Analysis Of Factors That Affect Forest Fires

2022-10-23

Introduction

More than 80% of earth's biodiversity can be found in forests and so forest fires poses a huge ecological threat. Not only are forest fires a threat to the ecosystem, they can also pose economic threat as well as lead to human fatalities.

One of the key ways of understanding forest fires is by understanding the underlying factors that cause these forest fires. To do this we will be working with a data set collected by `The Department of Information Systems/R&D Algoritmi Centre, University of Minho, 4800-058 Guimaraes, Portugal` used in a scientific research paper for predicting forest fires using modelling techniques. The data set contains the following attributes

- 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: Outside temperature in Celsius degrees: 2.2 to 33.30
- RH: Outside relative humidity in percentage: 15.0 to 100
- wind: Outside 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

The FWI system (Fire Weather Index) is a system that was created in the 1970s and it involves simple calculations using readings from meteorological such as temperature, relative humidity, rain and wind that could be manually collected in weather stations.

The goal of this analysis is to find out patterns between each of these factors and forest fires. Primarily, we are concerned with knowing:

- Which months forest fires occur the most.
- Which day of the week forest fires occur the most.
- Is there a relationship between th FWI factors and the months with the most forest fires?
- Which of these FWI factors affect the severity of the fire the most.

```
# loading the libraries
library(tidyverse)
library(kableExtra)
```

```
# creating function to display tables
render_table <- function(table, scale_down=F){
  if(scale_down == T){
     rendered_table <- kbl(table) %>% kable_styling(
    latex_options = c("stripe", "HOLD_position", "scale_down")
  )
  } else{
   rendered_table <- kbl(table) %>% kable_styling(
    latex_options = c("stripe", "HOLD_position")
  )

  }

  return(rendered_table)
}
```

Data Exploration

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

```
# getting an overview of the data
forest_fires %>% glimpse()
```

```
## 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~
```

The data set has 517 rows and 13 columns. The glimpse function also shows us the data type of each of the column. Next we are going to check each column for null values.

```
column_names <- colnames(forest_fires)

# checking for null values
for(col in column_names){
  null_value <- forest_fires %>% pull(col) %>% is.na() %>% sum()
  paste(col, ":", null_value) %>% print()

}
```

```
## [1] "X : 0"
## [1] "Y : 0"
## [1] "month : 0"
## [1] "day : 0"
## [1] "FFMC : 0"
## [1] "DMC : 0"
## [1] "DC : 0"
## [1] "ISI : 0"
## [1] "temp : 0"
## [1] "RH : 0"
## [1] "wind : 0"
## [1] "rain : 0"
## [1] "area : 0"
```

We have no null value in any of the columns.

```
forest_fires %>% head() %>% render_table()
```

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

Converting Day and Month Column to Categorical Variables

```
# getting the unique values in the month and day columns
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"
```

One thing we have to do is convert the month and day column to categorical variables so they can follow a specified order when we plot them.

