Lab 2: Exploratory Data Analysis and Causal Model Bulding

Team 2: Savita Chari, Denny Lehman, Tymon Silva

December 7, 2021

```
library(tidyverse)
library(dplyr)
library(lmtest)
library(sandwich)
library(stargazer)
library(magrittr)
library(sandwich)
library(gridExtra)
library(funModeling)
library(cowplot)
library(MASS)
library(ggplot2)
library(car)
```

Data

```
fire_raw <- read_csv(file = '../src/data/forestfires.csv')


# add BUI variable
BUI_less <- 0.8*((fire_raw$DMC*fire_raw$DC)/(fire_raw$DMC+(0.4*fire_raw$DC)))
BUI_great <- fire_raw$DMC-(1-((0.8*fire_raw$DC)/(fire_raw$DMC+0.4*fire_raw$DC)))*(0.92+(0.0114*fire_raw$

fire_data <- fire_raw %>%
  mutate(
    BUI = case_when(
      DMC <= 0.4*DC ~ BUI_less,
      DMC > 0.4*DC ~ BUI_great,
    )
  )

# by adding BUI variable, we will remove DMC and DC variables that were used in
# the BUI calculation
fire_data <- fire_data %>%
  dplyr::select(wind, rain, temp, RH, FFMC, BUI, area)
```

First look

```
summary(fire_raw)
```

```
##        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
```

Divide the data

```
# Split the data into training and testing sets
# We split the data into an EDA and a Prod dataset because we had a large enough dataset.
# We kept 30% data for EDA set and 70% for the Prod dataset.

sample_size = floor(0.7*nrow(fire_data))
set.seed(777)

# randomly split data in r
picked = sample(seq_len(nrow(fire_data)),size = sample_size)
Prod = fire_data[picked,] # testing data set
print("Dimension of testing data set row/column")
```

```
## [1] "Dimension of testing data set row/column"
```

```
dim(Prod)
```

```
## [1] 361   7
```

```
print("Dimension of training data set row/column")
```

```
## [1] "Dimension of training data set row/column"
```

```
EDA = fire_data[-picked,] # training data set
dim(EDA)
```

```
## [1] 156   7
```

EDA section

EDA Part I - Insight into variables available to us

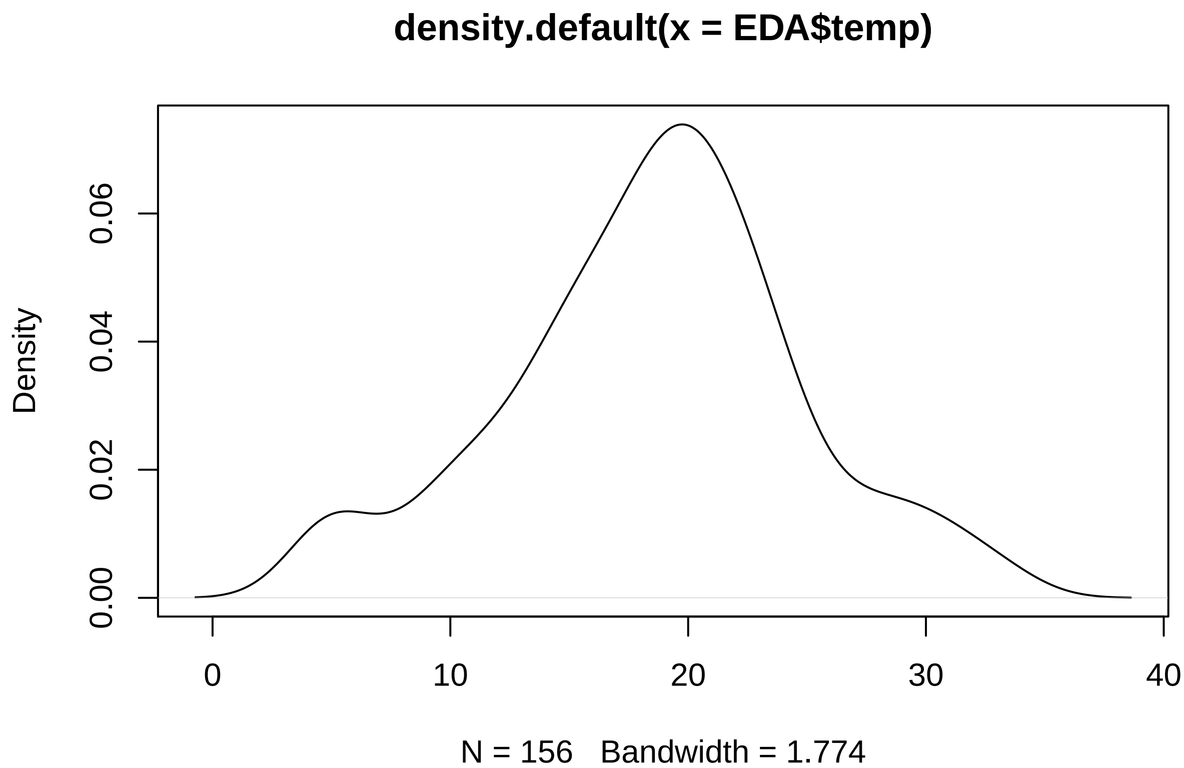
Behavior of individual variable

```
describe(EDA)
```

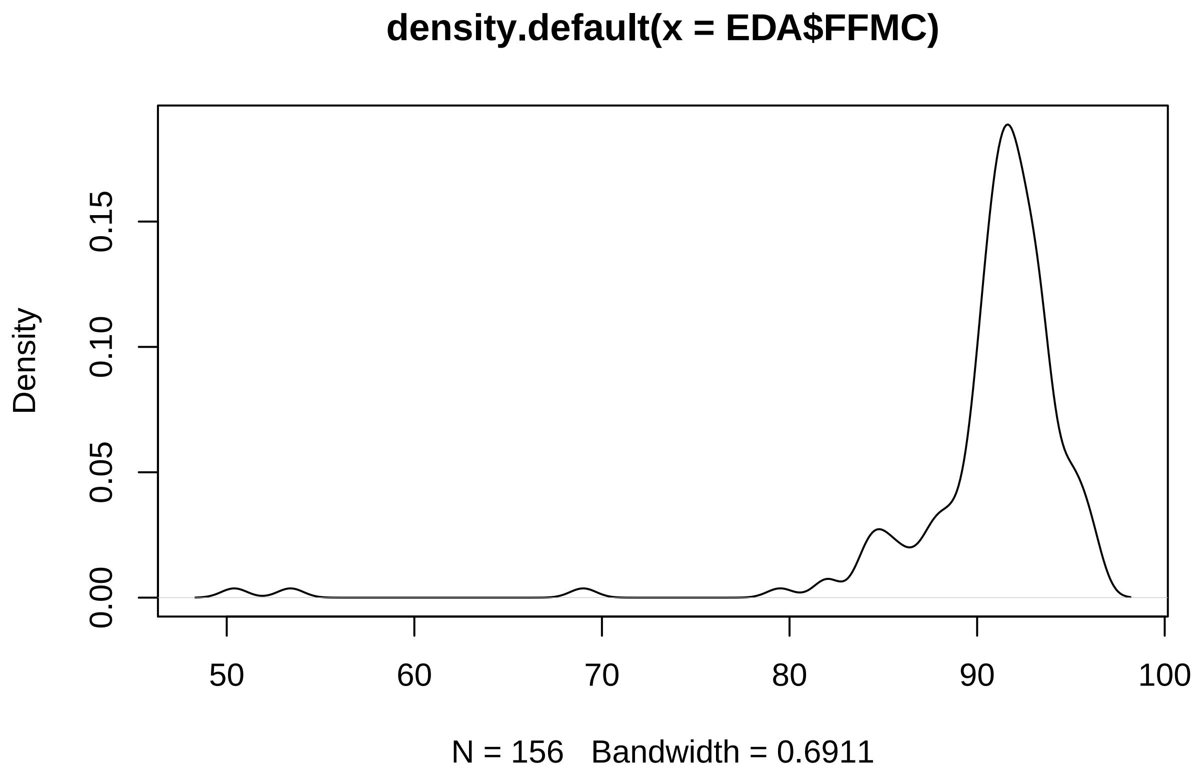
```
## EDA
##
##  7  Variables      156  Observations
## --------------------------------------------------------------------------------
## wind
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      156        0       19    0.992    4.153    1.994      1.3      2.2
##      .25      .50      .75      .90      .95
##      2.7      4.0      5.4      6.3      7.6
##
## lowest : 0.9 1.3 1.8 2.2 2.7, highest: 7.2 7.6 8.0 8.5 9.4
##
## Value        0.9   1.3   1.8   2.2   2.7   3.1   3.6   4.0   4.5   4.9   5.4
## Frequency      6     3     4    12    20    11    12    16    15    15    16
## Proportion 0.038 0.019 0.026 0.077 0.128 0.071 0.077 0.103 0.096 0.096 0.103
##
## Value        5.8   6.3   6.7   7.2   7.6   8.0   8.5   9.4
## Frequency      5     8     1     2     4     1     4     1
## Proportion 0.032 0.051 0.006 0.013 0.026 0.006 0.026 0.006
## --------------------------------------------------------------------------------
## rain
##        n  missing distinct     Info     Mean      Gmd
##      156        0        3    0.038  0.01154  0.02294
##
## Value        0.0   0.8   1.0
## Frequency    154     1     1
```

```
## Proportion 0.987 0.006 0.006
## ---------------------------------------------------------------------------
## temp
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      156        0      106        1    18.55     6.96     6.40    10.40
##      .25      .50      .75      .90      .95
##    14.70    19.10    21.95    26.30    29.07
##
## lowest :  4.6  4.8  5.1  5.3  5.5, highest: 30.8 31.0 32.3 33.1 33.3
## ---------------------------------------------------------------------------
## RH
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      156        0       54    0.999    43.62    18.25    23.50    25.00
##      .25      .50      .75      .90      .95
##    30.75    41.50    53.25    66.00    75.00
##
## lowest : 15 19 21 22 24, highest: 77 78 79 82 90
## ---------------------------------------------------------------------------
## FFMC
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      156        0       67    0.999    90.52    4.331    84.18    86.05
##      .25      .50      .75      .90      .95
##    90.20    91.60    93.03    94.55    95.12
##
## lowest : 50.4 53.4 69.0 79.5 81.9, highest: 95.1 95.2 95.9 96.0 96.1
## ---------------------------------------------------------------------------
## BUI
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      156        0      112        1    139.4    85.33    15.02    27.23
##      .25      .50      .75      .90      .95
##    75.41   152.71   184.21   216.30   264.95
##
## lowest :   3.077922   3.460465   3.709973   4.533333   8.297248
## highest: 272.394414 294.624433 300.788508 311.260078 313.911283
## ---------------------------------------------------------------------------
## area
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      156        0       83    0.906    9.624    15.54    0.000    0.000
##      .25      .50      .75      .90      .95
##    0.000    1.225    8.613   25.180   47.368
##
## lowest :   0.00   0.21   0.24   0.47   1.01, highest:  86.45  88.49  95.18 103.39 154.88
## ---------------------------------------------------------------------------
```

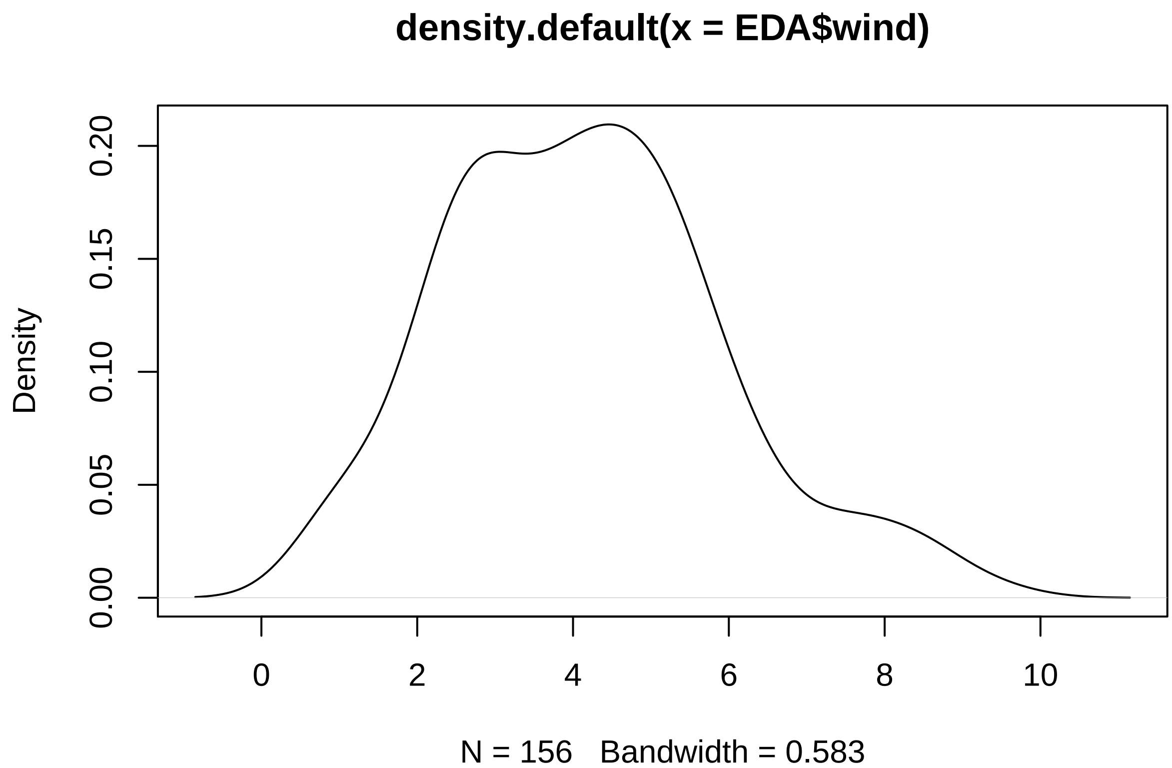
```r
# Get the basic density curve for the variables of interest to see the distribution of their values

plot(density(EDA$temp))
```



