Capstone Project CKME 136

Student: Walter Cios; Student ID 500791945

# Introduction

Building enclosures have become more energy efficient as energy codes have become more stringent in recent years, highlighting the importance of improvements such as better windows, orientation, or insulation. The heating and cooling load calculation is the first step of the HVAC design procedure. Heating and cooling load are used to right-size the HVAC equipment of a building.

The scope of this project is to build a statistical model to predict the heating and cooling load and to select the best performing model that provides the highest accuracy. The quantitative research will also investigate which variables are influencing the most the cooling efficiency and will explore the trade-offs between accuracy and interpretability of the model. Last, the energy efficiency impact when increasing the window area or house orientation will also be investigated as part of this study. This model and analysis will help HVAC equipment selection and prevent purchasing of oversized equipment.

# Dataset

The dataset analyzed is a public dataset, created by Angeliki Xifara & Athanasios The data is a simulation of 768 residential seven person buildings, assumed to be in Athens, Greece and has eight input variables (surface area, wall area, roof area, orientation, glazing area.) and two output variables (heating and cooling load). ...

# Step 1: Data Cleaning and Preparation

* Load the package to read the raw data from Excel;
* load also ggplot2 for graphs;
* load dplyr for applying transformations on the dataframes;
* load package GGally to be used for plotting
* load package reshape 2

library(readxl)  
library(ggplot2)  
library(dplyr)  
library(GGally)  
library(reshape2)

### Loading the data

After examining the original excel file, I noticed that only the first sheet had data, so I will only read that from the file

raw\_data <- read\_excel("./Data/ENB2012\_data.xlsx", sheet = 1)

### Display structure of "raw\_data"

I reviewed that all 768 observations are loaded

str(raw\_data)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 768 obs. of 10 variables:  
## $ X1: num 0.98 0.98 0.98 0.98 0.9 0.9 0.9 0.9 0.86 0.86 ...  
## $ X2: num 514 514 514 514 564 ...  
## $ X3: num 294 294 294 294 318 ...  
## $ X4: num 110 110 110 110 122 ...  
## $ X5: num 7 7 7 7 7 7 7 7 7 7 ...  
## $ X6: num 2 3 4 5 2 3 4 5 2 3 ...  
## $ X7: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ X8: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Y1: num 15.6 15.6 15.6 15.6 20.8 ...  
## $ Y2: num 21.3 21.3 21.3 21.3 28.3 ...

### Check for NA in the whole dataset and found none

....

sum(sapply(raw\_data, is.na))

## [1] 0

As an alternative we could also have done an NA check per columns and as expected found none

sapply(raw\_data, function(x) { sum(is.na(x)) })

## X1 X2 X3 X4 X5 X6 X7 X8 Y1 Y2   
## 0 0 0 0 0 0 0 0 0 0

### Change column name to real variable names

....

data <- raw\_data  
names(data) <- c("Relative Compactness", "Surface Area", "Wall Area", "Roof Area", "Overall Height",  
 "Orientation", "Glazing Area", "Glazing Area Distribution", "Heating Load", "Cooling Load")

## Converting categorical variables to factors

As per the dataset description the variable "Orientation" has only four values N,W,S,E The variable "Glazing Area Distribution"" has also the above and in addition "0" and "1"

toCategorical <- Vectorize(function(column) {  
 switch(as.character(column),  
 "0" = "No Glazing Area",  
 "1" = "Uniform",  
 "2" = "North",  
 "3" = "East",  
 "4" = "South",  
 "5" = "West",  
 NA)  
})  
  
data$`Glazing Area Distribution` <- as.factor(toCategorical(data$`Glazing Area Distribution`))  
data$Orientation <- as.factor(toCategorical(data$Orientation))  
  
data$`Glazing Area Distribution` <- relevel(data$`Glazing Area Distribution`, "No Glazing Area")

## Review again structure, this time for "data"

summary(data)

## 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 East :192 Min. :0.0000 No Glazing Area: 48   
## 1st Qu.:3.50 North:192 1st Qu.:0.1000 East :144   
## Median :5.25 South:192 Median :0.2500 North :144   
## Mean :5.25 West :192 Mean :0.2344 South :144   
## 3rd Qu.:7.00 3rd Qu.:0.4000 Uniform :144   
## Max. :7.00 Max. :0.4000 West :144   
## 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

From the summmary we can see that Orientation and Glazing Area Distribution, are qualitative-categorical variables, the other variables are quantitaive-continuous

