Forecasting Electrical Consumption of Commercial Buildings in Singapore

40.000 Engineering Systems and Design Project

Group 8

1000906 Wang Yiran

1000909 Liu Sidian

1000931 Lee Jisu

1001132 Tong Keng Cheong Daniel

1001080 Kok Jin Xin Alexius

Executive Summary

In this project with Accenture, the factors of energy consumption in commercial buildings were analyzed to formulate linear regression models that allow the forecasting of monthly energy consumption. After studying the possible contributors to energy consumption, factors that affect the HVAC system were considered as independent variables. Data sets were either given by Accenture or gathered from reliable online public sources. To establish the regression models, Excel's Linear Regression Tool, RStudio and SPSS were utilized. To identify the significant variables, multicollinearity tests and statistical significance tests such F-test, p-value and t-statistics were conducted. The coefficients of the independent variables were also considered to verify if the variable logically fit into the model. The linear regression models showed weather related variables such as monthly temperature, humidity and rainfall played small roles in affecting the HVAC system; due to the fact that Singapore

has no distinct four seasons, resulting in the weather variables being generally constant. The size of the building and the number of people in the building were found to be the main contributors towards energy consumption. This study concluded that the properties of the building rather than the surroundings of the building were the main cause of the energy consumption for Building 1 (Mixed Development). However, Building 2 (Commercial Mall) results showed that further study was needed as the significant variables were artificial variables even though the linear equation was valid. This may be improved by transforming the variables. Additional data points will be also be necessary as the number of data points involved in this study were too few to validate the regression models. However, within the assumptions and scope of the project, the objective to develop the 9 regression models was achieved.

Introduction

Accenture is a leading global professional services company. It provides a broad range of services and consultation in strategy, consulting, security, technology and operations. Accenture deploys analytics to recommend solutions to problems proposed by the clients from different areas. The project that the team collaborated with Accenture was about energy consumption in commercial buildings located in Singapore. The main aim of the project was to develop linear regression models for two commercial buildings to forecast the energy consumption, given the values of the independent variables in the model. These models may be used in future to calculate the energy savings in a building by comparing the forecasted energy consumption against the actual energy consumption. Three regression models were developed for each building to predict the total electricity consumption, landlord's electricity consumption and tenant's electricity consumption. After constructing three disparate models for each building with satisfiable statistics significance test values, the team further generalized the linear model to accommodate two commercial buildings in one regression model by using a binary variable as a toggle. The project was approached by carefully selecting the independent variables based on the past studies conducted [2]. Among all the related literatures reviewed, three articles which were the most relevant served as primary references.

Data and Methods

Motivation for Specific Modelling Technique

The data provided by Accenture covered two buildings, each of which included GFA, NLA, Human Traffic, Occupancy Rate, Opening Hours and Electricity Consumption. Government sources provided the weather parameters such as monthly average temperature, rainfall, cooling degree days and humidity.

A previous study [2] utilized linear regression to predict a building's energy consumption. It was found that GFA was strongly significant, which made logical sense as energy consumption had to relate to the amount of space in some capacity. Based on the proven methods, the team intended to formulate a multivariable linear model to predict future electricity consumption. This is because a multivariable linear model can be easily applied, easily analyzed and easily modified. Ultimately, it allows the modelling of the electrical consumption to be generalized and thus can be applied to similar buildings. The use of a multivariable linear model also allowed the team to generate a residual plot to observe the linearity and normality of the data.

Methodology

Two types of decision variables were taken into the model: variables with values specific to each building and the variables with values common to all buildings (Refer to Appendix A). The variables that were building specific included GFA, NLA, Human Traffic, Occupancy Rate and Opening Hours. Variables with values common to both buildings were temperature, humidity, rainfall and Cooling Degree Days (CDD). These weather parameters were thought to be relevant as they relate to the energy usage in some way. The values of the variables common to all buildings were gathered from a reputable online source (Weather Underground).

To capture the seasonality, binary variables were used to account for the variables representing the months. 11 monthly binary variables were used while the final missing month acted as a basis and thus omitted from the regression. The month of April was deemed to the most typical month and hence was used as the basis. To reflect the overall trend of the data, the variable Time Series was used. The variable Time Series Squared was added to compensate for the large contribution to the dependent variable in the event of long term forecasts. The variable Cooling Degree Days(CDD) squared was added for the same reason to combat the large variations found in CDD.

