final_project

December 6, 2020

Statistical Data Analysis - Project

0.0.1 Group Members:

1. Ashutosh Chauhan (S20180010017)
2. Pradum Singh (S20180010136)
3. Utkarsh Ajay Aditya (S20180010182)
4. Vipul Rawat (S20180010192)

Contents

1 Abstract

This study looked into assessing the heating load and cooling load requirements of buildings (that is, energy efficiency) as a function of building parameters. The study included EDA of the dataset, univariate and multivariate analysis of the parameters and regression model for the dataset. The conclusion includes the preffered values of the given parameters which would help in having a lower heating and cooling load requirements of buildings.

2 Introduction

2.1 Dataset Source

The dataset was created by Angeliki Xifara (angxifara '@' gmail.com, Civil/Structural Engineer) and was processed by Athanasios Tsanas (tsanasthanasis '@' gmail.com, Oxford Centre for Industrial and Applied Mathematics, University of Oxford, UK).

2.2 Dataset Information

The dataset has energy analysis using 12 different building shapes simulated in Ecotect. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. It consists of simulations with various settings as functions of the afore-mentioned characteristics to obtain 768 building shapes. The dataset comprises 768 samples and 8 features, aiming to predict two real valued responses. It can also be used as a multi-class classification problem if the response is rounded to the nearest integer.

2.3 Attribute Information

The dataset contains eight attributes (or features, denoted by X1...X8) and two responses (or outcomes, denoted by y1 and y2). The aim is to use the eight features to predict each of the two responses.

Specifically:

Symbol	Parameter	Description
X1	Relative Compactness	It is the measure of compactness of the closure or building
X2	Surface Area	Surface area of the Building
X3	Wall Area	area of the building covered by width of the wall.
X4	Roof Area	Area covered under roofs.
X5	Overall Height	Overall height of building
X6	Orientation	Orientation of building based on direction like (North facing, South facing and others)
X7	Glazing Area	means the total area of the wall which is glass
X8	Glazing Area Distribution	How Glazing Area is distributed within the whole building.

Symbol	Parameter	Description
y1	Heating Load	How much heating load is required to heat the building.
y2	Cooling Load	How much load is required to cool the building.

2.4 Tasks

1. EDA
2. Univariate Analysis of parameters
3. Bivariate Analsis of parameters
4. Regression Analysis

3 Methodology

3.1 EDA

Understanding and analyzing the given data set by summarizing their main characteristics, by plotting them visually.

3.1.1 EDA Objectives

1. We have to study the paramenters which has the most effect on heating load requirements of buildings.
2. We have to study the paramenters which has the most effect on cooling load requirements of buildings.
3. To find the relation between the parameters and its combined effect on the heating load and cooling load requirements of buildings.

3.1.2 Preliminary Data Processing

A simple preview and statistics of the dataset under study.

```
Parameter Names

X1 :  Relative Compactness
X2 :  Surface Area
X3 :  Wall Area
X4 :  Roof Area
X5 :  Overall Height
X6 :  Orientation
X7 :  Glazing Area
X8 :  Glazing Area Distribution
Y1 :  Heating Load
Y2 :  Cooling Load


<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 768 entries, 0 to 767
Data columns (total 10 columns):
X1    768 non-null float64
X2    768 non-null float64
X3    768 non-null float64
X4    768 non-null float64
X5    768 non-null float64
X6    768 non-null int64
X7    768 non-null float64
X8    768 non-null int64
Y1    768 non-null float64
Y2    768 non-null float64
dtypes: float64(8), int64(2)
memory usage: 60.1 KB
```

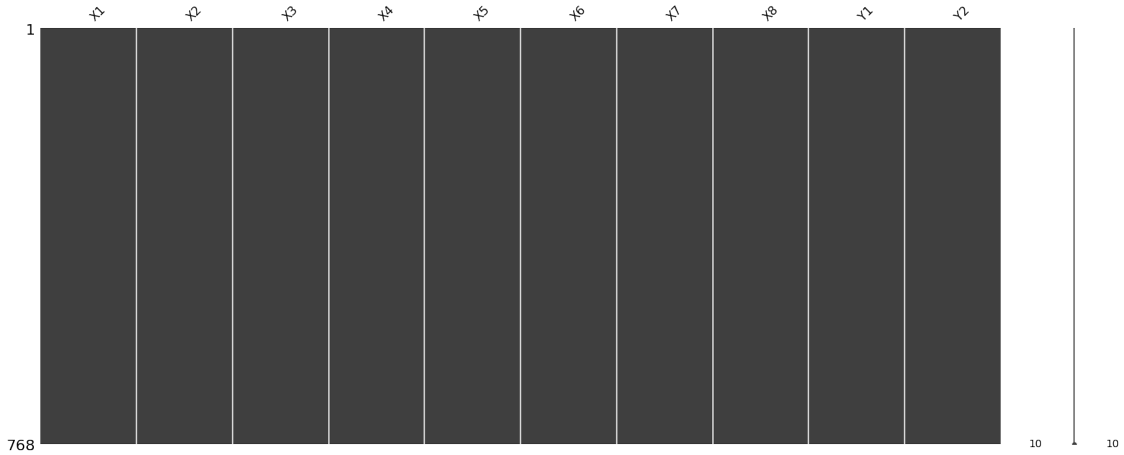
```
[6]:      X1      X2      X3      X4    X5  X6    X7  X8     Y1     Y2
     0 0.980 514.500 294.000 110.250 7.000   2 0.000   0 15.550 21.330
     1 0.980 514.500 294.000 110.250 7.000   3 0.000   0 15.550 21.330
     2 0.980 514.500 294.000 110.250 7.000   4 0.000   0 15.550 21.330
     3 0.980 514.500 294.000 110.250 7.000   5 0.000   0 15.550 21.330
     4 0.900 563.500 318.500 122.500 7.000   2 0.000   0 20.840 28.280
```

The output shows that we have 768 entries with 10 columns. Out of all the parameters, “Orientation” and “Glazing Area Distribution” are of integer type, rest of the parameters are of float type. The dataset parameters are in correct format.

3.1.3 Missing Values and Duplicated Entries

Checking for missing values and duplicated entries in the given dataset.

