Forest Fire Data Analysis

Adithyan Subramanian

In this project, I analyzed data from Paolo Cortez and Anibal Morais' paper, "A Data Mining Approach to Predict Forest Fires using Meteorological Data," to do my own exploratory data analysis of Forest fires in Portugal. I used R and Rstudio to perform some basic exploratory analyses within this project.

The data may be found at this link:
https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/

Initial exploration: Days and Months

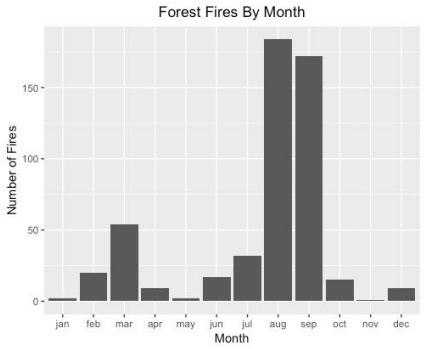

Fig 1

Our initial exploration started with assessing which months and days had the most fires. To do this, we used a histogram to visualize the data. As shown by *fig. 1,* the majority of forest fires occurred in August and September. This intuitively makes sense, as August and September are both hot, dry summer months. It would make sense for the most fires to happen in these months, as the climate is ripe for fire. The months of November, December, and January seem to have the least amount of fires. This also makes sense, as these winter months would not be ideal conditions for fire starting. It is interesting to see March have so many more fires than February, or April. The data, or intuition does not provide a clear reason as to why this is. A future analysis could lend light onto this anomaly.

When looking at the number of fires by the day of the week (*fig 2*), we see that the data did not specify what day the fire started for a sizable portion of the data. With that in mind, the rest of the data indicates that the most fires started on Friday, Saturday, or Sunday. This also makes sense as these days comprise the weekend when many people would spend time in the forests hiking, and camping, which increases the risk of a fire starting due to human negligence or error.

