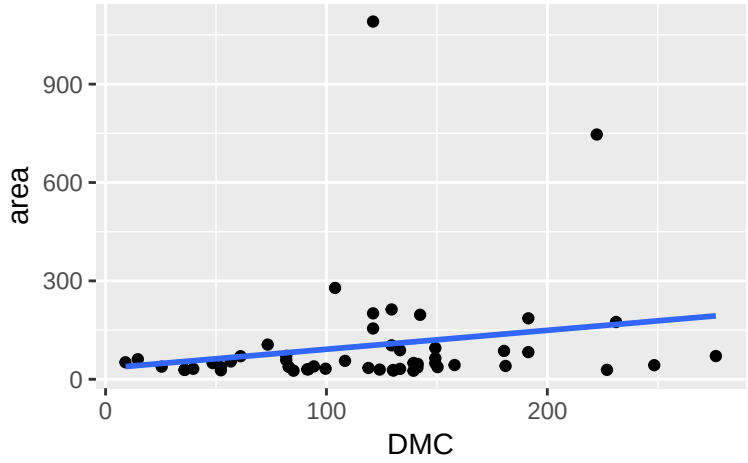
```
ggplot(Med_High_fire,
       aes(x=log(area), y=DMC^2) ) +geom_point()+ labs(x="Log of area", y="Square of DMC")+ geom_smooth
```



Even after scaling the variables, slope is barely perceptible

Observation: We don't see a very impressive relationship here between DMC and forest fire.

```
ggplot(pdfire, aes(x=DMC, y=area)) + geom_point() + geom_smooth(method = 'lm', se = F)
```



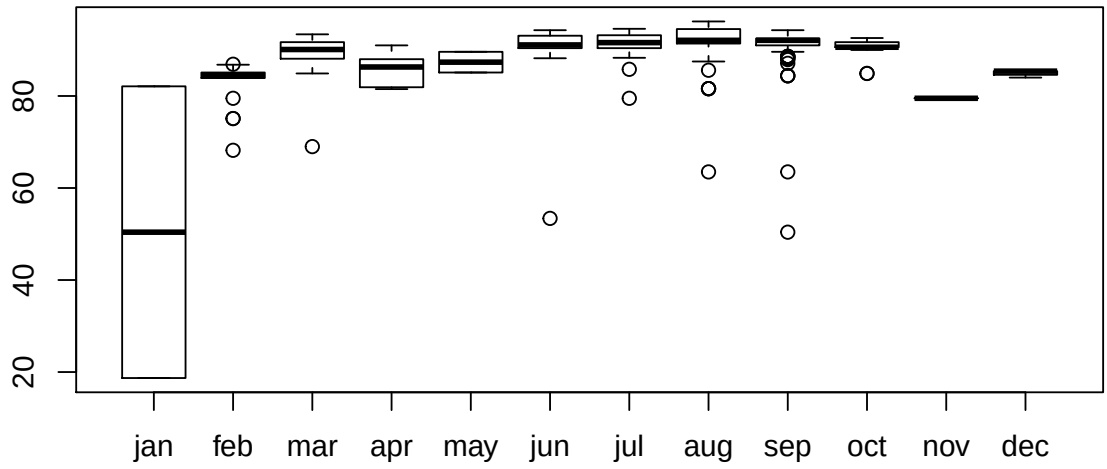

```
cor(pdfire$DMC, pdfire$area,use = "complete.obs")
```

```
## [1] 0.1963789
```

From the above plot and correlation it looks like DMC and higher burned area positively corelated.

Now let's see what is the approximate value of DMC during August and September and whether that belongs to the range of 80-140 (per our observation in Key Variable Analysis) :

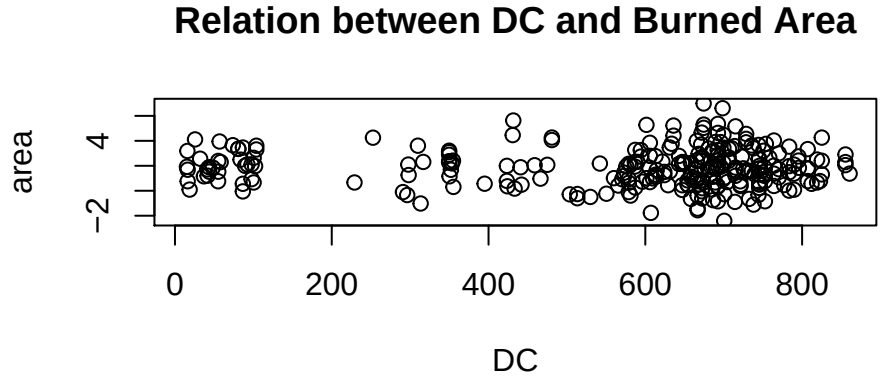
```
plot(forest_fire$month, forest_fire$FFMC)
```



Plot confirms that during August and September DMC value is greater than 80.

