Beginner Project: Exploratory Visualization of Forest Fire Data

Sunsun Chanakarn

2025-08-22

Exploring Data Through Visualizations: Independent Investigations

Load the packages and data we'll need for the project

```
# load library
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- 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 error
```

```
# load data csv file
forest_fires <- read_csv("forestfires.csv")
```

```
## Rows: 517 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): month, day
## dbl (11): X, Y, FFMC, DMC, DC, ISI, temp, RH, wind, rain, area
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

The Importance of Forest Fire Data

```
# what columns are in the dataset?
colnames(forest_fires)
```

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

We know that the columns correspond to the following information:

- **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

A single row corresponds to the location of a fire and some characteristics about the fire itself. Higher water presence is typically asssociated with less fire spread, so we might expect the water-related variables (`DMC` and `rain`) to be related with `area`.

Data Processing

`month` and `day` are character vartiables, but we know that there is an inherent order to them. We'll convert these variables into factors so that they'll be sorted into the correct order when we plot them.

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

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

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

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

```
## order month - Day
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")

forest_fires <- forest_fires %>%
  mutate(
    month = factor(month, levels = month_order),
    day = factor(day, levels = day_order)
  )

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, 5~
## $ Y     <dbl> 5, 4, 4, 6, 6, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4~
## $ month <fct> mar, oct, oct, mar, mar, aug, aug, aug, sep, sep, sep, sep, aug,~
## $ day   <fct> fri, tue, sat, fri, sun, sun, mon, mon, tue, sat, sat, sat, fri,~
## $ 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    <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, 0~
```

When Do Most Forest Fires Occur?

We need to create a summary tibble that counts the number of fires that appears in each month. Then, we'll be able to use this tibble in a visualization. We can consider **month** and **day** to be different grouping variable, so our code to produce the tibbles and plots will look similar.

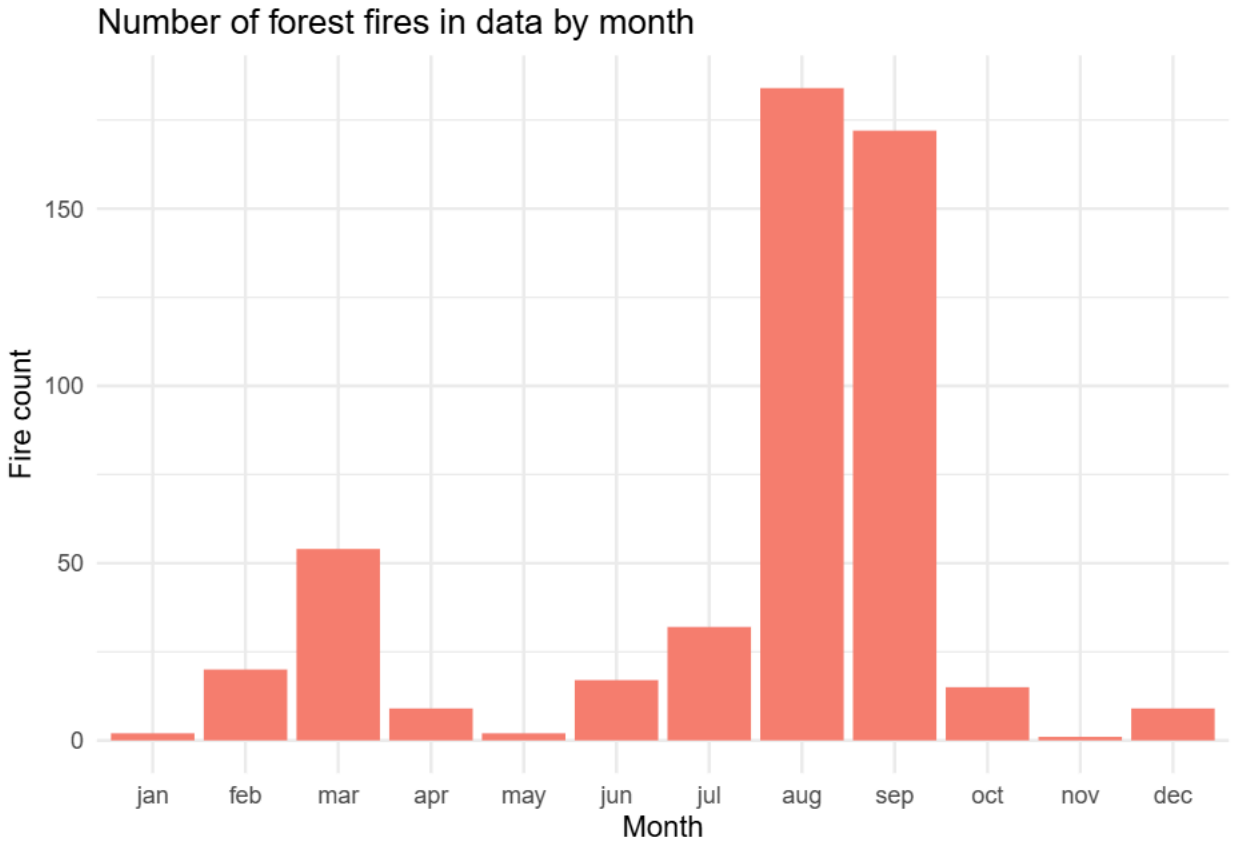
Month Level