The team attempted to transform the data so as to generate variables that would fit the linear regression better. Firstly, the variable Occupied Area values were calculated by multiplying the occupancy rate with net lettable area (NLA) to incorporate both data parameters into a single variable. Secondly, a dummy variable, testing, was added to check if the outliers contributed significantly to the regression model. Potential outliers were assigned the value of 1 while normal points were assigned the value of zero. Thirdly, for the combined regression model, a binary variable was used as a toggle (on/off) between the 2 building types. Building 1 was given a value of 1 while building 2 was given a value of 0. For the dependent variables, three types of electricity consumption were used depending on which variable the team was trying to forecast: landlord's electricity consumption, tenant's electricity consumption and total electricity consumption. Summation of the tenant's electricity consumption and landlord's electricity consumption were taken as the values for the total electricity consumption variable.

After setting up all the necessary variables which the team intended to use for the regressions, the team outlined their assumptions based on the prerequisites for running a linear regression with a particular data set. The following table indicates 6 criteria that needed to be assessed and passed before a linear regression could be validated. The first 3 were essential criteria to justify running a linear regression while the remaining assumptions determine the accuracy of the linear regression.

Assumptions	Criteria
Sample Size	For each independent variable in the model, there should be at least 10-20 independent data observations.
Linearity of Variables	All independent variables should relate to the dependent variable in a linear fashion. This can be verified through the use of individual scatterplots.
Multicollinearity	Every variable should be independent of each other and this can be verified through the use of

	Pearson's Correlation Analysis (Linearity) or Spearman's Rank Correlation Analysis (Monotonic Relationships).
Normality, linearity and Homoscedasticity of Residuals	After running the regression, the analysis of the residual plot can verify if a linear regression model is the best approach.
Normality of Variables	The accuracy of the coefficients for the independent variables in the regression model are determined by the normality of the variables when the sample size is small. This can be checked from the Stem & Leaf Plots, Q-Q Plots and Sharpiro-Wilk Tests.
Outliers in the Data	Outliers in the data might be detrimental to the linear equations as such points will affect the coefficients of the independent variables and the accuracy of the linear equation to a great extent.

Table 1: Assumptions for Linear Regression

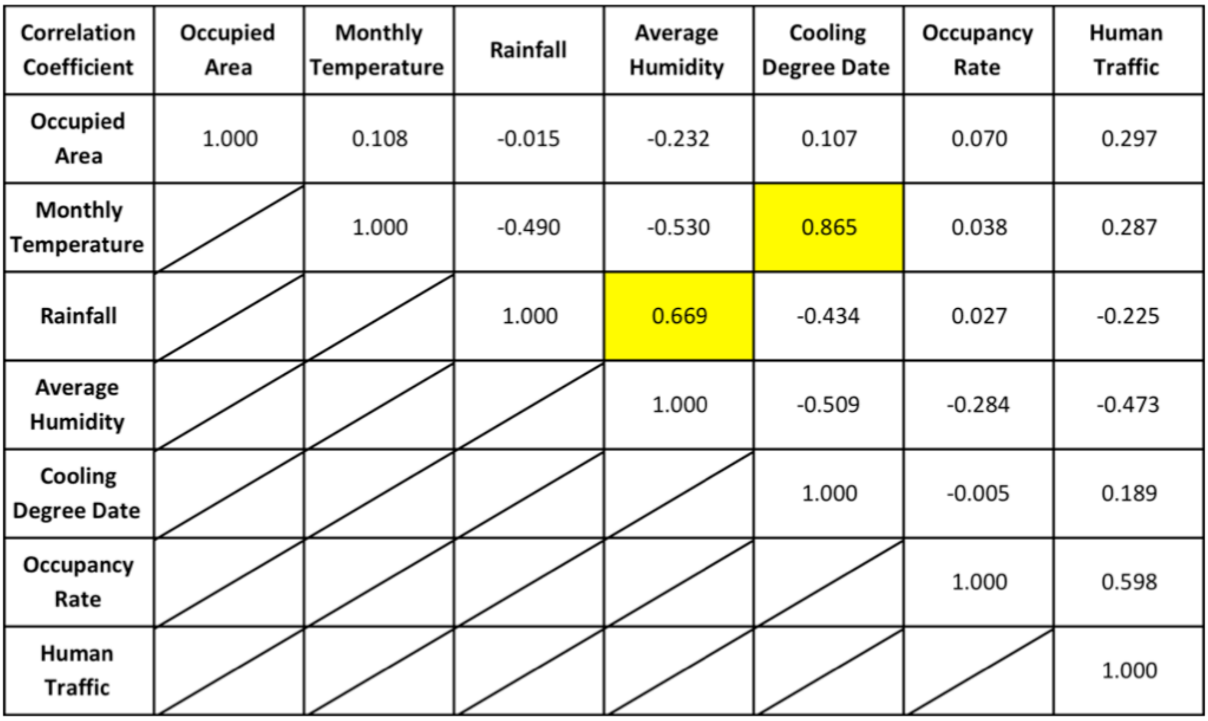
The team also used correlation analysis between the independent variables to further weed out variables that may affect the integrity of the multivariable linear model. The threshold of correlation above 0.65 suggested that the two variables were not independent. Subsequently, one of the correlated variables was removed or substituted after each iteration of the correlation analysis and the regression was re-generated to determine if the model has improved. By plotting the data against time, the team managed to determine if there was a clear trend and seasonality in the data sets through graphical analysis.

