UW 350: HW1

Anish Mohan

January 15, 2017

1. Summary and/or Key Conclusions

- Overall Height of the building has a significant impact on the Heating and Cooling load. Heating/Cooling load is lower when the Overall Height is smaller than 4 units. As the Overall Height increases above 7 units the average Heating/Cooling Load increases significantly.
- Roof Area of the building impacts the Heating and the Cooling Load. Heating/Cooling load seems to be higher when the Roof Areas is below 150 Units. When the Roof Area is above 200 units, the average Heating load is lower
- Surface Area of the building impacts the Heating and Cooling Load. Heating/Cooling loads seems to be higher when the Surface Area is smaller than 675 units. As the Surface Area increases beyond 675 Units, the Average Heating/Cooling load is lower.
- Even though the correlation between Heating/Cooling load and Glazing area is small; it was noted that as value of Glazing Area increases the Heating/Cooling load increases.
- Heating and Cooling Load distribution has the same distribution for all Orientations.

2. Reviewing Data

Source Data:- A. Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, 2012

```
  # Read the dataset
  energy = read.csv("EnergyEfficiencyData.csv", header = TRUE, stringsAsFactors = FALSE)
```

```
  # Review the names of the columns or the features
  names(energy)
```

```
##  [1] "Relative.Compactness"      "Surface.Area"
##  [3] "Wall.Area"                 "Roof.Area"
##  [5] "Overall.Height"            "Orientation"
##  [7] "Glazing.Area"              "Glazing.Area.Distribution"
##  [9] "Heating.Load"              "Cooling.Load"
```

```
  # Print a summary of the datatypes in the data.
  str(energy)
```

```
## 'data.frame':    768 obs. of  10 variables:
##  $ Relative.Compactness     : num  0.98 0.98 0.98 0.98 0.9 0.9 0.9 0.9 0.86 0.86 ...
##  $ Surface.Area             : num  514 514 514 514 564 ...
##  $ Wall.Area                : num  294 294 294 294 318 ...
##  $ Roof.Area                : num  110 110 110 110 122 ...
##  $ Overall.Height           : num  7 7 7 7 7 7 7 7 7 7 ...
##  $ Orientation              : int  2 3 4 5 2 3 4 5 2 3 ...
```

```
##  $ Glazing.Area             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Glazing.Area.Distribution: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Heating.Load             : num  15.6 15.6 15.6 15.6 20.8 ...
##  $ Cooling.Load             : num  21.3 21.3 21.3 21.3 28.3 ...
```

The data here shows that there are 768 rows (sample) and 10 columns (features).

```
# Quick check to see if there are any NA's in the data
summary(is.na(energy))
```

```
##  Relative.Compactness Surface.Area    Wall.Area       Roof.Area
##  Mode :logical        Mode :logical   Mode :logical   Mode :logical
##  FALSE:768            FALSE:768       FALSE:768       FALSE:768
##  NA's :0              NA's :0         NA's :0         NA's :0
##  Overall.Height  Orientation     Glazing.Area    Glazing.Area.Distribution
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:768       FALSE:768       FALSE:768       FALSE:768
##  NA's :0         NA's :0         NA's :0         NA's :0
##  Heating.Load    Cooling.Load
##  Mode :logical   Mode :logical
##  FALSE:768       FALSE:768
##  NA's :0         NA's :0
```

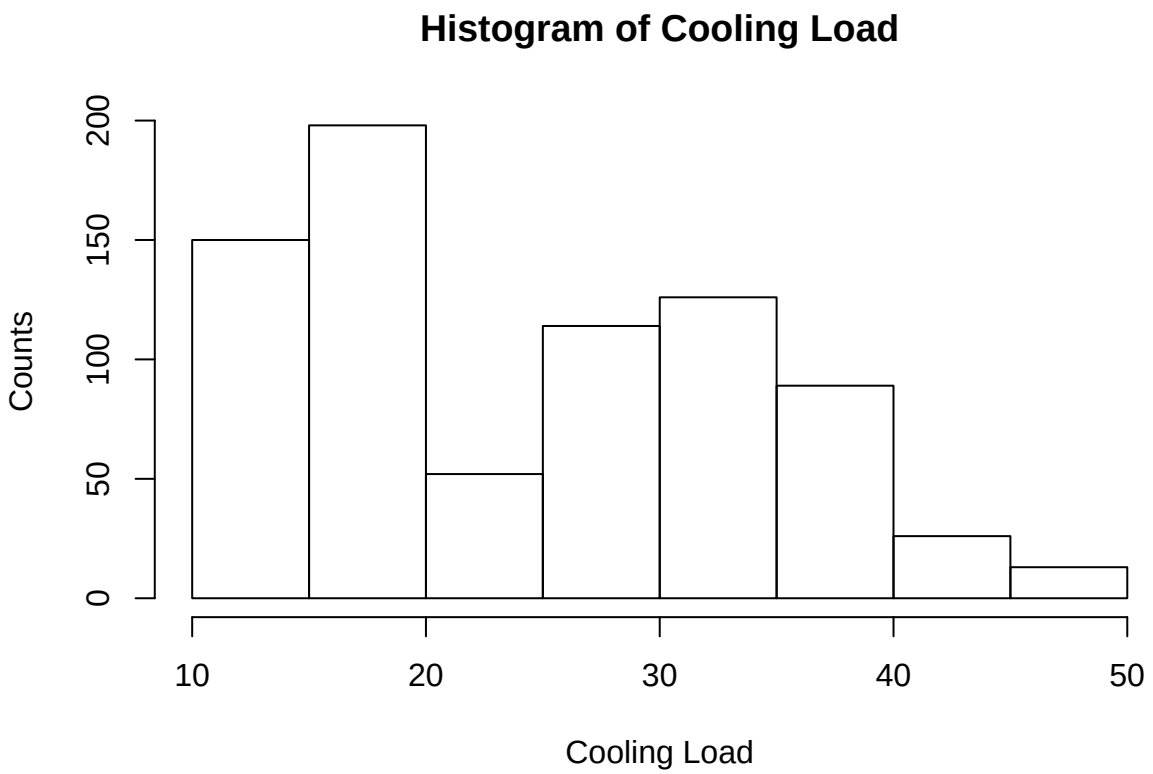
Per this chart, none of the columns have any NAs.

3. Exploratory Data Analysis

As a next step, look at each of the feature columns and review the distribution of the data

Histogram of Cooling Load

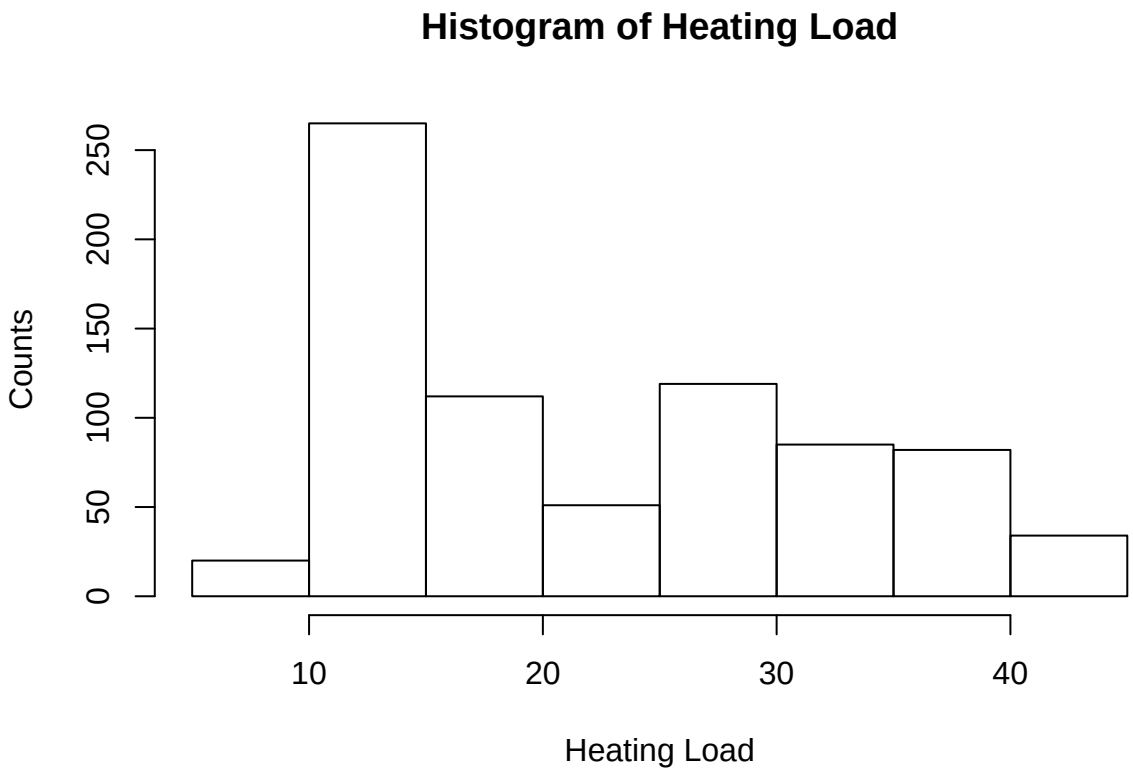
```
hist(energy$Cooling.Load, main = paste("Histogram of Cooling Load"), xlab = "Cooling Load", ylab = "C
```



The Cooling Load is varying from 10-50 units with a distribution that is unimodal and skewed towards lower values

