

Predicting the Size of Forest Fires

A machine learning approach

Mier Chen | STAT 4995 | 4/24/2017

## Executive Summary

Forest fires is a major environmental issue that endangers the human lives and the ecological system. We use the meteorological data collected in Montesinho Natural Park from January 2000 to December 2003 to predict the size of a staring forest fire. The model will benefit the firefighters to estimate how much manpower and resources they may need in the future forest fires. In the analysis, we try a couple of machine learning techniques, such as logistic regression and support vector machine. The result indicates that only the meteorological data is not enough for an accurate prediction. Human activities, the altitude where forest fires started and the vegetation coverage in each area are also needed for a better model.

## Introduction

**Fire Detection**

Forest fires is a major environmental issue. A large forest fire can destroy over 100 forest hectares, damage the habitat of endangered wildlife, cause dramatic human suffering, and increase the release of carbon dioxide into the air which leads to climate change.

Montesinho Natural Park, one of the largest natural park in Canada, has over a hundred forest fires annually. The forest fires threaten a great variety of wildlife, as well as the 9000 inhabitants living in the park.

We use the meteorological data in this report to predict the size of the forest fires. Weather conditions, such as temperature, wind speed, and air humidity, are well acknowledged as the keys to fire occurrence and fire spread. Moreover, it is of low cost to collect the meteorological data since automatic meteorological stations are often available. In our case, Portugal has 162 official stations that can obtain the real-time weather information with low costs.

Fire detection is crucial in the fight with the wildfires. We may be able to use the meteorological data to warn the public of a potential wildfire danger, support the fire department to predict the occurrence of forest fires and even the possible size of the wildfires.

The data set contains 517 instances and 13 variables (1 dependent variable, four discrete attributes and 8 continuous attributes). The data is cleaned. We don’t have any missing value.

### **Montesinho natural park**

The data we have are collected from the Montesinho Natural Park, from January 2000 to December 2003. Montesinho Natural Park is one of the country’s largest natural parks. The park, located in the Tras-os-Montes northeast region of Portugal has 74.23 hectares. Created in 1979, Montesinho not only contains a great diversity of wildlife but also has 92 villages and 9000 inhabitants living in the park. Once a forest fire occurs, it will cause human suffering and environmental threats.

The data collectors divide the park into 36 parts. I create a variable called “location” to denote each part (e.g：X, Y). Summer is the peak tourist season, especially July and August. The only campground in the natural park, Cepo Verde, is located at the edge of (5,6) and (5,5). I mark the camp site with red location icon. In the data exploration, I pay more attention in this area.



### **The forest Fire Weather Index (FWI)**

Our data also contains the variables: FFMC, DMC, DC, and ISI. They are the numeric ratings under the forest fire weather index. The Forest Fire Weather Index (FWI) is the Canadian system for rating fire danger, and it includes six components, shown in the following figure.

[[1]](#footnote-1)Fig. 1. The Fire Weather Index Structure [1]

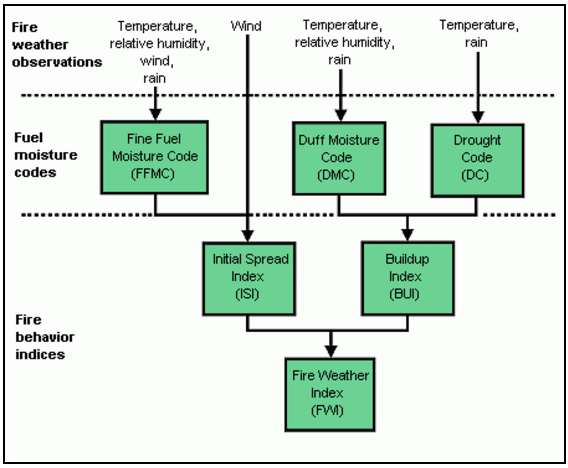
The **Fine Fuel Moisture Code (FFMC)** is a numeric rating of the moisture content of litter and other cured fine fuels. This code is an indicator of the relative ease of ignition and the flammability of fine fuel.

The **Duff Moisture Code (DMC)** is a numeric rating of the average moisture content of loosely compacted organic layers of moderate depth. This code gives

an indication of fuel consumption in moderate duff layers and medium-size woody material.

The **Drought Code (DC)** is a numeric rating of the average moisture content of deep, compact organic layers. This code is a useful indicator of seasonal drought effects on forest fuels and the amount of smoldering in deep duff layers and large logs.

The **Initial Spread Index (ISI)** is a numeric rating of the expected rate of fire spread. It combines the effects of wind and the FFMC on rate of spread without the influence of variable quantities of fuel. [2]



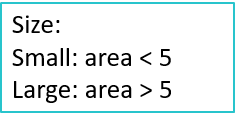
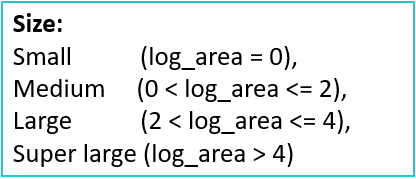
According to the above information from Natural Resources Canada, FFMC can be correlated with the relative humidity, wind speed, the amount of rain and temperature at that day. ISI can be correlated with wind speed. DMC is correlated with temperature, relative humidity, and rain. DC is correlated with DC. I checked their correlation, in the later analysis. Their correlations are not too high to be concerned.

## Approach

1. Data Exploration:
   1. Does the camp site, Cepo Verde (5,6), have the most number of forest fires?
   2. Some Important Variables (eg. RH, wind speed, location)
2. Build a Regression Model to predict the burned area
3. Machine Learning method to predict the size of forest fires (2 sizes & 4 sizes):
   1. Logistic Regression
   2. Support Vector Machine
4. Deep Learning Method

The distribution of area (burned area) is crazily skewed to the right since we have 217 observations with burned area equals to zero. To make its distribution closer to a normal distribution, I take log to the area.

I also use two ways to group the forest fires so that the area of the large forest fires’ distribution is closer to normal distribution. Each group has at least 100 observations, except the super large one, has 67 forest fires.

