TO PREDICT BURNED AREA IN THE FOREST

PUSHKAR SHARMA

April 28, 2022

**Statement of the research question and of the objective of regression modeling**

We are trying to check whether there is a linear relationship between the metrological factors and the forest area burned.

**Description of the data**

Number of Instances: 517

Number of Attributes: 13

Output variable: 1

**Attribute Description**

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: ‘Jan’ to ‘Dec’
4. day - day of the week: ‘Mon’ to ‘Sun’
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20 (Fine Fuel Moisture Code) :- FFMC denotes the moisture content surface litter and influences ignition and fire spread.
6. DMC - DMC index from the FWI system: 1.1 to 291.3

(Duff Moisture Code):- Represents the moisture content of shallow and deep organic layers.

1. DC - DC index from the FWI system: 7.9 to 860.6
2. ISI - ISI index from the FWI system: 0.0 to 56.10(Initial Spread Index):-The ISI is a score that correlates with fire velocity spread.
3. temp - temperature in Celsius degrees: 2.2 to 33.30
4. RH - relative humidity in %: 15.0 to 100
5. wind - wind speed in km/h: 0.40 to 9.40
6. rain - outside rain in mm/m2 : 0.0 to 6.4
7. area - the burned area of the forest (in ha): 0.00 to 1090.84 (TO BE PREDICTED)

## [1] 517 13

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

## LINEAR MODEL FROM SCRATCH

## [,1]  
## [1,] -5.986986920  
## [2,] 1.266091800  
## [3,] 1.492543205  
## [4,] 0.016041925  
## [5,] 0.005889101

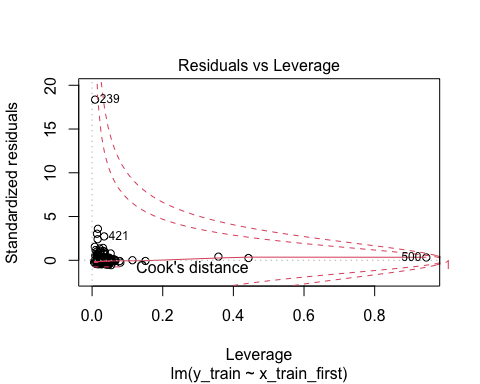
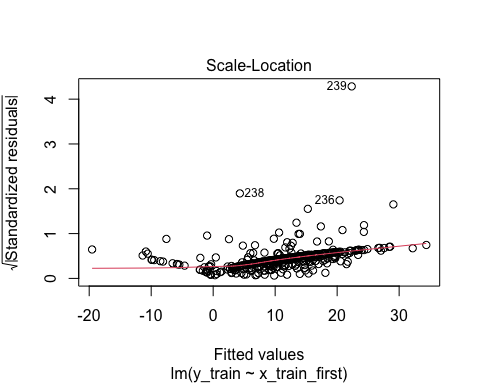
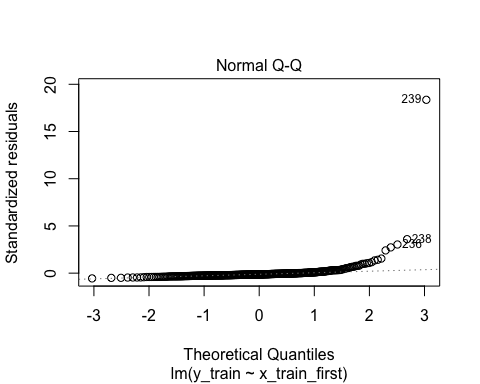
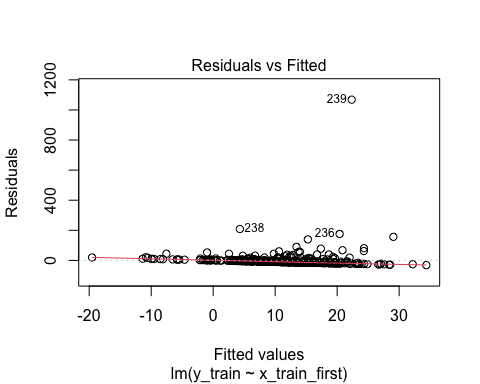
## [,1]  
## 1 9.300487  
## 3 12.878928  
## 6 20.161854  
## 9 14.364330  
## 17 12.801588

## [,1]  
## 1 0  
## 3 0  
## 6 0  
## 9 0  
## 17 0

## [1] 66.35069

## LINEAR MODEL USING lm

##   
## Call:  
## lm(formula = y\_train ~ x\_train\_first)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.63 -13.20 -7.66 -0.56 1068.49   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -5.986987 60.391147 -0.099 0.921  
## x\_train\_first1 1.266092 1.469720 0.861 0.390  
## x\_train\_first2 1.492543 2.793150 0.534 0.593  
## x\_train\_first3 0.016042 0.621356 0.026 0.979  
## x\_train\_first4 0.005889 0.069348 0.085 0.932  
## x\_train\_first5 0.007638 0.016890 0.452 0.651  
## x\_train\_first6 -0.680190 0.764015 -0.890 0.374  
## x\_train\_first7 0.537856 0.816969 0.658 0.511  
## x\_train\_first8 -0.233606 0.245170 -0.953 0.341  
## x\_train\_first9 1.145858 1.751315 0.654 0.513  
## x\_train\_first10 -1.168485 9.105273 -0.128 0.898  
##   
## Residual standard error: 58.45 on 402 degrees of freedom  
## Multiple R-squared: 0.0165, Adjusted R-squared: -0.007965   
## F-statistic: 0.6744 on 10 and 402 DF, p-value: 0.7484



coef(linear\_model) w and coefficient are coming out same so the result from the scratch function and the lm is same

**As P-value are very high for all the features vector so this seams that the basic linear model is not a good option to train the data so we will try to do EDA and then after seeing the plot we will try to do some transformation and try to improve the model**

## [1] 0

## X Y month day   
## Min. :1.000 Min. :2.0 Length:517 Length:517   
## 1st Qu.:3.000 1st Qu.:4.0 Class :character Class :character   
## Median :4.000 Median :4.0 Mode :character Mode :character   
## Mean :4.669 Mean :4.3   
## 3rd Qu.:7.000 3rd Qu.:5.0   
## Max. :9.000 Max. :9.0   
## FFMC DMC DC ISI   
## Min. :18.70 Min. : 1.1 Min. : 7.9 Min. : 0.000   
## 1st Qu.:90.20 1st Qu.: 68.6 1st Qu.:437.7 1st Qu.: 6.500   
## Median :91.60 Median :108.3 Median :664.2 Median : 8.400   
## Mean :90.64 Mean :110.9 Mean :547.9 Mean : 9.022   
## 3rd Qu.:92.90 3rd Qu.:142.4 3rd Qu.:713.9 3rd Qu.:10.800   
## Max. :96.20 Max. :291.3 Max. :860.6 Max. :56.100   
## temp RH wind rain   
## Min. : 2.20 Min. : 15.00 Min. :0.400 Min. :0.00000   
## 1st Qu.:15.50 1st Qu.: 33.00 1st Qu.:2.700 1st Qu.:0.00000   
## Median :19.30 Median : 42.00 Median :4.000 Median :0.00000   
## Mean :18.89 Mean : 44.29 Mean :4.018 Mean :0.02166   
## 3rd Qu.:22.80 3rd Qu.: 53.00 3rd Qu.:4.900 3rd Qu.:0.00000   
## Max. :33.30 Max. :100.00 Max. :9.400 Max. :6.40000   
## area   
## Min. : 0.00   
## 1st Qu.: 0.00   
## Median : 0.52   
## Mean : 12.85   
## 3rd Qu.: 6.57   
## Max. :1090.84