Histogram of Heating Load

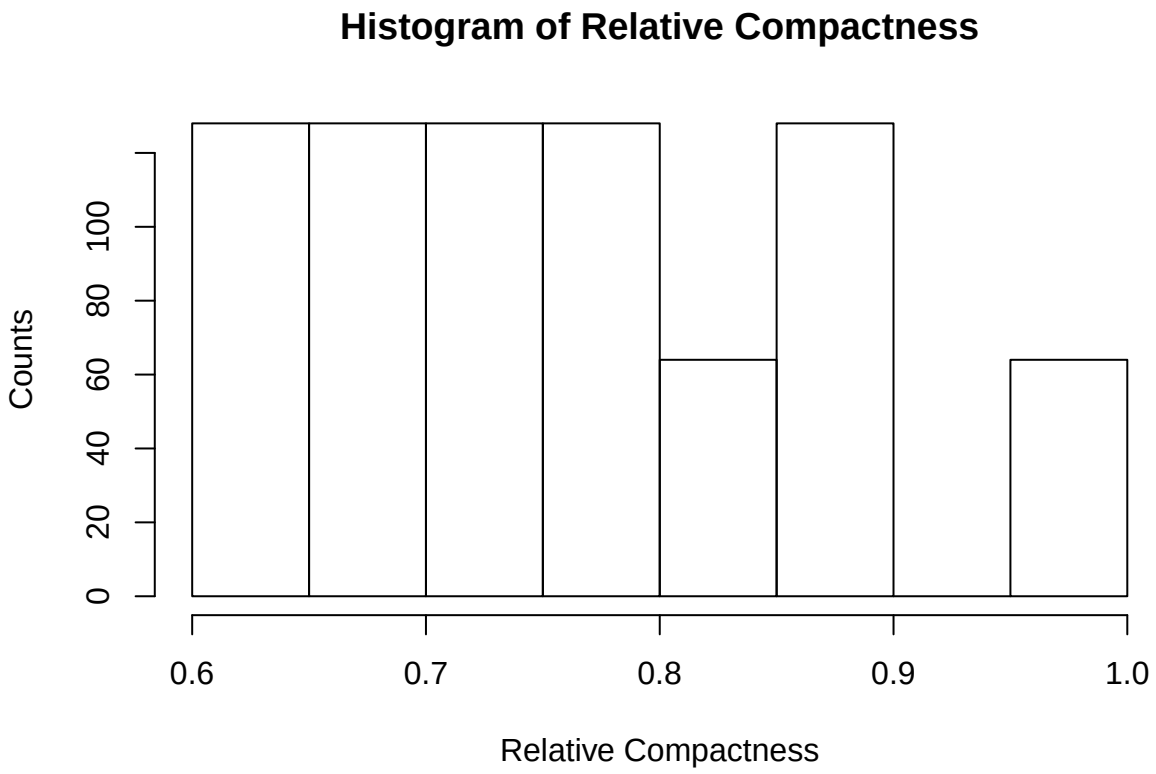
```
hist(energy$Heating.Load, main = paste("Histogram of Heating Load"), xlab = "Heating Load", ylab = "C
```



The Heating load varies from 0 to 50 units as well. Maximum counts for heating load is in the range [10-15] units with 250+ counts.

Histogram of Relative Compactness

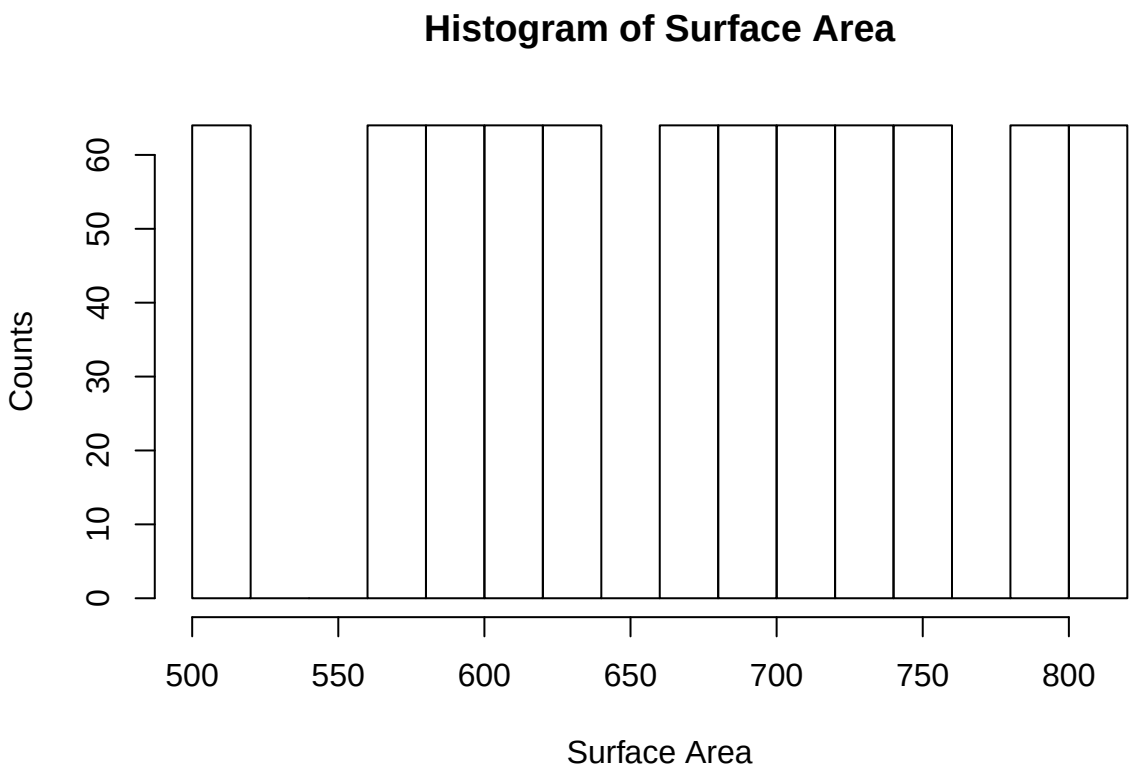
```
hist(energy$Relative.Compactness, main = paste("Histogram of Relative Compactness"), xlab = "Relative
```



Relative Compactness measure varies from 0.6 units to 1.0 units. Number of counts in each bin of the Relative Compactness measure is fairly similar i.e Most of the bins have about 125 counts and 2 remainng bins have a count of 60

Histogram of Surface Area

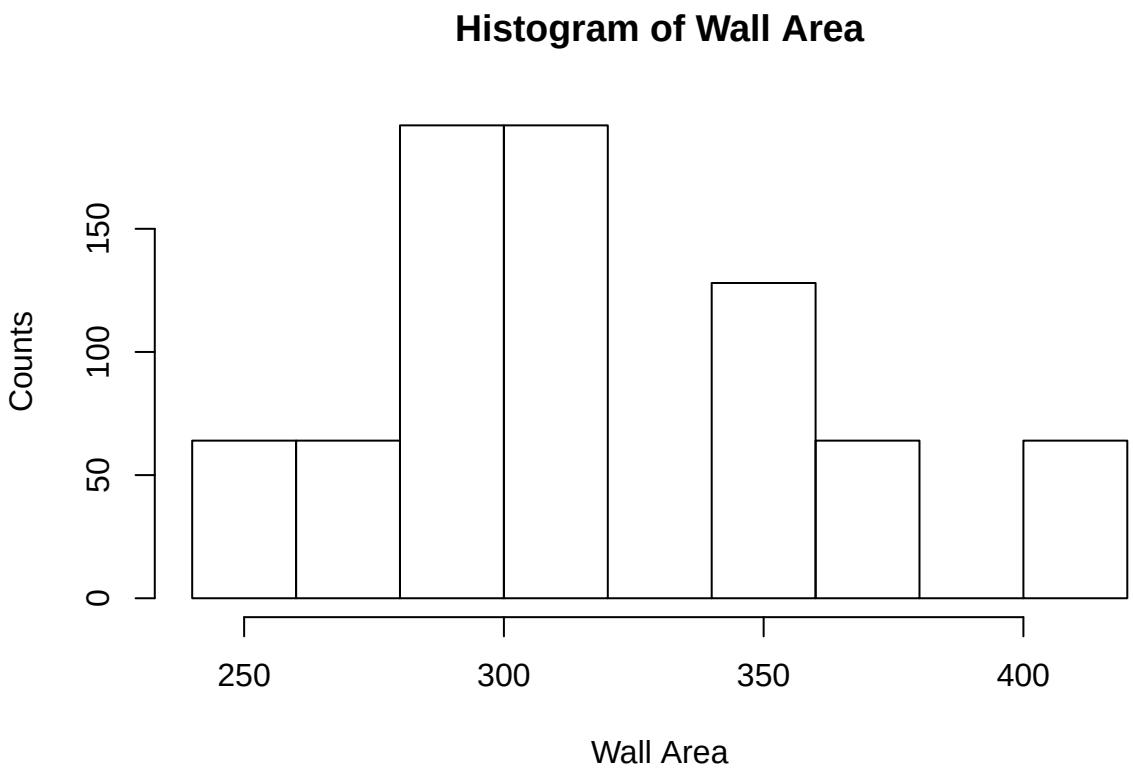
```
hist(energy$Surface.Area, main = paste("Histogram of Surface Area"), xlab = "Surface Area", ylab = "C
```



Surface Area measures from 500 to 820 Units. The counts for most of the bins are equal with each bin having about 70 entries

Histogram of Wall Area

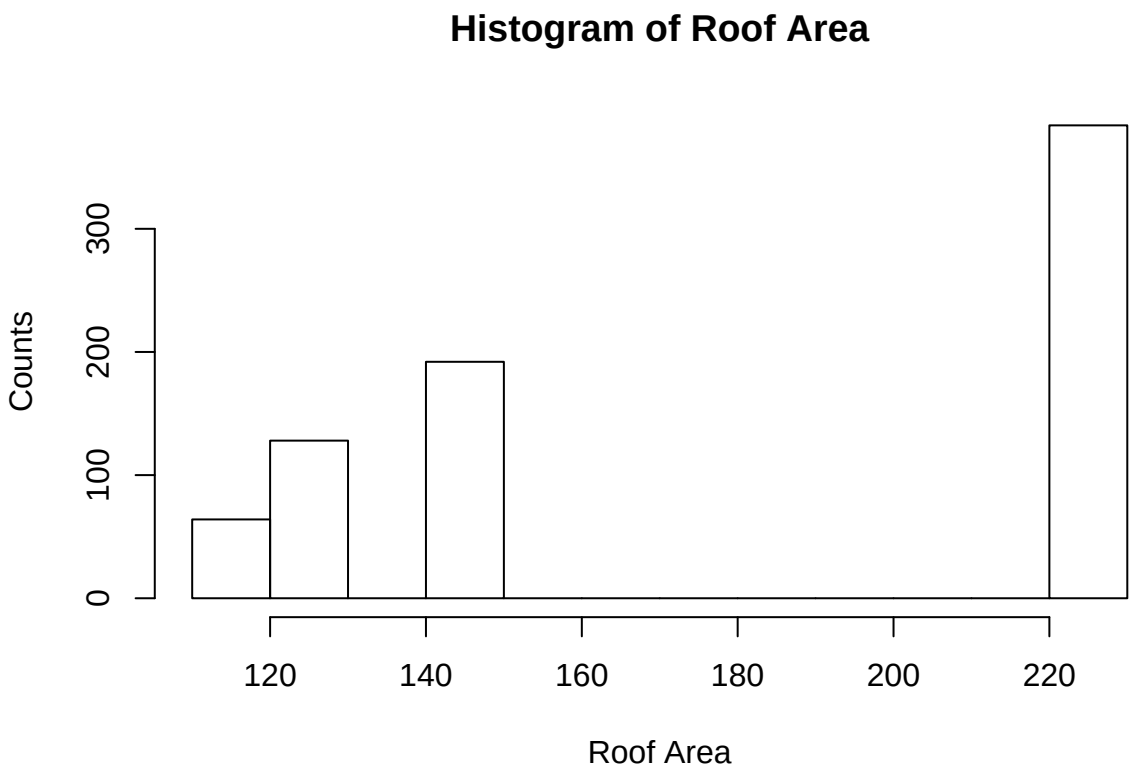
```
hist(energy$Wall.Area, main = paste("Histogram of Wall Area"), xlab = "Wall Area", ylab = "Counts")
```



