Data 621 - Final Project

Devin Teran, Atina Karim, Tom Hill, Amit Kapoor

5/23/2021

Contents

Abstract

For the final assignment, the team will be looking at a dataset of forest fires identified in Montesinho Natural Park, located in the mountainous northeast of Portugal. The original paper was published by a nearby

university with the intention of offering low-cost prediction based on available meteorological measurements. Their model used information collected in realtime to identify fires without more specialized equipment - sattelite imaging and smoke scanners.

The objective is to build a model that predicts the burned area of the forest given the explanatory variables. Additionally, a temporal model could be developed to understand which day and weather patterns offer the highest risk conditions.

Keywords

Regression modeling, forest fires, meteorological measurements

Introduction

Every year forest fires cause enormous amounts of destruction that results in the loss of human and animal lives and economical and ecological loss. In addition, unpredictable forest fires make it very difficult for fire departments to plan in order to have an appropriate amount of resources in order to effectively fight all the fires. By creating a model that uses meteorological measurements such as wind and relative humidity, fire departments would be able to prepare better in order to fight fires more effectively and ideally decrease the amount of destruction that is caused each year by forest fires.

Literature Review

A Data Mining Approach to Predict Forest Fires using Meterological Data by Paulo Cortez and Anibal Morais focused on using meterological data to predict smaller forest fires. This study focused on meterological data because it can easily be captured through weather stations in real-time for low costs. Other methods of infrared scanning and satellite are more expensive. Their work used a SVM, neural network, multiple linear regression and other data mining methods in order to predict the weather conditions that created small fires. While small fires are much more common than large fires, a drawback to this study is it doesn't predict when large fires will occur. The outcome of this study helps fire department manage resources by reaching fires more quickly.

The **Wildfire Burn Area Prediction paper by Adam Standford-Moore and Ben Moore**] aimed to predicate the burn area of wildfires. They used Kaggle data of historical wildfires in the United States from 1992-2015 and UCI dataset of wildfires in Portugal including info on first start and end date, longitude, latitude, year, and fire cause. The UCI dataset included weather features, which proved to be important in the modeling. Unlike other groups, this paper took all of the historical fires and binned them into 10 groups according to the fire size. They then sampled 4,000 fires from each bin in order to complete their modeling. Some of the bins with smaller fire sizes were heavily overrepresented in the data and they did this in an effort to balance the data. This group used various modeling approaches including SVM, Neural Networks, K-Nearest Neighbors, Decision Trees, Linear Regression and more. The best performance came from the SVM. The smaller fire sizes proved easier to predict, which alines with some of the other studies. This study found the features related to weather were more predictive over the historical fire data.

The article, **Predicting forest fires burned area and rate of spread from pre-fire multispectral satellite measurements by Carmine Maffeia and Massimo Menenti**] had an interesting take away that their focus was almost solely on moisture in the plant life on the ground. They focused on weather conditions in an attempt to determine how it would affect the moisture levels in plant life since plants will fuel fire. Plants with less moisture with burn quickly. They were attempting to come up with their own function to measure the moisture content by focusing on plant moisture at different levels on the ground (surface vs deeper levels of organic matter). They used various satellite imaging resources which was different from other studies in that this data collection is more expensive. It'd be interesting to see how this studies findings impacts towns and their ability to obtain data through more expensive collection routes. Our analysis will

not include these expensive data sources and instead will focus on meterological measurements as inexpensive data sources.

Methodology

Data Overview

Below is the description and abbreviations of the variables of interest in the data set. These abbreviations will be used throughout our paper.

VARIABLE NAME	DEFINITION
X	x-axis spatial coordinate within the Montesinho park map: 1 to 9
Y	y-axis spatial coordinate within the Montesinho park map: 2 to 9
month	month of the year: "jan" to "dec"
day	day of the week: "mon" to "sun"
FFMC*	Fine Fuel Moisture Code index from the FWI system: 18.7 to 96.20
DMC*	Duff Moisture Code index from the FWI system: 1.1 to 291.3
DC*	Drought Code index from the FWI system: 7.9 to 860.6
ISI*	Initial Spread Index from the FWI system: 0.0 to 56.10
temp	temperature in Celsius degrees: 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	outside rain in mm/m2 : 0.0 to 6.4
area	the burned area of the forest (in ha): 0.00 to 1090.84

*The FWI, or Fire Weather Index system, is an estimator developed in Canada for assessing fire risk. It ranges 0-20 and considers weather and fire conditions withing a critical period prior at the start of the fire. The FFMC is a composite of rain, humidity, temperature and wind, DMC is rain, humidity and temperature, DC is rain and temperature, and finally ISI is a fire behavior index. One element of FWI, the buildup index, was not included as this is an indicator of the bulk fuel on the ground and not ascertainable from meteorology.

Source : Cortez, Paulo & Morais, A.. (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data.

Data Overview & Exploration

Our dataset has 517 observations and 13 columns. None of the variables have any missing values. All of the data are numeric except for day and month. It is also interesting to note that our response variable, area, has a large range and the difference between the median (0.52) and mean (~12) seems to suggest there maybe outliers in our data.

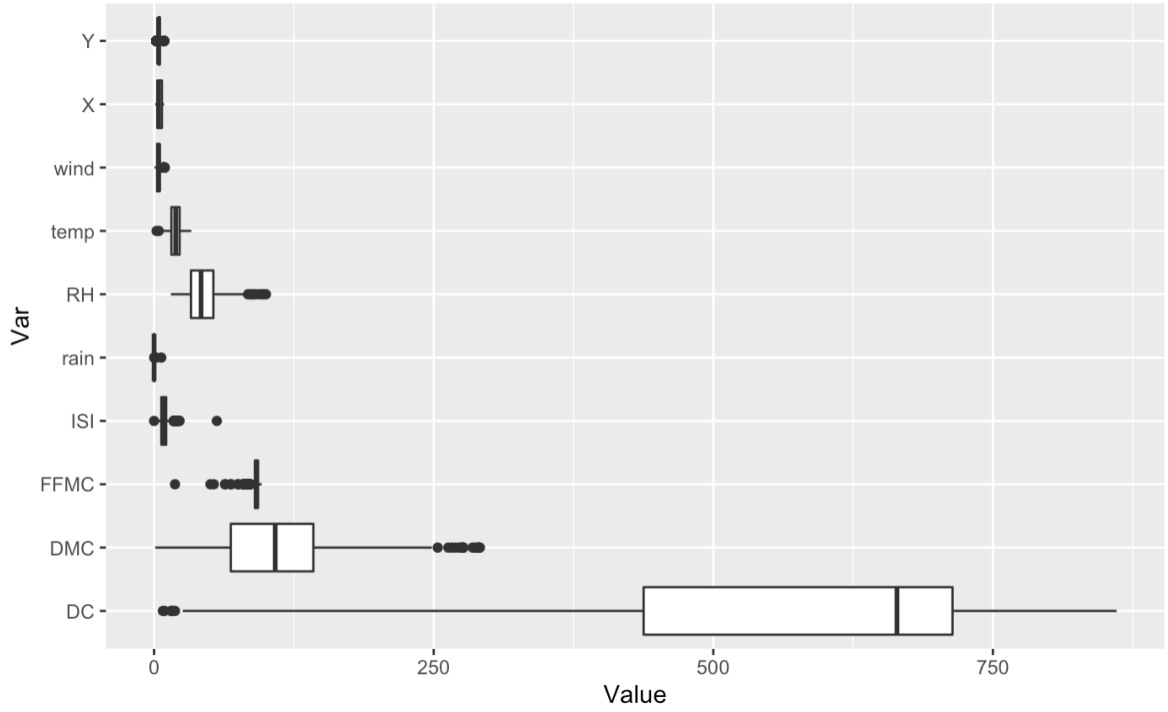
There are 247 observations were our target variable, area, is 0. This is not good for the assumptions of the linear model but we have little choice but to proceed. We assume that since the area is in hectare, 0 hectare means that the burned area is less than 1 hectare (107639 sqft.).

The majority of fires that occurred happened during the months of August and September. Weekend days (Friday, Saturday and Sunday) were the most common days for fires to occur. Looking at our variable rain, we were also able to see that this area in Portugal did not get much rainfall.

Using `corrplot`, it is apparent that variables `X` and `Y` are positively correlated, which is perhaps spurious or owing to the lack of distinct values for these variables. `temp` is positively correlated with several of the indices. `temp` and `RH` are negatively correlated. It seems like our dataset has multicollinearity between the independent variables.

In order to look further into the relationships between our response variable, area, and each of the predictor variables (See appendix for graphs) we imposed a linear fit (with 95% confidence band in grey) to each of the pairs of variables to understand the relationships. We see some outliers in the data. Please note, that there are instances where the response variable seems to fall below 0 - such as for area and FFMC. Once again, this may mean that the burned area was below 1 ha.

Using boxplots, we were also able to determine that the variables relative humidity(RH), Fine Fuel Moisture Code(FFMC), Duff Moisture Code(DMC) and Drought Code(DC) appear to have outliers in the data.



Handing Missing Values

At first glance, there are no missing values in this dataset. However, only days with information about fires are added to the original dataset.The original paper mentions that readings spanned between January 2000 to December 2003, which would be a span of 1460 days as opposed to the 517 observations in the dataset. Since forest fires tend to be seasonal during a dry season, there may be a relationship to the data available for each month.

The most observations are available for the month of August by a large margin. March, July and February are the next most common months. January, May, and November have 2, 2, and 1 fires reported.

Looking at the August number, it's evident that multiple fires were happening per day during this time period. This creates an interesting twist as multiple fires could strain firefighting resources and offers the prospect of considering a logistical or count as response variable if weather data were available for the missing dates. It's also possible that the extra August observations there are duplicate entries as the observations were collected from two sources throughout the study.

Data Preparation

Handling missing and outliers The very first in data preparation we will perform is handling missing data and outliers through imputation. We will use mice package to perform imputation here. MICE (Multivariate Imputation via Chained Equations) is one of the commonly used package for this activity. It creates multiple imputations for multivariate missing data. Also we will perform `nearZeroVar` to see if a

variable has very little change or variation and not useful for prediction. The variable **rain** was removed during process as it was seen to not be useful for prediction.

The variable month and day are categorical variables, having 12 and 7 classes. For modeling, we got to convert into set of dummy variables. We will use `dummyVars` function for this purpose that creates a full set of dummy variables.

Data Tranformations We use the package caret `preprocess` method using transformation as `YeoJohnson` which applies Yeo-Johnson transformation, like a BoxCox, but values can be negative as well. This provides optimal data transformations for modeling.

Testing & Training Separation Finally in this step for data preparation we partition the training dataset for training and test using `createDataPartition` method from `caret` package. We reserve 75% for training and rest 25% for validation purpose.

Experimentation and Results

We tried various models including linear regression, robut regression, ridge regression and using various techniques. Here is an overview of each model:

Linear Modeling

We have transformed our data to fulfill the assumptions of linear regression. We will first test out linear regression on our data.
Model 1: Using all of the variables this model returned an R^2 value of approximately 10% meaning the model account for 10% of the variance in the data.
Model 2: Since we had some outliers of large forest fires we ran a model on only fires where the area burned is under 2 hectares.
Model 3: Next we remove fires where burn area is 0 hectares. This results in an R^2 of 19%. While we have a large number of records where area=0, this subset would be interesting to examine if the fire department is looking to predict large forest fires that would need substantial resources to deal with.
Model 4: Finally we used the *leaps()* package, which performs an exhaustive search for determing the best predictors to use to predict our target variable. According to leaps, **best model** includes the following predictors - Thursday, the months of August, December, January, July, June, March, October, September, spatial coordinates X & Y, relative humidity, temperature and duff moisture code. However, the adjusted R^2 for this model is still low at 5.7%, and is significant. Residual SE is also low at 0.6344.

So far, our actual best model is when we leave out the records with burned area less than 0. We will proceed with this model. But before we entirely dismiss the second best model, produced by leaps, we will try to see if there are any other methods we could try using the predictors suggested by leaps to get a better model fit.

Robust Regression

First Robust Regression

Our dataset also had outliers - which we handled through imputation during data preparation. However, we also built a robust regression model to fully ensure that outliers and influential observations are not affecting our model. Robust regression is an iterative procedure that seeks to identify outliers and minimize their impact on the coefficient estimates.

Upon comparing the RSE from our original linear model (linear regression model #4 - best.model) and the robust regression model, keeping all variables constant, it seems that our original linear model (linear regression model #4 - best.model) performance was better. From here on, we will dismiss the best model produced by leaps and move forward with the subset where burned area is greater than 0 ha.

Large Fires Model

Upon revising the coefficients of our model (linear regression model #4 - best.model), it seems like the only

signficant predictor for large fires at p <.05 level is DMC - a moisture code index that takes rates moisture content that takes into account rain, humidity and temperature.

We tried to refine this model a little bit with the leaps package to see if we can get a better fit just for this subset. The leaps package suggested using 12 predictors to get the best model. This resulted in our adjusted R^2 decreasing slightly, however we are seeing several significant predictors. We had flagged the dataset for multicollinearity previously and upon running the car() package to test for multicollinearity in our chosen regression mode - it seemed like there were aliased coefficients in the model suggesting perfect collinearity between certain IVs. Therefore, despite the lower adjusted R^2, we are choosing to move forward with this model as our best to avoid overfitting.

Regularized Model

Model 1:
One thing to note from our linear model, is the difference between the adjusted R^2 and R^2 - this indicates that there is still multicollinearity in our data (which we has observed with some of the IVs during EDA). This may result in overfitting i.e. overestimating some of the coefficients assigned to our predictors. Such a model, will not perform well in the unseen test data. Therefore, we perform regularization through ridge regression to overcome this issue.

