Forecasting Forest Fires

Salvatore Porcheddu

2 1 2021

Contents

Introduction

In this project we will explore and visualize information coming from a dataset on forest fires.

The data is associated with a scientific research paper on forest fire prediction in Portugal and can be found here.

Importing the necessary libraries and the dataset

```
library(dplyr)
library(ggplot2)
library(readr)
library(tidyr)

fires <- read_csv("forestfires.csv")
```

Column description and data exploration

All but two of the variables in the dataset have a numeric format. Some of them are self-explanatory, others require additional clarification.

The `X` and `Y` columns represent spatial coordinates.

The four columns `FFMC` (Fine Fuel Moisture Code), `DMC` (Duff Moisture Code), `DC` (Drought Code) and `ISI` (Initial Spread Index) all represent indexes that measure various risk and danger factors for forest fires: the higher their value, the higher the risk/danger.

The variable `RH` is a measure of the relative humidity (in percentage).

```
head(fires)
```

```
## # A tibble: 6 x 13
##       X     Y month day    FFMC   DMC    DC   ISI  temp    RH  wind  rain  area
##   <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     7     5 mar   fri    86.2  26.2  94.3   5.1   8.2    51   6.7   0       0
## 2     7     4 oct   tue    90.6  35.4 669.    6.7  18      33   0.9   0       0
## 3     7     4 oct   sat    90.6  43.7 687.    6.7  14.6    33   1.3   0       0
## 4     8     6 mar   fri    91.7  33.3  77.5   9     8.3    97   4     0.2     0
## 5     8     6 mar   sun    89.3  51.3 102.    9.6  11.4    99   1.8   0       0
## 6     8     6 aug   sun    92.3  85.3 488    14.7  22.2    29   5.4   0       0
```

```
glimpse(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 <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    <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~
```

Converting date variables to categorical variables

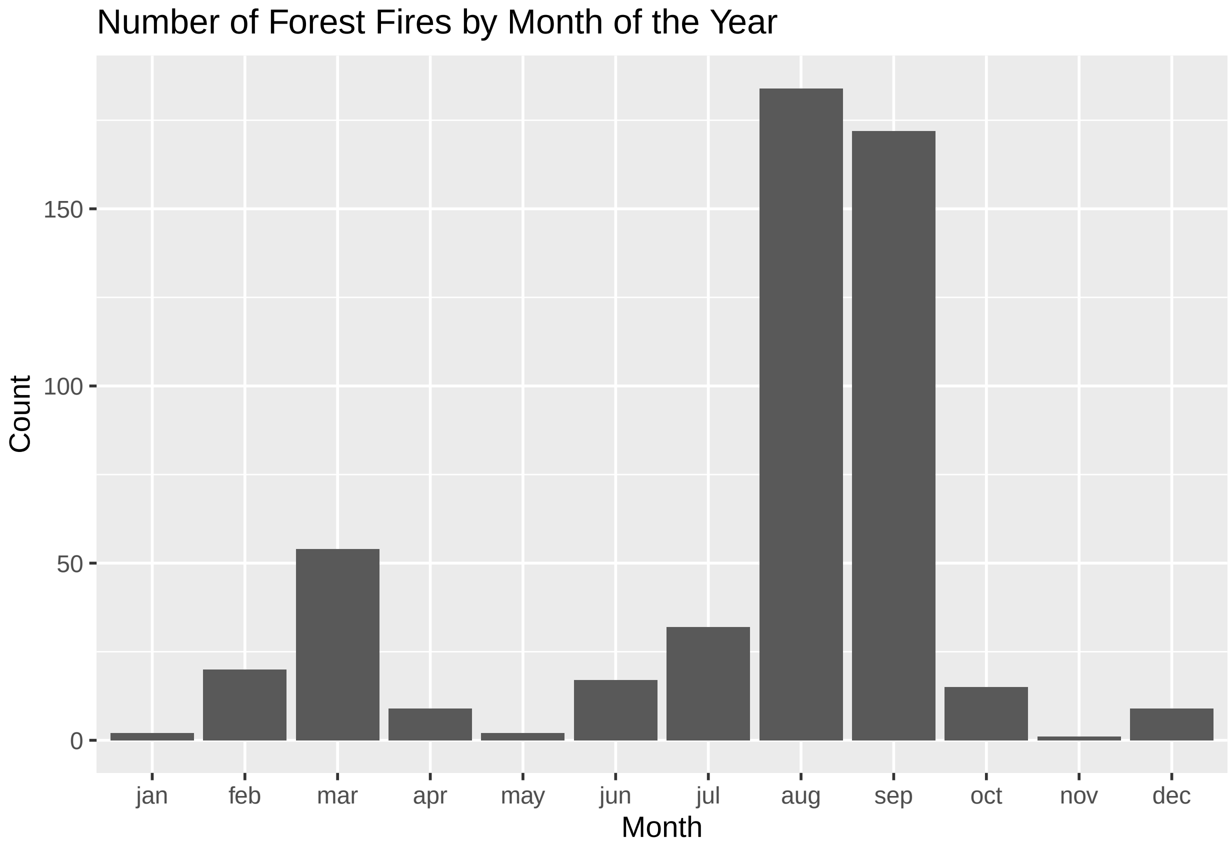
When using the `month` and `day` variables for our analysis and visualization needs, we want to avoid them being automatically sorted by alphabetical order. We can do this by converting these variables to categorical variables using the factor data type.

```
months <- c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep",
            "oct", "nov", "dec")
days <- c("mon", "tue", "wed", "thu", "fri", "sat", "sun")
```

```
firesf <- fires %>%
  mutate(month = factor(month, levels = months), day = factor(day,
                                                              levels = days))
```