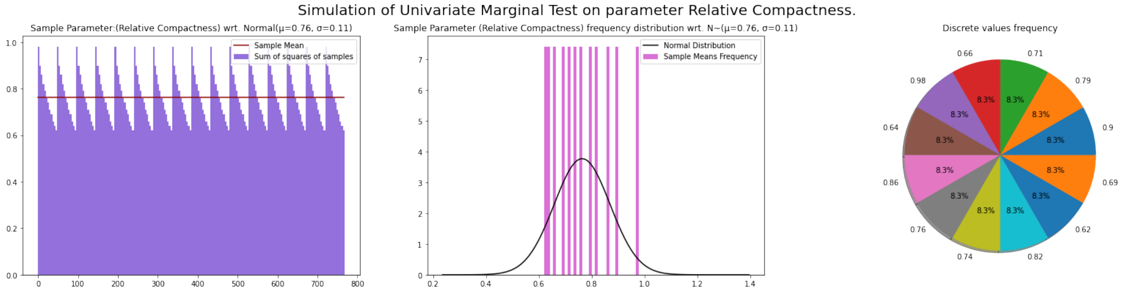
```
No duplicated entries(rows) found.

Preview of data with null values:
Empty DataFrame
Columns: [X1, X2, X3, X4, X5, X6, X7, X8, Y1, Y2]
Index: []
```



As there are no white spaces in the graph, none of the values are missing.

3.1.4 Uni-variate Analysis of all the parameters

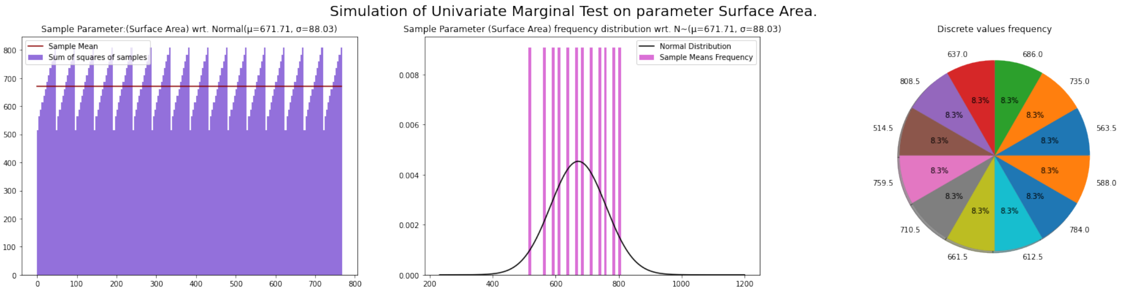
1. Relative Compactness(X1)



```
No.of.unique values : 12
Max Value:  0.98        Min Value:  0.62
Mean :  0.7641666666666677      Variance : 0.011188874402433754
```

X1: Relative Compactness The relative compactness has 12 distinct values each with the same probability of 0.8. The minimum and maximum values for the parameter are 0.62 and 0.98 respectively. The distribution seems to be showing a Uniform Distribution. The mean is very close to the average value of the max and min values.This parameter has a low variance of 0.011.

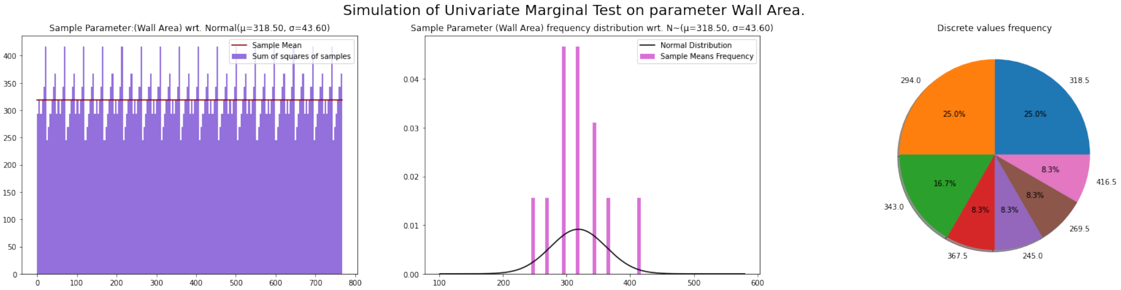
2. Surface Area(X2)



```
No.of.unique values : 12
Max Value:  808.5       Min Value:  514.5
Mean :  671.7083333333334       Variance : 7759.163841807892
```

X2: Surface Area The surface area has 12 distinct values each with the same probability of `0.8`. Same as that of Relative Compactness(X1)
The minimum and maximum values for the parameter are `808.5` and `514.5` respectively. The distribution seems to be showing a Uniform Distribution. The mean is very close to the average value of the max and min values. This parameter has a high variance of `7759.16`.

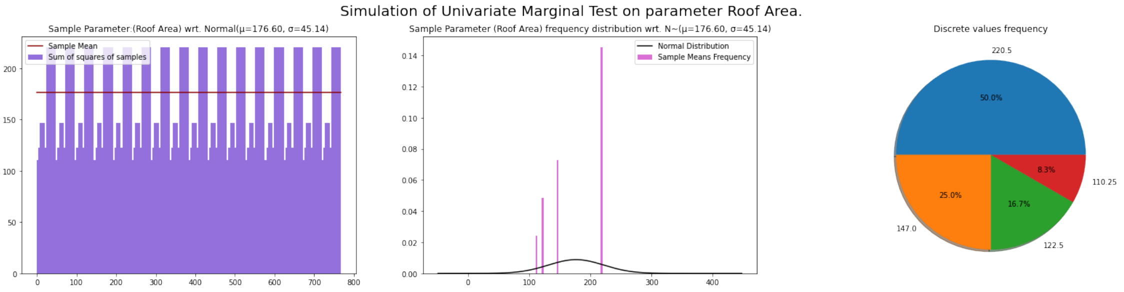
3. Wall Area(X3)



```
No.of.unique values : 7
Max Value:  416.5       Min Value:  245.0
Mean :  318.5   Variance : 1903.2698826597132
```

X3: Wall Area The wall area has 7 distinct values among which 2 of them have a probability of `0.25` .
The minimum and maximum values for the parameter are `416.5` and `245.0` respectively. The distribution seems to be showing a normal Distribution. The mean is very close to the average value of the max and min values and most of the values are coming from around the mean. This parameter has a high variance of `1903.27`.

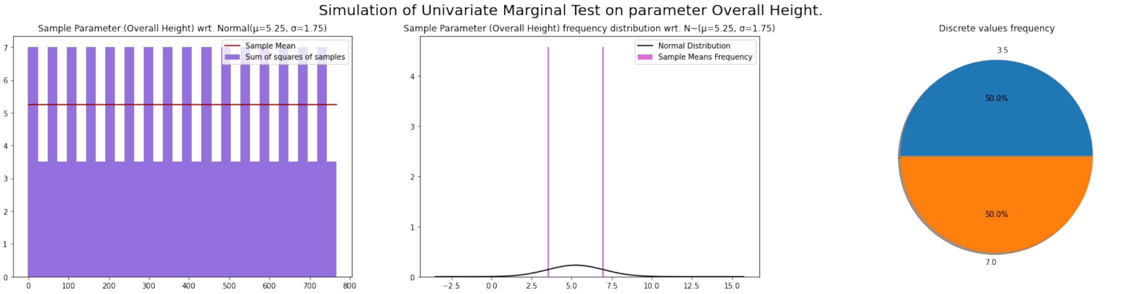
4. Roof Area(X4)



```
No.of.unique values : 4
Max Value:  220.5       Min Value:  110.25
Mean :  176.60416666666666      Variance : 2039.9630595393185
```

X4: Roof Area The roof area has only 4 distinct values, out of which 1 of them has a probability of 0.50 .
The minimum and maximum values for the parameter are 220.5 and 110.25 respectively. The mean is close to the average value of the max and min values. This parameter has a high variance of 2039.96.