Wall Area varies from ~230 units to 420 Units. It is a unimodal distribution with peak of about 150+ counts around 280-320 units of wall-area

Histogram of Roof Area

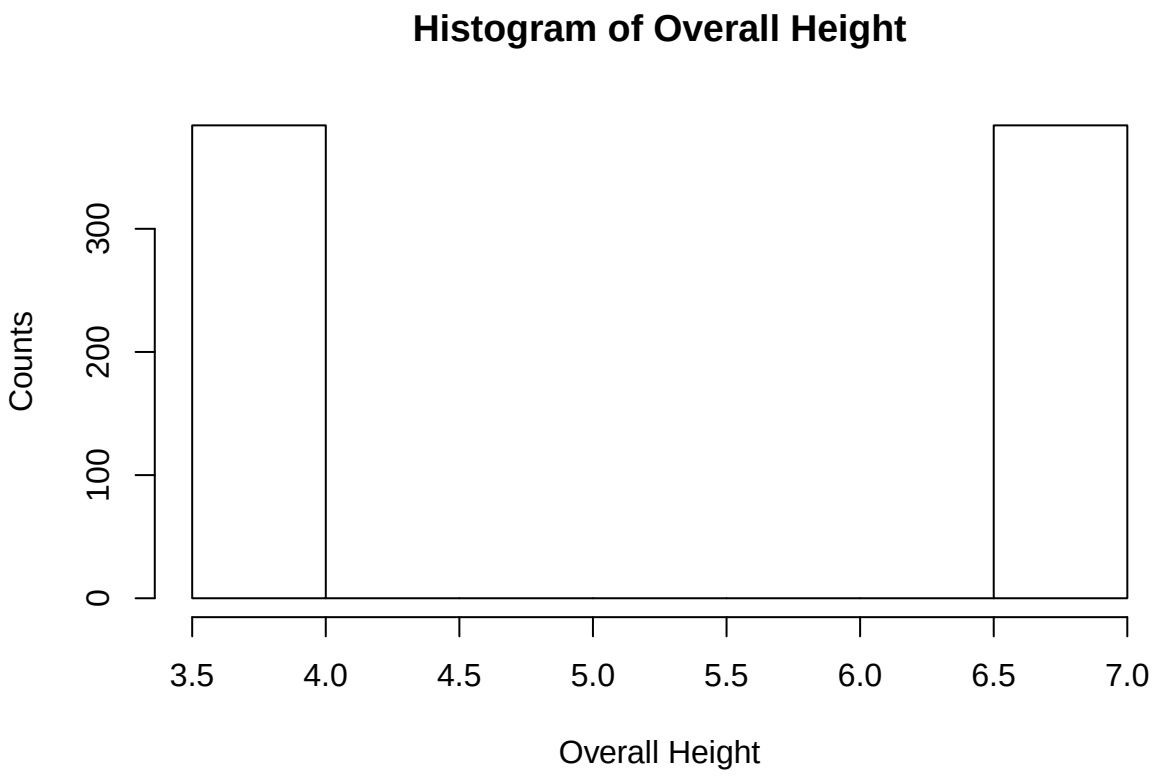
```
hist(energy$Roof.Area, main = paste("Histogram of Roof Area"), xlab = "Roof Area", ylab = "Counts")
```



Roof Area has a skewed distribution with most of the houses having Roof Area above 220 Units. Some of the houses have roof area below 150 units, but there are no units with Roof Area between 150-220 units.

Histogram of Overeall Height

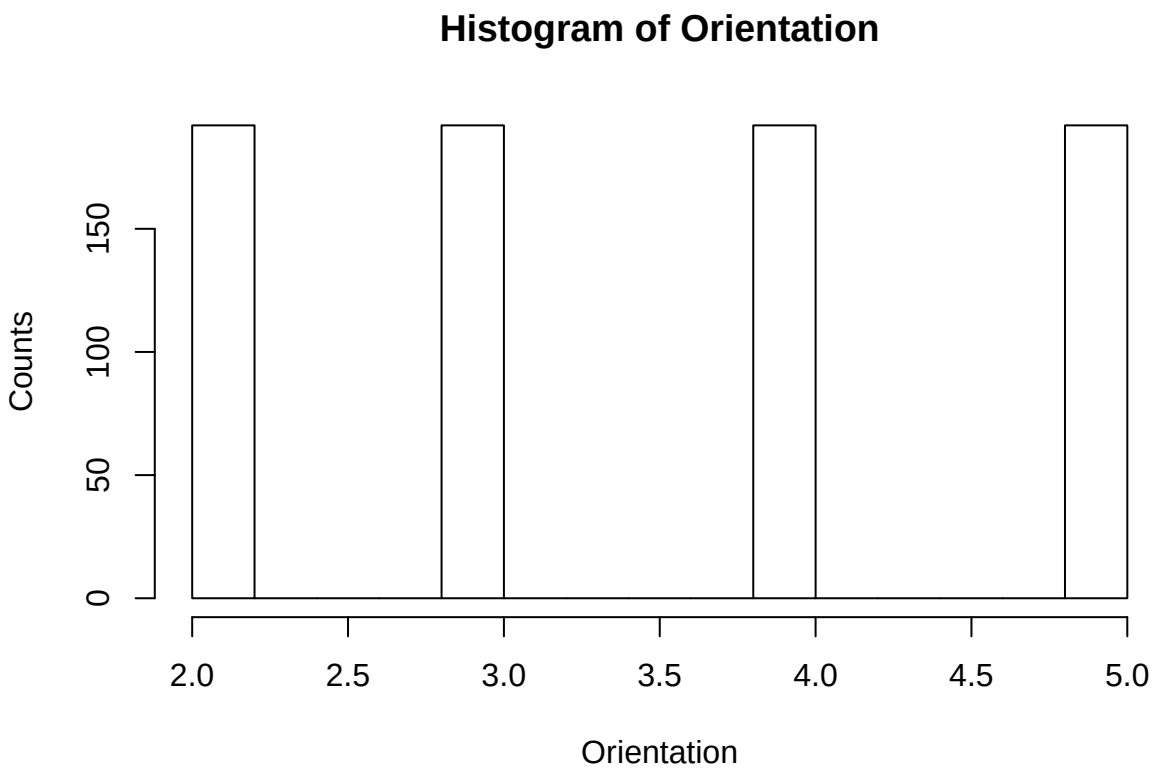
```
hist(energy$Overall.Height, main = paste("Histogram of Overall Height"), xlab = "Overall Height", ylab
```



All the measurements of the Overall Height are either 3.5 units or 7.0 units. The distribution is equal for these measurements

Histogram of Orientation

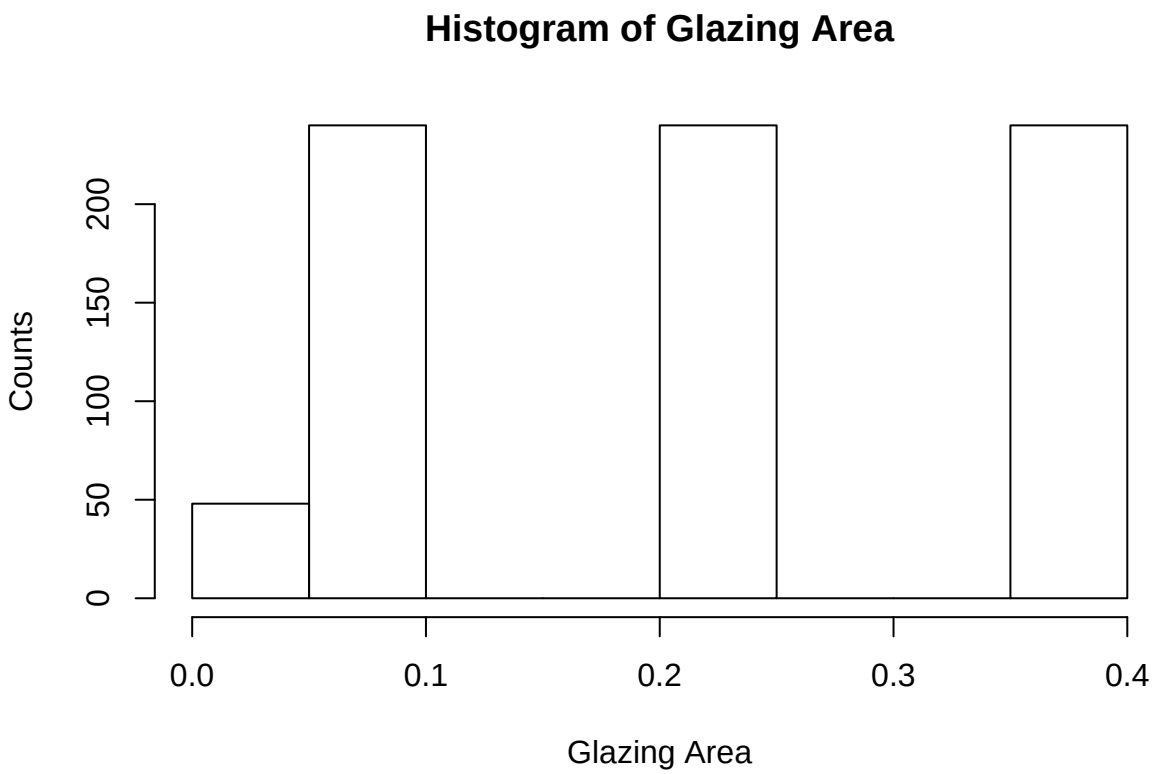
```
hist(energy$Orientation, main = paste("Histogram of Orientation"), xlab = "Orientation", ylab = "Coun
```



Orientation takes only 4 values and the there seems to be equal distribution of those values of orientation.

