# NYC Buidling Energy Benchmark

## A K-Means Clustering Approach

Ian Xuan Xiao
Center of Urban Progress & Science, New York University
New York, United States
ian.xxiao@nyu.edu

***Abstract*—This paper presents a K-Means Clustering approach to finding peer buildings for energy benchmark by using Local Law 84 and PLUTO datasets. The objective is to develop a simple and fair benchmarking method to encourage building energy reduction in New York City.**

***Keywords—Benchmarking; Energy performance; Unsupervised learning; K-Means Clustering; OpenNYC***

## I. Introduction

New York City committed to an ambitious energy reduction plan. The city wants to reduce 80% of the Green House Gas (GHG) emission by 2050. As part of this effort, the city launched programs, such as Local Law 84, to mandate building owners to report energy usage. The hope is to encourage energy reduction through benchmarking and transparency.

However, developing a simple, yet fair, benchmarking metric to evaluate building energy performance is difficult. Many factors, such as occupancy, building age, and tenant type and behavior, may require more intensive energy usage given the same building size and function (e.g. multifamily, office, hotel, hospital, etc.).

The key to developing a simple and fair benchmarking metric is finding the right buildings for peer comparison. Traditionally, building owners could not do so because of the lack of public data to qualify and quantify similarity. Local Law 84 on OpenNYC offers an opportunity, especially as the data accumulates throughout future years.

In this paper, I present an approach of using 5 years of Local Law 84 (LL84) and Pluto data to find similar buildings given a set of characteristics. Then, I conducted various statistical and visual analysis to assess the validity of such approach, demonstrate its strengths and weaknesses, and share next steps and alternative approaches.

Policymaker, building owners, environmental advocates, and investors are the intended audience. With this in mind, I tried to avoid any complicated machine learning models and statistical analysis.

## II. Literature Review

There are many existing benchmarking approaches: simulation models, points systems, end-use metrics, and regression models. Many of these approaches lack comprehensive consideration of key building factors, not flexible for post-construction changes, and require a large amount of data to be statistically significant (Kontokosta).

In addition, many modelling approaches require a subjective definition of high, medium, and low energy consumption (Chung). Doing so tends to invite debates about fairness for building with different characteristics and operating conditions.

## III. Data

This section covers the overview of data sources and descriptive analysis of the final integrated dataset.

Data in this analysis consists of 5 years of Local Law 84 (LL84) published on OpenNYC from 2012 - 2016. LL84 data provides the following key, but not exhaustive, information:

- Building Identifier
- Energy Usage Intensity (EUI)
- Floor Area
- Building Function (e.g. multifamily, office, hotel)

In addition, data from PLUTO provides detailed building information, such as the number of units, breakdown by residential, retail, office, and commercial, building age, and lot size. These variables provide richer context to define similarity among buildings.

There are 17,942 raw samples in LL84 from 2016 and 858,370 in Pluto (both were accessed on November 7, 2017).

My final integrated dataset for analysis consists of 9,852 samples. Samples with missing values were excluded. Details of data integration, cleaning, and filtering can be found in the Method section.

The majority of the building in my analysis are Multifamily homes, offices, and university buildings. They made up about 80%, 8.42%, and 1.9% of the total buildings respectively.

This analysis covers buildings across all five boroughs: Manhattan, Queens, Bronx, Brooklyn, and Staten Island. Buildings in Staten Island is under-represented.

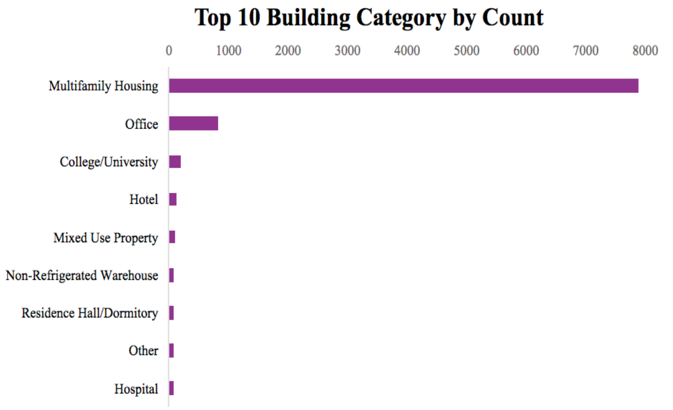


***Figure1:*** *Count Breakdown by Building Category. Multifamily Housing made up about 80% of the sample population. The dataset excluded samples with any missing values. Total sample in the final dataset = 9,852*

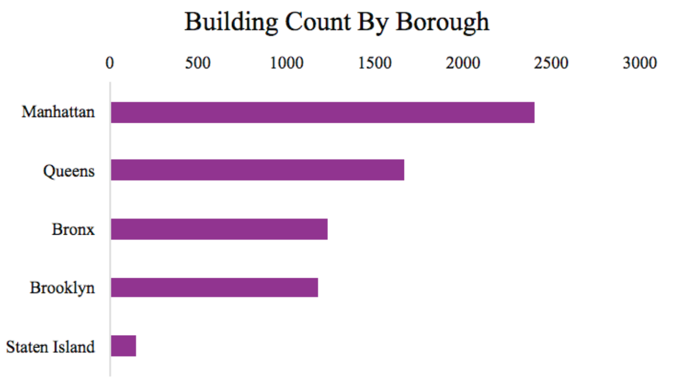


***Figure 2:*** *Building Count by Borough; Staten Island is under-represented while there are even representation across other boroughs. Total sample in the final dataset = 9,852*

