CS506 Project Final Report

Topic: Mapping the Intensity of Energy Use Across Boston University
Group Members: Chengyu Deng, Xiaotong Niu, Qian Zhang
Due Date: May.2nd, Wednesday

Project Description

Well known for its extreme weathers, the city of Boston apparently neutralizes outdoor and indoor temperature with high energy consumption, either in winter by using a large amount of gas for heating or in summer by using a large amount of electricity for cooling. As a University located in the center of Boston, Boston University indeed follows the energy consumption pattern of the city of Boston. In this project, we are dedicated to analyzing electricity and gas consumption data for buildings at BU, and then use Linear Regression model, Quadratic Nonlinear Regression model and Logistic Regression model to see how much weather contributes to energy consumptions and find factors that largely contribute to energy consumptions besides weather among the choices of water usage intensity, greenhouse gas emission intensity, building age and property type. In the end, we expect to have several factors that contribute to energy consumption the most and suggest our project partner at Sustainability@BU to take action in order to reach the goal of Carbon Neutral in the future.

Based on historical records and common senses, we made some assumptions that weather contributes the most to both gas and electricity consumption while the effect of other factors remains unknown. So far, after a great amount of effort of data cleaning and method implementing, we have obtained some conclusions that besides temperature, greenhouse gas emission and building age affect a buildings' electricity and gas energy consumption, while building types affect buildings' gas energy consumption only.

Data Description

We came along a distant way of getting feasible datasets and using them. Initially, we planned to use the building energy consumption data directly monitored by the University. However, we were told that the ideal dataset, which contains nearly everything we might need on a monthly basis, was not accessible to us due to privacy reasons. Hence, we need to find data from somewhere else. In order to get building data of BU, we were suggested to use Building Energy Reporting and Disclosure Ordinance (BERDO) dataset provided by Analyze Boston. Being used as our main dataset, BERDO contains reported buildings' energy usage data and other useful information of the year 2015, 2016 and 2017 in Boston area. To be more specific, in each year's BERDO dataset, it has around 1500 data entries(buildings); each entry contains information such as site Energy Use Intensity(EUI), the percentage of electricity/natural gas/steam used, building age, property usage etc.

Besides BERDO dataset, there are three more minor datasets we made use of for getting reasonable data regarding temperature, water intensity distribution by month, and greenhouse gas emission distribution by month. For temperature, we extracted monthly average temperatures' data of years 2015, 2016 and 2017 from the website U.S. Climate Data. Then, for water intensity distribution by month, we made use of monthly distribution percentage listed in water usage reports for each month from the website MWRA Online. And for greenhouse gas emission distribution by month, we used the CO2 emission monthly distribution from Earth

System Research Library (ESRL) since the data of greenhouse gas emission listed in the BERDO dataset only counted specifically CO2 emission according to Data Dictionary of BERDO.

The first thing we did was to find out BU properties and their corresponding information in the BERDO dataset. In order to achieve this goal, we combined both BERDO Data Map and Boston University Map to select buildings manually. Then we filtered buildings that don't belong to BU campus and finally we obtained around 40 buildings' entries in the datasets in three years. Then we are suggested to convert both yearly based electricity and gas consumption data in BERDO to a monthly basis by using the distribution of monthly consumption of electricity and gas provided by Sustainability@BU. To achieve this, we did the following things:

- We used Site Energy Use Intensity (EUI, unit: kBTU/sq) as total energy consumption containing electricity, gas, and steam(we deleted steam because, according to BERDO, there is only a very small amount of buildings in BU that reported steam percentage)
- Calculate the energy use for electricity and gas by using EUI, the energy consumption percentage of electricity and gas. (Electricity Consumption = EUI * electricity percentage, Gas Consumption = EUI * gas percentage)
- Calculate the monthly energy consumption of electricity and gas by using the distribution provided by Sustainability@BU

Next, we proceeded to the most important part, which is to extract the independent factors from the BERDO that we would like to inspect and concatenate them with temperature data, monthly water intensity distribution data and CO2 emission monthly distribution data that we got from U.S. Climate Data, MWRA Online and ESRL. The factors we selected are:

- Monthly Average Temperature in Boston
- Water Intensity (the total water usage divided by gross floor area)
    - In order to get water intensity data for each month, we multiplied monthly distribution of water usage provided by MWRA Online with annual water intensity in order to achieve the consistency of using monthly data.
    - For the purpose of Quadratic Nonlinear Regression, we squared the results we got from previous steps.
- Greenhouse Gas Emission Intensity (greenhouse gas emission divided by gross floor area)
    - We did similar action with greenhouse gas emission intensity as we did for water intensity. That is, we multiplied monthly distribution of CO2 gas emission provided by ESRL with annual greenhouse gas intensity in order to achieve the consistency of using monthly data.
    - Also, for the purpose of Quadratic Nonlinear Regression, we also squared the results we got from previous steps.
- Building Age (the year in which the building is built)
- Property Type (the purpose of the building)

Upon finishing of processing data, we discarded a few buildings that have water intensity and greenhouse gas emission intensity values as either 0(zero), None or inf(infinity). Since there are only a few of them, deleting would not make any significant differences in our final results.

Approaching to the end of data cleaning, we stacked up data we processed earlier from the year of 2015 to the year of 2017 and made each data entry under monthly basis so that each building can have multiple entries in the dataset. This can be considered as a data augmentation technique since the more data we use in our models, the more accurate prediction we will get.

