Analyzing Forest Fire Data

Forest fires can create ecological problems and endanger human lives and property. Understanding when they occur and what causes them is important for managing them.

The data in this project is associated with a scientific research paper on predicting the occurrence of forest fires in Portugal using modeling techniques.

```
library(readr)
library(dplyr)
```

```
##
## 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(ggplot2)
library(purrr)
```

```
forest_fires <- read_csv("forestfires.csv")
```

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

```
head(forest_fires)
```

```
## # A tibble: 6 x 13
##       X     Y month day     FFMC   DMC    DC   ISI  temp     RH  wind  rain
area
```

```
##   <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
<dbl>
## 1     7     5 mar   fri    86.2  26.2  94.3   5.1   8.2    51   6.7   0
0
## 2     7     4 oct   tue    90.6  35.4 669.    6.7  18      33   0.9   0
0
## 3     7     4 oct   sat    90.6  43.7 687.    6.7  14.6    33   1.3   0
0
## 4     8     6 mar   fri    91.7  33.3  77.5   9     8.3    97   4     0.2
0
## 5     8     6 mar   sun    89.3  51.3 102.    9.6  11.4    99   1.8   0
0
## 6     8     6 aug   sun    92.3  85.3 488    14.7  22.2    29   5.4   0
0
```

Here are descriptions of the variables in the data set and the range of values for each taken from the paper:

- **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
- **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

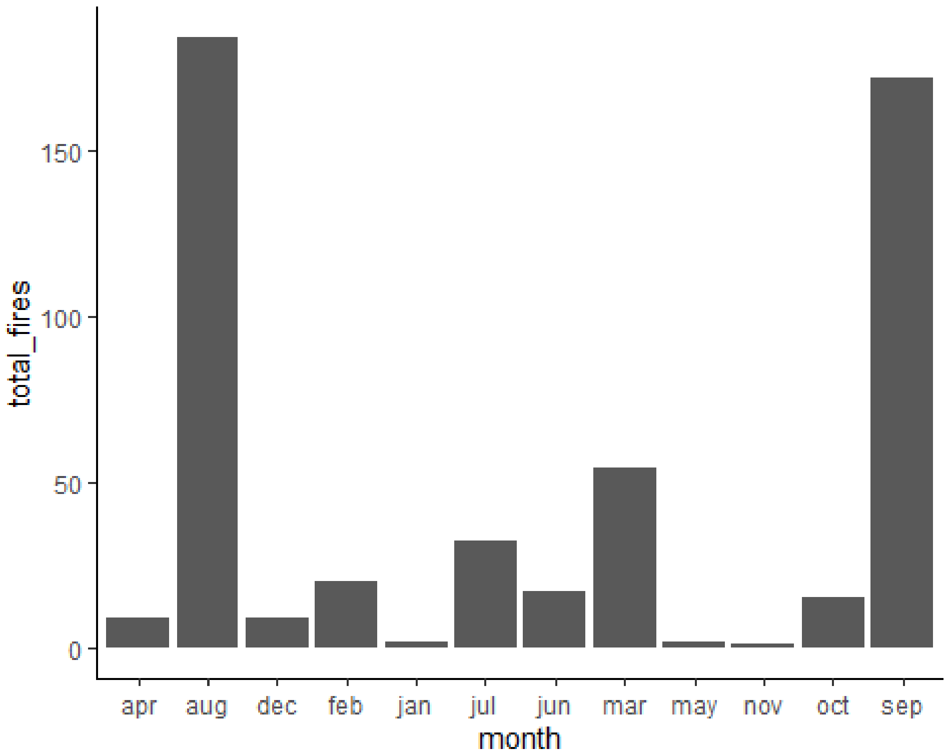
The acronym FWI stands for fire weather index, a method used by scientists to quantify risk factors for forest fires.

There are two variables that have to do with time: month and day. We can ask the questions:

- During which months are forest fires most common?
- On which days of the week are forest fires most common?

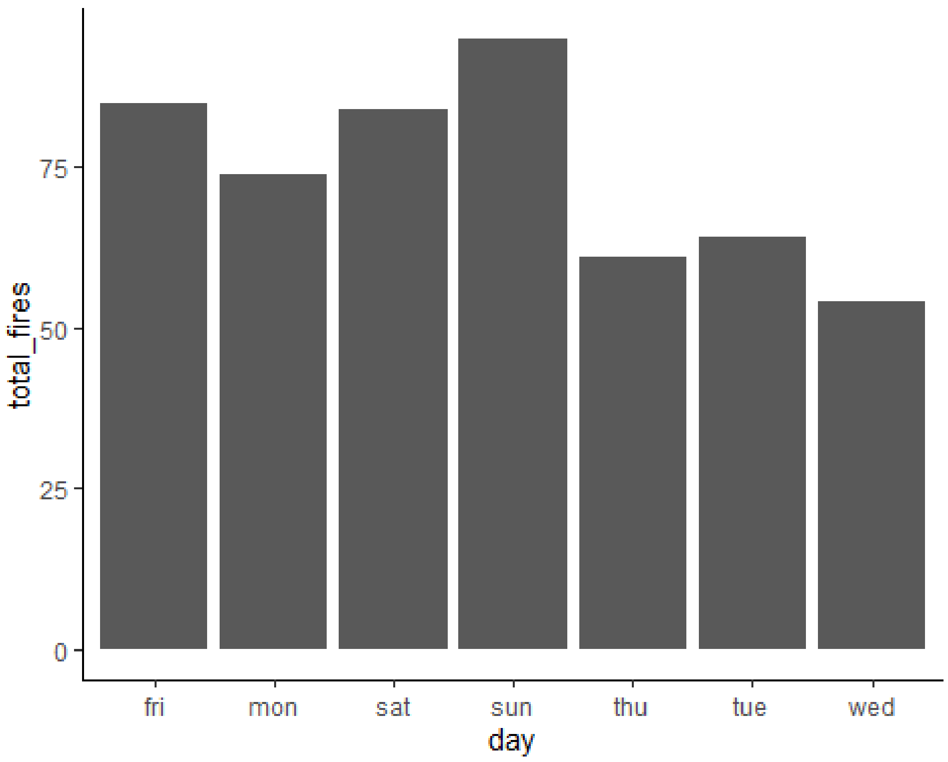
Fires by Month

```
fires_by_month <- forest_fires %>%
  group_by(month) %>%
  summarize(total_fires = n())
ggplot(data = fires_by_month) +
  aes(x = month, y = total_fires) +
  geom_bar(stat = "identity")  +
  theme(panel.background = element_rect(fill = "white"),
        axis.line = element_line(size = 0.25,
                                 colour = "black"))
```



Fires by Day of the Week

```
fires_by_DOW <- forest_fires %>%
  group_by(day) %>%
  summarize(total_fires = n())
ggplot(data = fires_by_DOW) +
  aes(x = day, y = total_fires) +
  geom_bar(stat = "identity") +
  theme(panel.background = element_rect(fill = "white"),
        axis.line = element_line(size = 0.25,
                                 colour = "black"))
```



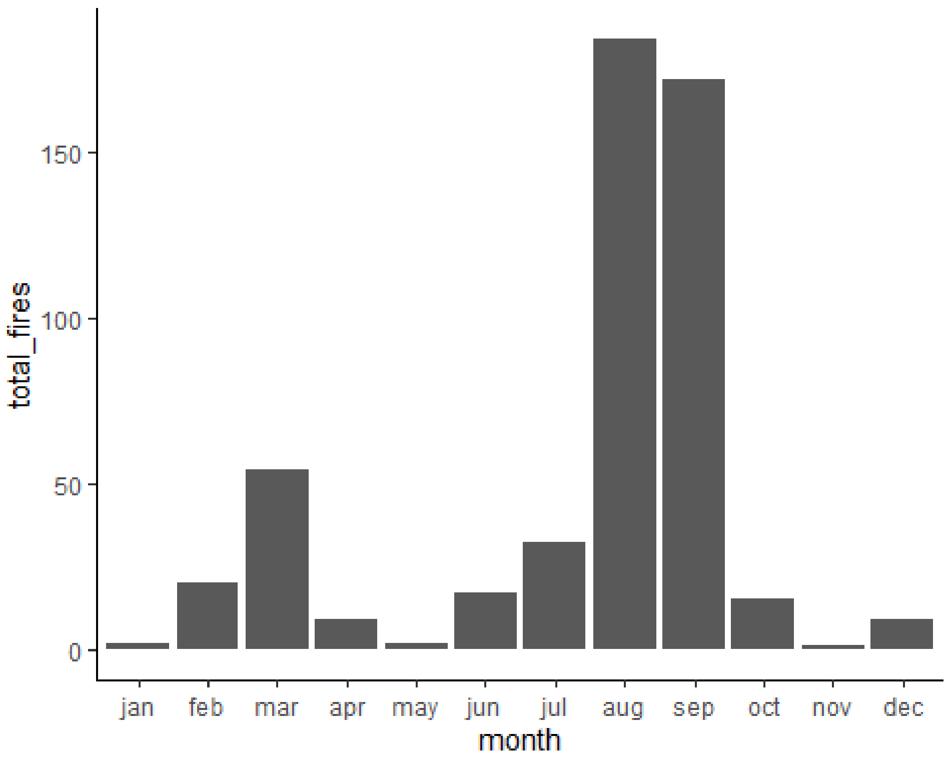
These charts let us see some trends. August and September certainly have the most fires, and Wednesdays seems to have fewer fires than other days of the week.