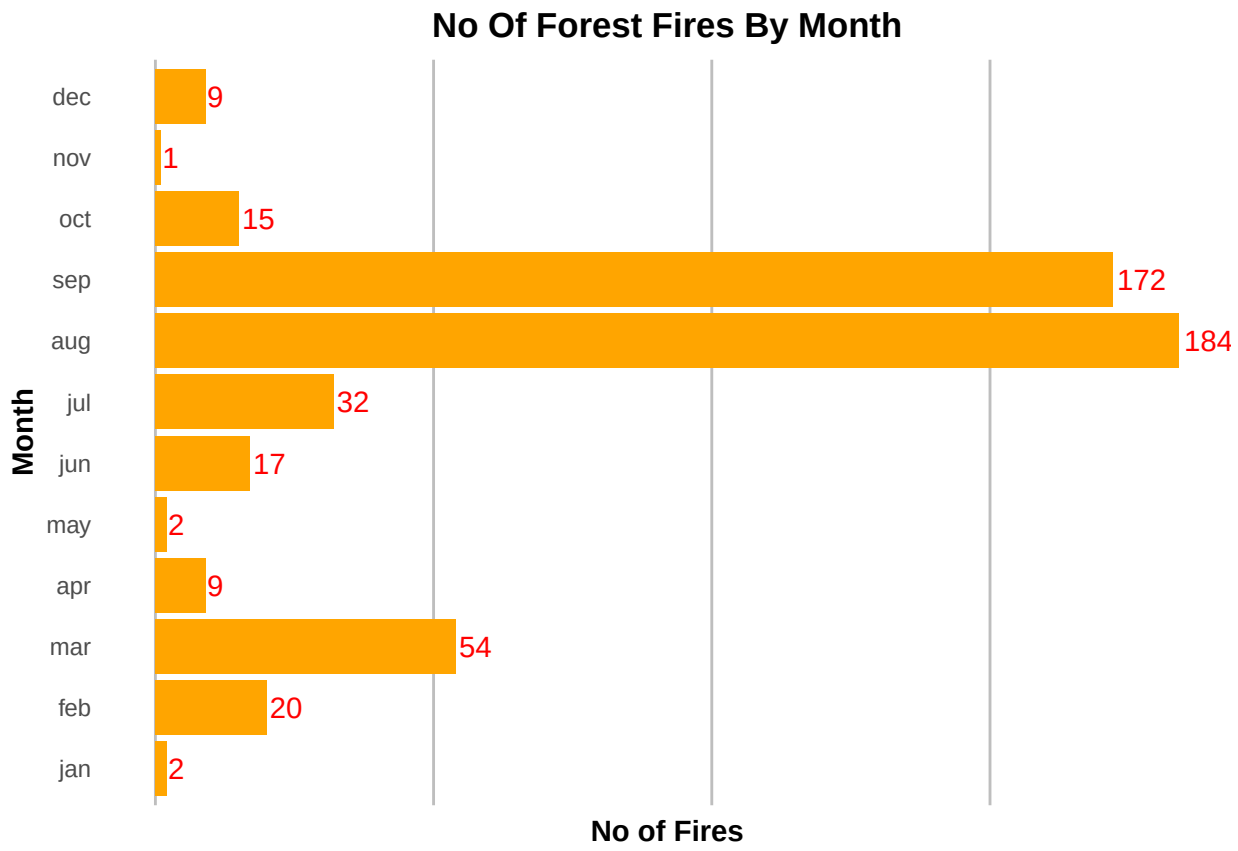
```
# converting the month and day columns to categorical variables

months = c("jan", "feb", "mar", "apr",
```

```
            "may", "jun", "jul", "aug",
            "sep", "oct", "nov", "dec")

days = c("mon", "tue", "wed", "thu",
         "fri", "sat", "sun")

forest_fires <- forest_fires %>% mutate(
  day = factor(day, levels = days),

  month = factor(month, levels = months)
)
```

Which Months Do Forest Forest Fires Occur The Most Frequently?

To figure out which of the months have the most forest fire occurrences, we first have to group the data by month and then take the count for each month using the `n()` function.

```
# grouping the data by month
month_summary <- forest_fires %>% group_by(month) %>% summarise(
  no_of_fires = n()
) %>% arrange(no_of_fires)

# creating plot showing forest fire relationship with month
month_summary %>% ggplot(aes(x= no_of_fires, y = month)) +
geom_col(fill = "orange") +
labs(
  title = "No Of Forest Fires By Month",
  x = "No of Fires",
  y = "Month"
) +
geom_text(aes(label = no_of_fires), hjust = -0.1, color="red") +
theme(
  plot.title = element_text(face = "bold", hjust = 0.5),
  axis.title = element_text(face = "bold"),
  axis.ticks = element_blank(),
  axis.text.x = element_blank(),
  panel.background = element_rect(fill = "white"),
  panel.grid.major.x = element_line(color="gray", size=0.5)
)
```



