Project 2 - DS-600

Introduction

Every year, there are many forest fires occur worldwide. To protect the natural wildlife, numerous attempts has been made to prevent forests from getting fire. In this project, we will be outlining a summary in about a forest fire that occurred in a region at Portugal. The data is collected by using metheorological and other data. For more information about the dataset, this link can be followed.

The following is the dataset information.

```
# reading
forestfires = read.csv('forestfires.csv')
```

Data Cleaning

With the dataset, we will first be investigating the variables' data type and general shape of occurance. We will be examining data in terms of the following objectives:

1. Variable data types
2. Empty or NA values

Below, we are presenting the general outline of the dataset to understand the structure of our dataset. To use the some functions, we have to install the library "dplyr" first.

```
head(forestfires)

##   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   NA 35.4 669.1  6.7 18.0 33  0.9  0.0    0
## 3 7 4   oct sat   NA 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

class(forestfires)

## [1] "data.frame"

dim(forestfires)

## [1] 517  13

names(forestfires)

##  [1] "X"     "Y"     "month" "day"   "FFMC"  "DMC"   "DC"    "ISI"
##  [9] "temp"  "RH"    "wind"  "rain"  "area"
```

```
glimpse(forestfires)
```

```
## Observations: 517
## Variables: 13
## $ X     <int> 7, 7, 7, 8, 8, 8, 8, 8, 8, 7, 7, 7, 6, 6, 6, 6, 5, 8, 6,...
## $ Y     <int> 5, 4, 4, 6, 6, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4,...
## $ month <fct> mar, oct, oct, mar, mar, aug, aug, aug, sep, sep, sep, s...
## $ day   <fct> fri, tue, sat, fri, sun, sun, mon, mon, tue, sat, sat, s...
## $ FFMC  <dbl> 86.2, NA, NA, 91.7, 89.3, 92.3, 92.3, 91.5, 91.0, 92.5, ...
## $ DMC   <dbl> 26.2, 35.4, 43.7, 33.3, 51.3, 85.3, 88.9, 145.4, 129.5, ...
## $ DC    <dbl> 94.3, 669.1, 686.9, 77.5, 102.2, 488.0, 495.6, 608.2, 69...
## $ ISI   <dbl> 5.1, 6.7, 6.7, 9.0, 9.6, 14.7, 8.5, 10.7, 7.0, 7.1, 7.1,...
## $ temp  <dbl> 8.2, 18.0, 14.6, 8.3, 11.4, 22.2, 24.1, 8.0, 13.1, 22.8,...
## $ RH    <int> 51, 33, 33, 97, 99, 29, 27, 86, 63, 40, 51, 38, 72, 42, ...
## $ wind  <dbl> 6.7, 0.9, 1.3, 4.0, 1.8, 5.4, 3.1, 2.2, 5.4, 4.0, 7.2, 4...
## $ rain  <dbl> 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0...
## $ area  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

```
summary(forestfires)
```

```
##        X               Y            month          day          FFMC
##  Min.   :1.000   Min.   :2.0   aug    :184   fri:85   Min.   :18.70
##  1st Qu.:3.000   1st Qu.:4.0   sep    :172   mon:74   1st Qu.:90.20
##  Median :4.000   Median :4.0   mar    : 54   sat:84   Median :91.65
##  Mean   :4.669   Mean   :4.3   jul    : 32   sun:95   Mean   :90.65
##  3rd Qu.:7.000   3rd Qu.:5.0   feb    : 20   thu:61   3rd Qu.:92.90
##  Max.   :9.000   Max.   :9.0   jun    : 17   tue:64   Max.   :96.20
##                                (Other): 38   wed:54   NA's   :7
##       DMC              DC             ISI              temp
##  Min.   :  1.1   Min.   :  7.9   Min.   : 0.000   Min.   : 2.20
##  1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500   1st Qu.:15.50
##  Median :108.3   Median :664.2   Median : 8.400   Median :19.30
##  Mean   :110.9   Mean   :547.9   Mean   : 9.022   Mean   :18.89
##  3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800   3rd Qu.:22.80
##  Max.   :291.3   Max.   :860.6   Max.   :56.100   Max.   :33.30
##
##        RH              wind            rain              area
##  Min.   : 15.00   Min.   :0.400   Min.   :0.00000   Min.   :   0.00
##  1st Qu.: 33.00   1st Qu.:2.700   1st Qu.:0.00000   1st Qu.:   0.00
##  Median : 42.00   Median :4.000   Median :0.00000   Median :   0.52
##  Mean   : 44.29   Mean   :4.018   Mean   :0.02166   Mean   :  12.85
##  3rd Qu.: 53.00   3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.:   6.57
##  Max.   :100.00   Max.   :9.400   Max.   :6.40000   Max.   :1090.84
##
```

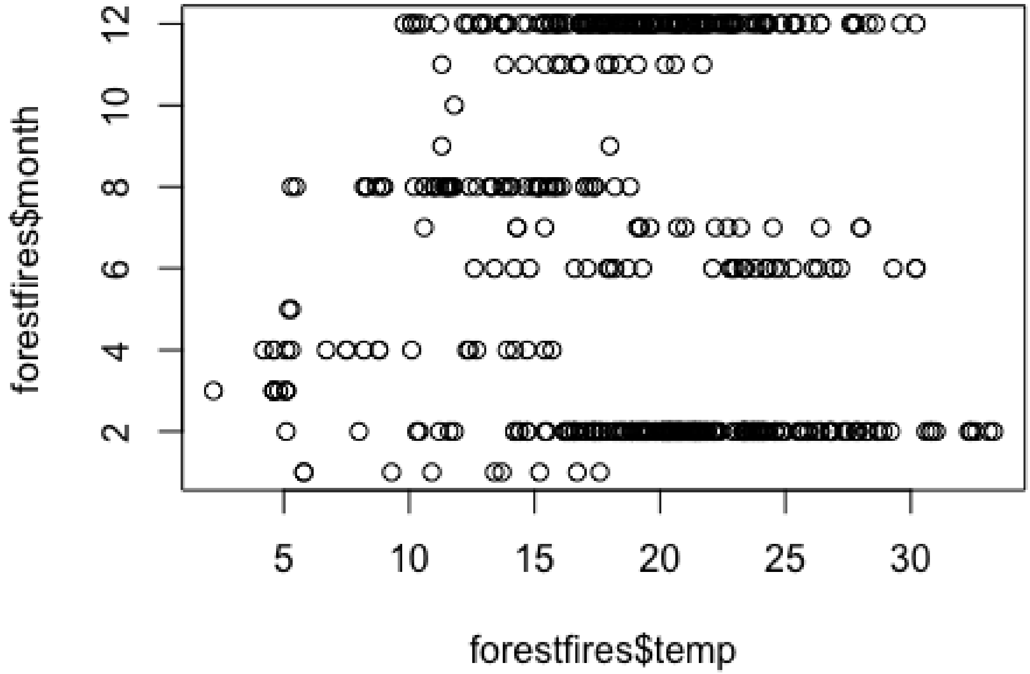
We observed that FFMC variable has some missing values. Therefore, we decided to update values with replacing NAs with mean of it.