```r
## month level
fires_by_month <- forest_fires %>%
  group_by(month) %>%
  summarize(total_fires = n())

print(fires_by_month)
```

```
## # A tibble: 12 x 2
##    month total_fires
##    <fct>       <int>
##  1 jan             2
##  2 feb            20
##  3 mar            54
##  4 apr             9
##  5 may             2
##  6 jun            17
##  7 jul            32
##  8 aug           184
##  9 sep           172
## 10 oct            15
## 11 nov             1
## 12 dec             9
```

```r
fires_by_month %>%
  ggplot(aes(x = month, y = total_fires)) +
  geom_col(fill = "salmon") +
  theme_minimal() +
  labs(
    title = "Number of forest fires in data by month",
    x = "Month",
    y = "Fire count"
  )
```

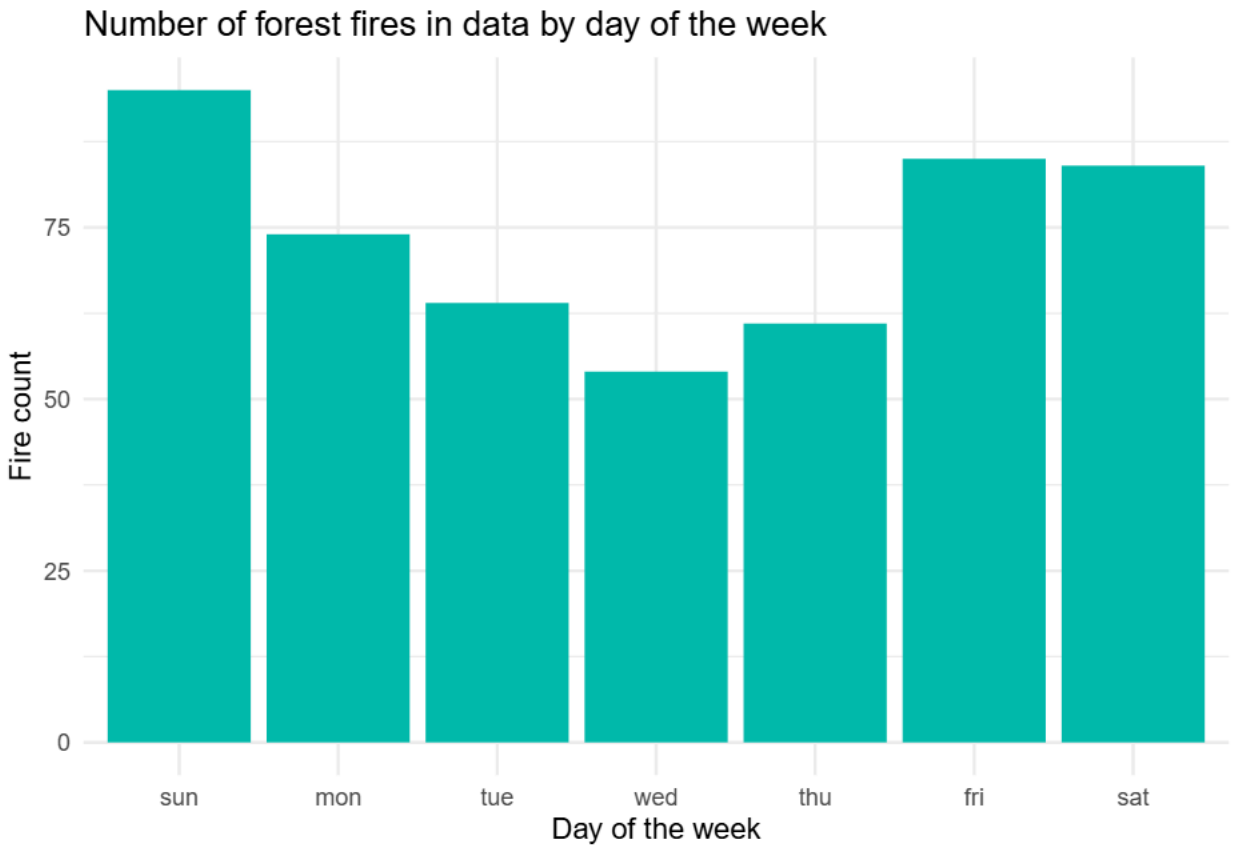



