



# NYC Buildings' Energy Efficiency

Climate Change Whisperers

---

Ayush Madhogaria  
Ben Douglas Glotzer  
David Zheng  
Jaqueline Pinheiro  
Yingxue Liu

## Introduction

In 2018, New York City passed Local Law 33, which required buildings to obtain energy efficiency scores that measure energy consumption [1]. To comply, buildings covered by the law must use the United States Environmental Protection Agency's online benchmarking tool, Energy Star Portfolio Manager, to submit data to the NYC Department of Buildings by May 1 of each year. Then, on October 1 of each year, the department releases Building Energy Efficiency Rating labels, including a 1-100 Energy Star score and a corresponding letter grade, which building owners must display near each public entrance within 30 days. Buildings that fail to submit the required data receive an F as the letter grade. This program is part of the Climate Mobilization Act, the most comprehensive package of legislation of its type in any city in the world, with the goal of dramatically reducing carbon emissions [2].

This ambitious initiative is a crucial step towards combating climate change, but the rating system may also be an indicator of inequality in the city. It is undeniable that human behavior must change immediately to avoid catastrophe, but those with the least amount of influence in society are often powerless to make significant changes to the environmental repercussions of their lifestyles. In the early 2000s, BP launched a propaganda campaign to promote the notion of a “carbon footprint” and convince the American public that the onus was entirely on them to mitigate the risk of climate change [3]. It was a clever and deceitful ploy that, to this day, has helped distract from the fact that the company has done little to reduce their own abuse of the environment. Clearly, structural change, more so than individual actions, is needed to address the disastrous outlook of the only known habitable planet in the universe.

In our project, we analyzed data on a collection of demographic features to determine whether they were predictive of the energy efficiency of buildings in New York City. We were interested in this topic for two main reasons:


1. If we found that certain demographic characteristics were strong predictors of efficiency scores, this could help direct urgently needed government intervention. This would be particularly important if we found that lower socioeconomic features were correlated with lower scores, as these populations most often lack the wherewithal to address this problem.
2. Residents of energy inefficient buildings likely pay more in utilities, so a correlation between lower socioeconomic status and lower efficiency scores would also reveal financial injustice that must be rectified.

## Data Preprocessing

A complete list of the data we used in our project and corresponding sources is provided in the appendix.

### Building Data

The main dataset, `DOB_Sustainability_Compliance_Map__Local_Law_33.csv`, contains the Energy Star and Letter score of over 21,000 buildings in the five boroughs of New York City. This dataset also included the physical address and borough that the buildings are located at but did not contain zip code for each



building. The zip codes would be required to connect this dataset with the socioeconomic features that are summarized by NYC zip codes. Utilizing the power of Google's Geocoding API, we were able to loop through all the addresses to retrieve the zip codes and, in addition, the longitude and latitude coordinates. The coordinates were supplementary data points used in creating data visuals.

### **Building Classification**

Buildings in NYC are classified in 218 Building Types according to their main characteristics related to structure, facilities, size, or even purpose. For instance, some building types would be: "Large Suburban Residence", "Metal Frame Warehouse", "Cemetery", "Garden Apartments". It is worth mentioning that the "DOB\_Sustainability\_Compliance\_Map\_\_ Local\_Law \_33.csv" dataset has only the Building Type Code, but not its description. To get the Building Type Description from the website mentioned on point 3 of Data Sources topic, we have used web scraping. Finally, in order to have a better aggregated understanding of our dataset, we classified those types of buildings in residential and commercial.

### **Demographic Data**

All of the demographic data we used in our project came from the US Census American Community Survey (ACS). After brainstorming features we were interested in analyzing, we navigated the website to identify the pertinent surveys and applied filters to select the relevant data. Then, we spent a considerable amount of time using Pandas for data cleaning and feature engineering. Some examples of this work include identifying desired columns and dropping the rest (some of the files had hundreds of columns, most of which were irrelevant), extracting numerical zip codes from longer geographic identifiers in string format in order to perform joins, and creating new features by applying arithmetic operations on other columns.