The graphical plots (Appendix B & C) allowed for ease of observation and preliminary analysis. Multiple linear regressions were run in Microsoft Excel, RStudio and SPSS to develop the final regression models that achieved the significant thresholds.

Significant variables were determined through statistical tests; based on the lower and upper bounds of the 95% confidence interval, t-Stats (greater than 2), p-value (less than 0.05) and sig-F (less than 0.1). For insignificant variables in each regression, a justification process based on 3 conditions dictate if they were kept or discarded. Firstly, if the variable contributed to a better R Square value for the linear regression, then the variable will be kept. Secondly, if the coefficient of the variable made logical sense with respect to how the variable relates to energy consumption, then the variable will be kept. Lastly, if they had a positive impact on the significance of the other variables, then the variable will be kept.

Results

The multicollinearity analysis was conducted for all independent variables except the artificial variables. The correlation coefficient matrix is shown in the table below. The coefficients that failed the threshold of 0.65 are highlighted in the table. There were two sets of correlated independent variables: Rainfall and Humidity, as well as Temperature and Cooling Degree Days. Each linear regression was conducted by selecting one of the two highly correlated variables within each set and using them as an independent variables. The linear regressions were then compared and the final regressions chosen were those that had the highest R Square value and number of significant variables.



Correlation Coefficient	Occupied Area	Monthly Temperature	Rainfall	Average Humidity	Cooling Degree Date	Occupancy Rate	Human Traffic
Occupied Area	1.000	0.108	-0.015	-0.232	0.107	0.070	0.297
Monthly Temperature		1.000	-0.490	-0.530	0.865	0.038	0.287
Rainfall			1.000	0.669	-0.434	0.027	-0.225
Average Humidity				1.000	-0.509	-0.284	-0.473
Cooling Degree Date					1.000	-0.005	0.189
Occupancy Rate						1.000	0.598
Human Traffic							1.000

Table 2: Correlation Matrix

Analysis of the multivariable linear regression model was conducted for both buildings individually. The combination of independent variables chosen for final regression models was based on the R Square value, the number of significant variables, the degree of significance for each variable as well as a logical sign of the coefficients of the insignificant variables. For example, humidity was insignificant in the regression model for Building 1, Tenant's Electricity, but it was kept in the model as the coefficient was positive. This is justifiable; higher humidity results in higher energy consumption to dehumidify the air, outweighing the savings due to the reduction in temperature. Monthly binary variables and time series variables were always included in the linear regressions in order to capture the seasonality and the trend of the dependent variable.