Bivariate Relationship between DC and Burned Area

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

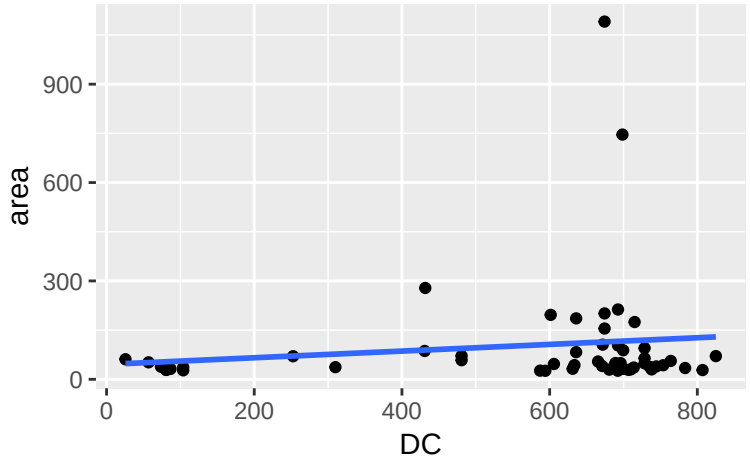



```
legend=levels(forest_fire$fire_size)
```

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

Let's plot only against the heavy fire incidents:

```
ggplot(pdfire, aes(x=DC, y=area)) + geom_point() + geom_smooth(method = 'lm', se = F)
```



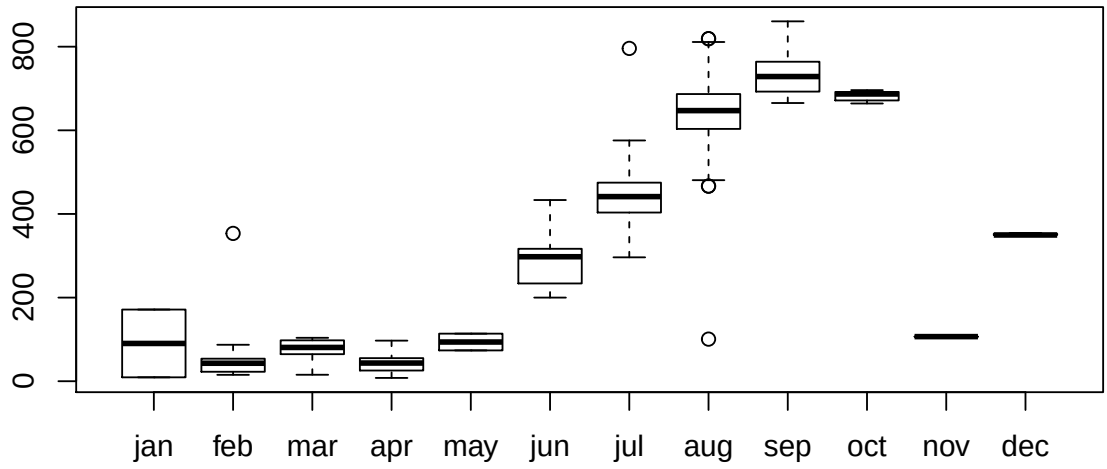

```
cor(pdfire$area,pdfire$DC)
```

```
## [1] 0.1358095
```

From the above plot and correlation it looks like DC and higher burned area positively corelated.

Now let's see what is the approximate value of DC during August and September:

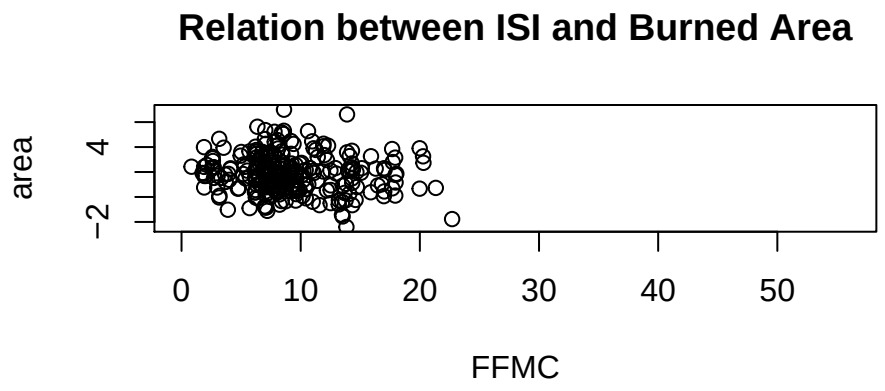
```
plot(forest_fire$month, forest_fire$DC)
```



And the plot shows that DC belongs to the range of 600 - 800 during this time, which is one of the instigating condition of forest fire.

Bivariate Relationship between ISI and Burned Area

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

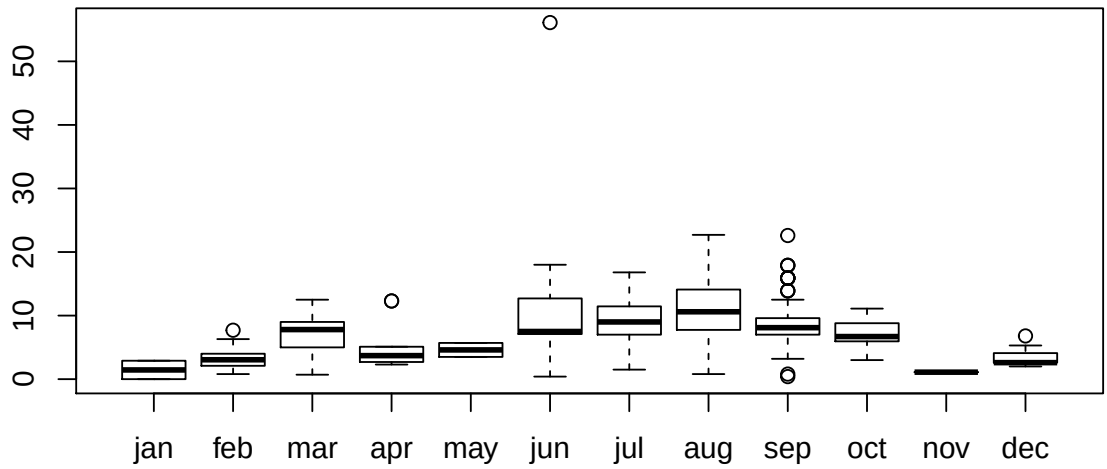



```
legend=levels(forest_fire$fire_size)
```

