A Project Report on Design of Experiment and Factor Analysis

Gurtej Khanooja

April 2017

Contents

Chapter 1

Introduction

1.1 Objective

- To carry out sequential factorial analysis methods on energy effiency dataset.
- To carry out one factorial analysis, multiple factorial analysis and 2^k factorial analysis on the dataset implementing all the techniques learned in class.
- Provide proper conclusion for each factorial analysis technique used on the dataset.

1.2 Understanding the Dataset

```
data<-read.csv("/Users/Gurtej/Documents/NEU/Course Material/Semester 2/SME/SME - Project 1/ENB2012_data
head(data)
```

```
##     X1    X2    X3     X4 X5 X6 X7 X8    Y1
## 1 0.98 514.5 294.0 110.25  7  2  0  0 15.55
## 2 0.98 514.5 294.0 110.25  7  3  0  0 15.55
## 3 0.98 514.5 294.0 110.25  7  4  0  0 15.55
## 4 0.98 514.5 294.0 110.25  7  5  0  0 15.55
## 5 0.90 563.5 318.5 122.50  7  2  0  0 20.84
## 6 0.90 563.5 318.5 122.50  7  3  0  0 21.46
```

The dataset is borrowed from UC Irvine Machine Learning Repository. It's an energy effiency dataset with 768 observations.There are total of 8 attributes and 1 respose variable in the dataset. The different attributes collectively are responsible for the energy consumption (the response variable). Each factor contributes towards the energy consumption in its own way.

Following are the attributes and the dependent variable in the data set:

List of independent variables:

- X1 Relative Compactness
- X2 Surface Area
- X3 Wall Area
- X4 Roof Area
- X5 Overall Height
- X6 Orientation
- X7 Glazing Area
- X8 Glazing Area Distribution

Dependent Variable:

- y1 Heating Load

Our goal is to perform factor analysis on the energy effiency data and get hands on experience of concepts learned in the class. Using stastical results we would like to understand, which are the most influential factors that affect the heating load. We will start with one factor analysis by randomly selecting a factor from our dataset and seeing it influence on the dependent variable by applying factor analysis. This would be further extended to two and three factor anlaysis. Moving on we will see what is $ 2^k $ factor analysis and see how to design the experiment for it and try to see $ 2^k $ factor analysis in action by considering multiple factors.

1.3 Loading the dataset

```
data <- read.csv("/Users/Gurtej/Documents/NEU/Course Material/Semester 2/SME/SME - Project 1/ENB2012_da
head (data)
```

```
##     X1    X2    X3     X4 X5 X6 X7 X8    Y1
## 1 0.98 514.5 294.0 110.25  7  2  0  0 15.55
## 2 0.98 514.5 294.0 110.25  7  3  0  0 15.55
## 3 0.98 514.5 294.0 110.25  7  4  0  0 15.55
## 4 0.98 514.5 294.0 110.25  7  5  0  0 15.55
## 5 0.90 563.5 318.5 122.50  7  2  0  0 20.84
## 6 0.90 563.5 318.5 122.50  7  3  0  0 21.46
```

```
attach (data)
```

Chapter 2

One-Factor Designs and Analysis of Variance

2.1 Preparing data for one factor analysis

For one-factor design we are intrested in seeing how the heating load is affected by varying one factor in the dataset.

For the purpose of project here, the attribute that we will consider for the one-factor experiment is orientation of building affecting the heating load. The treatment i.e. the orentation is divided in four levels (2,3,4 and 5). Each level signifies a different orentation.

```
one_factor <- data[,c(6,9)]
one_factor$X6 <- factor(one_factor$X6)
head(one_factor[order(one_factor$X6),], n=50)
```

```
##     X6    Y1
## 1    2 15.55
## 5    2 20.84
## 9    2 19.50
## 13   2 17.05
## 17   2 28.52
## 21   2 24.77
## 25   2  6.07
## 29   2  6.37
## 33   2  6.85
## 37   2  7.18
## 41   2 10.85
## 45   2  8.60
## 49   2 24.58
## 53   2 29.03
## 57   2 26.28
## 61   2 23.53
## 65   2 35.56
## 69   2 32.96
## 73   2 10.36
## 77   2 10.71
## 81   2 11.11
## 85   2 11.68
## 89   2 15.41
## 93   2 12.96
## 97   2 24.29
## 101  2 28.88
## 105  2 26.48
## 109  2 23.75
## 113  2 35.65
## 117  2 33.16
## 121  2 10.42
## 125  2 10.64
## 129  2 11.45
## 133  2 11.45
## 137  2 15.41
```

```
## 141  2 12.88
## 145  2 24.28
## 149  2 28.07
## 153  2 25.41
## 157  2 22.93
## 161  2 35.78
## 165  2 32.52
## 169  2 10.39
## 173  2 10.77
## 177  2 11.22
## 181  2 11.59
## 185  2 15.16
## 189  2 12.68
## 193  2 24.38
## 197  2 29.06
```

In the above data we have extracted two rows from the initial data, the two rows being X6 (orentation of the building) and Y1 (heating load).

Second, we have sorted the data according to ascending order of X6 to begain with our analysis.

2.2 Graphical interpretation of data

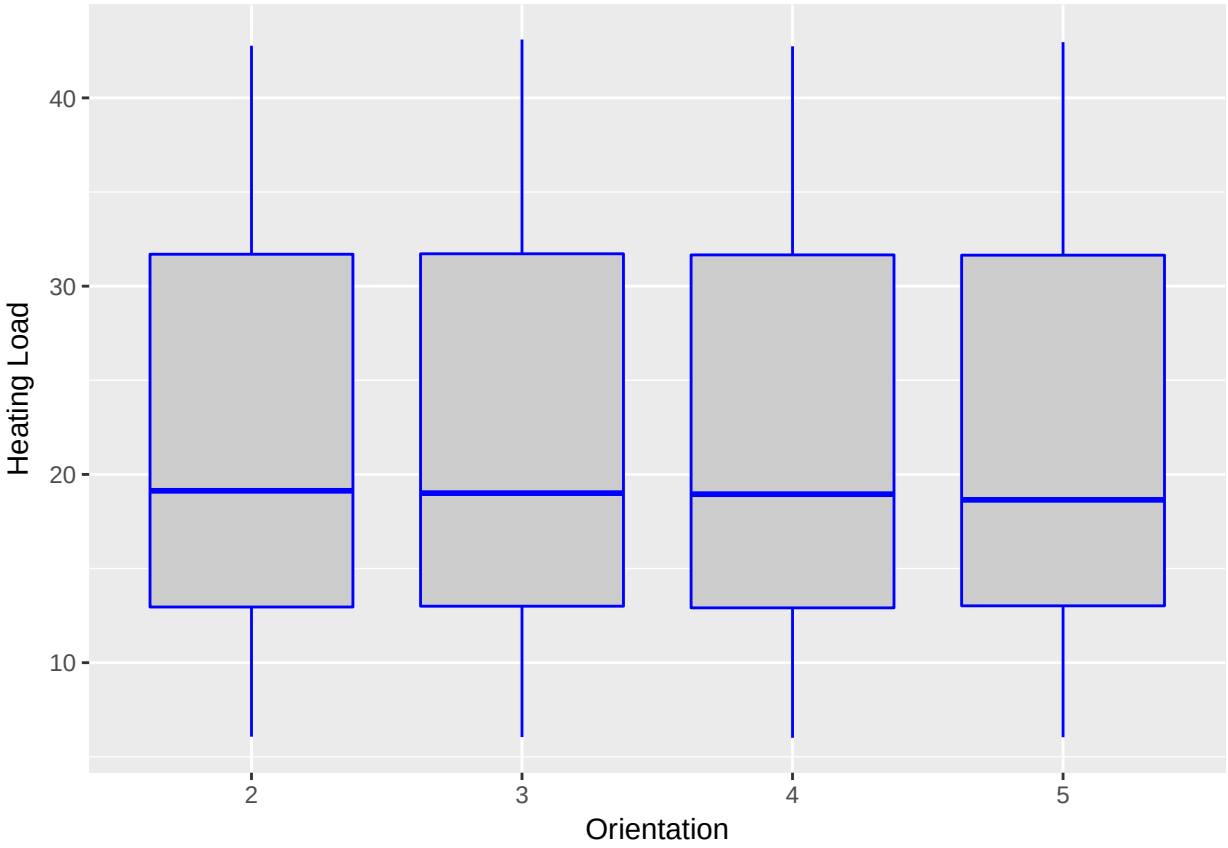
2.2.1 Box Plot

```
require(ggplot2)
```

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

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

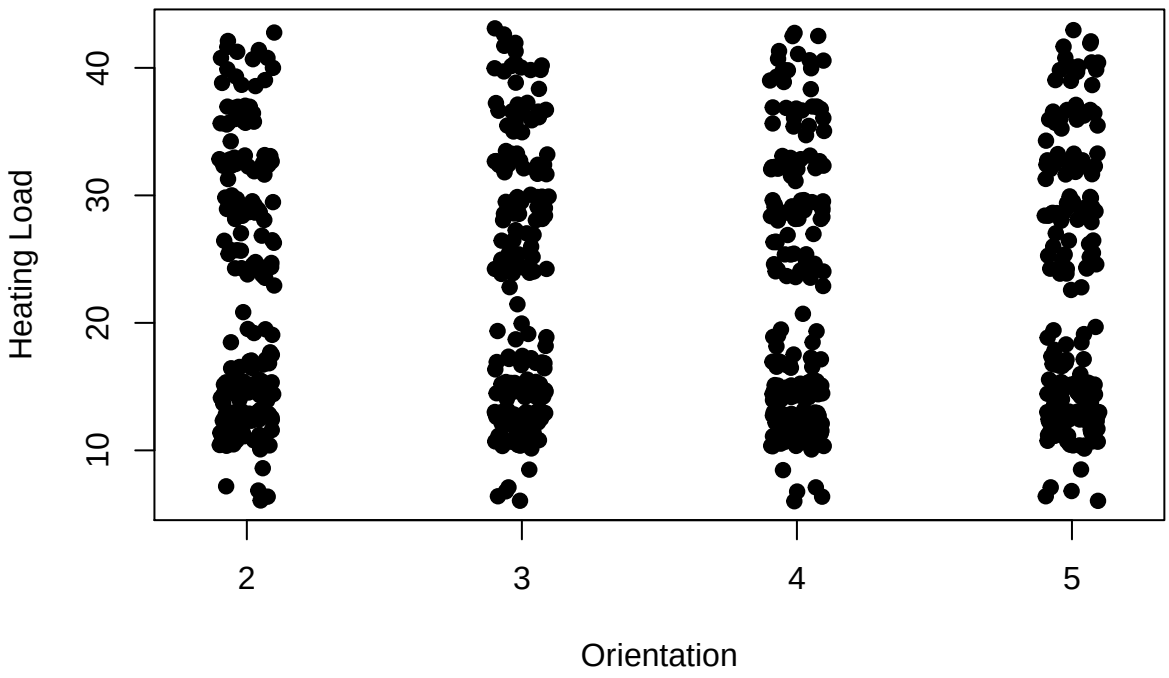
```
ggplot(one_factor, aes(x = X6, y = Y1)) +
geom_boxplot(fill = "grey80", colour = "blue") +
scale_x_discrete() + xlab("Orientation") +
ylab("Heating Load")
```



The above box plot shows variation of heating load with the orientation for various level of orientation factor. We can see from the plot that the values of heating load are uniformly distributed for all different kinds of orentation level suggesting a strong possiblity of same mean observation amoung them.