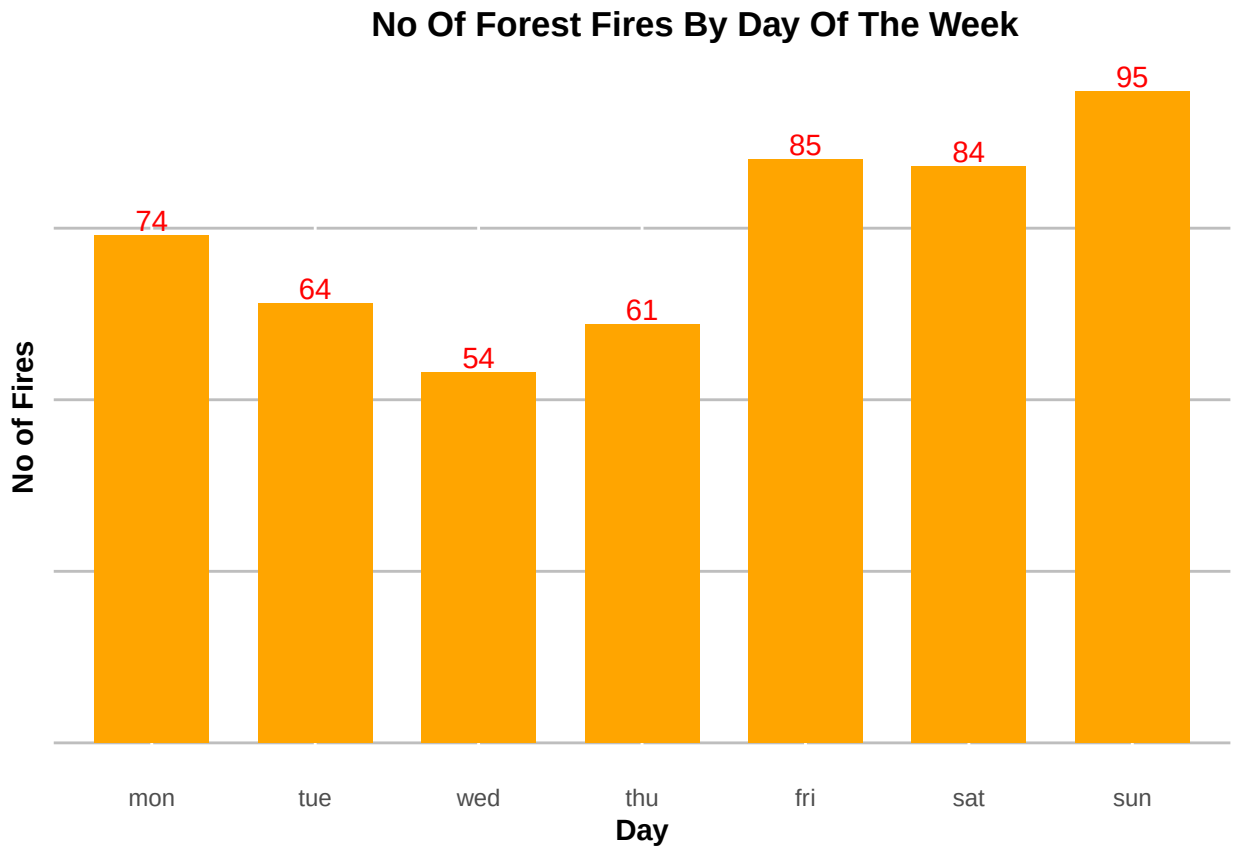
Fires were more likely to occur in August and September with each month having 184 and 172 forest fires respectively.

Which Days Do Forest Fires Occur The Most Frequently?

We are going to repeat the same process like we did for the months , just that this time, we will be grouping the data set by day instead of month.

```
# grouping the data by day
day_summary <- forest_fires %>% group_by(day) %>% summarise(
  no_of_fires = n()
) %>% arrange(no_of_fires)

# creating plot showing forest fire relationship with day of the week
day_summary %>% ggplot(aes (x = day, y= no_of_fires)) +
  geom_col(width = 0.7, fill = "orange") +
  labs(
    title = "No Of Forest Fires By Day Of The Week",
    x =  "Day",
    y = "No of Fires"
  ) +
  geom_text(aes(label = no_of_fires), vjust = -0.2, color="red") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.6),
    axis.title = element_text(face = "bold"),
    axis.ticks = element_blank(),
    axis.text.y = element_blank(),
```

```
    panel.background = element_rect(fill = "white"),
    panel.grid.major.y = element_line(color="gray", size=0.5)
  )
```