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

Now let's see what is the approximate value of ISI during August and September:

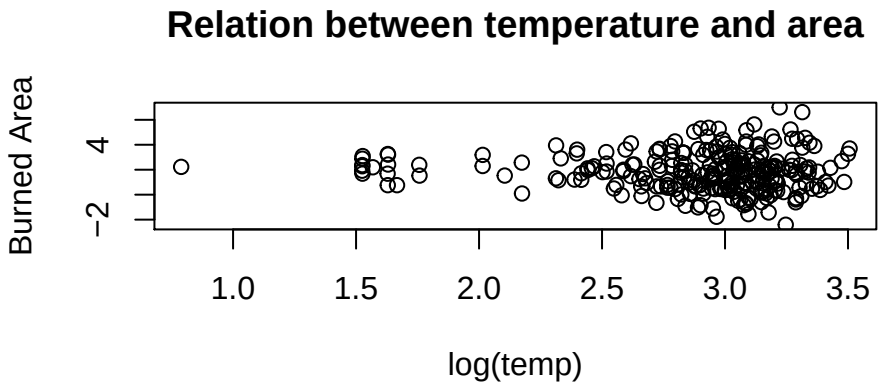
```
plot(forest_fire$month, forest_fire$ISI)
```



And the plot shows that ISI belongs to 5-15 range during this time, which is another provoking condition of forest fire.

Bivariate Relationship between Temperature and Burned Area

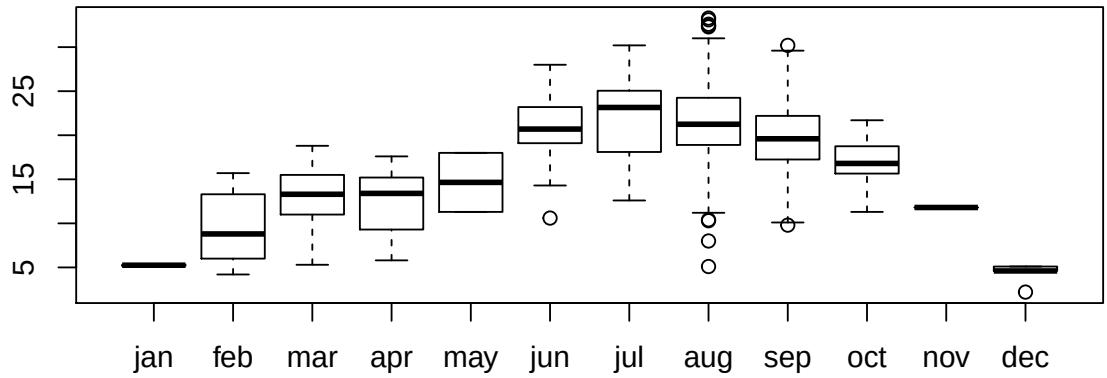
```
plot(jitter(log(forest_fire$temp), factor=2),jitter(forest_fire$logarea, factor=2 ),
     xlab = "log(temp)", ylab = "Burned Area",
     main = "Relation between temperature and area")
```




```
legend=levels(forest_fire$fire_size)
```

Observation: Maximum fire incident occurs when temperature is in between 15 - 25. This is much more distinct when we use log of temperature Now let's see what is the approximate value of Temperature during August and September:

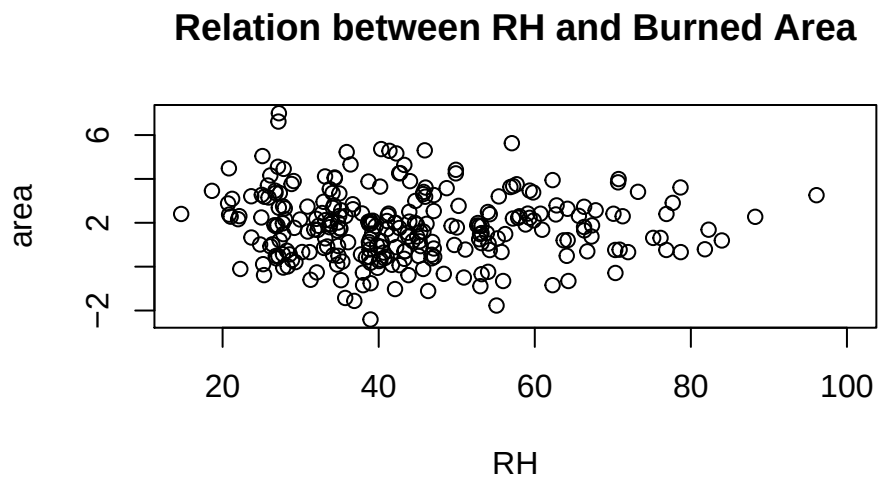
```
plot(forest_fire$month, forest_fire$temp)
```



And the plot shows that temperature hovers between 15 - 25 degree celcius during this time, which is onother ideal condition of forest fire per our analysis.