The results for 9 linear regression models are shown Table 3 (Appendix F), where the coefficients of independent variables and R squared values are included. Human Traffic was found to be the significant variable for all three categories of electrical consumption while Occupied Area was found to be the significant variable for all three categories of electrical consumption in the combined building regression. In addition, the variable, Testing, was strongly significant in the models with Building 1 landlord electricity data involved. This proved the existence of outliers through the statistical analysis and supported by the graphical analysis of the data. The landlord and total electricity usage for January 2012 and November 2012 were highly abnormal compared to the rest of the data. (Appendix D)

Analysis and Discussion

In an effort to build a regression model that accounted for variation in electricity usage accurately, the team attempted to involve all possible factors and tried different combinations and transformations of variables. From the team's observation of the result, GFA, Human Traffic and Occupied Area were the main factors that contributed to electricity consumption. These variables formalised the team's expectation that the size and crowd do matter significantly when attempting to model the energy consumption of a building. Secondly, the impact of temperature and Cooling Degree Days were minimal due to the relatively constant temperature in Singapore throughout the year. The climate remains almost invariable, with monthly temperature variation of 4 degree celsius and changes of humidity within a 5% margin.

Models for building 2 pertaining to the total and landlord electricity usage were less reliable due to the lack of significant variables aside from the artificial variables. The team's analysis showed that this resulted from the violated assumptions of a linear relationship between the dependent variable and every independent variable. From the scatter plot (Appendix E) of energy consumption against "Occupied Area" and "Human Traffic" respectively, a linear pattern was not found while linearity is the underlying requirement for a regression model. Hence, transformation of variables could have resulted in a better fit. However, since there were no observable pattern in scatter plots, satisfiable transformations with stronger significance were not found and the transformations attempted failed to improve the model significantly.

The number of data points provided by Accenture was less than the minimum number of data points required. Hence the first assumption for sample size was violated and more data should be collected to validate and improve the model.

Referring to Appendix (E), for the variables of occupied area, the Shapiro-Wilk test Significant values were less than 0.05. This indicated that the variables were not normally distributed. The stem-and-leaf plot further supports this conjecture as the plots were not roughly symmetrical. Hence the assumption of the normality was violated. However the significance of violating this assumption only affected the accuracy of the coefficients in the linear equations generated.

Conclusion and Future Works

The objective of the team in this project was to develop nine regression models - three regression models each for two buildings and combined data of two buildings. Each regression model can calculate the total electricity consumption, the landlord's electricity consumption or the tenant's electricity consumption. To generate nine linear regression models, factors affecting the energy consumption of HVAC in commercial building were considered as independent variables. Different regression programs were used to come up with the linear regression models and these models were counter-checked studiously amongst the different outputs from these programs and the results are checked with various statistical significance test values. Among the 9 regression models, 7 models were considered highly reliable and robust enough to predict future electricity usage with strong significance level and high R square. However, the 2 regression models for Building 2 failed to show linear patterns, this makes the model unreliable thus usage should proceed with caution. Improvements can be made through an extension into nonlinear modeling and analysis. Moreover, limited data points and data collection error may also be considered as the main contributors to the inaccuracy of Building 2 models. Larger sample size and accurate data should result in stronger model for all buildings.

Reference

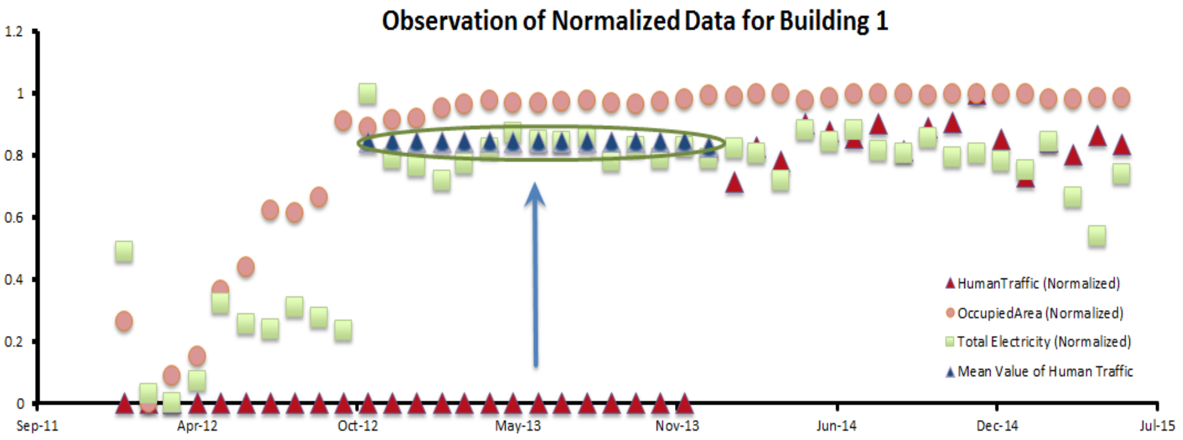
1. Lam, Joseph C., and Danny H.W. Li. 'Electricity Consumption Characteristics In Shopping Malls In Subtropical Climates'. *Energy Conversion and Management* 44.9 (2003): 1391-1398
2. . NATIONAL UNIVERSITY OF SINGAPORE, 2006.
3. Building Construction Authority,. *BCA Building Energy Benchmark Report*. Singapore. 2014.