Changing the order of the months.

```
forest_fires <- forest_fires %>%
  mutate(month = factor(month, levels = c("jan", "feb", "mar", "apr", "may",
"jun", "jul", "aug", "sep", "oct", "nov", "dec")),
         day = factor(day, levels = c("sun", "mon", "tue", "wed", "thu",
"fri", "sat")))
```

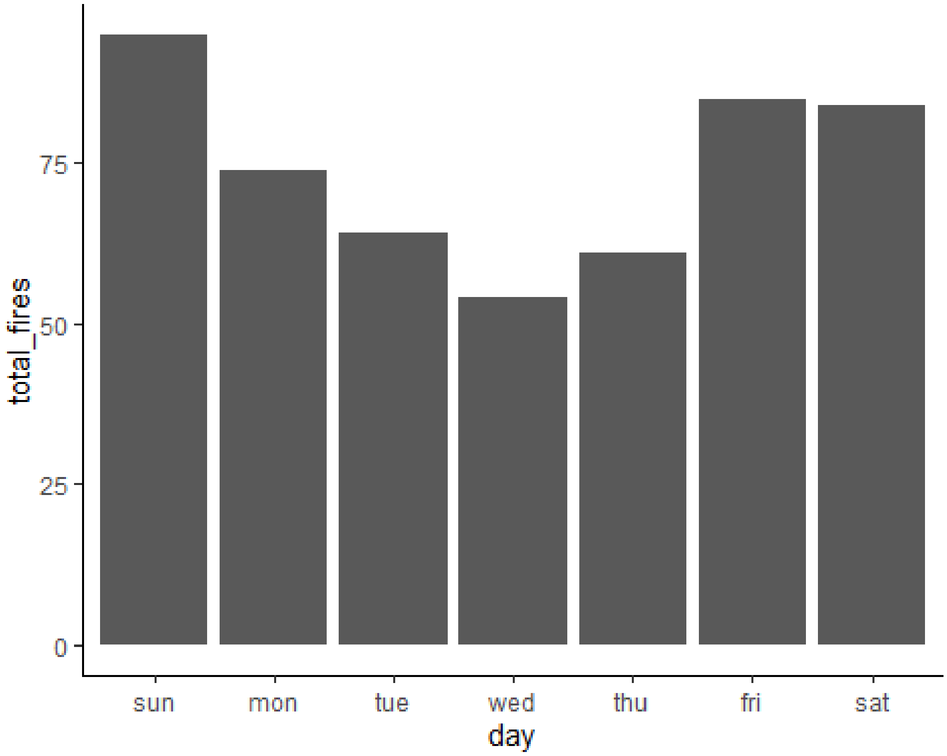
Fires by Month(Reordered)

```
fires_by_month <- forest_fires %>%
  group_by(month) %>%
  summarize(total_fires = n())
ggplot(data = fires_by_month) +
  aes(x = month, y = total_fires) +
  geom_bar(stat = "identity")  +
  theme(panel.background = element_rect(fill = "white"),
        axis.line = element_line(size = 0.25,
                                 colour = "black"))
```



Fired by Day of the Week(Reordered)

```
fires_by_DOW <- forest_fires %>%
  group_by(day) %>%
  summarize(total_fires = n())
ggplot(data = fires_by_DOW) +
  aes(x = day, y = total_fires) +
  geom_bar(stat = "identity") +
  theme(panel.background = element_rect(fill = "white"),
        axis.line = element_line(size = 0.25,
                                 colour = "black"))
```



It’s clear that August and September, late summer months in the Northern hemisphere, see more forest fires than other months.

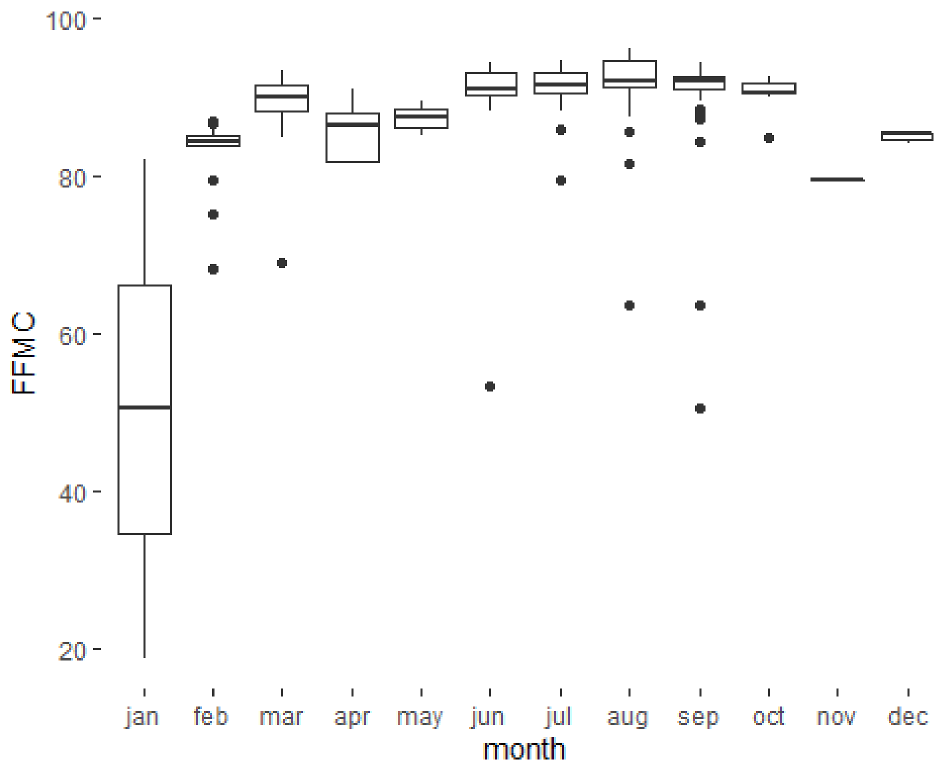
It also looks as though Friday, Saturday, and Sunday have more forest fires than days in the middle of the week.

To explore causes of the temporal patterns of forest fire occurrence the bar charts reveal, we can look more closely at how the variables that relate to forest fires vary by month and by day of the week.

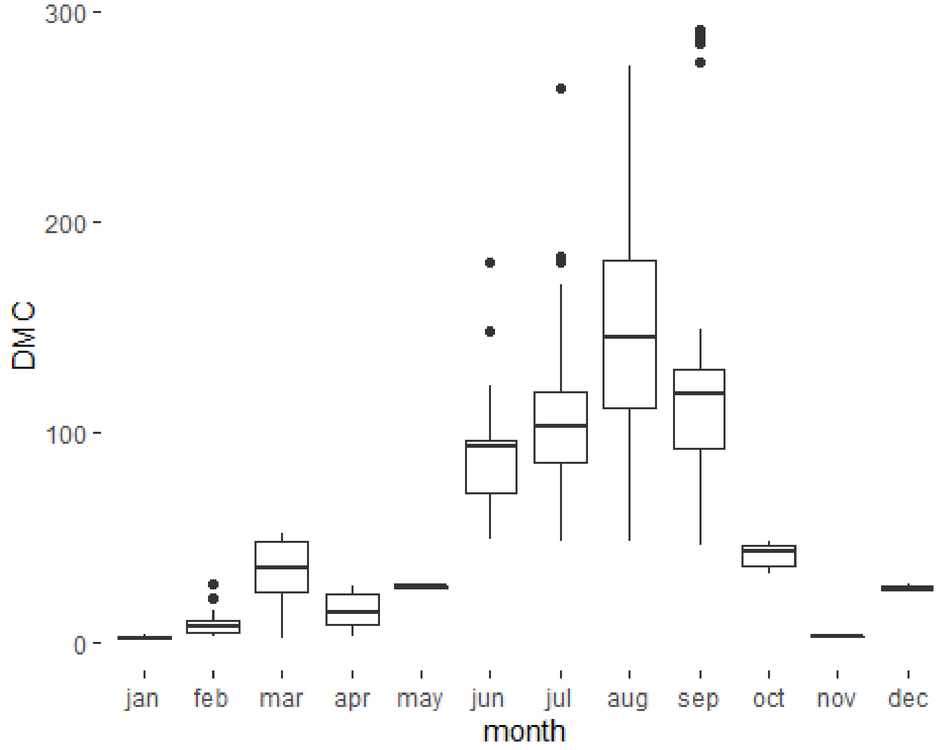
```
create_boxplots <- function(x, y) {
  ggplot(data = forest_fires) +
    aes_string(x = x, y = y) +
    geom_boxplot() +
    theme(panel.background = element_rect(fill = "white"))
}
x_var_month <- names(forest_fires)[3] ## month
y_var <- names(forest_fires)[5:12]

month_box <- map2(x_var_month, y_var, create_boxplots) ## visualize variables
by month
```