Determining and plotting the number of fires by month/day of the week

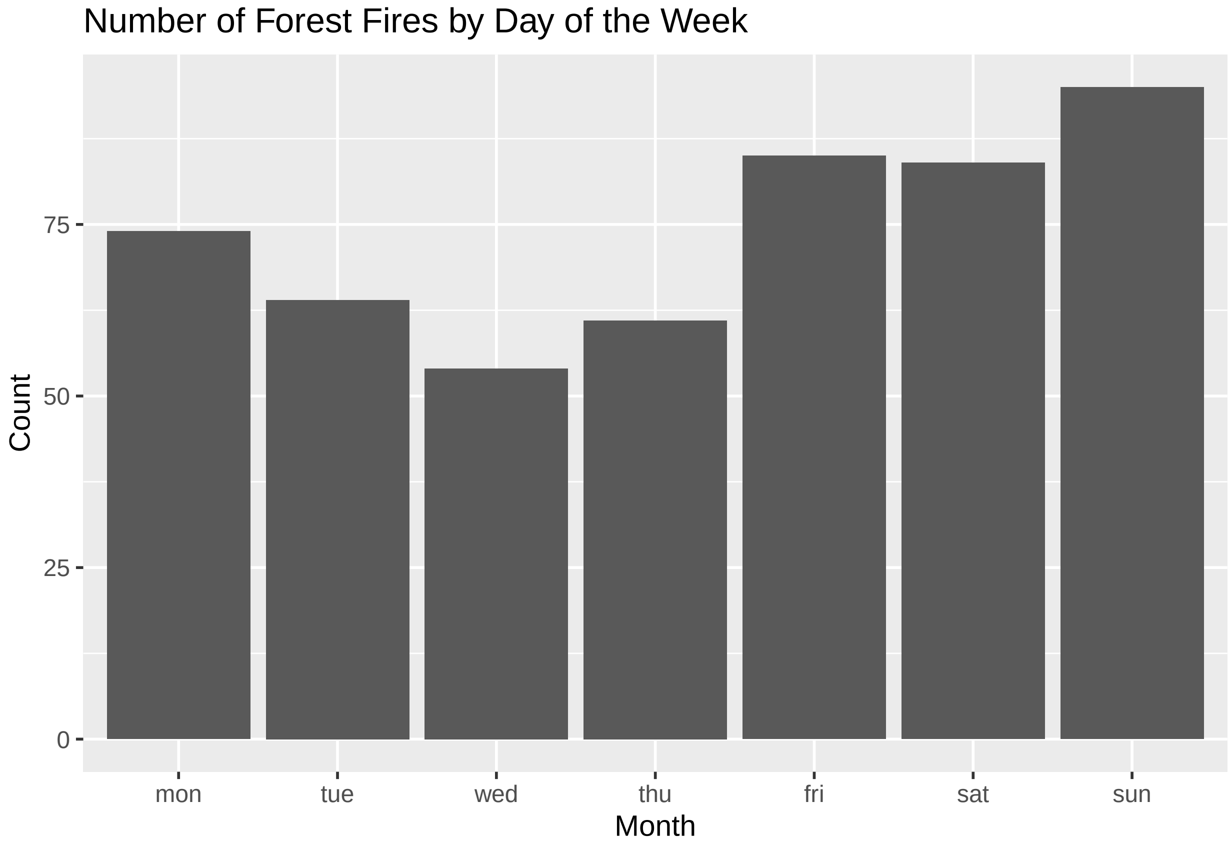
```
firesM <- firesf %>%
  group_by(month) %>%
  summarize(n = n())

firesD <- firesf %>%
  group_by(day) %>%
  summarize(n = n())

# we will start by plotting and describing the number of fires by month

firesM %>%
  ggplot(aes(x = month, y = n)) +
  geom_col() +
  labs(title = "Number of Forest Fires by Month of the Year", x = "Month",
       y = "Count")
```



