# Topic: "Energy Efficiency Project"

Harvard Extension School

Student: Huynh Vo \_ Fall 2017

Youtube URLs:

2 minute (short): <https://youtu.be/bIL9AZTkGKU>

15 minutes (long):  <https://youtu.be/2sjkyZTjbrs>

## Overview and summary

My topic, and this dataset, are related to Energy Efficiency, particularly the energy efficiency of housing. The dataset was obtained from [http://archive.ics.uci.edu/ml/datasets/energy+efficiency#](http://archive.ics.uci.edu/ml/datasets/energy+efficiency). The dataset relates how the size, shape, and other features of a building can help or hurt the energy required to keep the internal temperature of a building, such as a home, at a desired temperature. This is an important real world problem that our energy hungry society is only now beginning to look at in detail using data science tools. Some of the tools used to explore the data presented were different visualization graphs provided through RStudio tools and Azure Machine Learning for a more in depth analysis.

The machine learning tools were required to predict the responses of the heating and cooling loads for different building parameters such as relative compactness, surface area, glazing area, etc. Azure Machine Learning and R are the main tools for this analysis. Through exploration of the data using these tools I have concluded that many of the factors in the data help to determine the energy required to heat and cool a building, but not all of them do. I have used rcorr to see the correlation between given features. The Rcorr function showed that there is a positive correlation between heating load and cooling load, meaning if the house does not require much heat to warm up, it does not require much A/C to cool down either. Through histograms and box plots, created using the ggplot package, I also see that Relative Compactness, Wall Area, the number of stories of the building, and glazing area play a major role in determining the heating load and cooling load.

Besides visualization and after cleaning up the data, I tested out Azure Machine Learning using a Neural Network Regression model and a Boosted Decision Tree regression model to predict the outcome of heating load and cooling loads. The models were evaluated with a high Coefficient of Determination ( ~ 0.98). Azure Machine Learning is one of the more powerful tools for model prediction. Azure Machine is easy to use, the models are loaded automatically, are highly visualized, and most importantly, it requires minimal to no installation. Azure Studio is an online platform owned by Microsoft. They do very well in terms of organization and has a very gentle learning curve. It is open source. The only con is that for a free account, there is a limit for storage and speed when training the data. If you would like to analyze data that is bigger than 10MB, you will have to purchase the premium account and your data will be processed with faster speed.

Attributes for this dataset are listed below:

Relative Compactness - no unit

Surface Area - m² Wall Area - m²

Roof Area - m²

Overall Height - m

Orientation - 2:North, 3:East, 4:South, 5:West

Glazing Area - 0%, 10%, 25%, 40% (of floor area)

Glazing Area Distribution (Variance) - 1:Uniform, 2:North, 3:East, 4:South, 5:West

Heating Load - kWh/m²

Cooling Load - kWh/m²

*Note: For Data Cleansing and Data Exploration, I knit the file through Rstudio. For Azure ML, screen captures are provided as part of this document*.

## Data Cleansing and Exploration

library(readr)  
EnergyData <-read\_csv("C:/Users/Devon/Desktop/EnergyData.csv")

## Parsed with column specification:  
## cols(  
## X1 = col\_double(),  
## X2 = col\_double(),  
## X3 = col\_double(),  
## X4 = col\_double(),  
## X5 = col\_double(),  
## X6 = col\_integer(),  
## X7 = col\_double(),  
## X8 = col\_integer(),  
## Y1 = col\_double(),  
## Y2 = col\_double()  
## )

head(EnergyData)

## # A tibble: 6 x 10  
## X1 X2 X3 X4 X5 X6 X7 X8 Y1 Y2  
## <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <int> <dbl> <dbl>  
## 1 0.98 514.5 294.0 110.25 7 2 0 0 15.55 21.33  
## 2 0.98 514.5 294.0 110.25 7 3 0 0 15.55 21.33  
## 3 0.98 514.5 294.0 110.25 7 4 0 0 15.55 21.33  
## 4 0.98 514.5 294.0 110.25 7 5 0 0 15.55 21.33  
## 5 0.90 563.5 318.5 122.50 7 2 0 0 20.84 28.28  
## 6 0.90 563.5 318.5 122.50 7 3 0 0 21.46 25.38

Below, I rename columns so we can understand the data better:

colnames(EnergyData) <- c("RelativeCompactness", "SurfaceArea", "WallArea", "RoofArea", "OverallHeight", "Orientation", "GlazingArea", "GlazingAreaDistribution", "HeatingLoad", "CoolingLoad")  
head(EnergyData)

## # A tibble: 6 x 10  
## RelativeCompactness SurfaceArea WallArea RoofArea OverallHeight  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.98 514.5 294.0 110.25 7  
## 2 0.98 514.5 294.0 110.25 7  
## 3 0.98 514.5 294.0 110.25 7  
## 4 0.98 514.5 294.0 110.25 7  
## 5 0.90 563.5 318.5 122.50 7  
## 6 0.90 563.5 318.5 122.50 7  
## # ... with 5 more variables: Orientation <int>, GlazingArea <dbl>,  
## # GlazingAreaDistribution <int>, HeatingLoad <dbl>, CoolingLoad <dbl>

