Forest Fire Analysis

Jonathan

1/4/2021

Introduction

We will use a dataset on Portugal's forest fires to analyze the different patterns that may lead to or cause forest fires and attempt to predict the occurence of forest fires in Portugal using a variety of modeling techniques. For this specific project, I will be focusing on producing the visualizations of the results. Dataset Source: https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/ (https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/)

Dataset Keys:

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

Code:

Read data tables

```
# load packages and read csv of dataset
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ purrr   0.3.4
## ✓ tibble  3.0.4     ✓ dplyr   1.0.2
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.0
```

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

```
forest_fires <- 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()
## )
```

```
glimpse(forest_fires)
```

```
## Rows: 517
## Columns: 13
## $ X     <dbl> 7, 7, 7, 8, 8, 8, 8, 8, 8, 7, 7, 7, 6, 6, 6, 6, 5, 8, 6, 6, 6, …
## $ Y     <dbl> 5, 4, 4, 6, 6, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, …
## $ month <chr> "mar", "oct", "oct", "mar", "mar", "aug", "aug", "aug", "sep", …
## $ day   <chr> "fri", "tue", "sat", "fri", "sun", "sun", "mon", "mon", "tue", …
## $ FFMC  <dbl> 86.2, 90.6, 90.6, 91.7, 89.3, 92.3, 92.3, 91.5, 91.0, 92.5, 92.…
## $ DMC   <dbl> 26.2, 35.4, 43.7, 33.3, 51.3, 85.3, 88.9, 145.4, 129.5, 88.0, 8…
## $ DC    <dbl> 94.3, 669.1, 686.9, 77.5, 102.2, 488.0, 495.6, 608.2, 692.6, 69…
## $ ISI   <dbl> 5.1, 6.7, 6.7, 9.0, 9.6, 14.7, 8.5, 10.7, 7.0, 7.1, 7.1, 22.6, …
## $ temp  <dbl> 8.2, 18.0, 14.6, 8.3, 11.4, 22.2, 24.1, 8.0, 13.1, 22.8, 17.8, …
## $ RH    <dbl> 51, 33, 33, 97, 99, 29, 27, 86, 63, 40, 51, 38, 72, 42, 21, 44,…
## $ wind  <dbl> 6.7, 0.9, 1.3, 4.0, 1.8, 5.4, 3.1, 2.2, 5.4, 4.0, 7.2, 4.0, 6.7…
## $ rain  <dbl> 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0…
## $ area  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
```

