Exploratory Visualization of Forest Fire Data

Eda AYDIN

03 04 2021

Contents

Exploring Data Through Visualization: Independent Investigations

Load the package and we'll need for the project

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.0.4

## Warning: package 'tidyr' was built under R version 4.0.4

## Warning: package 'dplyr' was built under R version 4.0.4
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X = col_double(),
##   Y = col_double(),
##   month = col_character(),
##   day = col_character(),
##   FFMC = col_double(),
##   DMC = col_double(),
##   DC = col_double(),
##   ISI = col_double(),
##   temp = col_double(),
##   RH = col_double(),
##   wind = col_double(),
##   rain = col_double(),
##   area = col_double()
## )
```

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 associated with less fire spread, so we might expect the water-related variables ("DMC" and "rain") to be related with "area".

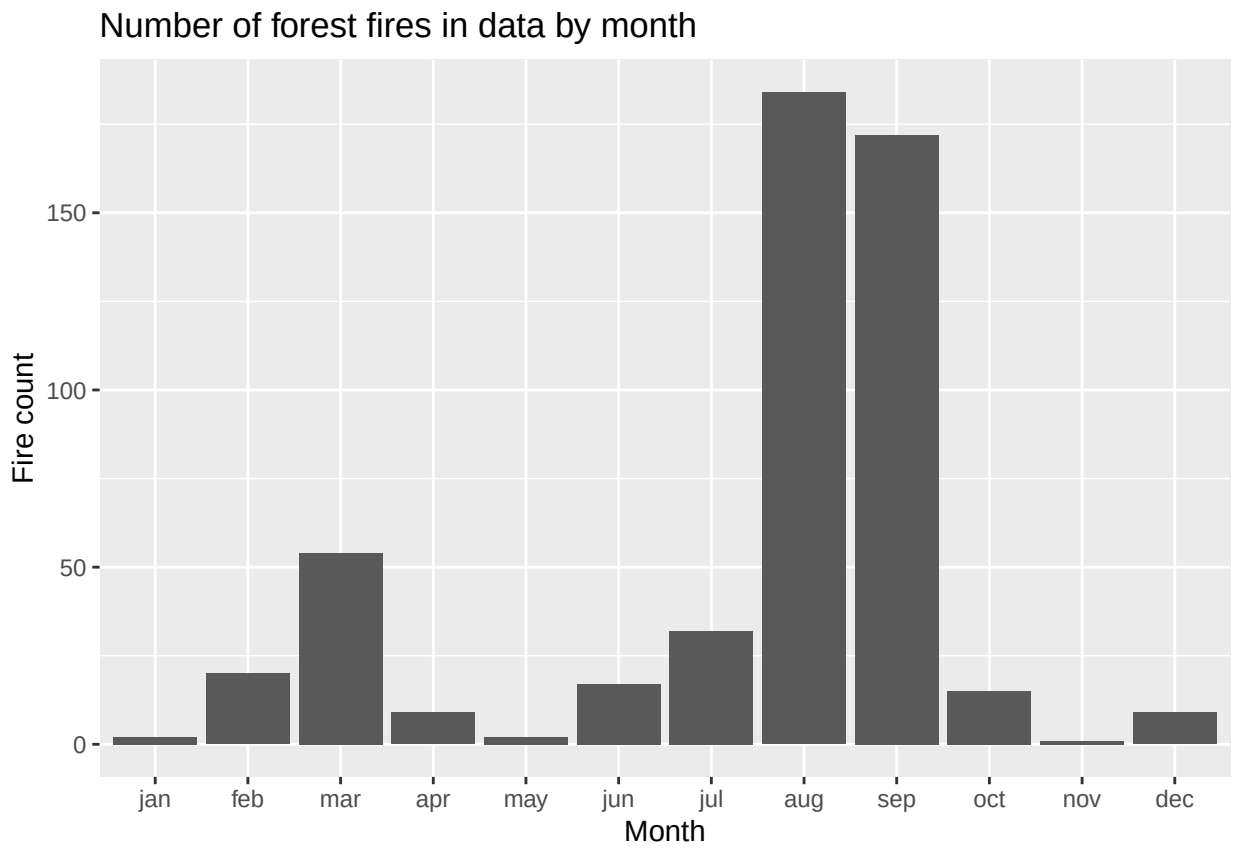
Data Processing

