IST 687 Final

Tajudeen Abdulazeez and Greg Miller

December 11, 2018

Load Data

```
library(tidyverse)
```

```
## -- Attaching packages -----------------------------------------------------------------------
-------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.2.1     v forcats 0.3.0
```

```
## -- Conflicts --------------------------------------------------------------------------------
-------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
#library(clusterSim) #normalization
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

```
library(ggplot2)
```

Load the dataset and view the first 5 rows

```
df <- read_excel('./ENB2012_data.xlsx')
head(df)
```

```
## # A tibble: 6 x 10
##      X1    X2    X3    X4    X5    X6    X7    X8    Y1    Y2
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  0.98  514.  294   110.     7     2     0     0  15.6  21.3
## 2  0.98  514.  294   110.     7     3     0     0  15.6  21.3
## 3  0.98  514.  294   110.     7     4     0     0  15.6  21.3
## 4  0.98  514.  294   110.     7     5     0     0  15.6  21.3
## 5  0.9   564.  318.  122.     7     2     0     0  20.8  28.3
## 6  0.9   564.  318.  122.     7     3     0     0  21.5  25.4
```

Summary Statistics of the Data Frame

```
summary(df)
```

```
##        X1               X2              X3              X4       
##  Min.   :0.6200   Min.   :514.5   Min.   :245.0   Min.   :110.2  
##  1st Qu.:0.6825   1st Qu.:606.4   1st Qu.:294.0   1st Qu.:140.9  
##  Median :0.7500   Median :673.8   Median :318.5   Median :183.8  
##  Mean   :0.7642   Mean   :671.7   Mean   :318.5   Mean   :176.6  
##  3rd Qu.:0.8300   3rd Qu.:741.1   3rd Qu.:343.0   3rd Qu.:220.5  
##  Max.   :0.9800   Max.   :808.5   Max.   :416.5   Max.   :220.5  
##        X5             X6             X7               X8       
##  Min.   :3.50   Min.   :2.00   Min.   :0.0000   Min.   :0.000  
##  1st Qu.:3.50   1st Qu.:2.75   1st Qu.:0.1000   1st Qu.:1.750  
##  Median :5.25   Median :3.50   Median :0.2500   Median :3.000  
##  Mean   :5.25   Mean   :3.50   Mean   :0.2344   Mean   :2.812  
##  3rd Qu.:7.00   3rd Qu.:4.25   3rd Qu.:0.4000   3rd Qu.:4.000  
##  Max.   :7.00   Max.   :5.00   Max.   :0.4000   Max.   :5.000  
##        Y1              Y2       
##  Min.   : 6.01   Min.   :10.90  
##  1st Qu.:12.99   1st Qu.:15.62  
##  Median :18.95   Median :22.08  
##  Mean   :22.31   Mean   :24.59  
##  3rd Qu.:31.67   3rd Qu.:33.13  
##  Max.   :43.10   Max.   :48.03  
```

From the summary statistics, we can see that there is no missing values. All columns are numerical.

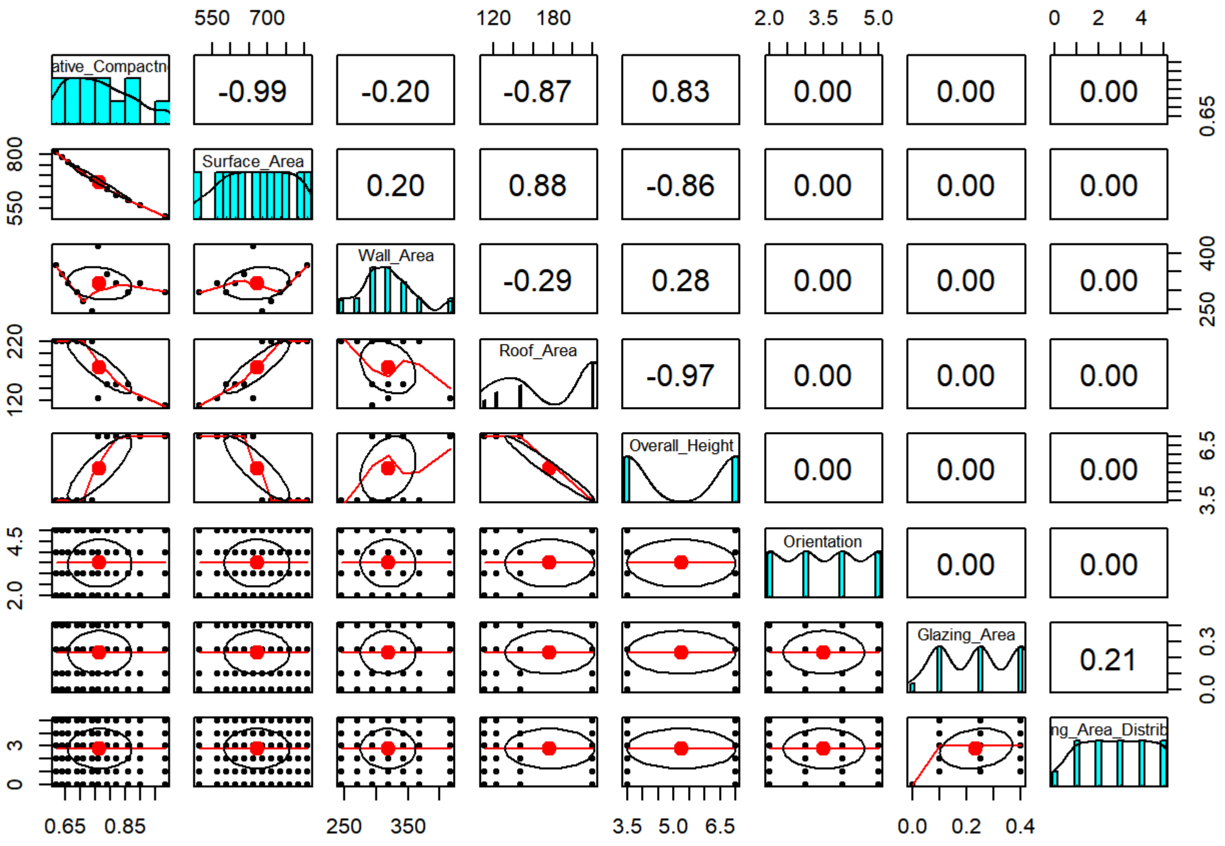
Change the column header names for easier reading

