Analyzing Forest Fire Data

Malik Williams

July 16, 2019

Analyzing Forest Fire Data

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.4.4
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

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

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

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

```
library(purrr)
```

```
## Warning: package 'purrr' was built under R version 3.4.4
```

```
library(ggplot2)
```

Forest fires are a danger to animals' and humans' lives. They happen one of two ways: naturally caused or human caused. Understanding the places that they start and what reasons cause them will allow us to take preventative measures to ensure our safety.

In this project, we will perform visual analysis on a data for forest fires.

About the Data

Below is a table detailing what data is in each column.

Column Name	Description
`X`	X-axis spatial coordinate within the Montesinho park map: 1 to 9
`Y`	Y-axis spatial coordinate within the Montesinho park map: 2 to 9
`month`	Month of the year: 'jan' to 'dec'
`day`	Day of the week: 'mon' to 'sun'
`FFMC`	Fine Fuel Moisture Code index from the FWI system: 18.7 to 96.20
`DMC`	Duff Moisture Code index from the FWI system: 1.1 to 291.3
`DC`	Drought Code index from the FWI system: 7.9 to 860.6
`ISI`	Initial Spread Index from the FWI system: 0.0 to 56.10
`temp`	Temperature in Celsius degrees: 2.2 to 33.30
`RH`	Relative humidity in percentage: 15.0 to 100
`wind`	Wind speed in km/h: 0.40 to 9.40

Column Name	Description
`rain`	Outside rain in mm/m2 : 0.0 to 6.4
`area`	The burned area of the forest (in ha): 0.00 to 1090.84

***Note: FWI stands for "Fire Weather Index", which is the method used by scientists to determine the risk factors involved in a forest fires. More information about the FWI system and variables in the dataset can be found here.**

Let's familiarize ourselves with the data by printing out the first 5 observations.

```r
fires <- read_csv("forestfires.csv")
```

```
## Parsed with column specification:
## cols(
##   X = col_integer(),
##   Y = col_integer(),
##   month = col_character(),
##   day = col_character(),
##   FFMC = col_double(),
##   DMC = col_double(),
##   DC = col_double(),
##   ISI = col_double(),
##   temp = col_double(),
##   RH = col_integer(),
##   wind = col_double(),
##   rain = col_double(),
##   area = col_double()
## )
```

```r
head(fires, 5)
```

```
## # A tibble: 5 x 13
##       X     Y month day    FFMC   DMC    DC   ISI  temp    RH  wind  rain
##   <int> <int> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <dbl>
## 1     7     5 mar   fri    86.2  26.2  94.3   5.1   8.2    51   6.7   0
## 2     7     4 oct   tue    90.6  35.4 669.    6.7  18      33   0.9   0
## 3     7     4 oct   sat    90.6  43.7 687.    6.7  14.6    33   1.3   0
## 4     8     6 mar   fri    91.7  33.3  77.5   9     8.3    97   4     0.2
## 5     8     6 mar   sun    89.3  51.3 102.    9.6  11.4    99   1.8   0
## # ... with 1 more variable: area <dbl>
```

Forest Fire Trends

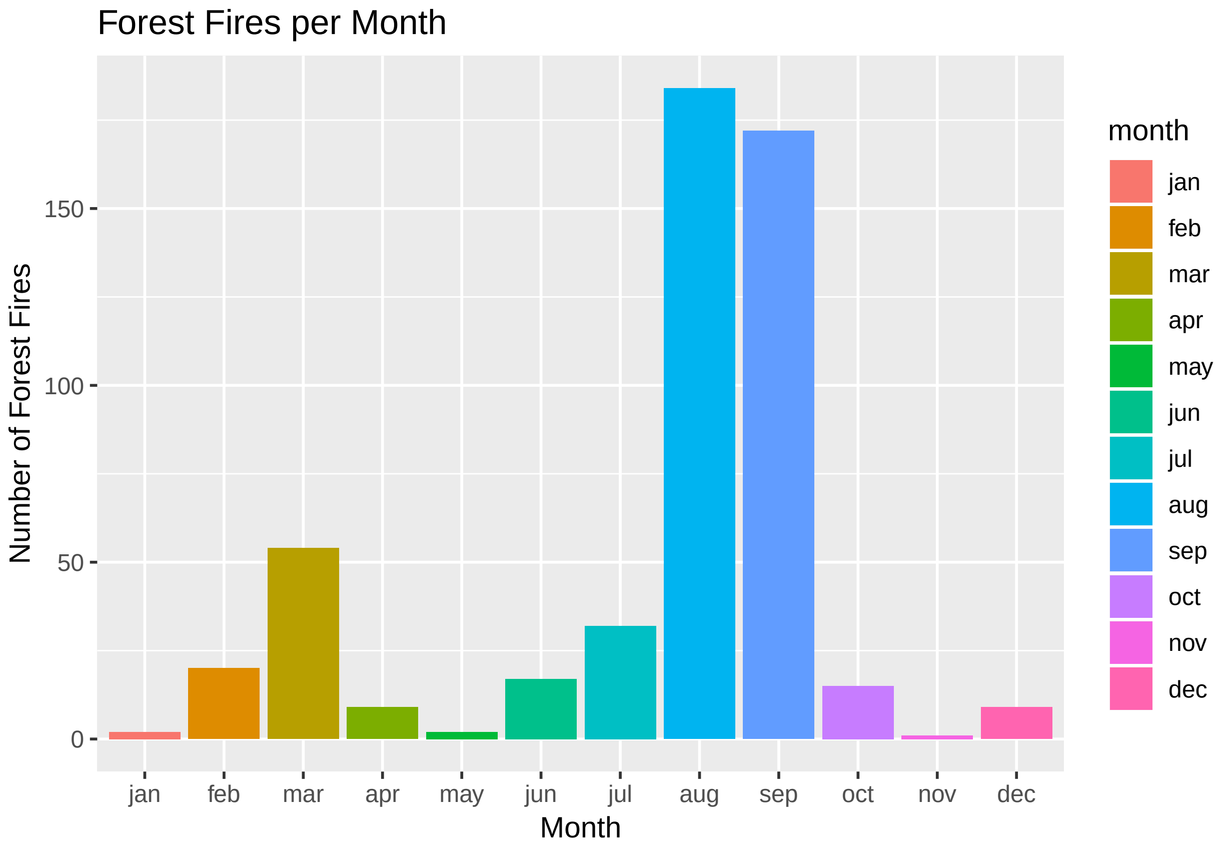
When discussing forest fires and what preventative measures can be taken to stop them, it is beneficial to know *when* they occur. Luckily, we have our `month` and `day` variables that can help. The two questions we can answer are as follows:

- Which months are forest fires the most common?
- Which days of the week are forest fires most likely to occur?

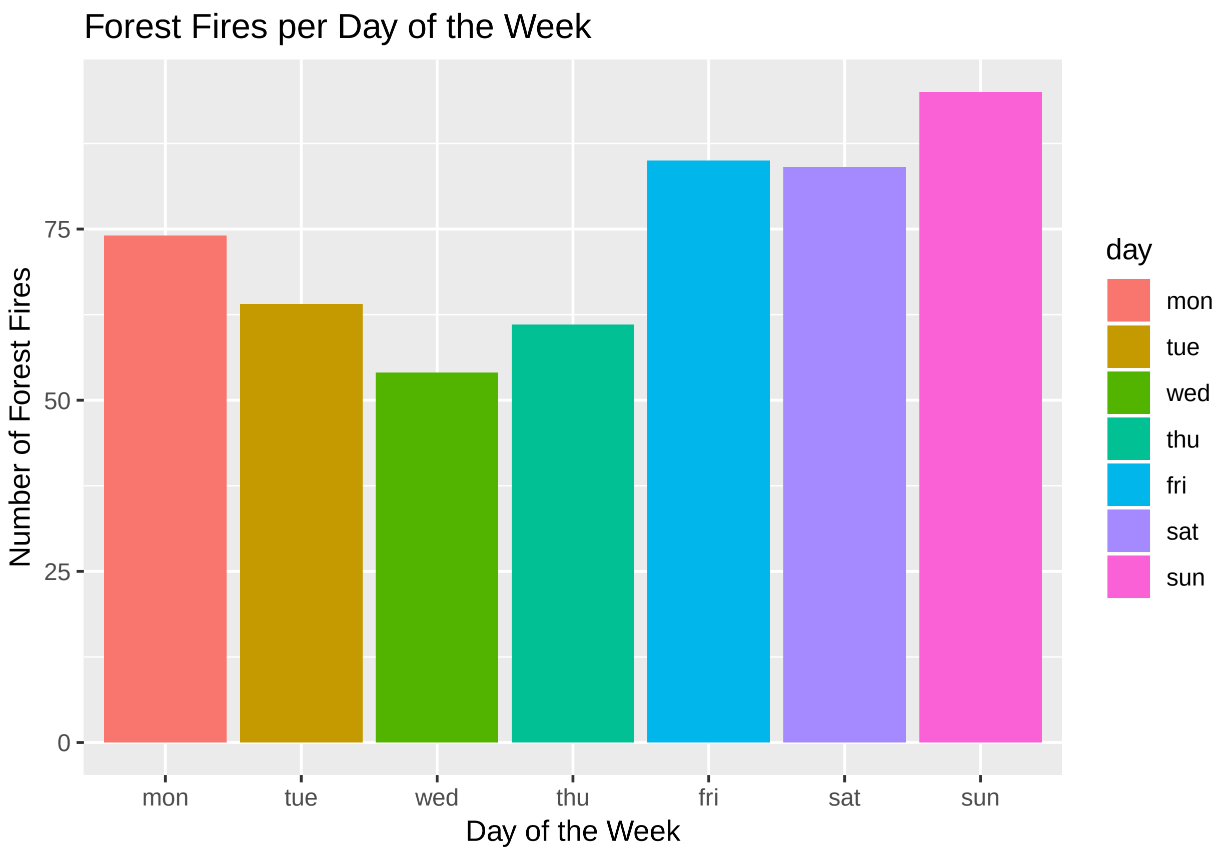
To answer this question, we will create two bar graphs: 1. One showing the number of forest fires for each month 2. One that shows the number of forest fires for each day of the week

First, we create two separate groupings (one by month, and the other by day), and we create a variable called "occurences" that represents the count of forest fires per each month/day.

```r
fires <- fires %>% mutate(month = factor(month, levels = c("jan", "feb", "mar", "apr", "may", "jun", "ju

fires_month <- fires %>% group_by(month) %>% summarize(occurences = n())
fires_day <- fires %>% group_by(day) %>% summarize(occurences = n())

ggplot(data = fires_month) +
  aes(x = month, y = occurences, fill = month) +
  geom_bar(stat = 'identity') +
  labs(y = 'Number of Forest Fires', x = 'Month', title = "Forest Fires per Month")
```



```r
ggplot(data = fires_day) +
  aes(x = day, y = occurences, fill = day) +
  geom_bar(stat = 'identity') +
  labs(y = 'Number of Forest Fires', x = 'Day of the Week', title = "Forest Fires per Day of the Week")
```



From the graphs, we see that forest fires are a lot more common in August and September, which are hotter months in the Northern Hemisphere. For forest fires per day of the week, the middle of the week seems to be when the smallest number of fires occur with Wednesday having the least number of fires. While the weekend tends to have the most fires with sunday having the most, followed by friday and saturday.

There could be a number of explanations for what we found. For instance, people are more likely to barbecue on weekends in forest preserves, which could explain why there are a higher number of fires on the weekend.

Analyzing Potential Causes of Forest Fires

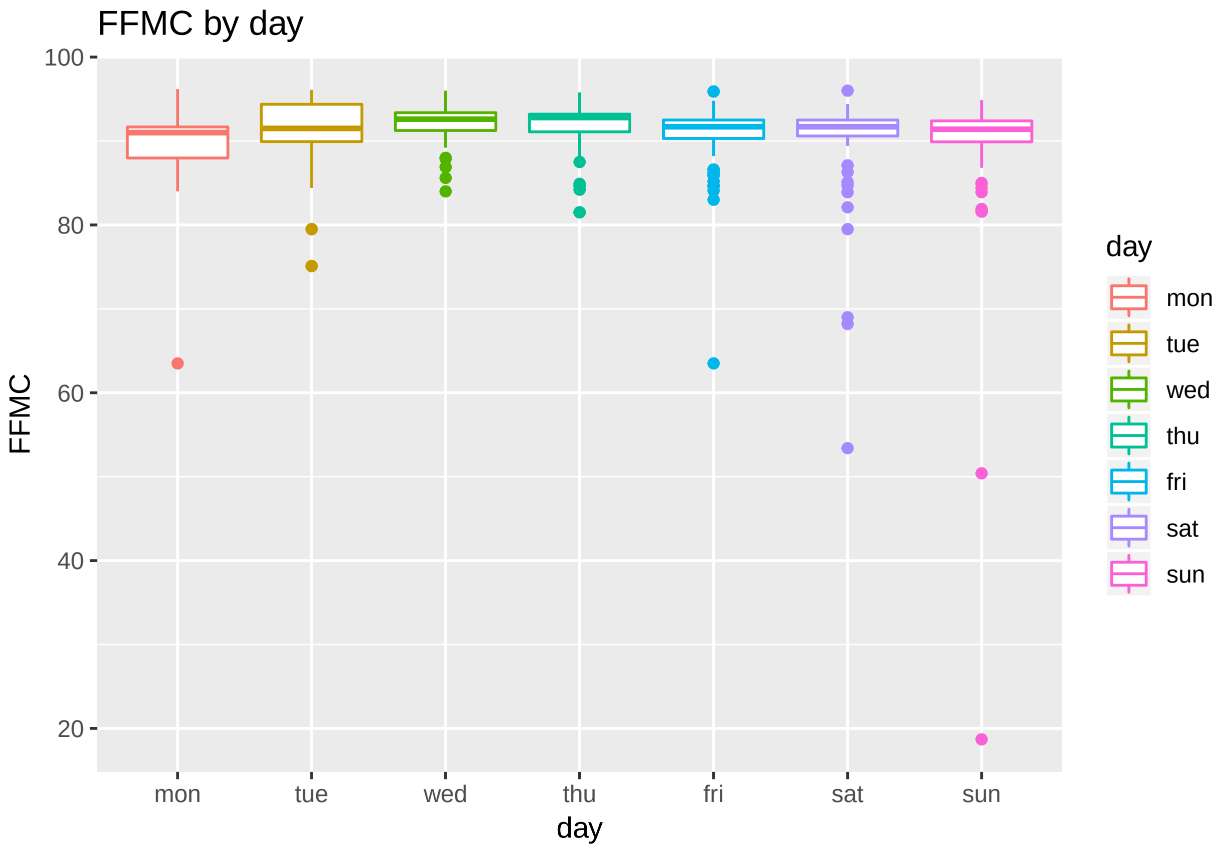
To explore potential causes of the temporal patterns of forest fire occurrence shown by the bar charts, we can take a closer look at how the variables that relate to forest fires vary by month and by day of the week. To do this, we will construct box plots to visualize the distribution of the following variables by month and by day of the year: * FFMC * DMC * DC * ISI * temp * RH * wind * rain

We create two functions, `create_month_boxplot` and `create_day_boxplot` to create a Month/Day vs Desired Variable plot for each variable.