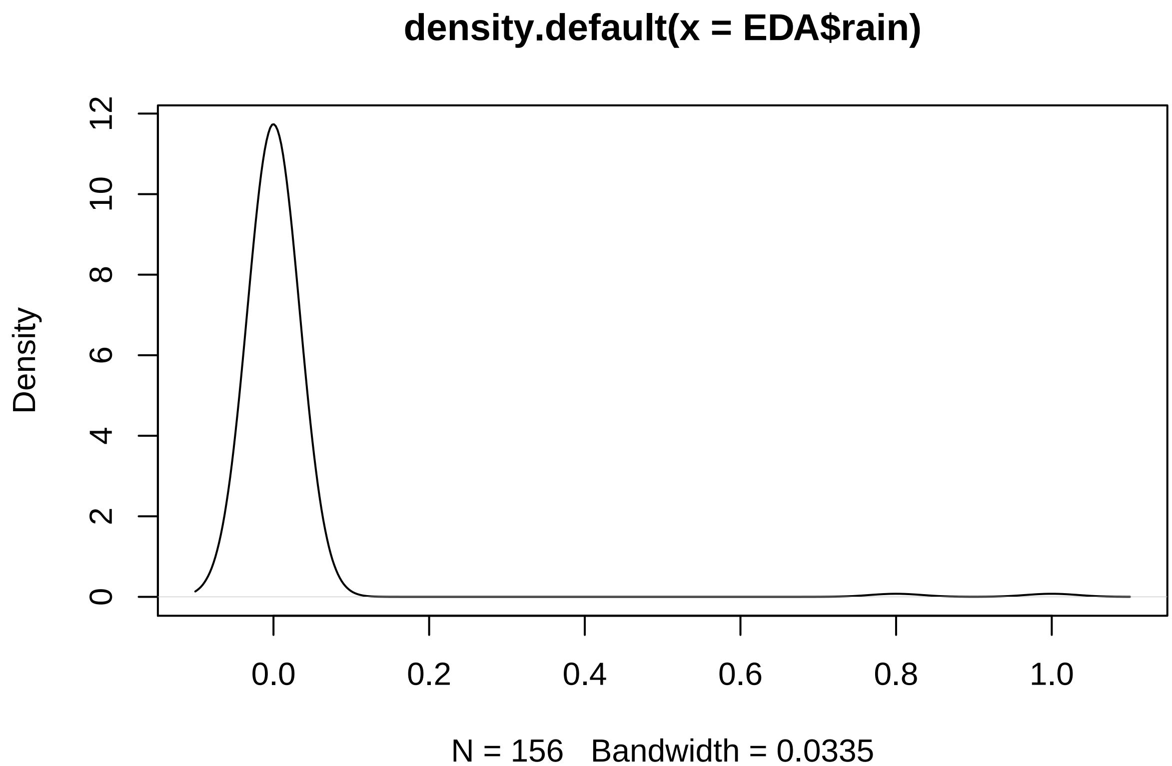

```
plot(density(EDA$FFMC))
```



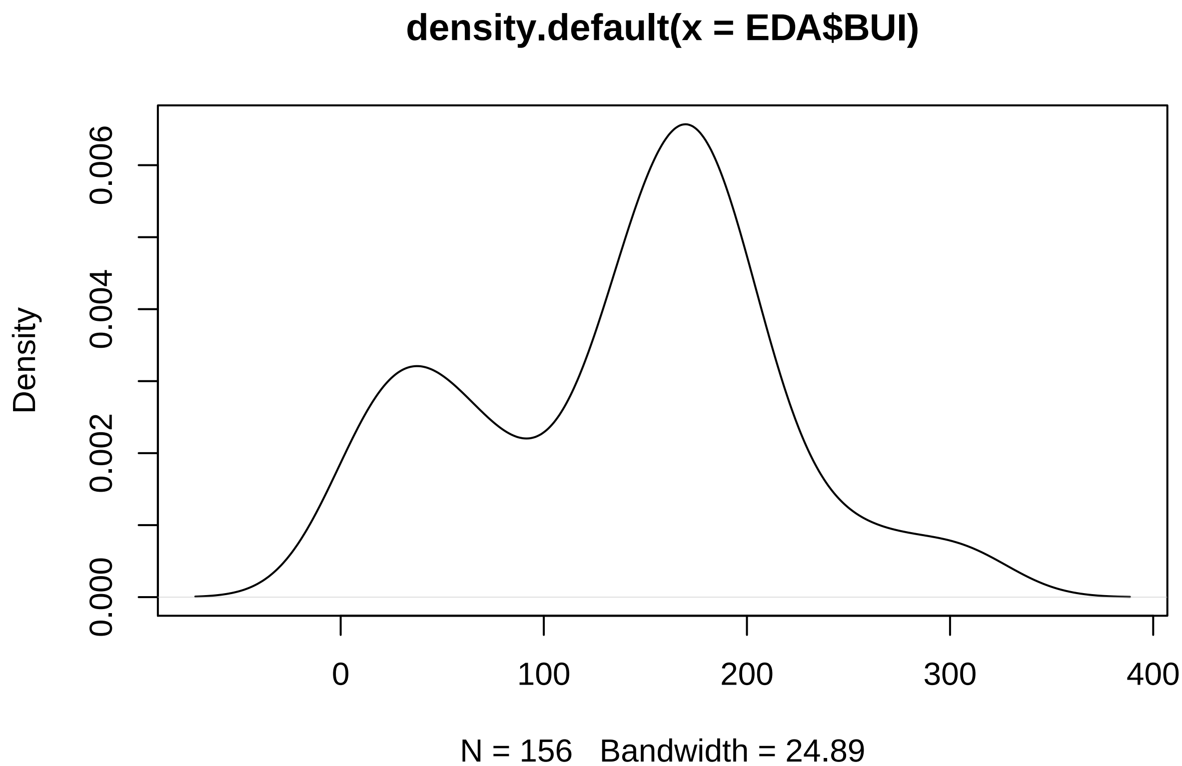

```
plot(density(EDA$wind))
```



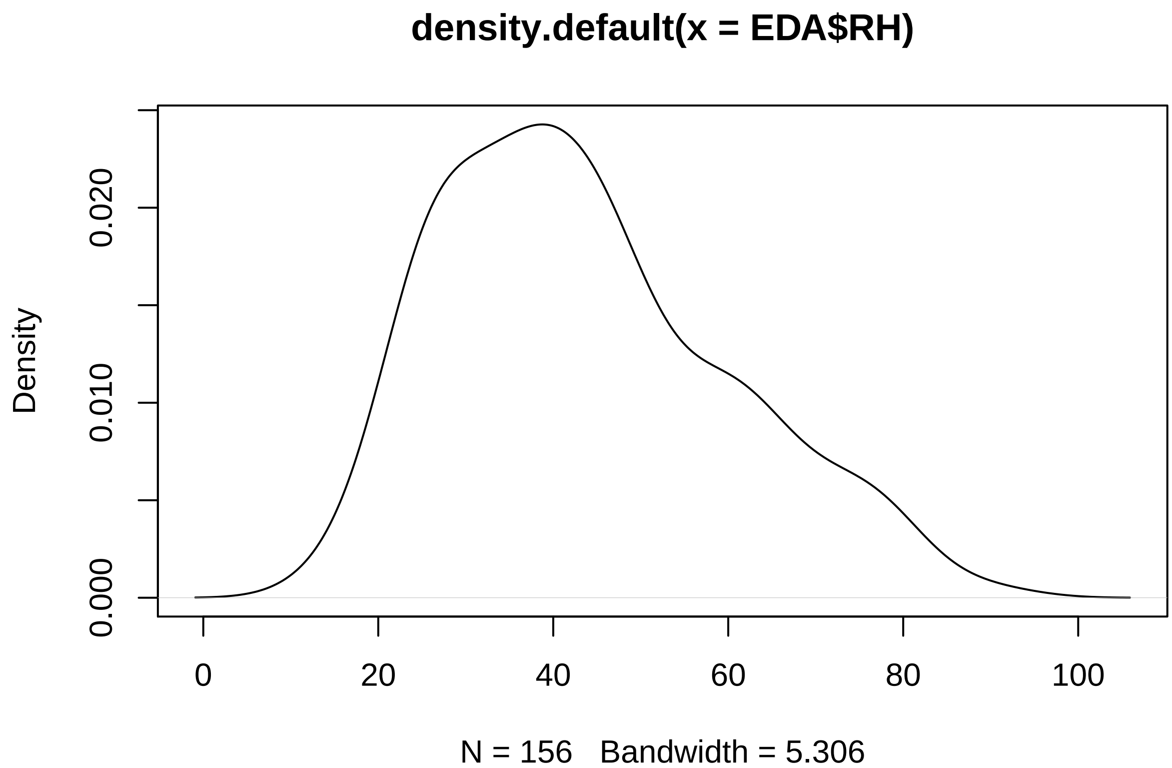

```
plot(density(EDA$rain))
```



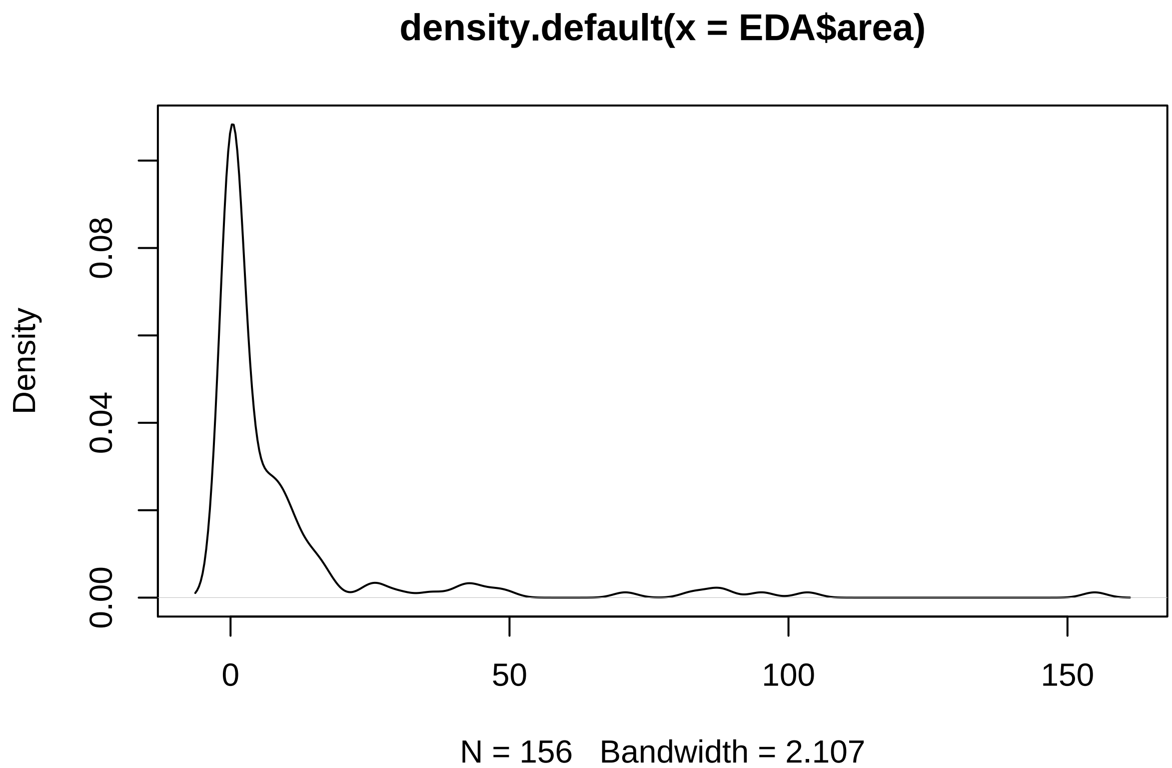

```
plot(density(EDA$BUI))
```



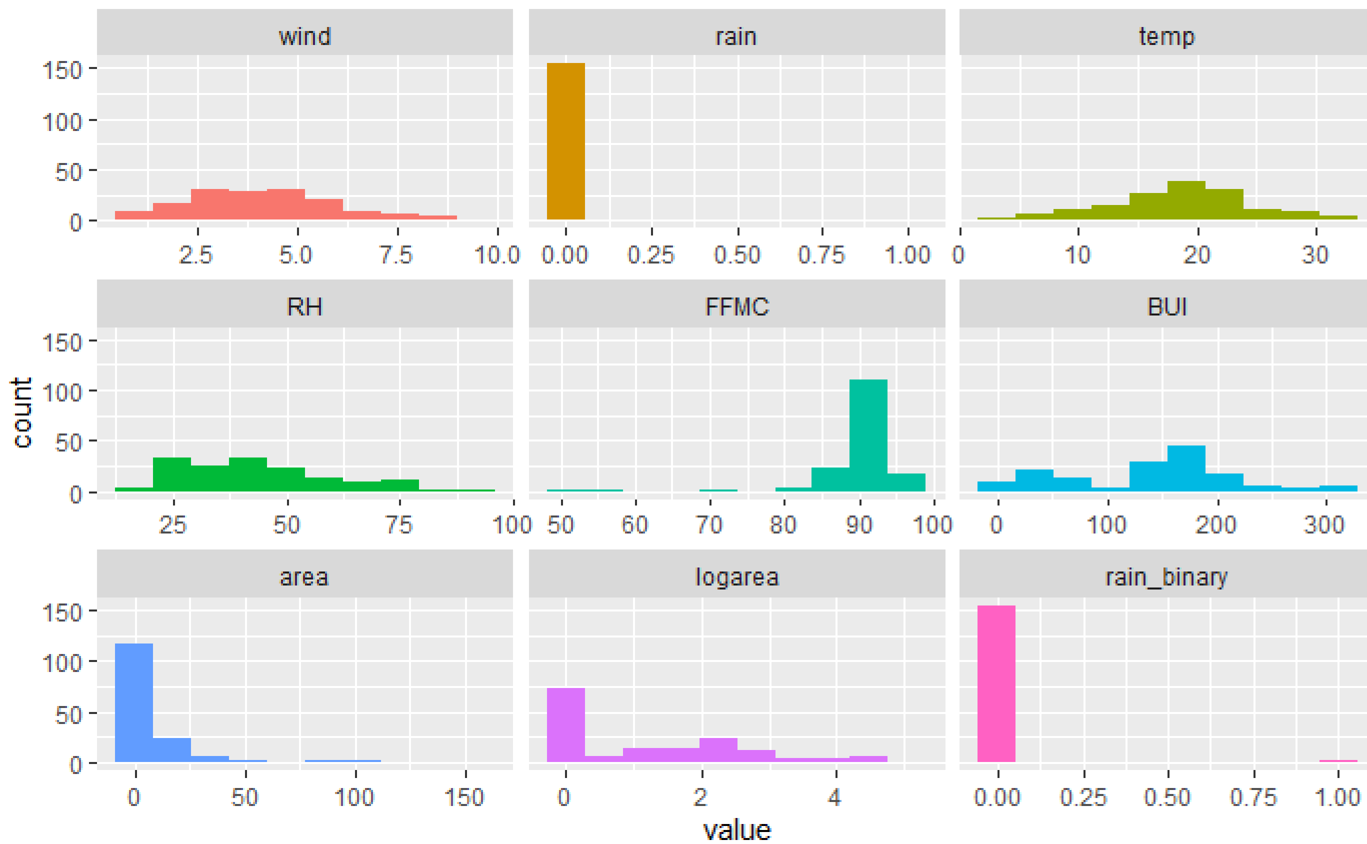

```
plot(density(EDA$RH))
```




```
plot(density(EDA$area))
```

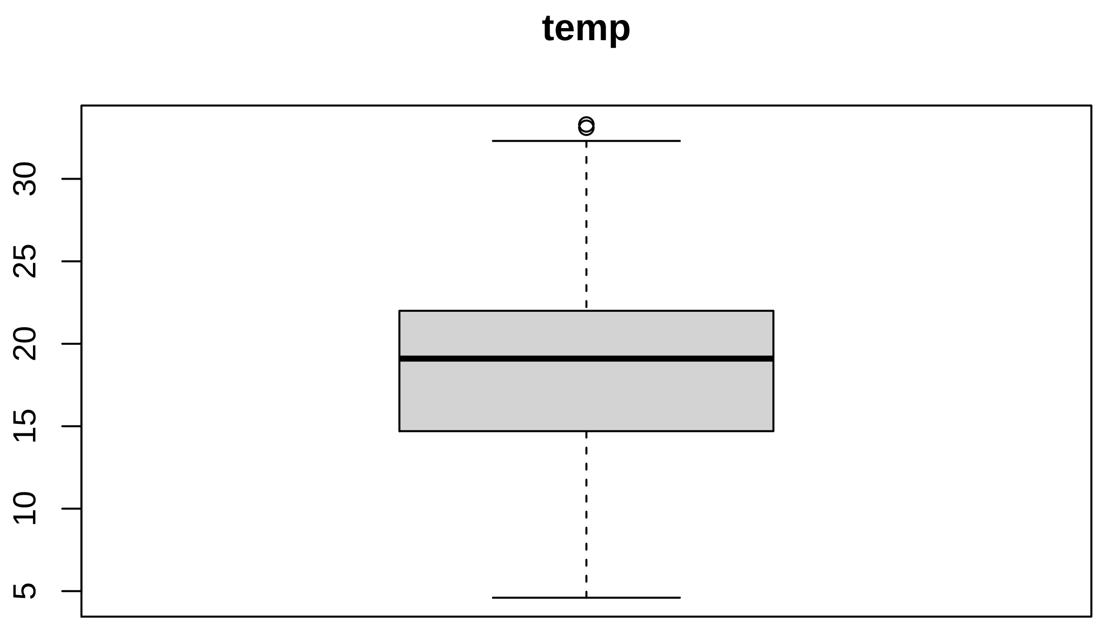


Observation of distribution Normal : Temp,FFMC, Wind, RH Some what Normal : FFMC, BUI Skewed : Rain, Area