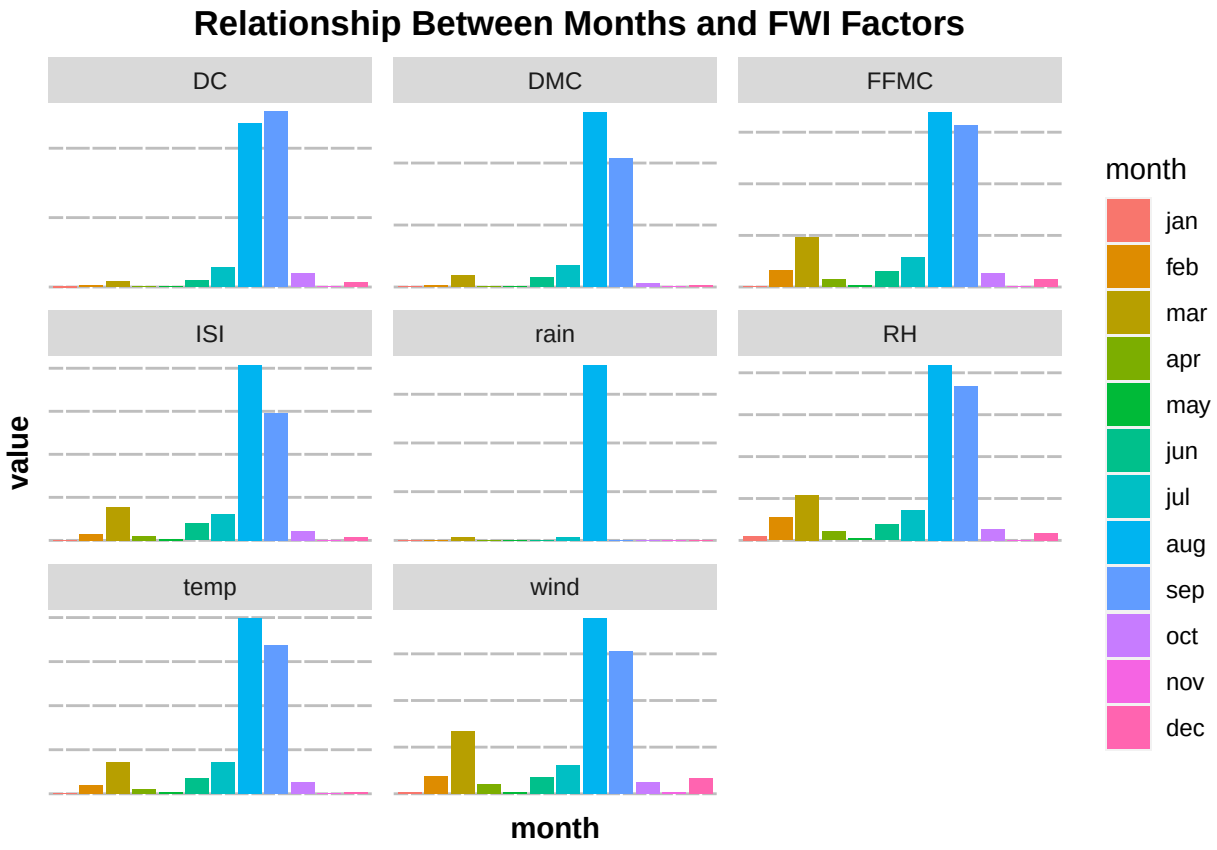
There tend to be more forest fires towards the end of the week. Friday, Saturday and Sunday had the most forest fires occurrences with 85, 84 and 95 forest fires respectively.

Relationship Between Months and FWI Factors

As we have previously seen, the months of August and September tend to have the most forest fire occurrences. Here we want to look at the values of the FWI factors for each of the month. To be able to visualize this, we re going to turn our table into a long format using the `pivot_longer()` function. We are also going to be using ggplot `facet_wrap()` function to create multiple subplots.

```
# converting the data to long format
forest_fires_long = forest_fires %>% pivot_longer(
  cols = c(FFMC, DMC, DC,
           ISI, temp, RH, wind, rain),
  names_to = "factors",
  values_to = "value"
)

# plot showing relationship between FWI factors And month
forest_fires_long %>% ggplot(
  aes(x=month, y=value, fill = month)
) +
geom_col() +
```

```
  labs(title = "Relationship Between Months and FWI Factors") +
facet_wrap(
    scales = "free_y",
    vars(factors)
  ) +

  theme(
    plot.title = element_text(face = "bold", hjust = 0.6),
    axis.title = element_text(face = "bold"),
    axis.ticks = element_blank(),
    axis.text = element_blank(),
    panel.background = element_rect(fill = "white"),
    panel.grid.major.y = element_line(color="gray", size=0.5)
  )
```