## [1] 103 13

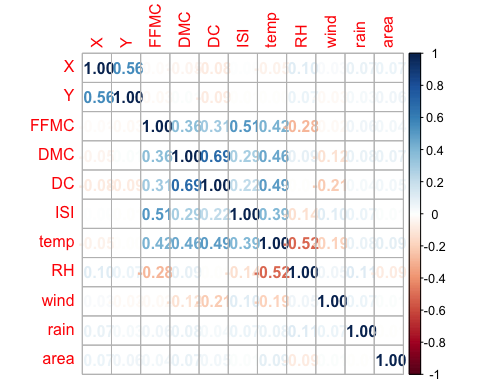
## [1] 414 13

## X Y month day   
## Min. :1.000 Min. :2.000 Length:414 Length:414   
## 1st Qu.:3.000 1st Qu.:4.000 Class :character Class :character   
## Median :4.000 Median :4.000 Mode :character Mode :character   
## Mean :4.652 Mean :4.316   
## 3rd Qu.:7.000 3rd Qu.:5.000   
## Max. :9.000 Max. :9.000   
## FFMC DMC DC ISI   
## Min. :18.70 Min. : 1.10 Min. : 15.3 Min. : 0.000   
## 1st Qu.:90.30 1st Qu.: 73.88 1st Qu.:460.7 1st Qu.: 6.525   
## Median :91.60 Median :108.30 Median :661.8 Median : 8.400   
## Mean :90.68 Mean :112.23 Mean :549.5 Mean : 9.132   
## 3rd Qu.:92.80 3rd Qu.:142.40 3rd Qu.:713.7 3rd Qu.:11.000   
## Max. :96.20 Max. :291.30 Max. :860.6 Max. :56.100   
## temp RH wind rain   
## Min. : 4.20 Min. : 15.00 Min. :0.400 Min. :0.0000   
## 1st Qu.:15.90 1st Qu.: 32.25 1st Qu.:2.700 1st Qu.:0.0000   
## Median :19.40 Median : 40.50 Median :4.000 Median :0.0000   
## Mean :19.12 Mean : 43.55 Mean :3.999 Mean :0.0256   
## 3rd Qu.:22.88 3rd Qu.: 53.00 3rd Qu.:5.400 3rd Qu.:0.0000   
## Max. :33.10 Max. :100.00 Max. :9.400 Max. :6.4000   
## area   
## Min. : 0.000   
## 1st Qu.: 0.000   
## Median : 0.385   
## Mean : 12.771   
## 3rd Qu.: 5.995   
## Max. :1090.840

No missing values. Month and day are as factors. From Summary FFMC, DMC and DC seem left skewed. ISI , Rain may be right skewed. Area heavily right skewed

## Exploratory data analysis

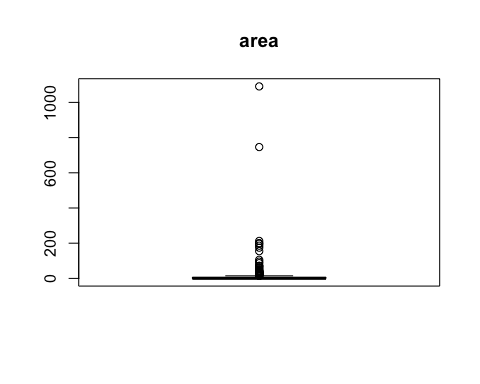
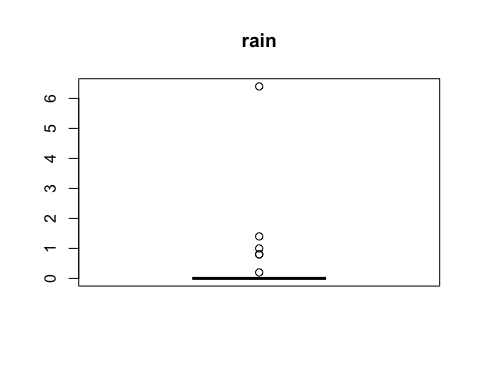
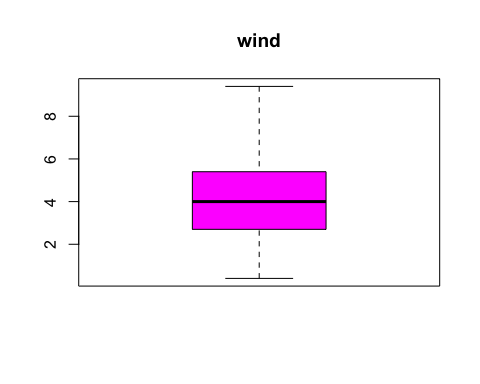
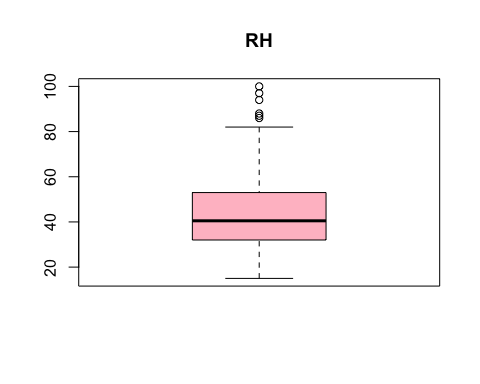
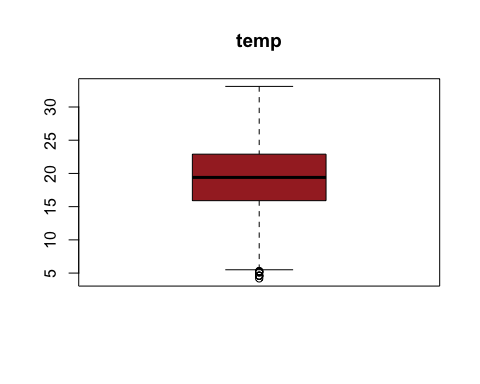
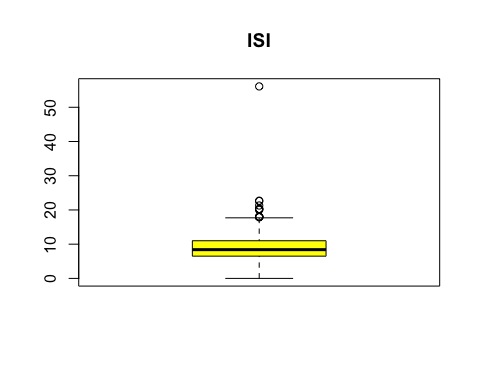
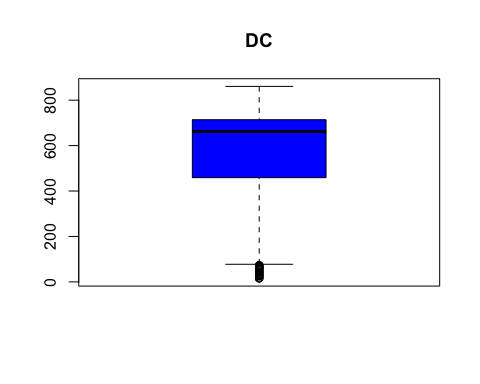
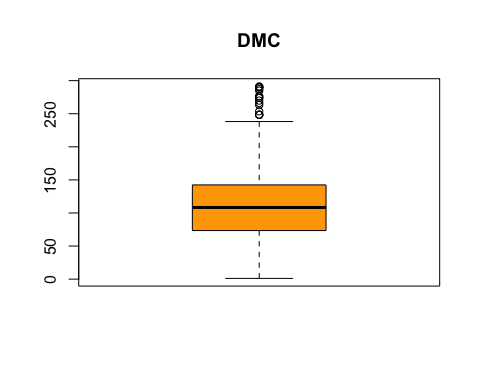
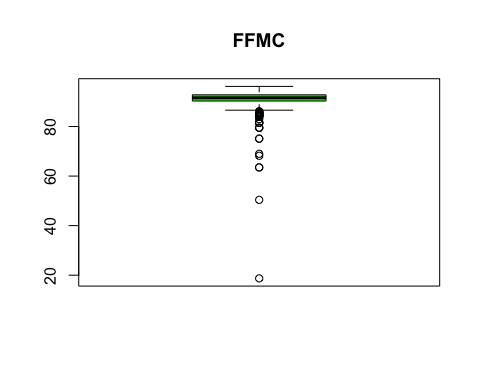
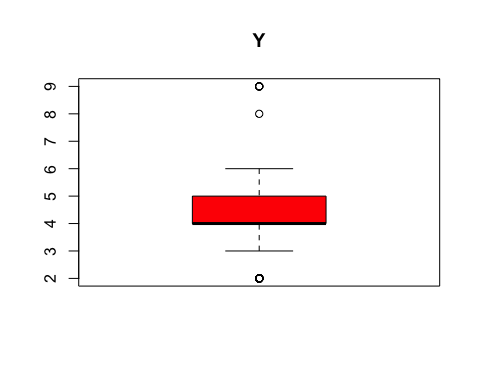
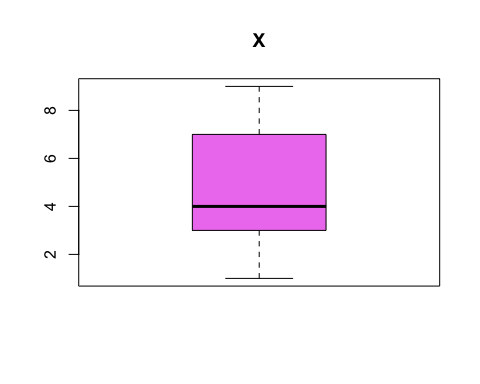
## X Y FFMC DMC DC  
## X 1.000000000 0.555815132 -0.01264825 -0.04748225 -0.081634076  
## Y 0.555815132 1.000000000 -0.02839400 0.01337154 -0.085468283  
## FFMC -0.012648249 -0.028393996 1.00000000 0.36450301 0.306235383  
## DMC -0.047482249 0.013371539 0.36450301 1.00000000 0.689149989  
## DC -0.081634076 -0.085468283 0.30623538 0.68914999 1.000000000  
## ISI 0.002332372 -0.009040472 0.51469812 0.28562713 0.215971018  
## temp -0.051680997 -0.001651361 0.41664954 0.45635976 0.492601704  
## RH 0.101135606 0.069925060 -0.28408131 0.08760536 -0.007500934  
## wind 0.027410477 -0.026852782 -0.02420261 -0.12119978 -0.214377146  
## rain 0.072293426 0.034257854 0.05869154 0.07724143 0.038985914  
## area 0.069143610 0.055872288 0.04259023 0.07424035 0.049340940  
## ISI temp RH wind rain  
## X 0.002332372 -0.051680997 0.101135606 0.02741048 0.072293426  
## Y -0.009040472 -0.001651361 0.069925060 -0.02685278 0.034257854  
## FFMC 0.514698121 0.416649538 -0.284081313 -0.02420261 0.058691539  
## DMC 0.285627128 0.456359760 0.087605359 -0.12119978 0.077241429  
## DC 0.215971018 0.492601704 -0.007500934 -0.21437715 0.038985914  
## ISI 1.000000000 0.393163956 -0.139210864 0.10138257 0.067244655  
## temp 0.393163956 1.000000000 -0.516526599 -0.19124628 0.078330132  
## RH -0.139210864 -0.516526599 1.000000000 0.05336059 0.111950702  
## wind 0.101382568 -0.191246278 0.053360592 1.00000000 0.066265873  
## rain 0.067244655 0.078330132 0.111950702 0.06626587 1.000000000  
## area 0.011720108 0.092980803 -0.085854528 0.01018596 -0.006688597  
## area  
## X 0.069143610  
## Y 0.055872288  
## FFMC 0.042590225  
## DMC 0.074240353  
## DC 0.049340940  
## ISI 0.011720108  
## temp 0.092980803  
## RH -0.085854528  
## wind 0.010185955  
## rain -0.006688597  
## area 1.000000000



