Lab 2

w203: Statistics for Data Science Project II

w203 Savita chari

```
setwd("C:/Users/savit/W203_lab_2")
fire_data <- read.csv("forestfires.csv")
glimpse(fire_data)
```

```
## Rows: 517
## Columns: 13
## $ X     <int> 7, 7, 7, 8, 8, 8, 8, 8, 8, 7, 7, 7, 6, 6, 6, 6, 5, 8, 6, 6, 6, 5~
## $ Y     <int> 5, 4, 4, 6, 6, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4~
## $ month <chr> "mar", "oct", "oct", "mar", "mar", "aug", "aug", "aug", "sep", "~
## $ day   <chr> "fri", "tue", "sat", "fri", "sun", "sun", "mon", "mon", "tue", "~
## $ FFMC  <dbl> 86.2, 90.6, 90.6, 91.7, 89.3, 92.3, 92.3, 91.5, 91.0, 92.5, 92.5~
## $ DMC   <dbl> 26.2, 35.4, 43.7, 33.3, 51.3, 85.3, 88.9, 145.4, 129.5, 88.0, 88~
## $ DC    <dbl> 94.3, 669.1, 686.9, 77.5, 102.2, 488.0, 495.6, 608.2, 692.6, 698~
## $ ISI   <dbl> 5.1, 6.7, 6.7, 9.0, 9.6, 14.7, 8.5, 10.7, 7.0, 7.1, 7.1, 22.6, 0~
## $ temp  <dbl> 8.2, 18.0, 14.6, 8.3, 11.4, 22.2, 24.1, 8.0, 13.1, 22.8, 17.8, 1~
## $ RH    <int> 51, 33, 33, 97, 99, 29, 27, 86, 63, 40, 51, 38, 72, 42, 21, 44, ~
## $ wind  <dbl> 6.7, 0.9, 1.3, 4.0, 1.8, 5.4, 3.1, 2.2, 5.4, 4.0, 7.2, 4.0, 6.7,~
## $ rain  <dbl> 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,~
## $ area  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
print(nrow(subset(fire_data,area == 0)))
```

```
## [1] 247
```

```
describe(fire_data)
```

```
## fire_data
##
##  13  Variables      517  Observations
## ---------------------------------------------------------------------------
## X
##        n  missing distinct     Info     Mean      Gmd
##      517        0        9    0.982    4.669    2.648
##
## lowest : 1 2 3 4 5, highest: 5 6 7 8 9
##
## Value          1     2     3     4     5     6     7     8     9
## Frequency     48    73    55    91    30    86    60    61    13
## Proportion 0.093 0.141 0.106 0.176 0.058 0.166 0.116 0.118 0.025
## ---------------------------------------------------------------------------
```

```
## Y
##        n  missing distinct     Info     Mean      Gmd
##      517        0        7     0.92      4.3    1.309
##
## lowest : 2 3 4 5 6, highest: 4 5 6 8 9
##
## Value          2     3     4     5     6     8     9
## Frequency     44    64   203   125    74     1     6
## Proportion 0.085 0.124 0.393 0.242 0.143 0.002 0.012
## ---------------------------------------------------------------------------
## month
##        n  missing distinct
##      517        0       12
##
## lowest : apr aug dec feb jan, highest: mar may nov oct sep
##
## Value        apr   aug   dec   feb   jan   jul   jun   mar   may   nov   oct
## Frequency      9   184     9    20     2    32    17    54     2     1    15
## Proportion 0.017 0.356 0.017 0.039 0.004 0.062 0.033 0.104 0.004 0.002 0.029
##
## Value        sep
## Frequency    172
## Proportion 0.333
## ---------------------------------------------------------------------------
## day
##        n  missing distinct
##      517        0        7
##
## lowest : fri mon sat sun thu, highest: sat sun thu tue wed
##
## Value        fri   mon   sat   sun   thu   tue   wed
## Frequency     85    74    84    95    61    64    54
## Proportion 0.164 0.143 0.162 0.184 0.118 0.124 0.104
## ---------------------------------------------------------------------------
## FFMC
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      517        0      106    0.999    90.64    4.053     84.1     85.9
##      .25      .50      .75      .90      .95
##     90.2     91.6     92.9     94.3     95.1
##
## lowest : 18.7 50.4 53.4 63.5 68.2, highest: 95.8 95.9 96.0 96.1 96.2
## ---------------------------------------------------------------------------
## DMC
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      517        0      215        1    110.9    71.27    14.92    25.70
##      .25      .50      .75      .90      .95
##    68.60   108.30   142.40   195.18   231.10
##
## lowest :   1.1   2.4   3.0   3.2   3.6, highest: 276.3 284.9 287.2 290.0 291.3
## ---------------------------------------------------------------------------
## DC
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      517        0      219        1    547.9    257.3    43.58    80.80
##      .25      .50      .75      .90      .95
```

```
##   437.70   664.20   713.90   758.10   795.30
##
## lowest :   7.9   9.3  15.3  15.5  15.8, highest: 825.1 844.0 849.3 855.3 860.6
## ---------------------------------------------------------------------------
## ISI
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      517        0      119        1    9.022    4.631      2.6      3.8
##      .25      .50      .75      .90      .95
##      6.5      8.4     10.8     14.3     17.0
##
## lowest :  0.0  0.4  0.7  0.8  1.1, highest: 20.3 21.3 22.6 22.7 56.1
## ---------------------------------------------------------------------------
## temp
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      517        0      192        1    18.89    6.494     8.20    11.20
##      .25      .50      .75      .90      .95
##    15.50    19.30    22.80    25.98    27.90
##
## lowest :  2.2  4.2  4.6  4.8  5.1, highest: 32.3 32.4 32.6 33.1 33.3
## ---------------------------------------------------------------------------
## RH
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      517        0       75    0.999    44.29    18.01       24       27
##      .25      .50      .75      .90      .95
##       33       42       53       68       77
##
## lowest :  15  17  18  19  20, highest:  94  96  97  99 100
## ---------------------------------------------------------------------------
## wind
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      517        0       21    0.994    4.018    2.007      1.3      1.8
##      .25      .50      .75      .90      .95
##      2.7      4.0      4.9      6.3      7.6
##
## lowest : 0.4 0.9 1.3 1.8 2.2, highest: 7.6 8.0 8.5 8.9 9.4
## ---------------------------------------------------------------------------
## rain
##        n  missing distinct     Info     Mean      Gmd
##      517        0        7    0.046  0.02166  0.04312
##
## lowest : 0.0 0.2 0.4 0.8 1.0, highest: 0.4 0.8 1.0 1.4 6.4
##
## Value        0.0   0.2   0.4   0.8   1.0   1.4   6.4
## Frequency    509     2     1     2     1     1     1
## Proportion 0.985 0.004 0.002 0.004 0.002 0.002 0.002
## ---------------------------------------------------------------------------
## area
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      517        0      251    0.891    12.85     22.7     0.00     0.00
##      .25      .50      .75      .90      .95
##     0.00     0.52     6.57    25.26    48.71
##
## lowest :    0.00    0.09    0.17    0.21    0.24
## highest:  200.94  212.88  278.53  746.28 1090.84
```

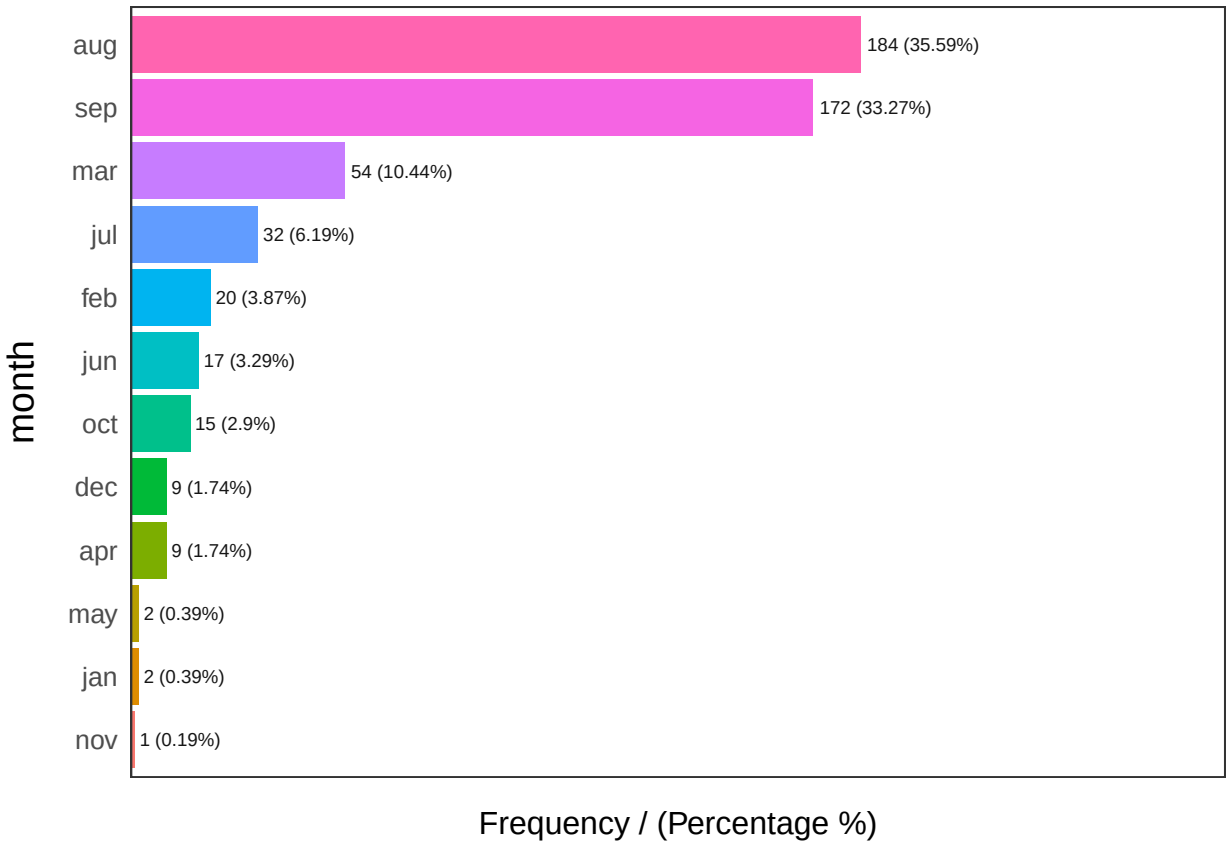
```
##
## Value          0    10    20    30    40    50    60    70    80    90   100
## Frequency    366    82    17    15     9     6     4     3     1     2     2
## Proportion 0.708 0.159 0.033 0.029 0.017 0.012 0.008 0.006 0.002 0.004 0.004
##
## Value        110   150   170   190   200   210   280   750  1090
## Frequency      1     1     1     1     2     1     1     1     1
## Proportion 0.002 0.002 0.002 0.002 0.004 0.002 0.002 0.002 0.002
##
## For the frequency table, variable is rounded to the nearest 10
## ------------------------------------------------------------------------------
```

```
fire_data =within(fire_data,{
    season=NA
    season[month %in% c("dec","jan","feb")]='1winter'
    season[month %in% c("oct","nov")]='4autumn'
    season[month %in%  c("jun","jul","aug", "sep")]='3summer'
    season[month %in%  c("mar","apr","may")]='2spring'
})
view(fire_data)
```