```
## day level
fires_by_day <- forest_fires %>%
  group_by(day) %>%
  summarize(total_fires = n())

print(fires_by_day)
```

```
## # A tibble: 7 x 2
##   day   total_fires
##   <fct>       <int>
## 1 sun            95
## 2 mon            74
## 3 tue            64
## 4 wed            54
## 5 thu            61
## 6 fri            85
## 7 sat            84
```

```
## create ggplot by day of week
fires_by_day %>%
  ggplot(aes(x = day, y = total_fires)) +
  geom_col(fill="#02BDAE") +
  theme_minimal() +
  labs(
    title = "Number of forest fires in data by day of the week",
    y = "Fire count",
    x = "Day of the week"
  )
```



We see a massive spike in fires in August and September, as well as a smaller spike in March. Fires seem to be more frequent on the weekend.

Plotting Other Variables Against Time

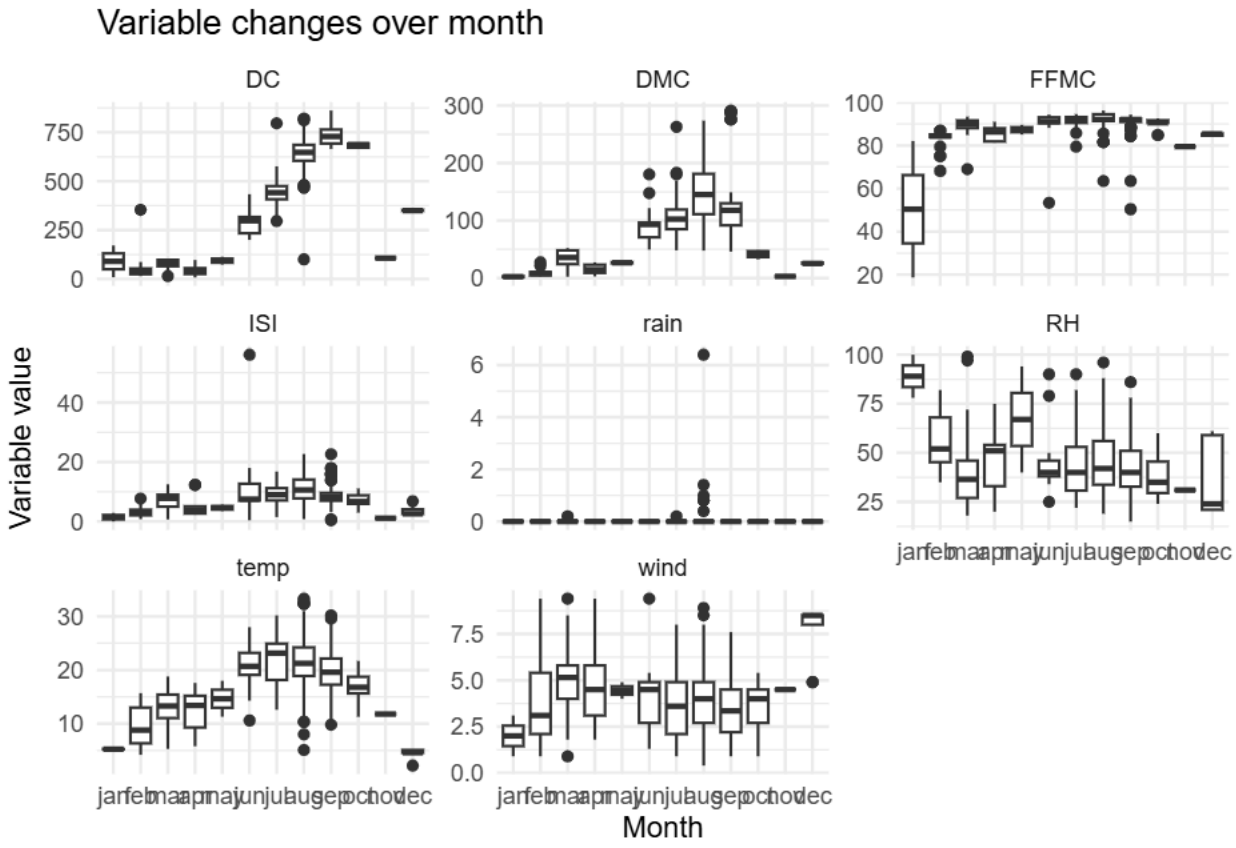
```
forest_fires_long <- forest_fires %>%
  pivot_longer(
    cols = c("FFMC", "DMC", "DC",
             "ISI", "temp", "RH",
             "wind", "rain"),
    names_to = "data_col",
    values_to = "value"
)

print(forest_fires_long)
```

```
## # A tibble: 4,136 x 7
##        X     Y month day    area data_col value
##    <dbl> <dbl> <fct> <fct> <dbl> <chr>    <dbl>
##  1     7     5 mar   fri       0 FFMC      86.2
##  2     7     5 mar   fri       0 DMC       26.2
##  3     7     5 mar   fri       0 DC        94.3
##  4     7     5 mar   fri       0 ISI        5.1
##  5     7     5 mar   fri       0 temp       8.2
##  6     7     5 mar   fri       0 RH        51
##  7     7     5 mar   fri       0 wind       6.7
```

```
##  8     7     5 mar   fri       0 rain       0
##  9     7     4 oct   tue       0 FFMC      90.6
## 10     7     4 oct   tue       0 DMC       35.4
## # i 4,126 more rows
```