```
# Convert the month variable into a categorical variable

month_order <- c("jan", "feb", "mar",
                 "apr", "may", "jun",
                 "jul", "aug", "sep",
                 "oct", "nov", "dec")

# convert the day variable into a categorical variable

day_order <- c("sun", "mon", "tue", "wed", "thu", "fri", "sat")
```

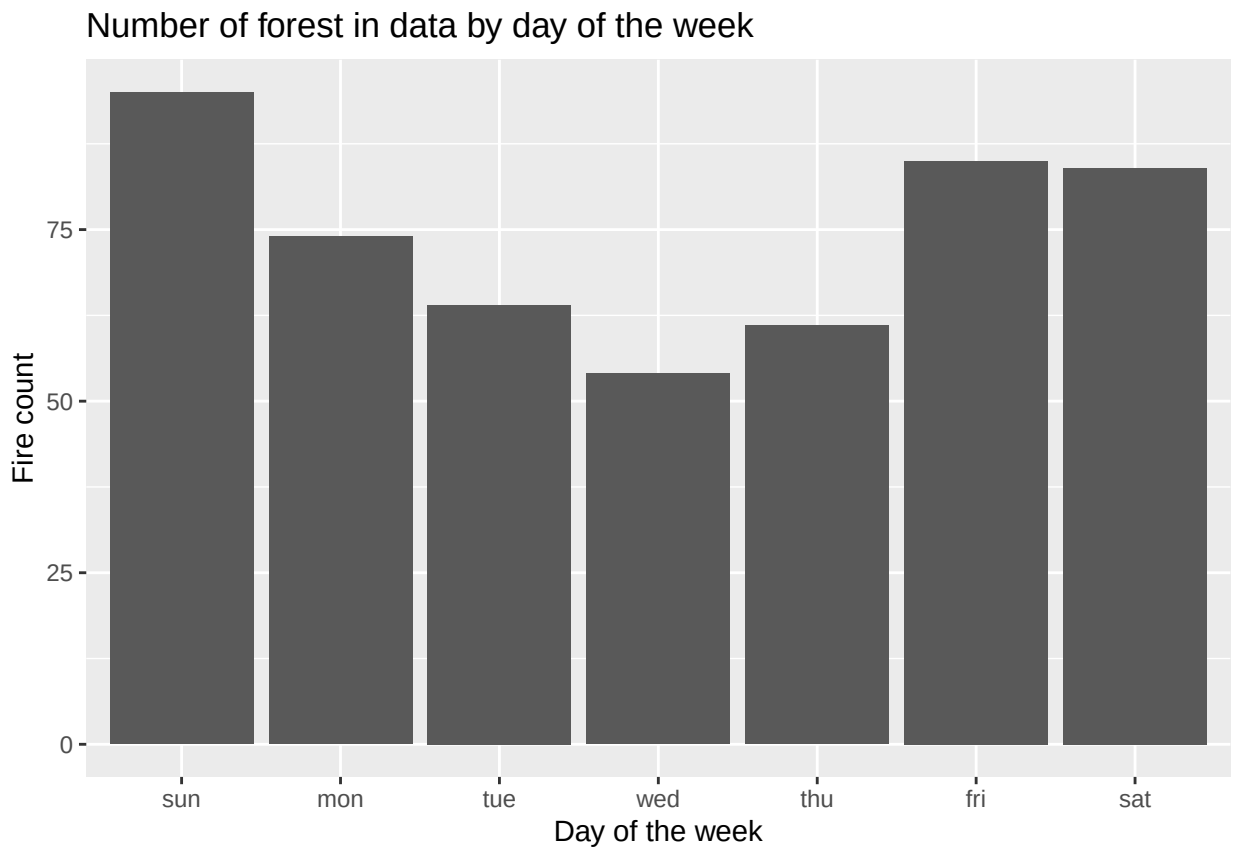
```
forest_fires <- forest_fires %>%
  mutate(
    month = factor(month, levels= month_order),
    day = factor(day, levels= day_order)
  )
```

When Do Most Forest Fires Occur?