The bar plot shows that the vast majority of fires occur in August and September; outside of the hottest months, we also register a significant amount of fires in March.

Let's now proceed and plot the number of fires by day of the week.

```
firesD %>%
  ggplot(aes(x = day, y = n)) +
  geom_col() +
  labs(title = "Number of Forest Fires by Day of the Week", x = "Month",
       y = "Count")
```



As we can see, although fires can of course occur any day of the week, Friday and the following weekend days seem to be particularly vulnerable.

Plotting the relationship between months and other variables of interest

Now that we have ascertained that August and September pose the highest risk for forest fires, we want to analyse this assumption on a deeper level by considering the reasons why these months show this heightened risk.

To do this, we will plot the relationship between the months of the year and several variables that have been shown by research to affect the risk/danger of forest fires; these are:

- Fine Fuel Moisture Code (`FFMC`), which represents the moisture content of surface litter and other cured fine fuels (higher values mean lower moisture content and so higher risk of forest fires);

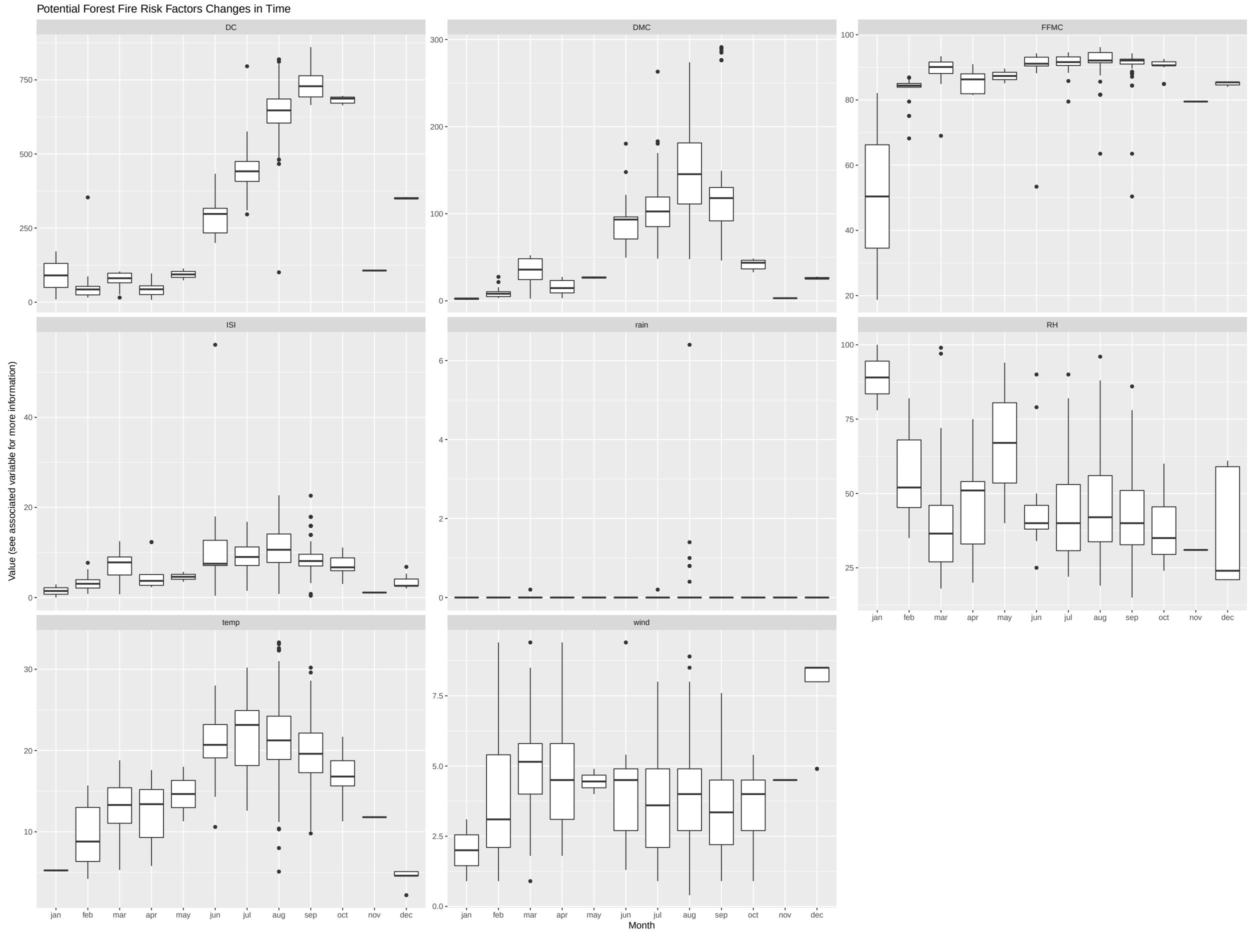
- Duff Moisture Code (DMC), which shows the average moisture content of loosely compacted organic layers of moderate depth (interpreted like the FFMC);
- Drought Code (DC), which gives the average moisture content of deep, compact, organic layers;
- Initial Spread Index (ISI), an indicator that shows the rate fire will spread in its early stages (higher values mean faster spreading);
- temperature (temp);
- humidity (RH);
- wind;
- rain.

