**Reasons Behind Forest Fire**



IE6600: Computation and Visualization for Analytic, Fall 2021

Northeastern University

Group Member: Ziqi Li, Qiyue Mao, Siyi Yu, Zihan Zhao

Date: 10/27/2021

**Table of contents**

1. **Background**
2. **Purpose & Objective**
3. **Modeling and Analysis**
4. Data Pre-processing
5. Correlation Matrix
6. Relationship between fire area and monthly tempreture
7. Relationship between fire area and monthly seasons
8. Relationship between fire area and relative humity
9. Relationship between fire frequency and position
10. Relationship between fire area and FFMC, DMC, DC and ISI
11. **Background**

In recent decades, wildfires have been happened frequently, particularly in forested regions that cause economic, ecological damage and impact substantial population areas significantly. In order to prevent such phenomenon, people should highlights the requirement to investigate the relationship between influencing factor and forest fire deeply. To achieve this goal, people need to data mining of datasets of historical forest-fire records provided by meteorological stations. These meteorological conditions, including temperature, wind, relative humidity and rain, influence forest fires mainly. Based on these factors, people also come up with several fire indexes, such as the forest Fire Weather Index (FWI). The forest Fire Weather Index (FWI) is the Canadian system for rating fire danger and it is formed from five components: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), and Buildup Index (BUI, which is not included in the dataset we are going to study). The first three are related to fuel codes: the FFMC denotes the moisture content surface litter and influences ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity. The ISI is a score that correlates with fire velocity spread, while BUI represents the amount of available fuel. The FWI index is an indicator of fire intensity and it combines the two previous components. To make the data more precise, people also recorded the X-Y coordinates, burned area, month and day-in-week when the fire outbreaks.

1. **Purpose & Objective:**

In this project, we will consider forest fire data from the Montesinho natural park and have two main purposes:

* A brief exploration about “forestfire.csv” to realized basic statistics and whether there is missing value.
* A thorough investigation about relationship between key factors (position, time, FWI, temp, humidity ect) and forest fire. Our analysis includes the correlation between factors, temperature's influence to burned area, the influence of wind and rain, relationship between relative humidity and burned area, relationship between x-y position and fire outbreak, and relationship between the 4 forest Fire Weather Index and fire outbreak.