```
#create a vector of the columns
df_column <- c('Relative_Compactness','Surface_Area', 'Wall_Area','Roof_Area','Overall_Height',
'Orientation','Glazing_Area','Glazing_Area_Distribution','Heating_Load','Cooling_Load')

colnames(df) <- df_column

summary(df)
```

```
##  Relative_Compactness  Surface_Area     Wall_Area       Roof_Area
##  Min.   :0.6200       Min.   :514.5   Min.   :245.0   Min.   :110.2
##  1st Qu.:0.6825       1st Qu.:606.4   1st Qu.:294.0   1st Qu.:140.9
##  Median :0.7500       Median :673.8   Median :318.5   Median :183.8
##  Mean   :0.7642       Mean   :671.7   Mean   :318.5   Mean   :176.6
##  3rd Qu.:0.8300       3rd Qu.:741.1   3rd Qu.:343.0   3rd Qu.:220.5
##  Max.   :0.9800       Max.   :808.5   Max.   :416.5   Max.   :220.5
##  Overall_Height  Orientation     Glazing_Area     Glazing_Area_Distribution
##  Min.   :3.50   Min.   :2.00   Min.   :0.0000   Min.   :0.000
##  1st Qu.:3.50   1st Qu.:2.75   1st Qu.:0.1000   1st Qu.:1.750
##  Median :5.25   Median :3.50   Median :0.2500   Median :3.000
##  Mean   :5.25   Mean   :3.50   Mean   :0.2344   Mean   :2.812
##  3rd Qu.:7.00   3rd Qu.:4.25   3rd Qu.:0.4000   3rd Qu.:4.000
##  Max.   :7.00   Max.   :5.00   Max.   :0.4000   Max.   :5.000
##   Heating_Load    Cooling_Load
##  Min.   : 6.01   Min.   :10.90
##  1st Qu.:12.99   1st Qu.:15.62
##  Median :18.95   Median :22.08
##  Mean   :22.31   Mean   :24.59
##  3rd Qu.:31.67   3rd Qu.:33.13
##  Max.   :43.10   Max.   :48.03
```

Pair plot for correlation

```
pairs.panels(df[,-c(9,10)])
```



- The Roof_Area and Overall_Height are highly correlated with correlation of -0.97
- The Relative_Compactness and Surface_Area are higly correlated with correlation of -0.99

Looking at Correlation between two pairs of variables

```
cor.test(df$Overall_Height, df$Roof_Area)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  df$Overall_Height and df$Roof_Area
## t = -115.59, df = 766, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9761011 -0.9683931
## sample estimates:
##        cor 
## -0.9725122
```

At 95% confidence the correlation is between -0.9761011 and -0.9683931. the two variables are strongly negatively correlated

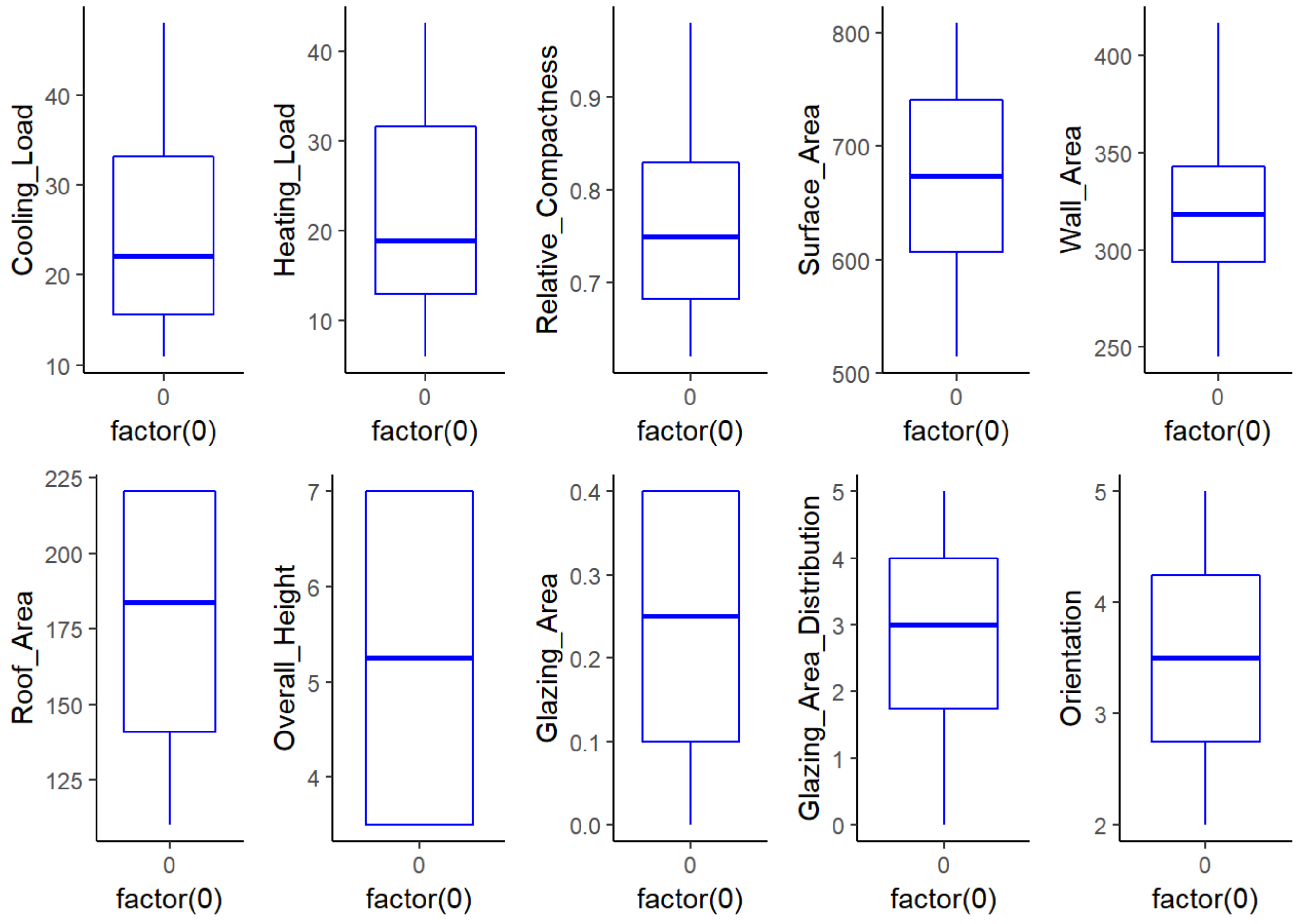
```
cor.test(df$Relative_Compactness, df$Surface_Area)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  df$Relative_Compactness and df$Surface_Area
## t = -216.15, df = 766, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9929678 -0.9906741
## sample estimates:
##        cor 
## -0.9919015
```

Relative_Compactness and Surface_Area are strongly negatively correlated with 95% confidence of between -0.9929678 and -.09906741

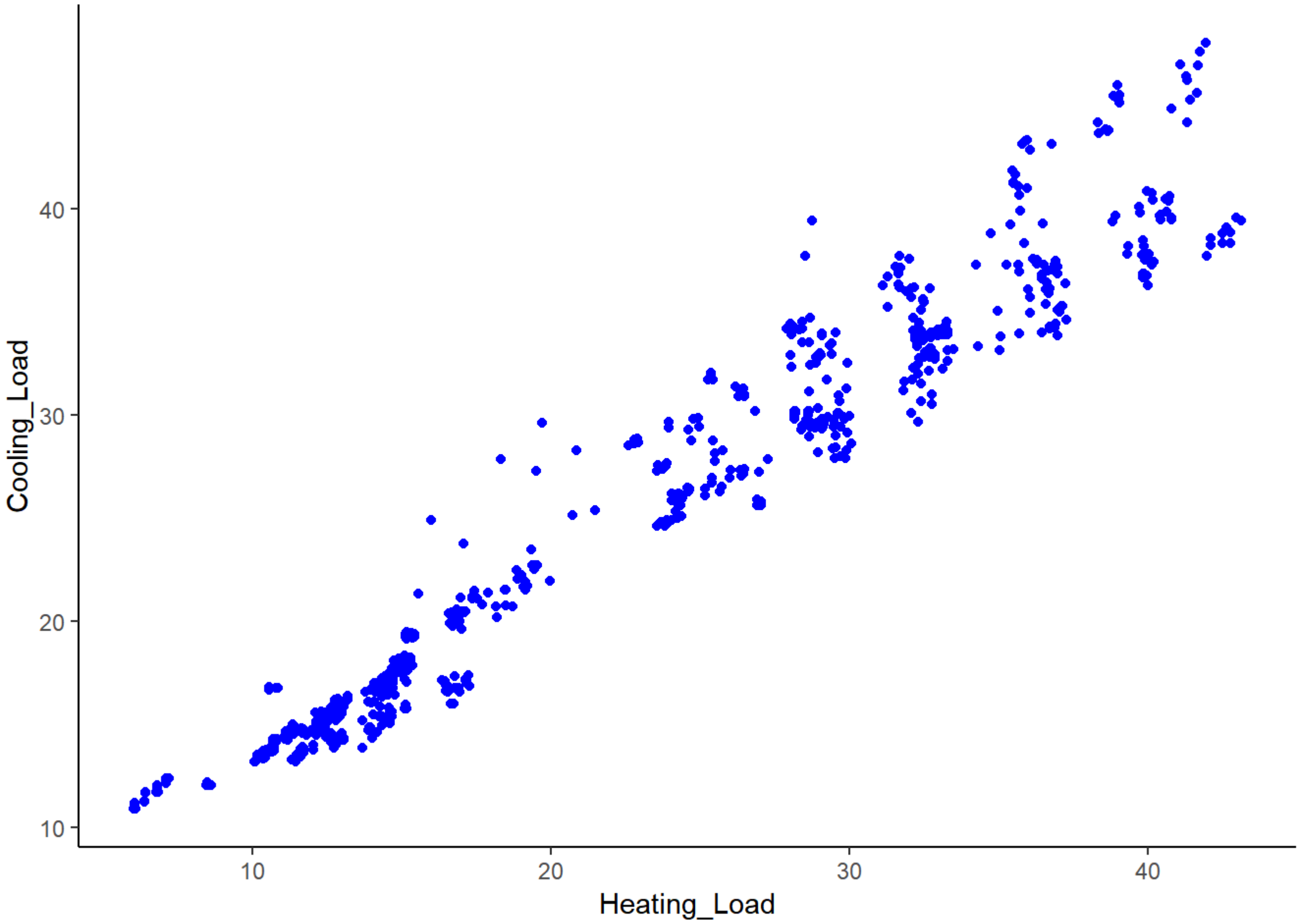
Data distribution