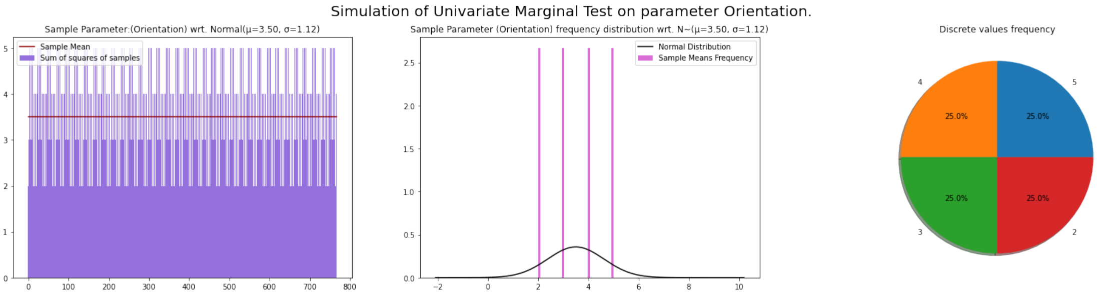
5. Overall Height(X5)



```
No.of.unique values : 2
Max Value:  7.0 Min Value:  3.5
Mean :  5.25    Variance : 3.0664928292046936
```

X5: Overall Height The Overall Height has only 2 distinct values, both of them having a probability of 0.50 each.
The minimum and maximum values for the parameter are 7.0 and 3.5 respectively. It seems to follow Uniform distribution. The mean is equal to the average value of the max and min values. This parameter has a low variance of 3.066, since we only have two values.

6. Orientation(X6)

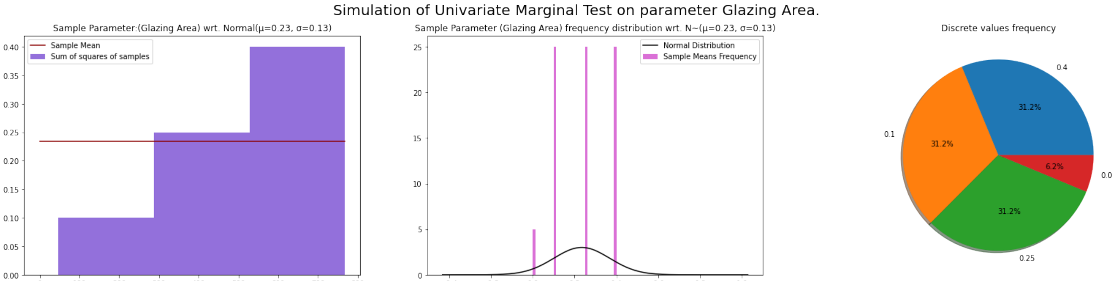


```
No.of.unique values : 4
Max Value:  5   Min Value:  2
Mean :  3.5     Variance : 1.2516297262059974
```

X6: Orientation The Orientation has only 4 distinct values, each of them having a probability of 0.25.

The minimum and maximum values for the parameter are 5 and 2 respectively. It seems to follow Uniform distribution. The mean is equal to the average value of the max and min values. This parameter has a low variance of 1.25, since we only have four close values.

7. Glazing Area(X7)

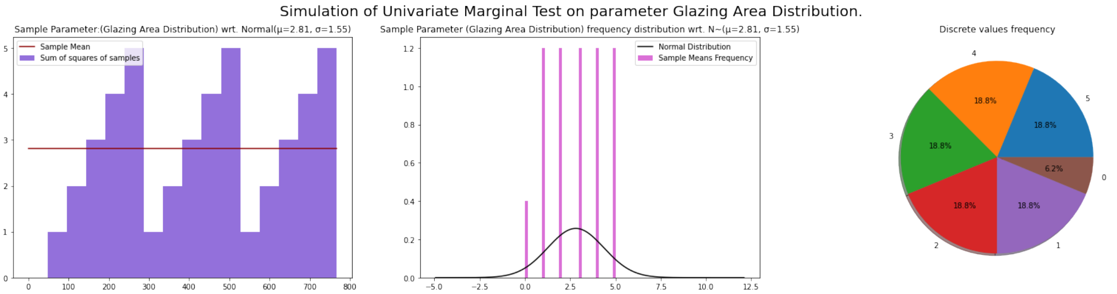


```
No.of.unique values : 4
Max Value:  0.4 Min Value:  0.0
Mean :  0.23437500000000186     Variance : 0.017747718383311874
```

X7: Glazing Area The Glazing Area has only 4 distinct values, three of them having a probability of 0.31 each.

The minimum and maximum values for the parameter are 0.4 and 0.0 respectively. The mean is close to the average value of the max and min values. This parameter has a very low variance of 0.017, since we only have four close values.

8. Glazing Area Distribution(X8)



```
No.of.unique values : 6
Max Value:  5   Min Value:  0
Mean :  2.8125  Variance : 2.405475880052151
```

X8: Glazing Area Distribution The Glazing Area Distribution has only 6 distinct values, five of them having a probability of `0.19` each.

The minimum and maximum values for the parameter are `5` and `0` respectively. The mean is close to the average value of the max and min values. This parameter has a very low variance of `2.40`, since we only have five close values.