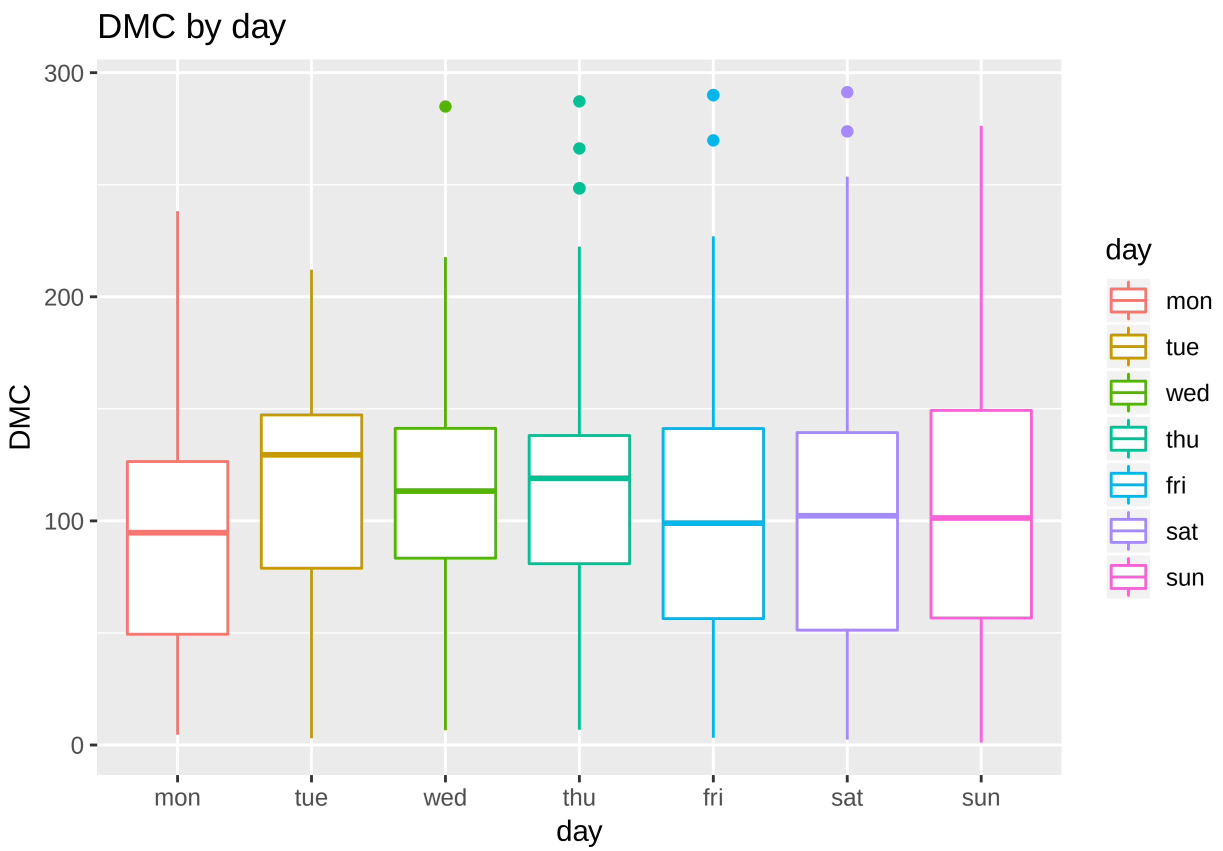
```r
# Printing all variables by days of the week
create_day_boxplot = function(x, y){
  ggplot(data = fires) +
    aes_string(x = x, y = y, color = x) +
    geom_boxplot() +
    labs(x = "day", y = y, title = paste(y, "by day"))
}

x_var2 <- names(fires[4])
```

```
y_var2 <- names(fires[5:12])
fire_stats_day <- map2(x_var2, y_var2, create_day_boxplot)
fire_stats_day
```

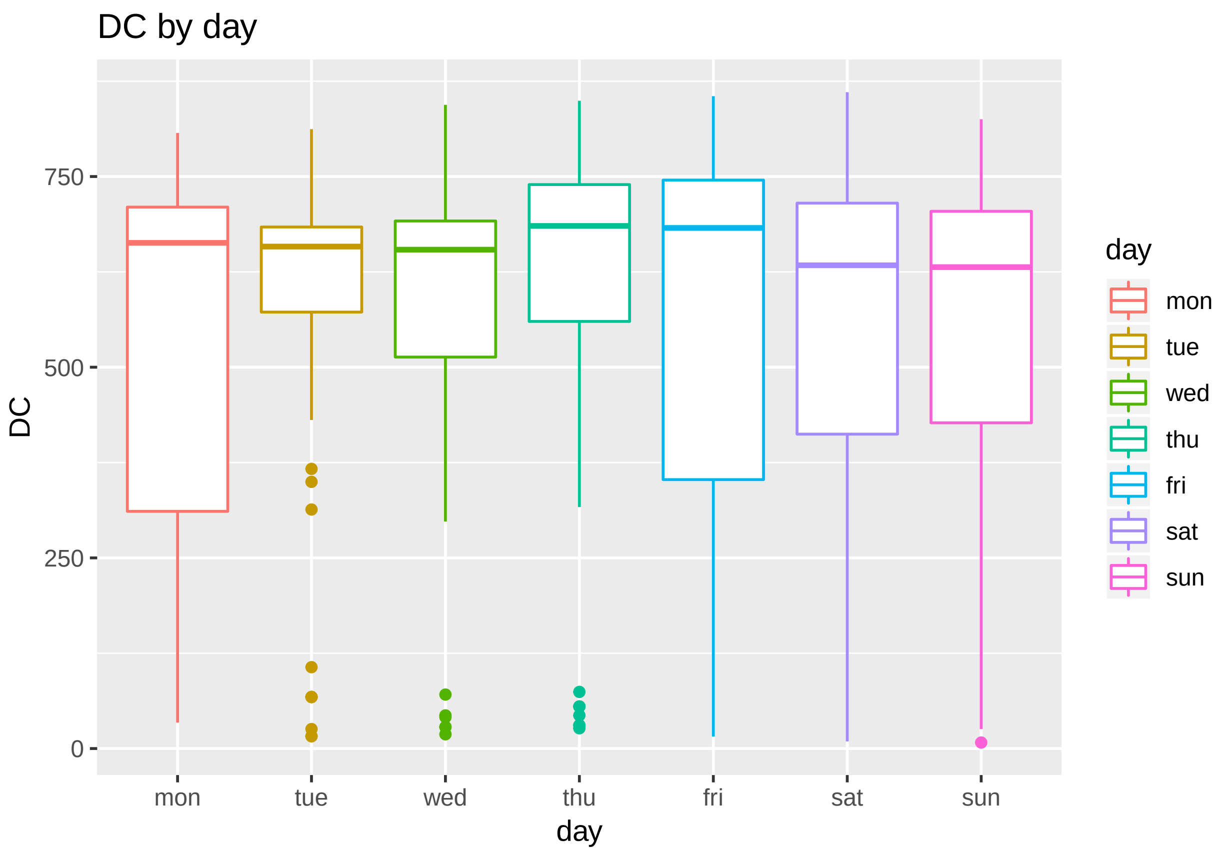
```
## [[1]]
```



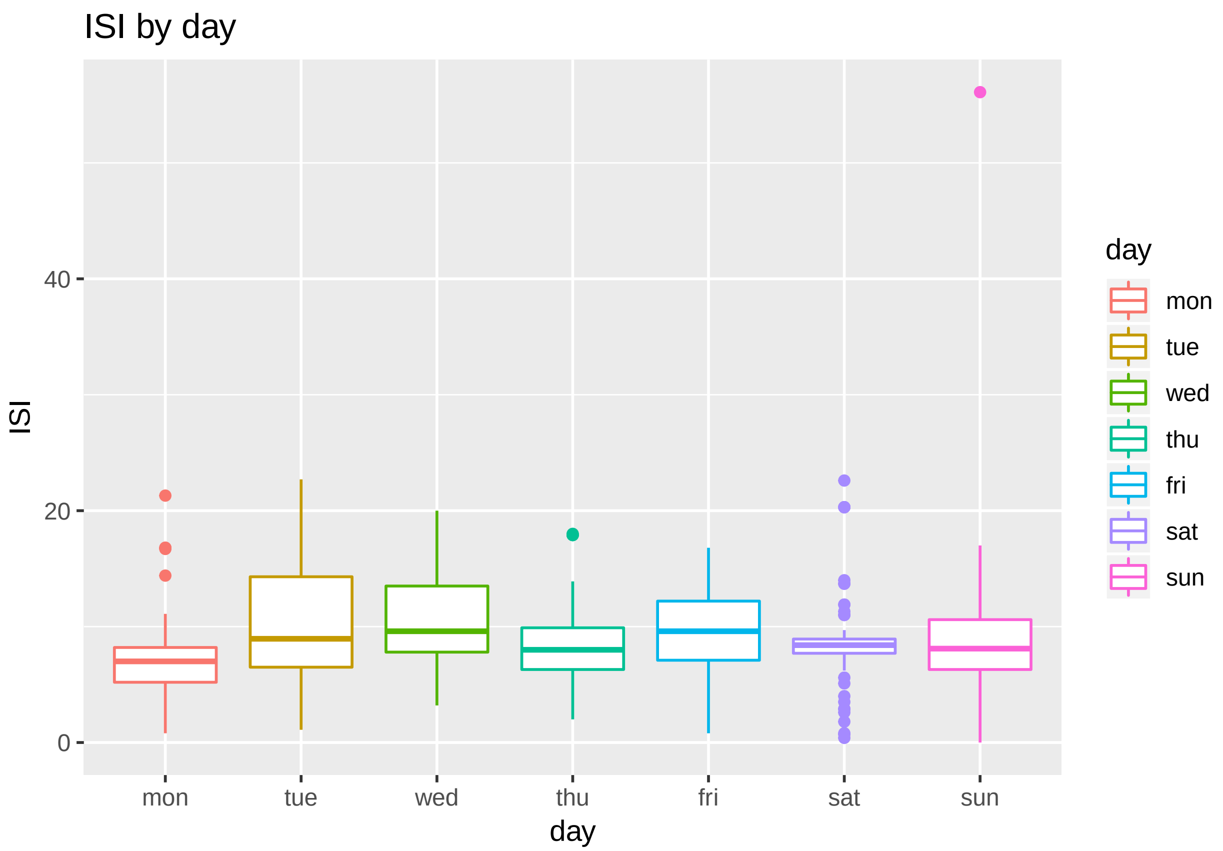
```
##
## [[2]]
```


DMC by day
DMC
300
200
100
0
mon
tue
wed
thu
fri
sat
sun
day
day
mon
tue
wed
thu
fri
sat
sun

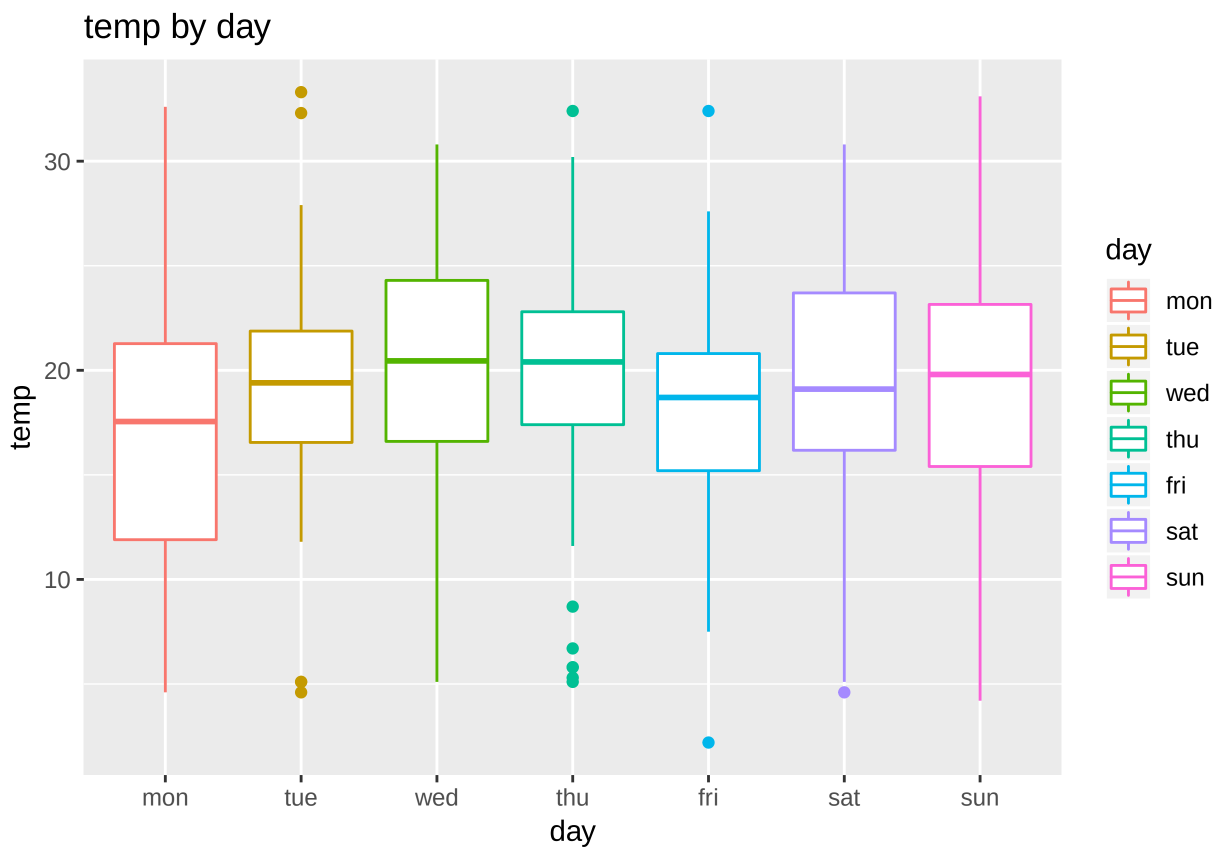
```
##
## [[3]]
```


DC by day
DC
750
500
250
0
mon
tue
wed
thu
fri
sat
sun
day
day
mon
tue
wed
thu
fri
sat
sun

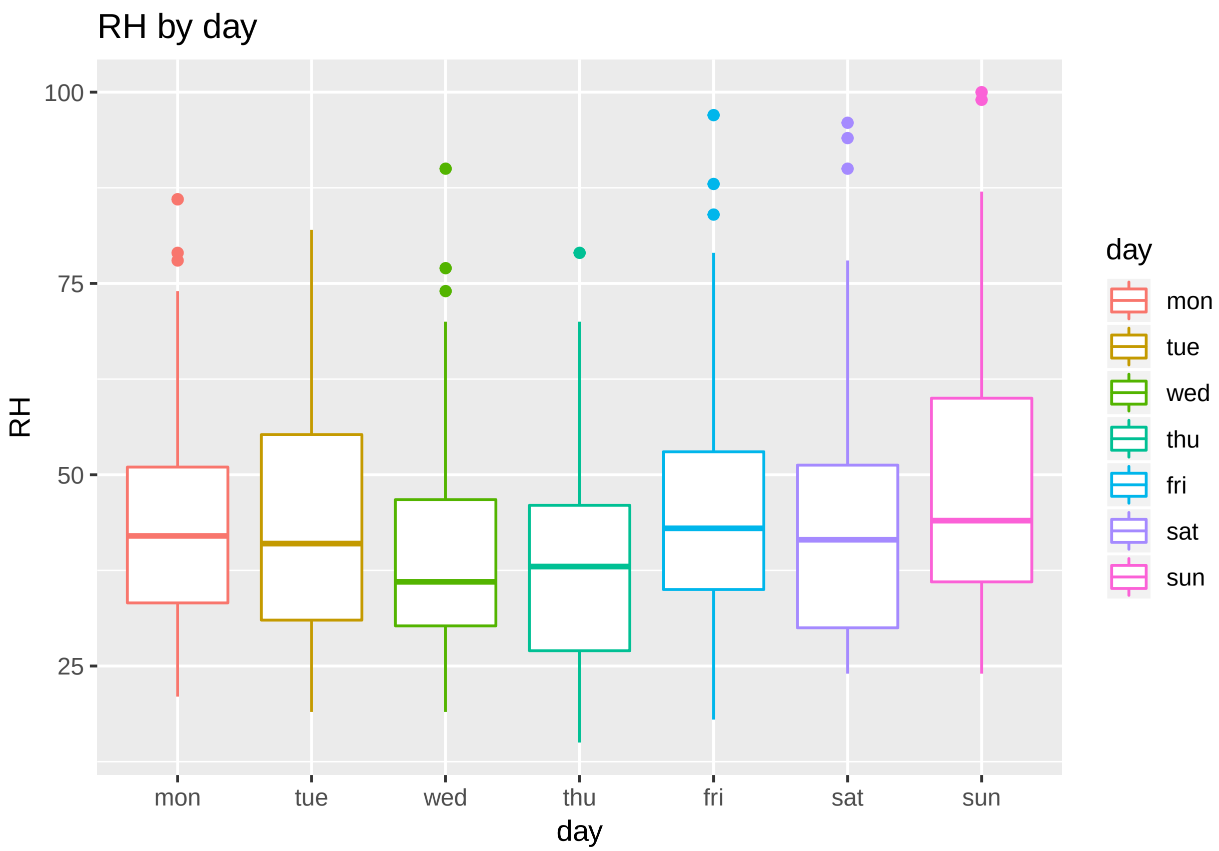
```
##
## [[4]]
```


ISI by day
ISI
40
20
0
mon
tue
wed
thu
fri
sat
sun
day
day
mon
tue
wed
thu
fri
sat
sun