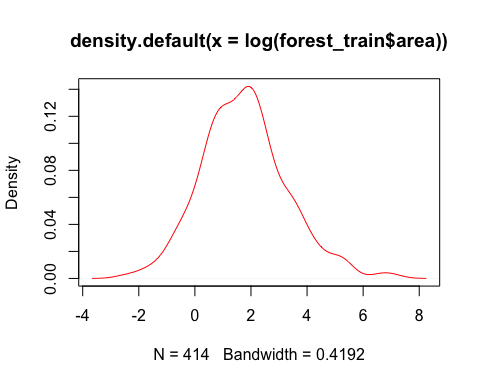
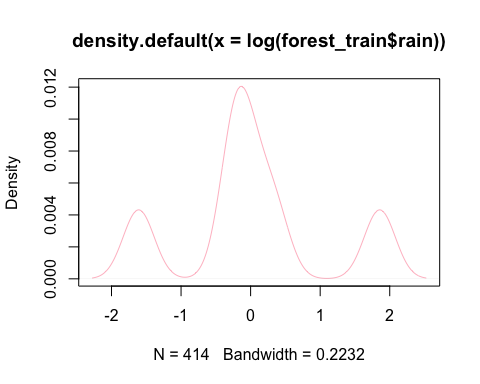
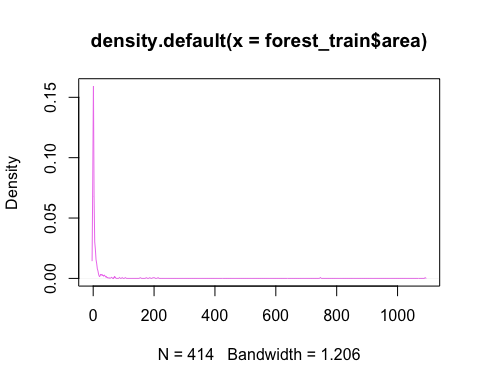
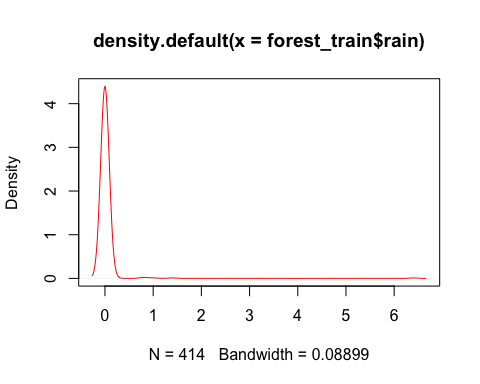
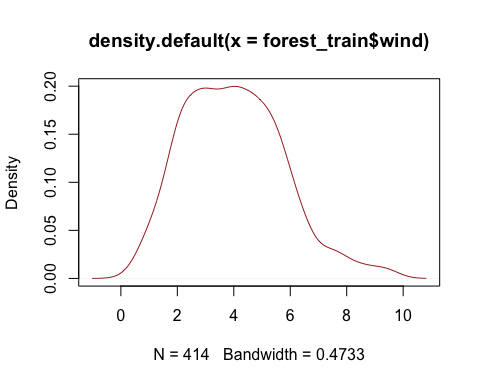
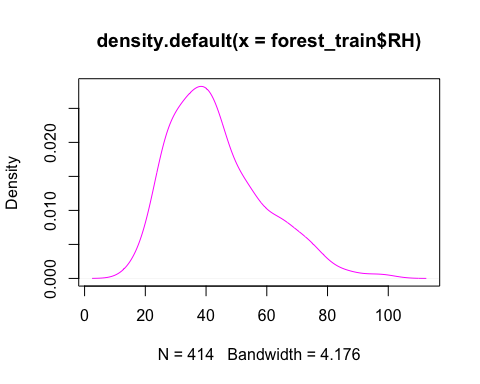
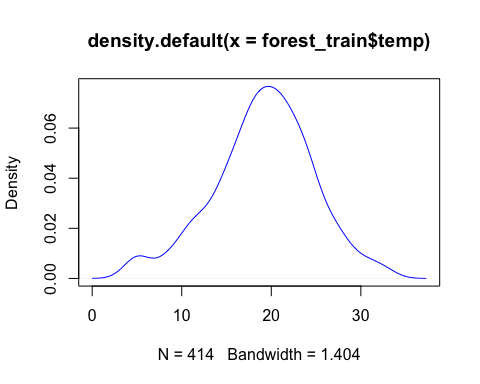
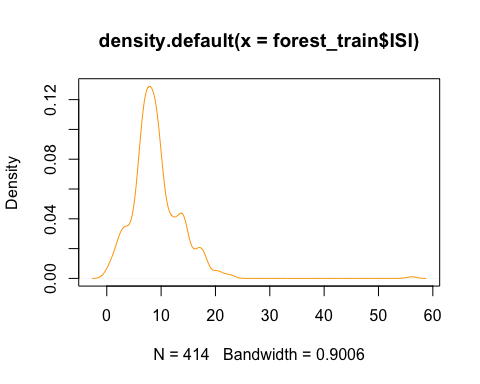
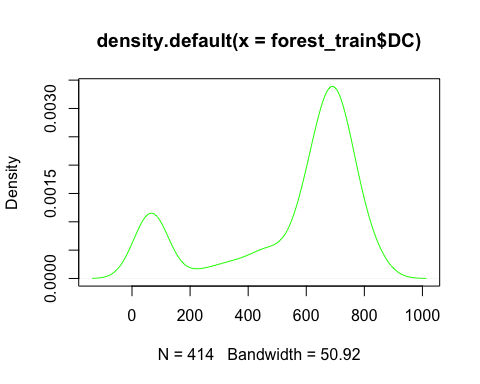
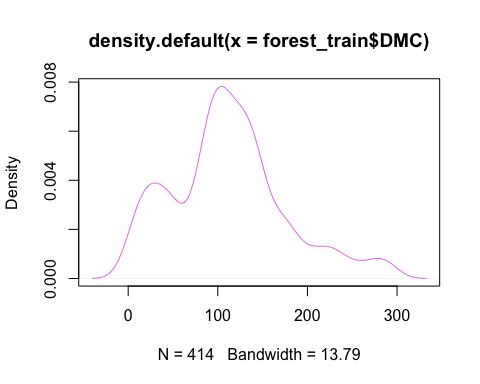
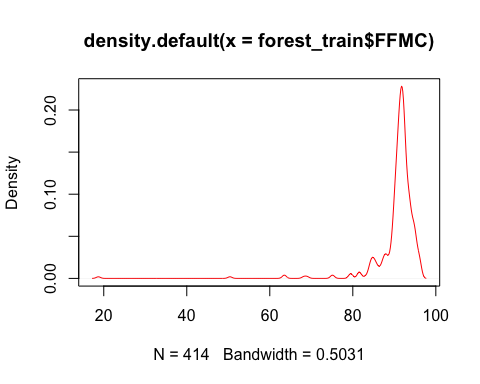
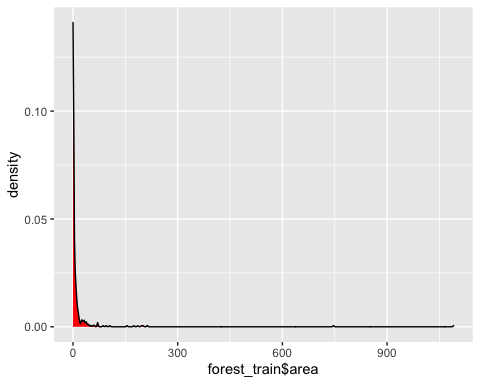
evident positive corr between DC & DMC,ISI & FFMC , X &Y , temp & DC. negative corr between RH & temp.

