Analyzing Forest Fire Data Through Data Visualizations

Mariby Cruz

2025-03-27

Introduction

The importance of forest fire data

This dataset was is associated with the scholarly article, **A Data Mining Approach to Predict Forest Fires using Meteorological Data** by Paulo Cortez and Anıbal Morais. The data was collected on forest fires in Portugal. It contains weather measurements such as relative humidity, wind speed, rain, burnt area, and additional Fire Weather Indices (FWI).
By collecting this data, a scientist can understand how each measurement influences forest fires - if there's any relationship. A single row of data represents a point in time on a single location where a fire occured with different weather condition measurements recorded. This guided project focuses on creating visuals to help us understand how forest fires have developed over time as these measurements change.

```
forestfires <- read.csv("~/git/DataQuest_guided_projects/forestfires.csv")
glimpse(forestfires)
```

```
## Rows: 517
## Columns: 13
## $ X     <int> 7, 7, 7, 8, 8, 8, 8, 8, 8, 7, 7, 7, 6, 6, 6, 6, 5, 8, 6, 6, 6, 5~
## $ Y     <int> 5, 4, 4, 6, 6, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 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, 92.5~
## $ DMC   <dbl> 26.2, 35.4, 43.7, 33.3, 51.3, 85.3, 88.9, 145.4, 129.5, 88.0, 88~
## $ DC    <dbl> 94.3, 669.1, 686.9, 77.5, 102.2, 488.0, 495.6, 608.2, 692.6, 698~
## $ 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, 0~
## $ temp  <dbl> 8.2, 18.0, 14.6, 8.3, 11.4, 22.2, 24.1, 8.0, 13.1, 22.8, 17.8, 1~
## $ RH    <int> 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, 0~
```

Data Processing

The months and days are stored as characters and would display in alphabetical order when outputted. Therefore, there needs to be a conversion and be put in the order that they occur.

```
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"
```

```
forestfires$month <- factor(forestfires$month, levels = c("jan", "feb", "mar", "apr", "may", "jun", "ju

forestfires$day <- factor(forestfires$day, levels = c("sun", "mon", "tue", "wed", "thu", "fri", "sat"))

#no missing data present
col_missing_counts <- c()
for (col_name in names(forestfires)) {
  #sum up the counts of NA for each column
  col_missing_counts[col_name] <- sum(is.na(forestfires[col_name]))
}
col_missing_counts
```

```
##     X     Y month   day  FFMC   DMC    DC   ISI  temp    RH  wind  rain  area
##     0     0     0     0     0     0     0     0     0     0     0     0     0
```