3.1.5 Bi-variate Data Analysis

Mean Vector

```
[21]: X1      0.764
      X2    671.708
      X3    318.500
      X4    176.604
      X5      5.250
      X6      3.500
      X7      0.234
      X8      2.812
      dtype: float64
```

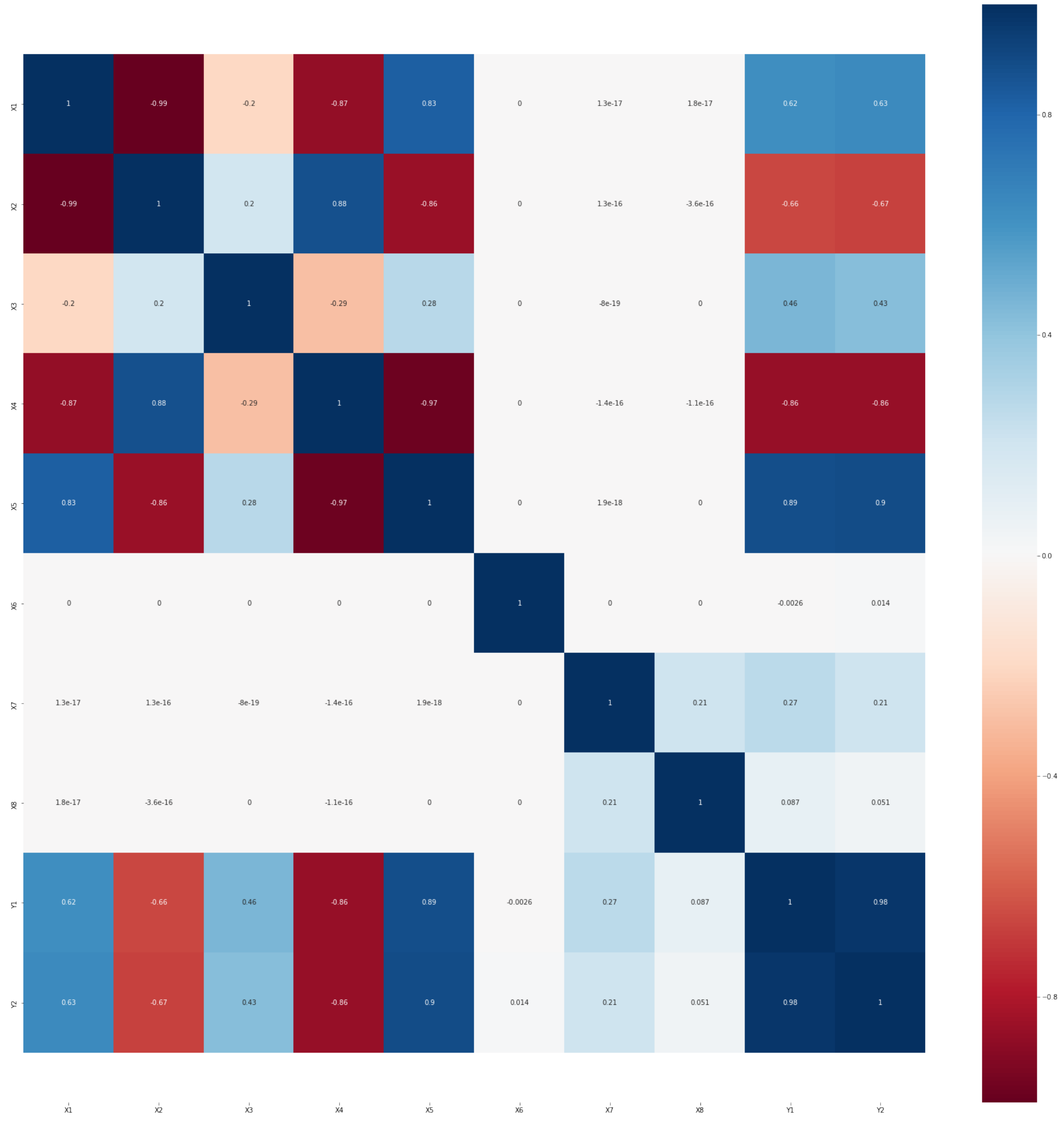
Covariance Matrix

```
[22]:        X1        X2        X3        X4       X5     X6     X7     X8
      X1  0.011    -9.242    -0.940    -4.151    0.153  0.000  0.000  0.000
      X2 -9.242 7,759.164   751.291 3,503.937 -132.370  0.000 -0.000 -0.000
      X3 -0.940   751.291 1,903.270  -575.990   21.465  0.000  0.000  0.000
      X4 -4.151 3,503.937  -575.990 2,039.963  -76.918  0.000 -0.000  0.000
      X5  0.153  -132.370    21.465   -76.918    3.066  0.000  0.000  0.000
      X6  0.000     0.000     0.000     0.000    0.000  1.252 -0.000  0.000
      X7  0.000    -0.000     0.000    -0.000    0.000 -0.000  0.018  0.044
      X8  0.000    -0.000     0.000     0.000    0.000  0.000  0.044  2.405
```

Correlation Matrix and Heat map

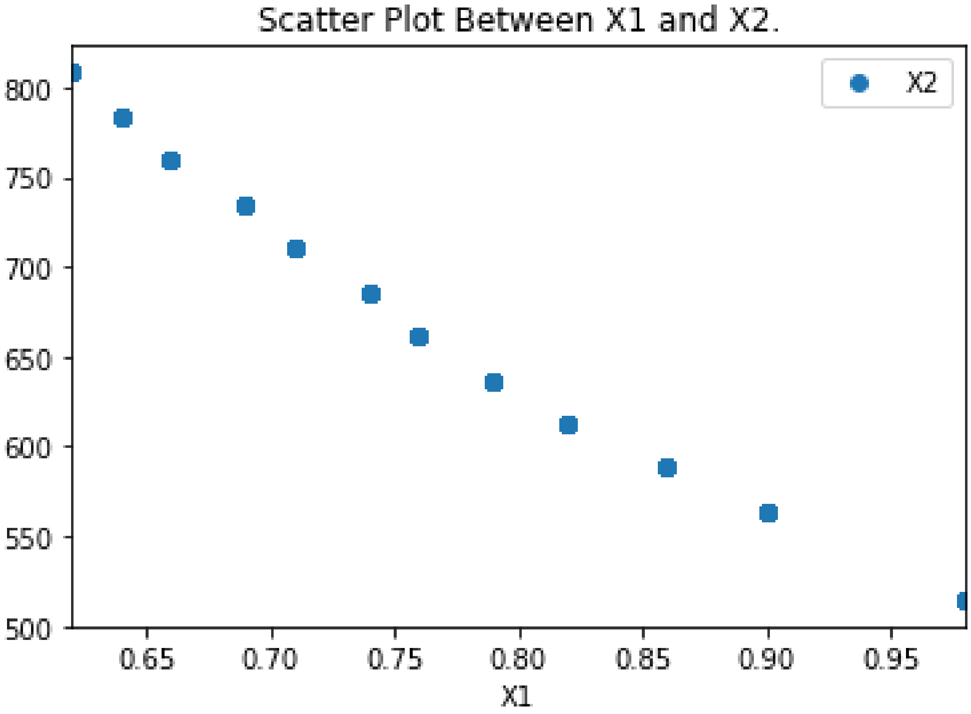
```
       X1     X2     X3     X4     X5     X6     X7     X8     Y1     Y2
X1  1.000 -0.992 -0.204 -0.869  0.828  0.000  0.000  0.000  0.622  0.634
X2 -0.992  1.000  0.196  0.881 -0.858  0.000  0.000 -0.000 -0.658 -0.673
X3 -0.204  0.196  1.000 -0.292  0.281  0.000 -0.000  0.000  0.456  0.427
X4 -0.869  0.881 -0.292  1.000 -0.973  0.000 -0.000 -0.000 -0.862 -0.863
X5  0.828 -0.858  0.281 -0.973  1.000  0.000  0.000  0.000  0.889  0.896
X6  0.000  0.000  0.000  0.000  0.000  1.000  0.000  0.000 -0.003  0.014
X7  0.000  0.000 -0.000 -0.000  0.000  0.000  1.000  0.213  0.270  0.208
X8  0.000 -0.000  0.000 -0.000  0.000  0.000  0.213  1.000  0.087  0.051
Y1  0.622 -0.658  0.456 -0.862  0.889 -0.003  0.270  0.087  1.000  0.976
Y2  0.634 -0.673  0.427 -0.863  0.896  0.014  0.208  0.051  0.976  1.000
```

