Montesinho Forest Fires

Fernando Anorve-Lopez and Christina Sousa

March 12, 2019

Introduction

The data are from $n = 517$ forest fires occuring in the Montesinho National Park in northeast Portugal between January 2000 and December 2003. They come to us by way of Paulo Cortez and Anibal Morais at the University of Minho, in Guimaraes, Portugal. Our goal is to use 12 predictors to model the total burned area of large forest fires within the park. This could be done using a multiple linear regression model. Additionally, we would like to identify which values of the predictors yield fires with burn area less than $100m^2$, and which ones lead to larger fires. This could be done using logistic regression.

At each fire occurrence, a fire inspector recorded the date and and time of the fire, as well as the spatial location of the fire within the park boundaries. They also recorded the type of vegetation, weather conditions, total burn area, and six components of the Fire Weather Index (FWI), a Canadian system for rating fire danger. These included Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) and FWI. The FFMC variable pertains to the moisture content of surface litter; the DMC and DC variable pertain to the "moisture content of shallow and deep organic layers" (Cortez and Morais 2017); the ISI variable pertains to fire velocity spread; and finally the BUI variable pertains to the amount of available fuel. These various codes and scores each contribute to the final FWI score. Although the individual elements are not measured on the same scale, higher values indicate greater danger of fire in all cases.

The researchers did not include BUI and FWI in the data, because these were strongly collinear with the other predictors.

The variables in the data (Cortez and Morais 2017) are given by:

1. **X** *(nominal)*: x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. **Y** *(nominal)*: y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. **month** *(nominal)*: month of the year: "jan" to "dec"
4. **day** *(nominal)*: day of the week: "mon" to "sun"
5. **FFMC** *(ordinal, continuous)*: FFMC index from the FWI system: 18.7 to 96.20
6. **DMC** *(ordinal, continuous)*: DMC index from the FWI system: 1.1 to 291.3
7. **DC** *(ordinal, continuous)*: DC index from the FWI system: 7.9 to 860.6
8. **ISI** *(ordinal, continuous)*: ISI index from the FWI system: 0.0 to 56.10
9. **temp** *(interval, continuous)*: temperature in Celsius degrees: 2.2 to 33.30
10. **RH** *(ordinal, discrete)*: relative humidity in %: 15.0 to 100
11. **wind** *(ratio, continuous)* wind speed in km/h: 0.40 to 9.40
12. **rain** *(ratio, discrete)*: outside rain in mm/m2 : 0.0 to 6.4
13. **area** *(ratio, continuous)*: the burned area of the forest (in ha): 0.00 to 1090.84

Exploratory Data Analysis

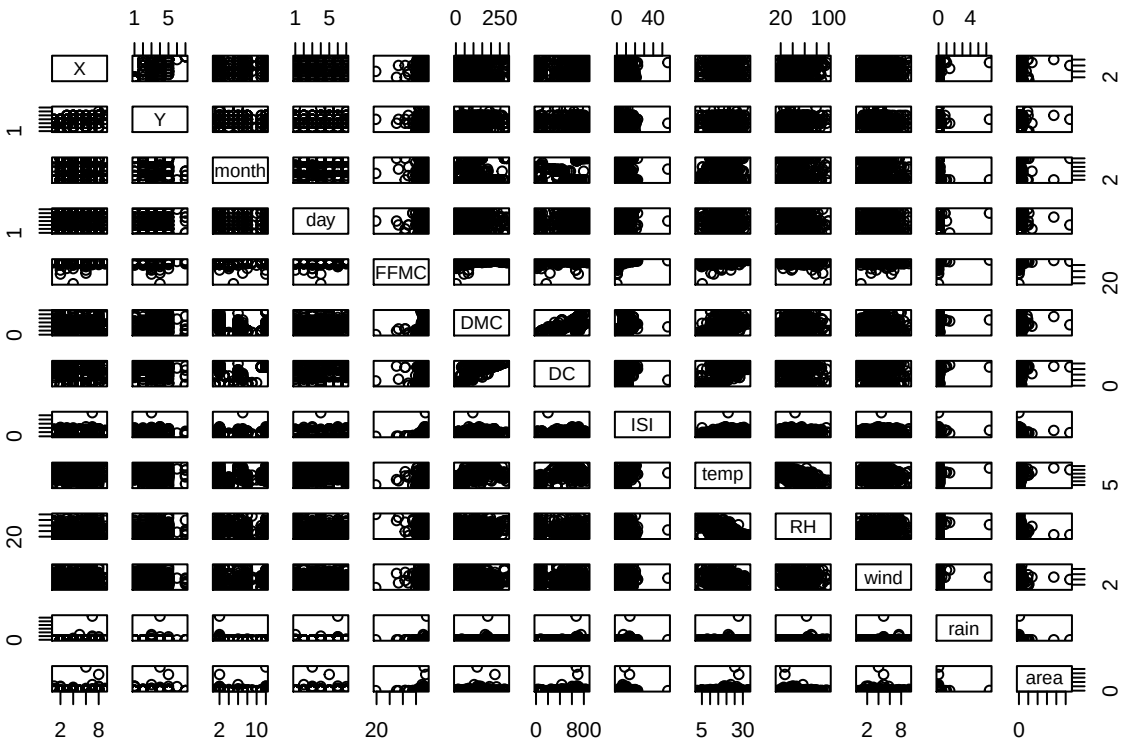
We begin the analysis by reading in the data, plotting it, and viewing relevant summary statistics.

```
#data file can be obtained at https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/
#read data into r
forest <-read.csv("forestfires.csv")
```