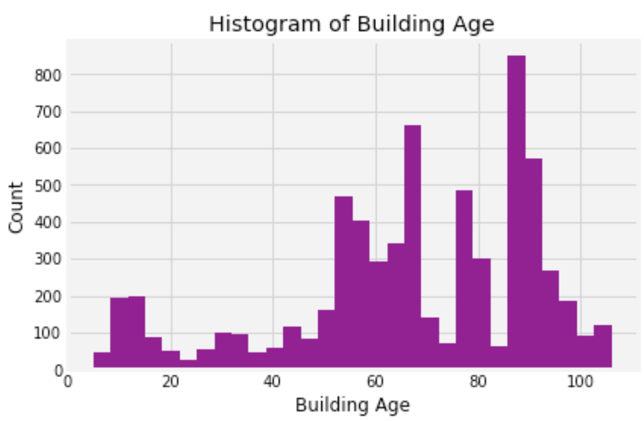


***Figure 3:*** *Histogram of Building Age; The majority of the buildings were built around the World Wars. Total sample in the final dataset = 9,852*

## IV. Method

**Data Integration**

LL84 from different years and Pluto data are joined by BBL.

**Data Processing and Cleaning**

String-to-Numeric conversions were applied to key columns such as BBL and EUI. EUIs in 2014 and 2015 were stored as strings, instead of numerical values, because some owners entered string values, such as "See Primary Building", when there are multiple buildings in their properties. Null values were placed when the conversion is not successful.

As a preliminary attempt, any sample with a missing value (NAN in Python) was removed. 8,062 samples out of 17,914 (~45%) in the integrated datasets were removed as a result. Alternatively, one can impute missing numerical value, such as EUI or building age, with population average if applicable.

**Feature Engineering**

The following features were created for each building to provide more context for subsequent modeling and identify outliers to ensure sample quality:

- 4-year EUI average from 2012-2015
- 3-year EUI average from 2013-2015
- EUI trend (up or down) from 2012-2015
- Standard deviation of 4-year EUI reporting
- Building age as of 2017
- Area of residential, retail, office, and commercial as percentages of total building area

**Outlier Removal**

I removed any sample that is above the 95 percentiles in EUI, the standard deviation of 4-year EUI reporting, building age, and percentage of residential, commercial, office, and retail usage. Extreme values in these cases may indicate data entry error.

**Modelling**

The objective of modeling is to group similar buildings and calculate group level benchmark metric.

First, I subset the buildings by their reported category (i.e. multifamily housing, office). In this analysis, I only focus on Multifamily Housing because they represent 80% of the population. Ideally, this process is performed across all building categories.

Then, I performed a K-Mean Clustering to create three sub-groups within Multifamily Housing using these key variables: building age, Zip Code, Building Area, Percentage of Residential, Commercial, and Office Usage. These factors can create a similarity measurement that captures detailed of and nuances in building configuration, condition, and neighborhood heuristics (i.e. social economics, population, building demands, etc.). The unsupervised learning mechanism of K-Means Clustering can theoretically group buildings without manual categorization (e.g. what building age is considered old/new or what high/low residential percentage means). This avoids subjective judgement that is often criticized in benchmarking exercises.

The final grouping was developed after using 100 random initial conditions to achieve consistency.

**Benchmarking**

The sub-group level energy statistics are then calculated to be used as benchmarking metrics.

| Group | Group Size | Mean EUI 2016 | Min. EUI 2016 | Max. EUI 2016 | STD. EUI 2016 |
|---|---|---|---|---|---|
| **0** | 5649 | 146.4 | 0.4 | 105263.6 | 1969.5 |
| **1** | 145 | 2968.0 | 18.7 | 24115.4 | 7733.8 |
| **2** | 650 | 274.0 | 0 | 49924 | 2323.5 |

***Table 1:*** *This is the EUI benchmarking statistic of each sub-group of Multifamily Housing based on building age, Zip Code, Building Area, and Percentage of Residential, Commercial, and Office Usage.*

Each building then can assess its relative energy performance based on the respective sub-group.

*The end-to-end script for this process can be found on at:*

*https://github.com/ianxxiao/civic_analytics/blob/master/NYC_Energy_Benchmarking.ipynb*

## V. Results

The map below is a visualization of Multifamily Housing based on the analysis. Each building is colored by the sub-group assigned by the K-Means Clustering. Due to limited credit on Carto, only 400 out of 9,852 buildings were geo-encoded and rendered. This visualization is a fair representation because of the dataset has no particular ordering.