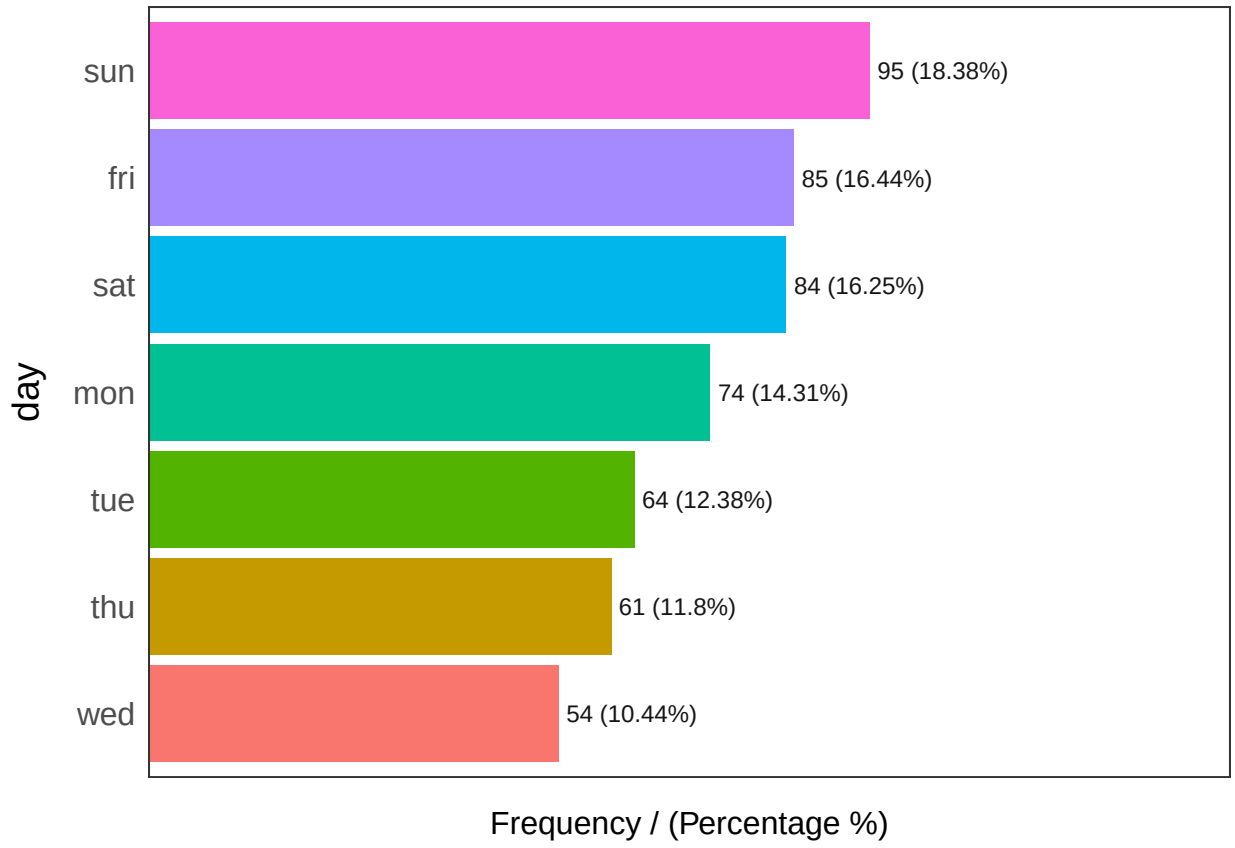
Transform area as log of area Around 50% of observations have 0 value. This skews the data. The amount of data is to large to drop so we will perform a transformation on the data by adding 1.1 and then perform log transformation on it

```
fire_data$logarea <- log(fire_data$area + 1.1)
view(fire_data)
```

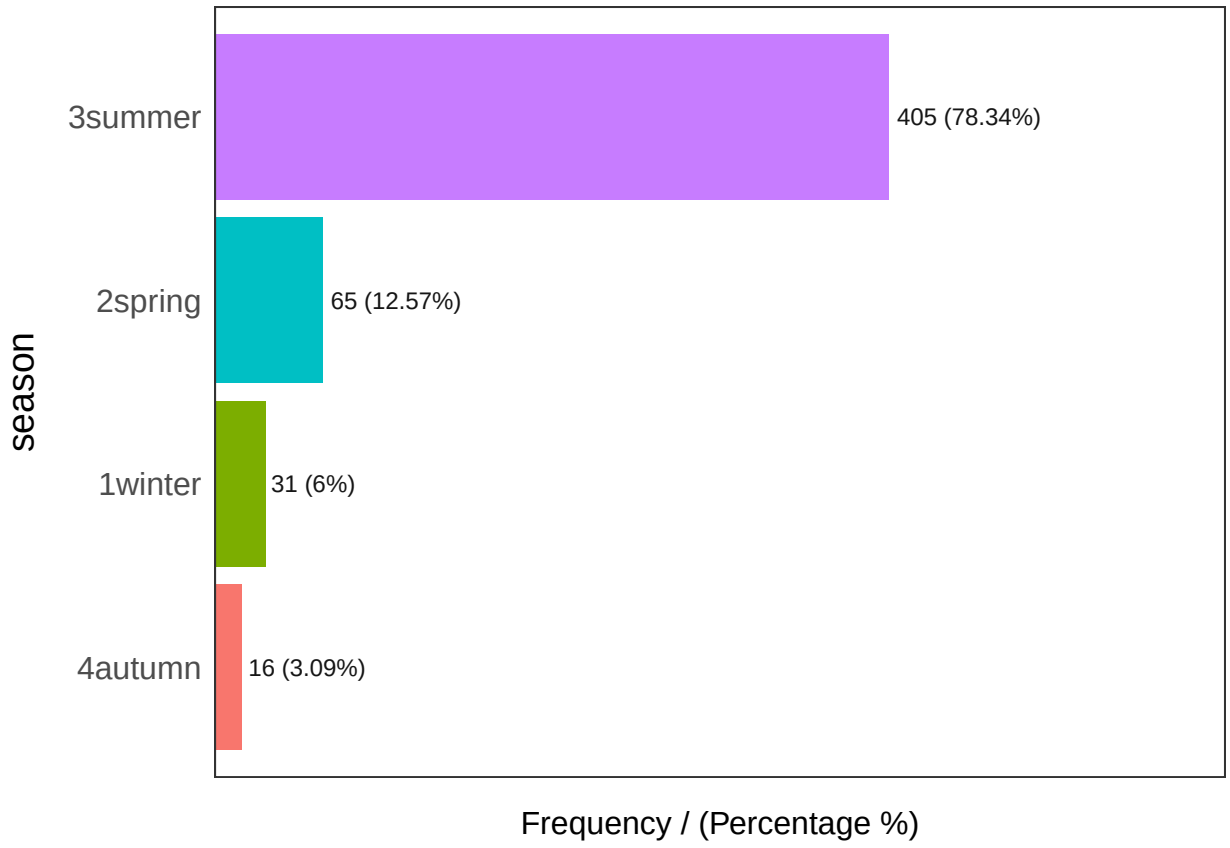
```
##          variable q_zeros     p_zeros q_na p_na q_inf p_inf      type unique
## X               X       0 0.000000000    0    0     0     0   integer      9
## Y               Y       0 0.000000000    0    0     0     0   integer      7
## month       month       0 0.000000000    0    0     0     0 character     12
## day           day       0 0.000000000    0    0     0     0 character      7
## FFMC         FFMC       0 0.000000000    0    0     0     0   numeric    106
## DMC           DMC       0 0.000000000    0    0     0     0   numeric    215
## DC             DC       0 0.000000000    0    0     0     0   numeric    219
## ISI           ISI       1 0.001934236    0    0     0     0   numeric    119
## temp         temp       0 0.000000000    0    0     0     0   numeric    192
## RH             RH       0 0.000000000    0    0     0     0   integer     75
## wind         wind       0 0.000000000    0    0     0     0   numeric     21
## rain         rain     509 0.984526112    0    0     0     0   numeric      7
## area         area     247 0.477756286    0    0     0     0   numeric    251
## season     season       0 0.000000000    0    0     0     0 character      4
## logarea   logarea       0 0.000000000    0    0     0     0   numeric    251
```



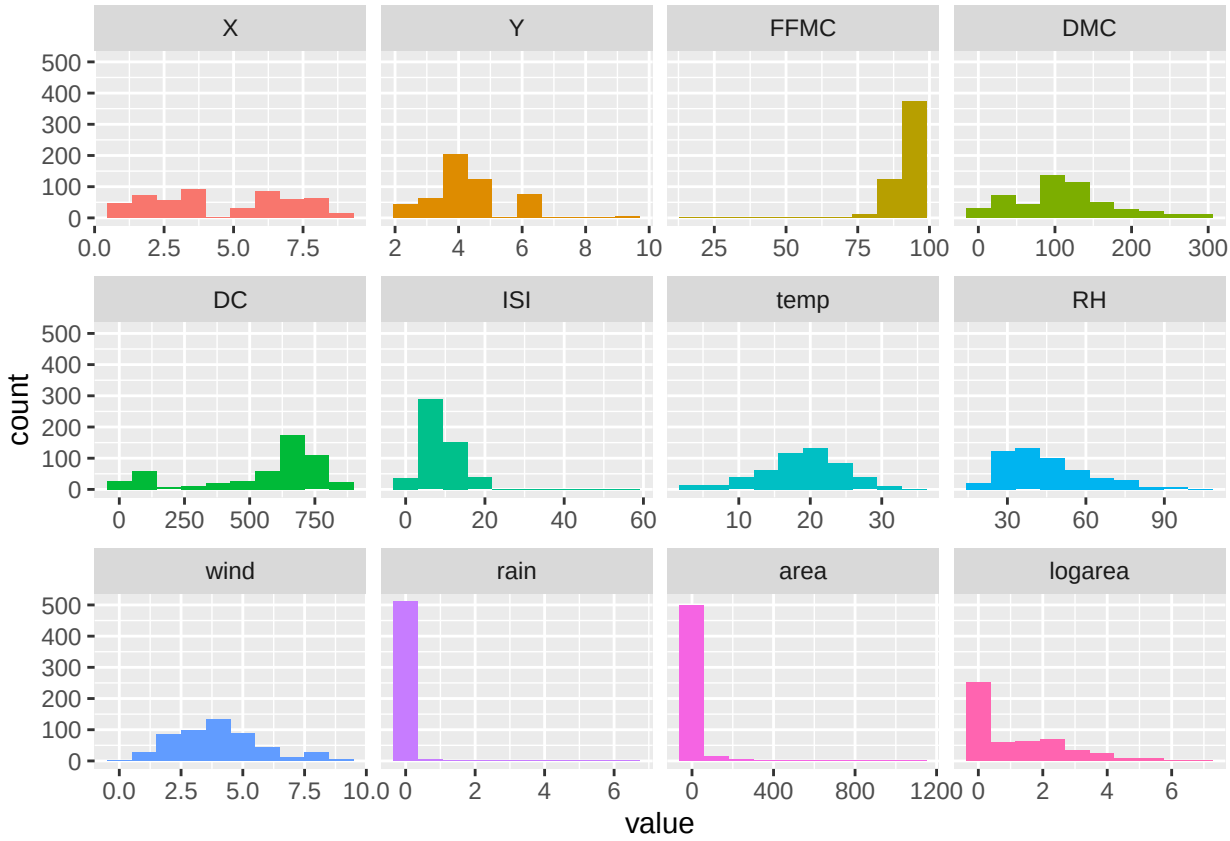

```
##    month frequency percentage cumulative_perc
## 1    aug       184      35.59           35.59
## 2    sep       172      33.27           68.86
## 3    mar        54      10.44           79.30
## 4    jul        32       6.19           85.49
## 5    feb        20       3.87           89.36
## 6    jun        17       3.29           92.65
## 7    oct        15       2.90           95.55
## 8    apr         9       1.74           97.29
## 9    dec         9       1.74           99.03
## 10   jan         2       0.39           99.42
## 11   may         2       0.39           99.81
## 12   nov         1       0.19          100.00
```