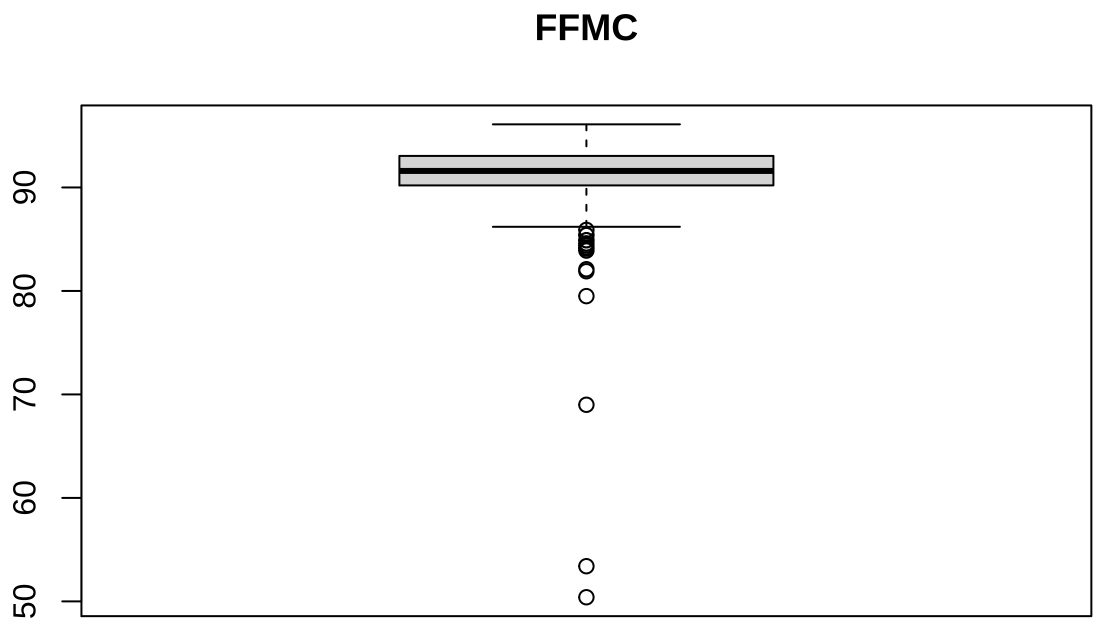


Observation: wind - Looks somewhat uniform Rain - Heavy right tail Temp - Looks Some what normal BUI - Looks Bimodal FFMC - has outliers which skews an otherwise normal looking distribution Area - Has a heavy right tail

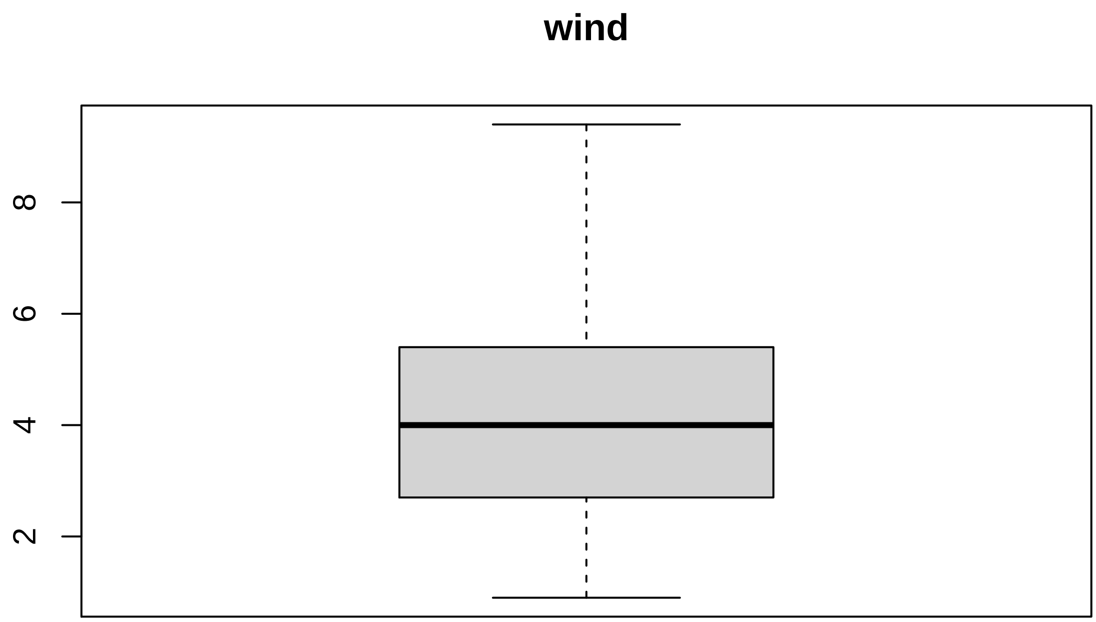
```r
# Inspect the distribution of each variable using a boxplot
# This gives us more information wrt outliers and variability
boxplot(EDA$temp, main='temp')
```



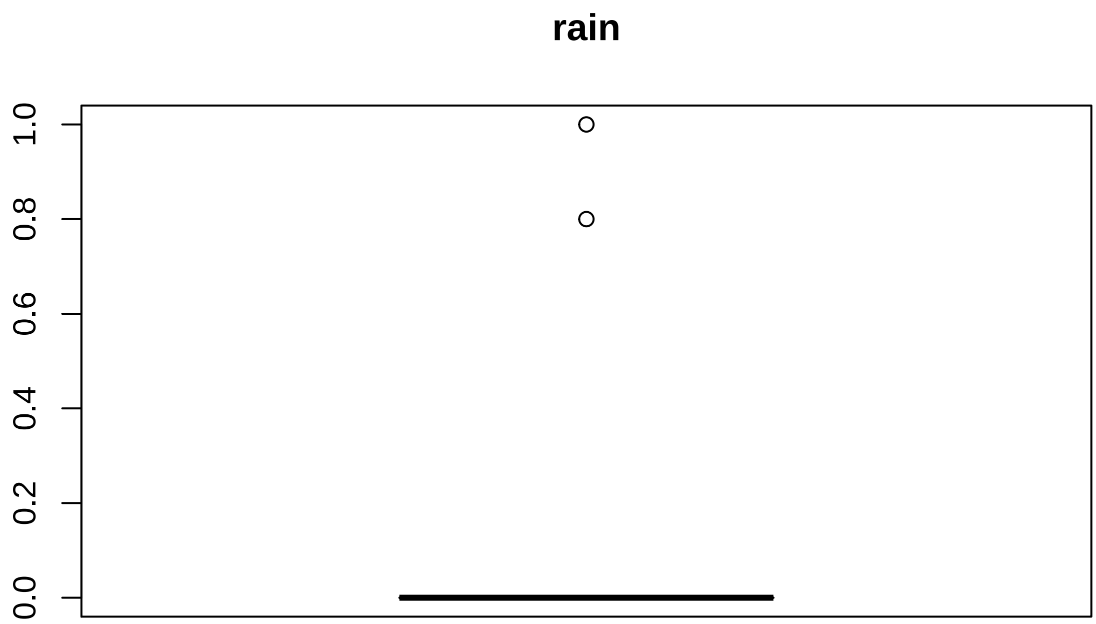

```r
boxplot(EDA$FFMC, main='FFMC')
```



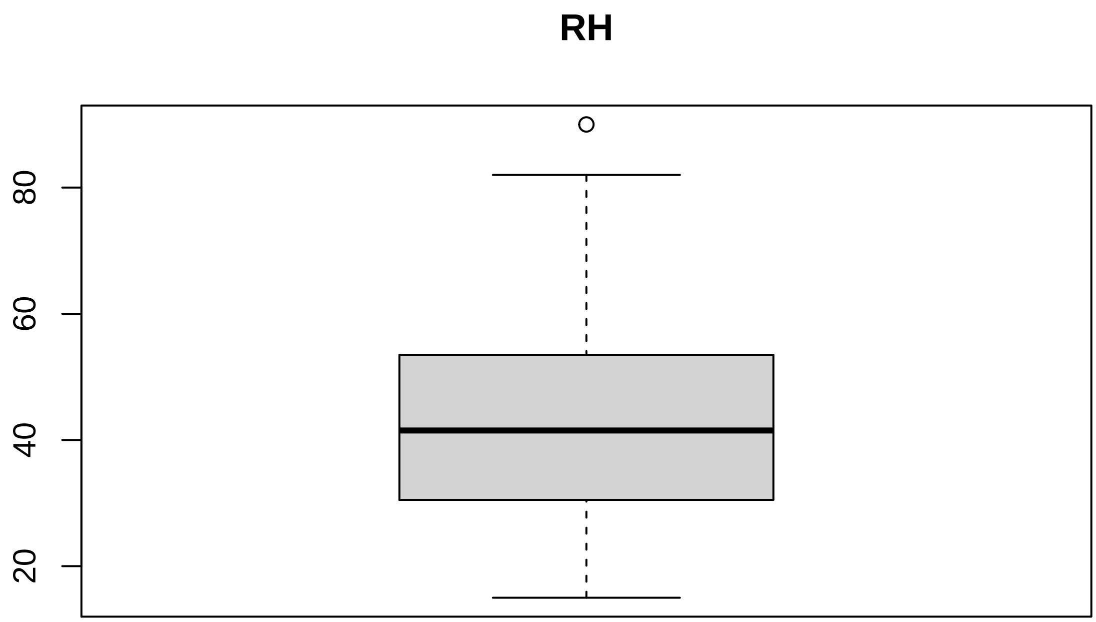

```
boxplot(EDA$wind, main='wind')
```



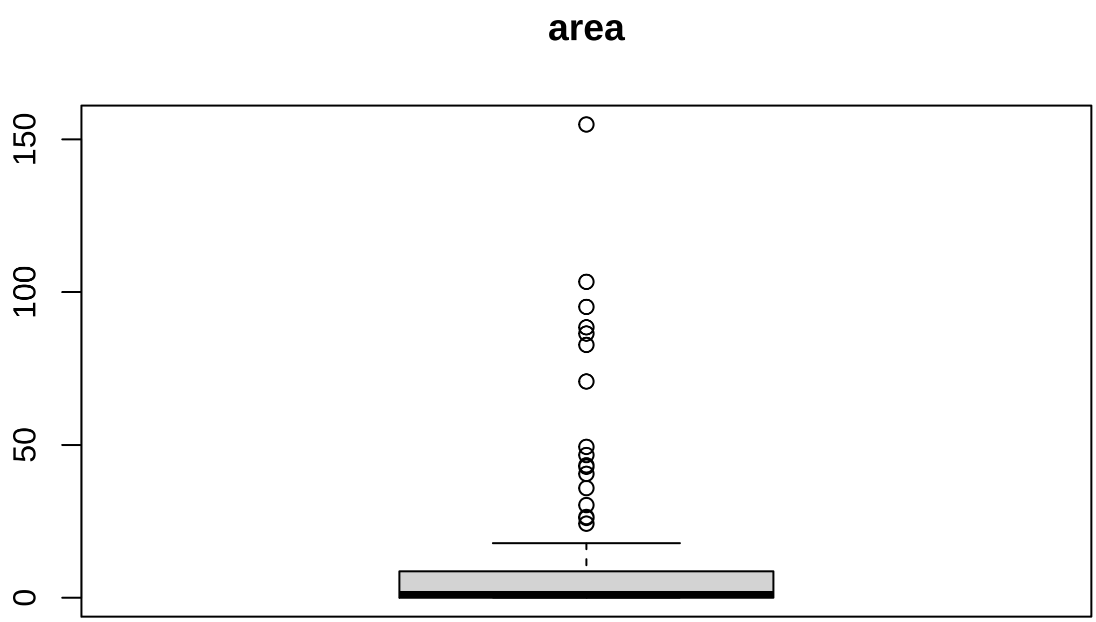

```
boxplot(EDA$rain, main='rain')
```




```
boxplot(EDA$RH,main="RH")
```




```
boxplot(EDA$area, main='area')
```



Observation Area - Has heavy outliers. The data has high variability Wind - The looks normal Rain - Has heavy outliers. The data has high variability FFMC - Has mnany outliers RH - has some outliers Temp - Has outliers

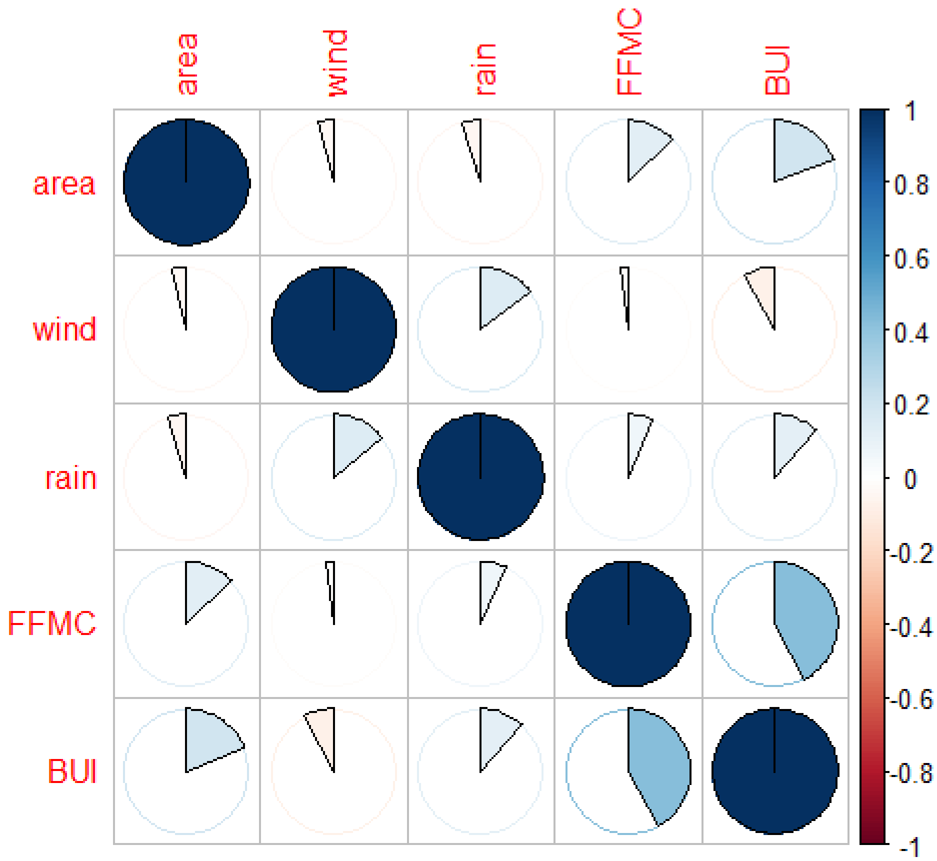
EDA Part II - Insight into a variable's relationship with other variables