Bivariate Relationship between RH and Burned Area

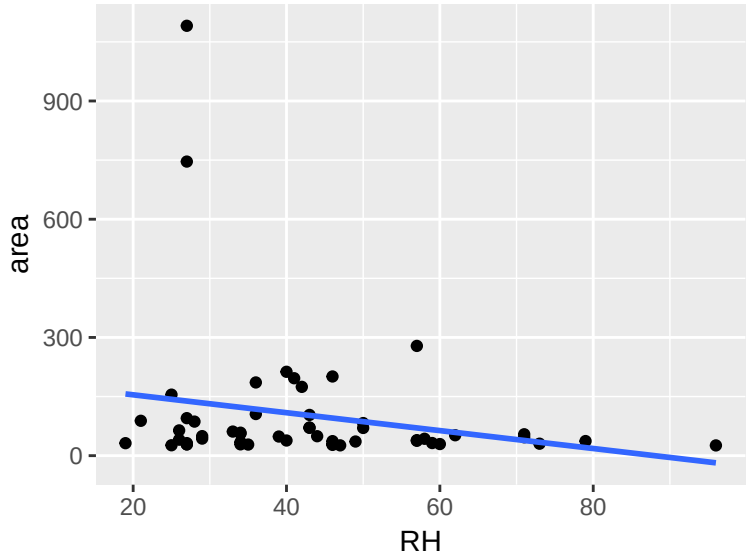
```
plot(jitter(forest_fire$RH, factor=2),jitter(forest_fire$logarea, factor=2 ),
     xlab = "RH", ylab = "area",
     main = "Relation between RH and Burned Area")
```




```
legend=levels(forest_fire$fire_size)
```

Observation: We don't see a specific linear relationship here. We will try to just take the subset of data against heavy fire incidents

```
ggplot(pdfire, aes(x=RH, y=area)) + geom_point() + geom_smooth(method = 'lm', se = F)
```



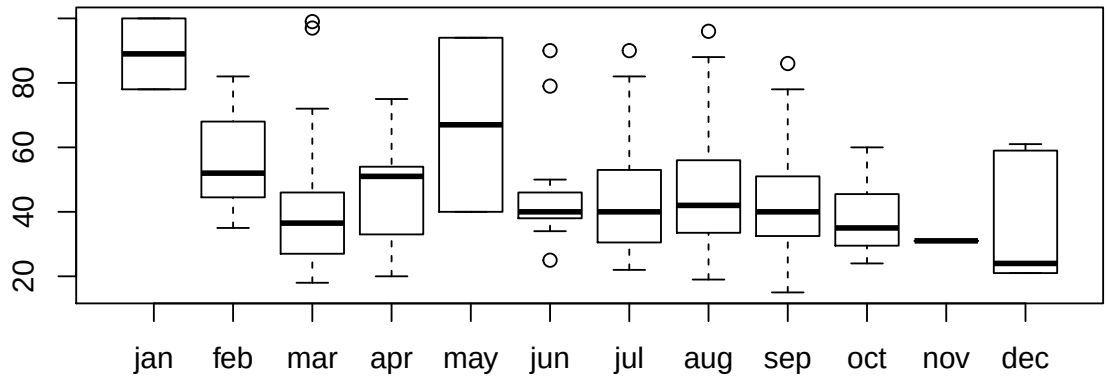

```
cor(pdfire$RH, pdfire$area)
```

```
## [1] -0.2081013
```

It looks like RH is negatively co-related with Burned area.

Now let's see what is the approximate value of RH during August and September:

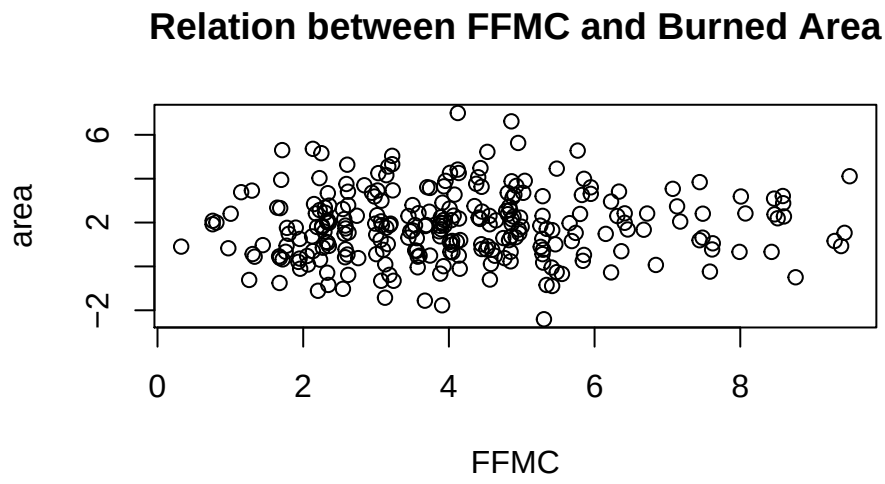
```
plot(forest_fire$month, forest_fire$RH)
```



And the plot shows that RH is in the range of 40-50 during this time, which coincides with forest fire instigating index.