Until now, the datasets are ready to use for training both Linear Model and Quadratic Nonlinear Model. For the Logistic Model, however, we further mapped electricity and gas usage into 5 categories(levels): very low usage, low usage, medium usage, high usage, and very high usage since a Multi-class Logistic Model predicts the probability of discrete categories. We achieved doing this by the following:

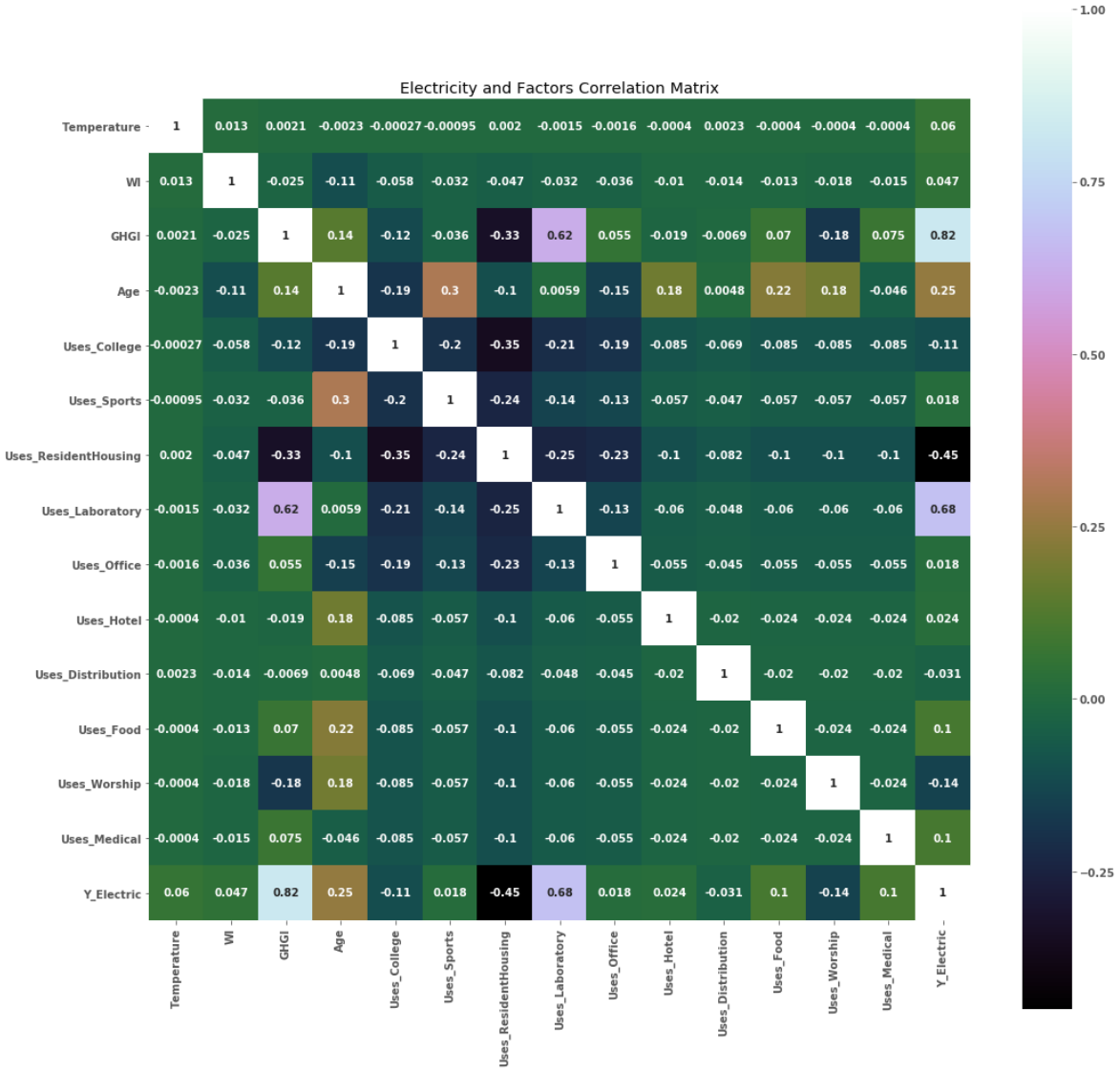
- Find out the minimum and the maximum value for all of the data entries of electricity and gas usage
- Divide the difference between the maximum value and the minimum value evenly into 5 parts
- Categorize them into 5 ranges and mapped usages into these 5 ranges accordingly