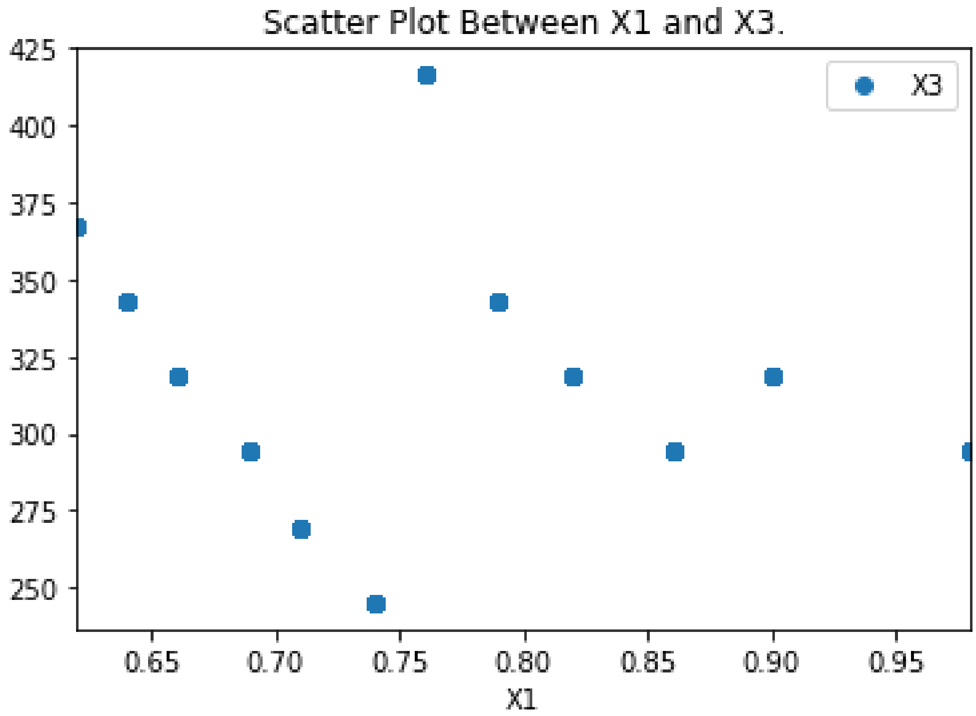
```
[23]: (10.5, -0.5)
```

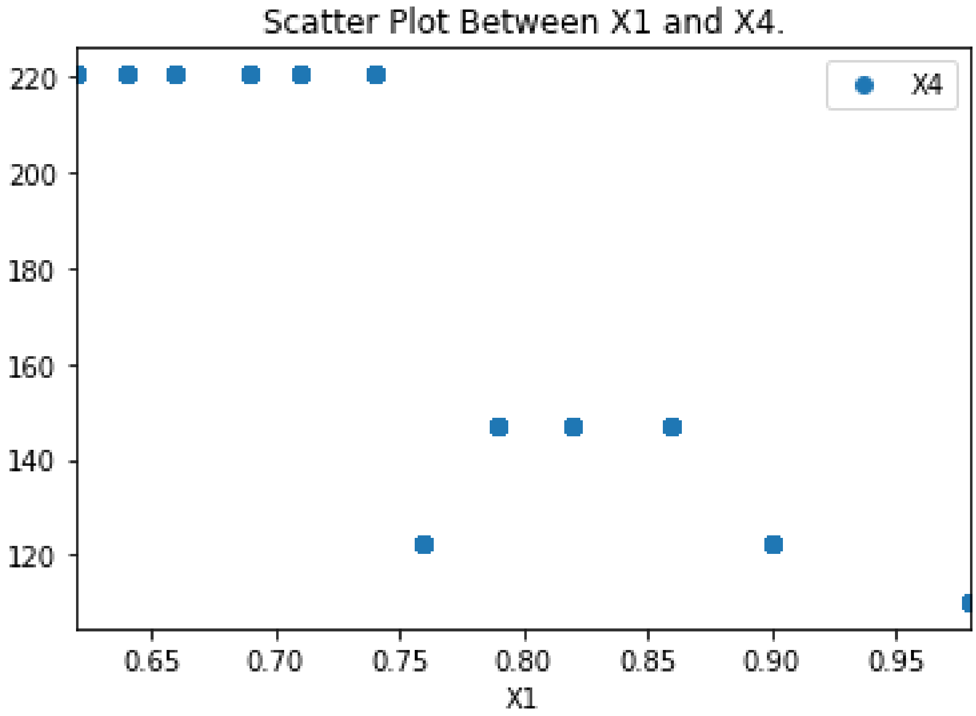


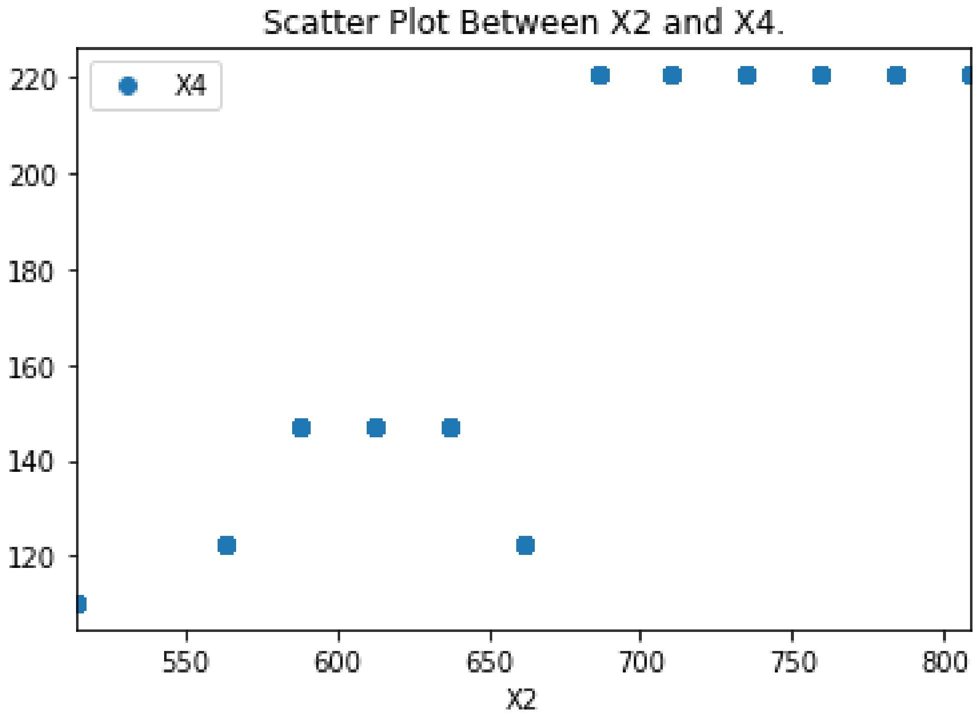
A correlation of r = 0.9 suggests a strong, positive association between two variables, whereas a correlation of r = -0.2 suggest a weak, negative association. A correlation close to zero suggests no linear association between two continuous variables.

