STAT151a Final Project

Amanda Wu

Apr 28, 2018

Source file ⇒ stat151a_final_project.Rmd (data:text/x-markdown;base64,LS0tCnRpdGxlOiAiU1RBVDE1MWEgRmluYWwgUHJvamVjdCIKYXV0aG9yOiAiQW1hbmRhIFd1IgpkYXRlOiAiQXByIDI4LCAyMDE4IgpvdXRwdXQ6I

Loading data

The dataset contains information on forest fire data from the Montesinho natural park in northeast Portugal. The data were collected from January 2000 to December 2003 and there are 517 observations of 13 variables,

```
## reading data into R
data = read.table("~/Desktop/forestfires.csv", header=TRUE, sep=",")
## double checking to see if the variables match with the descriptions given to us
head(data)
```

```
##   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
```

```
area = data$area

## Check to see if any missing data points are spotted
## and just get a general view of the data i'm working with
summary(data)
```

```
##        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 :91.60  
##  Mean   :4.669   Mean   :4.3   jul    : 32   sun:95   Mean   :90.64  
##  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                  
##       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  
## 
```

Explanatory Data Analysis

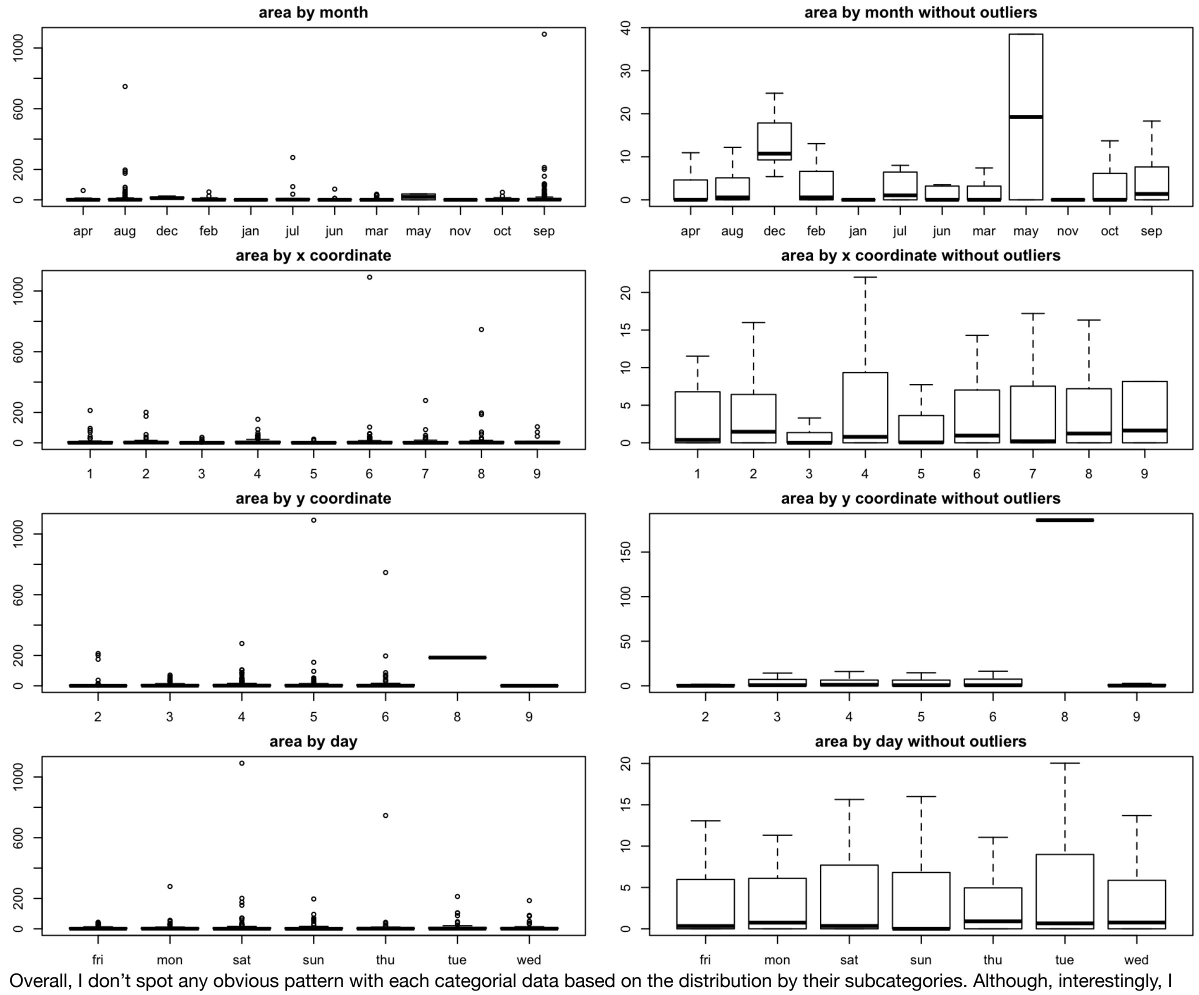
Before jumping into the analysis, I want to understand all the variables graphically. Specifically, I want to understanding the distribution of all the independent variables (predictors), as well as their relationships to each other (such as their correlations), which lead me to do a univariate analysis and a bivariate analysis.

The following plots are drawn to help me visualize their bahavior:
1) Box plot: to help me check for any outlier observations
2) Density plot: to help me see the distribution of the variable, and ideally I would prefer a bell shaped curve.
3) Scatter plot (correlation plot): to help me visualize the linear relationship between the predictor and the response as well as whether the covariates are collinear with each other.

Univariate Analysis

Categorical Data

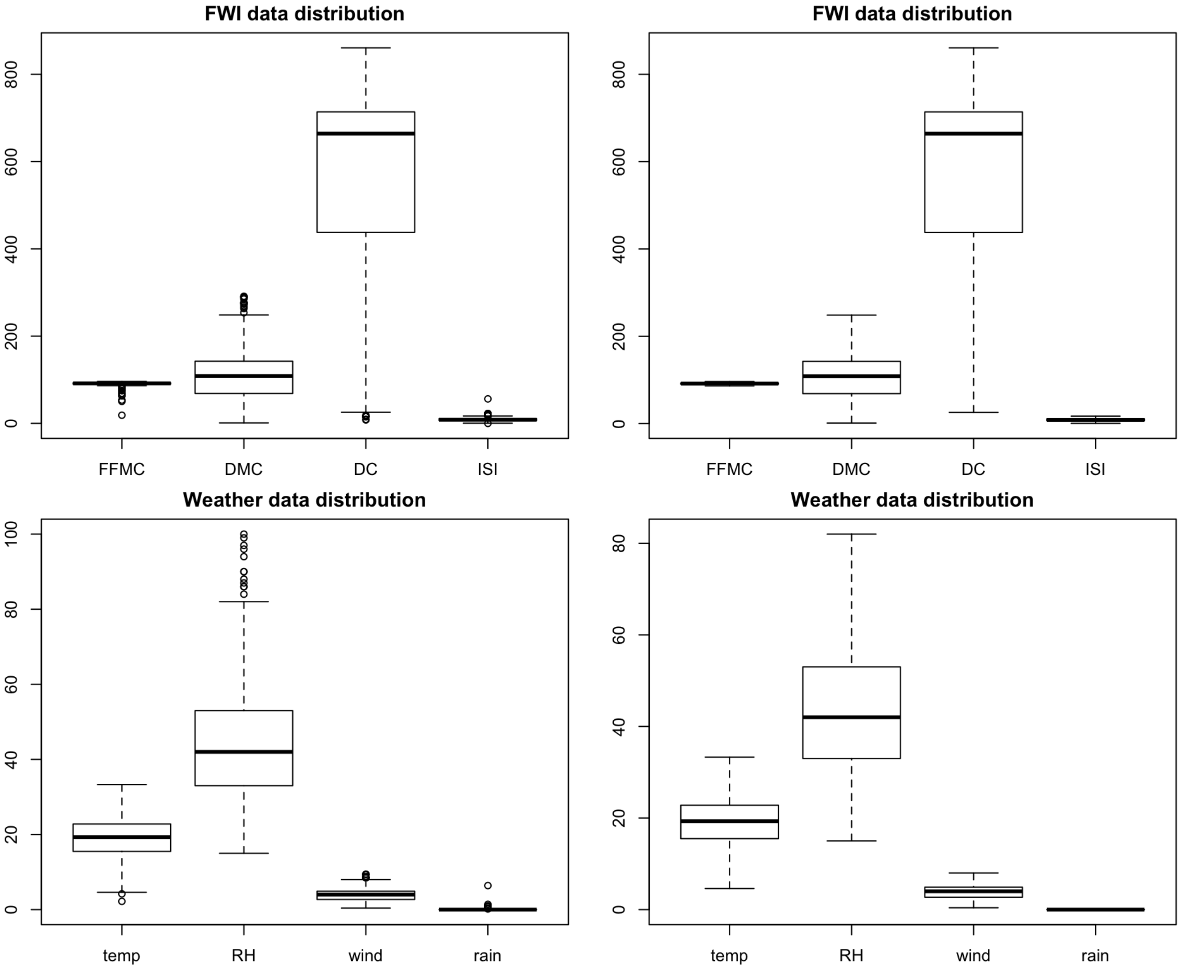
```
par(mfrow = c(4, 2))
par(mar = c(2,2,2,2))
## All the variables in this chunk are treated as categorical by me
## from the boxplots generated on the left, I spot many outliers that squish the data to the baseline
## I then got rid of the outliers to see if there's any obvious pattern to the data on the right

boxplot(area ~ data$month, data = data, main ="area by month", xlab = "Month", ylab = "area")
boxplot(area ~ data$month, data = data, main ="area by month without outliers", xlab = "Month", ylab = "area", ou
tline = FALSE)


boxplot(area ~ data$X, data = data, main ="area by x coordinate", xlab = "x coordinate", ylab = "area")
boxplot(area ~ data$X, data = data, main ="area by x coordinate without outliers", xlab = "x coordinate", ylab =
"area", outline = FALSE)


boxplot(area ~ data$Y, data = data, main ="area by y coordinate", xlab = "y coordinate", ylab = "area")
boxplot(area ~ data$Y, data = data, main ="area by y coordinate without outliers", xlab = "y coordinate", ylab =
"area", outline = FALSE)


boxplot(area ~ data$day, data = data, main ="area by day", xlab = "day", ylab = "area")
boxplot(area ~ data$day, data = data, main ="area by day without outliers", xlab = "day", ylab = "area", outline
= FALSE)
```



Overall, I don't spot any obvious pattern with each categorial data based on the distribution by their subcategories. Although, interestingly, I noticed that the fire area centered around y coordinate = 8 is significantly larger than the rest of the data points. Generally speaking, I think this is good for now, if I need more information, I will come back later.