positive DC & DMC - this as per definition makes sense positive temp & DMC - This somewhat does not makes much sense as moisture should decrease with temp negative RH & temp - This is also natural as temp increases humidity decreases.We then inspect the distribution of each variable in box plots

# We then inspect the distribution of each variable in box plots:-

 asymmetry also observed in variables like X,Y,DC,FFMC from box plot we can easily able to see the outliers and it is clearly seen that the rain and area has hogh variablity and very large outliersLet’s see the prob density distribution curve of response variable area

# Let’s see the prob density distribution curve of response variable area



## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.385 12.771 5.995 1090.840

## [1] 4790.204

## [1] 69.2113

Above box plots and density suggest reflected log transform of FFMC and **log transform** of rain and area, the response variable since it is highly concentrated near zero and as symmetrical

The variable distribution is very concentrated around 1 and 10, but we can some extreme outliers, even above 1000 !!

**Explanation of the modeling approach and choice of model. This should include the model type, model selection procedure, and the interaction terms that you will include in the final model**

*By seeing the EDA some variable show well with log transform now will use all the inference we collected from the EDA and try to make different model and try to increase the R-Square value of the model*

**MODEL SELECTION**

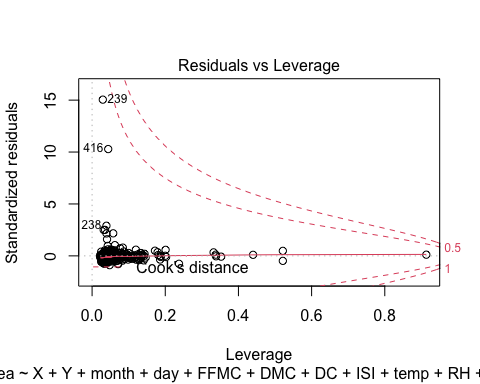
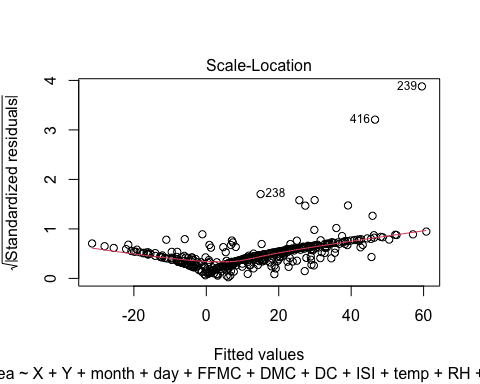
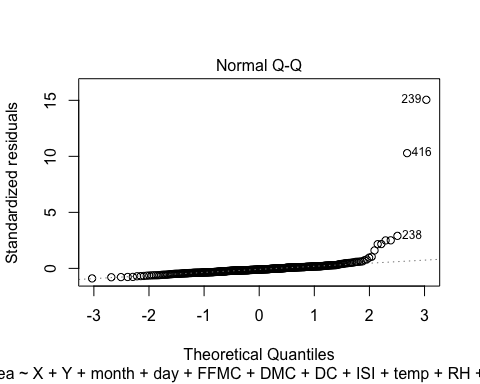
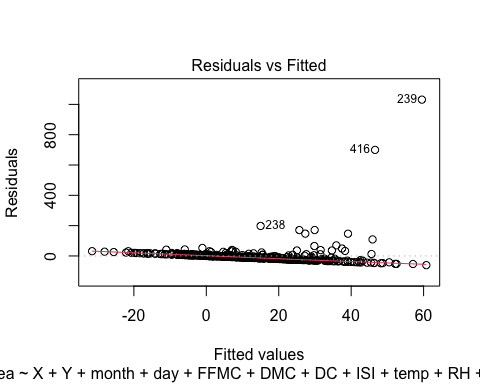
Inspite we have seen from EDA that the linear relationship is not there we will first apply the simple linear model without any modification after that I will apply the log transform as EDA is showing this will give better result so one my one I will apply and try to improve the R2 value as well the accuracy of the final Model

# Model 1

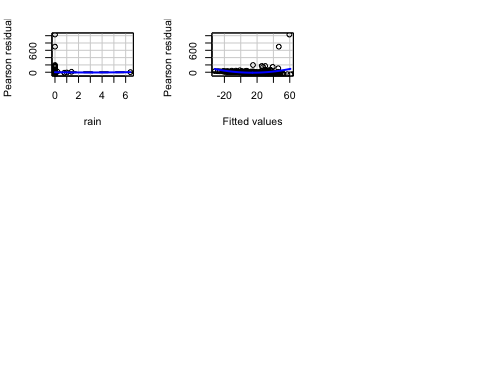
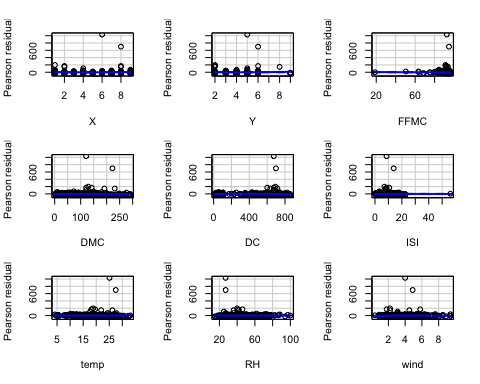
## We run the first Basic Model in this i have taken all the features given in the table

##   
## Call:  
## lm(formula = area ~ X + Y + month + day + FFMC + DMC + DC + ISI +   
## temp + RH + wind + rain, data = forest\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60.75 -18.62 -6.77 6.02 1031.31   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -38.61198 105.84300 -0.365 0.7155   
## X 2.41395 1.82637 1.322 0.1870   
## Y 1.07600 3.39825 0.317 0.7517   
## monthaug 41.21446 47.41200 0.869 0.3852   
## monthdec 42.26925 56.45415 0.749 0.4545   
## monthfeb 4.56170 34.12350 0.134 0.8937   
## monthjan 54.64431 103.94764 0.526 0.5994   
## monthjul 11.47397 42.69595 0.269 0.7883   
## monthjun 3.34894 39.51302 0.085 0.9325   
## monthmar -9.36092 30.69869 -0.305 0.7606   
## monthmay 10.37710 58.63214 0.177 0.8596   
## monthnov -7.77324 76.62911 -0.101 0.9193   
## monthoct 57.40928 57.26152 1.003 0.3167   
## monthsep 69.18937 52.64641 1.314 0.1895   
## daymon 1.84455 13.01591 0.142 0.8874   
## daysat 25.58773 12.30490 2.079 0.0382 \*  
## daysun 7.14457 11.98866 0.596 0.5516   
## daythu 13.85152 13.61441 1.017 0.3096   
## daytue 9.84628 13.23795 0.744 0.4575   
## daywed 4.49411 14.10407 0.319 0.7502   
## FFMC 0.25813 1.10098 0.234 0.8148   
## DMC 0.22324 0.10591 2.108 0.0357 \*  
## DC -0.13034 0.06987 -1.866 0.0629 .  
## ISI -0.64368 0.99822 -0.645 0.5194   
## temp 1.02558 1.27955 0.802 0.4233   
## RH -0.28047 0.37082 -0.756 0.4499   
## wind 1.96535 2.21564 0.887 0.3756   
## rain -2.30308 10.94804 -0.210 0.8335   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 69.57 on 386 degrees of freedom  
## Multiple R-squared: 0.05579, Adjusted R-squared: -0.01025   
## F-statistic: 0.8448 on 27 and 386 DF, p-value: 0.692

The R sq is very low at 5.5% and only DMC and DC seems significant regressors

**Diagnostic Plots** 

There is negative linear relation between Residuals and Fitted values QQ plot of residuals is also not linear. **this indicates there can be Collinearity problems**

**Residual Plots** 

## Test stat Pr(>|Test stat|)   
## X 0.6476 0.5176   
## Y 0.6194 0.5360   
## FFMC 0.3924 0.6950   
## DMC -1.1223 0.2624   
## DC -1.5760 0.1159   
## ISI -0.0561 0.9553   
## temp 0.8661 0.3870   
## RH 0.2239 0.8229   
## wind -0.6186 0.5366   
## rain 0.1291 0.8973   
## Tukey test 5.6224 1.884e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The residual plots suggests very significant pattern for fitted values and residuals. Some square transformations in wind, temp, rain, RH is suggested.