1. **Modeling and Analysis**

***Data Preprocessing:***

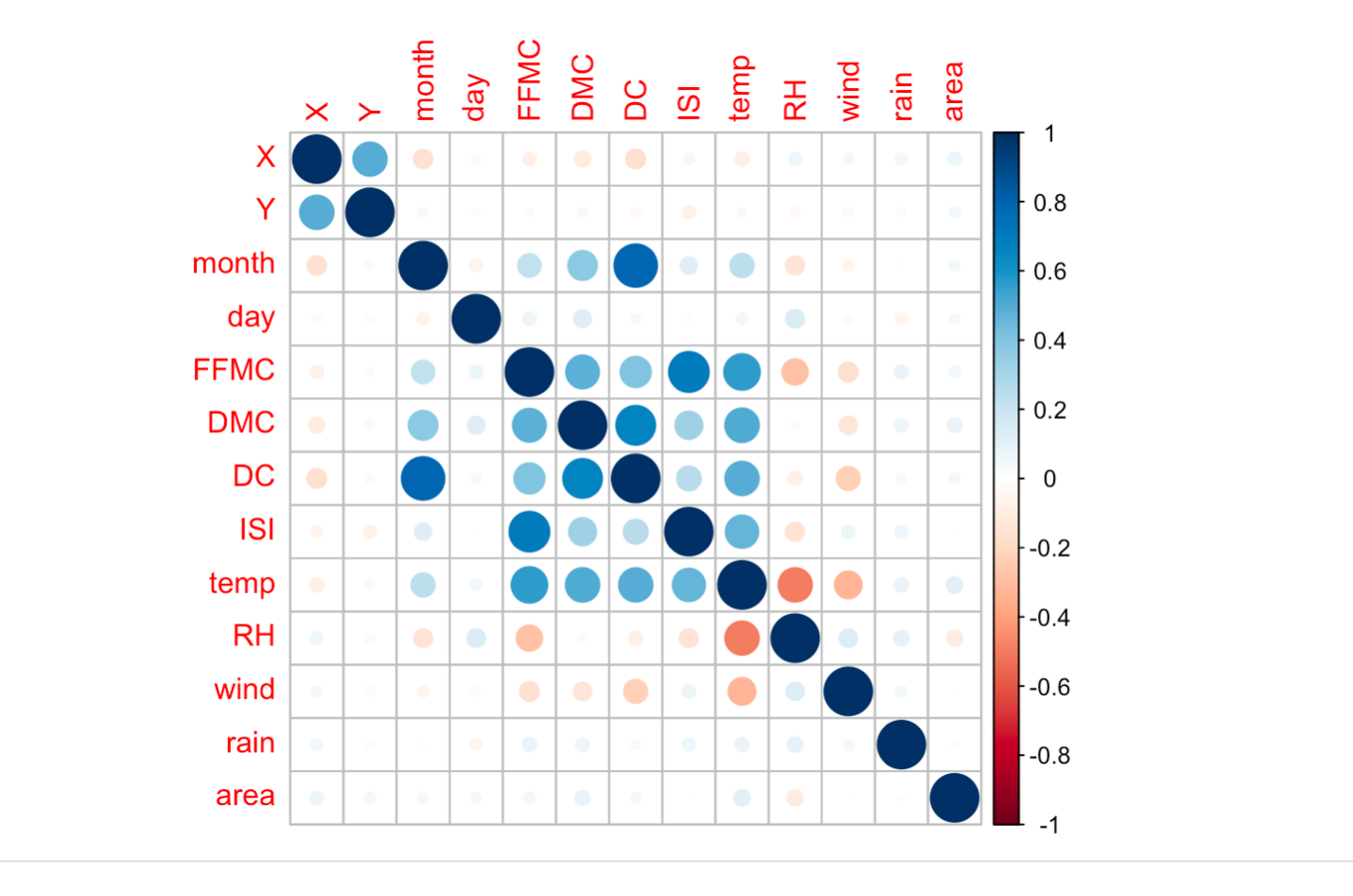
According to data exploration, we find that “forefire.csv” has 517 rows, 13 variables and no missing value, which is well presented. However, there exist some quirks here and there, such as the abbreviations of month and day is not in the right form. But they can be easily negated in R studio.

The first step should be changing month and day into numeric. Since the string in both columns does not match the system abbreviations. We used str\_to\_title to change the first letter to uppercase and matched month with month.abb thus change them into numbers. We also created a similar day.abb to match the day column.