Appendix

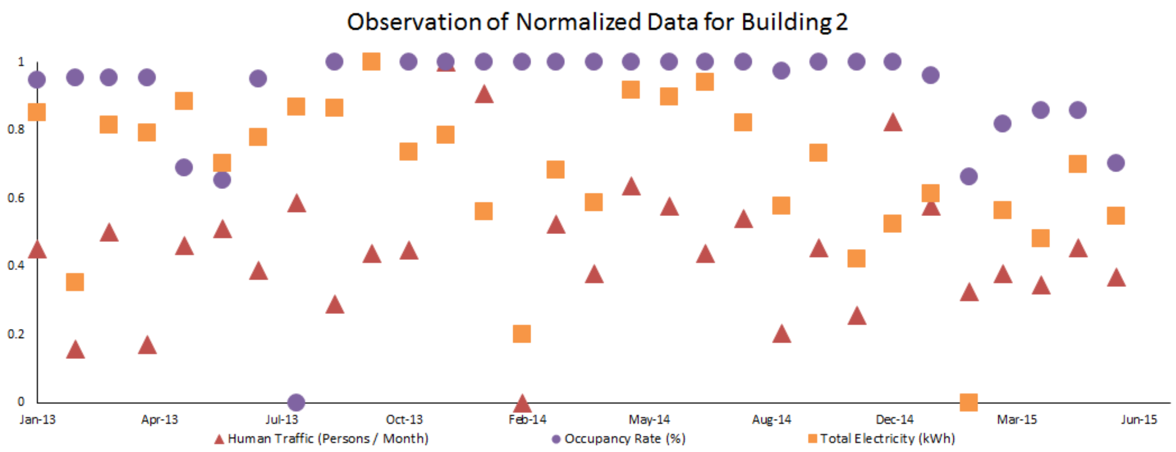
A. Independent Variable Justification Table

What are the variables?	Justification
Occupied Area	Larger area requires more energy for the cooling system to maintain the internal building temperature.
Human Traffic	Increase in number of people suggests more energy required for the HVAC system to combat body heat.
Rainfall	Increase in rainfall will reduce the temperature and but increase humidity. It is expected that the effect of increasing humidity outweighs the reduction in temperature.
Humidity	Increase in humidity level means more energy is required for ventilation systems to dehumidfy the air.
Temperature	Increase in outside temperature means the building will use more energy to cool the interior.
Cooling Degree Days	Increase in CDD indicates bigger tempreature difference between the interior and exterior, thus affect cooling system significantly.
Time Series	Used to capture the overall trend in the data.
Testing Variables	Used to test the significance of the outliers in the data
Binary Variable for Building	Used to denote the type of the building in combined regression model
Monthly Binary Variable	Used to capture seasonality in the data.