Below, we explore the correlation between features:

library(Hmisc)

## Warning: package 'Hmisc' was built under R version 3.4.3

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, round.POSIXt, trunc.POSIXt, units

cor1 <- rcorr(as.matrix(EnergyData))  
cor1

## RelativeCompactness SurfaceArea WallArea RoofArea  
## RelativeCompactness 1.00 -0.99 -0.20 -0.87  
## SurfaceArea -0.99 1.00 0.20 0.88  
## WallArea -0.20 0.20 1.00 -0.29  
## RoofArea -0.87 0.88 -0.29 1.00  
## OverallHeight 0.83 -0.86 0.28 -0.97  
## Orientation 0.00 0.00 0.00 0.00  
## GlazingArea 0.00 0.00 0.00 0.00  
## GlazingAreaDistribution 0.00 0.00 0.00 0.00  
## HeatingLoad 0.62 -0.66 0.46 -0.86  
## CoolingLoad 0.63 -0.67 0.43 -0.86  
## OverallHeight Orientation GlazingArea  
## RelativeCompactness 0.83 0.00 0.00  
## SurfaceArea -0.86 0.00 0.00  
## WallArea 0.28 0.00 0.00  
## RoofArea -0.97 0.00 0.00  
## OverallHeight 1.00 0.00 0.00  
## Orientation 0.00 1.00 0.00  
## GlazingArea 0.00 0.00 1.00  
## GlazingAreaDistribution 0.00 0.00 0.21  
## HeatingLoad 0.89 0.00 0.27  
## CoolingLoad 0.90 0.01 0.21  
## GlazingAreaDistribution HeatingLoad CoolingLoad  
## RelativeCompactness 0.00 0.62 0.63  
## SurfaceArea 0.00 -0.66 -0.67  
## WallArea 0.00 0.46 0.43  
## RoofArea 0.00 -0.86 -0.86  
## OverallHeight 0.00 0.89 0.90  
## Orientation 0.00 0.00 0.01  
## GlazingArea 0.21 0.27 0.21  
## GlazingAreaDistribution 1.00 0.09 0.05  
## HeatingLoad 0.09 1.00 0.98  
## CoolingLoad 0.05 0.98 1.00  
##   
## n= 768   
##   
##   
## P  
## RelativeCompactness SurfaceArea WallArea RoofArea  
## RelativeCompactness 0.0000 0.0000 0.0000   
## SurfaceArea 0.0000 0.0000 0.0000   
## WallArea 0.0000 0.0000 0.0000   
## RoofArea 0.0000 0.0000 0.0000   
## OverallHeight 0.0000 0.0000 0.0000 0.0000   
## Orientation 1.0000 1.0000 1.0000 1.0000   
## GlazingArea 1.0000 1.0000 1.0000 1.0000   
## GlazingAreaDistribution 1.0000 1.0000 1.0000 1.0000   
## HeatingLoad 0.0000 0.0000 0.0000 0.0000   
## CoolingLoad 0.0000 0.0000 0.0000 0.0000   
## OverallHeight Orientation GlazingArea  
## RelativeCompactness 0.0000 1.0000 1.0000   
## SurfaceArea 0.0000 1.0000 1.0000   
## WallArea 0.0000 1.0000 1.0000   
## RoofArea 0.0000 1.0000 1.0000   
## OverallHeight 1.0000 1.0000   
## Orientation 1.0000 1.0000   
## GlazingArea 1.0000 1.0000   
## GlazingAreaDistribution 1.0000 1.0000 0.0000   
## HeatingLoad 0.0000 0.9429 0.0000   
## CoolingLoad 0.0000 0.6926 0.0000   
## GlazingAreaDistribution HeatingLoad CoolingLoad  
## RelativeCompactness 1.0000 0.0000 0.0000   
## SurfaceArea 1.0000 0.0000 0.0000   
## WallArea 1.0000 0.0000 0.0000   
## RoofArea 1.0000 0.0000 0.0000   
## OverallHeight 1.0000 0.0000 0.0000   
## Orientation 1.0000 0.9429 0.6926   
## GlazingArea 0.0000 0.0000 0.0000   
## GlazingAreaDistribution 0.0154 0.1619   
## HeatingLoad 0.0154 0.0000   
## CoolingLoad 0.1619 0.0000

Taking the summary of the EnergyData shows a more complete picture of this raw data:

summary(EnergyData)

## RelativeCompactness SurfaceArea WallArea RoofArea   
## 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   
## OverallHeight Orientation GlazingArea GlazingAreaDistribution  
## Min. :3.50 Min. :2.00 Min. :0.0000 Min. :0.000   
## 1st Qu.:3.50 1st Qu.:2.75 1st Qu.:0.1000 1st Qu.:1.750   
## Median :5.25 Median :3.50 Median :0.2500 Median :3.000   
## Mean :5.25 Mean :3.50 Mean :0.2344 Mean :2.812   
## 3rd Qu.:7.00 3rd Qu.:4.25 3rd Qu.:0.4000 3rd Qu.:4.000   
## Max. :7.00 Max. :5.00 Max. :0.4000 Max. :5.000   
## HeatingLoad CoolingLoad   
## 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