## **Exploratory Data Analysis (EDA)**

The EDA was focused on the study of Energy Score distribution and its patterns per region and per building type. Find the figures related to the following analysis in the appendix.

In general, the Energy Score distribution has a point of mass at 0 and 100 scores, that corresponds to 12% of the total number of buildings, and that the other scores (from 2 to 99) are quite uniformly distributed. For the Letter Score distribution, we noted that it has an unequal distribution with a concentration of Letter Score D observations, which can lead to a risk of imbalanced classification (Figures 1 and 2).

### **Region Analysis**

The heatmap based on buildings' energy efficiency scores across NYC reveals that there is no clear concentration of scores in any specific region. Moreover, the attempt to cluster the scores among



boroughs reinforces that it is hard to suggest that there is a region with better scores. Going more granular to the Zip Code level, we can see some concentration of grades. However, the clusters are mainly formed for Zip Codes with scores below 55, which all have the same letter score D (Figures 3, 4 and 5).

### **Building Type Analysis**

The building types with highest mean scores are “WALK-UP CO-OP; CONVERSION FROM LOFT/WAREHOUSE” and “CONVERTED DWELLINGS OR ROOMING HOUSE”. However, those types are not really representative of our data set, as there are only one observation of each. Therefore, the type that has the best mean score as well as a balanced amount of observations (19%) is “ELEVATOR COOPERATIVE” (Figure 6). This building type should be better studied as well as its correlation with better scores. We believe that it would be useful to have a qualitative analysis to gather more insights about its structure and characteristics, and how they can lead to better energy efficiency.

## **ML Models & Results**

### **Classifier Models**


The premise of our project was to explore and analyze whether there is a correlation between socioeconomic indicators and the Energy Star score of buildings in the five boroughs of New York City. In general, we knew we were working with a multi-class classification model where we wanted to see if the socioeconomic features could predict buildings’ Energy Star letter score with some high level of accuracy. Based on several trials with different test dataset sizes, we found that 0.3 produced the best results, therefore the results provided here on out were from a 70/30 training-testing dataset split.

The first classifier type model we used to fit our training datasets was a Decision Tree classifier model. As with decision trees, our goal was to minimize entropy but also be aware of partitioning the data so much that we end up with single case subsets. In our initial model fitting, we initiated the model with a max depth of 5 based on an entropy criterion. The accuracy of the decision tree model fluctuated between 0.43 to 0.48. To finetune the hyperparameters of the Decision Tree Classifier, we passed into GridSearchCV on the Decision Tree Classifier, range of values for the max\_depth, min\_samples\_leaf, min\_samples\_split and criterion. The best score achieved was 0.478.

Following up the Decision Tree Classifier model, we decided to use the Random Forest Classifier machine learning model. Given our uncertainty on which socioeconomic features would best predict the four classes, we decided to use the Random Forest classifier because this model has the ability to create multiple different trees on subsets of features.

### **Regression Models**

After collecting various socio-economic and demographic features about the residential buildings in New York City and its residents, our goal was to try to use machine learning algorithms to try and see if we



could predict the energy score of these buildings, which would imply that these features do have an impact on the energy rating of buildings. Having done feature selection and trying different test-train sized datasets and using different classification models, we saw that we could not generate very accurate predictions. That being done, we also tried using regression models like Decision Tree Regressor, Logistic Regression, Random Forest Regression to try and see if we could get better predictions.

As we wanted to classify buildings based on energy letter scores which were divided into 1,2,3,4 classes and regression models have a continuous output, we rounded up the predicted values to allot the scores into respective classes.

We saw that the results for the Decision Tree Regression model were very poor and the accuracy was as low as 0.02, which could probably be because of the fact that Decision tree regressor model tries to fit a decision boundary to split the points into distinct regions and classify them accordingly, which might not be as effective when we have a huge number of features. However, the Logistic regression model produced better results and achieved accuracy levels of about 0.49.