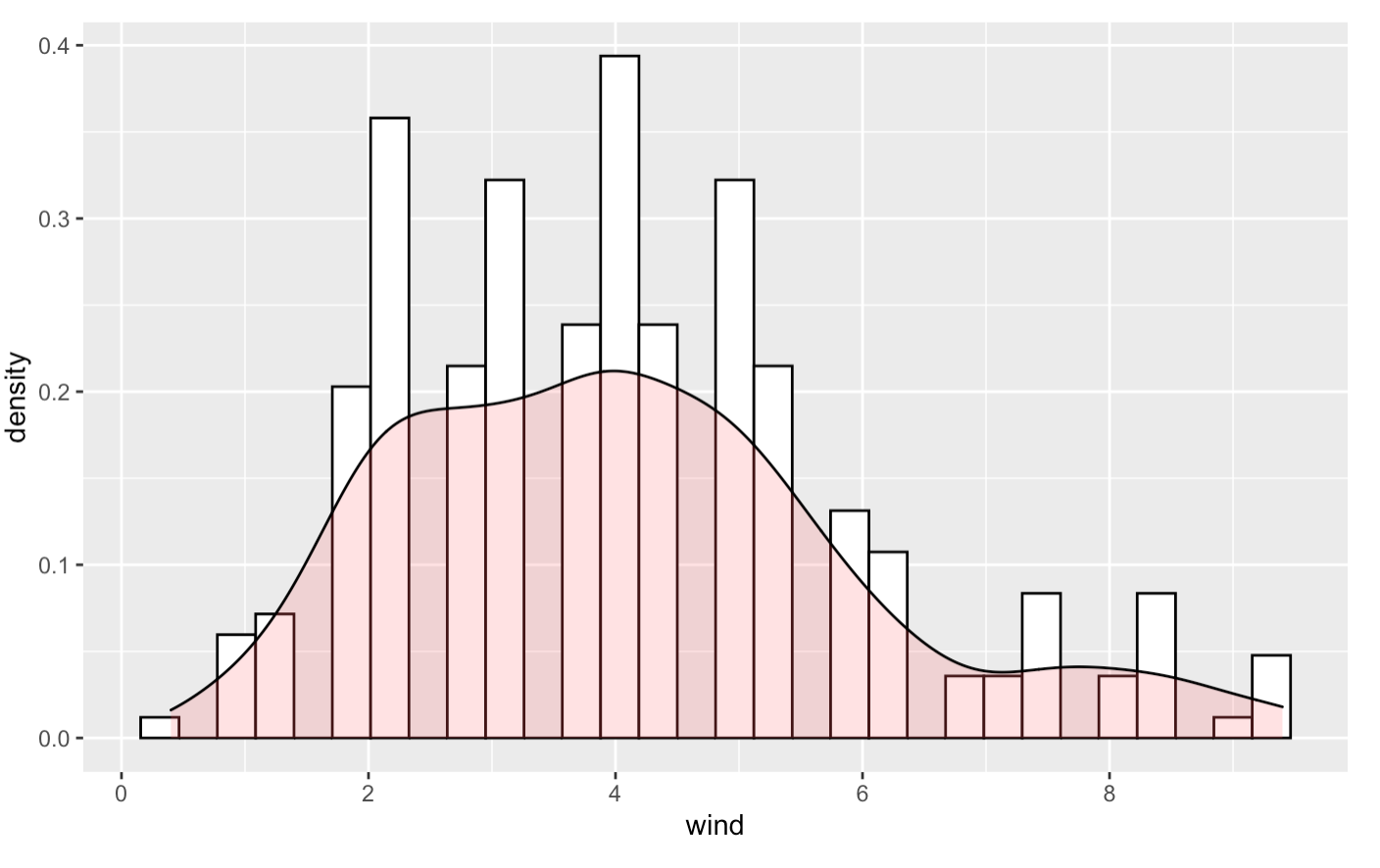
Then we cleaned the data through using filter to delete the data which burned area=0.