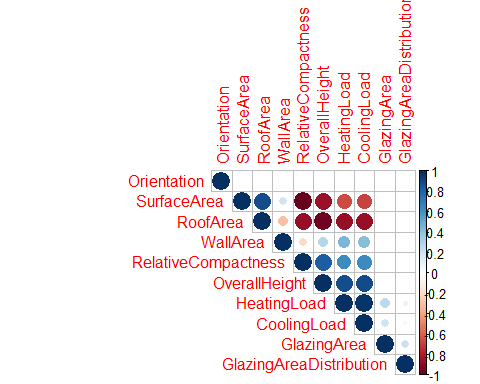
Interpretation from summary and correlation data: Heating load has a positive correlation with Overall Height, Relative Compactness, Wall Area, Glazing Area and Glazing Area Distribution. Heating Load has negative correlation with RoofArea and SurfaceArea. HeatingLoad has no correlation with Orientation. Cooling Load has positive correlation with Orientation and Glazing Area Distribution. Cooling Load has no correlation with the rest of the data. Overall, the following variables have a correlation to Heating Load and Cooling Load. They are listing in order from greatest impact to least impact : Relative Compactness,Wall Area, Overall Height, Glazing Area and Glazing Area Distribution. The Correlation matrix can be visualized as well through a correlogram:

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.4.3

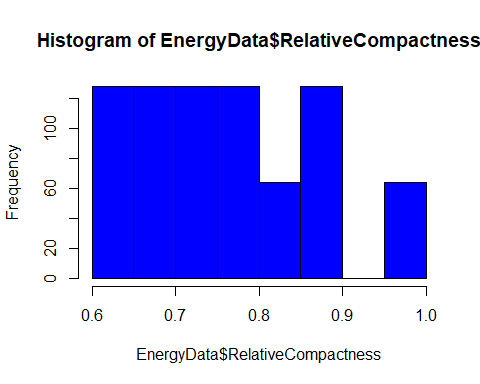
## corrplot 0.84 loaded

m <- cor(EnergyData)  
corrplot(m, order = "hclust", addrect =2, type = "upper")

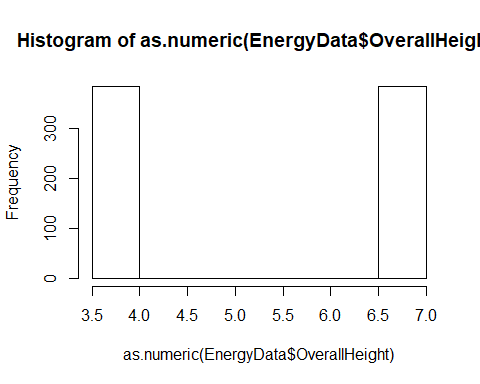


For visualization purposes, below graphs show feature distribution throughtout the dataset:

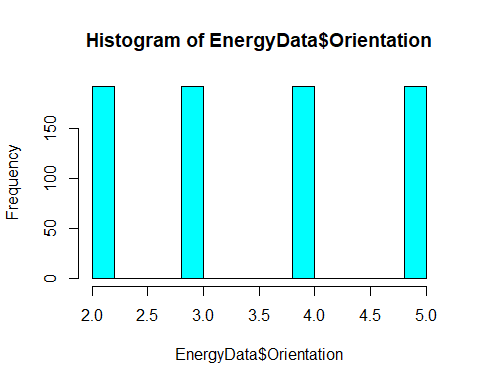
hist(EnergyData$RelativeCompactness, col = "blue")



hist(as.numeric(EnergyData$OverallHeight))

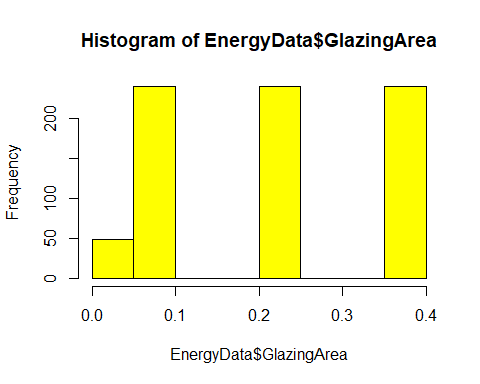


hist(EnergyData$Orientation, col = "cyan")

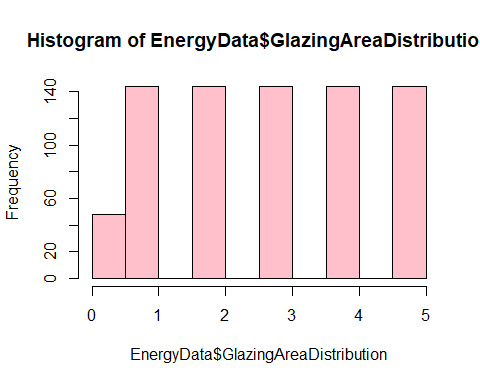


For the above graph, Orientation - 2:North, 3:East, 4:South, 5:West

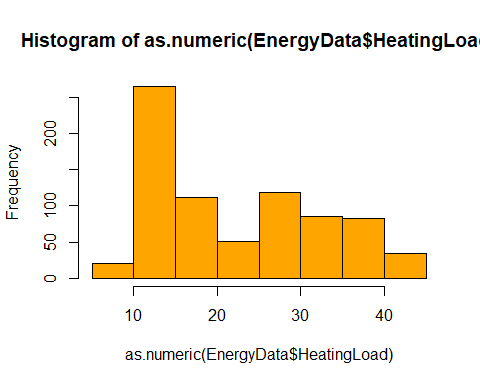
hist(EnergyData$GlazingArea, col = "yellow")



