# Guided Project: Analyzing Forest Fire Data

Husen Wahyu

6/7/2021

## Page 1 Load the Dataset

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

```
##
## -- 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()
## )
```

## Page 2 Getting to know the data about forest fire

```r
# What does a single row represent?
colnames(forestfires)
```

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

```r
# With what I know about fires, how might each of the variables related to fires themselves?
```

- 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

## Page 3 Date Preprocessing

```r
# Convert the month variable into a categorical variable, and make sure that the months in the data are
forestfires %>% pull(month) %>% unique()
```

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

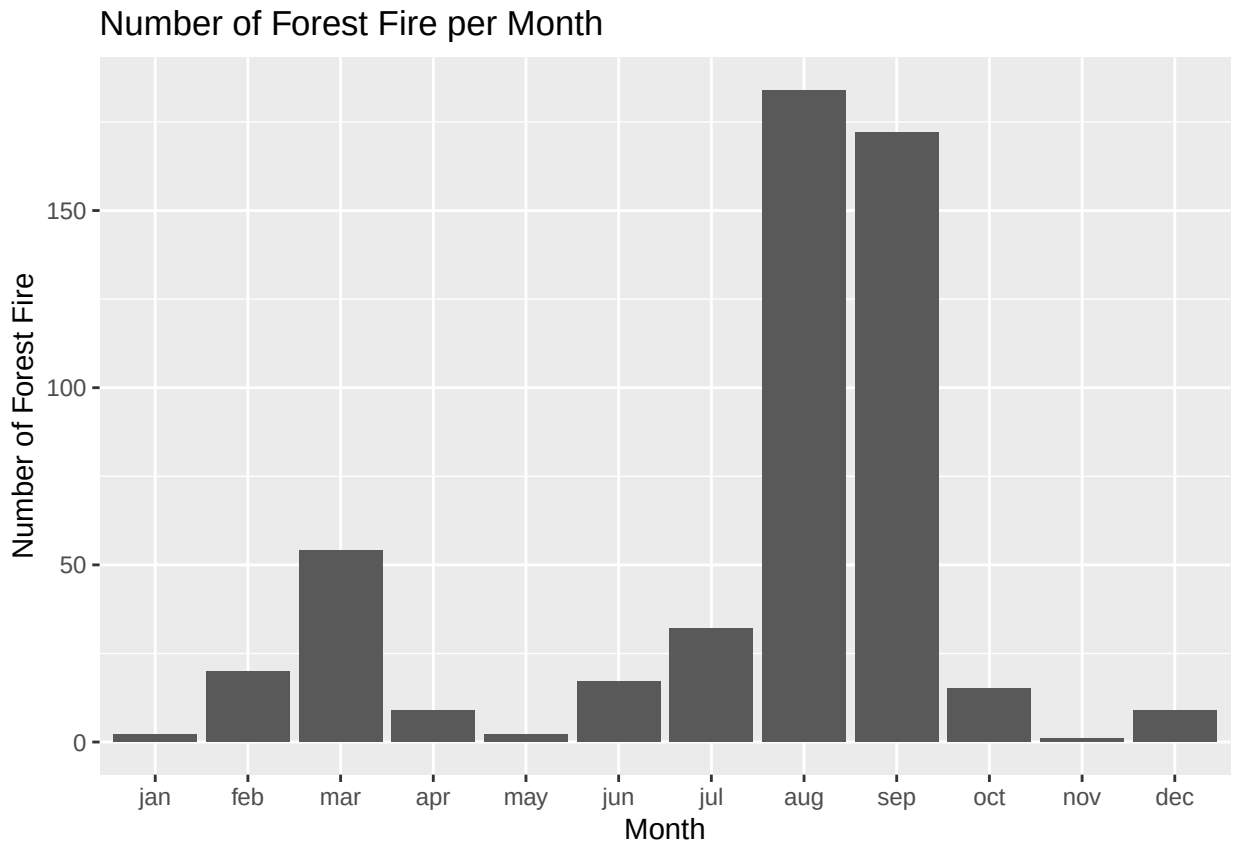
```r
month_order <- c("jan", "feb", "mar", "apr", "may", "jun",
                 "jul", "aug", "sep", "oct", "nov", "dec")

day_order <- c("sun", "mon", "tue", "wed", "thu", "fri", "sat")
forestfires %>% pull(day) %>% unique()
```