Data Processing

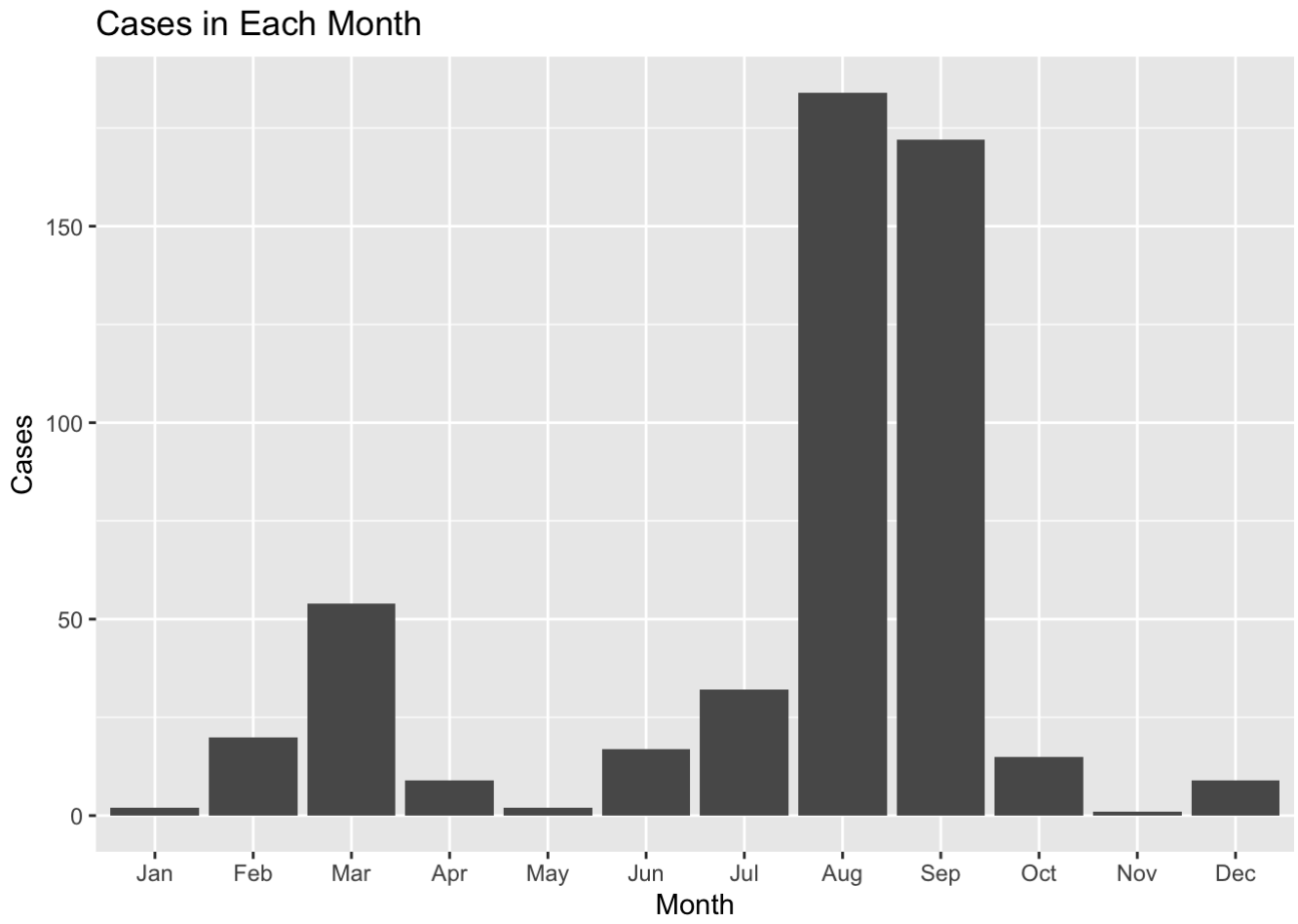
```
# change month and day to numerical values for easy organization later on
dayToNumber <- function(days) {
  result <- c()
  for (day in days) {
    number <- switch(day, "sun" = 7L, "mon" = 1L, "tue" = 2L, "wed" = 3L,
                     "thu" = 4L, "fri" = 5L, "sat" = 6L, -1L)
    result <- c(result, number)
  }
  return(result)
}

monthToNumber <- function(months) {
  result <- c()
  for (month in months) {
    number <- switch(month, "jan" = 1L, "feb" = 2L, "mar" = 3L, "apr" = 4L,
                     "may" = 5L,"jun" = 6L, "jul" = 7L, "aug" = 8L, "sep" = 9L,
                     "oct" = 10L, "nov" = 11L, "dec" = 12L, -1L)
    result <- c(result, number)
  }
  return(result)
}

forest_fires <- forest_fires %>% mutate(
  month_number = monthToNumber(month),
  day_number = dayToNumber(day))

#forest_fires %>% pull(dayNumber) %>% unique
```