***Analysis:***

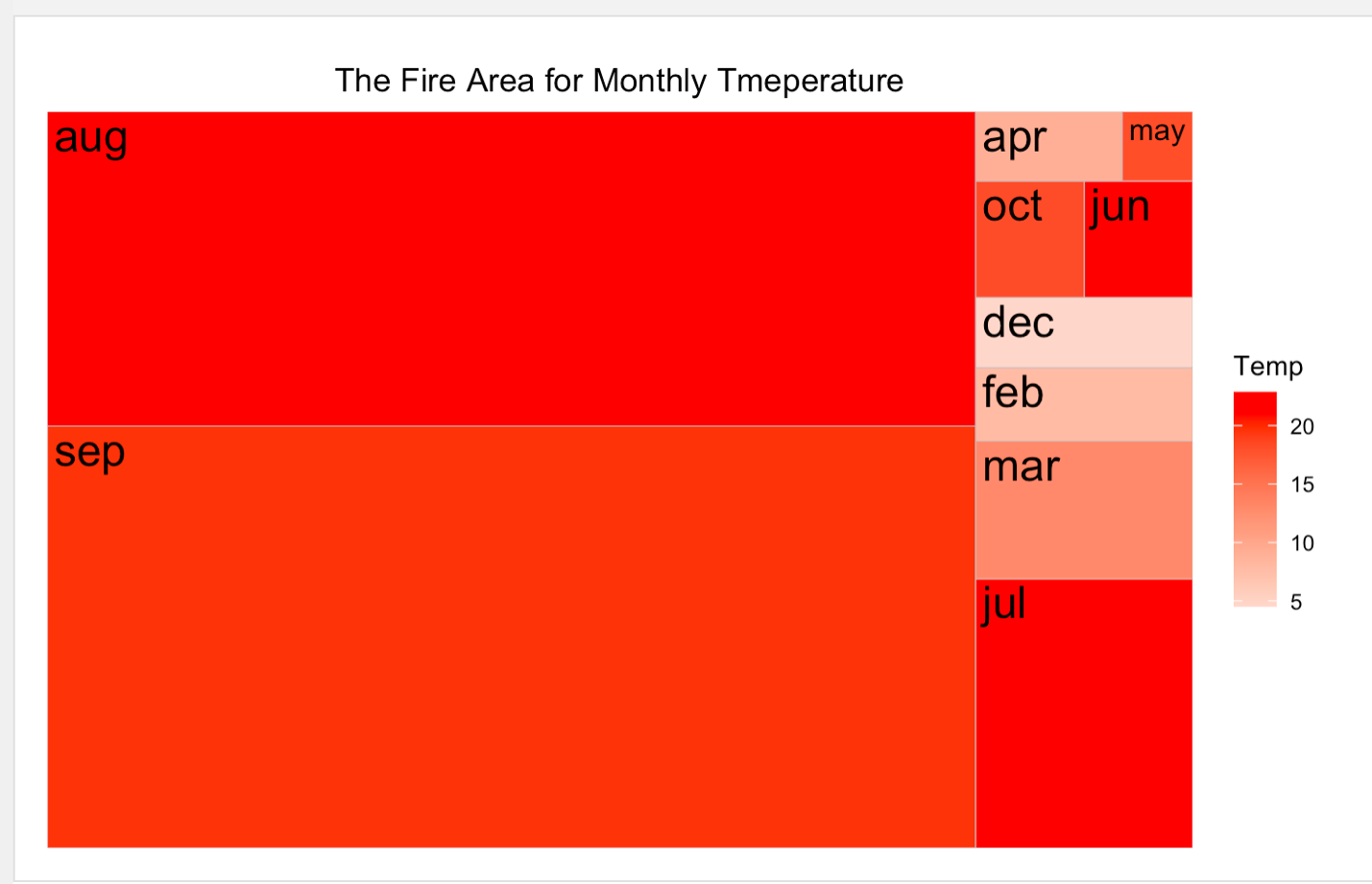
In order to have a brief understanding about the dataset, we first created a correlation matrix. According to the matrix, it is obvious that there exists relatively high Pearson correlation among the 4 forest Fire Weather Index. Also, it is clear that these 4 factors and temperature, DC and month, X and Y, temperature and Relative Humidity are highly correlated with each other.