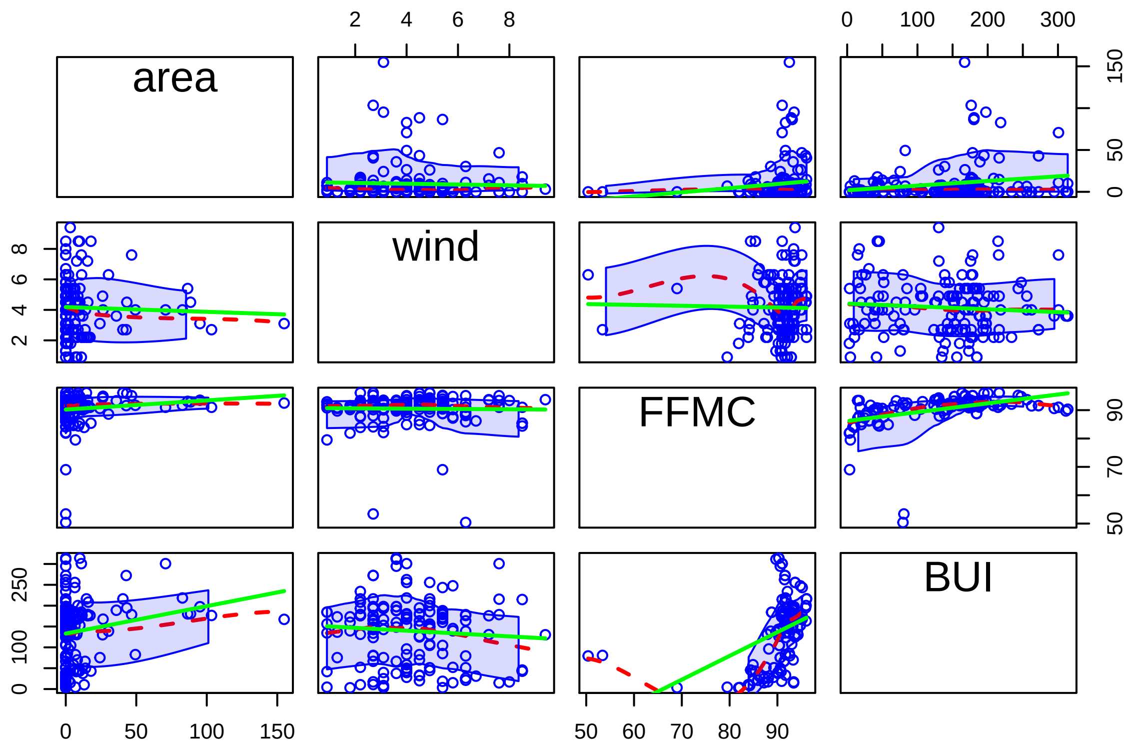
```
# Check the correlation  between all the basic variable of interest in the training data set based on o
# Create correlation matrix
env_var <- EDA[, c('area','wind', 'rain', 'FFMC',  'BUI')]
cor_EDA <- cor(env_var)
cor_EDA
```

```
##             area        wind        rain        FFMC         BUI
## area  1.00000000 -0.03856520 -0.04965709  0.12632617  0.19078954
## wind -0.03856520  1.00000000  0.14353432 -0.01783621 -0.07999498
## rain -0.04965709  0.14353432  1.00000000  0.06458313  0.11440311
## FFMC  0.12632617 -0.01783621  0.06458313  1.00000000  0.42124532
## BUI   0.19078954 -0.07999498  0.11440311  0.42124532  1.00000000
```

```
# Correlation plot to visualize the correlation between variables in training data set
#corrplot(cor_EDA,method='pie')
```

```r
# If you are knitting this notebook, comment out the plot_num command and use the following saved plot-
knitr::include_graphics("images/00000b.png")
```





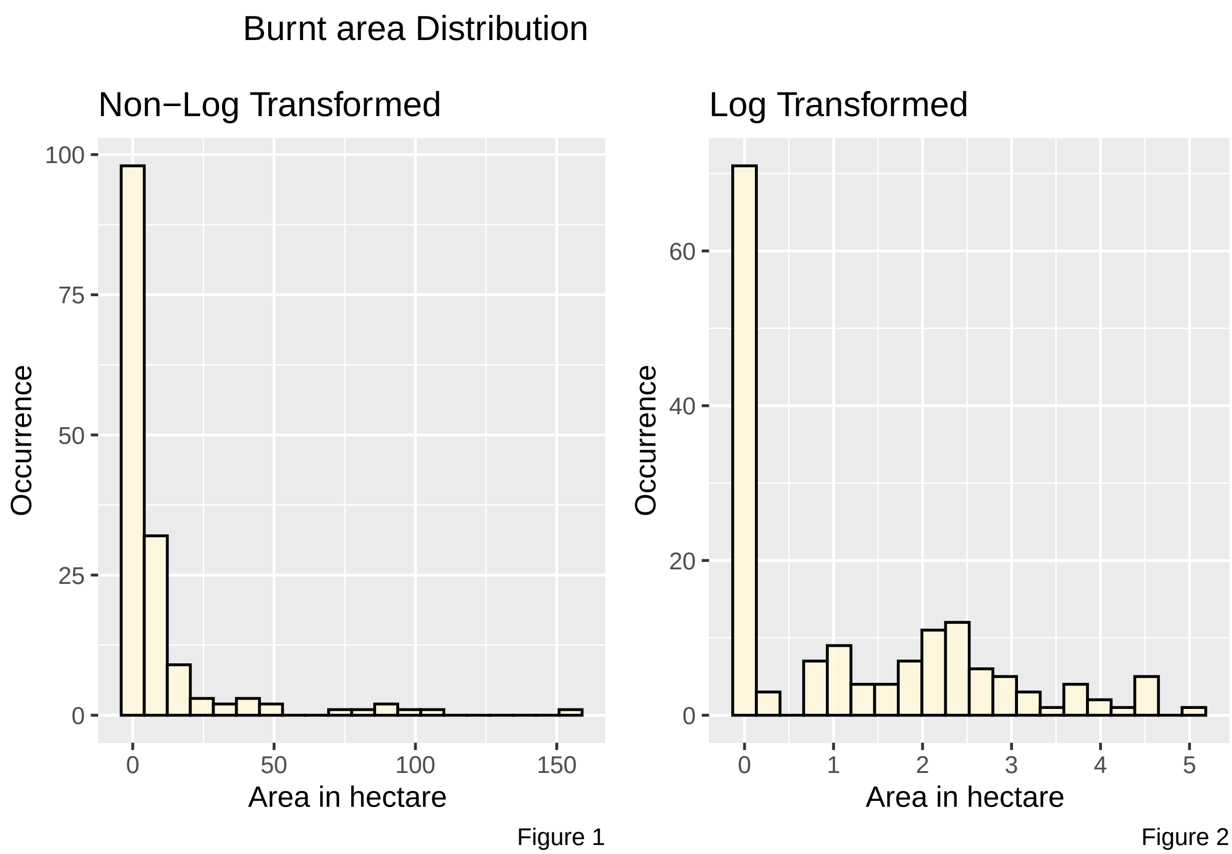
Focused Analysis and Transformations

Analysis of Outcome Variable : area

Area is very important for our study because it is our outcome variable. We want to find out what factors cause area to burn in forest fires. From the graph above, area seems to have a heavy left tail. We should use log transformation on area field to see if it helps improve a linear relationship with our predictor variables.

```
## Total observations in the dataset:  156
```

```
##
## Number of Observations where area is zero: 71
```



Observation:

The graphs above has a very heavy right tail and log transformation helped improve the distribution to look more normal. However, it is still skewed right.

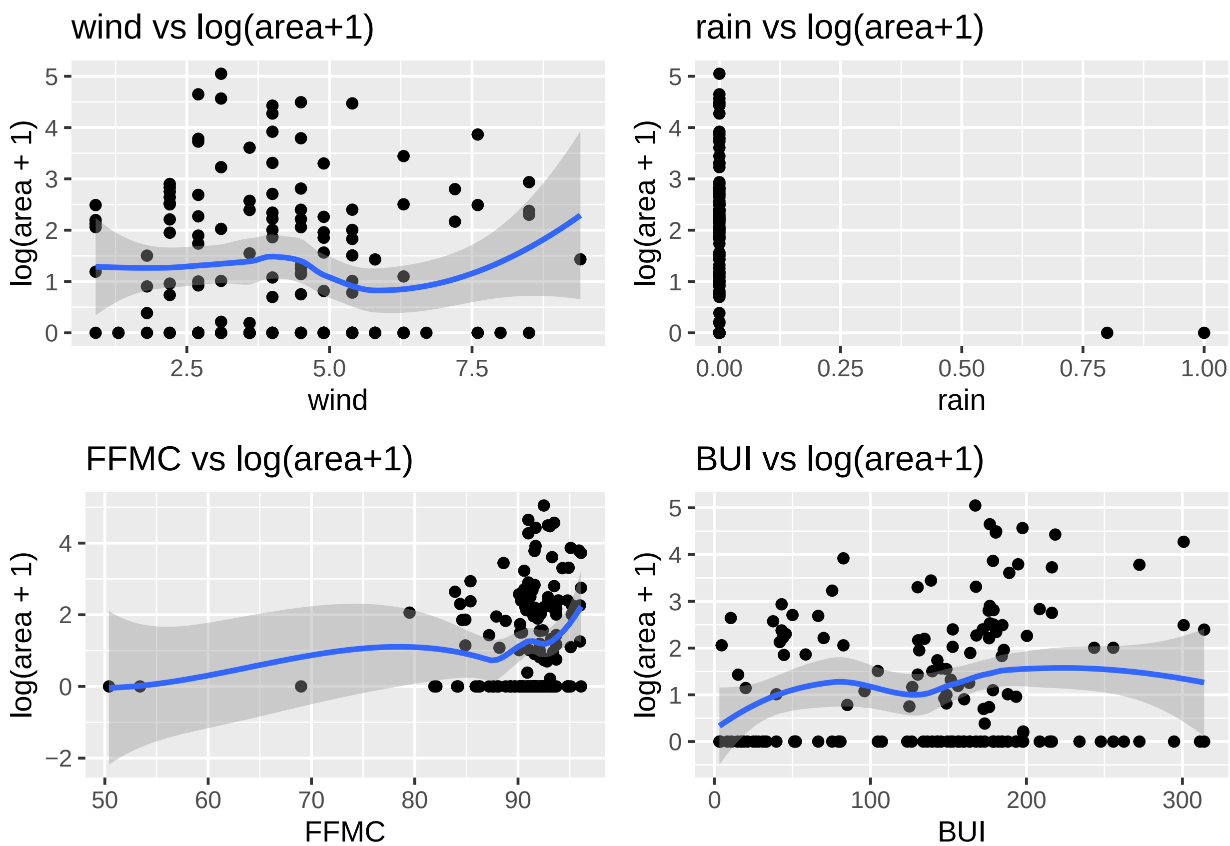
Analyzing of Linear Relationship of Predictors Variables with Outcome individually

Compare if transformation helps

In our causal diagram we have identified four predictor variables 1. rain - outside rain in mm/m2 : 0.0 to 6.4 2. wind - wind speed in km/h: 0.40 to 9.40 3. FFMC - FFMC index from the FWI system: 18.7 to 96.20 4. BUI - potential heat release in heavier fuels (total amount of fuel available for combustion): 0.0 to infinity

```
a <- ggplot(EDA, aes(x = wind, y = log(area+1))) +
  geom_point() +
  geom_smooth() +
  ggtitle("wind vs log(area+1)")

b <- ggplot(EDA, aes(x = rain, y = log(area+1))) +
  geom_point() +
```

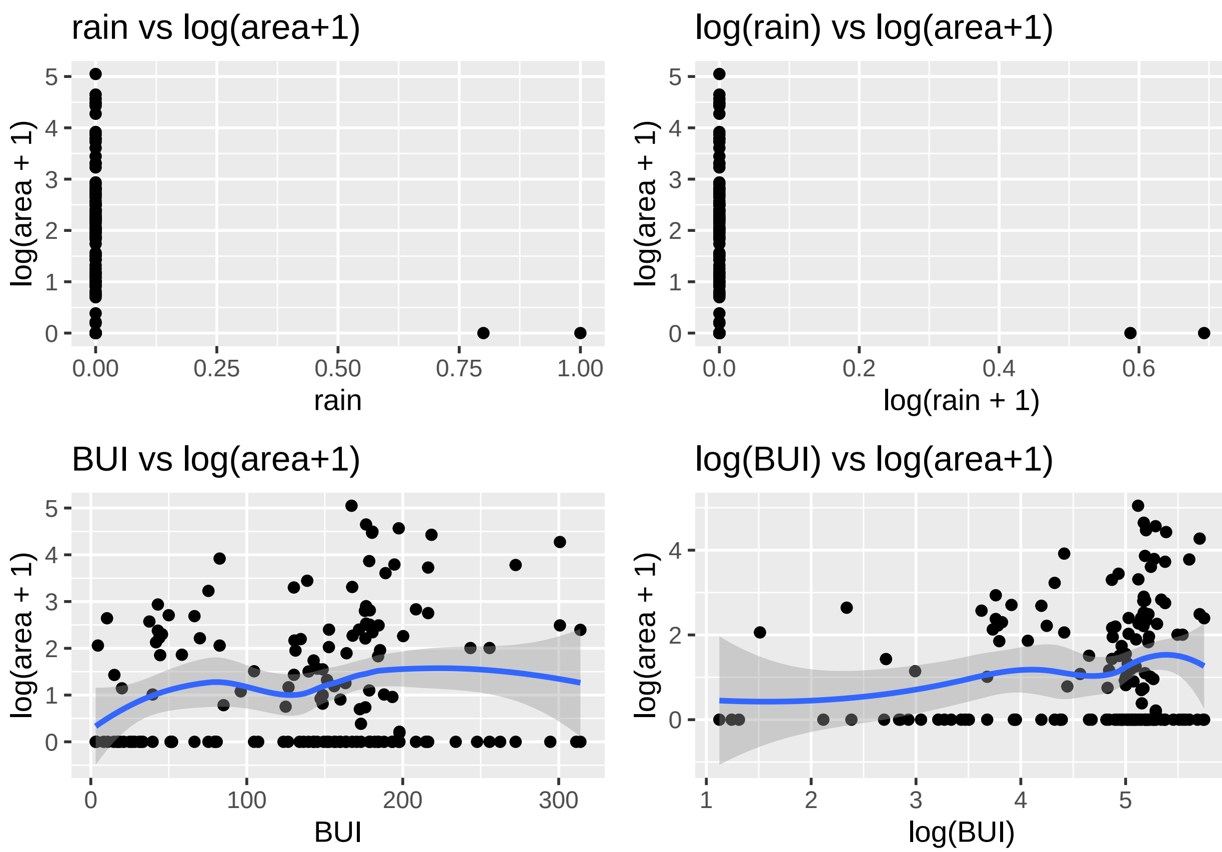
```
  geom_smooth() +
  ggtitle("rain vs log(area+1)")

c <- ggplot(EDA, aes(x = FFMC, y = log(area+1))) +
  geom_point() +
  geom_smooth() +
  ggtitle("FFMC vs log(area+1)")

d <- ggplot(EDA, aes(x = BUI, y = log(area+1))) +
  geom_point() +
  geom_smooth() +
  ggtitle("BUI vs log(area+1)")

grid.arrange(a, b, c, d, nrow = 2)
```



Observation: Rain and BUI variables have room for improvement. Will try log transforms to see if it helps improve linearity.

```
a <- ggplot(EDA, aes(x = rain, y = log(area+1))) +
  geom_point() +
  geom_smooth() +
  ggtitle("rain vs log(area+1)")

b <- ggplot(EDA, aes(x = log(rain+1), y = log(area+1))) +
  geom_point() +
  geom_smooth() +
```

```r
  ggtitle("log(rain) vs log(area+1)")

c <- ggplot(EDA, aes(x = BUI, y = log(area+1))) +
  geom_point() +
  geom_smooth() +
  ggtitle("BUI vs log(area+1)")

d <- ggplot(EDA, aes(x = log(BUI), y = log(area+1))) +
  geom_point() +
  geom_smooth() +
  ggtitle("log(BUI) vs log(area+1)")

grid.arrange(a, b, c, d, nrow = 2)
```



Conclusion rain predictor did not benefit from a log transform. However, log(BUI) seems to have slightly improved linear relationship with log(area+1).