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

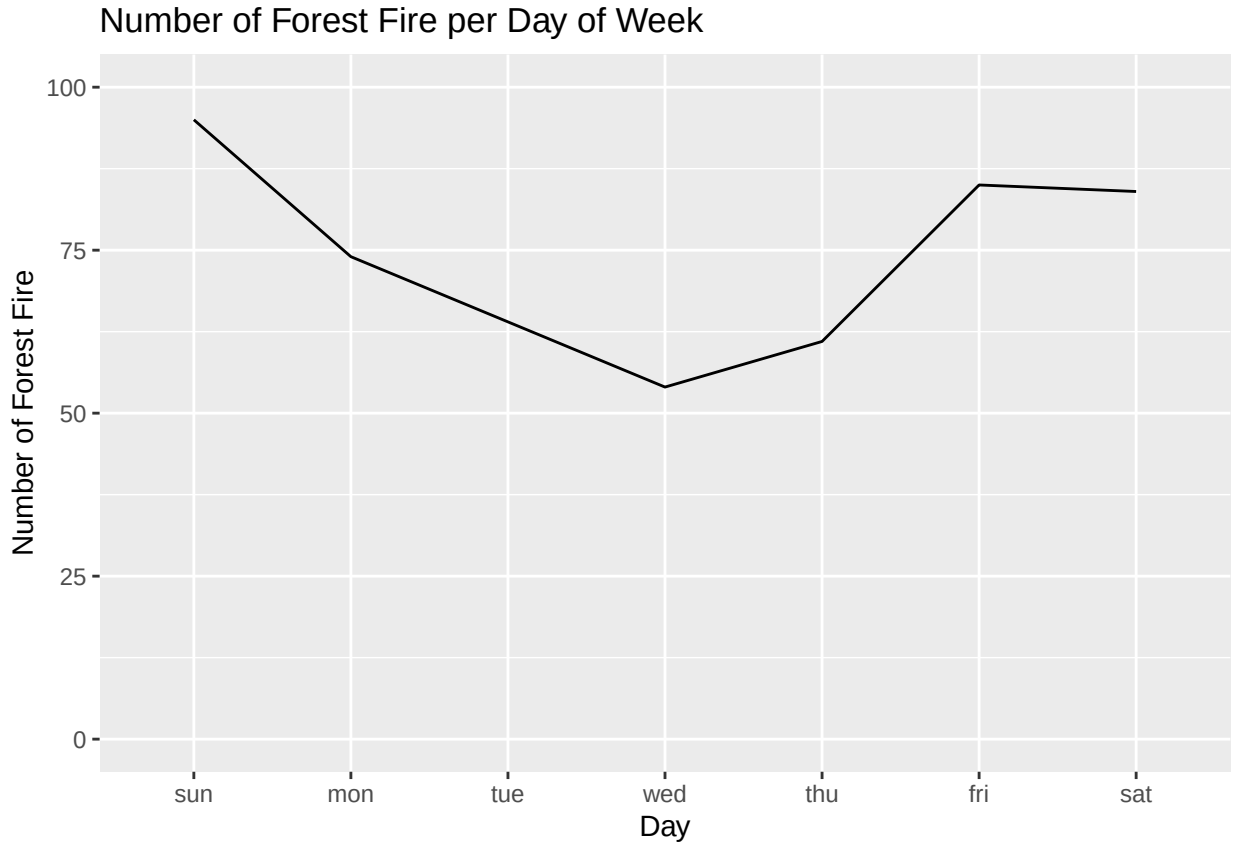
```r
forestfires <- forestfires %>%
  mutate(
    month = factor(month, month_order),
    day = factor(day, day_order)
  )
```

