Analyzing Forest Fires

This is a difficult regression task, where the aim is to predict the burned area of forest fires, in the northeast region of Portugal, by using meteorological and other data. The time period is from the entire year of 2007 and the recorded variables are as follow:

- **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

```r
require(tidyverse)
```

```
## Loading required package: tidyverse

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
require(ggplot2)

forestfires <- read.csv("forestfires.csv")
head(forestfires, 50)
```

```
##    X Y month day FFMC   DMC    DC  ISI temp RH wind rain area
## 1  7 5   mar fri 86.2  26.2  94.3  5.1  8.2 51  6.7  0.0    0
## 2  7 4   oct tue 90.6  35.4 669.1  6.7 18.0 33  0.9  0.0    0
## 3  7 4   oct sat 90.6  43.7 686.9  6.7 14.6 33  1.3  0.0    0
## 4  8 6   mar fri 91.7  33.3  77.5  9.0  8.3 97  4.0  0.2    0
## 5  8 6   mar sun 89.3  51.3 102.2  9.6 11.4 99  1.8  0.0    0
## 6  8 6   aug sun 92.3  85.3 488.0 14.7 22.2 29  5.4  0.0    0
## 7  8 6   aug mon 92.3  88.9 495.6  8.5 24.1 27  3.1  0.0    0
## 8  8 6   aug mon 91.5 145.4 608.2 10.7  8.0 86  2.2  0.0    0
## 9  8 6   sep tue 91.0 129.5 692.6  7.0 13.1 63  5.4  0.0    0
## 10 7 5   sep sat 92.5  88.0 698.6  7.1 22.8 40  4.0  0.0    0
## 11 7 5   sep sat 92.5  88.0 698.6  7.1 17.8 51  7.2  0.0    0
## 12 7 5   sep sat 92.8  73.2 713.0 22.6 19.3 38  4.0  0.0    0
## 13 6 5   aug fri 63.5  70.8 665.3  0.8 17.0 72  6.7  0.0    0
## 14 6 5   sep mon 90.9 126.5 686.5  7.0 21.3 42  2.2  0.0    0
## 15 6 5   sep wed 92.9 133.3 699.6  9.2 26.4 21  4.5  0.0    0
## 16 6 5   sep fri 93.3 141.2 713.9 13.9 22.9 44  5.4  0.0    0
## 17 5 5   mar sat 91.7  35.8  80.8  7.8 15.1 27  5.4  0.0    0
## 18 8 5   oct mon 84.9  32.8 664.2  3.0 16.7 47  4.9  0.0    0
## 19 6 4   mar wed 89.2  27.9  70.8  6.3 15.9 35  4.0  0.0    0
## 20 6 4   apr sat 86.3  27.4  97.1  5.1  9.3 44  4.5  0.0    0
## 21 6 4   sep tue 91.0 129.5 692.6  7.0 18.3 40  2.7  0.0    0
## 22 5 4   sep mon 91.8  78.5 724.3  9.2 19.1 38  2.7  0.0    0
## 23 7 4   jun sun 94.3  96.3 200.0 56.1 21.0 44  4.5  0.0    0
## 24 7 4   aug sat 90.2 110.9 537.4  6.2 19.5 43  5.8  0.0    0
## 25 7 4   aug sat 93.5 139.4 594.2 20.3 23.7 32  5.8  0.0    0
## 26 7 4   aug sun 91.4 142.4 601.4 10.6 16.3 60  5.4  0.0    0
## 27 7 4   sep fri 92.4 117.9 668.0 12.2 19.0 34  5.8  0.0    0
## 28 7 4   sep mon 90.9 126.5 686.5  7.0 19.4 48  1.3  0.0    0
## 29 6 3   sep sat 93.4 145.4 721.4  8.1 30.2 24  2.7  0.0    0
## 30 6 3   sep sun 93.5 149.3 728.6  8.1 22.8 39  3.6  0.0    0
## 31 6 3   sep fri 94.3  85.1 692.3 15.9 25.4 24  3.6  0.0    0
## 32 6 3   sep mon 88.6  91.8 709.9  7.1 11.2 78  7.6  0.0    0
## 33 6 3   sep fri 88.6  69.7 706.8  5.8 20.6 37  1.8  0.0    0
## 34 6 3   sep sun 91.7  75.6 718.3  7.8 17.7 39  3.6  0.0    0
## 35 6 3   sep mon 91.8  78.5 724.3  9.2 21.2 32  2.7  0.0    0
## 36 6 3   sep tue 90.3  80.7 730.2  6.3 18.2 62  4.5  0.0    0
## 37 6 3   oct tue 90.6  35.4 669.1  6.7 21.7 24  4.5  0.0    0
## 38 7 4   oct fri 90.0  41.5 682.6  8.7 11.3 60  5.4  0.0    0
## 39 7 3   oct sat 90.6  43.7 686.9  6.7 17.8 27  4.0  0.0    0
## 40 4 4   mar tue 88.1  25.7  67.6  3.8 14.1 43  2.7  0.0    0
## 41 4 4   jul tue 79.5  60.6 366.7  1.5 23.3 37  3.1  0.0    0
## 42 4 4   aug sat 90.2  96.9 624.2  8.9 18.4 42  6.7  0.0    0
## 43 4 4   aug tue 94.8 108.3 647.1 17.0 16.6 54  5.4  0.0    0
## 44 4 4   sep sat 92.5  88.0 698.6  7.1 19.6 48  2.7  0.0    0
## 45 4 4   sep wed 90.1  82.9 735.7  6.2 12.9 74  4.9  0.0    0
## 46 5 6   sep wed 94.3  85.1 692.3 15.9 25.9 24  4.0  0.0    0
## 47 5 6   sep mon 90.9 126.5 686.5  7.0 14.7 70  3.6  0.0    0
## 48 6 6   jul mon 94.2  62.3 442.9 11.0 23.0 36  3.1  0.0    0
## 49 4 4   mar mon 87.2  23.9  64.7  4.1 11.8 35  1.8  0.0    0
## 50 4 4   mar mon 87.6  52.2 103.8  5.0 11.0 46  5.8  0.0    0
```