Analysis of rain

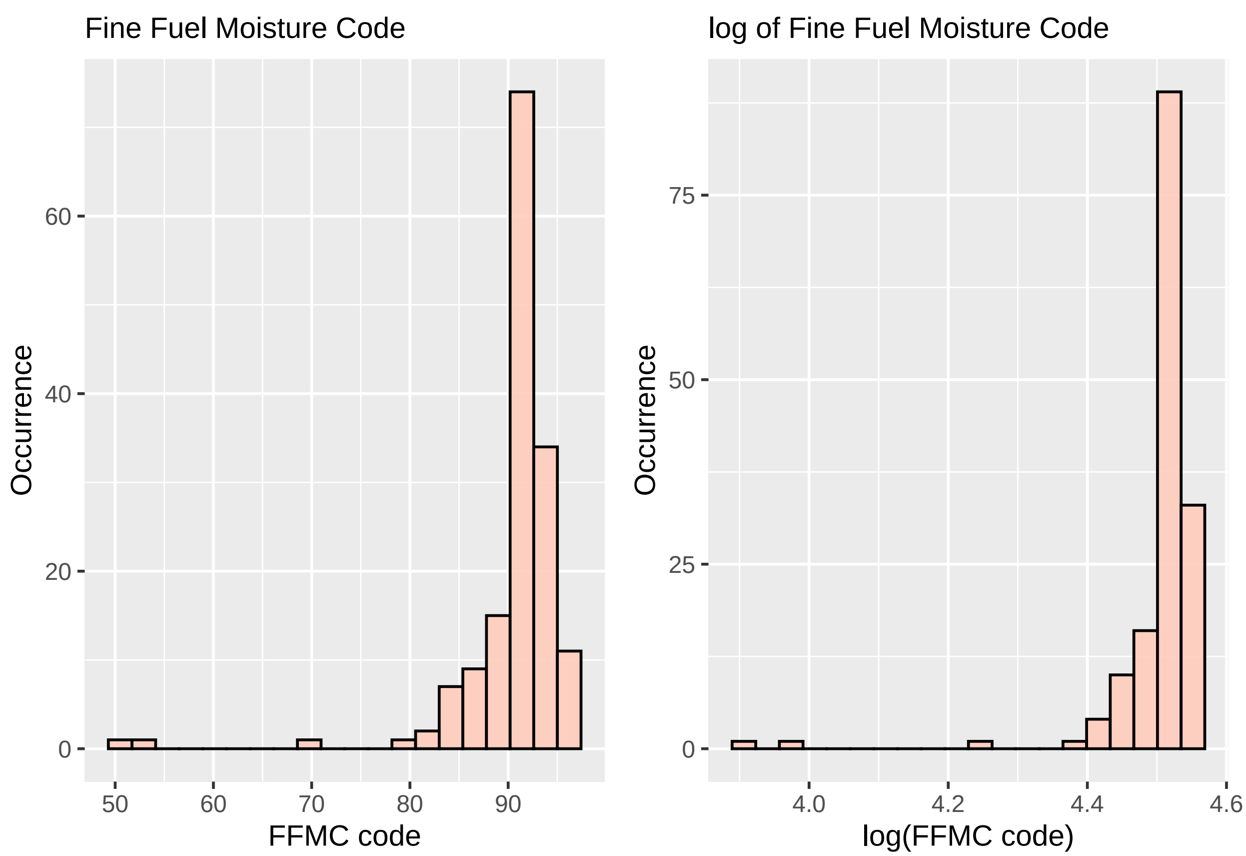
```r
# all transformations failed to improve the plot of rain and log(area+1)/
# so, we will convert our rain variable to binary, is not raining (0) or is raining (1).

Prod <- Prod %>%
```

```
  mutate(
    rain_binary = case_when(
      rain > 0 ~ 1,
      rain == 0 ~ 0,
      )
  )

EDA <- EDA %>%
  mutate(
    rain_binary = case_when(
      rain > 0 ~ 1,
      rain == 0 ~ 0,
      )
  )
```

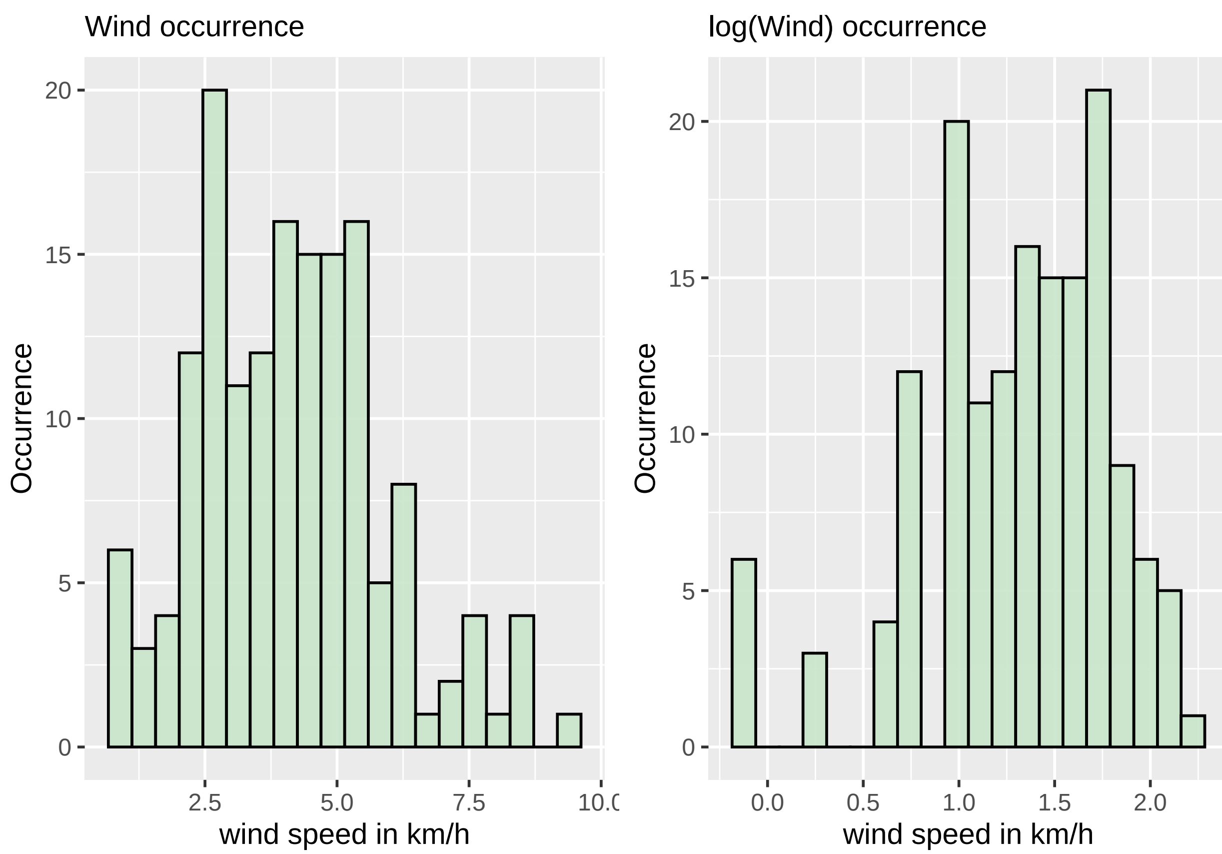
Analysis: wind, BUI and FFMC

```
hist_of_wind_dist <- EDA %>%
ggplot() + aes(x = wind ) +
geom_histogram( bins=20, fill="#C8E6C9", color="black", alpha=0.9) +
labs(
x = "wind speed in km/h", y="Occurrence",
subtitle = 'Wind occurrence'
)

hist_of_logwind_dist <- EDA %>%
ggplot() + aes(x = log(wind ) )+
geom_histogram( bins=20, fill="#C8E6C9", color="black", alpha=0.9) +
labs(
x = "wind speed in km/h", y="Occurrence",
subtitle = 'log(Wind) occurrence'
)
hist_of_BUI_dist <- EDA %>%
ggplot() + aes(x = BUI ) +
geom_histogram( bins=20, fill="#D1C4E9", color="black", alpha=0.9) +
labs(
x = "BUI", y="Occurrence",
subtitle = 'Build Up Index (BUI)'
)

hist_of_logBUI_dist <- EDA %>%
ggplot() + aes(x = log(BUI) ) +
geom_histogram( bins=20, fill="#D1C4E9", color="black", alpha=0.9) +
labs(
x = "log(BUI)", y="Occurrence",
subtitle = 'log of Build Up Index'
)

hist_of_FFMC_dist <- EDA %>%
ggplot() + aes(x = FFMC ) +
geom_histogram( bins=20, fill="#FFCCBC", color="black", alpha=0.9) +
```

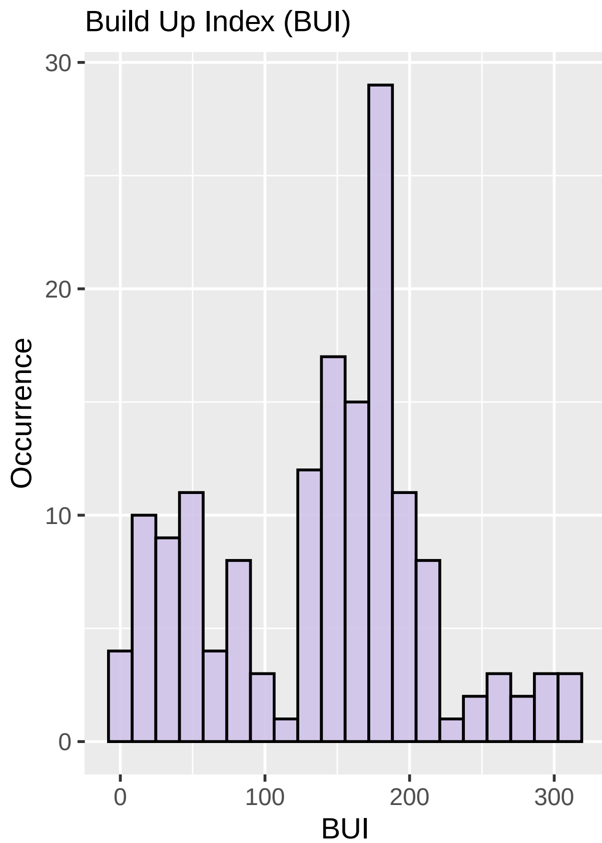
```
labs(
x = "FFMC code", y="Occurrence",
subtitle = 'Fine Fuel Moisture Code'
)
hist_of_logFFMC_dist <- EDA %>%
ggplot() + aes(x = log(FFMC) ) +
geom_histogram( bins=20, fill="#FFCCBC", color="black", alpha=0.9) +
labs(
x = "log(FFMC code)", y="Occurrence",
subtitle = 'log of Fine Fuel Moisture Code'
)
par(mfrow=c(1, 2))

plot_grid(hist_of_FFMC_dist, hist_of_logFFMC_dist)
```

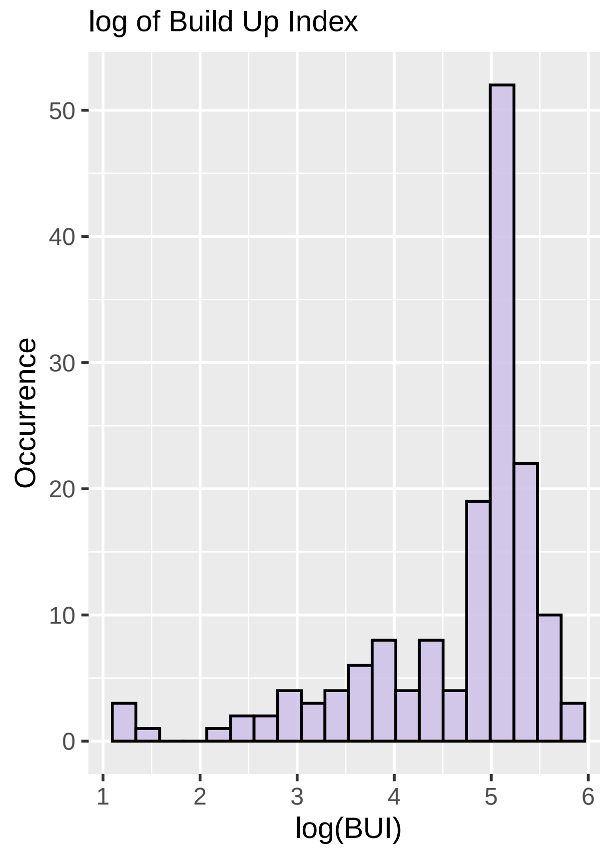



```
plot_grid(hist_of_wind_dist, hist_of_logwind_dist)
```




```
plot_grid(hist_of_BUI_dist, hist_of_logBUI_dist)
```






```
#plot_grid(hist_of_FFMC_dist, hist_of_logFFMC_dist, hist_of_wind_dist, hist_of_BUI_dist)
```

Observation:

FFMC - Log of FFMC did not make any difference in FFMC's distribution Wind - Log transformation makes the distribution worse BUI - Log transformation makes the distribution slightly normal

Building Our Linear Models

```
# a model that uses only the direct atmospheric metrics
model1 <- lm(log(area+1) ~ log(BUI) , data = Prod)
coeftest(model1, vcov=vcovHAC)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.946883   0.394191  2.4021  0.01681 *
## log(BUI)    0.024429   0.082624  0.2957  0.76766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# a model that uses all available variables in the causal diagram
model2 <- lm(log(area+1) ~ log(BUI) + wind + rain_binary + FFMC, data = Prod)
coeftest(model2, vcov=vcovHAC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.9629663  1.0575243  0.9106   0.3631
## log(BUI)     0.0954846  0.1002442  0.9525   0.3415
## wind         0.0961630  0.0402457  2.3894   0.0174 *
## rain_binary -0.6367068  0.4003307 -1.5905   0.1126
## FFMC        -0.0079667  0.0139260 -0.5721   0.5676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# our primary model used to answer our Research Question
model3 <- lm(log(area+1) ~ wind , data = Prod)
coeftest(model3, vcov=vcovHAC)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.735300   0.174045  4.2248 3.036e-05 ***
## wind        0.082652   0.039619  2.0862   0.03767 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Compare all the models in a tabular form
stargazer(
  model1,
  model2,
  model3,
  type = 'text', header = FALSE,
  star.cutoffs = c(0.05, 0.01, 0.001)
)
```

