Analysis of Forest Fires Data

Oksana Bihun

9/8/2023

Introduction

In this project, we analyze forest fires data taken from a scientific paper by Portuguese scientists and answer the following questions.

- What are months with exceptionally high quantity of fires?
- On which weekdays there were more fires?
- How are Canadian Forest Fire Weather Index (FWI) variables related to the number of fires in a given months?

The main focus of the project is visualization of the data and analysis of the plots to draw conclusions about the relationships of the variables.

Learning the Basic Properties of the Dataframe

We begin with uploading the data and finding out the dimensions of the data-frame, the names of the columns, the datatypes of the entries, and the unique values in each column.

```
library(tidyverse)
library(knitr)

# Reading the data
forest_fires <- read.csv("forestfires.csv")

# Learning basic properties of dataframe forest_fires
cat("Number of rows:",nrow(forest_fires),"\n")
cat("Number of columns:",ncol(forest_fires),"\n")
cat("Column names:",colnames(forest_fires),"\n")

head(forest_fires)%>%
kable( caption="The first few rows of the `forest_fires`
       dataframe:")

print("The data types of the entries in each column:")
sapply(forest_fires, typeof)
```

```
## Number of rows: 517
## Number of columns: 13
## Column names: X Y month day FFMC DMC DC ISI temp RH wind rain area
```

Table 1: The first few rows of the `forest_fires` dataframe:

X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0
7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0
7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0
8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0
8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0
8	6	aug	sun	92.3	85.3	488.0	14.7	22.2	29	5.4	0.0	0

```
## [1] "The data types of the entries in each column:"
##           X           Y       month         day        FFMC         DMC
##   "integer"   "integer" "character" "character"    "double"    "double"
##          DC         ISI        temp          RH        wind        rain
##    "double"    "double"    "double"   "integer"    "double"    "double"
##        area
##    "double"
```

The descriptions of the columns are the following:

- **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

From this exploration, we see that a single row in the data-frame represents data about the risk of a forest fire in a given location. The Canadian Forest Fire Weather Index (FWI) system provides information about the meaning of the data in each column.

Cleaning and Ordering

First, let us remove the rows that have missing data.

```
# Removing rows with empty entries
forest_fires <- forest_fires %>%
  filter(
    !is.na(month),
    !is.na(FFMC),
    !is.na(DMC),
    !is.na(DC),
    !is.na(ISI),
    !is.na(temp),
    !is.na(RH),
    !is.na(wind),
    !is.na(rain),
    !is.na(area),
    !is.na(day))
```

Next, we order the rows in the data-frame `forset_fires` according to months (from jan to dec) and days (from mon to sun). To this end, we factorize the month and day variables.

```
# Factorizing the month variable
forest_fires <- forest_fires %>%
  mutate(month=factor(month, levels = c("jan", "feb", "mar",
    "apr", "may", "jun", "jul", "aug", "sep", "oct", "nov",
    "dec"))) %>%
  arrange(month)

# Factorizing the day variable
forest_fires <- forest_fires %>%
  mutate(day=factor(day, levels = c("mon", "tue", "wed", "thu",
    "fri", "sat", "sun"))) %>%
  arrange(day)

# Ordering the forest_fires rows according to months and days
forest_fires <- forest_fires[ order( forest_fires$month, forest_fires$day ), ]

# Checking the results of the factorization and the ordering
head(forest_fires)%>%
kable( caption="The first few rows of the forest_fires
       dataframe after the ordering of the rows according to
       months and days:")
```

Table 2: The first few rows of the forest_fires dataframe after the ordering of the rows according to months and days:

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
339	2	4	jan	sat	82.1	3.7	9.3	2.9	5.3	78	3.1	0	0.00
423	4	5	jan	sun	18.7	1.1	171.4	0.0	5.2	100	0.9	0	0.00
1	2	2	feb	mon	84.0	9.3	34.0	2.1	13.9	40	5.4	0	0.00
2	7	4	feb	mon	84.7	9.5	58.3	4.1	7.5	71	6.3	0	9.96
3	6	5	feb	mon	84.1	4.6	46.7	2.2	5.3	68	1.8	0	0.00
75	6	5	feb	tue	75.1	4.4	16.2	1.9	4.6	82	6.3	0	5.39

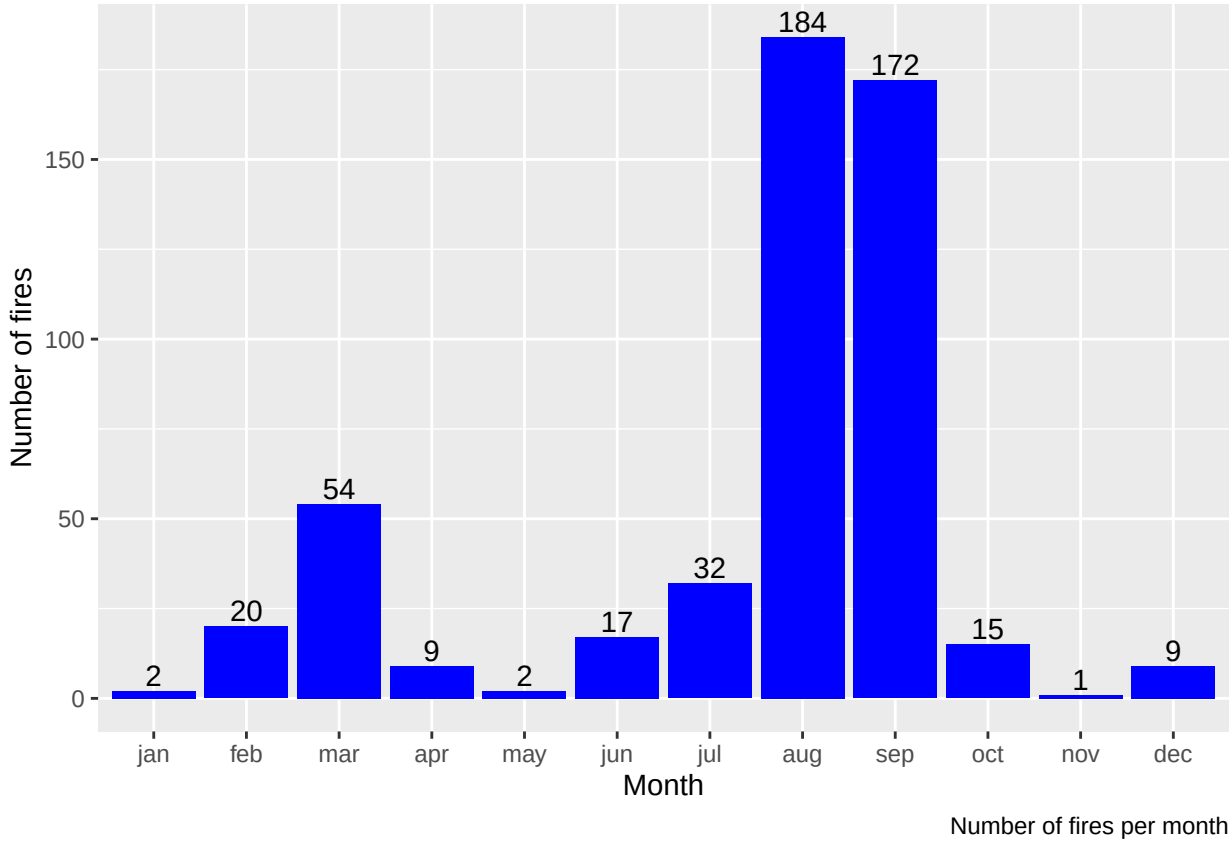
Data Analysis

In this section, we analyze the data to answer the questions posed in the Introduction.