Histogram of Glazing Area

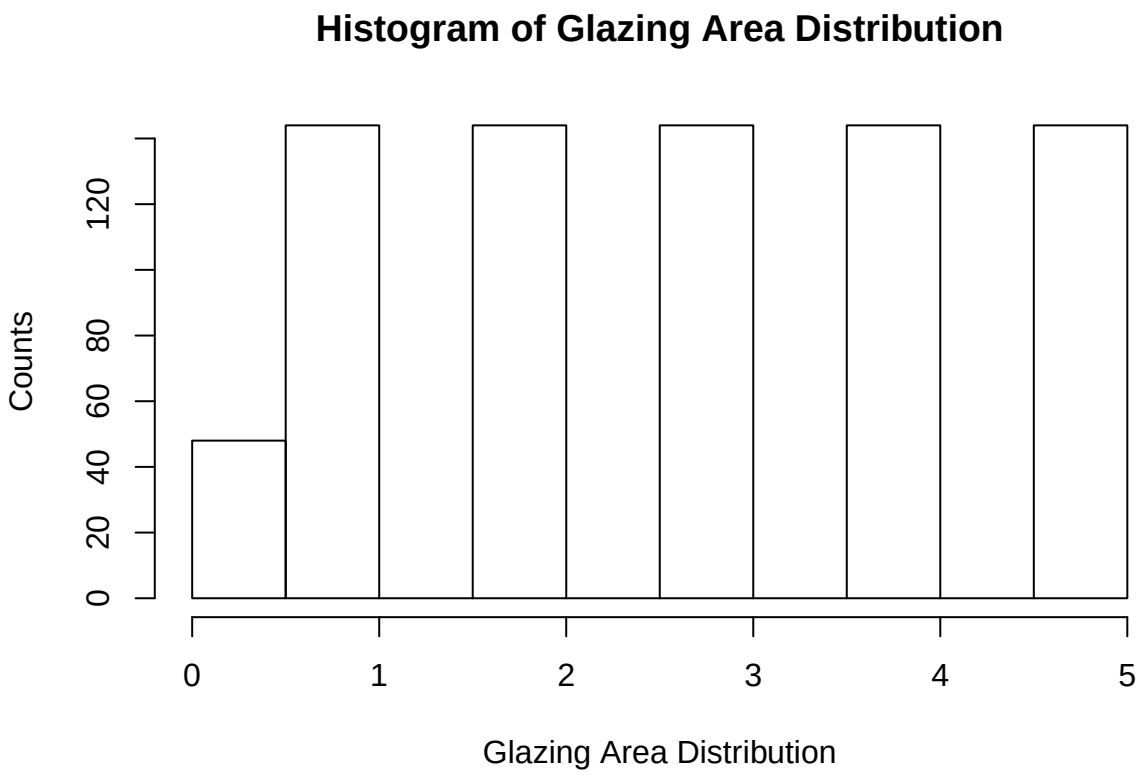
```
hist(energy$Glazing.Area, main = paste("Histogram of Glazing Area"), xlab = "Glazing Area", ylab = "C
```



Glazing Area also has only 4 values. Three non-zero values (0.1,0.25 and 0.4) have equal distribution for the counts.

Histogram of Glazing Area Distribution

```
hist(energy$Glazing.Area.Distribution, main = paste("Histogram of Glazing Area Distribution"), xlab =
```



Glazing Area distribution takes 5 integer value (0 to 5). The distribution across non-zero values are equal.

4. Analysis of factors contributing to Heating and Cooling Load

Analyzing Correlation

```
# Analyzing Correlation:
cor(energy)
```

```
##                           Relative.Compactness  Surface.Area  Wall.Area
## Relative.Compactness              1.000000e+00 -9.919015e-01 -0.2037817
## Surface.Area                     -9.919015e-01  1.000000e+00  0.1955016
## Wall.Area                        -2.037817e-01  1.955016e-01  1.0000000
## Roof.Area                        -8.688234e-01  8.807195e-01 -0.2923165
## Overall.Height                    8.277473e-01 -8.581477e-01  0.2809757
## Orientation                       0.000000e+00  0.000000e+00  0.0000000
## Glazing.Area                      7.617400e-20  4.664140e-20  0.0000000
## Glazing.Area.Distribution         0.000000e+00  0.000000e+00  0.0000000
## Heating.Load                      6.222722e-01 -6.581202e-01  0.4556712
## Cooling.Load                      6.343391e-01 -6.729989e-01  0.4271170
##                              Roof.Area Overall.Height  Orientation
## Relative.Compactness     -8.688234e-01      0.8277473  0.000000000
## Surface.Area              8.807195e-01     -0.8581477  0.000000000
## Wall.Area                -2.923165e-01      0.2809757  0.000000000
## Roof.Area                 1.000000e+00     -0.9725122  0.000000000
## Overall.Height           -9.725122e-01      1.0000000  0.000000000
```

```
## Orientation                0.000000e+00      0.0000000  1.000000000
## Glazing.Area              -1.197187e-19      0.0000000  0.000000000
## Glazing.Area.Distribution  0.000000e+00      0.0000000  0.000000000
## Heating.Load              -8.618283e-01      0.8894307 -0.002586534
## Cooling.Load              -8.625466e-01      0.8957852  0.014289598
##                            Glazing.Area Glazing.Area.Distribution
## Relative.Compactness       7.617400e-20                0.00000000
## Surface.Area               4.664140e-20                0.00000000
## Wall.Area                  0.000000e+00                0.00000000
## Roof.Area                 -1.197187e-19                0.00000000
## Overall.Height             0.000000e+00                0.00000000
## Orientation                0.000000e+00                0.00000000
## Glazing.Area               1.000000e+00                0.21296422
## Glazing.Area.Distribution  2.129642e-01                1.00000000
## Heating.Load               2.698410e-01                0.08736759
## Cooling.Load               2.075050e-01                0.05052512
##                           Heating.Load Cooling.Load
## Relative.Compactness       0.622272179   0.63433907
## Surface.Area              -0.658120227  -0.67299893
## Wall.Area                  0.455671157   0.42711700
## Roof.Area                 -0.861828253  -0.86254660
## Overall.Height             0.889430674   0.89578517
## Orientation               -0.002586534   0.01428960
## Glazing.Area               0.269840996   0.20750499
## Glazing.Area.Distribution  0.087367594   0.05052512
## Heating.Load               1.000000000   0.97586181
## Cooling.Load               0.975861813   1.00000000
```

- Some observations from the Correlation information:
  - Heating/Cooling Load seems to have high coorrelation (>0.8) with Overall Height.
  - Heating/Cooling Load seems to have some coorrelation (>0.5) with Relative Compactness.
  - Heating/Cooling Load seems to have high anti-coorrelation (<-0.8) with Roof Area.
  - Heating/Cooling Load seems to have some anti-coorrelation (<-0.5) with Surface Area.
  - Orientation does not seem to be correlated to any other variables. Additionally, Heating/Cooling load seems to have identifical distribution in all Orientations.

Plot variables that impact Heating/Cooling load

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

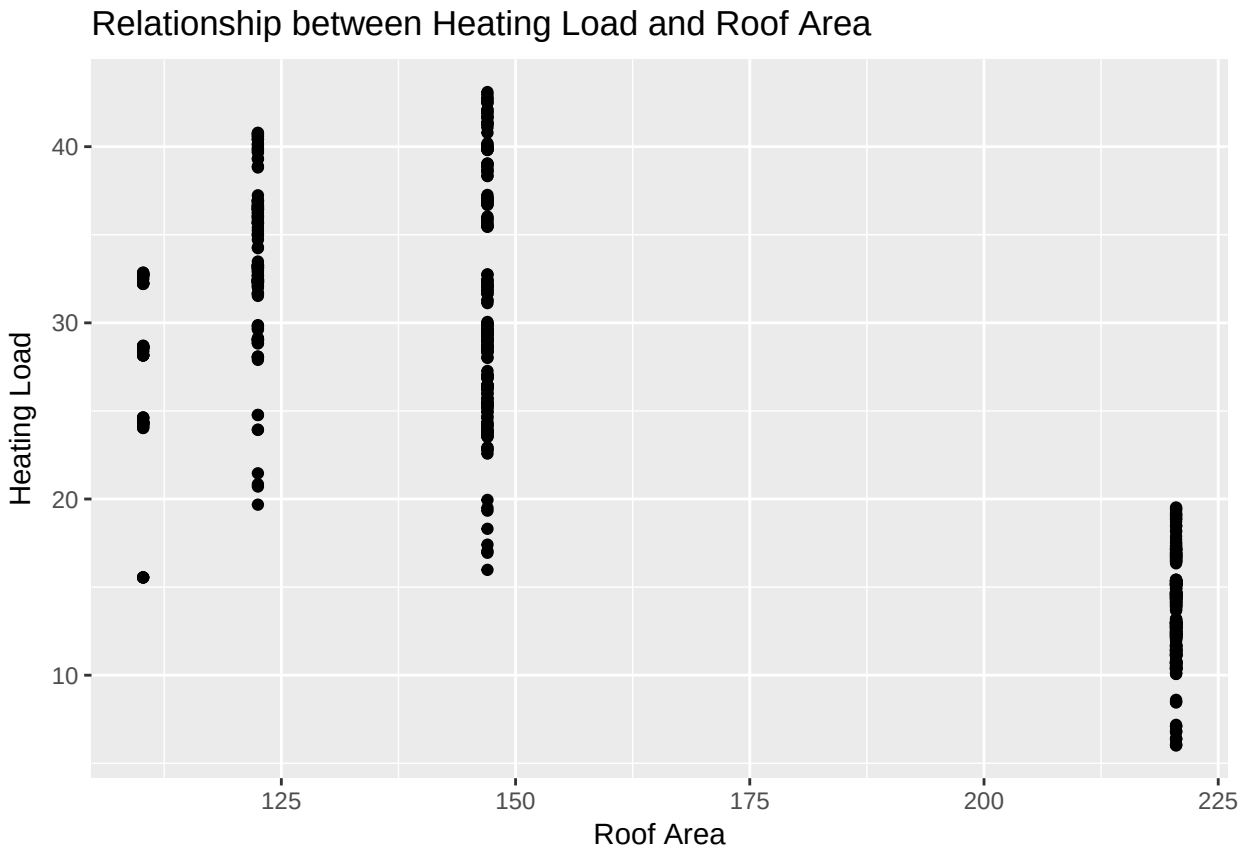
```
# Correlation between Heating/Cooling Load and Roof Area:
cor(energy$Heating.Load,energy$Roof.Area)
```

```
## [1] -0.8618283
```