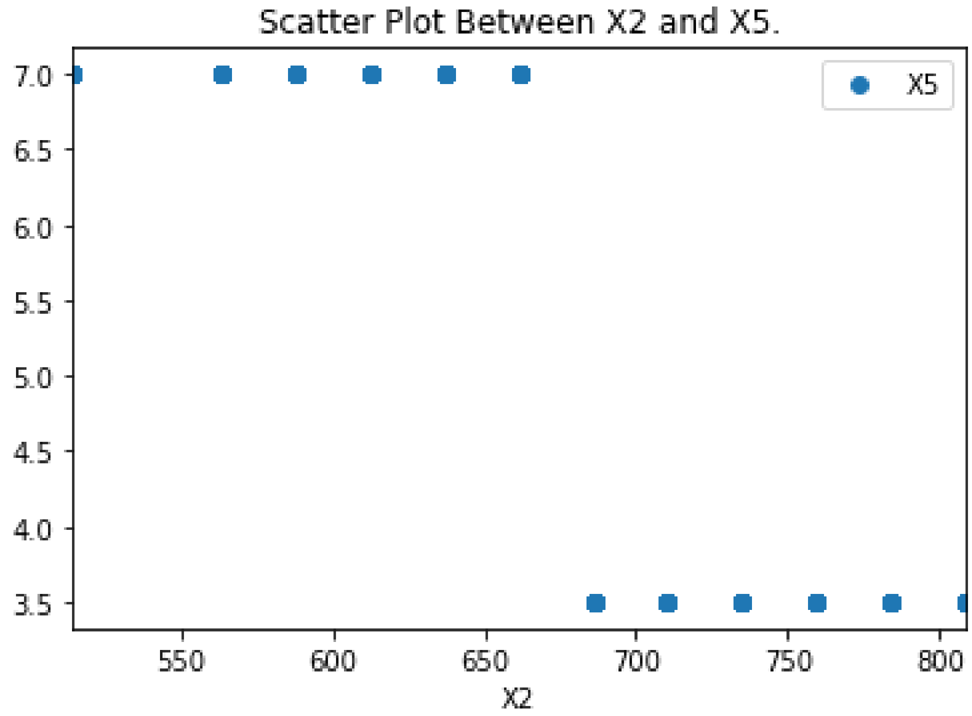
Scatter Plots between two parameters

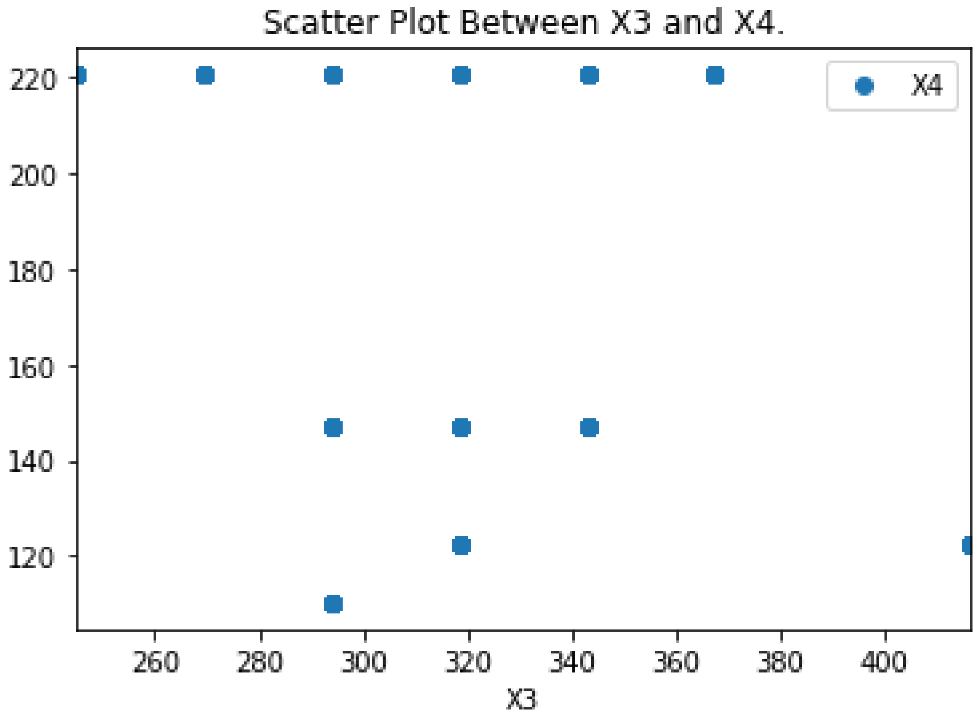


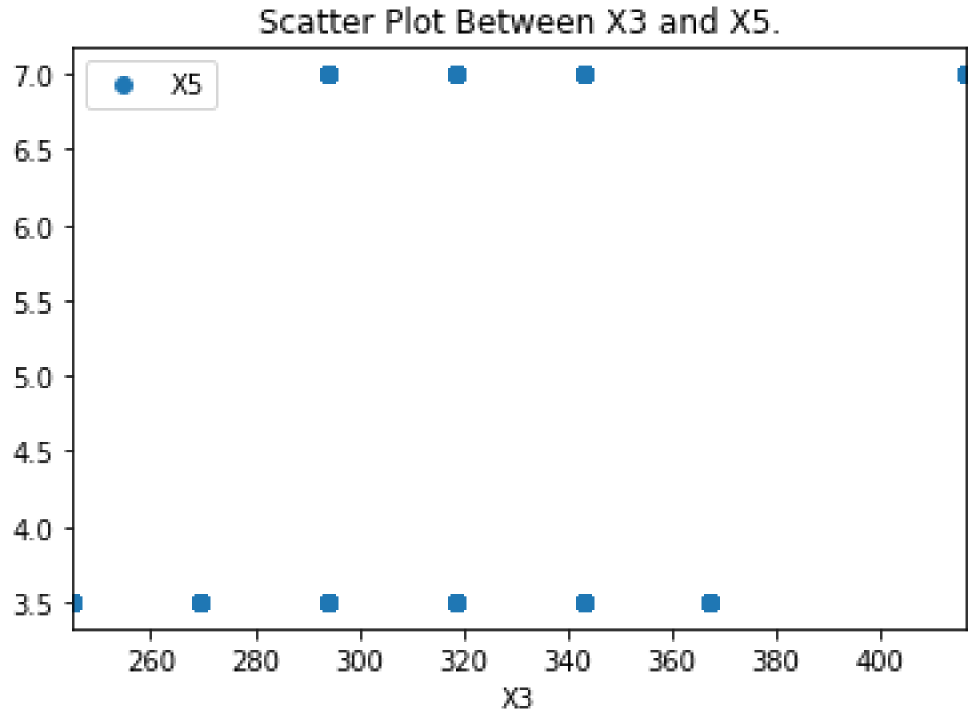

Scatter Plot Between X1 and X3.
425
400
375
350
325
300
275
250
0.65
0.70
0.75
0.80
0.85
0.90
0.95
X1
● X3

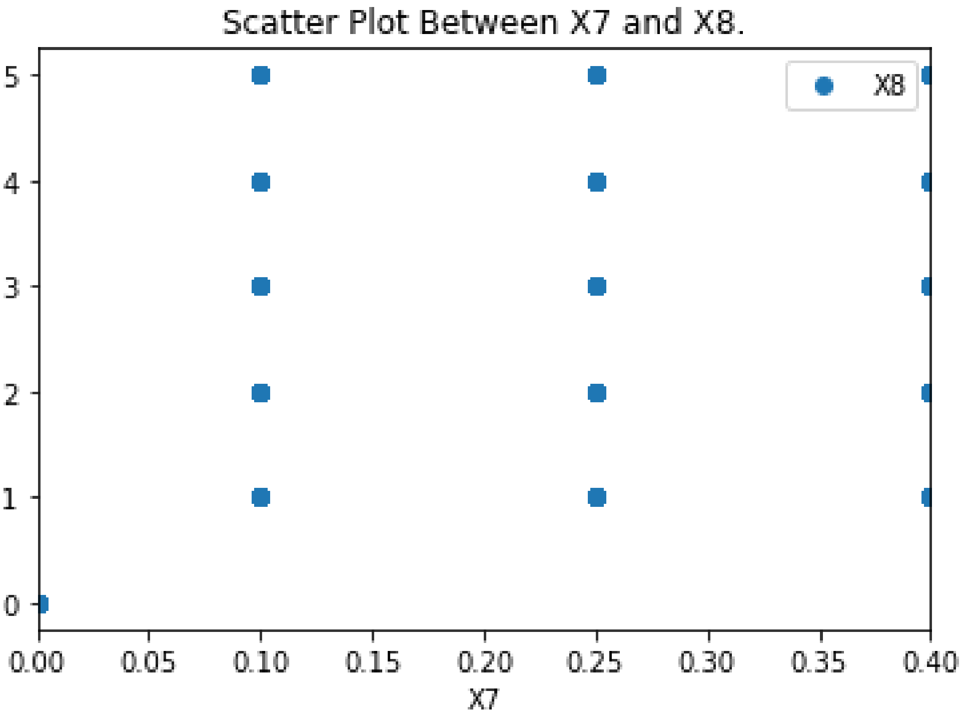

Scatter Plot Between X1 and X4.
220
200
180
160
140
120
0.65
0.70
0.75
0.80
0.85
0.90
0.95
X1
● X4


Scatter Plot Between X2 and X4.