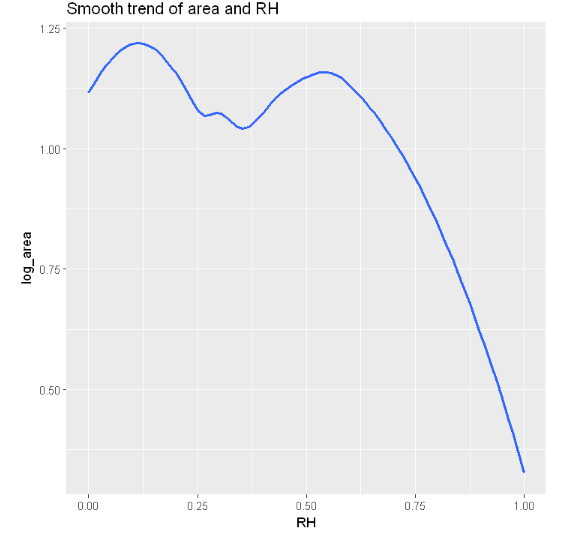
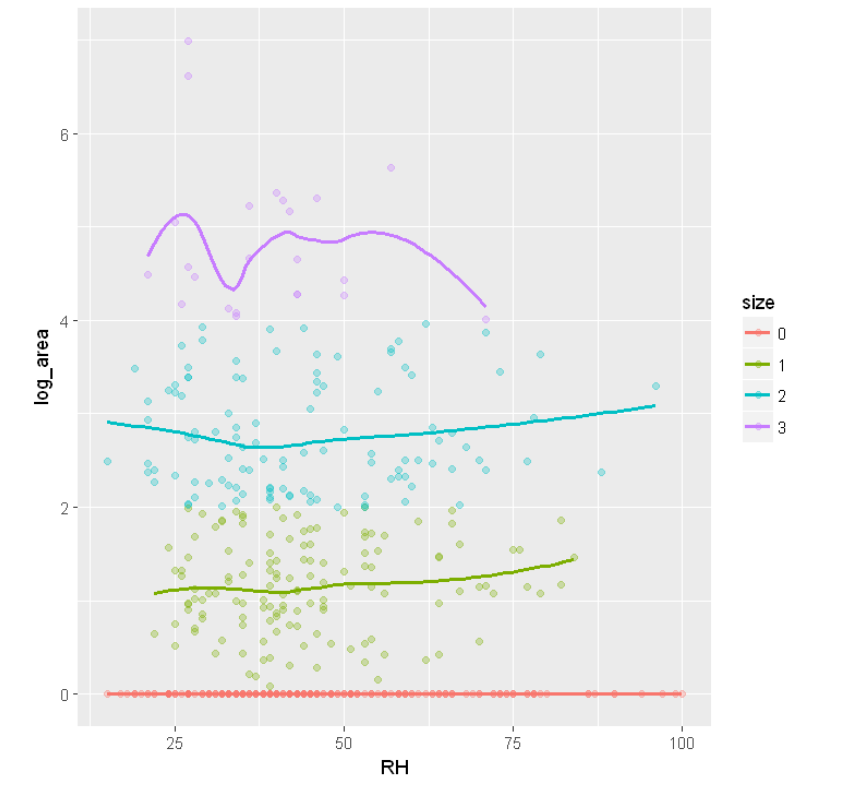


## Result

**Does the camp site has the most number of forest fires?**

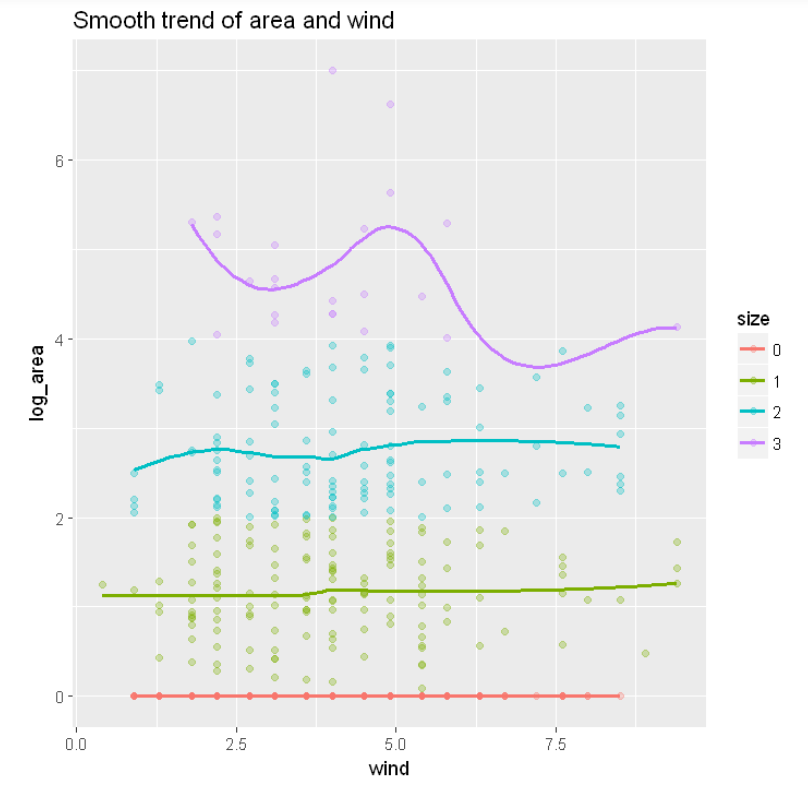
The misuse of campfires may cause the forest fires. Hence, I pay close attention to the campground. The only campground in the natural park, Cepo Verde, is located at the edge of (5,6) and (5,5). However, to my surprise, I found zero cases of forest fires in that area.