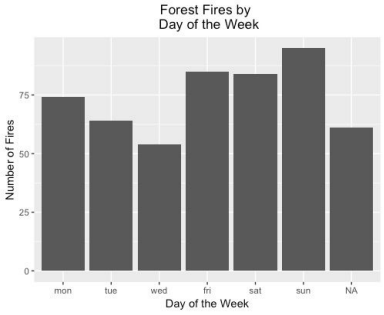

Fig 2

Analyzing variables based on months and days

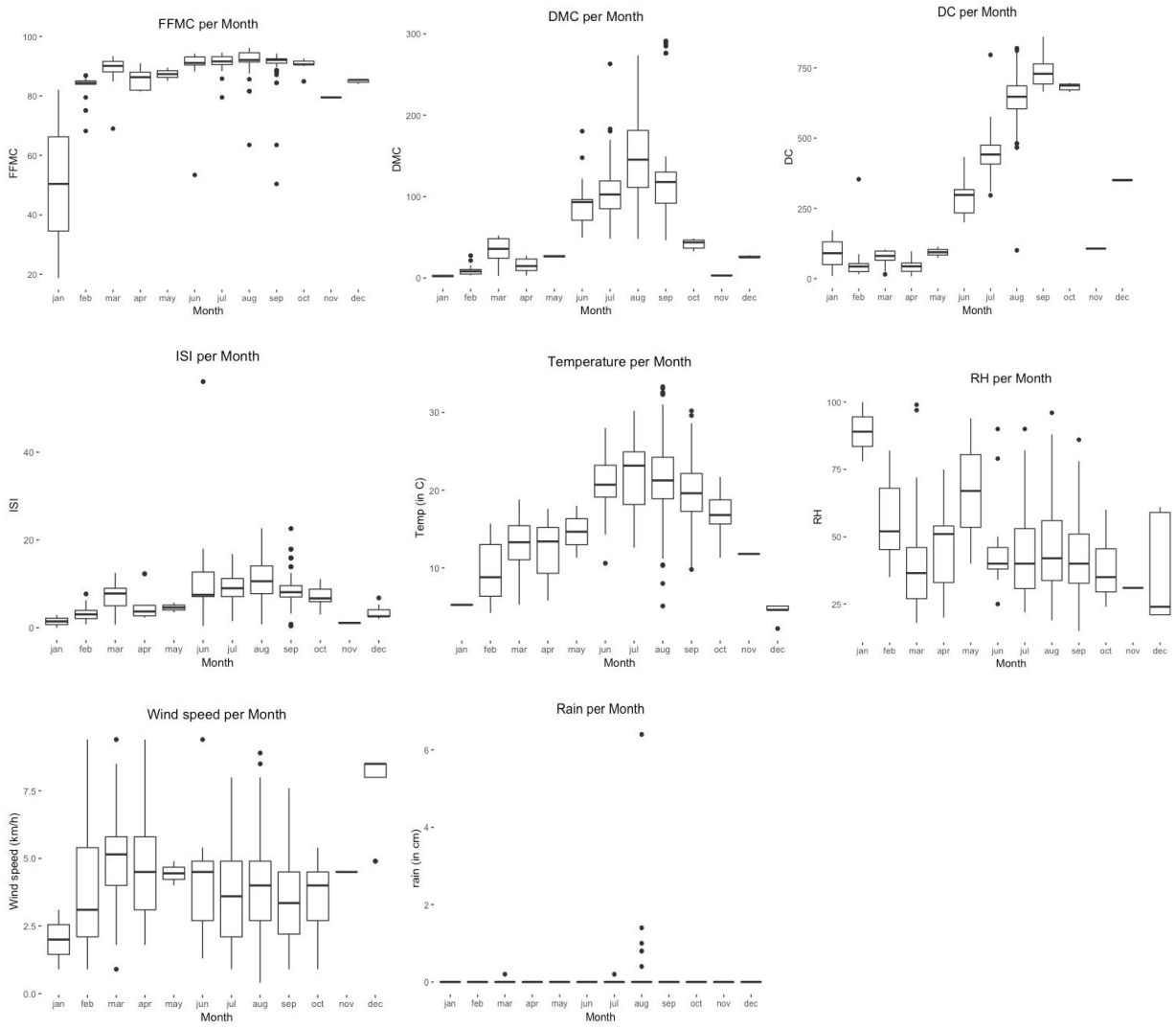

Fig 3

The dataset records various variables Scientists use to quantify risk factors of forest fires. A full explanation of these variables is beyond the scope of this project, but more information can be found at http://cwfis.cfs.nrcan.gc.ca/background/summary/fwi.
The first thing I did was visualize these variables based on month using boxplots, as seen on *fig 3*. FFMC, DMC, and DC variables all assess how quickly a fire can start given how dry and flammable fuel sources are. This provides some reason as to why August and September have so many fires. The months are dry and fuel sources are easily flammable, allowing fires to start quickly. The temperature also shows that August

and September have higher than average temperatures, which probably yield to higher FFMC, DMC, and DC ratings. The RH figure, analyzes humidity during the months. Humidity seems to fluctuate throughout the entire year, and it seems that August and September have somewhat average rates of humidity, so it is difficult to correlate humidity with rate of forest fire starting.

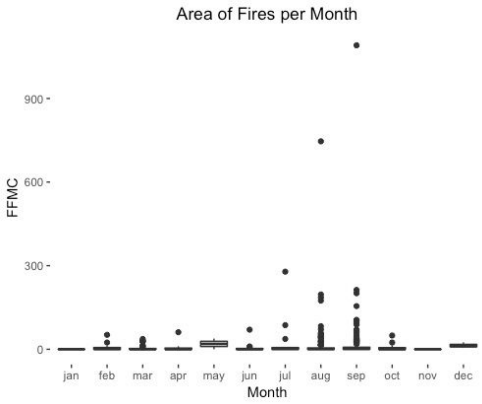

fig . 5

ISI measures how well a fire will spread. One would judge that a high ISI would mean a fire will spread to a high amount of area. Although ISI seems to be relatively stable throughout the year, there does seem to be a uptick during the months of July, August, and September. This leads me to believe that not only did august and September have the most fires, they also had the largest fires. Another boxplot would help us confirm this. Looking at *fig. 5*, it is clear that most fires are small, and burn close to 0 hectares. It is however, interesting to note that 3 largest fires occur in September, August, and July, confirming that ISI does accurately predict fire area.