```
# finding indices of na
ind = which(is.na(forestfires$FFMC))
```

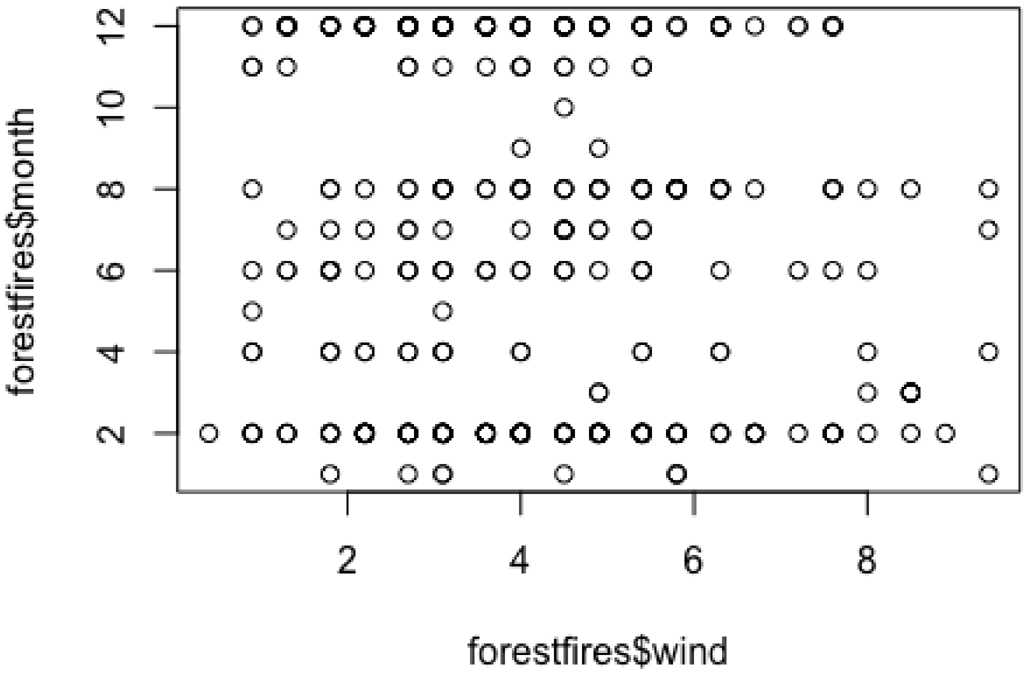
```
#replacing empty data with 0
forestfires$FFMC[ind]=0

#replacing empty data with 0
forestfires$FFMC[ind]=mean(forestfires$FFMC)