X4
220
200
180
160
140
120
550
600
650
700
750
800
X2


Scatter Plot Between X2 and X5.
X5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
550
600
650
700
750
800
X2


Scatter Plot Between X3 and X4.
X4
220
200
180
160
140
120
260
280
300
320
340
360
380
400
X3

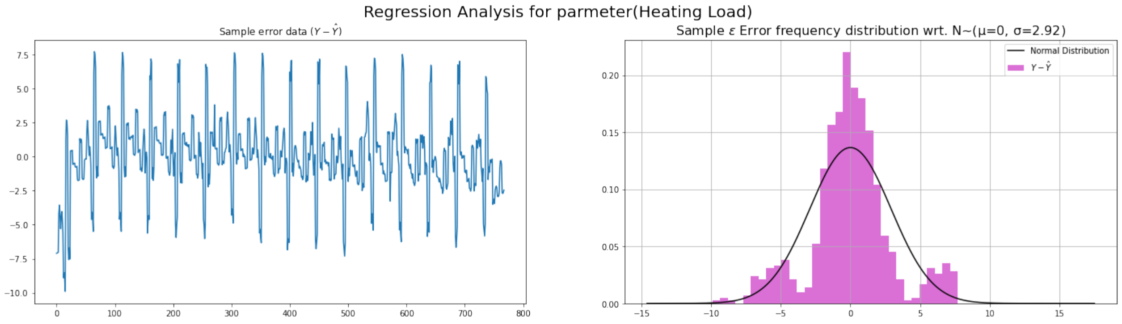

Scatter Plot Between X3 and X5.
X5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
260
280
300
320
340
360
380
400
X3