Plotting the Number of Fires in Each Month and Each Day

In this subsection, we plot the number of fires in a given month and in a given day. We begin with plotting the number of fires each month.

```
# Creating a tibble that contains the number of fires each
# month
forest_fires_n_by_month <-
forest_fires  %>% group_by(month) %>% summarise(n_fires = n())

# Creating a bar chart of forest_fires_n_by_month
forest_fires_n_by_month %>%
  ggplot(aes(x=month,y=n_fires))+
  geom_bar(position='dodge',stat="identity",fill='blue')+
  labs(caption = "Number of fires per month", x="Month",
      y="Number of fires")+
  geom_text(aes(label=n_fires),
      position=position_dodge(width=0.9), vjust=-0.25)
```



Number of fires per month

Next, we verify our visual observation that the number of fires that occurred in August and September is larger than the number of fires that occurred during all the remaining months.

```
print("The number of fires in August and September:")
```

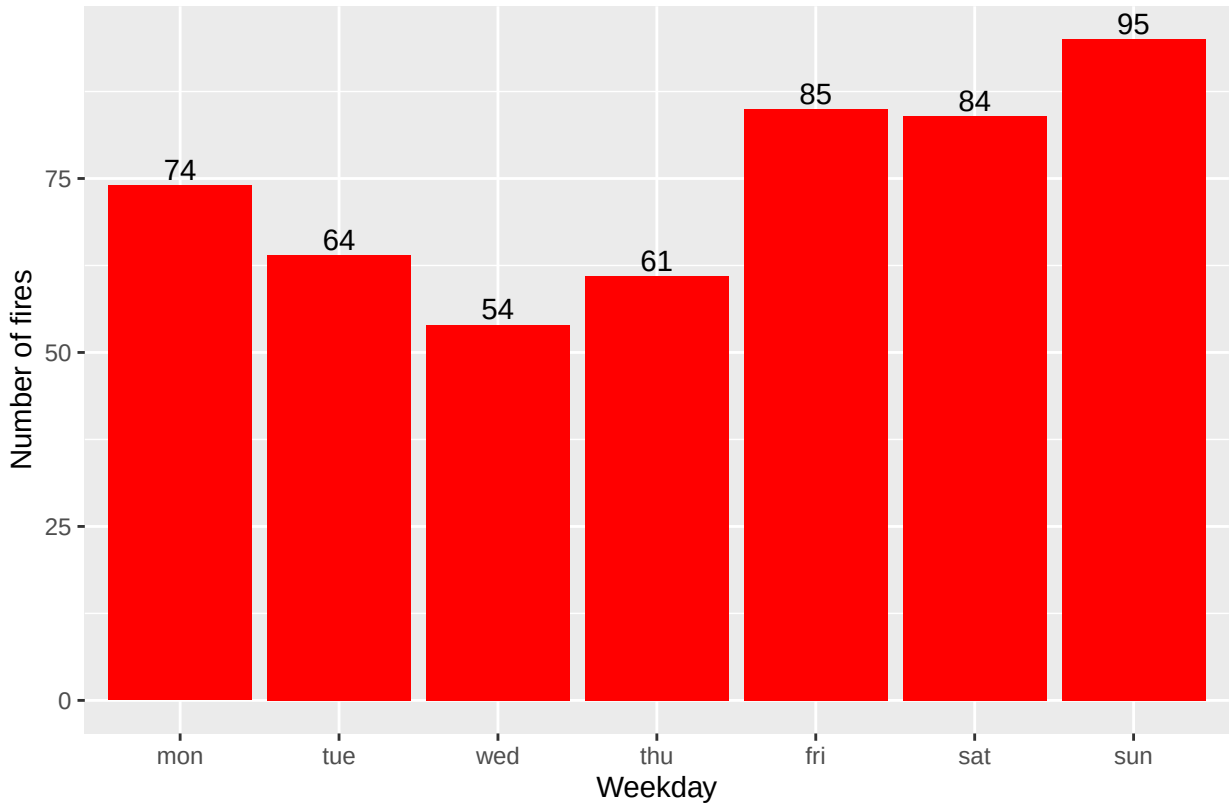
```r
# Calculating the number of fires in months 8 and 9
forest_fires_n_by_month$n_fires[[8]]+forest_fires_n_by_month$n_fires[[9]]

print("The number of fires in the remaining 10 months:")

# Calculating the number of fires in months 1 through 12
# except 8,9
indices=1:11
s_fires=0
for(i in indices){
  if(!(i %in% 8:9)){s_fires<-s_fires+forest_fires_n_by_month$n_fires[[i]]}
  i<-i+1}
s_fires
```

```
## [1] "The number of fires in August and September:"
## [1] 356
## [1] "The number of fires in the remaining 10 months:"
## [1] 152
```

We conclude that **there were more fires in August and September than during all the other months. The majority of fires occurred during the months of August and September**.

Let us now uncover which days of the weeks have more fires. To this end, we plot the number of fires that occurred on each weekday.

```r
# Creating a tibble that contains the number of fires that
# occurred on each weekday
forest_fires_n_by_weekday <-
forest_fires %>% group_by(day) %>% summarise(n_fires = n())

# Creating a bar chart of forest_fires_n_by_weekday
forest_fires_n_by_weekday %>% ggplot(aes(x=day,y=n_fires))+
geom_bar(position='dodge',stat="identity",fill='red')+
labs(caption = "Number of fires on a given weekday",
     x="Weekday", y="Number of fires")+
geom_text(aes(label=n_fires),
          position=position_dodge(width=0.9), vjust=-0.25)
```



Number of fires on a given weekday

From the last bar chart, it appears that **there were more fires on weekends**. Below, we calculate the percentage of the total number of the fires that occurred on Fridays, Saturdays, and Sundays.

```
print("The number of fires on Friday through Sunday:")

# Calculating the number of fires that occurred on
# weekdays 5 through 7
n_weekend_fires <- forest_fires_n_by_weekday$n_fires[[5]]+
forest_fires_n_by_weekday$n_fires[[6]]+
forest_fires_n_by_weekday$n_fires[[7]]

n_weekend_fires

print("The number of fires on the remaining weekdays:")

# Calculating the number of fires that occurred on
# weekdays 1 through 4
indices=1:7
s_fires=0
for(i in indices){
  if(!(i %in% 5:7)){s_fires<-s_fires+forest_fires_n_by_weekday$n_fires[[i]]}
  i<-i+1}
s_fires

print("The percentage of the fires that occurred on Fridays, Saturdays, and Sundays:")

# Calculating the percentage of the fires that occurred
```

```
# on weekends
n_weekend_fires/(n_weekend_fires+s_fires)*100
```

```
## [1] "The number of fires on Friday through Sunday:"
## [1] 264
## [1] "The number of fires on the remaining weekdays:"
## [1] 253
## [1] "The percentage of the fires that occurred on Fridays, Saturdays, and Sundays:"
## [1] 51.06383
```

We conclude that **51% of all the fires occurred on Fridays, Saturdays, and Sundays**.