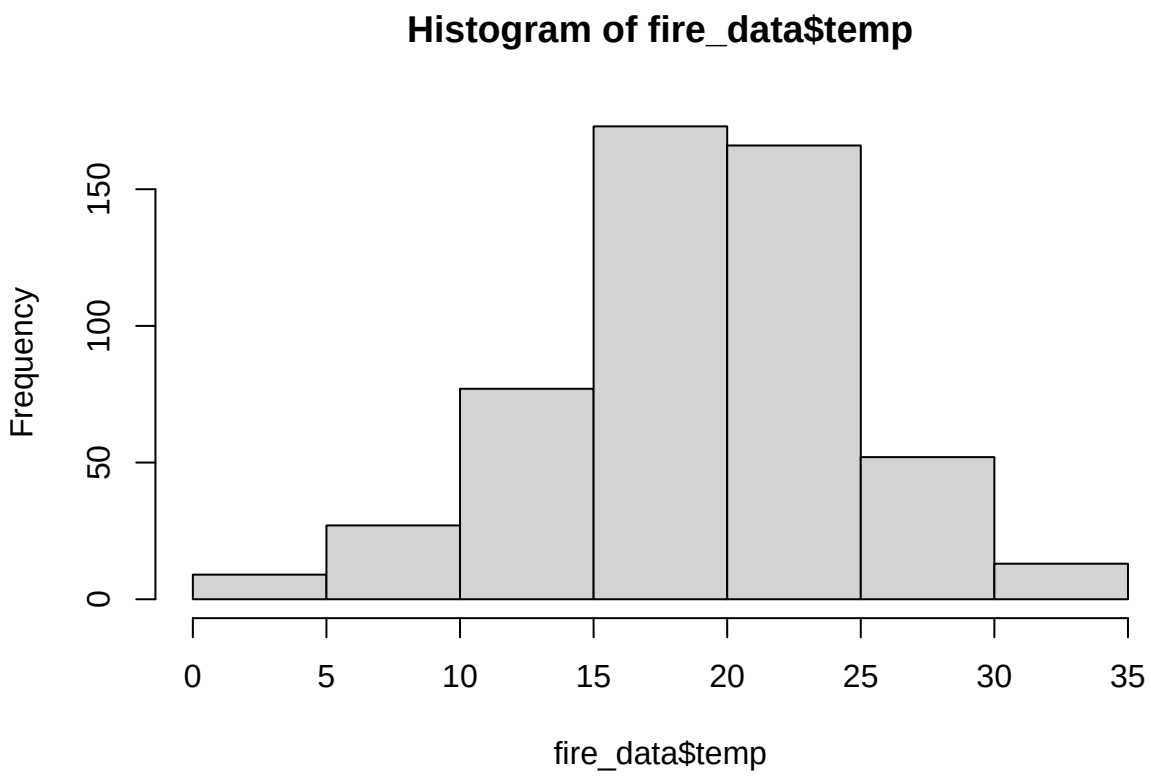

```
##   day frequency percentage cumulative_perc
## 1 sun        95      18.38           18.38
## 2 fri        85      16.44           34.82
## 3 sat        84      16.25           51.07
## 4 mon        74      14.31           65.38
## 5 tue        64      12.38           77.76
## 6 thu        61      11.80           89.56
## 7 wed        54      10.44          100.00
```



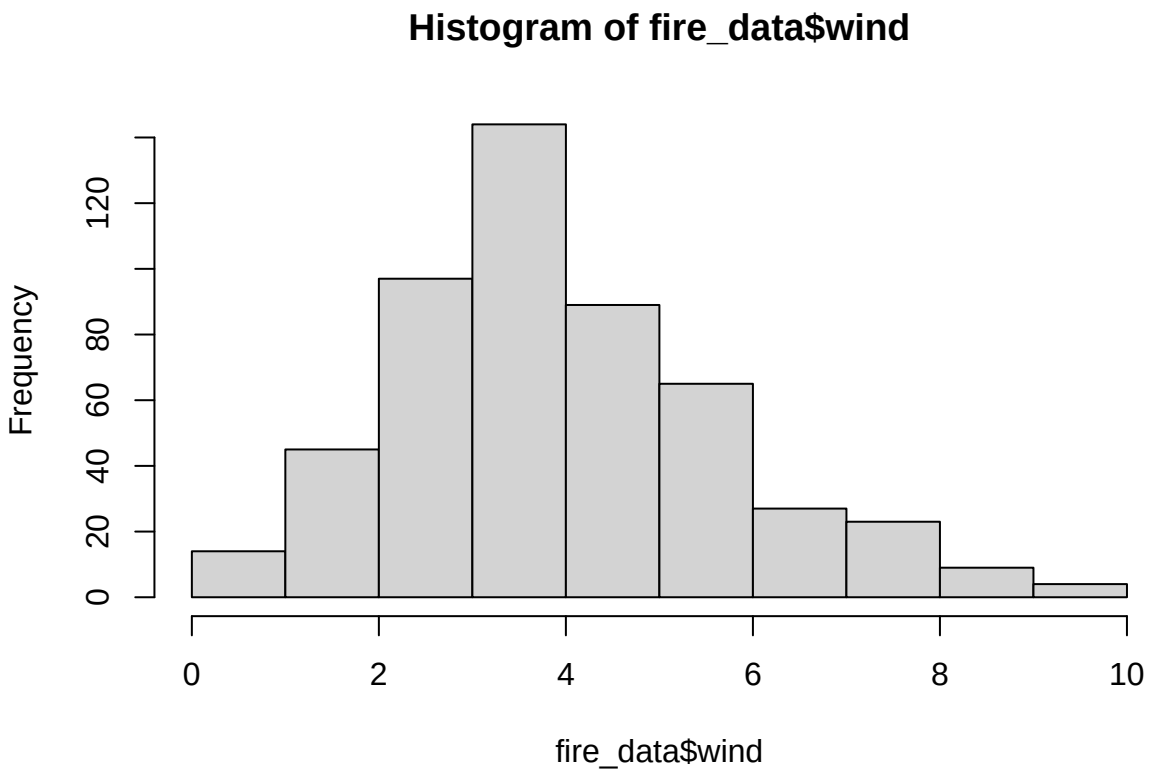

```
##    season frequency percentage cumulative_perc
## 1 3summer       405      78.34           78.34
## 2 2spring        65      12.57           90.91
## 3 1winter        31       6.00           96.91
## 4 4autumn        16       3.09          100.00

## [1] "Variables processed: month, day, season"
```



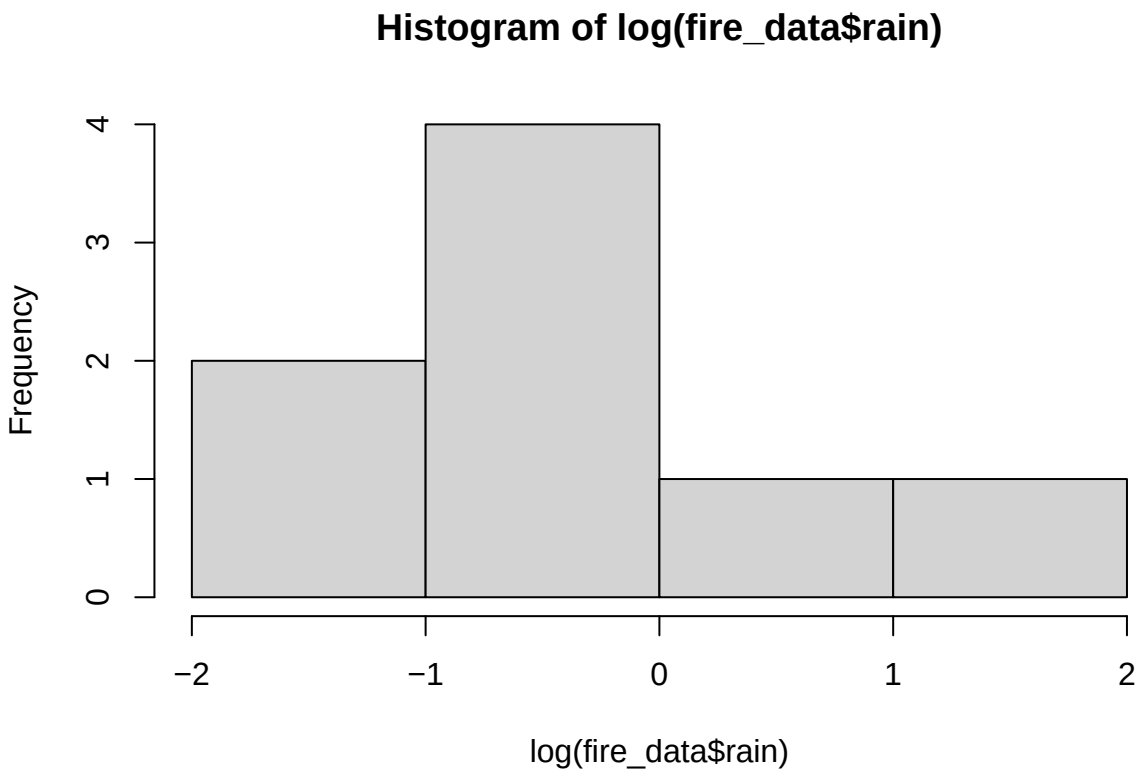

```
hist(fire_data$temp)
```




```
hist(fire_data$wind)
```