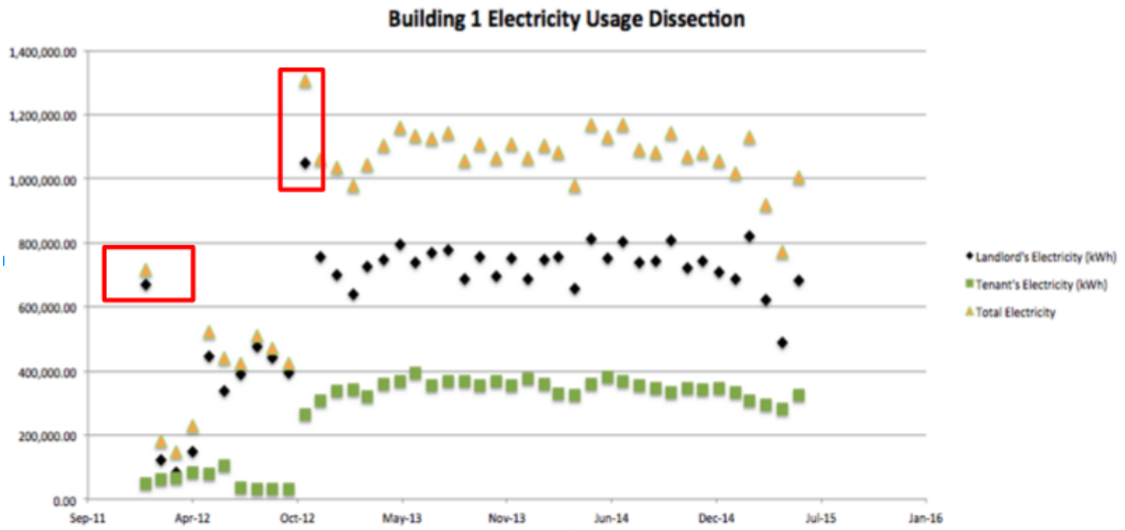
B. Graphical Interpretation of Data (Building 1)



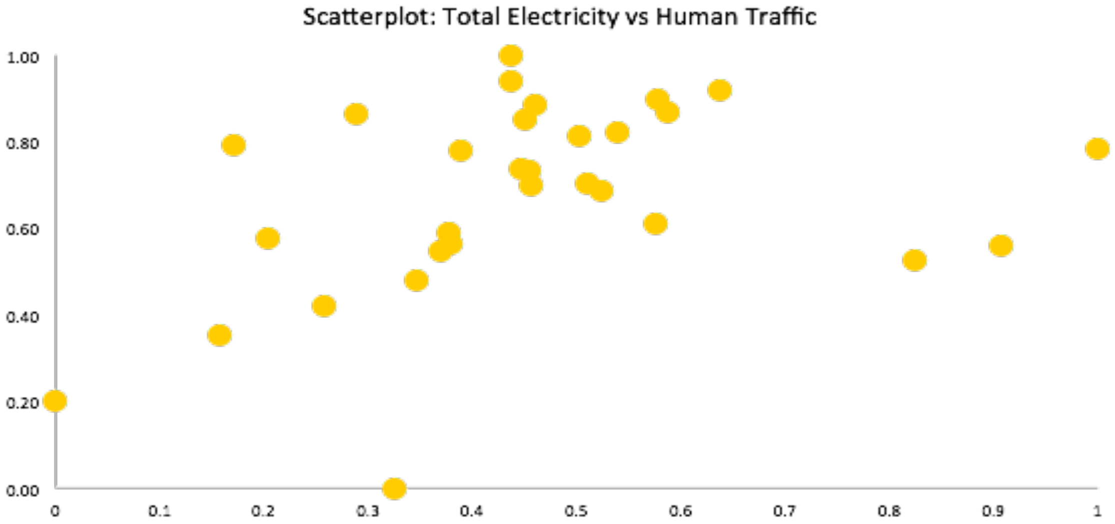
C. Graphical Interpretation of Data (Building 2)