# removing outliers
forestfires <- filter(forestfires, area < 150)
```

Now, we want to see how is the behavior of the weather through the time. Let's take a look what are the values of some variables, as temperature, humidity, wind and rain in the months of this year. Doing this, we can know more about the weather of this zone of Portugal.

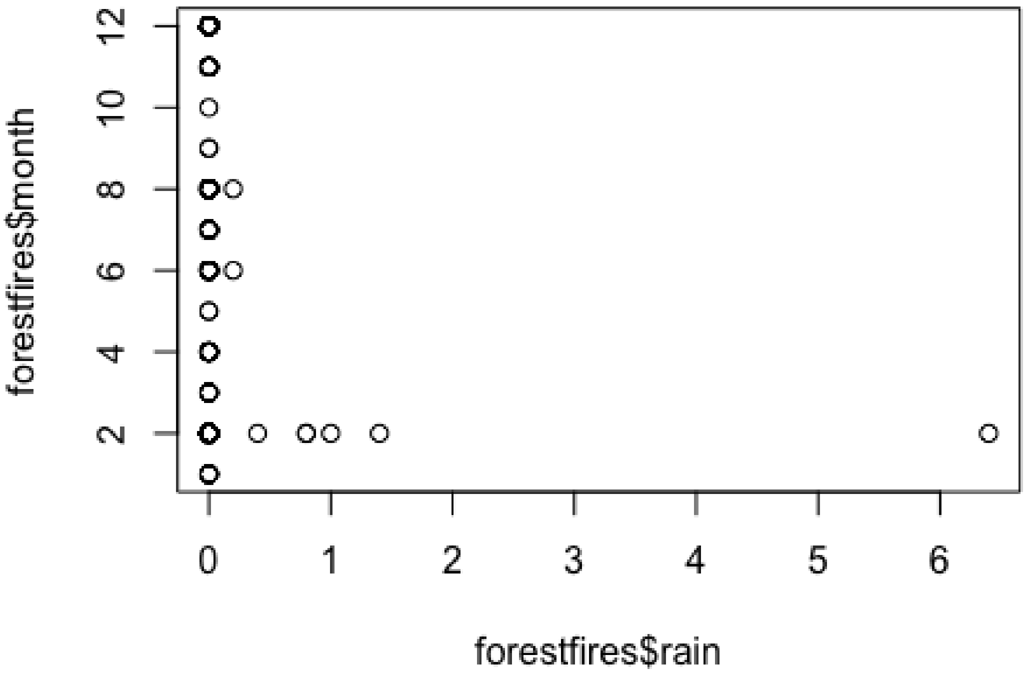
```
# making plots of temperature in Celsius, wind in km/h, rain in mm/m2, and
humidty (RH) in %
plot(forestfires$temp, forestfires$month)
```



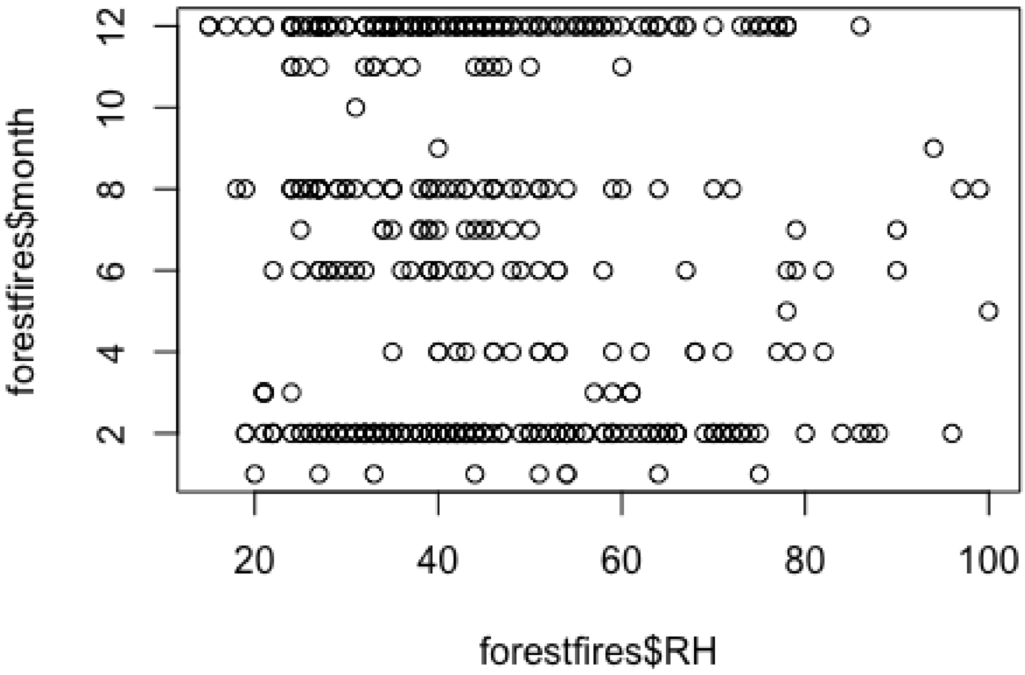
```
plot(forestfires$wind, forestfires$month)
```



```r
plot(forestfires$rain, forestfires$month)
```

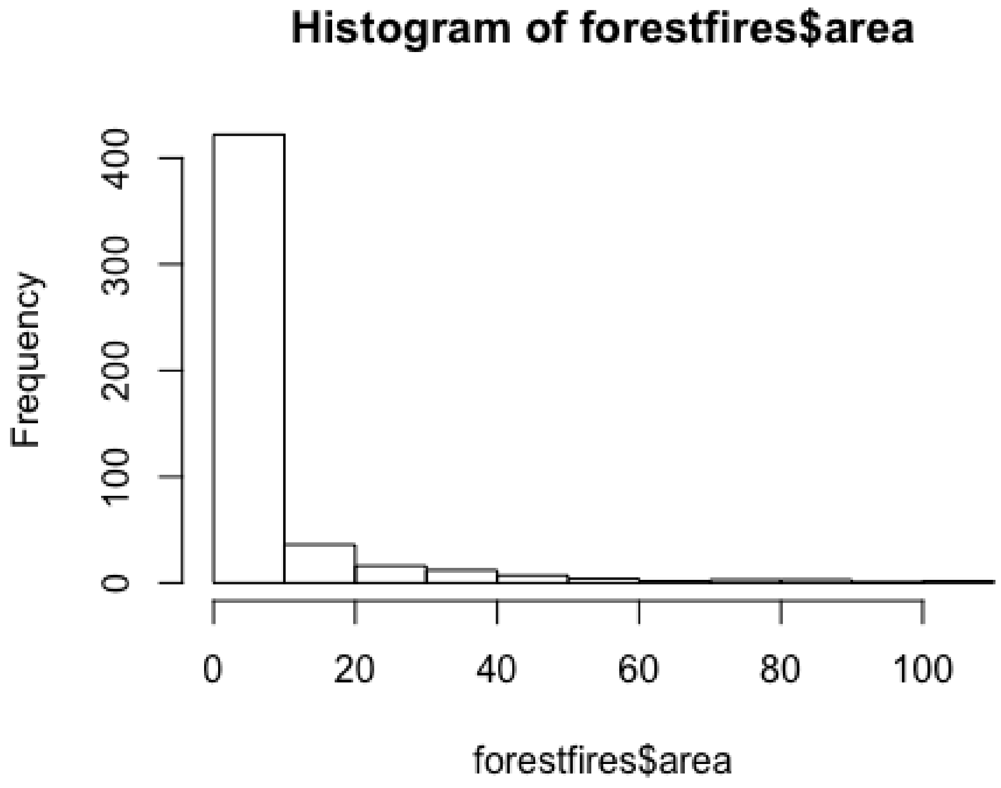