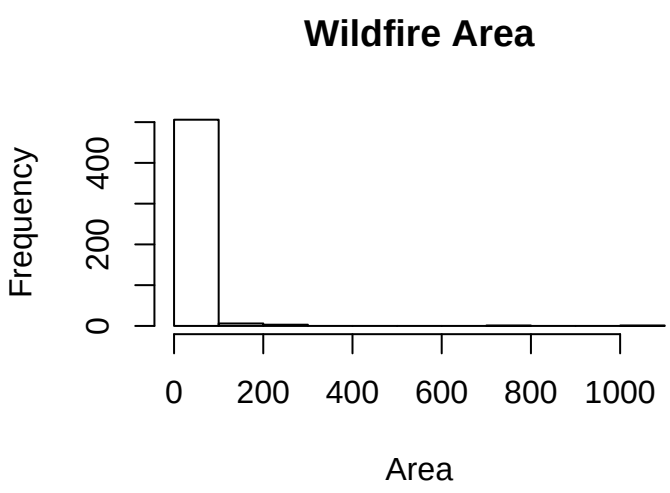
```r
#change spatial variables X and Y to factors
forest$X<-as.factor(forest$X)
forest$Y<-as.factor(forest$Y)

#change RH variable to numeric rather than integer (don't think this is needed)
#forest$RH<-as.numeric(forest$RH)

head(forest)
```

```
##   X Y month day FFMC  DMC    DC  ISI temp RH wind rain area
## 1 7 5   mar fri 86.2 26.2  94.3  5.1  8.2 51  6.7  0.0    0
## 2 7 4   oct tue 90.6 35.4 669.1  6.7 18.0 33  0.9  0.0    0
## 3 7 4   oct sat 90.6 43.7 686.9  6.7 14.6 33  1.3  0.0    0
## 4 8 6   mar fri 91.7 33.3  77.5  9.0  8.3 97  4.0  0.2    0
## 5 8 6   mar sun 89.3 51.3 102.2  9.6 11.4 99  1.8  0.0    0
## 6 8 6   aug sun 92.3 85.3 488.0 14.7 22.2 29  5.4  0.0    0
```