## Page 4 Analyzing and Visualizing the month and day of week the fireforest mostly occurs

```r
# Create a tibble that counts the number of forest fires by month.
forestfires_bymonth <- forestfires %>%
  group_by(month) %>%
    summarise(
      num_of_ff = n()
    )

# Create a tibble that counts the number of forest fires by day of the week.

forestfires_bydow <- forestfires %>%
  group_by(day) %>%
    summarise(
      num_of_ff_dow = n()
    )

# Using each of the tibbles that you created, create a visualization that allows us to answer the quest
forestfires_bymonth %>%
  ggplot(aes(x=month, y= num_of_ff, group=1)) +
  geom_col() +
  labs(
    title = "Number of Forest Fire per Month",
    x = "Month",
    y= "Number of Forest Fire"
  )
```



```r
# For line graphs, the data points must be grouped so that it knows which points to connect. In this ca
forestfires_bydow %>%
  ggplot(aes(x=day, y=num_of_ff_dow, group=1)) +
  geom_line() +
  labs(
    title = "Number of Forest Fire per Day of Week",
    x = "Day",
    y= "Number of Forest Fire"
  ) +
  ylim(0,100) # to set the y axis size on the chart
```




```r
(colnames(forestfires))[c(-1, -2, -3, -4)]
```

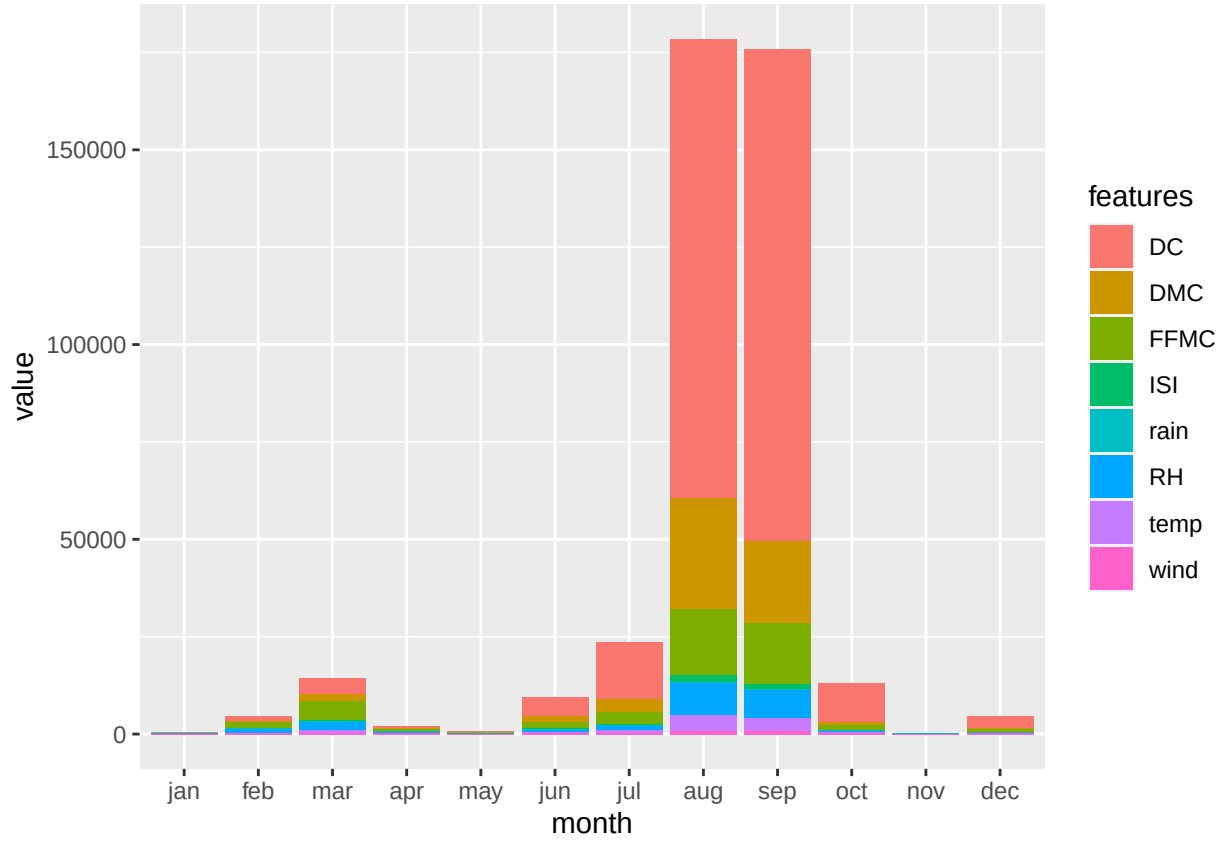
```
## [1] "FFMC" "DMC"  "DC"   "ISI"  "temp" "RH"   "wind" "rain" "area"
```