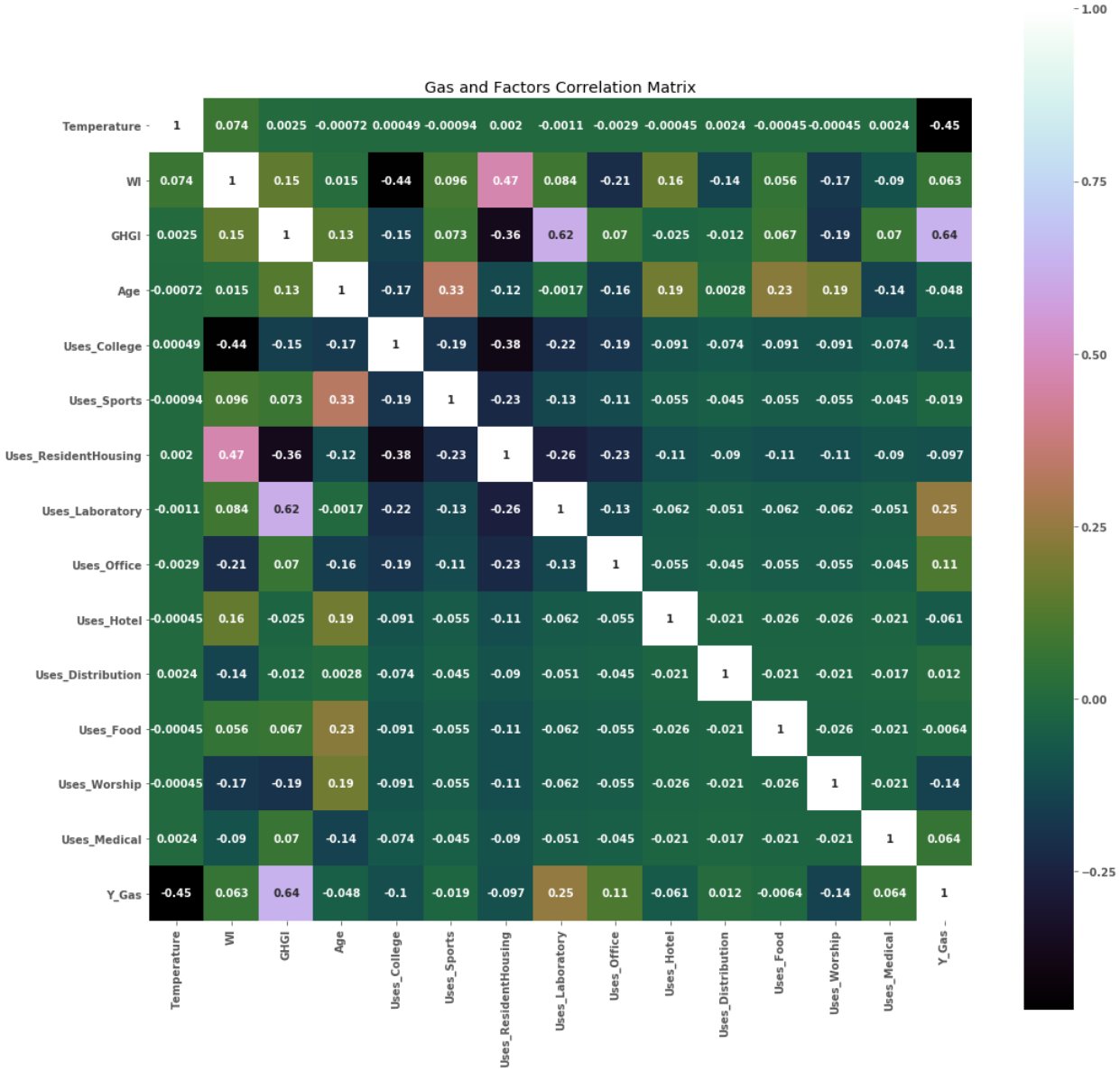
Finally, we created two files to store the cleaned and ready-to-use data, one for electricity and one for gas, under the names **df_201x_BU_analysis_E_Final.csv** and **df_201x_BU_analysis_G_Final.csv**.

Data Analysis

After cleaning the data, we did several data analysis by visualization.

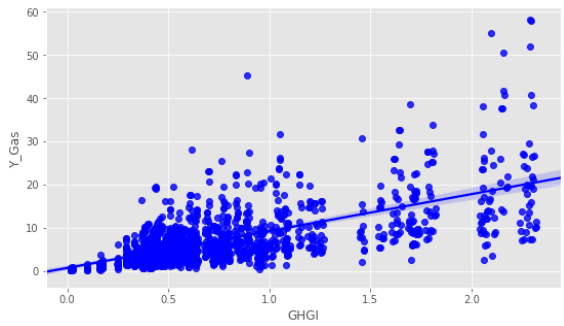
First of all, to find the relationship between factors for both electricity consumption and gas consumption, we made two heatmaps of Pearson's Correlation for our data.