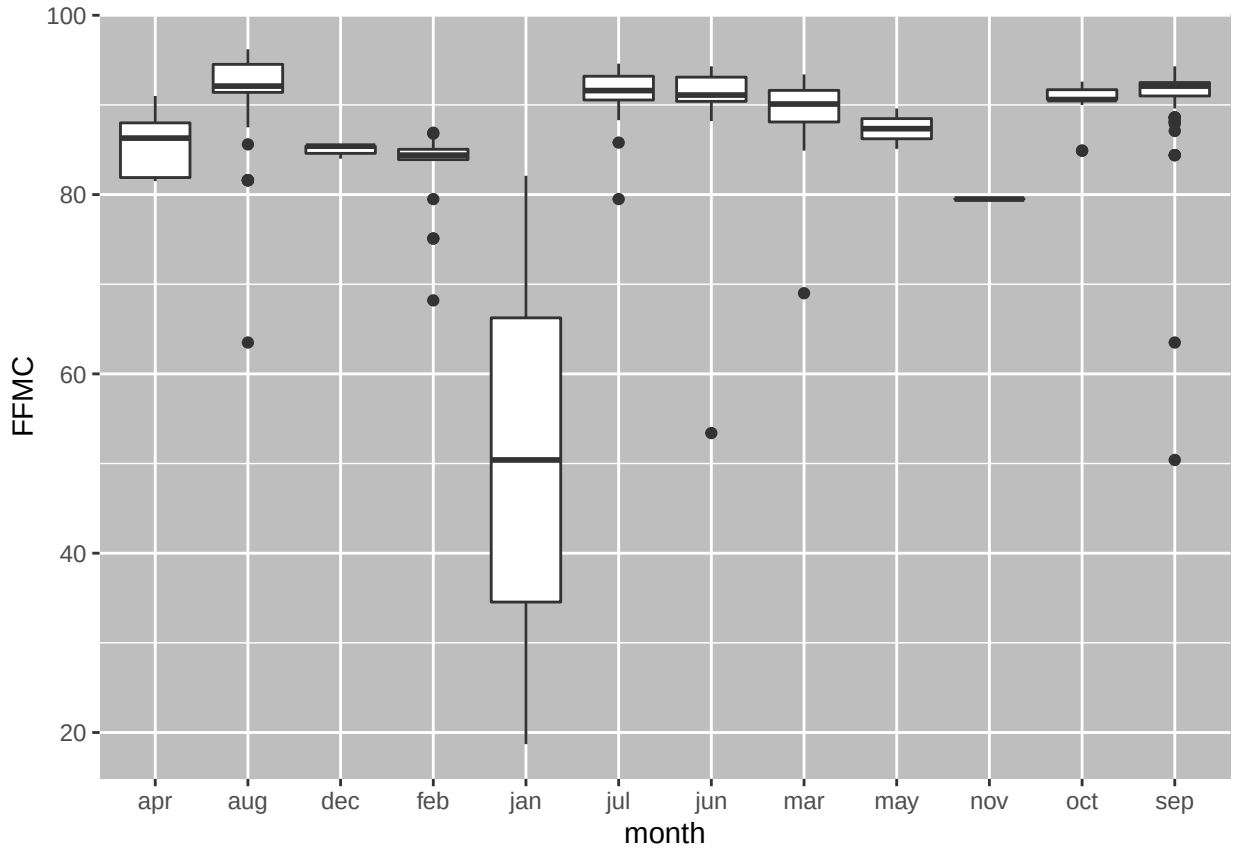



```
hist(log(fire_data$rain ))
```

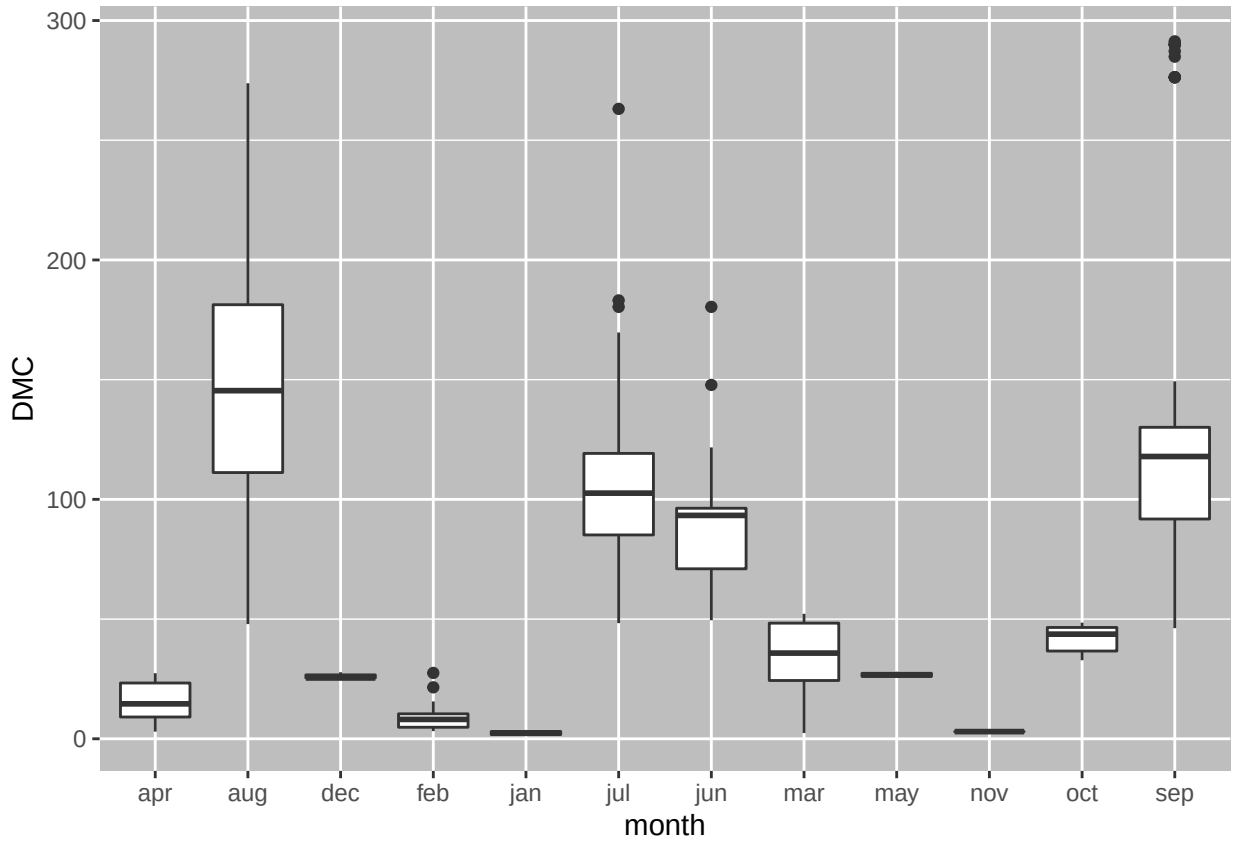



```
# Common function for Box plot so that it can be used with multiple parameters (aka columns)
bx_plt_func <- function (x, y){
                ggplot(data = fire_data) +
                aes_string(x = x, y =y) +
                geom_boxplot() +
                theme(panel.background = element_rect(fill = "grey"))
}
x_var_month <- names(fire_data[3]) ## 3rd column is month
x_var_day <- names(fire_data[4]) ## 4th column is day
y_var <- names(fire_data[5:12]) ## Column 5 onwards are being used for analysis
z_var <- "Monthwise Data Analysis"
# Box plots by month for all variables earmarked for analysis
month_box <- map2(x_var_month, y_var, bx_plt_func)

print(month_box)
```