Fires by Month(Box Plot)

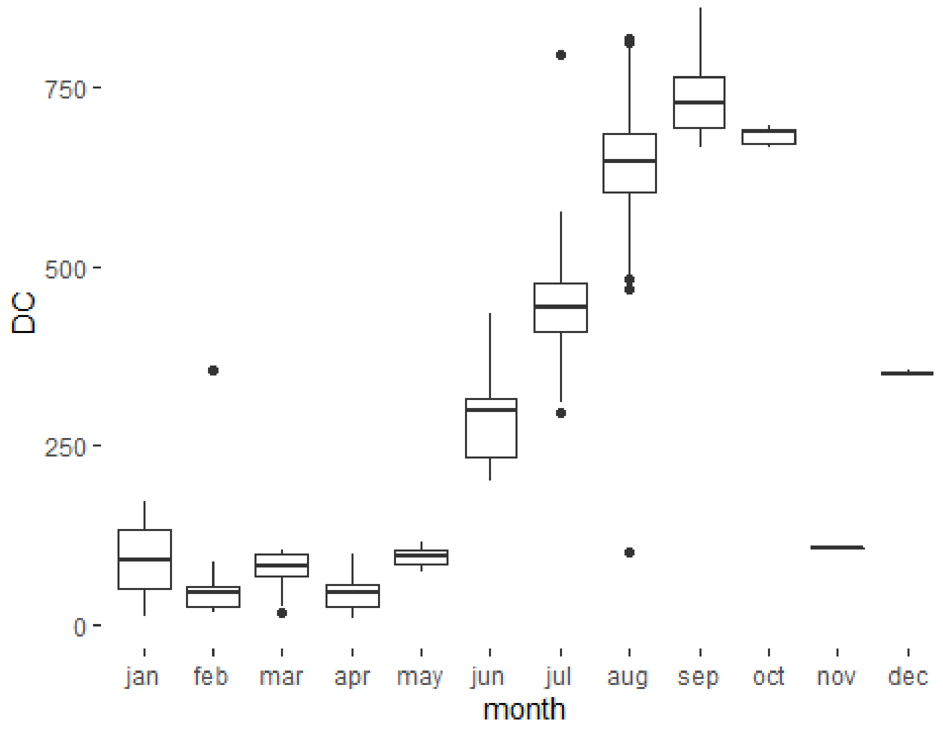
```
month_box

## [[1]]
```


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

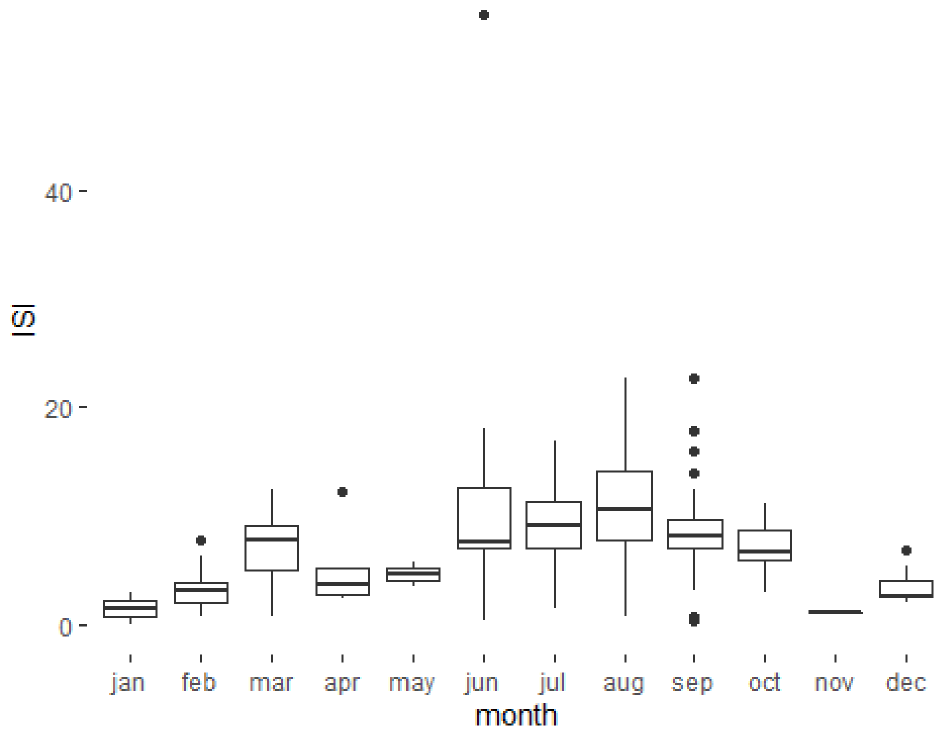
```
##
## [[2]]
```


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

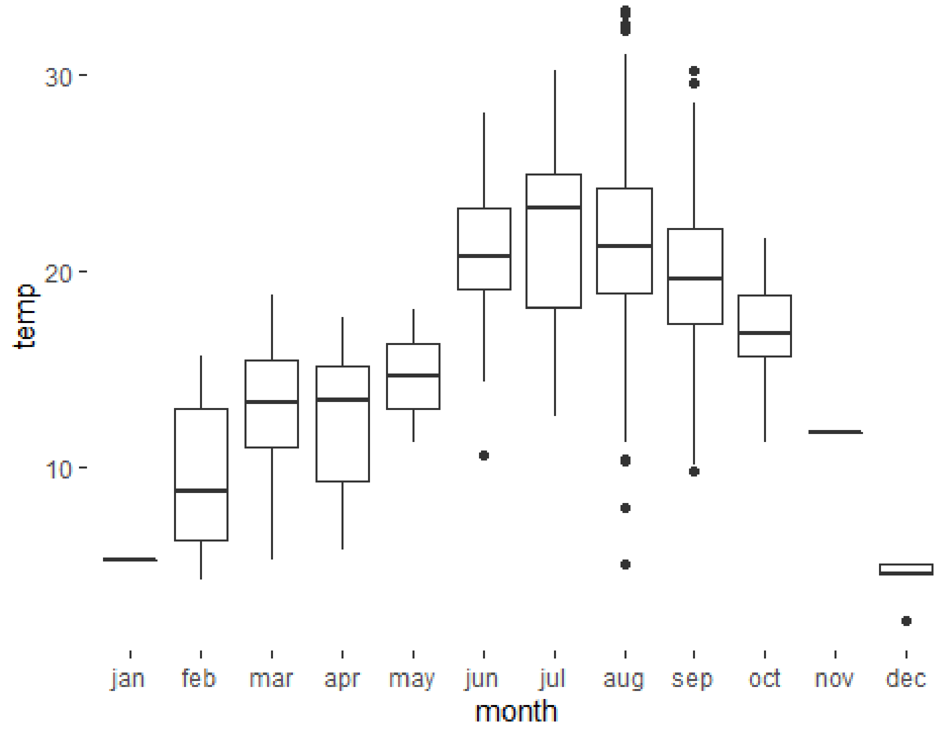
```
##
## [[3]]
```


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

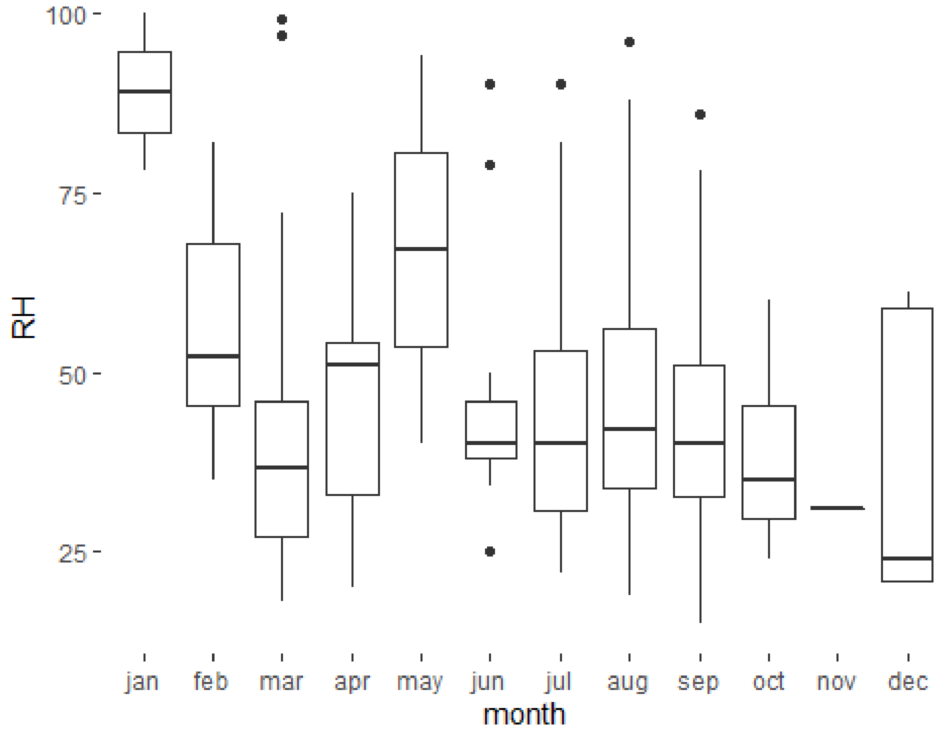
```
##
## [[4]]
```


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

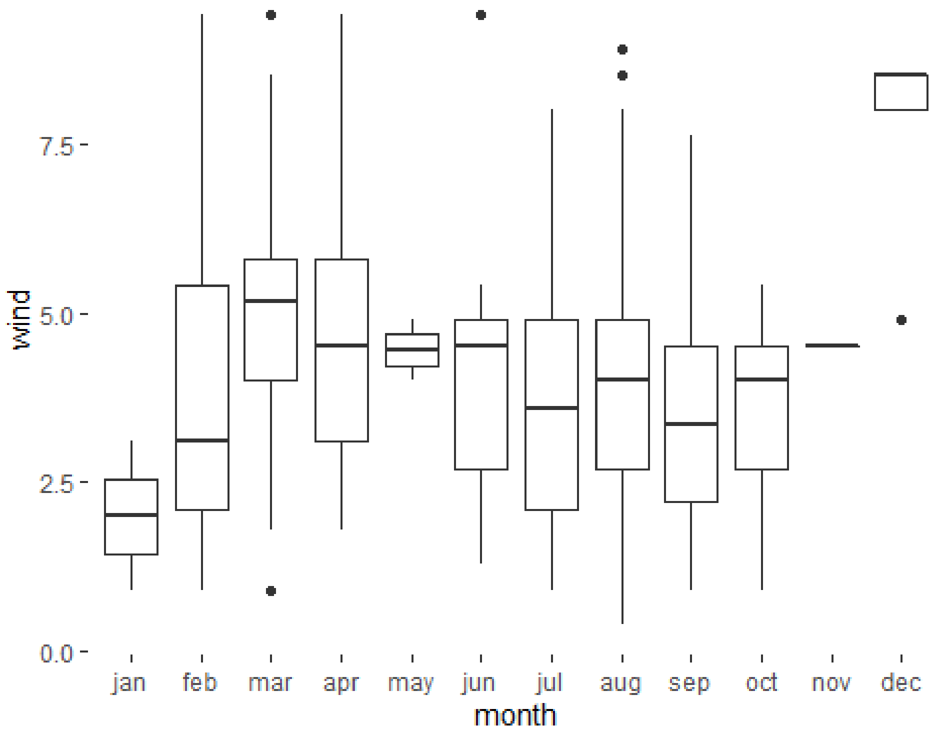
```
##
## [[5]]
```


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

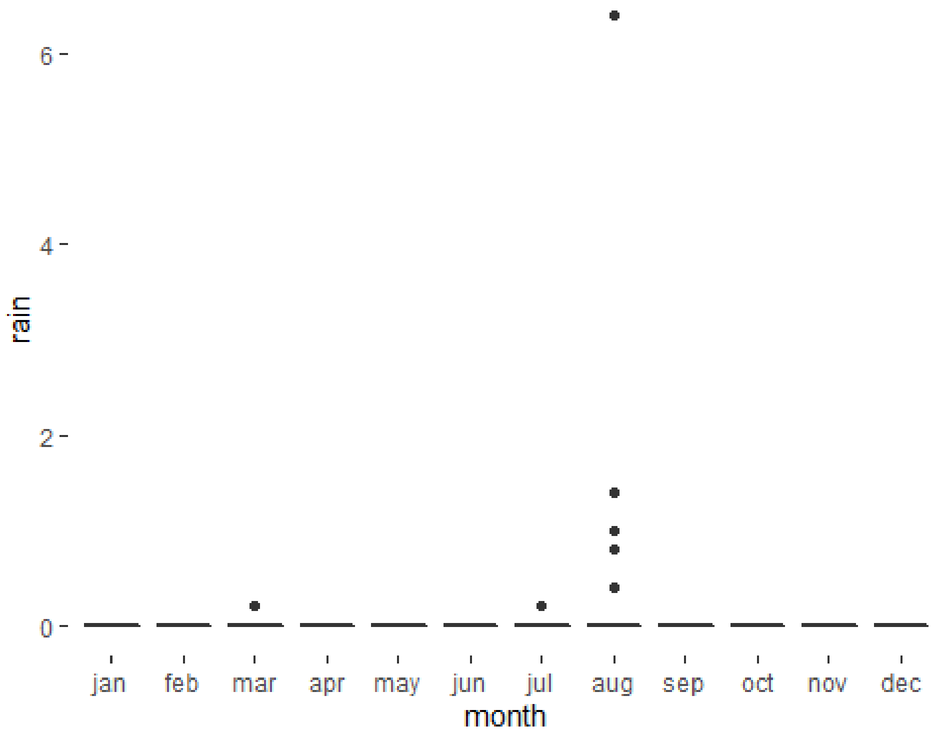