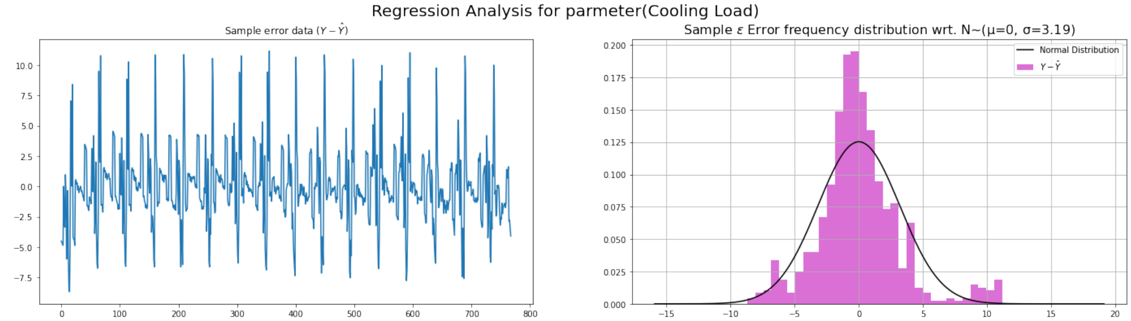
3.2 Linear Regression Analysis

Regression Analysis is a statistical tool for investigating the relationship between a dependent variable and one or more independent variables.

```
Parameter: Heating Load
Beta: [ 8.40145211e+01 -6.47739914e+01 -6.25815607e-02  3.61046266e-02
 -4.94174272e-02  4.16993882e+00 -2.33281250e-02  1.99326802e+01
  2.03771772e-01]
SSE: 6543.766185145473
SSR: 71546.0760945892
SST: 78089.84228112083
R Squared 0.9162020821738339
```



```
Parameter: Cooling Load
Beta: [ 9.72457492e+01 -7.07877069e+01 -6.60622283e-02  2.24993789e-02
 -4.43654509e-02  4.28384333e+00  1.21510417e-01  1.47170683e+01
  4.06972598e-02]
SSE: 7788.2049089323955
SSR: 61627.58283723099
SST: 69415.78774791502
R Squared 0.8878035506999205
```



```
[28]: '97.246 \\ -70.788 \\ -0.066 \\ 0.022 \\ -0.044 \\ 4.284 \\ 0.122 \\ 14.717 \\
      0.041'
```

4 Results

4.0.1 1. Important Factors for Energy Efficiency

After understanding all the parameters, below parameters can affect Energy Efficiency: 1. Roof Area 2. Overall Height 3. Relative Compactness 4. Surface Area 5. Wall Area

4.0.2 2.Factors

A. Relative Compactness It is the measure of compactness of the closure or building.
More compact the build less will be the empty area inside which needs to heated or cooled.

B. Surface Area It refers to the Surface area of the Building. More the surface area of the building less will be the energy required to heat the building.

C. Glazing Area Glazing Area is the proportion of floor area which is covered by windows, Glass walls, glass roofs etc.
Since it is exposed to external factors like sun, snow, wind and others, this may affect the heating or cooling conditions of the building.

D. Orientation Orientation of a building. The fact the sun is lower in the sky in Winter than in Summer allows us to plan and construct buildings that capture that free heat in Winter and reject

the heat in Summer.
The orientation of the whole building plays an important part in ensuring such a ‘passive’ process works. There are 4 orientations present which are 2,3,4,5. These may be representing north facing, South facing, East Facing, West Facing.

E. Roof Area Roof area is the actual area where cooling or heating would be required, that is, inside the building. So more the roof area, less will be the energy required.

4.0.3 3.Correlation between the features

1. Relative Compactness There is high negative correlation between X1 and X2, X1 and X4, which depicts relative compactness has an inverse relation with surface area and roof area.

There is average positive correlation X1 and y1, X1 and y2, X1 and X5 which depicts relative compactness has an direct proportional relational with heating load, cooling load and Overall height.

There is no or very less correlation between X1 and X6, X1 and X7 , X1 and X8 which depicts relative compactness has no relation with orientation, glazing area and glazing area distribution

2. Surface Area There is high positive correlation between X2 and X4 which depicts Surface Area has a directly proportional relationship with Roof area

There is moderate negative correlation between X2 and y1, X2 and y2 which depicts Surface Area has an inverse relationship with the heating load and cooling load

There is no or very less correlation between X2 and X6, X2 and X7 , X2 and X8 which depicts Surface Area has no relation with orientation, glazing area and glazing area distribution

3. Wall Area There is moderate positive correlation between X2 and y1, X2 and y2 which depicts Surface Area has an directly proportional(with postive slope) relationship with the heating load and cooling load

There is no or very less correlation between X3 and X6, X3 and X7 , X3 and X8 which depicts Wall Area has no relation with orientation, glazing area and glazing area distribution

4. Roof Area There is high negative correlation between X4 and y1, X4 and y2, X4 and X5 which depicts Roof Area has an inverse relation with heating load, cooling load and overall height.

There is no or very less correlation between X4 and X6, X4 and X7 , X4 and X8 which depicts Roof Area has no relation with orientation, glazing area and glazing area distribution

5. Overall Height There is high postive correlation between X5 and y1, X5 and y2 which depicts Overall height is directly proportional with heating load and cooling load.

6. Orientation Orienation hardly has any correlation with any of the parameters which depicts it has no relation with the other parameters.

7. Glazing Area Glazing Area has a low postive correlation with y1 and y2 which depicts it has a small effect on the heating load and cooling load

8. Glazing Area Distribution Glazing Area Distribution has any correlation with any of the parameters which depicts it has no relation with the other parameters.

9. Heating Load and 10. Cooling Load Heating Load and Cooling Load has very close correlation pattern with the other parameters which depict that both of them are very close to each other in behaviour.

Both of them are highly correlated with Roof Area and Overall Height Both of them are moderately correlated with relative compactness, Surface Area and wall Area. Both of them has very low correaltion with Orientation, Glazing area and glazing area distribution

4.0.4 4. Regression Anaylsis

A. Heating Load The response (Heating Load) fit the Regression model quite well. The R^2 parameter has a value of `0.9162` which is closer to 1, which is good.
The $\hat{\beta}$ vector is

$$\hat{\beta}_{(R+1)\times 1} = \begin{bmatrix} 84.015 \\ -64.774 \\ -0.063 \\ 0.036 \\ -0.049 \\ 4.170 \\ -0.023 \\ 19.933 \\ 0.204 \end{bmatrix}$$

B. Cooling Load The response (Cooling Load) fit the Regression model quite well. The R^2 parameter has a value of `0.8878` which is closer to 1, which is good.
The $\hat{\beta}$ vector is

$$\hat{\beta}_{(R+1)\times 1} = \begin{bmatrix} 97.246 \\ -70.788 \\ -0.066 \\ 0.022 \\ -0.044 \\ 4.284 \\ 0.122 \\ 14.717 \\ 0.041 \end{bmatrix}$$

5 Conclusion

Building with below attributes would have higher energy efficiency: 1. Relative Compactness should be lesser than 0.75
2. Roof Area should be around 220.5 sqft
3. Glazing Area should less be less than 0.234 units
4. Orientation should be Orientation number 4 5. Overall Height should be less than equal to 3.5
6. Surface area should be around 750 sqft or more.

6 References

1. 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
2. Dataset - Dataset used was created by Angeliki Xifara