```
g <- ggplot(df, aes(x=factor(0),Cooling_Load)) + geom_boxplot(color = 'blue') + theme_classic()
h <- ggplot(df, aes(x=factor(0),Heating_Load)) + geom_boxplot(color = 'blue') + theme_classic()
i <- ggplot(df, aes(x=factor(0),Relative_Compactness)) + geom_boxplot(color = 'blue') + theme_classic()
j <- ggplot(df, aes(x=factor(0),Surface_Area)) + geom_boxplot(color = 'blue') + theme_classic()
k <- ggplot(df, aes(x=factor(0),Wall_Area)) + geom_boxplot(color = 'blue') + theme_classic()
l <- ggplot(df, aes(x=factor(0),Roof_Area)) + geom_boxplot(color = 'blue') + theme_classic()
m <- ggplot(df, aes(x=factor(0),Overall_Height)) + geom_boxplot(color = 'blue') + theme_classic()
n <- ggplot(df, aes(x=factor(0),Glazing_Area)) + geom_boxplot(color = 'blue') + theme_classic()
o <- ggplot(df, aes(x=factor(0),Glazing_Area_Distribution)) + geom_boxplot(color = 'blue') + theme_classic()
p <- ggplot(df, aes(x=factor(0),Orientation)) + geom_boxplot(color = 'blue') + theme_classic()

plot_grid(g,h,i,j,k,l,m,n,o,p,nrow = 2, ncol = 5)
```



Is there a relationship between the two output variables?

```
ggplot(df, aes(x=Heating_Load, y=Cooling_Load)) + geom_point(color='blue') + theme_classic()
```

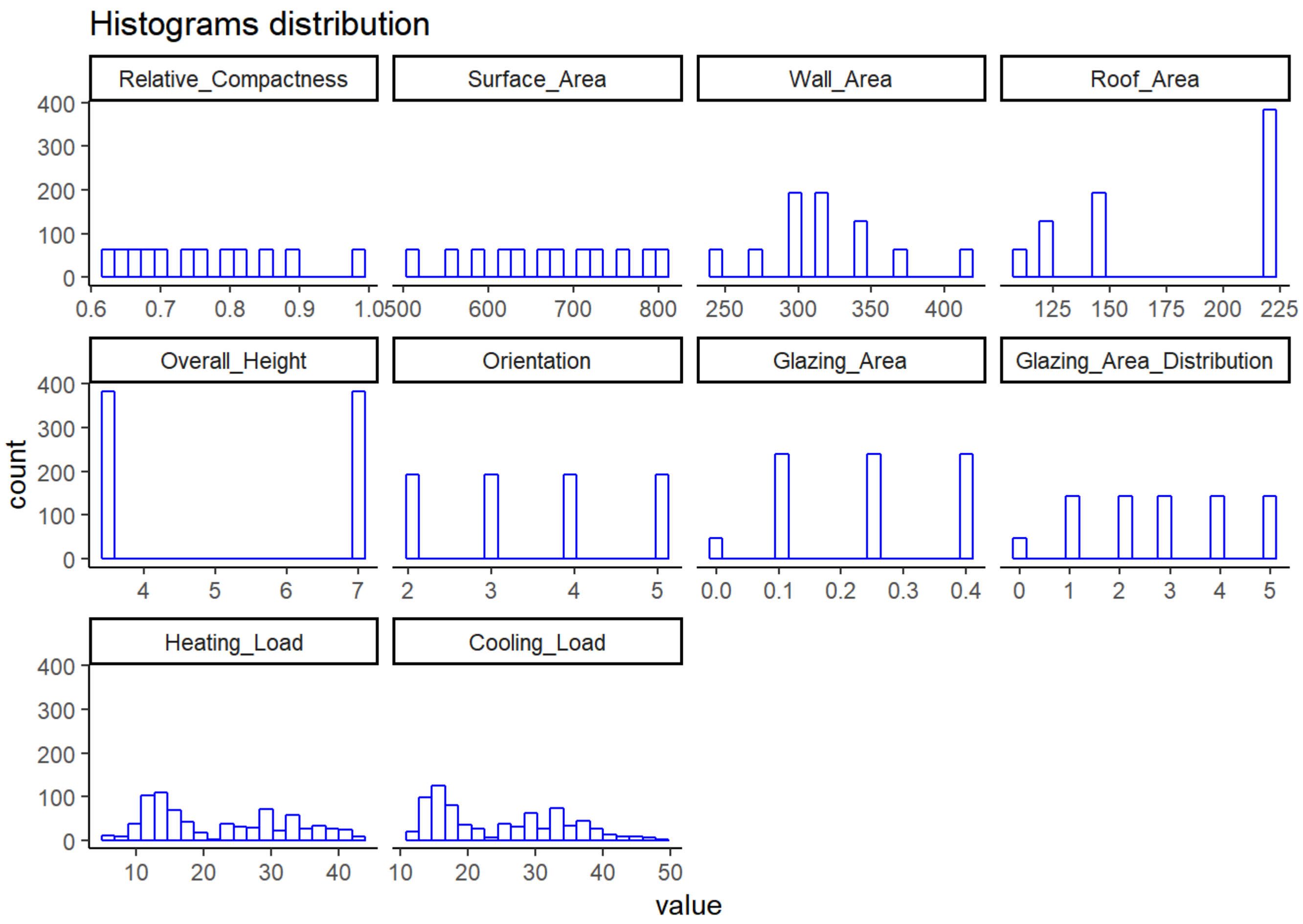


There is clearly a linear relationship between the two output variables.

