**ENEC490.003.SP24 Spring 2024: Assignment 1**

**Objective:** For this assignment, the objective is to use the data from the [Boston Buildings Inventory](https://data.boston.gov/dataset/boston-buildings-inventory) to create a field called Site EUI, which will be our dependent variable, and then to select some independent (or predictor) variables to estimate Site EUI using regression analysis.

**Data Files:** I have provided two data files in the Assignment 1 folder in Canvas for this assignment:

1. **boston\_building\_inventory\_021020.csv**: this is the Boston buildings inventory data set. Per the website *“This dataset pulls from many different data sources to identify individual building characteristics of all buildings in Boston. It also identifies high-potential retrofit options to reduce carbon emissions in multifamily buildings, using the best available data and assumptions from building experts.”* I have already removed rows from the data set where the field total\_site\_energy\_kbtu was 0 or missing.
2. **boston\_building\_inventory\_fields-definitions.csv**: this file has the field definitions for all the columns in the data set. It should help you interpret and choose predictor variables.

**Deliverables:** submit a jupyter notebook with your python code and comments showing your approach to the assignment and a word document answering the questions below on Canvas.

**Steps to include in your analysis:** This is a data science project, so use your critical thinking skills to investigate the data and complete the objective, which means you have some latitude in how you approach it and how far you choose to go in finding a model to predict Site EUI or which predictor variables you use. The steps I expect to be in your jupyter notebook file generally include:

1. Load the data set
2. Create the Site EUI variable, which is total\_site\_energy\_kbtu/gross\_area (Or you could also use sqft. They are mostly the same but sometimes gross\_area is bigger and the definitions don’t make it clear exactly why they vary so it’s up to you)
3. Explore the data for # of observations, missing values, data types, etc.
4. Clean the data
5. Visualize the data (scatter plots, box plots, pair plots, etc.) and examine the relationship between potential predictor variables and Site EUI
6. Feature engineering including converting categorical variables to dummy variables, using natural log or other transformations you think might help improve the model
7. Build the regression model and output the model summary data (I recommend statsmodels but you can use another library if preferred)

**Questions (submit answers to these with your jupyter notebook on Canvas):**

1. What steps did you take to explore and clean the data? What did you find out?
2. Which independent or predictor variables did you initially choose? Why?
3. How did you choose to visualize the data? Did the visual analysis impact the predictors you chose to continue with?
4. What feature engineering steps did you take to prepare the data for analysis?
5. What are the key model outputs from your final regression model and explain how you interpret them and evaluate the model’s performance. Did you choose to adjust your variables and re-run the analysis and, if so, did that impact your results?

**Answers:**

*I first used Excel to get an overview of what the data is like overall. While exploring, I highlighted the columns of variables that I think are relevant, including yr\_built, gross\_area, num\_floors, num\_bldgs, and building\_typology. The building age (calculated using the year built) will reflect the age of the energy system and thus might possibly reflect the efficiency and thus the amount of energy used. Gross area determines the space that is heated, the larger the space, the more energy is needed to support it. Apart from the two discussed in class example, I also chose a number of floors because there could be behavioral changes in terms of energy use with the increase in the house floor levels. And the final data does support the positive relationship to the calculated EUI. For num\_bldgs and building\_typology, I just used them to filter out some of the data. I restricted the number of buildings to one and the building typology to the residential buildings only (including category multifamily and residential). For filtering data, I also tried using unit\_con, which is the unit of commercial buildings. However, when I set it to 0, almost all year built are 1999, so I guess many of the data are blank making this step problematic. Then, I just ignored unit\_con. The dataset, after filtering, has 274 rows before further cleaning.*

*I tried to visualize the data in many of the steps. I first visualized the predictor variables alone and then filtered out some apparent outliers. I didn’t have a standard in which percentile to set, but in later steps, I will plot the variables against site\_EUI to see how the outliers for the correlation. Then, I plot the graph between the predictor variables and site\_EUI. Then, if I see the graph has EUI significantly higher than the others, I will filter out those. Finally, the pair plot visualization gives an overall view of the models, which is really helpful. This is where I can see visually how different model differ in their ability. I think visualizations in different parts of the data modeling process each contribute differently, and all positively contribute to shaping how the data will look.*

*I run the 2 models separately and also the one with the original data df\_res to compare. The two are with gross area and floor numbers. Overall, the gross area one performs the best with an r-square of 0.089, though still not great but better than the other two. The floor number one is 0.039, and the original one is 0.008. All models do better than the original one. During the process, I also tried adjusting the filter for both the variables for outliers. And I find it the most efficient to just adjust it where I plot the variables against the EUI because I can clearly see the outlier from the rest of the data that is interfering with the regression.*