COSC 757 Data Mining Assignment 1

Daphne McWilliams
Department of Computer Science
Towson University
Towson, MD
dmcwil2@students.towson.edu

***Abstract*—This paper applies exploratory data analysis, data preprocessing techniques, and regression analysis to the Forest Fire Dataset from UCI's Machine Learning Repository.**

I. Introduction

Wildfires are a natural occurrence in some forest ecosystems; on a large enough scale, however, these fires can cause significant ecological and economical damage and widespread suffering for the many living things caught in their wake, thus becoming a major environmental concern. Fires have a variety of causes, ranging from lightning to human error, and are best mitigated by quick detection. The detection and prevention of wildfires is of particular importance to us today, at a time when climate change and forest degradation have led to hotter and more fire-prone conditions globally, and fires are more likely to burn for longer and over larger areas.[1]

In the past, meteorological data has been incorporated into numerical indices, which are used for prevention and to support fire management decisions. A notable example is the Fire Weather Index system (FWI), which rates fire danger on a numerical scale. This system was designed in Canada in the 1970s and required calculations based on readings from four meteorological observations (temperature, relative humidity, rain, and wind) that could be manually collected in weather stations. This index is still in use today, not only in Canada, but in many other countries around the world, and is correlated with fire activity in these countries despite their differing climates.[2]

In more recent years, advances in computing and information technology have enabled the storage of vast quantities of data in sophisticated databases. This data holds valuable information, such as trends and patterns, which can be accessed through data mining techniques and can then be used to improve human decision-making. Meteorological data is abundant and readily available, and so the detection of forest fires is an area where data mining techniques may be able to deliver a clear advantage to humans hoping to mitigate or prevent damage altogether.[3]

II. Dataset Description

The creators of this dataset collected data based on the FWI from northeastern Portugal, with the goal of predicting the burned area, or size, of forest fires. The FWI rates fire danger on scale from 0 to 20 and depends on five components, which are calculated from consecutive daily observations of temperature, relative humidity, wind speed, and 24-hour precipitation. Further details about the five factors are below, although the last factor—the Buildup Index, or BUI—is not included as an attribute in this data set.

The Fine Fuel Moisture Code (FFMC) denotes the moisture content of surface litter and influences ignition and fire spread. It is intended to represent moisture conditions for litter fuels under the shade of a forest canopy and has the equivalent of a 16-hour time lag, or memory of past weather conditions. It ranges from 0–101, and is calculated from observations of temperature, relative humidity, wind speed, percipitation, and the previous day's FFMC. FFMC scores of 70 and above are considered "high" and correspond to increased risk of fire.

The Duff Moisture Code (DMC) is a numerical rating of the average moisture content of loosely compacted organic layers of moderate depth (2–10 cm). It affects fire intensity and represents the moisture conditions for the equivalent of 15-day (or 360-hour) time lag fuels. It is unitless and open ended, and is calculated from observations of temperature, relative humidity, precipitation, and the previous day's DMC. A DMC rating of more than 30 is considered high, and a rating above 40 indicates elevated risk of burning

The Drought Code (DC) is a numerical rating of the moisture contenet of deep, compact organic layers (> 10 cm). It is a useful indicator of seasonal drought and shows the likelihood of fire involving the deep layers of the forest floor. A long period of dry weather is need to dry out these layers. The DC approximates moisture conditions for the equivalent of 53-day (1272-hour) time lag fuels. It is calculated from observations of temperature, precipitation, and the previous day's DC. The DC rating is unitless, with a maximum value of 1000. A DC of 200 is considered high, and 300 or more indicates elevated risk of fire involving deep subsurface and heavy fuels.

The Initial Spread Index (ISI) integrates fuel moisture for fine dead fuels and surface wind speed to estimate spread potential. It is a key input for fire behavior predictions and is unitless and open ended. Ratings of 10 and above indiate a high risk of spread shortly after fire ignition. A rating of 16 or more indicates potential for extremely rapid spread.

Lastly, the Buildup Index (BUI) combines the current DMC and DC to produce an estimate of potential heat release in heavier fuels. It is unitless and open ended. [4]

The data comes from the Montesinho Natural Park in the northeast region of Portugal and was collected from January 2000 to December 2003. It comes from two sources: The first database was collected by the inspector that was responsible for the Montesinho fire occurrences. Every time

[1] https://www.globalforestwatch.org/topics/fires/#slides

[2] https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system

[4] https://www.malagaweather.com/fwi-txt.htm

a forest fire occurred, several features were registered, such as the time, date, spatial location within a nine-by-nine grid, the type of vegetation involved, the six components of the FWI system, and the total burned area. The second database contained several weather observations (e.g., wind speed) that were recorded within a 30-minute period by a meteorological station located in the center of the park. Information from these two databases were integrated into a single dataset with a total of 517 entries.

The dataset contains a total of 13 attributes, which are listed in the table below along with their units and ranges (Table 1).

TABLE I. DATASET ATTRIBUTES

Name	Description
x	x-axis coordinate (1-9)
y	y-axis coordinate (1-9)
month	January to December
day	Monday to Sunday
FFMC	FFMC code (18.7 to 96.20)
DMC	DMC code (1.1 to 291.3)
DC	DC code (7.9 to 860.6)
ISI	ISI index (0.0 to 56.10)
temp	Temperature in C° (in C°: 2.2 to 33.30)
RH	Relative humidity(in %: 15.0 to 100)
wind	Wind speed (in km/h: 0.40 to 9.40)
rain	Rain (in mm/m^2: 0.0 to 6.4)
area	Total burned area (in hectares: 0.00 to 1090.84)

The first four attributes of the data set denote the spatial and temporal attributes. Only two geographic features were included—the x and y-axis values where the fire occurred within the Montesinho park map (Fig. 1). Month ('jan' to 'dec') and day of the week ('mon' to 'sun') were selected as temporal variables.