2.2.2 Jitter Plot

```
stripchart(Y1~as.factor(X6),vertical=T,pch=19,data = data, xlab= "Orientation", ylab = "Heating Load",m
```



The jitter plot also suggest simalar infrences like box plot above. The data seems to be uniformly distributed amoungst various level of orientation suggesting a strong possiblity of equal means amoungst the various levels graphically.

2.3 Model Analysis

2.3.1 One way ANOVA model

$$Ho : \mu 1 = \mu 2 = \mu 3 = \mu 4$$
$$H1 : \mu 1 \neq \mu 2 \neq \mu 3 \neq \mu 4$$

```
analysis <- lm(Y1 ~ X6, data = one_factor)
anova(analysis)
```

```
## Analysis of Variance Table
##
## Response: Y1
##            Df Sum Sq Mean Sq F value Pr(>F)
## X6          3      2   0.556  0.0054 0.9994
## Residuals 764  78088 102.210
```

From the above analysis the sum of square corrosponding to X6 is the treatment sum of squares (SSA) and the sum of square corrosponding to Residual is the sum o of squares of error (SSE).

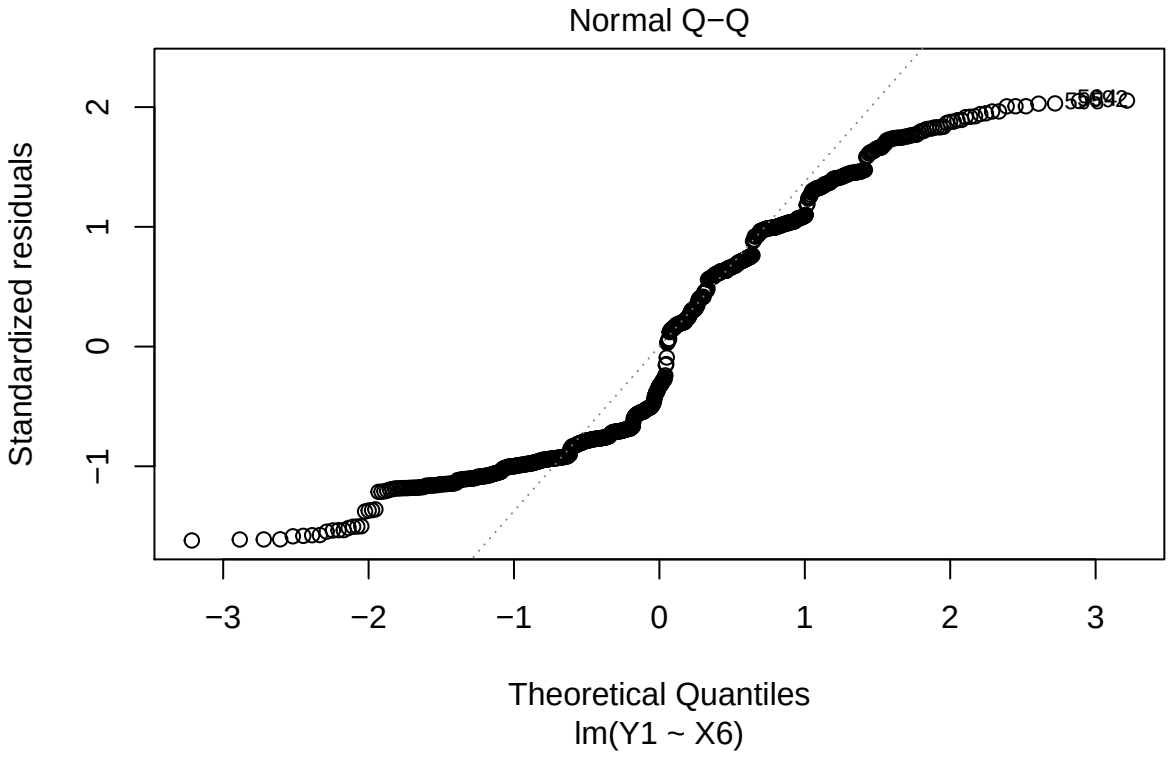
There are two estimates of population variance that we get from above result, one judged from variablity between the orientation of building (i.e. 0.556) and other judged from variablity within the orientation of building (i.e. 102.210).

From above the F-value is 0.0054 and corrosponding P-value is 0.9994 suggesting that we fail to reject the null hypothesis.

2.3.2 Checking Normality Condition

For doing ANOVA for 1 factor design it is important that the normality condition is satisfied. We will try visualizing it using the Q-Q plot. The Q-Q plot for this analysis is below:

```
plot(analysis, which=2)
```



We see that the plotted residuals does not consides the straigth line suggesting violation of the normality condition. The ANOVA is robust provided that the normality conditions are followed. Here we are performing the analyis for one factor analysis but since the normality conditions are not followed we cannot be sure about the results.

2.3.3 Multiple Comparsion Test

Multiple comparsion test is usually performed when we reject Ho and we want to observe what the relation between other means and see that if they are some way connected to each other. The sole purpose of doing this is because the ANOVA does'nt tell us, in case we reject we still don't know which population means are equal and which are not.

(a) Tukey's Test

```
TukeyHSD(aov(analysis))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = analysis)
##
## $X6
##            diff       lwr      upr     p adj
## 3-2  0.06781250 -2.588848 2.724473 0.9998981
## 4-2 -0.05296875 -2.709629 2.603691 0.9999514
## 5-2 -0.03750000 -2.694160 2.619160 0.9999828
## 4-3 -0.12078125 -2.777441 2.535879 0.9994263
## 5-3 -0.10531250 -2.761973 2.551348 0.9996193
## 5-4  0.01546875 -2.641191 2.672129 0.9999988
```

Though it is not of use to perform the Tukey's test in this particular case. But from the above result we can see that all the p-values are very high of range 0.999 suggesting that there is no significant difference between between means of different orentation.

(b) Duncan's multiple-range test

```
library(agricolae)
duncan.test(aov(analysis), "one_factor$X6", 0.05, console = TRUE)
```

```
## Name:  one_factor$X6
##  X6
```

There is no group formation in the Duncan Test suggesting that the mean of all the levels within the treatment is stastically same according to Ducan Test.

(c) Dunnett's Test

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 3.3.2

## Loading required package: survival

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 3.3.2

## Loading required package: MASS

##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
Dunnett <- glht(aov(analysis), linfct = mcp(X6 ="Dunnett"))
summary(Dunnett)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = analysis)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 3 - 2 == 0  0.06781    1.03184   0.066        1
## 4 - 2 == 0 -0.05297    1.03184  -0.051        1
## 5 - 2 == 0 -0.03750    1.03184  -0.036        1
## (Adjusted p values reported -- single-step method)
```

The P-value is equal to 1 for all observations in the table. This suggest that there is no significant difference between the levels within the treatment X6.

(d) Bonferroni's test

```
pairwise.t.test(one_factor$Y1, one_factor$X6, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  one_factor$Y1 and one_factor$X6
##
##   2 3 4
## 3 1 - -
## 4 1 1 -
## 5 1 1 1
##
## P value adjustment method: bonferroni
```

The high p-value between all paiirs indicates that the the levels within treatment are not significantly different.

Chapter 3

Multiple-Factor Designs and Analysis of Variance

3.1 Two Factor design and Two way ANOVA

3.1.1 Preparing data for two factor analysis

For two factor analysis we create a experiment such that we would like to see the effects that 2 factors play on the dependent variable. For the project purpose we continue taking X6 (orientation) as one factor having 3 levels and X4 (Roof Area) with 4 levels.

Levels for X6 are 2, 3, 4 and 5.

Levels for X4 are 110.25, 122.5, 147, 220.50.

```
two_factor <- data[,c(4,6,9)]
two_factor$X6 <- factor(two_factor$X6)
two_factor$X4 <- factor(two_factor$X4)
head(two_factor[order(two_factor$X6),], n=50)
```