All things considered, Random Forest classifier model outperformed all the other models with use of a little hyperparameter tuning, which could be due to the fact that the model uses an ensemble based machine learning algorithm which allows it to consider all the features and select the best combination of features to predict the final output.

## Conclusion

Despite using different regression and classifier models and performing feature engineering and selection, the overall accuracy of our models did not exceed 0.5. Therefore, we did not find evidence of correlation between socioeconomic indicators and building efficiency scores. However, we cannot definitively conclude that such a correlation does not exist.

A major limitation of our project is the granularity of the data available. While we have Energy Star scores for individual buildings in all five boroughs of New York City, the socioeconomic data we were able to retrieve was only specific down to the zip code level. Given the blend of new and older buildings across the zip codes in NYC, there would be a large range of scores within a given zip code. If socioeconomic data at a more granular level was available, potentially at the building level, it is possible that the models would more accurately predict Energy Star scores, so we cannot conclude that there is no correlation.

## References

1. *Local Law 33 as Amended by LL95 of 2019 - Steps to Compliance*  
[https://www1.nyc.gov/assets/buildings/pdf/ll33\\_compliance\\_steps.pdf](https://www1.nyc.gov/assets/buildings/pdf/ll33_compliance_steps.pdf)
2. *The Climate Mobilization Act, 2019*  
<https://www1.nyc.gov/site/sustainability/legislation/climate-mobilization-act-2019.page>
3. *"The carbon footprint sham"*  
<httphttps://mashable.com/feature/carbon-footprint-pr-campaign-sham>

## Appendix

### Data Sources

Description	Source
csv file with efficiency score and rating for all buildings in New York City	<a href="https://data.cityofnewyork.us/Business/DOB-Sustainability-Compliance-Map-Local-Law-33/355w-xvp2">https://data.cityofnewyork.us/Business/DOB-Sustainability-Compliance-Map-Local-Law-33/355w-xvp2</a>
Building classification table (building code and description)	<a href="https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html">https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html</a>
Google Maps Geocoding API	<a href="https://developers.google.com/maps/documentation/geocoding">https://developers.google.com/maps/documentation/geocoding</a>
csv file with median income per zip code	<a href="https://data.census.gov/cedsci/table?q=B19013%3A%20MEDIAN%20HOUSEHOLD%20INCOME%20IN%20THE%20PAST%2012%20MONTHS%20%28IN%202020%20INFLATION-ADJUSTED%20DOLLARS%29&amp;g=0400000US36%248600000">https://data.census.gov/cedsci/table?q=B19013%3A%20MEDIAN%20HOUSEHOLD%20INCOME%20IN%20THE%20PAST%2012%20MONTHS%20%28IN%202020%20INFLATION-ADJUSTED%20DOLLARS%29&amp;g=0400000US36%248600000</a>
csv file with median housing value per zip code	<a href="https://data.census.gov/cedsci/table?t=Housing%20Value%20and%20Purchase%20Price&amp;g=0400000US36%248600000&amp;tid=ACSDT5Y2020.B25077&amp;more=false">https://data.census.gov/cedsci/table?t=Housing%20Value%20and%20Purchase%20Price&amp;g=0400000US36%248600000&amp;tid=ACSDT5Y2020.B25077&amp;more=false</a>
csv file with median age per zip code	<a href="https://data.census.gov/cedsci/table?q=median%20age&amp;g=0400000US36%248600000&amp;tid=ACSDT5Y2020.B0100">https://data.census.gov/cedsci/table?q=median%20age&amp;g=0400000US36%248600000&amp;tid=ACSDT5Y2020.B0100</a>