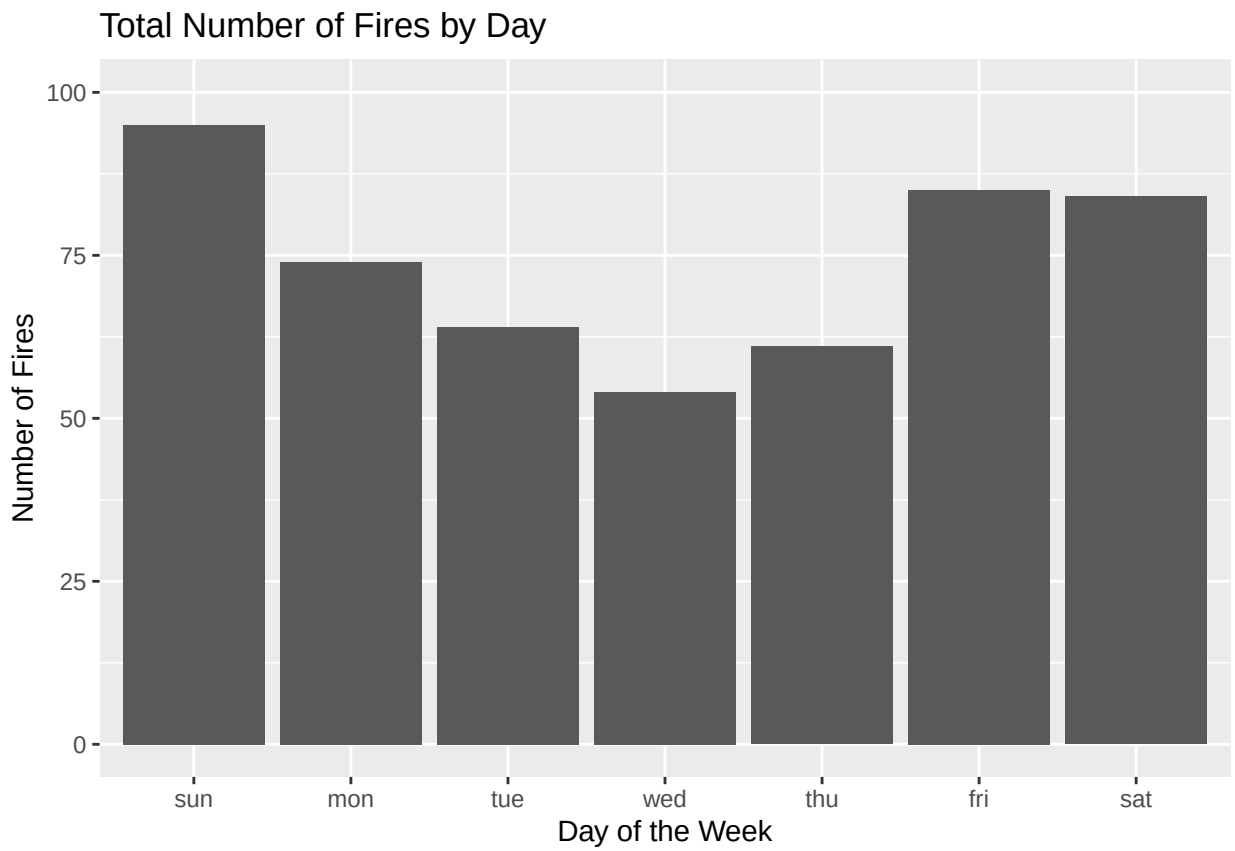
When do forest fires occur?

It seems that there may be less fires during the middle of the week, specifically on Wednesdays and Thursdays. I would wildly guess that there are fewer fires in the middle of the week because there are less people lighting fires near forests (ie. camping fires).
Moreover, there is a higher frequency of fires in the months of August and September. This may be because this is the peak of summer heat in Portugal making conditions drier and easier to ignite. Coupled with summer travelling and higher probability of accidental fires.