Month Levels

```
# Clean and organize the data, plot most frequent cases on graph
months <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov",
"Dec")
forest_fires_by_month <- forest_fires %>% group_by(month_number) %>% summarise(cases = n
())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

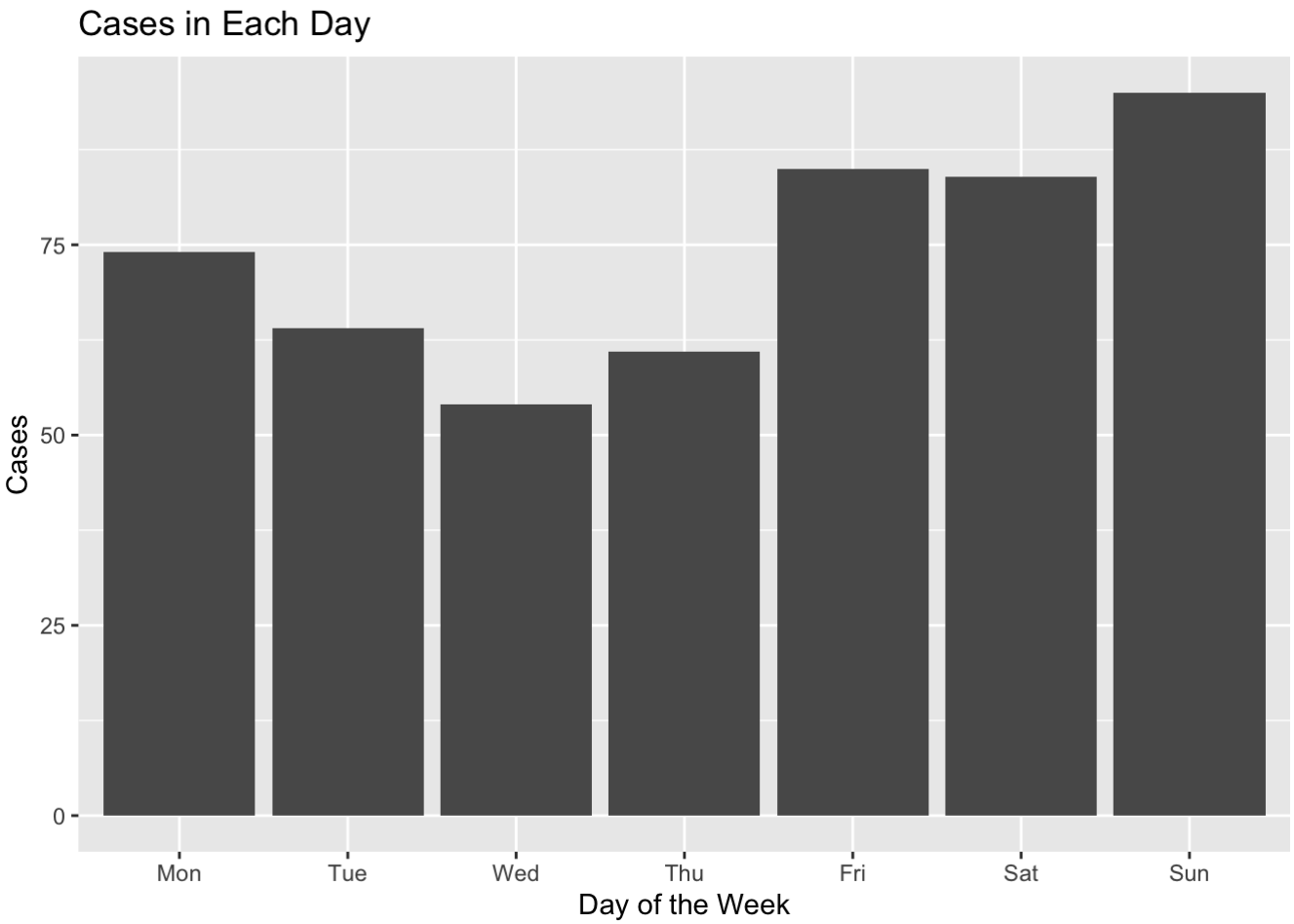
```
forest_fires_by_month %>% ggplot(aes(x = month_number, y = cases)) + geom_col() + scale_
x_discrete(limits = months) + labs(title = "Cases in Each Month", x = "Month", y = "Case
s")
```