csv file with population per zip code	<a href="https://data.census.gov/cedsci/table?t=Age%20and%20Sex&amp;g=0400000US36%248600000&amp;tid=ACSS T5Y2020.S0101">https://data.census.gov/cedsci/table?t=Age%20and%20Sex&amp;g=0400000US36%248600000&amp;tid=ACSS T5Y2020.S0101</a>
csv file with occupancy rate per zip code	<a href="https://data.census.gov/cedsci/table?t=Vacancy%20Rates&amp;g=0400000US36%248600000 1600000US 3651000&amp;tid=ACSDT5Y2020.B25002">https://data.census.gov/cedsci/table?t=Vacancy%20Rates&amp;g=0400000US36%248600000 1600000US 3651000&amp;tid=ACSDT5Y2020.B25002</a>
csv file with median year structure built per zip code	<a href="https://data.census.gov/cedsci/table?t=Year%20Structure%20Built&amp;g=0400000US36%248600000&amp;tid=ACSDT5Y2020.B25035">https://data.census.gov/cedsci/table?t=Year%20Structure%20Built&amp;g=0400000US36%248600000&amp;tid=ACSDT5Y2020.B25035</a>
csv file with count per highest level of educational attainment per zip code	<a href="https://data.census.gov/cedsci/table?t=Educationa l%20Attainment&amp;g=0400000US36%248600000">https://data.census.gov/cedsci/table?t=Educationa l%20Attainment&amp;g=0400000US36%248600000</a>
csv file with count and rate by marital status per zip code	<a href="https://data.census.gov/cedsci/table?t=Marital%20Status%20and%20Marital%20History&amp;g=0400000 US36%248600000&amp;tid=ACSST5Y2020.S1201">https://data.census.gov/cedsci/table?t=Marital%20Status%20and%20Marital%20History&amp;g=0400000 US36%248600000&amp;tid=ACSST5Y2020.S1201</a>
csv file with count per race per zip code	<a href="https://data.census.gov/cedsci/table?t=Race%20and%20Ethnicity&amp;g=0400000US36%248600000 1600 000US3651000&amp;tid=ACSDT5Y2020.B02001">https://data.census.gov/cedsci/table?t=Race%20and%20Ethnicity&amp;g=0400000US36%248600000 1600 000US3651000&amp;tid=ACSDT5Y2020.B02001</a>

## Exploratory Data Analysis

Figure 1 - Energy Score Distribution, and Energy Score Distribution without 0 and 100 scores