Loading tidyverse and ggplot2 packages and importing the forest fire csv file into a dataframe. We use head to look at the first several rows in order to get a sense of the structure and formatting of the data.

```
forestfires %>% pull(month) %>% unique()
```

```
##  [1] "mar" "oct" "aug" "sep" "apr" "jun" "jul" "feb" "jan" "dec" "may" "nov"
```

```
forestfires %>% pull(day) %>% unique()
```

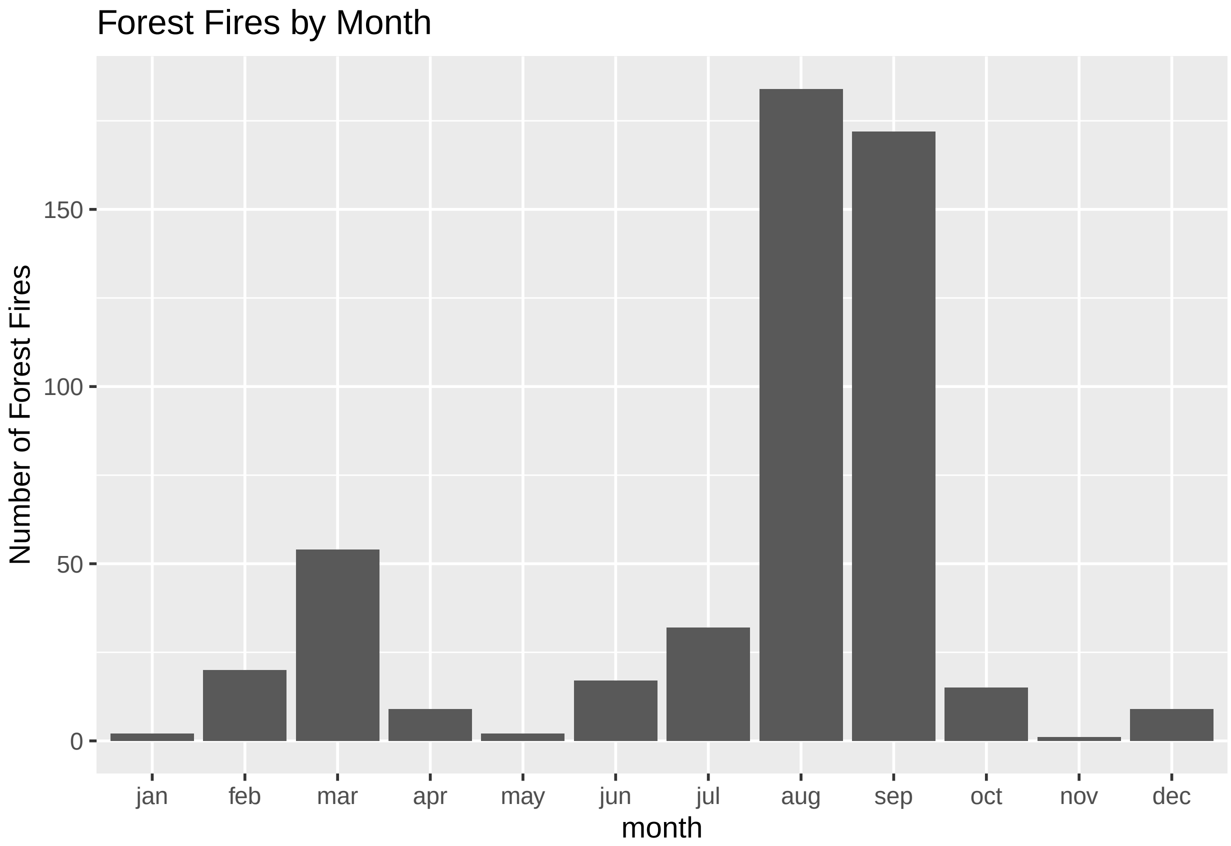
```
## [1] "fri" "tue" "sat" "sun" "mon" "wed" "thu"
```