More boxplots side by side

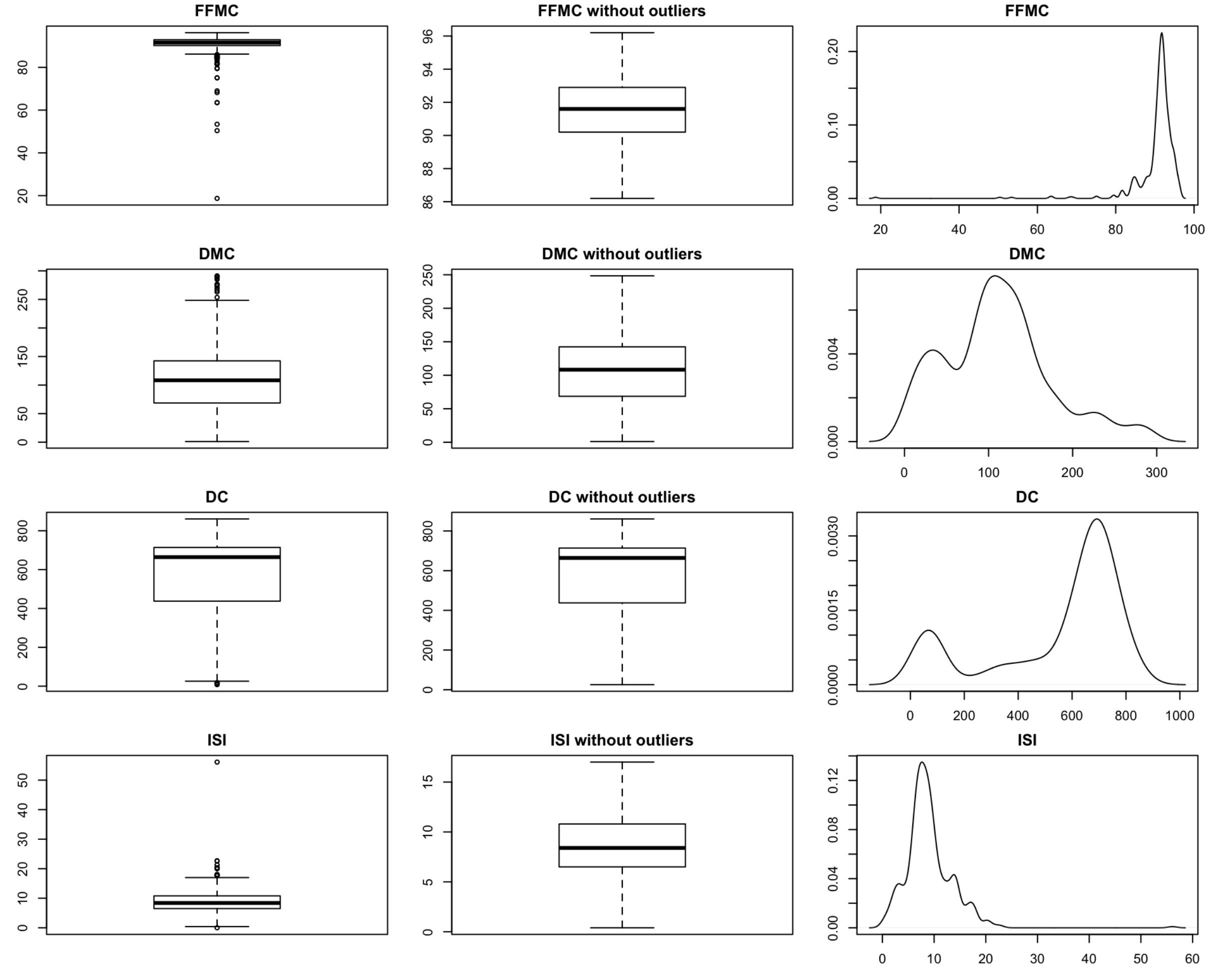
```
par(mfrow = c(2, 2))
par(mar = c(2,2,2,2))

## Now, lets move on to our continuous variables
fwi = data[,c(5:8)]
weather = data[, c(9:12)]

boxplot(fwi, main = "FWI data distribution")
boxplot(fwi, main = "FWI data distribution", outline = FALSE)
boxplot(weather, main = "Weather data distribution")
boxplot(weather, main = "Weather data distribution", outline = FALSE)
```



The boxplot distributions shown above are very interesting. First of all I immediately spot many outliers in all of the variables here, which I will go into detail later. And even after I got rid of all the outliers, the the spread of DC is still very wide with DMC being the second widest within the FWI dataset, while the spread of FFMC and ISI stay very concentrated to its median. I might have to think about the transformation of data based on what I observe later one by one. Next, similarly, the weather data also show a wide spread in RH(humidity) and relatively wide spread in temperature, while wind and rain (exp rain) are concentrated towards their median.

FWI Data in detail

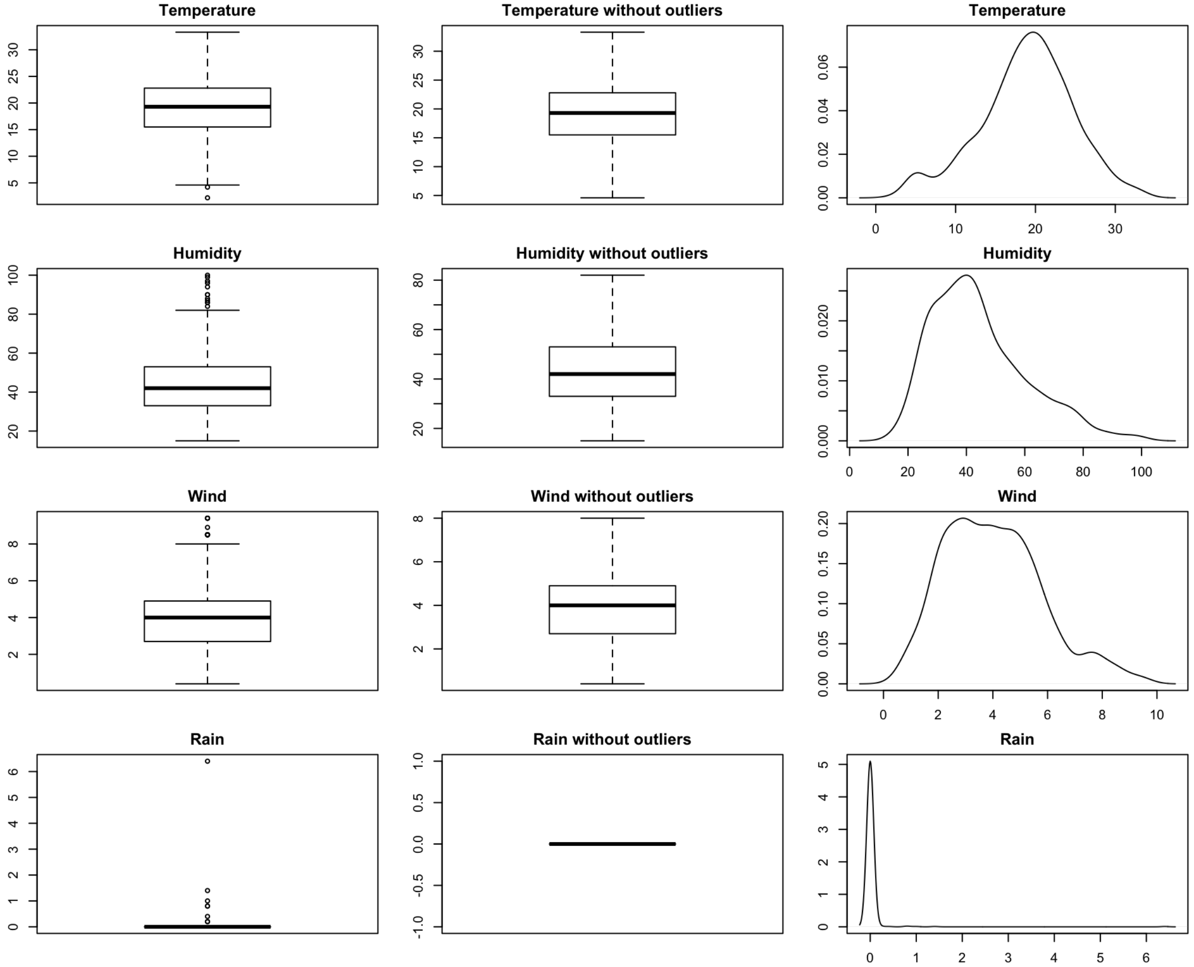
```
par(mfrow = c(4, 3))
par(mar = c(2,2,2,2))

## Now, lets get a even closer look on FWI variables' distribution through their density plots

boxplot(data$FFMC, data = data, main ="FFMC") #outliers
boxplot(data$FFMC, data = data, main ="FFMC without outliers", outline = FALSE)
plot(density(data$FFMC), main = "FFMC")

boxplot(data$DMC, data = data, main ="DMC") #outliers
boxplot(data$DMC, data = data, main ="DMC without outliers", outline = FALSE)
plot(density(data$DMC), main = "DMC")

boxplot(data$DC, data = data, main ="DC") #outliers
boxplot(data$DC, data = data, main ="DC without outliers", outline = FALSE)
plot(density(data$DC), main = "DC")

boxplot(data$ISI, data = data, main ="ISI") #outliers
boxplot(data$ISI, data = data, main ="ISI without outliers", outline = FALSE)
plot(density(data$ISI), main = "ISI")
```



Overall, like what I have expected, all 4 of the variables show very skewed distribution. Specifically, FFMC is left skewed with a long tail, DMC is somewhat right skewed, DC is somewhat left skewed with a dip in the middle of the data, and lastly, ISI is pretty right skewed with a long tail on the right. Based on what we have learned in chapter 4, and transformation might be needed for all these data.