Before plotting the variables, we need to prepare the data by formatting it long instead of wide.

```
firesL <- firesf %>%
  pivot_longer(cols = !c(X, Y, month, day, area), names_to = "variable",
               values_to = "value")

head(firesL, 8)
```

```
## # A tibble: 8 x 7
##       X     Y month day    area variable 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
```

```
firesL %>%
  ggplot(aes(x = month, y = value)) +
  geom_boxplot() +
  facet_wrap(vars(variable), scales = "free_y") +
  labs(title = "Potential Forest Fire Risk Factors Changes in Time",
       x = "Month", y = "Value (see associated variable for more information)")
```



Four of the variables show significant spikes during summer and in August and September in particular: the DC, DMC and FFMC indexes and the temperature. Conversely, humidity rates tend to be lower in summer, which further increases the risk of forest fires.

These findings are summarised by the ISI values, which clearly indicate that fires spread more rapidly in summer, with a median peak in August.

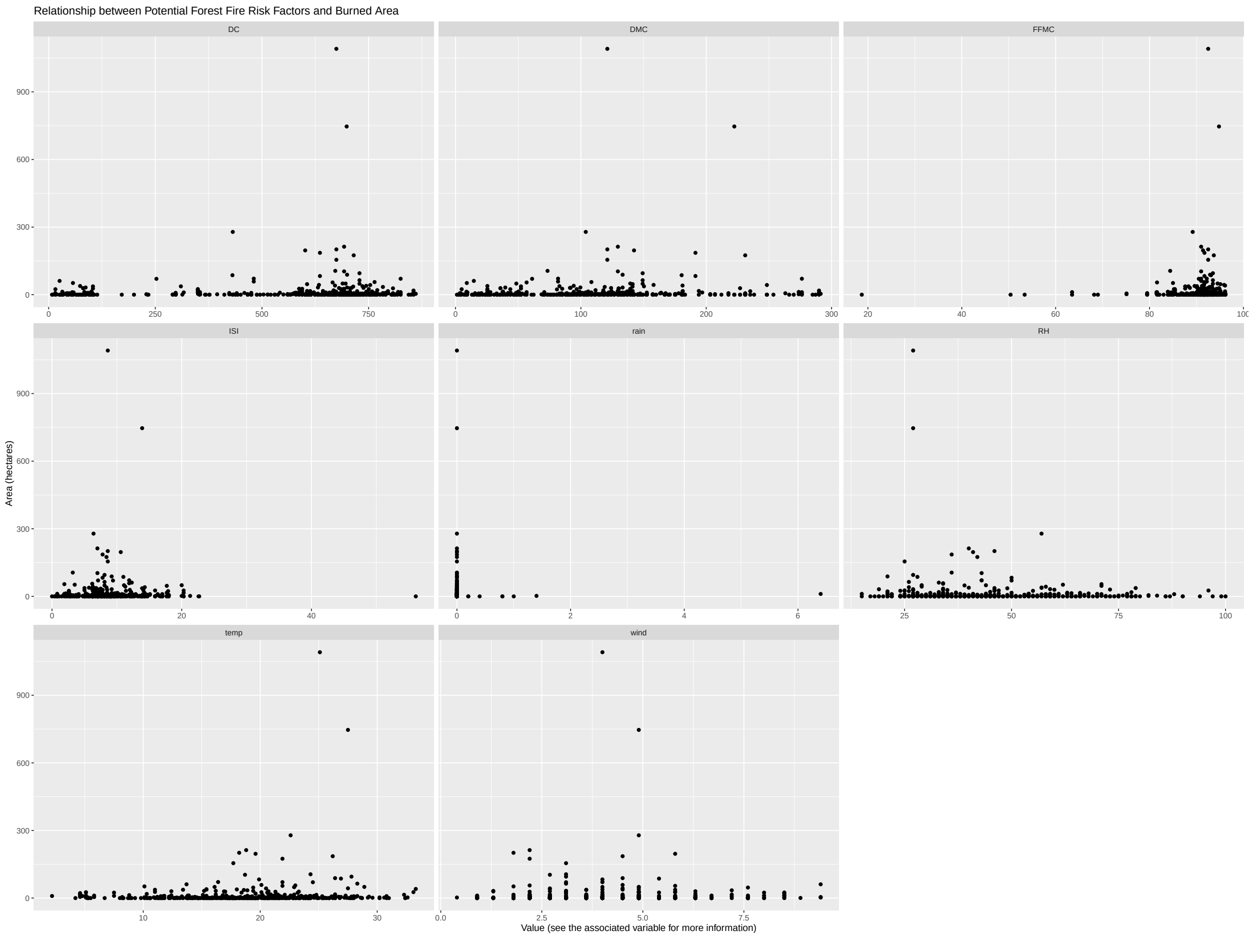
Plotting the relationship between the burned area and other variables of interest

A forest fire severity can be often summarised by the size of the area that it burns: this information is available to us thanks to the `area` variable, which is measured in hectares.

Given the results of our previous analysis, we might assume that the severity of one forest fire can be mostly explained by means of the DC, DMC, FFMC, ISI, the temperature and the humidity.

To corroborate this assumption it is useful to plot the relationship between these variables and the burned area, which we will do now with scatter plots.

```
firesL %>%
  ggplot(aes(x = value, y = area)) +
  geom_point() +
  facet_wrap(vars(variable), scales = "free_x") +
  labs(
    title = "Relationship between Potential Forest Fire Risk Factors and Burned Area",
    x = "Value (see the associated variable for more information)",
    y = "Area (hectares)")
```