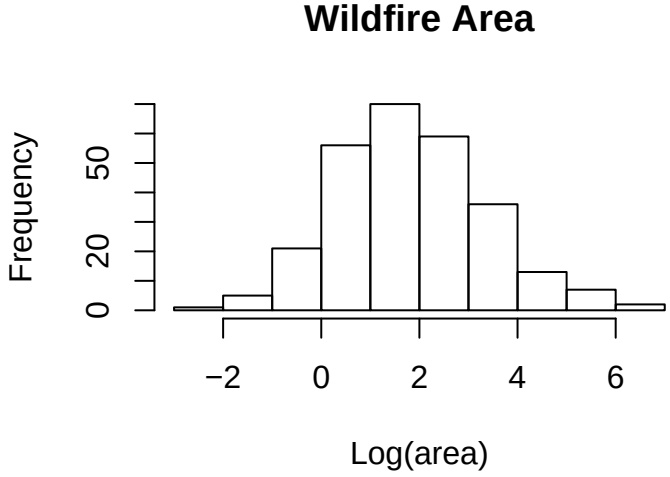
```r
pairs(forest)
```



Out of 517 records, there are 247 records with an `area` value of 0. The `area` data is considerably right-skewed. This is explained by Cortez and Morais as a general trend with forest fire data observed in various locations: either the fires are extinguished quickly and burn less than $100m^2$. of area, or they are not easily contained and end up burning a large area. Fires that burn less than $100m^2$ of area are all catalogued as zero. This poses a missing data problem and prevents us from doing a complete analysis on these types of fires. To face this problem we can apply the MLR model to the data conditioned to nonzero burn area only. We remove these data for now, however, we will make use of them later to compare small fires and large fires using logistic regression.



The values of `area` also range over several orders of magnitude. The positive skewness suggests the need of a transformation. Trying log transformation seems suitable. We log transform `area` of the nonzero-valued observations and note that the histogram looks more symmetric now.

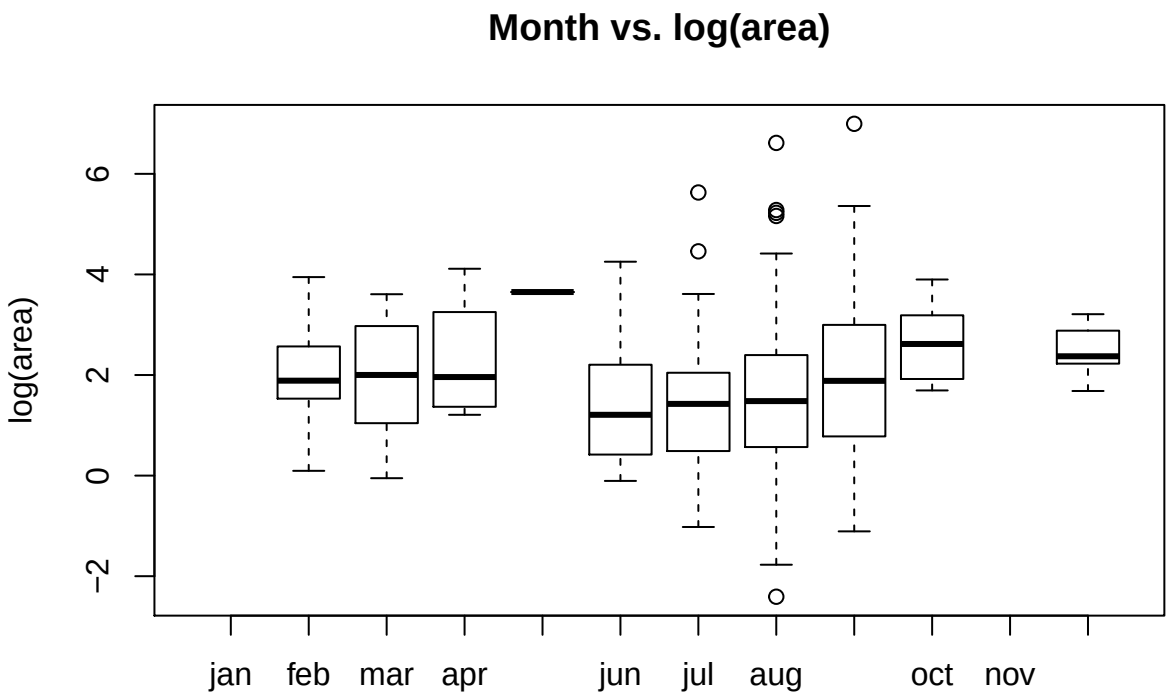


We now turn out attention to the other relevant summary statistics of the data. We notice that certain days and months appear to have more fires. Also, the distributions of `FFMC`, `DMC`, `DC`, and `ISI` appear to be somewhat skewed. Finally, `rain` does not seem to offer much information for forest fires with large areas.