Taking a look at the month and day variables we can see how they are abbreviated as well as how they are ordered.

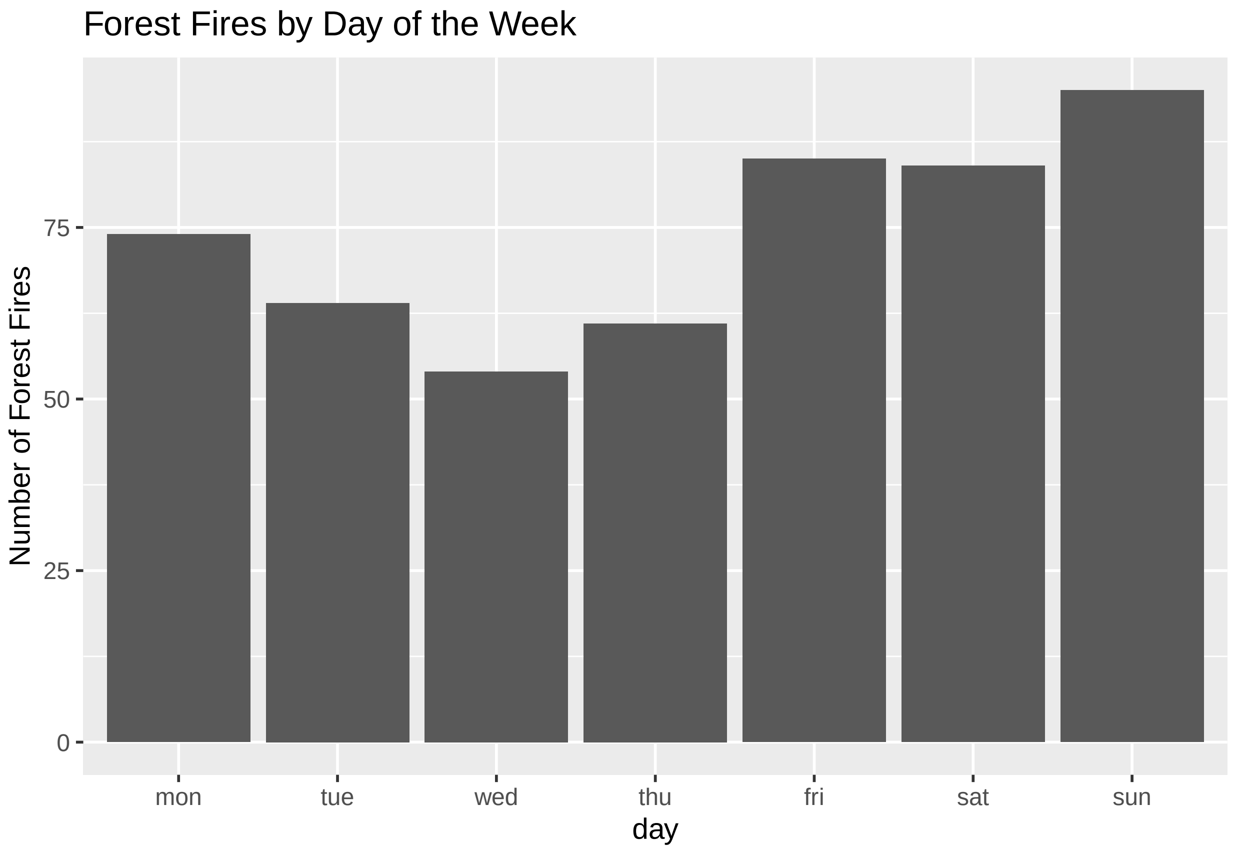
```
forestfires <- forestfires %>%
    mutate(
      month = factor(month, levels = c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "
      day = factor(day, levels = c("mon", "tue", "wed", "thu", "fri", "sat", "sun"))
    )
```