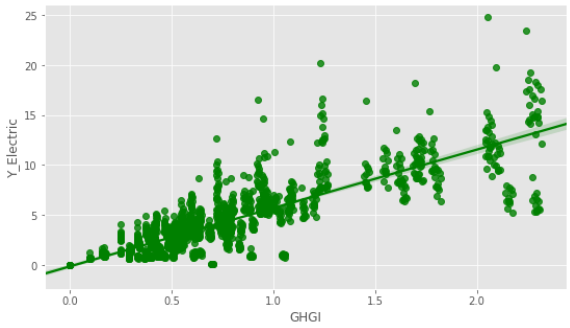

Electricity and Factors Correlation Matrix
Temperature 1 0.013 0.0021 -0.0023 -0.00027 -0.00095 0.002 -0.0015 -0.0016 -0.0004 0.0023 -0.0004 -0.0004 -0.0004 0.06
WI 0.013 1 -0.025 -0.11 -0.058 -0.032 -0.047 -0.032 -0.036 -0.01 -0.014 -0.013 -0.018 -0.015 0.047
GHGI 0.0021 -0.025 1 0.14 -0.12 -0.036 -0.33 0.62 0.055 -0.019 -0.0069 0.07 -0.18 0.075 0.82
Age -0.0023 -0.11 0.14 1 -0.19 0.3 -0.1 0.0059 -0.15 0.18 0.0048 0.22 0.18 -0.046 0.25
Uses_College -0.00027 -0.058 -0.12 -0.19 1 -0.2 -0.35 -0.21 -0.19 -0.085 -0.069 -0.085 -0.085 -0.085 -0.11
Uses_Sports -0.00095 -0.032 -0.036 0.3 -0.2 1 -0.24 -0.14 -0.13 -0.057 -0.047 -0.057 -0.057 -0.057 0.018
Uses_ResidentHousing 0.002 -0.047 -0.33 -0.1 -0.35 -0.24 1 -0.25 -0.23 -0.1 -0.082 -0.1 -0.1 -0.1 -0.45
Uses_Laboratory -0.0015 -0.032 0.62 0.0059 -0.21 -0.14 -0.25 1 -0.13 -0.06 -0.048 -0.06 -0.06 -0.06 0.68
Uses_Office -0.0016 -0.036 0.055 -0.15 -0.19 -0.13 -0.23 -0.13 1 -0.055 -0.045 -0.055 -0.055 -0.055 0.018
Uses_Hotel -0.0004 -0.01 -0.019 0.18 -0.085 -0.057 -0.1 -0.06 -0.055 1 -0.02 -0.024 -0.024 -0.024 0.024
Uses_Distribution 0.0023 -0.014 -0.0069 0.0048 -0.069 -0.047 -0.082 -0.048 -0.045 -0.02 1 -0.02 -0.02 -0.02 -0.031
Uses_Food -0.0004 -0.013 0.07 0.22 -0.085 -0.057 -0.1 -0.06 -0.055 -0.024 -0.02 1 -0.024 -0.024 0.1
Uses_Worship -0.0004 -0.018 -0.18 0.18 -0.085 -0.057 -0.1 -0.06 -0.055 -0.024 -0.02 -0.024 1 -0.024 -0.14
Uses_Medical -0.0004 -0.015 0.075 -0.046 -0.085 -0.057 -0.1 -0.06 -0.055 -0.024 -0.02 -0.024 -0.024 1 0.1
Y_Electric 0.06 0.047 0.82 0.25 -0.11 0.018 -0.45 0.68 0.018 0.024 -0.031 0.1 -0.14 0.1 1
Temperature
WI
GHGI
Age
Uses_College
Uses_Sports
Uses_ResidentHousing
Uses_Laboratory
Uses_Office
Uses_Hotel
Uses_Distribution
Uses_Food
Uses_Worship
Uses_Medical
Y_Electric
1.00
0.75
0.50
0.25
0.00
−0.25



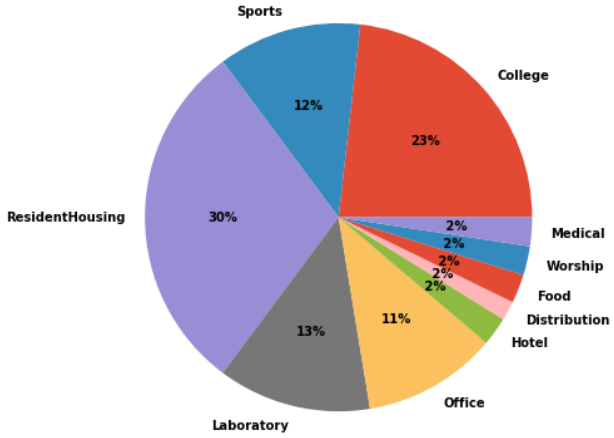
The correlation between greenhouse gas emission and electricity consumption is 0.82, which is relatively high compared to other correlations. Similarly, the correlation between greenhouse gas emission(GHGI) and gas consumption is also higher than the most of other correlations, which lead us to make an attention to GHGI and assume it to be a potentially significant factor on energy consumption.

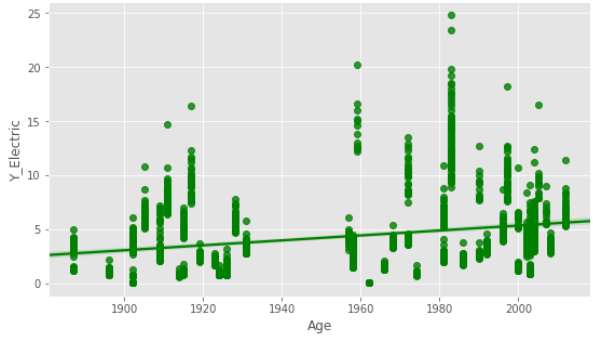
Then we plotted the trend of relationship for both gas consumption and electricity consumption with GHGI. Through the plots, we can clearly find out that the relationship between energy consumption and GHGI is positive.

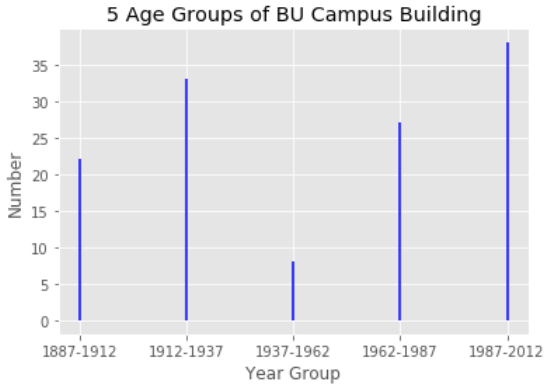




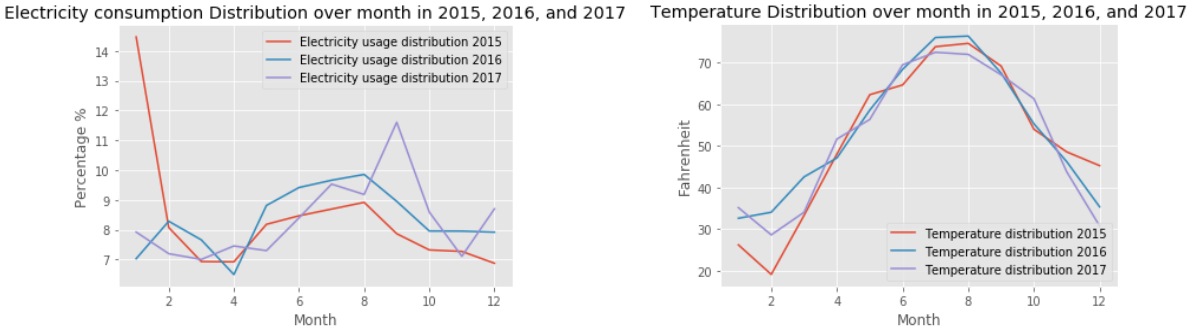
Type of building is another factor that we assume important to energy consumption. In correlation heatmap matrix, the correlation between electricity consumption and laboratory is 0.68, which interests us to explore the relationship between the type of building and electricity consumption in training model.