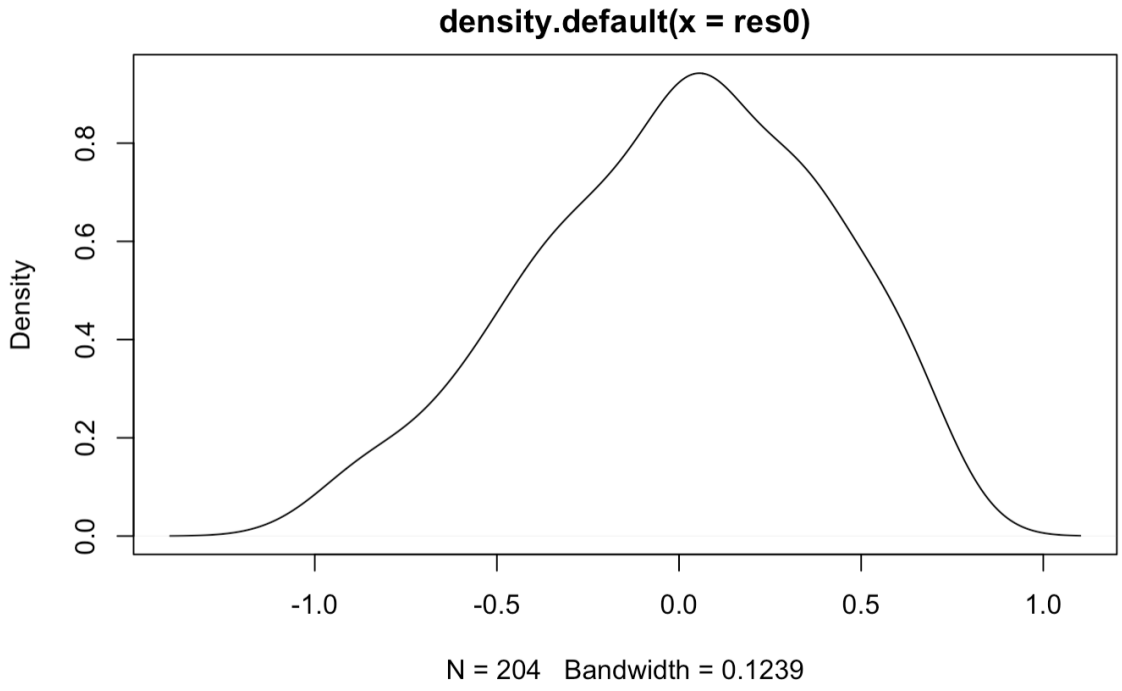
Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model (Singh,2019). We used the **glmnet()** package to build the regularized model.

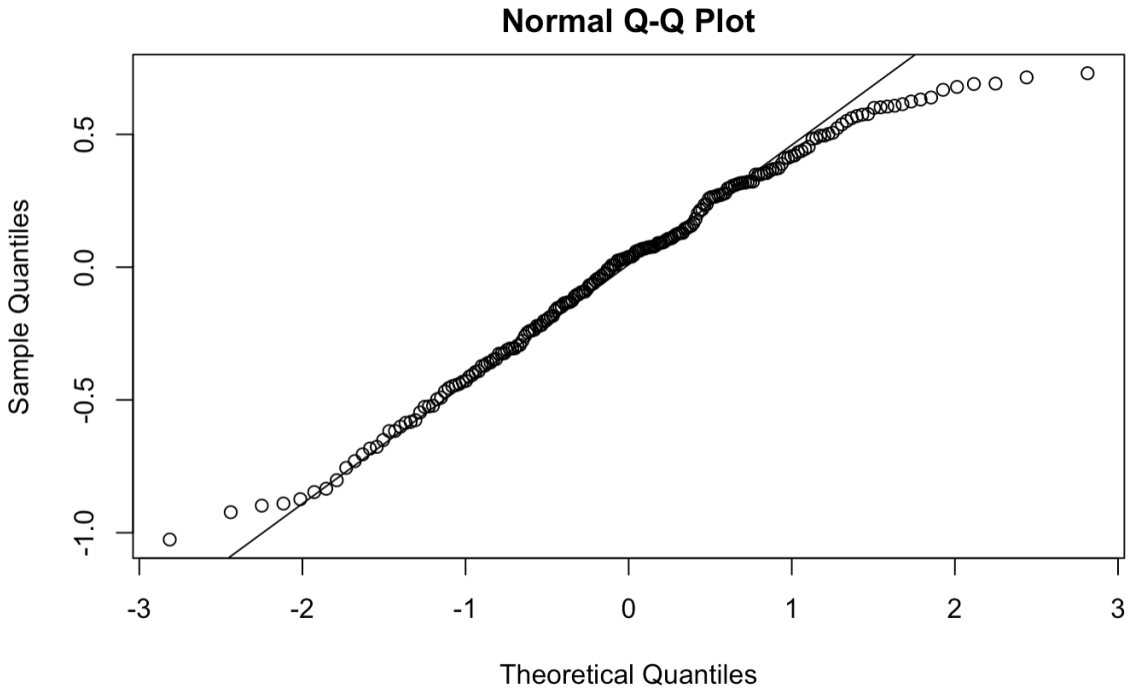
One of the major differences between linear and regularized regression models is that the latter involves tuning a hyperparameter, lambda. for our model, we will see what the best lambda is.

We can see that our R^2 for the regularized model has decreased which suggests that the ridge model has not improved the original model, rather deteriorated its fit.

Select Model

We have chosen the **Large Fires Model** to be our best model. Before finalizing this choice, we will analyze the diagnostic plots for our chosen model. The diagnostic plots suggest that the residuals are normally fitted. Therefore, this is our final model.





Predicting on the Test Data

We will use out final model to predict the burned area from forest fires on our fire test, and view the results for the top rows to see if the numbers make sense:

```
       4        11        12        13        19        20
1.213208 1.153766 1.143095 0.991248 1.226917 1.087829
```

Discussion and Conclusion

The purpose of this study is to identify forest fires in Montesinho Natural Park and predict the area (in ha) that is likely to be affected, given weather conditions. Because our dataset categorized all small forest fires in this region(anything less than 1 ha) as 0, we decided to limit our study to only evaluate the area for large fires (area above or close to 1 ha).

Our model indicates that the most significant predictor of larget forest fires (in terms of forest fires) is DMC. The Duff Moisture Code Index,is a numeric rating of the average moisture content of loosely compacted organic layers of moderate depth. In addition, it also seems from our model that large forest fires may also be likely to occur during the months of December and October in this region. This was surprising initially given that we were seeing such large numbers of forest fires in August during EDA. However, our final models seems to predict a significant but a weak negative relationship between large forest fires for August. This maybe because that there are higher numbers of small forest fires in the month of August rather than large ones. This is definitely an interesting insight that could be investigated in a future study.

We hope this study could be used to improve disaster management and prevent significant damage from forest fires in Montesinho Natural Park in Portugal.

```
Call:
lm(formula = area ~ day.thu + month.aug + month.dec + month.feb +
    month.jul + month.jun + month.mar + month.oct + X + DMC +
    temp + RH, data = fires_over0)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02563 -0.29392  0.03845  0.31345  0.73017

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.1975994  0.6192037   1.934   0.0546 .
day.thu     -0.0128295  0.0968186  -0.133   0.8947
month.aug   -0.1281100  0.0764066  -1.677   0.0952 .
month.dec    0.3105683  0.2030811   1.529   0.1278
month.feb    0.1501138  0.1903242   0.789   0.4312
month.jul   -0.0212590  0.1307957  -0.163   0.8711
month.jun   -0.0935115  0.1983496  -0.471   0.6379
month.mar    0.1428452  0.1359841   1.050   0.2948
month.oct    0.3224931  0.2506387   1.287   0.1998
X            0.0009583  0.0214136   0.045   0.9644
DMC          0.0064152  0.0038252   1.677   0.0952 .
temp        -0.0009754  0.0038757  -0.252   0.8016
RH          -0.0505237  0.1534175  -0.329   0.7423
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4112 on 191 degrees of freedom
Multiple R-squared:  0.05405,   Adjusted R-squared:  -0.005382
F-statistic: 0.9094 on 12 and 191 DF,  p-value: 0.5386
```

References

1. Cortez, Paulo & Morais, A.. (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data.

2. Stanford-Moore, A., & Moore, B. (n.d.). Wildfire Burn Area Prediction. http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26582553.pdf.

3. Carmine Maffei, Massimo Menenti, Predicting forest fires burned area and rate of spread from pre-fire multispectral satellite measurements, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 158, 2019, Pages 263-278, ISSN 0924-2716, https://doi.org/10.1016/j.isprsjprs.2019.10.013. (https://www.sciencedirect.com/science/article/pii/S0924271619302515)
4.Singh, Deepika (2019). Linear, Lasso, and Ridge Regression with R.https://www.pluralsight.com/guides/linear-

Appendices

Supporting Graphs - Data Exploration

Statistics for Variables

```
##         vars   n   mean     sd median trimmed    mad  min     max   range  skew
## X          1 517   4.67   2.31   4.00    4.67   2.97  1.0    9.00    8.00  0.04
## Y          2 517   4.30   1.23   4.00    4.31   1.48  2.0    9.00    7.00  0.41
## month*     3 517   6.76   4.37   7.00    6.72   7.41  1.0   12.00   11.00  0.08
## day*       4 517   3.74   1.93   4.00    3.67   2.97  1.0    7.00    6.00  0.16
## FFMC       5 517  90.64   5.52  91.60   91.45   1.93 18.7   96.20   77.50 -6.54
## DMC        6 517 110.87  64.05 108.30  106.52  51.74  1.1  291.30  290.20  0.54
## DC         7 517 547.94 248.07 664.20  578.69 118.90  7.9  860.60  852.70 -1.09
## ISI        8 517   9.02   4.56   8.40    8.73   3.11  0.0   56.10   56.10  2.52
## temp       9 517  18.89   5.81  19.30   19.09   5.34  2.2   33.30   31.10 -0.33
## RH        10 517  44.29  16.32  42.00   42.71  14.83 15.0  100.00   85.00  0.86
## wind      11 517   4.02   1.79   4.00    3.90   1.93  0.4    9.40    9.00  0.57
## rain      12 517   0.02   0.30   0.00    0.00   0.00  0.0    6.40    6.40 19.70
## area      13 517  12.85  63.66   0.52    3.18   0.77  0.0 1090.84 1090.84 12.77
##         kurtosis    se
## X         -1.18  0.10
## Y          1.38  0.05
## month*    -1.72  0.19
## day*      -1.11  0.08
## FFMC      66.14  0.24
## DMC        0.18  2.82
## DC        -0.27 10.91
## ISI       21.15  0.20
## temp       0.11  0.26
## RH         0.41  0.72
## wind       0.03  0.08
## rain     415.60  0.01
## area     191.50  2.80
```