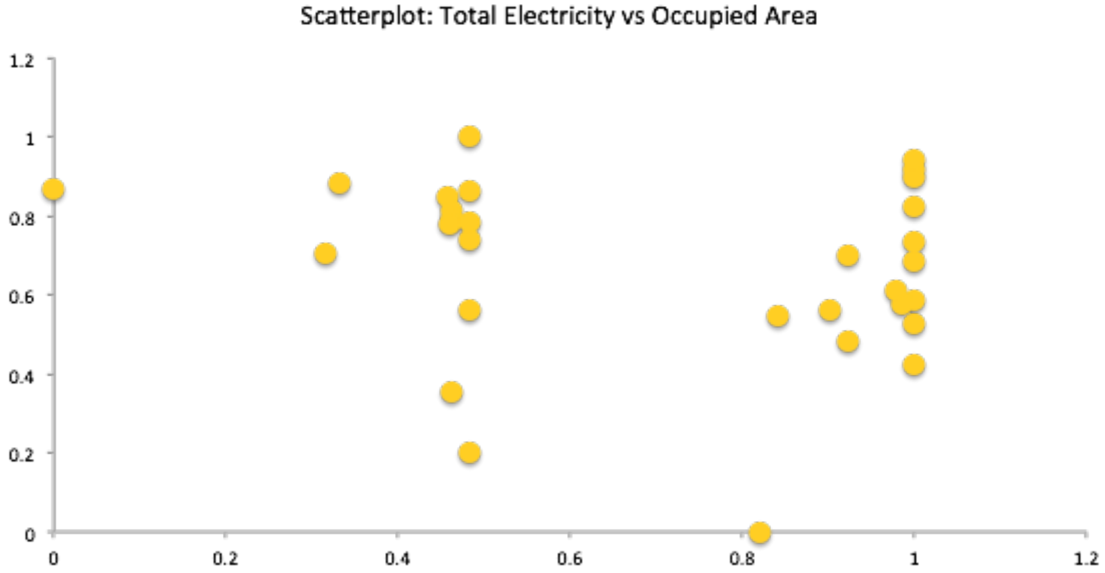


D. Outlier Identification Graph (Building 1)

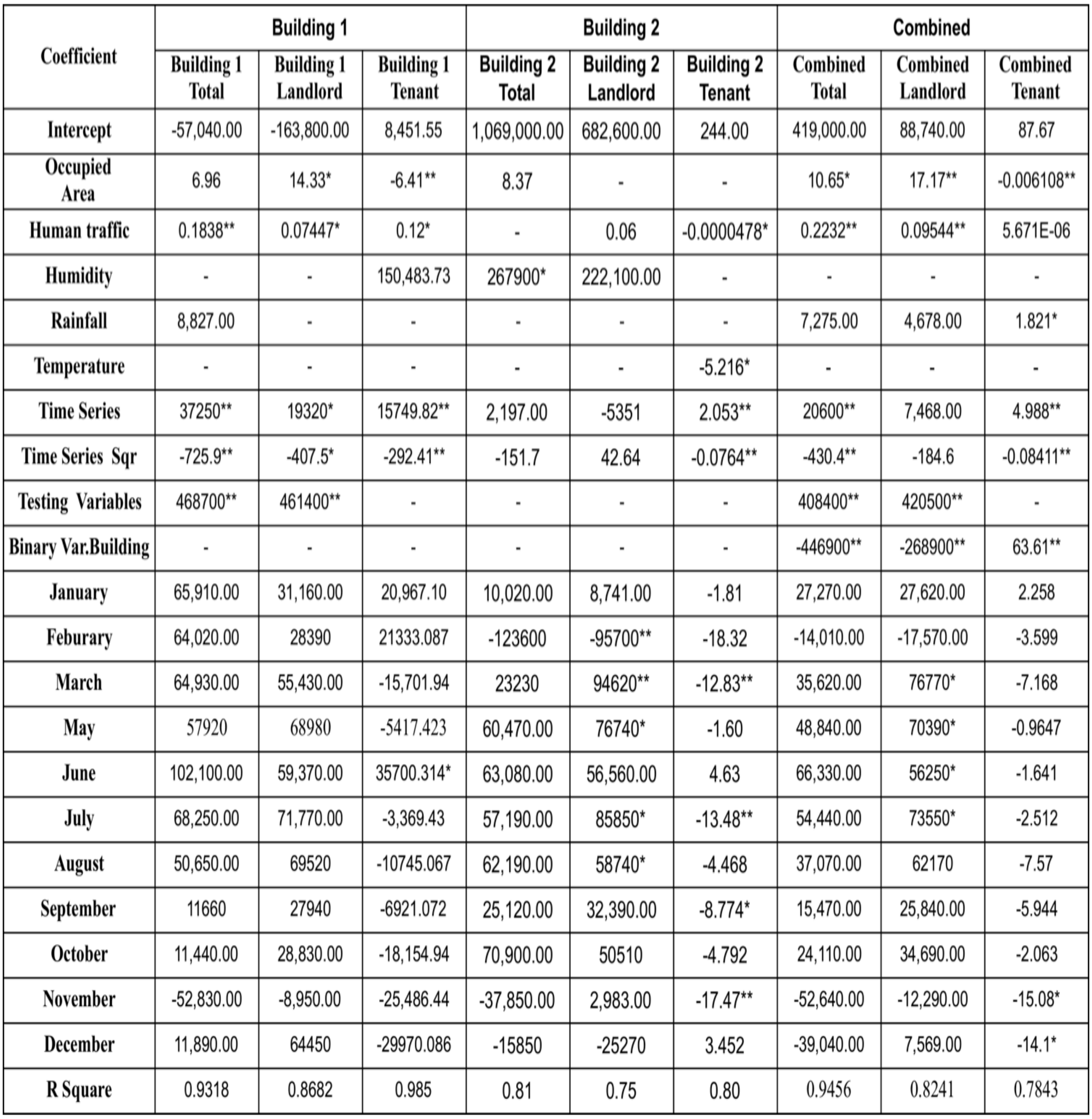


E. Scatter Plot of Electricity Usage vs Independent Variable





F. Final Results for Nine Regression Models



Coefficient	Building 1			Building 2			Combined		
	Building 1 Total	Building 1 Landlord	Building 1 Tenant	Building 2 Total	Building 2 Landlord	Building 2 Tenant	Combined Total	Combined Landlord	Combined Tenant
Intercept	-57,040.00	-163,800.00	8,451.55	1,069,000.00	682,600.00	244.00	419,000.00	88,740.00	87.67
Occupied Area	6.96	14.33*	-6.41**	8.37	-	-	10.65*	17.17**	-0.006108**
Human traffic	0.1838**	0.07447*	0.12*	-	0.06	-0.0000478*	0.2232**	0.09544**	5.671E-06
Humidity	-	-	150,483.73	267900*	222,100.00	-	-	-	-
Rainfall	8,827.00	-	-	-	-	-	7,275.00	4,678.00	1.821*
Temperature	-	-	-	-	-	-5.216*	-	-	-
Time Series	37250**	19320*	15749.82**	2,197.00	-5351	2.053**	20600**	7,468.00	4.988**
Time Series Sqr	-725.9**	-407.5*	-292.41**	-151.7	42.64	-0.0764**	-430.4**	-184.6	-0.08411**
Testing Variables	468700**	461400**	-	-	-	-	408400**	420500**	-