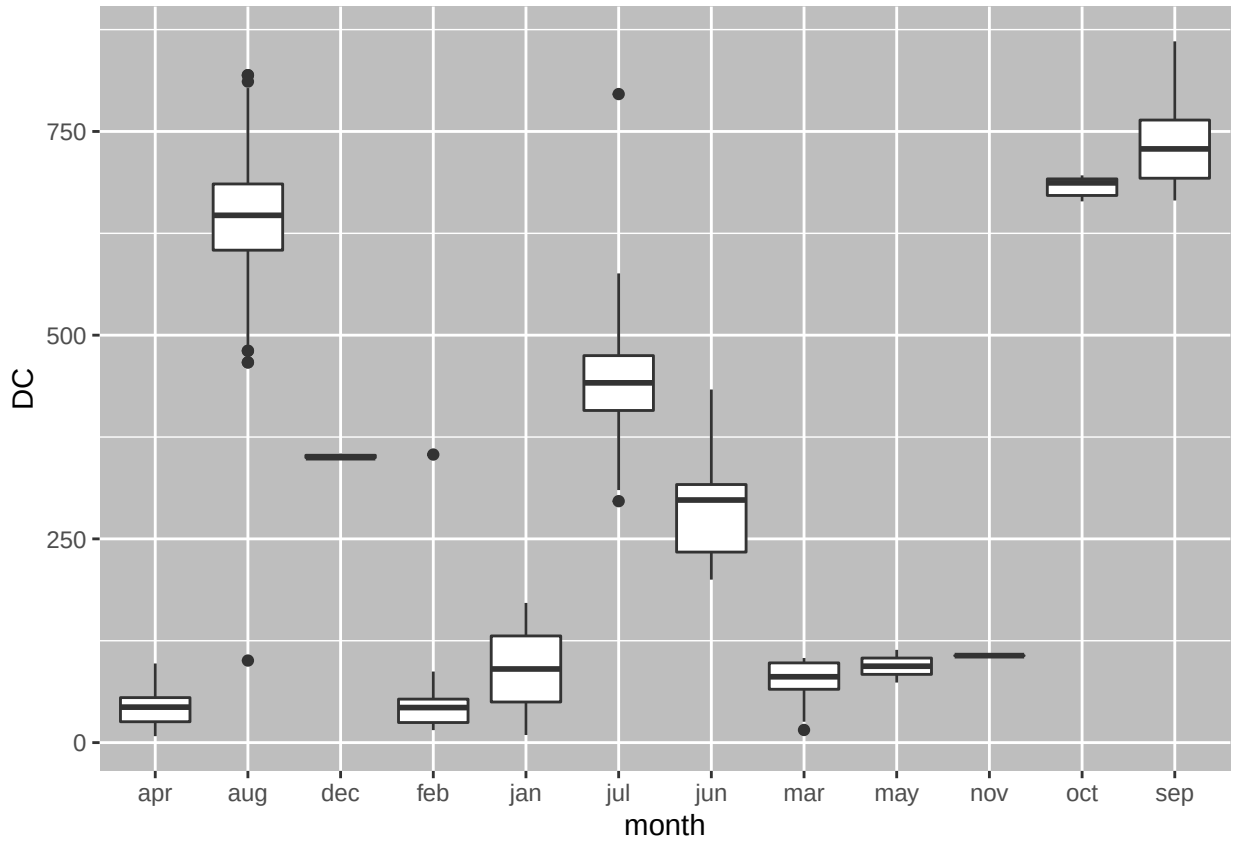
```
## [[1]]
```


100
80
60
40
20
FFMC
apr
aug
dec
feb
jan
jul
jun
mar
may
nov
oct
sep
month

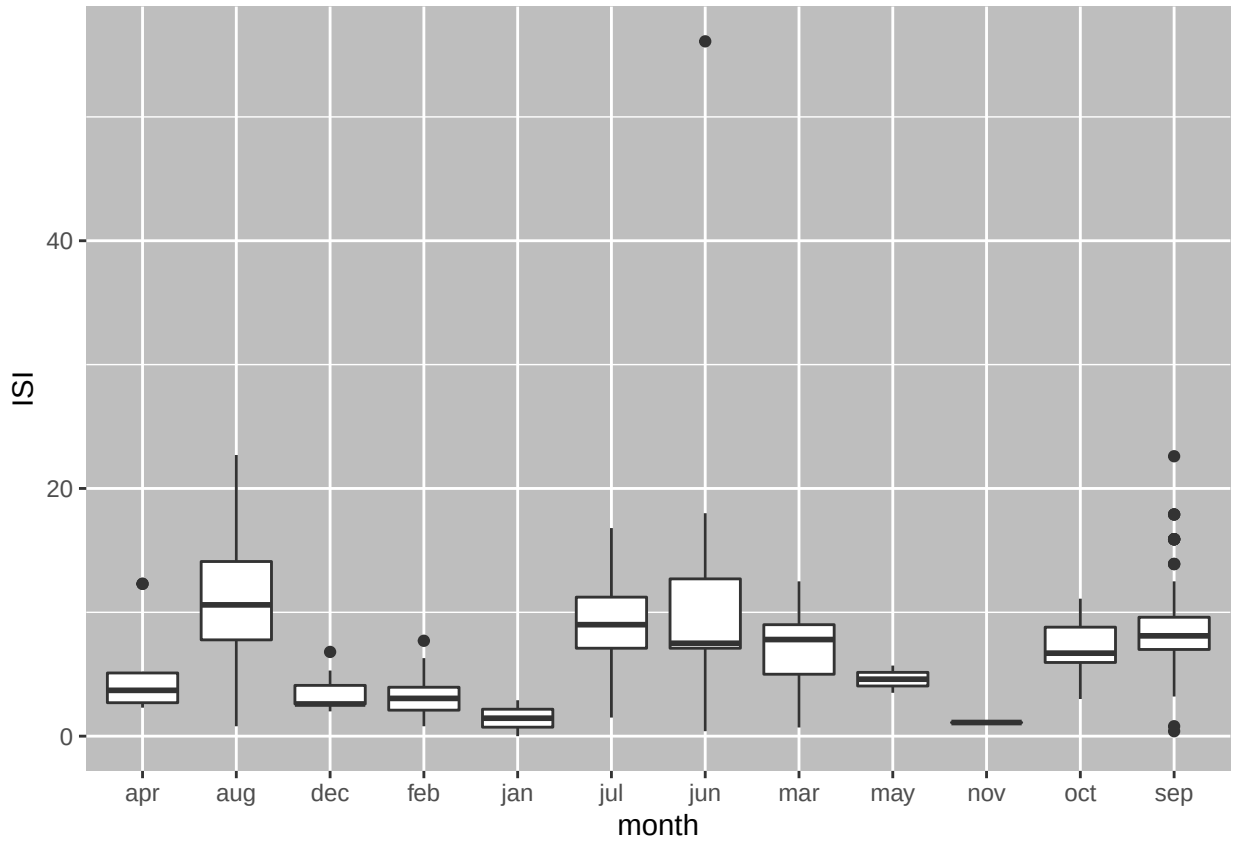
```
##
## [[2]]
```


300
200
100
0
DMC
apr
aug
dec
feb
jan
jul
jun
mar
may
nov
oct
sep
month

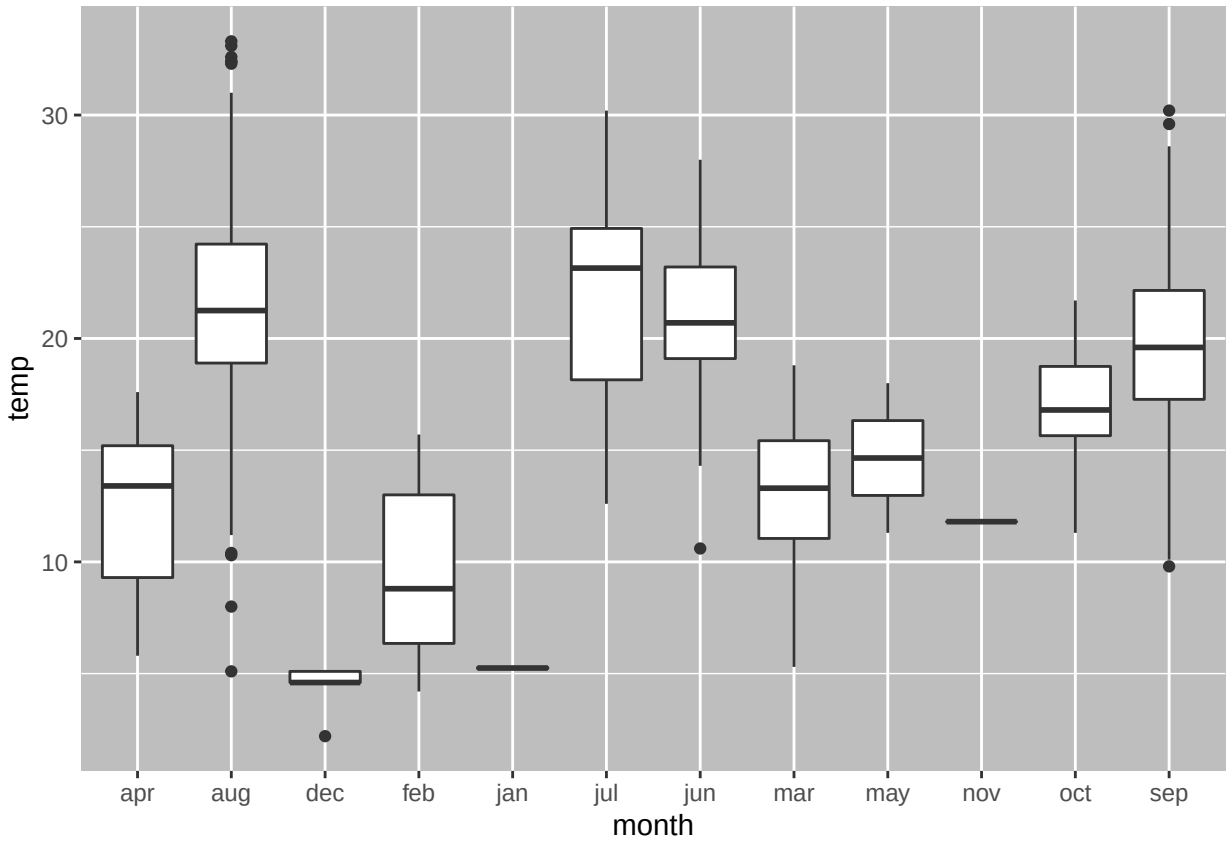
```
##
## [[3]]
```


750
500
250
0
DC
apr
aug
dec
feb
jan
jul
jun
mar
may
nov
oct
sep
month

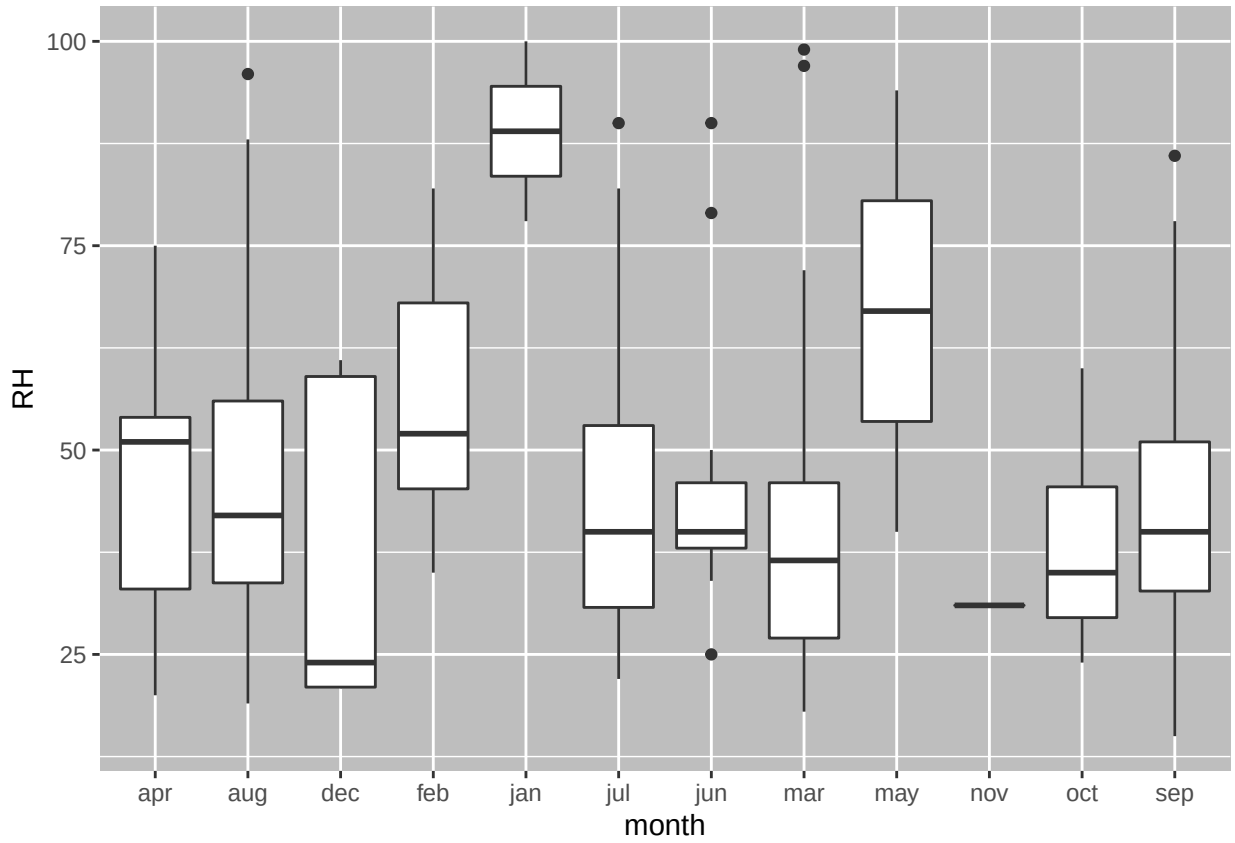
```
##
## [[4]]
```



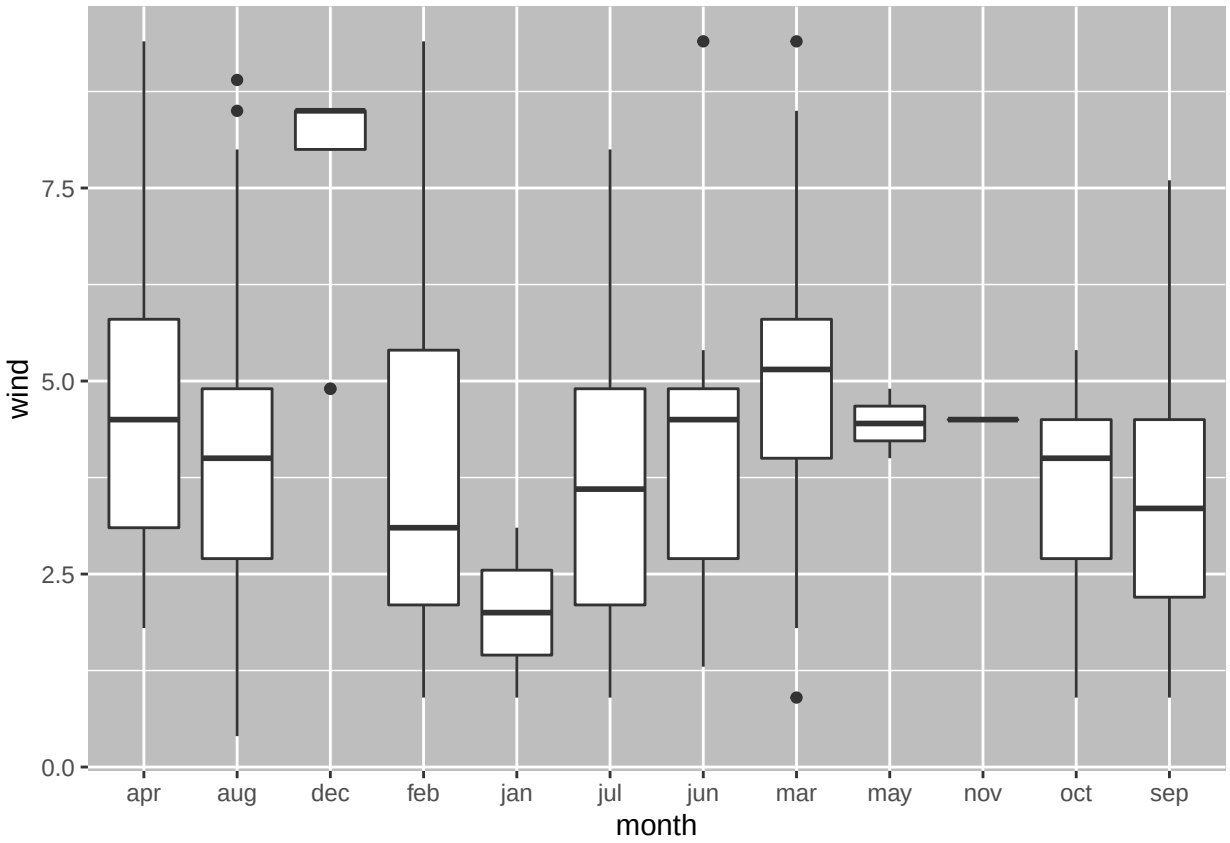

```
##
## [[5]]
```


30
20
10
temp
apr
aug
dec
feb
jan
jul
jun
mar
may
nov
oct
sep
month

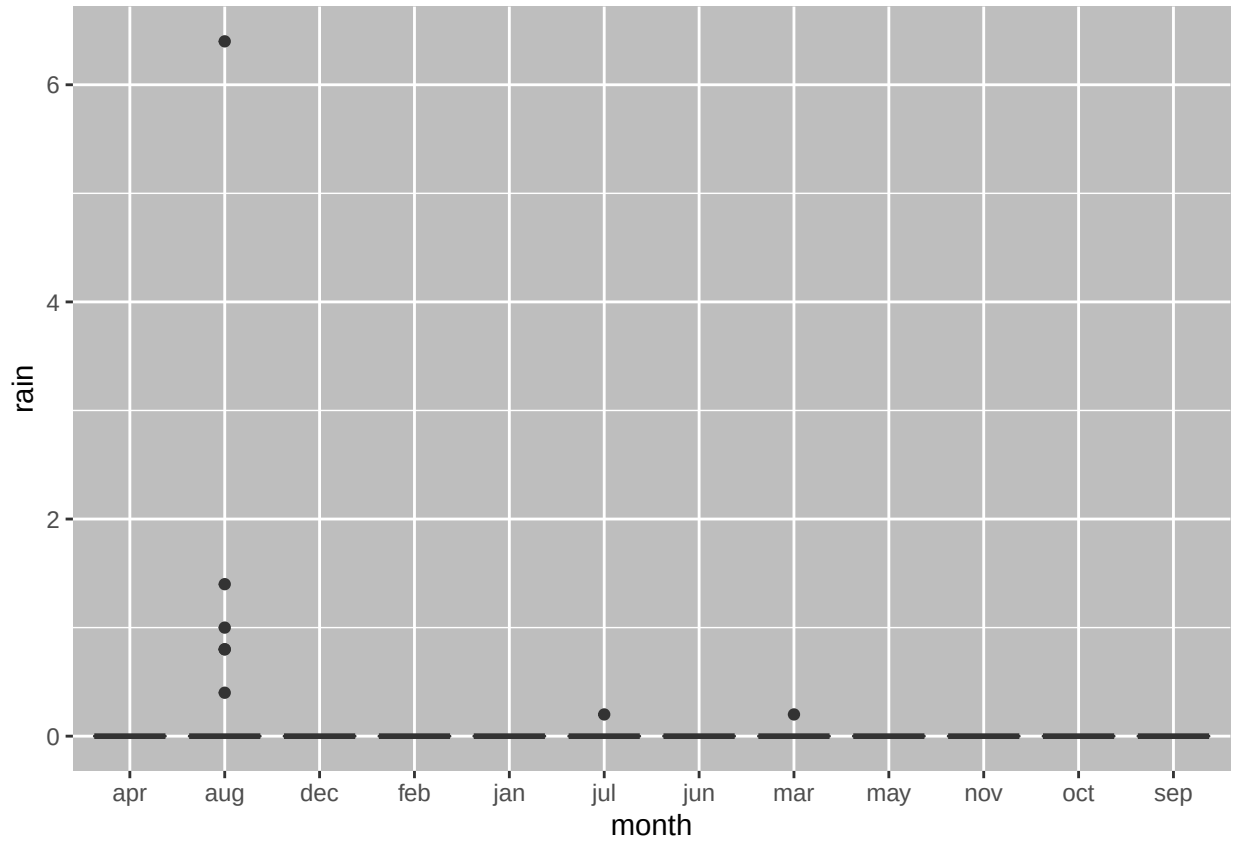
```
##
## [[6]]
```


100
75
50
25
RH
apr
aug
dec
feb
jan
jul
jun
mar
may
nov
oct
sep
month

```
##
## [[7]]
```


wind
7.5
5.0
2.5
0.0
apr
aug
dec
feb
jan
jul
jun
mar
may
nov
oct
sep
month

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

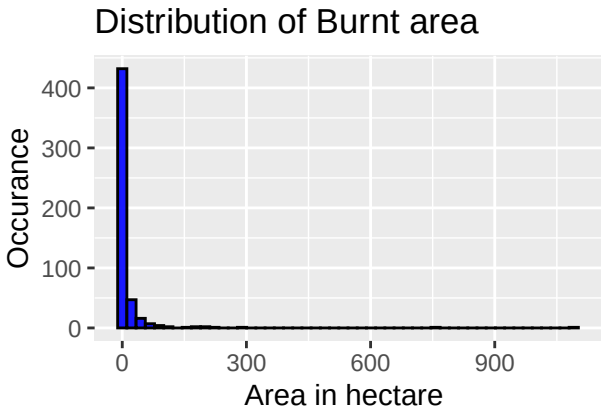


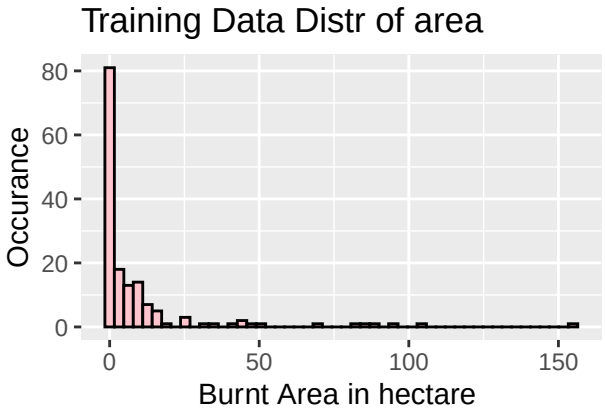

```
sample_size = floor(0.7*nrow(fire_data))
set.seed(777)
# randomly split data in r
picked = sample(seq_len(nrow(fire_data)),size = sample_size)
development =fire_data[picked,]
holdout =fire_data[-picked,]
view(holdout)
view(development)
```

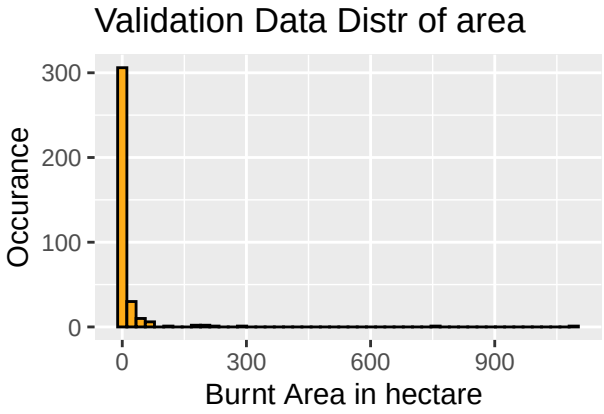
```
hist_of_area_dist <- fire_data %>%
ggplot() + aes(x = area) +
geom_histogram( bins=50, fill="blue", color="black", alpha=0.9) +
labs(
x = "Area in hectare ", y="Occurance",
title = 'Distribution of Burnt area'