```
##
## [[6]]
```


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

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

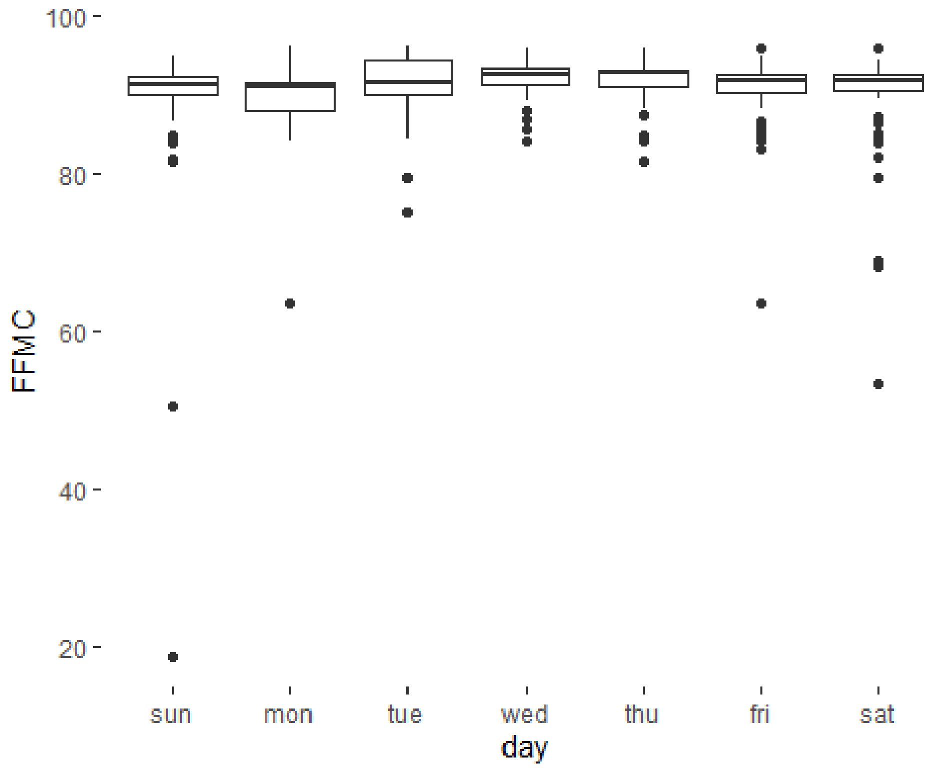

wind
0.0
2.5
5.0
7.5
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month

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

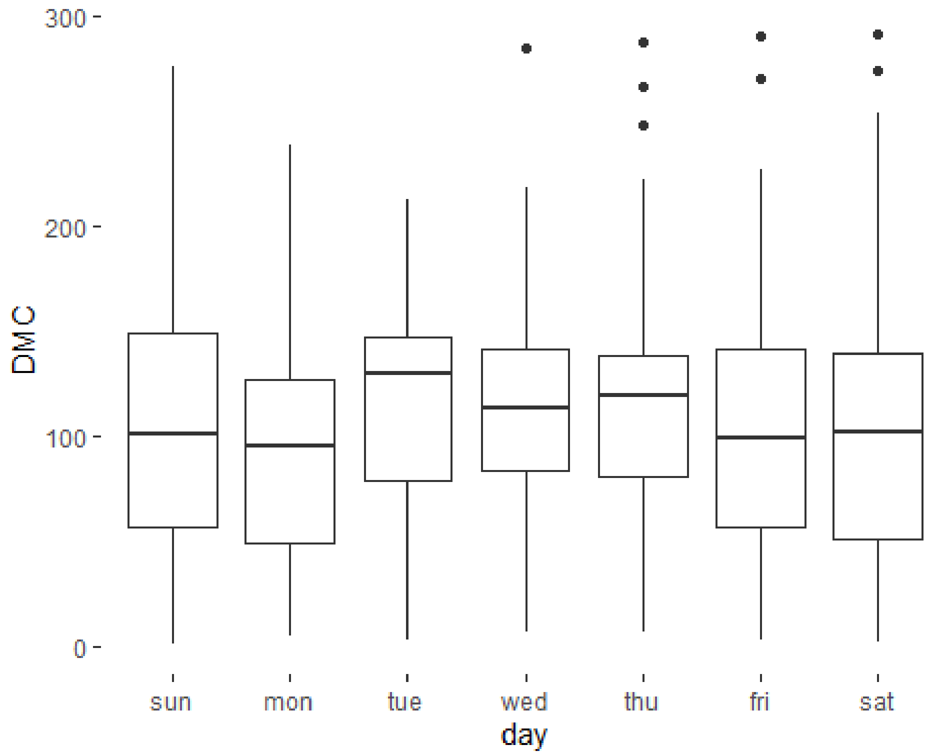

rain
0
2
4
6
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month

Fires by Day of the Week(Box Plot)

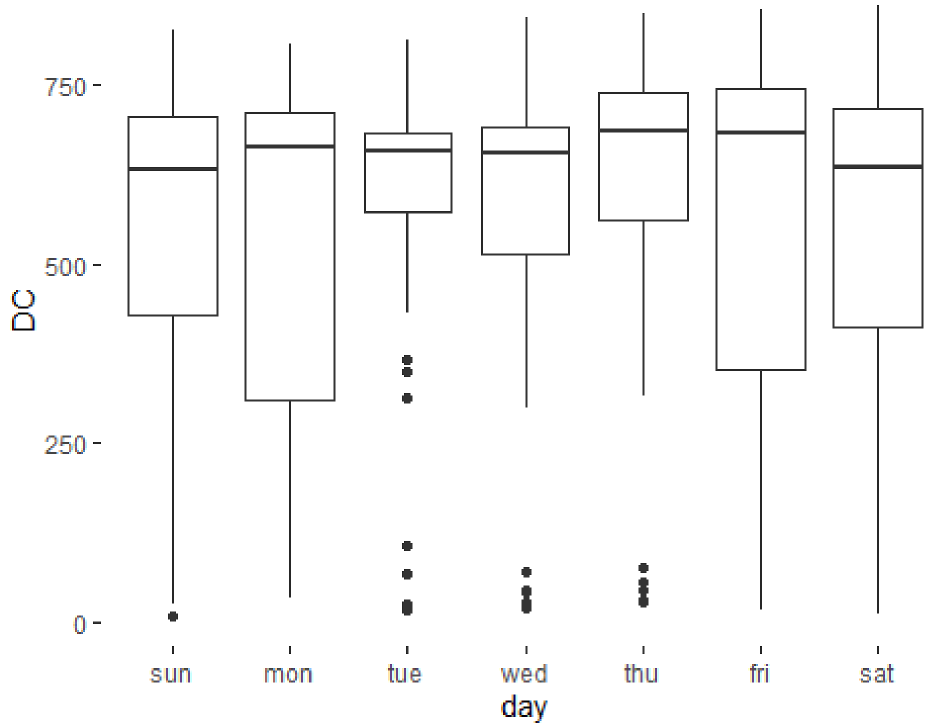
```
x_var_day <- names(forest_fires)[4] ## day

day_box <- map2(x_var_day, y_var, create_boxplots) ## visualize variables by
day

day_box

## [[1]]
```