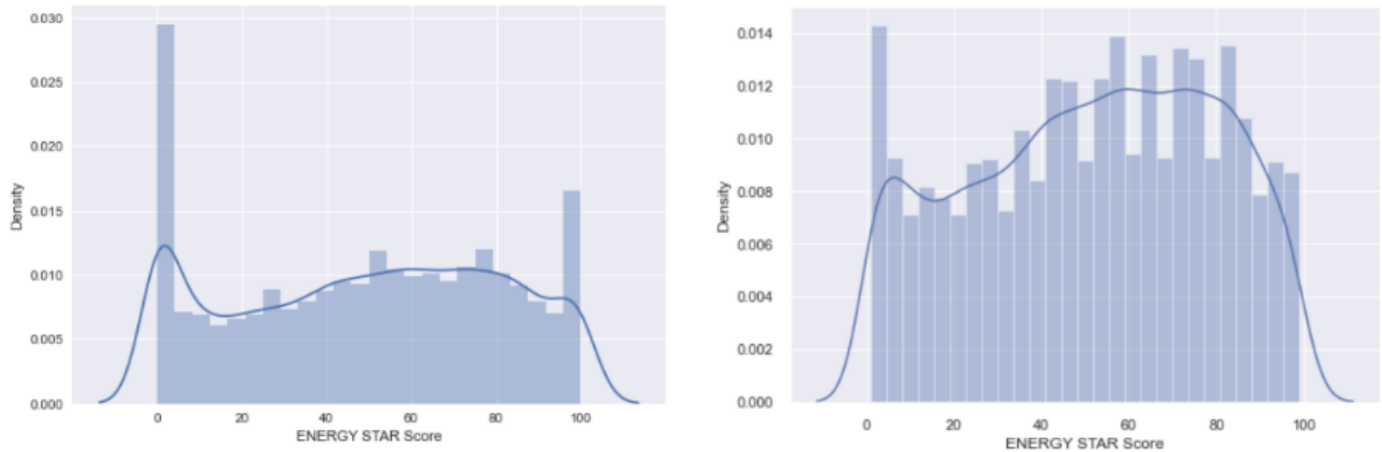


Figure 2: Letter Score Histogram

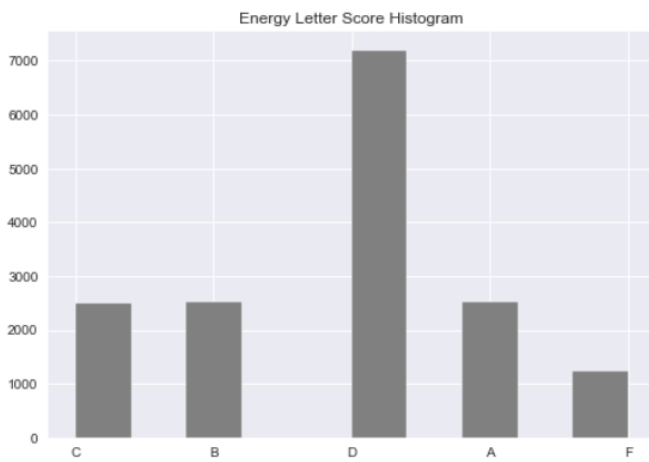