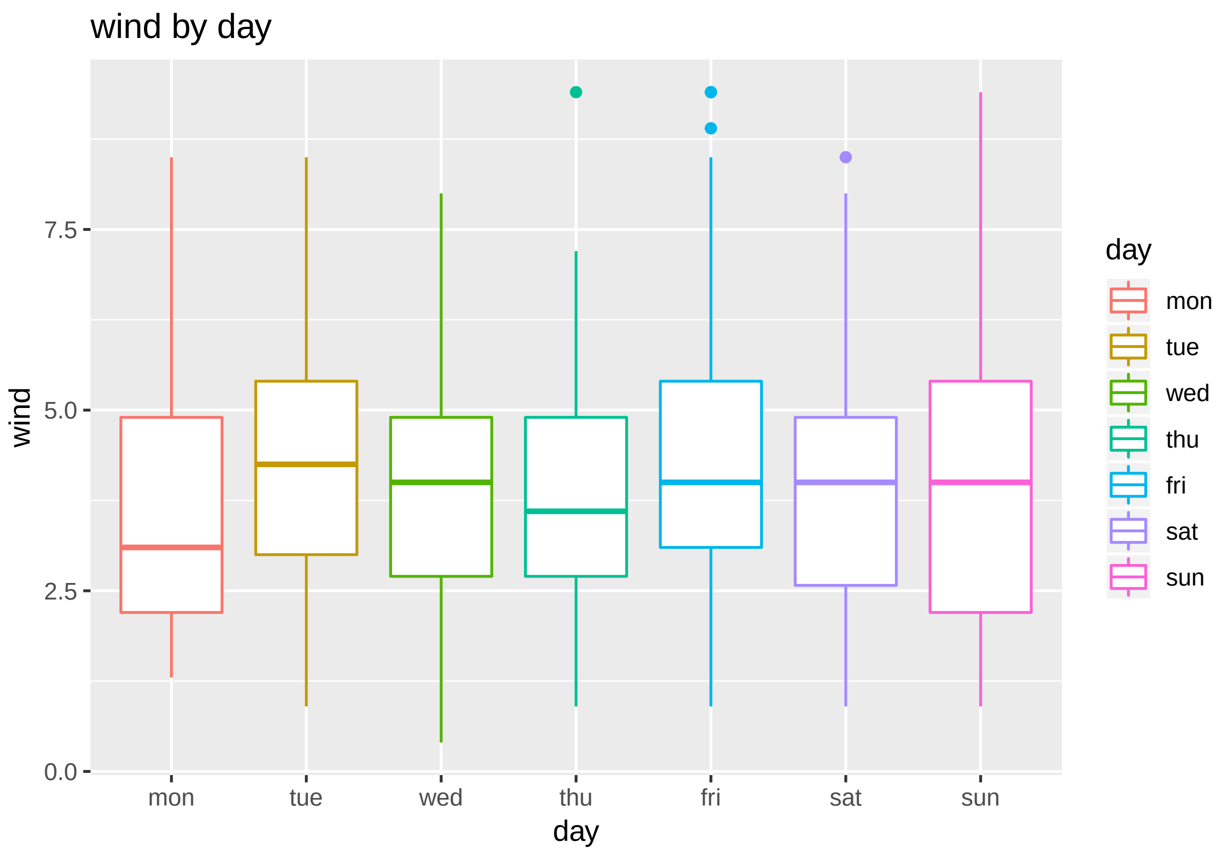
```
##
## [[5]]
```


temp by day
temp
30
20
10
mon
tue
wed
thu
fri
sat
sun
day
day
mon
tue
wed
thu
fri
sat
sun

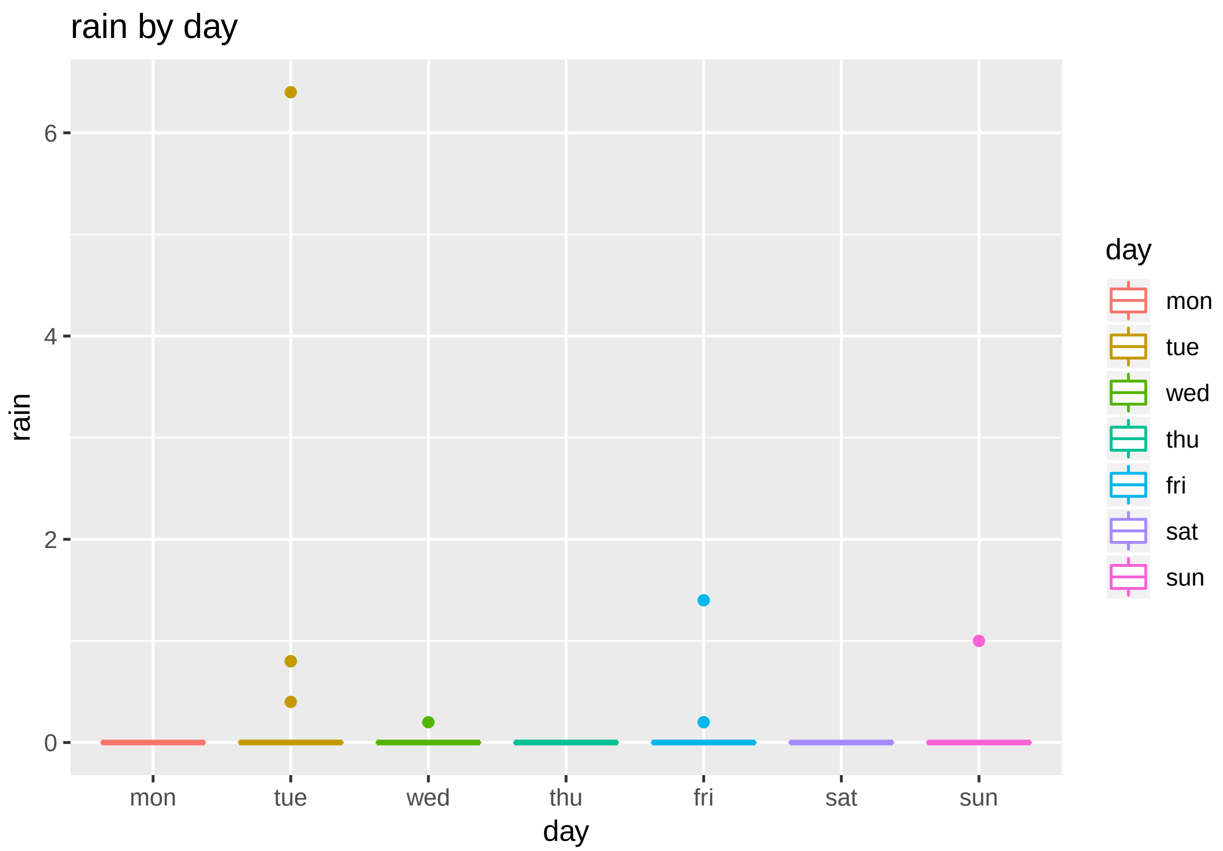
```
##
## [[6]]
```


RH by day
100
75
50
25
RH
mon
tue
wed
thu
fri
sat
sun
day
day
mon
tue
wed
thu
fri
sat
sun

```
##
## [[7]]
```


wind by day
wind
7.5
5.0
2.5
0.0
mon
tue
wed
thu
fri
sat
sun
day
day
mon
tue
wed
thu
fri
sat
sun

```
##
## [[8]]
```



Let's look at each variable by each day of the week to see if anything stands out.

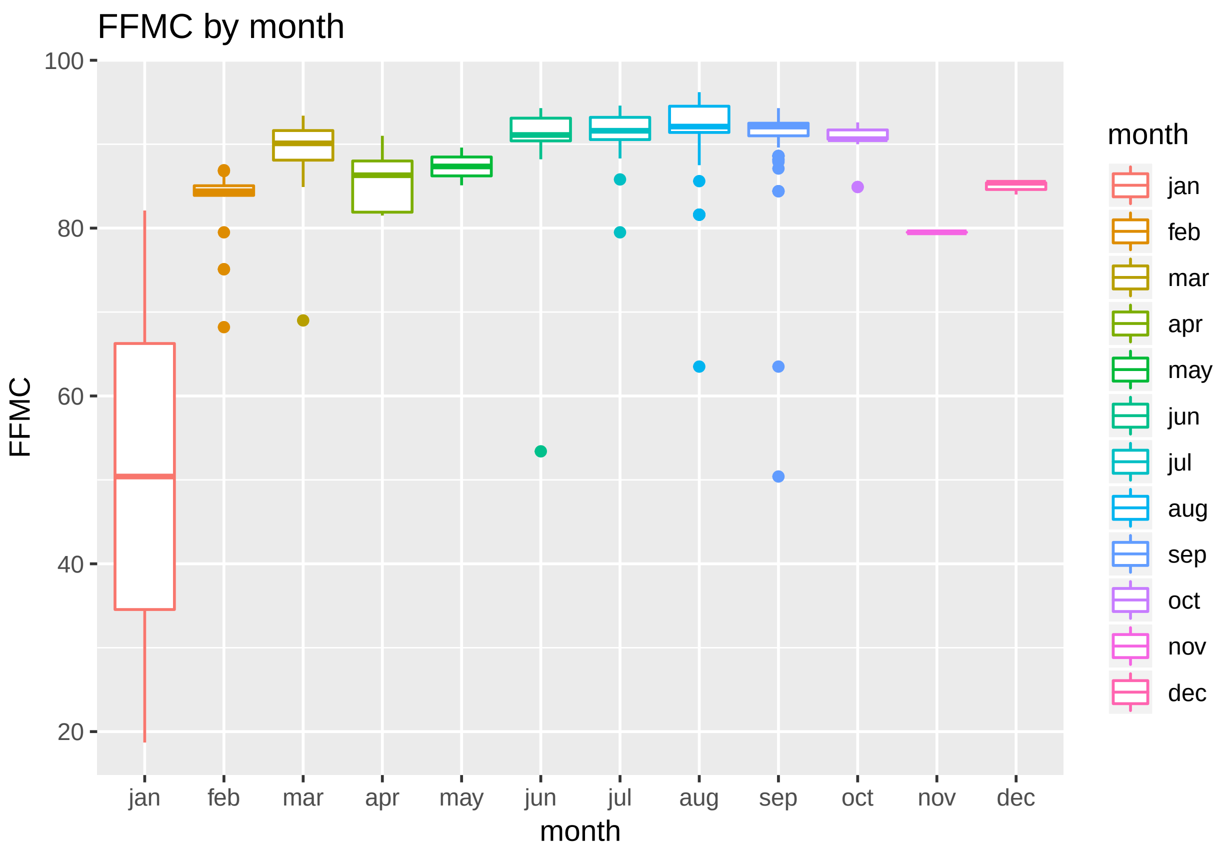
First, We see that the medians for each variable are pretty consistent across the days of the week. The lengths of the boxes are also similar, meaning that the ranges of data are consistent across days of the week.

The number of outliers and length of boxplot whiskers vary from day to day, but there does not appear to be a discernable pattern. Even though, we know that the number of forest fires is higher over the weekends, it is difficult to pinpoint a probable cause to explain this.

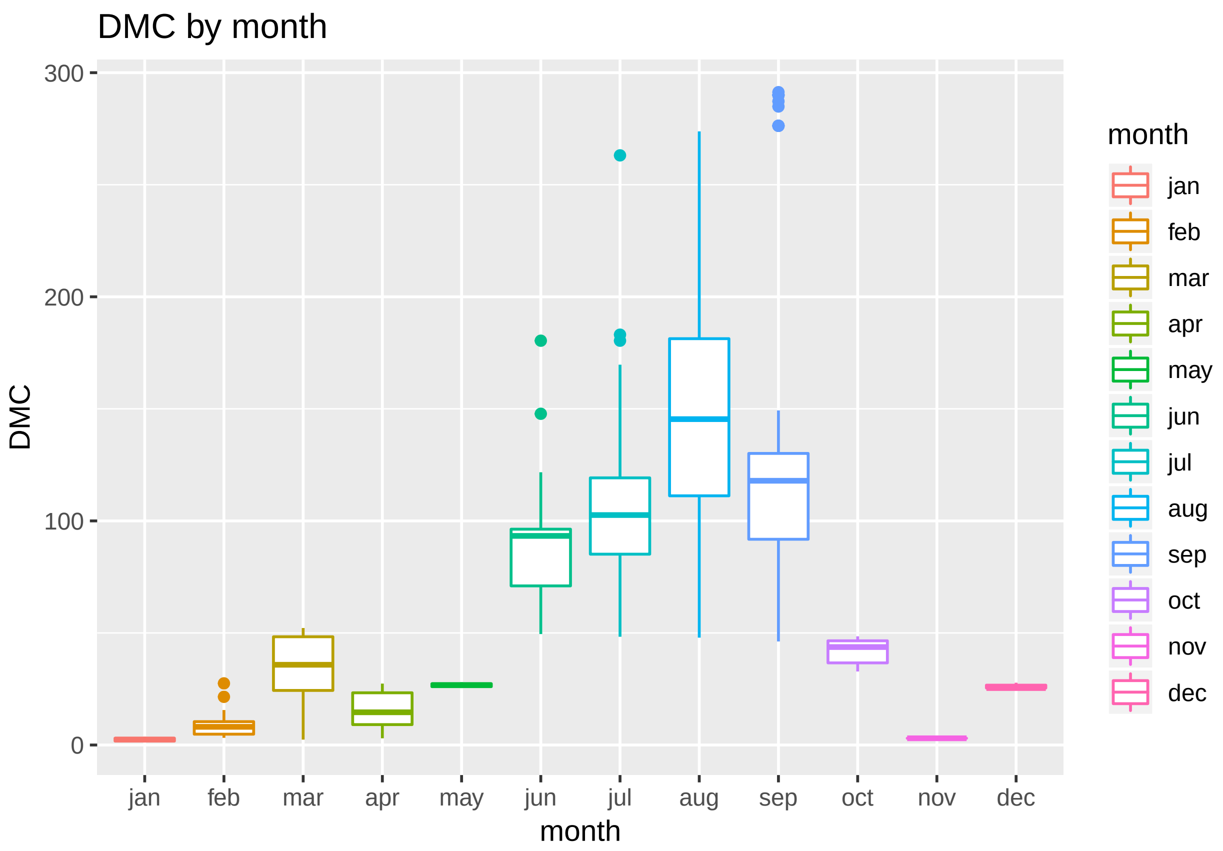
Interestingly, the `rain` by `day` plot's boxes are all flat lines at 0 with no box whiskers and a few outliers. This means that it is rare that forest fires occur when it is raining. This makes sense because rain puts out fires.

```
# Printing all variables by month
 create_month_boxplot = function(x, y){
  ggplot(data = fires) +
    aes_string(x = x, y = y, color = x) +
    geom_boxplot() +
    labs(x = "month", y = y, title = paste(y, "by month"))
}

x_var <- names(fires[3])
y_var <- names(fires[5:12])
fire_stats_month <- map2(x_var, y_var, create_month_boxplot)
fire_stats_month
```