```
## plot boxplot
forest_fires_long %>%
  ggplot(aes(x = month, y = value)) +
  geom_boxplot() +
  facet_wrap(vars(data_col), scale = "free_y") +
  theme_minimal() +
  labs(
    title = "Variable changes over month",
    x = "Month",
    y = "Variable value"
  )
```

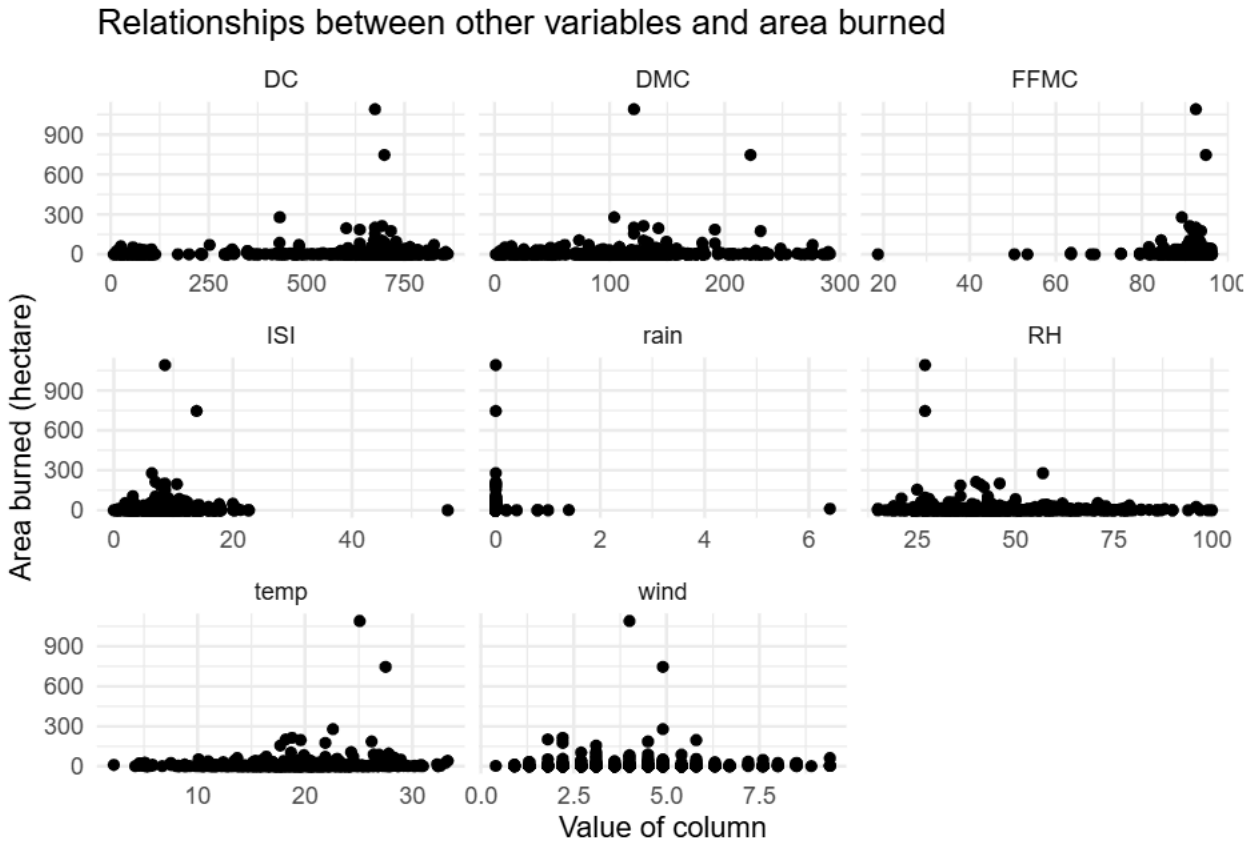


Examining Forest Fire Severity

We are trying to see how each of the variables in the dataset relate to `area`. We can leverage the long format version of the data we created to use with `facet_wrap()`.