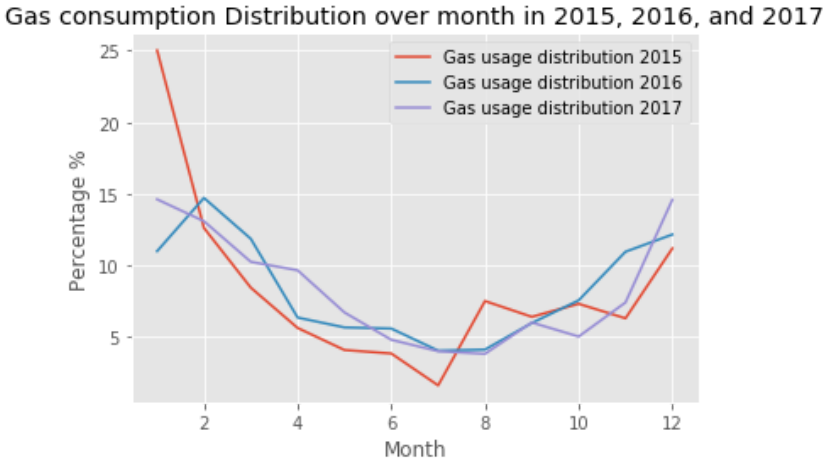




To our surprise, the relationship between age and electricity consumption is positive, which means a building with younger age consumes more electricity than a building with older age.



Temperature distribution follows the change of seasons, and as the temperature increases, the consumption of gas declines due to less use of heating, and gas is the main energy source of heating furnaces.



Methodologies and Algorithms

The approaches we use to predict the importance of factors are Multinomial Linear Regression, Quadratic Nonlinear Regression, and Multinomial Logistic Regression. We train all models through our prepared datasets, compare the performances and then select a specific model that shows the best performance among all and analyze the importance of the independent factors through it.

In order to construct the models, we use the following data as factors that potentially contribute to the BU electricity/gas consumption:

- x_{TEMP} : Monthly Average Temperature in Boston
- x_{WI} : Monthly Water Intensity per square feet of the Building
- x_{GHGI} : Monthly Greenhouse Gas Emission Prediction of the building
- x_{AGE} : The Construction Year of the Building

The following variables are binary variables which indicate whether the building is in a certain type of use.

- x_{COL} : A variable indicating whether the building is for college teaching purpose (classrooms, libraries etc.)
- x_{SPORT} : A variable indicating whether the building is for sports/fitness purpose (stadium, gyms, etc.)
- x_{RES} : A variable indicating whether the building is a university residential hall.
- x_{LAB} : A variable indicating whether the building is a laboratory.
- x_{OFF} : A variable indicating whether the building is an office building.
- x_{HOT} : A variable indicating whether the building is a hotel.
- x_{DIS} : A variable indicating whether the building is a distribution center.
- x_{FOOD} : A variable indicating whether the building is a food service building.
- x_{WOR} : A variable indicating whether the building is a worship/multi-purpose building.
- x_{MED} : A variable indicating whether the building is a medical purpose building.

The reason why we choose ten variables as an indicator vector to express the using purpose of a building is that in this way, after running the regression, we can directly observe which type of usage of the building will impact the electricity/gas consumption and how significant would a specific type impact the energy usage by simple observe the value of its weight, while only using one variable to indicate all the types fails to do that.

Multinomial Linear Regression:

Being as the most basic and interpretable model to classify and predict our data, Linear Regression is the first model we try to fit our data and figure out the appropriateness of this implementation. We define the following two linear functions of electricity and gas consumption:

$$Y'_E = w_0 + w_1 x_{TEMP} + w_2 x_{WI} + w_3 x_{GHGI} + w_4 x_{AGE} + [w_5 x_{COL} + w_6 x_{SPORT} + w_7 x_{RES} + w_8 x_{LAB} + w_9 x_{OFF} + w_{10} x_{HOT} + w_{11} x_{DIS} + w_{12} x_{FOOD} + w_{13} x_{WOR} + w_{14} x_{MED}]$$

$$Y'_G = \theta_0 + \theta_1 x_{TEMP} + \theta_2 x_{WI} + \theta_3 x_{GHGI} + \theta_4 x_{AGE} + [\theta_5 x_{COL} + \theta_6 x_{SPORT} + \theta_7 x_{RES} + \theta_8 x_{LAB} + \theta_9 x_{OFF} + \theta_{10} x_{HOT} + \theta_{11} x_{DIS} + \theta_{12} x_{FOOD} + \theta_{13} x_{WOR} + \theta_{14} x_{MED}]$$

where Y'_E is the predicted Site Energy Use Intensity (EUI) of electricity (Unit: kBTU/sf), and Y'_G is the predicted Site EUI of gas (Unit: kBTU/sf). Notice that the units of Y'_E and Y'_D are kBTU/sqft so that we don't need to worry about the size of each building which can affect the total amount of energy use for each building.