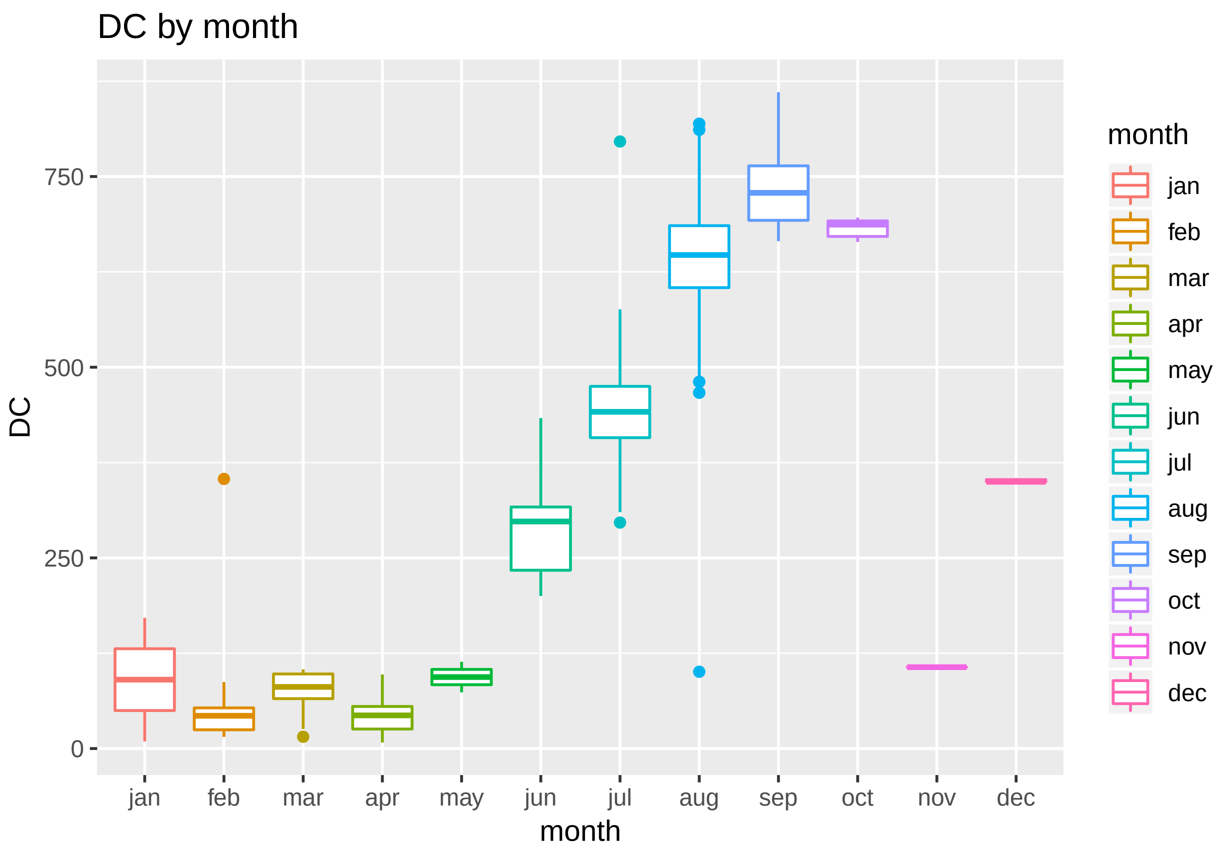
```
## [[1]]
```


FFMC by month
FFMC
100
80
60
40
20
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec

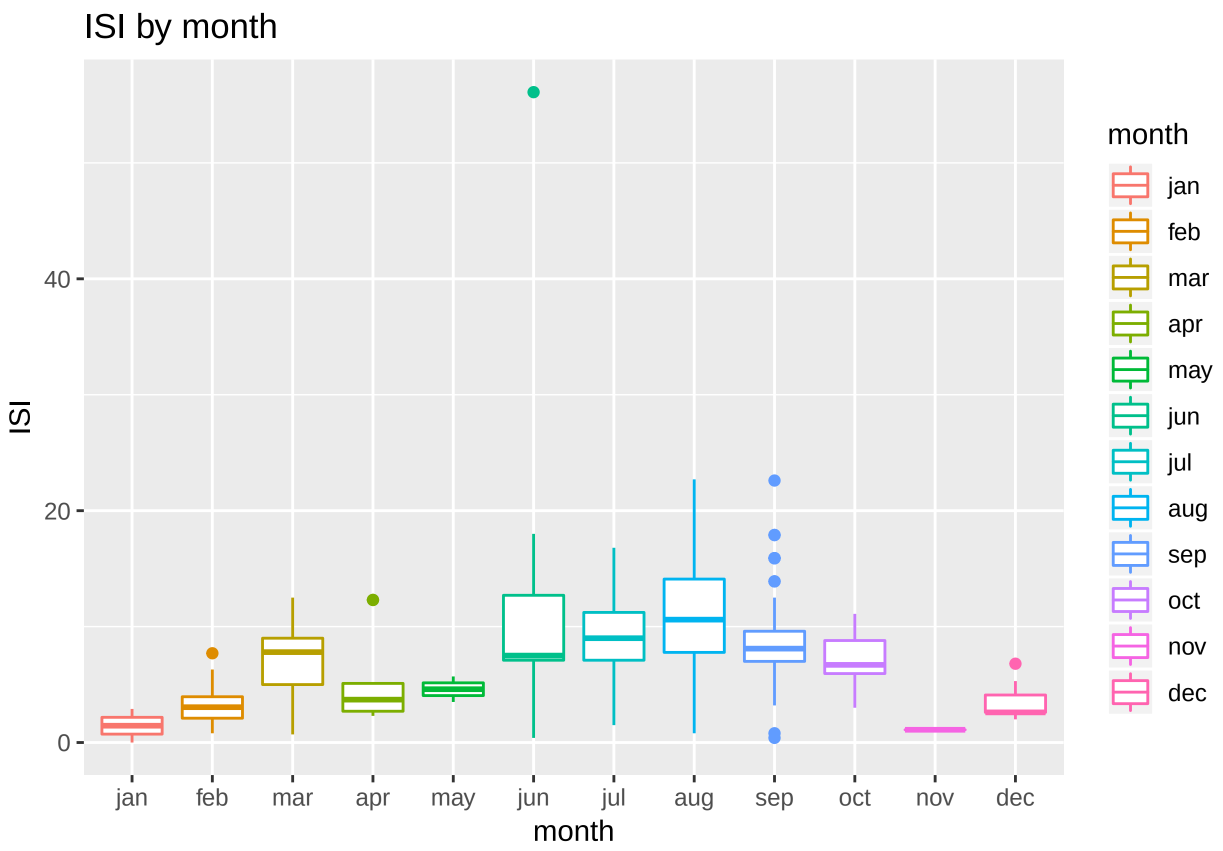
```
##
## [[2]]
```


DMC by month
DMC
300
200
100
0
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec

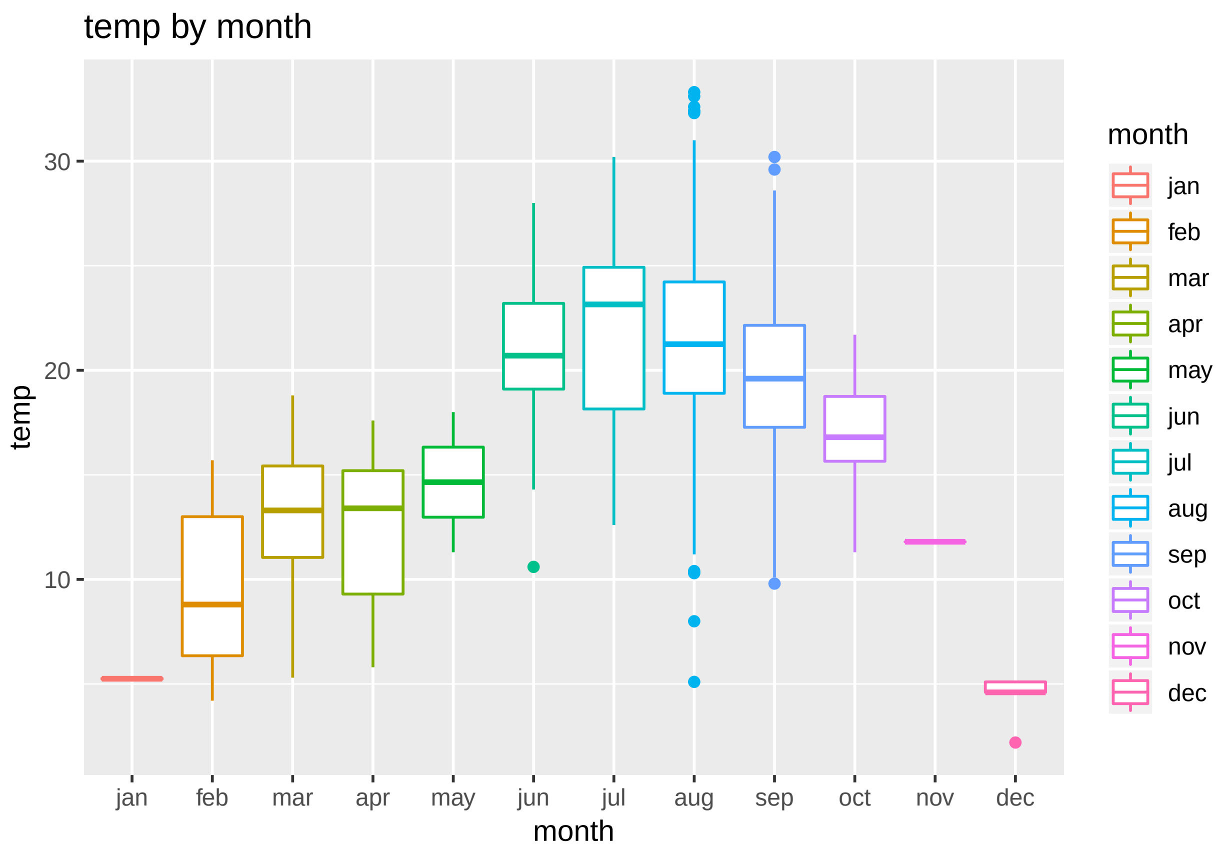
```
##
## [[3]]
```


DC by month
DC
0
250
500
750
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec

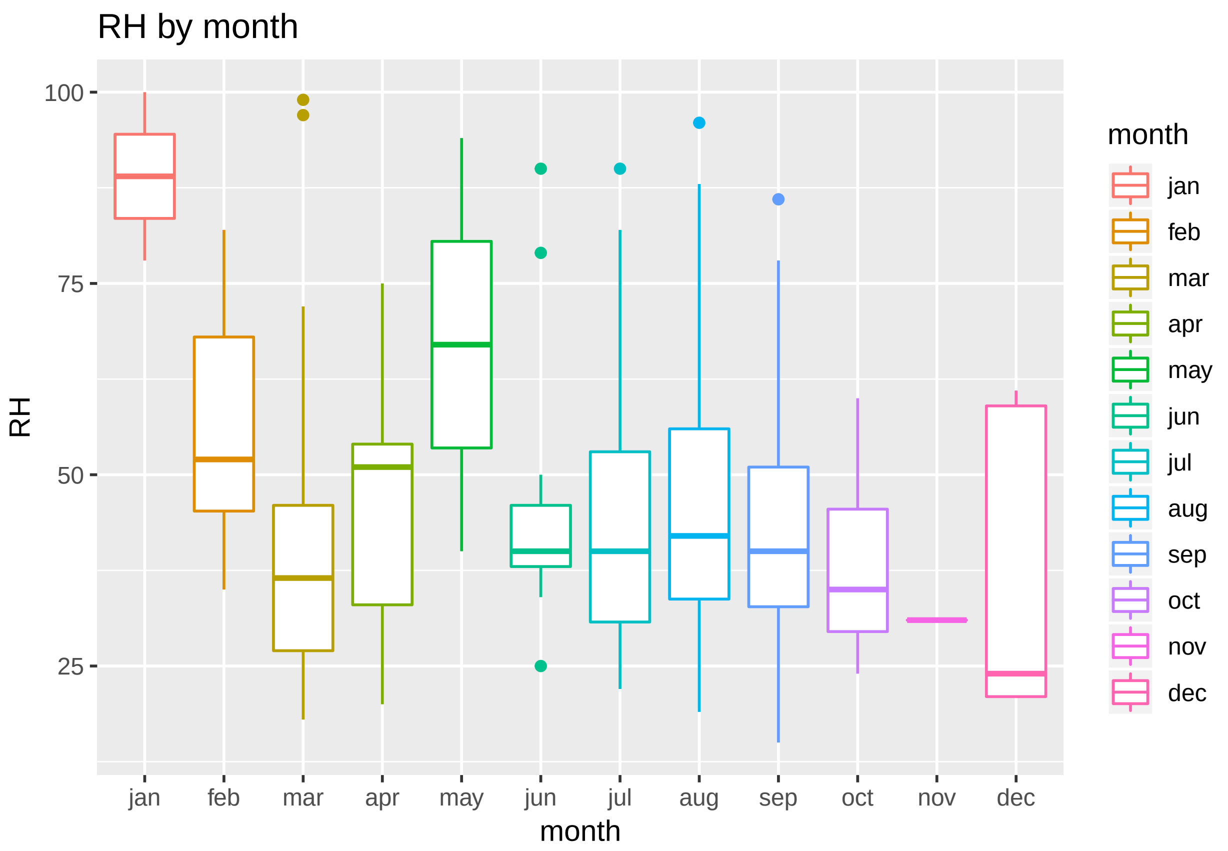
```
##
## [[4]]
```


ISI by month
ISI
0
20
40
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec

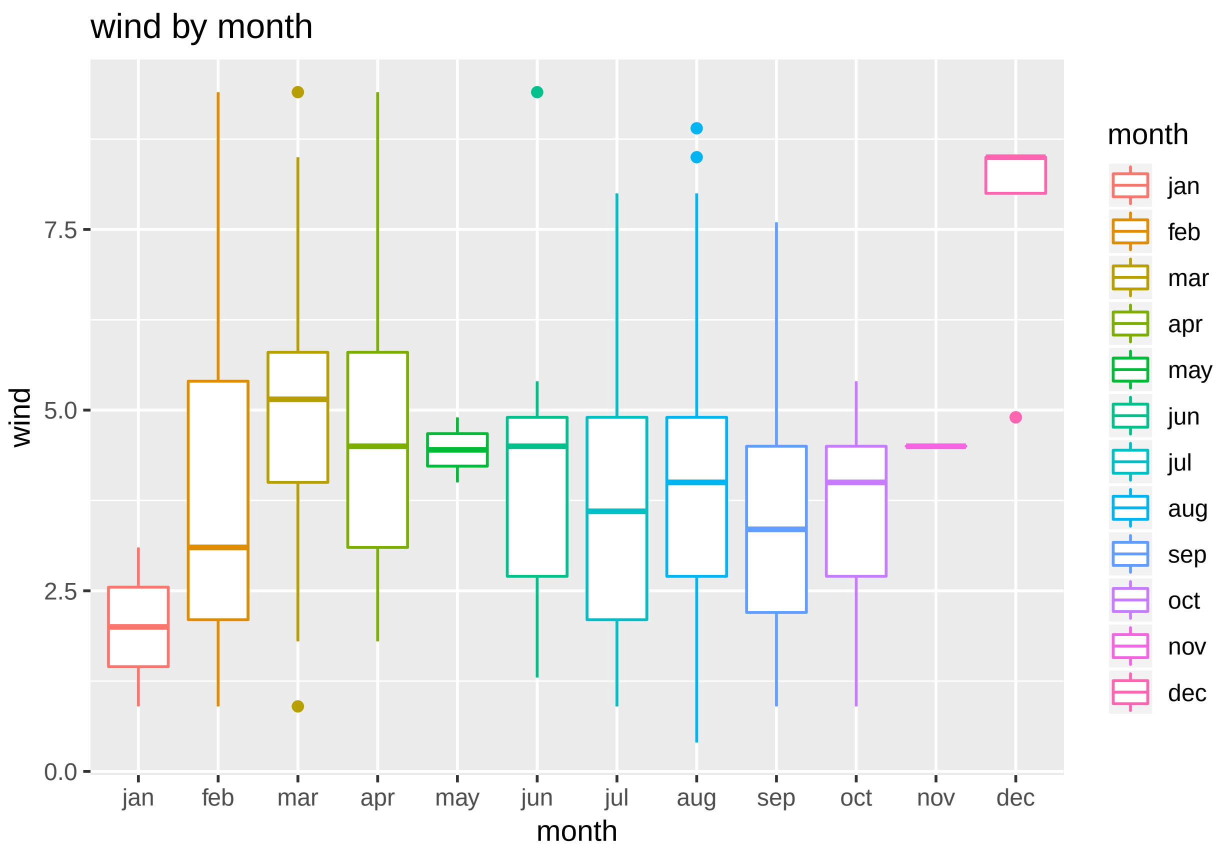
```
##
## [[5]]
```


temp by month
temp
30
20
10
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec

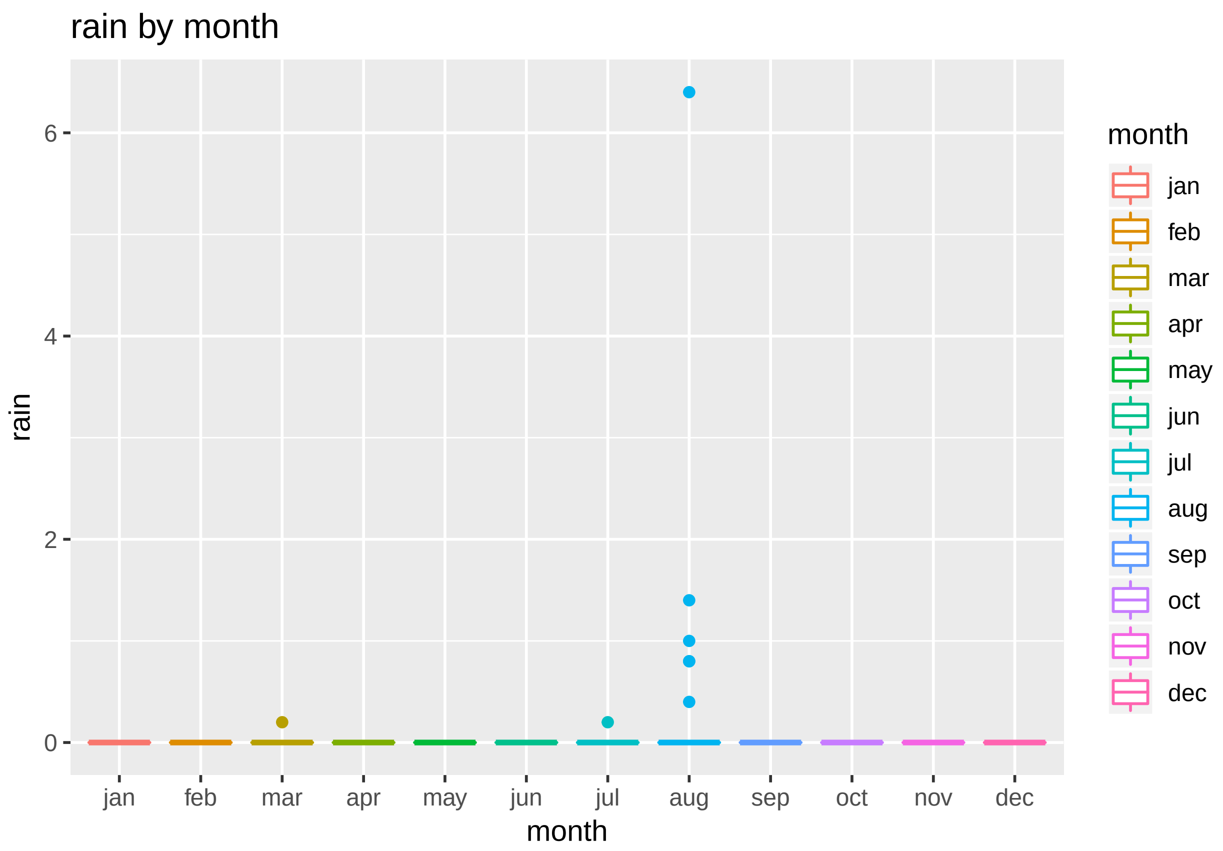
```
##
## [[6]]
```


RH by month
100
75
50
25
RH
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec

```
##
## [[7]]
```


wind by month
wind
7.5
5.0
2.5
0.0
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month
month
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec

```
##
## [[8]]
```



When looking at the variable distributions by month, almost all the variables display clear differences among months (nnlike the variable distributions by days of the week).

The `temp` variable is noticeably higher in the summer months (jun-sep), which means these months are hotter than other months. The drought code (`DC`), a measure that indicates how dry it is, is much higher in August, September, and October. Having a drought code above 300 is an extreme indication that fire will involve deep sub-surface and heavy fuels.Since it is dry, there is a higher chance of starting a fire. The Duff Moisture Code (`DMC`), the average moisture content, is very high in the summer months, with median values over 100. A DMC rating of more than 30 is dry, and above 40 indicates that intensive burning will occur in the duff and medium fuels.This means that big fires are more likely to occur in summer months.

We have analyzed patterns of variables across days of the week and months. While interesting patterns cannot be found for variables by day of the week, there are clear differences between months that explain why there are more forest fires in the summer months.

Analyzing Forest Fire Severity I

Let's take this a step further. Which variables are related to the severity of forest fires?

To do this, we will use the `area` variable as a measure of severity of forest fires because it represents the number of hectares of forest that were burned by the fire.

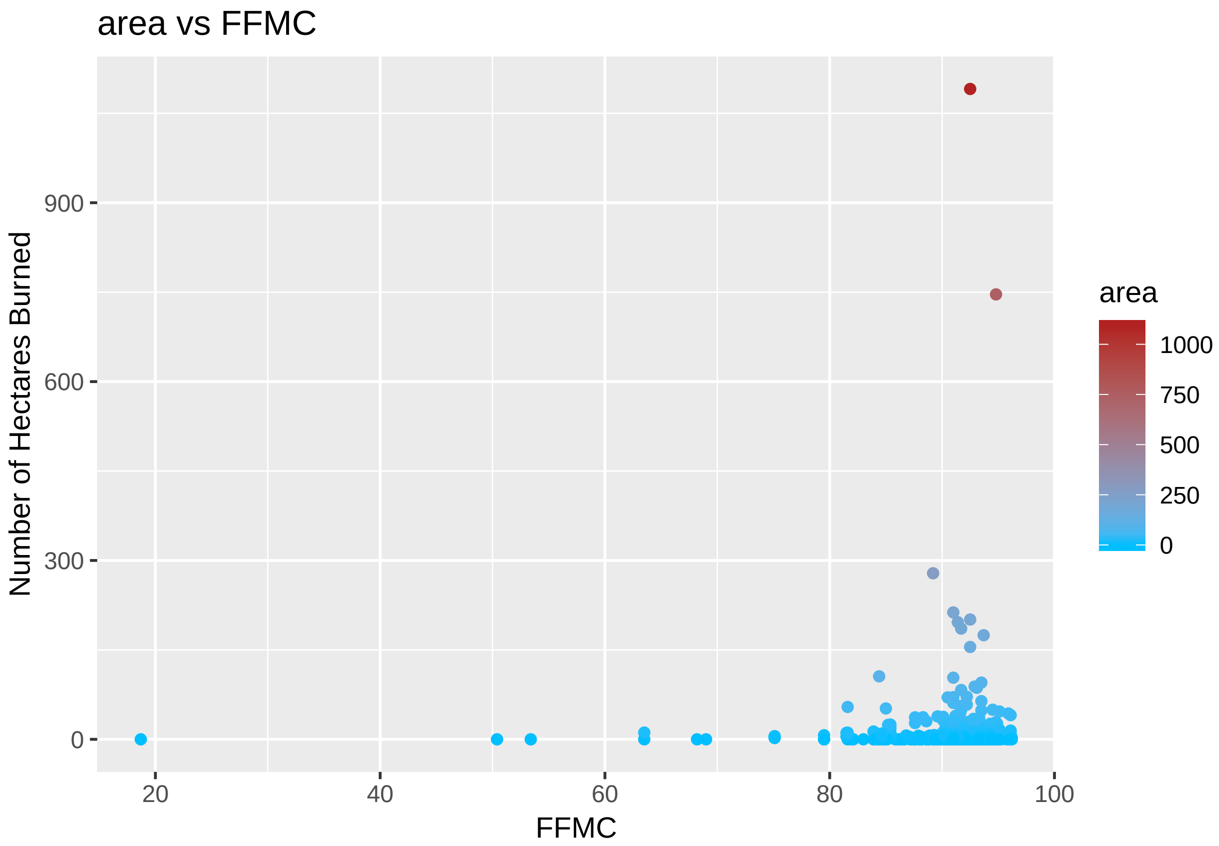
We will create scatterplots to see the relationship between the area (Y) and the following variables (X):

- FFMC: Fine Fuel Moisture Code index from the FWI system: 18.7 to 96.20
- DMC: Duff Moisture Code index from the FWI system: 1.1 to 291.3
- DC: Drought Code index from the FWI system: 7.9 to 860.6
- ISI: Initial Spread Index from the FWI system: 0.0 to 56.10

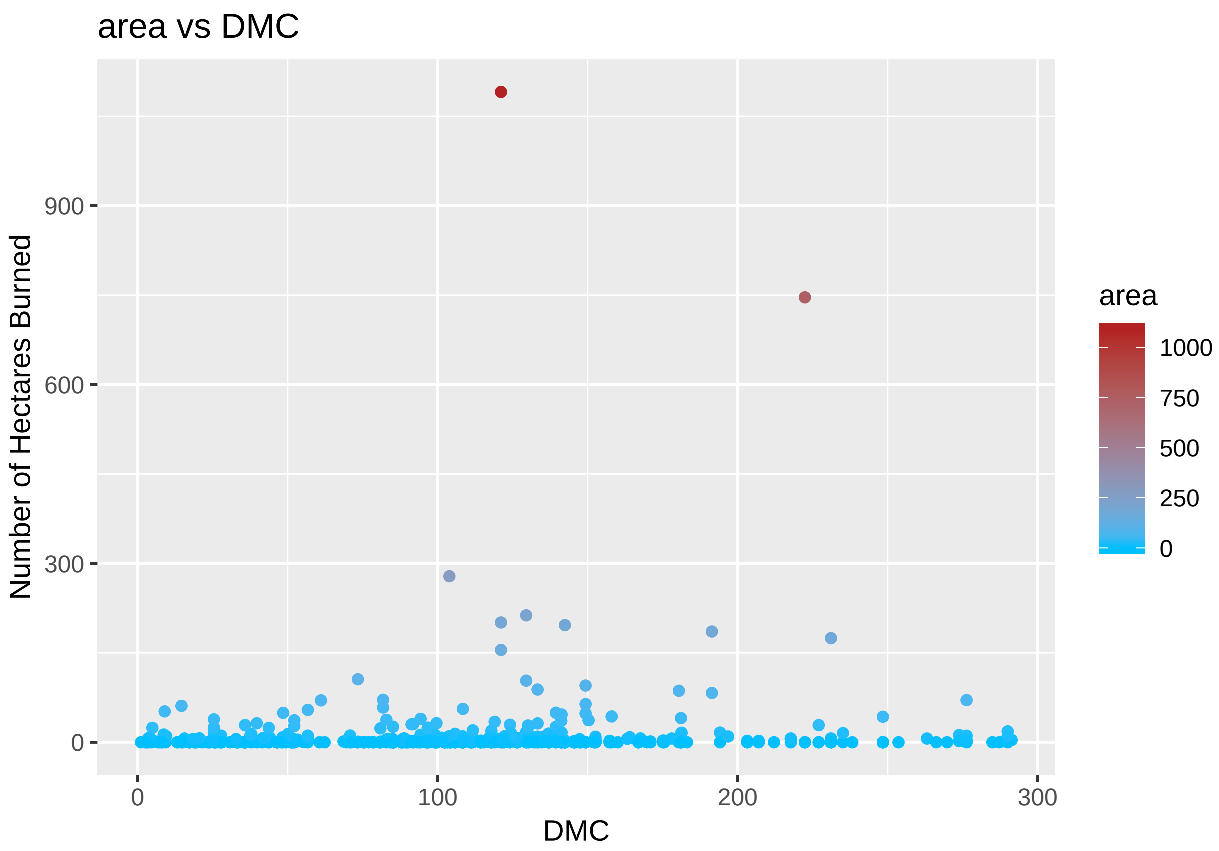
- temp: Temperature in Celsius degrees: 2.2 to 33.30
- RH: Relative humidity in percentage: 15.0 to 100
- wind: Wind speed in km/h: 0.40 to 9.40
- rain: Outside rain in mm/m2 : 0.0 to 6.4

```
create_scatter = function(x, y){
  ggplot(data = fires) +
    aes_string(x = x, y = y, color = y) +
    geom_point() +
    labs(x = x, y = "Number of Hectares Burned", title = paste(y, "vs", x)) +
    scale_color_gradient(low = "deepskyblue", high = "firebrick")
}

x_var <- names(fires[5:12])
y_var <- names(fires[13])
fire_area_vs_var <- map2(x_var, y_var, create_scatter)
fire_area_vs_var
```

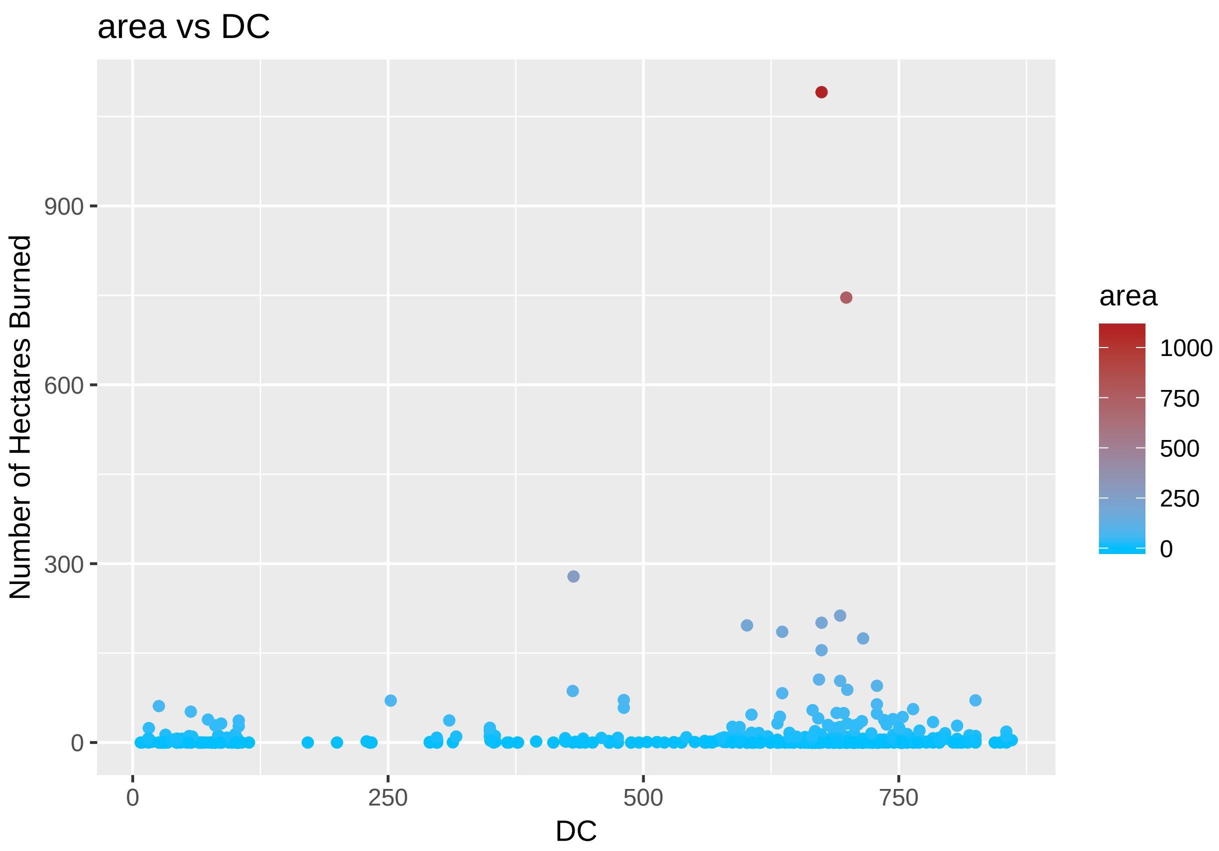
```
## [[1]]
```



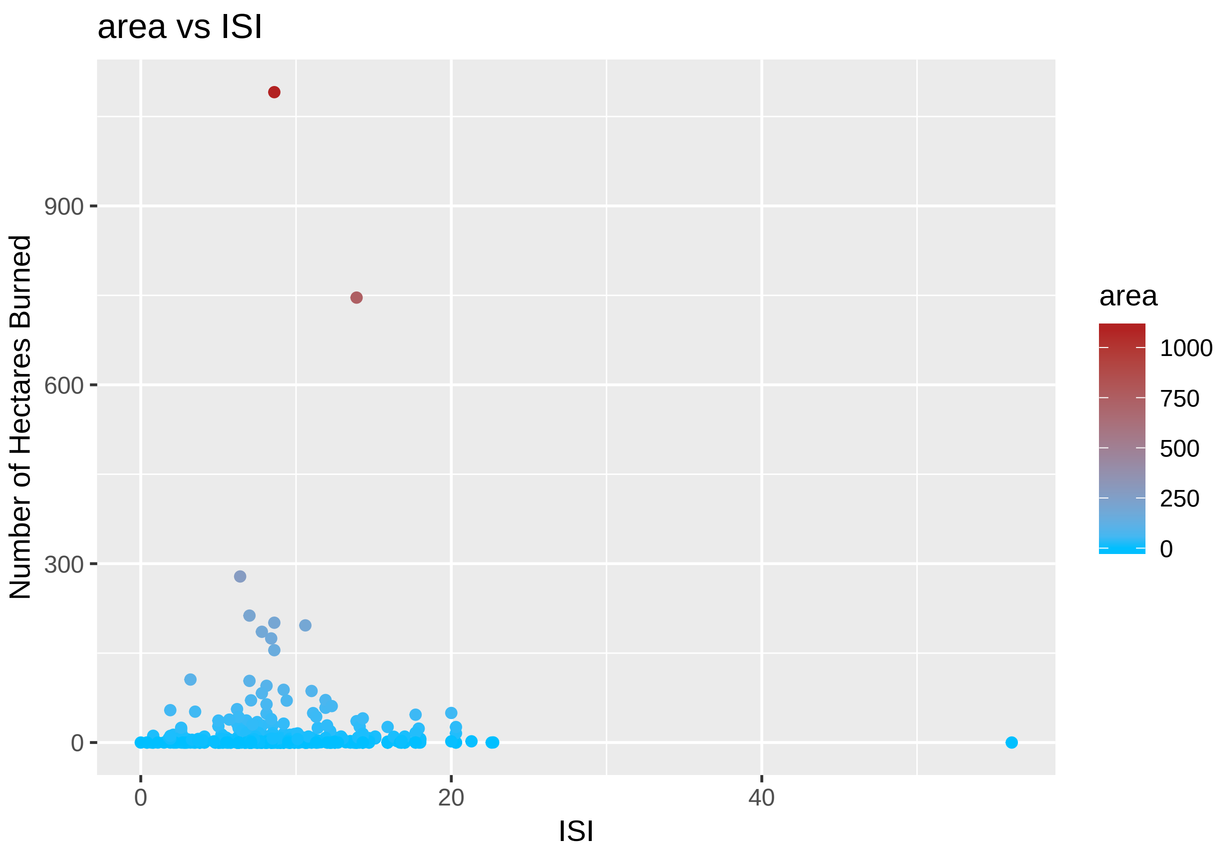
```
##
## [[2]]
```


area vs DMC
Number of Hectares Burned
900
600
300
0
0
100
200
300
DMC
area
1000
750
500
250
0