```
days <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat","Sun")
forest_fires_by_day <- forest_fires %>% group_by(day_number) %>% summarise(cases = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
forest_fires_by_day %>% ggplot(aes(x = day_number, y = cases)) + geom_col() + labs(title
= "Cases in Each Day", x = "Day of the Week", y = "Cases") + scale_x_discrete(limits = d
ays)
```

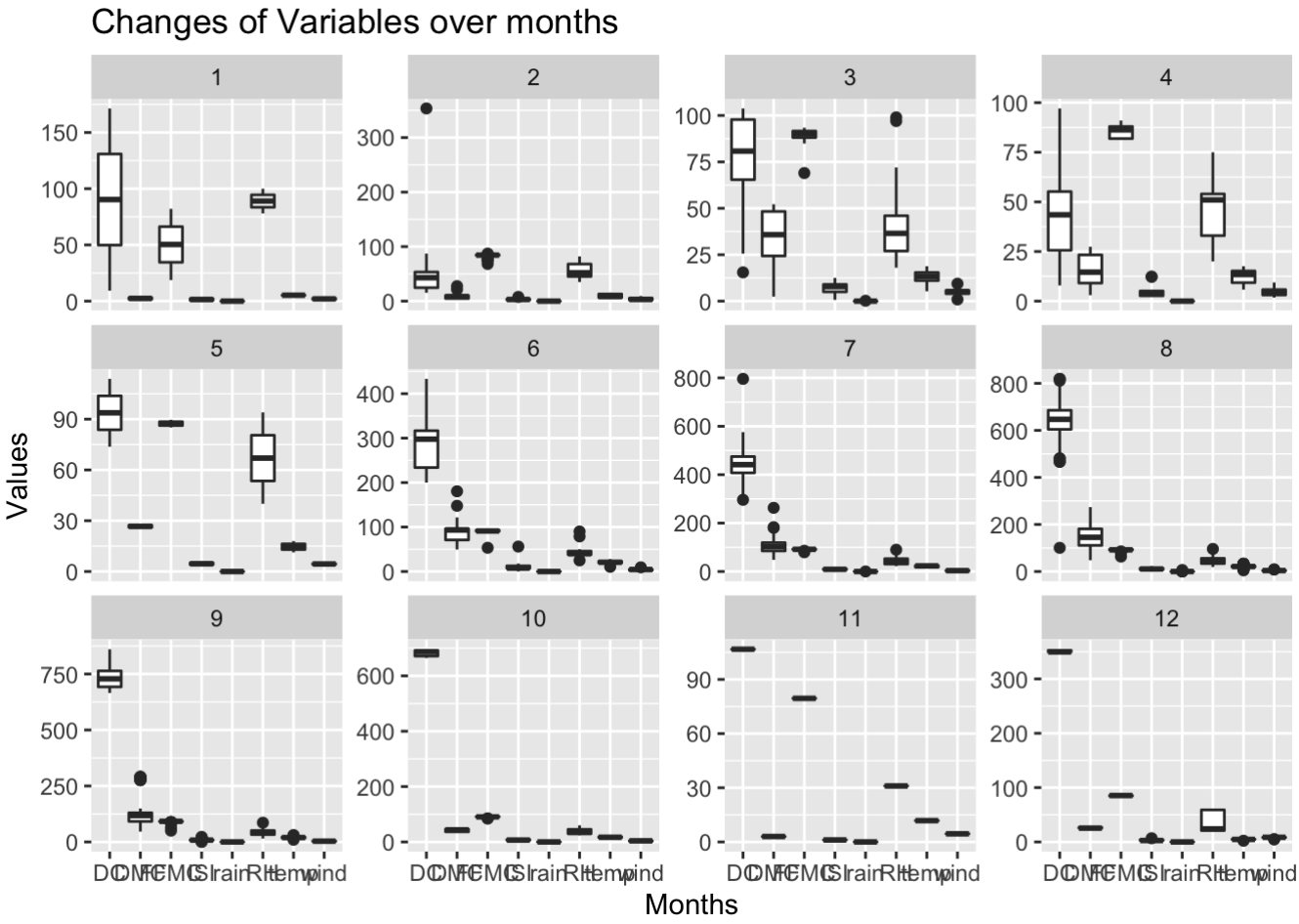


Checkpoint

From our monthly cases chart, we will notice that the most forest fire cases occur in August and September along with March being relatively common, whereas months January, May, and November have the least amount of forest fires happening. If we look at the figure showing the amount of cases in each day, we will notice how the weekend days (Friday, Saturday, Sunday) has the most amount of forest fire cases, whereas the weekdays have lesser cases.

Plotting variables against time