We can change the order of the month and day variables as well as changing them into categorical variables.

```
numff.month <- forestfires %>%
  group_by(month) %>%
  summarize(
   n = n()
  )
numff.day <- forestfires %>%
  group_by(day) %>%
  summarize(
   n = n()
  )
numff.month %>%
      ggplot(aes(x = month, y = n)) +
      geom_col() +
      labs(
        title = "Forest Fires by Month",
        y = "Number of Forest Fires")
```



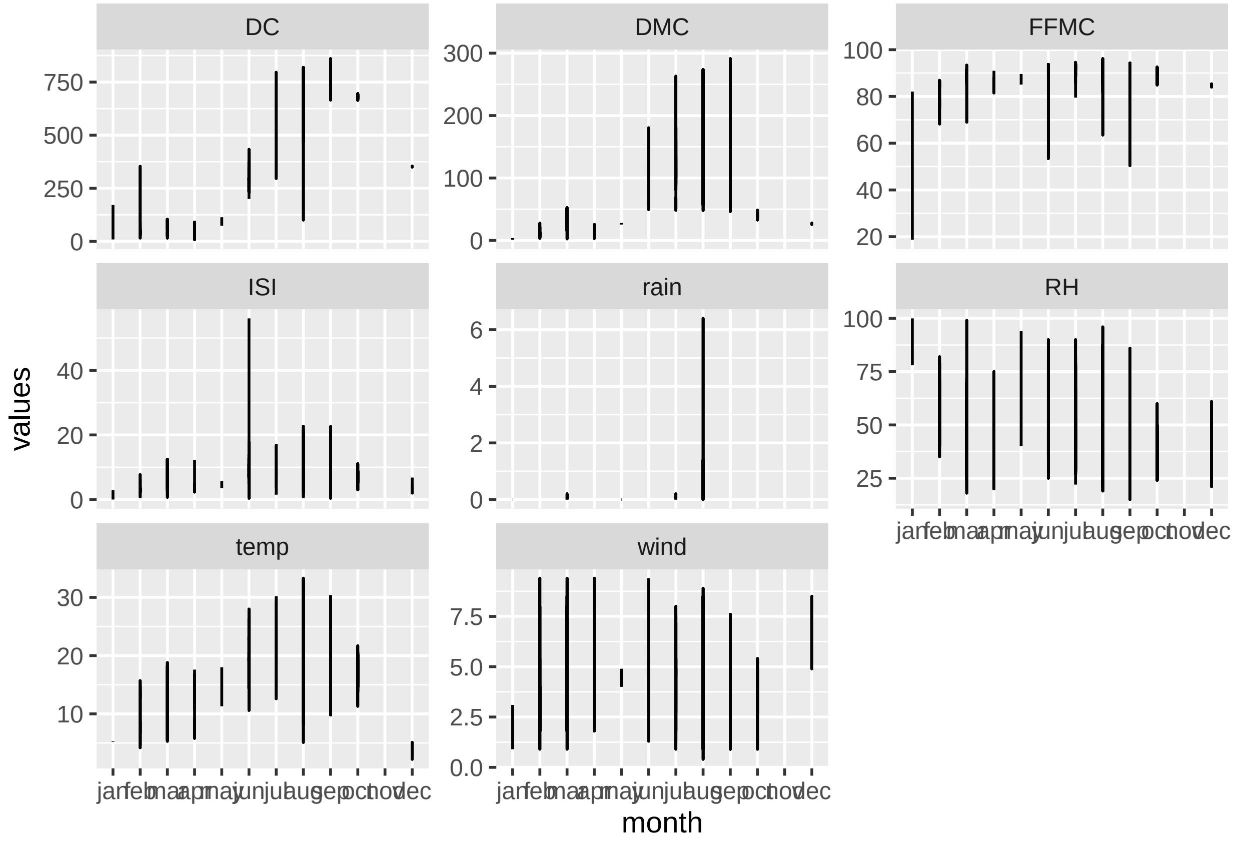
```
numff.day %>%
      ggplot(aes(x = day, y = n)) +
      geom_col() +
      labs(
        title = "Forest Fires by Day of the Week",
        y = "Number of Forest Fires")
```