```
plot(forestfires$RH, forestfires$month)
```



Another way to see the variables is through the histograms. Let's see how is the behavior of some variables, as burned area of the forest.

```
# making histogram of burned area of the forest
hist(forestfires$area)
```



Gathering the Information

The data presents the events of fires in x, y spatial coordinates within the Montesinho map. Other columns represents the features of each observed variable. For the sake of analysis, we may need to do some operations within the data. The following will present some manipulation examples that may occur during data analysis.

```
# unite coordinates
forestfires_t1 <- unite(forestfires, "coordinates", c(X, Y))

# changing data
forestfires_t2 <- mutate(forestfires_t1, coordinates = paste("coor",
coordinates, sep = "_"))

# filtering the data with some parameters
forestfires_t3 <- filter(forestfires_t2, month == "mar", day == "fri")

# summarize all the columns with mean
summarise_if(forestfires_t3, is.numeric, funs(round(mean(., na.rm = T), 2)))

##    FFMC   DMC    DC  ISI temp    RH wind rain area
## 1 90.01 34.46 81.19 8.36   14 38.36 5.86 0.02 0.99
```

```
# doing multiple things at once
forestfires %>%
  mutate(ISIandTemp = ISI * temp, rain = exp(rain), area = log(2)) %>%
  select(X:day,ISIandTemp, ISI, temp, area) %>%
  arrange(desc(ISIandTemp)) %>%
  filter(temp > 32)
```

```
##   X Y month day ISIandTemp  ISI temp      area
## 1 4 5   aug mon     547.68 16.8 32.6 0.6931472
## 2 6 5   aug tue     476.19 14.3 33.3 0.6931472
## 3 2 5   aug sun     466.71 14.1 33.1 0.6931472
## 4 3 4   aug tue     461.89 14.3 32.3 0.6931472
## 5 4 4   aug thu     447.12 13.8 32.4 0.6931472
## 6 1 3   aug fri     366.12 11.3 32.4 0.6931472
```

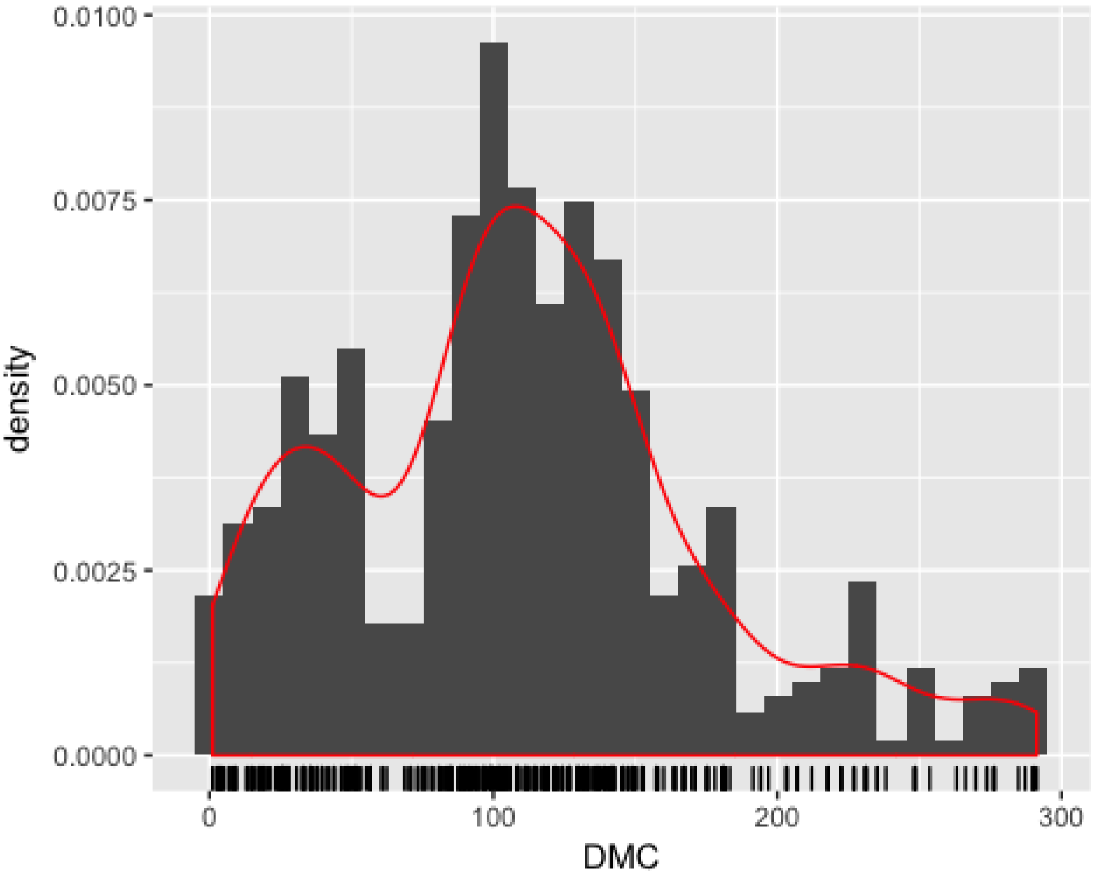
```
# grouping the data
forestfires %>%
  group_by(month) %>%
  summarise(
    n = n(),
    RainTotal = sum(rain),
    WindAverage = mean(wind),
    areaTotal = sum(area)
  )
```