# Step 2: Exploratory Data Analysis

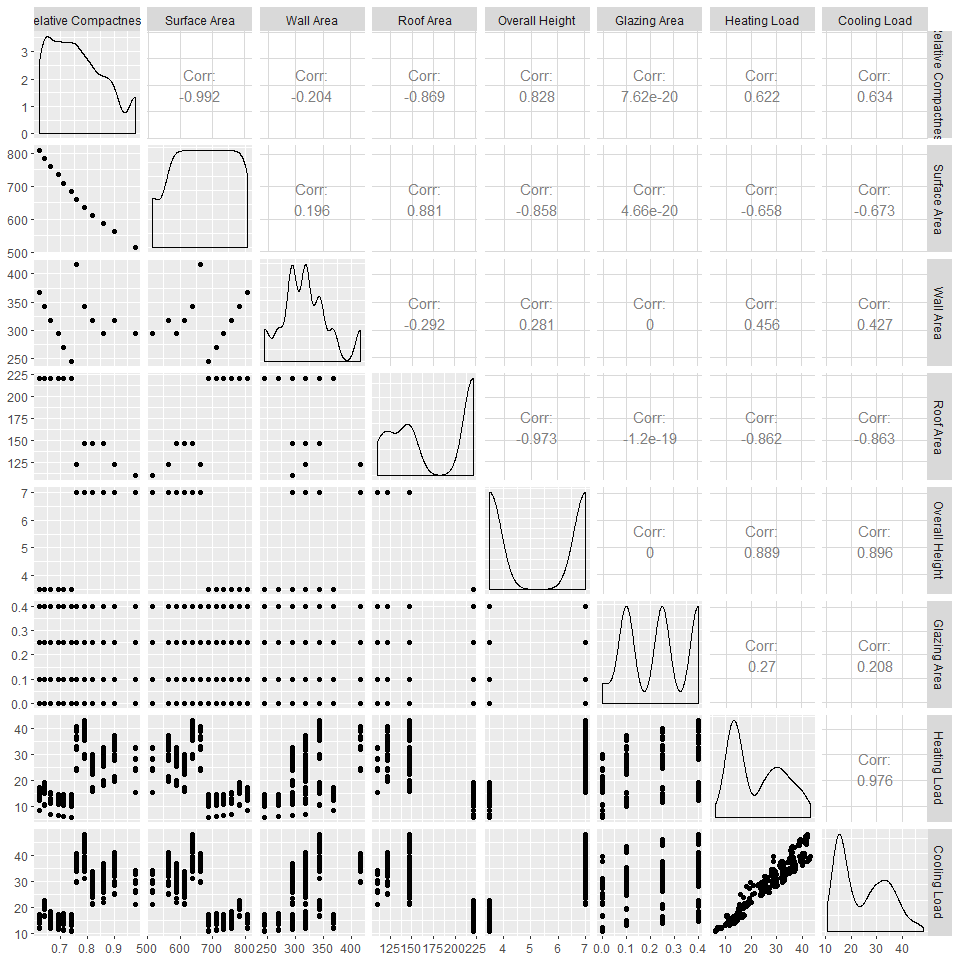
Split up categorical variables from the continuous ones, into a new dataframe "data\_categorical and keep the continuos ones in "data\_numerical"

data\_numeric <- select(data, -`Orientation`, -`Glazing Area Distribution`)  
data\_categorical <- select(data, `Orientation`, `Glazing Area Distribution`)

## Paired Scatterplots

I looked for a better display of plot and found the package "GGally" that is being used for scatterplots between all pairs of variables

ggpairs(data\_numeric)



## Outlier detection

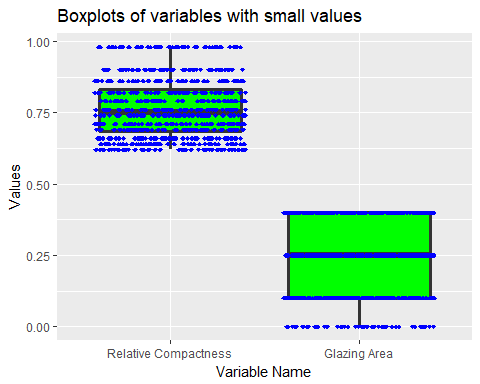
Review if there are outliers by using boxplots.Going forward I will try to use only ggplot for all graphs to create an consistent view

Looking at the Excel Table "Variable Review for Analysis" that has the Min and Max, I decided to split the boxplots into three categories by cluster of data: small (values 0-1); medium (values >1-50); large (>50)