Weather data in detail

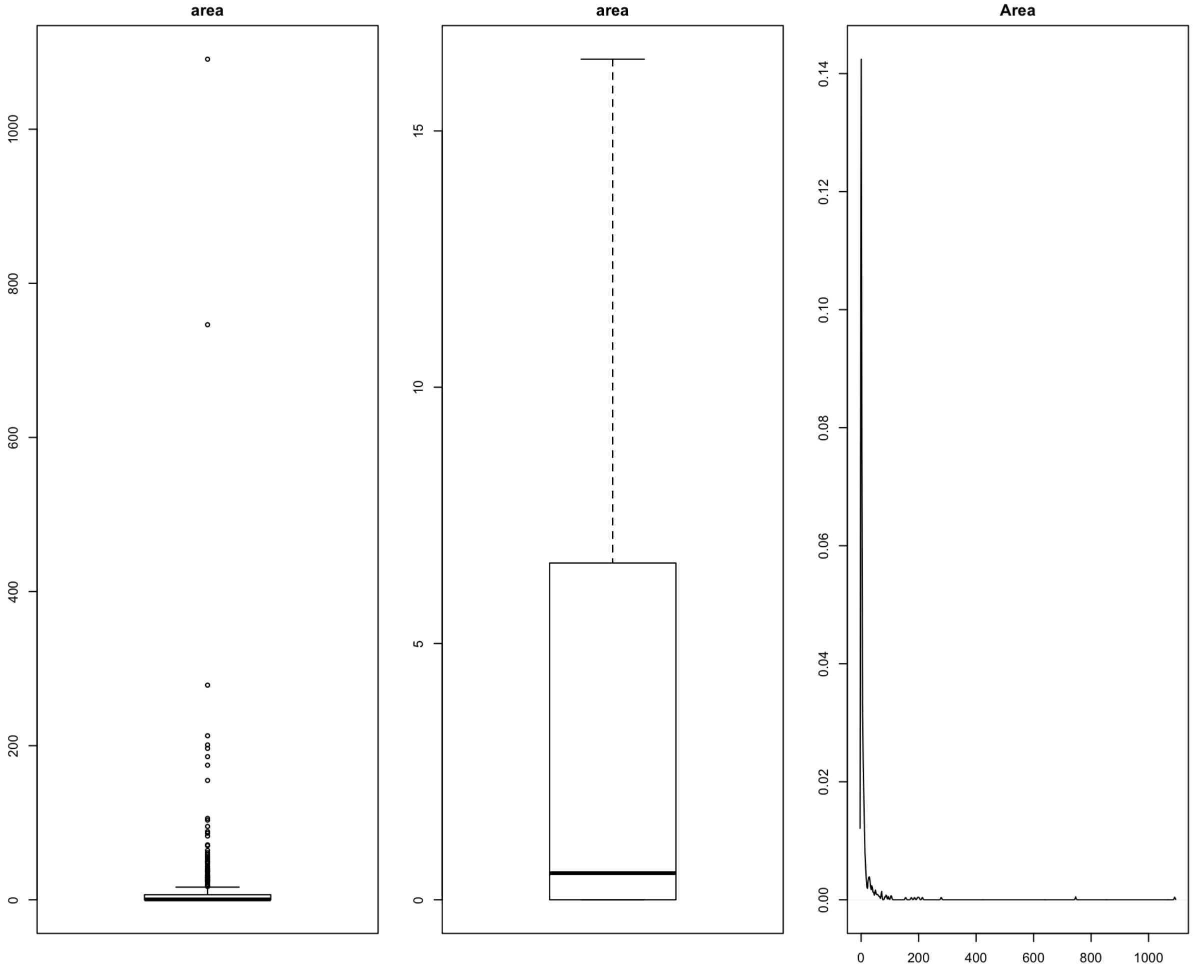
```
par(mfrow = c(4, 3))
par(mar = c(2,2,2,2))

## lets continue to take a closer look on weather variables' distribution through their density plots

boxplot(data$temp, data = data, main ="Temperature")
boxplot(data$temp, data = data, main ="Temperature without outliers", outline = FALSE)
plot(density(data$temp), main ="Temperature")


boxplot(data$RH, data = data, main ="Humidity") #outliers
boxplot(data$RH, data = data, main ="Humidity without outliers", outline = FALSE)
plot(density(data$RH), main = "Humidity")

boxplot(data$wind, data = data, main ="Wind")
boxplot(data$wind, data = data, main ="Wind without outliers", outline = FALSE)
plot(density(data$wind), main = "Wind")

boxplot(data$rain, data = data, main ="Rain") #outliers
boxplot(data$rain, data = data, main ="Rain without outliers", outline = FALSE)
plot(density(data$rain), main = "Rain")
```



Ok, now I am looking at weather variables. Good that they are not as bad. Both temperature and Wind distribution are relatively bell curved, whereas humidity is a little right skewed and rain is extremely right skewed centered around 0 with a super long tail. Again, transformation might be applied to those skewed data as I investigate more.

Response variable in detail

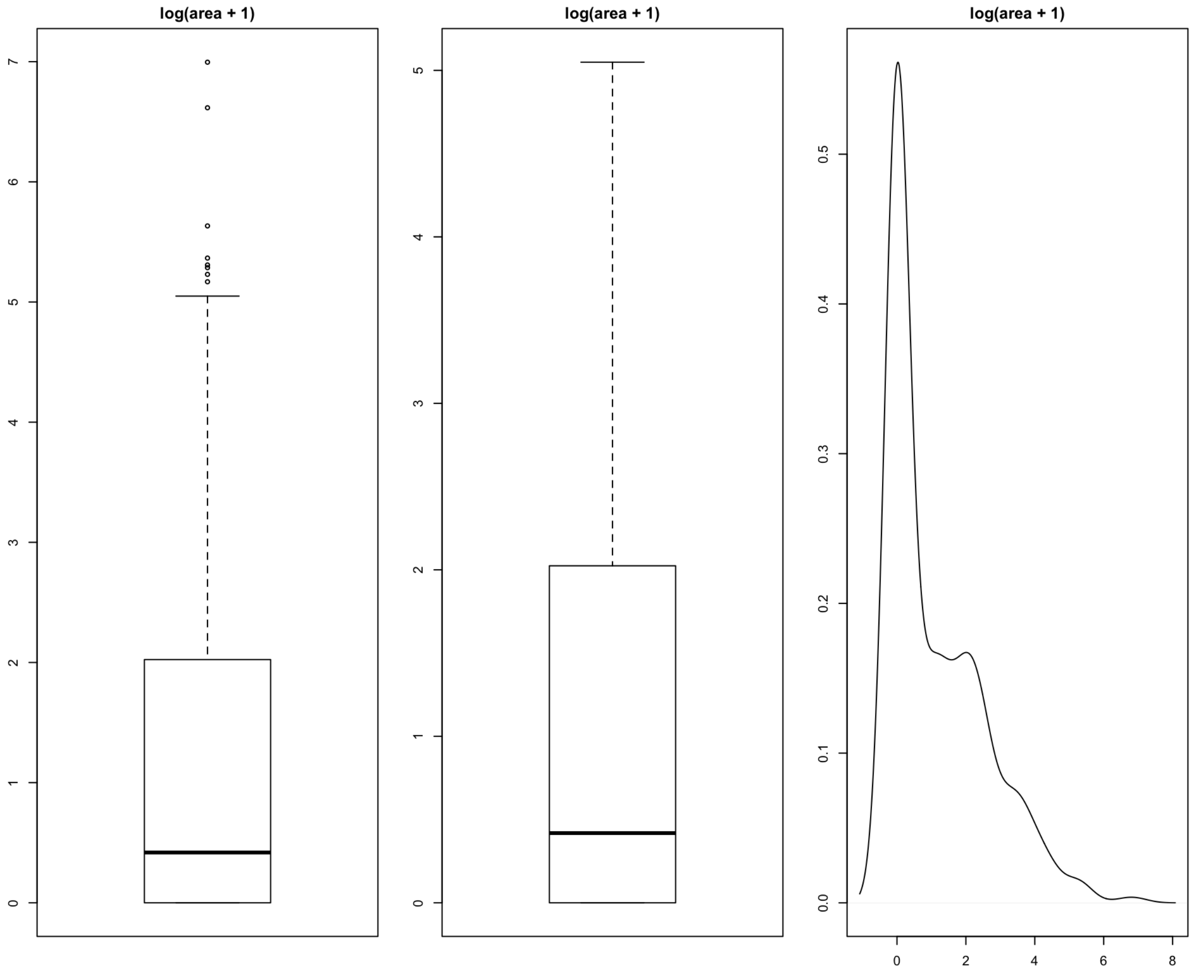
```
par(mfrow = c(1, 3))
par(mar = c(2,2,2,2))

## Ok enough with our explanatory variables
## lets now get a look at our response variable y, which is the area of the fires
boxplot(area, main = "area") #outliers
boxplot(area, main = "area", outline = FALSE)
plot(density(area), main = "Area")
```



Sadly, from the density plot above, I spot an extremely right skewed distribution centered around 0. With that being said, we will definitely need to tranform our response variable. Based on the transformation rule we learned in the beginning of the semester, I decide to use a log. However, I remember I also need to make sure that our transformed y is interpretable, so its logged result cannot be less than 0. Addtionally, as our y variable has a lot of dat points centered around 0, so if I just log(0), it would gives errors since it's negative infinity. Therefore, I tranformed it as log(area + 1).

y-transformation

```
par(mfrow = c(1, 3))
par(mar = c(2,2,2,2))

## now lets take a look at our transformed y
## and see if it actually helps with the skewed distribution
boxplot(log(area + 1), main = "log(area + 1)")
boxplot(log(area + 1), main = "log(area + 1)", outline = FALSE)
data$area <- log(data$area + 1)
plot(density(data$area), main = "log(area + 1)")
```



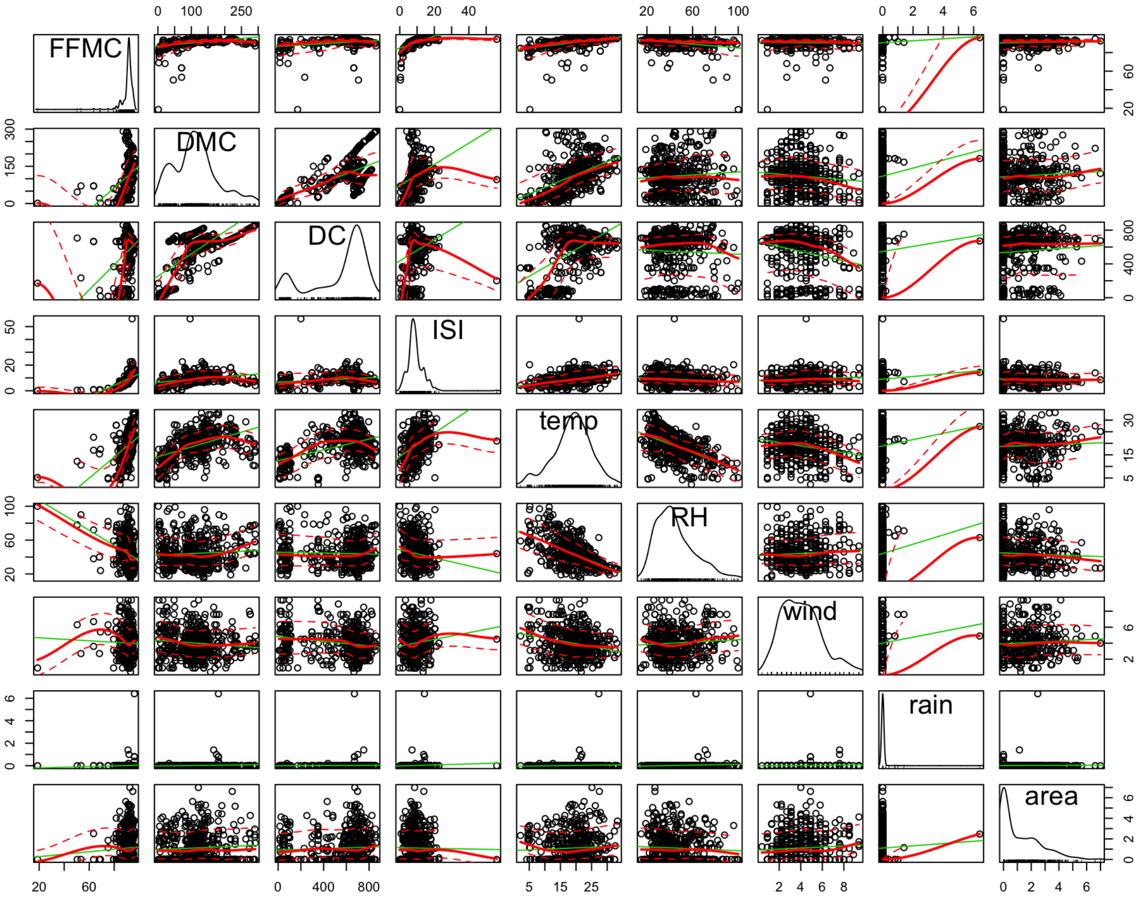
It seems like it did help with the skewness, (I also tried other log transformation, but none of them look as nice as this one, so I will go with this one) , and it does look somewhat bell-curved, though it is still a little skewed to the right, we can ignore it for now.