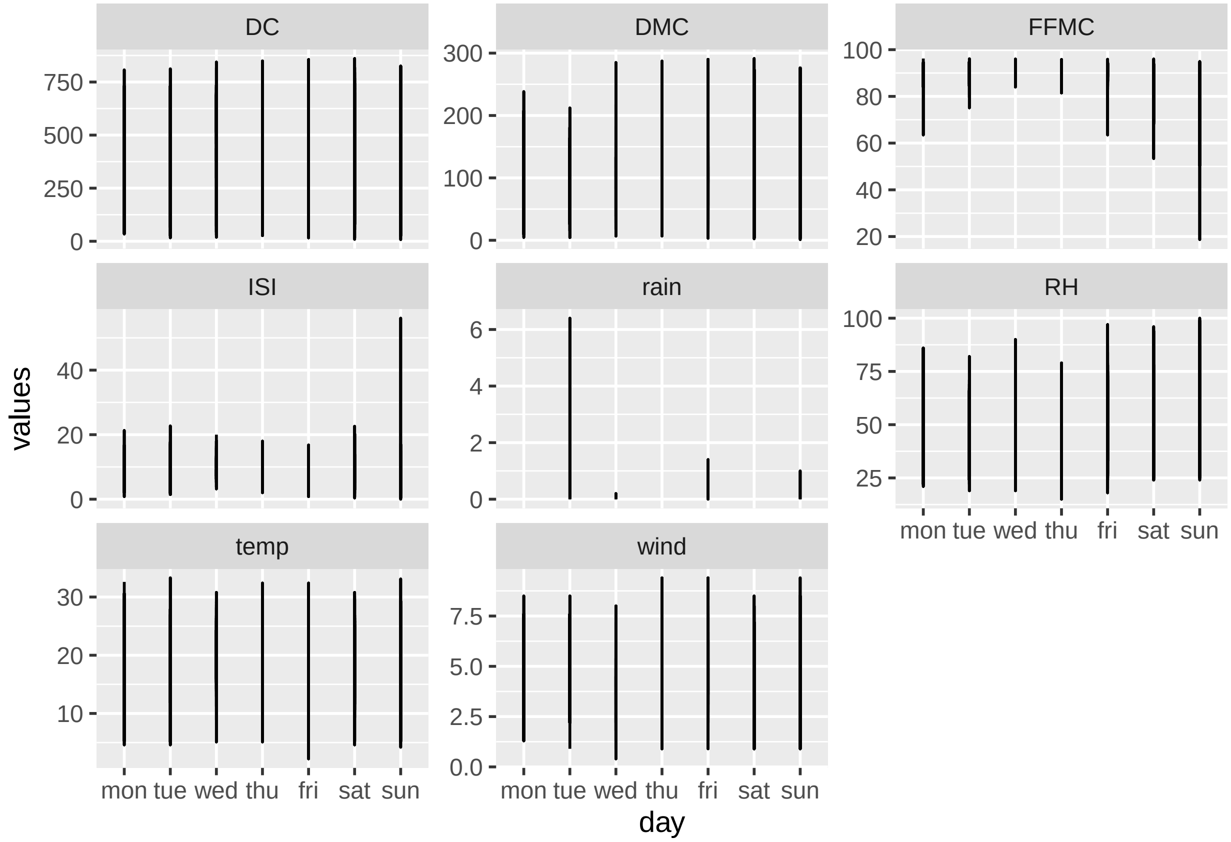
We group the data by both month and day, then summarize the dataframes by counting the number of observations for each month/day and load those summaries into a new data frame, respectively. Thanks to the categoricalization of the variables 'month' and 'day' we are able to have the designated order in our graphs.