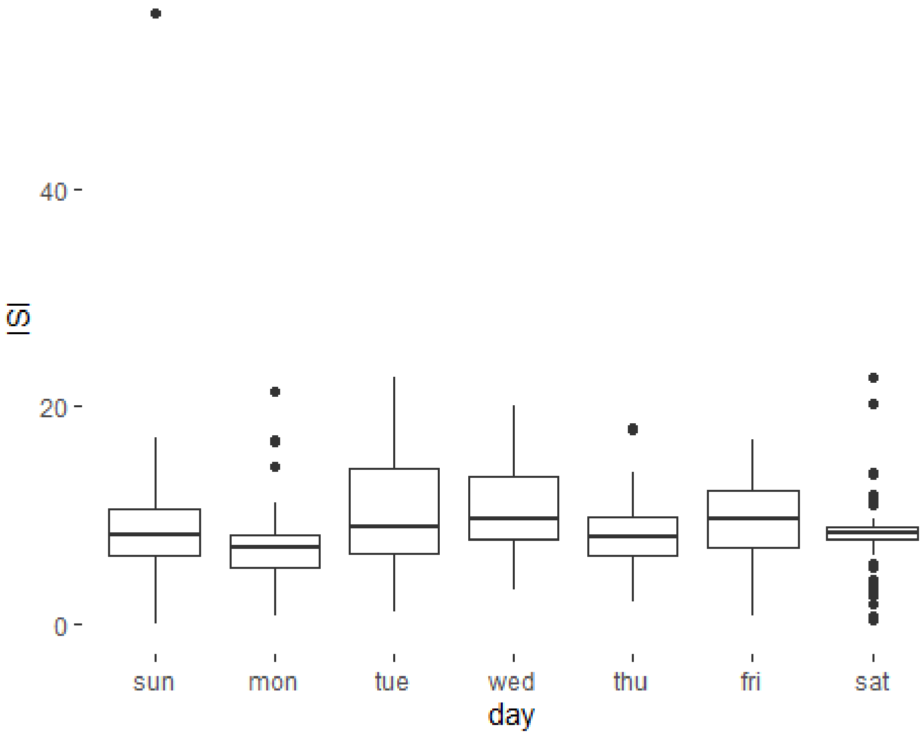
```
##
## [[2]]
```


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

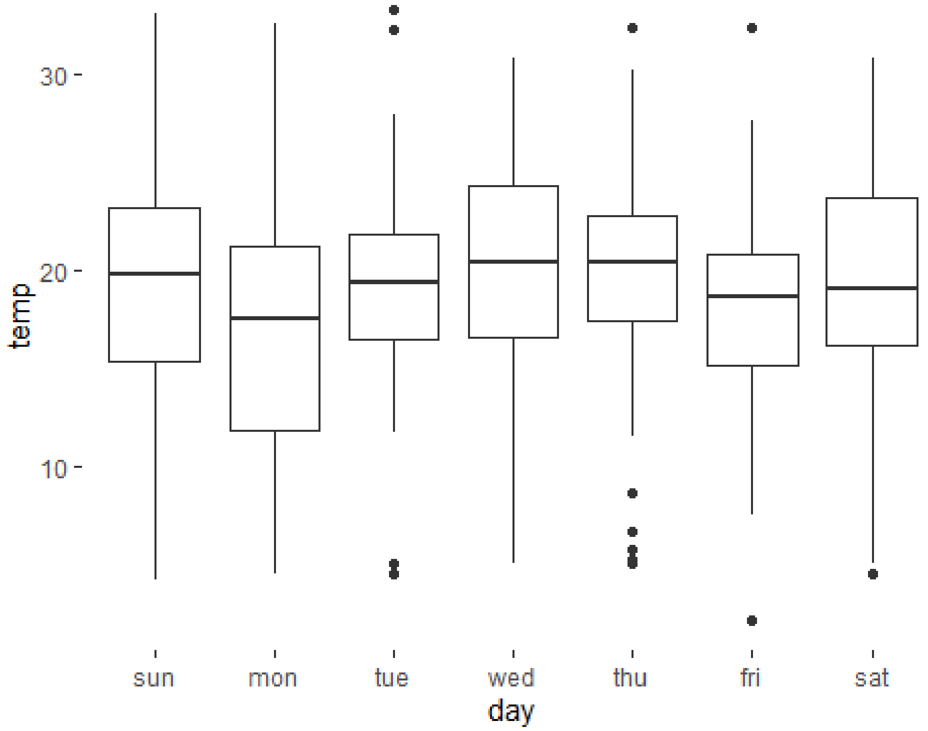
```
##
## [[3]]
```


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

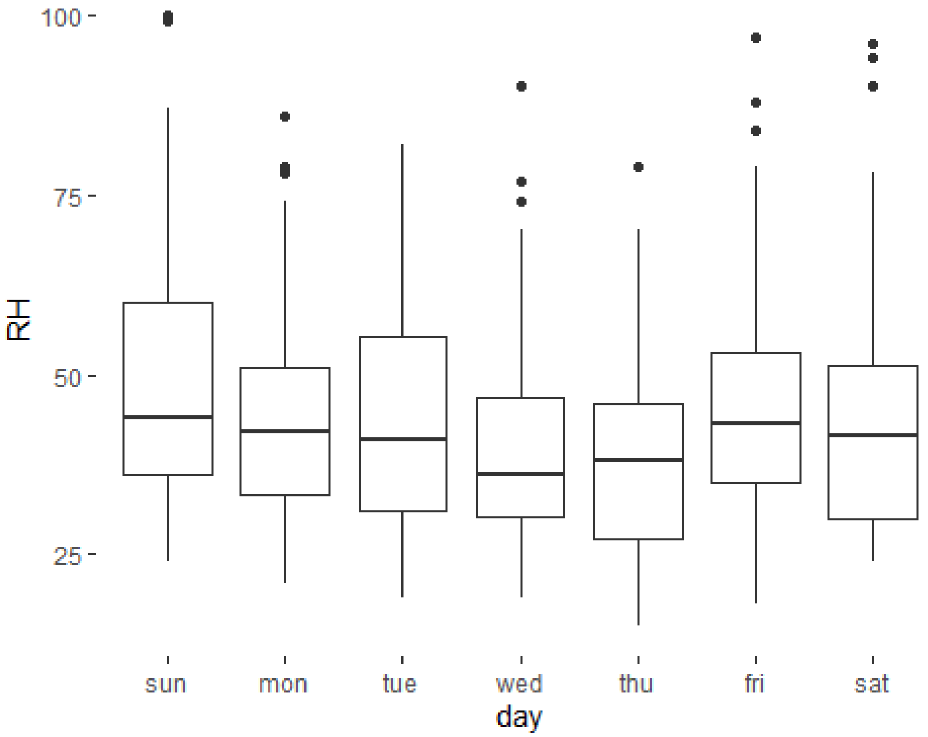
```
##
## [[4]]
```



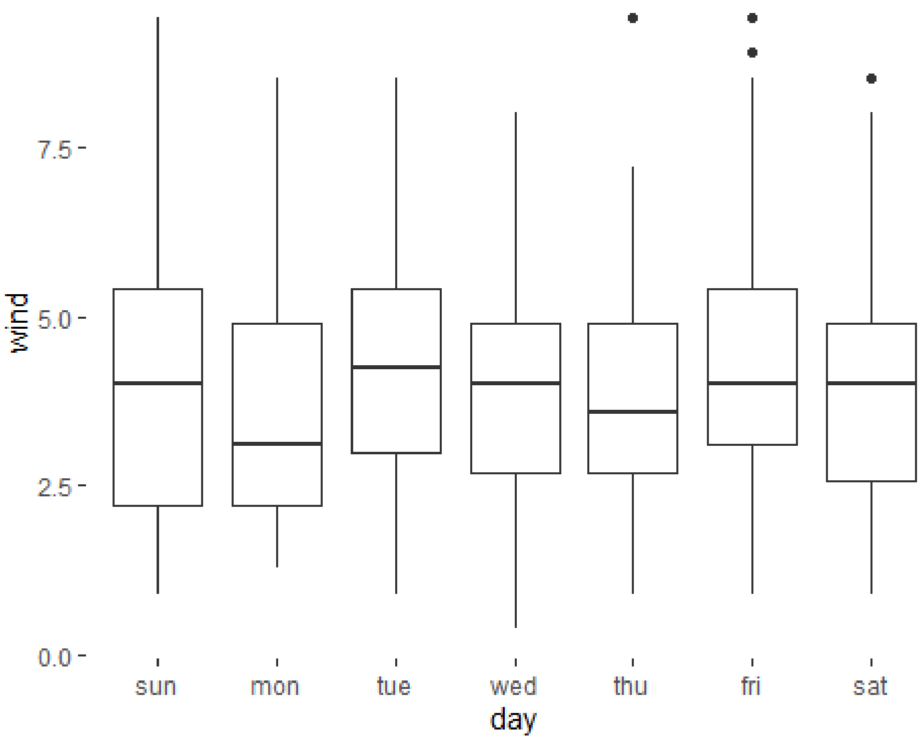
```
##
## [[5]]
```


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

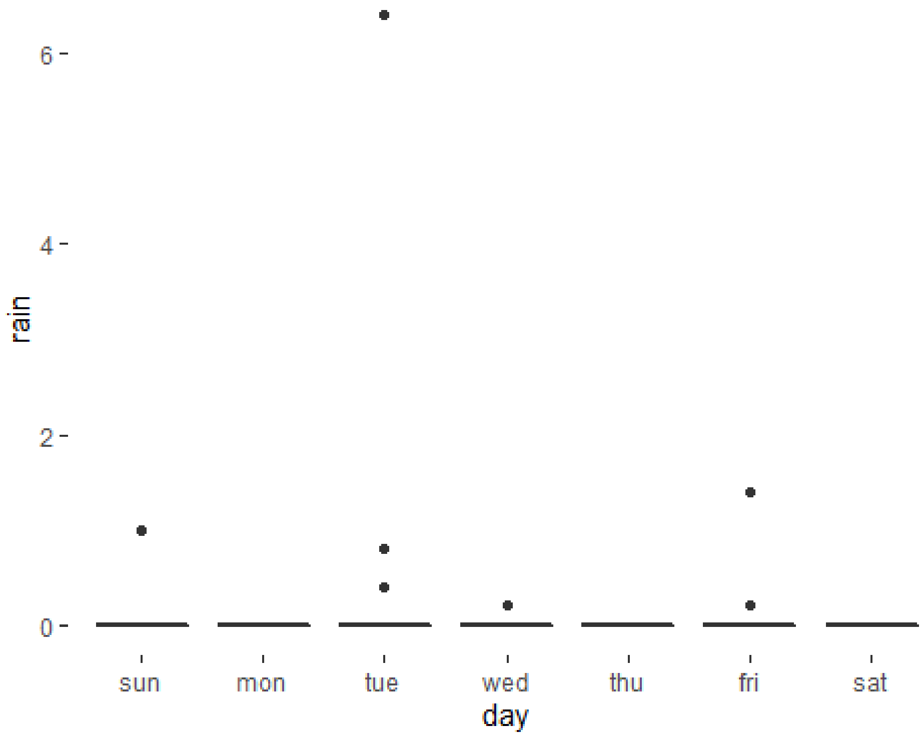
```
##
## [[6]]
```


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