Overall, I transformed y to achieve approximate symmetry and homoscedasticity of the residuals. Transformations of the independent variables have a different purpose: after all, in this regression all the independent variables are taken as fixed like we did in class, not random, so “normality” is inapplicable. The independent variables don’t need to be normally distributed. The real issue with transforming the independent variables is whether the effect is linear. Thus, I decided to leave them as they are, which can also be more easy to interpret later for our model.

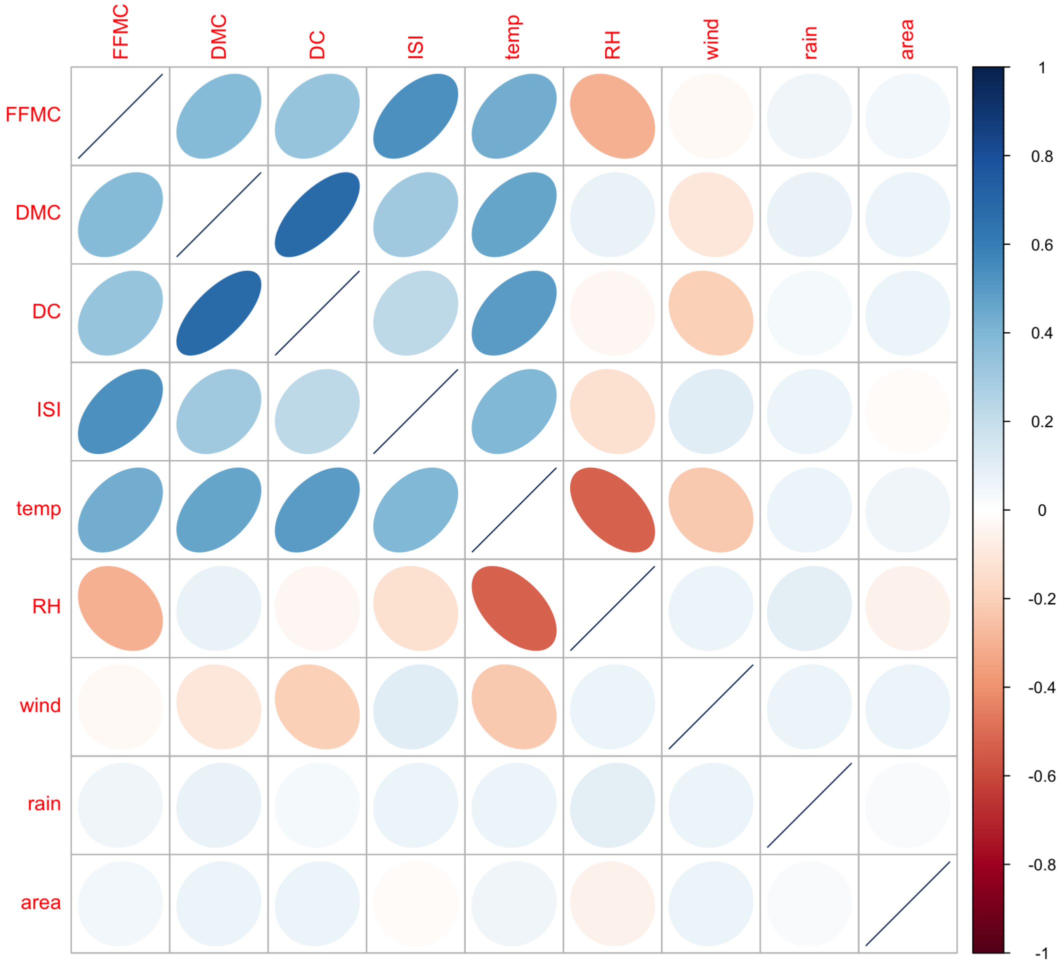
The main objective in these transformations is to achieve linear relationships with the response variable.

Bivariate Analysis

```
## lets try to spot if theres any relationships between our continous explanatory variables with each other and withour response variable
forest = data[,c(5:13)]
scatterplotMatrix(forest)
```



```
corrplot(cor(forest), method = "ellipse")
```



Based on visualizing scatter plots and corrrelation plots correlation, while there are many outliers all over the plots, something else I found interesting is that: There is a positive correlation between DC & DMC, and There is a positive correlation between temp & DMC, and There is a negative correlation between temp & RH All the above, suggest that there is probility collinearity involved, and I need to consider the interaction terms as I build my model.

Initial Modeling

lets first try add all the variables I have now

```
par(mfrow=c(2,2))
par(mar = c(2,2,2,2))
# Basic model with all continunous variables included
full_lm = lm(area ~ day+month+FFMC+DMC+DC+ISI+temp+RH+wind+rain, data = data)
summary(full_lm)
```

```
##
## Call:
## lm(formula = area ~ day + month + FFMC + DMC + DC + ISI + temp +
##     RH + wind + rain, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9115 -1.0352 -0.5310  0.7901  5.2962
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.432917   1.642356  -0.264  0.79220
## daymon       0.161628   0.226854   0.712  0.47651
## daysat       0.323862   0.217820   1.487  0.13770
## daysun       0.205031   0.211974   0.967  0.33390
## daythu       0.084256   0.240235   0.351  0.72595
## daytue       0.321098   0.235737   1.362  0.17379
## daywed       0.205553   0.246564   0.834  0.40487
## monthaug     0.177713   0.817020   0.218  0.82790
## monthdec     2.100714   0.786458   2.671  0.00781 **
## monthfeb     0.149695   0.561305   0.267  0.78982
## monthjan    -0.492862   1.217493  -0.405  0.68579
## monthjul     0.004581   0.711150   0.006  0.99486
## monthjun    -0.311975   0.655635  -0.476  0.63440
## monthmar    -0.403309   0.504648  -0.799  0.42457
## monthmay     0.653480   1.102336   0.593  0.55358
## monthnov    -1.050017   1.479920  -0.710  0.47834
## monthoct     0.746366   0.977212   0.764  0.44537
## monthsep     0.835290   0.915135   0.913  0.36182
## FFMC         0.007537   0.016660   0.452  0.65118
## DMC          0.004073   0.001863   2.187  0.02921 *
## DC          -0.001876   0.001258  -1.491  0.13665
## ISI         -0.013408   0.017982  -0.746  0.45622
## temp         0.037199   0.022252   1.672  0.09523 .
## RH           0.001599   0.006212   0.257  0.79699
## wind         0.061519   0.038424   1.601  0.11001
## rain         0.051398   0.214711   0.239  0.81091
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.384 on 491 degrees of freedom
## Multiple R-squared:  0.06842,    Adjusted R-squared:  0.02098
## F-statistic: 1.442 on 25 and 491 DF,  p-value: 0.07785
```

```
# plot(full_lm)


# add some interaction terms
interact <- cbind(data)
# again the interaction terms I added here are all based on my observations above
interact$rh_temp <- data$RH*data$temp
interact$temp_dmc <- data$temp*data$DMC
interact$isi_ffmc <- data$ISI*data$FFMC
interact$dmc_dc <- data$DMC*data$DC
lm_interact <- lm(area ~ day+FFMC+DMC+DC+ISI+temp+RH+wind+rain + rh_temp + temp_dmc+ isi_ffmc + dmc_dc, data = in
teract)
summary(lm_interact)
```

```
##
## Call:
## lm(formula = area ~ day + FFMC + DMC + DC + ISI + temp + RH +
##     wind + rain + rh_temp + temp_dmc + isi_ffmc + dmc_dc, data = interact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8812 -1.0648 -0.5838  0.8414  5.5522
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.709e-01  1.596e+00  -0.170   0.8653
## daymon       1.110e-01  2.300e-01   0.483   0.6294
## daysat       2.807e-01  2.183e-01   1.285   0.1992
## daysun       1.650e-01  2.128e-01   0.775   0.4385
## daythu      -2.255e-02  2.427e-01  -0.093   0.9260
## daytue       2.333e-01  2.357e-01   0.990   0.3229
## daywed       1.360e-01  2.475e-01   0.550   0.5828
## FFMC         1.571e-02  1.519e-02   1.034   0.3015
## DMC         -4.921e-03  6.030e-03  -0.816   0.4149
## DC           7.955e-04  4.954e-04   1.606   0.1089
## ISI          7.751e-02  4.894e-01   0.158   0.8742
## temp        -3.779e-02  4.032e-02  -0.937   0.3491
## RH          -8.188e-05  1.128e-02  -0.007   0.9942
## wind         7.714e-02  3.755e-02   2.054   0.0405 *
## rain         5.755e-02  2.217e-01   0.260   0.7953
## rh_temp     -2.506e-04  6.765e-04  -0.370   0.7112
## temp_dmc     4.442e-04  1.947e-04   2.281   0.0230 *
## isi_ffmc    -1.100e-03  5.141e-03  -0.214   0.8307
## dmc_dc      -3.549e-06  6.560e-06  -0.541   0.5887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 498 degrees of freedom
## Multiple R-squared:  0.03696,    Adjusted R-squared:  0.00215
## F-statistic: 1.062 on 18 and 498 DF,  p-value: 0.3885
```

As it is pretty easy to tell, the model fitting is terrible for my original base model, transformed base model, and transformed based model with interaction terms, though R-square did improve a tiny bit. This not necessarily because the model fit is bad, it might also be the predictors that we current have, do not have sufficient information to explain our response variable, which is fire area.

The model fit above really made me thinking, because based on my intuition and common sense, a lot of the FWI indexes and weather variables should have an impact on the area of the fires. And this made me go back to my original univariate and bivariate analysis, so I double checked to see if there's further transformation that I missed or any correlations between variables that I didnt spot in the beginning.

So I went back to my univariate analysis, I noticed that I have really dont much with my categorical variables yet, and just based on the graphs, I have no ideas how many data points exactly fall into each subcatories. Especially I would assume that when its hotter, there should be more forest fires. Thus, I did a summary of each category.

```
# getting a summary count of forest fire over all our categorical data
summary(data$day)
```

```
## fri mon sat sun thu tue wed
##  85  74  84  95  61  64  54
```

```
summary(data$month)
```

```
## apr aug dec feb jan jul jun mar may nov oct sep
##   9 184   9  20   2  32  17  54   2   1  15 172
```

```
summary(data$X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   4.669   7.000   9.000
```

```
summary(data$Y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     4.0     4.0     4.3     5.0     9.0
```