Quick takeaways; Peak forest fire season is within the months of the August and September. Additionally, forest fires tend to occur during the weekend i.e. Friday-Sunday.

```
forestfires_long <- forestfires %>%
  pivot_longer(
    cols = c(FFMC, DMC, DC, ISI, temp, RH, wind, rain),
    names_to = "variables",
    values_to = "values"
  )
forestfires_long %>%
  ggplot(aes(x=month, y=values)) +
  geom_line() +
  facet_wrap(vars(variables),
  scales = "free_y")
```

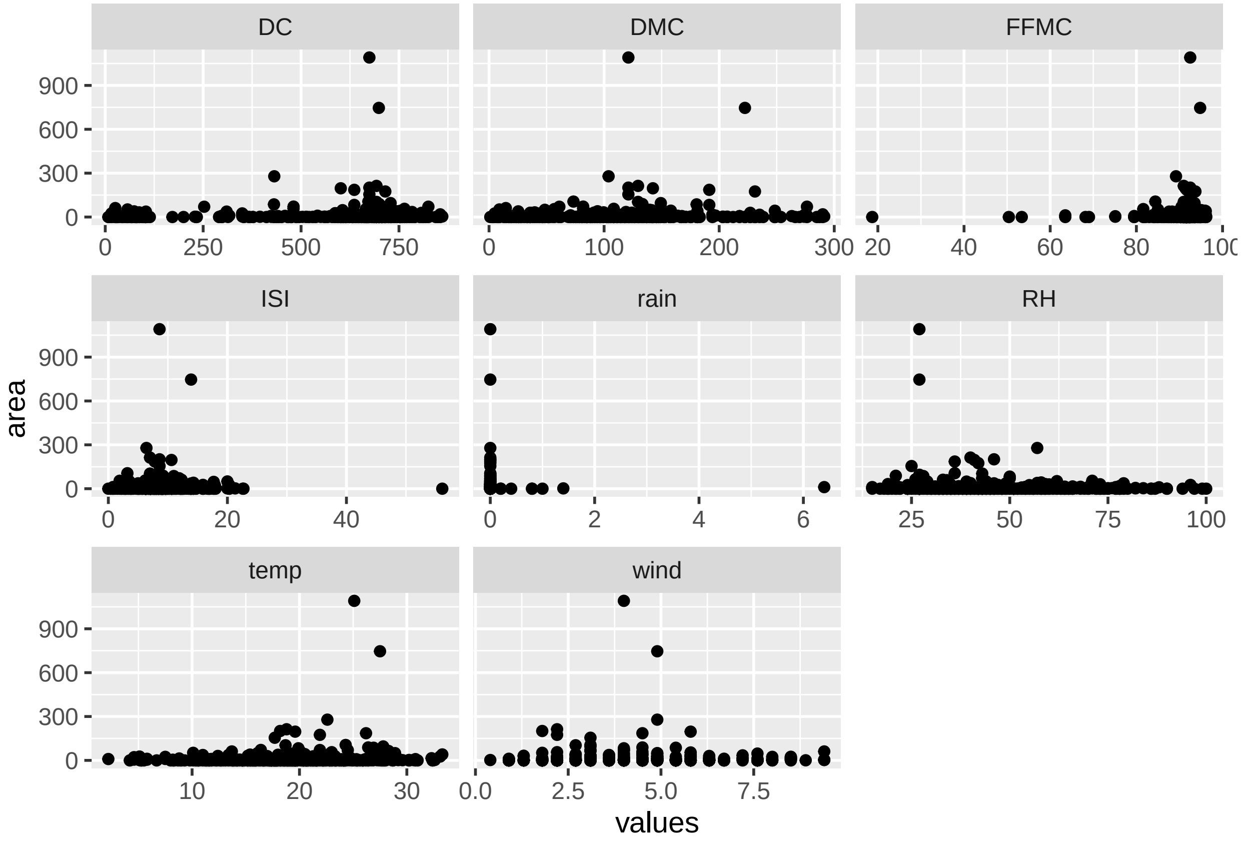


```
forestfires_long %>%
  ggplot(aes(x=day, y=values)) +
  geom_line() +
  facet_wrap(vars(variables),
  scales = "free_y")
```



Values for the following indices have positive correlation with the number of forest fires by month; DC, DMC, and Temperature. This shows that Temperature, Rain, and Relative humidity are significantly contributing factors as they factor into the calculation for DC, and DMC.

```
forestfires_long %>%
  ggplot(aes(x=values, y=area)) +
  geom_point() +
  facet_wrap(vars(variables) ,
  scales = "free_x")
```