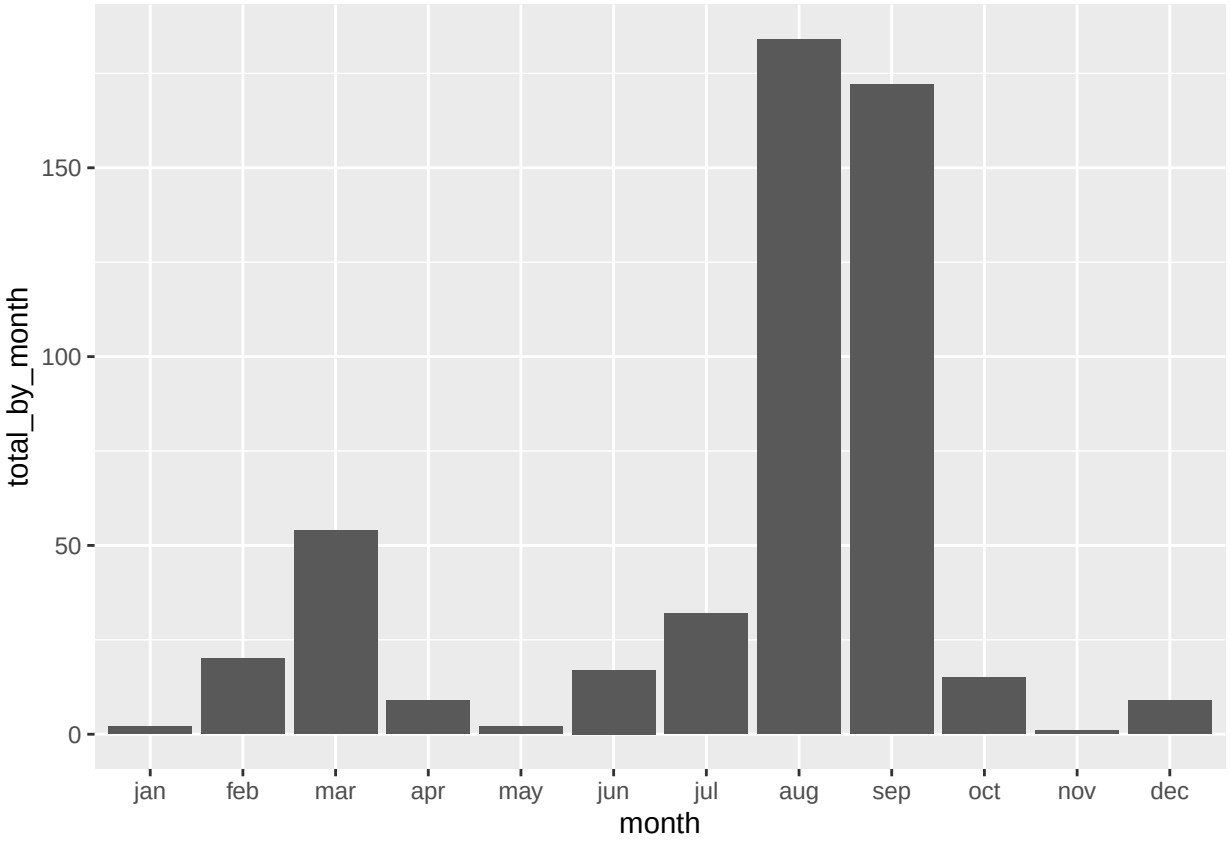
Overview of Variables

Unique Area Values

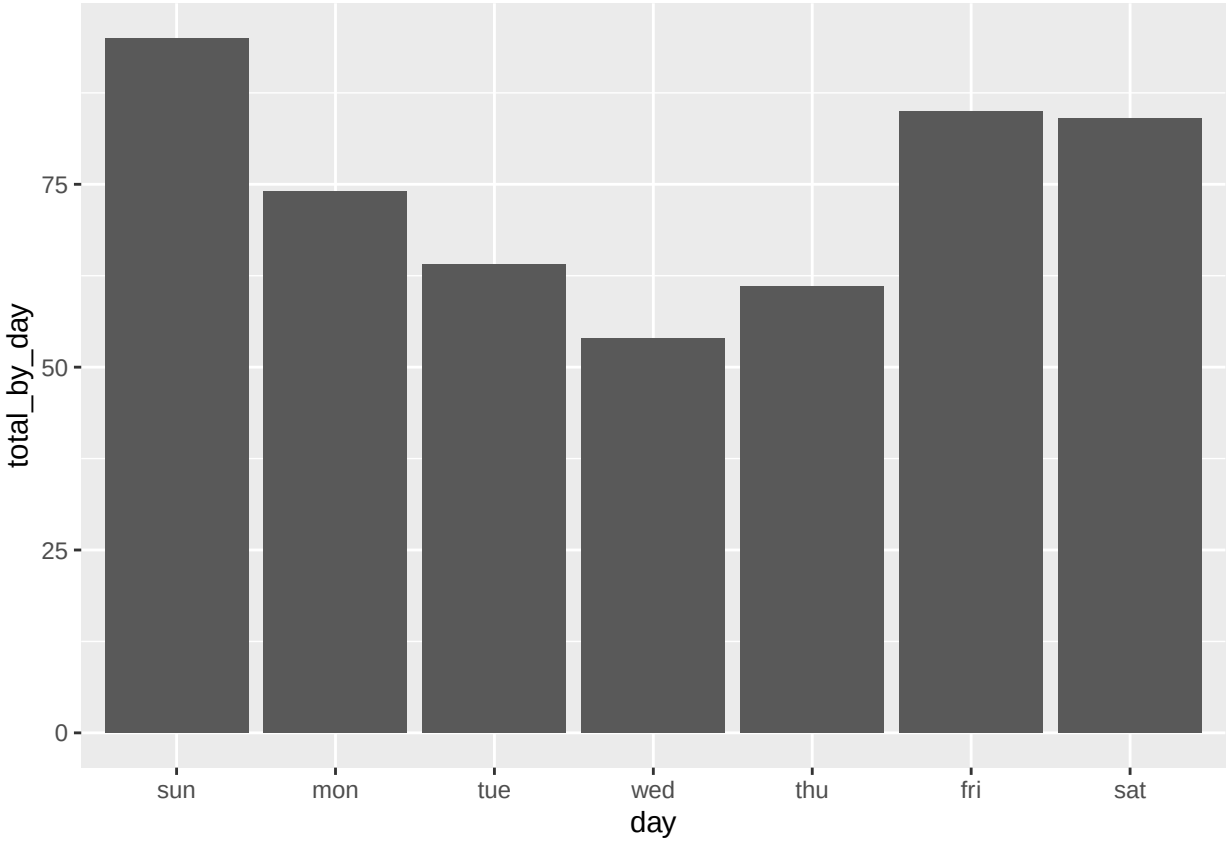
```
## # A tibble: 251 x 2
##     area area_0
##    <dbl>  <int>
##  1  0       247
##  2  0.09      1
##  3  0.17      1
##  4  0.21      1
##  5  0.24      1
##  6  0.33      1
##  7  0.36      1
##  8  0.41      1
##  9  0.43      2
## 10  0.47      1
```

```
## # ... with 241 more rows
```