But first we observe that there are many zero values in area which is giving very irregular results.

**FEATURE SELECTION**

## Relative Importances:

## temp X DMC RH Y DC ISI wind FFMC rain   
## 0.006 0.004 0.004 0.004 0.001 0.001 0.001 0.001 0.000 0.000

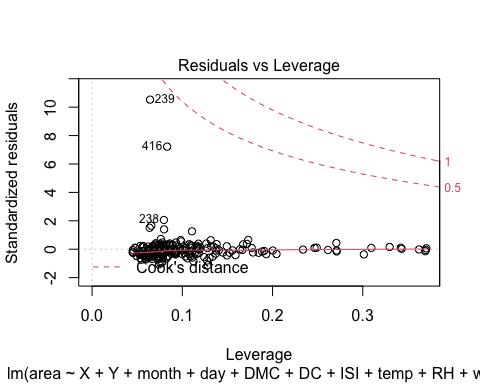
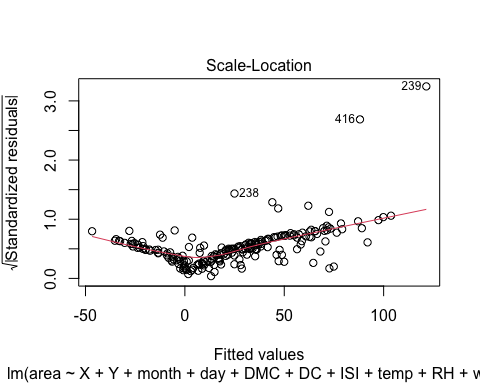
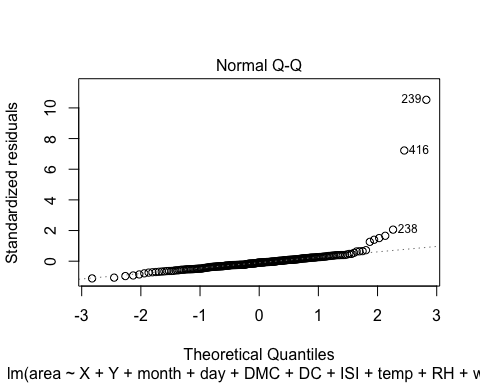
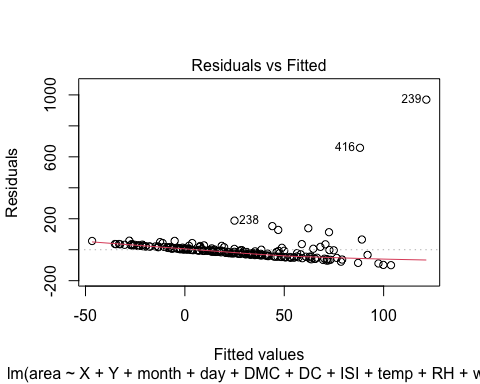
**as the seen from the above the importance of all the features are less so the linear regression is not the better option to fit the data but we can improve this by log transformation**

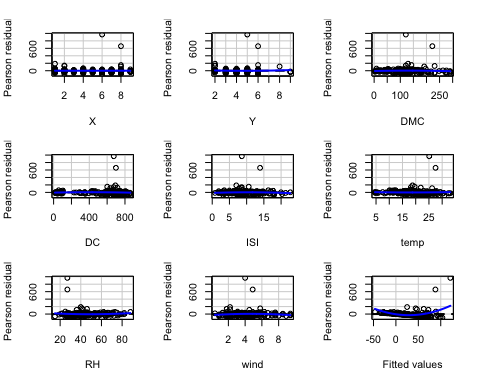
**now we will remove the rain FFMC and rain and let us what will happen**

## Model 2

##   
## Call:  
## lm(formula = area ~ X + Y + month + day + DMC + DC + ISI + temp +   
## RH + wind, data = forest\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -99.75 -29.97 -8.16 11.99 969.37   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -33.4071 84.6906 -0.394 0.6937   
## X 4.8457 3.3443 1.449 0.1490   
## Y 0.6256 6.7218 0.093 0.9259   
## monthaug 115.8738 93.7933 1.235 0.2182   
## monthdec 99.8117 95.5214 1.045 0.2974   
## monthfeb 14.7853 60.3165 0.245 0.8066   
## monthjul 46.7812 80.3632 0.582 0.5612   
## monthjun 27.0996 73.1833 0.370 0.7116   
## monthmar -9.3514 54.8011 -0.171 0.8647   
## monthmay 25.6491 111.1294 0.231 0.8177   
## monthoct 158.3102 120.4913 1.314 0.1905   
## monthsep 182.6161 106.4742 1.715 0.0880 .  
## daymon -4.3030 26.3626 -0.163 0.8705   
## daysat 45.8951 24.9156 1.842 0.0670 .  
## daysun 8.1596 24.4134 0.334 0.7386   
## daythu 25.4539 26.9919 0.943 0.3469   
## daytue 7.4864 25.8380 0.290 0.7723   
## daywed 5.8147 27.5697 0.211 0.8332   
## DMC 0.4633 0.2047 2.263 0.0247 \*  
## DC -0.3303 0.1504 -2.196 0.0294 \*  
## ISI -1.0701 2.2040 -0.486 0.6279   
## temp 2.7184 2.5230 1.077 0.2827   
## RH -0.5431 0.7181 -0.756 0.4504   
## wind 4.1113 4.1455 0.992 0.3226   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 95.17 on 188 degrees of freedom  
## Multiple R-squared: 0.1103, Adjusted R-squared: 0.001445   
## F-statistic: 1.013 on 23 and 188 DF, p-value: 0.4508

in this we have remove the zero value rows and reduce the data set and actually run only on data where there is a burn area. we run model 2 on reduced subset of data but having all the features which are given in the question R sq has significantly improved 11.15 %

**Diagnostic Plots** 

**Residual Plots** 

## Test stat Pr(>|Test stat|)   
## X 0.0249 0.9802   
## Y 0.5117 0.6095   
## DMC -0.9818 0.3275   
## DC -1.4277 0.1551   
## ISI -0.6950 0.4879   
## temp 0.4686 0.6399   
## RH 0.8527 0.3949   
## wind -1.1753 0.2414   
## Tukey test 7.0513 1.772e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# the fitted values plot and QQ plot has improved, now we proceed with other transformations

# Model 3

**Residual Plot** **Diagnostic Plots**

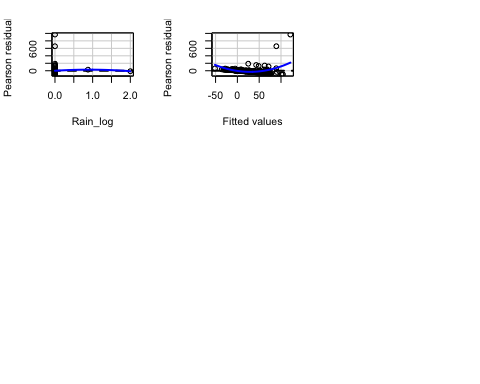
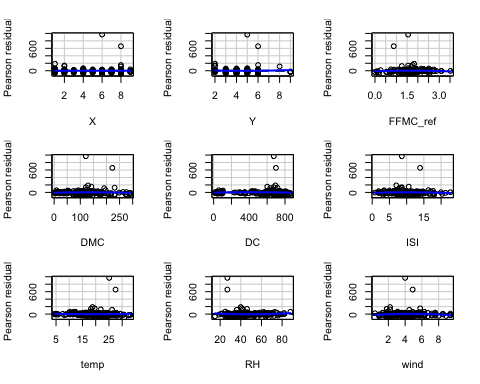
##   
## Call:  
## lm(formula = area ~ X + Y + month + day + FFMC\_ref + DMC + DC +   
## ISI + temp + RH + wind + Rain\_log, data = forest\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -99.17 -30.30 -7.58 12.73 969.34   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -14.6600 107.8415 -0.136 0.8920   
## X 4.9274 3.3702 1.462 0.1454   
## Y 0.6309 6.7582 0.093 0.9257   
## monthaug 112.6104 94.8394 1.187 0.2366   
## monthdec 97.2123 96.6271 1.006 0.3157   
## monthfeb 14.9785 60.6429 0.247 0.8052   
## monthjul 44.1689 81.4053 0.543 0.5881   
## monthjun 25.1029 73.9745 0.339 0.7347   
## monthmar -10.2425 55.1434 -0.186 0.8528   
## monthmay 24.5369 111.7792 0.220 0.8265   
## monthoct 154.7677 121.5446 1.273 0.2045   
## monthsep 178.9410 107.6419 1.662 0.0981 .  
## daymon -4.1637 26.5460 -0.157 0.8755   
## daysat 45.5726 25.0806 1.817 0.0708 .  
## daysun 8.3409 24.6986 0.338 0.7360   
## daythu 24.8622 27.1822 0.915 0.3616   
## daytue 8.1601 26.0629 0.313 0.7546   
## daywed 5.1507 27.7710 0.185 0.8531   
## FFMC\_ref -7.8292 23.9940 -0.326 0.7446   
## DMC 0.4512 0.2086 2.163 0.0318 \*  
## DC -0.3253 0.1519 -2.142 0.0335 \*  
## ISI -1.7005 2.9214 -0.582 0.5612   
## temp 2.6519 2.6598 0.997 0.3201   
## RH -0.5020 0.7463 -0.673 0.5020   
## wind 4.3596 4.2247 1.032 0.3034   
## Rain\_log -10.4013 48.4518 -0.215 0.8303   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 95.65 on 186 degrees of freedom  
## Multiple R-squared: 0.1109, Adjusted R-squared: -0.008607   
## F-statistic: 0.928 on 25 and 186 DF, p-value: 0.5667

we did the log transform of the FFMC and Rain because we conclude that after doing EDA