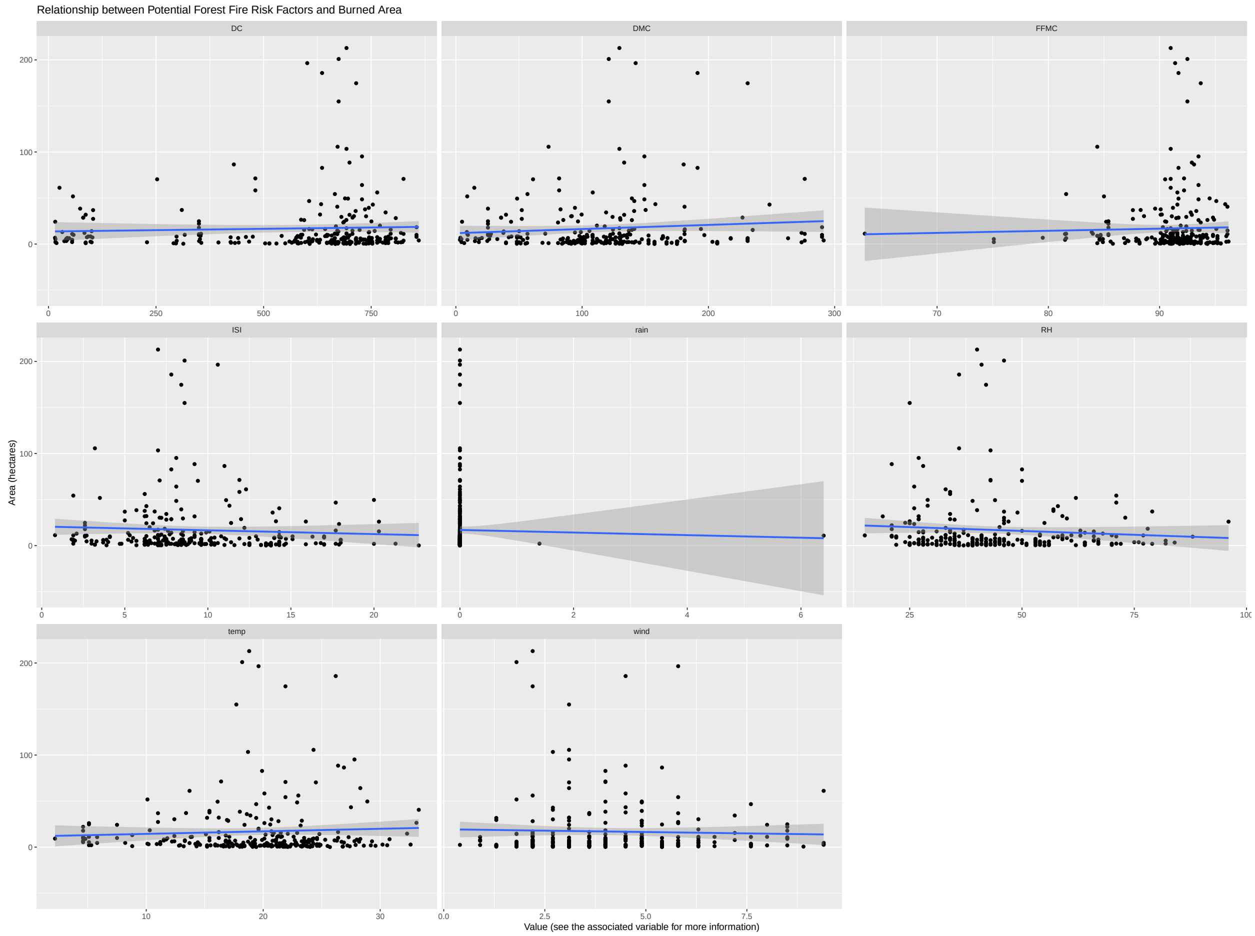
Unfortunately, the final result does not help us in determining if there is correlation between the variables, because of mainly two reasons:

- there are some outliers with very high values of `area` that cause a distortion of the y-axis limits;
- conversely, there are many more values with an `area` value of 0 that generate another distortion of the axis.