## Page 5 Finding the relationship of each column along the year

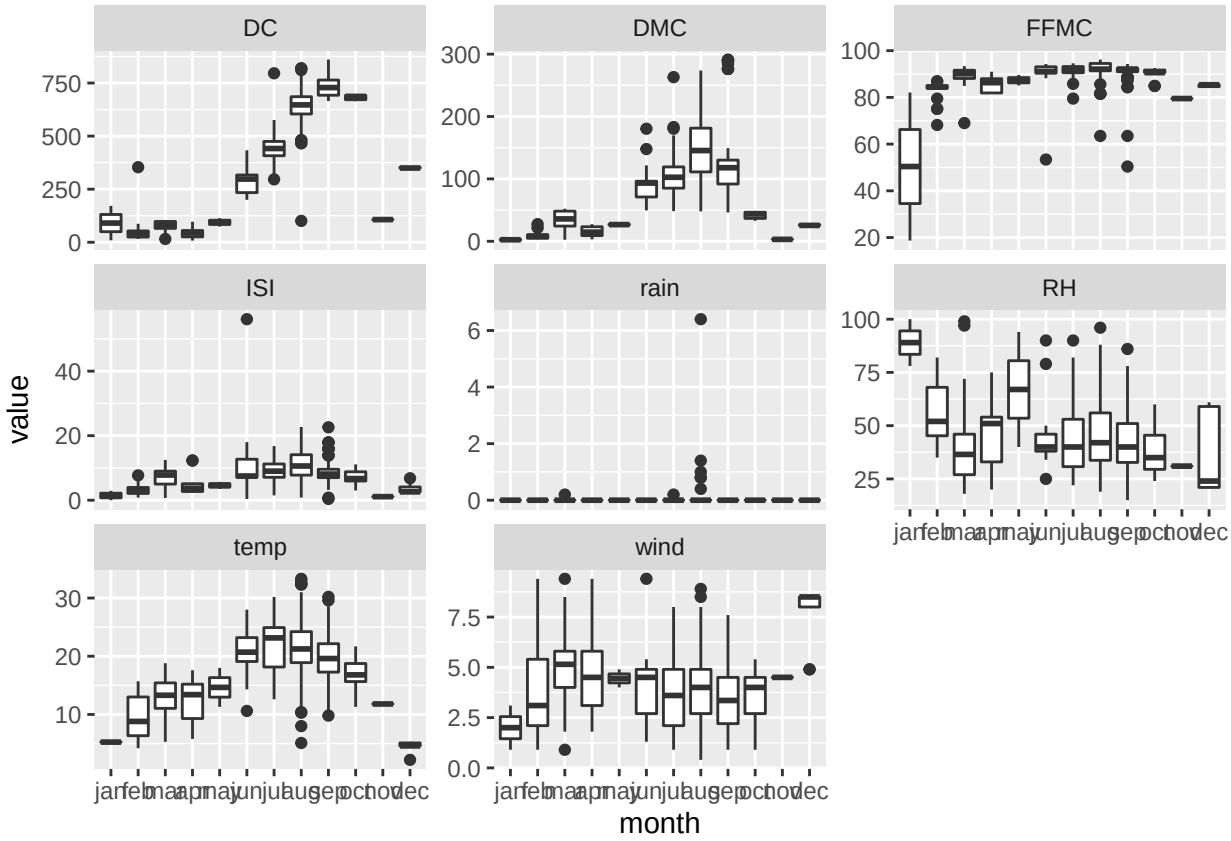
```r
# Create the new dataframe of long version of dataset with the features value in one column (for visual

forest_fires_long <- forestfires %>%
  pivot_longer(
    cols = (colnames(forestfires))[c(-1, -2, -3, -4, -13)],
```

```
    names_to = "features",
    values_to = "value"
  )
```

```
# col plot
forest_fires_long %>%
  ggplot(aes(x=month, y=value, fill=features)) +
  geom_col()
```