We want to predict the weights $W = [w_0, ..., w_{14}]$ and $\Theta = [\theta_0, ..., \theta_{14}]$ of both linear functions above by using Ordinary Least Squares Linear Regression. To predict W and Θ, we calculate the residuals between the function outputs, and the actual electricity/gas consumption values (EUI), Y_E and Y_G, and then find that W and Θ that minimize the residuals which are:

$arg\ min_w \| Y_E - Y'_E \|$ for electricity model

$arg\ min_\theta \| Y_G - Y'_G \|$ for gas model

($\|.\|$ is the L2 norm)

In the code, we used the OLS from Statsmodels framework to handle Linear Regression and calculated the optimized coefficients W and Θ for the two linear functions. By using this approach, we can obtain the optimal W and Θ. Then we use the 5-fold Cross-validation technique to evaluate our linear model and evaluate the average mean-square error as the model performance reference.

Then we improve our model by adding the regularized term to see if our regular multinomial linear model has overfitting problem (regularization). We also use 5-fold Cross-validation technique to evaluate its performance.

Quadratic Nonlinear Regression:

We also hypothesize that our data could potentially have a quadratic relationship between energy consumptions and independent factors. So it's good to construct a Quadratic Nonlinear Model to figure out if it fits our data.

Quadratic Nonlinear Model has similar architecture to the Multinomial Linear Regression. The difference is that we define some factors to have quadratic in the formulae. We define:

$$Y''_E = ß_0 + ß_1 x_{TEMP} + ß_2 x^2_{WI} + ß_3 x^2_{GHGI} + ß_4 x_{AGE} + [ß_5 x_{COL} + ß_6 x_{SPORT} + ß_7 x_{RES} + ß_8 x_{LAB} + ß_9 x_{OFF} + ß_{10} x_{HOT} + ß_{11} x_{DIS} + ß_{12} x_{FOOD} + ß_{13} x_{WOR} + ß_{14} x_{MED}]$$

$$Y''_G = \pi_0 + \pi_1 x_{TEMP} + \pi_2 x^2_{WI} + \pi_3 x^2_{GHGI} + \pi_4 x_{AGE} + [\pi_5 x_{COL} + \pi_6 x_{SPORT} + \pi_7 x_{RES} + \pi_8 x_{LAB} + \pi_9 x_{OFF} + \pi_{10} x_{HOT} + \pi_{11} x_{DIS} + \pi_{12} x_{FOOD} + \pi_{!3} x_{WOR} + \pi_{14} x_{MED}]$$

The approach to calculate optimal $ß = [ß_0, ..., ß_{14}]$ and $\pi = [\pi_0, ..., \pi_{14}]$ is the same as the one in the linear model. Once we obtain the optimal $ß$ and π, we also use the 5-fold Cross-validation to evaluate its performance. We didn't regularize this model since regularization will offset the effect brought by quadratic terms in the model.

Multinomial Logistic Regression:

We were also suggested to use Logistic Regression model for our data. Previously, we put our electricity and gas consumption into 5 levels based on the usage magnitude. Now we want our Logistic Regression model to predict the correct level of energy consumption given the independent factors. We define:

$$P(Y'_E) = 1/(1 + e^{-Y'_E})$$
$$Y'_E = w_0 + w_1 x_{TEMP} + w_2 x_{WI} + w_3 x_{GHGI} + w_4 x_{AGE} + [w_5 x_{COL} + w_6 x_{SPORT} + w_7 x_{RES} + w_8 x_{LAB} + w_9 x_{OFF} + w_{10} x_{HOT} + w_{11} x_{DIS} + w_{12} x_{FOOD} + w_{13} x_{WOR} + w_{14} x_{MED}]$$

$$P(Y'_G) = 1/(1 + e^{-Y'_G})$$
$$Y'_G = \theta_0 + \theta_1 x_{TEMP} + \theta_2 x_{WI} + \theta_3 x_{GHGI} + \theta_4 x_{AGE} + [\theta_5 x_{COL} + \theta_6 x_{SPORT} + \theta_7 x_{RES} + \theta_8 x_{LAB} + \theta_9 x_{OFF} + \theta_{10} x_{HOT} + \theta_{11} x_{DIS} + \theta_{12} x_{FOOD} + \theta_{13} x_{WOR} + \theta_{14} x_{MED}]$$

Where $P(Y'_E)$ and $P(Y'_G)$ probability of the energy consumption is at a certain level. (Level: very low, low, medium, high, very high)

Since we have five classes of level and a single logistic regression only predicts one level. We construct five logistic one-versus-rest regression models in total for handling five levels.

Once we trained our model, again, we use 5-fold Cross-validation to evaluate its performance. Since the energy consumption is a level, we use prediction accuracy instead of a mean-square error to measure the performance.

Experimental Results

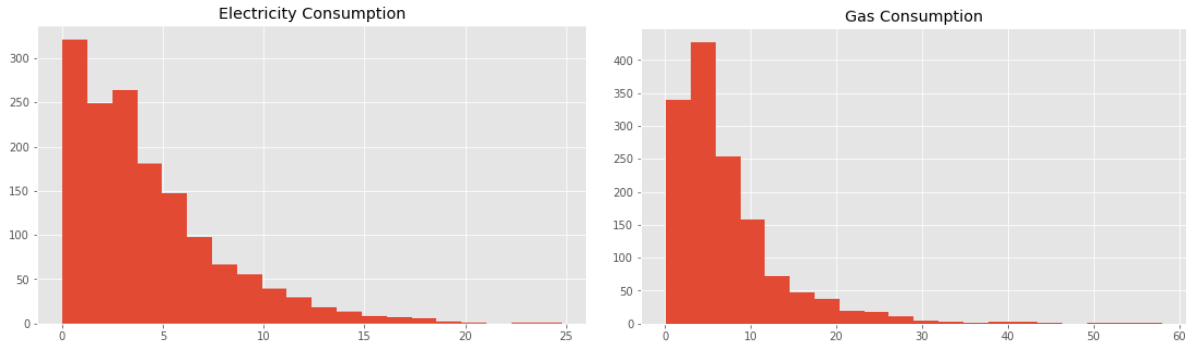
From all the 5-fold Cross-validations we ran in each model, we obtained the following:

Model Name	Electricity Model Average Mean-Square Error / Accuracy	Gas Model Average Mean-Square Error / Accuracy
Multinomial Linear Model	31.7607323327	33.6829730576
Regularized Multinomial Linear Model	41.5251496989	62.8668535059
Quadratic Nonlinear Model	2.66772530378	14.7329440932
Multinomial Logistic Regression (5 Logistic Regressions)	92%, 78%, 94%, 98%, 99% (Very Low, Low, Medium, High, Very High)	91%, 86%, 97%, 99%, 99% (Very Low, Low, Medium, High, Very High)

From the table above, we found that the average mean-square error from Quadratic Nonlinear Model is the lowest and the average mean-square error values from Multinomial Linear Model and its regularized version is quite large. This phenomenon explains that our electricity and gas datasets both have a quadratic relationship between the energy consumption and the independent factors we listed, so a linear model fails to decently describe them. The failure of the linear model means that there is no overfitting problem for the linear model. Hence, it's reasonable that the Regularized Multinomial Linear Model has a larger average mean-square error than the regular Multinomial Linear Model.

For the Multinomial Logistic Regression, although the performance (accuracy for each level) is quite high for both electricity and gas consumption data, we found that the sample size for "High" level logistic regression and "Very High" logistic regression are extremely small. Two graphs below demonstrate that the number of data for "High" and "Very High" level is below 5, and there are lots of data that lay upon "Very Low" and "Low" level. Since we used separate Logistic Regression model for each level, there might be too few data for logistic models of "High" level and "Very High" level. Therefore, due to lack of data, our Multinomial Logistic

Regression model might not perfectly describe the general relationship between energy consumption and those independent factors.



The histogram of number of data based on electricity consumption and gas consumption.
(x-axis: magnitude of electricity/gas usage, y-axis: number of data)
Electricity: VeryLow: [0, 5), Low: [5, 10), Medium: [10, 15), High: [15, 20), VeryHigh: [20, 25]
Gas: VeryLow: [0, 12), Low: [12, 24), Medium: [24, 36), High: [36, 48), VeryHigh: [48, 60)

Thus, we choose the Quadratic Nonlinear Model to find the significance of the listed factors. The following is the Quadratic Nonlinear Model OLS Regression Results for Electricity Consumption:

OLS Regression Results

Dep. Variable:	Y_Electric	R-squared:	0.768
Model:	OLS	Adj. R-squared:	0.766
Method:	Least Squares	F-statistic:	354.4
Date:	Wed, 02 May 2018	Prob (F-statistic):	0.00
Time:	18:17:21	Log-Likelihood:	-2986.1
No. Observations:	1512	AIC:	6002.
Df Residuals:	1497	BIC:	6082.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-31.5816	4.296	-7.351	0.000	-40.009	-23.155
Temperature	0.0130	0.003	4.696	0.000	0.008	0.018
np.power(WI, 2)	2.686e-06	3.09e-06	0.869	0.385	-3.38e-06	8.75e-06
np.power(GHGI, 2)	1.3881	0.052	26.764	0.000	1.286	1.490
Age	0.0175	0.001	12.873	0.000	0.015	0.020
Uses_College	-0.2947	3.433	-0.086	0.932	-7.028	6.439
Uses_Sports	-0.5125	3.435	-0.149	0.881	-7.251	6.226
Uses_ResidentHousing	-1.8923	3.433	-0.551	0.582	-8.626	4.841
Uses_Laboratory	3.5505	3.436	1.033	0.302	-3.189	10.290
Uses_Office	-0.1648	3.434	-0.048	0.962	-6.901	6.572
Uses_Hotel	-0.0489	3.446	-0.014	0.989	-6.808	6.710
Uses_Distribution	-0.7702	3.451	-0.223	0.823	-7.539	5.998
Uses_Food	0.8891	3.446	0.258	0.796	-5.871	7.649
Uses_Worship	-3.1703	3.446	-0.920	0.358	-9.930	3.590
Uses_Medical	2.0718	3.444	0.602	0.548	-4.684	8.828

Omnibus:	486.757	Durbin-Watson:	1.898
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3097.927
Skew:	1.347	Prob(JB):	0.00
Kurtosis:	9.474	Cond. No.	2.50e+07

From the table above, we found that there are 3 factors whose p-values are less than 5%, which are Temperature, GHGI (squared), and Age. Moreover, these three factors' confidence intervals don't contain any 0, and this means that these three factors are significant for impacting the electricity consumption in our model. The coefficient of temperature, in this case, is positive and it follows the trend that the higher temperature is, the more electricity is consumed. GHGI also plays an important role for energy consumption due to its relatively high coefficient value. The impact of building age also follows the trend in our data visualization. Since the coefficient is positive, it means that, in our model, the newer the building is, the higher electricity consumption of that building will have. This might be explained by the fact that lots of advanced technologies are introduced to a new building which will consume more electricity.