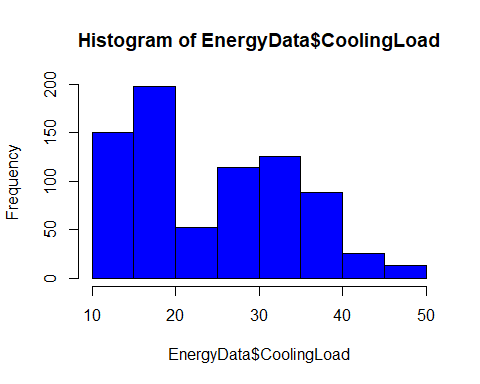
hist(EnergyData$GlazingAreaDistribution, col = "pink")



hist(as.numeric(EnergyData$HeatingLoad), col = "orange")



hist(EnergyData$CoolingLoad, col = "blue")



Now that I see how the values are distributed. I would like to convert value of overall height to the number of stories in a particular building. 3.5 meters represents a 1 story building, while 7 meters represents a 2 story building. These are categorical values.

StoriesFUN <- function(x) {  
 ifelse((x!=3.5), "two", "one")  
}  
EnergyData$Stories<-lapply(EnergyData$OverallHeight,StoriesFUN)  
EnergyData$Stories<-as.character(EnergyData$Stories)  
head(EnergyData)

## # A tibble: 6 x 11  
## RelativeCompactness SurfaceArea WallArea RoofArea OverallHeight  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.98 514.5 294.0 110.25 7  
## 2 0.98 514.5 294.0 110.25 7  
## 3 0.98 514.5 294.0 110.25 7  
## 4 0.98 514.5 294.0 110.25 7  
## 5 0.90 563.5 318.5 122.50 7  
## 6 0.90 563.5 318.5 122.50 7  
## # ... with 6 more variables: Orientation <int>, GlazingArea <dbl>,  
## # GlazingAreaDistribution <int>, HeatingLoad <dbl>, CoolingLoad <dbl>,  
## # Stories <chr>

According to the dataset source, the Orientation and Glazing Area Distribution variables follow below indications: 0 = No glazing 1 = Uniform 2 = North 3 = East 4 = South 5 = West

CompasFUN <- function(x) {  
 ifelse(( x == 0), "no glazing",  
 ifelse((x == 1), "uniform",  
 ifelse((x == 2), "north",  
 ifelse((x == 3), "east",  
 ifelse((x == 4), "south", "west")))))  
}  
EnergyData$Orientation1<-lapply(EnergyData$Orientation,CompasFUN)  
EnergyData$Orientation1<-as.character(EnergyData$Orientation1)  
EnergyData$GlazingAreaDistribution1<-lapply(EnergyData$GlazingAreaDistribution,CompasFUN)  
EnergyData$GlazingAreaDistribution1<-as.character(EnergyData$GlazingAreaDistribution1)  
head(EnergyData)

## # A tibble: 6 x 13  
## RelativeCompactness SurfaceArea WallArea RoofArea OverallHeight  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.98 514.5 294.0 110.25 7  
## 2 0.98 514.5 294.0 110.25 7  
## 3 0.98 514.5 294.0 110.25 7  
## 4 0.98 514.5 294.0 110.25 7  
## 5 0.90 563.5 318.5 122.50 7  
## 6 0.90 563.5 318.5 122.50 7  
## # ... with 8 more variables: Orientation <int>, GlazingArea <dbl>,  
## # GlazingAreaDistribution <int>, HeatingLoad <dbl>, CoolingLoad <dbl>,  
## # Stories <chr>, Orientation1 <chr>, GlazingAreaDistribution1 <chr>

Now, lets see what boxplots can show us more about the data:

library(dplyr)

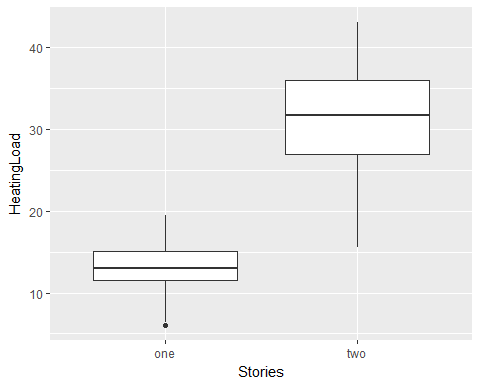
##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:Hmisc':  
##   
## combine, src, summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

