# Main Report

### Project Description

The goal of this project is to analyze BU's building energy usage intensity (EUI) and discover relationships between driving factors to help the University achieve its goal of becoming carbon neutral by 2040. Ideally, the analysis would help inform decisions made by sustainability@BU, the Climate Action Plan and the buildings team of Carbon Free Boston.

The main question we are trying to answer is “How much does BU's building energy use intensity vary with property type/year built/ temperature?”. Several specific questions are raised to help answering the major question. Taking temperature as an example, “how does temperature varies among 2015-2017”, and “how does EUI of each type of building varies among 2015-2017”, etc.

The methods we used includes data scraping, classification, linear regression, Probability and statistics.

Up to now, we integrated two data sets - Building Energy Reporting and Disclosure Ordinance of Boston(BERDO) 2015-2017, and Property Assessment of Boston 2015-2017; analyzed how (much) property type/year built/temperature affect EUI; made some effort on predicting EUI of a building with given information.

Most of the results shows what we were expecting, although part of them won’t give much interesting conclusion.

### Data Description

The data sets we are using are retrieved directly from official websites.

BERDO has our core attribute of each building - EUI. While, we may also need several other attributes in different data sets to find out how they are related. So, when processing the data, we combined two data sets using the address in BERDO to find the matching building in Property Assessment data set, and generated a single .csv file which have all attributes we need for each building. However, in this process, the total number of buildings dropped from around 1800 to 746, since most of the buildings don’t match with the other one. However this may not cause severe consequences due to the scarcity of data, because we didn’t do much of our analysis solely upon the integrated data set.

When cleaning BERDO 2015-2017, we detected and removed most of outliers based on our observation. We made this tentative conclusion that, “normally EUI should never be over 800 kBTU/sf”. This can be proved by listing all the buildings of same property type, and only one or two tops of the data would be at least 10 times greater then others, which is obviously a contaminated data.

**how clean and prepare the data for processing**

**how address missing values or inconsistent entries**

**How large were your data before and after cleaning**

**What are your data? Briefly describe your datasets, along with the attributes that you are using and their types.**