The following is the Quadratic Nonlinear Model OLS Regression Results for Gas Consumption:

OLS Regression Results

Dep. Variable:	Y_Gas	R-squared:	0.661
Model:	OLS	Adj. R-squared:	0.658
Method:	Least Squares	F-statistic:	208.4
Date:	Wed, 02 May 2018	Prob (F-statistic):	5.30e-315
Time:	18:17:06	Log-Likelihood:	-3945.1
No. Observations:	1404	AIC:	7918.
Df Residuals:	1390	BIC:	7992.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	60.6697	6.018	10.082	0.000	48.865	72.474
Temperature	-0.1906	0.007	-28.722	0.000	-0.204	-0.178
np.power(WI, 2)	-0.0177	0.008	-2.312	0.021	-0.033	-0.003
np.power(GHGI, 2)	4.7293	0.130	36.473	0.000	4.475	4.984
Age	-0.0272	0.003	-8.084	0.000	-0.034	-0.021
Uses_College	5.8679	0.588	9.971	0.000	4.713	7.022
Uses_Sports	6.5608	0.799	8.215	0.000	4.994	8.128
Uses_ResidentHousing	7.7840	0.631	12.333	0.000	6.546	9.022
Uses_Laboratory	1.3249	0.645	2.053	0.040	0.059	2.591
Uses_Office	5.8382	0.597	9.772	0.000	4.666	7.010
Uses_Hotel	6.5587	0.986	6.650	0.000	4.624	8.494
Uses_Distribution	8.0753	0.953	8.473	0.000	6.206	9.945
Uses_Food	6.6855	0.980	6.822	0.000	4.763	8.608
Uses_Worship	4.9846	0.974	5.116	0.000	3.073	6.896
Uses_Medical	6.9898	0.868	8.057	0.000	5.288	8.692

Omnibus:	618.876	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7148.334
Skew:	1.733	Prob(JB):	0.00
Kurtosis:	13.497	Cond. No.	2.40e+19

From this table, we surprisingly found that all factors' p-values are below 5% and none of those factors' confidence intervals contains any 0. This means that all factors we hypothesized at the beginning are significant. Therefore, temperature, water usage intensity, age, and building type will affect the gas consumption pretty well. The coefficient of temperature, in this case, is negative and it follows the trend that the lower temperature is, the more gas is consumed. GHGI also plays an important role for energy consumption due to its relatively high coefficient value. Moreover, types also determine the magnitude of gas consumption: a lot of types will have

impacts on the gas consumption such as distribution center, residence halls, medical buildings, and sports-using buildings. Surprisingly, type LAB has relatively small coefficient value compared to other types, which means that being a laboratory doesn't necessarily means it will consume much gas energy.

Conclusions

By looking at our results, we generated conclusions as following:

- **Temperature**: same as we expected, temperature indeed plays big roles in both **electricity** and **gas** consumptions. Also, the coefficient value of temperature for electricity is positive and for gas is negative, which means the hotter the weather, the more electricity will be used since people need to turn AC on, and the fewer people are going to use gas to heat up. However, since temperature is not something that human can control, we would suggest our project partner to worry anything about it.
- **Water Intensity**: the impact of water intensity exists with the explanation of the possibility of using **gas** to heat water in winter or to cool off in summer. However, it does not affect much on the usage of **electricity**. Based on such results, we would suggest our project partner to propose some water saving campaigns on using hot water in winter or the overuse of cold water in summer.
- **Greenhouse Gas Emission Intensity**: Compared to other factors, the effect of GHGI is quite large from the perspective of both **electricity** and **gas**. In other words, the major amount of electricity and gas consumption at least emits some form of greenhouse gases. Consider the large impact of this factor, we would recommend our project partner to propose the using of clean energy all across the university.
- **Age**: age seems to play significant roles in both **electricity** and **gas** consumptions. In our opinion, it might be explained that although we know that some buildings are in the process of renovating, the finish of renovation may not lead to an installation of more efficient equipment. Observing this, we may suggest the University to install some up-to-date equipment in buildings in order to increase energy efficiency.
- **Property Type**: taking a look at all of those various types of buildings we found, types really is an important factor in the consumption of **gas**, while for **electricity** it is does not make much contribution. In details, properties that are functioned as a distribution center or residential housing will be likely to affect more of gas consumption. This can be explained that students will need to turn on heating system during winter. Based on the results, we would suggest the University to renovate its resident housing since many of BU's resident halls are brownstones which were built years ago.

Code Link

Notice:
For each model (Linear, Quadratic Nonlinear, Logistic Model), we implemented code and obtained summary results about coefficients, p-values, and confidence intervals. We only use the summary of Quadratic Nonlinear Model since we selected it as the best model. For the summary of other models, please check our Jupyter Notebook file in the code link.

References

BERDO Dataset:
https://data.boston.gov/dataset/building-energy-reporting-and-disclosure-ordinance

BERDO Data Map:
http://boston.maps.arcgis.com/apps/webappviewer/index.html?id=049576c7287f4ee09bcb0a062e43b55c

Boston Weather:
https://www.usclimatedata.com/climate/boston/massachusetts/united-states/usma0046

MWRA Online:
http://www.mwra.com/monthly/wsupdat/archivecomwatuse.htm

Earth System Research Laboratory:
https://www.esrl.noaa.gov/gmd/ccgg/trends/data.html

Boston University Map:
http://www.bu.edu/maps/

Boston University Sustainability:
http://www.bu.edu/sustainability/

Carbon Neutrality:
https://en.wikipedia.org/wiki/Carbon_neutrality

SkLearn Linear Regression:
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

StatsModels Linear Regression:
http://www.statsmodels.org/dev/regression.html

Wikipedia Linear Regression:
https://en.wikipedia.org/wiki/Linear_regression

OLS Regression:
https://en.wikipedia.org/wiki/Ordinary_least_squares

Statsmodels OLS:
http://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

Statsmodels Logit Regression:
http://www.statsmodels.org/dev/generated/statsmodels.discrete.discrete_model.Logit.html

Mean Square Error:
https://en.wikipedia.org/wiki/Mean_squared_error

Sklearn mean square error:
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html