```
# Which months do forest fires happen the most?

fires_by_month <- forest_fires %>%
  group_by(month) %>%
  summarize(total_fires = n())

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

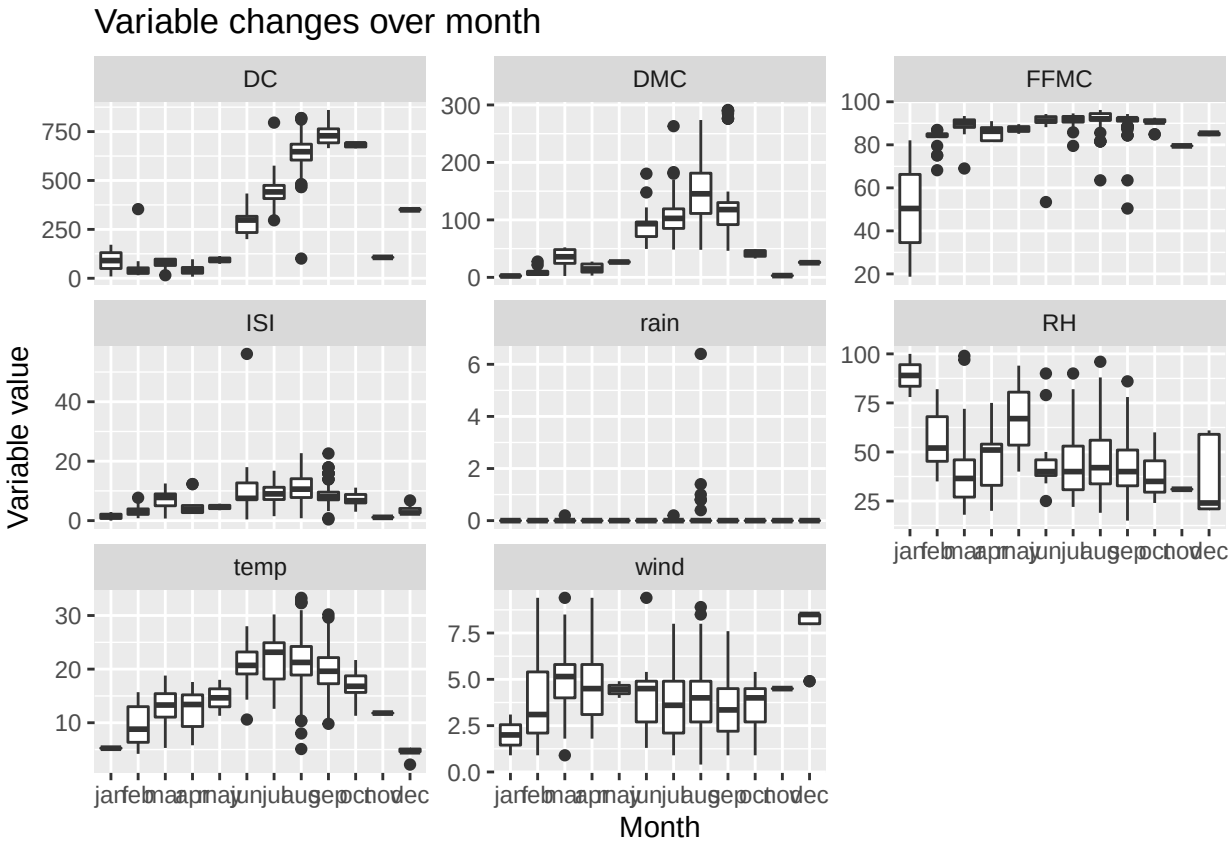



```
# Which days of the week do forest fires happen the most?

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

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