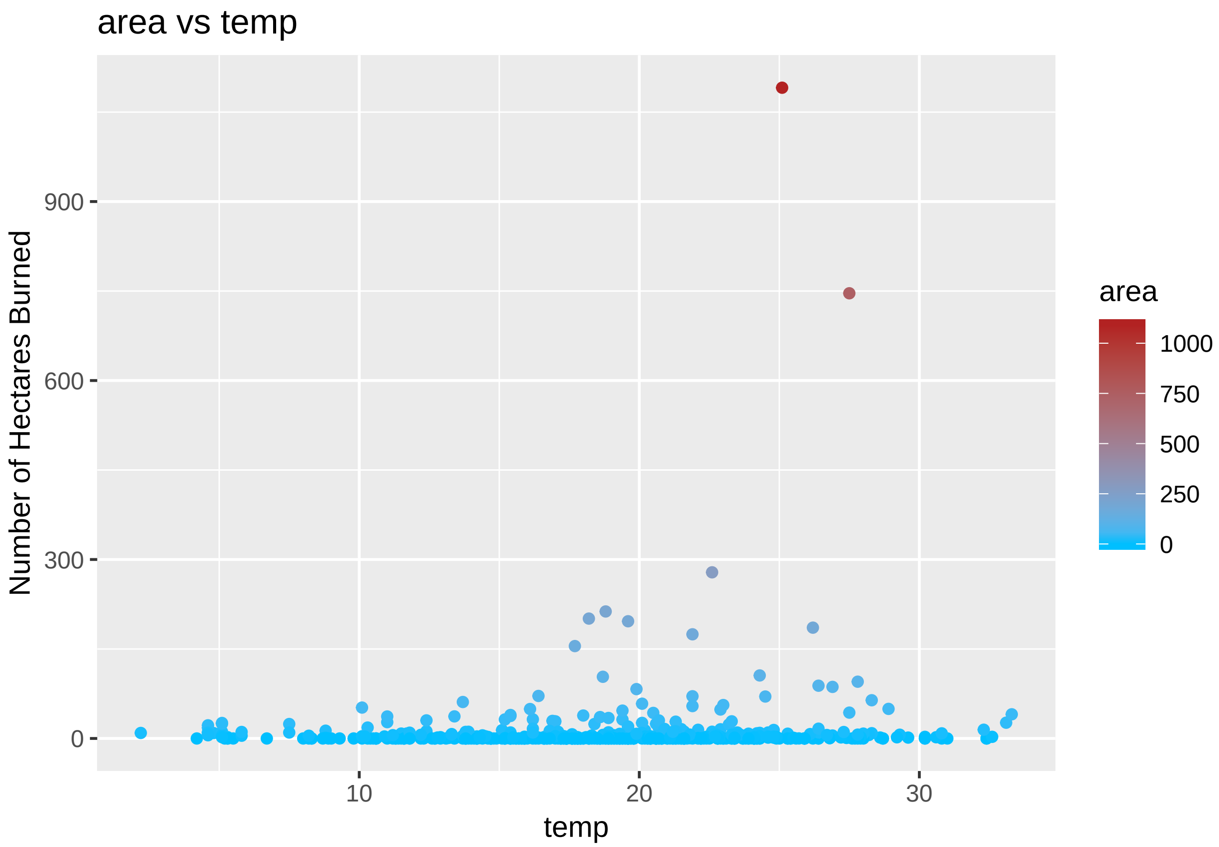
```
##
## [[3]]
```


area vs DC
Number of Hectares Burned
900
600
300
0
0
250
500
750
DC
area
1000
750
500
250
0

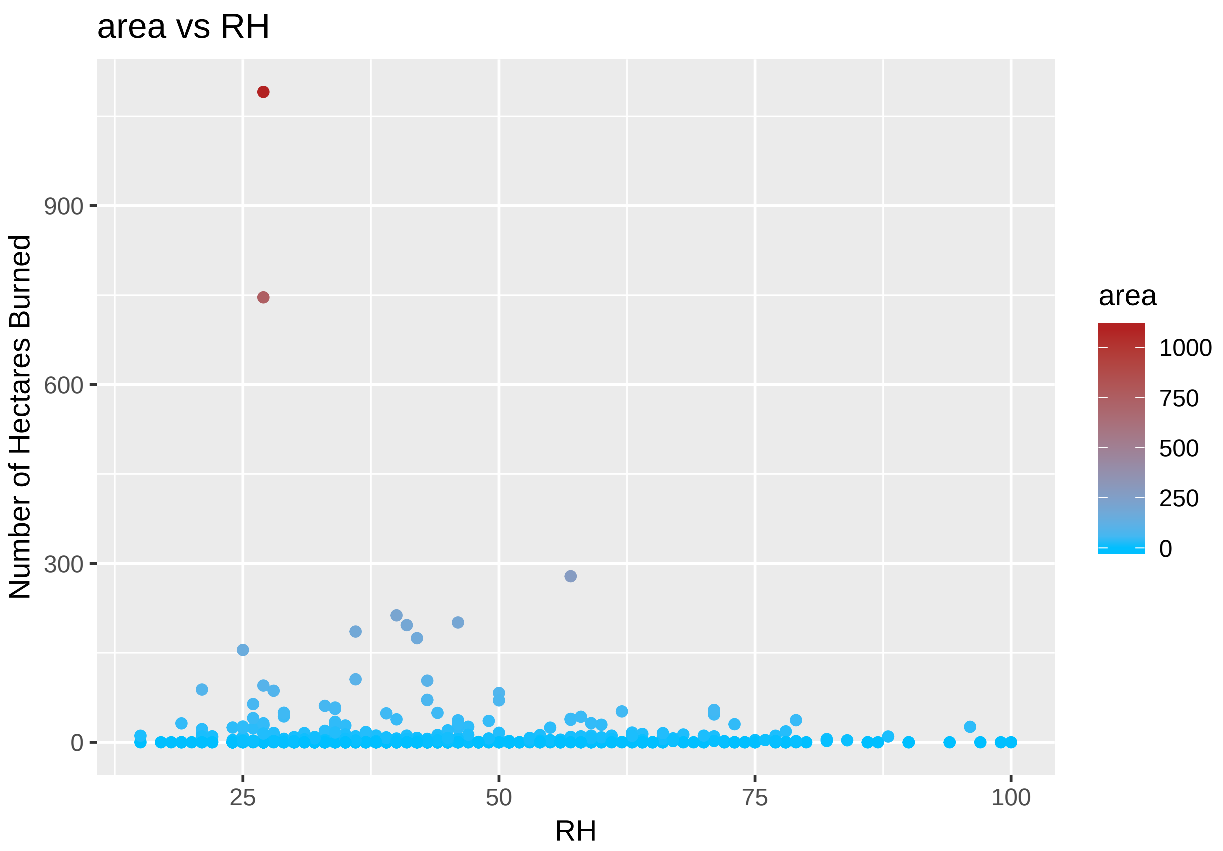
```
##
## [[4]]
```


area vs ISI
Number of Hectares Burned
900
600
300
0
0
20
40
ISI
area
1000
750
500
250
0

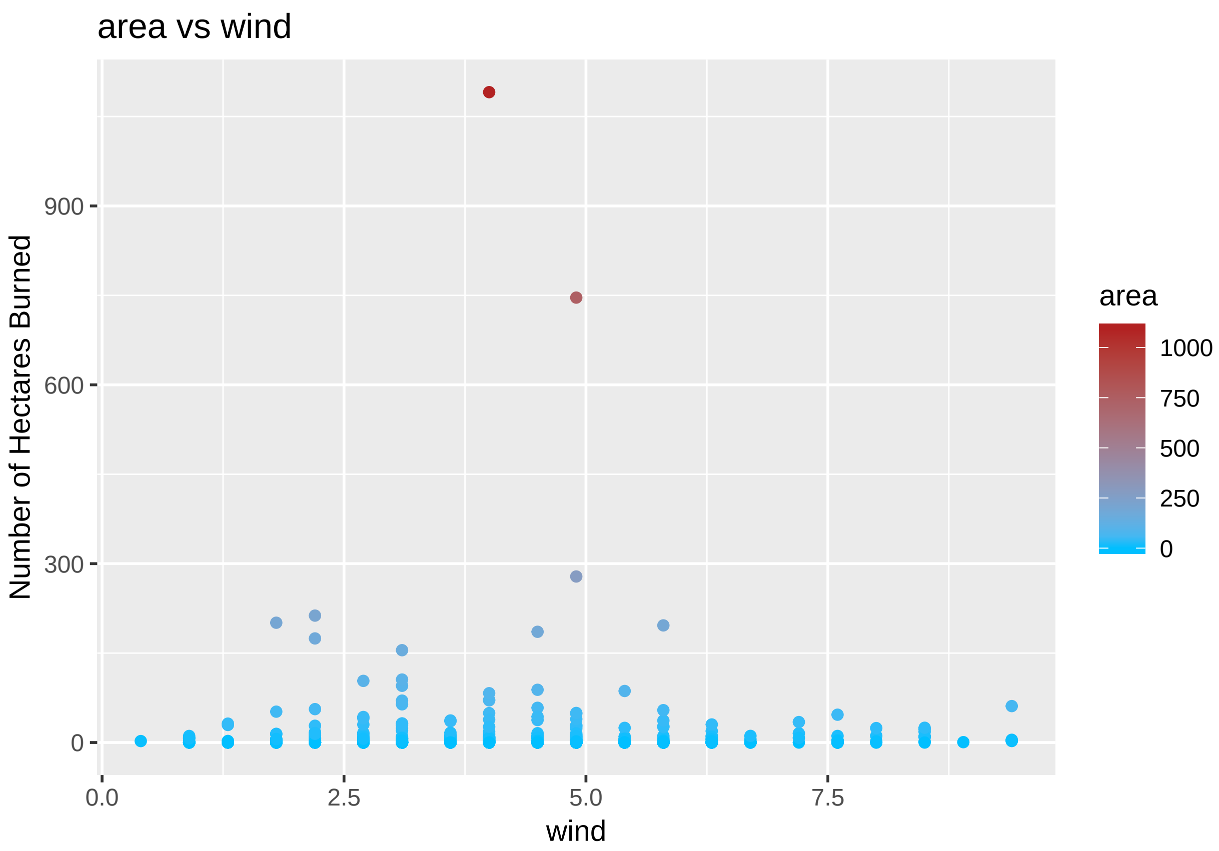
```
##
## [[5]]
```


area vs temp
Number of Hectares Burned
900
600
300
0
10
20
30
temp
area
1000
750
500
250
0

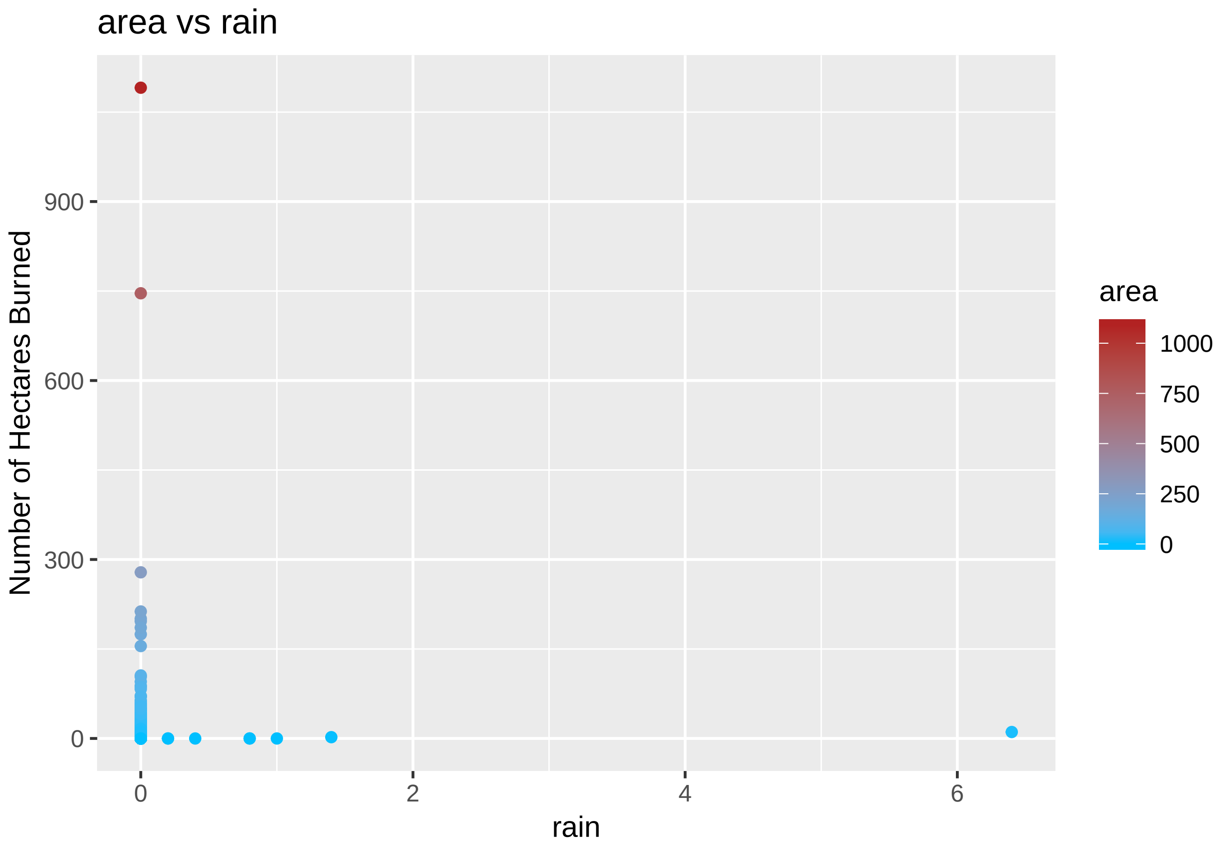
```
##
## [[6]]
```


area vs RH
Number of Hectares Burned
900
600
300
0
25
50
75
100
RH
area
1000
750
500
250
0

```
##
## [[7]]
```