```
##         X4 X6    Y1
## 1   110.25  2 15.55
## 5    122.5  2 20.84
## 9      147  2 19.50
## 13     147  2 17.05
## 17     147  2 28.52
## 21   122.5  2 24.77
## 25   220.5  2  6.07
## 29   220.5  2  6.37
## 33   220.5  2  6.85
## 37   220.5  2  7.18
## 41   220.5  2 10.85
## 45   220.5  2  8.60
## 49  110.25  2 24.58
## 53   122.5  2 29.03
## 57     147  2 26.28
## 61     147  2 23.53
## 65     147  2 35.56
## 69   122.5  2 32.96
## 73   220.5  2 10.36
## 77   220.5  2 10.71
## 81   220.5  2 11.11
## 85   220.5  2 11.68
## 89   220.5  2 15.41
## 93   220.5  2 12.96
## 97  110.25  2 24.29
## 101  122.5  2 28.88
## 105    147  2 26.48
## 109    147  2 23.75
## 113    147  2 35.65
## 117  122.5  2 33.16
## 121  220.5  2 10.42
```

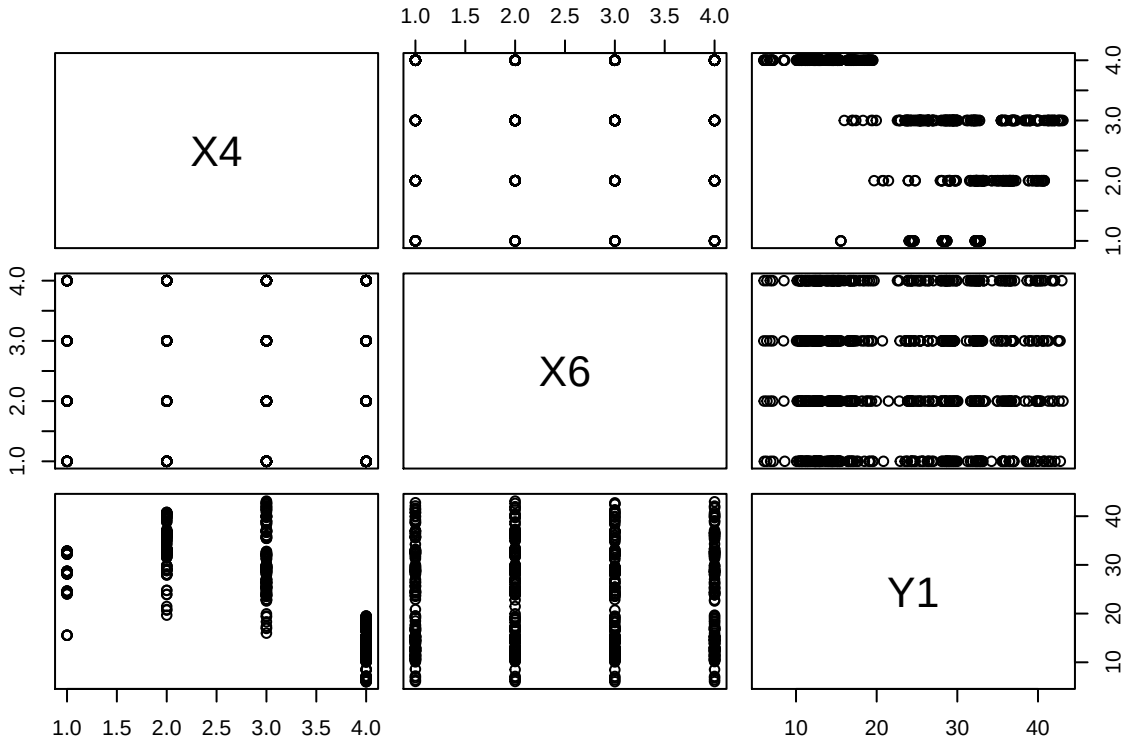
```
## 125  220.5  2 10.64
## 129  220.5  2 11.45
## 133  220.5  2 11.45
## 137  220.5  2 15.41
## 141  220.5  2 12.88
## 145 110.25  2 24.28
## 149  122.5  2 28.07
## 153    147  2 25.41
## 157    147  2 22.93
## 161    147  2 35.78
## 165  122.5  2 32.52
## 169  220.5  2 10.39
## 173  220.5  2 10.77
## 177  220.5  2 11.22
## 181  220.5  2 11.59
## 185  220.5  2 15.16
## 189  220.5  2 12.68
## 193 110.25  2 24.38
## 197  122.5  2 29.06
```

So now we have data ready for two factor analysis of variables. We can use this data frame now to carry out the factorial analysis.

3.1.2 Graphical Analysis

Comparision Plots

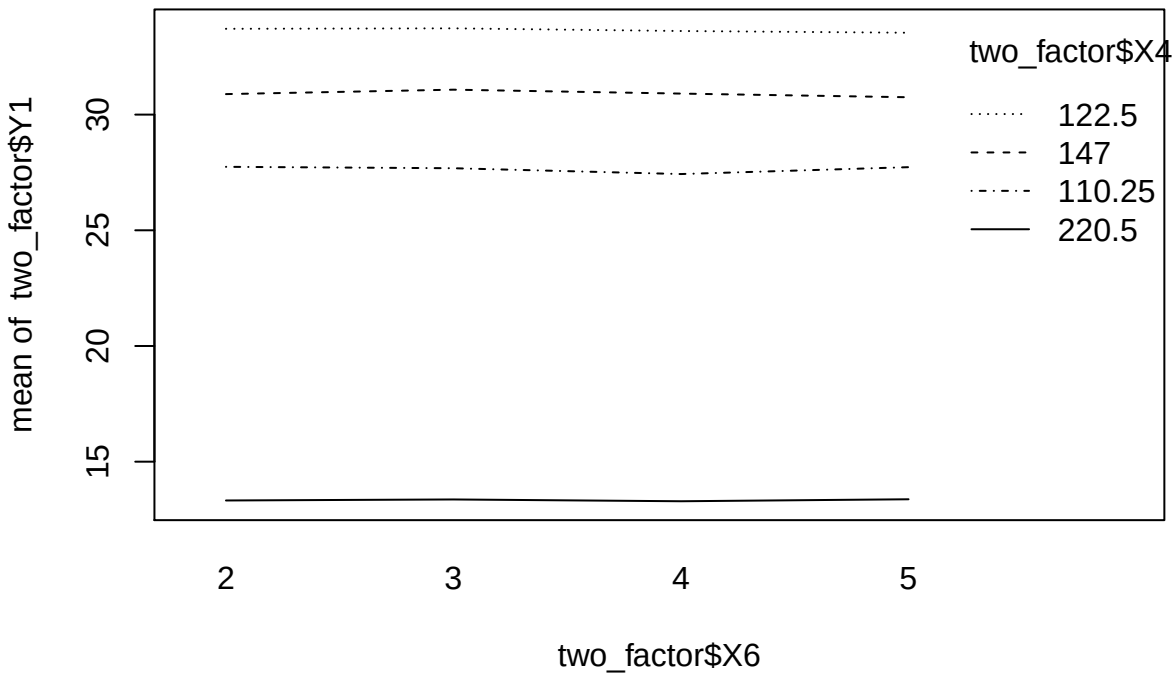
```
plot(two_factor)
```



The above graph will get us started with a little bit exploratory analysis for two factor analysis. Dependency of heat load on X6 (orentation of building) looks preety uniform, but we can clearly see that for lower roof area the heating load is comparatively higer as compared to higher roof area.

Interaction Plots

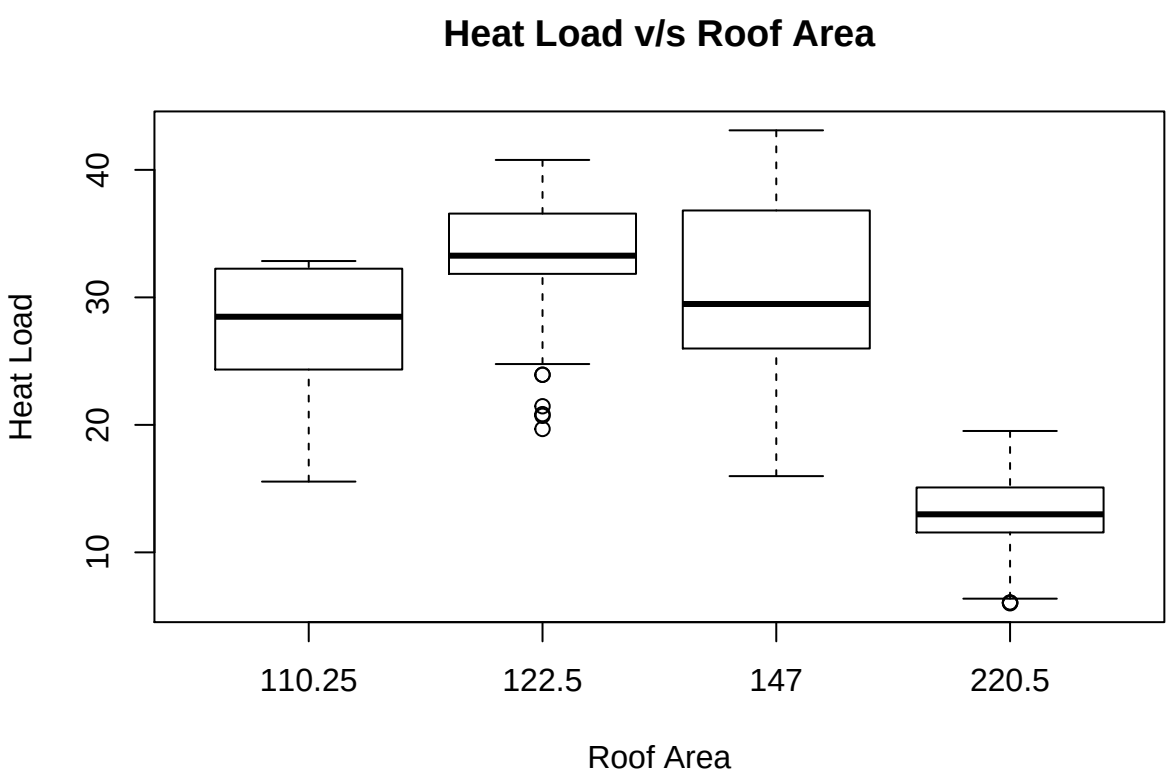
```
interaction.plot(two_factor$X6, two_factor$X4, two_factor$Y1)
```



Factor X4 is at four different level. The interaction plot shows that there is no interaction effect between the different levels of X4 on th dependent variable as the levels of other independent variable changes. There are four factor levels for both X4 and X6.

Box Plots

```
boxplot(Y1~X4, data=two_factor, main = "Heat Load v/s Roof Area", xlab = "Roof Area", ylab= "Heat Load"
```



We had already seen earlier during one factor design of experiment a box plot between heat load and orentation of building. Here we now analyze through boxplot variation of heat load with roof area.

We can clearly see that there is variation of heat load with the roof area. At different factor level of roof area there is different mean heat load and it drastically decreases for factor level of 220.5

3.1.3 Model Analysis

Two way ANOVA Model

The general model for two way anova is as follows:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

We will test 3 different hypothesis using the two-way anova:

(i) (a) $H_0^{'} : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$
(b) $H_0^{''} : \beta_1 = \beta_2 = \beta_3 = \beta_4$
(c) $H_0^{'''} : (\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{13} = ...(\alpha\beta)_{44}$

(ii) (a) At least one of α_i is not equal to zero.
(b) At least one of β_j is not equal to zero.
(c) At least one of $(\alpha\beta)_{ij}$ is not equal to zero.

–> The first hypothesis is to study whether there is significant difference in the mean heat load due different orientation of the building.

–> The second hypothesis is to study whether there is significant difference in the mean heat load due roof area of the building.

–> The third hypothesis is to study whether there is significant difference in the mean heat load due to interaction of roof area and orientation of the building.