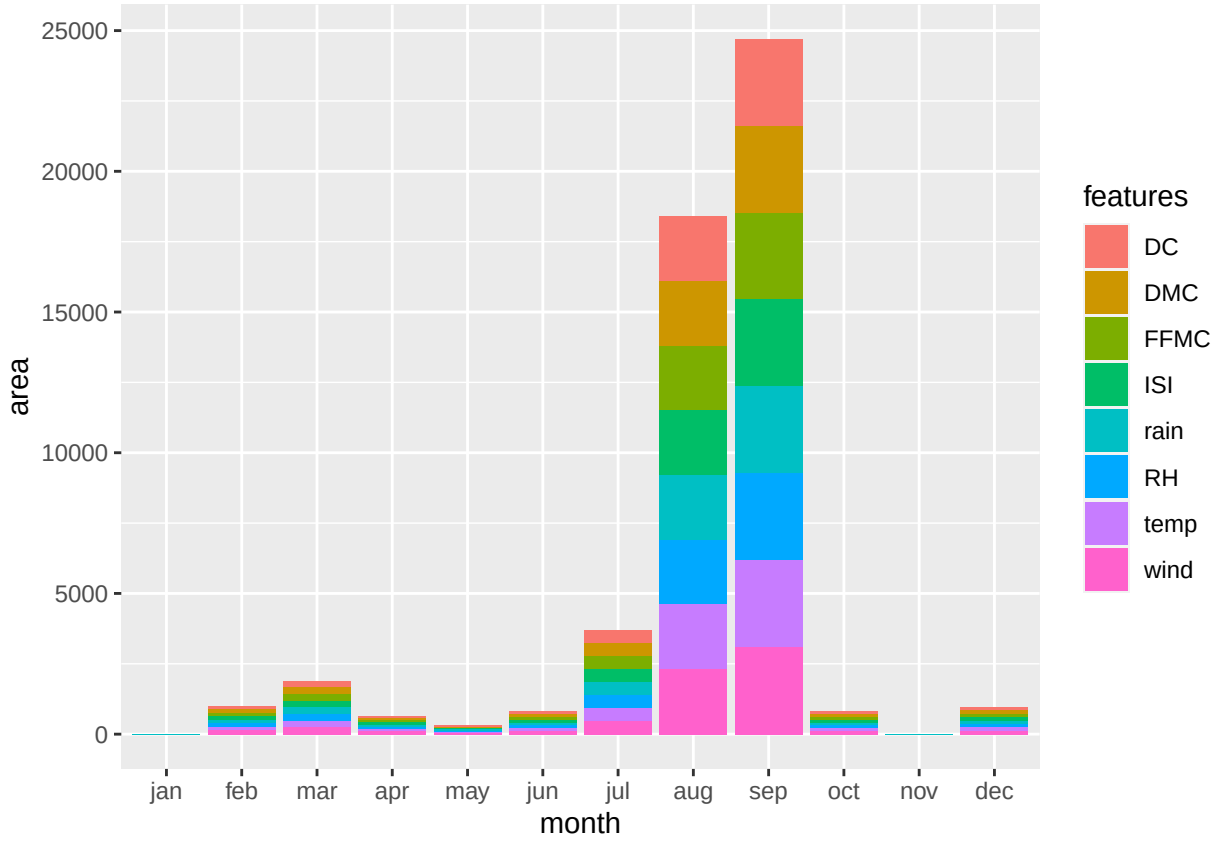

```
# boxplot
forest_fires_long %>%
  ggplot(aes(x=month, y=value)) +
  geom_boxplot()+
  facet_wrap(vars(features), scale="free_y")
```




All variables are vary. The DC, DMC, and temp are having the same pattern with the forest fire pattern