Figure 3: Energy Score Heatmap



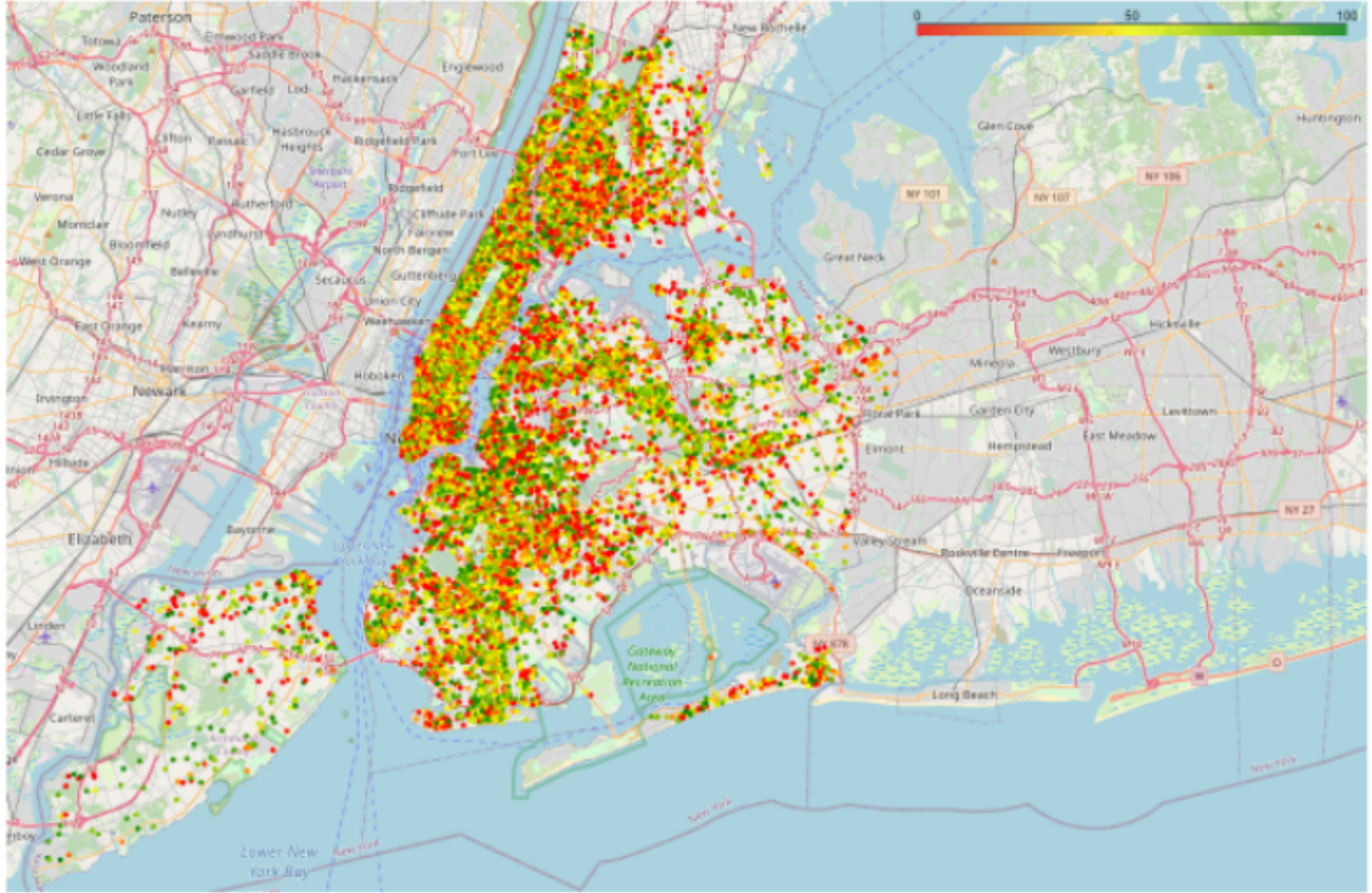


Figure 4: Mean Energy Score by Borough

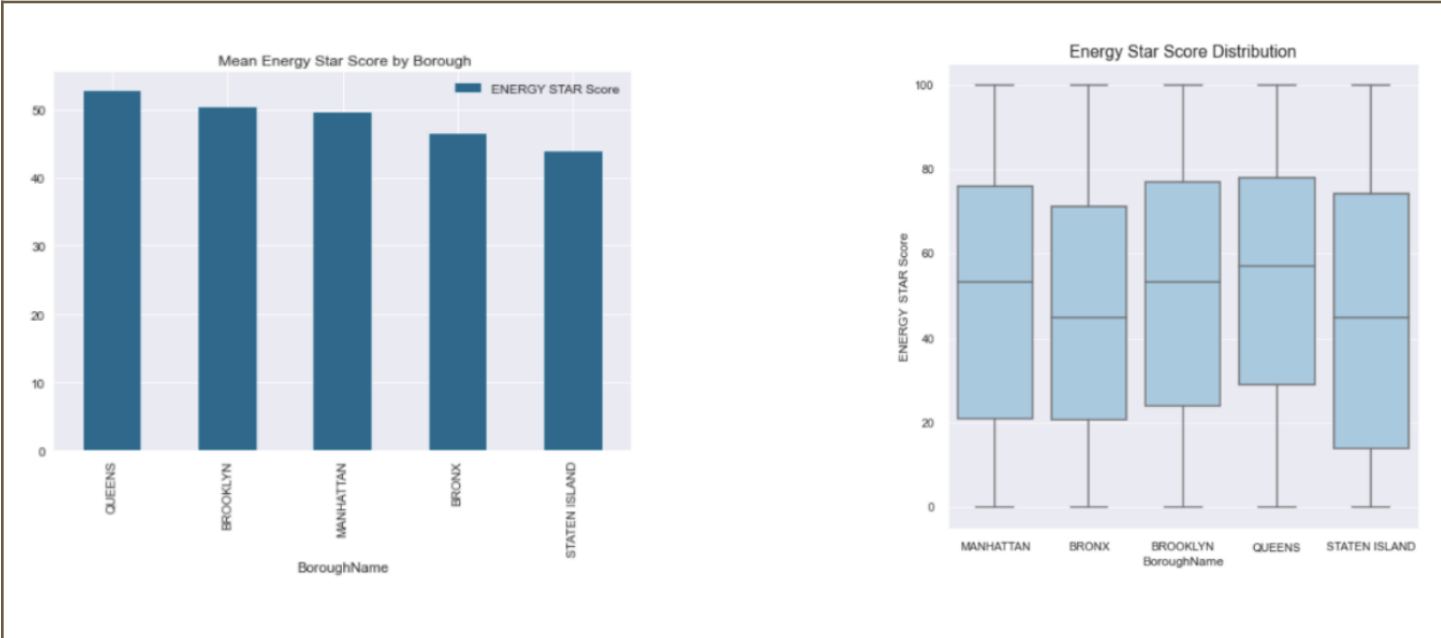