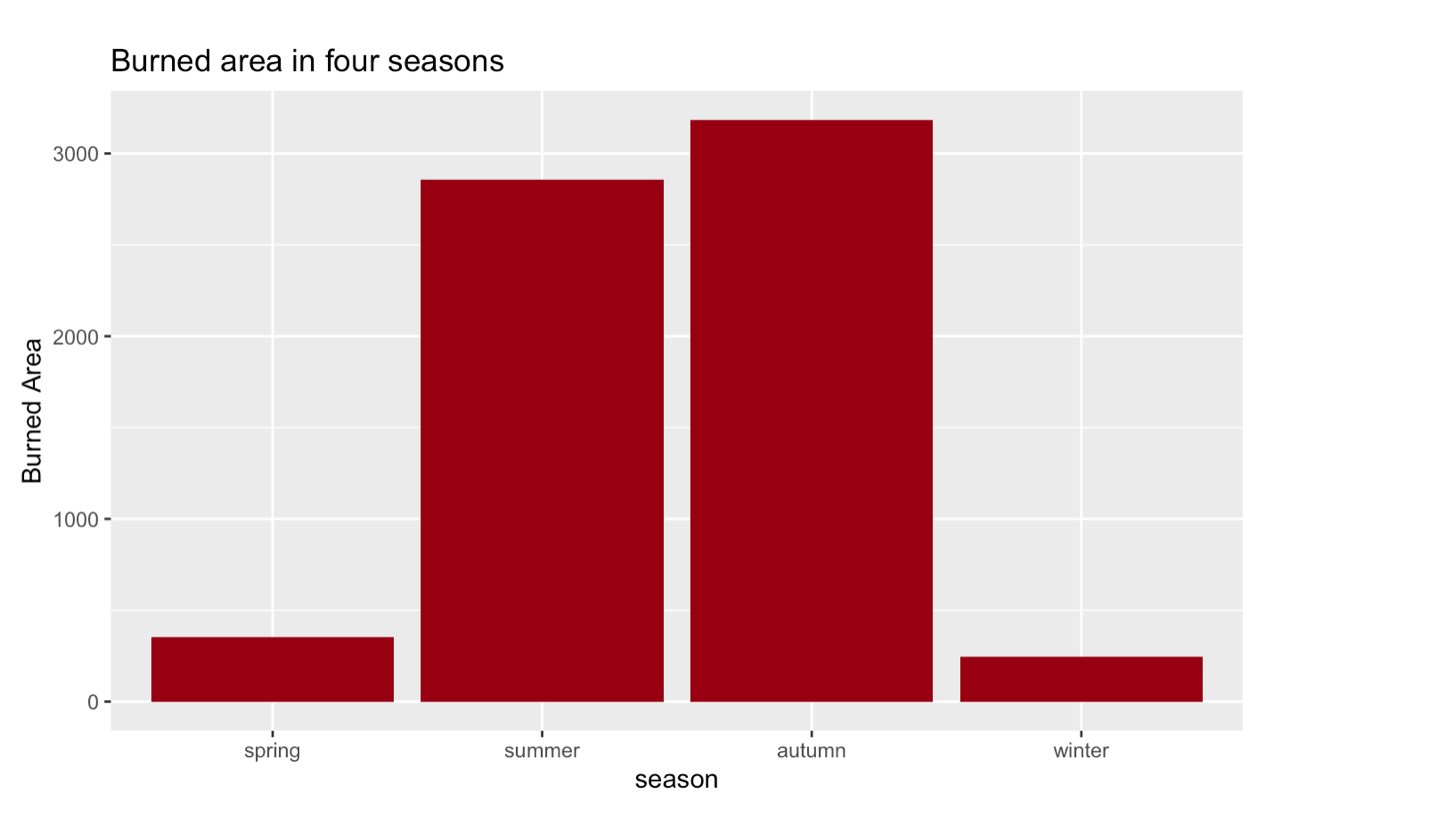
Then we analyzed the influence of wind on forest fire. We want to figure out which windspeed has the highest risk of causing forest fire. To solve our problem, we decided to draw a histogram chart. According to the histogram, most forest fire outbreak happened when the windspeed is 4.2km/h. There also exist many forest fire when the windspeed is 2.2. Besides, in the chart we realized wind and fire outbreaks might be normally distributed, to prove this, we added a density line. However, unlike what we guessed, the chart is right skewed.