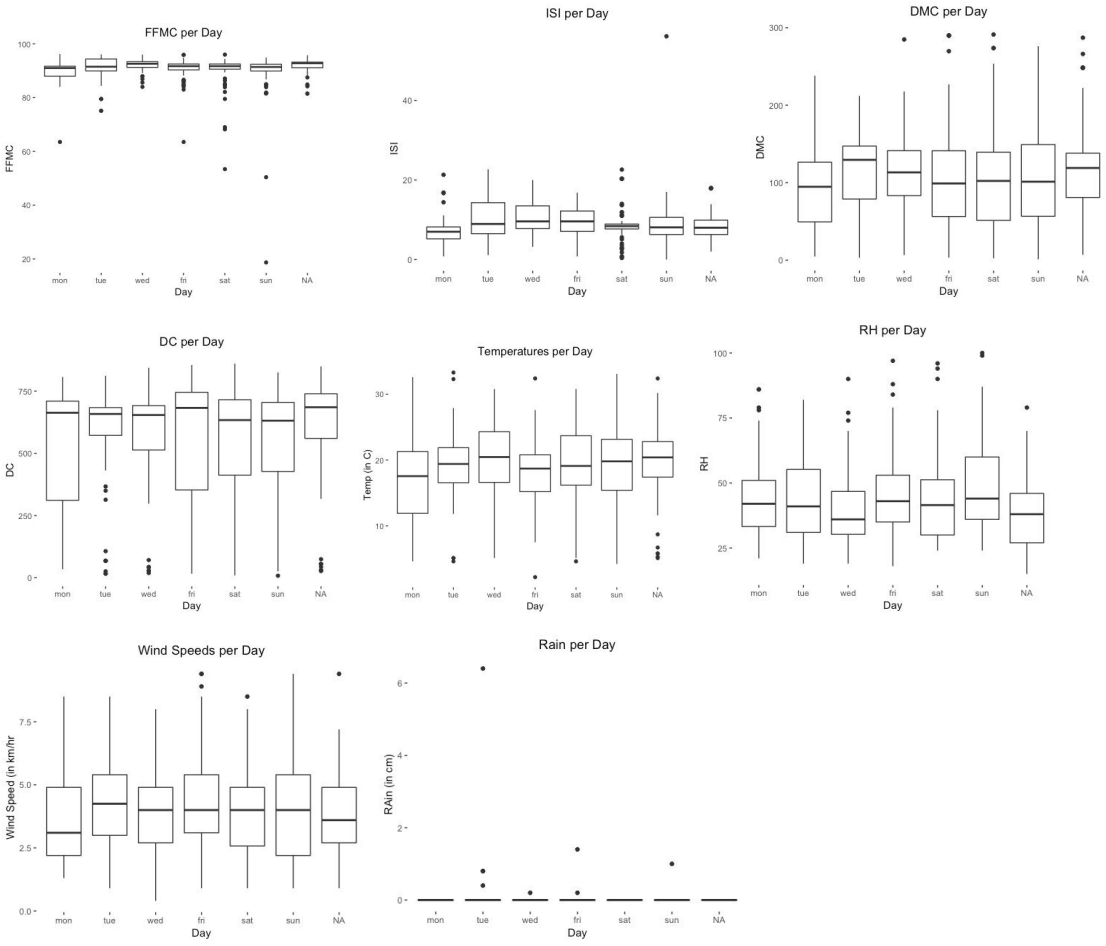

Fig 6

Looking at how the variables compare based on days of the week (*Fig 6),* we see that there is really not much variance with any of the variables based on the days of the week. This makes sense because weather, temperature and other fire variables are not going to change based on what day of the week it is.

How variables affect the size of fires

The next step in our analysis was to analyze how the variables affected the size of the forest fires. To do so I created scatter plots of each variable, based on how large the fire grew to.