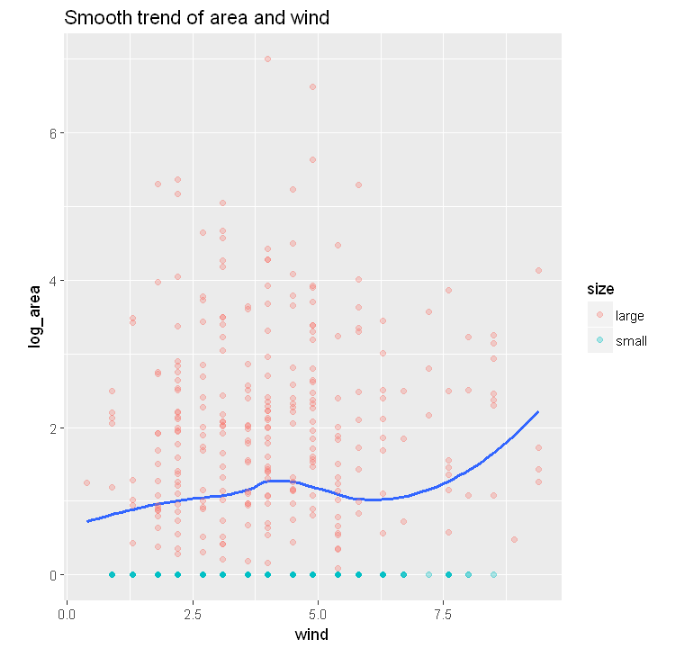
**RH**

The first graph is created by grouping the size of forest fires into four levels. I construct a smoothed line in each group to illustrate their relationship. The second graph is created using all the log\_area, without grouping. It shows a stronger relationship. However, the second chart is more likely to be affected by outliers (super large forest fires).

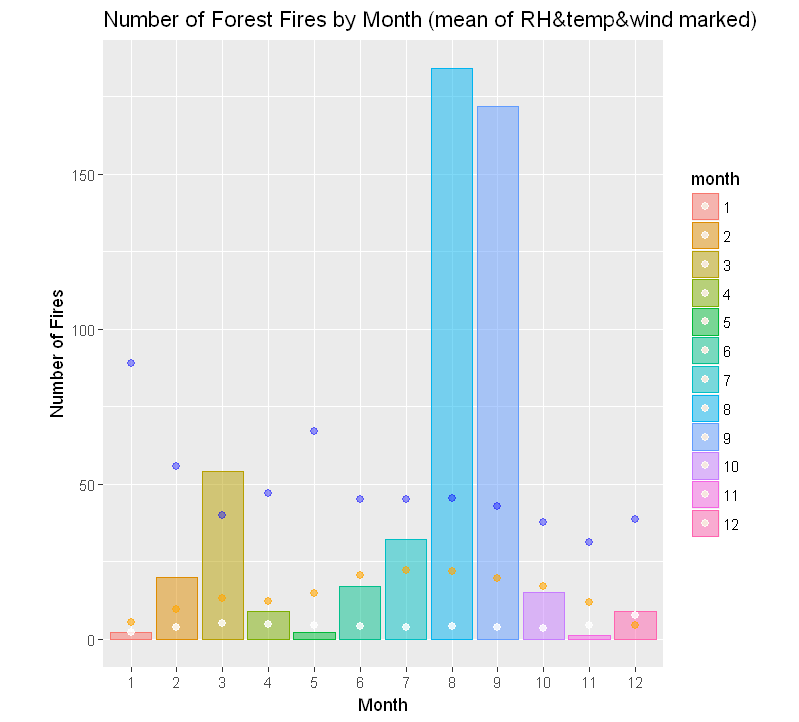
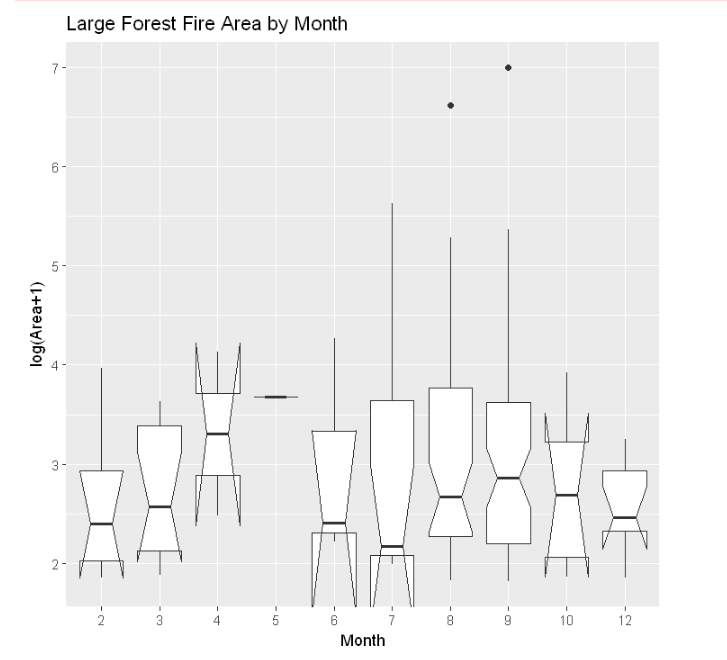
The above graph shows a relationship between the burned area and the relative humidity. When the RH is small, the burned area is more likely to be small. It peaks at RH equals to 25%. As RH increase, the burned area decreases. The top 2 largest forest fires happened when the RH is around 28. It explains the reason that the RH peaks at around 25 in the second graph.

**Wind** **speed**

****

I first group the forest fires into four groups. I notice that it is very hard to see a relationship between the area(log) and the wind speed. Hence, I created another graph by grouping the large and small forest fires (greater than 0 or not).