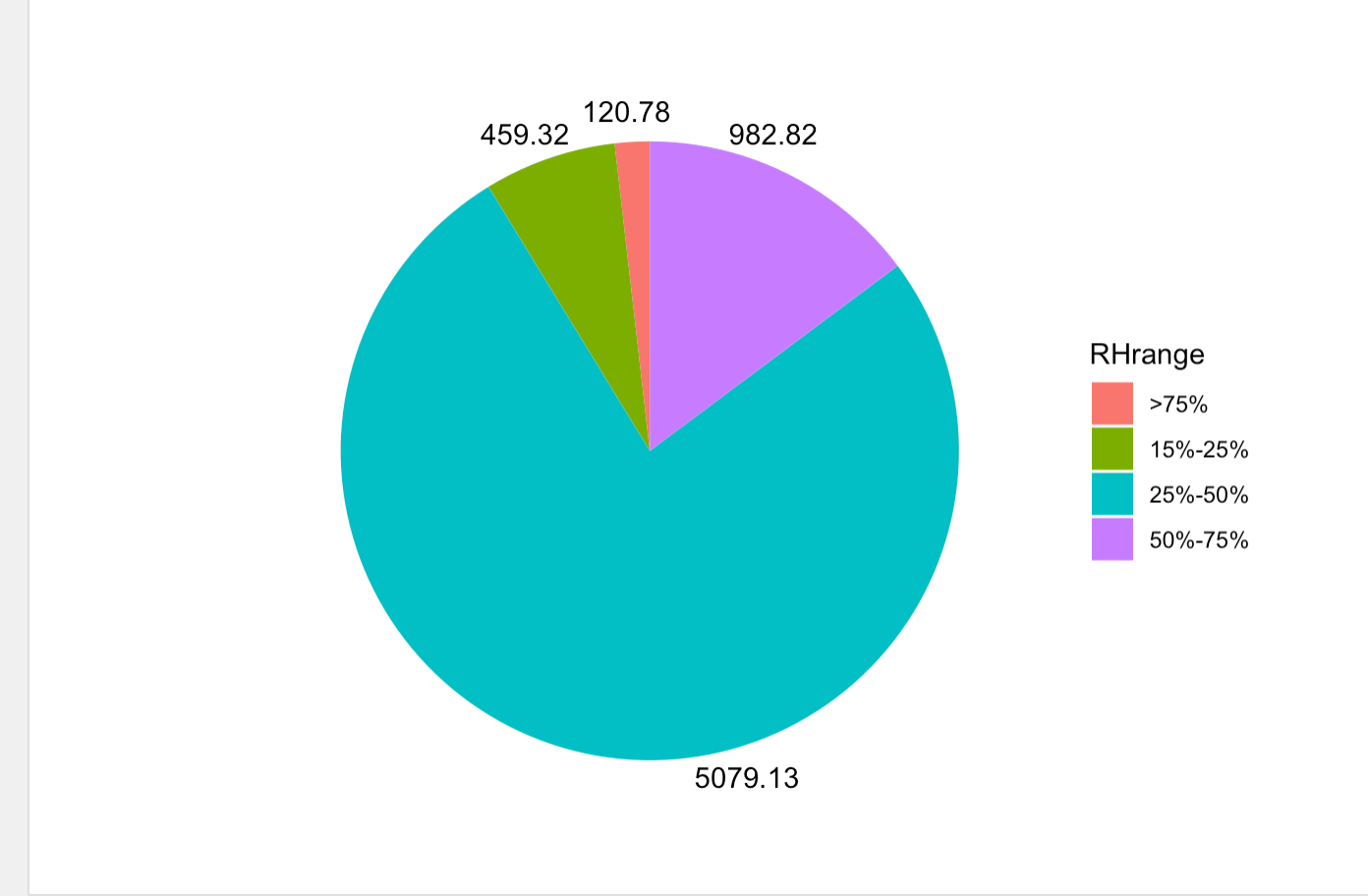
The next question we worked with is the influence of monthly average temperature's on burned area. In this graph, we tried to find the relationship between temperature and area that forest fires burned. We decided to use tree map for the data visualization. We first is simply compute monthly average temperature as X axis, and set the sum of burned area each month as Y axis and then make the filter as the average temperature every month. The area of each element shows the size of burned area each month, which can be easily distinguished. From the graph, it is easily to find that the months with higher temperature tend to have larger burned area than others, especially August and September.



In order to have another perspective on this problem, we divide 12 months as four seasons. With each season's average temperature, the data will be easier to recognize. The obvious differences of temperature and burned area among seasons can help present the data better. For example, the monthly result of July, August and September in the previous chart matches the summer and autumn part in this chart, which shows the largest burning area among seasons.

