forestfires

Hayden Fu

5/29/2021

##Project Objective

The forestfires data consists of attributes of forest fires in the northeast region of Portugal.The goal is to predict the burned areas of forest fires.The output 'area' was first transformed with a ln(x+1) function. Then, several Data Mining methods were applied. After fitting the models, the outputs were processed with the inverse of the ln(x+1) function. In my project I will try to find the correlation between the different variables of the forestfires data; specifically relating the temperature/rain/wind and the area.

Variables 1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9 2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9 3. month - month of the year: 'jan' to 'dec' 4. day - day of the week: 'mon' to 'sun' 5. FFMC - FFMC index from the FWI system: 18.7 to 96.20 6. DMC - DMC index from the FWI system: 1.1 to 291.3 7. DC - DC index from the FWI system: 7.9 to 860.6 8. ISI - ISI index from the FWI system: 0.0 to 56.10 9. temp - temperature in Celsius degrees: 2.2 to 33.30 10. RH - relative humidity in %: 15.0 to 100 11. wind - wind speed in km/h: 0.40 to 9.40 12. rain - outside rain in mm/m2 : 0.0 to 6.4 13. area - the burned area of the forest (in ha): 0.00 to 1090.84 Note: (this output variable[area] is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

##Importing Required Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

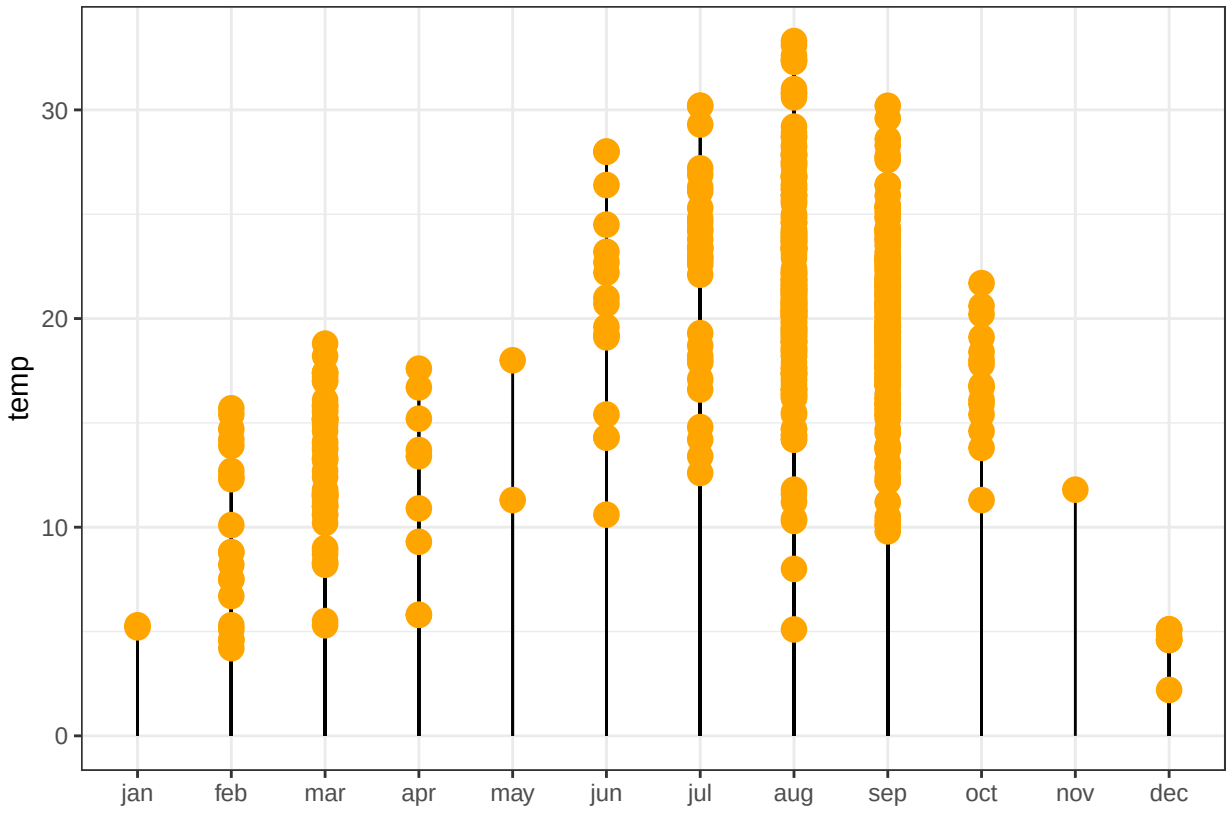
```
library(dplyr)
library(ggplot2)
forestfires <- read.csv("C:\\Users\\hayden\\Downloads\\forestfires.csv")
```