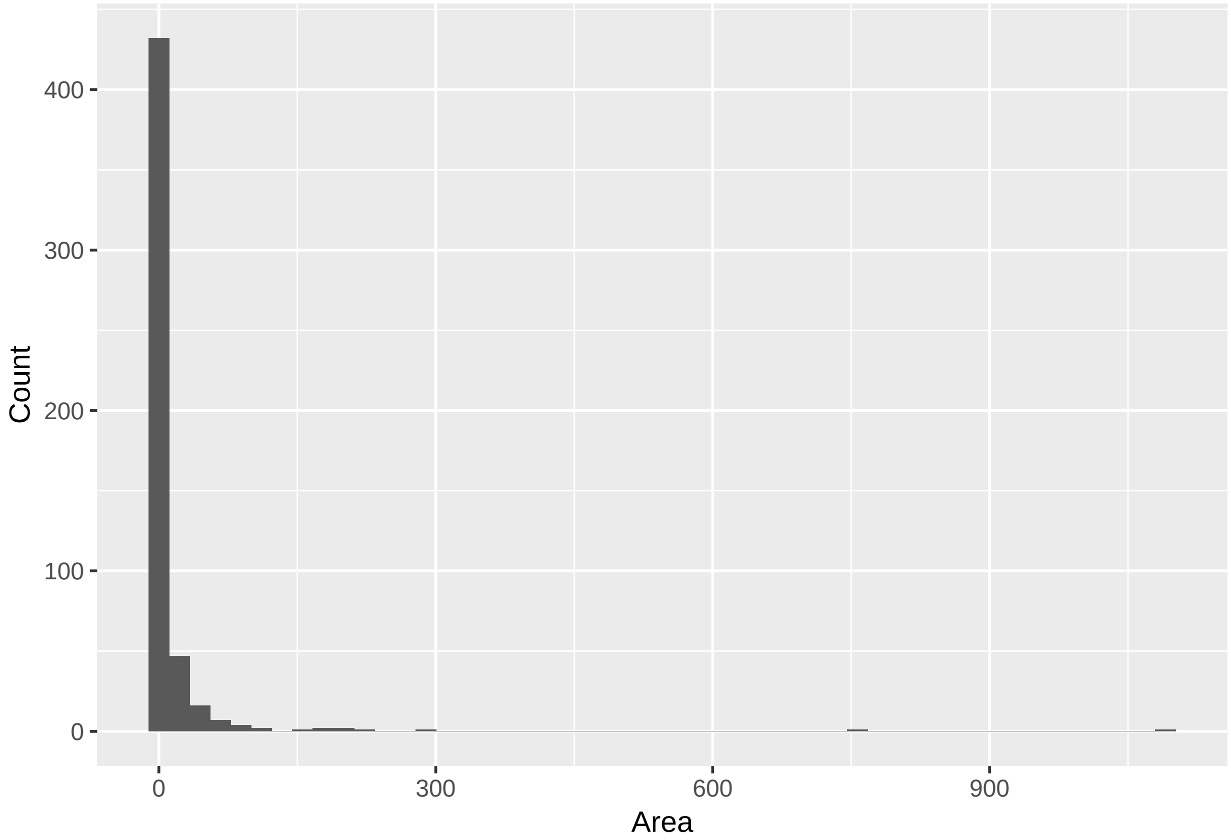

area vs wind
Number of Hectares Burned
900
600
300
0
0.0
2.5
5.0
7.5
wind
area
1000
750
500
250
0

```
##
## [[8]]
```



These graphs aren't very good in depicting the relationship between area and the respective variables. This is because there are only a few observations where there was significant severity of hectares burned. Most observations have close to 0 hectares burned. The distribution of the `area` variable can better be shown through the histogram below.

```
ggplot(data = fires) +
  aes(x = area, color = area) +
  geom_histogram(bins = 50) +
  labs(Title = "Distribution of Area", x = "Area", y = "Count")
```



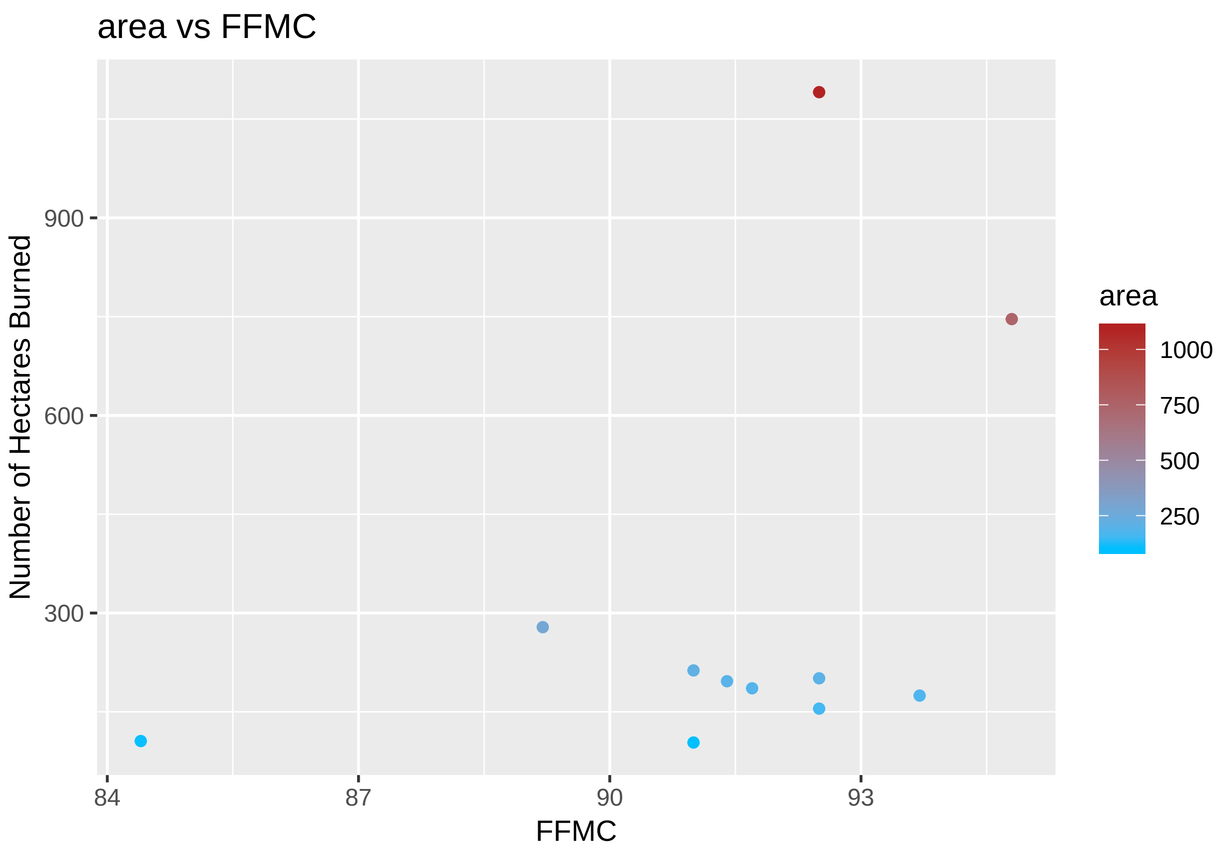
There are over 400 observations with a very low number of hectares burned as mentioned.There are too few data points with large area values to judge the correlation. One solution to this is by filtering the data to only look at observations where the area is sufficiently large.

Analyzing Forest Fire Severity II

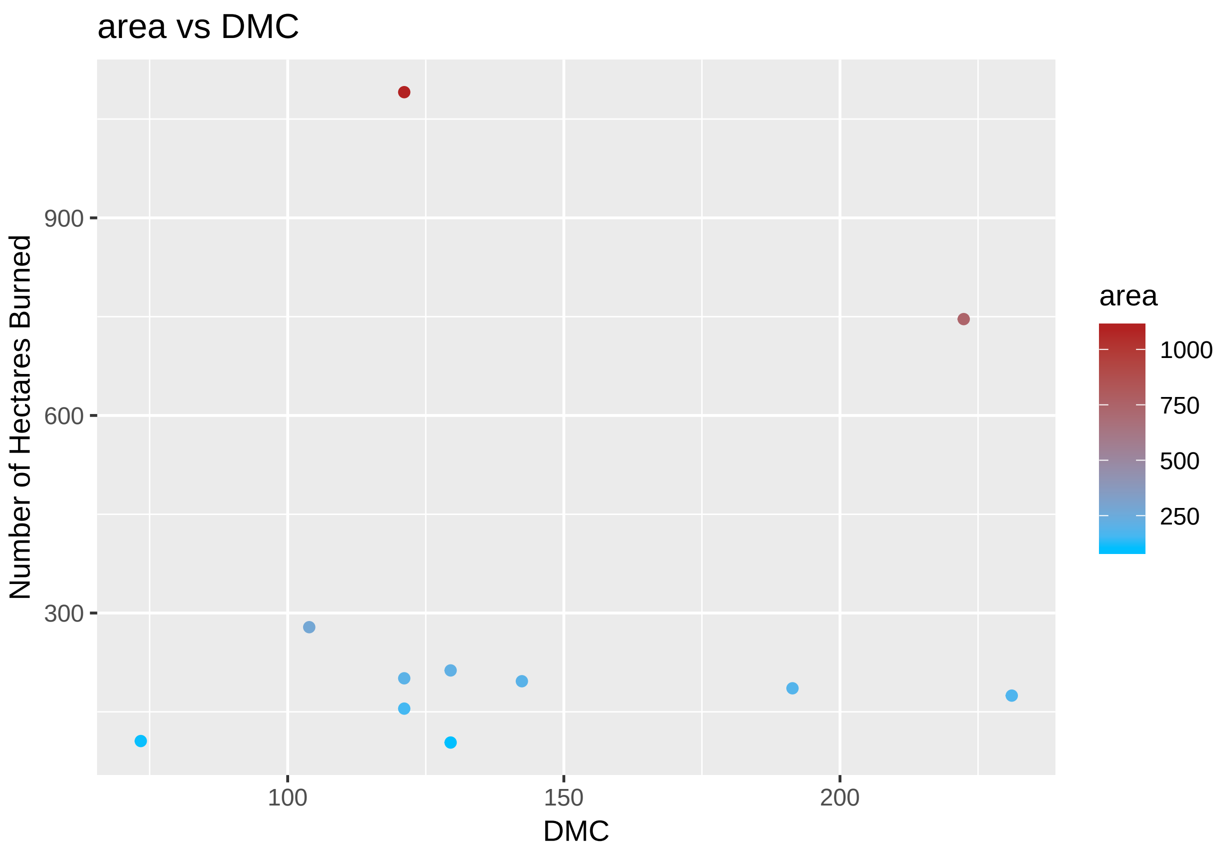
Since there were too many points where the area was really small, we could not see a correlation between area and any other variables. To remedy this, we will stratify the data to only include observations where the area is 100 or greater.

```
fires_area_over_100 <- fires %>% filter(area >= 100)

#Printing Scatterplots
create_scatter_100 = function(x, y){
  ggplot(data = fires_area_over_100) +
    aes_string(x = x, y = y, color = y) +
    geom_point() +
    labs(x = x, y = "Number of Hectares Burned", title = paste(y, "vs", x)) +
    scale_color_gradient(low = "deepskyblue", high = "firebrick")
}

x_var <- names(fires[5:12])
y_var <- names(fires[13])
fire_area_vs_var <- map2(x_var, y_var, create_scatter_100)
fire_area_vs_var
```

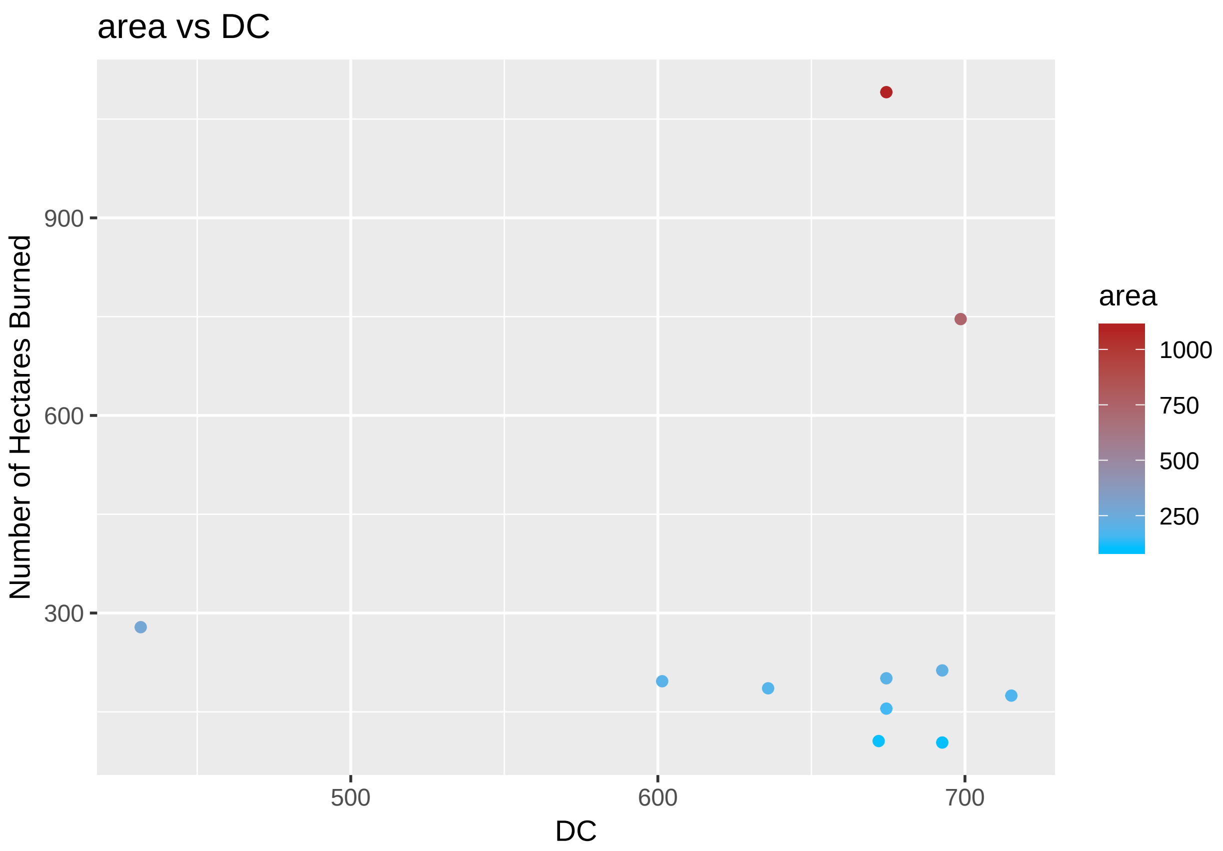
```
## [[1]]
```


area vs FFMC
Number of Hectares Burned
900
600
300
84
87
90
93
FFMC
area
1000
750
500
250