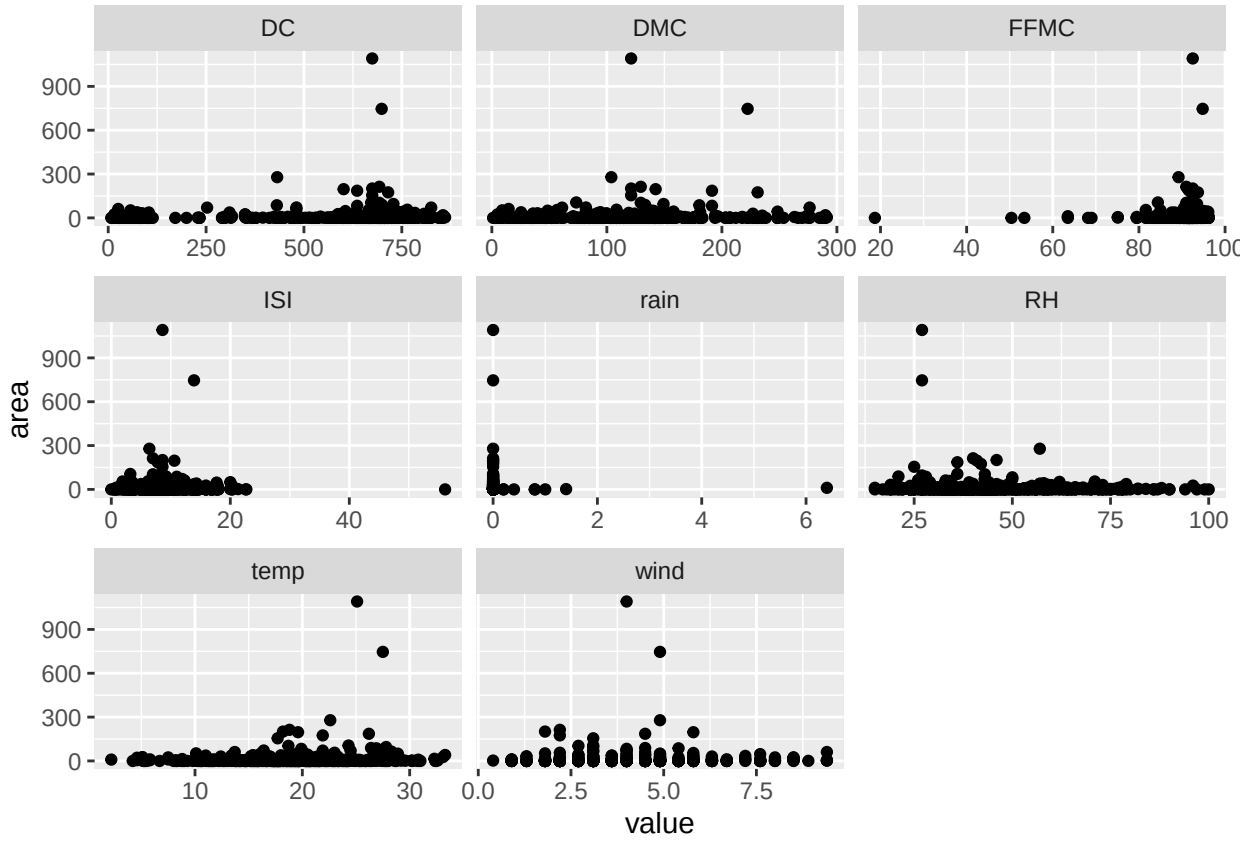
## Page 6 Finding realtionship of each features with area of forest fires

```
forest_fires_long %>%
  ggplot(aes(x=month, y= area, fill=features)) +
  geom_col()
```




The plot above is not showing any insight

```
forest_fires_long %>%
  ggplot(aes(x=value, y=area)) +
  geom_point() +
  facet_wrap(vars(features), scale="free_x")
```