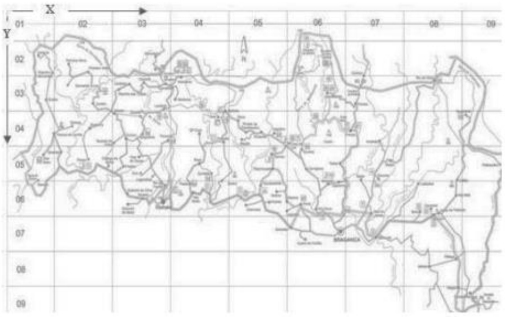

Fig. 1. Map of the Montesinho Park (Cortez and Morais 2007, pg. 4)

The next four attributes are the four FWI components that are affected directly by the weather conditions: the FFMC index, the DMC index, the DC index, and the ISI index. The BUI and FWI index itself were discarded since they are dependent on the previous values. The next four variables are temperature, relative humidity (RH), wind speed, and outside rain, from which the four FWI components are derived.

The final variable (area) refers to the burned area of the forest. It exhibits a strong skew, with the majority of the fires being of a small size. The creators of this dataset note that this skewed trait is also present in fire data from other countries, such as Canada In the dataset under investigation, there are 247 samples with a zero value. However, all entries denote fire occurrences, and zero value means that an area lower than 1ha/100 = 100m^2 was burned.[5]

III. Exploratory Data Analysis

Exploratory data analysis of this dataset was conducted using RStudio, an open-source integrated development environment (IDE) for the R programming language.

A. *Distribution of Attributes*

Below are a series of histograms showing the distribution of various attributes in the dataset.

1) X and Y-Axis Coordinates

These variables indicate the location of fires within the park. Location is plotted on a nine-by-nine grid (as shown in Fig. 1). The histograms below show the distribution of fires along the x (Fig. 2) and y axes (Fig. 3) separately. Taken together and compared against the map of the park, the histograms show a slight tendency of fires to cluster in the northwestern region of the park (x values between 0 and 5, and y values between 2 and 4).

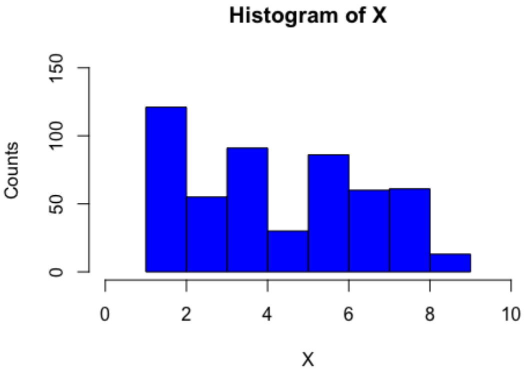

Fig. 2. Histogram of Fire Lcoation on X Axis (West to East)

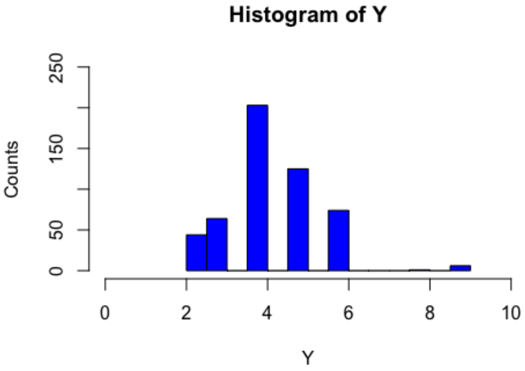

Fig. 3. Histogram of Fire Location on the Y Axis (North to South)

[5] Cortez and Morais 2007, pg. 5

2) *Month and Day of the Week*

Month (‘jan’ to ‘dec’) and day (‘sun’ to ‘mon’) are given as categorical variables in the dataset, and they are represented here with bar charts.

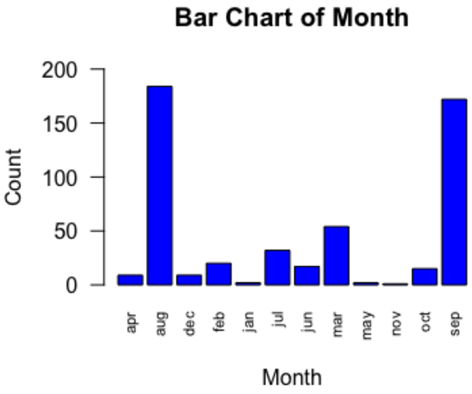

Fig. 4. Month

RStudio arranged the variables on the x axis alphabetically, but it’s still possible to discern a trend. Clearly, the vast majority of fires occurred in the months of August and September, which are the hottest and dryest months of the year.