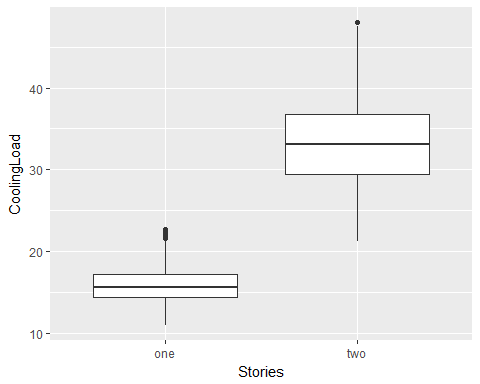
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

ggplot(EnergyData, aes(x=Stories, y=HeatingLoad))+  
 geom\_boxplot(aes(group=Stories))



Interpertation for above boxplot: The boxplot shows me the median HeatingLoad value for 1 story buildings is approximately less than half of amount required comparing to 2 stories buildings in general.

ggplot(EnergyData, aes(x=Stories, y=CoolingLoad))+  
 geom\_boxplot(aes(group=Stories))



Interpertation of the above boxplot: The boxplot shows the median Cooling Load value for a 1 story building is approximately less than half of the amount required compared to 2 story buildings in general. Now, let's see how the orientation of buildings affect Heating Load and Cooling Load through graphs below:

ggplot(EnergyData, aes(x=Orientation1, y=HeatingLoad))+  
 geom\_boxplot(aes(group=Orientation1))



Lets see how boxplot looks like for CoolingLoad:

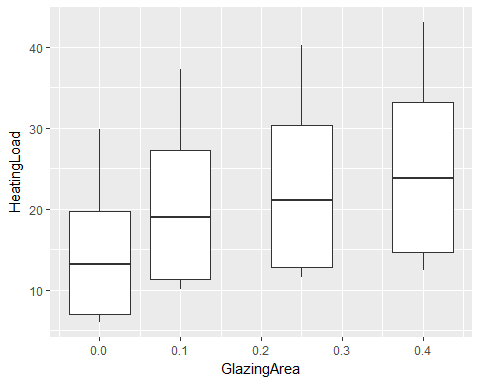
ggplot(EnergyData, aes(x=Orientation1, y=CoolingLoad))+  
 geom\_boxplot(aes(group=Orientation1))



Interpretation for the above bloxplots: Above boxplots tell me that Orientation of the building does not affect Heating Load and Cooling Load in this case since they show similar result.

I also would like to see how the glazing area affecting Heating Load and Cooling Load of buildings:

ggplot(EnergyData, aes(x=GlazingArea, y=HeatingLoad))+  
 geom\_boxplot(aes(group=GlazingArea))



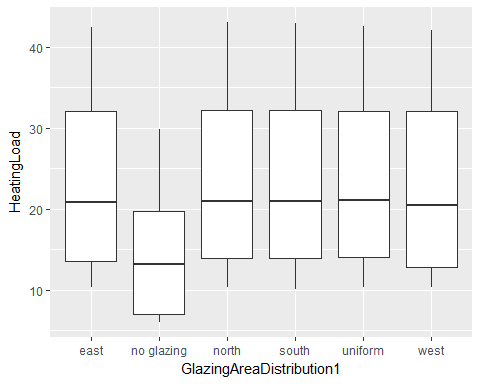
ggplot(EnergyData, aes(x=GlazingArea, y=CoolingLoad))+  
 geom\_boxplot(aes(group=GlazingArea))



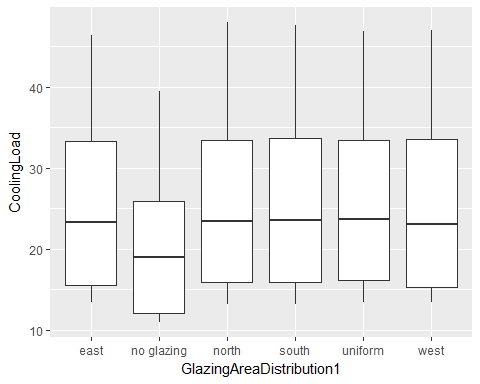
Interpretation for above boxplots: Glazing area is measured as a percentage of the facade area and the reflective factor of glass. For example, a glazing area of 40% means the glass windows take an equivalent 40% of the floor area.  
The above boxplots show that the more glazing area a building has the worse effect it has on the Heating Load and Cooling Load. Its an important factor for energy efficiency. Less windows is better for the interior temperature control of a building.