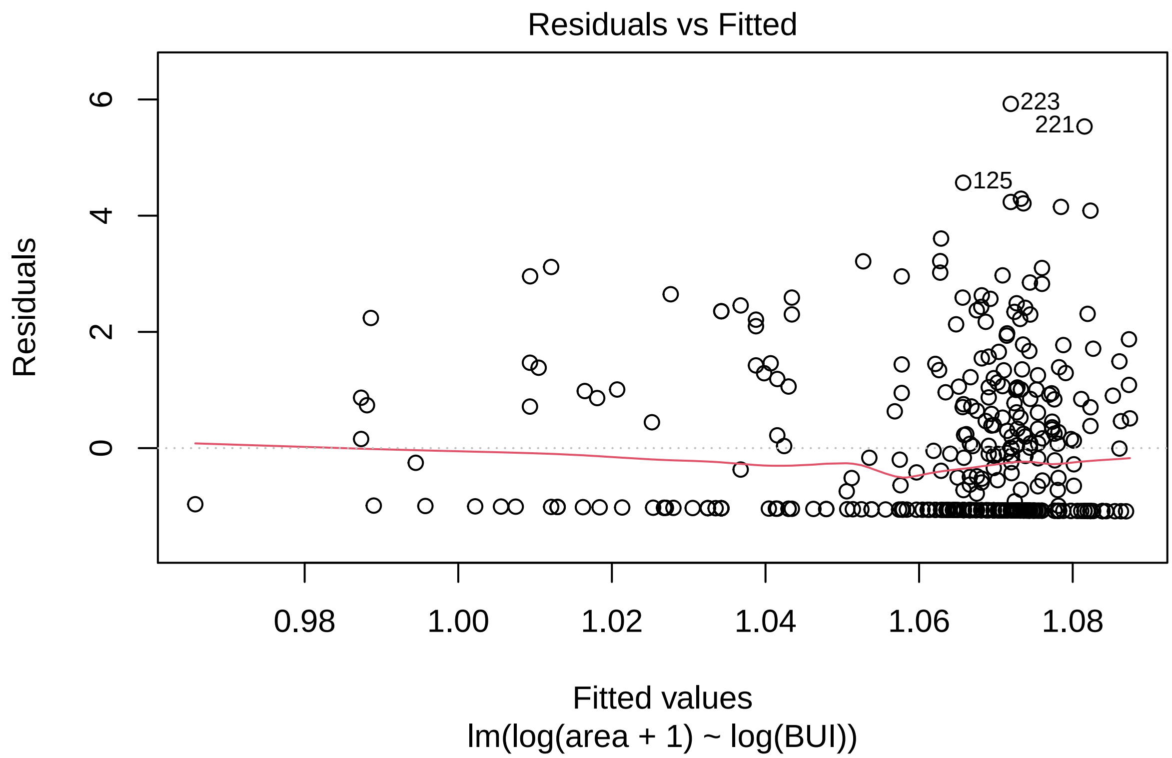
```
##
## ================================================================================
##                                          Dependent variable:
##                      -----------------------------------------------------------
##                                             log(area + 1)
##                              (1)                  (2)                  (3)
## --------------------------------------------------------------------------------
## log(BUI)                    0.024                0.095
##                            (0.087)              (0.105)
##
## wind                                             0.096*               0.083*
##                                                 (0.042)              (0.041)
##
## rain_binary                                     -0.637
```

```
##                                          (0.581)
##
## FFMC                                     -0.008
##                                          (0.016)
##
## Constant                  0.947*          0.963             0.735***
##                           (0.416)         (1.267)           (0.178)
##
## ----------------------------------------------------------------------------------
## Observations                361             361                361
## R2                        0.0002          0.017              0.011
## Adjusted R2               -0.003          0.006              0.008
## Residual Std. Error  1.402 (df = 359)    1.396 (df = 356)     1.394 (df = 359)
## F Statistic         0.080 (df = 1; 359) 1.509 (df = 4; 356) 4.083* (df = 1; 359)
## ==================================================================================
## Note:                                             *p<0.05; **p<0.01; ***p<0.001
```

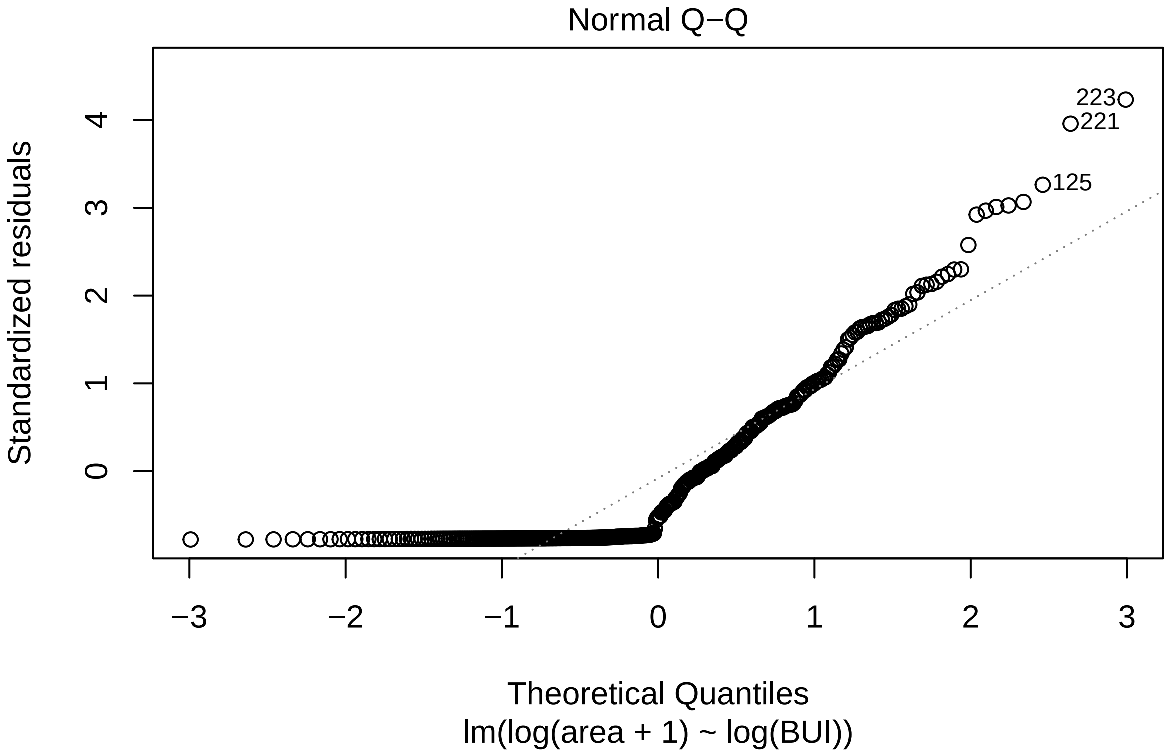
```r
anova(model1, model2, model3, test="F")
```

```
## Analysis of Variance Table
##
## Model 1: log(area + 1) ~ log(BUI)
## Model 2: log(area + 1) ~ log(BUI) + wind + rain_binary + FFMC
## Model 3: log(area + 1) ~ wind
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    359 705.79
## 2    356 694.18  3   11.6138 1.9853 0.1158
## 3    359 698.01 -3   -3.8326 0.6552 0.5802
```

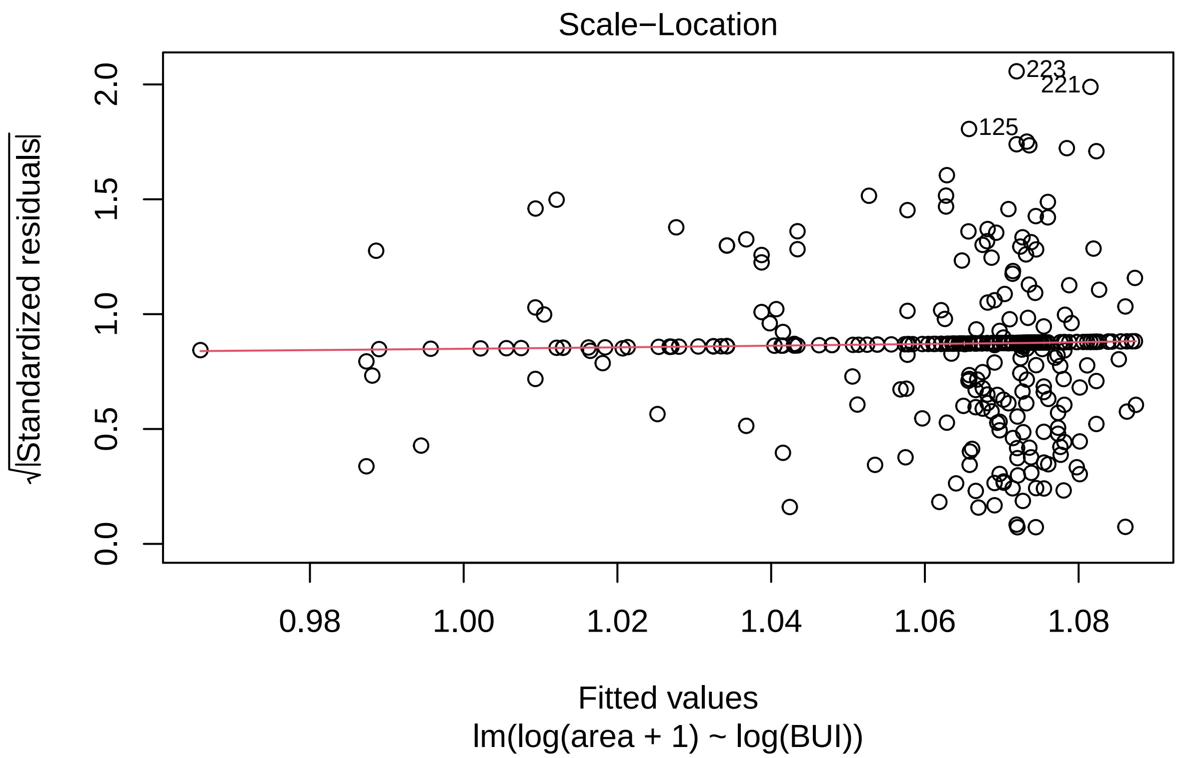
```r
plot(model1)
```

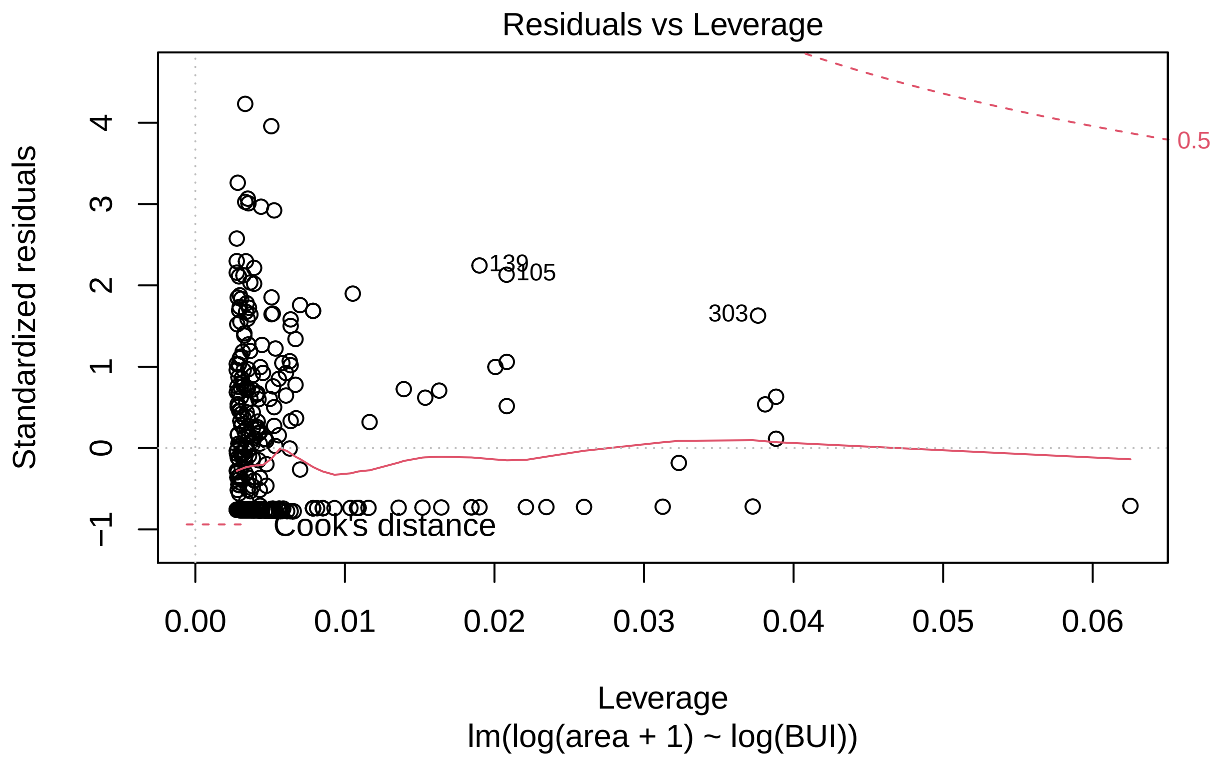


Residuals vs Fitted
Residuals
Fitted values
lm(log(area + 1) ~ log(BUI))
223
221
125
0
2
4
6
0.98
1.00
1.02
1.04
1.06
1.08



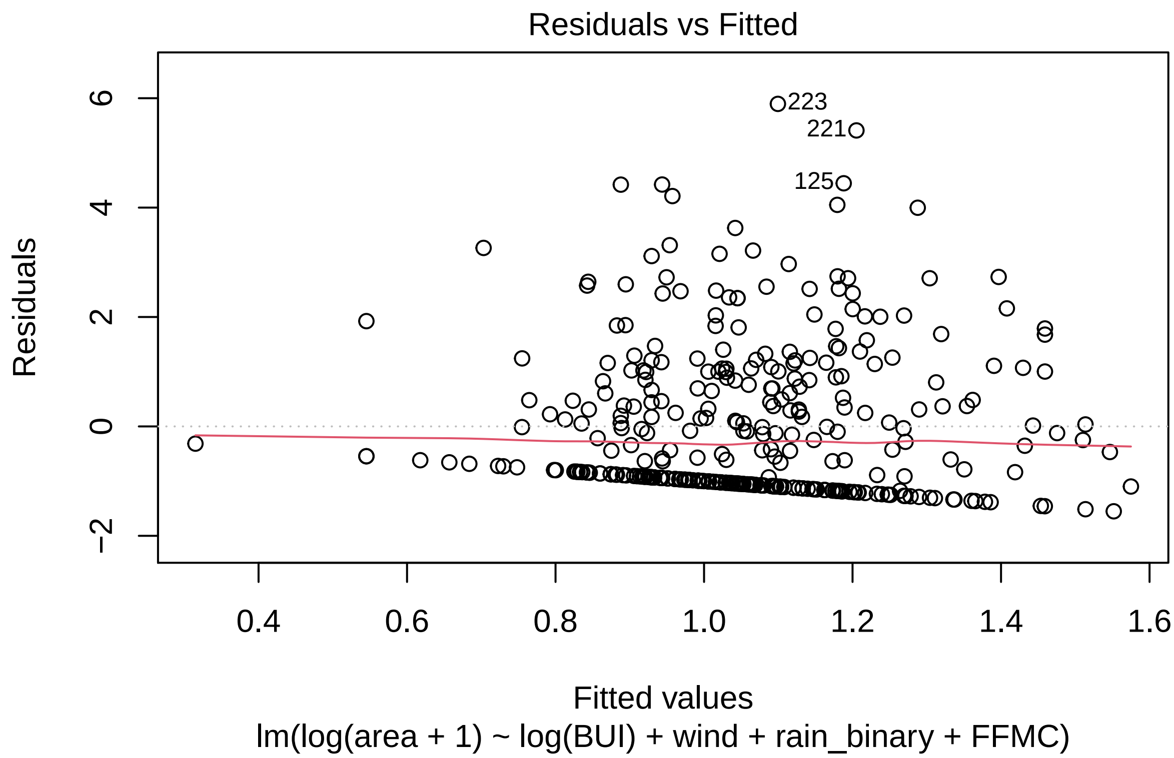
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(area + 1) ~ log(BUI))
223
221
125
4
3
2
1
0
−3
−2
−1
0
1
2
3