```
## # A tibble: 12 x 5
##    month     n RainTotal WindAverage areaTotal
##    <fct> <int>     <dbl>       <dbl>     <dbl>
##  1 apr       9     0            4.67      80.0
##  2 aug     180    10.8          4.08     995
##  3 dec       9     0            7.64     120
##  4 feb      20     0            3.76     126
##  5 jan       2     0            2.00       0
##  6 jul      31     0.200        3.70     181
##  7 jun      17     0            4.14      99.3
##  8 mar      54     0.200        4.97     235
##  9 may       2     0            4.45      38.5
## 10 nov       1     0            4.50       0
## 11 oct      15     0            3.46      99.6
## 12 sep     168     0            3.58    1427
```

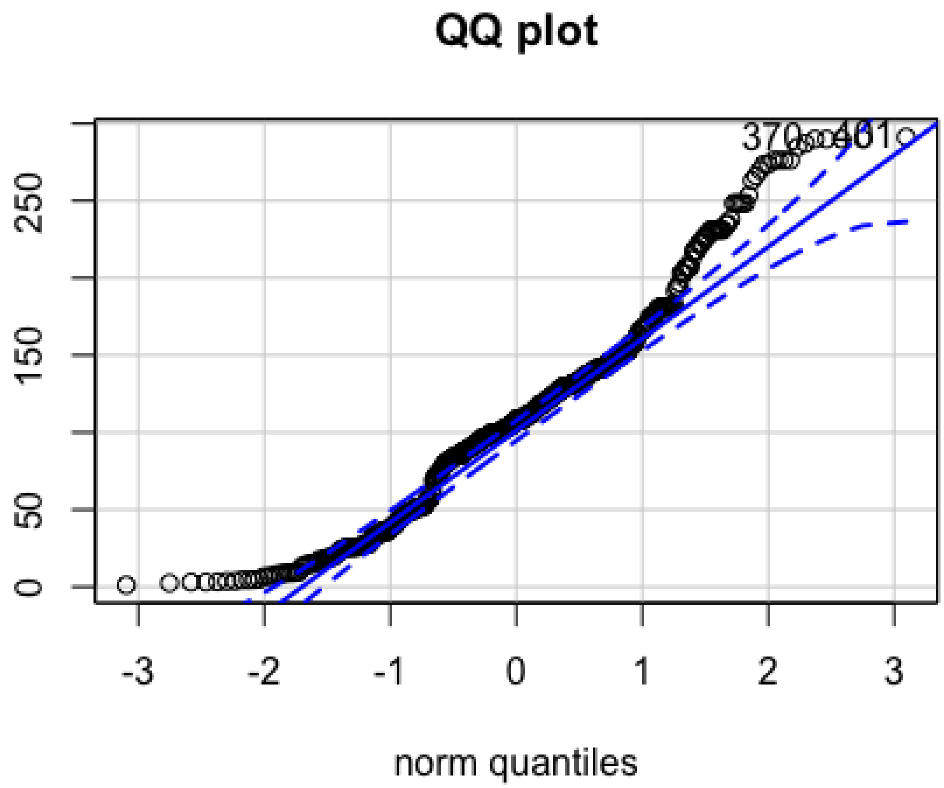
Using GGPlot

```
ggplot(forestfires, aes(x=DMC)) +
  geom_histogram(aes(y = ..density..)) +
  geom_density(color = "red") +
  geom_rug()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