Now, I want to study how the glazing area distribution affects the Heating Load and Cooling Load:

ggplot(EnergyData, aes(x=GlazingAreaDistribution1, y=HeatingLoad))+  
 geom\_boxplot(aes(group=GlazingAreaDistribution1))

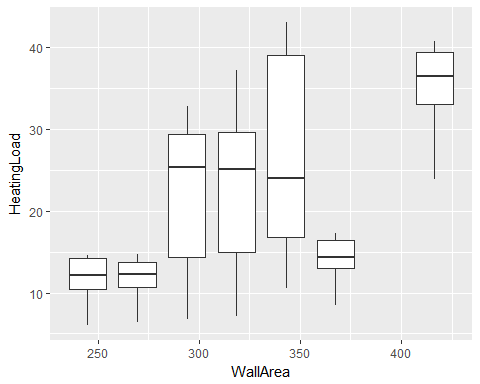


ggplot(EnergyData, aes(x=GlazingAreaDistribution1, y=CoolingLoad))+  
 geom\_boxplot(aes(group=GlazingAreaDistribution1))

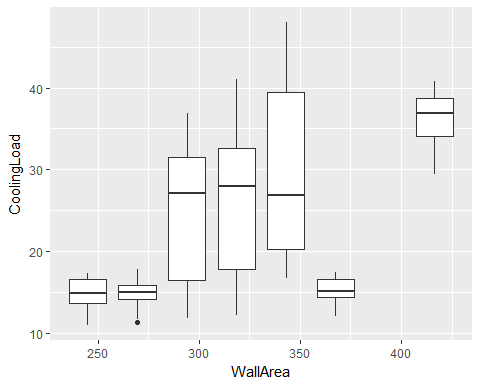


Interpretation for the above boxplots: Above boxplots tell me that buildings without glaze minimize the Heating Load and Cooling Load. This not only agrees with the previous data, but also agrees with qualitative data because air travels through windows/glass area easier compared to solid walls. Once again, the glazing distribution in terms of orientation does not affect the Heating Load and Cooling Load. Below, we want to see how the Surface Area affects HeatingLoad and CoolingLoad:

ggplot(EnergyData, aes(x=WallArea, y=HeatingLoad))+  
 geom\_boxplot(aes(group=WallArea))



ggplot(EnergyData, aes(x=WallArea, y=CoolingLoad))+  
 geom\_boxplot(aes(group=WallArea))



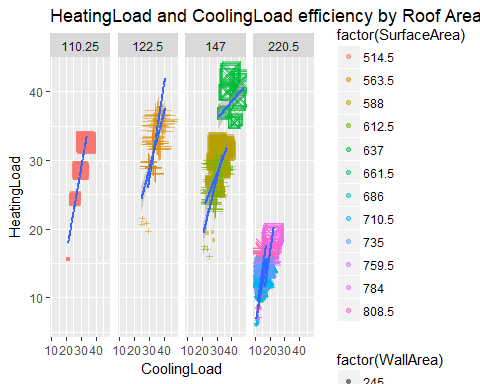
Interpretation for the above box plots: The above plot shows that wall area has a non-linear (inconsistent) relation to HeatingLoad and CoolingLoad in this data. It means that there are other factors that need to be taken into account, not just wall area. And it is totally ok to encounter this type of data because in real world, we need to count multiple factors into the calculation. Maybe the WallArea material determines how efficient these walls are for the buildings. This is the purpose of data exploration

Up to this point in terms of exploration, I would like to see how the overall effect of RoofArea, Wall Area, Surface Area and Glazing Area to HeatingLoad and CoolingLoad all together:

ggplot(EnergyData,  
 aes(x = CoolingLoad, y = HeatingLoad, group = factor(WallArea),   
 size = GlazingArea,  
 shape = factor(WallArea)))+  
 geom\_point(aes(colour= factor(SurfaceArea)), alpha = 0.5)+  
 geom\_smooth(method = "lm",se = TRUE )+  
 facet\_grid(~ RoofArea ) +  
 ggtitle('HeatingLoad and CoolingLoad efficiency by Roof Area, Wall Area, Surface Area and Glazing Area')

## Warning: The shape palette can deal with a maximum of 6 discrete values  
## because more than 6 becomes difficult to discriminate; you have 7.  
## Consider specifying shapes manually if you must have them.

## Warning: Removed 64 rows containing missing values (geom\_point).

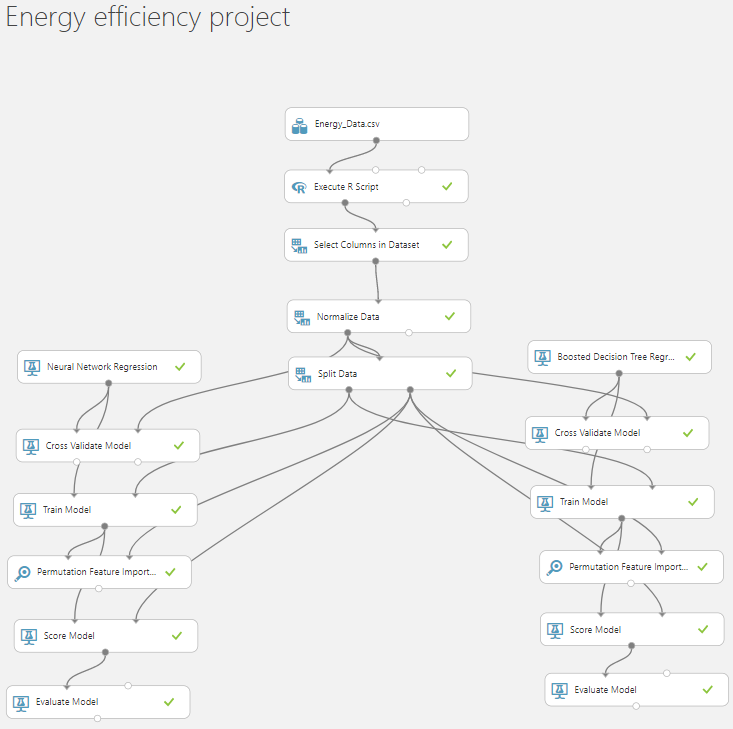


Interpretation for the above graph: The title of this plot was not completed printed out for some reasons. Looking at the graph, we see that it is HeatingLoad and CoolingLoad are influenced by RoofArea, Wall Area, SurfaceArea and GlazingArea in combination.