Bivariate Relationship between Wind and Burned Area

```
plot(jitter(forest_fire$wind, factor=2),jitter(forest_fire$logarea, factor=2 ),
     xlab = "FFMC", ylab = "area",
     main = "Relation between FFMC and Burned Area")
```

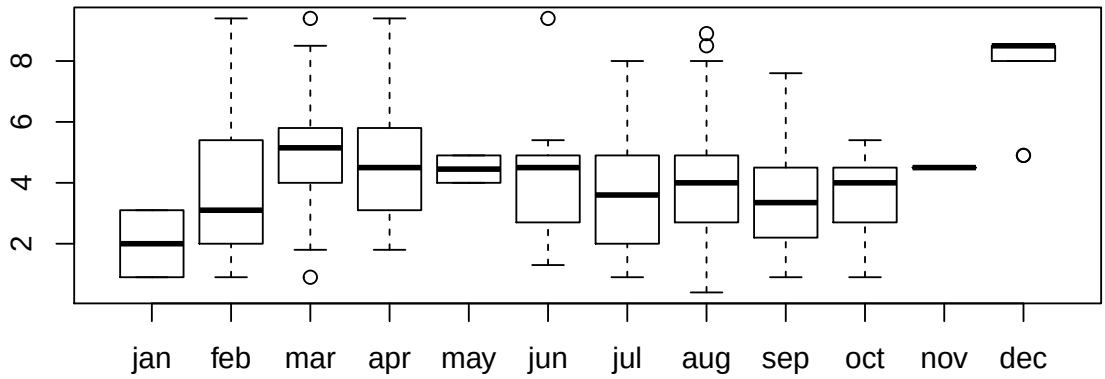



```
legend=levels(forest_fire$fire_size)
```

Observation: We don't see a specific linear relationship here.

Now let's see what is the approximate value of FFMC during August and September:

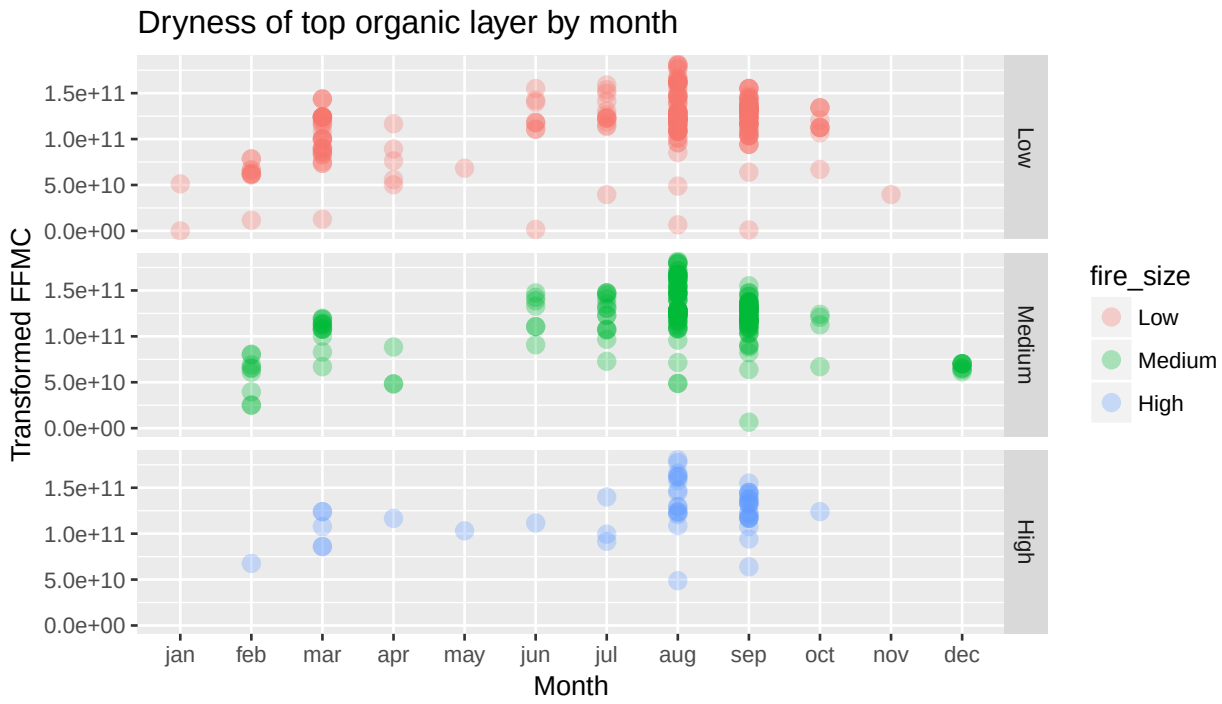
```
plot(forest_fire$month, forest_fire$wind)
```



And the plot shows that wind speed is in the range of 3 - 5 km/hr during this time, which is also a familiar condition of forest fire.