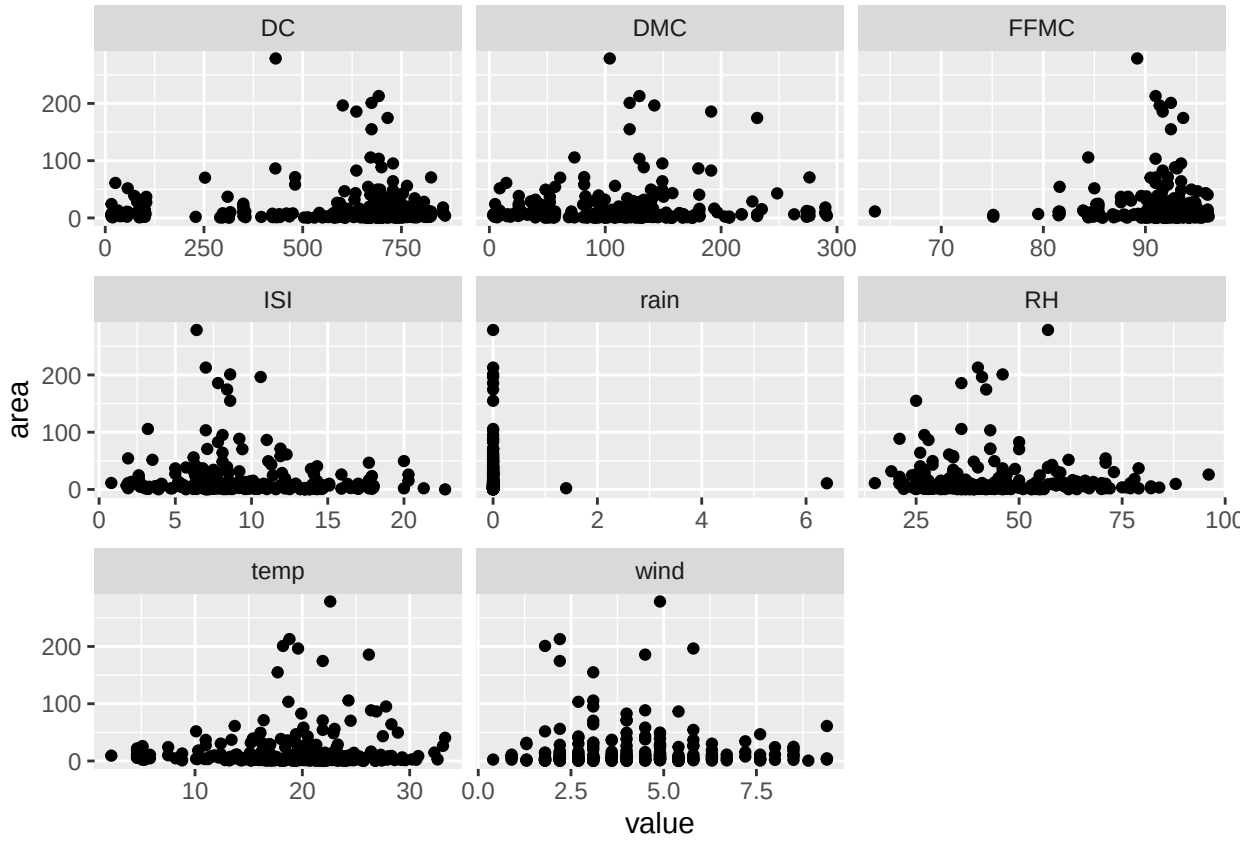



from the graph above, we can condlude that there are 3 categories of each variable in the scope of the relationship with area of forest fire

- Positive relationship : The wider the forest fire area, the higher the value of variable, they are: DC, FFMC, and temp
- Negative relationship : The higher the variable, the narrower the area of forest fire is. They are : ISI, rain, and
- Neutral : DMC and wind

## Page 6, Dealing with Outliers of Area Data

```
# FIltering with area below 300 and not zero (0)

forest_fires_long %>%
  filter(area < 300 & area != 0) %>%
  ggplot(aes(x=value, y=area)) +
  geom_point() +
  facet_wrap(vars(features), scale="free_x")
```




After the outlier is removed, we can see the true relationship of the features and the area of forest fire

Category of relationship with forest fire area (revised):

1. Positive strong : FFMC
2. Positive weak : DC and temp
3. Neutral : DMC and wind
4. Negative weak : RH and ISI
5. Negative strong : Rain