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



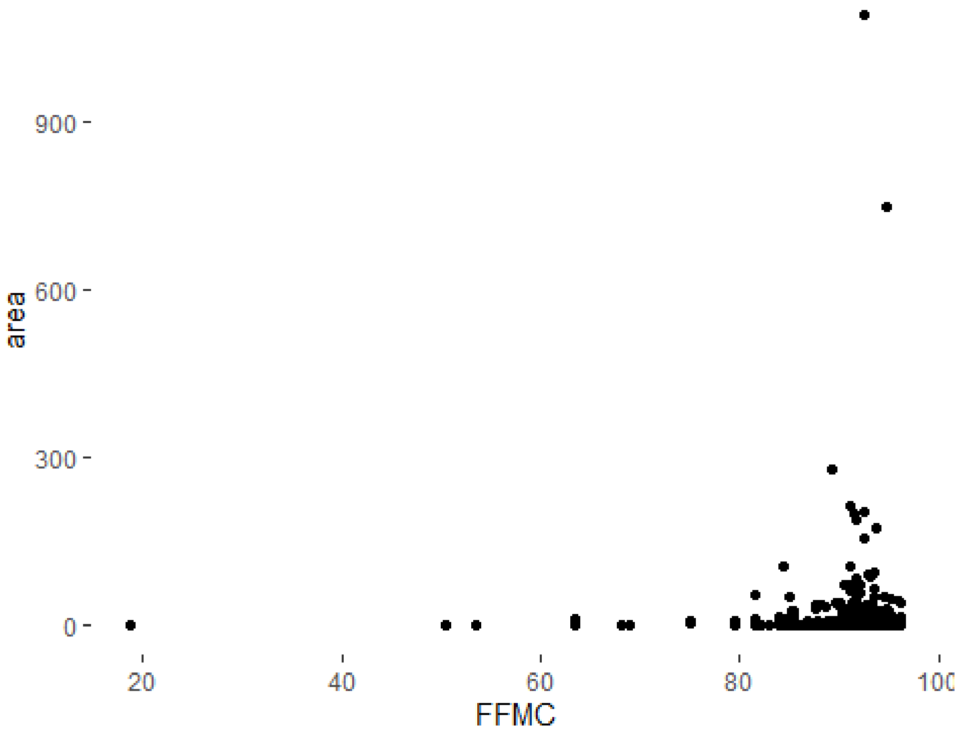
```
##
## [[8]]
```



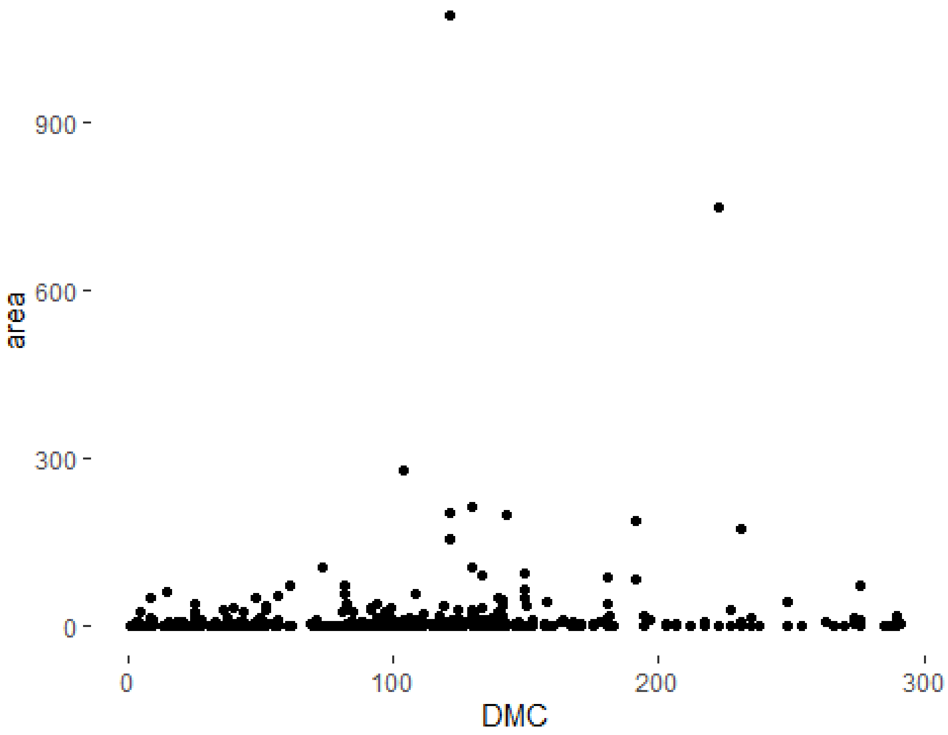
```
create_scatterplots = function(x, y) {
  ggplot(data = forest_fires) +
    aes_string(x = x, y = y) +
    geom_point() +
    theme(panel.background = element_rect(fill = "white"))
}

x_var_scatter <- names(forest_fires)[5:12]
y_var_scatter <- names(forest_fires)[13]

scatters <- map2(x_var_scatter, y_var_scatter, create_scatterplots)
```