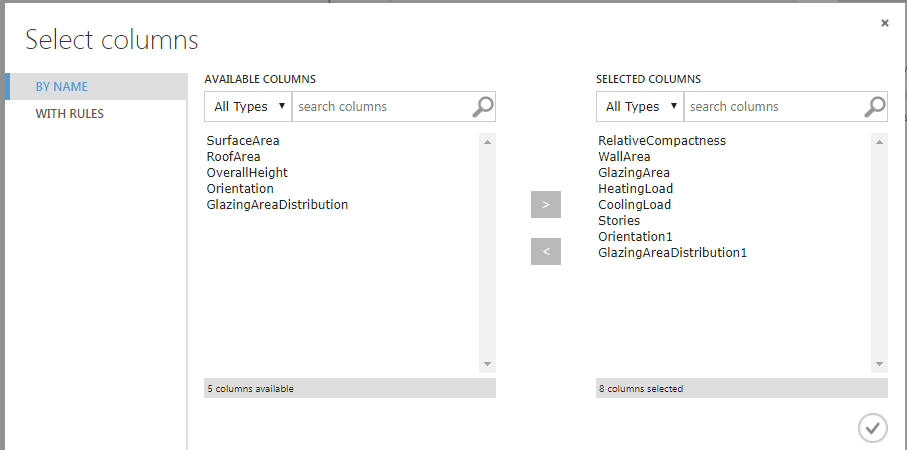
# Azure Machine Learning: Predictive Analytics

With this set of data, we can predict the responses of heating load and cooling load based on given features. In Azure Machine Learning, I used Neural Network Regression to predict heating load, and I use Boosted Decision Tree Regression to predict cooling load. I purposely use 2 different models to demonstrate Azure ML has different modelling options and we can choose a model that fits our data for our purposes. Please take a look at the HVFinalProject Word document for captures of my results. For both models, after cleansing steps and understand the data better, I normalize Wall Area and Glazing Area, split the data into 70%/30% for training model. I also include "Cross Validate Model" to increase Evaluation of models: Neural Network Regression has a Coefficient of Determination of 0.98135 which is very good. Boosted Decision Tree Regression provides Coefficient of Determination is 0.984267 which is also very good.

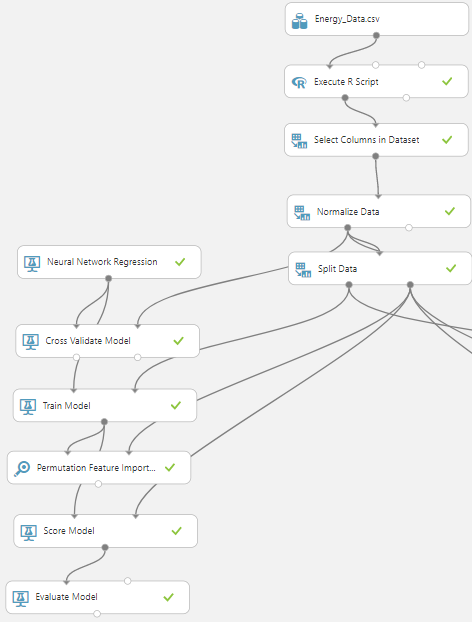
Below is the screen captures for both predictive models:

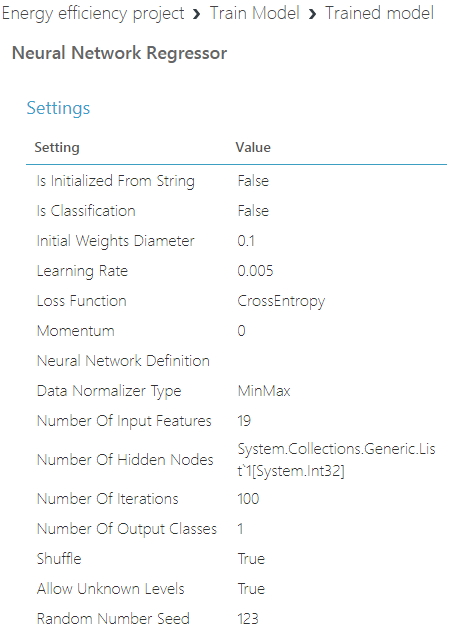


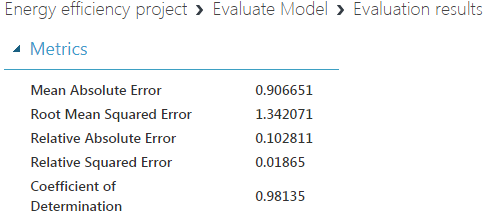
I have captured the select columns to train model on. I believe these features have direct impact on the HeatingLoad and Cooling Load:



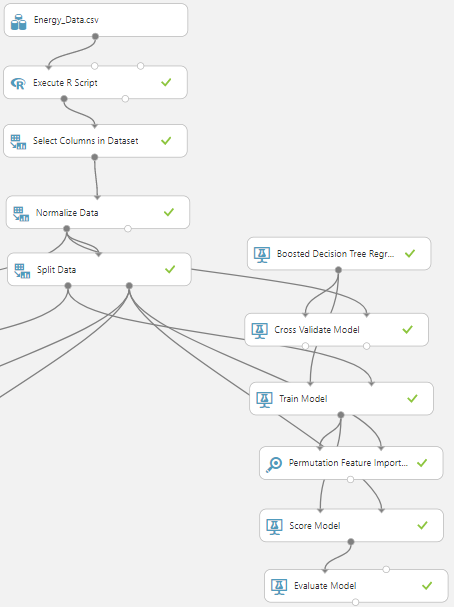
Below is the setting of the trained model for Neural Network Regression, predicting HeatingLoad values; in addition to model evaluation information.

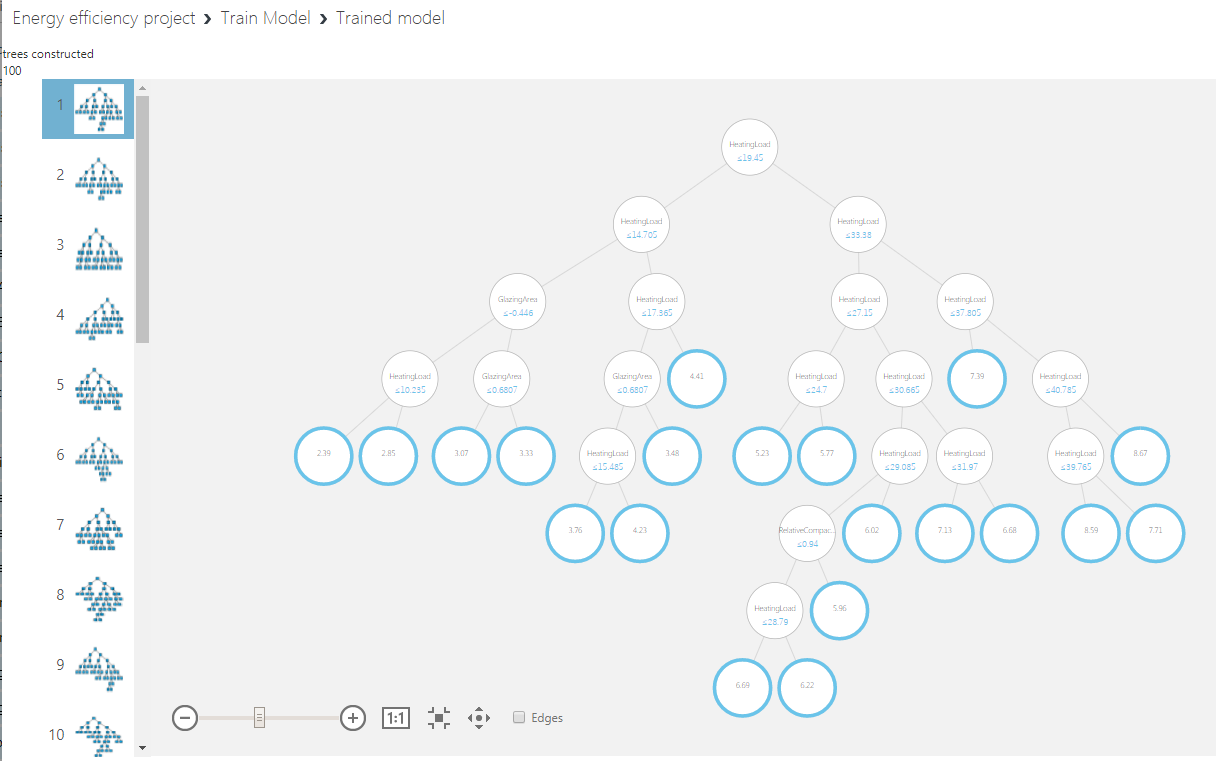


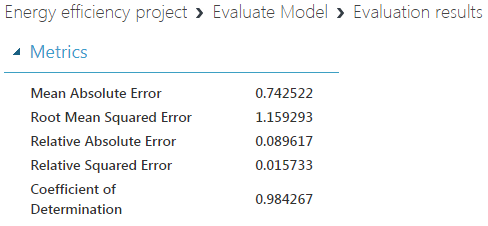




Below is the setting of the trained model for Boosted Decision Tree Regression, predicting CoolingLoad values, in addition to model evaluation information.







Overall, I have demonstrated Big Data related technology and include software demonstration using a Big Data source. I have learned a lot of interesting knowledge through this class. I appreciate your guidance throughout the course and thank you for your time. Happy holidays!