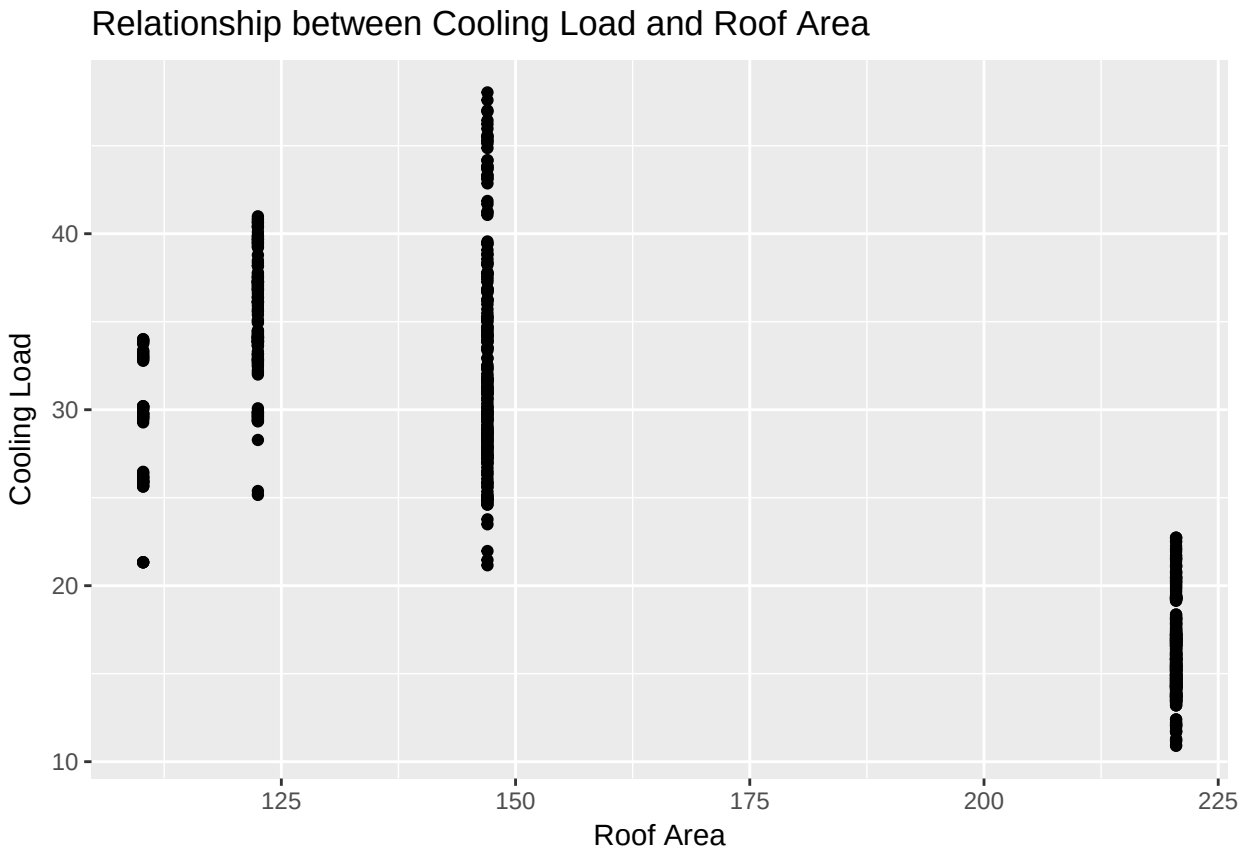
```
cor(energy$Cooling.Load,energy$Roof.Area)
```

```
## [1] -0.8625466
```

```
#Scatter plot of the Heating/Cooling Load and Roof Area
ggplot(energy, aes(y = Heating.Load, x = Roof.Area))+ geom_point() + xlab("Roof Area") + ylab("Heatin
```




```
ggplot(energy, aes(y = Cooling.Load, x = Roof.Area))+ geom_point() + xlab("Roof Area") + ylab("Coolin
```



- Correlation between Heating Load and Roof Area is : -0.862. Hence there is strong anti-correlation between Roof Area and Heating Load.
- Correlation between Cooling Load and Roof Area is : -0.863. Hence there is strong anti-correlation between Roof Area and Cooling Load.
- The Scatter plot shows that Heating/Cooling load seems to be higher when the Roof Areas is below 150 Units. When the Roof Area is above 200 units the Heating/Cooling load is lower

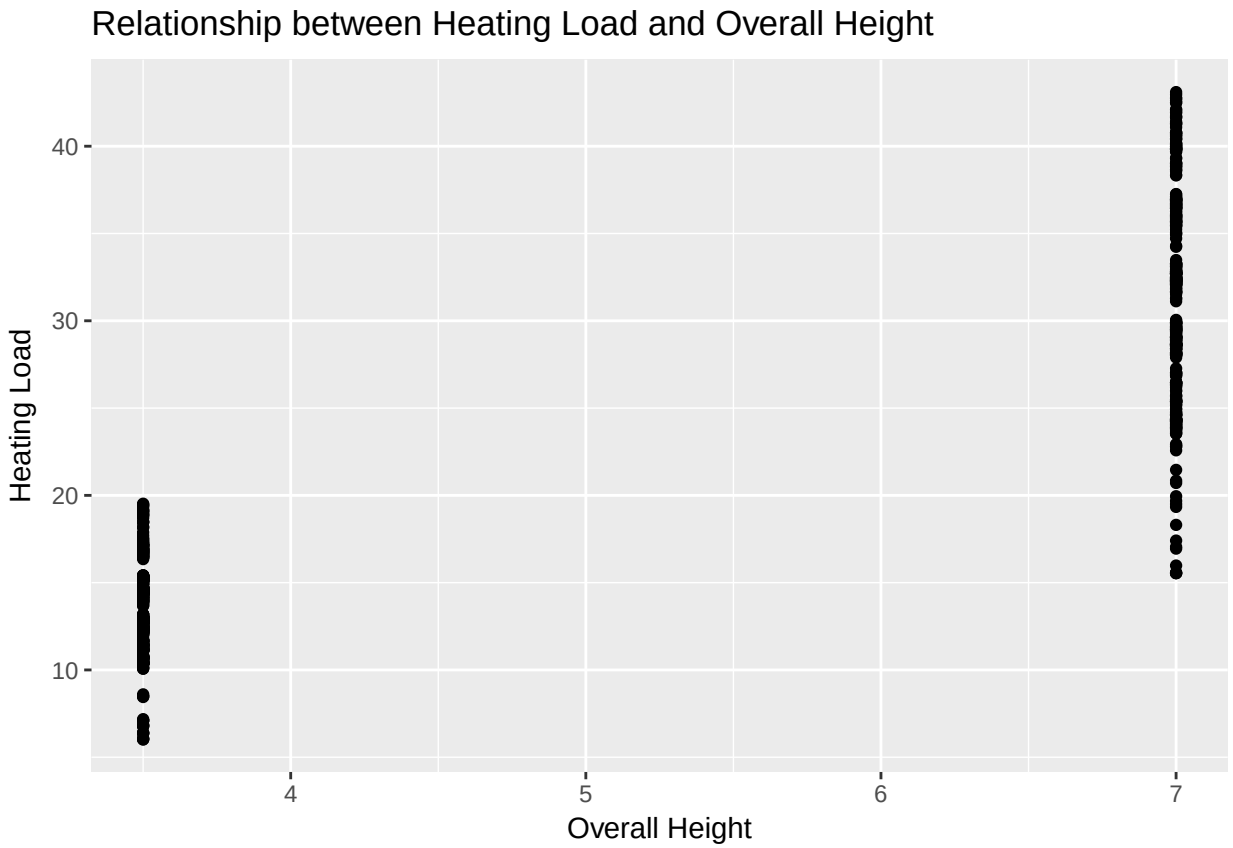
```
# Correlation between Heating/Cooling Load and Overall Height:
cor(energy$Heating.Load,energy$Overall.Height)
```

```
## [1] 0.8894307
```

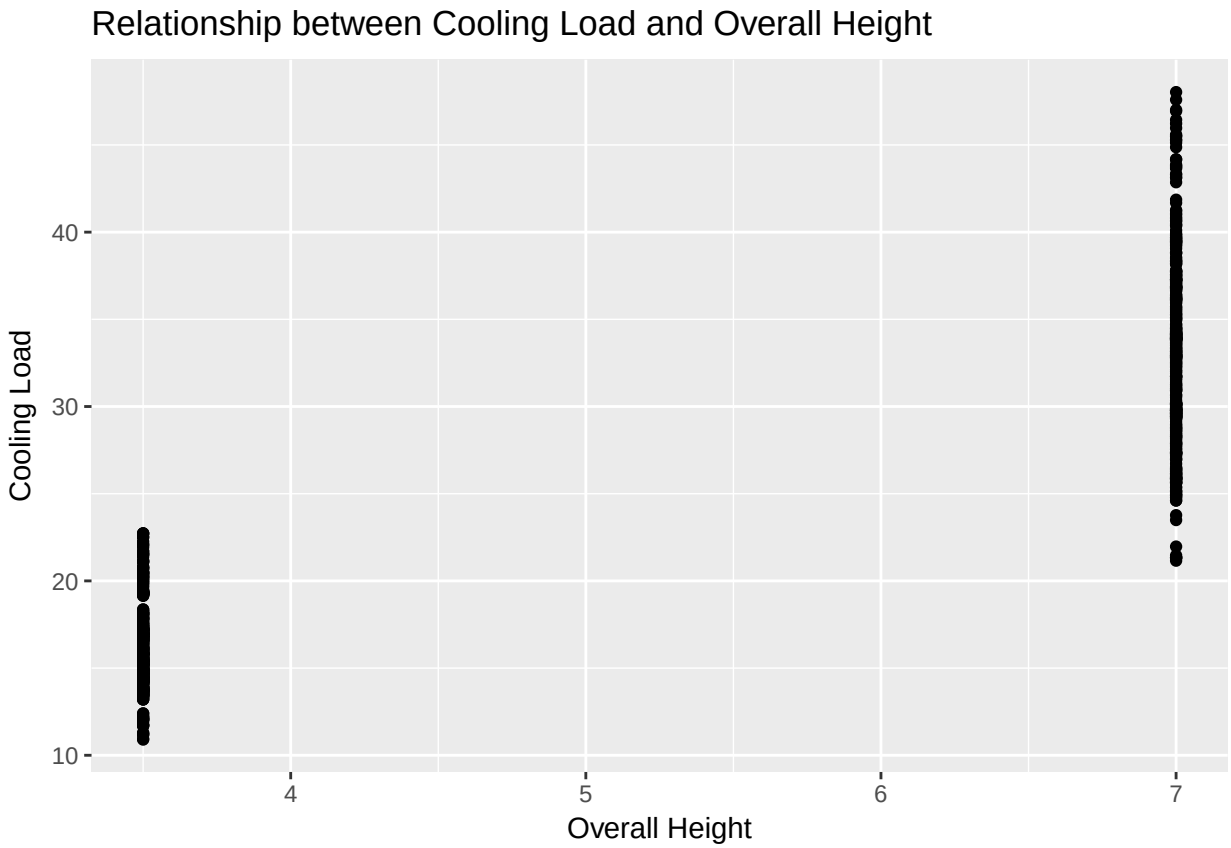
```
cor(energy$Cooling.Load,energy$Overall.Height)
```

```
## [1] 0.8957852
```

```
#Scatter plot of the Heating/Cooling Load and Overall Height
ggplot(energy, aes(y = Heating.Load, x = Overall.Height))+ geom_point() + xlab("Overall Height") + yla
```




```
ggplot(energy, aes(y = Cooling.Load, x = Overall.Height))+ geom_point() + xlab("Overall Height") + yla
```



- Correlation between Heating Load and Roof Area is : 0.889. Hence there is strong correlation between Overall Height and Heating Load.
- Correlation between Cooling Load and Roof Area is : 0.896. Hence there is strong correlation between Overall Height and Cooling Load.
- The Scatter plot shows that Heating/Cooling load is lower when the Overall Height is smaller than 4 units. As the Overall Height increases above 7 units the Heating Load increases.