Scatter Plots

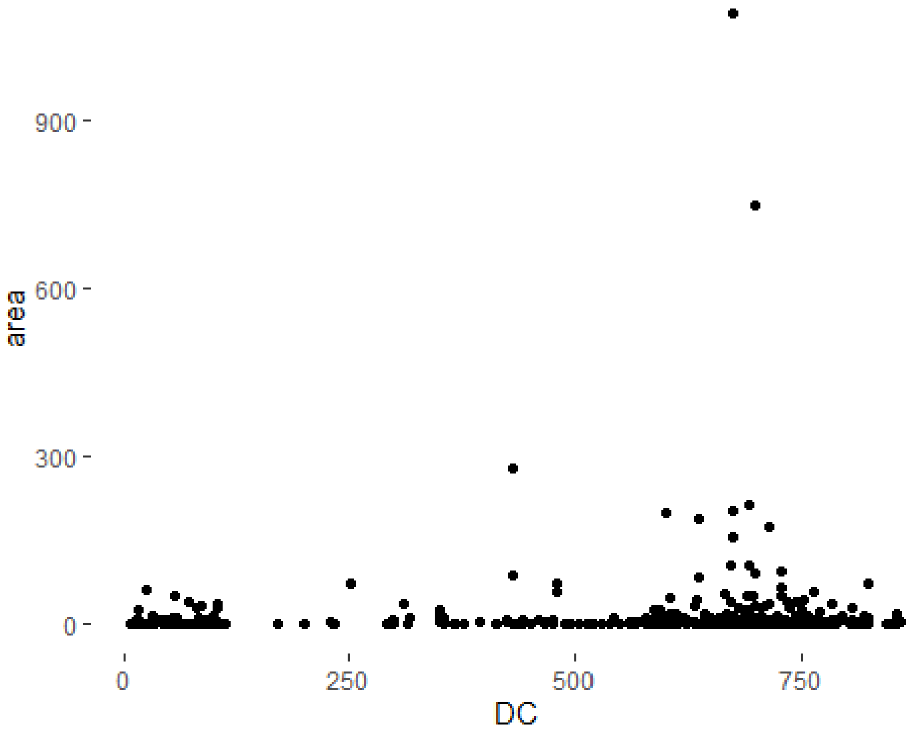
```
scatters

## [[1]]
```


900
600
300
0
area
20
40
60
80
100
FFMC

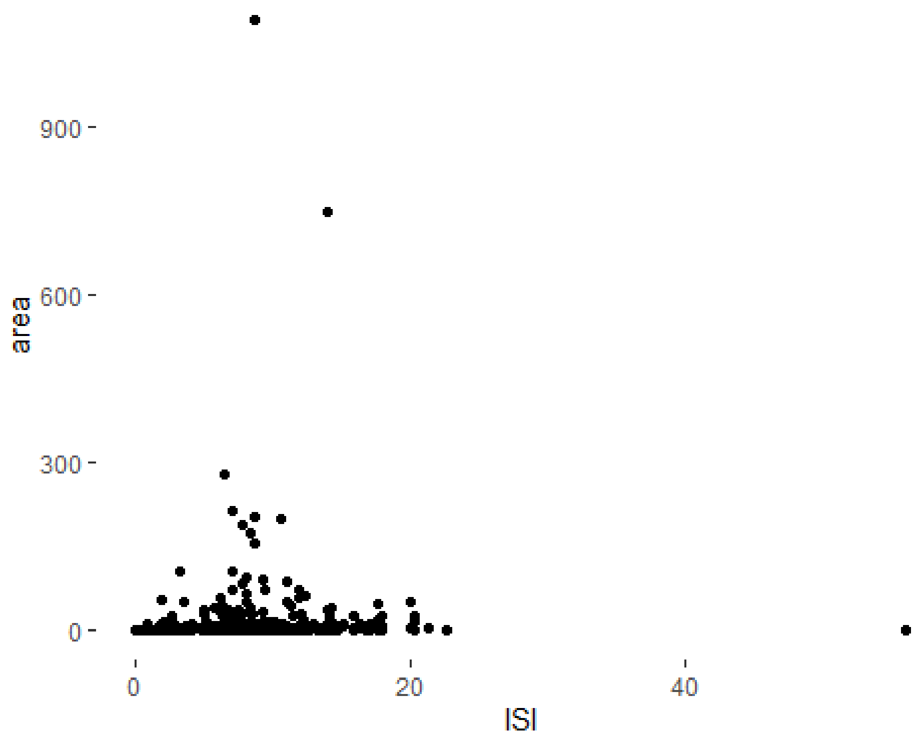
```
##
## [[2]]
```


900
600
300
0
area
0
100
200
300
DMC

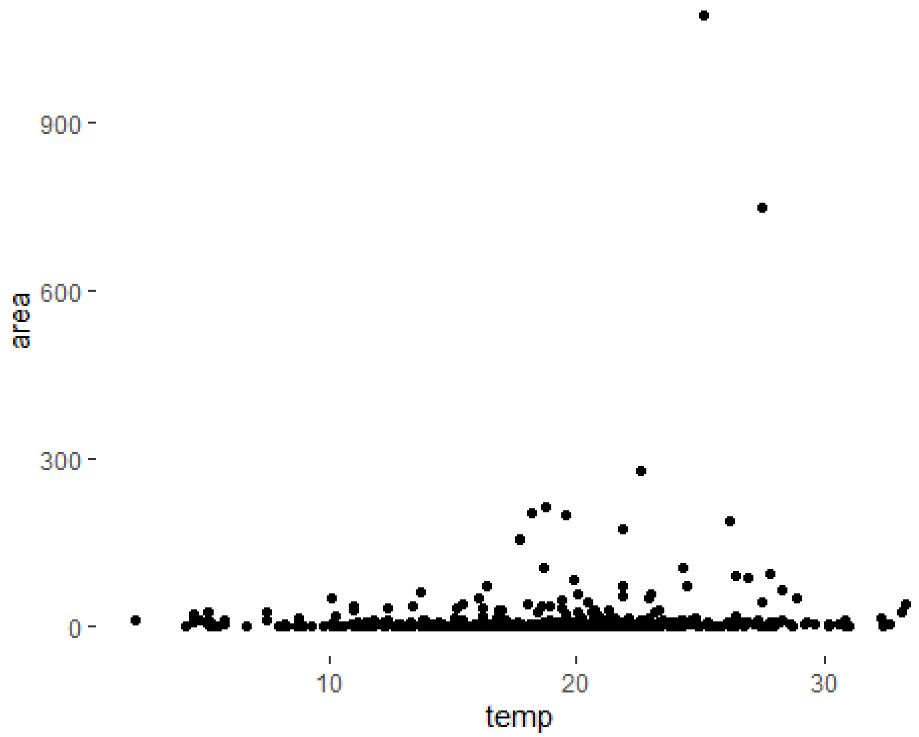
```
##
## [[3]]
```



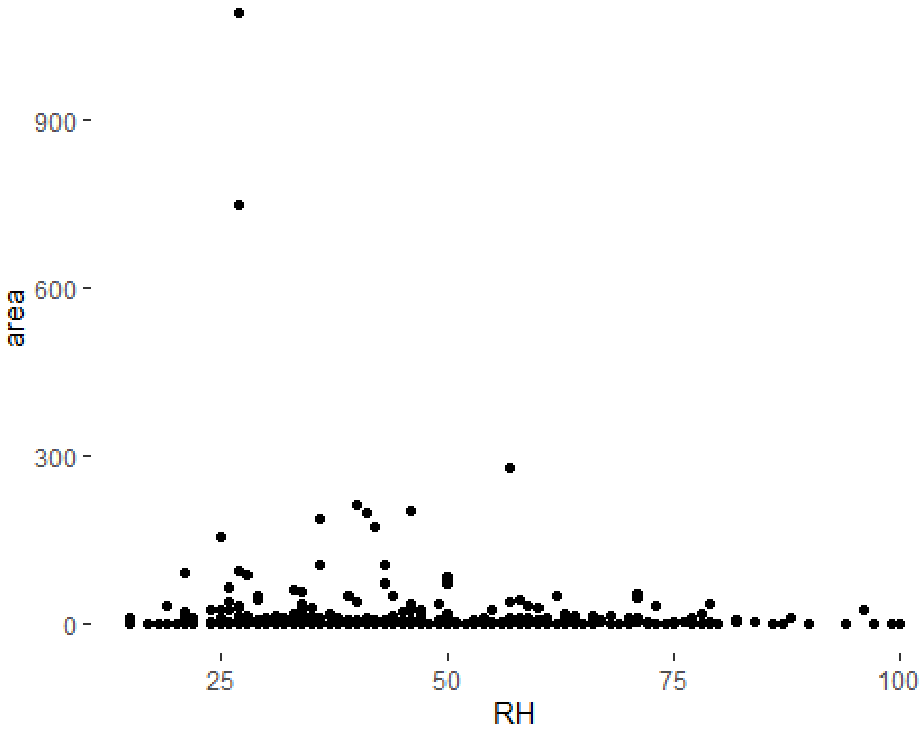
```
##
## [[4]]
```



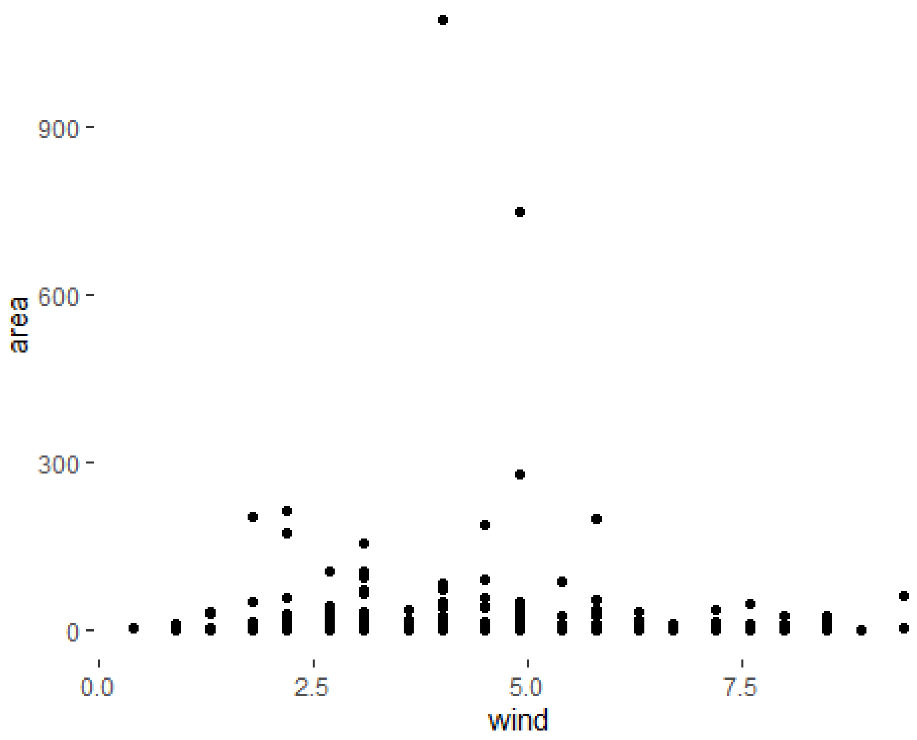
```