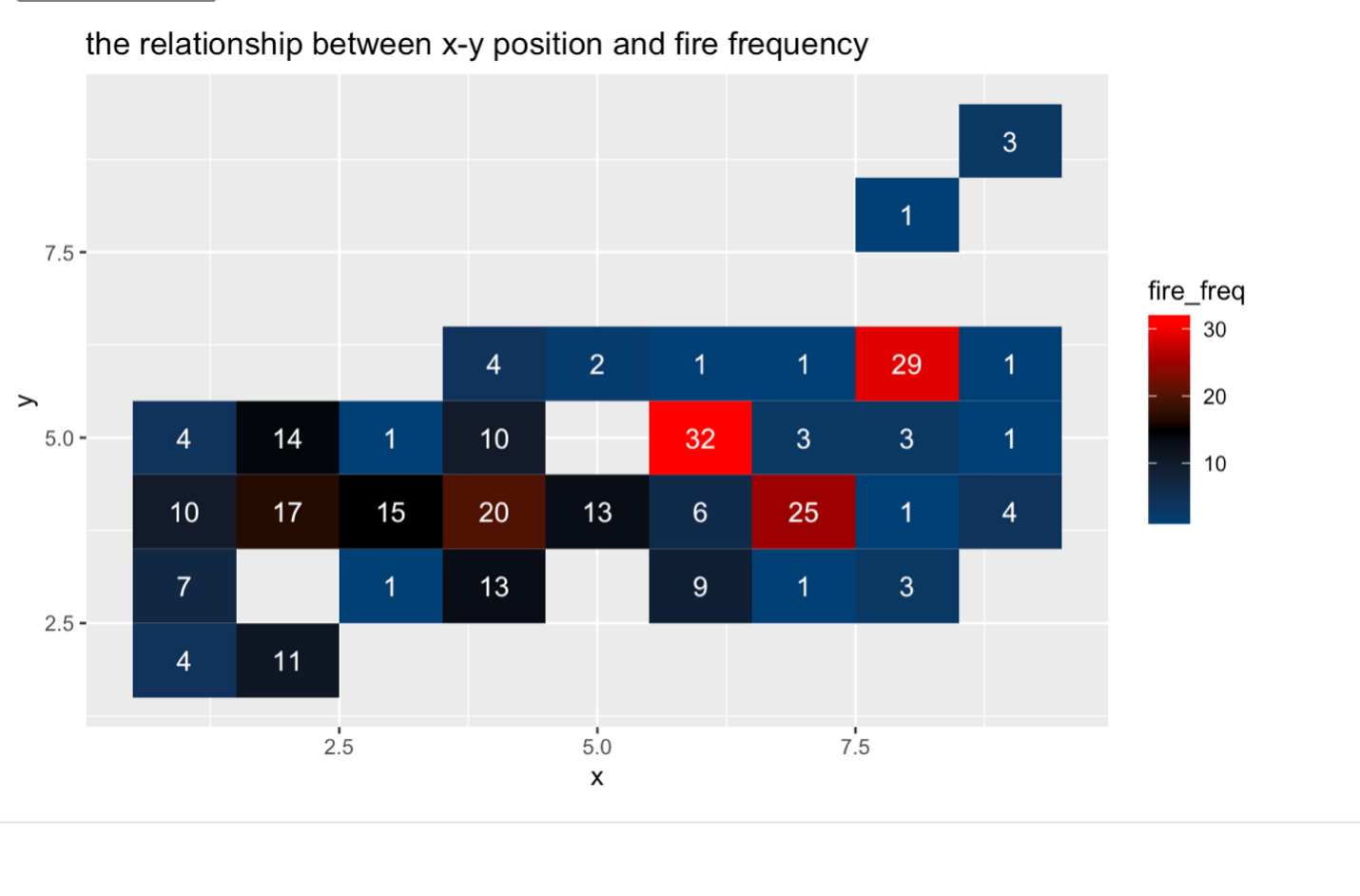


Besides temperature, we are also interested in the relationship between relative humidity and burned area. In order to achieve our goal, we make a new column called Relative Humidity Range to classify the data. We separated the data in four parts: 15%-25% RH, 25%-50% RH, 50%-75% RH, >75% RH. According to the chart, when the relative humidity over 75%, there is only a tiny area is burned. From the result, it is also clearly showed that most of burned area have the relative humidity between 25%-50%.