The Relationship between the Number of Fires and the Average Values of the Canadian Forest Fire Weather Index (FWI) Variables

Let us now discover the relationship between the variables in the original tibble and the number of fires in a given month. To this end, we create a tibble that provides the number of fires and the average values of all the FWI variables for each month.

In order to produce piecewise linear graphs, we create a new column, `month_n`, with numbers that correspond to the months.

```
# Creating column month_n that contains the numbers of
# the months in each row
forest_fires <- forest_fires %>% mutate(
  month_n=case_when(
    month=='jan' ~1,
    month=='feb' ~2,
    month=='mar' ~3,
    month=='apr' ~4,
    month=='may' ~5,
    month=='jun' ~6,
    month=='jul' ~7,
    month=='aug' ~8,
    month=='sep' ~9,
    month=='oct' ~10,
    month=='nov' ~11,
    month=='dec' ~12
  )
)
```

Next, we create a table with the columns that contain the average values of the Canadian Forest Fire Weather Index (FWI) variables over each month as well as the number of fires that occurred each month.

```
# Creating tibble fires_avg_by_month that, in each row,
# contains the number of the month and the average
# values of the FWI variables
fires_avg_by_month <- forest_fires %>%
group_by(month_n) %>% summarise(n_fires = n(),
  avg_FFMC=mean(FFMC),
  avg_DMC=mean(DMC),
  avg_DC=mean(DC),
  avg_ISI=mean(ISI),
  avg_temp=mean(temp),
  avg_RH=mean(RH),
```

```
  avg_wind=mean(wind),
  avg_rain=mean(rain),
  avg_area=mean(area)
)

# Verifying the results
head(fires_avg_by_month)%>%
kable( caption="The first few rows of the `fires_avg_by_month`
       dataframe that contains the number of the fires each
       month and the averages of the FWI variables over each
       month:")
```

Table 3: The first few rows of the `fires_avg_by_month` dataframe that contains the number of the fires each month and the averages of the FWI variables over each month:

month	n_fires	avg_FFMC	avg_DMC	avg_DC	avg_ISI	avg_temp	avg_RH	avg_wind	avg_rain	avg_area
1	2	50.40000	2.40000	90.35000	1.450000	5.25000	89.00000	2.000000	0.0000000	0.000000
2	20	82.90500	9.47500	54.67000	3.350000	9.63500	55.70000	3.755000	0.0000000	6.275000
3	54	89.44444	34.54259	75.94259	7.107407	13.08333	40.00000	4.968519	0.0037037	4.356667
4	9	85.78889	15.91111	48.55556	5.377778	12.04444	46.88889	4.666667	0.0000000	8.891111
5	2	87.35000	26.70000	93.75000	4.600000	14.65000	67.00000	4.450000	0.0000000	19.240000
6	17	89.42941	93.38235	297.70588	11.776471	20.49412	45.11765	4.135294	0.0000000	5.841177

Next, we pivot tibble `fires_avg_by_month` to be able to plot the average values of each of the FWI variables in different colors.

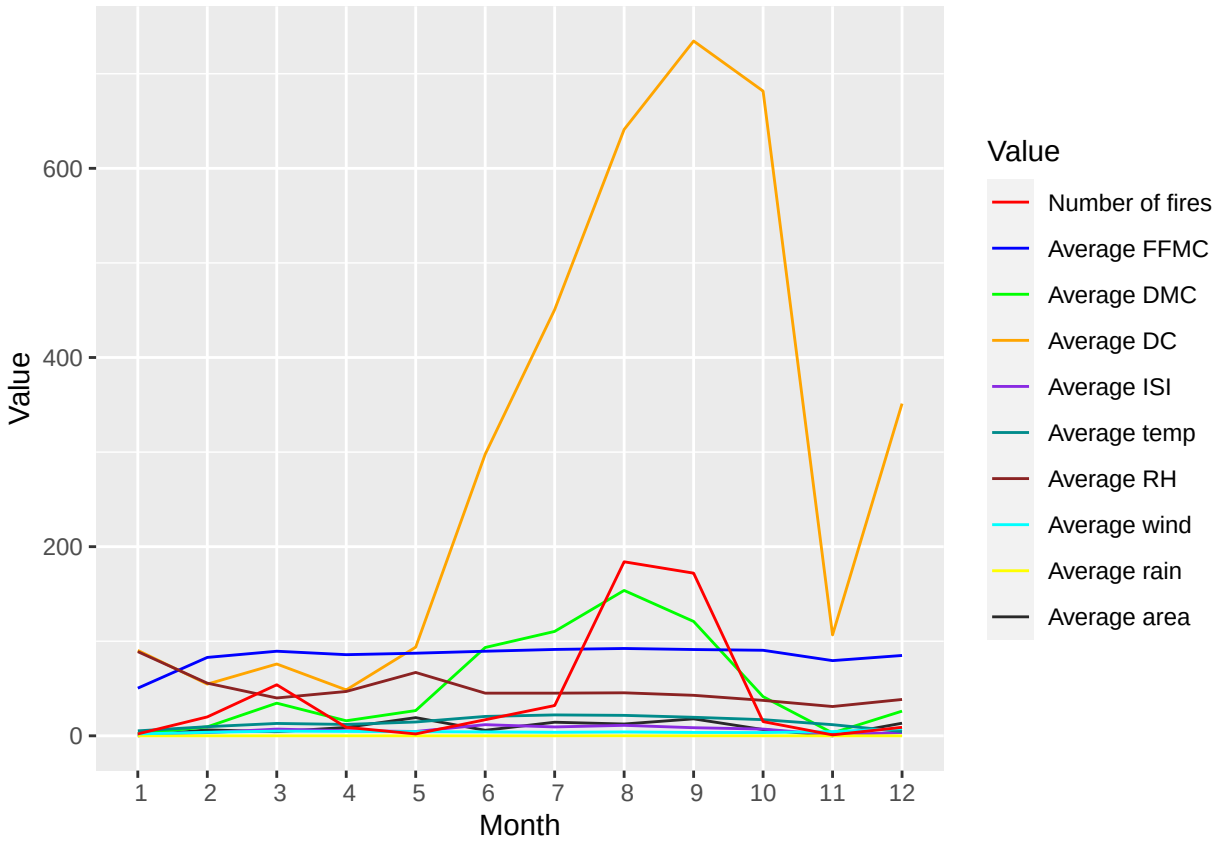
```
#Pivoting tibble fires_avg_by_month in preparation for
# using ggplot
fires_avg_by_month_pivoted <- fires_avg_by_month %>% pivot_longer(cols=c(
  n_fires,
  avg_FFMC,
  avg_DMC,
  avg_DC,
  avg_ISI,
  avg_temp,
  avg_RH,
  avg_wind,
  avg_rain,
  avg_area                                                    ),
names_to="FWI_var",
values_to='n_or_avg_value')

# Verifying the result of the pivoting
head(fires_avg_by_month_pivoted) %>%
  kable( caption="The first few rows of the pivoted tibble
         `fires_avg_by_month_pivoted` obtained from the tibble
         `fires_avg_by_month`:")
```

Table 4: The first few rows of the pivoted tibble `fires_avg_by_month_pivoted` obtained from the tibble `fires_avg_by_month`:

month_n	FWI_var	n_or_avg_value
1	n_fires	2.00
1	avg_FFMC	50.40
1	avg_DMC	2.40
1	avg_DC	90.35
1	avg_ISI	1.45
1	avg_temp	5.25

Finally, we plot the number of fires as well as the average values of the FWI variables for each month, all on the same plot but in different colors.