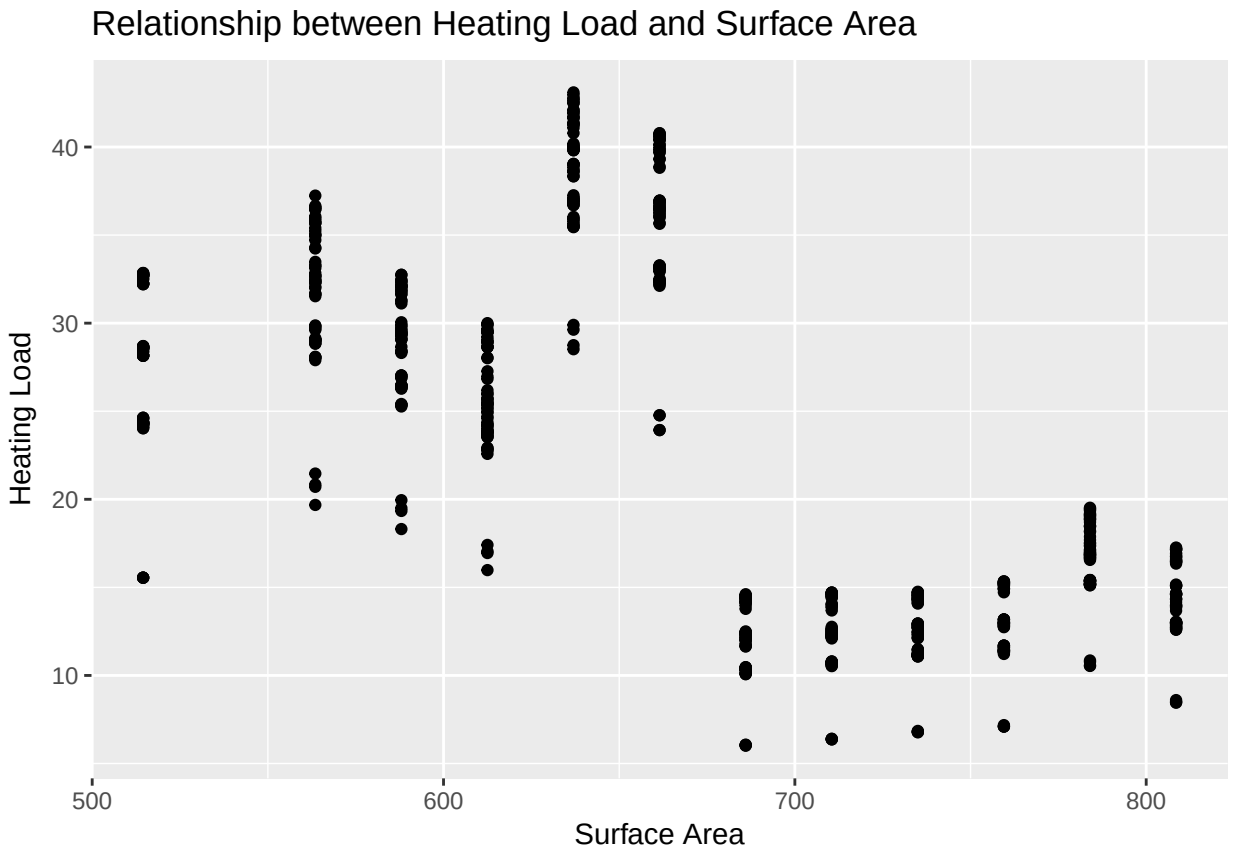
```
# Correlation between Heating/Cooling Load and Surface Area:
cor(energy$Heating.Load,energy$Surface.Area)
```

```
## [1] -0.6581202
```

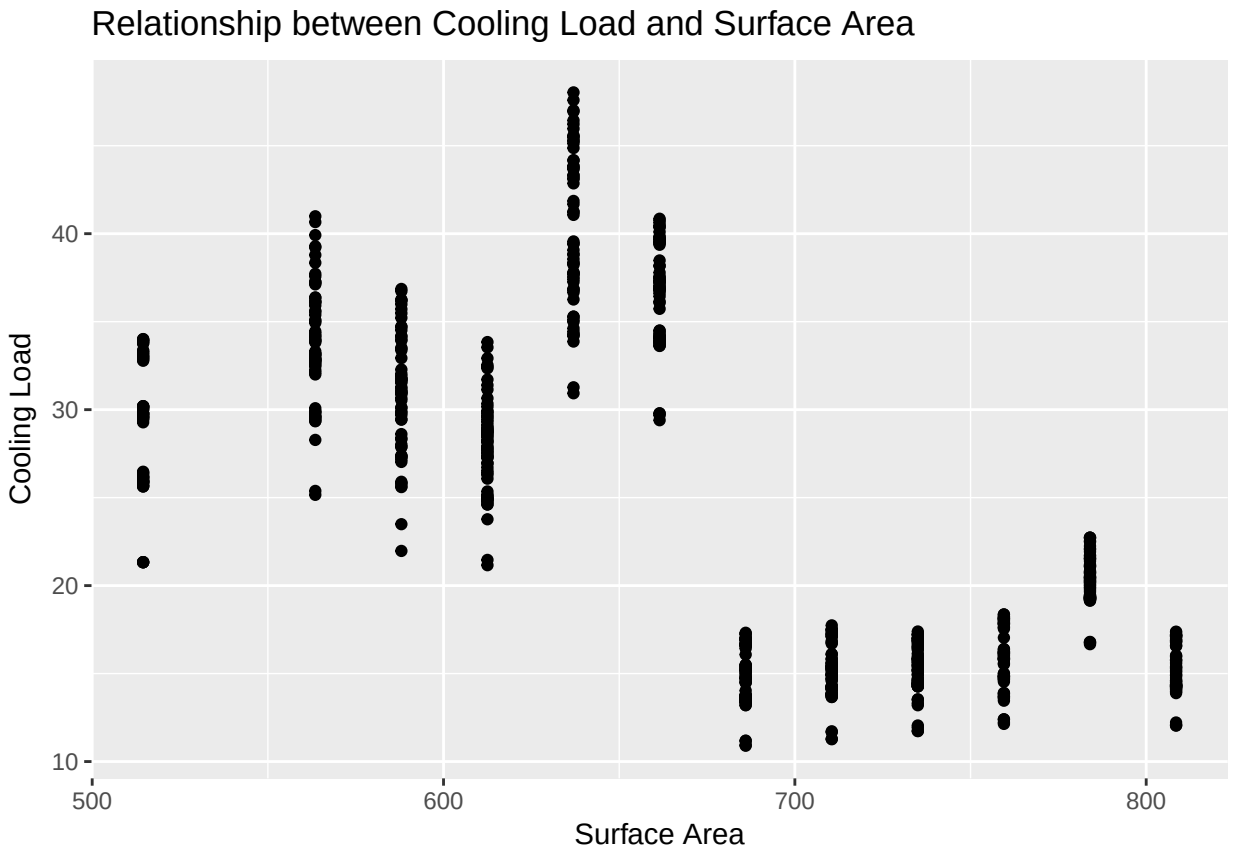
```
cor(energy$Cooling.Load,energy$Surface.Area)
```

```
## [1] -0.6729989
```

```
#Scatter plot of the Heating/Cooling Load and Surface Area:
ggplot(energy, aes(y = Heating.Load, x = Surface.Area))+ geom_point() + xlab("Surface Area") + ylab("
```




```
ggplot(energy, aes(y = Cooling.Load, x = Surface.Area))+ geom_point() + xlab("Surface Area") + ylab("
```



- Correlation between Heating Load and Roof Area is : -0.658. Hence there is strong anti-correlation between Surface Area and Heating Load.
- Correlation between Cooling Load and Roof Area is : -0.673. Hence there is strong anti-correlation between Surface Area and Cooling Load.
- The Scatter plot shows that Heating/Cooling loads seems to be higher when the Surface Area is smaller than 675 units. As the Surface Area increases beyond 675 Units, the Heating/Cooling load starts decreasing.