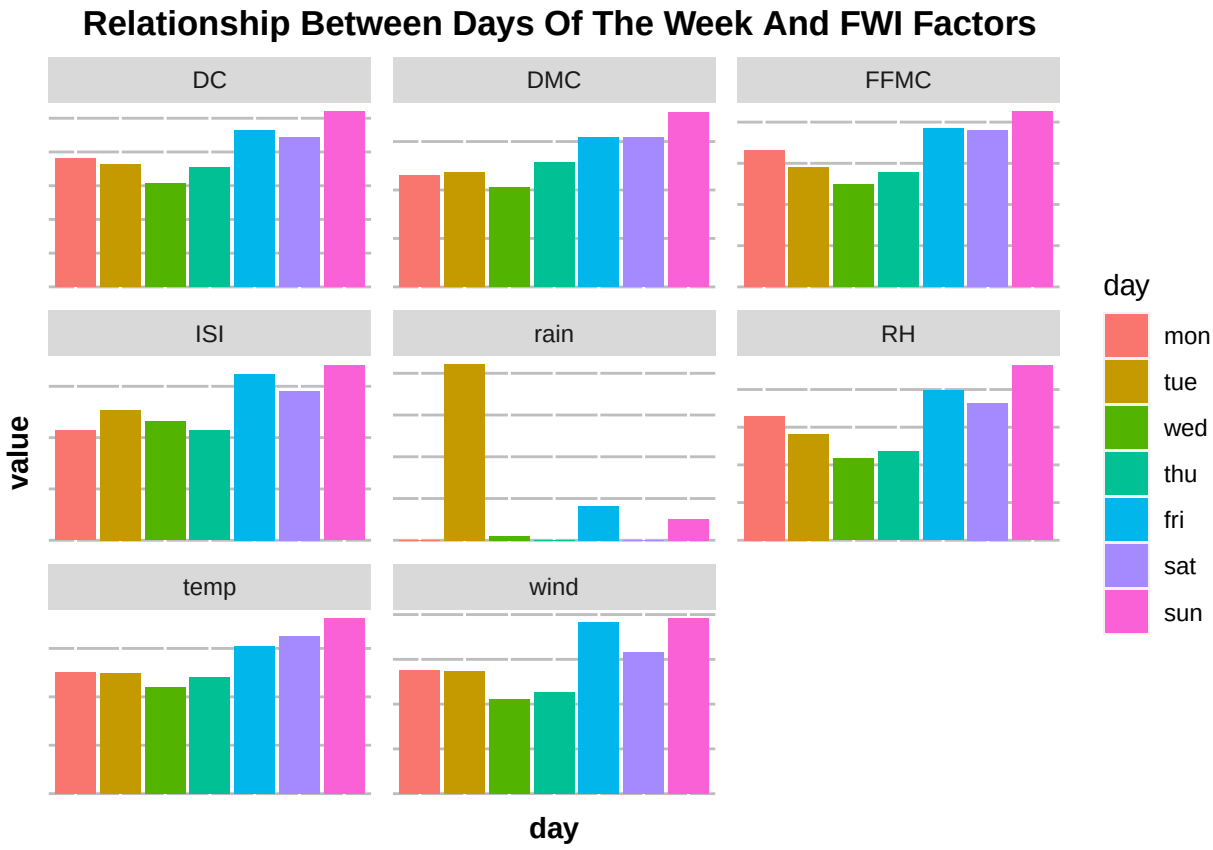
Looking at each individual factors, we can see that we have the highest values in the months of August and September which are the months that most forest fires occur in.

Relationship Between Days Of The Week And FWI Factors

Just as we have seen with the months, the months with the highest number of forest fires tend to have the highest values for the FWI factors. We will investigate the days of the week and see if it holds true too.

```
# plot showing relationship between FWI factors and days of the week
forest_fires_long %>% ggplot(
  aes(x = day, y = value, fill = day)
) +
```

```
geom_col() +
  labs(title = "Relationship Between Days Of The Week And FWI Factors") +
facet_wrap(
    scales = "free_y",
    vars(factors)
  ) +

  theme(
    plot.title = element_text(face = "bold", hjust = 0.6),
    axis.title = element_text(face = "bold"),
    axis.ticks = element_blank(),
    axis.text = element_blank(),
    panel.background = element_rect(fill = "white"),
    panel.grid.major.y = element_line(color="gray", size=0.5)
  )
```