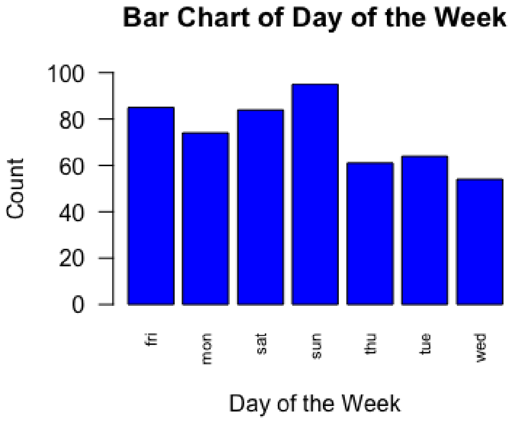

Fig. 5. Day of the Week

In Fig. 5, a relationship can be seen between day of the week and liklihood of forest fire, with the majority of fires occuring on Friday, Saturday, and Sunday. Fires often have human causes, so it makes a certain amount of sense that the majority would occur over the weekends, when human presence in the park would be at its most concentrated.

3) *FFMC, DMC, DC, ISI*

These variables correspond to the components of the FWI included in the data set. Their distribution is show in Fig. 6.

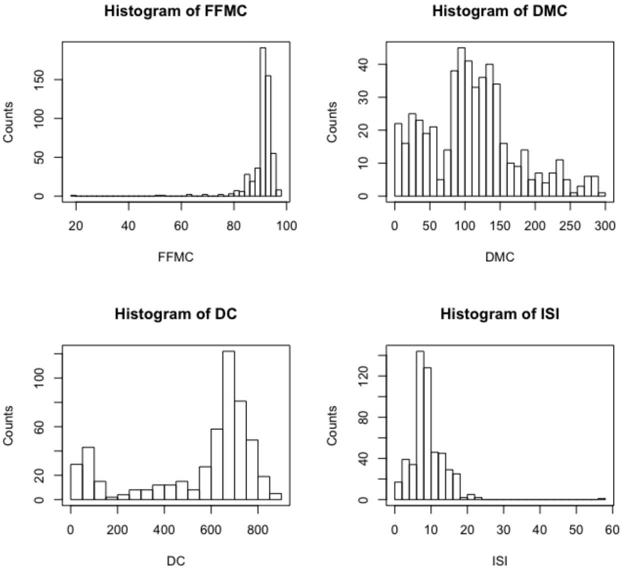

Fig. 6. Historgrams for FFMC, DMC, DC, and ISI

The majority of the FFMC data is grouped towards the high end of the scale, between 80 and 100. As mentioned earlier, FFMC scores of 70 and above are considered high and correspond to increased risk of fire. A high FFMC score indicates that small surface fuels, such as litter, leaves, needles, and small twigs, are quite dry and naturally more prone to combustion. The moisture content of these fuels is very sensitive to the weather, so even a day of rainy, dry, or windy will significantly affect their flammability. The mean value for for this attribute is quite high at 90.64, indicating that the matter on the surface of the forest floor is very dry.

Values for the DMC variable show a more even distribution than those for the FFMC variable, with the majority of data points being grouped towards the middle of the histogram (80–150). However, as mentioned earlier, a DMC rating of more than 30 is dry, and above 40 indicates elevated risk of burning. The mean value for this attribute is 110.9, which is very high and indicates that the organic matter 2–10 cm below the surface of the forest floor is extremely dry.

Values for the DC variable show a strong tendency towards the higher end of the scale, with the majority of datapoints being gouped from 600-800. As mentioned earlier, a DC of 200 is considered high, and 300 or more indicates elevated risk of fire involving deep subsurface fuels. The mean value for this attribute is 547.9, which is very high and indicates extreme dryness of the soil even in its deeper layers.

Values for the ISI variable are grouped towards the lower end of the scale. As mentioned earlier, ratings of 10 and above indiate a high risk of spread shortly after fire ignition. A rating of 16 or more indicates extremely rapid spread. The mean value for this attribute in Montesinho Park is 8.40, which indicates a moderate risk of spread.

4) *Temperature, RH, Wind, and Rain*

These variables correspond to weather conditions in the park. Their distribution is show in Fig. 7.

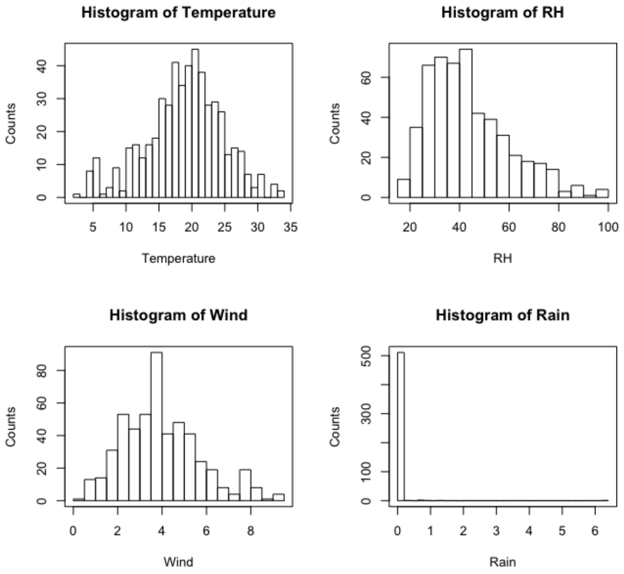

Fig. 7. Historgrams for Temperature, RH, Wind, and Rain

As can be seen in the figure above, relative humidity (RH) for the park tends towards the low end of the scale, which is likely a factor in the high average FFMC and DMC ratings. Additionally, it can be seen that Montesinho Natural Park is an area that gets very little rain, which may contribute to its relatively high average FFMC, DMC, and DC values.

5) *Area*

The distribution of the area attribute is shown in Fig. 8 below. As noted earlier, it is highly skewed towards zero, meaning that the majority of fires recorded were smaller than 1ha/100 = 100m^2.