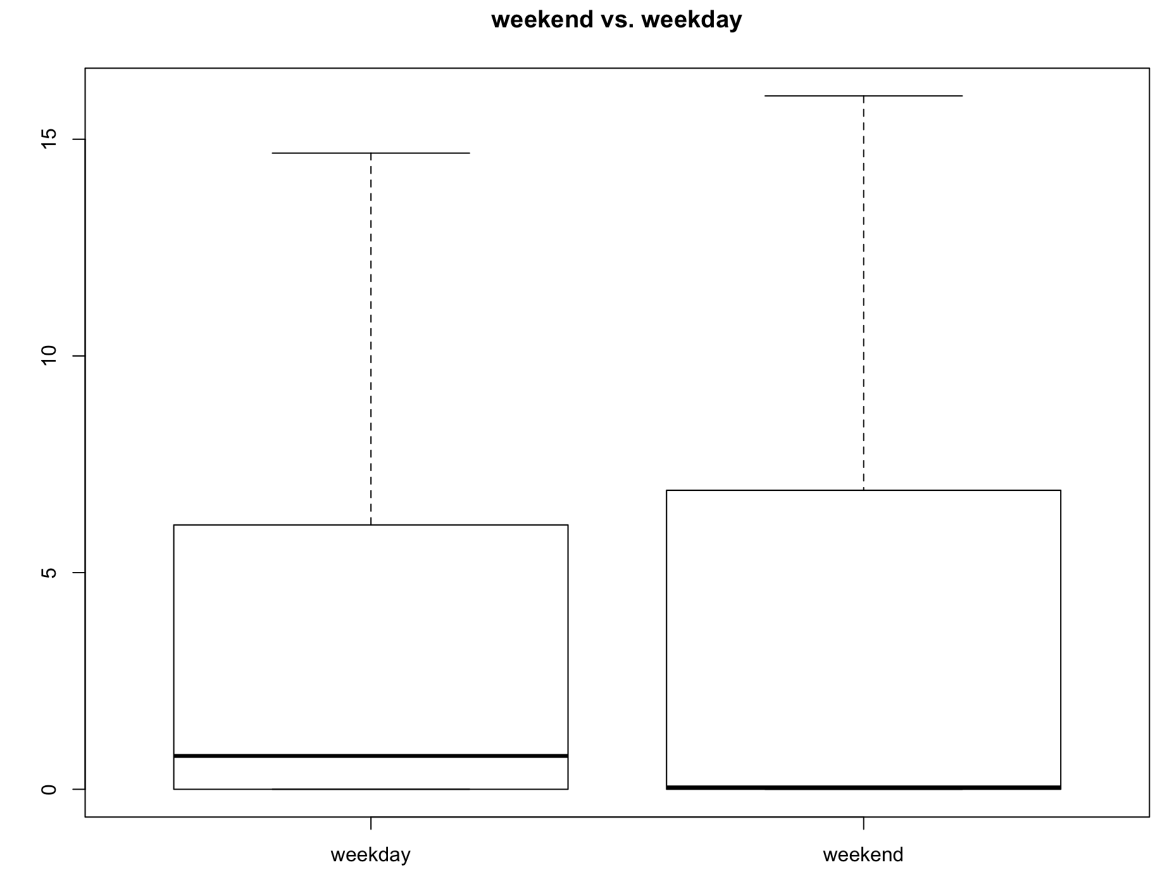
Based on my summary above, I found out that there's more fires in the weekend (fri + sat + sun), and there is significantly more fires in the summer(aug + sep). So I want to categorized them based on weekend vs weekday, and different seasons to see if this added variables can improve my model.

Adding weekend catogory to our dataset

```
# create a column called weekend
data$weekend <- rep("empty", 517)
# group friday, saturday, and sunday into weekend within the weekend column
# group the rest into the weekdays
for (i in 1:517){
  if (data$day[i] %in% c("fri", "sat", "sun")) data$weekend[i] <- "weekend"
  if (data$day[i] %in% c("mon", "tue", "wed", "thu")) data$weekend[i] <- "weekday"
}
data$weekend <- as.factor(data$weekend)
# get rid of the old explanatory variable so it wont affect our model
data$day <- NULL
head(data$weekend)
```

```
## [1] weekend weekday weekend weekend weekend weekend
## Levels: weekday weekend
```

```
# plot it newly created variable in boxplot
boxplot(area ~ data$weekend, outline  = FALSE, main = "weekend vs. weekday")
```

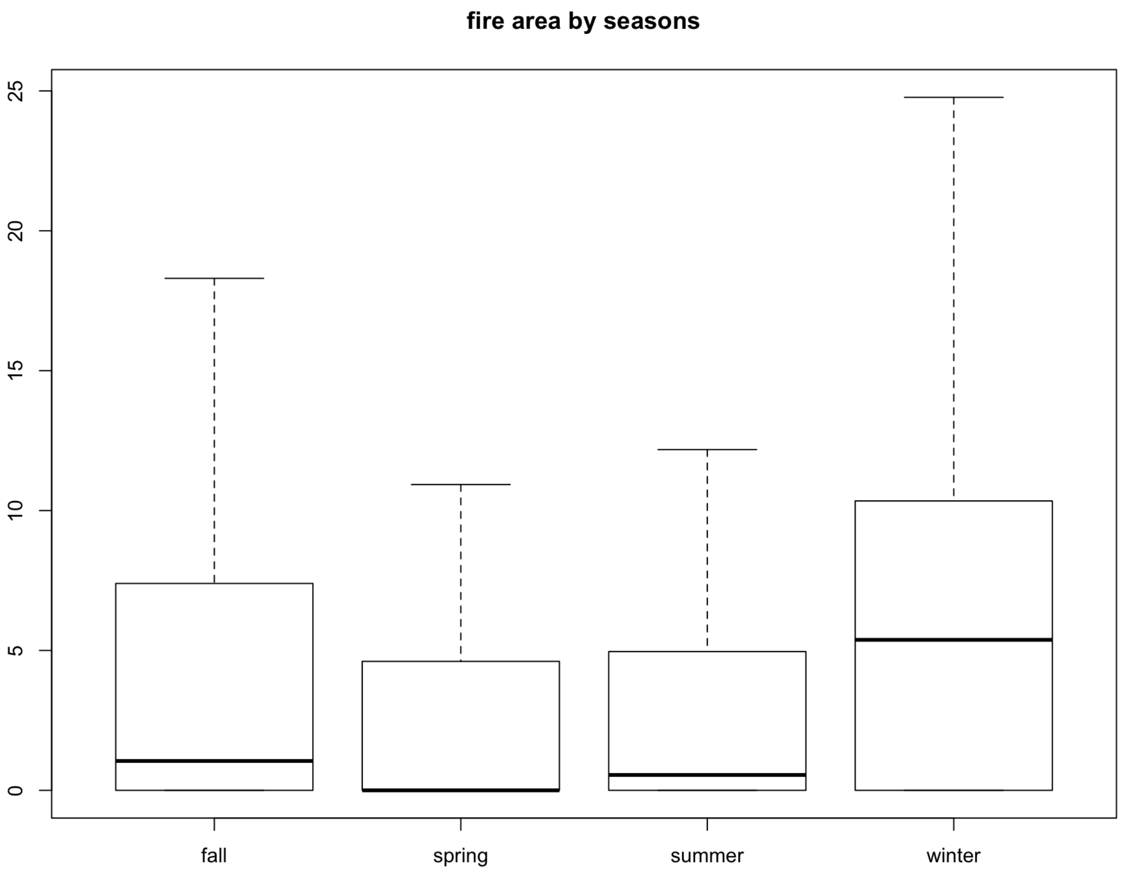


##### Adding season category to my dataset

```
# create a new season column
data$season <- rep("empty", 517)
# group the data points into their corresponding season based on their month
for (i in 1:517){
  if (data$month[i] %in% c("jan", "feb", "dec")) data$season[i] <- "winter"
  if (data$month[i] %in% c("mar", "apr", "may")) data$season[i] <- "spring"
  if (data$month[i] %in% c("jun", "jul", "aug")) data$season[i] <- "summer"
  if (data$month[i] %in% c("sep", "oct", "nov")) data$season[i] <- "fall"
}
data$season <- as.factor(data$season)
# again, get rid of the old data
data$month <- NULL
head(data$season)
```

```
## [1] spring fall   fall   spring spring summer
## Levels: fall spring summer winter
```

```
# plot it in boxplot
boxplot(area ~ data$season, outline  = FALSE, main = "fire area by seasons")
```



Although, both the weekend and season data here dont seem to provide me any extra information about fire area, I still hope to keep it to see it contributes to the model

New model with added variables

```
## lets try building our initial model again with my newly added variables
lm_new <- lm(area ~ weekend+season+FFMC+DMC+DC+ISI+temp+RH+wind+rain + RH:temp + FFMC:ISI+ DMC:DC + FFMC:DMC + te
mp:DMC, data = data)
summary(lm_new)
```

```
##
## Call:
## lm(formula = area ~ weekend + season + FFMC + DMC + DC + ISI +
##     temp + RH + wind + rain + RH:temp + FFMC:ISI + DMC:DC + FFMC:DMC +
##     temp:DMC, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8905 -1.0706 -0.5254  0.8071  5.4325
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.606e+00  2.111e+00  -0.761   0.4472
## weekendweekend -1.320e-02  1.278e-01  -0.103   0.9178
## seasonspring    1.988e-02  5.557e-01   0.036   0.9715
## seasonsummer   -5.771e-01  2.146e-01  -2.689   0.0074 **
## seasonwinter    9.114e-01  6.530e-01   1.396   0.1634
## FFMC            2.021e-02  1.981e-02   1.020   0.3083
## DMC             3.049e-02  3.417e-02   0.892   0.3727
## DC              9.132e-04  9.087e-04   1.005   0.3154
## ISI            -1.221e-01  5.649e-01  -0.216   0.8290
## temp           -8.910e-03  4.458e-02  -0.200   0.8417
## RH             -5.183e-04  1.138e-02  -0.046   0.9637
## wind            8.052e-02  3.733e-02   2.157   0.0315 *
## rain            3.810e-02  2.193e-01   0.174   0.8621
## temp:RH         1.136e-04  6.785e-04   0.167   0.8671
## FFMC:ISI        1.092e-03  5.978e-03   0.183   0.8551
## DMC:DC         -1.815e-05  8.199e-06  -2.214   0.0273 *
## FFMC:DMC       -2.199e-04  3.748e-04  -0.587   0.5576
## DMC:temp        3.194e-04  2.321e-04   1.376   0.1695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.382 on 499 degrees of freedom
## Multiple R-squared:  0.05491,    Adjusted R-squared:  0.02271
## F-statistic: 1.705 on 17 and 499 DF,  p-value: 0.03838
```

YAY, I see that r-squared has improved significantly, most importantly, I have a F-statistic for the model and I see some variables with very small p-values, which means that their coefficients should not be 0 and they might contribute to explaining our model. And, now I can finally move on to the actual modeling part.

Here is a good place for me to go back to our client's question: 1. What are the variables that influence the total burned area in forest fires? 2. What is a good predictive equation for predicting the total burned area in terms of given variables (do you need all of them?)

and really think about how I want to approach these 2 questions.

The first thing, that came to my mind is that these 2 questions are asking about very different things, though they seem very similar. The first question is asking for inferetial model, while the second questions asks me to predict.

Therefore, after going over what we have been learning in class through out the sememster, I chose ANOVA for my inferential model and stepwise selection for my predictive model, and here are my reasons: 1. ANOVA measures the relevance of features by their correlation with dependent variable while Stepwise Selection measure the usefulness of a subset of feature by actually training a model on it.
2. ANOVA methods might fail to find the best subset of features in many occasions but Stepwise Selection methods can always provide the best subset of predictor features. 3. Using the subset of features from ANOVA make the model more prone to overfitting, which I absolutely need to avoid for my predictive model, as compared to using subset of features from the Stepwise Selection.

And now LETS START:

Explanatory Modeling

Finding the most importatnt variables (or features) that explains major part of variance of the response variable is key to identify and build high performing models.