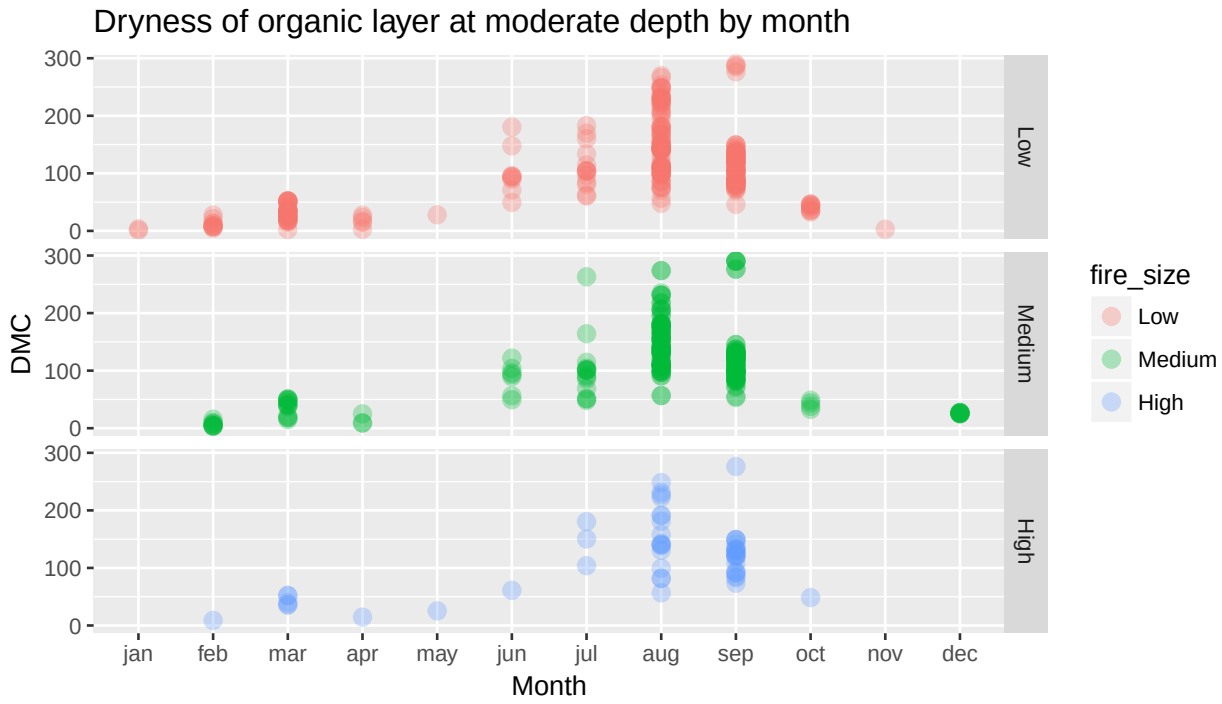
Analysis of Secondary Effects

```
ggplot(forest_fire,
       aes(x=month, y=FFMC^8/factorial(8))) + geom_point(size=3,
       aes(color = fire_size), alpha=0.3) + facet_grid (fire_size~.) + labs(x="Month",
       y="Transformed FFMC",
       title = "Dryness of top organic layer by month")
```

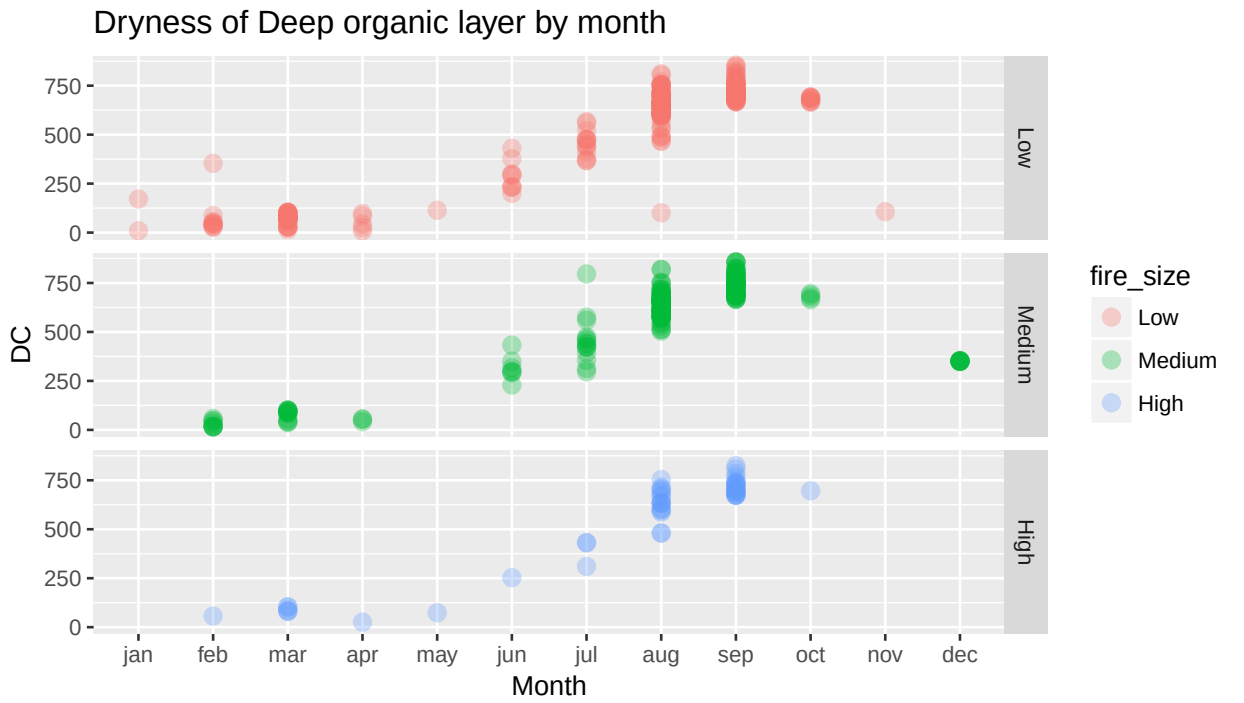


Top layer is dry for most months where substantial data is available. Summer months July, Aug and Sep are definitely dry. Not enough data in winter months to make a determination.

```
ggplot(forest_fire,
       aes(x=month, y=DMC))  + geom_point(size=3,
       aes(color = fire_size), alpha=0.3) + facet_grid (fire_size~.)+labs(x="Month",
       y="DMC", title = "Dryness of organic layer at moderate depth by month")
```