)

# Distribution of Burnt Area
 hist_of_logarea_dist <- (fire_data %>%
ggplot() + aes(x = logarea) +
geom_histogram( aes(y = ..density..), fill="blue", color="black")
+ labs(
x = "Area in hectare ", y="Occurance",
title = 'Distr of Log Transformed Burnt-area'
) )
```

```r
#  holdout - Area distribution for Training Data

holdout_hist_of_area_dist <- holdout %>%
ggplot() + aes(x = area) +
geom_histogram( bins=50, fill="pink", color="black", alpha=0.9) +
labs(
x = "Burnt Area in hectare ", y="Occurance",
title = 'Training Data Distr of area'
)

 holdout_hist_of_logarea_dist <- (holdout %>%
ggplot() + aes(x = logarea) +
geom_histogram( aes(y = ..density..), fill="pink", color="black") +  labs(
x = "Burnt Area in hectare ", y="Occurance",
title = 'Training Data Distr of Log(area)'
) )


# development - Area distribution for Training Data

development_hist_of_area_dist <- development %>%
ggplot() + aes(x = area) +
geom_histogram( bins=50, fill="orange", color="black", alpha=0.9) +
labs(
x = "Burnt Area in hectare ", y="Occurance",
title = 'Validation Data Distr of area'
)

 development_hist_of_logarea_dist <- (development %>%
ggplot() + aes(x = logarea) +
geom_histogram( aes(y = ..density..), fill="orange", color="black") +  labs(
x = "Burnt Area in hectare ", y="Occurance",
title = 'Validation Data Distr of Log(area)'
) )
```

```r
plot_grid(hist_of_area_dist, NULL, holdout_hist_of_area_dist,development_hist_of_area_dist )
```