```
analysis1 <- lm(Y1 ~ X6*X4, data = two_factor)
anova(analysis1)
```

```
## Analysis of Variance Table
##
## Response: Y1
##            Df Sum Sq Mean Sq   F value Pr(>F)
## X6          3      2     0.6    0.0284 0.9936
## X4          3  63365 21121.6 1079.0169 <2e-16 ***
## X6:X4       9      3     0.4    0.0179 1.0000
## Residuals 752  14720    19.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

–> From the above result for analysis of variance we see that p-value corrosponding to X6 is 0.9959 suggesting that we fail to to reject Ho.

–> For X4 the p-value is very small suggesting that we reject Ho and conclude that the means for different level of roof area are not equal and they have varying effect in the heat load.

–> From the above result for analysis of variance we see that p-value corrosponding to interaction of X4 and X6 is 0.9981 suggesting that we fail to to reject Ho.

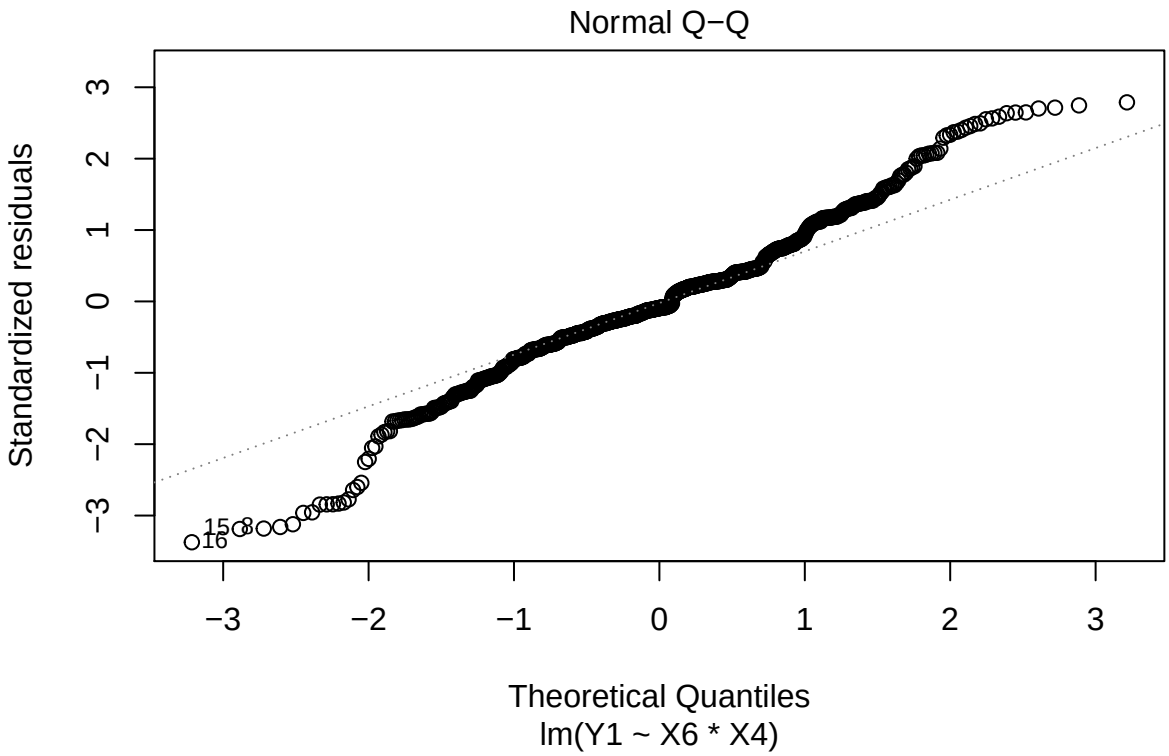
Apart from above conclusions we can derive following observations from the output above:

- SSA = 2
- SSB = 58001
- SS(AB) = 0
- SSE = 20086

As such there is no use of doing a multiple comparison test here as there is just one significant factor here i.e. X4. Doing multiple comparision test will not give us any significant inferences.

Checking Normality for two way ANOVA

```
plot(analysis1, which=2)
```



From above line we can see that majority of the observation follows the the dashed line sequence providing a proof for normality and suggesting that performing ANOVA would be a good decesion to use as an analysis tool.

Now we will extend our process to 3 factor design and analysis of experiment and see how the interaction amougst the variable at different treatment level varies in that case.

3.2 Three Factor design and Three way ANOVA

Now to further extend our analysis for three factor analysis we will take into consideration X7 (glazing area of building) in addition to X4 (roof area) and X6 (orientation)

3.2.1 Preparing data for three factor analysis

Levels for X6 are 2, 3, 4 and 5.

Levels for X4 are 110.25, 122.5, 147, 220.50.

Levels for X7 are 0.00, 0.10, 0.25, 0.40

```
three_factor <- data[,c(4,6,7,9)]
three_factor$X6 <- factor(three_factor$X6)
three_factor$X4 <- factor(three_factor$X4)
three_factor$X7 <- factor(three_factor$X7)
head(three_factor[order(three_factor$X6),], n=50)
```

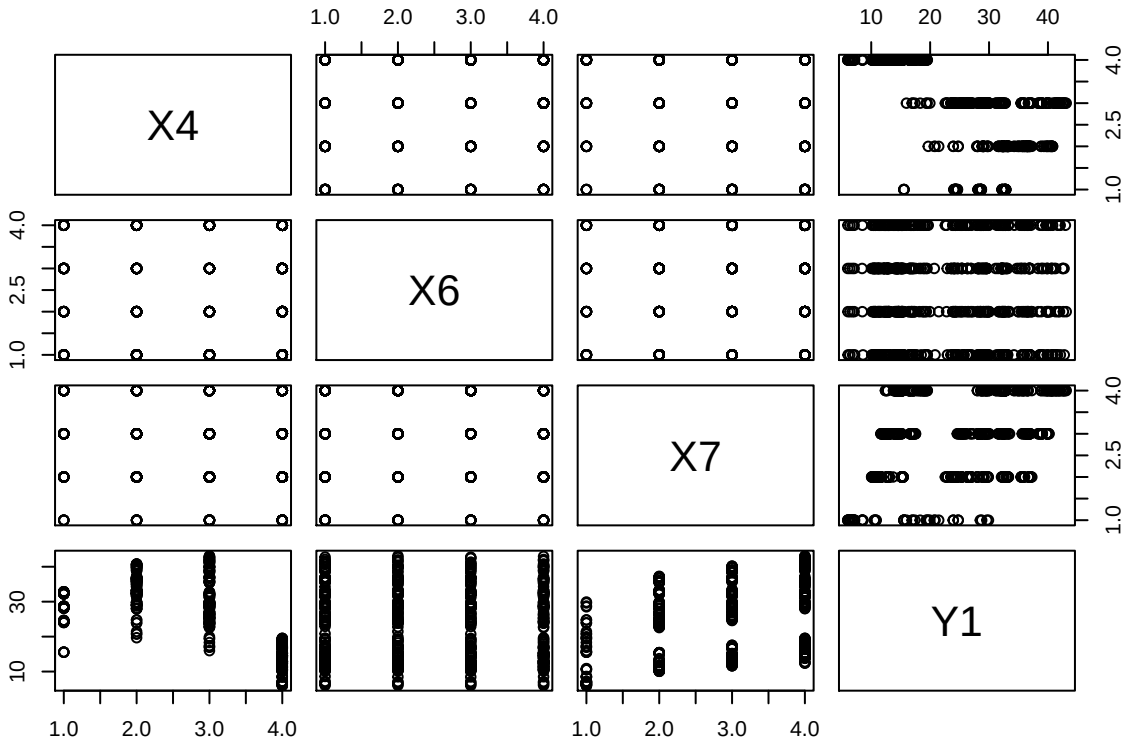
```
##         X4 X6  X7    Y1
## 1   110.25  2   0 15.55
## 5    122.5  2   0 20.84
## 9      147  2   0 19.50
## 13     147  2   0 17.05
## 17     147  2   0 28.52
## 21   122.5  2   0 24.77
## 25   220.5  2   0  6.07
## 29   220.5  2   0  6.37
## 33   220.5  2   0  6.85
## 37   220.5  2   0  7.18
## 41   220.5  2   0 10.85
## 45   220.5  2   0  8.60
## 49  110.25  2 0.1 24.58
## 53   122.5  2 0.1 29.03
## 57     147  2 0.1 26.28
## 61     147  2 0.1 23.53
## 65     147  2 0.1 35.56
## 69   122.5  2 0.1 32.96
## 73   220.5  2 0.1 10.36
## 77   220.5  2 0.1 10.71
## 81   220.5  2 0.1 11.11
## 85   220.5  2 0.1 11.68
## 89   220.5  2 0.1 15.41
## 93   220.5  2 0.1 12.96
## 97  110.25  2 0.1 24.29
## 101  122.5  2 0.1 28.88
## 105    147  2 0.1 26.48
## 109    147  2 0.1 23.75
## 113    147  2 0.1 35.65
## 117  122.5  2 0.1 33.16
## 121  220.5  2 0.1 10.42
## 125  220.5  2 0.1 10.64
## 129  220.5  2 0.1 11.45
## 133  220.5  2 0.1 11.45
## 137  220.5  2 0.1 15.41
## 141  220.5  2 0.1 12.88
## 145 110.25  2 0.1 24.28
## 149  122.5  2 0.1 28.07
## 153    147  2 0.1 25.41
## 157    147  2 0.1 22.93
## 161    147  2 0.1 35.78
## 165  122.5  2 0.1 32.52
## 169  220.5  2 0.1 10.39
## 173  220.5  2 0.1 10.77
## 177  220.5  2 0.1 11.22
## 181  220.5  2 0.1 11.59
## 185  220.5  2 0.1 15.16
## 189  220.5  2 0.1 12.68
## 193 110.25  2 0.1 24.38
## 197  122.5  2 0.1 29.06
```

So now we have data ready for three factor analysis of variables. We can use this data frame now to carry out the factorial analysis.