Binary Var.Building	-	-	-	-	-	-	-446900**	-268900**	63.61**
January	65,910.00	31,160.00	20,967.10	10,020.00	8,741.00	-1.81	27,270.00	27,620.00	2.258
Feburary	64,020.00	28390	21333.087	-123600	-95700**	-18.32	-14,010.00	-17,570.00	-3.599
March	64,930.00	55,430.00	-15,701.94	23230	94620**	-12.83**	35,620.00	76770*	-7.168
May	57920	68980	-5417.423	60,470.00	76740*	-1.60	48,840.00	70390*	-0.9647
June	102,100.00	59,370.00	35700.314*	63,080.00	56,560.00	4.63	66,330.00	56250*	-1.641
July	68,250.00	71,770.00	-3,369.43	57,190.00	85850*	-13.48**	54,440.00	73550*	-2.512
August	50,650.00	69520	-10745.067	62,190.00	58740*	-4.468	37,070.00	62170	-7.57
September	11660	27940	-6921.072	25,120.00	32,390.00	-8.774*	15,470.00	25,840.00	-5.944
October	11,440.00	28,830.00	-18,154.94	70,900.00	50510	-4.792	24,110.00	34,690.00	-2.063
November	-52,830.00	-8,950.00	-25,486.44	-37,850.00	2,983.00	-17.47**	-52,640.00	-12,290.00	-15.08*
December	11,890.00	64450	-29970.086	-15850	-25270	3.452	-39,040.00	7,569.00	-14.1*
R Square	0.9318	0.8682	0.985	0.81	0.75	0.80	0.9456	0.8241	0.7843

Table 3 Nine Regression Models (** Strongly Significant *Significant)