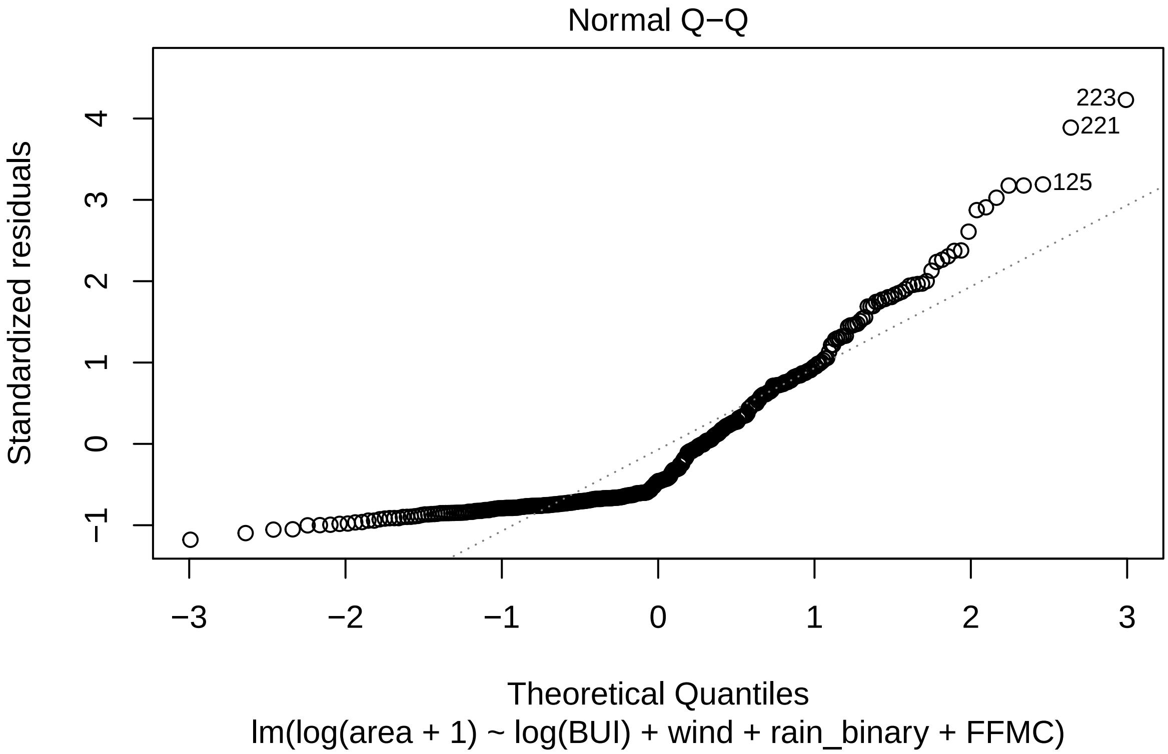

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.98
1.00
1.02
1.04
1.06
1.08
223
221
125
Fitted values
lm(log(area + 1) ~ log(BUI))

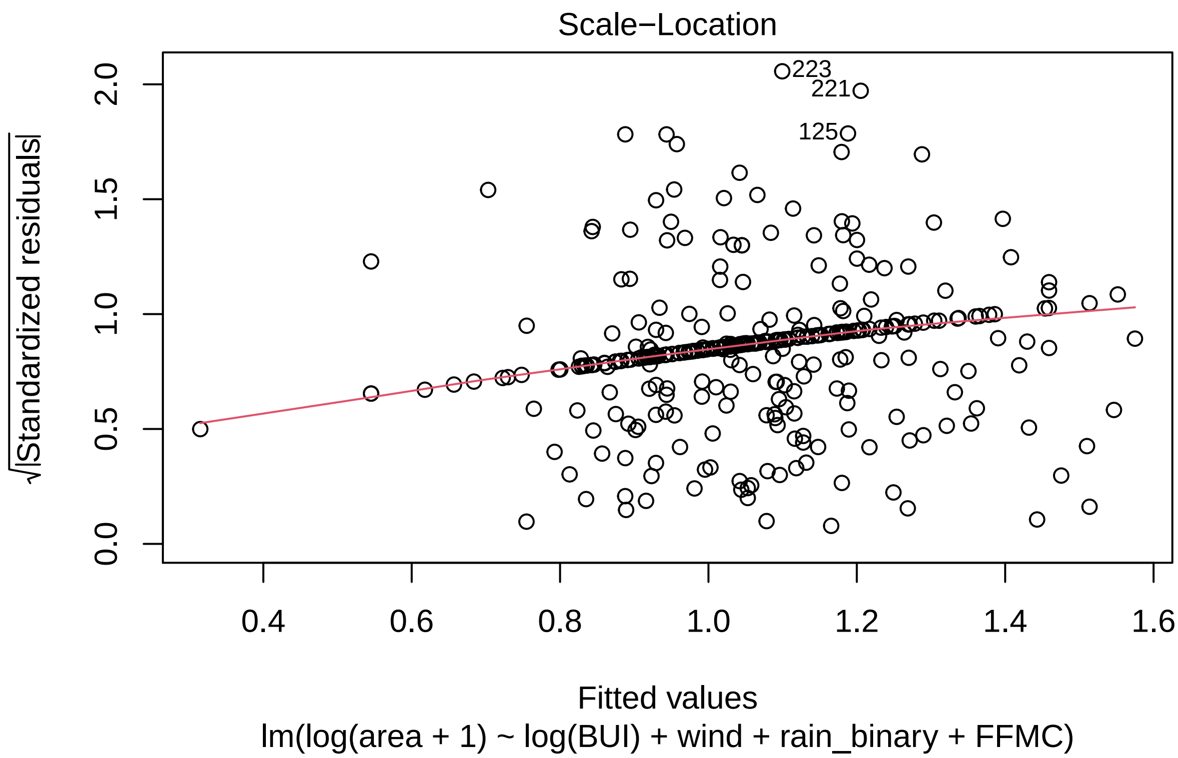


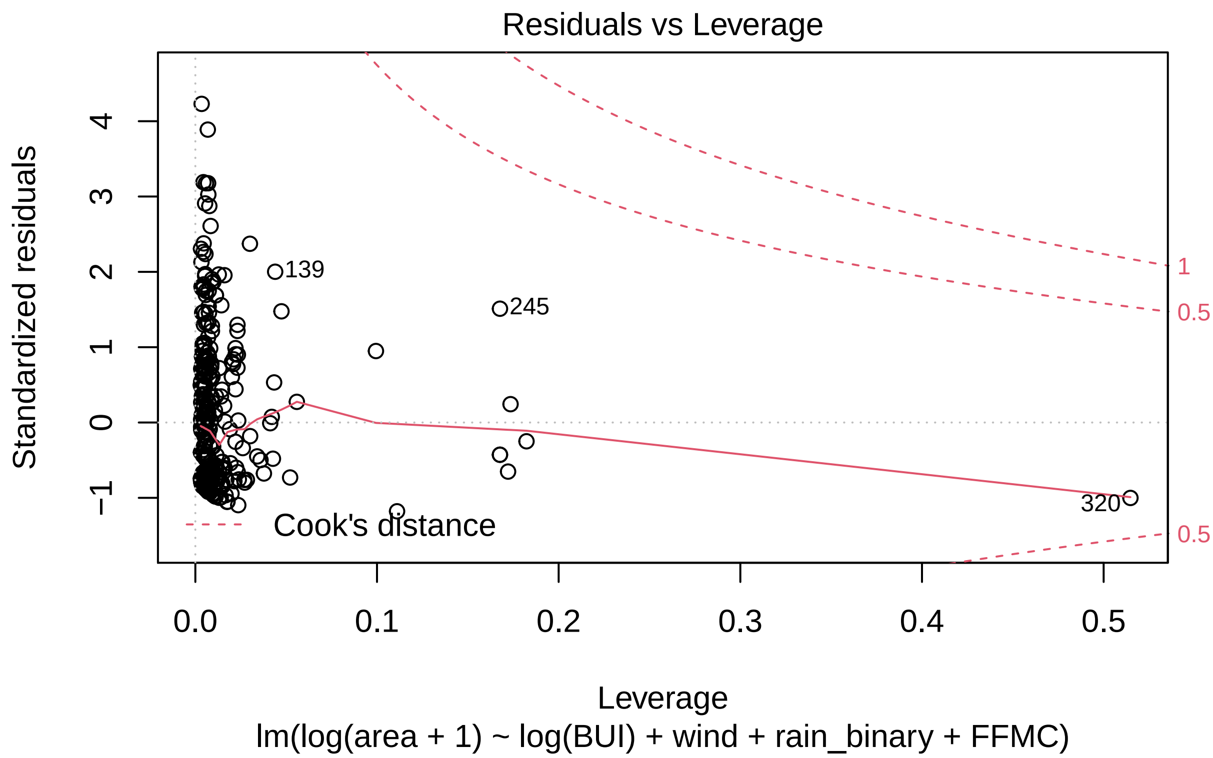
Residuals vs Leverage
Standardized residuals
Leverage
lm(log(area + 1) ~ log(BUI))
Cook's distance
0.5
139
105
303
4
3
2
1
0
−1
0.00
0.01
0.02
0.03
0.04
0.05
0.06


```
plot(model2)
```

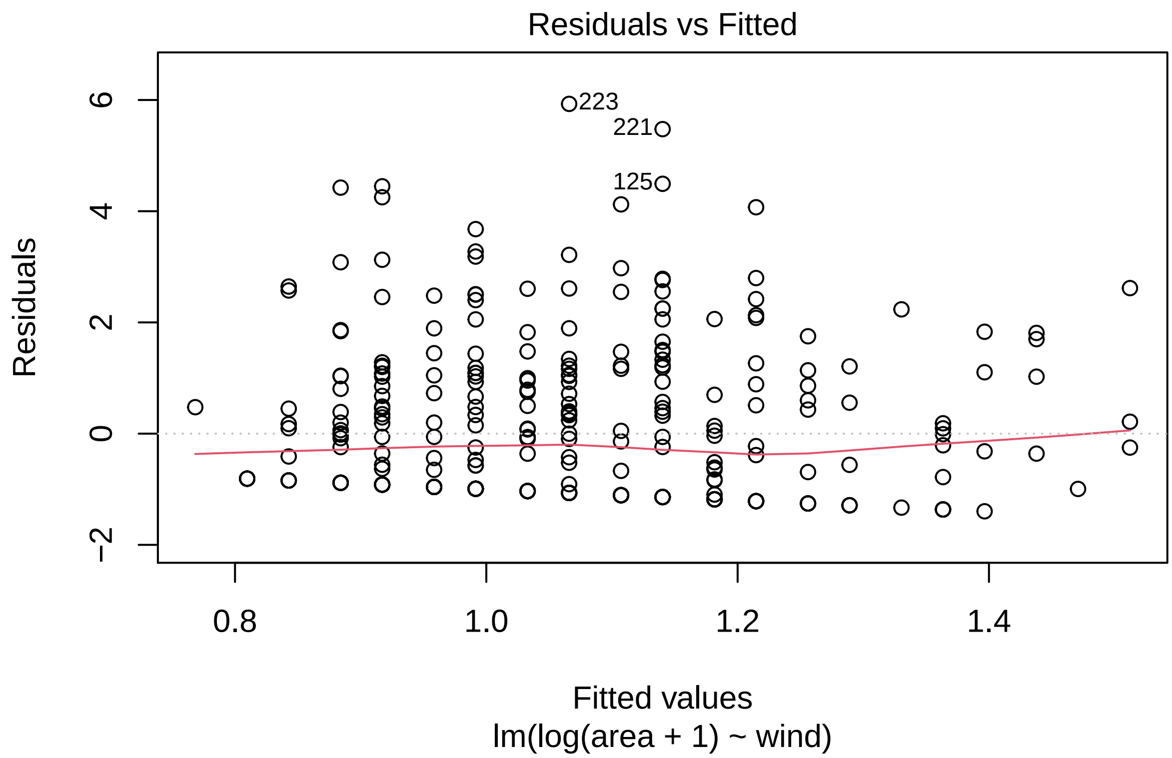

Residuals vs Fitted
Residuals
Fitted values
lm(log(area + 1) ~ log(BUI) + wind + rain_binary + FFMC)
223
221
125
6
4
2
0
−2
0.4
0.6
0.8
1.0
1.2
1.4
1.6

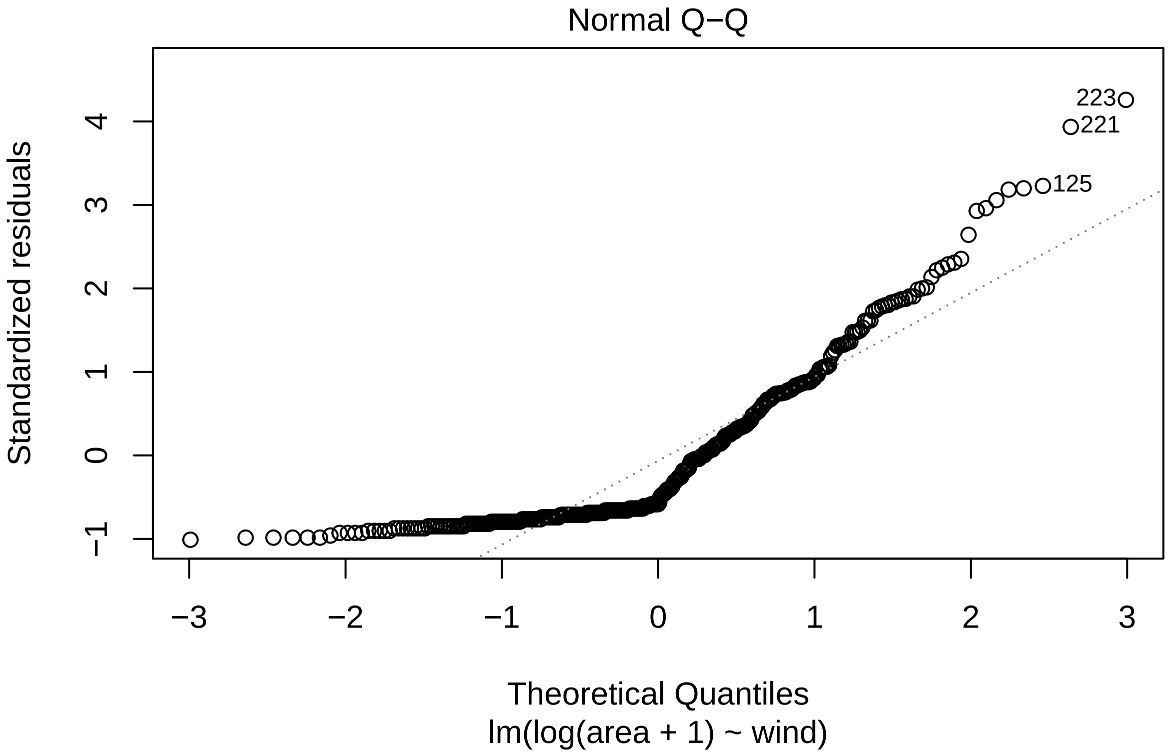

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(area + 1) ~ log(BUI) + wind + rain_binary + FFMC)
223
221
125
−3
−2
−1
0
1
2
3
4