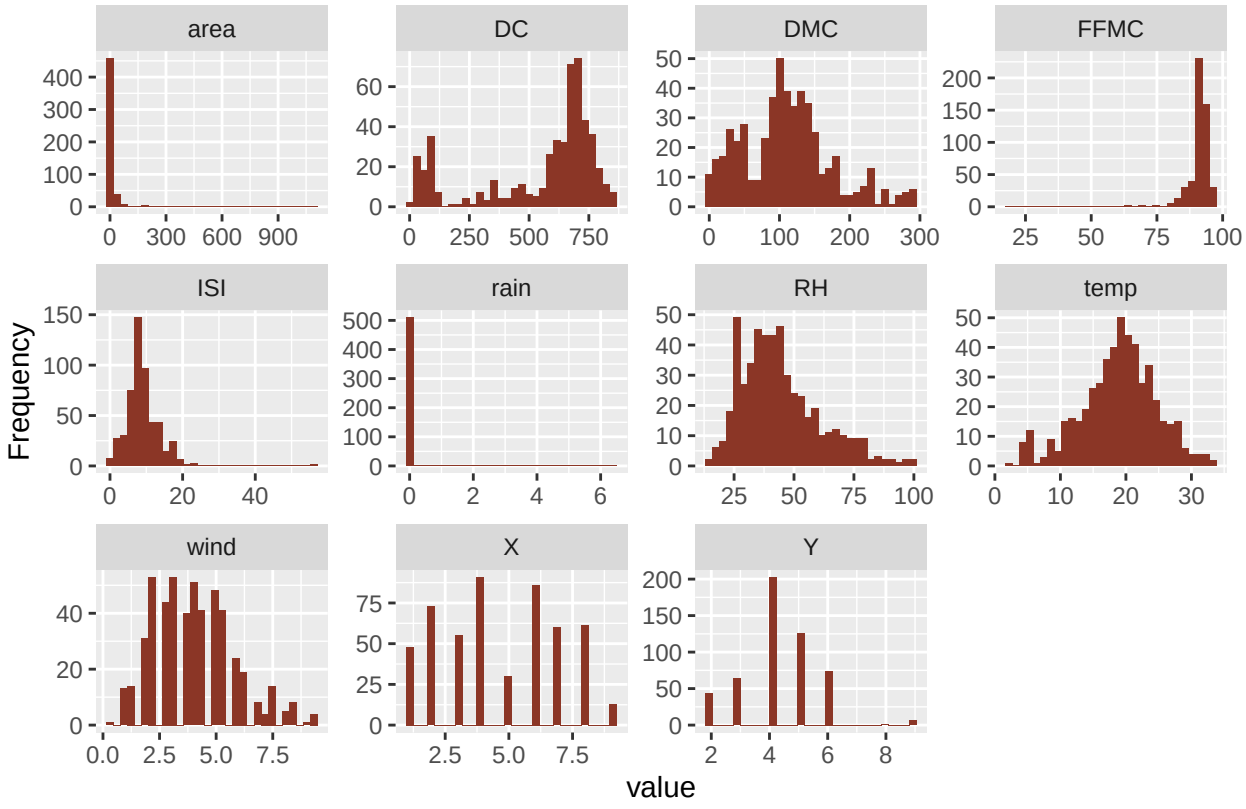
Unique Months



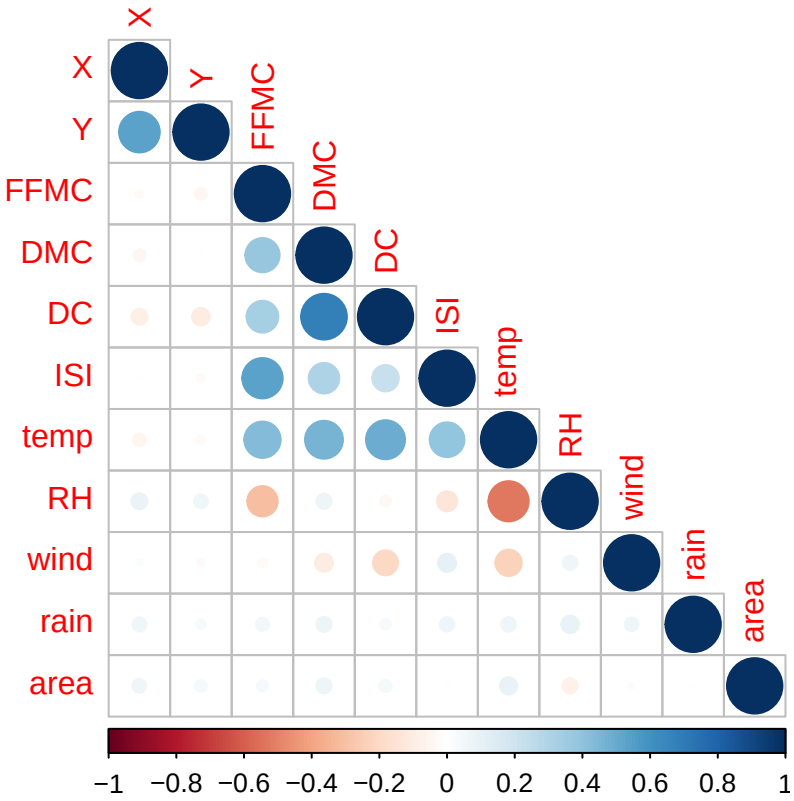
Unique Days



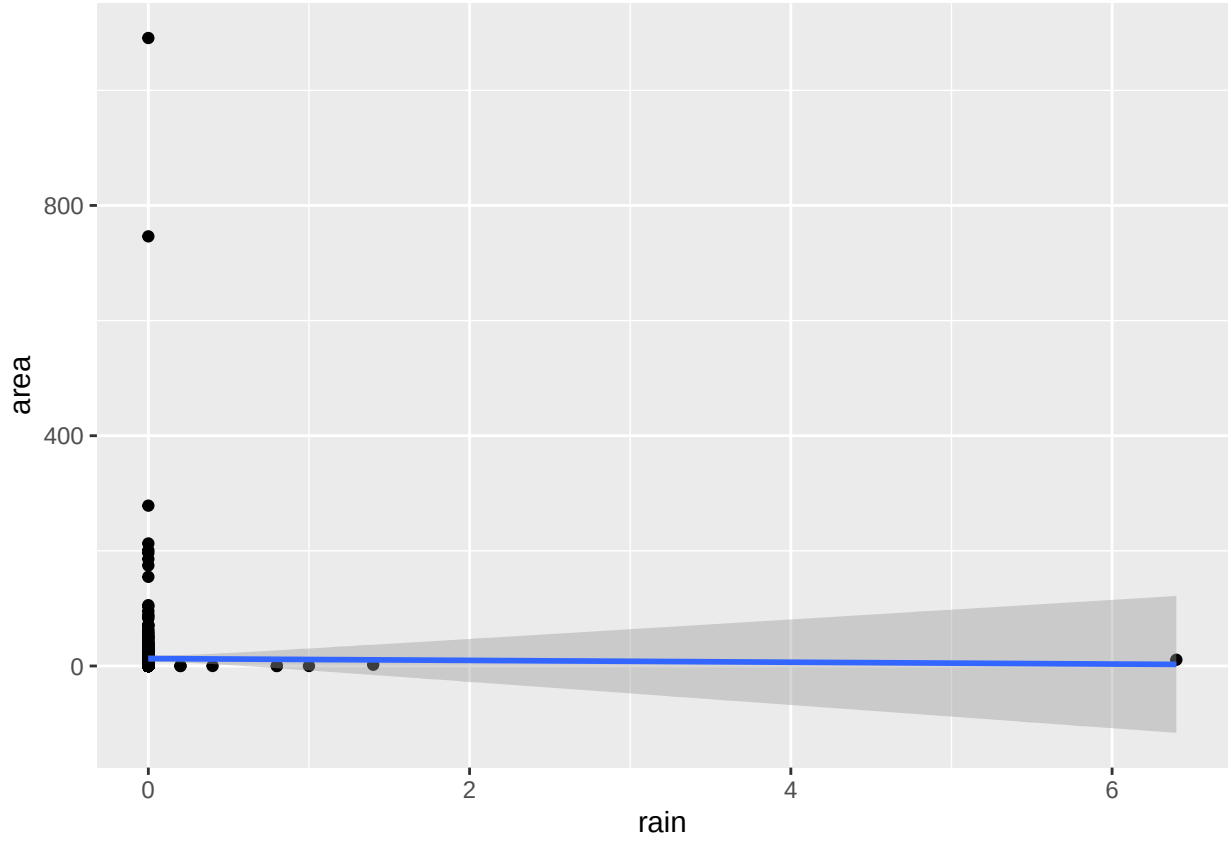
Numeric Variables

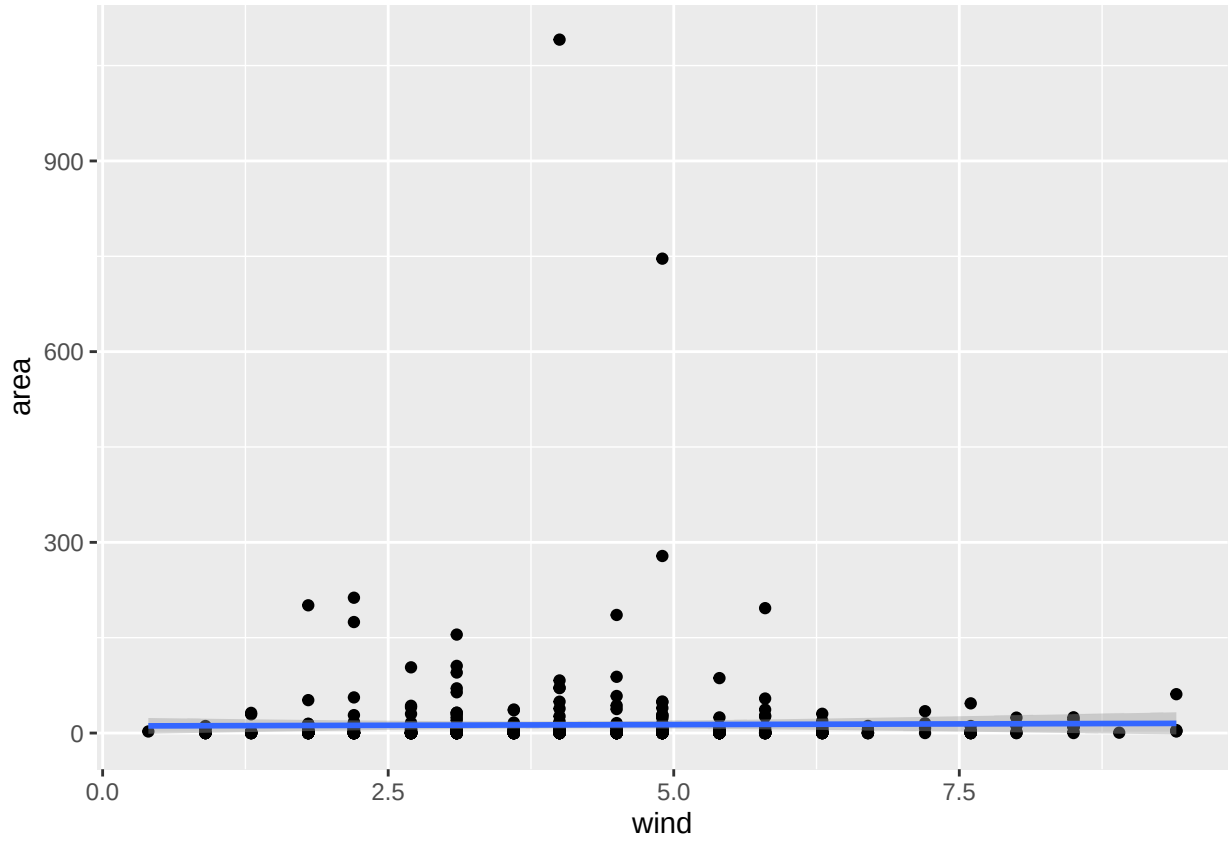


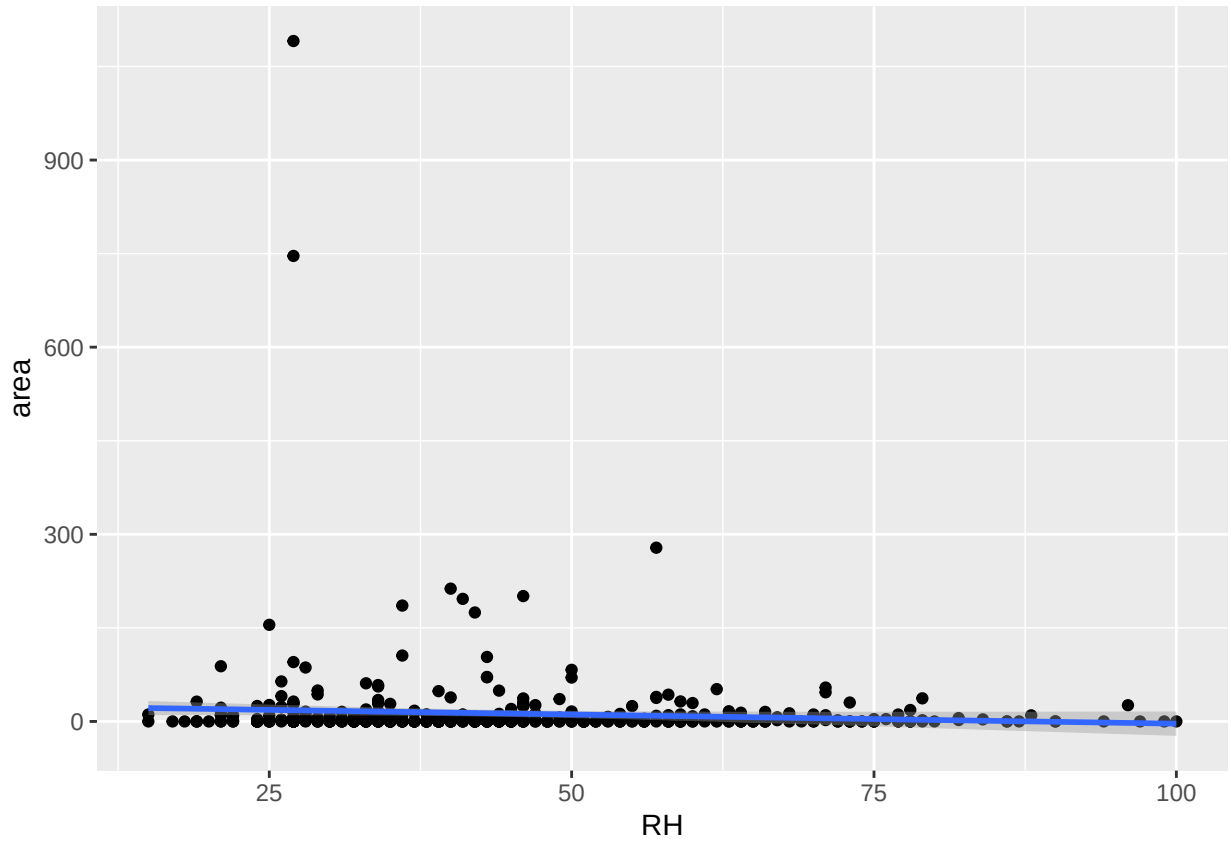
Correlations The `corrplot` below shows the correlation between predictor variables by ignoring the missing entries.

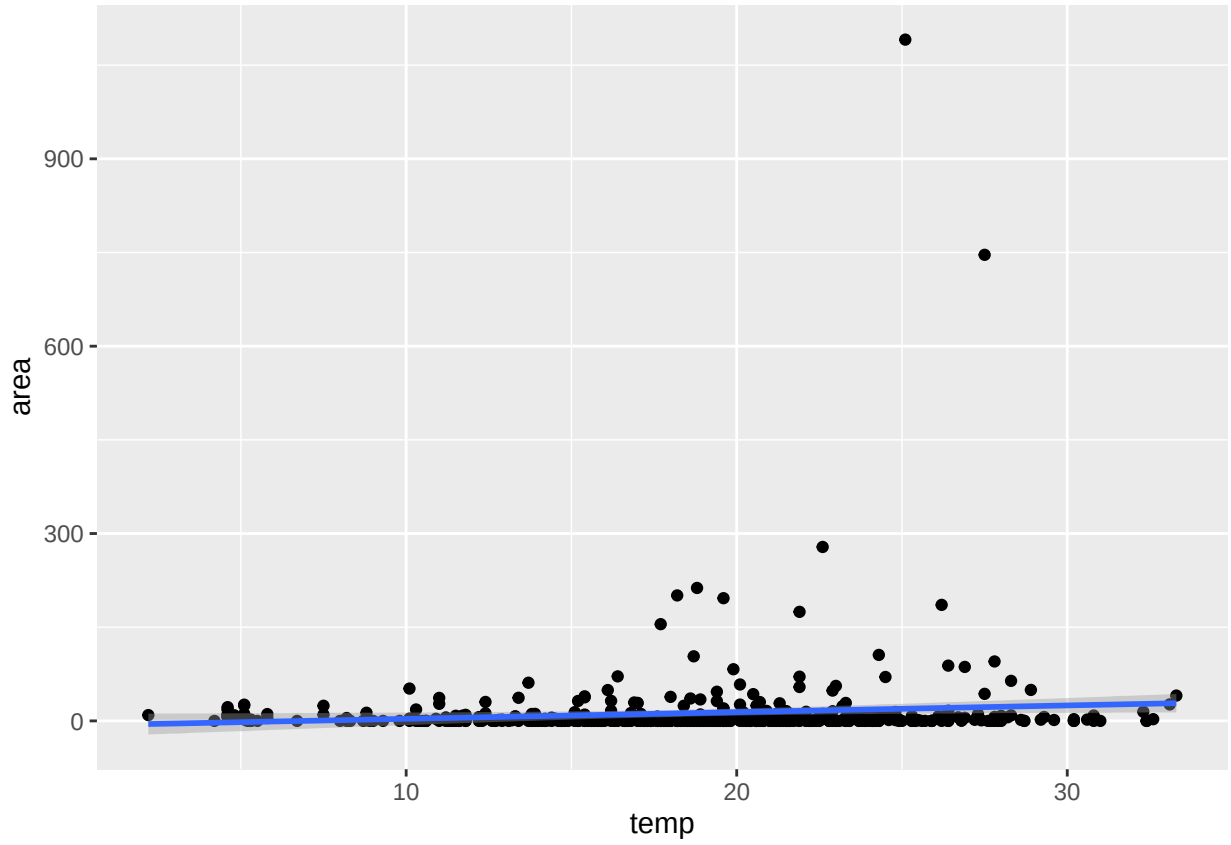


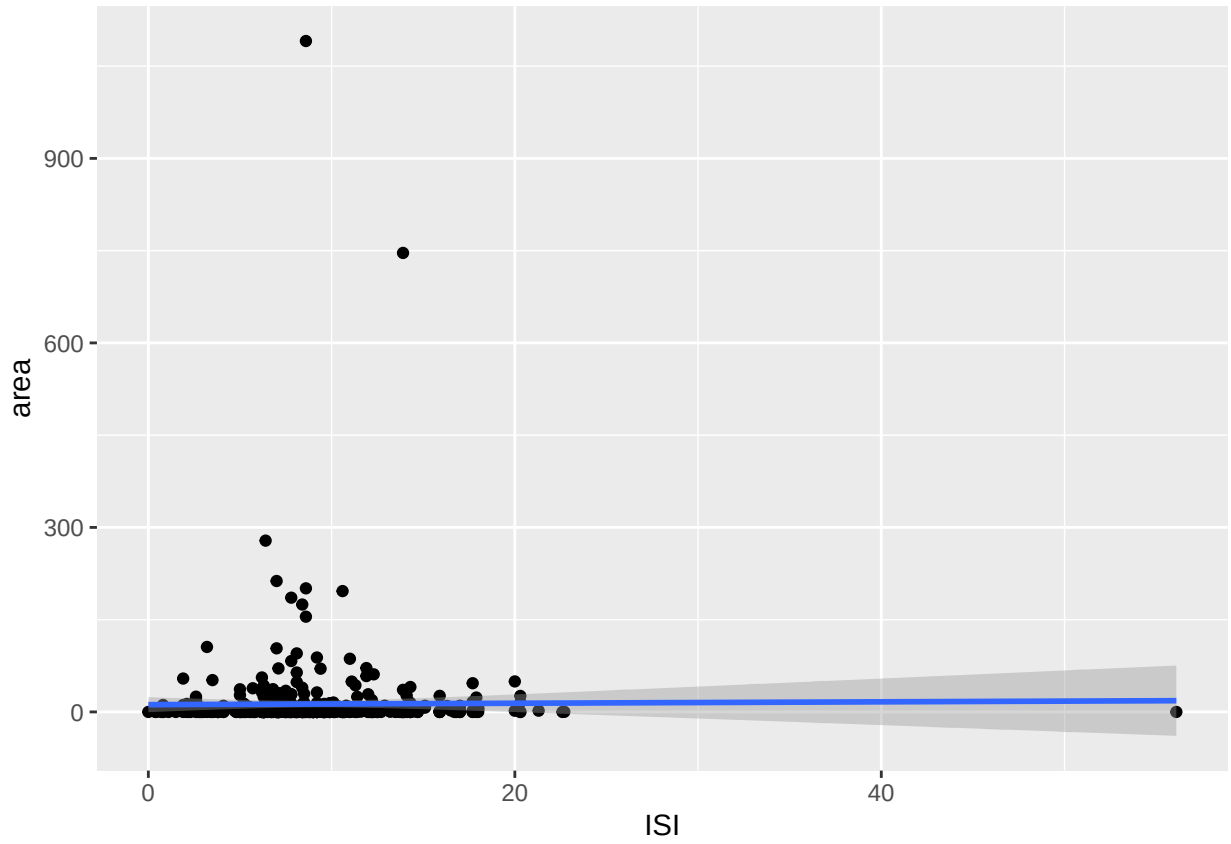
Relationship between Response and Predictor Variables