Figure 5: Energy Score by Zip Code

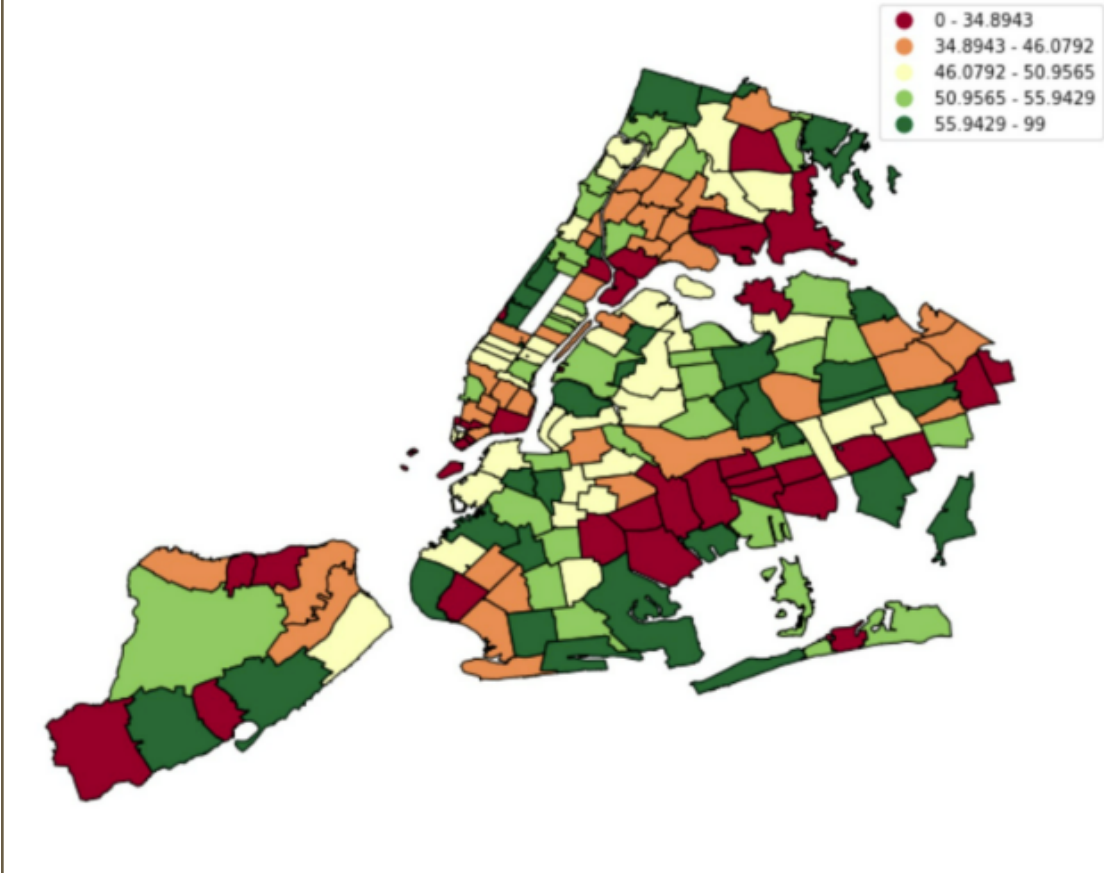


Figure 6: Mean Energy Star Score by Building Type