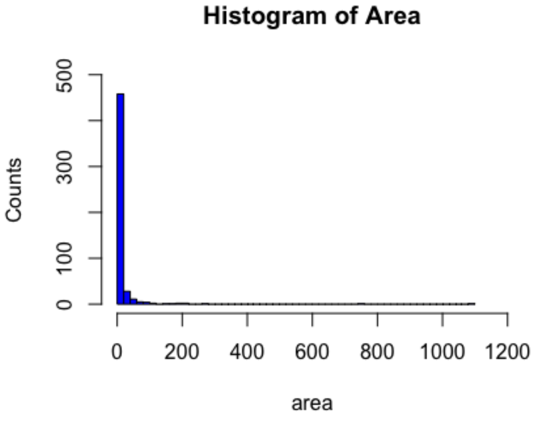

Fig. 8. Area

B. *Relationship Between Attributes*

Below are a series of scatterplots representing the relationship between area and various other variables in the dataset. I chose not to plot the FWI variables against the weather variables, since many components of the FWI index are calculated directly from measurements of temperature, relative humidity, rain, and wind, as detailed earlier.

1) *Area vs. Month and Day of Week*

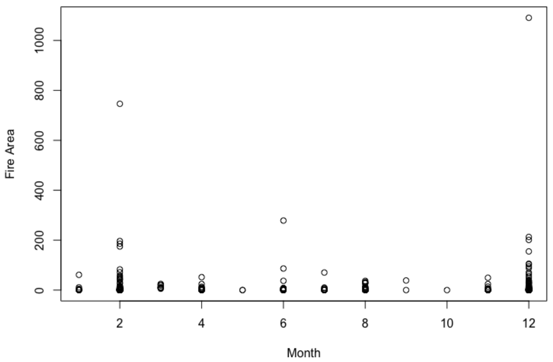

Fig. 9. Scatterplot of Area vs. Month (1 = April, 2 = August, 3 = December, 4 = February, 5 = January, 6 = June, 7 =July, 8 = March, 9 = May, 10 = November, 11 = October, 12 = September)

Fig. 9 shows that the largest fires seem to occur in August, June, and September, which tend to be hotter and drier. They also coincide with summer break for schools, and probably also with higher numbers of people visiting the park.

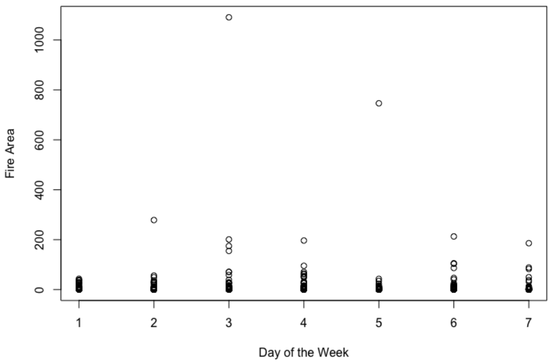

Fig. 10. Scatterplot of Area vs. Day of Week (1 = Friday, 2 = Monday, 3 = Saturday, 4 = Sunday, 5 = Thursday, 6 = Tuesday, 7 = Wednesday)

As can be seen in Fig. 10, the largest fires occurred on Thursday, Friday, and Saturday.

2) *Area vs. FFMC, DMC, DC, and ISI*

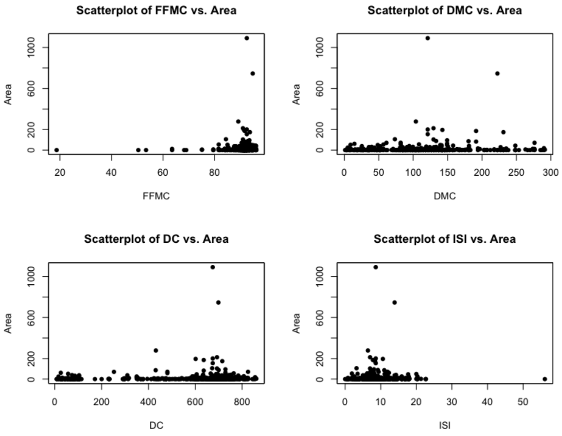

Fig. 11. Scatterplot of Area vs. FFMC, DMC, DC, and ISI

Looking at the scatterplots in Fig. 9, it seems that the larger fires occurred when the FFMC, DMC, DC ratings were all relatively high. High ratings on these scales correspond to very dry conditions in the surface, middle, and deep layers of the soil.

3) *Area vs. Temperature, RH, Wind, and Rain*

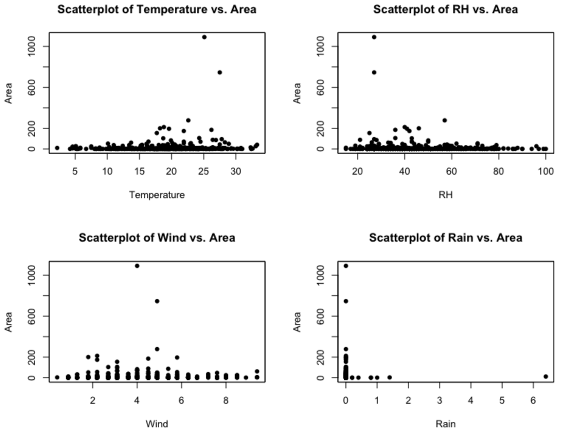

Fig. 12. Scatterplot of Area vs. Temperatre, RH, Wind, and Rain

Looking at the scatter plots in Fig. 10, it seems that larger fires occur when the temperature is high and the relative humidity is low.