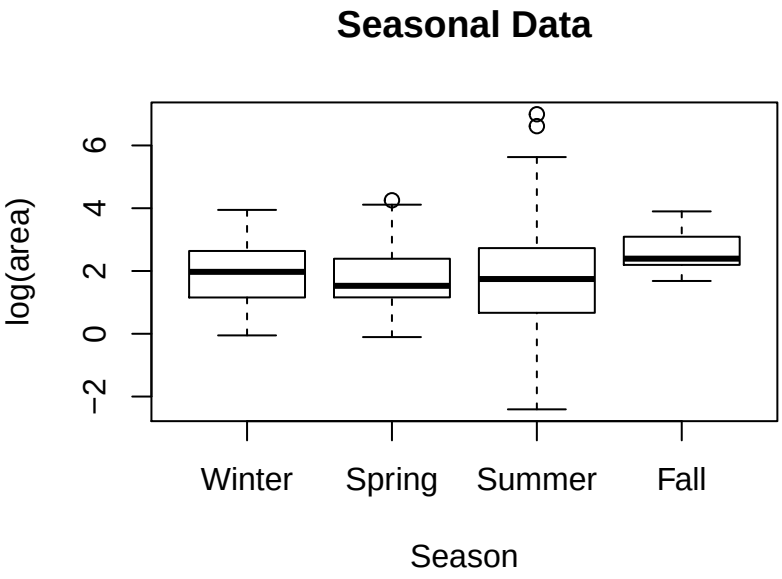
```
##        X      Y           month     day          FFMC
##  6      :48   2: 15   aug    :99   fri:43   Min.   :63.50
##  4      :47   3: 34   sep    :97   mon:39   1st Qu.:90.33
##  2      :42   4:111   mar    :19   sat:42   Median :91.70
##  8      :37   5: 68   jul    :18   sun:47   Mean   :91.03
##  7      :30   6: 38   feb    :10   thu:31   3rd Qu.:92.97
##  1      :25   8:  1   dec    : 9   tue:36   Max.   :96.20
##  (Other):41   9:  3   (Other):18   wed:32
##       DMC              DC             ISI              temp
```

```
##  Min.   :  3.2   Min.   : 15.3   Min.   : 0.800   Min.   : 2.20
##  1st Qu.: 82.9   1st Qu.:486.5   1st Qu.: 6.800   1st Qu.:16.12
##  Median :111.7   Median :665.6   Median : 8.400   Median :20.10
##  Mean   :114.7   Mean   :570.9   Mean   : 9.177   Mean   :19.31
##  3rd Qu.:141.3   3rd Qu.:721.3   3rd Qu.:11.375   3rd Qu.:23.40
##  Max.   :291.3   Max.   :860.6   Max.   :22.700   Max.   :33.30
##
##        RH              wind            rain             Larea
##  Min.   :15.00   Min.   :0.400   Min.   :0.00000   Min.   :-2.4079
##  1st Qu.:33.00   1st Qu.:2.700   1st Qu.:0.00000   1st Qu.: 0.7608
##  Median :41.00   Median :4.000   Median :0.00000   Median : 1.8516
##  Mean   :43.73   Mean   :4.113   Mean   :0.02889   Mean   : 1.8448
##  3rd Qu.:53.00   3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.: 2.7358
##  Max.   :96.00   Max.   :9.400   Max.   :6.40000   Max.   : 6.9947
##
```

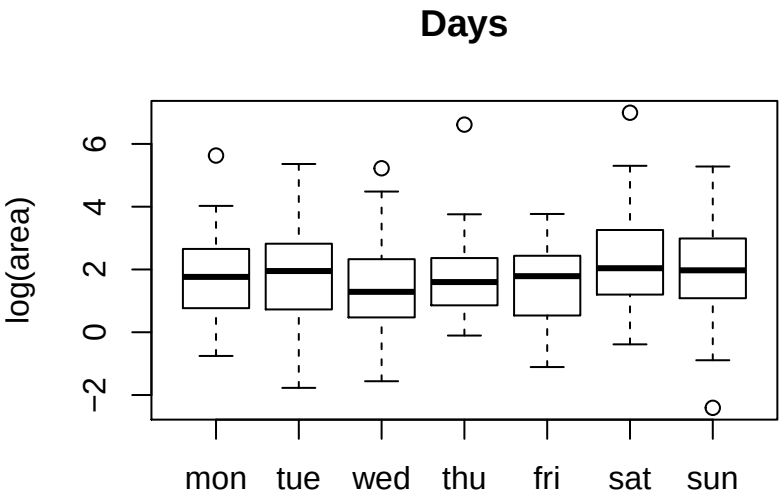
As categorical data, the `month` variable can be divided into subsets defined by seasons.

