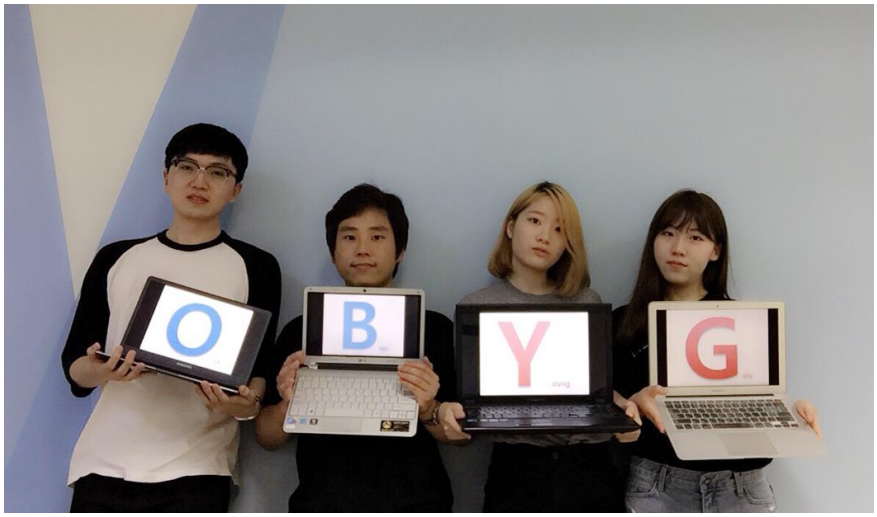
Group Project 3, Preliminary Analysis

:: Which factors affect to house energy efficiency? ::

OBYG – SeungHwan_Kim/ DongHyuk_Kim/YuJeong_Lee/Hyerin_Kim



Contents

1. How many unique observations to you have?

We have 2 independent variables and 8 independent variables. Each variable have 768 observations

2. What information/features/characteristics do you have for each observation?

There are totally 10 variables in the dataset. We change the concepts of variables into simple letters(x and y) in order to make more convenient analysis. X denote input variables and Y denote output variables.

Table 1. Variables and explanation

Dependent Variable	X1	Relative Compactness	12 types of value
	X2	Surface Area	12 types of value
	X3	Wall Area	7 types of value
	X4	Roof Area	4 types of value
	X5	Overall Height	2 types of value
	X6	Orientation	4 types of value
	X7	Glazing area	4 types of value
	X8	Glazing area distribution	6 types of value
Independent Variable	Y1	Heating Load	
	Y2	Cooling Load	

Heating Load(Y1) is the quantity of heat per unit time that must be supplied to maintain the temperature in a building or portion of a building at a given level(Merriam-Webster dictionary), and the Cooling Load(Y2) is total amount of heat energy that must be removed from a system by a cooling mechanism in a unit time, equal to the rate at which heat is generated by people, machinery, and processes, plus the net flow of heat into the system not associated with the cooling machinery. (McGraw-Hill Dictionary)

As independent variables we used the building characteristics about walls, floors, roofs and windows. Relative Compactness(X1) has 12 types of value. Surface Area(X2) has 12 types of value. Wall Area(X3) consists with 7 types of value. Roof Area(X4) has 4 types of value.
Overall Height(X5) has 2 types of values : 0, 1. Orientation(X6) has 4 types of value. 2) the value 2 means north oriented house, means 55% on the north side and 15% on each of the other sides. 3) east : 55% on the east side and 15% on each of the other sides. 4) south : 55% on the south side and 15% on each of the other sides, 5) west : 55% on the west side and 15% on each of the other sides. 1)

and the value 1 is uniform oriented house with 25% glazing on each side. Glazing(X7) areas expressed as the percentages of glazing areas. It has four types data: 0%, 10%, 25% and 40%. Glazing area distribution(X8) with 6 types of value : 0,1,2,3,4,5

3. What are the min/max/mean/median/sd values for each of these features? What is the distribution of the core features (show a histogram)?

Table 2. descriptive statistics of dataset

	x1	x2	x3	x4	y1	y2
median	0.75	673.75	318.5	183.75	18.95	22.08
max	0.98	808.5	416.5	220.5	43.1	48.03
min	0.62	514.5	245	110.25	6.01	10.9
mean	0.764167	671.7083	318.5	176.6042	22.3072	24.58776
sd	0.105777	88.08612	43.62648	45.16595	10.0902	9.513306

Figure 1. The distribution of the core features(output variables)