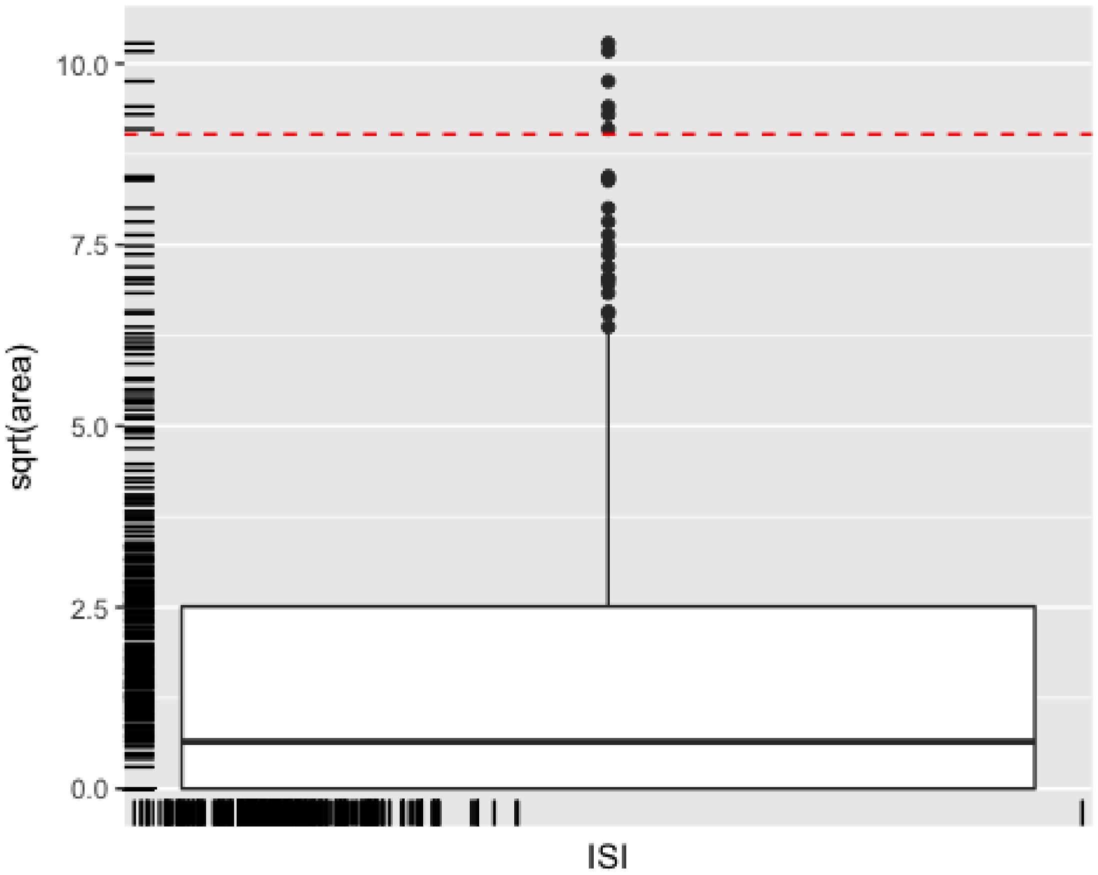
```
qqPlot(forestfires$DMC, main = "QQ plot", ylab = "")
```



```
## [1] 401 370
```

```
ggplot(forestfires, aes(x=ISI, y=sqrt(area))) +
  geom_boxplot() +
  geom_rug() +
  geom_hline(aes(yintercept=mean(forestfires$ISI, na.rm = T)), linetype = 2,
color = "red") +
  scale_x_discrete(breaks = NULL)
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



Model Assesment

Up to this point, we analyzed our dataset of forestfires in terms of NA values and data structures. Everything we made to this point was data cleaning and tidying the dataset so as to do analysis on the data.

Here, we are presenting our analysis on the dataset. We made linear analysis and examine feature’s relationships within the features and the with the predictor variable.

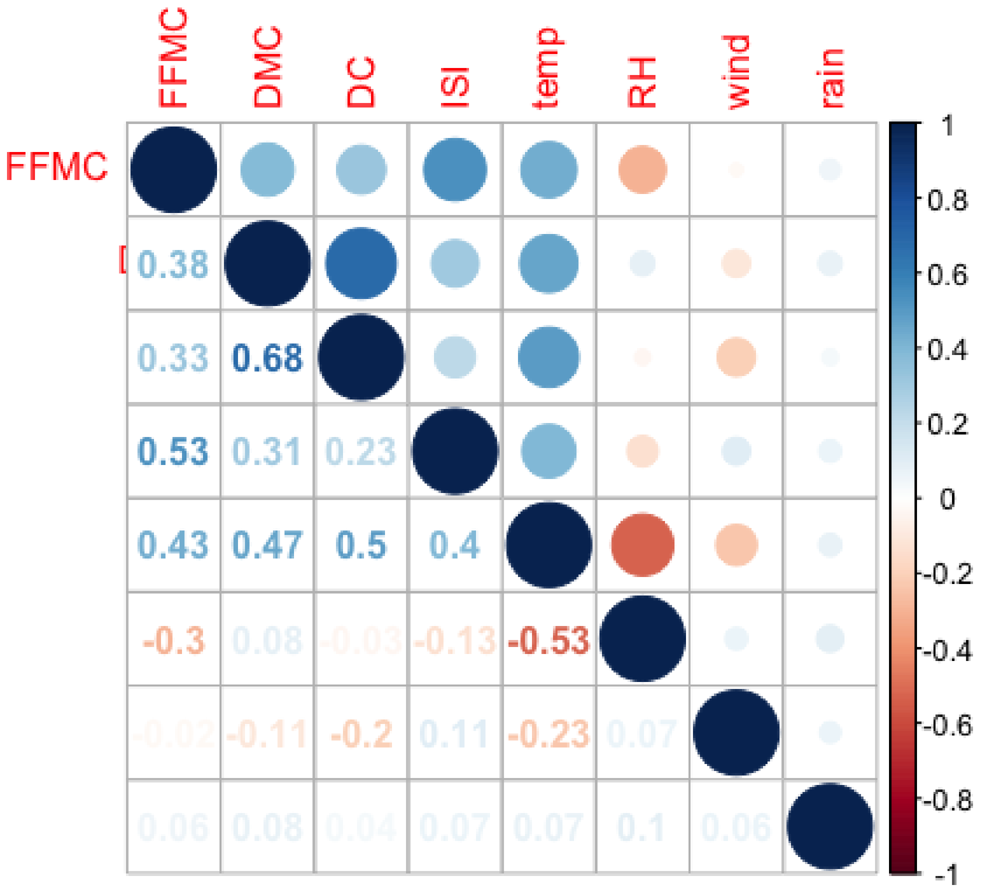
Correlation between features