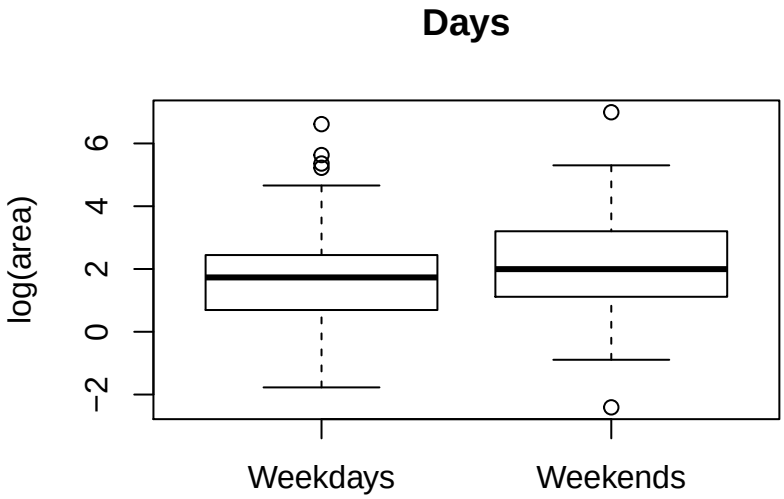


We can also redefine Jan, Feb, Mar as Winter; Apr, May, Jun as Spring; Jul, Aug, Sep as Summer, and Oct, Nov, Dec, as Fall, because seasonal weather conditions can have an effect on incidence of wildfires. In this case, Summer records show more variability than any other season.

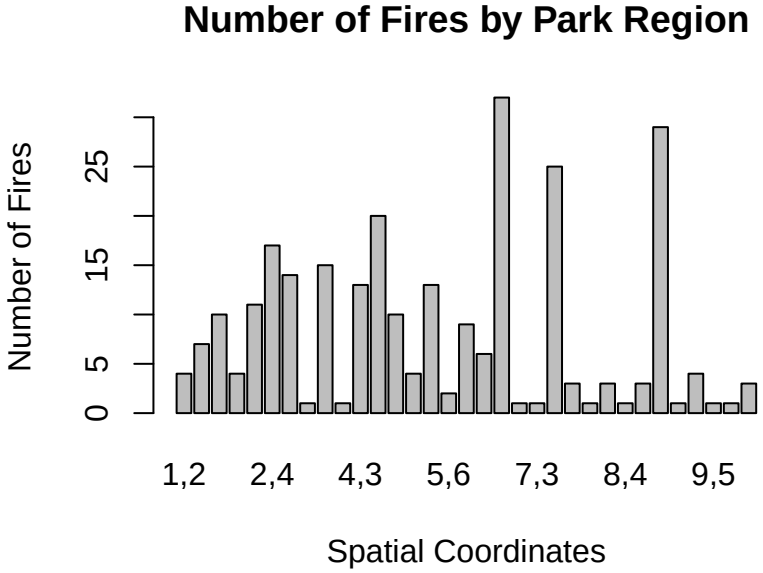


The `day` variable can be divided into subsets "weekdays" and "weekends", since forest fires could be explained by exposure to human beings who visit the park on the weekends.