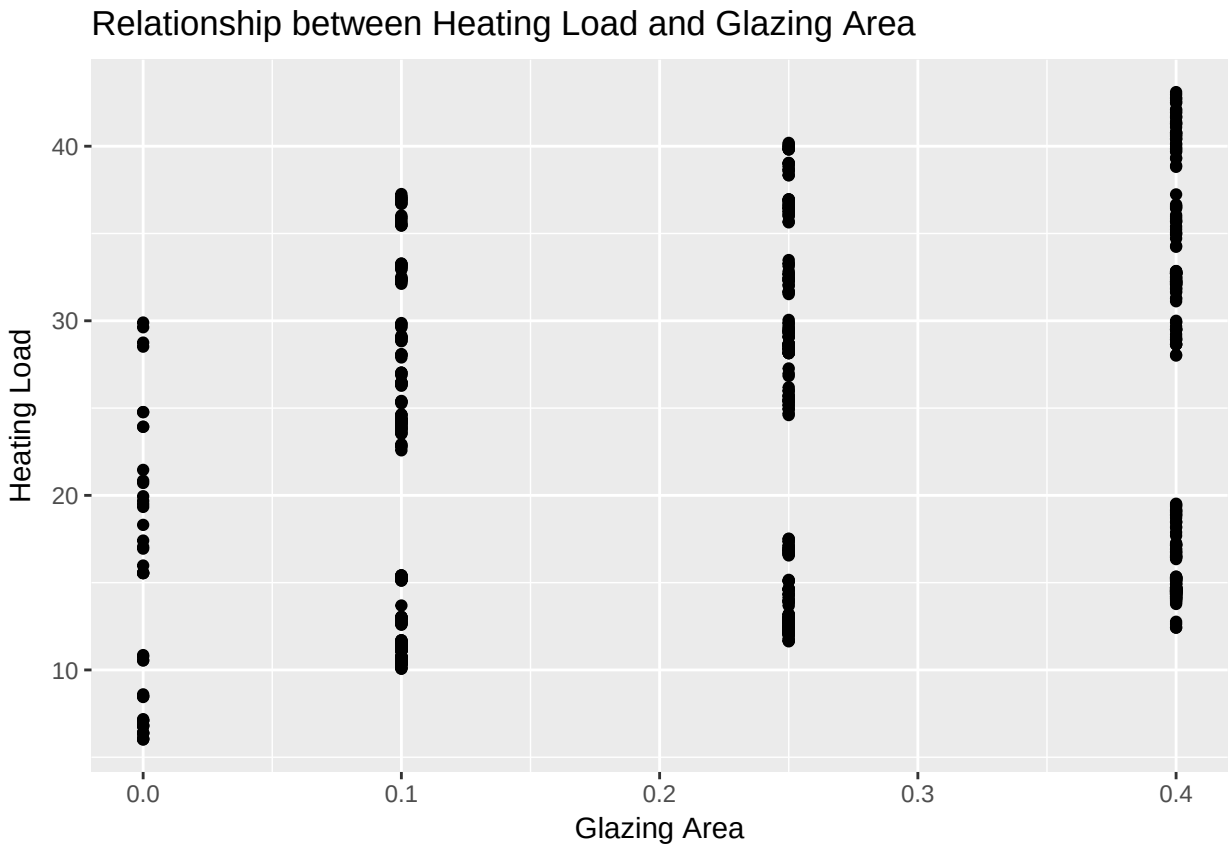
```
# Correlation between Heating/Cooling Load and Glazing Area:
cor(energy$Heating.Load,energy$Glazing.Area)
```

```
## [1] 0.269841
```

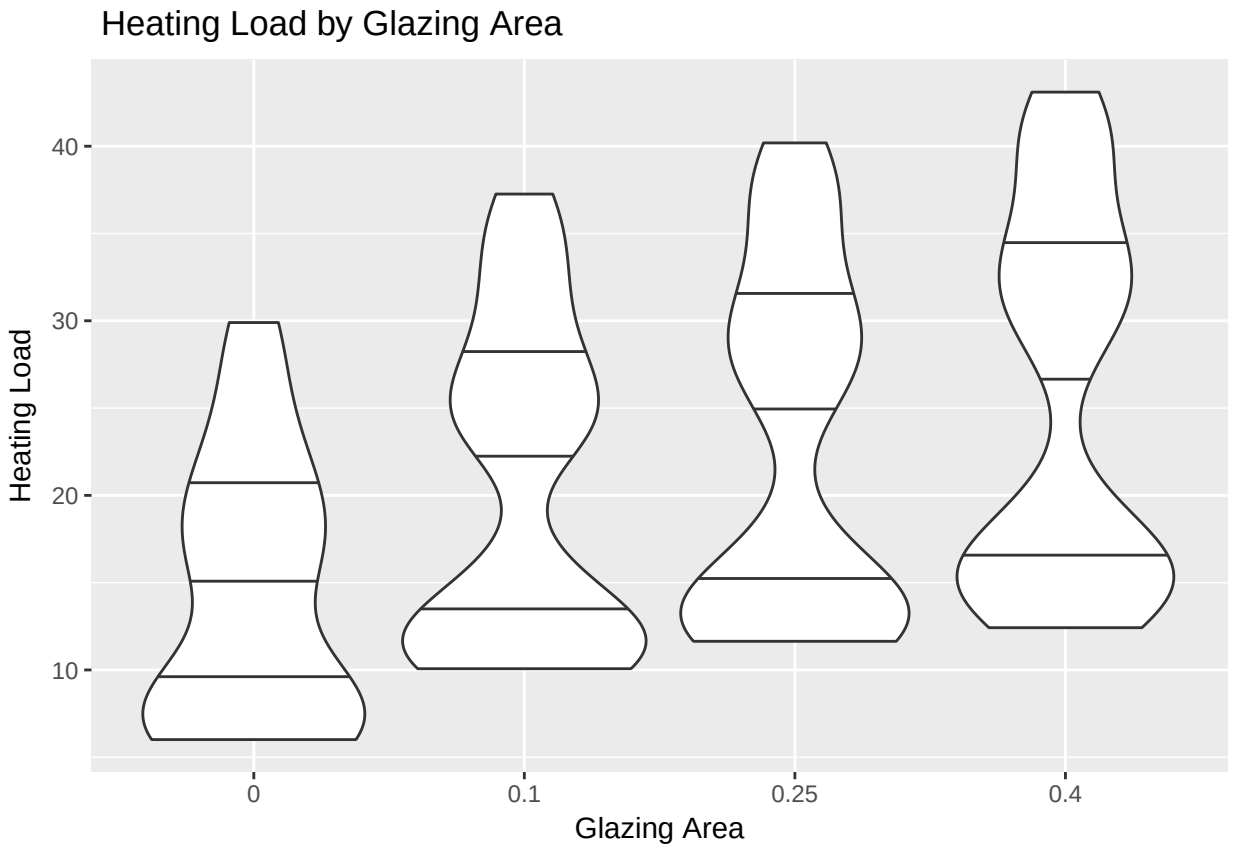
```
cor(energy$Cooling.Load,energy$Glazing.Area)
```

```
## [1] 0.207505
```

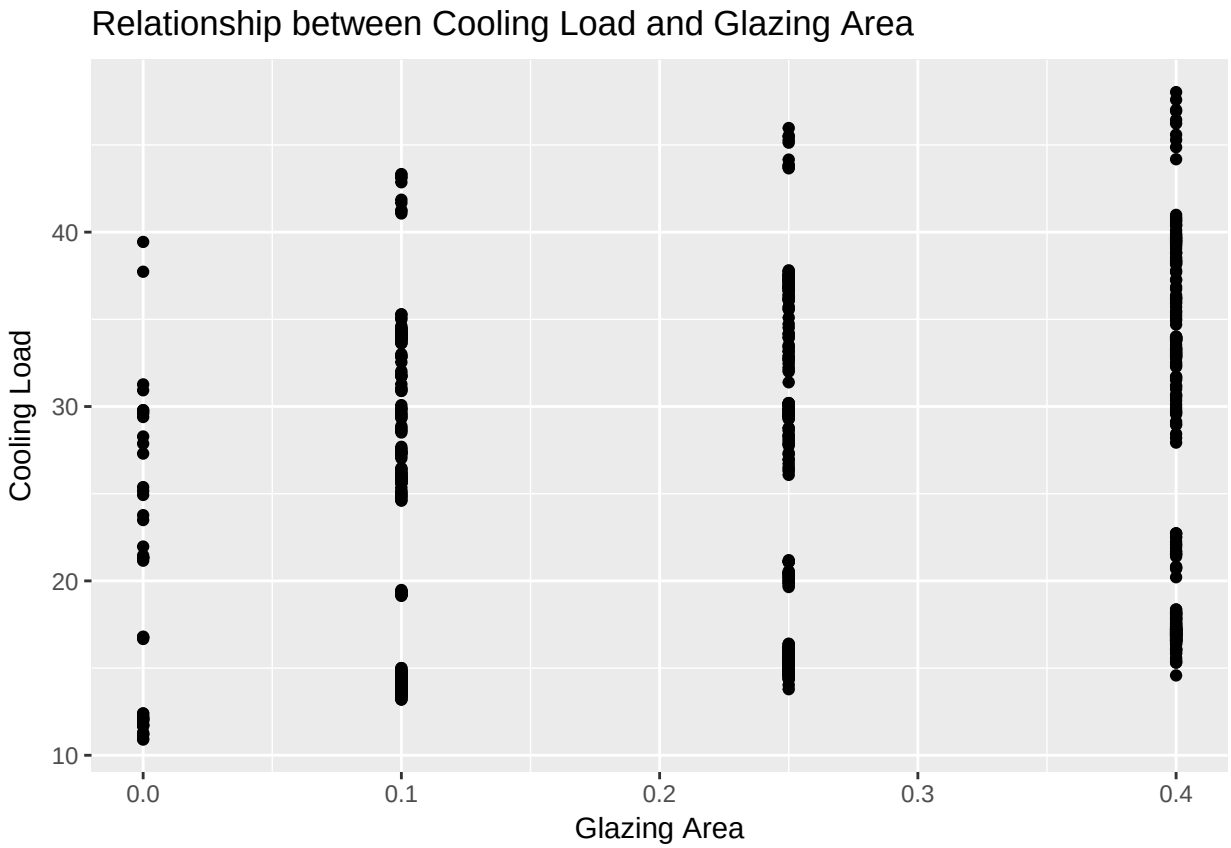
```
#Scatter and Violin plot of the Heating/Cooling Load and Glazing Area
ggplot(energy, aes(y = Heating.Load, x = Glazing.Area))+ geom_point() + xlab("Glazing Area") + ylab("
```