```
# convert original rom wide to long format, find patterns in months
forest_fires_long <- forest_fires %>% pivot_longer(
  cols = c(FFMC, DMC, DC, ISI, temp, RH, wind, rain),
  names_to = "column",
  values_to = "values"
) %>% arrange(month_number)


forest_fires_long %>% ggplot(aes(x = column, y = values)) + geom_boxplot() + facet_wrap
(vars(month_number), scales = "free_y") + labs(
  title = "Changes of Variables over months", x = "Months", y = "Values")
```

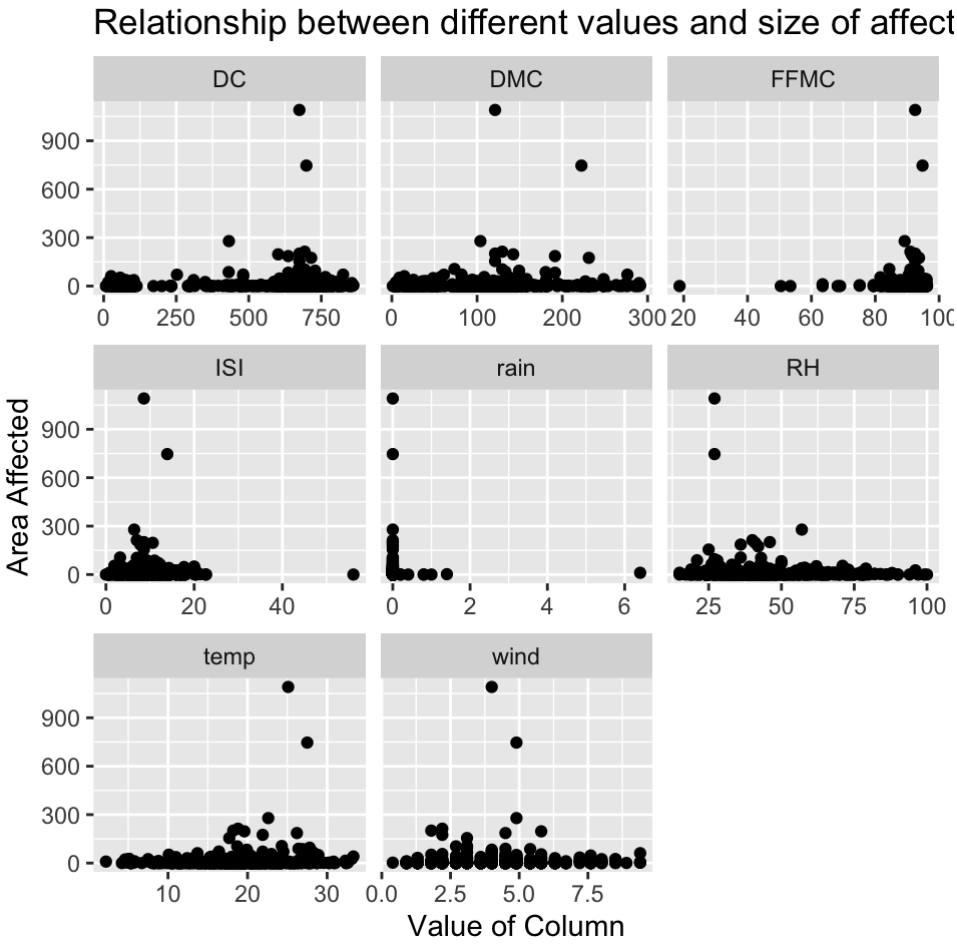


Checkpoint:

For August and September, there were allarming larger values of FFMC than any other month.

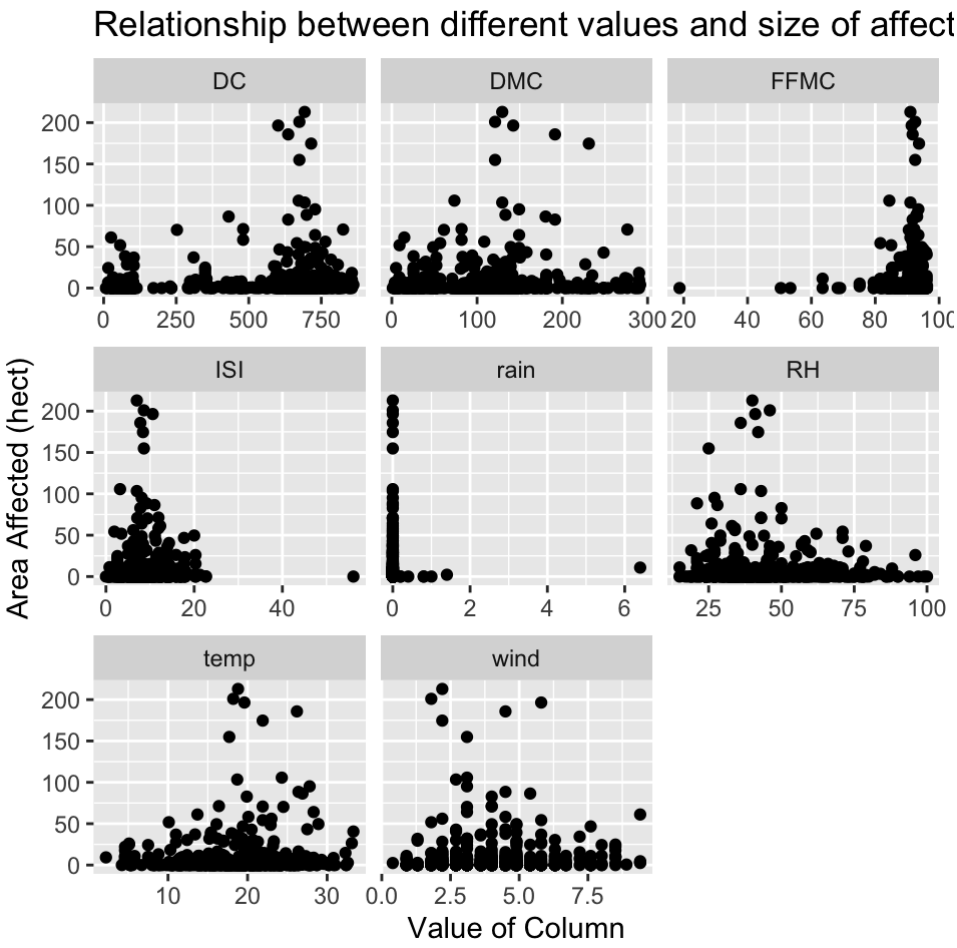
Examine Forest Fire's Serevity

```
forest_fires_long %>% ggplot(aes(x = values, y = area)) + geom_point() + facet_wrap(vars
(column), scales = "free_x") + labs(
  title = "Relationship between different values and size of affected area", x = "Value
 of Column", y = "Area Affected")
```



Examine the plots again without outlying areas

```
forest_fires_long %>% filter(area < 250) %>% ggplot(aes(x = values, y = area)) + geom_po
int() + facet_wrap(vars(column), scales = "free_x") + labs(
  title = "Relationship between different values and size of affected area", x = "Value
 of Column", y = "Area Affected (hect)")
```



Findings and Conclusions

We can see that there are relatively normal relationships between size of area and values for DMC, wind, temperature, and ISI (ignoring outliers). RH exhibits a positively skewed graph whereas DC and FFMC exhibits a negatively skewed diagram.