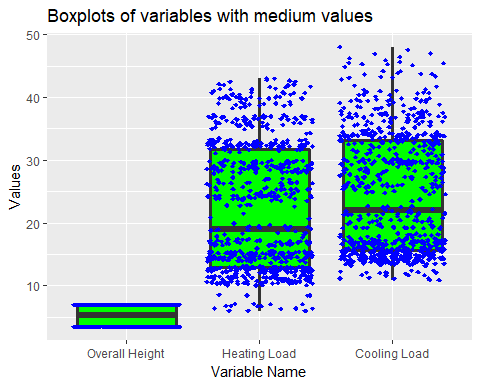
small <-select (data\_numeric,`Relative Compactness`,`Glazing Area`)  
  
medium <-select (data\_numeric,`Overall Height`,`Heating Load`,`Cooling Load`)  
  
large <-select (data\_numeric,`Surface Area`,`Roof Area`,`Wall Area`)

Creating boxplots for the grouped variables into small, medium and large

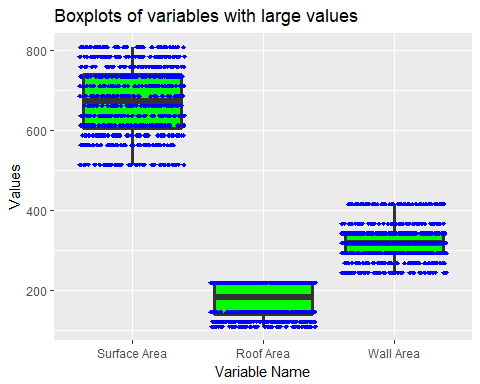
small\_long <- melt(small)  
ggplot(data=small\_long, aes(x = variable, y = value)) +  
 geom\_boxplot(fill = "green", lwd = 1.2) +  
 geom\_jitter(height = 0, width = 0.4, shape = 18, color = "blue") +  
 ggtitle("Boxplots of variables with small values") +  
 xlab("Variable Name") +  
 ylab("Values")



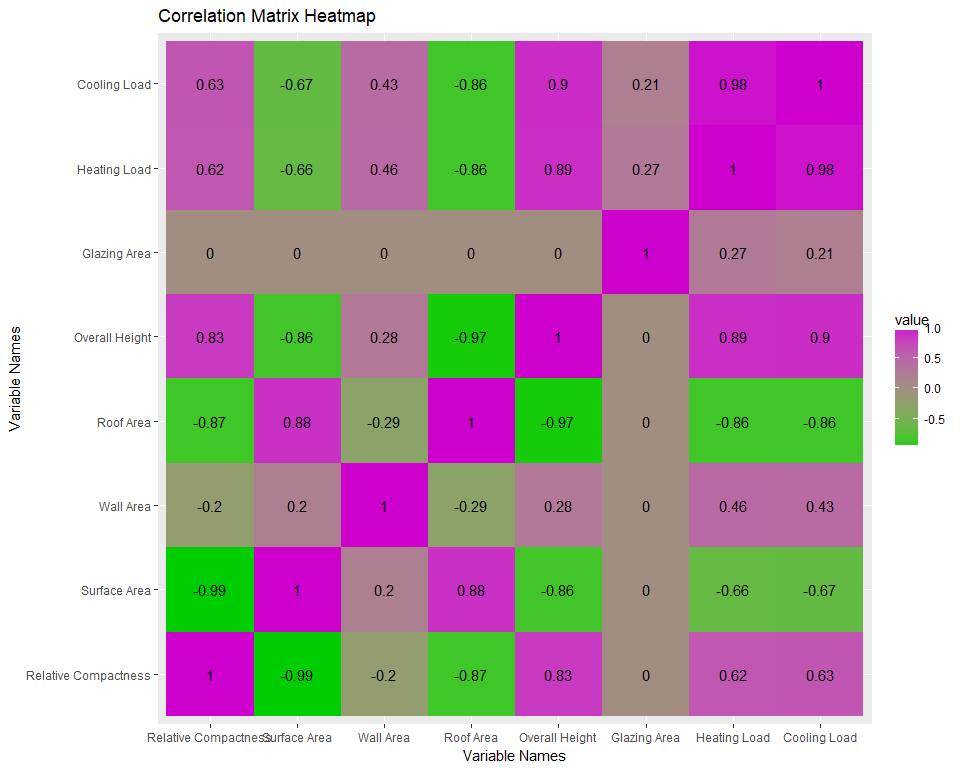
medium\_long <- melt(medium)  
  
ggplot(data=medium\_long, aes(x = variable, y = value)) +  
 geom\_boxplot(fill = "green",lwd = 1.2) +  
 geom\_jitter(height = 0, width = 0.4, shape = 18, color = "blue") +  
 ggtitle("Boxplots of variables with medium values") +  
 xlab("Variable Name") +  
 ylab("Values")