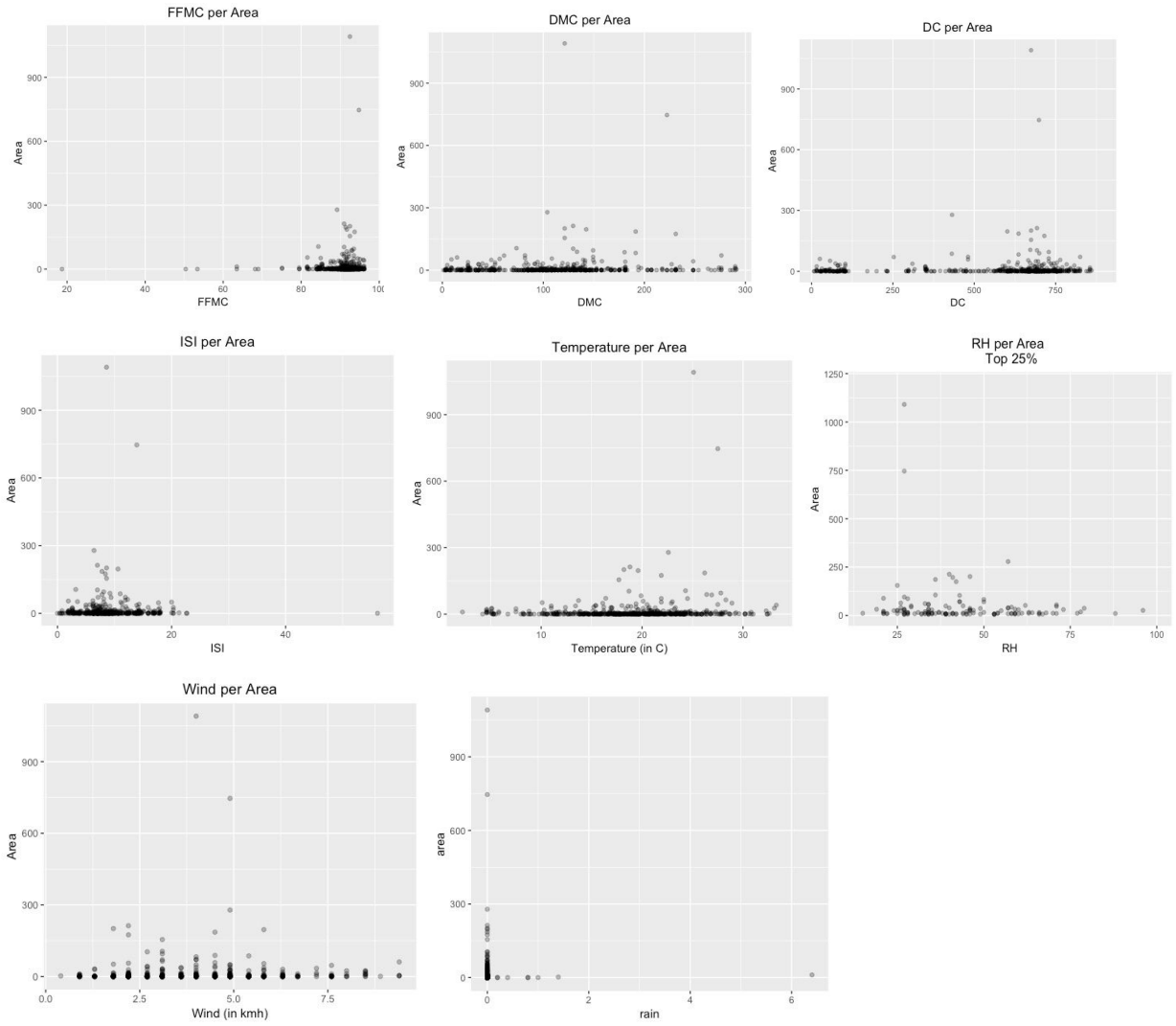

Fig 7

When looking at the scatterplots, it is glaring that the vast majority of forest fires were very small in area. With that said, it is hard to gather any information from these scatter plots, as all the data is bunched up around the zero mark. To gain any info, I decided to only focus on the highest quarter of the data.