Anova Approach

```
## I first created a base model with all the independent variables and interaction terms included.
baseMod <-lm(area ~ weekend+season+FFMC+DMC+DC+ISI+temp+RH+wind+rain + RH:temp + FFMC:ISI+ DMC:DC + FFMC:DMC + te
mp:DMC, data = data)

## I then set up a model with only intercepts
null_mod <- lm(area ~ 1, data = data)

## Now I use anova to see whether all the variables and interactions terms I have make sense
## That is, whether they have enough explanatory power than just using intercept
anova(baseMod, null_mod)
```

```
## Analysis of Variance Table
## 
## Model 1: area ~ weekend + season + FFMC + DMC + DC + ISI + temp + RH + 
##     wind + rain + RH:temp + FFMC:ISI + DMC:DC + FFMC:DMC + temp:DMC
## Model 2: area ~ 1
##   Res.Df     RSS  Df Sum of Sq      F  Pr(>F)  
## 1    499  953.69                               
## 2    516 1009.10 -17    -55.41 1.7054 0.03838 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on my anova test between my full model and my restricted intercept model, I see that the p-value is less then 0.05, which is the normal significance level, thus, the statistic is significant enough that we can reject our null hypothesis and claim that my explanatory variables do have some power in explaining forest fire area. YAY, so now I can continue to select features that actually contribute to the explanation of the data.

```
## Here, I will create multiple subsets of my full model
## And use anova() to check if the additional variable contribute to the explanatory ability of the model.
## Below, I have baseMod with all 10 explanatory variables and 4 interaction terms
## While, mod1 through mod14 contain one predictor less than the previous model

## With principle or marginality in mind
## I first got rid of FFMC:ISI
mod_1 <- lm(area ~ weekend+season+FFMC+DMC+DC+ISI+temp+RH+wind+rain + RH:temp + DMC:DC + FFMC:DMC + temp:DMC, dat
a = data)

## took off ISI
mod_2 <- lm(area ~ weekend+season+FFMC+DMC+DC+temp+RH+wind+rain + RH:temp + DMC:DC + FFMC:DMC + temp:DMC, data = 
data)

## took off DMC:DC
mod_3 <- lm(area ~  weekend+season+FFMC+DMC+DC+temp+RH+wind+rain + RH:temp  + FFMC:DMC + temp:DMC, data = data)

## took off DC
mod_4 <- lm(area ~  weekend+season+FFMC+DMC+temp+RH+wind+rain + RH:temp  + FFMC:DMC + temp:DMC, data = data)

## took off RH:temp
mod_5 <- lm(area ~ weekend+season+FFMC+DMC+temp+RH+wind+rain+ FFMC:DMC + temp:DMC, data = data)

## took off RH
mod_6 <- lm(area ~ weekend+season+FFMC+DMC+temp+wind+rain + FFMC:DMC + temp:DMC, data = data)

## took off FFMC:DMC
mod_7 <- lm(area ~ weekend+season+FFMC+DMC+temp+wind+rain + temp:DMC, data = data)

## took off FFMC
mod_8 <- lm(area ~ weekend+season+DMC+temp+wind+rain + temp:DMC, data = data)

## took off weekend
mod_9 <- lm(area ~ season+DMC+temp+wind+rain + temp:DMC, data = data)

## took off temp:DMC
mod_10 <- lm(area ~ season+DMC + temp+wind+rain, data = data)

## took off DMC
mod_11 <- lm(area ~ season+temp+wind + rain, data = data)

## took off rain
mod_12 <- lm(area ~ season+temp+wind, data = data)

## took off temp
mod_13 <- lm(area ~ season+wind, data = data)

## took off wind
mod_14 <- lm(area ~ season, data = data)

## took off season
mod_15 <- lm(area ~ 1, data = data)


anova(baseMod, mod_1, mod_2, mod_3, mod_4, mod_5, mod_6, mod_7, mod_8, mod_9, mod_10,mod_11, mod_12, mod_13, mod_
14, mod_15)
```

```
## Analysis of Variance Table
## 
## Model  1: area ~ weekend + season + FFMC + DMC + DC + ISI + temp + RH + 
##     wind + rain + RH:temp + FFMC:ISI + DMC:DC + FFMC:DMC + temp:DMC
## Model  2: area ~ weekend + season + FFMC + DMC + DC + ISI + temp + RH + 
##     wind + rain + RH:temp + DMC:DC + FFMC:DMC + temp:DMC
## Model  3: area ~ weekend + season + FFMC + DMC + DC + temp + RH + wind + 
##     rain + RH:temp + DMC:DC + FFMC:DMC + temp:DMC
## Model  4: area ~ weekend + season + FFMC + DMC + DC + temp + RH + wind + 
##     rain + RH:temp + FFMC:DMC + temp:DMC
## Model  5: area ~ weekend + season + FFMC + DMC + temp + RH + wind + rain + 
##     RH:temp + FFMC:DMC + temp:DMC
## Model  6: area ~ weekend + season + FFMC + DMC + temp + RH + wind + rain + 
##     FFMC:DMC + temp:DMC
## Model  7: area ~ weekend + season + FFMC + DMC + temp + wind + rain + FFMC:DMC + 
##     temp:DMC
## Model  8: area ~ weekend + season + FFMC + DMC + temp + wind + rain + temp:DMC
## Model  9: area ~ weekend + season + DMC + temp + wind + rain + temp:DMC
## Model 10: area ~ season + DMC + temp + wind + rain + temp:DMC
## Model 11: area ~ season + DMC + temp + wind + rain
## Model 12: area ~ season + temp + wind + rain
## Model 13: area ~ season + temp + wind
## Model 14: area ~ season + wind
## Model 15: area ~ season
## Model 16: area ~ 1
##    Res.Df     RSS Df Sum of Sq      F  Pr(>F)  
## 1     499  953.69                                
## 2     500  953.76 -1   -0.0638 0.0334 0.85514  
## 3     501  955.70 -1   -1.9432 1.0168 0.31378  
## 4     502  964.07 -1   -8.3677 4.3782 0.03691 *
## 5     503  964.20 -1   -0.1301 0.0681 0.79427  
## 6     504  964.31 -1   -0.1120 0.0586 0.80879  
## 7     505  964.32 -1   -0.0070 0.0036 0.95186  
## 8     506  966.02 -1   -1.7054 0.8923 0.34531  
## 9     507  967.37 -1   -1.3522 0.7075 0.40068  
## 10    508  967.44 -1   -0.0701 0.0367 0.84823  
## 11    509  970.92 -1   -3.4779 1.8198 0.17795  
## 12    510  978.61 -1   -7.6922 4.0248 0.04538 *
## 13    511  978.88 -1   -0.2643 0.1383 0.71017  
## 14    512  989.51 -1  -10.6325 5.5633 0.01873 *
## 15    513  996.70 -1   -7.1948 3.7645 0.05291 .
## 16    516 1009.10 -3  -12.3969 2.1621 0.09161 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I know I built a lot of models here. The reason why I am doing this is because I want to be extra careful when I am dealing with explanatory models, as it cares a lot about minimizing bias as to where predictive models want to minimize variance. For every variable that I decide to eliminate, I need to have solid evidence to show that this variable indeed doesnt contribute to the model. And this is basically why I decided to check it one by one.

With the anova chart R produced above, I noticed that DMC:DC, DMC, temp, wind, season, all have p-values way below the significance level of 0.05 or close to it. This basically means, that there's sufficient evidence that we need rejust the null for this variables that their coefficient is not equal to 0 and they contribute to the model. Another thing to pay attention to is that, I also need to think about principle of marginality since there is an interaction term DMC:DC involved, I only have DMC here, so I also need to include DC. This won't really affect the accuracy of my model that much, as this is my model here is for explaining the data, it doesnt penalize if I have too many variables, as long as it has something to contribute to the model.

To summarize, here I have answered the first question of my client, that is which variables can influence the total burned area in the forest. Based on my analysis those variables are: Temperature, Wind, Season, DMC, DC and DMC:DC. Later in this report I will also validate this answer.

Predictive Modeling

Step-wise Regression

Again, I picked step-wise regression (also since Professor recommended not to use LASSO and Ridge ahaha) because this can be a very effective method since I want to:
1) be highly selective about discarding valuable predictor variables
2) build multiple models on the response variable for later validation

```
# base intercept only model
base.mod <- lm(area ~ 1, data = data)
# full model with all predictors
all.mod <- lm(area ~ weekend+season+FFMC+DMC+DC+ISI+temp+RH+wind+rain + RH:temp + FFMC:ISI+ DMC:DC + FFMC:DMC + t
emp:DMC, data = data)
# perform step-wise algorithm
stepMod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction = "both", trace = 0, steps =
 1000)
# get the shortlisted variable
shortlistedVars <- names(unlist(stepMod[[1]]))
# remove intercept
shortlistedVars <- shortlistedVars[!shortlistedVars %in% "(Intercept)"]
# lets see what variables got picked from this method
print(shortlistedVars)
```

```
## [1] "seasonspring" "seasonsummer" "seasonwinter" "DMC"
## [5] "temp"         "wind"
```

```
# summary
summary(stepMod)
```

```
##
## Call:
## lm(formula = area ~ season + DMC + temp + wind, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8019 -1.0471 -0.5753  0.8487  5.5190
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.041620   0.359305   0.116  0.90783
## seasonspring -0.080146   0.242688  -0.330  0.74135
## seasonsummer -0.390222   0.145059  -2.690  0.00738 **
## seasonwinter  0.740523   0.335333   2.208  0.02767 *
## DMC           0.002589   0.001278   2.026  0.04329 *
## temp          0.032042   0.014280   2.244  0.02527 *
## wind          0.079298   0.035522   2.232  0.02603 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.38 on 510 degrees of freedom
## Multiple R-squared:  0.0377, Adjusted R-squared:  0.02638
## F-statistic:  3.33 on 6 and 510 DF,  p-value: 0.003165
```

```
# lets see their coefficients and other statistics
step(stepMod)
```

```
## Start:  AIC=339.89
## area ~ season + DMC + temp + wind
##
##          Df Sum of Sq    RSS    AIC
## <none>                971.06 339.89
## - DMC     1    7.8151 978.88 342.03
## - wind    1    9.4886 980.55 342.92
## - temp    1    9.5870 980.65 342.97
## - season  3   26.2162 997.28 347.66
```

```
##
## Call:
## lm(formula = area ~ season + DMC + temp + wind, data = data)
##
## Coefficients:
##  (Intercept)  seasonspring  seasonsummer  seasonwinter           DMC
##     0.041620     -0.080146     -0.390222      0.740523      0.002589
##         temp          wind
##     0.032042      0.079298
```

```
lm.final <- stepMod
```

Stepwise Seletion, similarly suggests that I includ season, DMC, temperature and wind in my predictiv model, and a descent AIC score of 340.