To make the variance of the area values more visible, I made the modified_area variable as well as the modified_rain variable:

```
modified_area <- 100 * log10(forestfires$area +1)
modified_rain <- 10 * log(forestfires$rain +2, base = 2)
```

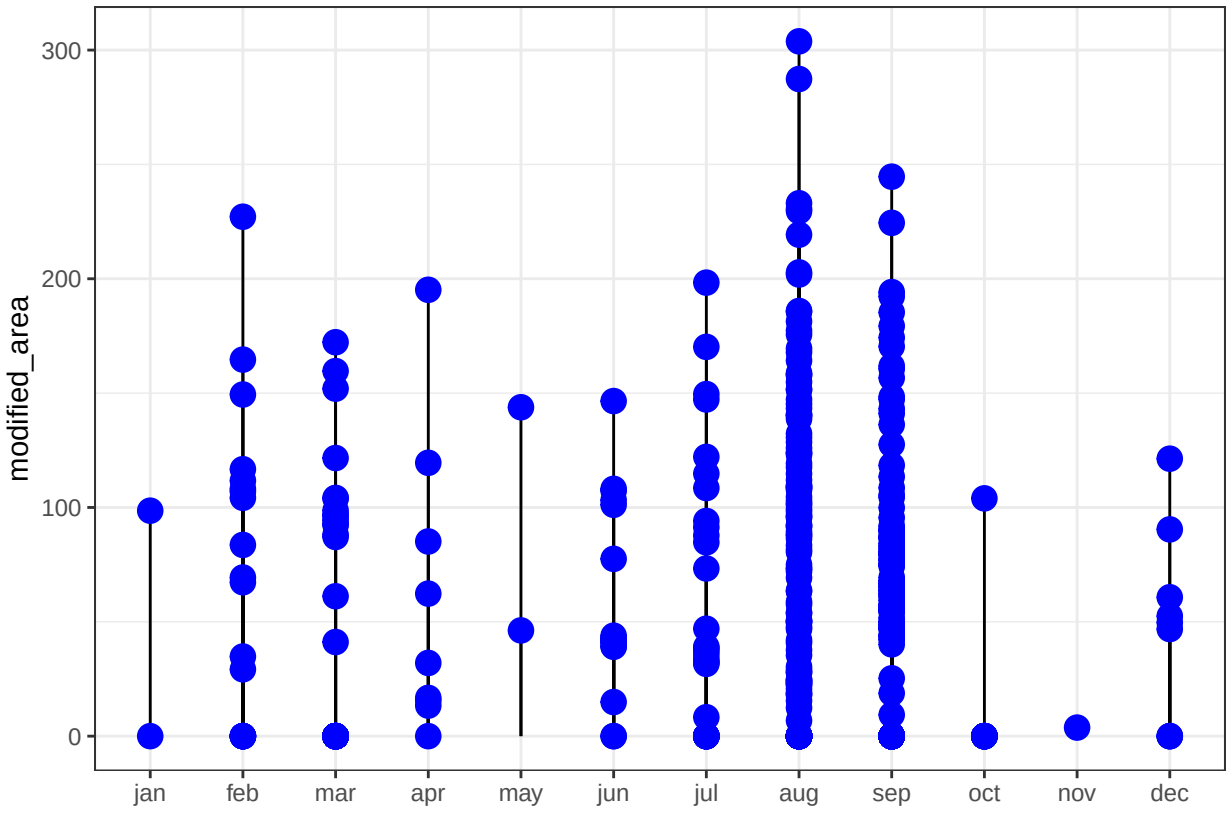
We look to examine the correlation between the temperature and month:

```
forestfires %>%
  arrange(temp) %>%
  mutate(name=factor(month, levels=c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov",
  ggplot( aes(x=name, y=temp)) +
  geom_segment( aes(xend=name, yend=0)) +
  geom_point( size=4, color="orange") +
  theme_bw() +
  xlab("")
```