# Step 2   
large\_long <- melt(large)  
  
ggplot(data=large\_long, aes(x = variable, y = value)) +  
 geom\_boxplot(fill = "green", lwd = 1.2) +  
 geom\_jitter(height = 0, width = 0.4, shape = 18, color = "blue") +  
 ggtitle("Boxplots of variables with large values") +  
 xlab("Variable Name") +  
 ylab("Values")



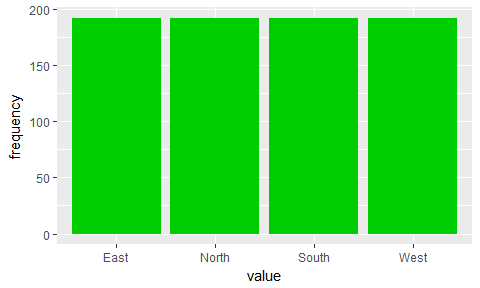
correlationMatrix <- cor(data\_numeric, method = "pearson")  
correlationsLong <- melt(correlationMatrix)  
  
ggplot(data = correlationsLong, aes(x = Var1, y = Var2)) +  
 geom\_tile(aes(fill = value)) +  
 geom\_text(aes(label = round(value, 2))) +  
 scale\_fill\_gradient(low = "green3", high = "magenta3") +  
 ggtitle("Correlation Matrix Heatmap") +  
 xlab("Variable Names") +  
 ylab("Variable Names")



table(data\_categorical$Orientation)

##   
## East North South West   
## 192 192 192 192

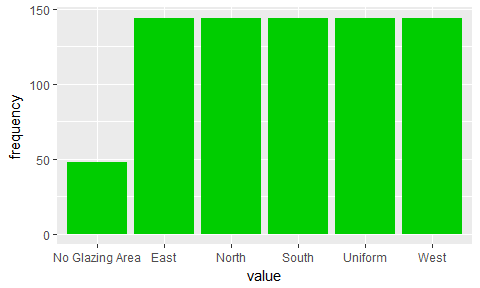
table1 <- as.data.frame(table(data\_categorical$Orientation))  
table1 <- rename(table1, value = Var1, frequency = Freq)  
  
ggplot(data = table1, aes(x = value, y = frequency)) +  
 geom\_bar(stat = "identity", fill = "green3")



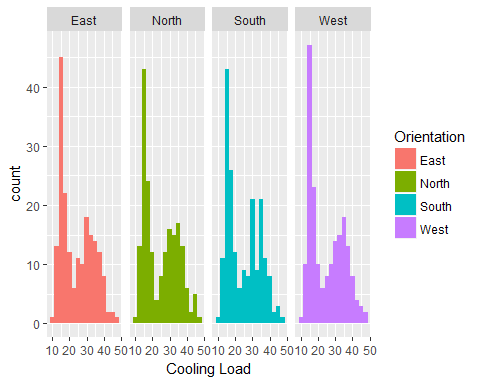
table(data\_categorical$`Glazing Area Distribution`)

##   
## No Glazing Area East North South   
## 48 144 144 144   
## Uniform West   
## 144 144

table2 <- as.data.frame(table(data\_categorical$`Glazing Area Distribution`))  
table2 <- rename(table2, value = Var1, frequency = Freq)  
  
ggplot(data = table2, aes(x = value, y = frequency)) +  
 geom\_bar(stat = "identity", fill = "green3")



plotData <- select(data, `Cooling Load`, `Orientation`)  
  
ggplot(plotData, aes(x = `Cooling Load`, fill = `Orientation`)) +  
 geom\_histogram(binwidth = 2.5) +  
 facet\_grid(~ Orientation)



plotData <- select(data, `Cooling Load`, `Glazing Area Distribution`)  
  
ggplot(plotData, aes(x = `Cooling Load`, fill = `Glazing Area Distribution`)) +  
 geom\_histogram(binwidth = 2.5) +  
 facet\_grid(~ `Glazing Area Distribution`)