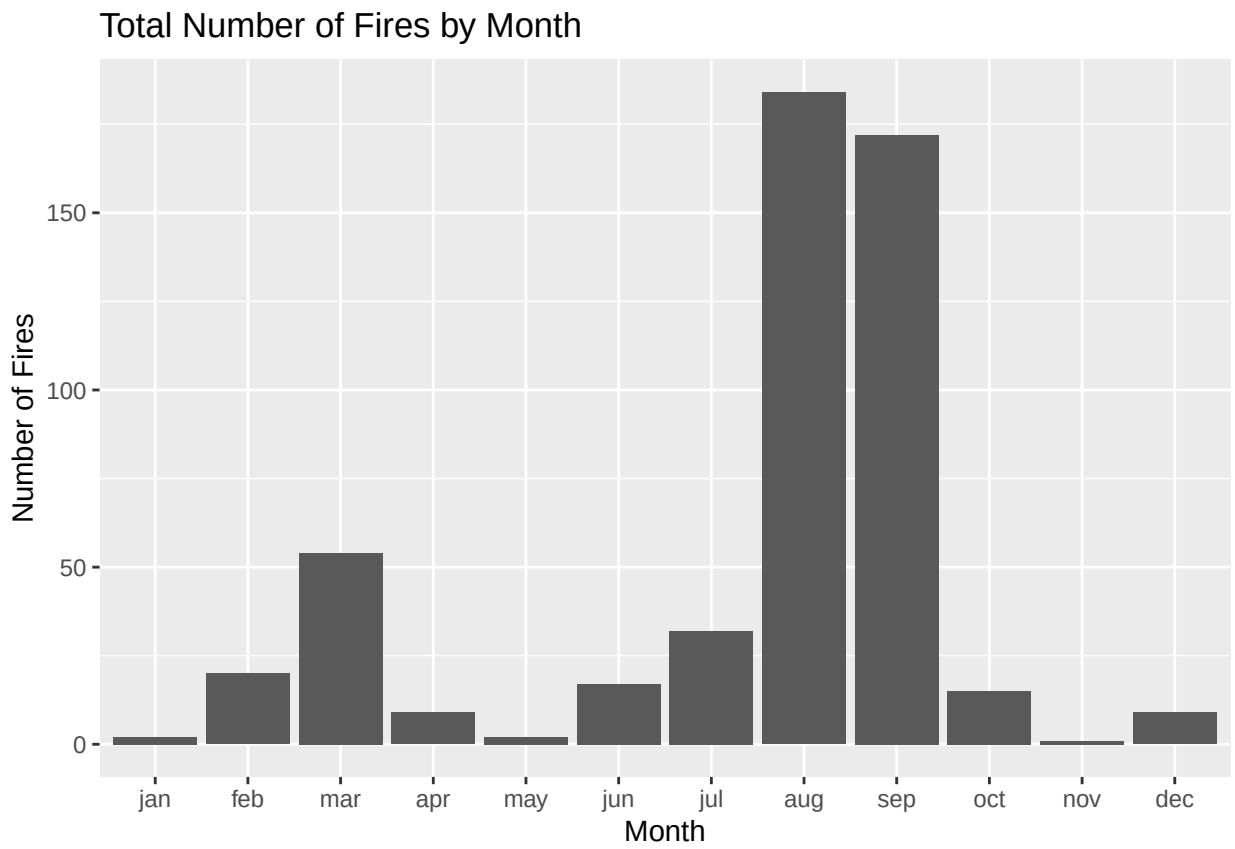
```
#created tibbles of the number of fires on each day and in each month.
fire_counts_month <- forestfires %>%
    group_by(month) %>%
    summarize(
      fire_counts = n()
    )

fire_counts_day <- forestfires %>%
    group_by(day) %>%
    summarize(
      fire_counts = n()
    )

#Creating barcharts for the counts above
fire_counts_day %>%
  ggplot(aes(x = day, y = fire_counts)) +
  geom_col() +
  ylim(0, 100) +
  labs(
    title = "Total Number of Fires by Day",
    x = "Day of the Week",
    y = "Number of Fires"
  )
```




```
fire_counts_month %>%
  ggplot(aes(x = month, y = fire_counts)) +
  geom_col() +
  labs(
    title = "Total Number of Fires by Month",
    x = "Month",
    y = "Number of Fires"
  )
```