**FEATURE IMPORTANCE**

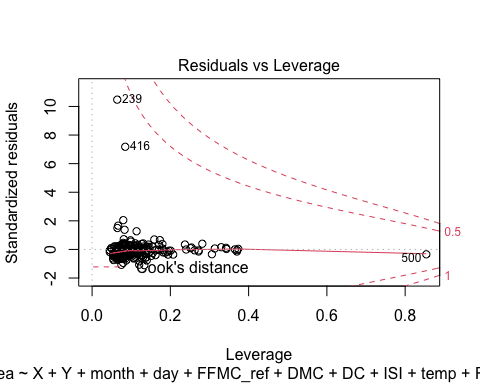
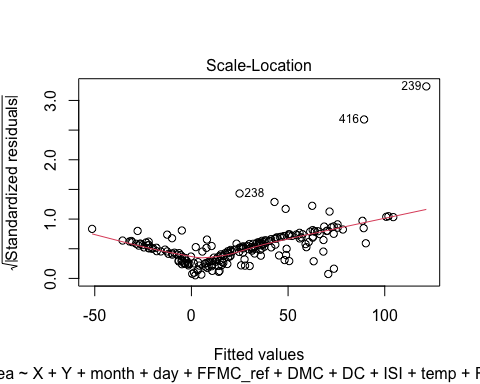
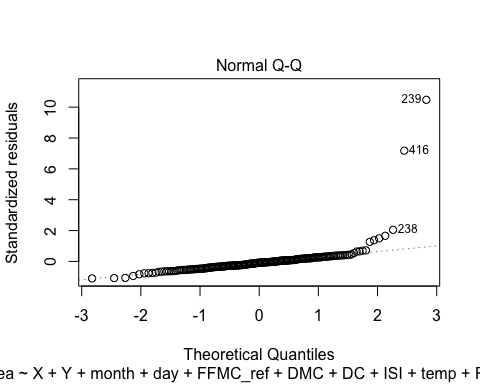
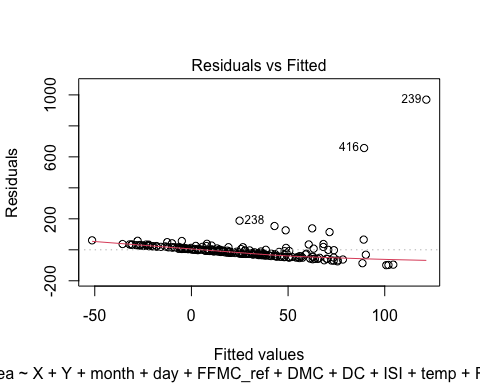
## Relative Importances:

## RH DMC temp FFMC\_ref ISI DC wind Rain\_log   
## 0.014 0.006 0.006 0.002 0.002 0.001 0.001 0.000

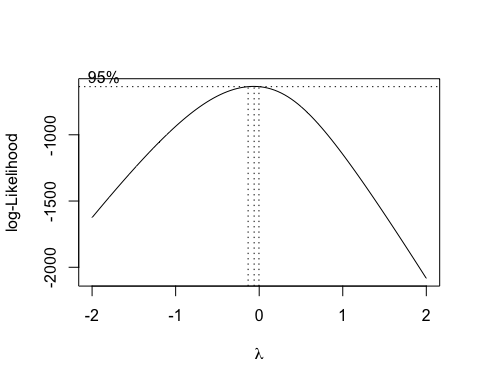
**Residual Plot** 

## Test stat Pr(>|Test stat|)   
## X 0.0746 0.9406   
## Y 0.5250 0.6002   
## FFMC\_ref -0.5856 0.5589   
## DMC -0.9457 0.3455   
## DC -1.3968 0.1641   
## ISI -0.6369 0.5250   
## temp 0.4339 0.6649   
## RH 0.8371 0.4036   
## wind -1.1452 0.2536   
## Rain\_log -0.3419 0.7328   
## Tukey test 6.9591 3.423e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## the model is improved on R sq and residuals also.DMC and DC have emerged as significant

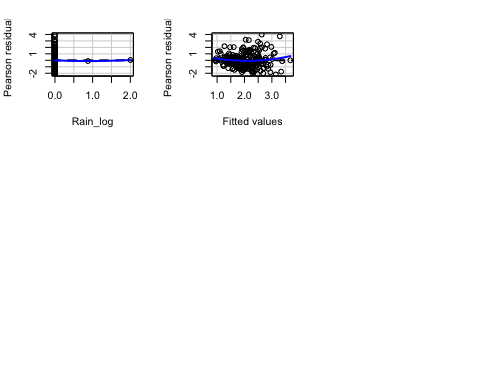
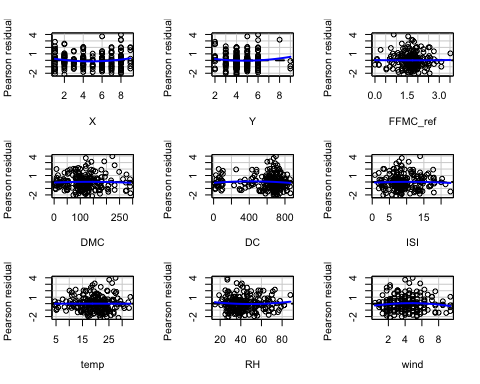
**Diagnostic Plots**  We still need to improve patterns in fitted values and residuals.

## Model 4

**Residual Plot** **Diagnostic Plots** 

## [1] -0.06060606

##   
## Call:  
## lm(formula = log(area + 1) ~ X + Y + month + day + FFMC\_ref +   
## DMC + DC + ISI + temp + RH + wind + Rain\_log, data = forest\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1848 -0.8309 -0.1009 0.4934 3.9868   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.021539 1.359333 2.223 0.02743 \*   
## X 0.062177 0.042482 1.464 0.14498   
## Y -0.045320 0.085186 -0.532 0.59536   
## monthaug 0.726861 1.195443 0.608 0.54391   
## monthdec 1.819511 1.217977 1.494 0.13690   
## monthfeb -0.227503 0.764399 -0.298 0.76632   
## monthjul -0.222130 1.026107 -0.216 0.82885   
## monthjun -0.456538 0.932443 -0.490 0.62498   
## monthmar -0.295299 0.695078 -0.425 0.67144   
## monthmay 1.160481 1.408968 0.824 0.41120   
## monthoct 2.572316 1.532060 1.679 0.09483 .   
## monthsep 1.879358 1.356817 1.385 0.16767   
## daymon -0.237153 0.334610 -0.709 0.47937   
## daysat 0.585469 0.316140 1.852 0.06562 .   
## daysun 0.276868 0.311324 0.889 0.37498   
## daythu 0.185932 0.342630 0.543 0.58801   
## daytue 0.299868 0.328521 0.913 0.36254   
## daywed -0.206063 0.350052 -0.589 0.55680   
## FFMC\_ref -0.301008 0.302442 -0.995 0.32090   
## DMC 0.007985 0.002629 3.037 0.00274 \*\*  
## DC -0.004604 0.001914 -2.405 0.01713 \*   
## ISI -0.033315 0.036824 -0.905 0.36679   
## temp 0.021487 0.033527 0.641 0.52239   
## RH -0.005938 0.009407 -0.631 0.52865   
## wind 0.042177 0.053252 0.792 0.42935   
## Rain\_log -0.052836 0.610731 -0.087 0.93115   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.206 on 186 degrees of freedom  
## Multiple R-squared: 0.1799, Adjusted R-squared: 0.06971   
## F-statistic: 1.632 on 25 and 186 DF, p-value: 0.03587



## Test stat Pr(>|Test stat|)   
## X 1.9013 0.05882 .  
## Y 1.2159 0.22558   
## FFMC\_ref 0.1209 0.90389   
## DMC -0.8522 0.39520   
## DC -2.4454 0.01540 \*  
## ISI -0.4464 0.65583   
## temp -0.0103 0.99178   
## RH 0.7860 0.43285   
## wind -1.2524 0.21200   
## Rain\_log 0.1061 0.91563   
## Tukey test 1.6733 0.09427 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The fitted values vs residual is random now and model prediction is also improved. Let’s transform other variables. #R-square value=0.1799