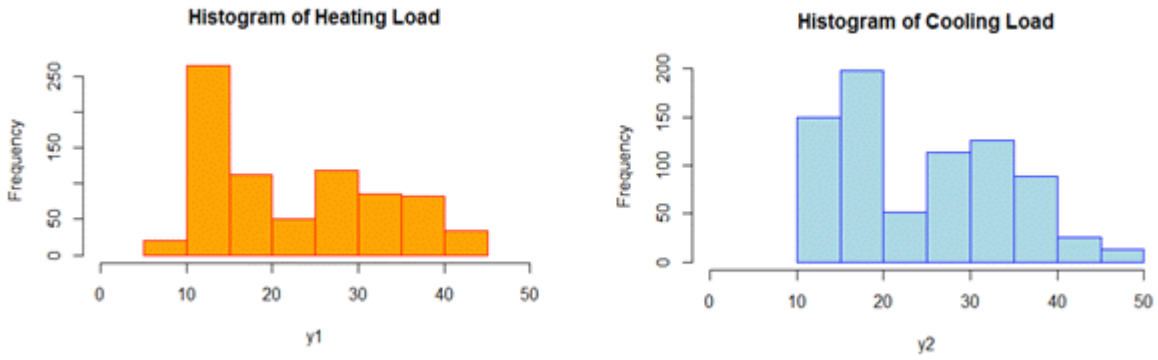
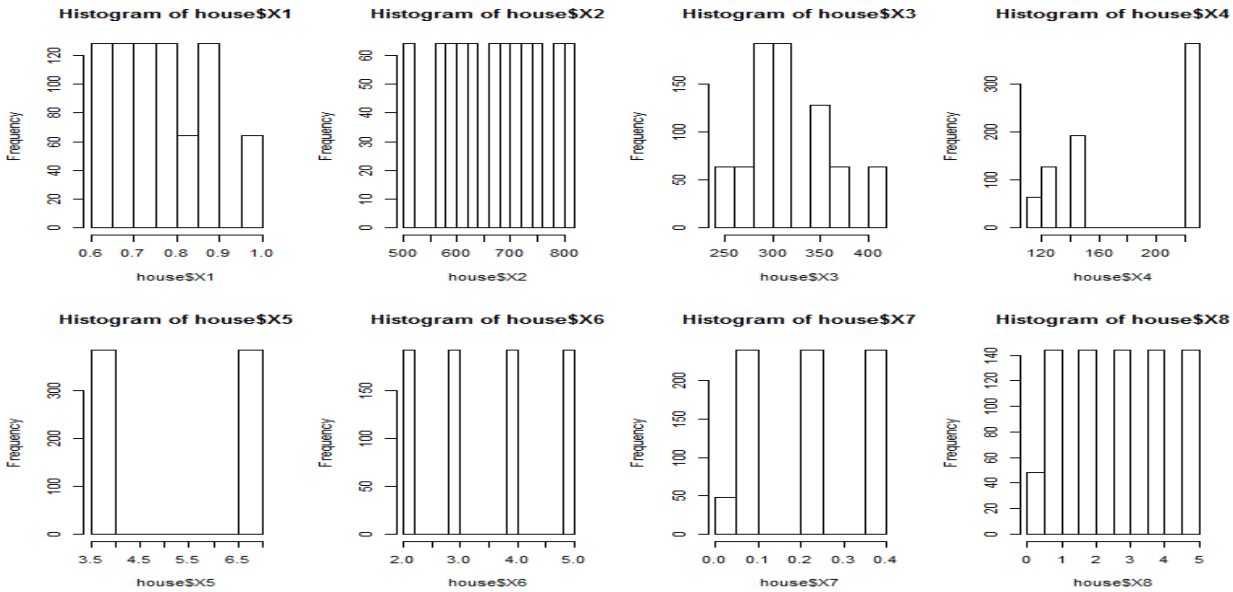


Figure 2. The distribution of the core features(input variables)

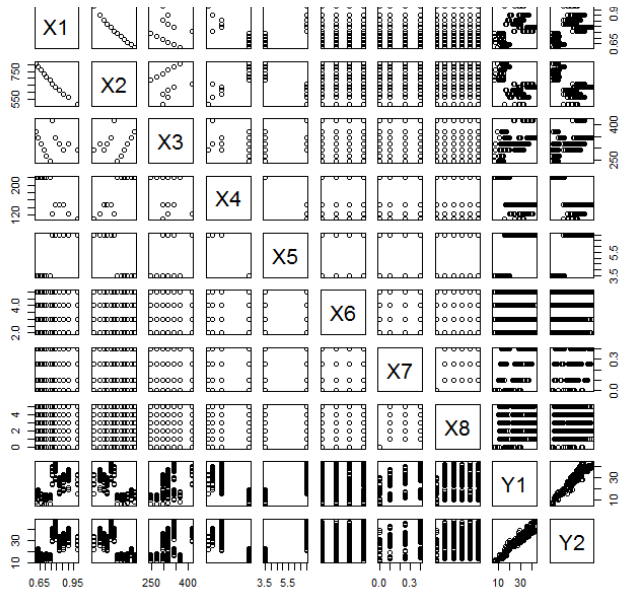


4. Are there obvious trends in the data (over time, across subgroups, etc.), and are the differences statistically significant?

We can find out trends in the certain data which is time series. A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.

But, our data is not a time series. As you can see, our data is collected in particular time. We cannot find out some trends or changes in our data.

Figure 3. Correlation between variables



By drawing a scatter plot, we can roughly know about the relationship of variables. As you can see, there is a linear relationship between Y1 and Y2. These are response variables. So, we can

understand that there is a high correlation between both of them. For the exact analyze, we make a decision to analyze only one response variable. The same method will be apply to the other response variable. Moreover, From X1 to X8, there is no linear relationship each other.

Because it is not time-series data, there are not trend.

5. What are the other salient aspects of the data (e.g. geospatial factors, text content, etc.)

We expect that X3(Wall Area) and X7(Glazing Area) has relatively powerful effect on Thermal load.

6. Provide a bullet-list of the next 5-10 tasks you will perform in analyzing your dataset.

We are going to perform ① Date cleaning, ② Correlation Analysis, ③ t-test above variables, ④ Predicting, ⑤ Regression Analysis.