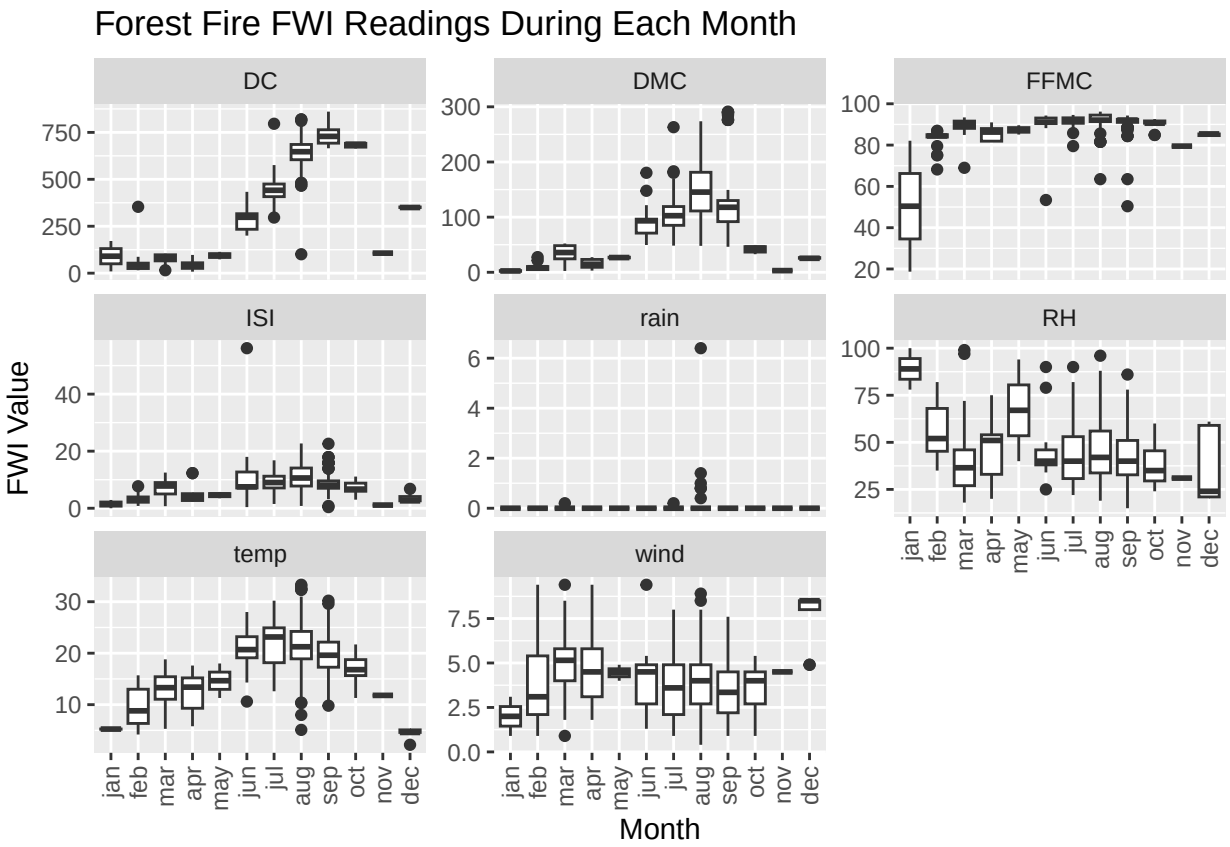
Plotting Other Variables Against Time

When FWI Index measurements are plotted for each month, there is a relationship with the the number of forest fires in the months of August and September. For instance, there are higher temperature values meaning it's warmer, higher DMC (Duff Moisture Code) meaning more moisture in loose leaf organic matter, and higher DC (Drought Code) meaning more moisture in deep compacted organic layers.

```
#pivot data into long format to facet_wrap them in the plot
forestfires_long <- forestfires %>%
  pivot_longer(
    cols = c(FFMC, DMC, DC, ISI, temp, RH, wind, rain),
    names_to = "Index",
    values_to = "Value"
  )

#created boxplots to see the distribution of index values throughout the month
forestfires_long %>%
  ggplot(aes(x = month, y = Value)) +
  geom_boxplot() +
  facet_wrap(