Plotting Other Variables Against Time

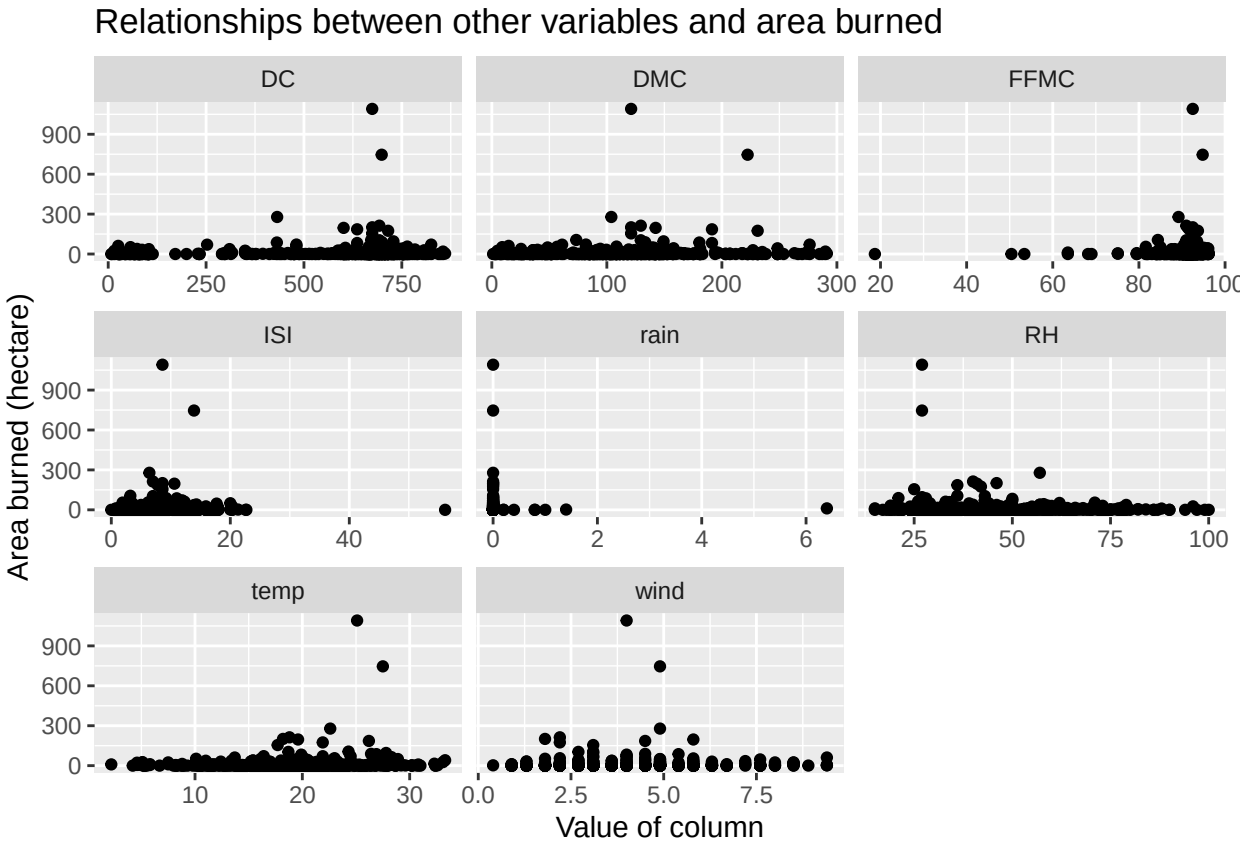
```
forest_fires_long <- forest_fires %>%
  pivot_longer(
    cols = c("FFMC", "DMC", "DC", "ISI", "temp", "RH", "wind", "rain"),
    names_to = "data_col",
    values_to = "value"
  )

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



Examining Forest Fire Severity

We're 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()".