Model Diagnostics

Basic Assumptions:

1. Error Assumtions:

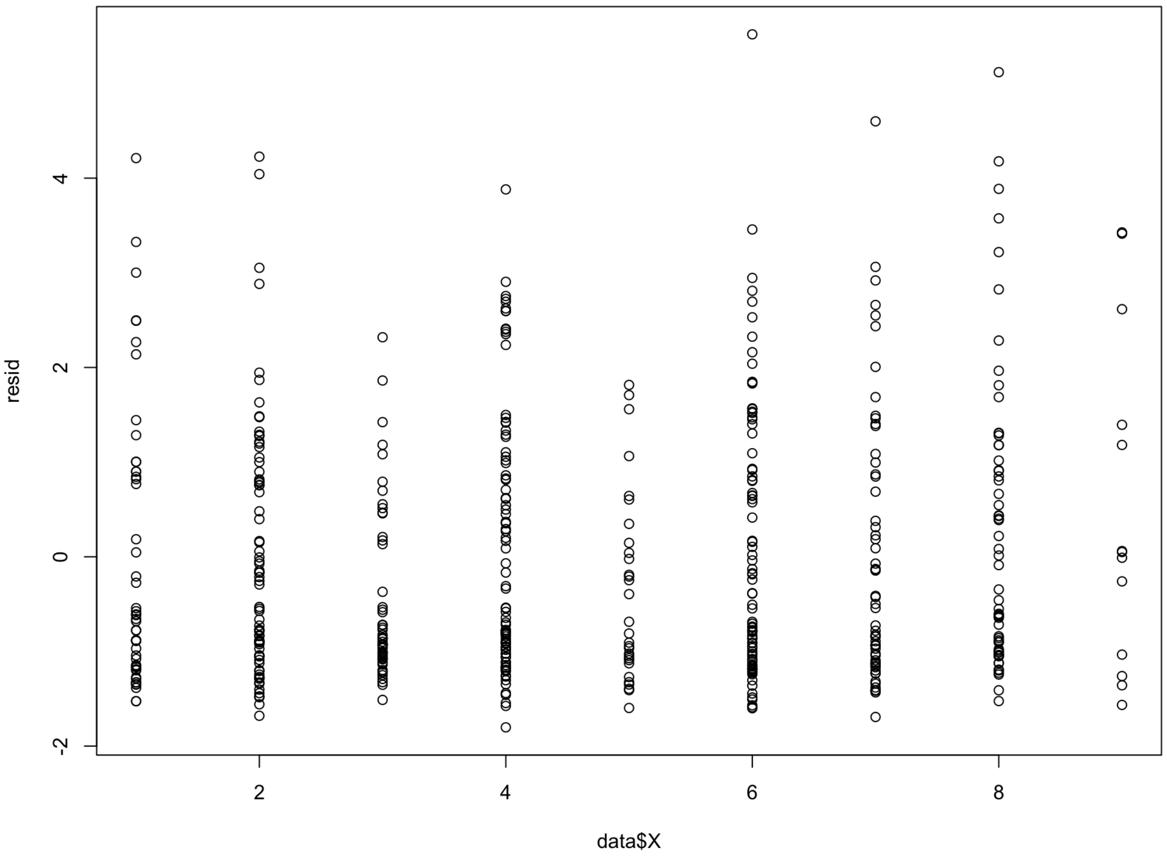
1. Independence: this is pretter hard to measure
2. Homoscedasticity - Constant Variance: check if residuals are independent of y, I plot the errors against the fitted value y-hat. I expect to see a relatively constant spread of points centered around 0
3. Normality: Q-Q plot

2. Structure of the Model: Linearity

I will check the relationship between my explanatory variables and my response variable. I have to make sure that there is no obvious nonlinearity that would invalidate my model, which might require me further transformation. And to visualize these relationships, I will use partial residual plot, one for each explanatory variable

1) Independence

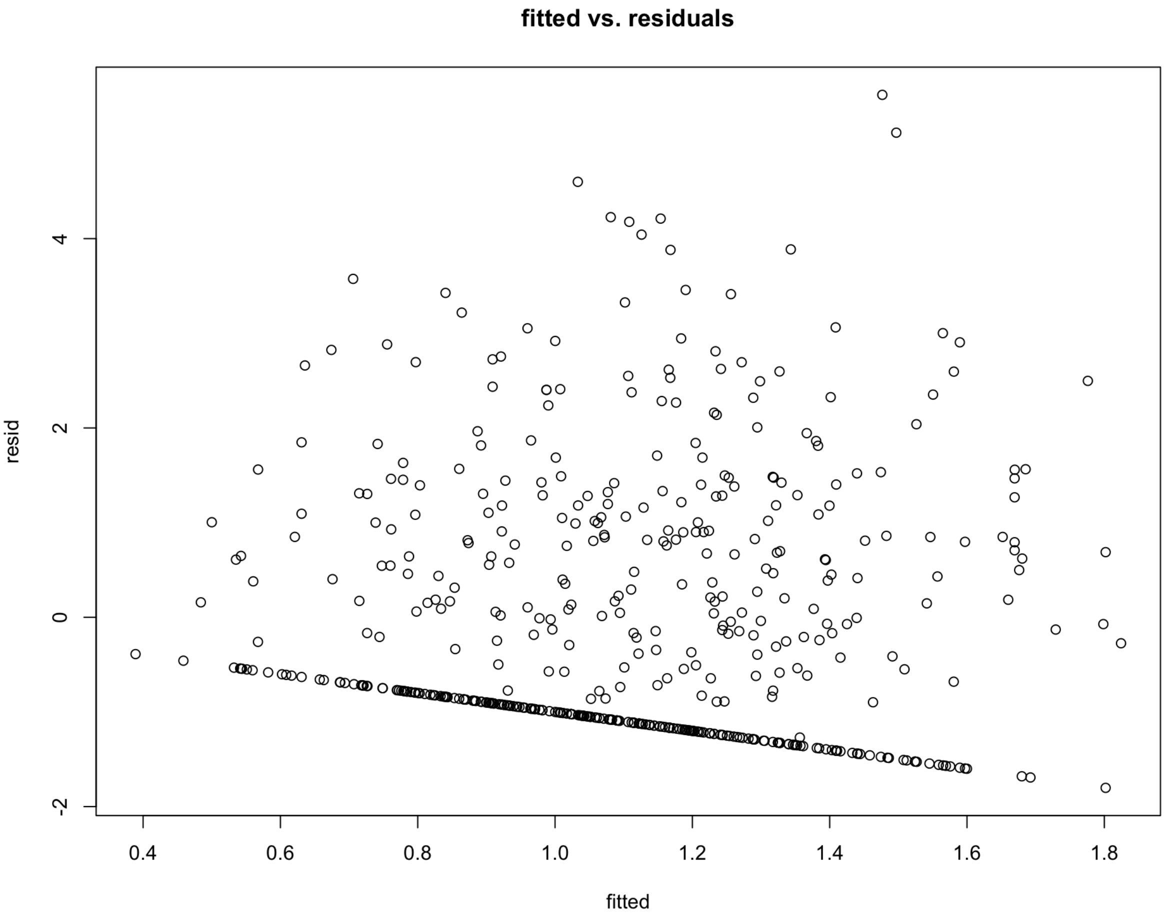
```
## Independence
resid = lm.final$residuals
## plot residuals against any spatial variables present
plot(data$X, resid)
```



Plotted residuals against the spatial data I have, and the dots look pretty random to me, which indicates that the independence criteria is met. I used this plot because dependence on spatial varaiables are common source of lack of independence.

2) Homoscedasticity

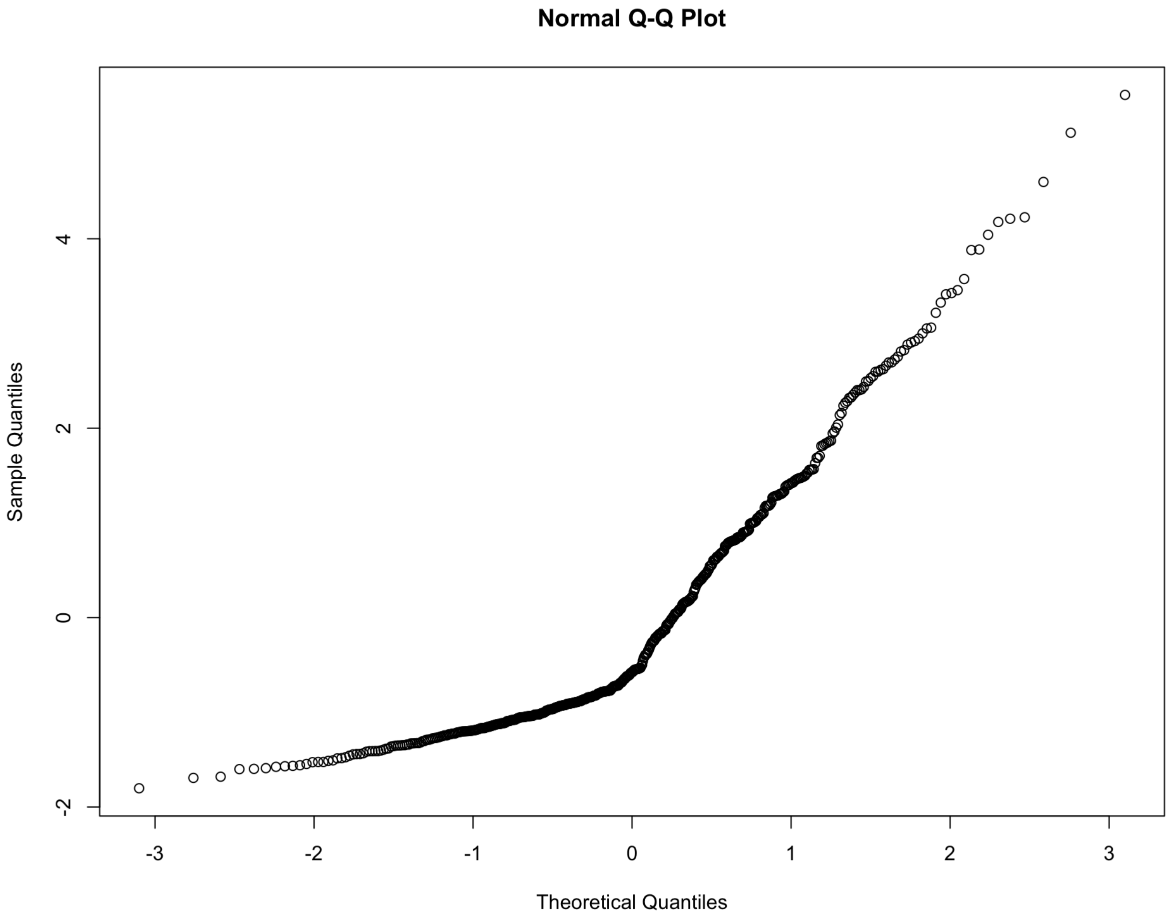
```
## Homoscedasticity
resid = lm.final$residuals
fitted = lm.final$fitted.values
plot(fitted, resid, main = "fitted vs. residuals")
```



Indeed, I see a mostly constant spread of residuals centered around 0. Most residuals are in the range of -2 to 4 (with an exception of the line in the bottom, which I assume to be the result of my transformation). Thus, I think this is sufficient for me to assume that the requirement of homoscedasticity is fulfilled