```
symnum(cor(forestfires[5:12], use = "complete.obs"))
```

```
##      F DM DC I t R w r
## FFMC 1
## DMC  . 1
## DC   . ,  1
## ISI  . .     1
## temp . .  .  . 1
## RH              . 1
## wind              1
## rain                1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

Correlation Plot

```
cm <- cor(forestfires[,5:12], use="complete.obs")
corrplot(cm, type="upper", tl.pose="d")

## Warning in text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt
## = tl.srt, : "tl.pose" is not a graphical parameter

## Warning in text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col
## = tl.col, : "tl.pose" is not a graphical parameter

## Warning in title(title, ...): "tl.pose" is not a graphical parameter

corrplot(cm,add=TRUE, type="lower", method="number",diag=FALSE, tl.pos="n",
cl.pos="n")
```



###

Dividing the dataset

Here, we are preparing our dataset for out-of-sample analysis. We are dividing the dataset into two parts, training and testing datasets for future analysis.

```
# removing outliers
forestfires_test <- forestfires[c(401:508), ]
forestfires <- forestfires[c(1:400), ]
```

Linear Model

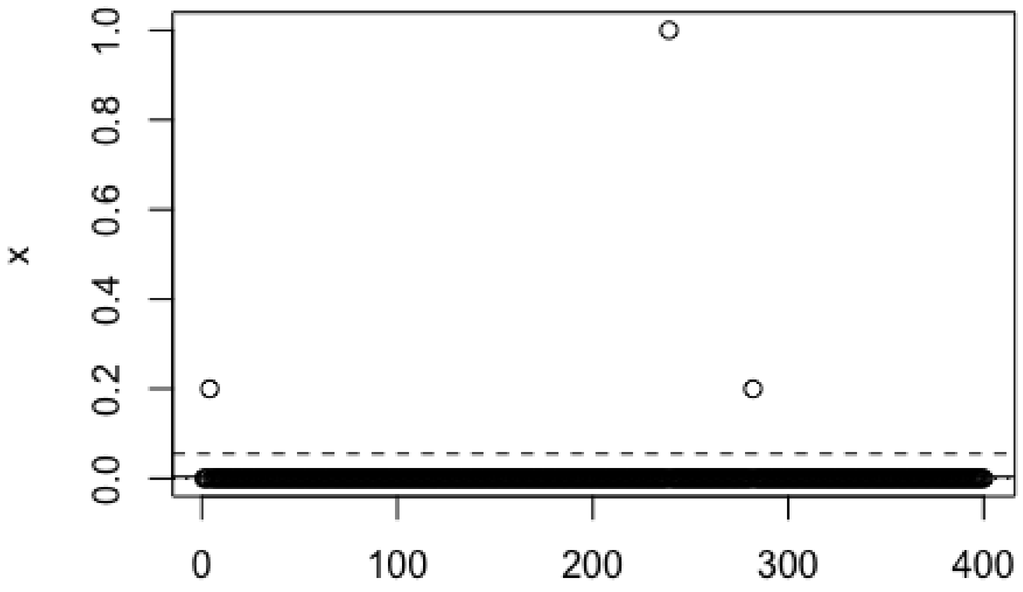
Implementations of models are represented in below.

```
fit.m0 <- lm(area ~ ., forestfires)
fit.m1 <- lm(area ~ day + DMC + temp + RH, forestfires)
fit.m2 <- lm(area ~ month + day + DC + temp, forestfires)
summary(fit.m2)

## 
## Call:
## lm(formula = area ~ month + day + DC + temp, data = forestfires)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -12.593  -7.053  -3.926   0.145  92.077 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.35445    6.77950  -0.200   0.8418  
## monthaug     4.42181   10.82910   0.408   0.6833  
## monthdec    15.40462    9.19492   1.675   0.0947 .
## monthfeb     4.85361    7.38139   0.658   0.5112  
## monthjan    -2.01206   12.40329  -0.162   0.8712  
## monthjul     5.46573    9.25313   0.591   0.5551  
## monthjun    -2.08078    8.54205  -0.244   0.8077  
## monthmar     0.93991    6.58904   0.143   0.8866  
## monthmay    -2.80110   16.37175  -0.171   0.8642  
## monthoct     8.10454   12.03876   0.673   0.5012  
## monthsep    10.58704   12.16251   0.870   0.3846  
## daymon       1.68135    2.67983   0.627   0.5308  
## daysat       2.03481    2.69789   0.754   0.4512  
## daysun       3.56401    2.56463   1.390   0.1654  
## daythu       0.31122    2.94353   0.106   0.9159  
## daytue       4.55673    2.86511   1.590   0.1126  
## daywed       1.42290    3.03022   0.470   0.6389  
## DC          -0.01116    0.01512  -0.738   0.4611  
## temp         0.30005    0.19768   1.518   0.1299  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 15.03 on 381 degrees of freedom
## Multiple R-squared:  0.0466, Adjusted R-squared:  0.001559 
## F-statistic: 1.035 on 18 and 381 DF,  p-value: 0.4193
```

Distribution of the Values of Columns

```
plot1 = function(x) {plot(x,xlab="")
  abline(h=mean(x,na.rm=T),lty=1)
  abline(h=mean(x,na.rm=T)+sd(x,na.rm=T),lty=2)
  abline(h=median(x,na.rm=T),lty=3)}
```

```
plot1(forestfires$rain)
```