IV. Data Preprocessing

Data preprocession for this dataset was also conducted using RStudio. In the sections that follow, the results of normalization on the FFMC, DMC, DC, and ISI variables are displayed alongside histograms of the non-normalized data. Additionally, the DMC variable is binned according to the equal width and k-means clustering methods. Finally, the area variable is transformed through natural log, inverse square root, and square root methods.

A. *Normalization*

1) *Min-Max Normalization*

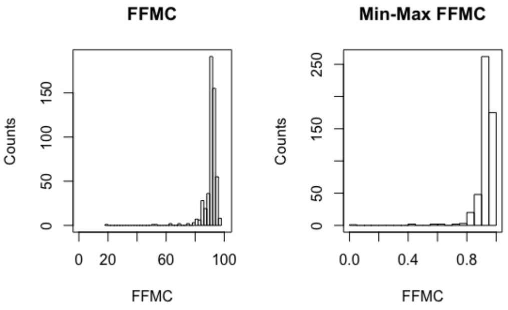

Fig. 13. Min-Max Normalization of FFMC

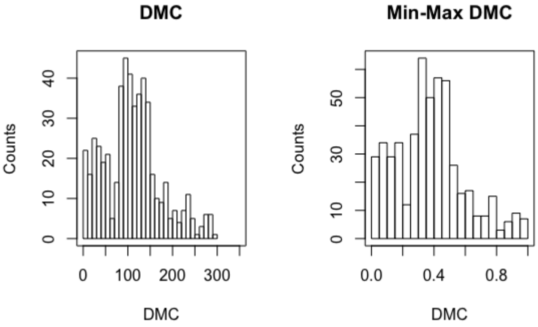

Fig. 14. Min-Max Normalization of DMC

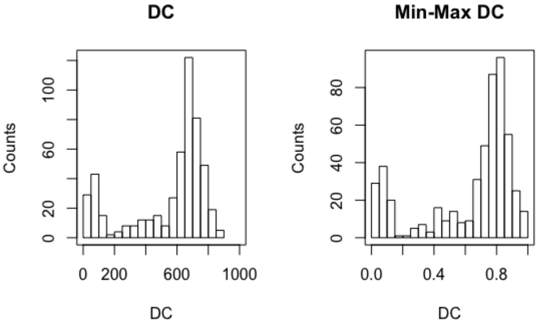

Fig. 15. Min-Max Normalization of DC

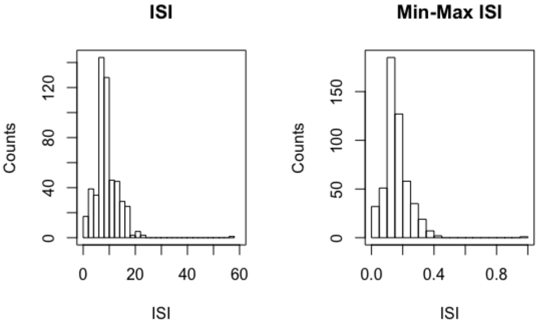

Fig. 16. Min-Max Normalization of ISI

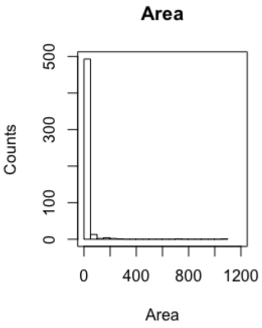

Fig. 17. Min-Max Normalization of Area

2) Z-Score Normalization

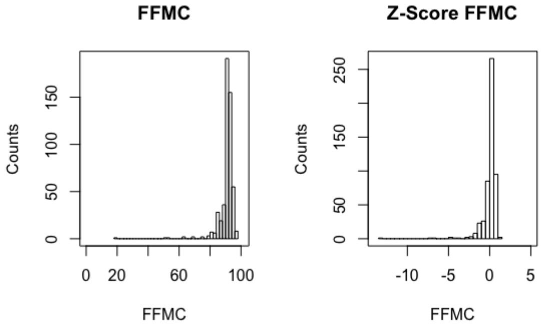

Fig. 18. Z-Score Normalization of FFMC

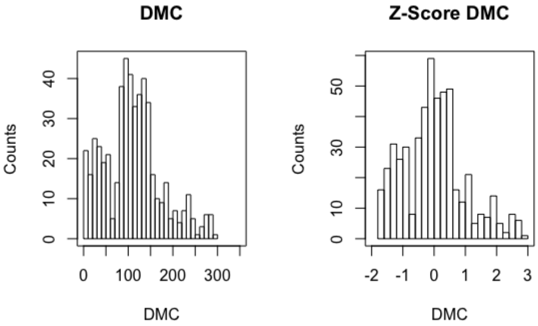

Fig. 19. Z-Score Normalization of DMC

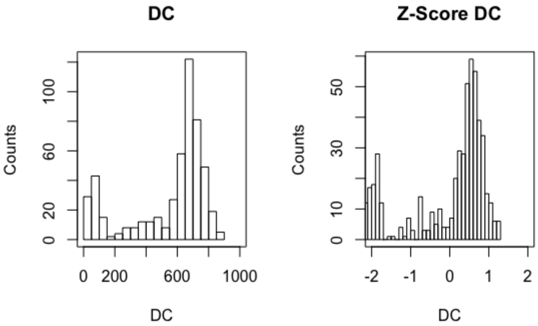

Fig. 20. Z-Score Normalization of DC

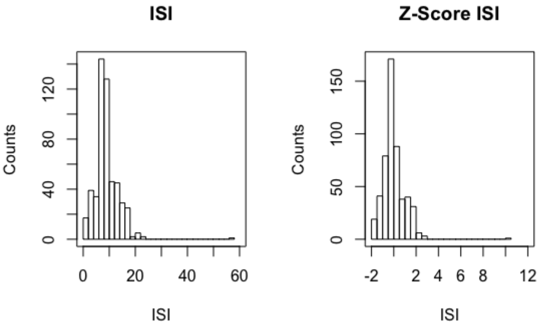

Fig. 21. Z-Score Normalization of ISI

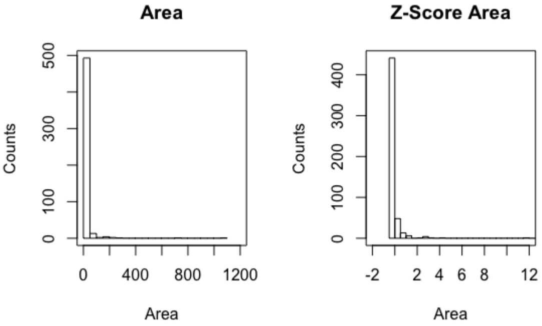