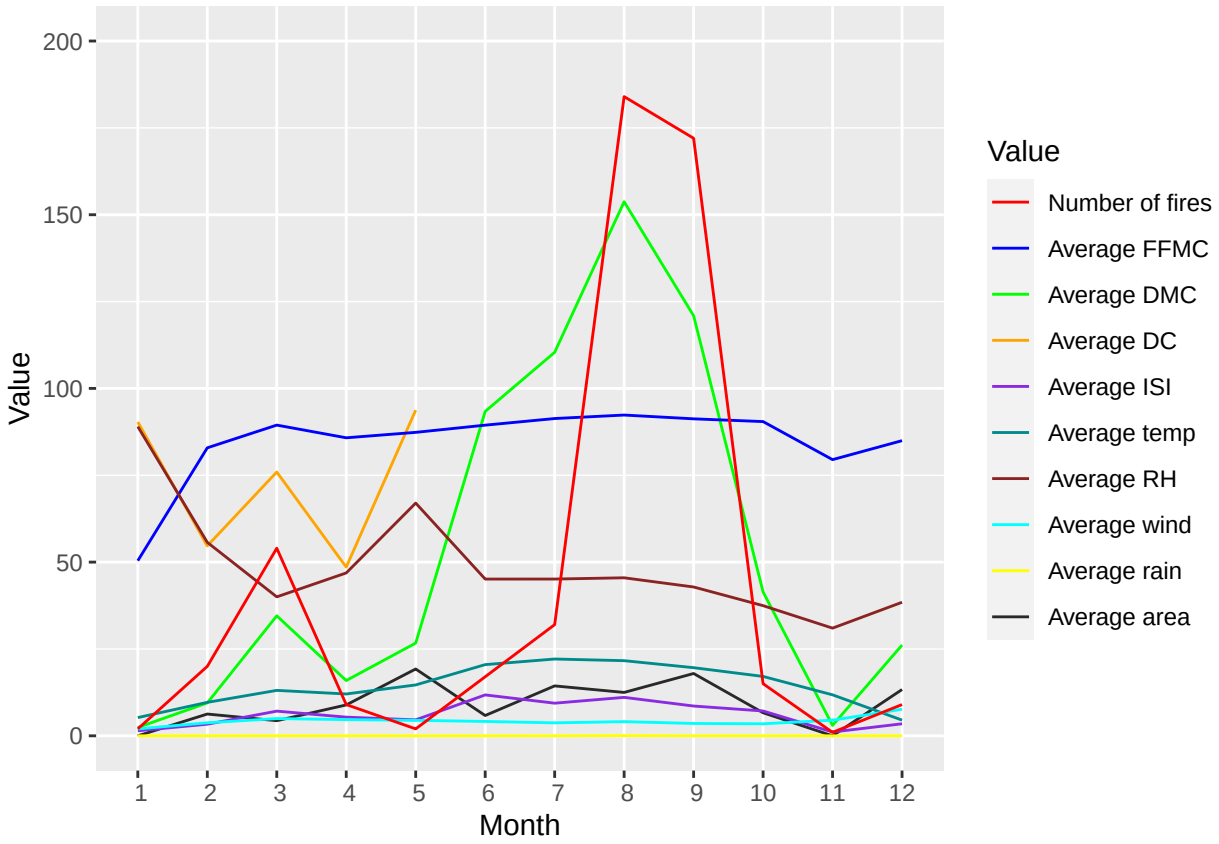
```
#Plotting fires_avg_by_month_pivoted
plot_1 <- fires_avg_by_month_pivoted %>%
ggplot(aes(x=month_n,y=n_or_avg_value, color=FWI_var))+
scale_x_discrete("Month",
  limits=c(1,2,3,4,5,6,7,8,9,10,11,12))+
geom_line()+
labs(x='Month',y='Value')+
scale_color_manual(
  name = "Value",
  values = c("red","blue","green","orange","blueviolet",
     "cyan4","brown4", "cyan","yellow","gray16"),
  breaks = c(
    "n_fires",
    "avg_FFMC",
    "avg_DMC",
    "avg_DC",
    "avg_ISI",
   "avg_temp",
   "avg_RH",
    "avg_wind",
   "avg_rain",
    "avg_area"),
  labels = c(
   "Number of fires",
    "Average FFMC",
    "Average DMC",
    "Average DC",
   "Average ISI",
    "Average temp",
    "Average RH",
   "Average wind",
    "Average rain",
    "Average area")
)

# Displaying the plot
plot_1
```



From the plot, it appears that the number of the fires in a given month and the average DC (Drought Code index) in that month are related. Indeed, the average DC index had higher values in August and September when there were more fires.