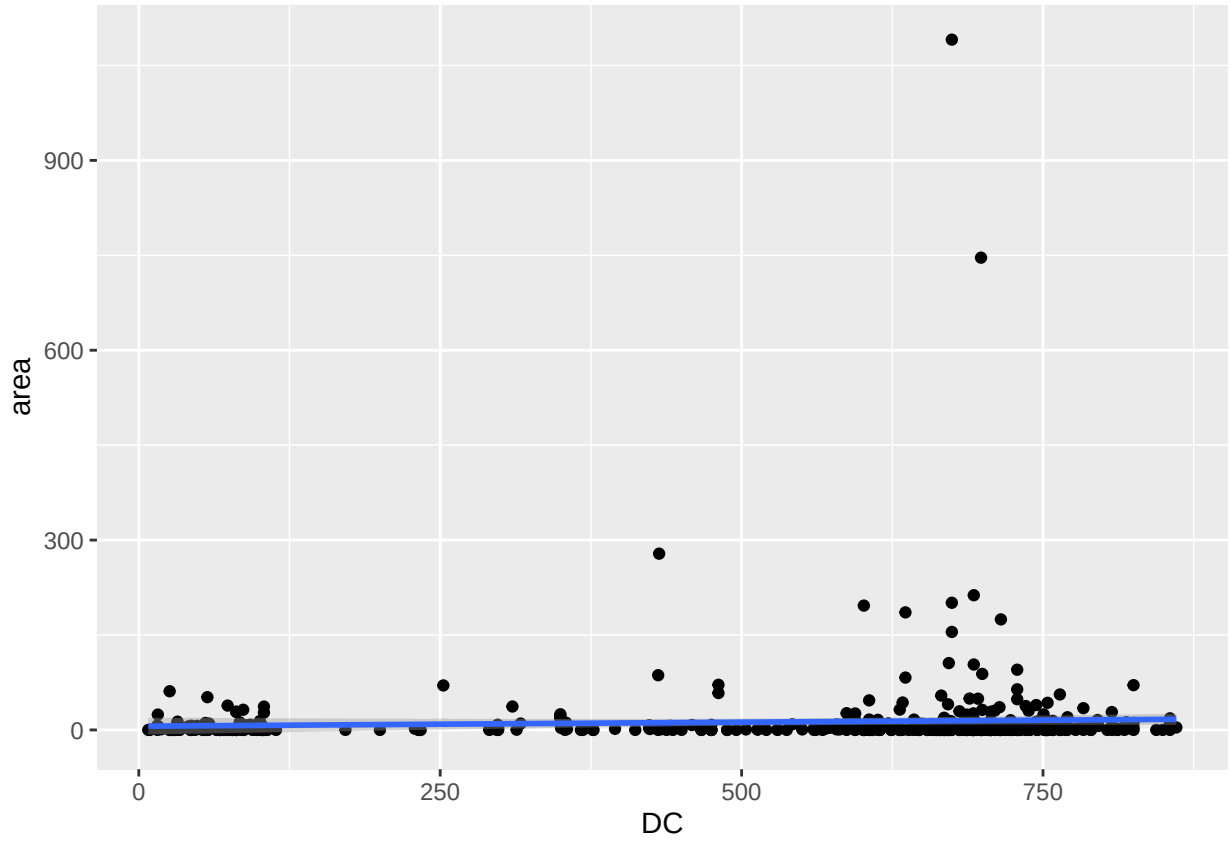

area
800
400
0
0
2
4
6
rain

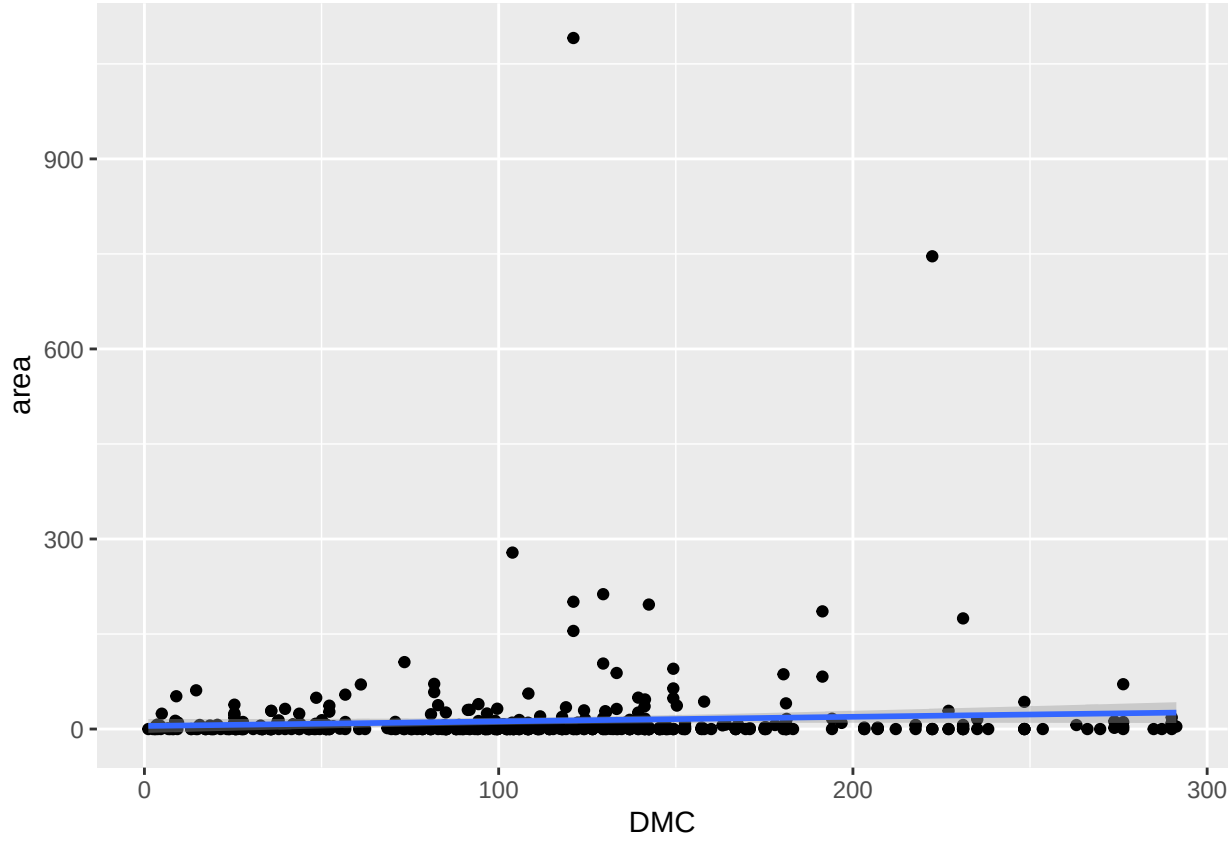

area
900
600
300
0
0.0
2.5
5.0
7.5
wind

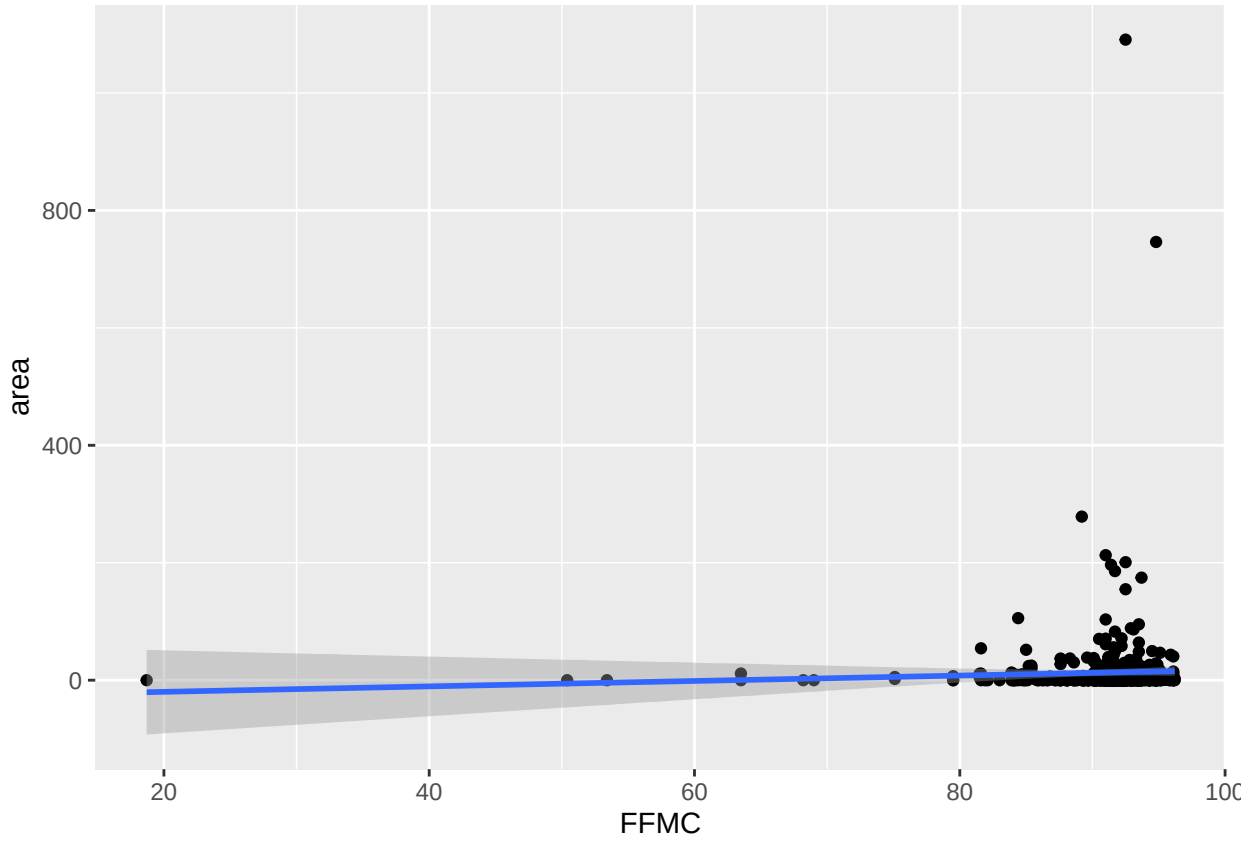

area
900
600
300
0
25
50
75
100
RH

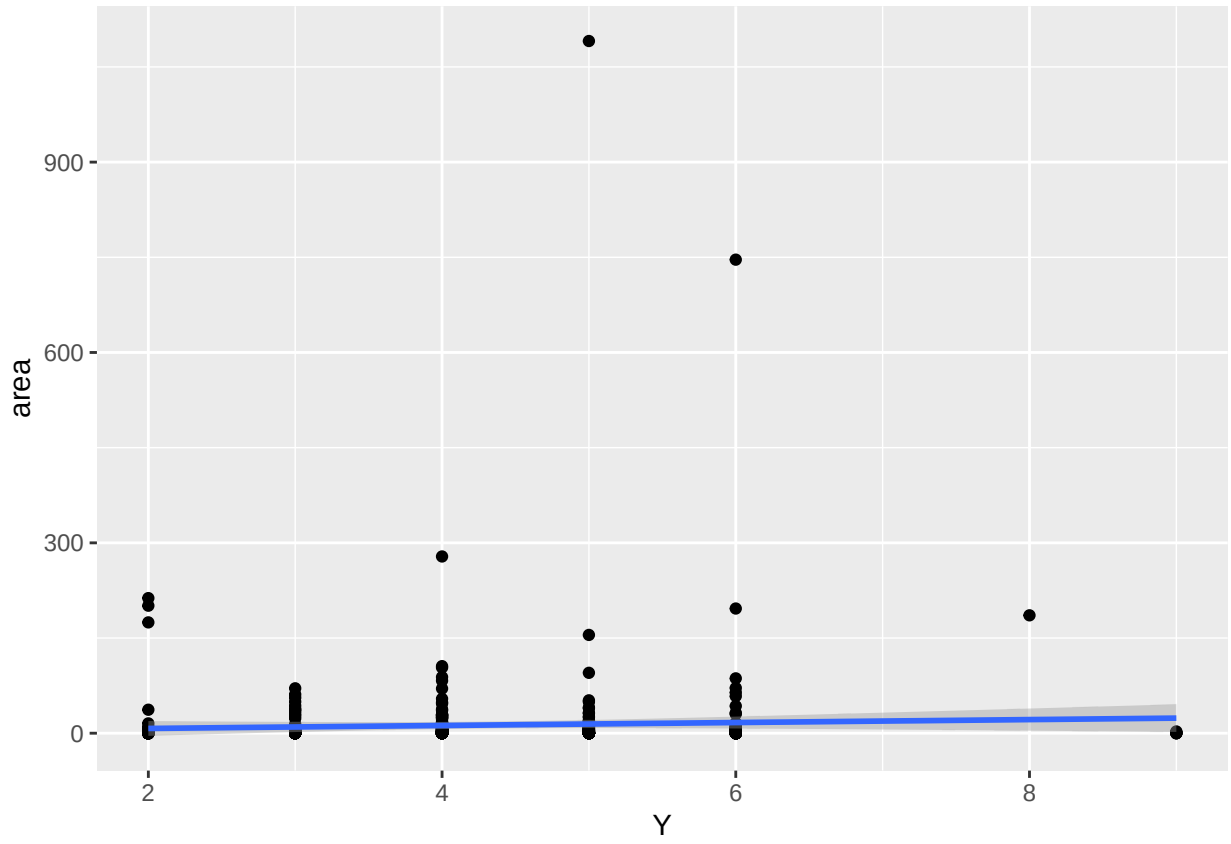

area
900
600
300
0
10
20
30
temp

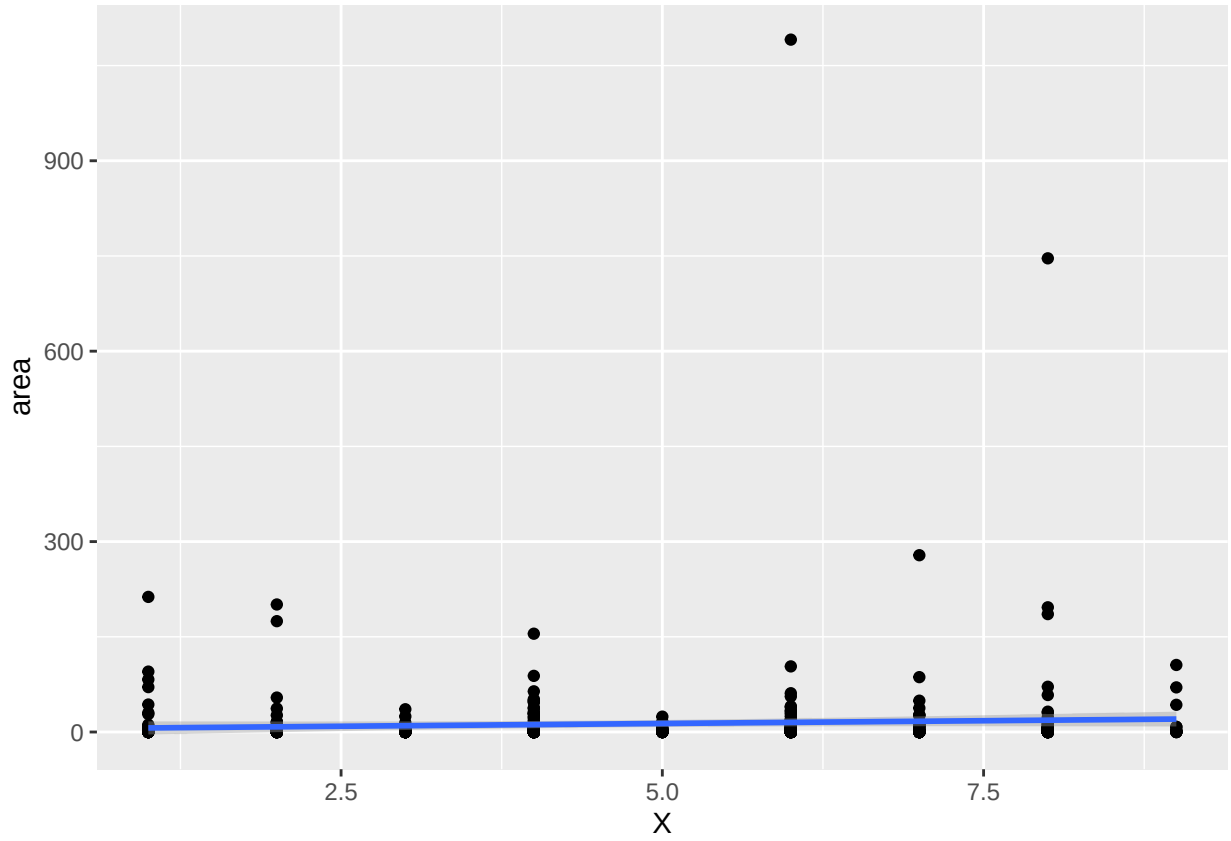

900
600
300
0
area
0
20
40
ISI


900
600
300
0
area
0
250
500
750
DC

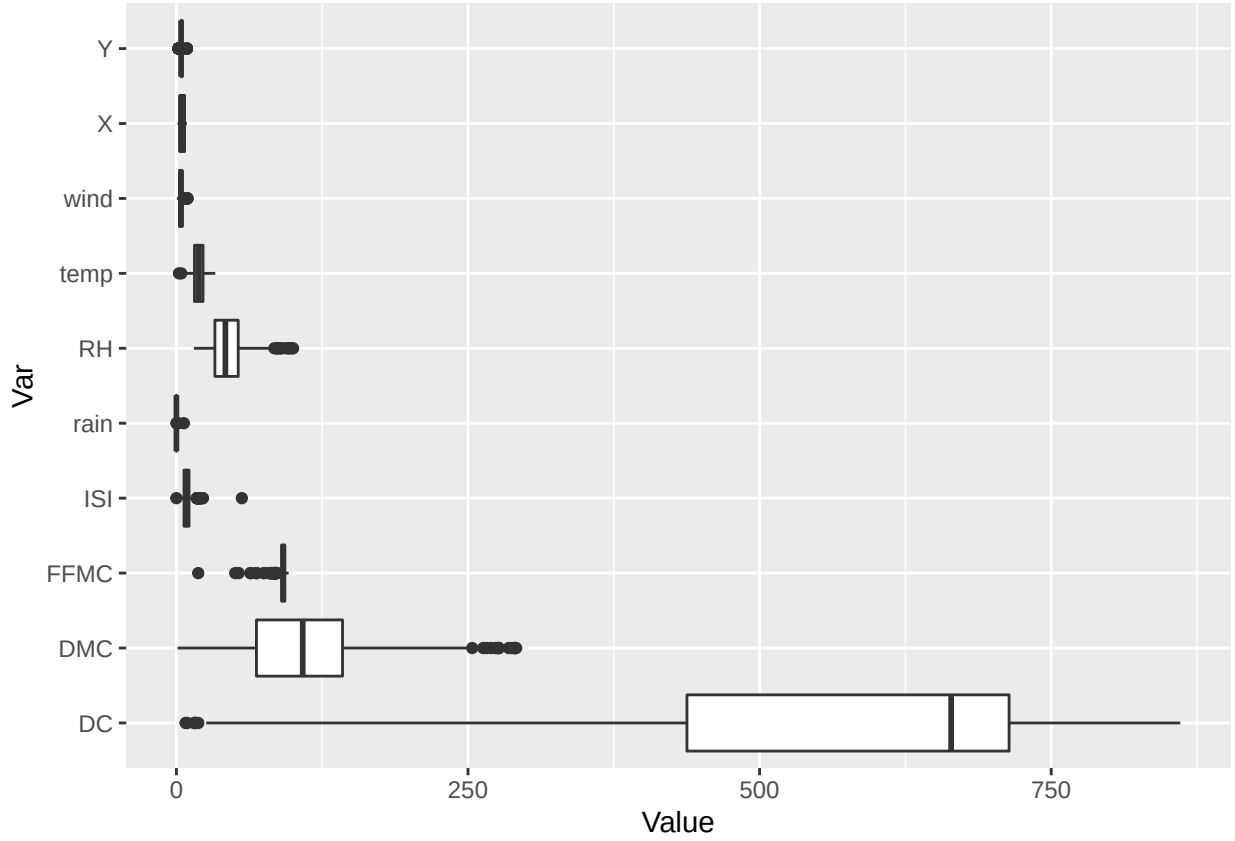

900
600
300
0
area
0
100
200
300
DMC


800
400
0
area
20
40
60
80
100
FFMC


900
600
300
0
area
2
4
6
8
Y


900
600
300
0
area
2.5
5.0
7.5
X

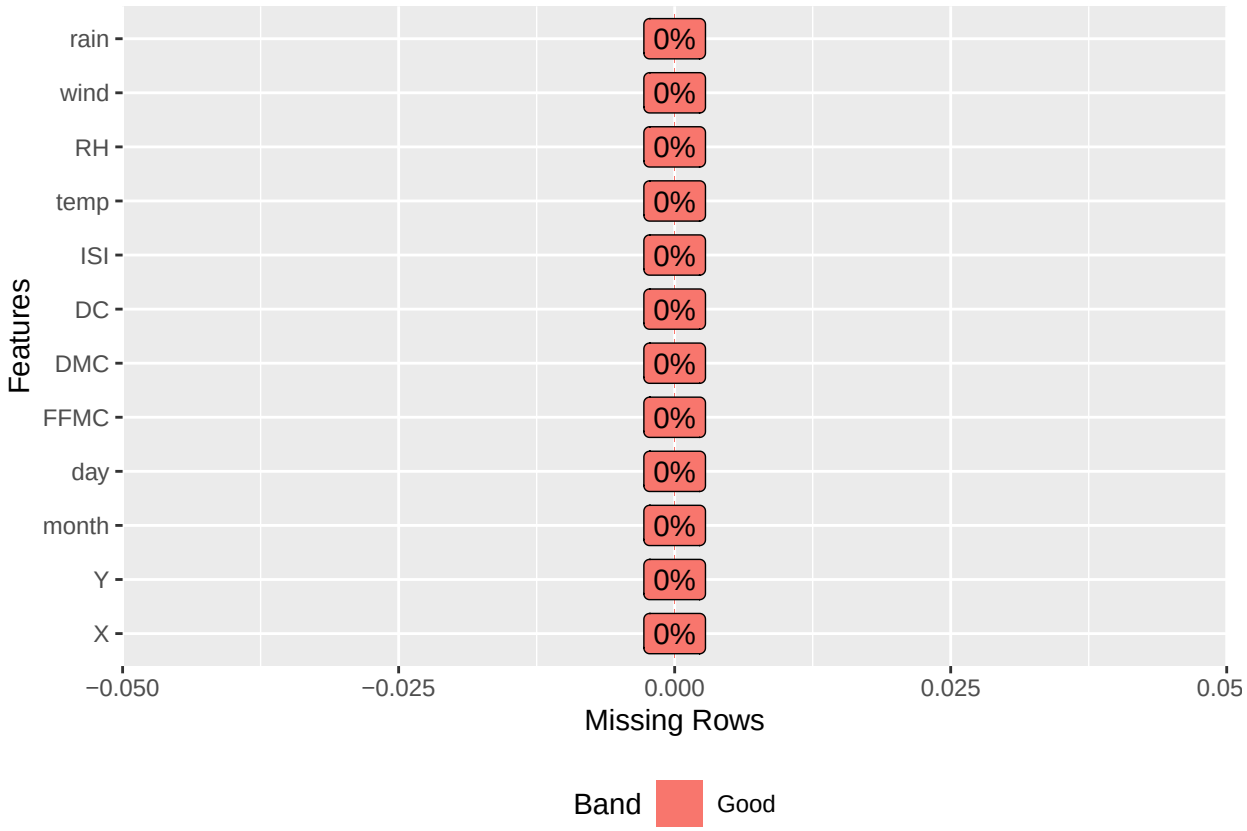
Boxplot Outliers Variables such as RH, FFMC, DMC and DC appear to have outliers in the data.



Supporting Graphs - Handling Missing Values

Missing Values

```
## # A tibble: 12 x 2
##    month `n()`
##    <fct> <int>
##  1 apr       9
##  2 aug     184
##  3 dec       9
##  4 feb      20
##  5 jan       2
##  6 jul      32
##  7 jun      17
##  8 mar      54
##  9 may       2
## 10 nov       1
## 11 oct      15
## 12 sep     172
```




```
##   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
```

Supporting Graphs & Code - Data Preparation

Missing Values

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

Creating Dummy Variables

```
##   day.mon day.sat day.sun day.thu day.tue day.wed month.aug month.dec month.feb
## 1       0       0       0       0       0       0         0         0         0
## 2       0       0       0       0       1       0         0         0         0
## 3       0       1       0       0       0       0         0         0         0
## 4       0       0       0       0       0       0         0         0         0
## 5       0       0       1       0       0       0         0         0         0
```

```
## 6       0       0       1       0       0       0         1         0         0
##   month.jan month.jul month.jun month.mar month.may month.nov month.oct
## 1         0         0         0         1         0         0         0
## 2         0         0         0         0         0         0         1
## 3         0         0         0         0         0         0         1
## 4         0         0         0         1         0         0         0
## 5         0         0         0         1         0         0         0
## 6         