Fig. 22. Z-Score Normalization of Area

3) Decimal Scale Normalization

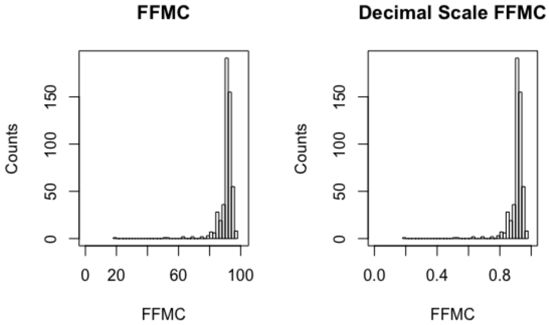

Fig. 23. Decimal Scale Normalization of FFMC

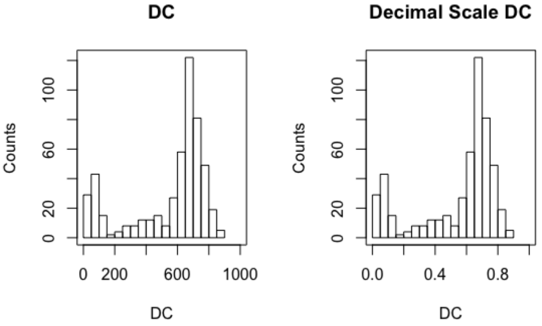

Fig. 24. Decimal Scale Normalization of DMC

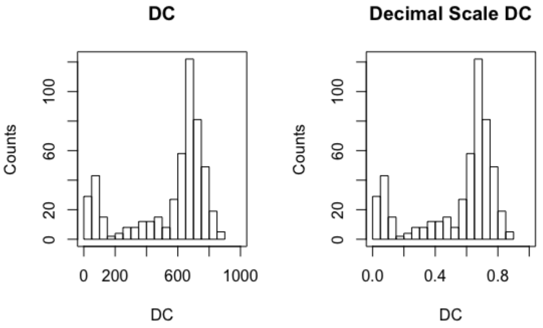

Fig. 25. Decimal Scale Normalization of DC

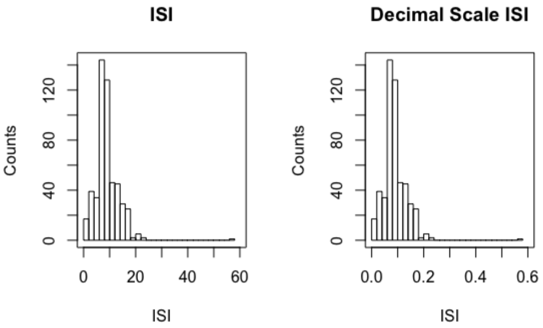

Fig. 26. Decimal Scale Normalization of ISI

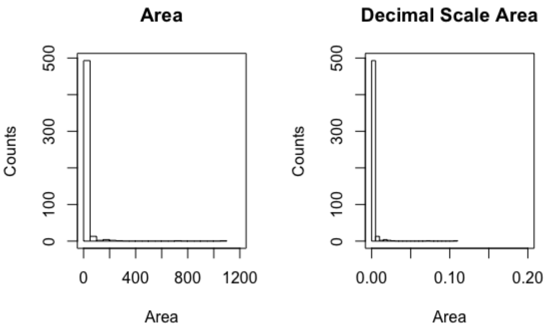

Fig. 27. Decimal Scale Normalization of Area

B. *Binning*

1) Equal Width Binning

Fig. 28 shows the DMC variable after equal width binning. The data for this variable ranged from 1.1 to 291.3 and was binned in 10 intervals of 29.13. This bin width was chosen in order to approximate the scale intervals for the DMC[6]. This binning method was unsuccesful, as the intervals on the DMC rating scale are not uniform. For example, the first bin on the histogram below encompasses both the low (0–21) and moderate (22–27) risk rating intervals on the DMC scale, and the remaining nine bins correspond to high to exteme risk of fire. Moreover, it does not add much to our understanding of the DMC variable, as it is already clear from the histrogram for DMC that the majority of the values fall towards the high end of the scale.