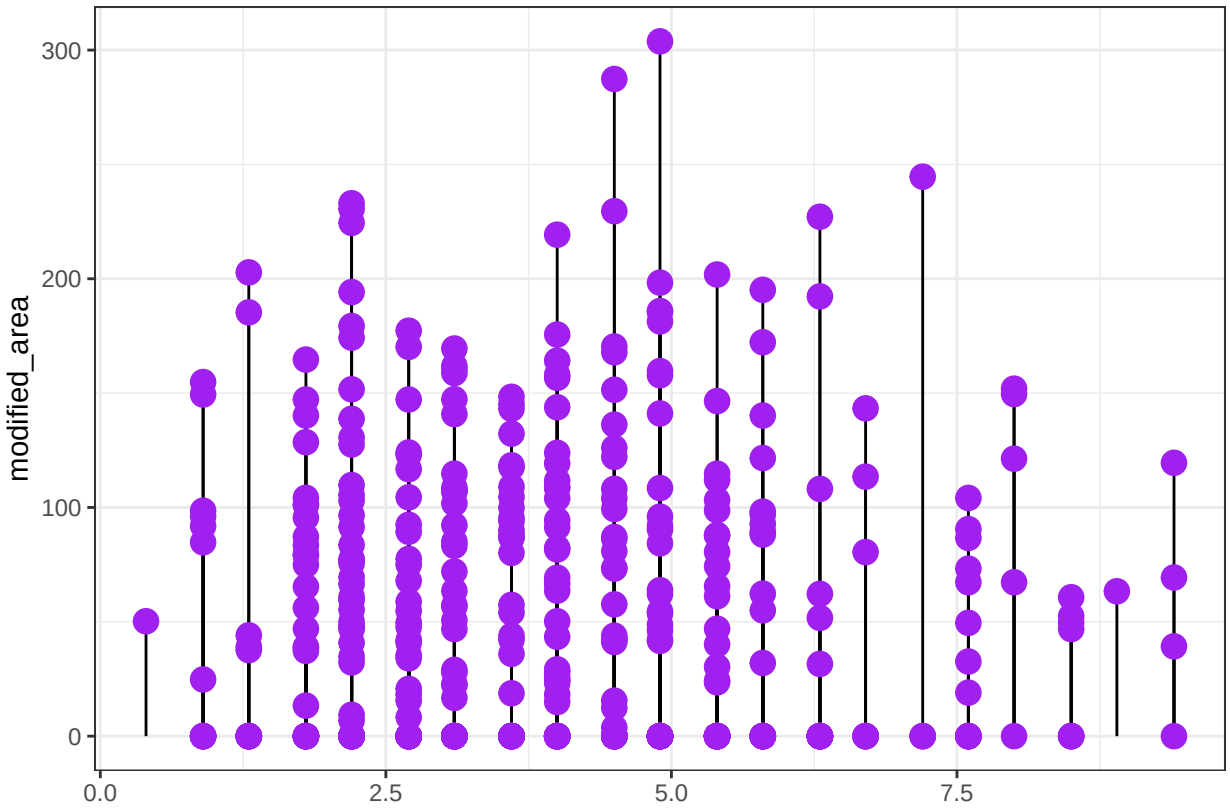
From this, we see that in the months of July, August, and September, the temperature is highest. Next, we will examine the correlation between the month and area, temp and area, wind and area, and rain and area.

```
forestfires %>%
  arrange(modified_area) %>%
  mutate(name=factor(month, levels=c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov",
  ggplot( aes(x=name, y=modified_area)) +
  geom_segment( aes(xend=name, yend=0)) +
  geom_point( size=4, color="blue") +
  theme_bw() +
  xlab("")
```