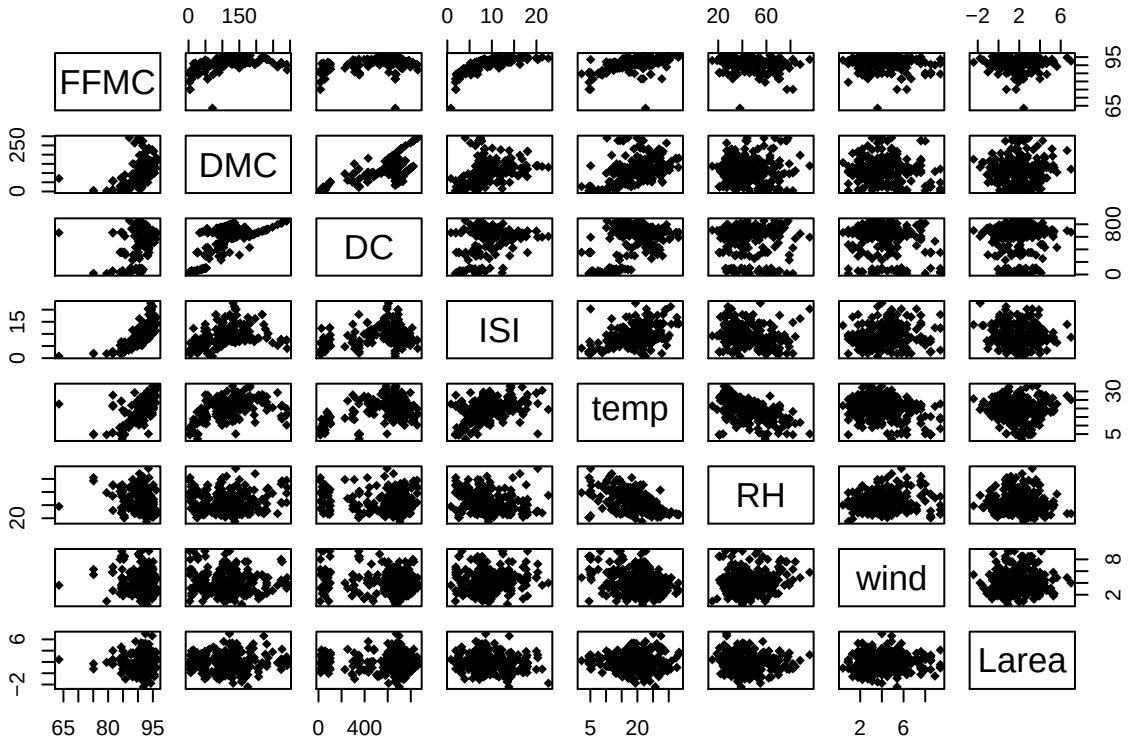




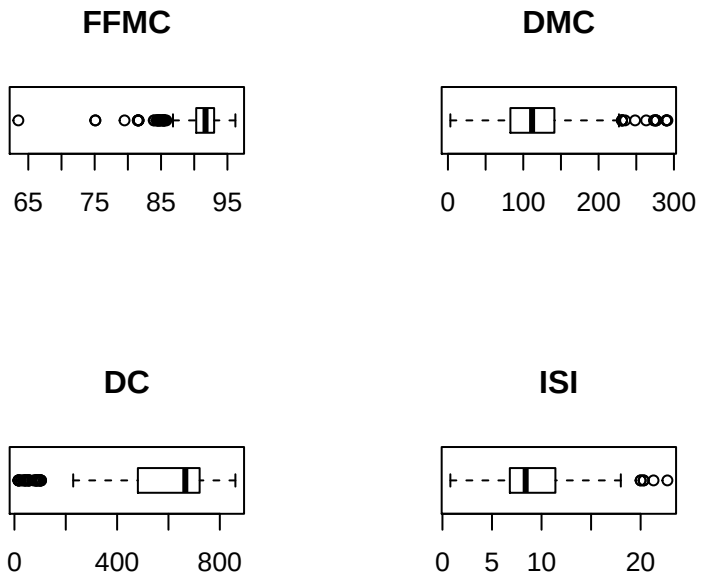
With respect to the spacial coordinates, it seems that the `X` and `Y` coordinates would be better explained by assigning each grid of the map its own category, to see if one area of the map appears more than others. Indeed, there are 4 regions (out of 81) of the map that had 20 or more fires in this 3-year period. It may be interesting to compare the sizes of fires in these areas, and whether these fires are bigger than fires in other areas of the park.



We also note that, after removing the forest fires with zero area, there are only 2 records with nonzero values for `rain`. Hence it does not make sense to include this predictor in our current analysis, and we remove it. We may take another look at this later in the logistic regression setting. We now view the continuous predictors after removal of this variable:



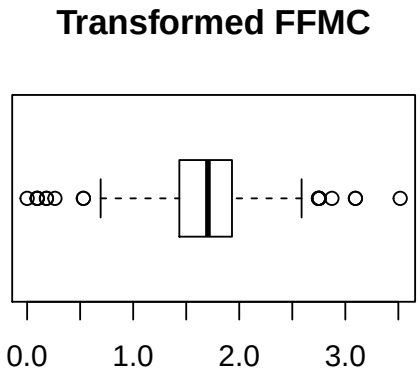
For the fire index variables, we note that the quantiles appear to be a bit skewed, particularly for `FFMC` and `DMC`, so we suspect that transformation of these variables may be in order. The boxplots of the fire index variables appear below.