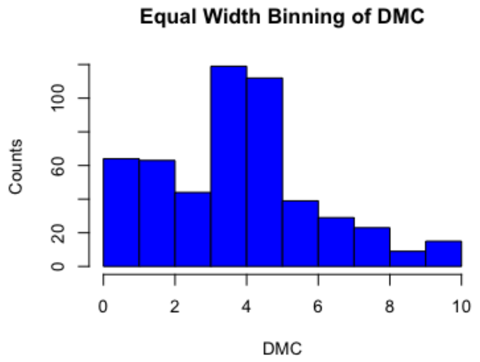

Fig. 28. Equal Width Binning of DMC Variable

2) K-Means Clustering

The figure below shows the DMC variable binned by k-means clustering. The values for this varialbe were grouped into eight bins through this clustering method.

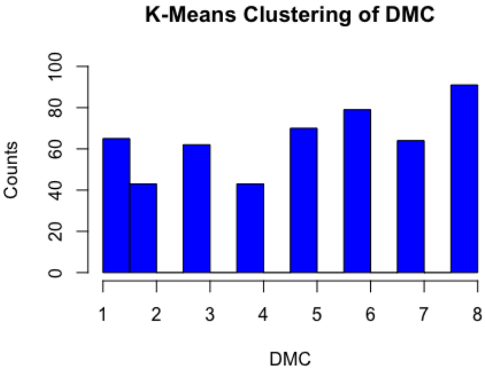

Fig. 29. K-Means Clusering of DMC Variable

C. *Transformation*

Many of the variables associated with the dataset exhibit some degree of skew. For the purposes of this assignment, I’ve chosen to work with the area attribute. As mentioned earlier, the area attribute of the dataset exhibits a strong skew towards the lower end of the scale, with the vast majority of fires recorded being of a small size.

[6] https://wildfire.alberta.ca/wildfire-status/fire-weather/understanding-fire-weather.aspx

1) Natural Log Transformation

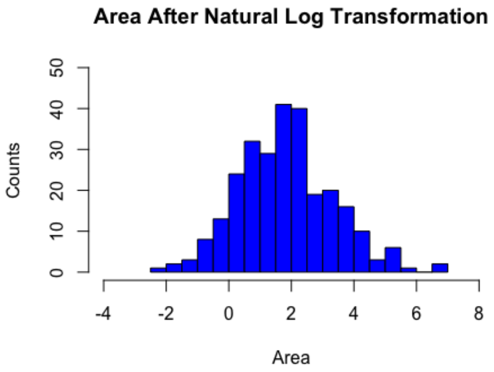

Fig. 30. Natural Log Transformation of Area

2) Inverse Square Root Transformation

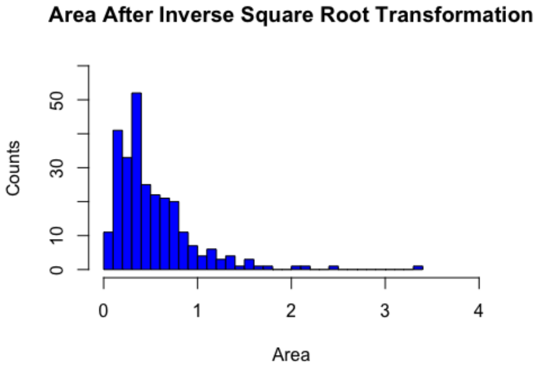

Fig. 31. Inverse Square Root Transformation of Area

3) Square Root Transformation

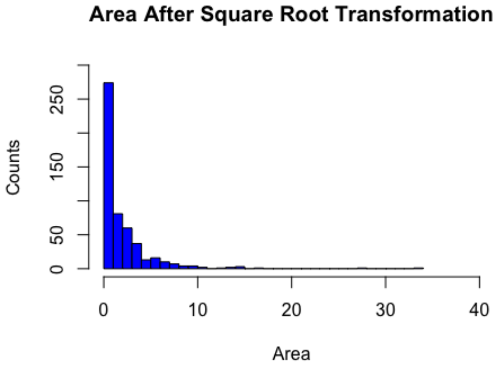

Fig. 32. Square Root Transformation of Area

As can clearly be seen, the natural log transformation in Fig. 28 is the most effective at transforming the highly skewed area variable into a more normalized data set.

V. Regression Analysis

I chose to do a regression analysis of area vs. the four FWI variables (FFMC, DMC, DC, ISI) (Figs. 33-36). I also did a regression analysis for area vs. temperature and relative humidity (Figs. 37-38). I was interested in whether the FWI ratings or direct weather data had a stronger bearing on fire size. Overall, larger fires are so few and far between in this data set that in the end, I did not find that any of the variables I chose gave a reliable estimate for fire size, which is what I expected.