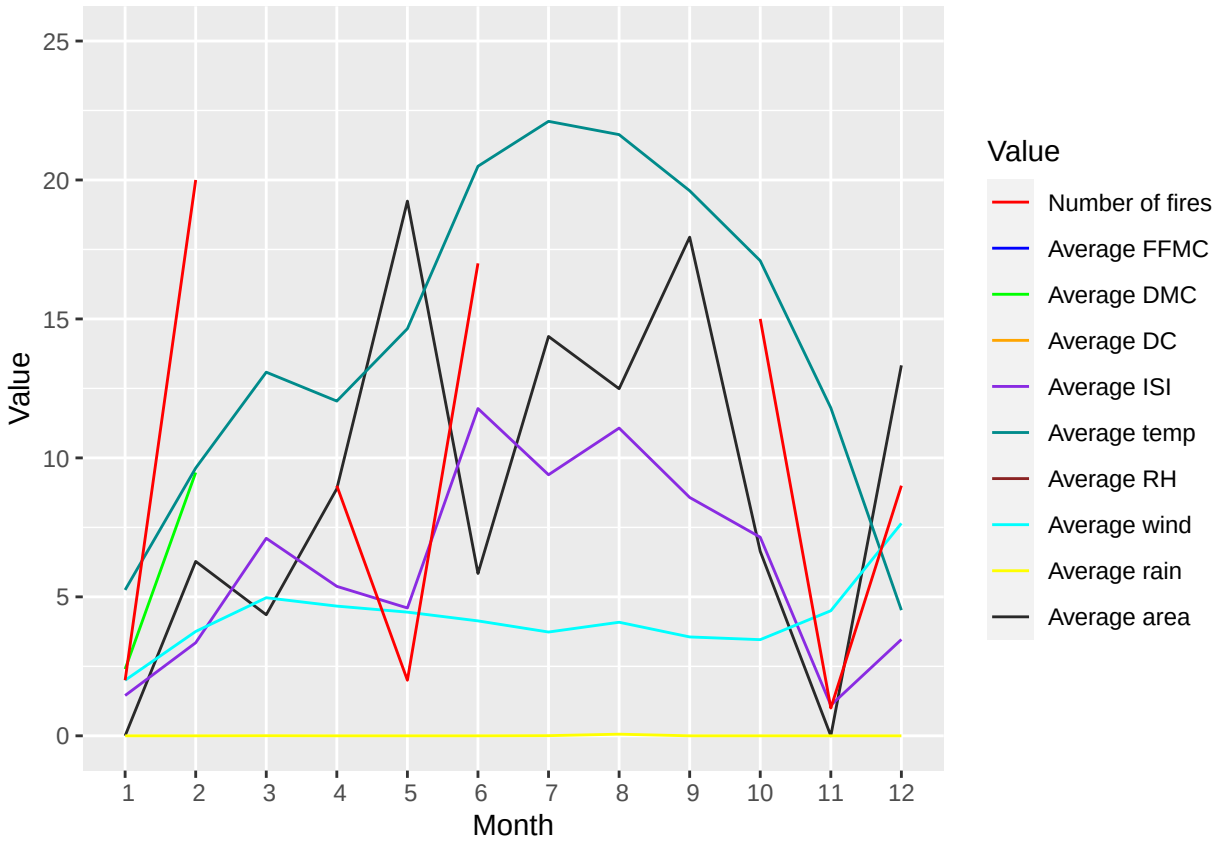
Because the average DC values are much higher than the values of all the other variables plotted above, we will restrict the y-axis values plotted to the range from 0 to 200.

```
# Restricting the y-values to the range from 0 to 200
plot_1+ylim(0, 200)
```



From the last plot, we see that the average DMC (Duff Moisture Code index) is higher during the months with higher numbers of fires. We further restrict the displayed values on the y-axis to 0-25.

```
# Restricting the y-values to the range from 0 to 25
plot_1+ylim(0, 25)
```



From the last plot, we see that the values of the average temperature, the average ISI (Initial Spread Index), and the average area were higher during the months with higher numbers of fires.

In summary, **the average DC (Drought Code index), DMC (Duff Moisture Code index), ISI (Initial Spread Index), temperature, and area were higher during the months with higher numbers of fires**.

Analyzing Scatter Plots of Each FWI Variable

In this subsection, we produce box plots of each of the variables. Recall that each box in a box plot shows the minimum and the maximum values of the given variable, the first and the third quartile, and the median.

We begin with pivoting the cleaned tibble forest_fires so as to transform it into a long format.

```
# Pivoting forest_fires in preparation for plotting the
# relationships of the fire areas and the remaining FFMC
# variables
forest_fires_pivoted <- forest_fires %>%
pivot_longer(cols=c(
    FFMC,
    DMC,
    DC,
    ISI,
    temp,
    RH,
    wind,
    rain,
    area
  ),
  names_to="FWI_var",
```

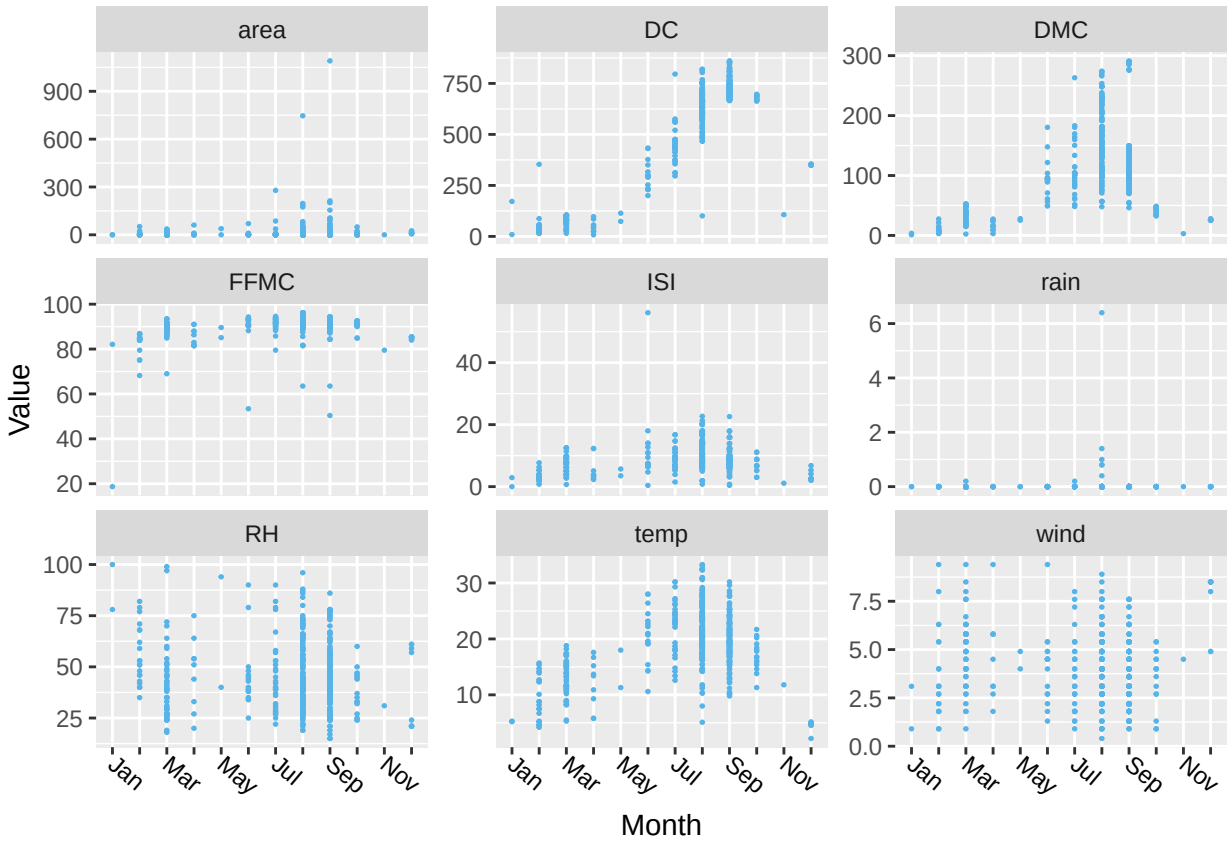
```
  values_to='Value')

# Verifying the results of the pivoting
head(forest_fires_pivoted) %>%
kable( caption="The first few rows of the pivoted tibble
       `forest_fires_pivoted` obtained from the cleaned
       original tibble `forest_fires`:")
```

Table 5: The first few rows of the pivoted tibble `forest_fires_pivoted` obtained from the cleaned original tibble `forest_fires`:

X	Y	month	day	month_n	FWI_var	Value
2	4	jan	sat	1	FFMC	82.1
2	4	jan	sat	1	DMC	3.7
2	4	jan	sat	1	DC	9.3
2	4	jan	sat	1	ISI	2.9
2	4	jan	sat	1	temp	5.3
2	4	jan	sat	1	RH	78.0

We use the `facet_wrap` function to produce all the desired plots at once.

```
# Plotting forest_fires_pivoted
plot_2 <- forest_fires_pivoted %>%
ggplot(aes(x=month, y=Value))+
geom_point(color="#56B4E9",size=0.3)+
facet_wrap(ncol=3,vars(FWI_var), scales = "free_y")+
scale_x_discrete(labels=c("Jan", "", "Mar", "", "May",
  "","Jul", "",  "Sep", "", "Nov", "")) +
ylab("Value") +
xlab("Month") +
theme(axis.text.x = element_text(size=9, angle = -40,
  vjust = 0.5, hjust = 0, color = "black")
  # rotate and move x tick labels (months)
  )

# Displaying the plot
plot_2
```



From the above scatter plots, we conclude that the **DC, DMC, ISI, and temperature values seem to be higher in August and September, the two months with more fires**.

Relationships between the FWI Variables and the Severity of the Fires

In this section, we analyze the relationships between the variables represented by the columns in the original dataframes and the severity of the fires. We deem fires with larger area burned more severe. To perform this analysis, we produce scatter plots that have the area of the fire on the x-axis and another FWI variable on the y-axis.

Before plotting all the above relationships, we pivot the cleaned tibble forest_fires accordingly.

```
# Pivoting forest_fires in preparation for plotting the
# relationship between the burned area and the other FFMC
# variables
forest_fires_pivoted <- forest_fires %>%
pivot_longer(cols=c(
    FFMC,
    DMC,
    DC,
    ISI,
    temp,
    RH,
    wind,
    rain
  ),
  names_to="FWI_var",
  values_to='Value')
```

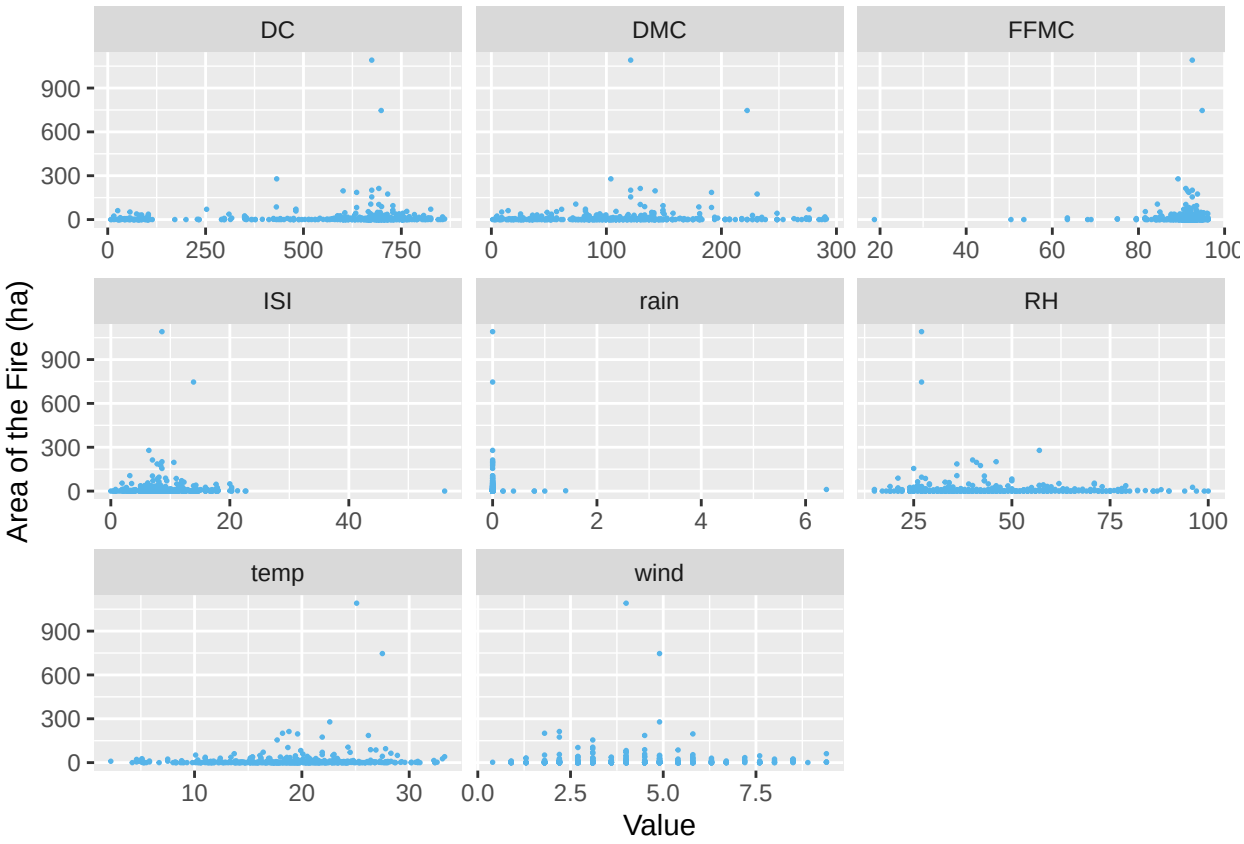
```
# Verifying the result of the pivoting
head(forest_fires_pivoted)%>%
kable( caption="The first few rows of the
       `forest_fires_pivoted` data-frame created from the
       cleaned original `forest_fires` data-frame:")
```

Table 6: The first few rows of the `forest_fires_pivoted` data-frame created from the cleaned original `forest_fires` data-frame:

X	Y	month	day	area	month_n	FWI_var	Value
2	4	jan	sat	0	1	FFMC	82.1
2	4	jan	sat	0	1	DMC	3.7
2	4	jan	sat	0	1	DC	9.3
2	4	jan	sat	0	1	ISI	2.9
2	4	jan	sat	0	1	temp	5.3
2	4	jan	sat	0	1	RH	78.0

As in the previous subsection, we use the `facet_wrap` function to produce all the desired plots at once.