## Model 5

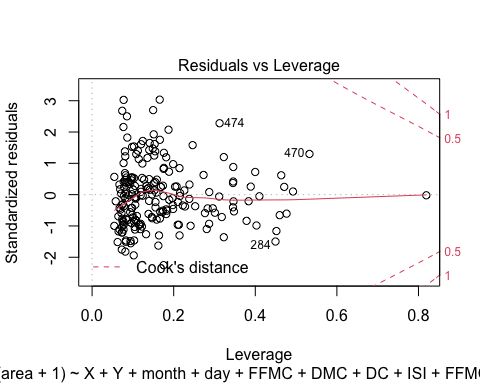
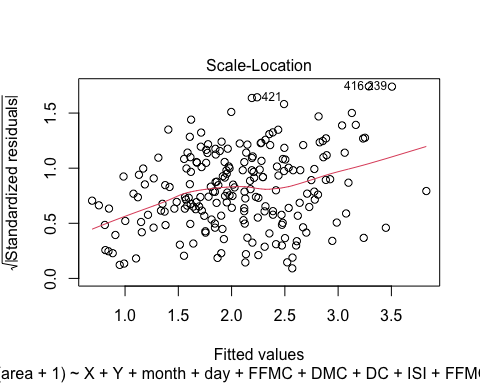
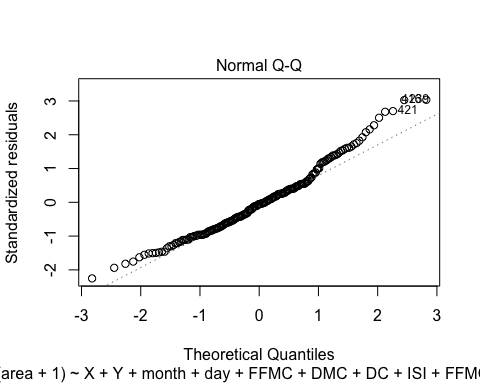
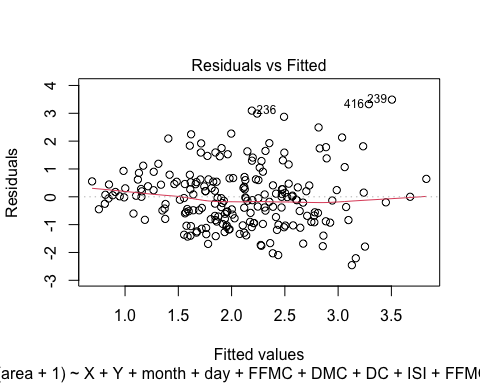
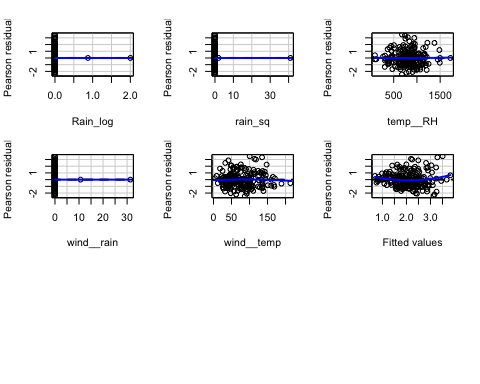
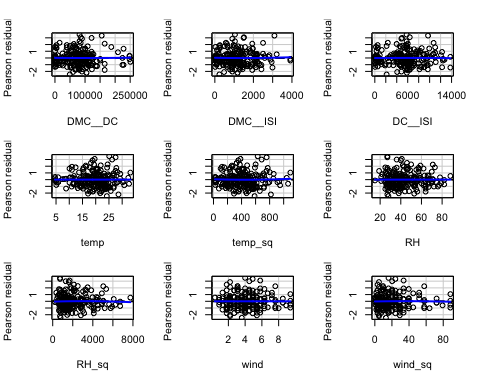
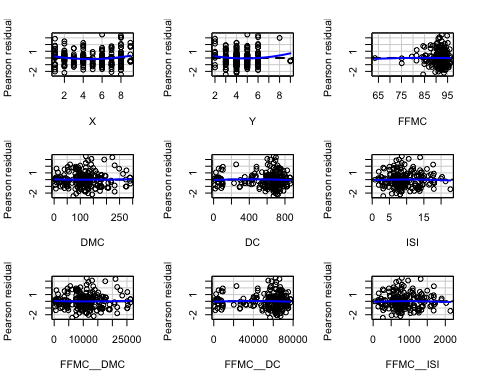
**FINAL MODEL** **Residual Plot** **Diagnostic Plots**

##   
## Call:  
## lm(formula = log(area + 1) ~ X + Y + month + day + FFMC + DMC +   
## DC + ISI + FFMC\_\_DMC + FFMC\_\_DC + FFMC\_\_ISI + DMC\_\_DC + DMC\_\_ISI +   
## DC\_\_ISI + temp + temp\_sq + RH + RH\_sq + wind + wind\_sq +   
## Rain\_log + rain\_sq + temp\_\_RH + wind\_\_rain + wind\_\_temp,   
## data = forest\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.4561 -0.7971 -0.0498 0.5414 3.4904   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.022e+00 9.984e+00 -0.804 0.4228   
## X 5.828e-02 4.380e-02 1.330 0.1851   
## Y -5.832e-02 8.679e-02 -0.672 0.5025   
## monthaug -1.469e+00 1.575e+00 -0.932 0.3524   
## monthdec 1.488e+00 2.022e+00 0.736 0.4628   
## monthfeb 1.279e-01 8.542e-01 0.150 0.8812   
## monthjul -2.139e+00 1.383e+00 -1.547 0.1238   
## monthjun -2.173e+00 1.291e+00 -1.684 0.0941 .  
## monthmar -1.240e+00 8.004e-01 -1.549 0.1232   
## monthmay 4.694e-01 1.497e+00 0.314 0.7542   
## monthoct 2.688e-01 1.825e+00 0.147 0.8830   
## monthsep -2.028e-01 1.703e+00 -0.119 0.9054   
## daymon -3.765e-01 3.440e-01 -1.095 0.2752   
## daysat 6.280e-01 3.221e-01 1.950 0.0528 .  
## daysun 2.226e-01 3.194e-01 0.697 0.4868   
## daythu 1.930e-01 3.559e-01 0.542 0.5883   
## daytue 2.280e-01 3.439e-01 0.663 0.5081   
## daywed -1.997e-01 3.607e-01 -0.554 0.5805   
## FFMC 9.643e-02 1.086e-01 0.888 0.3757   
## DMC -1.430e-01 1.120e-01 -1.277 0.2032   
## DC 3.350e-02 2.021e-02 1.657 0.0992 .  
## ISI 2.671e-01 9.211e-01 0.290 0.7722   
## FFMC\_\_DMC 1.867e-03 1.237e-03 1.510 0.1329   
## FFMC\_\_DC -3.801e-04 2.331e-04 -1.630 0.1049   
## FFMC\_\_ISI -2.431e-03 9.634e-03 -0.252 0.8011   
## DMC\_\_DC -2.455e-05 1.585e-05 -1.549 0.1232   
## DMC\_\_ISI 6.756e-06 7.943e-04 0.009 0.9932   
## DC\_\_ISI -1.270e-04 2.282e-04 -0.557 0.5785   
## temp 1.108e-01 2.705e-01 0.410 0.6827   
## temp\_sq -1.433e-03 4.669e-03 -0.307 0.7592   
## RH 9.288e-03 7.948e-02 0.117 0.9071   
## RH\_sq 1.087e-04 4.831e-04 0.225 0.8223   
## wind 3.307e-01 3.085e-01 1.072 0.2852   
## wind\_sq -3.651e-02 2.201e-02 -1.659 0.0989 .  
## Rain\_log 5.043e-01 1.734e+00 0.291 0.7716   
## rain\_sq -1.677e-02 8.855e-02 -0.189 0.8500   
## temp\_\_RH -1.601e-03 2.336e-03 -0.685 0.4940   
## wind\_\_rain NA NA NA NA   
## wind\_\_temp 1.550e-03 1.016e-02 0.153 0.8789   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.201 on 174 degrees of freedom  
## Multiple R-squared: 0.2392, Adjusted R-squared: 0.07743   
## F-statistic: 1.479 on 37 and 174 DF, p-value: 0.05027

## in this we did residual improvement of other variables. Hence we try the square of variables

## which are dense distributed and show some quadratic pattern.Also factors like wind, temp

## should have greater impact on fire spread and area

**Diagnostic Plots**  **Residual Plot** 

## Test stat Pr(>|Test stat|)   
## X 2.1110 0.03621 \*  
## Y 1.6252 0.10594   
## FFMC -0.3685 0.71296   
## DMC 0.6606 0.50974   
## DC -1.4151 0.15884   
## ISI -0.5387 0.59079   
## FFMC\_\_DMC 0.8187 0.41408   
## FFMC\_\_DC -1.1554 0.24950   
## FFMC\_\_ISI -0.4813 0.63093   
## DMC\_\_DC 0.3757 0.70760   
## DMC\_\_ISI 0.5062 0.61335   
## DC\_\_ISI -0.2107 0.83338   
## temp -0.9371 0.35001   
## temp\_sq 1.0366 0.30135   
## RH -0.8656 0.38790   
## RH\_sq -0.7541 0.45183   
## wind -0.7199 0.47256   
## wind\_sq -0.5096 0.61101   
## Rain\_log 0.5284 0.59790   
## rain\_sq -0.7944 0.42807   
## temp\_\_RH 0.5049 0.61424   
## wind\_\_rain 1.2527 0.21201   
## wind\_\_temp -0.8062 0.42121   
## Tukey test 1.9441 0.05188 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## the model is improved in terms of r sq and the residuals are also randomly distributed.