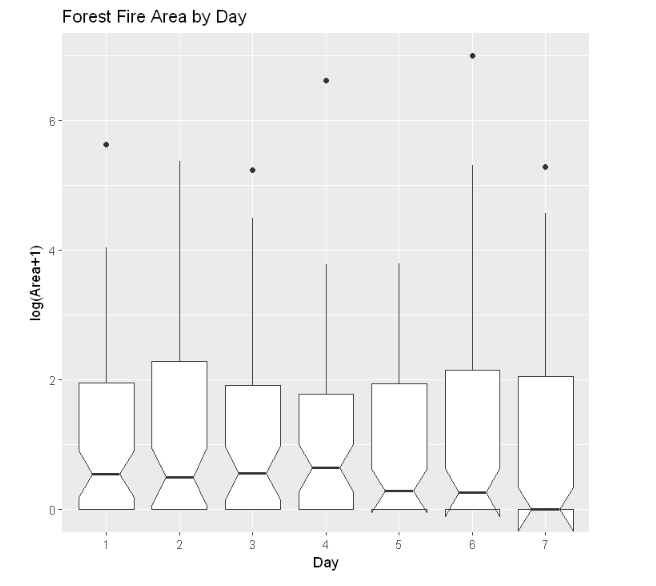
In the first picture, I found that most of the super large forest fires happened when the wind speed is from 3 to 6 km/h. In the second graph, the smoothed linear model between wind speed and log\_area shows that as the wind speed increase, the burned area increase as well. The figure shows a local maximum which is affected by a super large forest fire.

**Month**

The first figure is created using all the number of forest fires. I mark monthly mean of RH in blue, the wind speed in white and the temperature in orange in the plot. To better observe the huge forest fires, I remove all the data whose area is equal to 0.

August and September has the highest number of forest fires. The biggest forest fires happened in these two months as well. January, May, and December have almost 0 occurrences of forest fires. The relative humidity of January and May are higher than other months. The average of temperature is relatively low in December and January. It is interesting to observe that the number of fires declines suddenly from September to October. However, if we concern about the size of the forest fires, the average burned area in October are even larger than September.

The second boxplot is created by removing all the zero area observations. Although August has a slightly higher number of fires than September, September’s the average burned area is larger than August's. We also see that the super large forest fires mostly happened in Summer, especially in August.

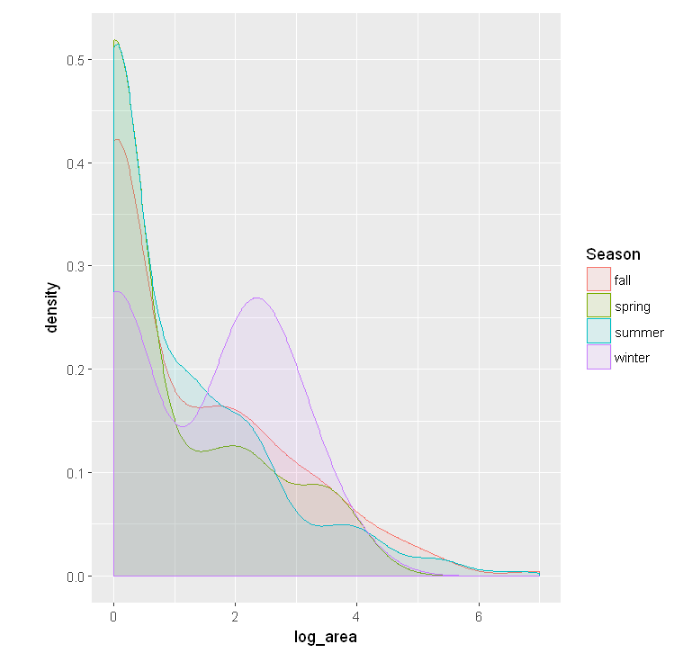
**Day**

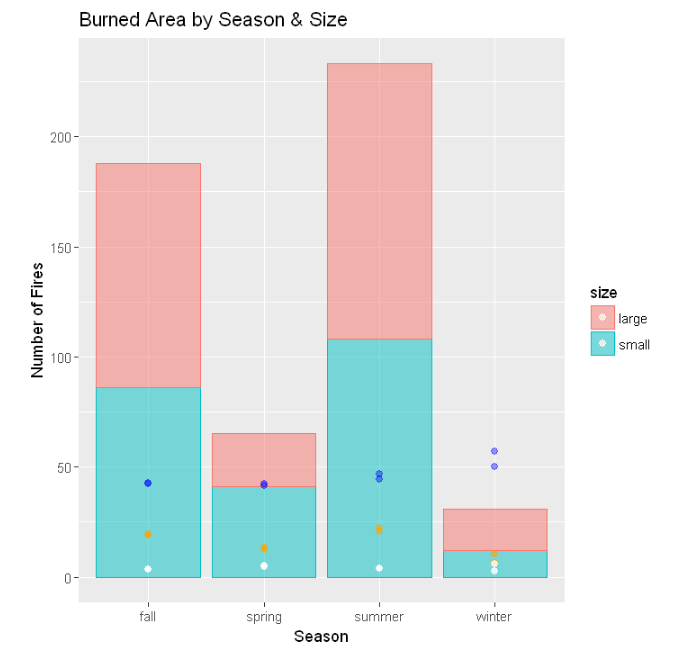
The super large forest fires happened on Monday, Wednesday, Thursday, Saturday, and Sunday. It does not indicate a direct relationship. From Monday to Thursday, the distributions of the fire size are similar. From Friday to Saturday, their distributions are similar. Sunday is special since it has much more small forest fires than other days.