3.2.2 Graphical Analysis

Comparision Plots

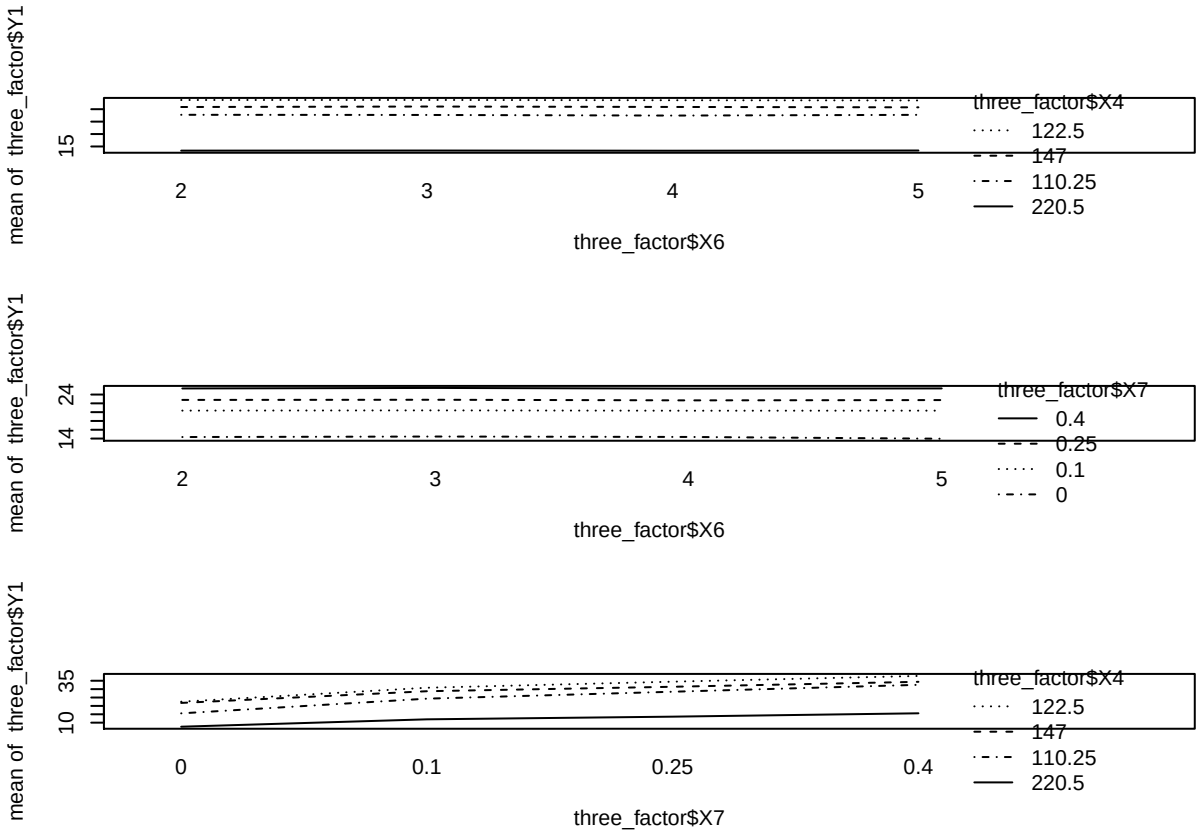
```
plot(three_factor)
```



The above graph will get us started with a little bit exploratory analysis for three factor analysis. There is noting special that can be observed here apart from already observed relation between parameters in case of two factor analysis. Though a new trend to observe would be mean increase in the heating load with increase in glazing area.

Interaction Plots

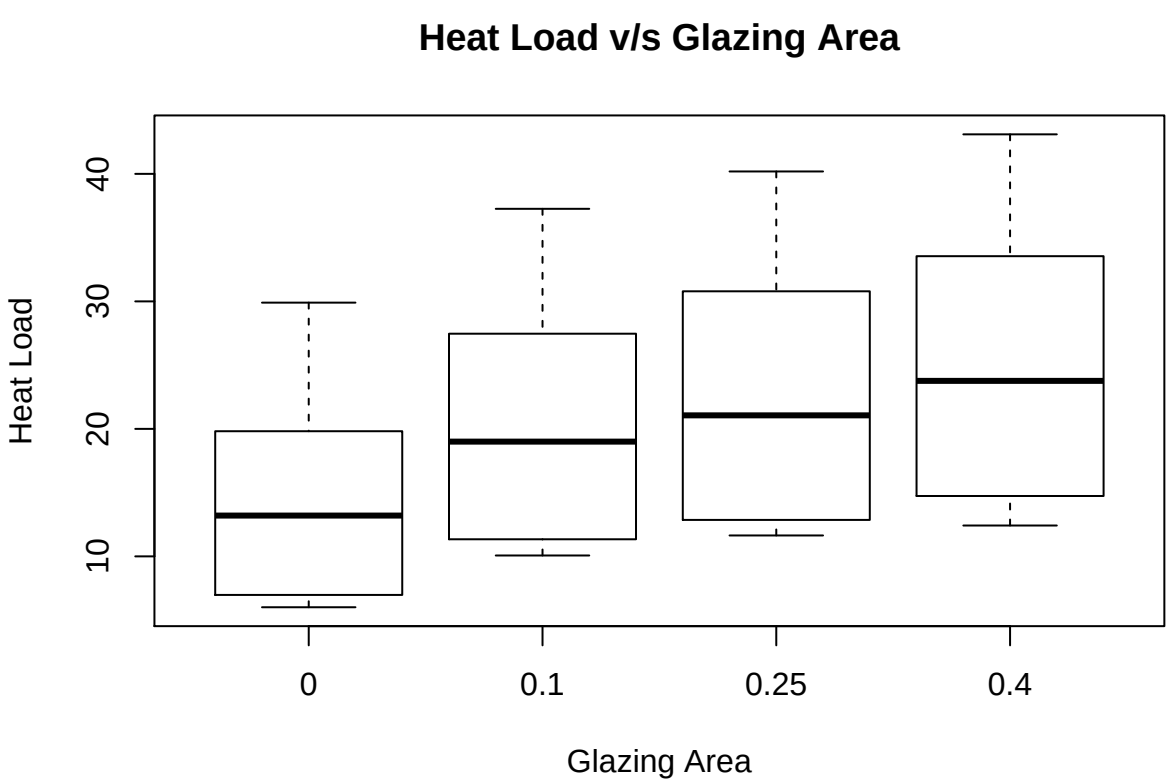
```
par(mfrow=c(3,1))
interaction.plot(three_factor$X6, three_factor$X4, three_factor$Y1)
interaction.plot(three_factor$X6, three_factor$X7, three_factor$Y1)
interaction.plot(three_factor$X7, three_factor$X4, three_factor$Y1)
```



The three interaction plots above shows that there is no interaction effect between the different levels of any treatment on the dependent variable as the treatment levels of other independent variable changes. The lines in all the graphs run parallely and does not show any sign of interaction. Graphically this means that factors emerging due to interaction of different main factors doesnot play a significant or no role at all on the heating load.

Box Plots

```
boxplot(Y1~X7, data=two_factor, main = "Heat Load v/s Glazing Area", xlab = "Glazing Area", ylab= "Heat
```



We had already seen earlier during one factor design of experiment a box plot between heat load with orentation and roof area of building. Here we now analyze through boxplot variation of heat load with glazing area.

We can clearly see that there is slight variation of heat load with the glazing area. At different factor level of glazing area there is different mean heat load and it slightly changes for every treatment, lowest being at 0 glazing area and maximum being at 0.4 glazing area.

3.1.3 Model Analysis

Three way ANOVA Model The general model for three way anova is as follows:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$$

We will test 7 different hypothesis using the two-way anova:

(i) (a) $H_0^{'} : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$
(b) $H_0^{''} : \beta_1 = \beta_2 = \beta_3 = \beta_4$
(c) $H_0^{'''} : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4$
(d) $H_0^{''''} : (\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{13} = ...(\alpha\beta)_{44}$
(e) $H_0^{'''''} : (\alpha\gamma)_{11} = (\alpha\gamma)_{12} = (\alpha\gamma)_{13} = ...(\alpha\gamma)_{44}$
(f) $H_0^{''''''} : (\beta\gamma)_{11} = (\beta\gamma)_{12} = (\beta\gamma)_{13} = ...(\beta\gamma)_{44}$
(g) $H_0^{'''''''} : (\alpha\beta\gamma)_{111} = (\alpha\beta\gamma)_{112} = (\alpha\beta\gamma)_{113} = ...(\alpha\beta\gamma)_{444}$

(ii) (a) $H_0^{'}$: At least one of α_i is not equal to zero.

(b) $H_0^{''}$: At least one of β_j is not equal to zero.
(c) $H_0^{'''}$: At least one of γ_k is not equal to zero.
(d) $H_0^{''''}$: At least one of $(\alpha\beta)_{ij}$ is not equal to zero.
(e) $H_0^{'''''}$: At least one of $(\alpha\gamma)_{jk}$ is not equal to zero.
(f) $H_0^{''''''}$: At least one of $(\beta\gamma)_{ik}$ is not equal to zero.
(g) $H_0^{'''''''}$: At least one of $(\alpha\beta\gamma)_{ijk}$ is not equal to zero.

Now we move on to prepare the ANOVA model to study the above hypothesis.

```
analysis2 <- lm(Y1 ~ X6*X4*X7, data = three_factor)
anova(analysis2)
```