3) Normality

```
## Normality - QQ plot
qqnorm(resid)
```

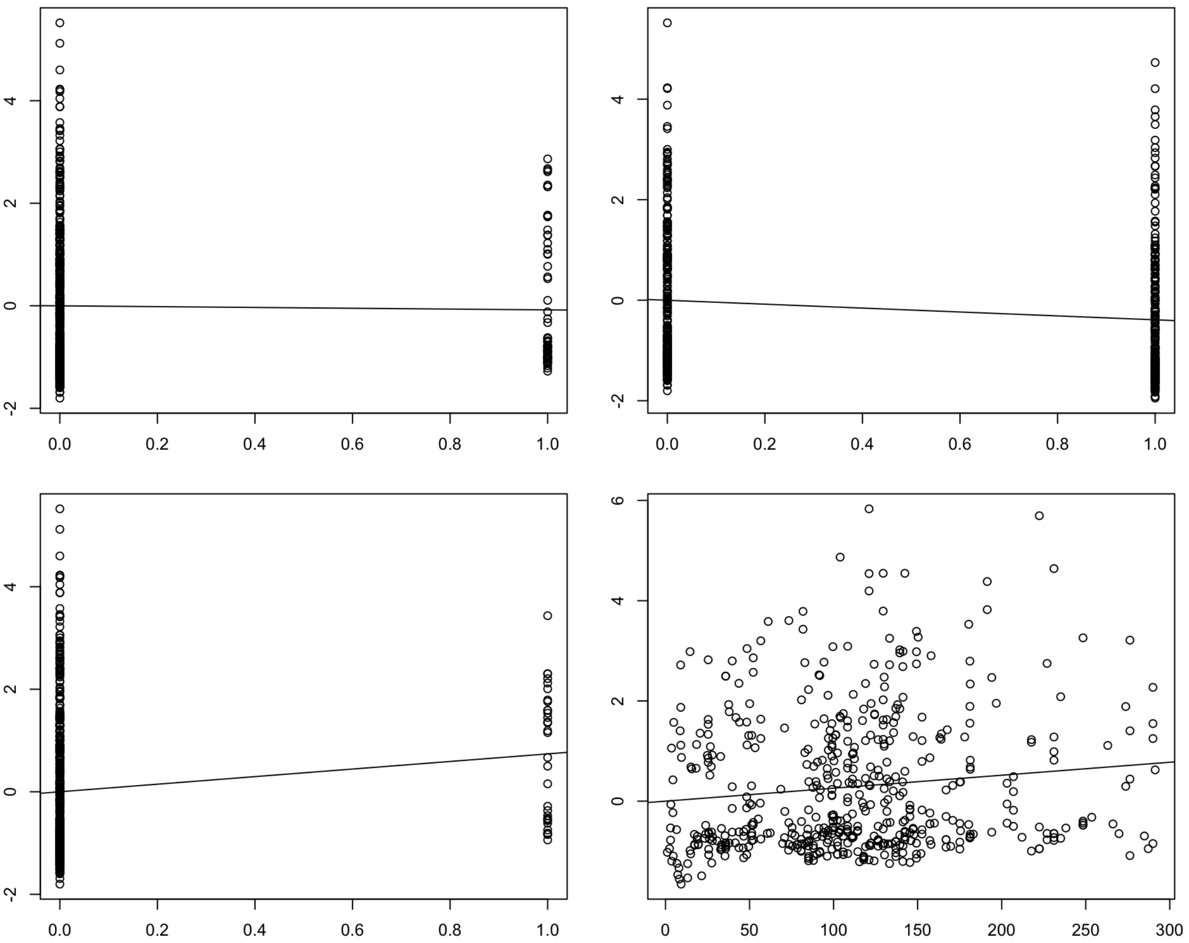


```
#qqline(resid)
```

For the most part, the residuals are somewhat close to a straight line as desired, though, there is still some points that deviated a little from th QQ line, I think this is enough to conclude that errors are normally distributed

4) Linearity

```
par(mfrow=c(2,2))
par(mar = c(2,2,2,2))
## Linearity - partial residual plots
prplot(lm.final, 1)
prplot(lm.final, 2)
prplot(lm.final, 3)
prplot(lm.final, 4)
```



Finally, I check the structure of my explanatory varialbles relationship with the response and make sure there is no obvious nonlinearity that would invalidate my model. To do this, I drew partial residual plot for each explanatory varialbe. Just by looking at the scatter plots, the correlation seems pretty weak since we have season as our categorical variable in the regression.

Outliers/Leverage Points/Influential Points

1. Outliers - Studentized residuals
2. Leverage Points - Mahalanobis Distance
3. Influential Points - Cook's Distance

1) Outliers

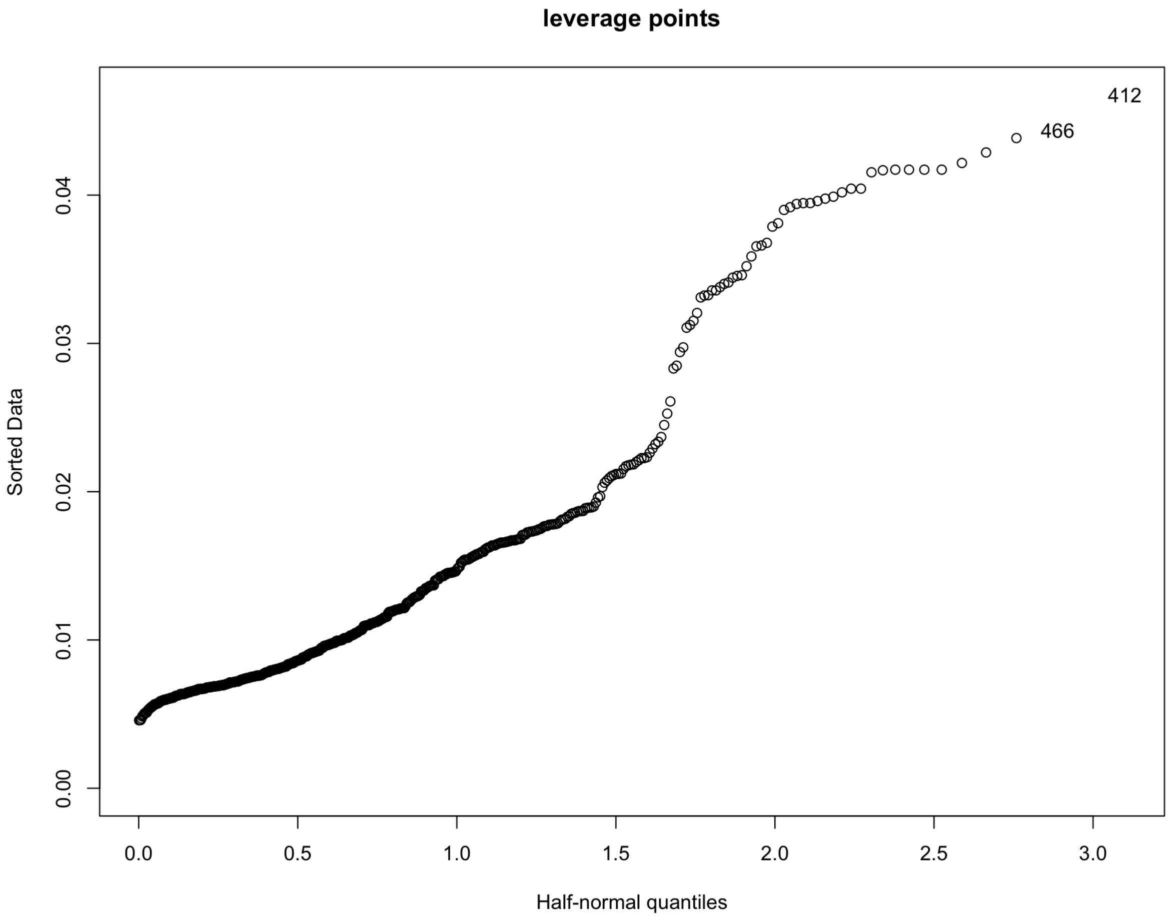
```
## Outliers
## Need to check particularly for outliers that are influential
## To do this, I calucate studentized residuals and perform a Bonferroni-corrected hypothesis test on them.
outlierTest(lm.final, cutoff = 0.05)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 239 4.079237         5.2436e-05      0.02711
```

From the chart, I see the p-value is 0.0271, so I should reject the hypothesis that is some of the points are influential outliers. My largest studentized residual point is 239: 4.08 which correspond to a Bonferonni p-value of 0.0271.This studentized residual test demonstrates that this point does in some way skew my regression plane enough to be classified as a problematic influential outlier.

2) Leverage Points

```
## Leverage Points
hat = lm.influence(lm.final)$hat
halfnorm(hat, main = "leverage points")
```



```
tail(sort(hat))
```

```
##        280        375         60        380        466        412 
## 0.04171454 0.04216657 0.04287593 0.04384624 0.04438233 0.04675813
```

As it is pretty easy to see in the half-norm plot and the hat matrix, data points 412, 466, 380 have the highest leverages.

3) Influential Points

```
## Influential Points
## Here, I will use Coook's distance to find points with large distance values
## Meaning that the coefficients of my regression would change significantly
## If i was to remove this point
hat.root = sqrt(1 - hat)
n = nrow(data)
k = 4
se = sqrt(sum(resid ^ 2) / (n - k - 1))
sigma.stdized = resid / (hat.root * se )
cook.dist = (sigma.stdized^2 / (k + 1)) * (hat / (1 - hat))
tail(sort(cook.dist))
```

```
##        378        212        470        239        396        416 
## 0.02349417 0.02641931 0.02857010 0.03001156 0.03152950 0.03885153
```

416 is most influential according to Cook’s distance. It is interesting to find out that point 239 which is also my outlier, also appears here as the third highest influential point.

Violations? Transformations?

As I would already expect from my univariate and multivariate analysis from the beginning, there are many outliers, leverage points and influential points detected through model diagnostics, and many of them overlapped too. This means that I need to consider whether I want to include these observations in the final model, which might depend on various factors and require me to take an even closer look at those specific data points to see if the errors are just an input error.

Model Evaluation

Model evaluation for explanatory model

```
# fit my explanatory model
explain <- lm(area ~ DMC + DC + DMC:DC + temp + wind + season, data = data)

# function for roo meak squared error
rmse <- function(error) {sqrt(mean(error ^ 2))}
rmse(explain$residuals)
```

```
## [1] 1.363455
```

For model evaluation for my predictive model, I decided to use K(10) fold cross validation to estimate my predictive model error. Cross Validation involves partitioning the data into an explicit training dataset used to prepare the model and an unseen test dataset used to evaluate the models performance un unseen data. Below, I used cross validation that I partitioned the forest fires into 10 sub dataset so that some is used for traning and the rest used for validating the model accuracy.

```
# define training control
train_control <- trainControl(method = "cv", number = 10)

# train the model
cv_mod <- train(area ~ DMC + temp + wind + season, data = data, trControl = train_control, method = "lmStepAIC")
```

and now with cross validation it train the data based on my intial predictive model, which resulted in many combinations of variables. And based on the biggest AIC, I can go ahead and pick the best model. Also, AIC is used here before it can reflect the bias and variance trade-off, as it is very sensitive in preditive models. We want to contrain the number of total predictors without hurting the model accuracy too much.

Final Model Inference

Overall, it was fun exploring this challenging dataset. I like how there are 2 different aspects to building the models for different use. For my explanatory model, temperature, wind, season, DMC, DMC:DC, DC resulted in relatively accurate goodness of fit, meaning that they do have a somewhat strong explanatory power on the burned area. On the other, I wish I had more time to play with the predictive model, as in the end, I did detect many influential points and outliers that seem like they need to be eliminate from the model for a higher accuracy, and also the linearity assumption was not met by my current model, in that sense, I also need ot go back and transform some independent variable and repeat all these steps until all my assumptions are somewhat met.