```
g <- ggplot(data=melt(df), aes(x=value)) + geom_histogram(bins = 20,color='blue', fill ='white')
 + facet_wrap(~variable, scales = 'free_x') + theme_classic()
```

```
## No id variables; using all as measure variables
```

```
g <- g + ggtitle('Histograms distribution')
g
```



Baseline Model using random forest

```
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
## 
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:psych':
## 
##     outlier
```

```
## The following object is masked from 'package:dplyr':
## 
##     combine
## The following object is masked from 'package:ggplot2':
## 
##     margin
```

```
#split datasets input features and target

M1 <- df[,-10]  #heating
m2 <- df[,-9] #cooling
```

Build the model

Heating Load

```
#heating load

rf_heating <- randomForest(Heating_Load ~ ., data = M1, mtry=3, importance = TRUE, na.action = n
a.omit)

rf_heating
```

```
## 
## Call:
##  randomForest(formula = Heating_Load ~ ., data = M1, mtry = 3,      importance = TRUE, na.act
ion = na.omit) 
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
## 
##           Mean of squared residuals: 0.4610571
##                     % Var explained: 99.55
```

Importance of Variables for Heating

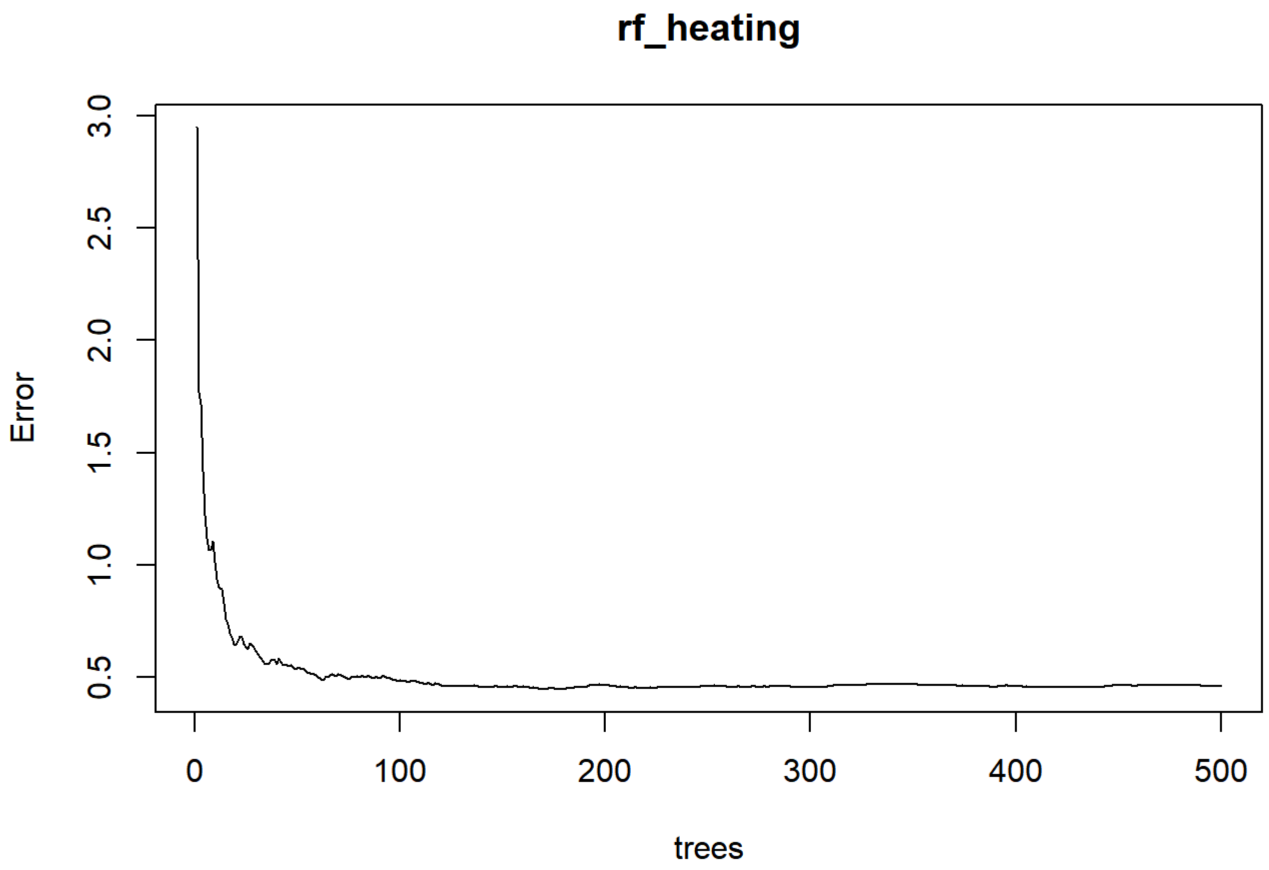
varImpPlot(rf_heating, pch = 20, main = “Importance of Variables”)

\`\`\`

```
round( importance( rf_heating ), 2 )
```

```
##                           %IncMSE IncNodePurity
## Relative_Compactness        17.90      22264.00
## Surface_Area                14.40      16183.88
## Wall_Area                   17.24       3644.49
## Roof_Area                   11.87      13887.40
## Overall_Height              12.46      15031.04
## Orientation                -16.71         56.70
## Glazing_Area                76.82       4385.87
## Glazing_Area_Distribution   34.82       1799.93
```