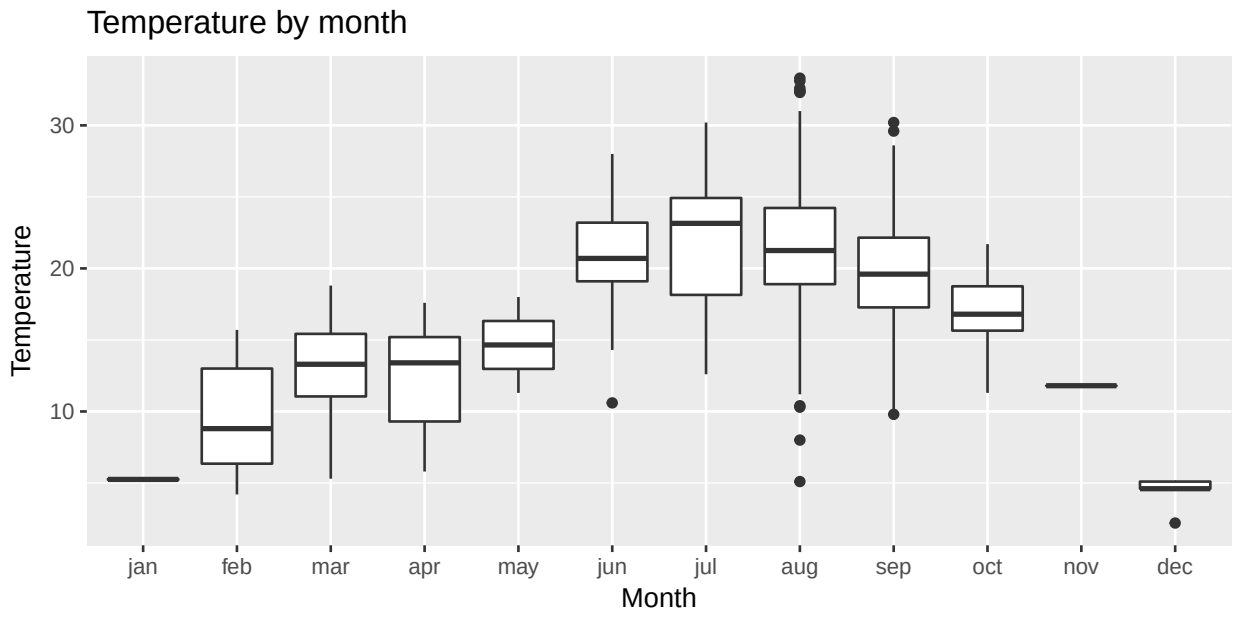


```
ggplot(forest_fire,
       aes(x=month, y=DC))  + geom_point(size=3,
       aes(color = fire_size), alpha=0.3) + facet_grid (fire_size~.) + labs(x="Month",
       y="DC", title = "Dryness of Deep organic layer by month")
```