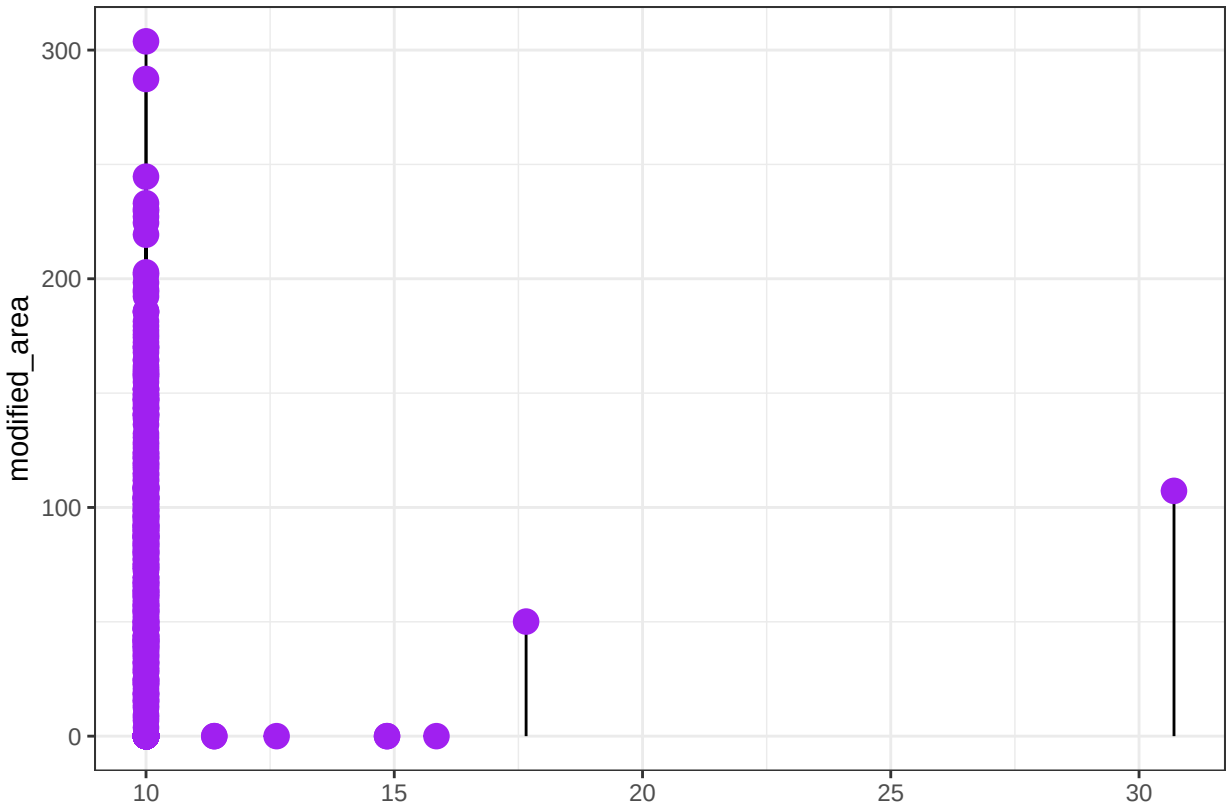
From the model, it is visible that the area was lowest in the months of August and September, which were both notably months with high temperatures. When we compare the wind to modified_area, we see that modified_area is highest at values between 4.5 and 7.5:

```
forestfires %>%
  arrange(modified_area) %>%
  ggplot( aes(x=wind, y=modified_area)) +
  geom_segment( aes(xend=wind, yend=0)) +
  geom_point( size=4, color="purple") +
  theme_bw() +
  xlab("")
```



Finally, when we compare rain to modified_area, although there are few samples of rain, those samples (except for one outlier) all show that the modified_area is significantly reduced when there is rain. Almost all large cases of large firespread is caused by a lack of rain.

```
forestfires %>%
  arrange(modified_area) %>%
  ggplot( aes(x=modified_rain, y=modified_area)) +
  geom_segment( aes(xend=modified_rain, yend=0)) +
  geom_point( size=4, color="purple") +
  theme_bw() +
  xlab("")
```



##Conclusion and Insight

From the different plots examined, we can see that during the months of August and September, the area that fire spreads to is greatest, as well as the temperature. It is also during this time that there is the least rain, and rain has been seen as inversely proportional to the fire spread. Most of the variables have a clear proportional or inverse proportional relationship except for wind, as wind has no clear correlation to wind, month, or area. The data gained is that during the months of August and September, people must be careful as there is very little rain and the temmperature is high, and forest fires are most dangerous and common during those two months.