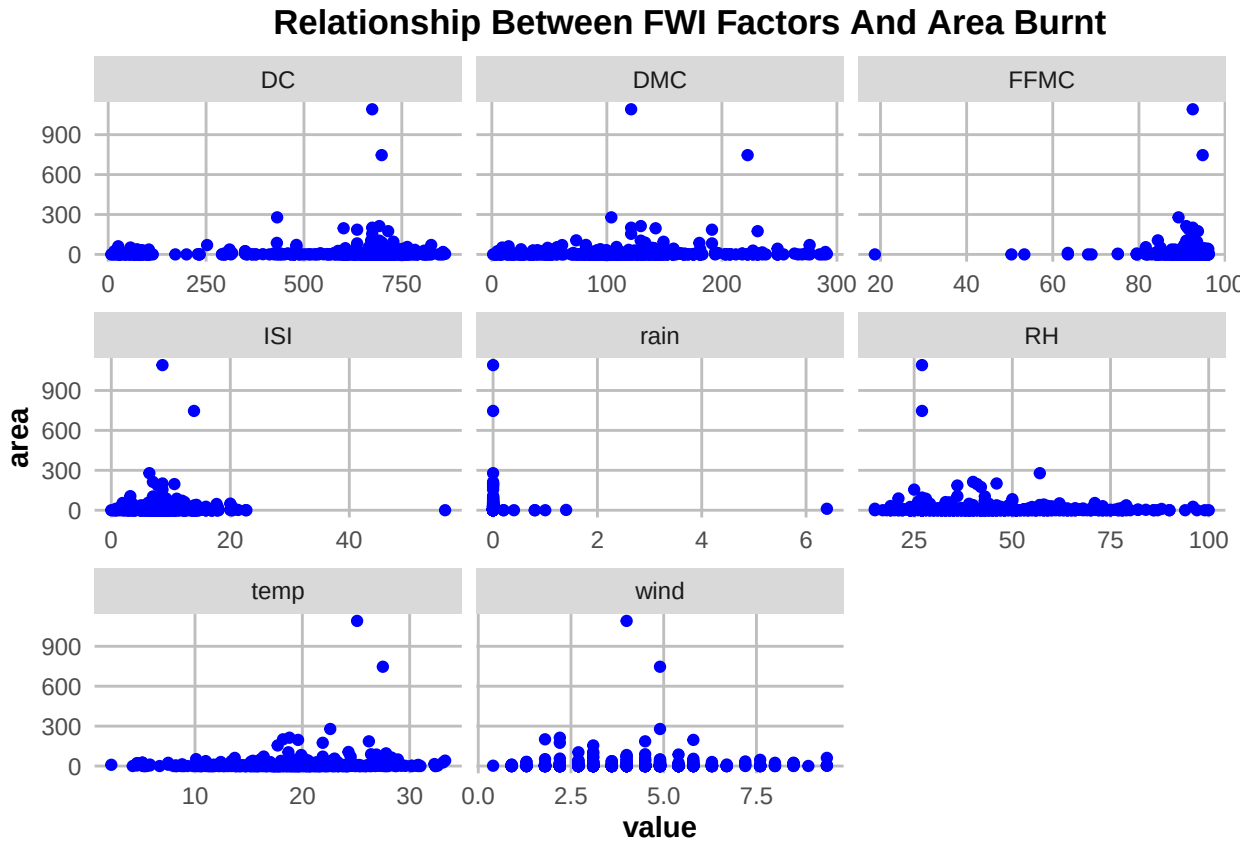
Just like the months, the days of the week that saw the most fires also had the highest values for most of the FWI factors.

Factors That Affect Forest Fire Intensity

In this case, we are going to define the intensity of the forest fire by the area of the forest burnt.

```
# scatter plot showing relationship between area burnt and FWI factors
forest_fires_long %>% ggplot(
  aes(x = value, y = area)
) +
```