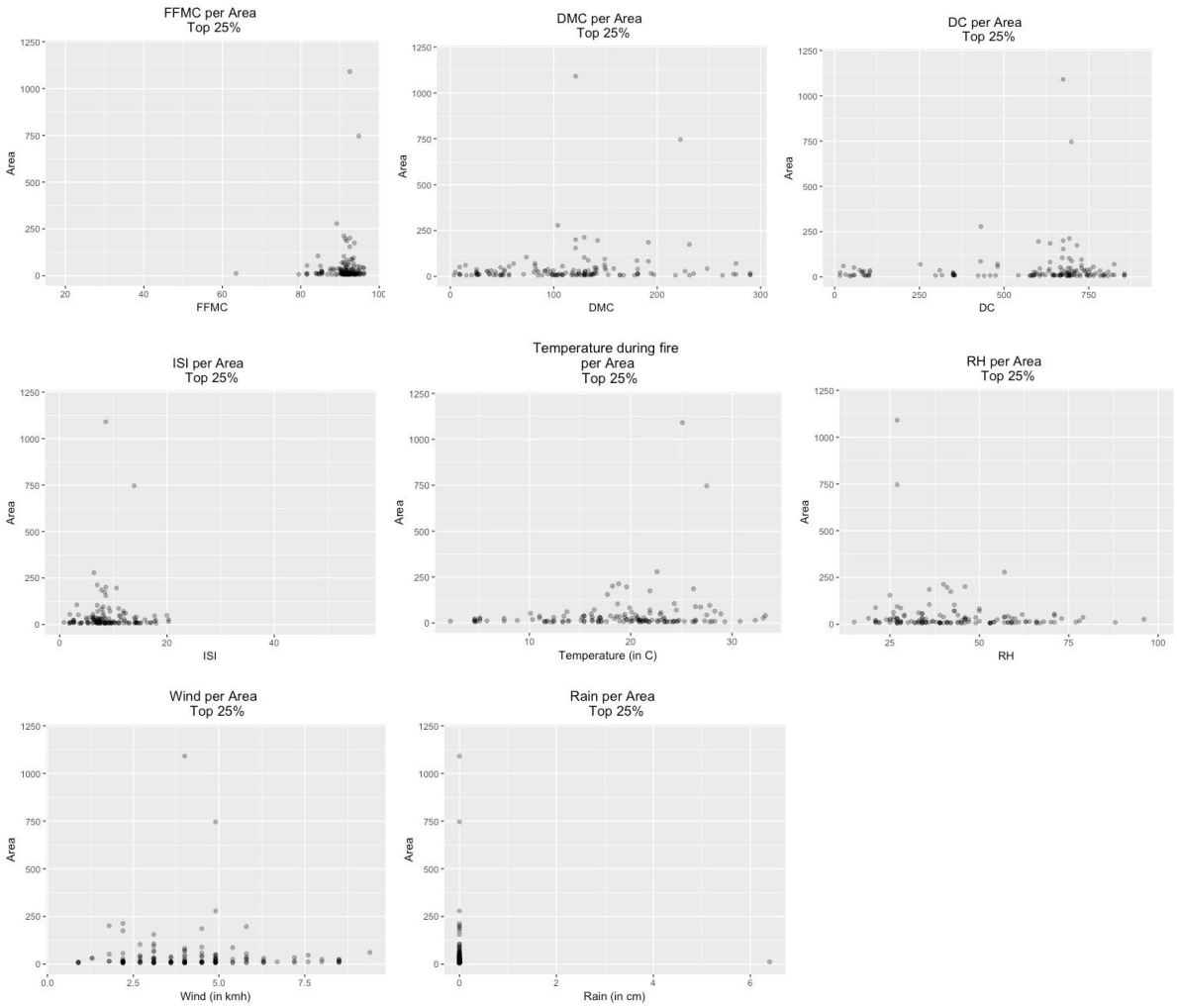

Fig. 8

Even looking at the top 25%, the data points are still very bunched together, but we can make some observations. First, regardless, of how large the fire is, FFMC must be high. This makes sense since FFMC measures how quickly a fire can start. It is also logical that the largest fires all occured when there was no rain. What is interesting is that DMC and DC values seem to be varied. A further study may dive into this further to find reasoning behind this. Unusually, ISI values for the largest fires seem to be relatively low. This is surprising because one would assume high ISI values would indicate larger fires. This is another fact that could be further studied. The rest of the variables do not yield any obvious results.

Conclusion

As the data shows, August and September have the most amount of fires and the largest fires due to high FFMC, DC, and DMC values. It is has been displayed that for a fire to be as large as possible the conditions should include no rain, and a high FFMC. All other variables do not show any obvious correlation.