    vars(Index),
    scales = "free_y"
  ) +
  labs(
      title = "Forest Fire FWI Readings During Each Month",
```

```
    y = "FWI Value",
    x = "Month"
  ) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



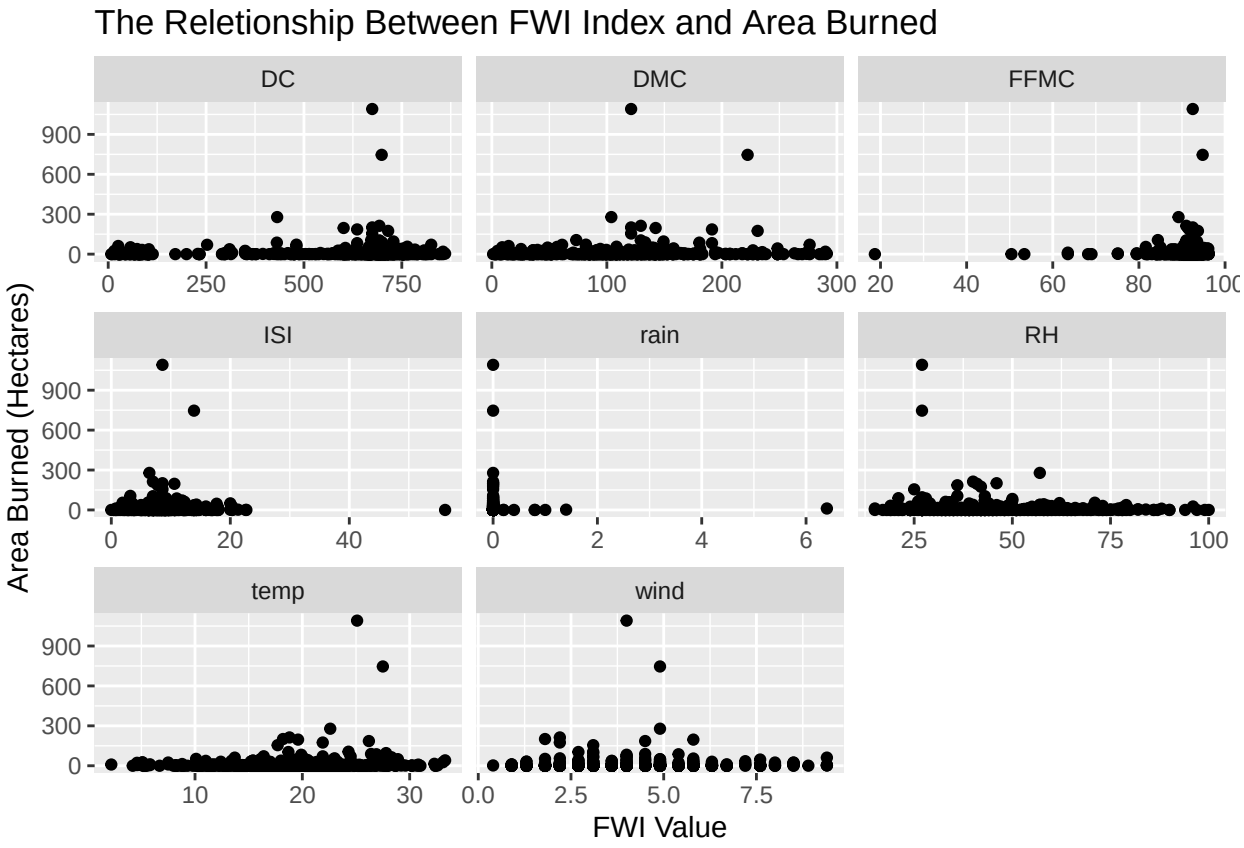
Examining Forest Fire Severity

Data doesn't always have a variable that measures severity. DataQuest shows a way to represent that with a proxy. In this dataset, the variable `area` measures the number of hectares burned during a forest fire. So, the assumption is that more severe fires mean larger area of forest burned.

The scatter plots below show that the largest forest fires burning about 600-900 ha, also had index recordings of high temperature, high FFMC, low ISI, low rain, and low humidity. These variables would be prioritized to investigate models that can help predict and mitigate forest fires.