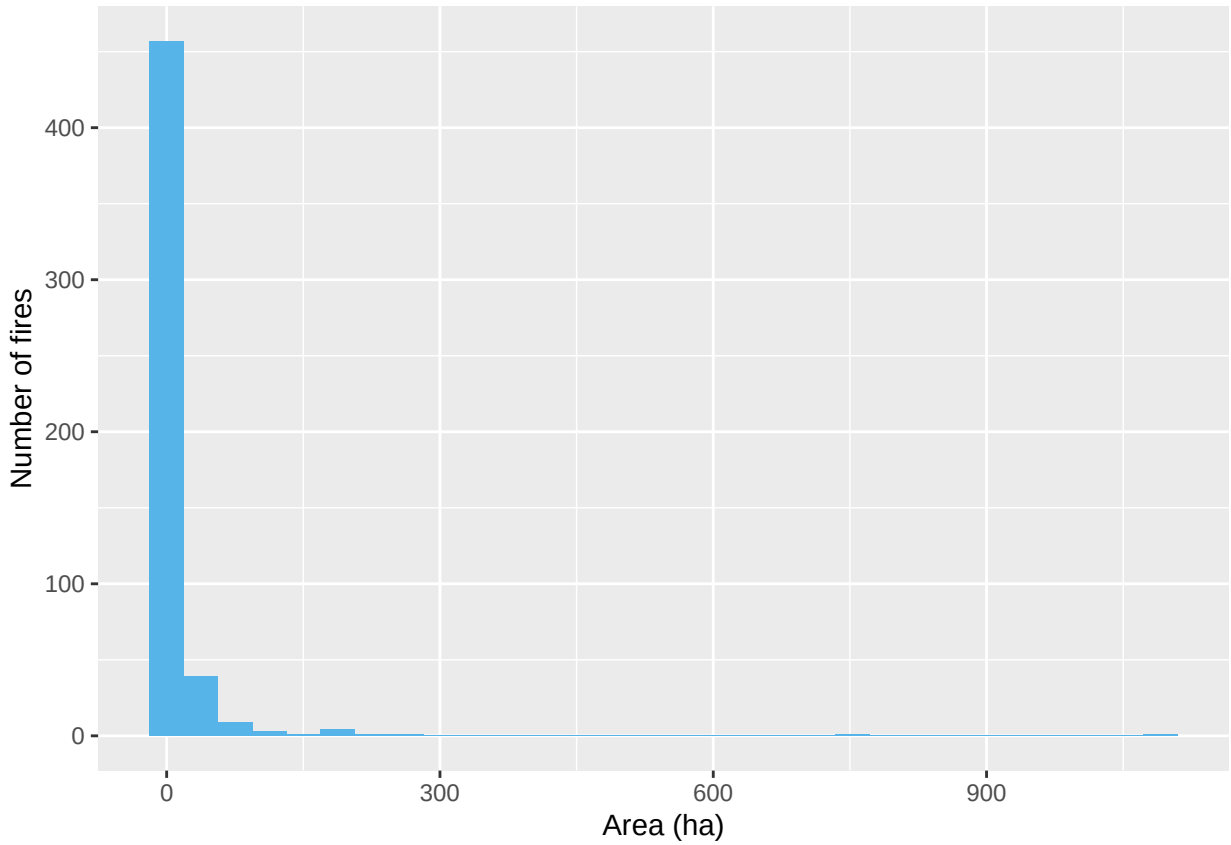
```
# Plotting forest_fires_pivoted using facet_wrap
plot_3 <- forest_fires_pivoted %>%
ggplot(aes(x=Value, y=area))+
geom_point(color="#56B4E9",size=0.3)+
facet_wrap(ncol=3,vars(FWI_var), scales = "free_x")+
ylab("Area of the Fire (ha)") +
xlab("Value")

# Displaying the plot
plot_3
```



There are two fires with a very large area such that the points on the plots corresponding to these fires obscure our ability to see the remaining points on the plots clearly.

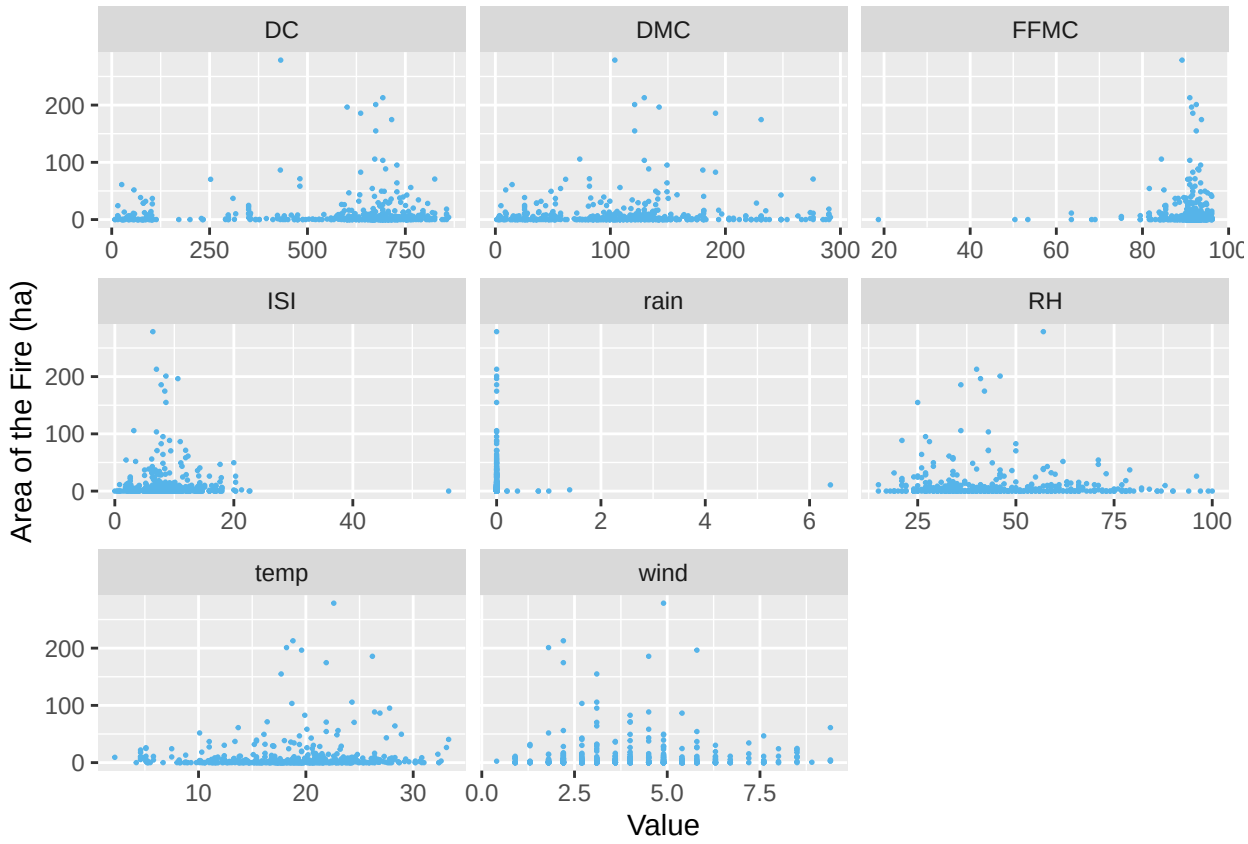
We confirm the obstruction created by the two outliers with a very high fire area by observing the following histogram. In the histogram, we plot the fire area on the x axis and the number of the fires with the given area on the y-axis.

```
# Plotting a histogram of the number of fires of given areas
plot_4 <- forest_fires %>%
ggplot(aes(x=area))+
geom_histogram(fill="#56B4E9",bins=30)+
ylab("Number of fires") +
xlab("Area (ha)")

# Displaying the plot
plot_4
```