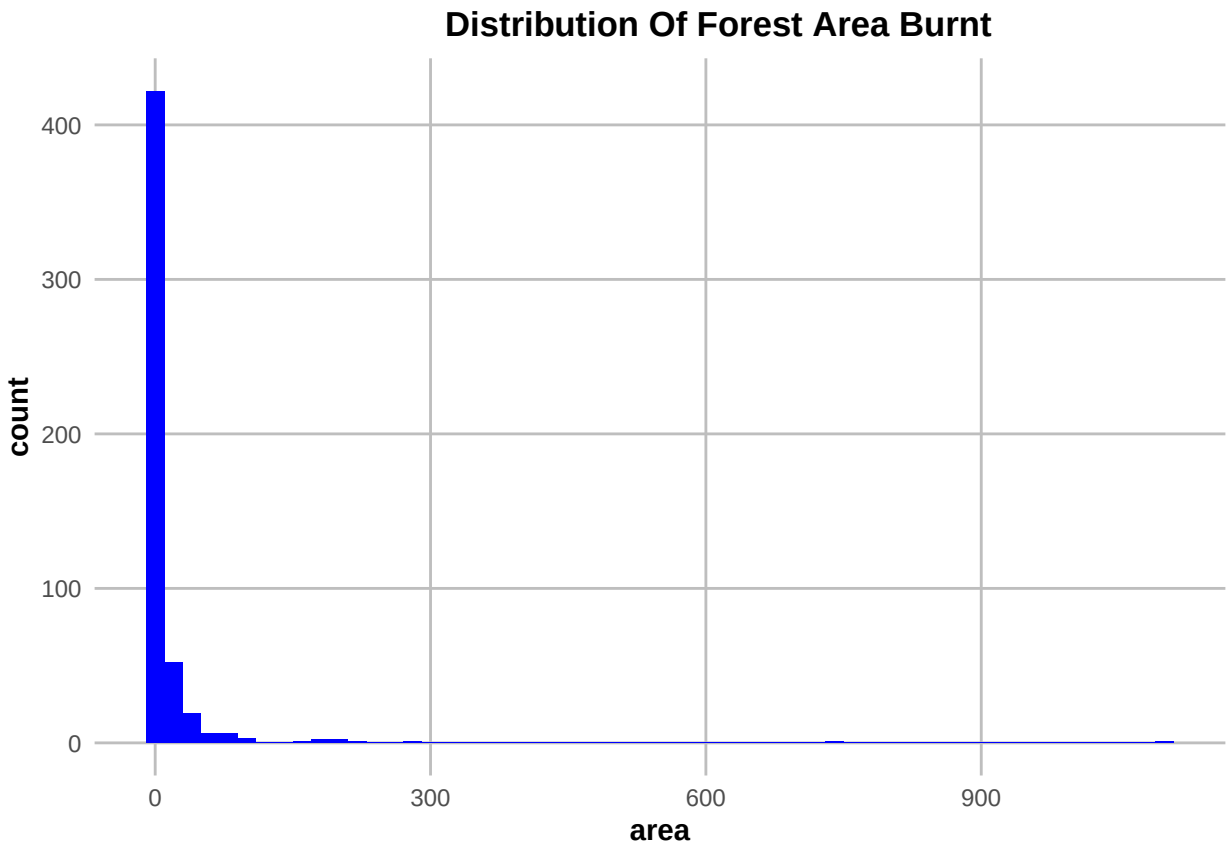
```
  labs( title = "Relationship Between FWI Factors And Area Burnt") +
geom_point(color = "blue") +
facet_wrap(
  scales = "free_x",
  vars(factors)
) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.6),
    axis.title = element_text(face = "bold"),
    axis.ticks = element_blank(),
    panel.background = element_rect(fill = "white"),
    panel.grid.major = element_line(color="gray", size=0.5)
  )
```



Most of the values are clustered towards the 0. This is probably because most of the forest fires didn't burn up to a hectare of the forest. We can visualize the area on a histogram to see how it is distributed.

```
# plotting histogram showing the distribution of area burnt
forest_fires %>% ggplot(
  aes(x = area)
) +
geom_histogram(fill = "blue", bins = 20, binwidth = 20) +
labs(title = "Distribution Of Forest Area Burnt") +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.6),
    axis.title = element_text(face = "bold"),
    axis.ticks = element_blank(),
```