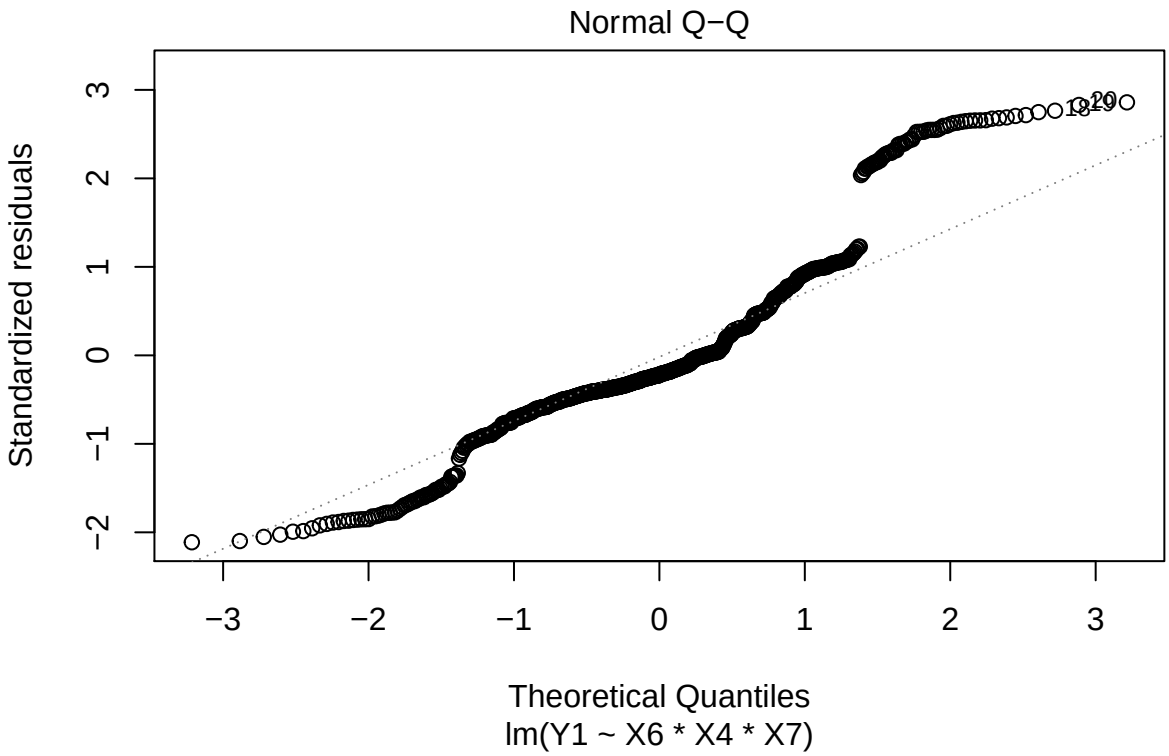
```
## Analysis of Variance Table
##
## Response: Y1
##             Df Sum Sq Mean Sq   F value    Pr(>F)
## X6           3      2     0.6    0.0506     0.985
## X4           3  63365 21121.6 1923.2314 < 2.2e-16 ***
## X7           3   6362  2120.7  193.1031 < 2.2e-16 ***
## X6:X4        9      3     0.4    0.0319     1.000
## X6:X7        9      2     0.2    0.0169     1.000
## X4:X7        9    623    69.2    6.2992 1.333e-08 ***
## X6:X4:X7    27      2     0.1    0.0076     1.000
## Residuals  704   7732    11.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

–> From the above output of ANOVA analysis we see that there are only 3 factors that are significantly affecting the heating load with change in treatment level. These factors are X4, X7 and interaction between X4 and X7. Other factors have very high p-value suggesting their negligible importance in heat load variance.

Checking Normality Condition

```
plot(analysis2, which=2)
```

```
## Warning: not plotting observations with leverage one:
##   2
```



From the above Q-Q plot we can say that in case of 3 factor design the normality is being followed for initial observations and later observations are being deviated from following the normality trend.

Multiple Comparision Test

(a) Tukey's Test

Since we rejected Ho here, now we will move on to perform mutiple comparision test to see that what factors tend to have a common mean condition and try to observe interaction of mean between various factors taken into consideration.

```
TT3 <- TukeyHSD(aov(analysis2))
head(TT3[[1]])
```

```
##            diff        lwr       upr     p adj
## 3-2  0.06781250 -0.8031892 0.9388142 0.9971457
## 4-2 -0.05296875 -0.9239705 0.8180330 0.9986317
## 5-2 -0.03750000 -0.9085017 0.8335017 0.9995122
## 4-3 -0.12078125 -0.9917830 0.7502205 0.9843907
## 5-3 -0.10531250 -0.9763142 0.7656892 0.9895340
## 5-4  0.01546875 -0.8555330 0.8864705 0.9999656
```

```
head(TT3[[4]])
```

```
##                          diff       lwr      upr        p adj
## 3:110.25-2:110.25 -0.061875 -4.090752 3.967002 1.000000e+00
## 4:110.25-2:110.25 -0.313125 -4.342002 3.715752 1.000000e+00
```

```
## 5:110.25-2:110.25 -0.015000 -4.043877 4.013877 1.000000e+00
## 2:122.5-2:110.25   5.964687  2.475578 9.453797 7.576015e-07
## 3:122.5-2:110.25   5.984062  2.494953 9.473172 6.785567e-07
## 4:122.5-2:110.25   5.868750  2.379640 9.357860 1.300493e-06
```

Considering 3 factors with 4 levels each there would be 12288 observations that whould be generated by applying this test. It is not useful neither feasible to study these many observations. A sample of output produced by test is depicted in the above data frame.

Other multiple comparsion test are generally useful while studying one factor analysis, therefore we won't be using those test here with 3 factor analyis.

Chapter 4

2^k Factorial Experiments and Analysis

4.1 Introduction

This is a 2^k (k=4 in this case) design which involves creation of a factorial design with exactly 2 levels.The reason behind using this approach is to keep experimental ???costs??? in control, which means that we need to take measurements (in an experiment) for a carefully choosen limited number of factor levels.This technique ensures that the main effects and low-order interaction effects can be estimated and tested, at the expense of high-order interactions.The scientific rationale behind this approach is that it is unlikely that there are significant complex interactions among various factors so we can assume that there are probably only main effects and a few low-order interactions. Now the first step in this analysis was to acquire data that fullfills this requirement. Since the data set available has multiple levels for each factor therefore we transform the data by finding the mean of the numerical values for each factor and then defining 2 levels (high and low) based on whether the value is above the average or below the average computed.

4.2 Prepration of Data

Let's do a initial ANOVA test and see what are the variables that we can choose to perform the 2^k factor analysis.

```
analysis3 <- lm(Y1 ~ X1*X2*X3*X4*X5*X6*X7*X8)
anova(analysis3)
```

```
## Analysis of Variance Table
##
## Response: Y1
##                    Df  Sum Sq Mean Sq    F value    Pr(>F)
## X1                  1 30238.2 30238.2 43827.2844 < 2.2e-16 ***
## X2                  1  8092.9  8092.9 11729.8193 < 2.2e-16 ***
## X3                  1 26144.8 26144.8 37894.3117 < 2.2e-16 ***
## X5                  1  1310.6  1310.6  1899.5693 < 2.2e-16 ***
## X6                  1     0.5     0.5     0.7572 0.3845117
## X7                  1  5686.0  5686.0  8241.3746 < 2.2e-16 ***
## X8                  1    73.1    73.1   106.0025 < 2.2e-16 ***
## X1:X2               1   468.3   468.3   678.7444 < 2.2e-16 ***
## X1:X3               1   242.4   242.4   351.2927 < 2.2e-16 ***
## X2:X3               1    50.2    50.2    72.8006 < 2.2e-16 ***
## X2:X4               1  3088.3  3088.3  4476.1998 < 2.2e-16 ***
## X3:X4               1   233.4   233.4   338.2324 < 2.2e-16 ***
## X1:X5               1   717.3   717.3  1039.6755 < 2.2e-16 ***
## X1:X6               1     1.1     1.1     1.5888 0.2079362
## X2:X6               1     0.1     0.1     0.1960 0.6581078
## X3:X6               1     0.0     0.0     0.0156 0.9007705
## X5:X6               1     0.1     0.1     0.0886 0.7660875
## X1:X7               1   419.4   419.4   607.9292 < 2.2e-16 ***
## X2:X7               1     0.2     0.2     0.2909 0.5898037
## X3:X7               1   135.8   135.8   196.7639 < 2.2e-16 ***
## X5:X7               1     9.0     9.0    13.0355 0.0003284 ***
## X6:X7               1     0.1     0.1     0.1297 0.7188841
```

```
## X1:X8              1    12.8    12.8    18.5455 1.906e-05 ***
## X2:X8              1     0.4     0.4     0.5523 0.4576443
## X3:X8              1     1.8     1.8     2.6649 0.1030551
## X5:X8              1     0.0     0.0     0.0009 0.9763869
## X6:X8              1     0.7     0.7     0.9430 0.3318634
## X7:X8              1   310.9   310.9   450.5916 < 2.2e-16 ***
## X1:X2:X3           1   345.1   345.1   500.1607 < 2.2e-16 ***
## X1:X2:X6           1     0.1     0.1     0.1075 0.7431277
## X1:X3:X6           1     0.1     0.1     0.1203 0.7288088
## X2:X3:X6           1     0.0     0.0     0.0029 0.9568003
## X2:X4:X6           1     1.2     1.2     1.6722 0.1964145
## X3:X4:X6           1     0.2     0.2     0.2411 0.6235757
## X1:X5:X6           1     0.1     0.1     0.0861 0.7693494
## X1:X2:X7           1     0.2     0.2     0.3380 0.5612061
## X1:X3:X7           1     4.7     4.7     6.8775 0.0089268 **
## X2:X3:X7           1     0.6     0.6     0.8762 0.3495736
## X2:X4:X7           1     0.4     0.4     0.5839 0.4450553
## X3:X4:X7           1     0.1     0.1     0.0746 0.7847973
## X1:X5:X7           1     1.5     1.5     2.1033 0.1474523
## X1:X6:X7           1     0.0     0.0     0.0046 0.9457418
## X2:X6:X7           1     0.0     0.0     0.0475 0.8276164
## X3:X6:X7           1     0.3     0.3     0.4628 0.4965656
## X5:X6:X7           1     0.2     0.2     0.2787 0.5977291
## X1:X2:X8           1     0.0     0.0     0.0002 0.9889222
## X1:X3:X8           1     0.0     0.0     0.0002 0.9896335
## X2:X3:X8           1     0.8     0.8     1.1568 0.2825101
## X2:X4:X8           1     0.2     0.2     0.2255 0.6350489
## X3:X4:X8           1     0.0     0.0     0.0047 0.9452495
## X1:X5:X8           1     0.0     0.0     0.0229 0.8798930
## X1:X6:X8           1     0.2     0.2     0.2521 0.6157966
## X2:X6:X8           1     0.5     0.5     0.6564 0.4181331
## X3:X6:X8           1     0.0     0.0     0.0399 0.8417808
## X5:X6:X8           1     0.1     0.1     0.1082 0.7423189
## X1:X7:X8           1     9.4     9.4    13.5649 0.0002490 ***
## X2:X7:X8           1     1.7     1.7     2.5209 0.1128172
## X3:X7:X8           1     5.7     5.7     8.2067 0.0043040 **
## X5:X7:X8           1     0.9     0.9     1.2538 0.2632296
## X6:X7:X8           1     0.4     0.4     0.6082 0.4357457
## X1:X2:X3:X6        1     0.0     0.0     0.0178 0.8940004
## X1:X2:X3:X7        1     0.3     0.3     0.4064 0.5240051
## X1:X2:X6:X7        1     0.1     0.1     0.1451 0.7033416
## X1:X3:X6:X7        1     0.1     0.1     0.1576 0.6915223
## X2:X3:X6:X7        1     0.1     0.1     0.1102 0.7400004
## X2:X4:X6:X7        1     0.2     0.2     0.2405 0.6239906
## X3:X4:X6:X7        1     0.0     0.0     0.0003 0.9854510
## X1:X5:X6:X7        1     0.0     0.0     0.0105 0.9184809
## X1:X2:X3:X8        1     0.2     0.2     0.2358 0.6273917
## X1:X2:X6:X8        1     0.0     0.0     0.0150 0.9025023
## X1:X3:X6:X8        1     0.1     0.1     0.1120 0.7380232
## X2:X3:X6:X8        1     0.3     0.3     0.3934 0.5307509
## X2:X4:X6:X8        1     2.8     2.8     4.0766 0.0438778 *
## X3:X4:X6:X8        1     0.4     0.4     0.5628 0.4534066
## X1:X5:X6:X8        1     1.1     1.1     1.5584 0.2123385
## X1:X2:X7:X8        1     0.2     0.2     0.3614 0.5479536
```