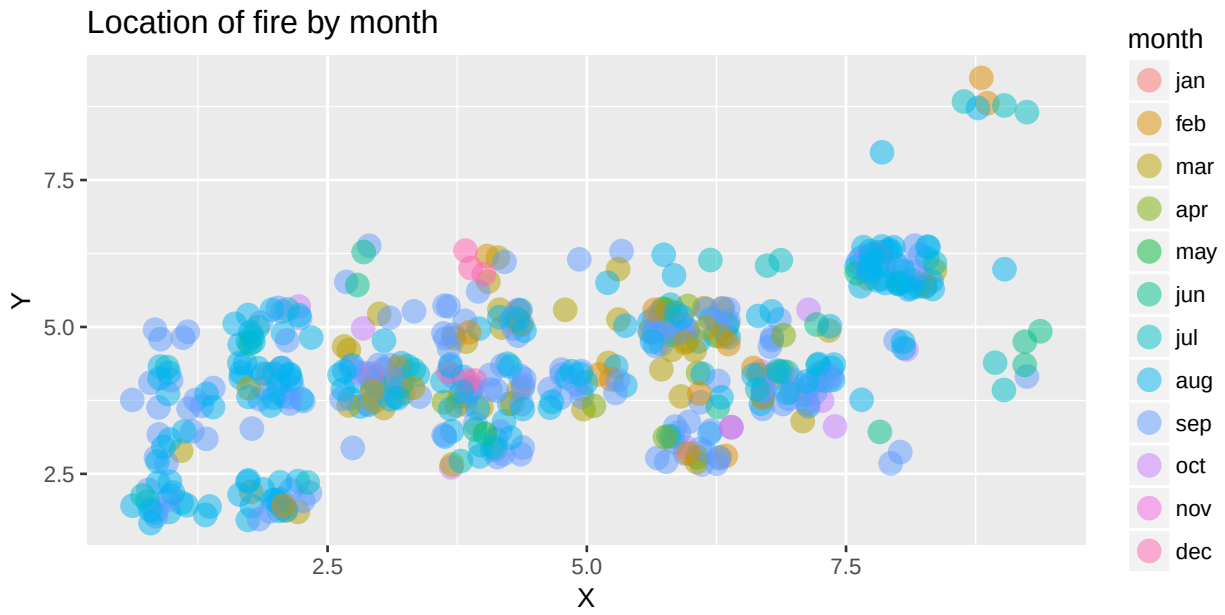


Data shows value of DC rising as summer approaches and peaks in the month of Aug and September. Flammability of deep organic layer is considered a reason for sustained forest fires.

```
ggplot(forest_fire,
       aes(x=month, y=temp))  + geom_boxplot() + labs(x="Month",
       y="Temperature", title = "Temperature by month")
```

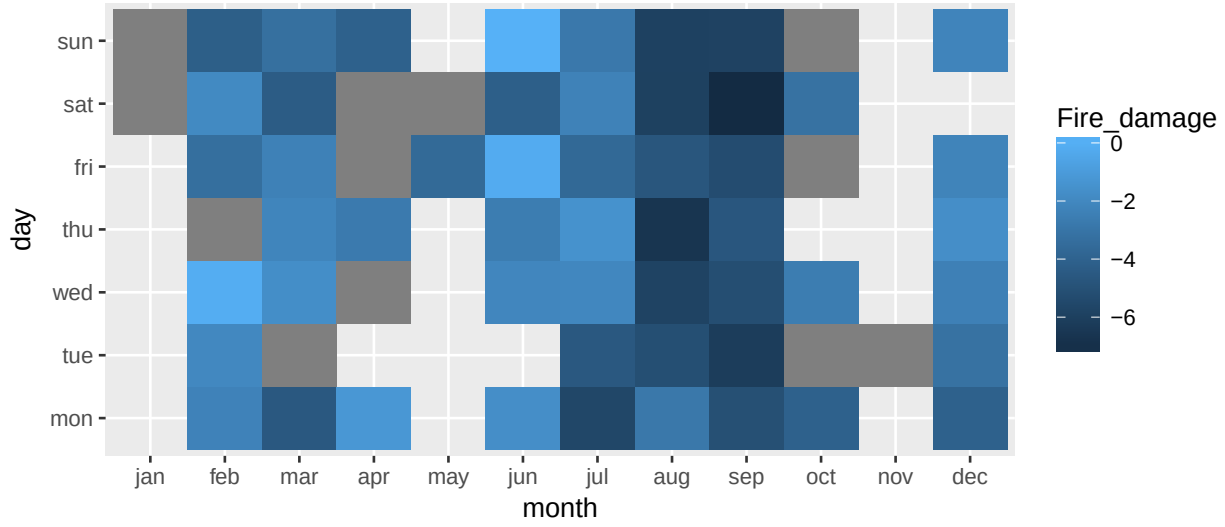


```
ggplot(forest_fire, aes(x=jitter(X, factor = 2),
       y=jitter(Y, factor = 2))) + geom_point(aes(color=month ),
       size = 4, alpha = 0.5) + labs(x="X",  y="Y", title = "Location of fire by month")
```



No fires seen in top left and bottom right of the park. Suspect this has to do with the shape of the park itself. Plot is predominantly (greenish) blue indicating most fires observed in summer months of Jul, Aug and September.

```
areadm = aggregate(forest_fire$area, by=list(forest_fire$sorted_day, forest_fire$month), FUN=sum)
areadm$x=-log(areadm$x)
colnames(areadm)[colnames(areadm)=="x"]="Fire_damage"
ggplot(areadm,
       aes(Group.2, Group.1)) + geom_tile(aes(fill = Fire_damage)) + scale_colour_gradient2()+labs(x="m
       y="day", title = "Total burnt area by month and day", subtitle = "Darker shades indicate more an
```



Conclusion

An analysis of the provided data points does not yield a strong evidence of any pattern of the relationship for the entire year. From the analysis above, we see that the fire season runs during August and September months with peak of activity on weekends (Friday, Saturday, Sunday). Though there seems to be a trend that shows maximum burned area during those months, we don't see a strong relationship between the key variables of interest.

Key relationship analysis for the months of August and September suggests that:
1. Fine Fuel Moisture Code (FFMC) value is greater than 80 during this time
2. Duff Moisture Code (DMC), Drought Code (DC) and high burned area is also positively correlated during the same time
3. Initial Spread Index (ISI) belongs to 5-15 range during this time, which can be another cause of forest fire.
5. On plotting high burned area versus temperature we see maximum burned area between temperature range of 15 to 25. This pattern is more evident when plotting against logarithmic scale of temperature
6. Relative humidity stays in the range of 40-50 during the time of high burned area
7. Wind Speed variable stays between 3-5 km/hr that familiar cause of forest fire

Secondary effects also show that that the top layer is dry for most months when substantial data is available. Summer months July, August and September are definitely dry. There is not enough data for winter months to see a definite pattern. Data shows value of DC rising as summer approaches and peaks in the month of Aug and September. Flammability of deep organic layer is considered a reason for sustained forest fires.

The provided sample size is not strong enough to generalize the relationship for a larger burned area as well. There could be other factors that can also cause forest fires and can influence the analysis provided they are available. An example are factors such as lightning, human factors like igniting of vegetation.