**Season**

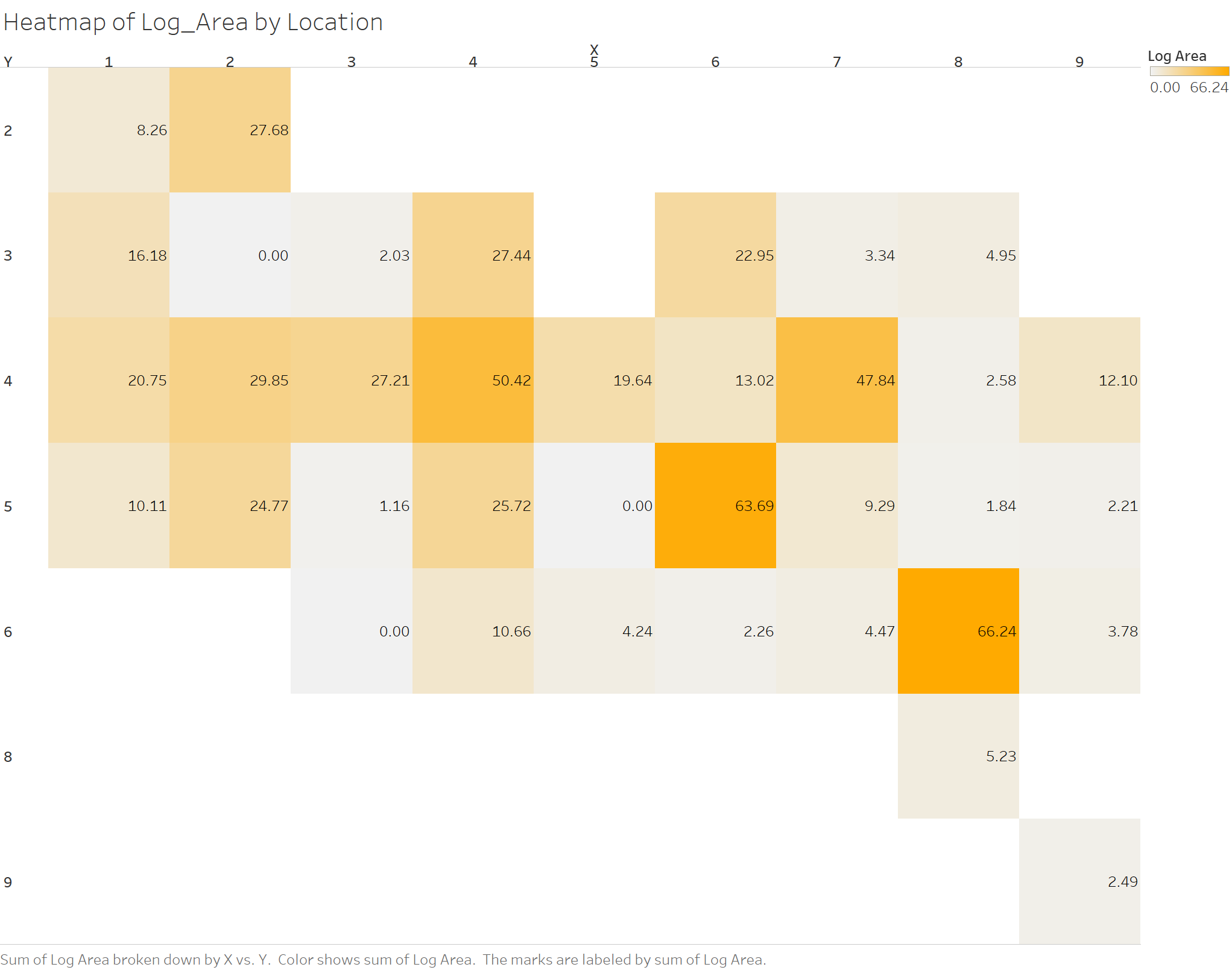
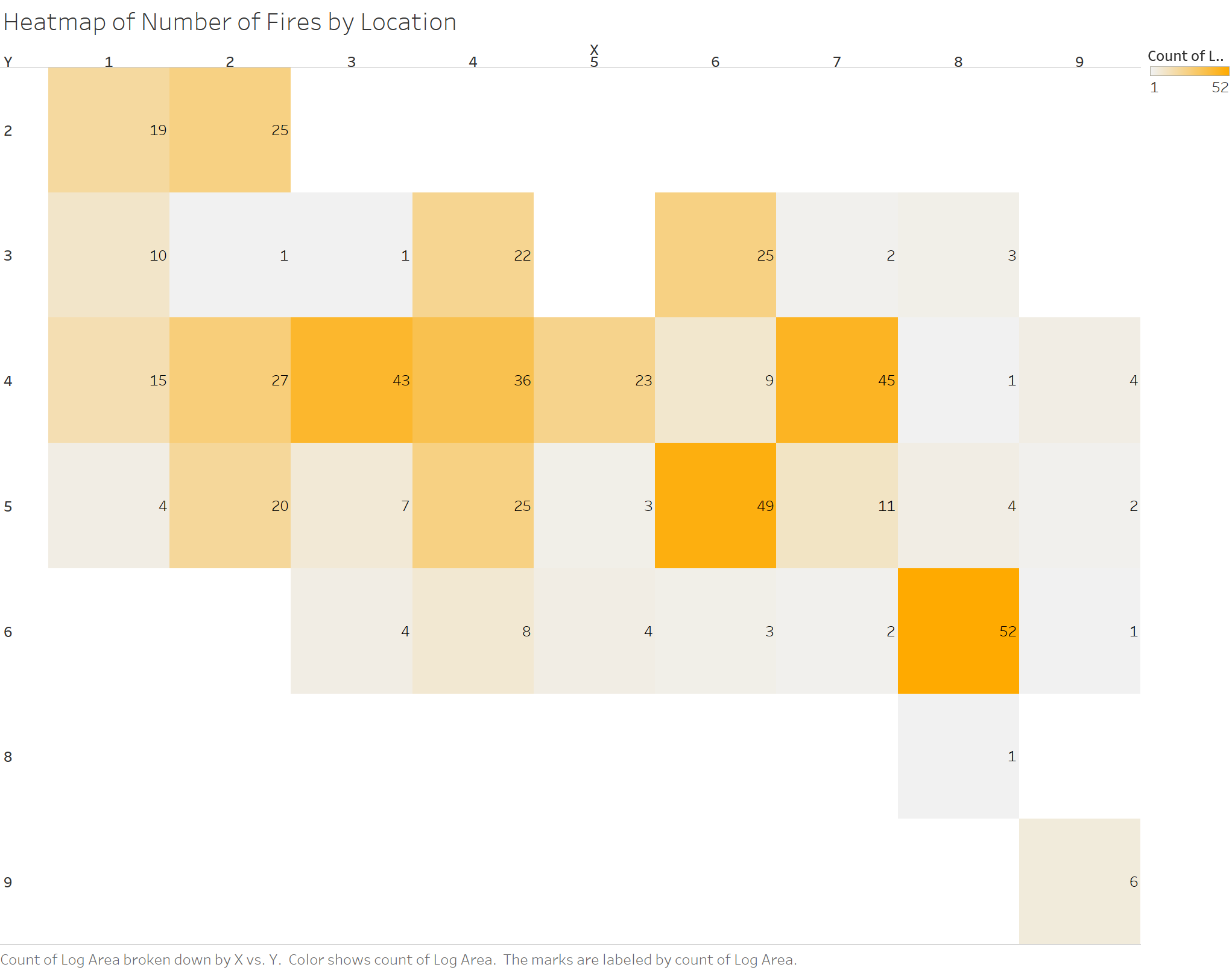
Since some months like January does not have enough observations. I group the month and create a variable called Season. I tried some different ways to group the month, but it does not result in a significant difference. Hence, I do it in a old-fashioned way. For example, spring is from March to May. I build a density plot and stacked bar plot to explore the impact of the seasons.