```
## create scatter plot
forest_fires_long %>%
  ggplot(aes(x = value, y = area)) +
  geom_point() +
  facet_wrap(vars(data_col), scales = "free_x") +
  theme_minimal() +
```

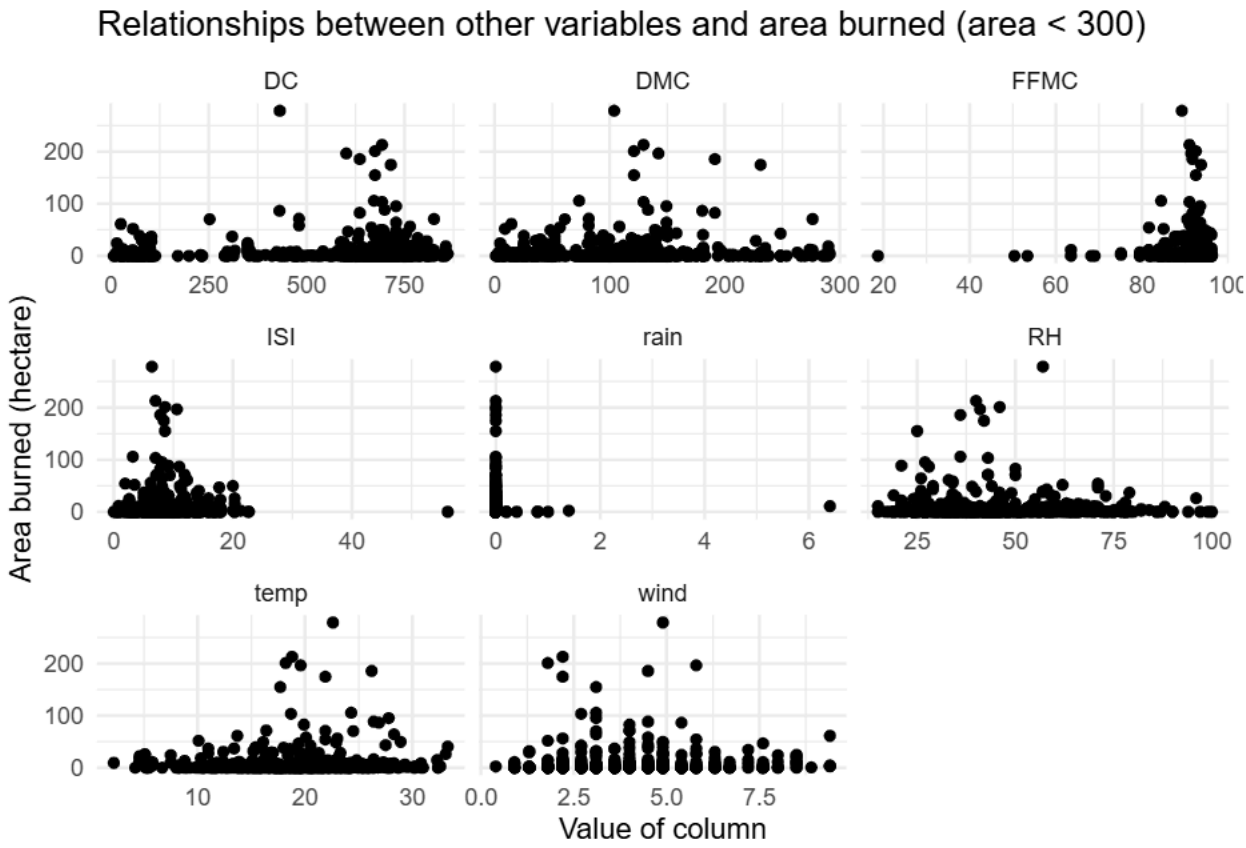
```
  labs(
    title = "Relationships between other variables and area burned",
    x = "Value of column",
    y = "Area burned (hectare)"
  )
```



Outlier Problems

It seems that there are two rows where **area** that still hurt the scale of the visualization. Let's make a similar visualization that excludes these observations so that we can better see how each variable relates to **area**.

```
## filter area < 300
forest_fires_long %>%
  filter(area < 300) %>%
  ggplot(aes(x = value, y = area)) +
  geom_point() +
  facet_wrap(vars(data_col), scales = "free_x") +
  theme_minimal() +
  labs(
    title = "Relationships between other variables and area burned (area < 300)",
    x = "Value of column",
    y = "Area burned (hectare)"
  )
```


Relationships between other variables and area burned (area < 300)
DC
DMC
FFMC
ISI
rain
RH
temp
wind
Area burned (hectare)
Value of column