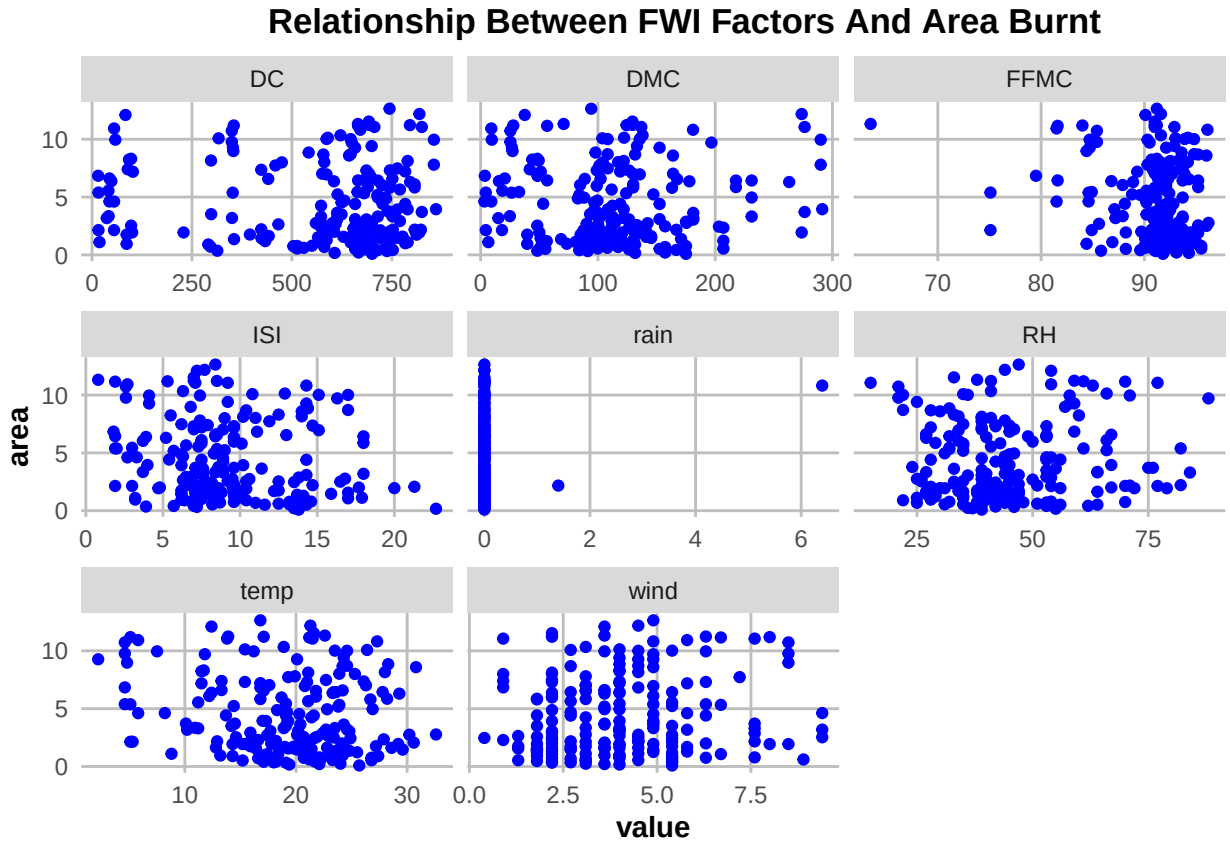
```
    panel.background = element_rect(fill = "white"),
    panel.grid.major = element_line(color="gray", size=0.5)
  )
```



The histogram confirms our assumption. Most of the values for the area are clustered around 0. We are going to recreate the scatter plot from before but this time we will only look at values greater than 0 and less than the mean area burnt.

```
mean_area <- forest_fires %>% pull(area) %>% mean()

# re-plotting relationship between FWI factors and area burnt
forest_fires_long %>% filter(area <= mean_area & area != 0) %>% ggplot(
  aes(x = value, y = area)
) +
  labs(title = "Relationship Between FWI Factors And Area Burnt") +
  geom_point(color = "blue") +
  facet_wrap(
    scales = "free_x",
    vars(factors)
  ) +
theme(
    plot.title = element_text(face = "bold", hjust = 0.6),
    axis.title = element_text(face = "bold"),
    axis.ticks = element_blank(),
    panel.background = element_rect(fill = "white"),
    panel.grid.major = element_line(color="gray", size=0.5)
```

```
)
```



After filtering and looking at the plot again, there doesn't seem to be any relationship that indicates that any of these factors affect the intensity of the fire when we look at it with respect to the area burnt.

Conclusion

Our goal with this analysis was to understand forest fires better and learn more about how different factors can lead to forest fires. We can draw the following conclusion from our analysis.

- Forest fires are most likely to occur in August and September. These months also have the highest values for the FWI factors.
- Forest fires were more prevalent on Fridays, Saturdays and Sundays which are the days with the highest values for the FWI factors.
- The FWI factors do not have any correlation to how intense a fire will be, rather they show the likelihood for a forest fire to occur. For example we can deduce from the scatter plot that a forest fire is most likely to start when the FFMC is above 80.