```
#created scatter plot to see the relationship between index values and area burnt
forestfires_long %>%
  ggplot(aes(x = Value, y = area)) +
  geom_point() +
  facet_wrap(
    vars(Index),
    scales = "free_x"
  ) +
  labs(
      title = "The Reletionship Between FWI Index and Area Burned",
```

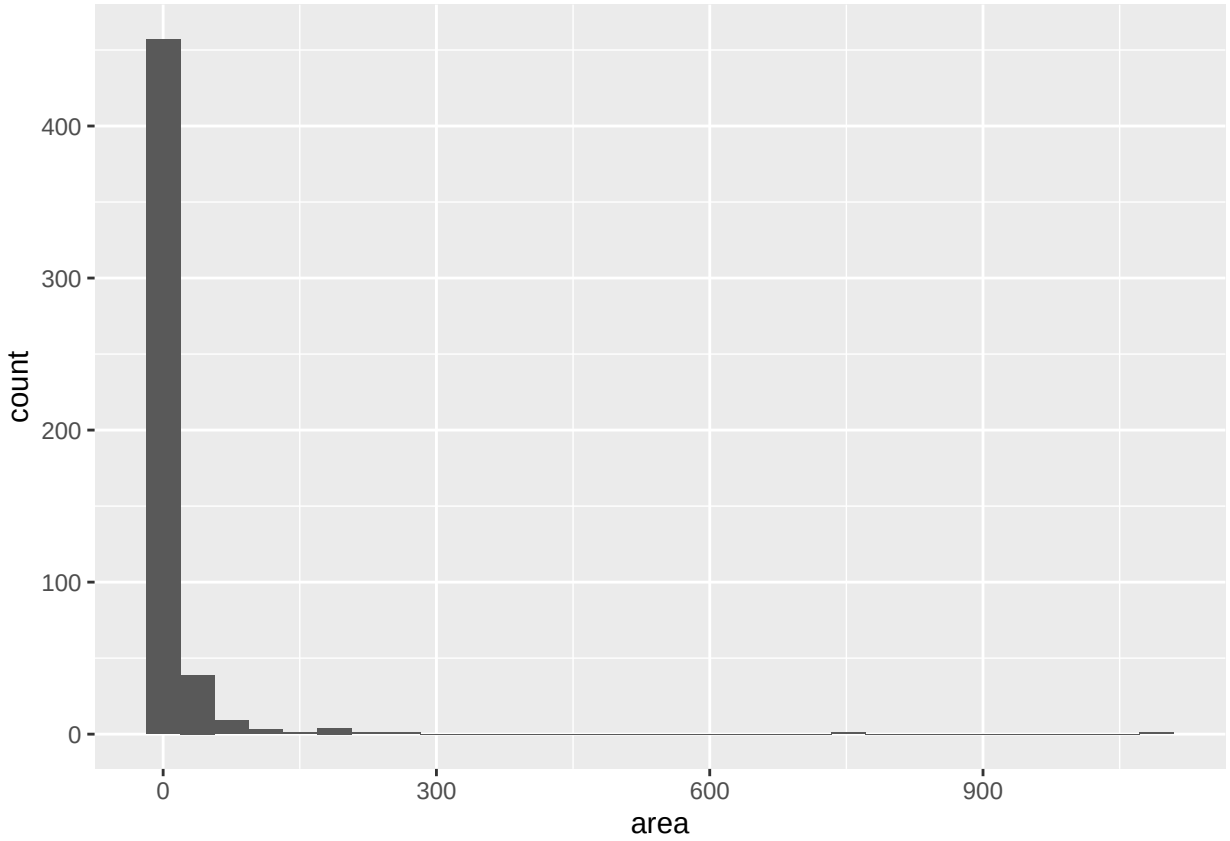
```
      x = "FWI Value",
      y = "Area Burned (Hectares)")
```



Outlier Problem