We observe that the vast majority of the fires have zero area and that there are two outlier fires with the fire area above 600 ha. Given the above, we remove the two fires with the highest areas from the tibble `forest_fires_pivoted` and plot the data again.

```
# Removing the rows in forest_fires_pivoted that correspond
# to two most severe fires (top two largest areas)
forest_fires_adjusted <- forest_fires_pivoted %>% arrange(-area) %>% tail(-16)

# Verifying the result of the removal
head(forest_fires_adjusted)

# Plotting forest_fires_adjusted
plot_5 <- forest_fires_adjusted %>%
ggplot(aes(x=Value, y=area))+
geom_point(color="#56B4E9",size=0.3)+
facet_wrap(ncol=3,vars(FWI_var), scales = "free_x")+
ylab("Area of the Fire (ha)") +
xlab("Value")

# Displaying the plot
plot_5
```



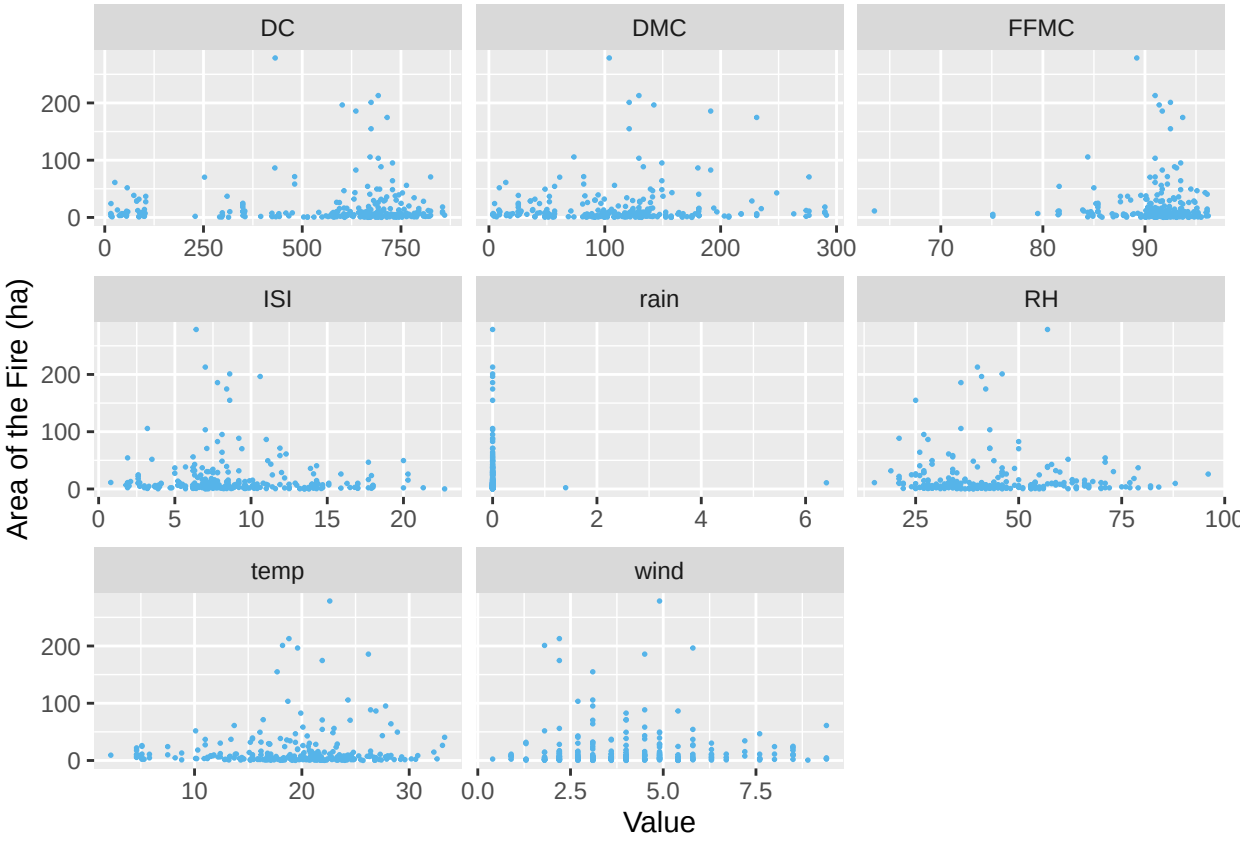

```
## # A tibble: 6 x 8
##       X     Y month day    area month_n FWI_var Value
##   <int> <int> <fct> <fct> <dbl>   <dbl> <chr>   <dbl>
## 1     7     4 jul   mon    279.       7 FFMC     89.2
## 2     7     4 jul   mon    279.       7 DMC     104.
## 3     7     4 jul   mon    279.       7 DC      432.
## 4     7     4 jul   mon    279.       7 ISI       6.4
## 5     7     4 jul   mon    279.       7 temp     22.6
## 6     7     4 jul   mon    279.       7 RH       57
```

From the above plots, we observe that **larger FFMC (Fine Fuel Moisture Code index) values are associated with more severe fires. The same holds for the DC (Drought Code index) and the temperature values. Lower rain values are associated with more severe fires. Counter-intuitively, quite severe fires occurred at average ISI (Initial Spread index) values. Severe fires occurred under low to average values of the wind speed**.

We would like to also filter out the fires with the zero area and observe the same relationships between the variables again. The two fires with the largest areas are still removed from the data.

```
# Removing rows with two largest burned areas as well as
# zero burned areas from forest_fires_pivoted
forest_fires_adjusted <- forest_fires_pivoted %>% arrange(-area) %>%
tail(-16) %>%
filter(area!=0) %>%
arrange(area)

# Plotting forest_fires_adjusted
plot_6 <- forest_fires_adjusted %>%
```

```
ggplot(aes(x=Value, y=area))+
geom_point(color="#56B4E9",size=0.3)+
facet_wrap(ncol=3,vars(FWI_var), scales = "free_x")+
ylab("Area of the Fire (ha)") +
xlab("Value")

# Displaying the plot
plot_6
```



From the above plots, we draw the same conclusions as before.

Conclusion

Here, we answer the questions posed in the Introduction.

- **Q**: What are months with exceptionally high quantity of fires?
- **A**: The vast majority of the fires occurred in August and September. More precisely, the number of fires in August and September was 356 and there were 152 fires in all the remaining months.That is, 87% of all the fires occurred in August and September.

- **Q**: On which weekdays there were more fires?
- **A**:There were more fires on the weekends. More precisely, there were 264 fires on Fridays, Saturdays, and Sundays, and 253 fires on the other weekdays. That is, 51% of all the fires occurred on weekends.

- **Q**:How are Canadian Forest Fire Weather Index (FWI) variables related to the number of fires in a given months?
- **A**: The average DC (Drought Code index), DMC (Duff Moisture Code index), ISI (Initial Spread Index), temperature, and area were higher during the months with higher numbers of fires. The two most severe fires with the area burned being larger than 600 ha occurred under the following conditions: high DC, FFMC (Fine Fuel Moisture Code index), and temperature values, low rain and RH (Relative Humidity) values, and medium DMC and wind speed values. Larger FFMC values are associated with more severe fires. The same holds for the DC (Drought Code index) and the temperature values. Lower rain values are associated with more severe fires. Counter-intuitively, quite severe fires occurred at average ISI (Initial Spread index) values. Severe fires occurred under low to average values of the wind speed.