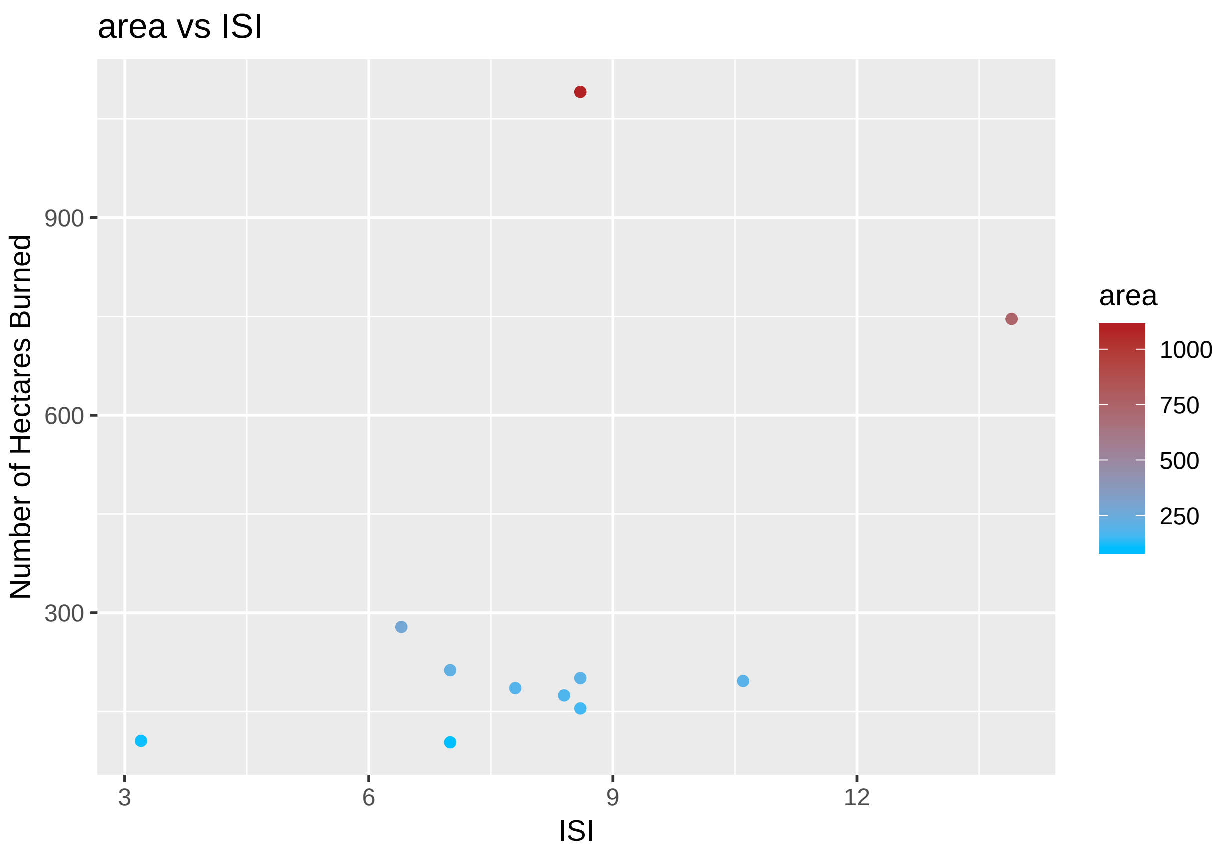
```
##
## [[2]]
```


area vs DMC
Number of Hectares Burned
900
600
300
100
150
200
DMC
area
1000
750
500
250

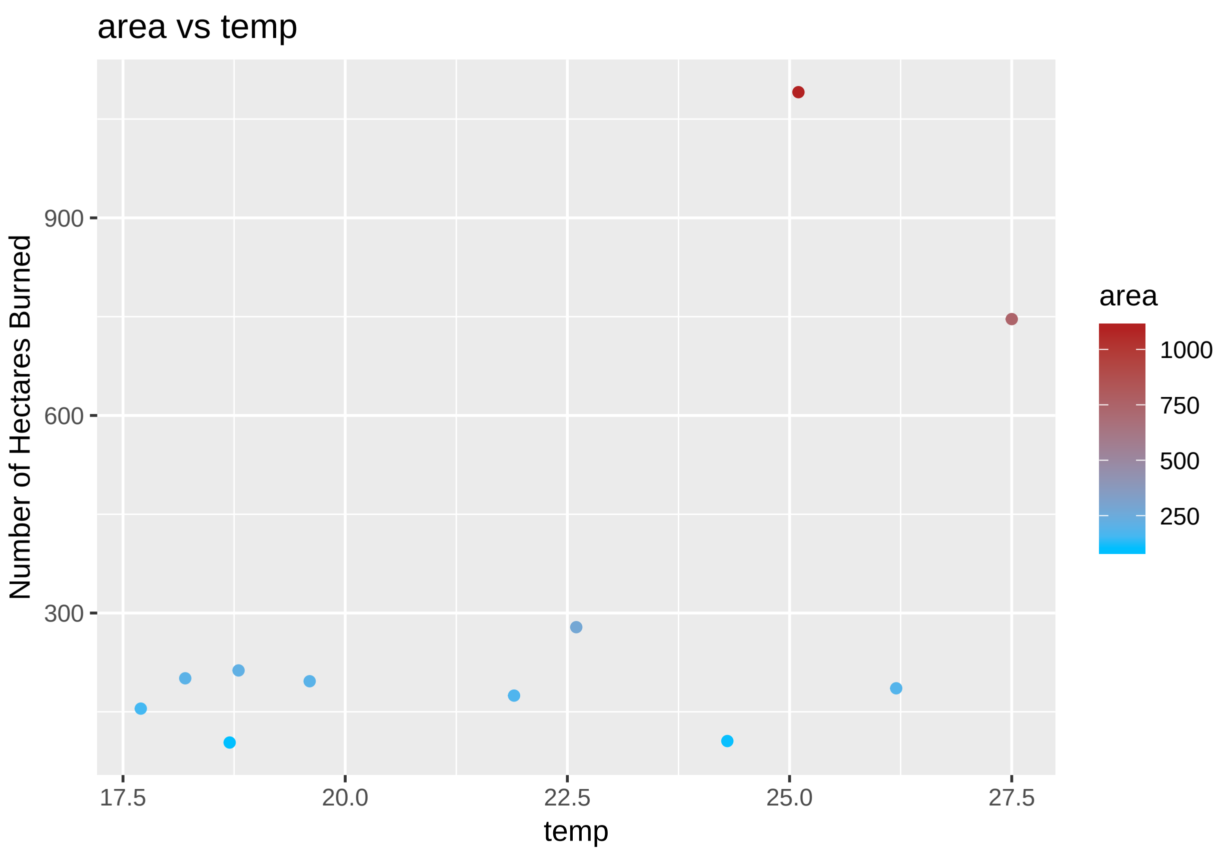
```
##
## [[3]]
```


area vs DC
Number of Hectares Burned
900
600
300
500
600
700
DC
area
1000
750
500
250

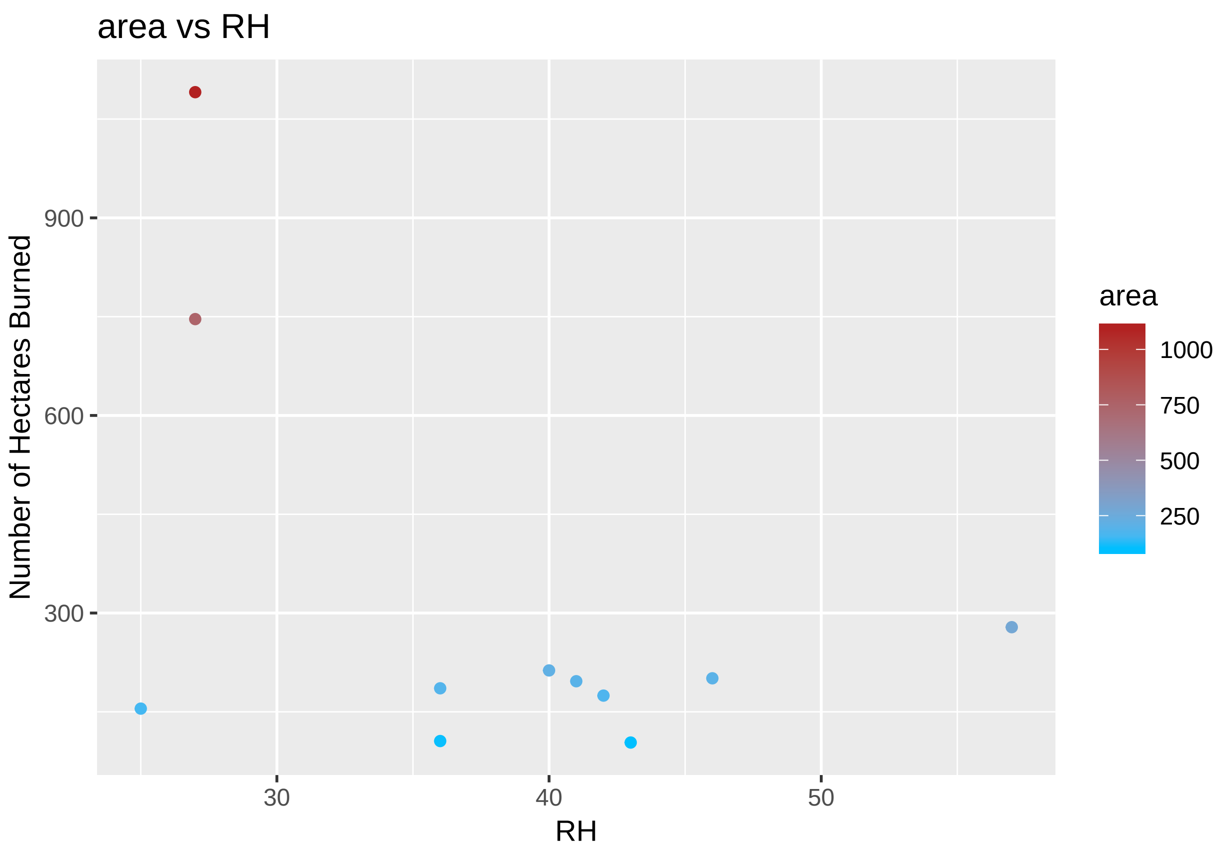
```
##
## [[4]]
```


area vs ISI
Number of Hectares Burned
900
600
300
3
6
9
12
ISI
area
1000
750
500
250

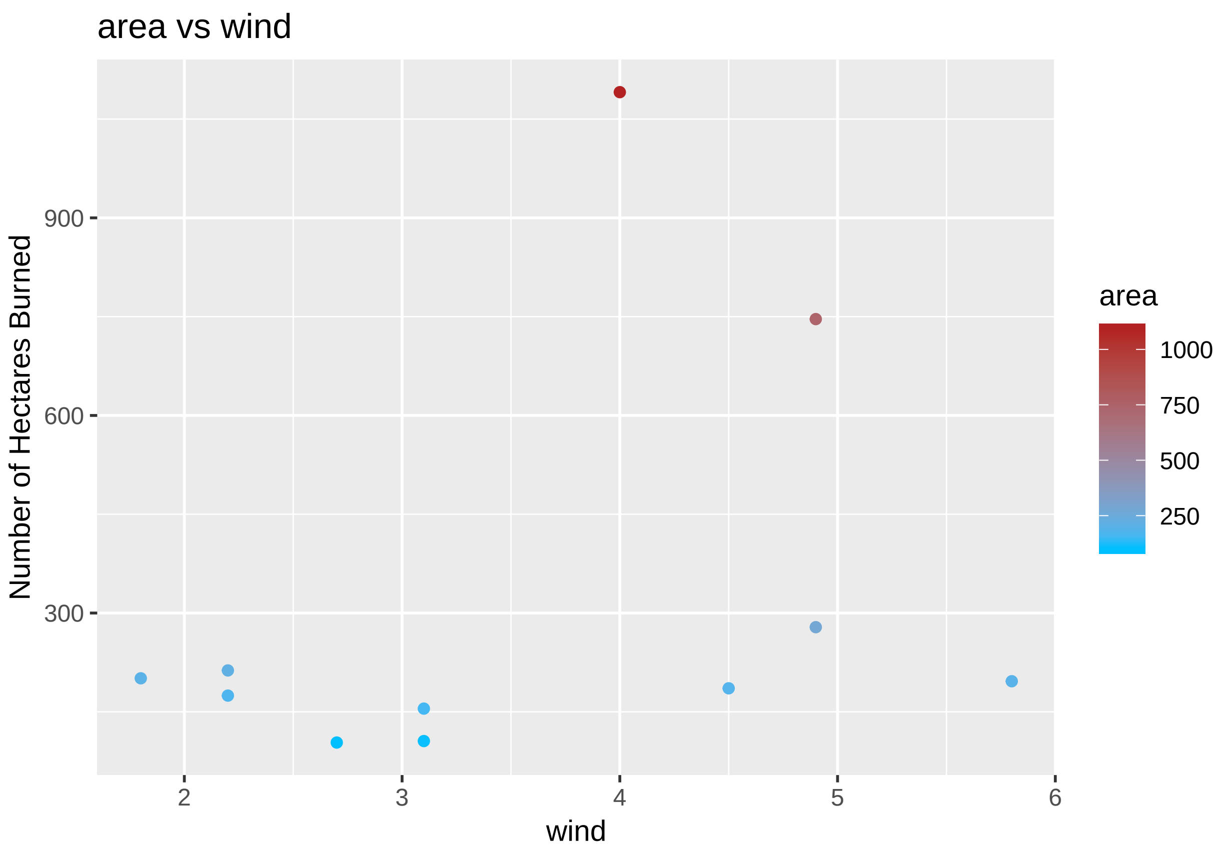
```
##
## [[5]]
```


area vs temp
Number of Hectares Burned
900
600
300
17.5
20.0
22.5
25.0
27.5
temp
area
1000
750
500
250

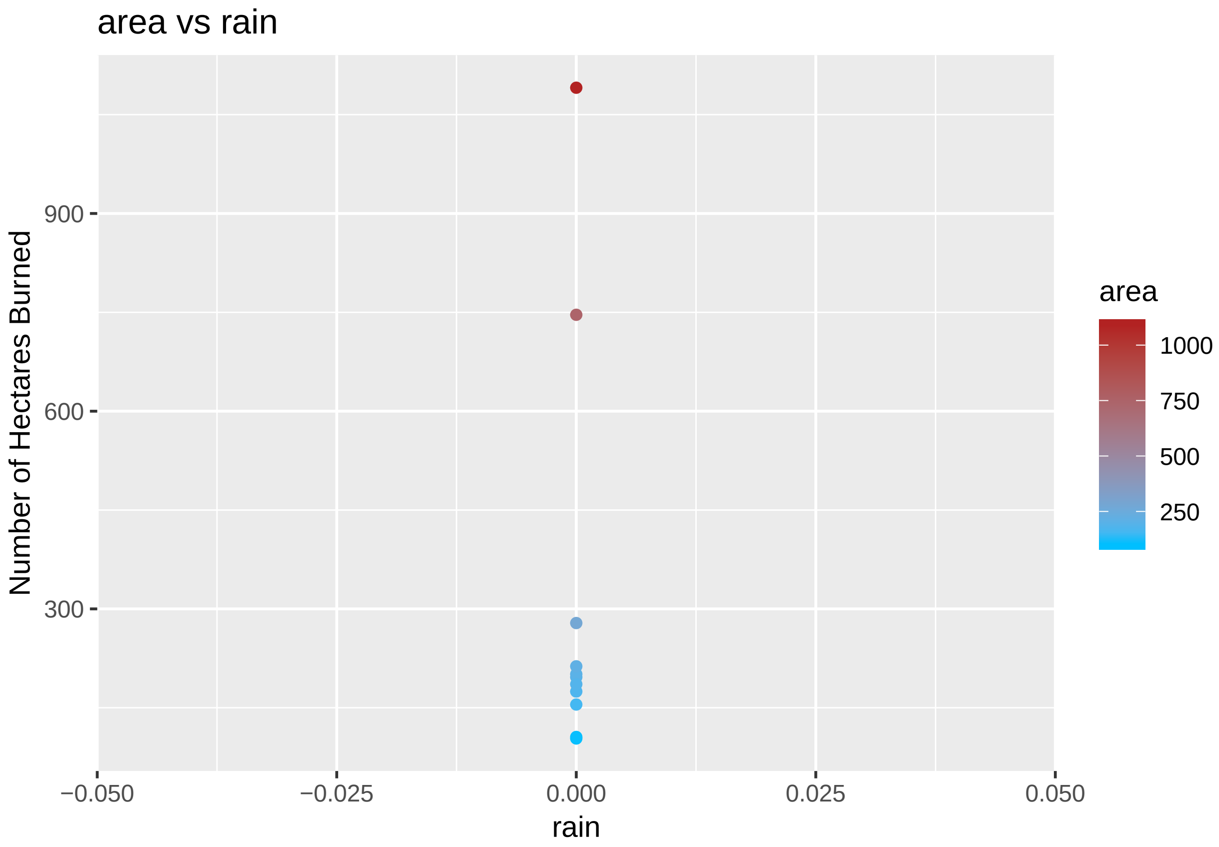
```
##
## [[6]]
```


area vs RH
Number of Hectares Burned
900
600
300
30
40
50
RH
area
1000
750
500
250

```
##
## [[7]]
```


area vs wind
Number of Hectares Burned
900
600
300
2
3
4
5
6
wind
area
1000
750
500
250

```
##
## [[8]]
```



There is a slight positive correlation between `area` and `temp` and a slight negative correlation between `area` and `RH`. This makes sense because that means that forest fires increase in severity with high temperatures and low moisture.

Conclusion

Through looking at the forest fires data, we have found that forest fires are most likely to occur on weekends and the months August and September. Looking deeper into the data, we were able to deduce that the increase in temperature and the decrease in humidity augment the severity of forest fires. One safety precaution when starting a fire of any kind in the forest is to always carry a sufficient amount of water with you just in case a fire breaks out.