## We are at optimal model.

## X Y month day FFMC DMC DC ISI temp RH wind rain area  
## 200 2 4 sep mon 63.5 70.8 665.3 0.8 22.6 38 3.6 0 11.32  
## 363 7 4 sep fri 88.2 55.2 732.3 11.6 15.2 64 3.1 0 0.52  
## 416 8 6 aug thu 94.8 222.4 698.6 13.9 27.5 27 4.9 0 746.28  
## 479 7 4 jul sun 93.7 101.3 423.4 14.7 18.2 82 4.5 0 2.21

## out f these only two are high outliers 416 and and 480. These seem to be burned

# due to some other factors, may be intentional !! So we remove two observations.

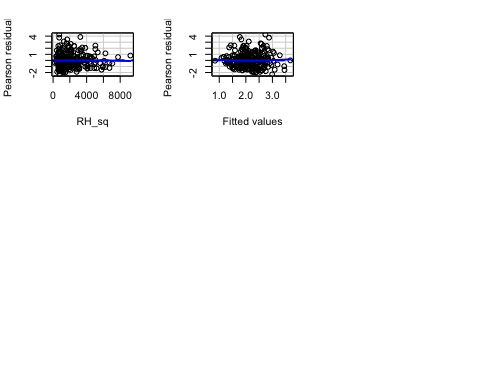
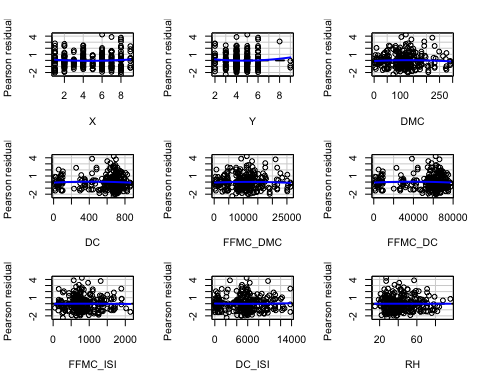
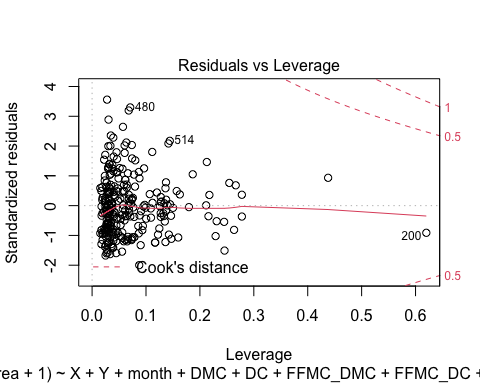
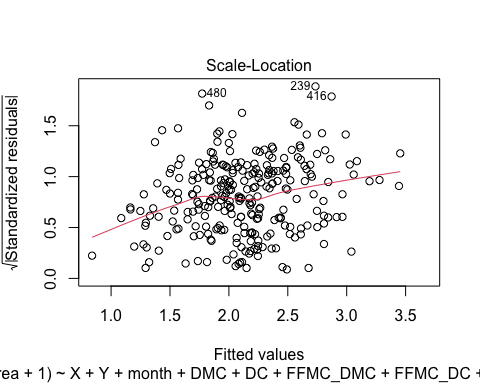
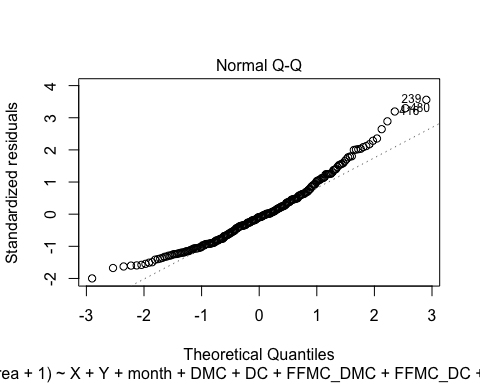
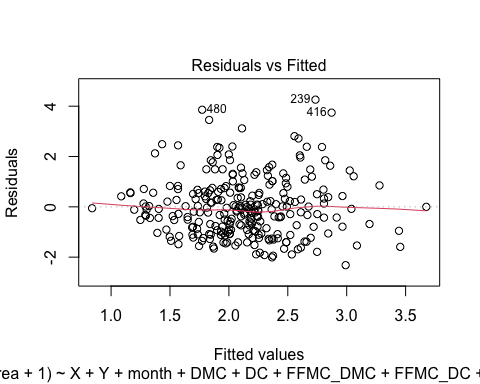
## Now we test the model

## [1] 0.3240162

##The predicted accuracy is 32.40% This is a considerably good fit as per the given data set.

## Running on original data

##   
## Call:  
## lm(formula = log(area + 1) ~ X + Y + month + DMC + DC + FFMC\_DMC +   
## FFMC\_DC + FFMC\_ISI + DC\_ISI + RH + RH\_sq, data = forest\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3183 -0.9042 -0.1093 0.6017 4.2608   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.3196163 1.0742334 3.090 0.00223 \*\*  
## X 0.0567522 0.0379537 1.495 0.13610   
## Y -0.1058543 0.0765299 -1.383 0.16784   
## monthaug 0.7700554 1.0885343 0.707 0.47996   
## monthdec 1.0513579 0.9107171 1.154 0.24943   
## monthfeb -0.1659198 0.7430294 -0.223 0.82348   
## monthjul 0.2423118 0.9171080 0.264 0.79183   
## monthjun -0.3465654 0.8764909 -0.395 0.69288   
## monthmar -0.4656705 0.6892276 -0.676 0.49989   
## monthmay 1.1707608 1.3635455 0.859 0.39138   
## monthoct 2.6457688 1.3411918 1.973 0.04963 \*   
## monthsep 1.8960656 1.2353813 1.535 0.12610   
## DMC -0.1477903 0.0611750 -2.416 0.01642 \*   
## DC 0.0176931 0.0092727 1.908 0.05752 .   
## FFMC\_DMC 0.0017165 0.0006717 2.556 0.01119 \*   
## FFMC\_DC -0.0002305 0.0001021 -2.258 0.02482 \*   
## FFMC\_ISI 0.0007411 0.0008568 0.865 0.38787   
## DC\_ISI -0.0001700 0.0001336 -1.272 0.20446   
## RH -0.0384053 0.0272716 -1.408 0.16030   
## RH\_sq 0.0003568 0.0002694 1.324 0.18657   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.214 on 250 degrees of freedom  
## Multiple R-squared: 0.1332, Adjusted R-squared: 0.0673   
## F-statistic: 2.022 on 19 and 250 DF, p-value: 0.00811



## Test stat Pr(>|Test stat|)   
## X 0.9252 0.35578   
## Y 1.0905 0.27653   
## DMC -1.1109 0.26770   
## DC -1.8439 0.06639 .  
## FFMC\_DMC -0.9931 0.32163   
## FFMC\_DC -1.5923 0.11259   
## FFMC\_ISI -0.1500 0.88091   
## DC\_ISI 0.5368 0.59186   
## RH -0.1024 0.91856   
## RH\_sq -0.4534 0.65065   
## Tukey test 0.4817 0.63001   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Interpretations and interesting findings**

1. *in this we do log transform of some features as shown in above models and also there are many outliers so we remove them and when i am making the model without any transformation then the result which is coming is very poor i.e r square value is very less*
2. *one more interpretation or finding is that, data which i initially assumed that it will best fit the regression line , this hypothesis or assumption is proved to be wrong because r2 square by all transformations and complex part increase to .32 only*

**Additional work such as analysis and models that you tried but did not include in the final model**

1. *I tried to remove the different features without log transformation but the result is coming out to be worst than my final model*
2. *I tried many classical model on this data they are giving best result compared to MLR*