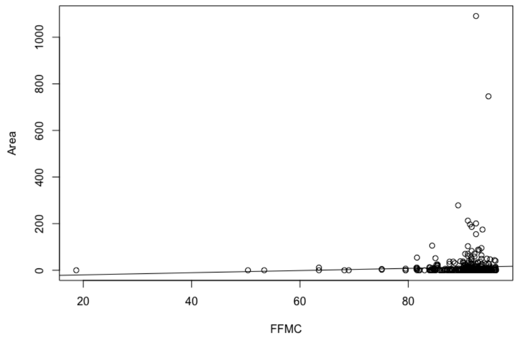

Fig. 33. Regression for FFMC vs. Area

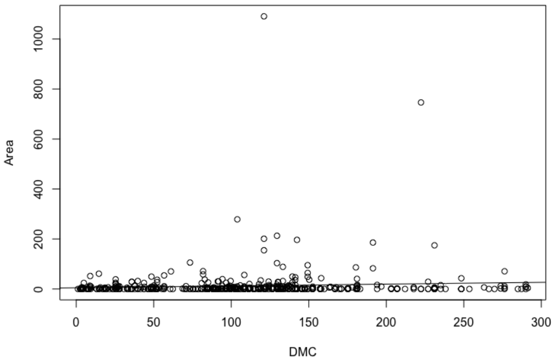

Fig. 34. Regression for DMC vs. Area

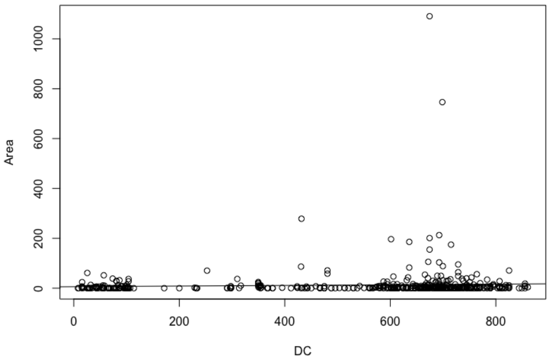

Fig. 35. Regression for DC vs. Area

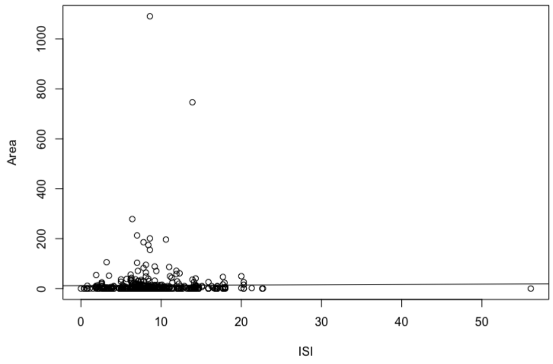

Fig. 36. Regression for ISI vs. Area

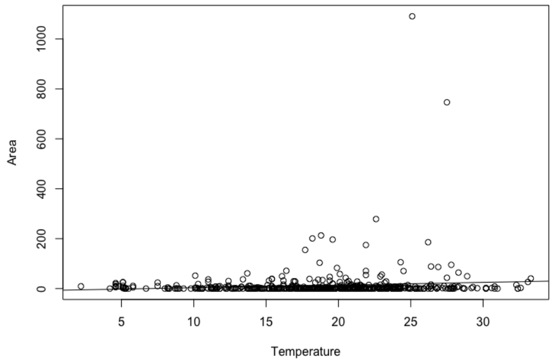

Fig. 37. Regression for Temperature vs. Area

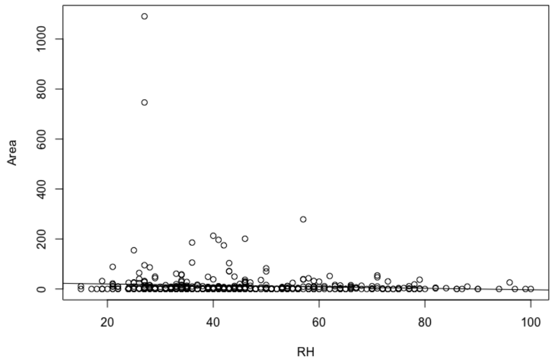

Fig. 38. Regression for Relative Humidity vs. Area

VI. CONCLUSION

Forest fires are a natural phenomenon that if not detected early can have an adverse impact on human life and the surrounding ecosystem. For this reason, the detection and prevention of fires is a pressing issue, especially in recent decades, when climate change and deforestation have led to hotter and more fire-prone conditions all over the world. The ability to use existing weather data to make reliable predictions about future fires could improve quality of life for many people and prevent environmental damage.

With my current data mining skill level, I found this to be a challenging dataset to work with. It does not seem that the dataset would be able to effectively predict large forest fires. It is possible that accurate predicitions could be made about smaller forest fires; however, as stated earler, any fire with an area smaller than $100m^2$ is recorded as a 0. The dataset raises the question of whether there are any factors in addition to the variables already included that might have an impact on fire size, such as vegetation, etc. Finally, the data was collected about 20 years ago, from 2000 to 2003. It would be interesting to examine the data from the years 2004 onward to see if there has been an increase in the frequency of larger fires in the park.

REFERENCES

[1] UCI Machine Learning Repository. (n.d.). *Forst fires data set* [Data set]. https://archive.ics.uci.edu/ml/datasets/Forest+Fires

[2] Cortez, P., & Morais, A. (2007). A data mining approach to predict forest fires using meteorological data. In J. Neves, M. F. Santos, & J. Machado (Eds.), *New trends in artificial intelligence, proceedings of the 13th EPIA 2007 - Portuguese conference on artificial intelligence* (pp. 512–523).

[3] Alberta Agriculture and Forestry. (n.d.). *Understanding fire weather*. https://wildfire.alberta.ca/wildfire-status/fire-weather/understanding-fire-weather.aspx

[4] Malaga Weather. (n.d.). *Fire weather*. https://www.malagaweather.com/fwi-txt.htm

[5] National Wildfire Coordinating Group. (n.d.). *Fire weather index (FWI) system*. https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system

[6] Global Forest Watch. (n.d.). *Fires*. https://www.globalforestwatch.org/topics/fires/#slides/3