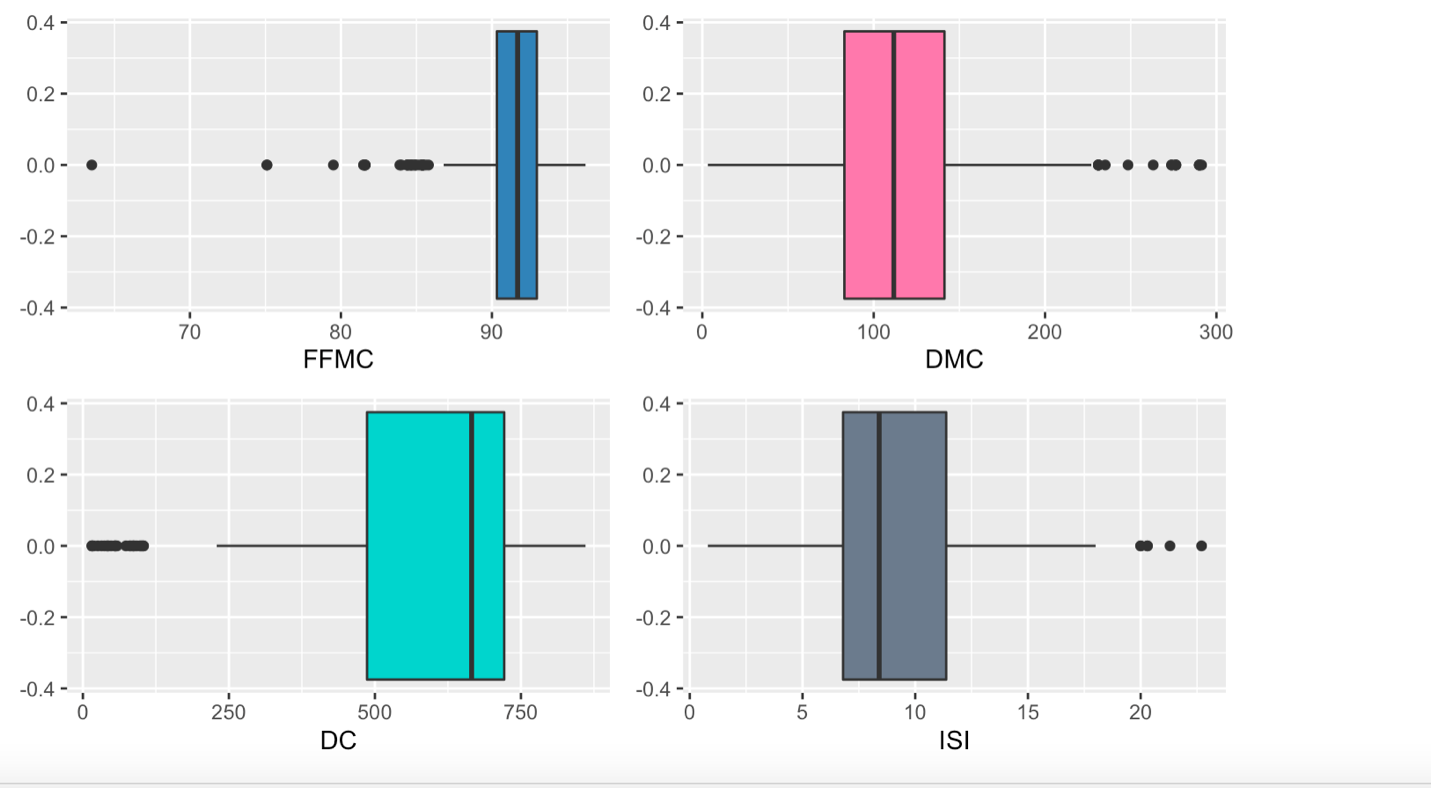


After that we started to analyze the relationship between location (X-Y coordinated) and fire outbreak number.

In order to better present the dataset, we decided to choose the heat map to illustrate. We simply used column X, the x-axis spatial coordinate, as x-axis and Y, which is y-axis spatial coordinate, as y axis. Then we fill the graph by the value of fire outbreak number. According to the heat map, we conclude that the frequency of forest fires in different place is totally different. These three locations, ((6.5),(8,6),(7,4)), have more fire outbreaks than other locations so that the workers in those areas need to pay more attention to prevent forest fires.



Through our study of the data, we gained interest in the relation between FFMC, DMC, DC and ISI. In order to find their relationship with the outbreak of forest fire, we built several boxplot to see the distribution.

According to the boxplot, there tend to be more fire outbreak with higher DC—it is obviously left skewed and have high medium lines, the median is around 666. FFMC is normally distributed, with its median around 92. The chart of ISI is slightly right skewed with median around 8. The chart of DMC is normally distributed with median around 111. 

Citations

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data.

In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence,

Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December,

Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.

Available at: http://www.dsi.uminho.pt/~pcortez/fires.pdf