To fix this issue we will plot the data again, but this time we will filter it to remove the outliers and the zeros (note that we could have also simply tweaked the y-axis limits to remove these values from the plot).

```
firesL %>%
  filter((area <= 250) & (area > 0)) %>%
  ggplot(aes(x = value, y = area)) +
  geom_point() +
  facet_wrap(vars(variable), scales = "free_x") +
  stat_smooth(method = "lm") +
  labs(
    title = "Relationship between Potential Forest Fire Risk Factors and Burned Area",
    x = "Value (see the associated variable for more information)",
    y = "Area (hectares)")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



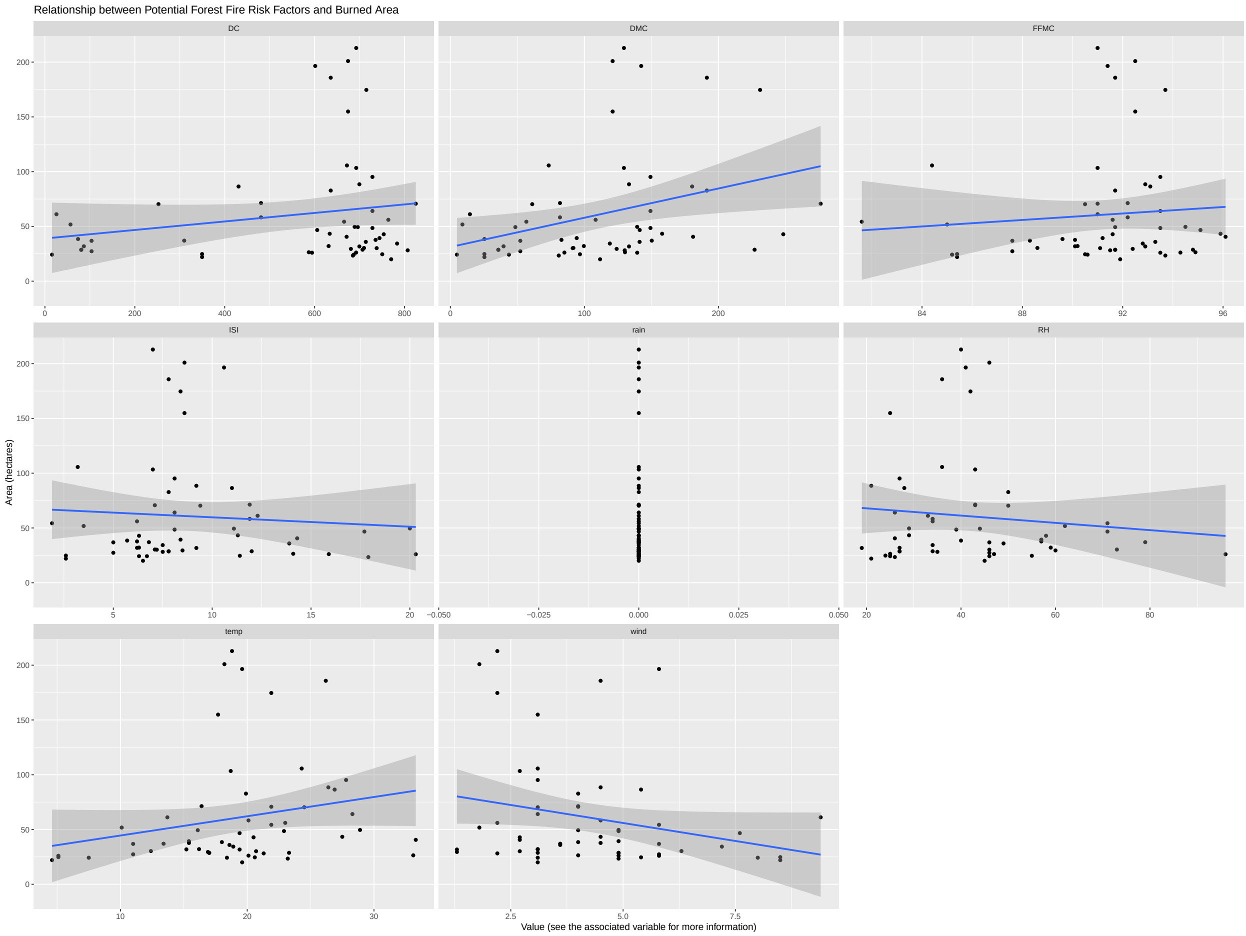
After removing the outliers and the zeros, we can now see that there seems to be a positive correlation between the variables DC, DMC, FFMC, temp and the area, and a negative correlation between the humidity RH and the area.

These correlations, though, really seem to be too slight to be of any significance, because there are still too many very low area values.

As a consequence, we will try to isolate the fires that have caused at least a certain number of burned hectares and see if there is a stronger correlation with the more severe fires (but we will still exclude the outliers).

```
firesL %>%
  filter((area >= 20) & (area <= 250)) %>%
  ggplot(aes(x = value, y = area)) +
  geom_point() +
  facet_wrap(vars(variable), scales = "free_x") +
  stat_smooth(method = "lm") +
  labs(
    title = "Relationship between Potential Forest Fire Risk Factors and Burned Area",
    x = "Value (see the associated variable for more information)",
    y = "Area (hectares)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



If we only consider forest fires which burned between 20 and 250 hectares the previous correlations become more significant.

In particular, the DMC and the temperature seem to have the strongest positive correlations with the number of hectares burned by the fire.

The humidity rate is also confirmed to be negatively correlated with the burned area; surprisingly enough, the wind speed seems also seems to have a negative correlation, which we probably would not have expected.

Conclusions

Based on our data exploration and visualization we can assert the following conclusions:

- forest fires occur the most in summer and in particular in the months of August and September;
- the FFMC, DMC and DC indexes, combined with the temperature and the humidity rate are good predictors of the occurrence of a forest fire and can also be decent predictors of its severity.