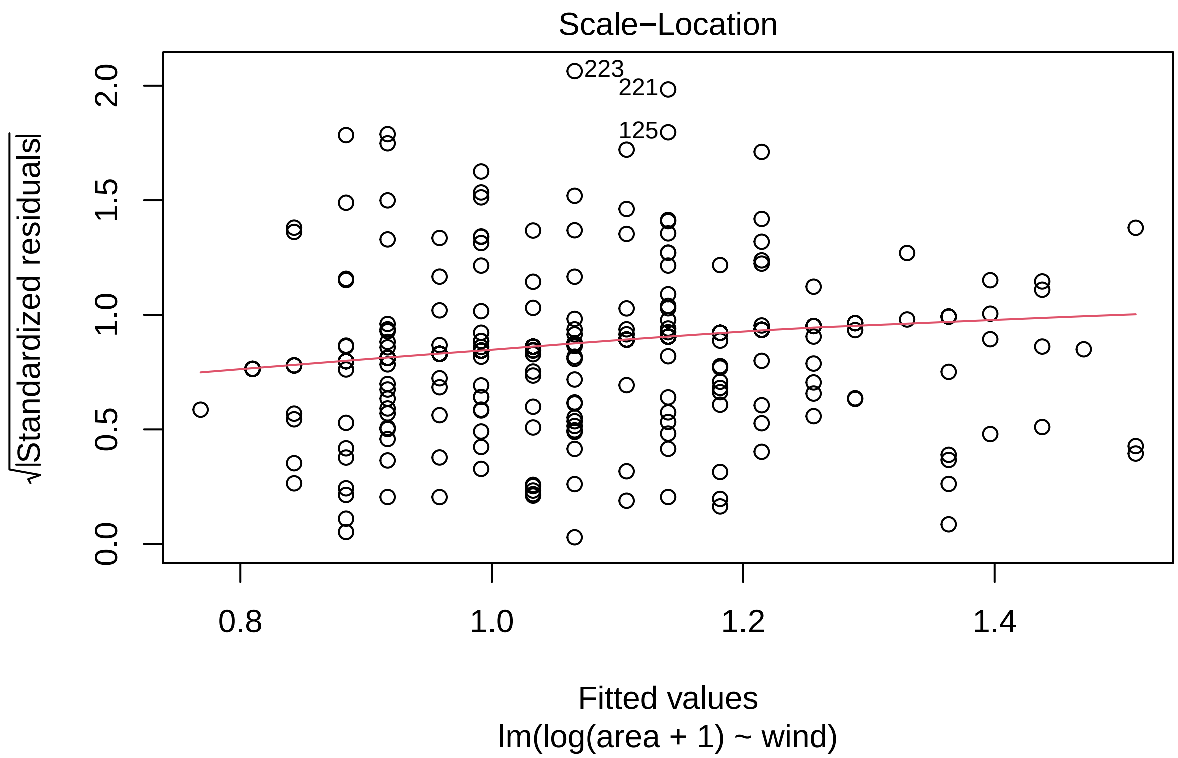

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
223
221
125
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Fitted values
lm(log(area + 1) ~ log(BUI) + wind + rain_binary + FFMC)

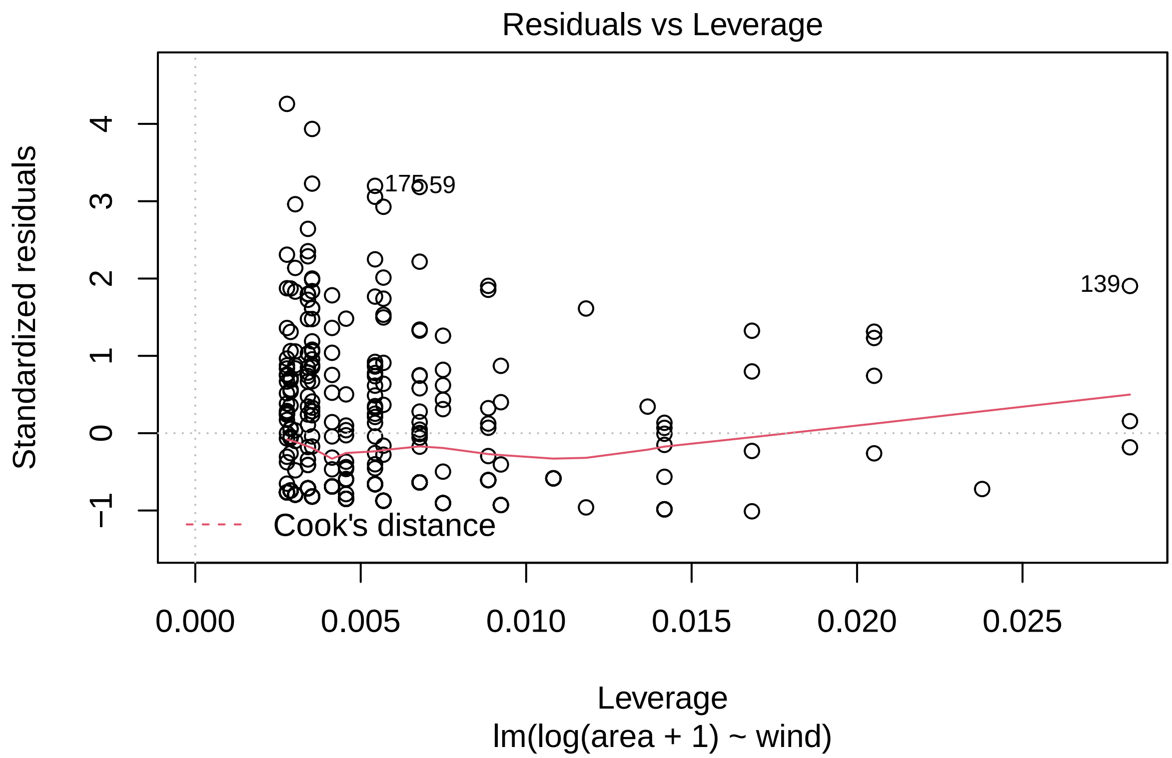

Residuals vs Leverage
Standardized residuals
Leverage
lm(log(area + 1) ~ log(BUI) + wind + rain_binary + FFMC)
Cook's distance
139
245
320
1
0.5
0.5
4
3
2
1
0
−1
0.0
0.1
0.2
0.3
0.4
0.5

```
plot(model3)
```

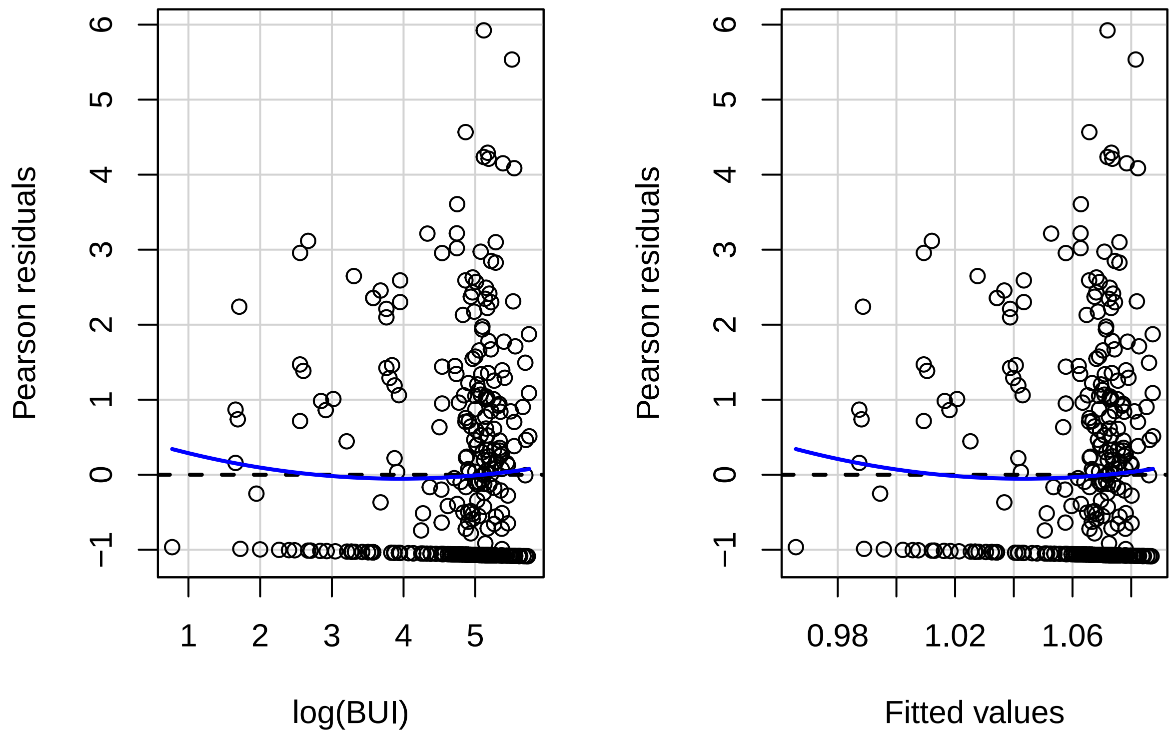

Residuals vs Fitted
Residuals
Fitted values
lm(log(area + 1) ~ wind)
223
221
125
−2
0
2
4
6
0.8
1.0
1.2
1.4


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
223
221
125
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log(area + 1) ~ wind)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.8
1.0
1.2
1.4
223
221
125
Fitted values
lm(log(area + 1) ~ wind)

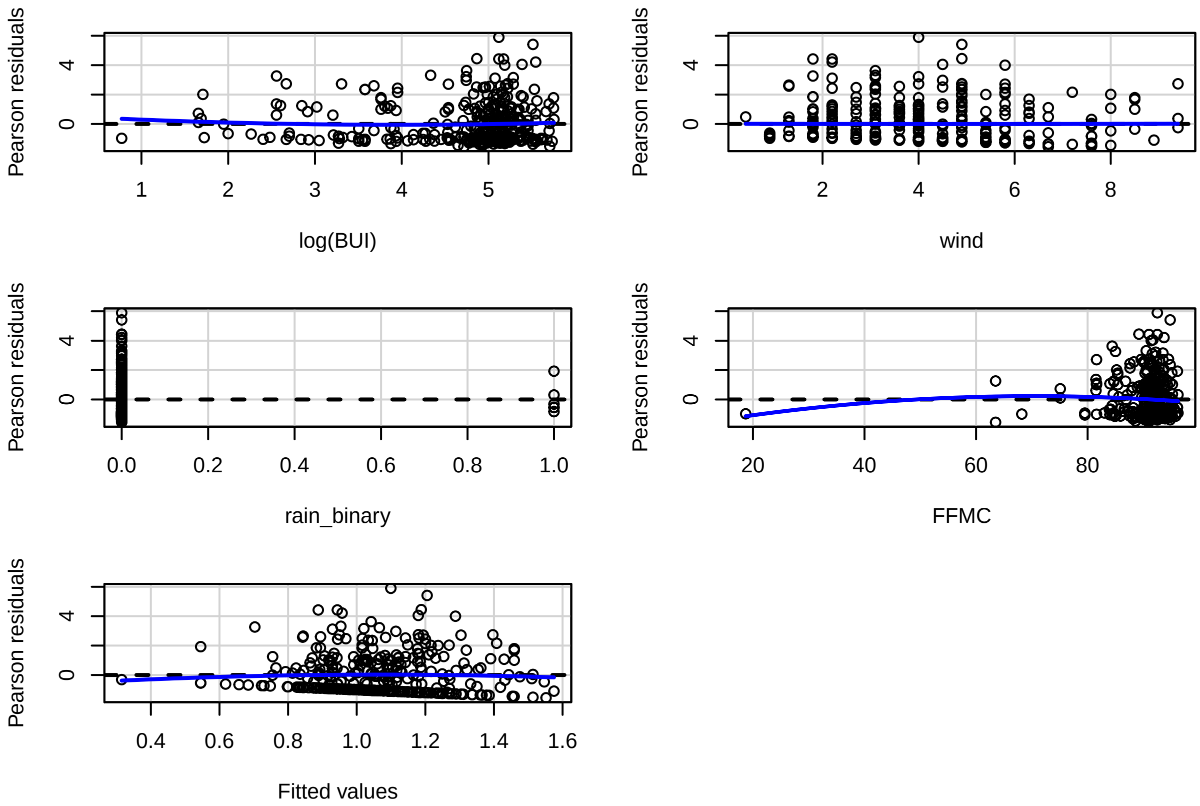



```
par(mfrow=c(1, 3))
residualPlots(model1)
```



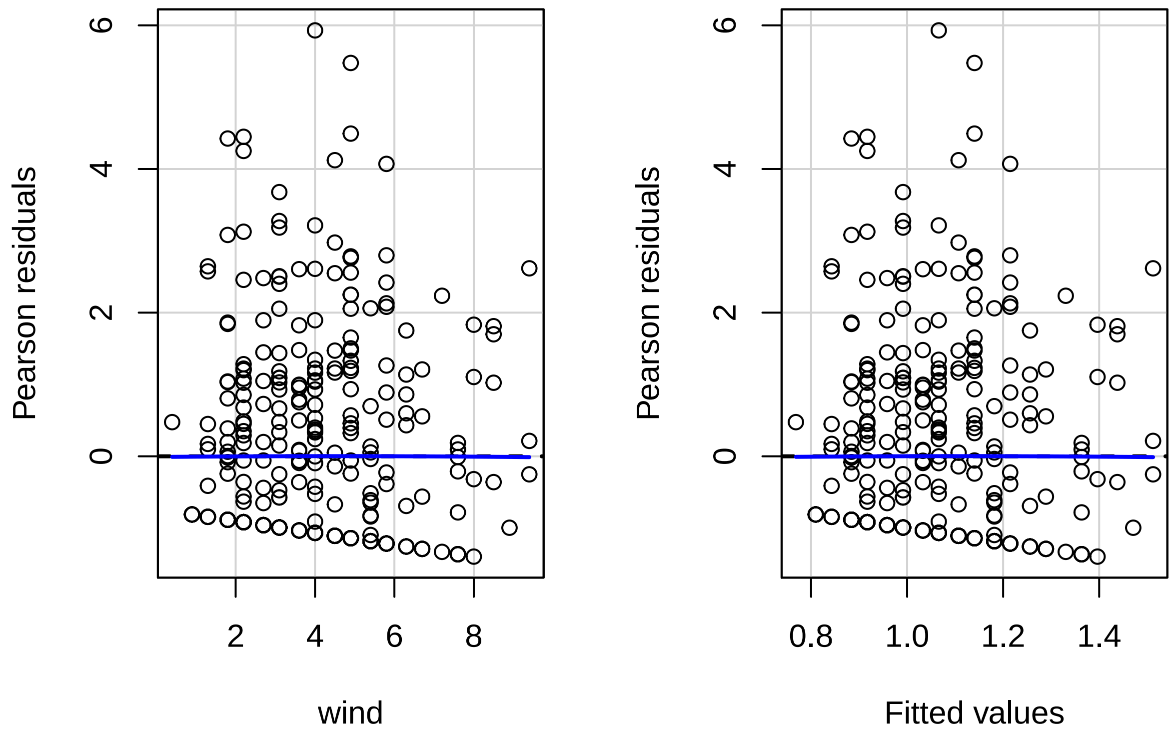

```
##              Test stat Pr(>|Test stat|)
## log(BUI)        0.5324           0.5948
## Tukey test      0.5324           0.5944
```

```
residualPlots(model2)
```




```
##              Test stat Pr(>|Test stat|)
## log(BUI)        0.5900           0.5556
## wind            0.0439           0.9650
## rain_binary    -0.4575           0.6476
## FFMC           -1.2377           0.2167
## Tukey test     -0.6567           0.5113
```

```
residualPlots(model3)
```




```
##              Test stat Pr(>|Test stat|)
## wind           -0.0329           0.9738
## Tukey test     -0.0329           0.9738
```

Observations

For detailed analysis of the models please refer to the project document