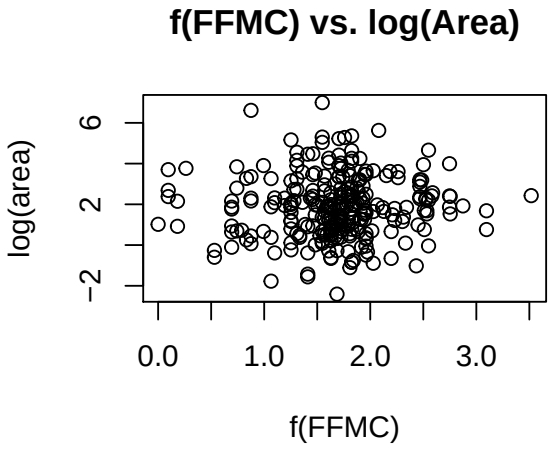


Noticing that `FFMC` is negatively skewed, we can try to normalize it by reflecting it, adding an appropriate value to make it positive, and finally applying a log transformation.

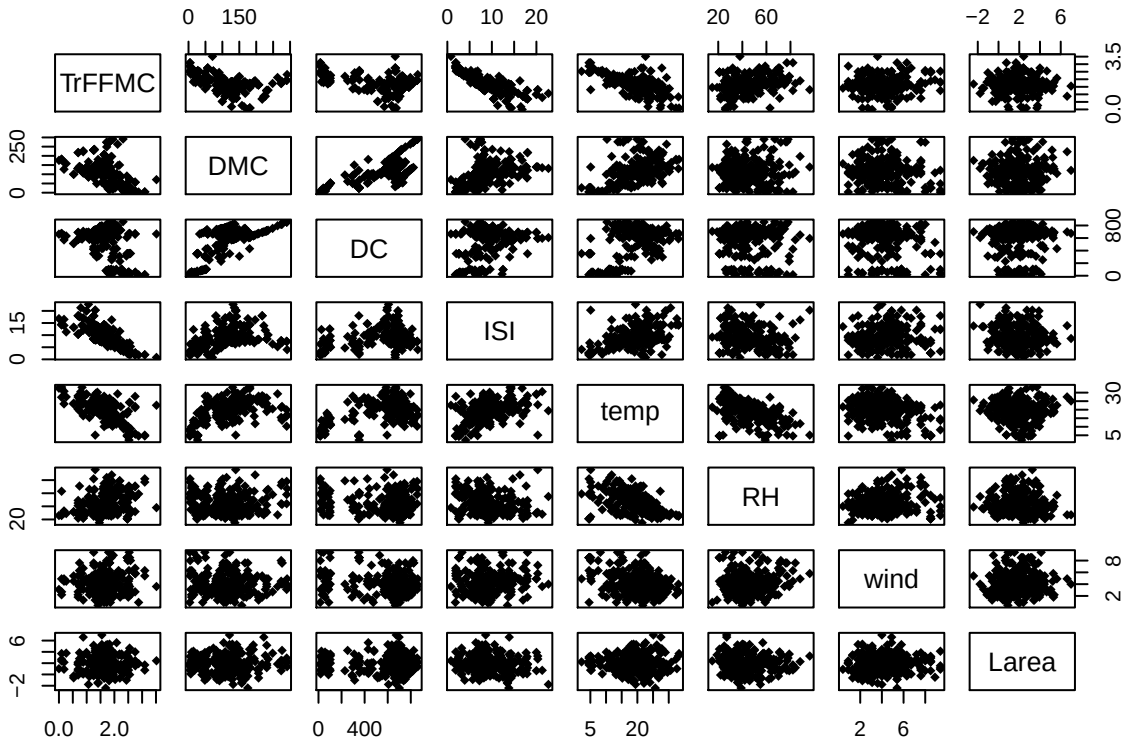
$$f(\text{FFMC}) = \log(-\text{FFMC} + \max(\text{FFMC}) + 1)$$

This seems to correct the skew.





The final cleaned continuous variables appear below.