0         0         0         0         0         0         0
##   month.sep X Y FFMC  DMC    DC  ISI temp RH wind area
## 1         0 7 5 86.2 26.2  94.3  5.1  8.2 51  6.7    0
## 2         0 7 4 90.6 35.4 669.1  6.7 18.0 33  0.9    0
## 3         0 7 4 90.6 43.7 686.9  6.7 14.6 33  1.3    0
## 4         0 8 6 91.7 33.3  77.5  9.0  8.3 97  4.0    0
## 5         0 8 6 89.3 51.3 102.2  9.6 11.4 99  1.8    0
## 6         0 8 6 92.3 85.3 488.0 14.7 22.2 29  5.4    0
```

Preprocess using transformation

```
##   day.mon day.sat day.sun day.thu day.tue day.wed month.aug month.dec month.feb
## 1       0       0       0       0       0       0         0         0         0
## 2       0       0       0       0       1       0         0         0         0
## 3       0       1       0       0       0       0         0         0         0
## 4       0       0       0       0       0       0         0         0         0
## 5       0       0       1       0       0       0         0         0         0
## 6       0       0       1       0       0       0         1         0         0
##   month.jan month.jul month.jun month.mar month.may month.nov month.oct
## 1         0         0         0         1         0         0         0
## 2         0         0         0         0         0         0         1
## 3         0         0         0         0         0         0         1
## 4         0         0         0         1         0         0         0
## 5         0         0         0         1         0         0         0
## 6         0         0         0         0         0         0         0
##   month.sep        X        Y FFMC      DMC        DC      ISI     temp
## 1         0 4.719169 3.212827 86.2 11.21099  510.7975 2.651478 12.89604
## 2         0 4.719169 2.707822 90.6 13.80252 8655.2357 3.154088 34.18492
## 3         0 4.719169 2.707822 90.6 15.93924 8990.6886 3.154088 26.30863
## 4         0 5.249877 3.684990 91.7 13.23445  385.3973 3.776615 13.08871
## 5         0 5.249877 3.684990 89.3 17.77076  573.4218 3.924486 19.34488
## 6         0 5.249877 3.684990 92.3 24.99253 5480.3976 5.016372 44.50093
##         RH      wind area
## 1 3.546170 2.9191850    0
## 2 3.201249 0.7153387    0
## 3 3.201249 0.9594060    0
## 4 4.045678 2.1271965    0
## 5 4.061313 1.2274007    0
## 6 3.098072 2.5660589    0
```