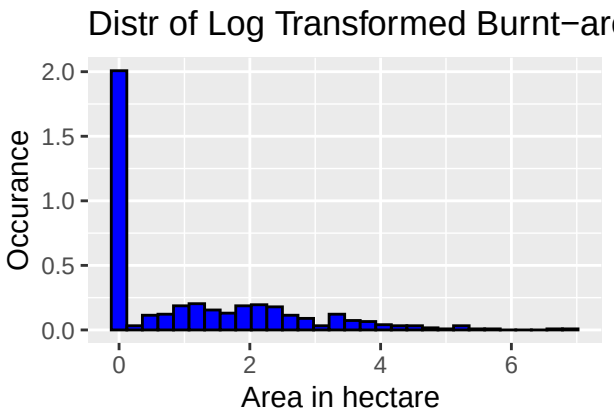


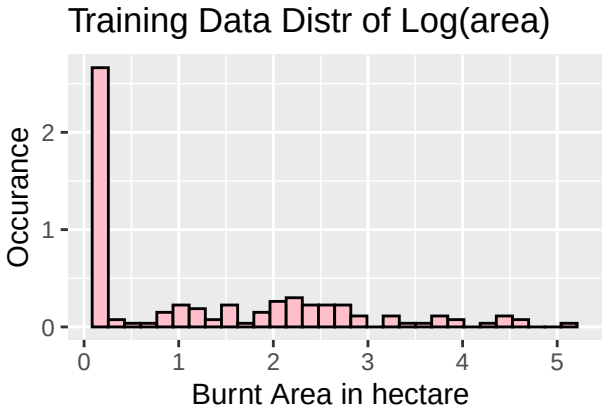


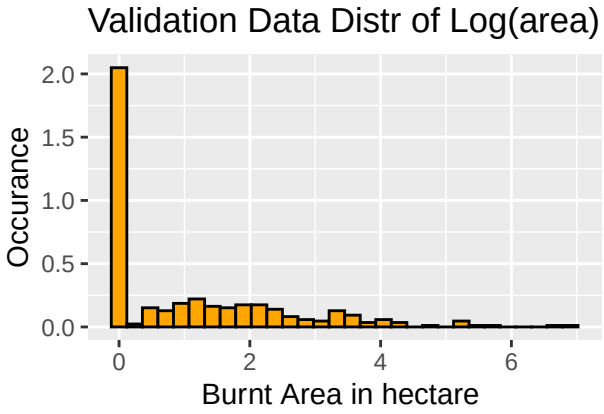


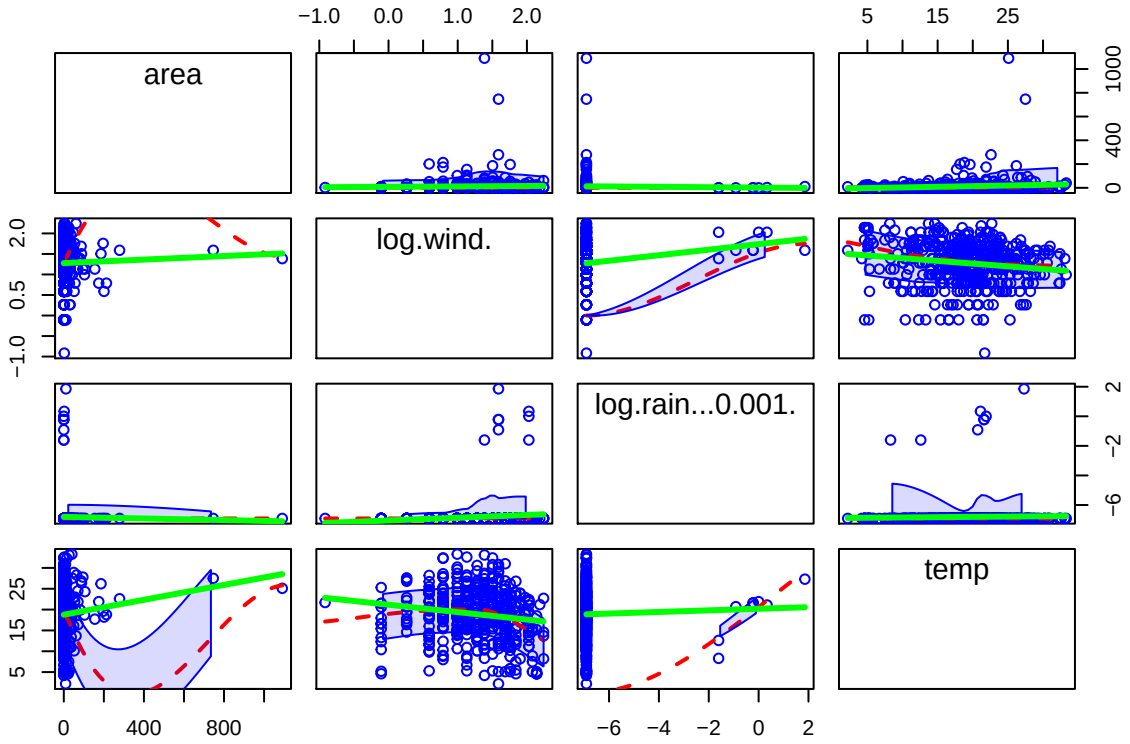

```
plot_grid(hist_of_logarea_dist, NULL, holdout_hist_of_logarea_dist, development_hist_of_logarea_dist)
```

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


Distr of Log Transformed Burnt−area
Occurance
2.0
1.5
1.0
0.5
0.0
0
2
4
6
Area in hectare


Training Data Distr of Log(area)
Occurance
2
1
0
0
1
2
3
4
5
Burnt Area in hectare


Validation Data Distr of Log(area)
Occurance
2.0
1.5
1.0
0.5
0.0
0
2
4
6
Burnt Area in hectare



#———————— Coeftest ————————————————————-

```
model_h_long <- lm(formula = area ~ log(rain+ 0.001) + temp + (wind) , data=holdout)
coeftest(model_h_long, vcov=vcovHAC)
```

```
##
## t test of coefficients:
##
##                       Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)       -13.6216905   6.6978440 -2.0337 0.0437169 *
## log(rain + 0.001)  -1.7299839   0.5111256 -3.3847 0.0009069 ***
## temp                0.6148402   0.2418863  2.5419 0.0120277 *
## wind                0.0094852   0.8615275  0.0110 0.9912301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
set.seed(5600)
shapiro.test(sample(model_h_long$residuals, size = 5000,replace=TRUE))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(model_h_long$residuals, size = 5000, replace = TRUE)
## W = 0.53835, p-value < 2.2e-16
```

```
model_d_long <- lm(formula = area ~ log(rain+0.001) + temp + log(wind) , data=development)
coeftest(model_d_long, vcov=vcovHAC)
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -40.50723   26.86839 -1.5076   0.1325
## log(rain + 0.001)  -2.41449    1.04558 -2.3092   0.0215 *
## temp                1.43894    0.90038  1.5981   0.1109
## log(wind)           8.66469    5.77227  1.5011   0.1342
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
set.seed(5600)
shapiro.test(sample(model_d_long$residuals, size = 5000,replace=TRUE))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(model_d_long$residuals, size = 5000, replace = TRUE)
## W = 0.20591, p-value < 2.2e-16
```

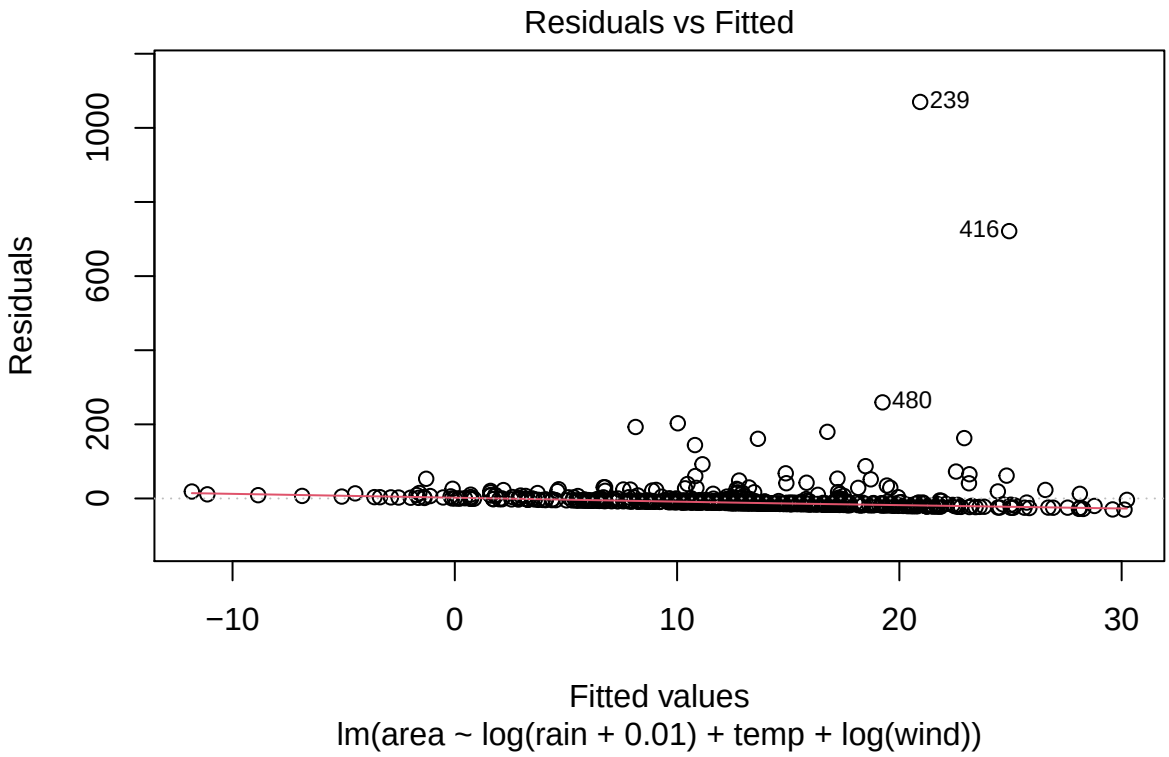
```
model_long <- lm(formula = area ~ log(rain+0.01) + temp + log(wind) , data=fire_data)
coeftest(model_long, vcov=vcovHAC)
```