```
## X1:X3:X7:X8         1     0.7     0.7     1.0569 0.3042824
## X2:X3:X7:X8         1     0.9     0.9     1.3213 0.2507749
## X2:X4:X7:X8         1     0.5     0.5     0.6978 0.4038216
## X3:X4:X7:X8         1     0.7     0.7     1.0801 0.2990389
## X1:X5:X7:X8         1     0.7     0.7     0.9932 0.3193250
## X1:X6:X7:X8         1     0.0     0.0     0.0035 0.9526025
## X2:X6:X7:X8         1     0.0     0.0     0.0184 0.8920090
## X3:X6:X7:X8         1     0.0     0.0     0.0010 0.9746284
## X5:X6:X7:X8         1     0.0     0.0     0.0242 0.8765189
## X1:X2:X3:X6:X7      1     0.1     0.1     0.0962 0.7565533
## X1:X2:X3:X6:X8      1     0.1     0.1     0.1872 0.6654168
## X1:X2:X3:X7:X8      1     3.2     3.2     4.6651 0.0311351 *
## X1:X2:X6:X7:X8      1     0.0     0.0     0.0021 0.9633780
## X1:X3:X6:X7:X8      1     0.0     0.0     0.0005 0.9818169
## X2:X3:X6:X7:X8      1     0.3     0.3     0.4082 0.5230904
## X2:X4:X6:X7:X8      1     0.3     0.3     0.4184 0.5179653
## X3:X4:X6:X7:X8      1     0.0     0.0     0.0127 0.9102857
## X1:X5:X6:X7:X8      1     0.1     0.1     0.2023 0.6530457
## X1:X2:X3:X6:X7:X8   1     0.0     0.0     0.0229 0.8797769
## Residuals         672   463.6     0.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above we see as far as single main factors are considered X1, X2, X3, X5 and X8 play a significant role on the varinace of dependent variable i.e. (Y1: Heating Load)

We randomly select 4 factors from above 5 and will try to segregate them in two levels as described by the above procedure.

The factors that we choose for the analysis are as follows:

- X1 Relative Compactness (Initial number of levels: 12)
- X2 Surface Area (Initial number of levels: 12)
- X3 Wall Area (Initial number of levels: 7)
- X5 Overall Height (Initial number of levels: 2)

Extracting the needed coloumn out of the initial data frame

```
two_k <- data[,c(1,2,3,5,9)]
head(two_k)
```

```
##     X1    X2    X3 X5    Y1
## 1 0.98 514.5 294.0  7 15.55
## 2 0.98 514.5 294.0  7 15.55
## 3 0.98 514.5 294.0  7 15.55
## 4 0.98 514.5 294.0  7 15.55
## 5 0.90 563.5 318.5  7 20.84
## 6 0.90 563.5 318.5  7 21.46
```

The above data fram cntains four attributes that are of our interest and further analysis will be done on it. Now we will convert these two make all the treatment two level by process as described above.

4.2.1 Converting the dataset in two level form

```
colSums(two_k)/768
```

```
##          X1          X2          X3          X5          Y1
##   0.7641667 671.7083333 318.5000000   5.2500000  22.3072005
```

The mean value of each coloum in as described above. In every case if the observation is bigger than the mean value it would be assigned value "1" and if the value is smaller than or equal to the mean value it will be assigned value "-1".

```
two_k$X1[two_k$X1 <= 0.7641667] <- -1
two_k$X1[two_k$X1 > 0.7641667] <- 1
two_k$X2[two_k$X2 <= 671.7083333] <- -1
two_k$X2[two_k$X2 > 671.7083333] <- 1
two_k$X3[two_k$X3 <= 318.5000000] <- -1
two_k$X3[two_k$X3 > 318.5000000] <- 1
two_k$X5[two_k$X5 == 3.5] <- -1 #Since here we just have two levels initially
two_k$X5[two_k$X5 == 7.0] <- 1  #Since here we just have two levels initially
head(two_k)
```

```
##   X1 X2 X3 X5    Y1
## 1  1 -1 -1  1 15.55
## 2  1 -1 -1  1 15.55
## 3  1 -1 -1  1 15.55
## 4  1 -1 -1  1 15.55
## 5  1 -1 -1  1 20.84
## 6  1 -1 -1  1 21.46
```

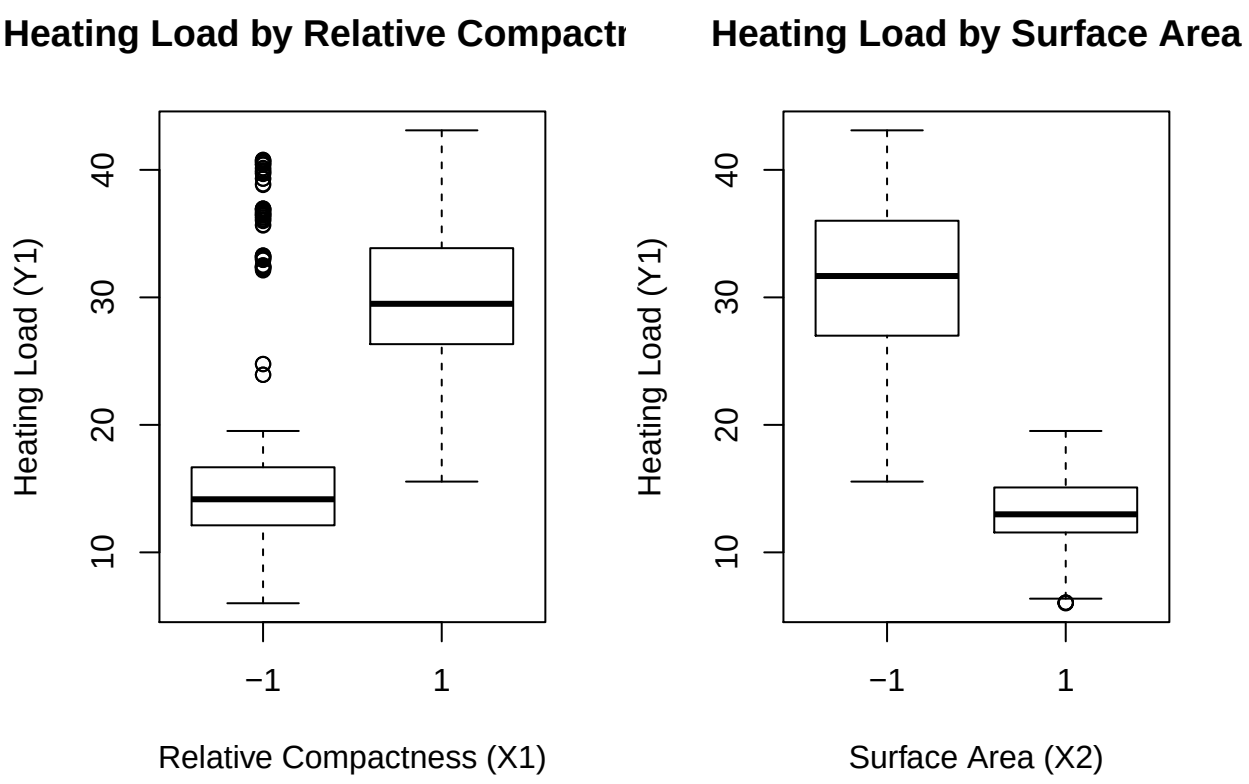
Now as seen above we have made our data ready to the 2^k factor analysis.

4.3 Graphical Analysis

Graphical Look at Factors taken into Consideration

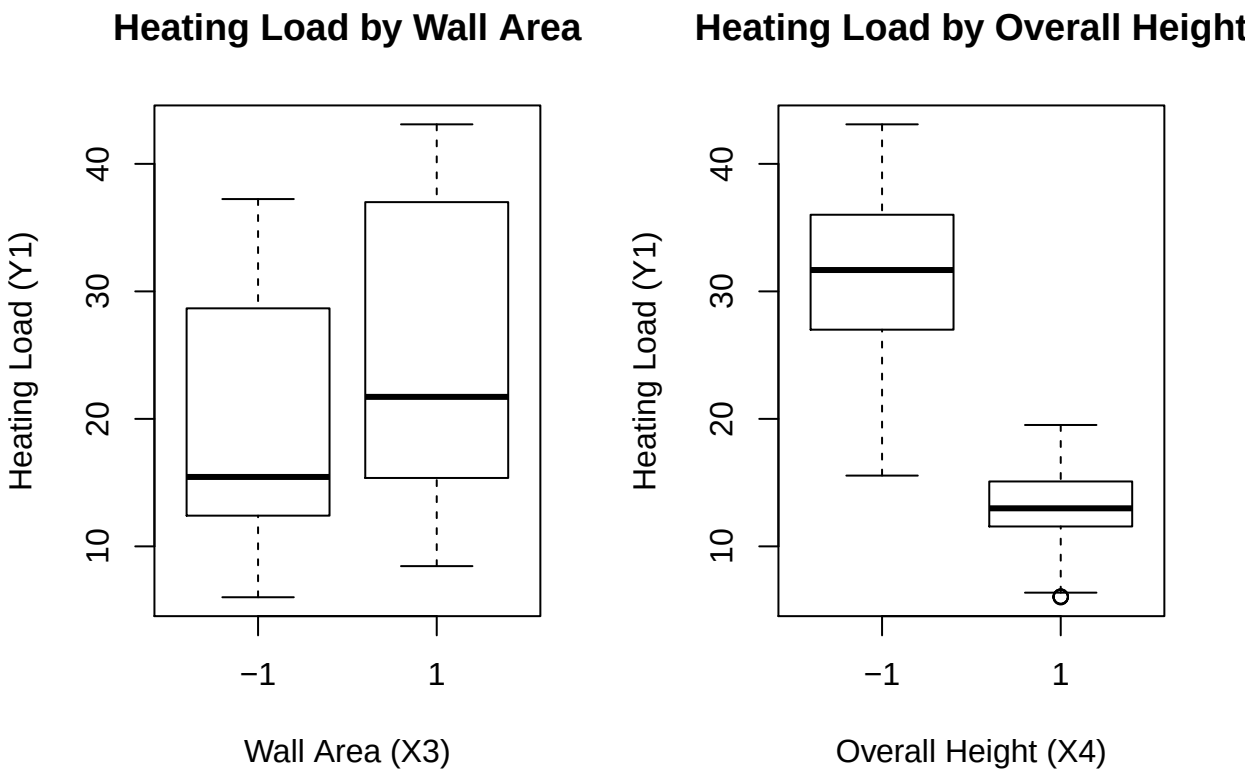
```
par(mfrow=c(1,2))
boxplot(Y1~X1, data=two_k, main="Heating Load by Relative Compactness",
        xlab="Relative Compactness (X1)",ylab="Heating Load (Y1)")

boxplot(Y1~X2, data=two_k, main="Heating Load by Surface Area",
        xlab="Surface Area (X2)",ylab="Heating Load (Y1)")
```




```
boxplot(Y1~X3, data=two_k, main="Heating Load by Wall Area",
        xlab="Wall Area (X3)",ylab="Heating Load (Y1)")

boxplot(Y1~X2, data=two_k, main="Heating Load by Overall Height",
        xlab="Overall Height (X4)",ylab="Heating Load (Y1)")
```