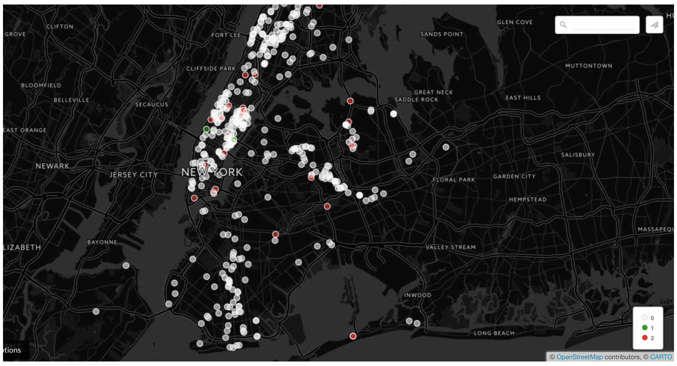


***Figure 4:*** *Geo-map of Multifamily Housing colored by sub-group from K-Means Clustering. Group 0 is white, Group 1 is green, and Group 2 is red.*

Although this map provides an intuitive view of where the buildings locate and how they are grouped, it does not indicate any validity in the K-Means Clustering approach.

Does the K-Means Clustering provide reasonable groupings based on building configuration, age, and neighborhood heuristics?

To answer this, one can first compare the standard deviation of EUI of the total population and each sub-group. The standard deviation of each sub-group should be smaller such that it indicates consistent and specific behavior.

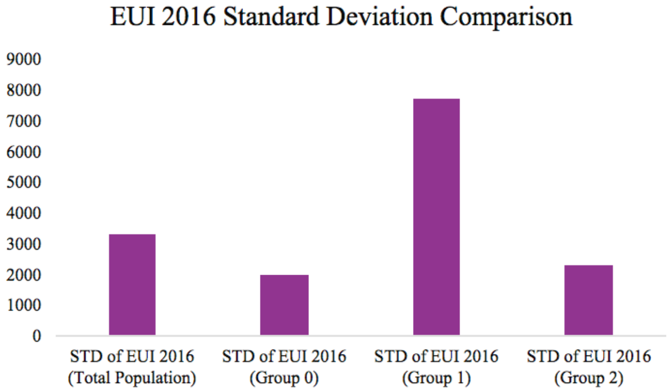


***Figure 5:*** *Comparison of Standard Deviation of EUI 2016 between total population and each K-Means group.*

Secondly, one can validate if the differences in the average EUI due to the grouping have any statistical significance. If so, it signifies that the K-Means Clustering is indeed grouping distinctive building characteristics that separate energy consumption behaviors. In other words, buildings owners can be more assured that they are comparing against peer buildings by separating out dissimilar ones.

T-tests with an alpha of 0.05 were conducted to test the following hypothesis.

| Hypothesis | T Score | Outcome |
|---|---|---|
| Group 1 has a higher average EUI in 2016 than group 0. | 4.39 | Hypothesis Confirmed |
| Group 1 has a higher average EUI in 2016 than group 2. | 1.40 | Hypothesis Not Confirmed |
| Group 2 has a higher average EUI in 2016 than group 0. | 4.19 | Hypothesis Confirmed |

***Table 2:*** *Hypothesis testing results at 95% confidence level; Null hypothesis can be rejected when z-score is larger than 2.*

## VI. Discussion and Conclusion

Based on Figure 5, group 1 show very high standard deviation among its samples. This is due to a relatively small group size. This is expected because this group may be capturing the fringe behaviors, which do not represent the majority and can have extreme ranges. Group 0 and 2 demonstrate expected results.

According to Table 2, there is a statistically significant distinction between group 0 with group 1 and 2. However, it is not conclusive for the grouping of 1 and 2. From the building owners' perspective, they can be confident that they are being benchmarked against similar peers - according to building age, configuration, and neighborhood heuristics - that is significantly different from other buildings even they are categorized as Multifamily Housing.

A further descriptive investigation is required to provide more specific policy insights and factual backing. For example, how do building age, building configuration, and social economics in each neighborhood differ between each grouping?

There are many ways to improve the K-Means Clustering. An option is to normalize all factors between 0 and 1 because distance-based clustering technique is sensitive to numerical scales.

K-Means Clustering can be perceived as a black-boxed approach because of its reliance on a random process that may lack human intervention when desired. An alternative and more descriptive approach is to re-purpose a technique called Collaborative Filtering, which is being used in the large scale product recommendations. Finding similar entities, such as similar customers, is the core concept.

Despite such shortcoming, using K-Means Clustering to find similar peers for benchmarking is better than a non-factual assignment of performance expectations.

## REFERENCES

[1] Kontokosta, Constantine E. A Market-Specific Methodology for a Commercial Building Energy Performance Index. Article. New York: Springerlink, 2014.

[2] Chung, W., Hui, Y. V., & Miu Lam, Y. (2006). Benchmarking the energy efficiency of commercial buildings.

## CODE ASSET

https://github.com/ianxxiao/civic_analytics/blob/master/NYC_Energy_Benchmarking.ipynb