```
plot(rf_heating)
```

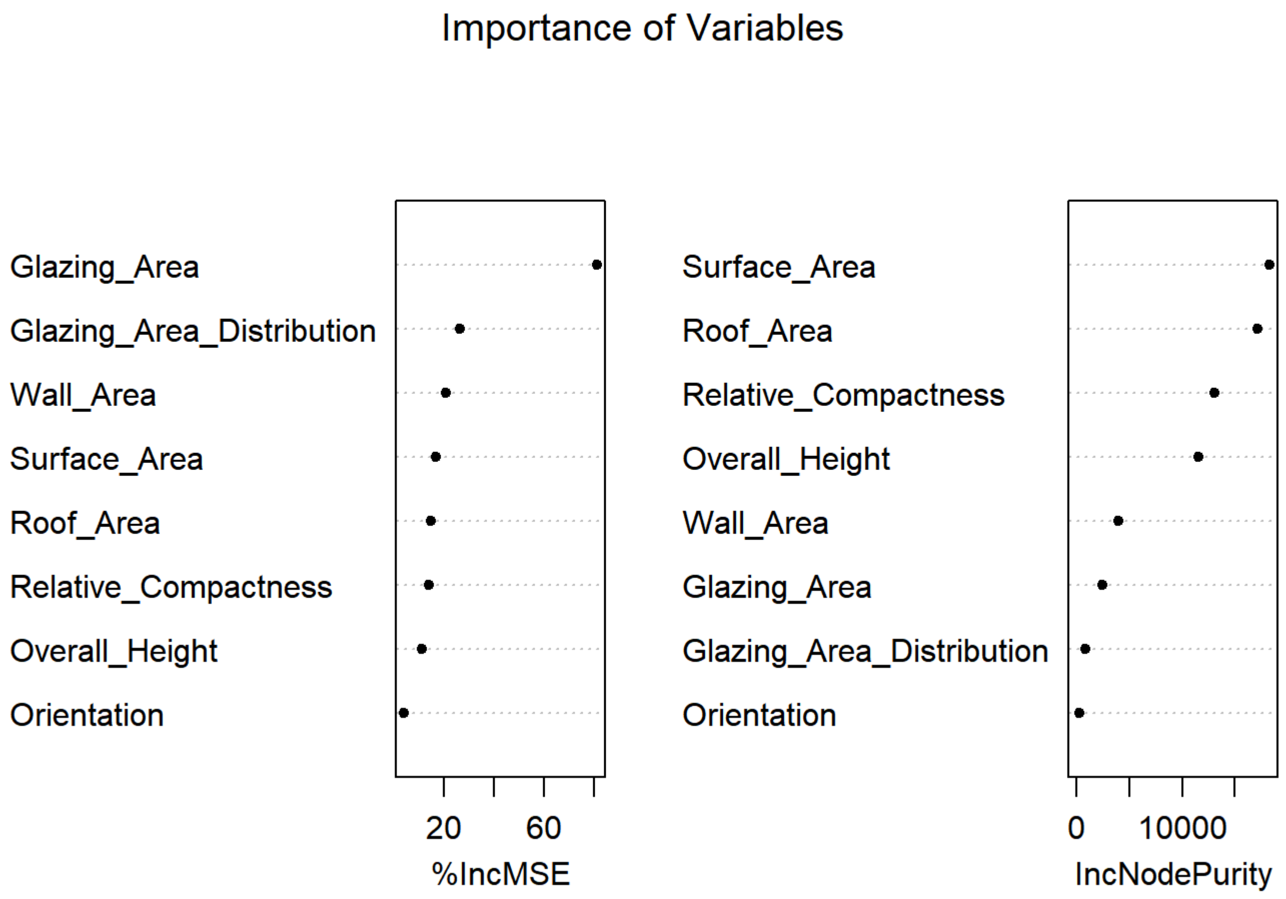


cooling

```
rf_cooling <- randomForest(Cooling_Load ~ ., data = m2, mtry=3, importance = TRUE, na.action = n
a.omit)
rf_cooling
```

```
##
## Call:
##  randomForest(formula = Cooling_Load ~ ., data = m2, mtry = 3,      importance = TRUE, na.act
ion = na.omit)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 3.007613
##                     % Var explained: 96.67
```

```
### Importance of Variables for Cooling

varImpPlot(rf_cooling, pch = 20, main = "Importance of Variables")
```



```
round( importance( rf_cooling ) ,2 )
```

```
##                           %IncMSE IncNodePurity
## Relative_Compactness        13.90      13031.66
## Surface_Area                16.71      18296.55
## Wall_Area                   20.76       3959.98
## Roof_Area                   14.89      17117.23
## Overall_Height              11.22      11589.65
## Orientation                  3.97        253.86
## Glazing_Area                81.55       2434.99
## Glazing_Area_Distribution   26.66        842.54
```

Build a baseline model using Linear Regression

heating

```
lm_heating <- lm(Heating_Load ~.,data = M1)

summary(lm_heating)
```

```
##
## Call:
## lm(formula = Heating_Load ~ ., data = M1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8965 -1.3196 -0.0252  1.3532  7.7052
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                84.013418  19.033613   4.414 1.16e-05 ***
## Relative_Compactness      -64.773432  10.289448  -6.295 5.19e-10 ***
## Surface_Area               -0.087289   0.017075  -5.112 4.04e-07 ***
## Wall_Area                   0.060813   0.006648   9.148  < 2e-16 ***
## Roof_Area                         NA         NA      NA       NA
## Overall_Height              4.169954   0.337990  12.338  < 2e-16 ***
## Orientation                -0.023330   0.094705  -0.246  0.80548
## Glazing_Area               19.932736   0.813986  24.488  < 2e-16 ***
## Glazing_Area_Distribution   0.203777   0.069918   2.915  0.00367 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.934 on 760 degrees of freedom
## Multiple R-squared:  0.9162, Adjusted R-squared:  0.9154
## F-statistic:  1187 on 7 and 760 DF,  p-value: < 2.2e-16
```