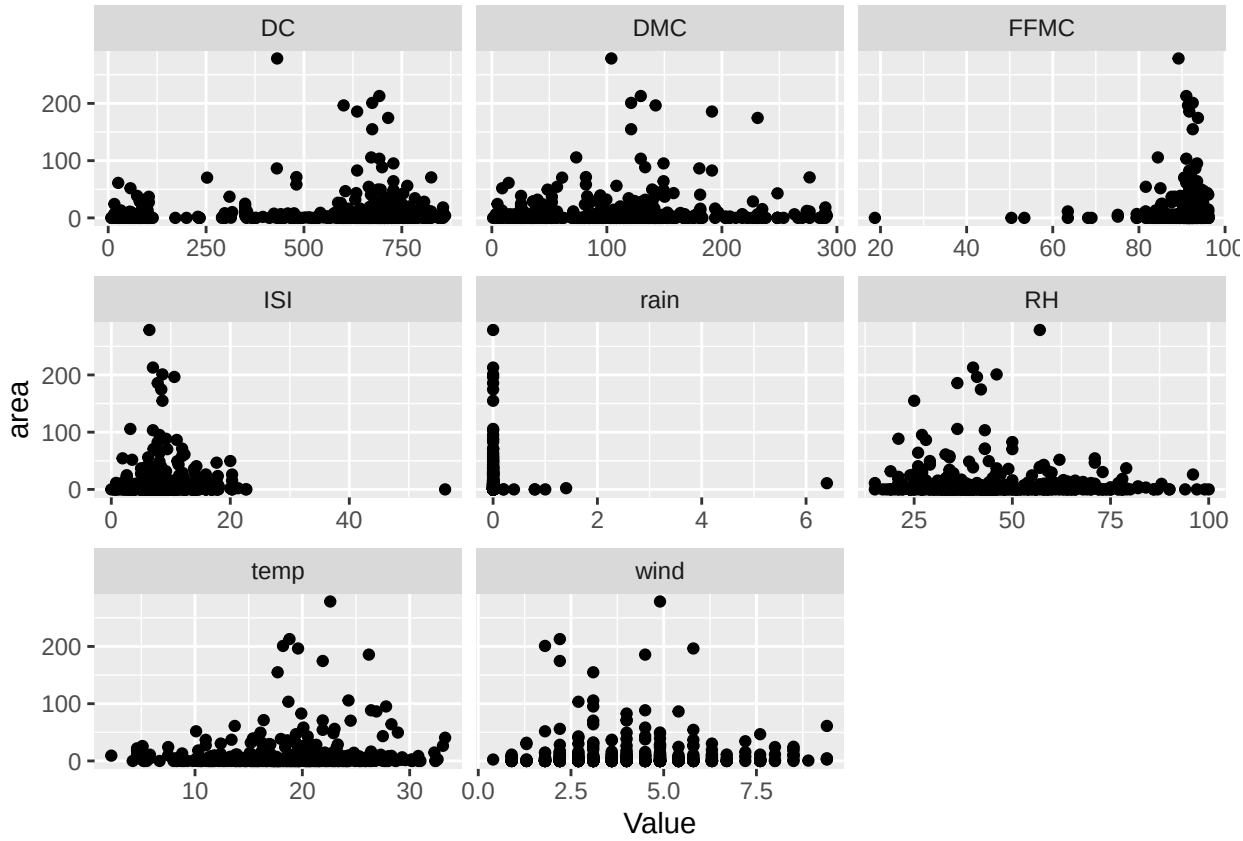
The graphs above have two high extreme values with large hectares burned, which obscures the relationship between index values and the distribution of smaller hectares burned. The graphs below show the clarity of the relationships between index and amount of hectares burnt when outlier data is filtered out. Ranges of the index measurements are easier to pinpoint and show when fires are more likely to become severe.