From the stacked bar plot, I found that Summer has the most occurrence of forest fires, followed by fall. In opposite, winter has the least number of wildfires, but winter has over fifty percent of the large forest fires.

The density plot illustrates the distribution of forest fires. For example, it indicates that Spring, Summer, and Fall have similar distributions of the area (log). However, Winter has a very different distribution with two prominent peaks. It has the other peak with log\_area equals to 2.5.

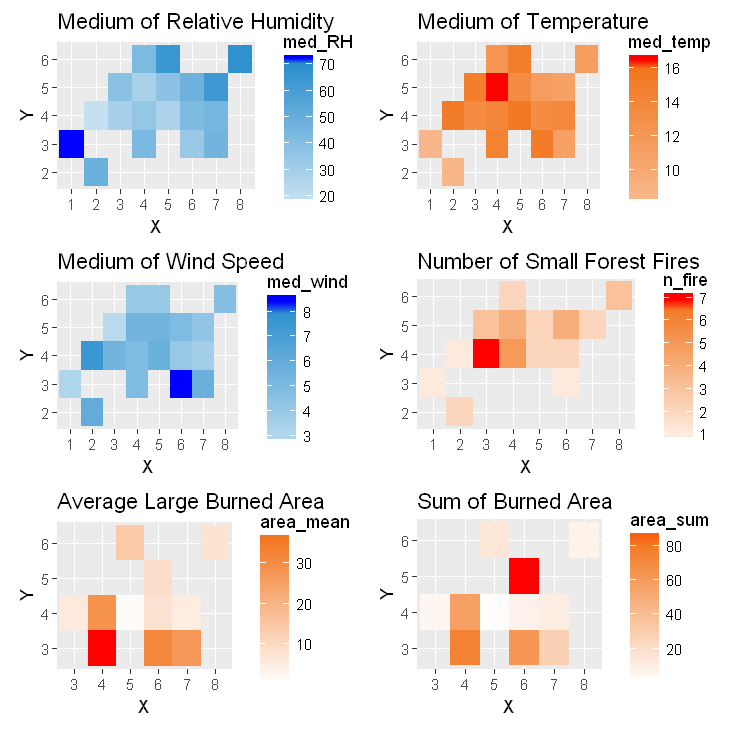
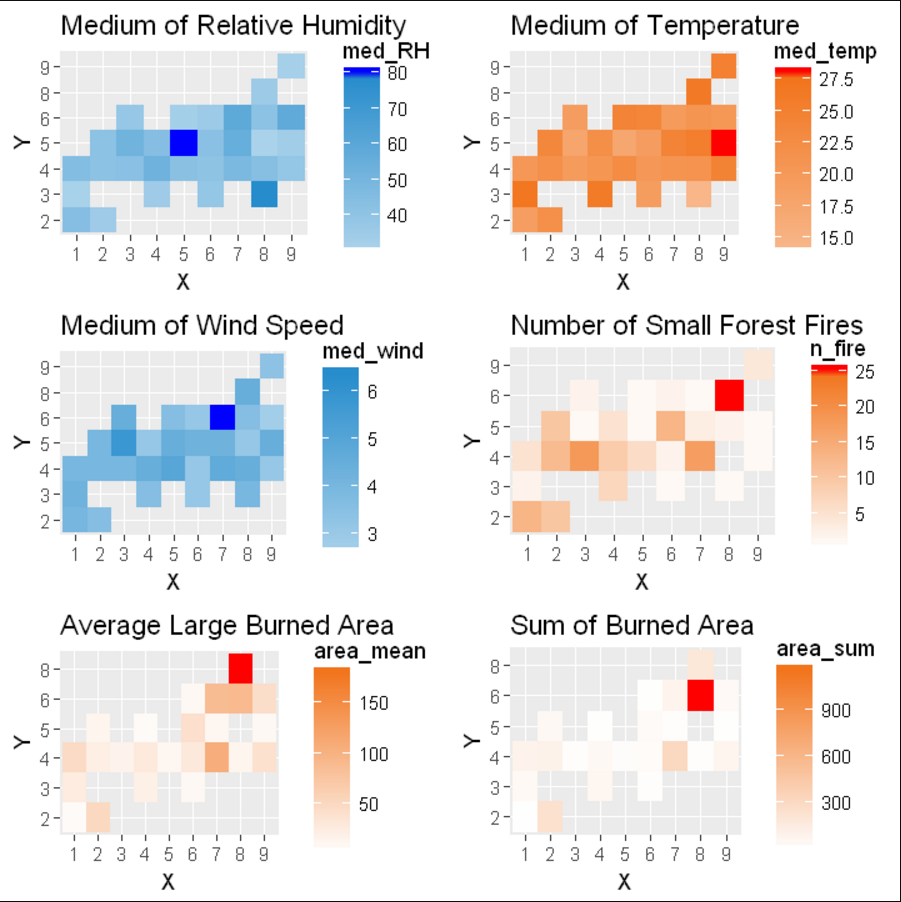
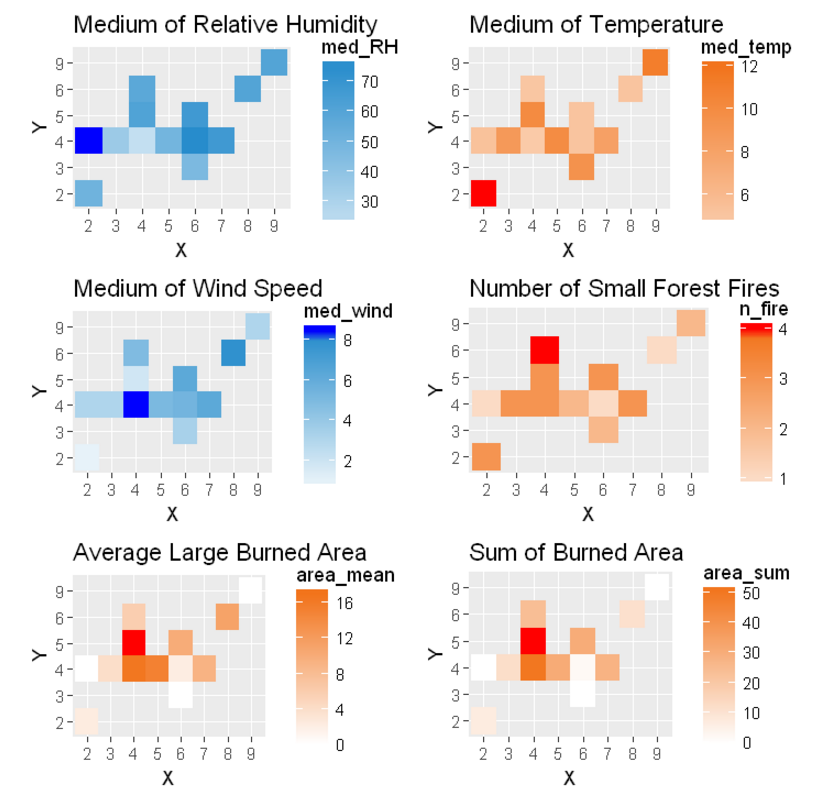