```
anova(lm_heating)
```

```
## Analysis of Variance Table
##
## Response: Heating_Load
##                            Df  Sum Sq Mean Sq   F value    Pr(>F)
## Relative_Compactness        1 30238.2 30238.2 3511.8926 < 2.2e-16 ***
## Surface_Area                1  8092.8  8092.8  939.9113 < 2.2e-16 ***
## Wall_Area                   1 26144.8 26144.8 3036.4862 < 2.2e-16 ***
## Overall_Height              1  1310.6  1310.6  152.2140 < 2.2e-16 ***
## Orientation                 1     0.5     0.5    0.0607  0.805480
## Glazing_Area                1  5686.1  5686.1  660.3875 < 2.2e-16 ***
## Glazing_Area_Distribution   1    73.1    73.1    8.4944  0.003667 **
## Residuals                 760  6543.8     8.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cooling

```
lm_cooling <- lm(Cooling_Load ~ ., data = m2)
summary(lm_cooling)
```

```
##
## Call:
## lm(formula = Cooling_Load ~ ., data = m2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6940 -1.5606 -0.2668  1.3968 11.1775
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               97.245749  20.764711   4.683 3.34e-06 ***
## Relative_Compactness     -70.787707  11.225269  -6.306 4.85e-10 ***
## Surface_Area              -0.088245   0.018628  -4.737 2.59e-06 ***
## Wall_Area                  0.044682   0.007253   6.161 1.17e-09 ***
## Roof_Area                        NA         NA      NA       NA
## Overall_Height             4.283843   0.368730  11.618  < 2e-16 ***
## Orientation                0.121510   0.103318   1.176    0.240
## Glazing_Area              14.717068   0.888018  16.573  < 2e-16 ***
## Glazing_Area_Distribution   0.040697   0.076277   0.534    0.594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.201 on 760 degrees of freedom
## Multiple R-squared:  0.8878, Adjusted R-squared:  0.8868
## F-statistic: 859.1 on 7 and 760 DF,  p-value: < 2.2e-16
```

```
anova(lm_cooling)
```

```
## Analysis of Variance Table
##
## Response: Cooling_Load
##                            Df  Sum Sq Mean Sq   F value Pr(>F)
## Relative_Compactness        1 27931.9 27931.9 2725.6959 <2e-16 ***
## Surface_Area                1  8254.2  8254.2  805.4720 <2e-16 ***
## Wall_Area                   1 21052.3 21052.3 2054.3534 <2e-16 ***
## Overall_Height              1  1383.2  1383.2  134.9739 <2e-16 ***
## Orientation                 1    14.2    14.2    1.3832 0.2399
## Glazing_Area                1  2988.9  2988.9  291.6699 <2e-16 ***
## Glazing_Area_Distribution   1     2.9     2.9    0.2847 0.5938
## Residuals                 760  7788.2    10.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Training for a Linear Model

```
Mh <- df[,-10]  # heating
Mc <- df[,-9] # cooling
```

Linear Model

Heating

```
lm_heating2 <- lm(Heating_Load ~., data = Mh)

summary(lm_heating2)
```

```
##
## Call:
## lm(formula = Heating_Load ~ ., data = Mh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8965 -1.3196 -0.0252  1.3532  7.7052
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               84.013418  19.033613   4.414 1.16e-05 ***
## Relative_Compactness     -64.773432  10.289448  -6.295 5.19e-10 ***
## Surface_Area              -0.087289   0.017075  -5.112 4.04e-07 ***
## Wall_Area                  0.060813   0.006648   9.148  < 2e-16 ***
## Roof_Area                        NA         NA      NA       NA
## Overall_Height             4.169954   0.337990  12.338  < 2e-16 ***
## Orientation               -0.023330   0.094705  -0.246  0.80548
## Glazing_Area              19.932736   0.813986  24.488  < 2e-16 ***
## Glazing_Area_Distribution   0.203777   0.069918   2.915  0.00367 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.934 on 760 degrees of freedom
## Multiple R-squared:  0.9162, Adjusted R-squared:  0.9154
## F-statistic:  1187 on 7 and 760 DF,  p-value: < 2.2e-16
```

```
anova(lm_heating2)
```

```
## Analysis of Variance Table
##
## Response: Heating_Load
##                            Df  Sum Sq Mean Sq   F value    Pr(>F)
## Relative_Compactness        1 30238.2 30238.2 3511.8926 < 2.2e-16 ***
## Surface_Area                1  8092.8  8092.8  939.9113 < 2.2e-16 ***
## Wall_Area                   1 26144.8 26144.8 3036.4862 < 2.2e-16 ***
## Overall_Height              1  1310.6  1310.6  152.2140 < 2.2e-16 ***
## Orientation                 1     0.5     0.5    0.0607  0.805480
## Glazing_Area                1  5686.1  5686.1  660.3875 < 2.2e-16 ***
## Glazing_Area_Distribution   1    73.1    73.1    8.4944  0.003667 **
## Residuals                 760  6543.8     8.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cooling

```
lm_cooling2 <- lm(Cooling_Load ~.,data = Mc)
summary(lm_cooling2)
```

```
##
## Call:
## lm(formula = Cooling_Load ~ ., data = Mc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6940 -1.5606 -0.2668  1.3968 11.1775
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                97.245749  20.764711   4.683 3.34e-06 ***
## Relative_Compactness      -70.787707  11.225269  -6.306 4.85e-10 ***
## Surface_Area               -0.088245   0.018628  -4.737 2.59e-06 ***
## Wall_Area                   0.044682   0.007253   6.161 1.17e-09 ***
## Roof_Area                         NA         NA      NA       NA
## Overall_Height              4.283843   0.368730  11.618  < 2e-16 ***
## Orientation                 0.121510   0.103318   1.176    0.240
## Glazing_Area               14.717068   0.888018  16.573  < 2e-16 ***
## Glazing_Area_Distribution   0.040697   0.076277   0.534    0.594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.201 on 760 degrees of freedom
## Multiple R-squared:  0.8878, Adjusted R-squared:  0.8868
## F-statistic: 859.1 on 7 and 760 DF,  p-value: < 2.2e-16
```