Finally, we comment on collinearity in this data. It appears that DMC and DC are linearly correlated. There may also be correlation between temp and ISI and RH and temp. In fact, temp appears to be linearly correlated with just about all the other predictors. There is also an interesting (and perhaps troubling!) curvilinear relationship between DMC and ISI as well as DMC and temp. To summarize, there appears quite a bit of multicollinearity in this data. This may be because the fire indexes are based on the various weather variables, and because weather variables such as temperature and humidity are related to one another.

Summary

At this point there do not appear to be any continuous variables that look like strong predictors for Larea. Perhaps the discrete variables will become more important as we attempt to model this response. We also think that a logistic regression approach might be more productive in determining which variables are significant to predict whether the burn area of a wildfire will be larger or smaller than $100m^2$.

We could also attempt to interpolate the missing data for the zero-area entries by randomly sampling areas between 0 and $100m^2$ according to a distribution consistent with the rest of the data, to see if this improves our results.

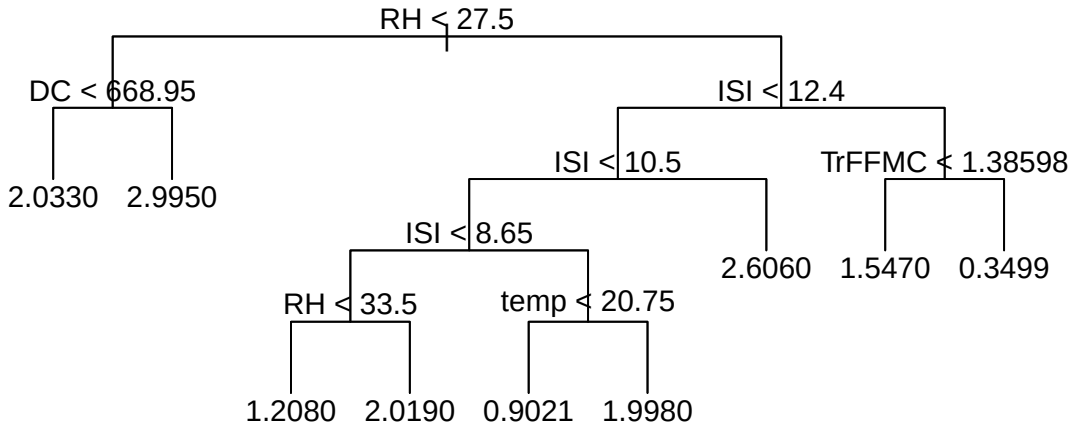
To model Larea, we also might proceed by seeking out suitable transformations of these variables by applying the Yeo-Johnson method. The car package function powerTransform suggests that the cleaned data could be normalized with a few reasonable transformations.

```
require(car)
summary(powerTransform(cbind(TrFFMC,DMC,DC,ISI,temp,RH,wind)~1,forest5,family="yjPower"))
```

```
## yjPower Transformations to Multinormality
##        Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## TrFFMC    2.0905        2.00       1.8008       2.3802
## DMC       0.3221        0.33       0.2270       0.4171
## DC        1.3926        1.39       1.1822       1.6029
## ISI       0.0516        0.00      -0.1214       0.2247
## temp      1.1356        1.00       0.8984       1.3729
## RH        0.1952        0.00      -0.1064       0.4968
## wind      0.3811        0.50       0.1109       0.6514
##
##  Likelihood ratio test that all transformation parameters are equal to 0
##                                          LRT df       pval
## LR test, lambda = (0 0 0 0 0 0 0) 643.3795  7 < 2.22e-16
```

Also, a regression tree suggests that RH, DC, ISI, and the transformed FFMC variables may still be significant in predicting Larea. These may be worth adding to the MLR model one-at-a-time to see if a desirable level of R^2 can be achieved (see figure on next page).

```
require(tree)
data.tree <- tree(Larea~., data=forest6[,6:13],mincut=15, minsize=35)
plot(data.tree, type="uniform")
text(data.tree)
```



Citations

Cortez, Paulo, and Anibal Morais. 2017. "A Data Mining Approach to Predict Forest Fires Using Meteorological Data." Edited by J Neves, M F Santos, and J Machado. *New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, December. Guimaraes, Portugal, 512–23. http://www.dsi.uminho.pt/~pcortez/fires.pdf.