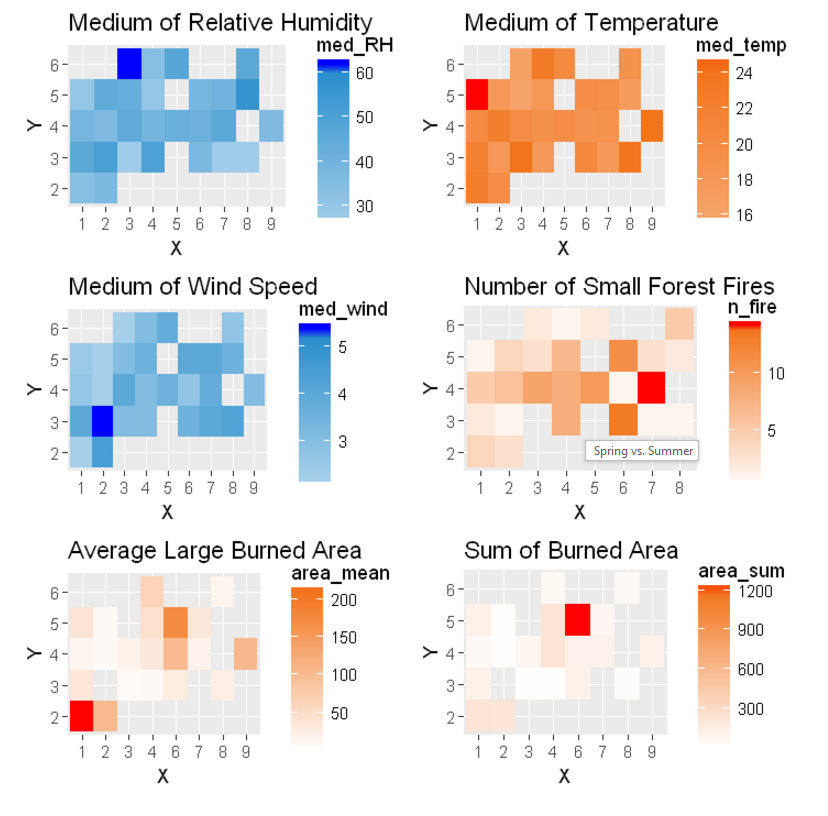
**Location**

**** ****

The park is divided into 36 parts, marked with X and Y. I create two heatmaps. In the first heatmap, the saturation of orange color represents the sum of burned area. In the second heatmap, the saturation of orange color represents the count of forest fires in each section.

The first picture indicates that (8, 6), (6,5), (7, 4) and (4, 4) have the top four largest burned area. From the second picture, we get a similar result. (8, 6), (6,5), (7, 4), (3, 4) and (4, 4) have the top 5 number of wildfires. It is interesting to see that (8, 6), (6,5) and (7, 4) have the highest frequency and severity of forest fires. It worth digging to figure out what factors cause these fires.

******Multi-heatmaps**

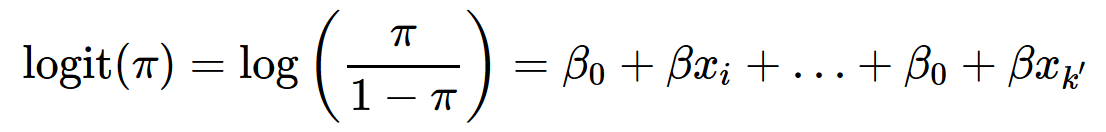


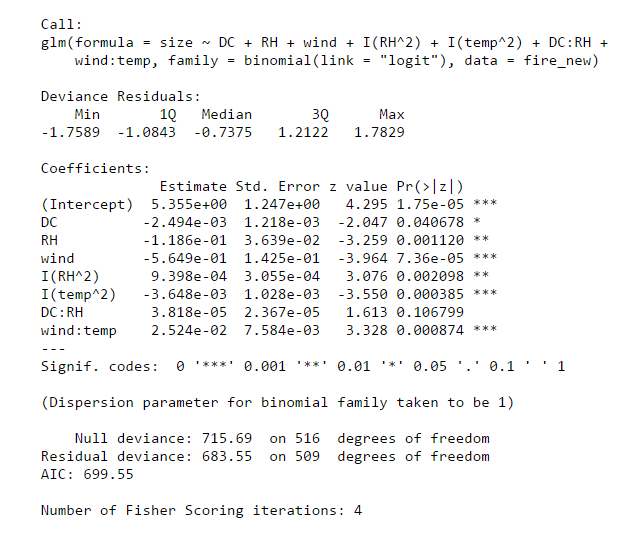
I use the above four multi-heatmaps to find some features in (8, 6), (6, 5) and (7, 4). These features may be the reasons that these locations have such high frequency and severity of forest fires.