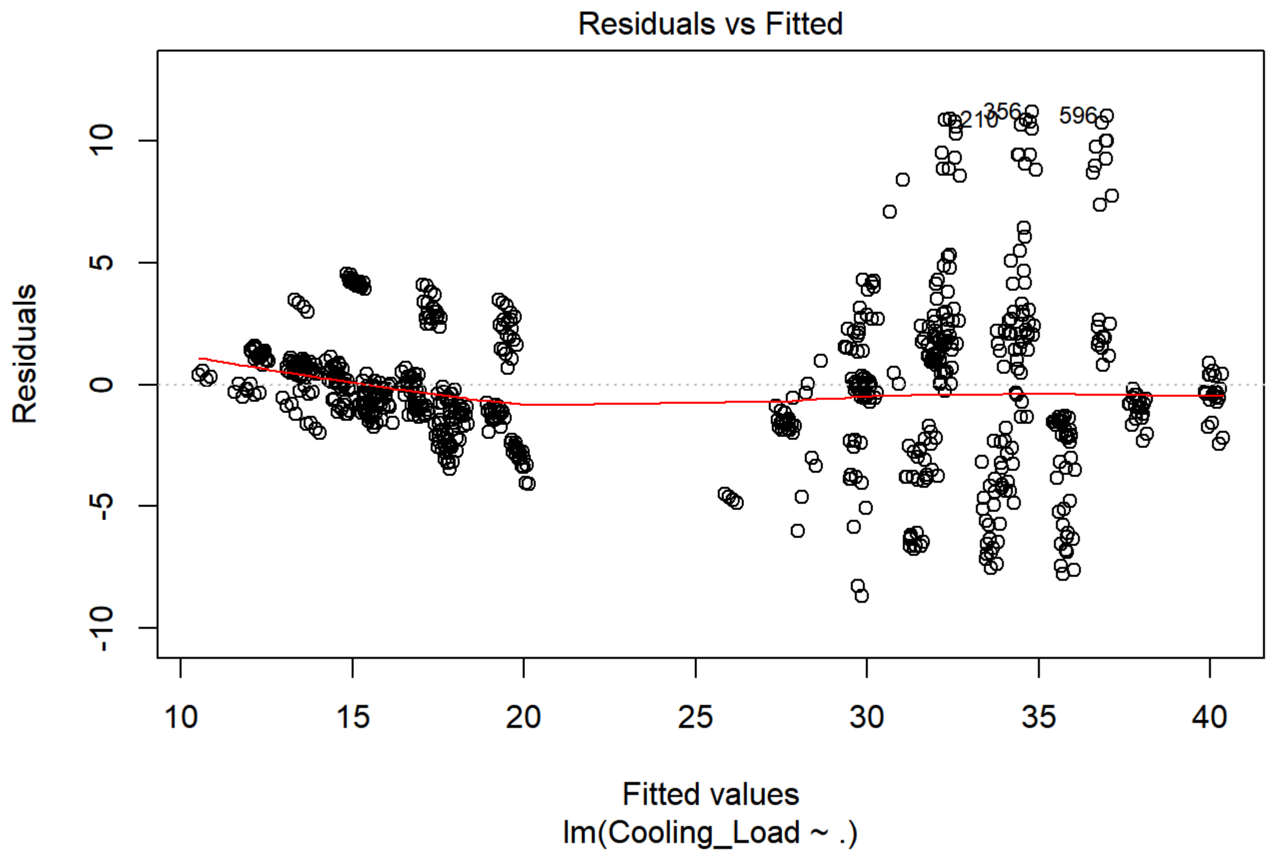
```
anova(lm_cooling2)
```

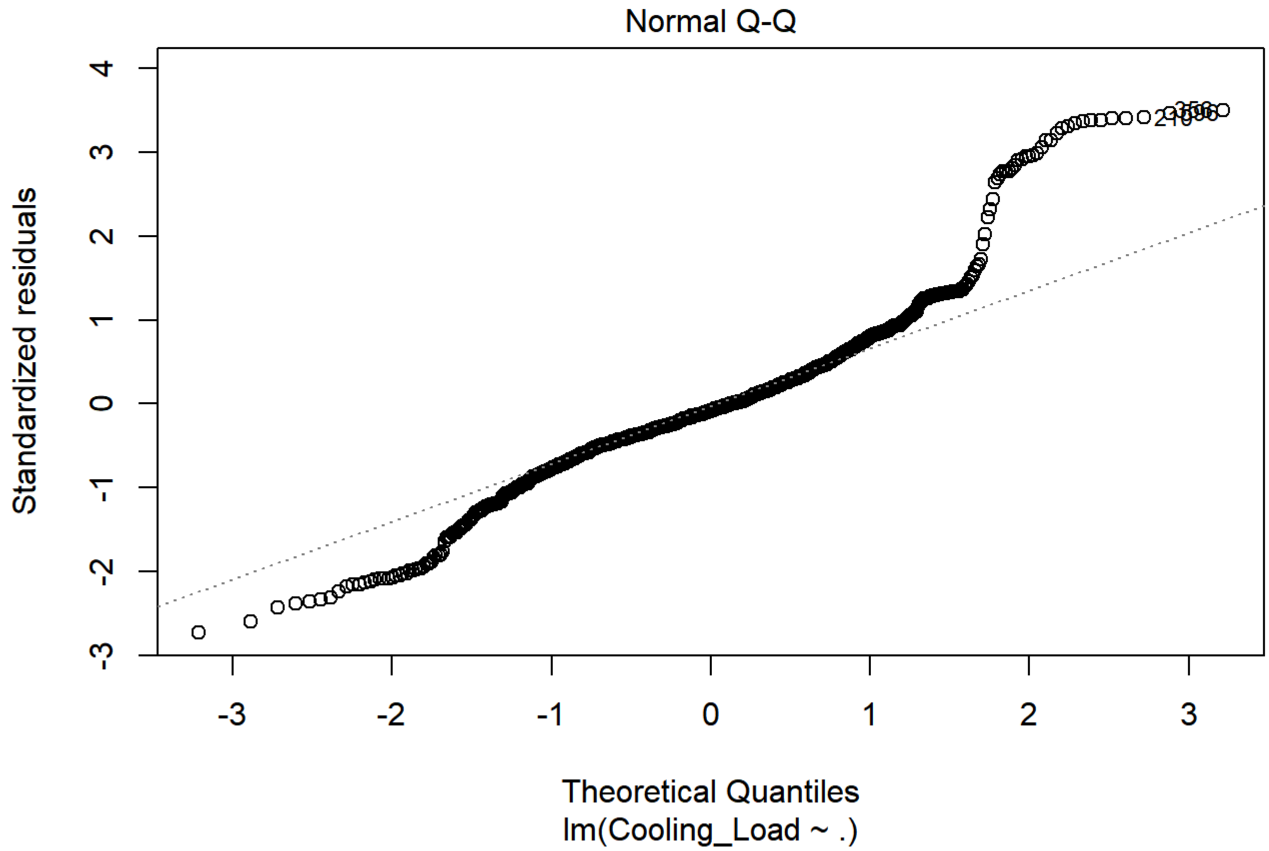
```
## Analysis of Variance Table
##
## Response: Cooling_Load
##                            Df  Sum Sq Mean Sq   F value Pr(>F)
## Relative_Compactness        1 27931.9 27931.9 2725.6959 <2e-16 ***
## Surface_Area                1  8254.2  8254.2  805.4720 <2e-16 ***
## Wall_Area                   1 21052.3 21052.3 2054.3534 <2e-16 ***
## Overall_Height              1  1383.2  1383.2  134.9739 <2e-16 ***
## Orientation                 1    14.2    14.2    1.3832 0.2399
## Glazing_Area                1  2988.9  2988.9  291.6699 <2e-16 ***
## Glazing_Area_Distribution   1     2.9     2.9    0.2847 0.5938
## Residuals                 760  7788.2    10.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

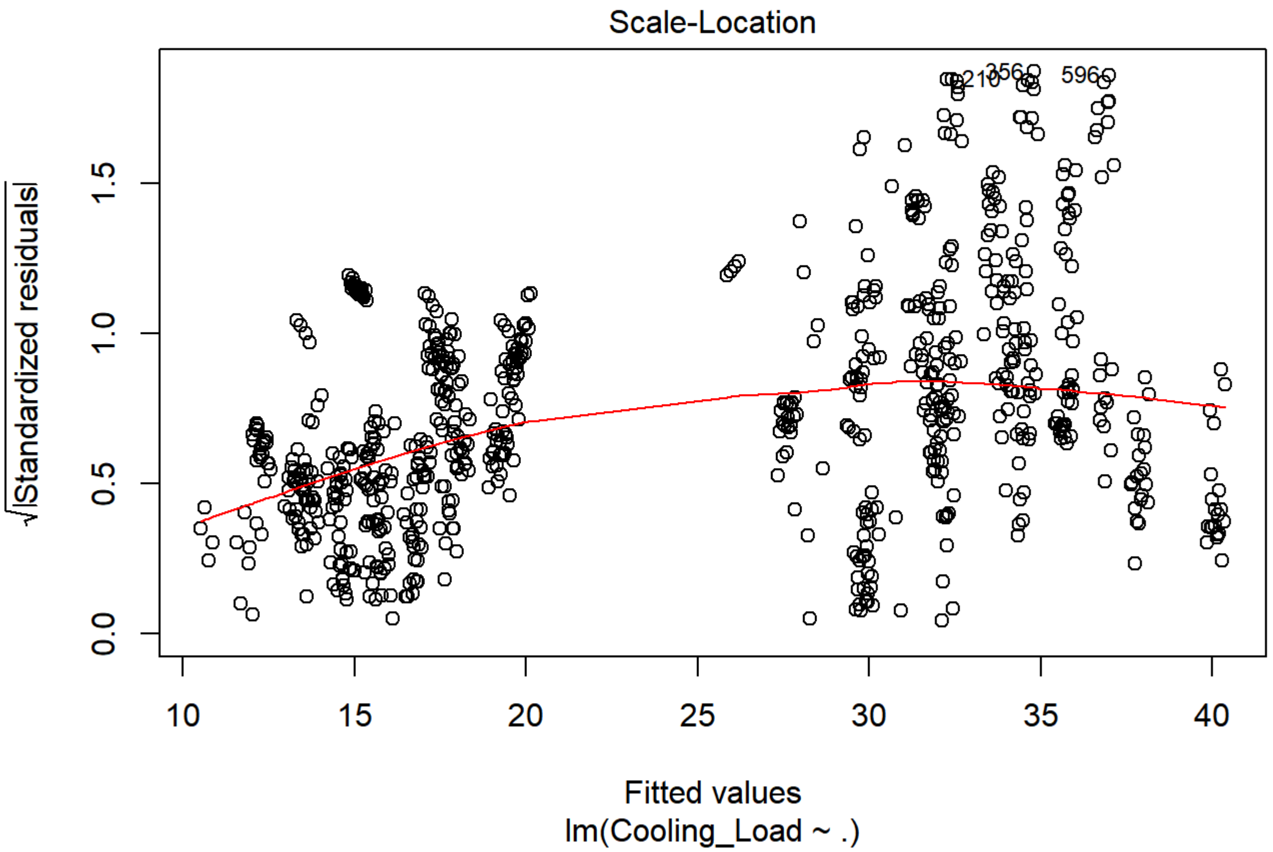
```
round(rf_heating$importance,2)
```

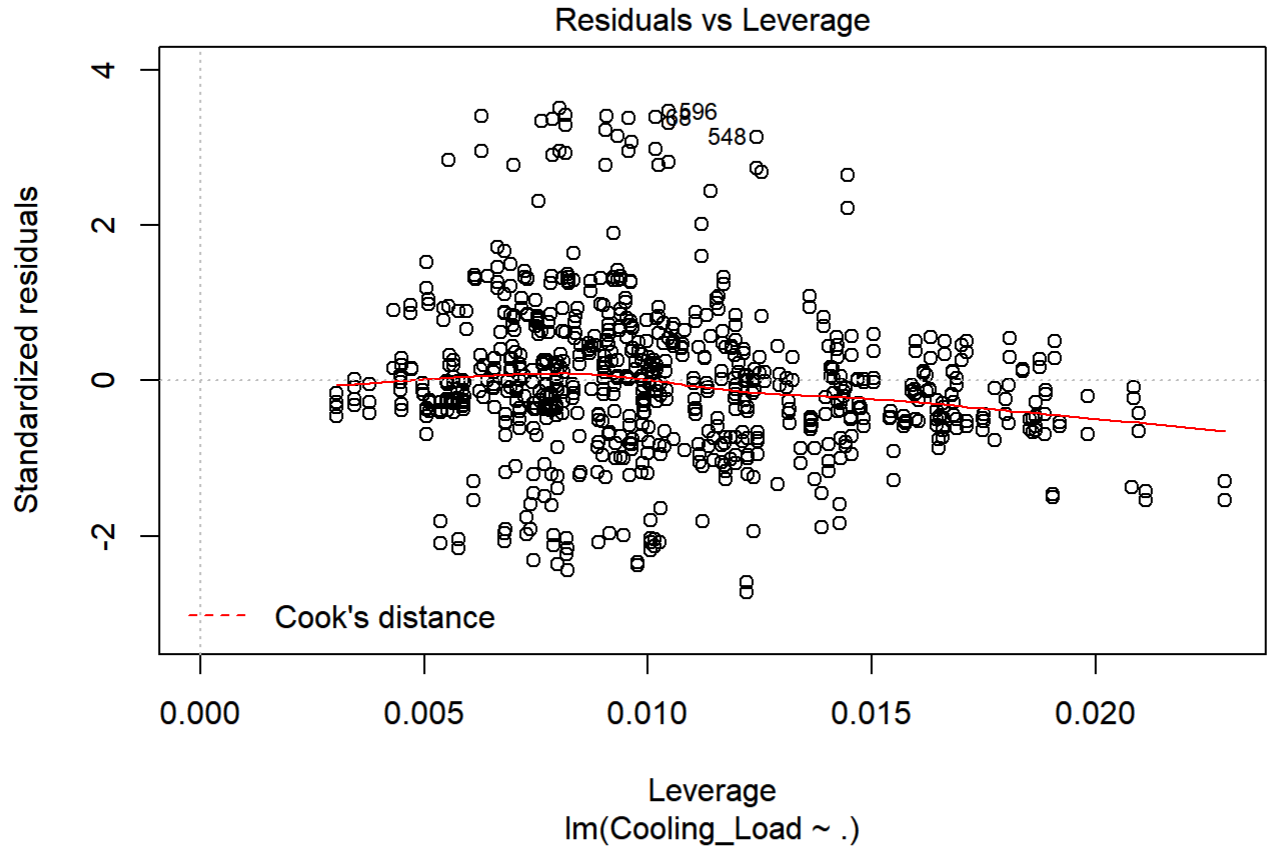
```
##                            %IncMSE IncNodePurity
## Relative_Compactness        66.87      22264.00
## Surface_Area                51.40      16183.88
## Wall_Area                   12.96       3644.49
## Roof_Area                   46.79      13887.40
## Overall_Height              51.50      15031.04
## Orientation                 -0.17         56.70
## Glazing_Area                11.35       4385.87
## Glazing_Area_Distribution    3.63       1799.93
```

```
plot(lm_cooling2)
```


Residuals vs Fitted
Residuals
10
5
0
-5
-10
210
356
596
10
15
20
25
30
35
40
Fitted values
lm(Cooling_Load ~ .)


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
-3
356
210
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Cooling_Load ~ .)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
30
35
40
210
356
596
Fitted values
lm(Cooling_Load ~ .)


Residuals vs Leverage
Standardized residuals
-2
0
2
4
0.000
0.005
0.010
0.015
0.020
68
596
548
Cook's distance
Leverage
lm(Cooling_Load ~ .)