```
#Distribution of area
forestfires %>%
    ggplot(aes(x = area)) +
    geom_histogram()
```



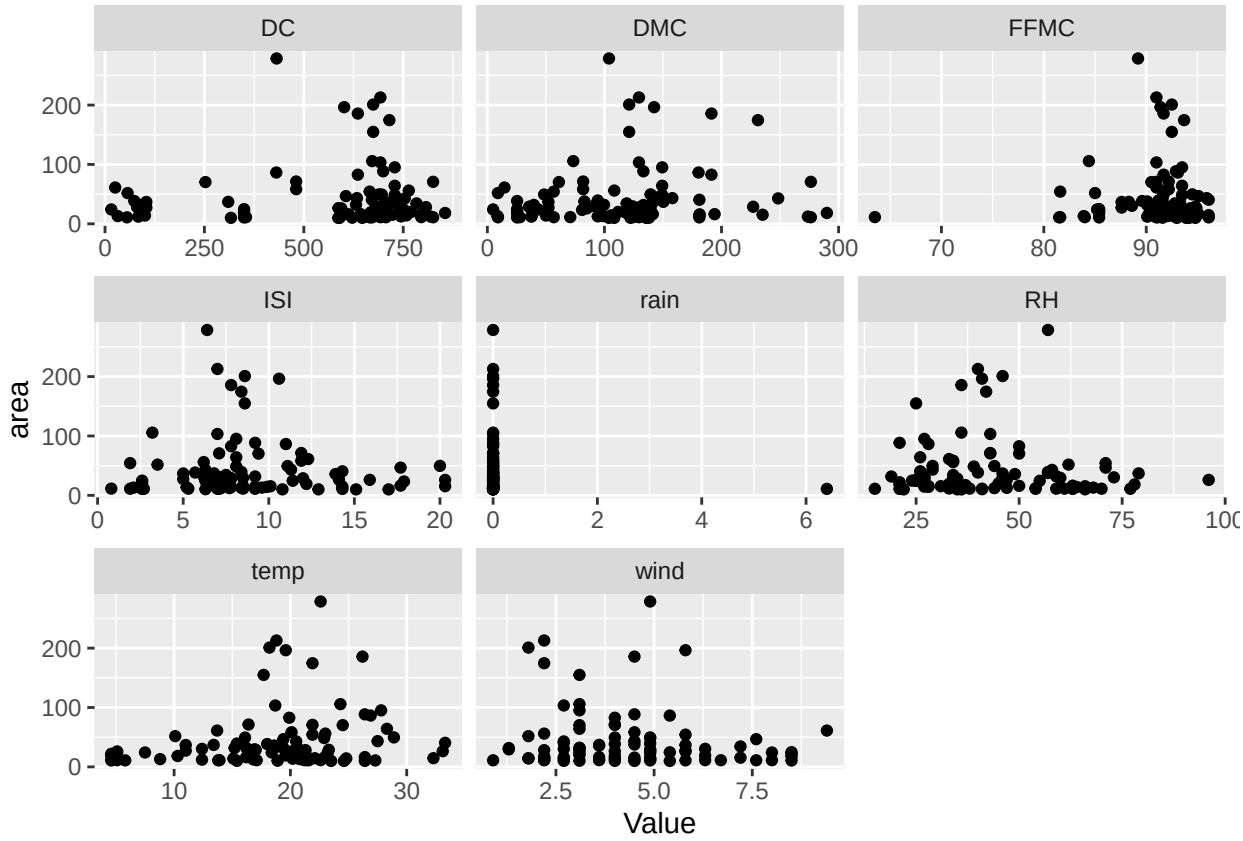

```
#filter for data to keep 300 ha or less burned
forestfires_area1 <- forestfires_long %>%
  filter(
    area < 300
  )

forestfires_area1 %>%
    ggplot(aes(x = Value, y = area)) +
    geom_point() +
    facet_wrap(
    vars(Index),
    scales = "free_x"
  )
```




```
#filter for data to keep data beteen 10 - 300 ha
forestfires_area2 <- forestfires_long %>%
  filter(
    area > 10 & area < 300
  )

forestfires_area2 %>%
    ggplot(aes(x = Value, y = area)) +
    geom_point() +
    facet_wrap(
    vars(Index),
    scales = "free_x"
  )
```


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