Most of the fires in (8, 6) occur in Summer. In summer, it has average relative humidity, average temperature, and relatively low wind speed. In opposite, (6, 5) has many forest fires happened in spring, fall, and winter. It is entirely different to (8, 6). Moreover, most of the forest fires in (7, 4) happen in autumn and winter. It is strange that it only has a few cases happened in Summer, and even less in spring.

Among these three locations, none of them have an extreme value of temperature, wind speed and relative humidity.

**Binary Logistic Regression**

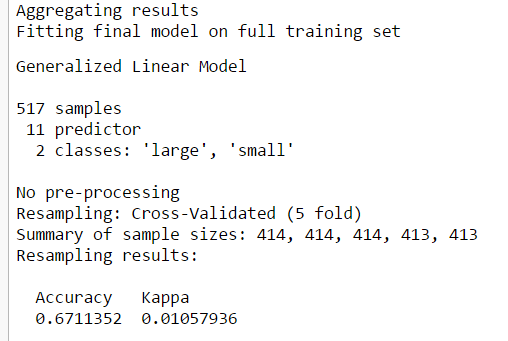




The model is used to predict the probability of large burned area (the size of total burned area is equal to or larger than 0 ha) when a forest fire occurs. I use the backward selection to obtain the above result. I finally choose to include 4 variables with interactions in our population maximal model: Drought Code (DC), Temp (temperature), Wind (wind speed), RH (relative humidity), the interaction between temperature and wind, RH^2 and wind^2.

Deviance is a measure of goodness of fit of a generalized linear model. We have a value of 715.69 on 516 degrees of freedom. Including the independent variables decreased the deviance to 683.55 points on 509 degrees of freedom, not significant reduction in deviance. The Residual Deviance has reduced by 32.14 with a loss of seven degrees of freedom. The model does not fit very well.

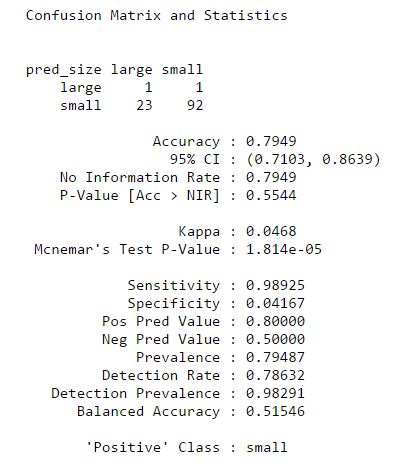
**Logistic Regression (5-fold cross validation)**



In this model, I group the size into 2 groups, large and small. The large forest fires are the ones with area larger than 5 hectares. The testing set’s accuracy score is around 67 percent. However, I realized that the model did a better job predicting the small forest fires. In opposite, it does not do a good job predicting the large forest fires.

If I group the size into 4 groups, small, medium, large and super large, the training accuracy is around 50.57%, and the testing accuracy is around 42.31 percent. The result is not very nice.

**Support Vector Machine**

I also use support vector machine to predict the size of the forest fires. I tried to fit an SVM using a non-linear kernel, the (Gaussian) radial basis and complicated sounding hyperbolic tangent sigmoid. Among these three methods, the Gaussian model works the best.

When I group the forest fires into two groups, large and small, the training accuracy is around 80 percent, and the testing accuracy is around 78 percent.

When I group the fires into four groups, the training accuracy reduces to 48.08 percent, and the testing accuracy reduces to 41.23 percent.

It is interesting to see that the support vector machine did a good job predicting the small forest fires.

**Deep Learning**

I also try the deep learning method. The training accuracy is 48.75 percent. The testing accuracy is 42.31 percent. While I change the parameters, the loss rates are often stuck at 1.22 and the accuracy rates were often stuck at 47.96 percent.

## Conclusion

Being able to forecast the size of a starting fire is crucial for the firefighters to estimate how much human resources and resources should be allocated when a fire just started or even before a fire started.

In this study, the size of burned forest area during a forest fire was estimated using different methods, such as Logistic regression and support vector machine. None of them did a good job predicting the size of wildfires. They all do a good job predicting the small fires.

However, there are too many noises in our data set that prevent us achieving a satisfying accuracy. For example, A wildfire that has the weather opportunity to grow huge were quickly reported and controlled by the fire department. In this case, the data collectors note down “0”. Another possibility is that the same area just had a super large forest fires last month, it is less likely to get another large forest fires although the meteorological condition is ideal for it. I also compared the heatmap by location with the Park map (the cover picture). It shows that many houses and a particular kind trees are located in location (8, 6). It is possible that the fires in that area are affected by human activities, the type of trees and some organic contents that can be burned as fuel.

The current data set we have doesn’t imply the human activities, altitude, and vegetation coverage. I believe that it is crucial to include some more information to this data for better prediction, such as the distribution of houses and population. Also, it may be useful to include the altitude since lightning causes some wildfires.

We need more variables to better predicting the size of the fires as long as we can afford the extra cost. As we mentioned before, we only use meteorological information in this analysis because they are easy and cheap to obtain. We can reapply our model to any location in the world. I believe that at least the distribution of the population can be easily obtained. In the most parts of the world, this information has already been collected.

# References

S. Taylor and M. Alexander. Science, technology, and human factors in fire danger rating:

The Canadian experience. International Journal of Wildland Fire, 15:121–135, 2006.

Natural Resources Canada, (2008, July 11). Canadian Wildland Fire Information System. http://cwfis.cfs.nrcan.gc.ca/background/summary/fwi.

Natureza, T. d. (2014). *montesinho.com - Nature Tourism*. Retrieved from https://www.montesinho.com/en/oparque/mapa

1. [↑](#footnote-ref-1)