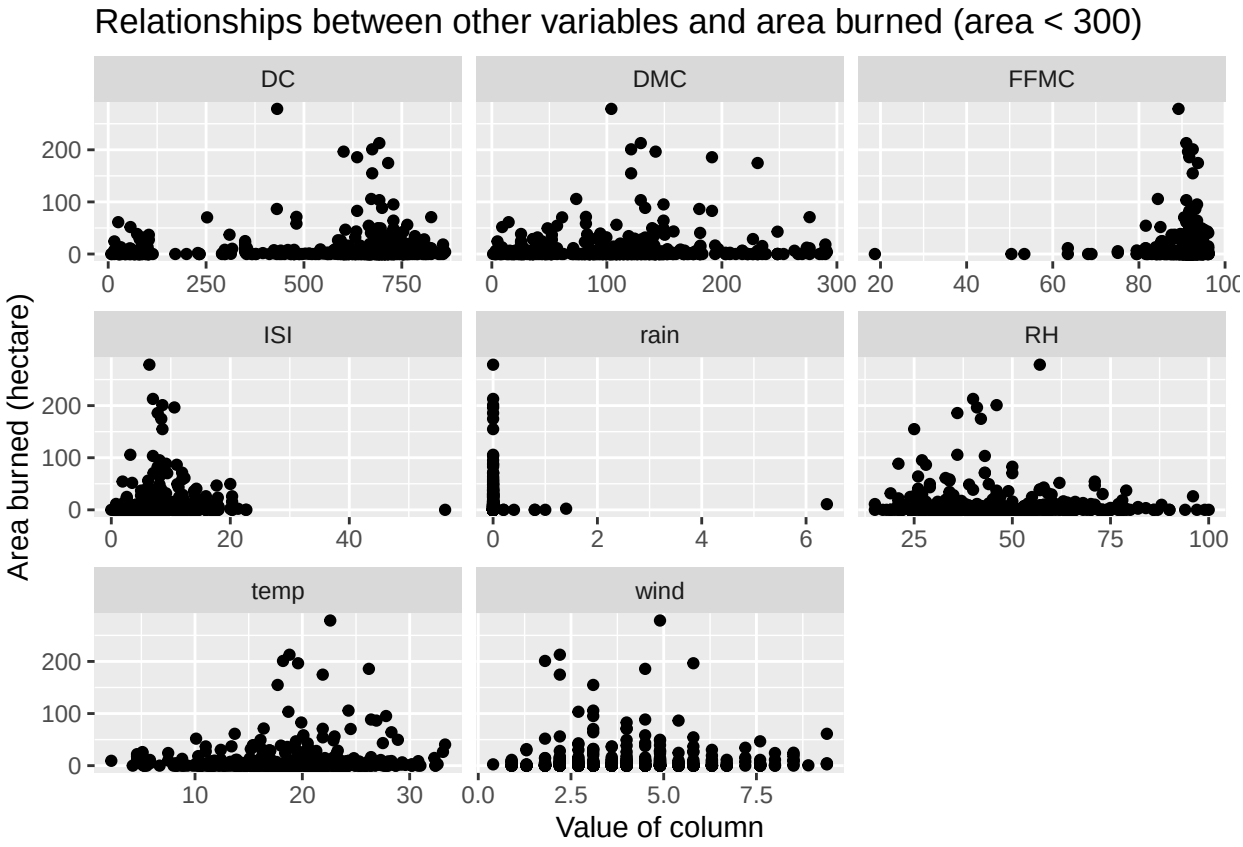
```
forest_fires_long %>%
  ggplot(aes(x=value, y=area)) +
  geom_point() +
  facet_wrap(vars(data_col), scales = "free_x") +
  labs(
    title = "Relationships between other variables and area burned",
    x = "Value of column",
    y = "Area burned (hectare)"
  )
```



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 variabe relates to "area".

```
forest_fires_long %>%
  filter (area <300) %>%
  ggplot(aes(x = value, y = area)) +
  geom_point() +
  facet_wrap(vars(data_col), scales = "free_x") +
  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