There are some outliers in all the factor plots, the number of outliers in case of "Relative Compactness" factor are way more. In all cases there is significant difference in the mean at both levels.

4.4 Design of Experiment and Analysis for 2^k Factor Analysis

Since we have 4 factors all at two levels, we will try to see the number and type of test that we have to perform.

```
library(AlgDesign)
FFA <- gen.factorial(levels = 2, nVars = 4, center = TRUE, varNames = c("X1", "X2", "X3", "X5"))
FFA
```

```
##    X1 X2 X3 X5
## 1  -1 -1 -1 -1
## 2   1 -1 -1 -1
## 3  -1  1 -1 -1
## 4   1  1 -1 -1
## 5  -1 -1  1 -1
## 6   1 -1  1 -1
## 7  -1  1  1 -1
## 8   1  1  1 -1
## 9  -1 -1 -1  1
## 10  1 -1 -1  1
## 11 -1  1 -1  1
## 12  1  1 -1  1
```

```
## 13 -1 -1  1  1
## 14  1 -1  1  1
## 15 -1  1  1  1
## 16  1  1  1  1
```

So as we see above we have the test matrix above. Total of 16 runs are requried for the full factorial analysis.

4.4.1 ANOVA Analysis

```
two.k <- lm(Y1~X1*X2*X3*X5, data=two_k)
anova(two.k)
```

```
## Analysis of Variance Table
##
## Response: Y1
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X1          1  35912   35912 3190.75 < 2.2e-16 ***
## X2          1  27343   27343 2429.39 < 2.2e-16 ***
## X3          1   4636    4636  411.92 < 2.2e-16 ***
## X1:X3       1   1611    1611  143.16 < 2.2e-16 ***
## Residuals 763   8588      11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Out of the factors in the above anova model only X1, X2, X3 and interaction between X1 and X3 tend to be significant. It can be observed from their low p-values. It means relative compactness, surface area and wall area along with interaction of relative compactness and wall area tend to affect the heating load the most. The influence on heating load due to the fourth factor taken into consideration i.e. the overall height of the building is either very less or is alredy been taken into consideration due to factors X1, X2, X3 and interaction between X1 and X3.

Now, we will try to look at the regression equation to estimate the heating load, which will mostly be built out of the factors as found out through analysis of variance.

4.4.2 Regression Model

```
summary(two.k)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 * X2 * X3 * X5, data = two_k)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7958  -1.7803   0.2331   2.1294   8.8942
##
## Coefficients: (11 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.6715     0.1483 159.657   <2e-16 ***
```

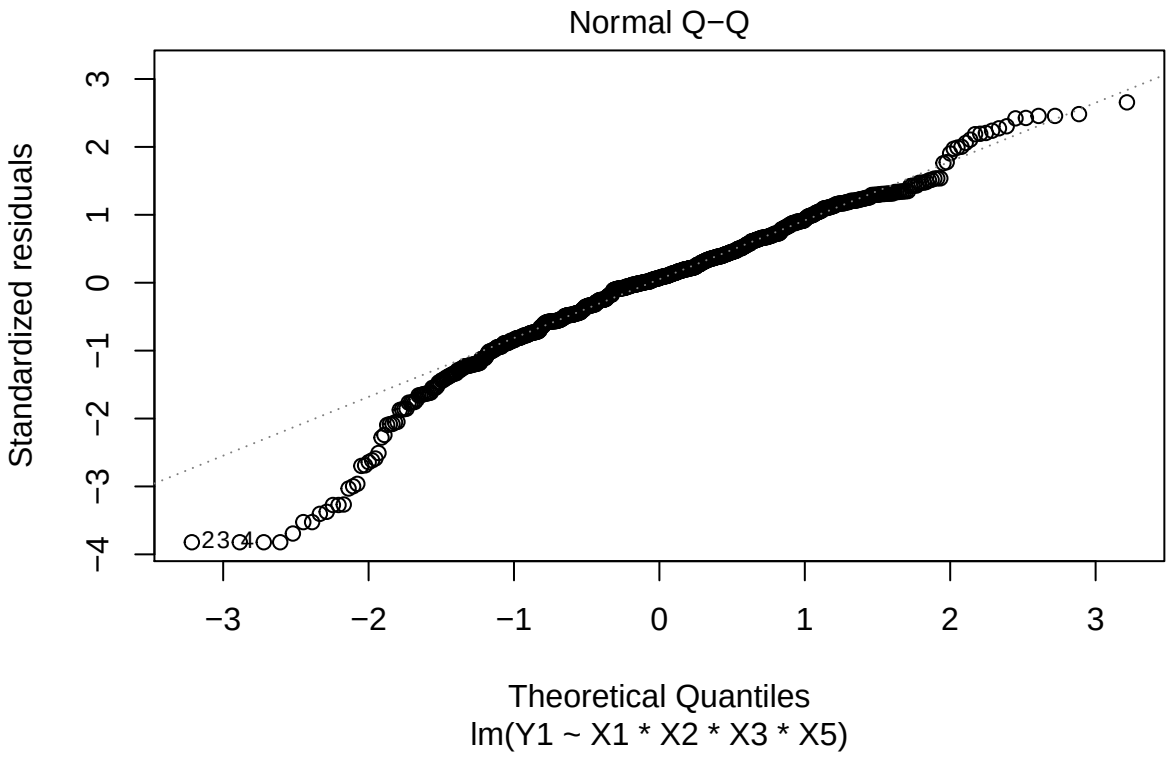
```
## X1           -0.3021     0.2568  -1.177     0.24
## X2          -10.1075     0.2568 -39.359   <2e-16 ***
## X3            3.3571     0.1483  22.643   <2e-16 ***
## X5                NA         NA      NA       NA
## X1:X2             NA         NA      NA       NA
## X1:X3          1.7740     0.1483  11.965   <2e-16 ***
## X2:X3             NA         NA      NA       NA
## X1:X5             NA         NA      NA       NA
## X2:X5             NA         NA      NA       NA
## X3:X5             NA         NA      NA       NA
## X1:X2:X3          NA         NA      NA       NA
## X1:X2:X5          NA         NA      NA       NA
## X1:X3:X5          NA         NA      NA       NA
## X2:X3:X5          NA         NA      NA       NA
## X1:X2:X3:X5       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.355 on 763 degrees of freedom
## Multiple R-squared:   0.89,  Adjusted R-squared:  0.8895
## F-statistic:  1544 on 4 and 763 DF,  p-value: < 2.2e-16
```

The regression model from above is:

Y1 = 23.6715 - 0.3021X1 - 10.1075X2 + 3.3571X3 + 1.7740X1X3

It is based on the factors that we got out of performing anova test earlier on the model. The regression equation is way of represting the heating load in terms of the important factors that we got out of the factor analysis or anova test.

```
plot(two.k, which = 2)
```



From the Q-Q plot above we can be sure about our results because the normality condition seem to be satisfied pretty well by it.

We can remove some variables and see different models but since we just have four factors under consideratin, doing that would not yield anything useful.

The last thing that we would like to calculate is the confidence interval for the coffecients in regression model.

```
confint(two.k)
```

```
##                    2.5 %     97.5 %
## (Intercept)  23.3804876 23.9625984
## X1           -0.8062517  0.2019939
## X2          -10.6115837 -9.6033382
## X3            3.0660344  3.6481453
## X5                   NA         NA
## X1:X2                NA         NA
## X1:X3         1.4829485  2.0650593
## X2:X3                NA         NA
## X1:X5                NA         NA
## X2:X5                NA         NA
## X3:X5                NA         NA
## X1:X2:X3             NA         NA
## X1:X2:X5             NA         NA
## X1:X3:X5             NA         NA
## X2:X3:X5             NA         NA
## X1:X2:X3:X5          NA         NA
```

The above table shows us the confidence interval for various coffecients considered in the regression model.

Chapter 5

Conclusion

Now in this section I would like to give a brief summary of the things that we covered as part of this project:

- We understood the dataset and defined objectives for the project.
- We choose one factor out of 8 given factors to start with the one factor analysis.
- Graphical analysis and ANOVA testing was carried out to do the analysis for one factor design.
- Following the map struture of one factor analysis, analysis was done in simalar fashion for two and three factor experiments.
- Lastly we ended up doing $ 2^k $ factorial analysis strting with converting the dataset to appropriate form, doing graphical analysis followed by model analysis (ANOVA testing and regression analysis).

NOTE: We did not included the concept of blocking in our project because we are not sure how the data was collected. Blocking is ususally done to reduce or subdue the effect of higly varying observations in data if it occoured due to differentiating factors that can influence the data collected (for example: if the data was collected during different times of day). We could have displayed the idea of blocking by just radomly distributing data in blocks without specifying the creteria or basis on which we are forming them, but the sole concept of blocking would have been lost. That's why we did'nt included it.