Training and Test Partition

Supporting Code - Experimentation and Results

Linear Regression

Model 1 - all variables

```
##
## Call:
## lm(formula = area ~ ., data = X.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9466 -0.5441 -0.2320  0.5879  1.2607
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.391e+00  1.093e+00  -1.272  0.20402
## day.mon      5.389e-02  1.242e-01   0.434  0.66466
## day.sat      1.670e-01  1.195e-01   1.398  0.16311
## day.sun      2.972e-02  1.135e-01   0.262  0.79360
## day.thu      7.460e-02  1.331e-01   0.560  0.57550
## day.tue      1.584e-01  1.240e-01   1.278  0.20213
## day.wed      8.410e-02  1.315e-01   0.639  0.52292
## month.aug   -1.068e-01  3.724e-01  -0.287  0.77442
## month.dec    1.003e+00  3.759e-01   2.668  0.00798 **
## month.feb    7.192e-02  3.008e-01   0.239  0.81119
## month.jan   -3.528e-01  5.964e-01  -0.592  0.55453
## month.jul   -2.349e-01  3.335e-01  -0.704  0.48166
## month.jun   -2.959e-01  3.453e-01  -0.857  0.39193
## month.mar   -2.089e-01  2.701e-01  -0.773  0.43987
## month.may    9.517e-01  6.964e-01   1.367  0.17259
## month.nov   -5.971e-01  7.011e-01  -0.852  0.39493
## month.oct   -2.364e-02  4.537e-01  -0.052  0.95848
## month.sep    1.165e-01  4.146e-01   0.281  0.77881
## X            1.248e-02  2.888e-02   0.432  0.66586
## Y           -1.516e-02  6.523e-02  -0.232  0.81635
## FFMC         4.947e-03  9.341e-03   0.530  0.59667
## DMC          6.636e-03  6.383e-03   1.040  0.29923
## DC          -3.907e-05  4.022e-05  -0.971  0.33198
## ISI         -2.196e-02  4.712e-02  -0.466  0.64151
## temp         1.065e-02  4.686e-03   2.272  0.02365 *
## RH           3.069e-01  1.782e-01   1.722  0.08589 .
## wind         1.145e-01  6.076e-02   1.885  0.06020 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6432 on 362 degrees of freedom
## Multiple R-squared:  0.08925,    Adjusted R-squared:  0.02383
## F-statistic: 1.364 on 26 and 362 DF,  p-value: 0.1128
```

Model 2 - This subset only runs the linear model for small forest fires.

```
##
## Call:
## glm.nb(formula = area ~ ., data = small_fires, init.theta = 15682.82901,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4442  -1.0342  -0.8406   0.7458   1.4755
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.558e+00  2.686e+00  -1.697   0.0897 .
## day.mon      1.035e-01  2.578e-01   0.402   0.6879
## day.sat      2.740e-01  2.396e-01   1.144   0.2528
## day.sun      3.677e-02  2.344e-01   0.157   0.8753
## day.thu      1.132e-01  2.747e-01   0.412   0.6804
## day.tue      2.509e-01  2.454e-01   1.022   0.3066
## day.wed      1.278e-01  2.679e-01   0.477   0.6334
## month.aug   -1.745e-01  7.366e-01  -0.237   0.8127
## month.dec    1.132e+00  6.646e-01   1.704   0.0884 .
## month.feb    1.277e-01  6.002e-01   0.213   0.8315
## month.jan   -3.580e+01  4.745e+07   0.000   1.0000
## month.jul   -3.843e-01  6.634e-01  -0.579   0.5624
## month.jun   -4.725e-01  7.090e-01  -0.666   0.5052
## month.mar   -4.169e-01  5.550e-01  -0.751   0.4525
## month.may    8.346e-01  9.492e-01   0.879   0.3792
## month.nov   -3.651e+01  6.711e+07   0.000   1.0000
## month.oct   -1.418e-01  9.456e-01  -0.150   0.8808
## month.sep    1.823e-01  8.156e-01   0.224   0.8231
## X            1.870e-02  5.707e-02   0.328   0.7431
## Y           -2.824e-02  1.312e-01  -0.215   0.8296
## FFMC         1.842e-02  2.644e-02   0.697   0.4860
## DMC          1.004e-02  1.273e-02   0.789   0.4304
## DC          -6.445e-05  7.949e-05  -0.811   0.4175
## ISI         -6.382e-02  1.045e-01  -0.611   0.5415
## temp         1.685e-02  9.172e-03   1.837   0.0662 .
## RH           4.936e-01  3.460e-01   1.427   0.1537
## wind         2.055e-01  1.253e-01   1.640   0.1011
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(15682.83) family taken to be 1)
##
##     Null deviance: 368.95  on 388  degrees of freedom
## Residual deviance: 345.35  on 362  degrees of freedom
## AIC: 792.5
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  15683
##           Std. Err.:  115361
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -736.495
```

Model 3 - Remove fires where burn area is 0 hectares.

```
##
## Call:
## lm(formula = area ~ ., data = fires_over0)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07467 -0.24693  0.03432  0.26510  0.78189
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.155e+00  1.418e+00   1.520  0.13035
## day.mon     -1.057e-01  1.179e-01  -0.897  0.37104
## day.sat      1.181e-01  1.088e-01   1.085  0.27935
## day.sun      4.843e-02  1.055e-01   0.459  0.64679
## day.thu     -3.274e-03  1.228e-01  -0.027  0.97875
## day.tue      3.338e-02  1.083e-01   0.308  0.75821
## day.wed     -9.990e-02  1.165e-01  -0.858  0.39222
## month.aug   -1.863e-01  3.515e-01  -0.530  0.59670
## month.dec    2.028e-01  3.136e-01   0.647  0.51858
## month.feb   -1.839e-01  2.919e-01  -0.630  0.52941
## month.jan           NA         NA      NA       NA
## month.jul   -2.474e-01  3.222e-01  -0.768  0.44356
## month.jun   -3.520e-01  3.375e-01  -1.043  0.29840
## month.mar   -2.215e-01  2.709e-01  -0.818  0.41469
## month.may    2.038e-01  4.909e-01   0.415  0.67850
## month.nov           NA         NA      NA       NA
## month.oct    6.091e-01  4.595e-01   1.325  0.18670
## month.sep    1.176e-01  3.906e-01   0.301  0.76369
## X            3.908e-03  2.455e-02   0.159  0.87370
## Y           -2.209e-02  5.933e-02  -0.372  0.71014
## FFMC        -8.499e-03  1.413e-02  -0.601  0.54830
## DMC          1.757e-02  5.550e-03   3.167  0.00181 **
## DC          -7.457e-05  3.579e-05  -2.084  0.03860 *
## ISI         -9.136e-03  5.202e-02  -0.176  0.86080
## temp         1.632e-03  4.143e-03   0.394  0.69405
## RH          -6.672e-02  1.636e-01  -0.408  0.68393
## wind         5.104e-02  5.582e-02   0.914  0.36173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.405 on 179 degrees of freedom
## Multiple R-squared:  0.1401, Adjusted R-squared:  0.02481
## F-statistic: 1.215 on 24 and 179 DF,  p-value: 0.234
```

Model 4 - leaps package Now, that we have run leaps through our dataset, let's see what the package recommends in terms of the number of predictors to use for our dataset.

```
## [1] 8
```

Seems like we have to use 13 predictors to get the best mode and the summary table below provides details on which predictors to use for the model. The best predictors are indicated by 'TRUE'.

```
## (Intercept)     day.mon     day.sat     day.sun     day.thu     day.tue
##        TRUE       FALSE        TRUE       FALSE       FALSE        TRUE
##     day.wed   month.aug   month.dec   month.feb   month.jan   month.jul
##       FALSE       FALSE        TRUE        TRUE        TRUE       FALSE
##   month.jun   month.mar   month.may   month.nov   month.oct   month.sep
##       FALSE       FALSE        TRUE        TRUE       FALSE        TRUE
##           X           Y        FFMC         DMC          DC         ISI
```

```
##       FALSE       FALSE       FALSE        TRUE        TRUE       FALSE
##        temp          RH        wind
##        TRUE        TRUE        TRUE
```

```
##
## Call:
## lm(formula = area ~ day.thu + month.aug + month.dec + month.jan +
##     month.jul + month.jun + month.mar + month.oct + month.sep +
##     X + Y + DMC + temp + RH, data = X.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8404 -0.5686 -0.2391  0.5846  1.3098
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.526351   0.672917  -0.782 0.434596
## day.thu     -0.027057   0.106356  -0.254 0.799326
## month.aug   -0.337794   0.220927  -1.529 0.127113
## month.dec    0.958873   0.282713   3.392 0.000769 ***
## month.jan   -0.621250   0.478684  -1.298 0.195146
## month.jul   -0.361440   0.232224  -1.556 0.120452
## month.jun   -0.388527   0.269563  -1.441 0.150330
## month.mar   -0.189568   0.170263  -1.113 0.266259
## month.oct   -0.360612   0.243269  -1.482 0.139087
## month.sep   -0.238568   0.196561  -1.214 0.225626
## X            0.010997   0.028423   0.387 0.699044
## Y           -0.014608   0.063351  -0.231 0.817762
## DMC          0.003299   0.004715   0.700 0.484586
## temp         0.010216   0.004584   2.229 0.026438 *
## RH           0.272472   0.170820   1.595 0.111538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.642 on 374 degrees of freedom
## Multiple R-squared:  0.06248,    Adjusted R-squared:  0.02739
## F-statistic:  1.78 on 14 and 374 DF,  p-value: 0.03967
```

Robust Regression

First Robut Regression

```
##
## Call: rlm(formula = area ~ day.thu + month.aug + month.dec + month.jan +
##     month.jul + month.jun + month.mar + month.oct + month.sep +
##     X + Y + DMC + temp + RH, data = X.train, psi = psi.bisquare)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8365 -0.5557 -0.2335  0.5921  1.3402
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept) -0.6484  0.7232    -0.8967
## day.thu     -0.0211  0.1143    -0.1848
## month.aug   -0.3587  0.2374    -1.5108
```

```
## month.dec    0.9861  0.3038     3.2457
## month.jan   -0.6145  0.5144    -1.1945
## month.jul   -0.3870  0.2496    -1.5507
## month.jun   -0.3995  0.2897    -1.3789
## month.mar   -0.2047  0.1830    -1.1189
## month.oct   -0.3934  0.2614    -1.5046
## month.sep   -0.2557  0.2112    -1.2103
## X            0.0109  0.0305     0.3558
## Y           -0.0150  0.0681    -0.2203
## DMC          0.0033  0.0051     0.6468
## temp         0.0111  0.0049     2.2452
## RH           0.3003  0.1836     1.6357
##
## Residual standard error: 0.8386 on 374 degrees of freedom
```

```
## [1] 0.6419812
```

```
## [1] 0.838557
```

Model 4 (modified)

```
## [1] 12
```

```
## (Intercept)     day.mon     day.sat     day.sun     day.thu     day.tue
##        TRUE       FALSE        TRUE       FALSE       FALSE        TRUE
##     day.wed   month.aug   month.dec   month.feb   month.jan   month.jul
##       FALSE       FALSE        TRUE        TRUE       FALSE       FALSE
##   month.jun   month.mar   month.may   month.nov   month.oct   month.sep
##       FALSE       FALSE        TRUE       FALSE       FALSE        TRUE
##           X           Y        FFMC         DMC          DC         ISI
##       FALSE       FALSE        TRUE        TRUE        TRUE       FALSE
##        temp          RH        wind
##        TRUE        TRUE        TRUE
```

```
##
## Call:
## lm(formula = area ~ day.thu + month.aug + month.dec + month.feb +
##     month.jul + month.jun + month.mar + month.oct + X + DMC +
##     temp + RH, data = fires_over0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02563 -0.29392  0.03845  0.31345  0.73017
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.1975994  0.6192037   1.934   0.0546 .
## day.thu     -0.0128295  0.0968186  -0.133   0.8947
## month.aug   -0.1281100  0.0764066  -1.677   0.0952 .
## month.dec    0.3105683  0.2030811   1.529   0.1278
## month.feb    0.1501138  0.1903242   0.789   0.4312
## month.jul   -0.0212590  0.1307957  -0.163   0.8711
## month.jun   -0.0935115  0.1983496  -0.471   0.6379
## month.mar    0.1428452  0.1359841   1.050   0.2948
## month.oct    0.3224931  0.2506387   1.287   0.1998
## X            0.0009583  0.0214136   0.045   0.9644
```

```
## DMC          0.0064152  0.0038252   1.677   0.0952 .
## temp        -0.0009754  0.0038757  -0.252   0.8016
## RH          -0.0505237  0.1534175  -0.329   0.7423
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4112 on 191 degrees of freedom
## Multiple R-squared:  0.05405,    Adjusted R-squared:  -0.005382
## F-statistic: 0.9094 on 12 and 191 DF,  p-value: 0.5386
```