Anova Analysis

After fitting a linear model, to understand model's success, we do anova analysis to the model's variance.

```
anova(fit.m2)

## Analysis of Variance Table
##
## Response: area
##            Df Sum Sq Mean Sq F value Pr(>F)
## month      10   2624  262.40  1.1609 0.3160
## day         6    920  153.41  0.6787 0.6670
## DC          1    144  144.29  0.6384 0.4248
## temp        1    521  520.74  2.3038 0.1299
## Residuals 381  86120  226.04
```

To find the best model within two models, we are planing to do anova analysis to both models. The result of analysis on variances will indicate which model will better in terms of variance.

```
anova(fit.m1, fit.m2)

## Analysis of Variance Table
##
## Model 1: area ~ day + DMC + temp + RH
## Model 2: area ~ month + day + DC + temp
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    390 88177
## 2    381 86120  9    2057.6 1.0114 0.4301
```

Here, we see that applying each model an anova analysis will take more computational time. Therefore, we will be using STEP function in order to find the best model. Below, we are presenting that we are improving the model with an AIC score of 2776.82 to 2759.06. our final model in below function indicates that the following is the best model that we should use, in order to predict area.

- formula = area ~ ISI + temp + wind

```
step(fit.m0, trace = F)

##
## Call:
## lm(formula = area ~ X + month + DMC + DC, data = forestfires)
##
## Coefficients:
## (Intercept)            X     monthaug     monthdec     monthfeb
##    -0.21571      0.67922     16.63969     24.19850      6.15527
##    monthjan     monthjul     monthjun     monthmar     monthmay
##     2.15854     13.73631      1.43427      1.51780     -1.41558
##    monthoct     monthsep          DMC           DC
##    30.50263     28.48066      0.10288     -0.04679
```

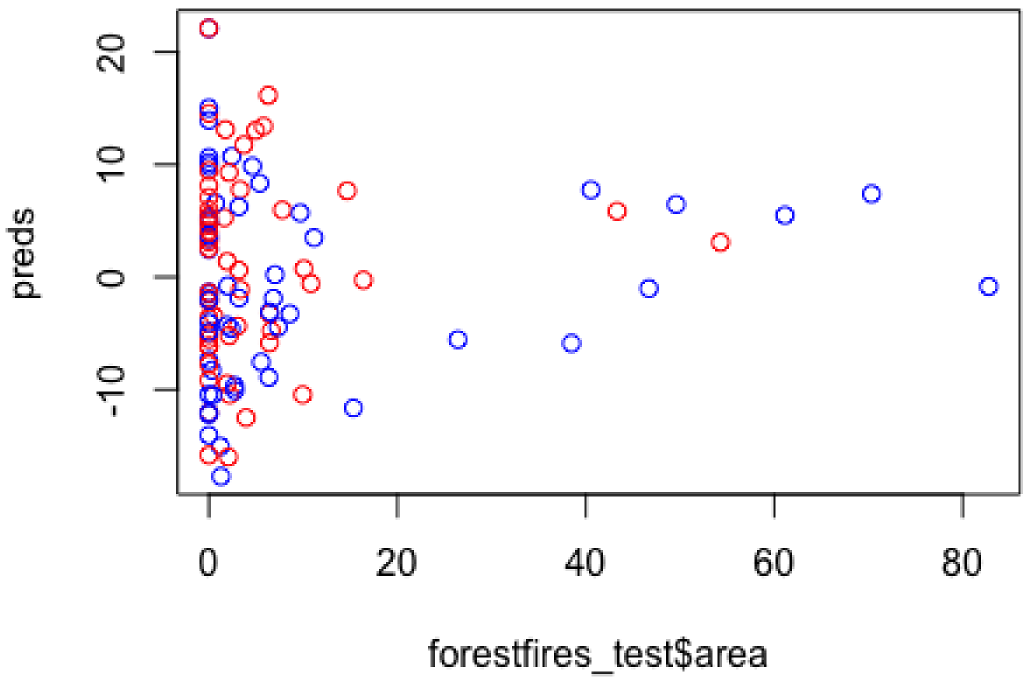
Predicting with Test Dataset using Model

The best model we acquired from step function is used to create the final linear model. We are using the final model to assess the success of our model with the testing dataset that we created above.

```
# final model
fit.m99 <- lm(formula = area ~ ISI + temp + wind, data = forestfires)

# prediction with final model in test dataset
preds <- predict(fit.m99, forestfires_test[-1], probability = T)

# summary
summary(fit.m99)

##
## Call:
## lm(formula = area ~ ISI + temp + wind, data = forestfires)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.147  -6.826  -4.976   0.027  96.775
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7829     3.5598   0.501   0.6168
## ISI          -0.3041     0.1821  -1.670   0.0957 .
## temp          0.2947     0.1577   1.869   0.0623 .
## wind          0.5371     0.4436   1.211   0.2267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.01 on 396 degrees of freedom
## Multiple R-squared:  0.01208,	Adjusted R-squared:  0.004599
## F-statistic: 1.615 on 3 and 396 DF,  p-value: 0.1854
```

According to the linear model, we expect the model's residual distribution to be normally distributed. Below, we are presenting the models' predicted values with testing dataset's 'area'.



The color red is representing the real data and the color blue, predicted data. Because the majority of the data is around zero, meaning that there are small fires in terms of area, that indicates our model represents a good model indicator.

Conclusion

The data after cleaning operations is ready for statistical analysis. After operations of tidying, restructing and removing outliers, we will have the data that is ready for further analysis.

After analysis on the report, we evaluate the model in terms of correlation between features. After, we run a linear model to see the results. Above, we represent our findings and plot features according to predictors.