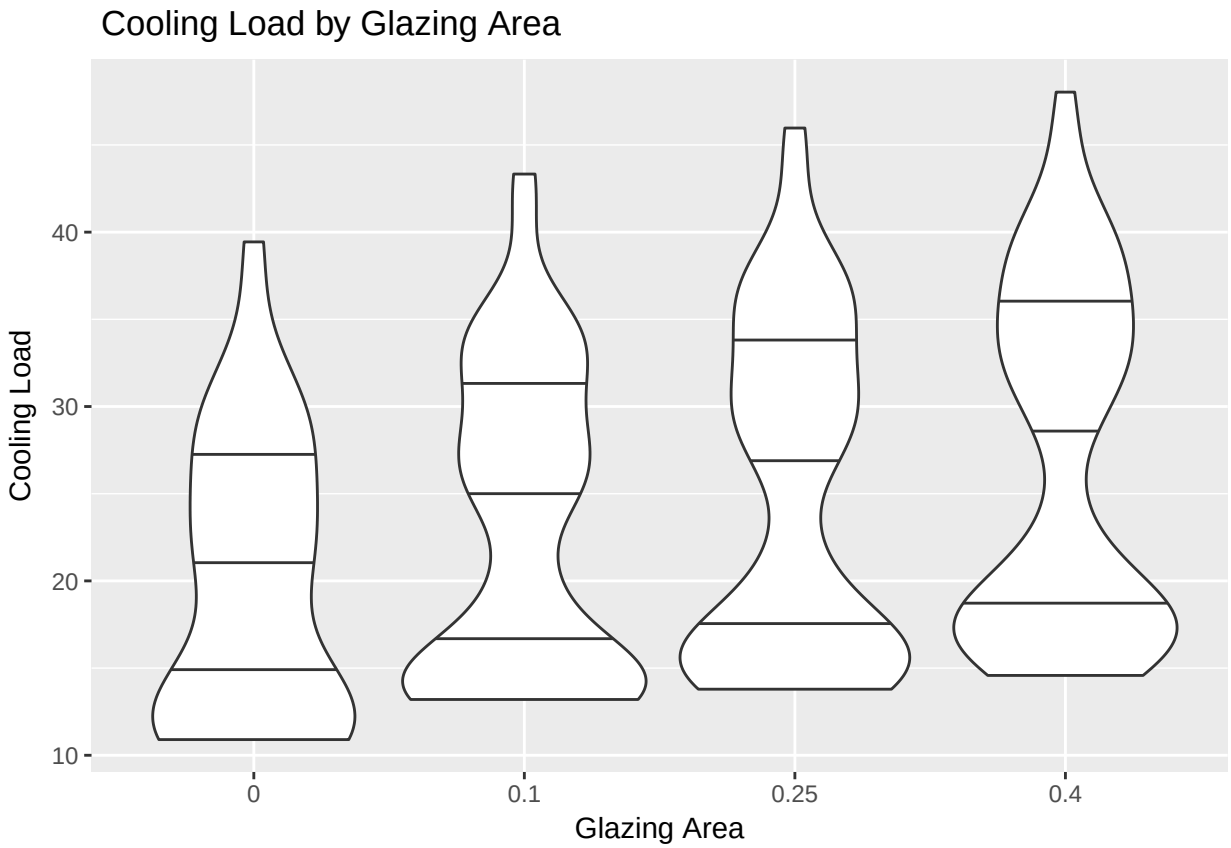

```
ggplot(energy, aes(x = factor(Glazing.Area), y = Heating.Load)) + geom_violin(trim = TRUE, draw_quanti
```




```
ggplot(energy, aes(y = Cooling.Load, x = Glazing.Area))+ geom_point() + xlab("Glazing Area") + ylab("
```

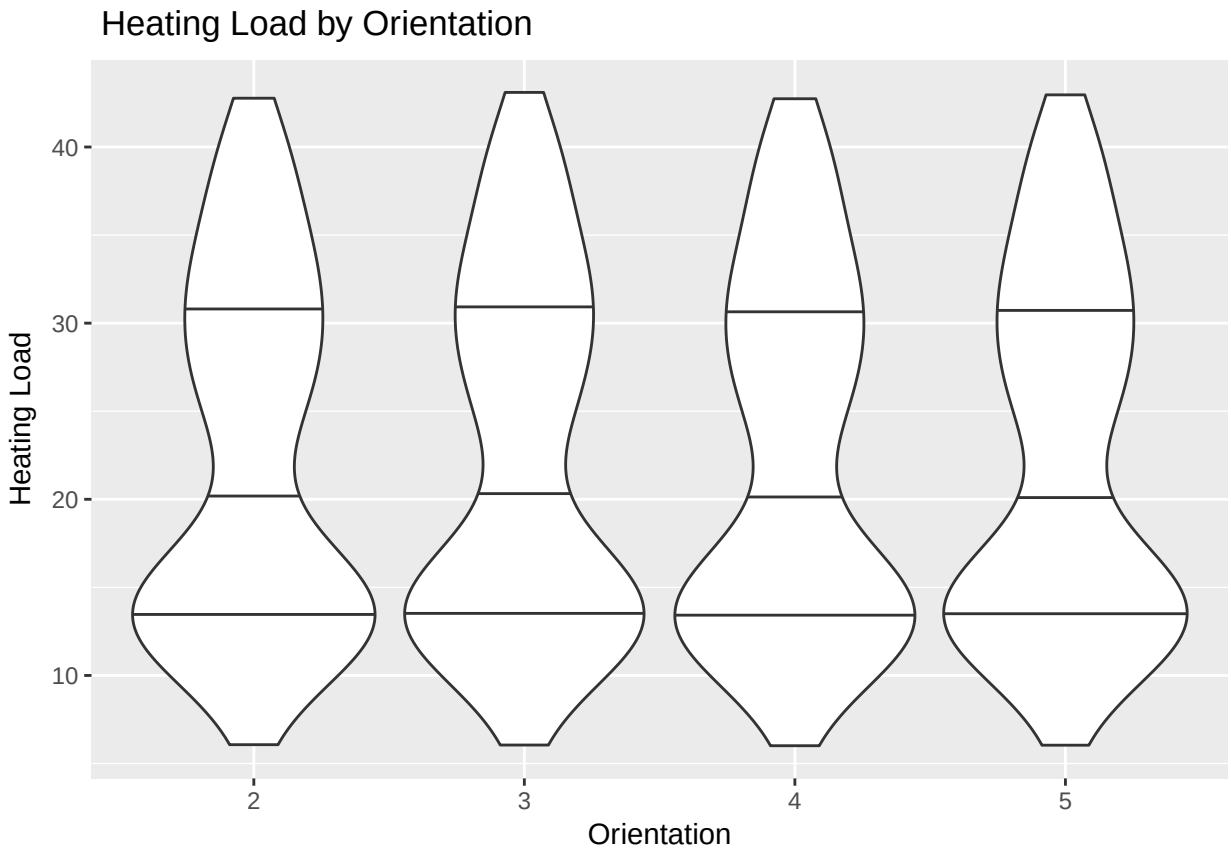



```
ggplot(energy, aes(x = factor(Glazing.Area), y = Cooling.Load)) + geom_violin(trim = TRUE, draw_quanti
```

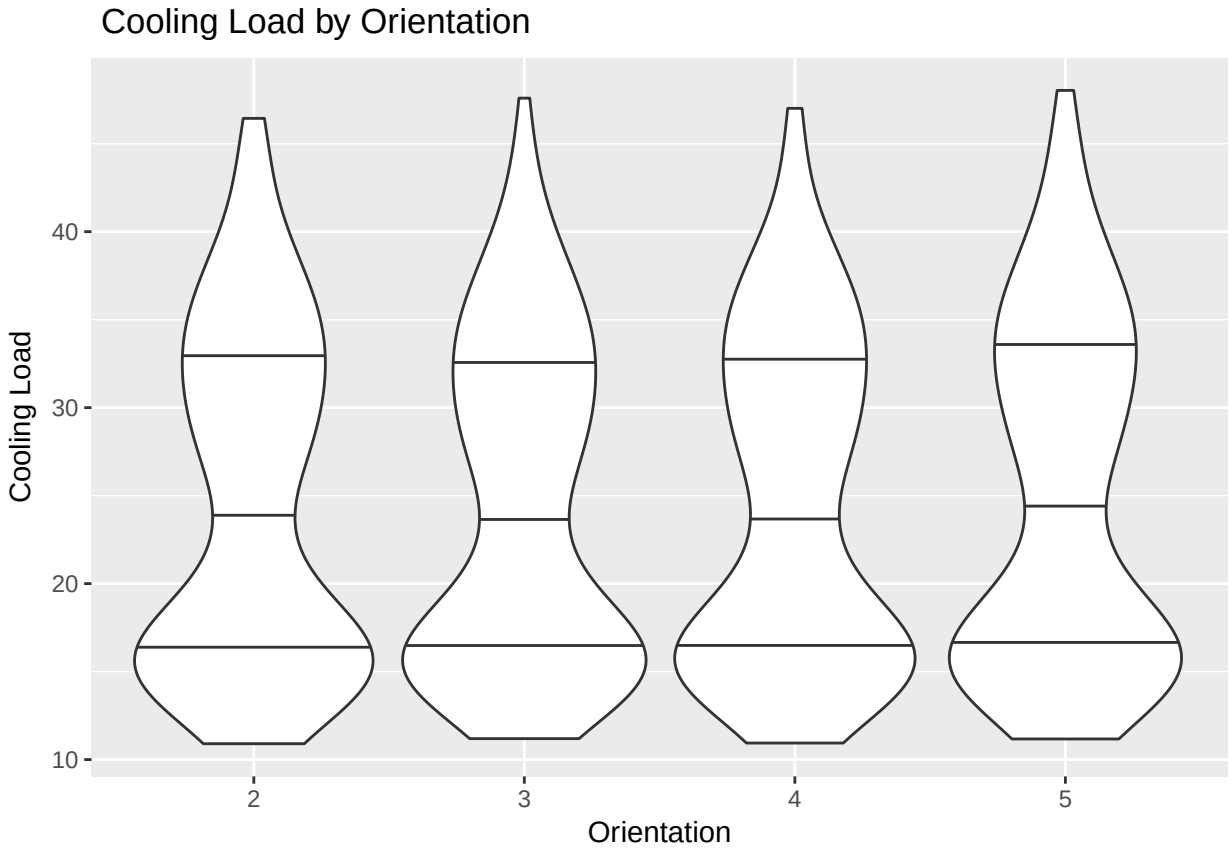


- Correlation between Heating Load and Glazing Area is : 0.270. Correlation coefficient is low implying not a strong correlation between Glazing Area and Heating Load.
- Correlation between Cooling Load and Glazing Area is : 0.208. Correlation coefficient is low implying not a strong correlation between Glazing Area and Heating Load.
- However, the violin plot shows that the distribution of Heating/Cooling moves towards higher values as the Glazing area changes from 0 units to 0.4 units.

```
# Violin plot of the Heating/Cooling Load and Orientation
ggplot(energy, aes(x = factor(Orientation), y = Heating.Load)) + geom_violin(trim = TRUE, draw_quanti
```




```
ggplot(energy, aes(x = factor(Orientation), y = Cooling.Load)) + geom_violin(trim = TRUE, draw_quanti
```



- The violin plot shows that the distribution of Heating/Cooling distribution is same for all Orientations.