##
## [[5]]
```



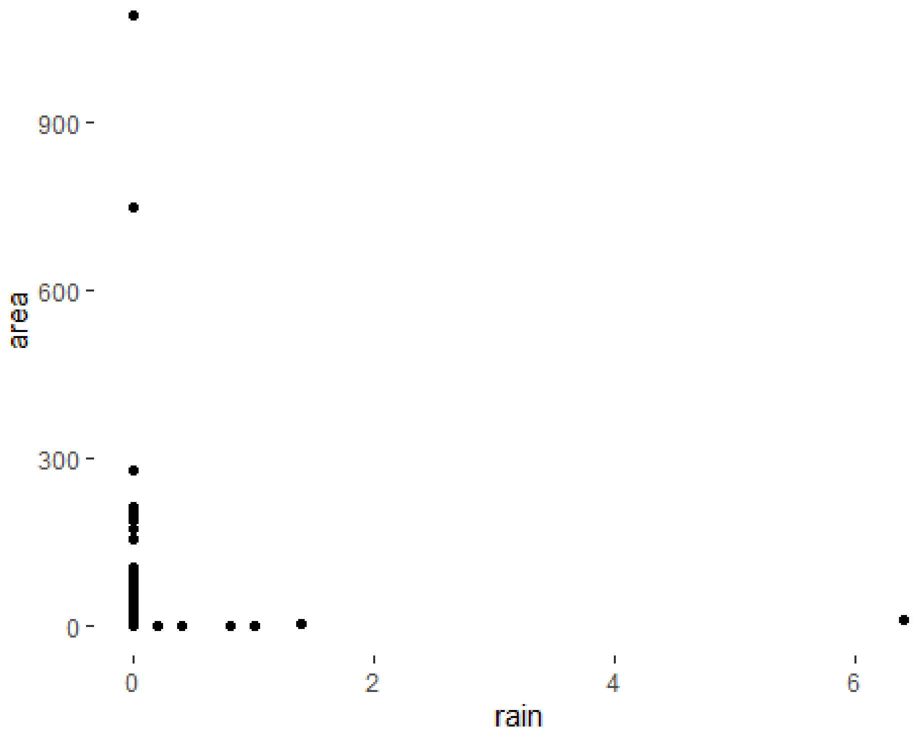
```
##
## [[6]]
```



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



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



There are a few points representing very large values of area, and many points representing values of area that are zero or close to zero. As a result, most points are clustered around the bottom of the plots.