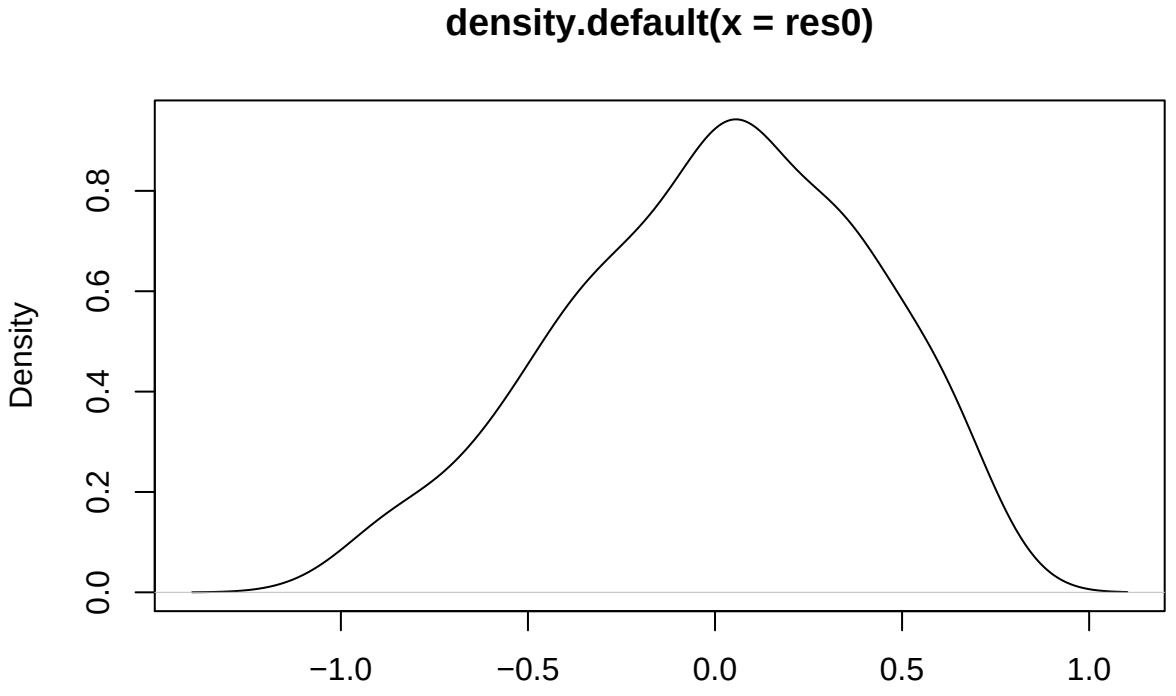
Regularized Model

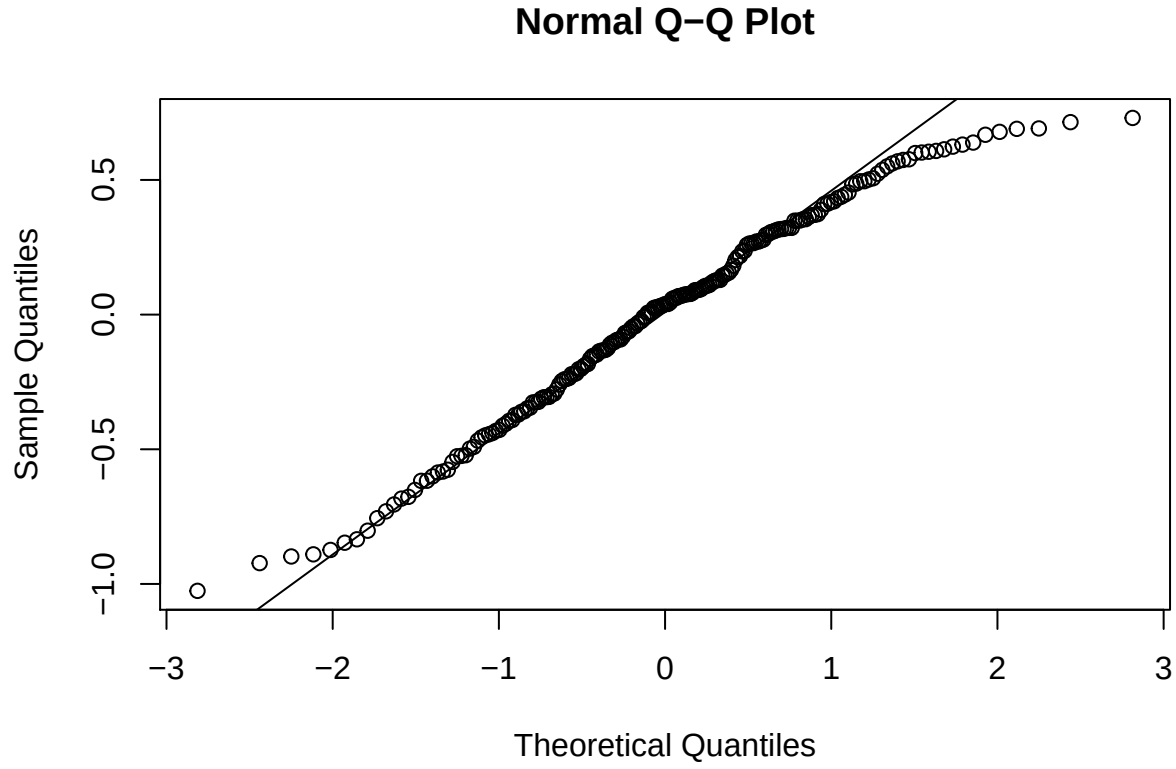
Model 1: Ridge Regression

```
##           Length Class     Mode
## a0         51    -none-    numeric
## beta      612    dgCMatrix S4
## df         51    -none-    numeric
## dim         2    -none-    numeric
## lambda     51    -none-    numeric
## dev.ratio  51    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
## call        6    -none-    call
## nobs        1    -none-    numeric
```

```
## [1] 2.511886
```

```
##        RMSE    Rsquare
## 1 0.2942954 0.01339165
```





The diagnostic plots suggest that the residuals are normally fitted. Therefore, this is our final model.

Predicting on the Test Data

We will use out final model to predict the burned area from forest fires on our fire test, and view the results for the top rows to see if the numbers make sense:

```
##        4       11       12       13       19       20
## 1.213208 1.153766 1.143095 0.991248 1.226917 1.087829
```

```
##
## Call:
## lm(formula = area ~ day.thu + month.aug + month.dec + month.feb +
##     month.jul + month.jun + month.mar + month.oct + X + DMC +
##     temp + RH, data = fires_over0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02563 -0.29392  0.03845  0.31345  0.73017
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.1975994  0.6192037   1.934   0.0546 .
## day.thu     -0.0128295  0.0968186  -0.133   0.8947
## month.aug   -0.1281100  0.0764066  -1.677   0.0952 .
## month.dec    0.3105683  0.2030811   1.529   0.1278
## month.feb    0.1501138  0.1903242   0.789   0.4312
## month.jul   -0.0212590  0.1307957  -0.163   0.8711
## month.jun   -0.0935115  0.1983496  -0.471   0.6379
## month.mar    0.1428452  0.1359841   1.050   0.2948
## month.oct    0.3224931  0.2506387   1.287   0.1998
## X            0.0009583  0.0214136   0.045   0.9644
## DMC          0.0064152  0.0038252   1.677   0.0952 .
## temp        -0.0009754  0.0038757  -0.252   0.8016
## RH          -0.0505237  0.1534175  -0.329   0.7423
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4112 on 191 degrees of freedom
## Multiple R-squared:  0.05405,    Adjusted R-squared:  -0.005382
## F-statistic: 0.9094 on 12 and 191 DF,  p-value: 0.5386
```

Appendix Code

```
knitr::opts_chunk$set(echo=FALSE, error=FALSE, warning=FALSE, message=FALSE, fig.align = "center")
# Libraries

library(DataExplorer)
library(visdat)
library(dplyr)
library(tidyr)
library(MASS)
library(psych)
library(AER)
library(mlr)
library(mice)
```

```
library(imputeTS)
library(ggplot2)
library(caret)
library (skimr)
set.seed(621)

fires <- read.csv('https://raw.githubusercontent.com/hillt5/DATA_621/master/Final%20Project/uci-forest-

fires %>%  describe()
skimr::skim(fires)
area_1 <- fires %>%
  group_by(area) %>%
  summarize(area_0=n())
area_1
# unique months
fires %>% dplyr::select(month) %>% unique

# fires by month
fires_by_month <- fires %>%
  mutate(month=factor(month, levels = c('jan', 'feb', 'mar','apr','may','jun','jul','aug','sep','oct','
  group_by(month) %>%
  summarise(total_by_month = n())

fires_by_month %>%
  ggplot(aes(x=month, y=total_by_month)) +
  geom_col()
# unique days
fires %>% dplyr::select(day) %>% unique
# fires by day
fires_by_day <- fires %>%
  mutate(day=factor(day, levels = c('sun', 'mon', 'tue','wed','thu','fri','sat'))) %>%
  group_by(day) %>%
  summarise(total_by_day = n())

fires_by_day %>%
  ggplot(aes(x=day, y=total_by_day)) +
  geom_col()
plot_histogram(fires, geom_histogram_args = list("fill" = "tomato4"))
tibble(fires %>% summarize_all(n_distinct))
fire_corr <- fires[-c(3,4)]
corrplot::corrplot(cor(fire_corr), type = 'lower')
ggplot(fires,aes(rain,area)) + geom_point() +stat_smooth(method="lm")
ggplot(fires,aes(wind,area)) + geom_point() +stat_smooth(method="lm")
ggplot(fires,aes(RH,area)) + geom_point() +stat_smooth(method="lm")
ggplot(fires,aes(temp,area)) + geom_point() +stat_smooth(method="lm")
ggplot(fires,aes(ISI,area)) + geom_point() +stat_smooth(method="lm")
ggplot(fires,aes(DC,area)) + geom_point() +stat_smooth(method="lm")
ggplot(fires,aes(DMC,area)) + geom_point() +stat_smooth(method="lm")
ggplot(fires,aes(FFMC,area)) + geom_point() +stat_smooth(method="lm")
ggplot(fires,aes(Y,area)) + geom_point() +stat_smooth(method="lm")
ggplot(fires,aes(X,area)) + geom_point() +stat_smooth(method="lm")

fires %>%
```

```
  dplyr::select(-month, -day,-area) %>%
  pivot_longer(everything(), names_to = 'Var', values_to='Value') %>%
  ggplot(aes(x = Var, y = Value)) +
  geom_boxplot() +
  coord_flip()
fires %>% group_by(month) %>% summarize(n())

plot_missing(fires %>% dplyr::select(-area))
head(fires)
set.seed(317)

# Training set
fires.clean <- mice(data.frame(fires), method = 'rf', m=2, maxit = 2, print=FALSE)
fires.clean <- complete(fires.clean)

nzv_preds <- nearZeroVar(fires.clean)
fires.clean <- fires.clean[,-nzv_preds]

head(fires.clean)
set.seed(317)

# for month
dum.month <- dummyVars(area ~ month, data = fires.clean,fullRank=T)
dum.mon.predict <- predict(dum.month, fires.clean)
fires.clean <- cbind(dum.mon.predict, fires.clean) %>% dplyr::select(-month)

# for day
dum.day <- dummyVars(area ~ day, data = fires.clean,fullRank=T)
dum.day.predict <- predict(dum.day, fires.clean)
fires.clean <- cbind(dum.day.predict, fires.clean) %>% dplyr::select(-day)

head(fires.clean)
set.seed(317)
preproc_traindf <- preProcess(fires.clean, method = "YeoJohnson")
fires.clean <- predict(preproc_traindf, fires.clean)

head(fires.clean)
set.seed(317)

partition <- createDataPartition(fires.clean$area, p=0.75, list = FALSE)

## training/test partition for independent variables
#X.train <- fires.clean[partition, ] %>% dplyr::select(-area)
#X.test <- fires.clean[-partition, ] %>% dplyr::select(-area)
#
## training/test partition for dependent variable area
#y.train <- fires.clean$PH[partition]
#y.test <- fires.clean$PH[-partition]

# training/test partition for independent variables
X.train <- fires.clean[partition, ]
X.test <- fires.clean[-partition, ]
```

```
lm1 <- lm(X.train,formula=area ~.)
summary(lm1)
small_fires <- X.train[X.train$area < 2,]
small_lm1 <- glm.nb(small_fires,formula = area ~.)
summary(small_lm1)
fires_over0 <- X.train[X.train$area > 0,]
lm0 <- lm(area ~ .,data=fires_over0)
summary(lm0)

library(leaps)
regsubsets.out <-
    regsubsets(area~.,
               data =X.train,
               nbest = 1,       # 1 best model for each number of predictors
               nvmax = NULL,    # NULL for no limit on number of variables
               force.in = NULL, force.out = NULL,
               method = "exhaustive")

summary.out <- summary(regsubsets.out)
as.data.frame(summary.out$outmat)
which.max(summary.out$adjr2)
summary.out$which[13,]
best.model <- lm(area ~ day.thu + month.aug + month.dec + month.jan + month.jul+ month.jun+month.mar+mo
summary(best.model)
rr.bisquare <- rlm(area~day.thu + month.aug + month.dec + month.jan + month.jul+ month.jun+month.mar+mo
summary(rr.bisquare)
summary(best.model)$sigma
summary(rr.bisquare)$sigma
regsubsets.out2 <-
    regsubsets(area~.,
               data =fires_over0,
               nbest = 1,       # 1 best model for each number of predictors
               nvmax = NULL,    # NULL for no limit on number of variables
               force.in = NULL, force.out = NULL,
               method = "exhaustive")

summary.out2 <- summary(regsubsets.out2)
as.data.frame(summary.out$outmat)
which.max(summary.out2$adjr2)
summary.out$which[12,]
lm0edit<- lm(area~day.thu + month.aug + month.dec + month.feb + month.jul+ month.jun+month.mar+month.oc
summary(lm0edit)
library(glmnet)
x = data.matrix(fires_over0[,c('day.thu', 'month.aug', 'month.dec','month.feb','month.jul', 'month.jun'

y_train = fires_over0$area # highlighting the dependent variable

lambdas <- 10^seq(2, -3, by = -.1)
ridge_reg = glmnet(x, y_train, nlambda = 25, alpha = 0, lambda = lambdas)

summary(ridge_reg)
# Using cross validation glmnet
cv_ridge <- cv.glmnet(x, y_train, alpha = 0, lambda = lambdas)
```

```r
optimal_lambda <- cv_ridge$lambda.min
optimal_lambda
# Compute R^2 from true and predicted values
eval_results <- function(true, predicted, df) {
  SSE <- sum((predicted - true)^2)
  SST <- sum((true - mean(true))^2)
  R_square <- 1 - SSE / SST
  RMSE = sqrt(SSE/nrow(df))


  # Model performance metrics
data.frame(
  RMSE = RMSE,
  Rsquare = R_square
)

}

# Prediction and evaluation on train data
predictions_train <- predict(ridge_reg, s = optimal_lambda, newx = x)
eval_results(y_train, predictions_train, X.train)

res0 <- resid(lm0edit)
plot(density(res0))
qqnorm(res0)
qqline(res0)
predictions <- predict(lm0edit,X.test)
head(predictions)
summary(lm0edit)
```