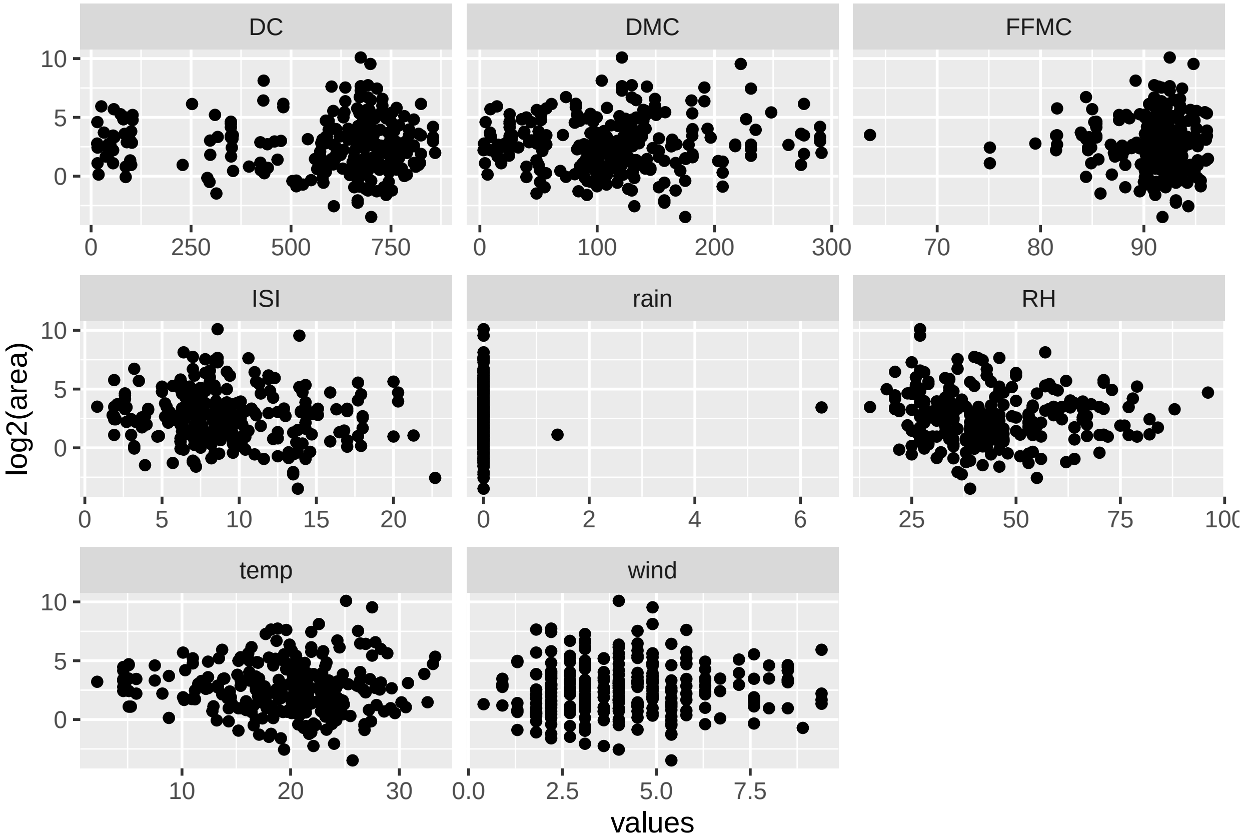


```r
cor_ff <- cor(forestfires[13], forestfires[5:12], method = "spearman")
head(cor_ff)
```

```
##            FFMC        DMC         DC         ISI       temp          RH
## area 0.02530046 0.07191967 0.06163303 0.01249593 0.07869596 -0.02422121
##            wind        rain
## area 0.05319584 -0.06407348
```

Higher values of FFMC, DC, and Temperature show more correlation with higher values of forest fire area (intensity). Running a correlative comparison shows temperature as having the highest p-value.

```r
forestfires_long_filter <- forestfires_long %>%
  filter(area > 0)
forestfires_long_filter %>%
  ggplot(aes(x=values, y=log2(area))) +
  geom_point() +
  facet_wrap(vars(variables) ,
  scales = "free_x")
```


DC
DMC
FFMC
ISI
rain
RH
temp
wind
log2(area)
values
10
5
0
0
250
500
750
0
100
200
300
70
80
90
0
5
10
15
20
0
2
4
6
25
50
75
100
10
20
30
0.0
2.5
5.0
7.5