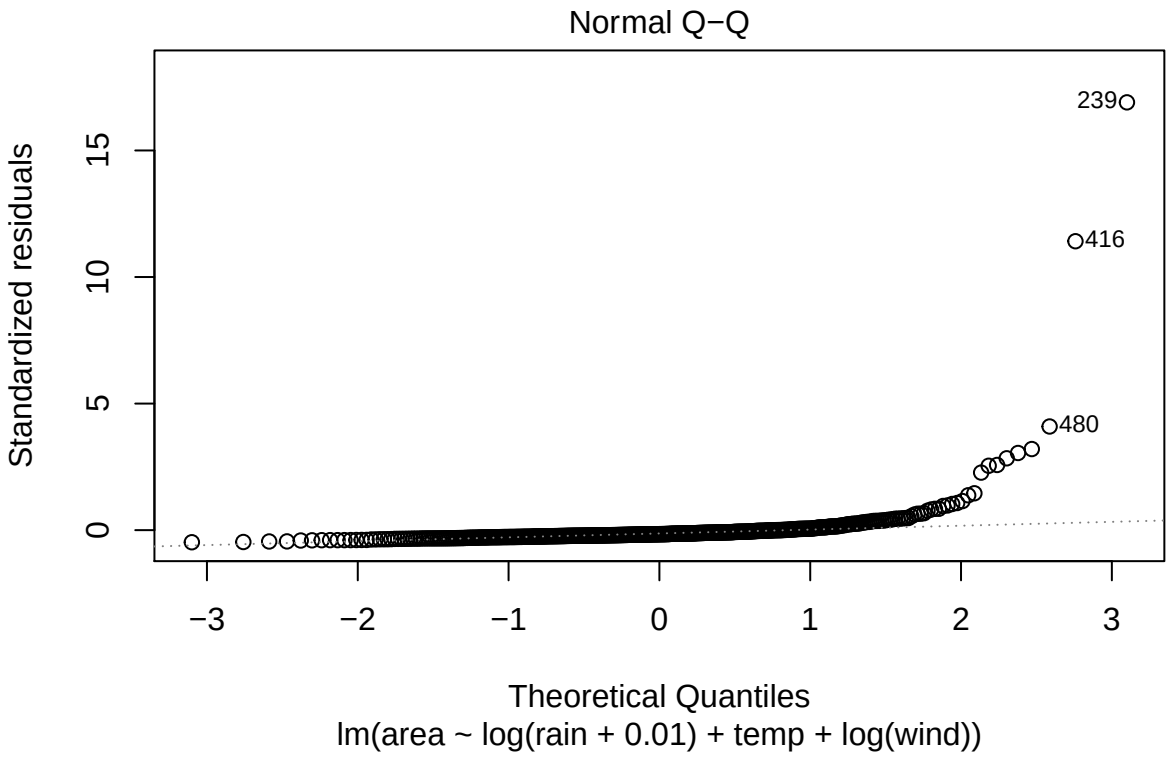
```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -32.04558   16.20735 -1.9772 0.048551 *
## log(rain + 0.01)  -3.36483    1.17604 -2.8612 0.004393 **
## temp               1.16293    0.57401  2.0260 0.043283 *
## log(wind)          5.98406    3.18027  1.8816 0.060454 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

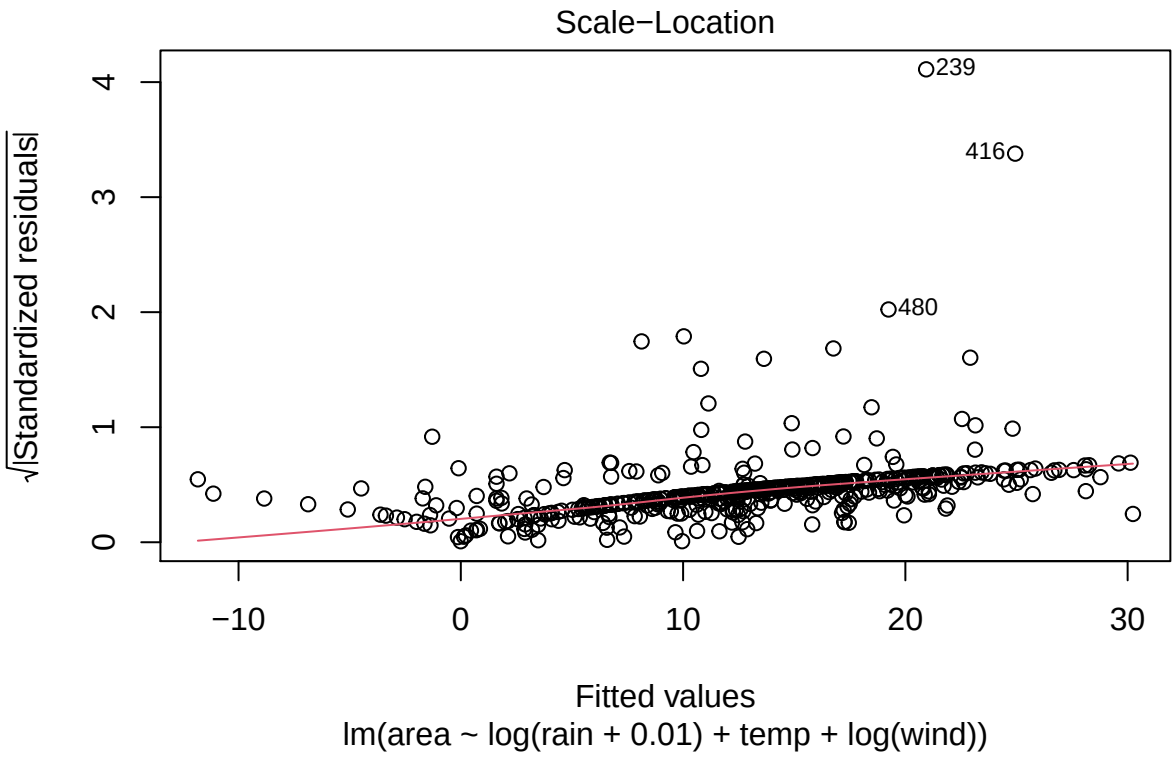
```
set.seed(5600)
shapiro.test(sample(model_long$residuals, size = 5000,replace=TRUE))
```

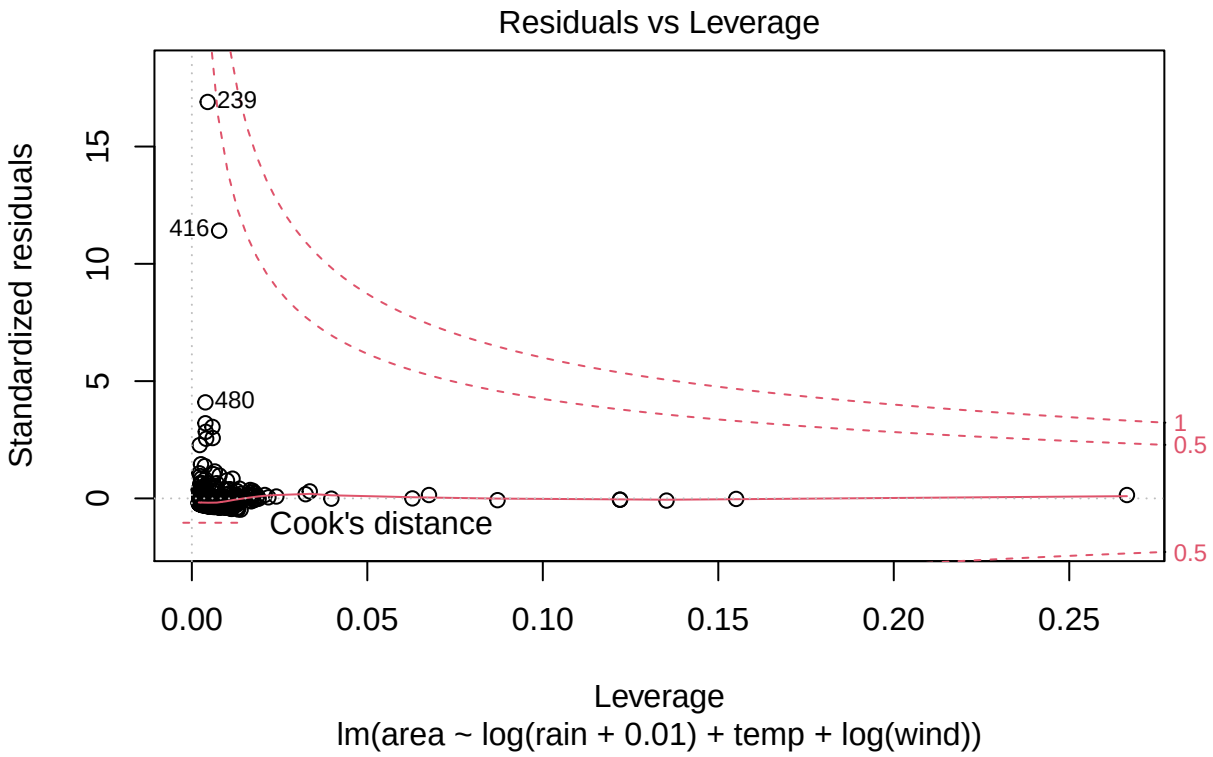
```
##
##  Shapiro-Wilk normality test
##
## data:  sample(model_long$residuals, size = 5000, replace = TRUE)
## W = 0.21347, p-value < 2.2e-16
```

```
plot(model_long)
```

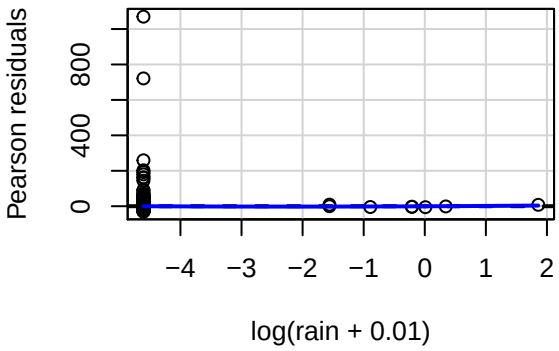

Residuals vs Fitted
Residuals
0
200
600
1000
239
416
480
−10
0
10
20
30
Fitted values
lm(area ~ log(rain + 0.01) + temp + log(wind))

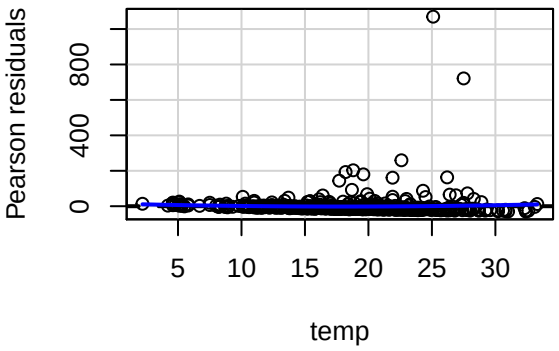

Normal Q−Q
Standardized residuals
0
5
10
15
239
416
480
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(area ~ log(rain + 0.01) + temp + log(wind))

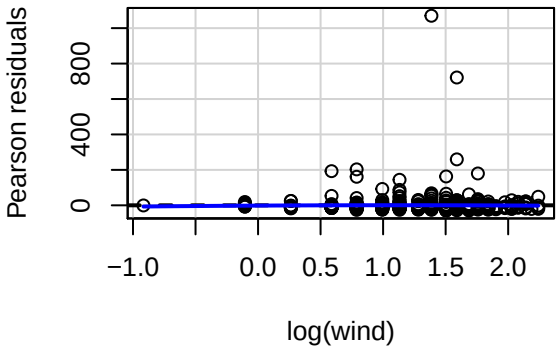

Scale−Location
√|Standardized residuals|
239
416
480
0
1
2
3
4
−10
0
10
20
30
Fitted values
lm(area ~ log(rain + 0.01) + temp + log(wind))

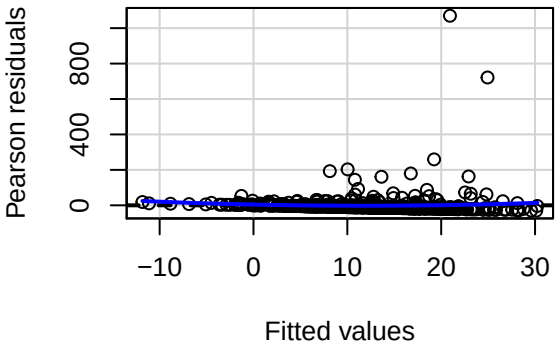



```
residualPlots(model_long)
```










```
##                  Test stat Pr(>|Test stat|)
## log(rain + 0.01)    0.0801           0.9362
## temp                0.8728           0.3832
## log(wind)          -0.2421           0.8088
## Tukey test          1.2501           0.2113
```