CS 5304 - Final Project
Where to live in NYC

Romane d'Oncieu de la Batie
Yingying Zheng

May 10th, 2019

1 Introduction

In this project, we want to help New Yorkers find the best neighborhood to buy an apartment. We believe that there is already a lot of information out there to help you choose by size of apartment but what about other features that would not be specific to apartment characteristics?

People today are more concerned with the environment and we want to understand which neighborhoods are the best to live in if you are concerned with energy efficiency. Price being also an important concern, we aim at showing users the best neighborhoods in terms of energy efficiency and combine that with the price you would pay to buy an apartment in that neighborhood. Therefore, the first part of this project will be to find the best neighborhoods of NYC for someone who wants to minimize selling price and is concerned with building energy efficiency.

Yet, we understand that energy efficiency is not the only criteria for someone to choose a neighborhood to buy an apartment in. This is why we want to combine this data with other information such as points of interests near you. This will be our second step to this project: we will provide visualization of the distribution of different types of points of interests.

We believe that this analysis could be picked upon by an external organization to help buyers make better informed decisions seamlessly. Indeed, by building a scoring system, it would be possible to take user preferences as input and then return the best neighborhood to live in.

We published our project on the following accessible website: http://wheretoliveinnyc.com/.

2 Data Presentation and Visualization

2.1 Data sets used

For the first part of our study, we used 3 datasets found on kaggle:

1/ New York City - Buildings Database
https://www.kaggle.com/new-york-city/nyc-buildings/version/2BK.csv

2/ NYC Municipal Building Energy Benchmarking Results
https://www.kaggle.com/new-york-city/nyc-municipal-building-energy-benchmarking-results

3/ NYC Property Sales
https://www.kaggle.com/new-york-city/nyc-property-sales

All of them contain different information about New York buildings. We were able to link these databases using the location information provided by each of them. More specifically, they all contain information about the Borough Block and Lot of the buildings. Therefore, when it was missing, we generated the BBL code which is a unique identifier of a building location. Using this BBL code, we were then able to link the databases for our analysis.

It is important to note that these 3 data sets have very different sizes and the pairwise overlap of them is not complete. We will build up on this later in this report.

2.2 Buildings Database

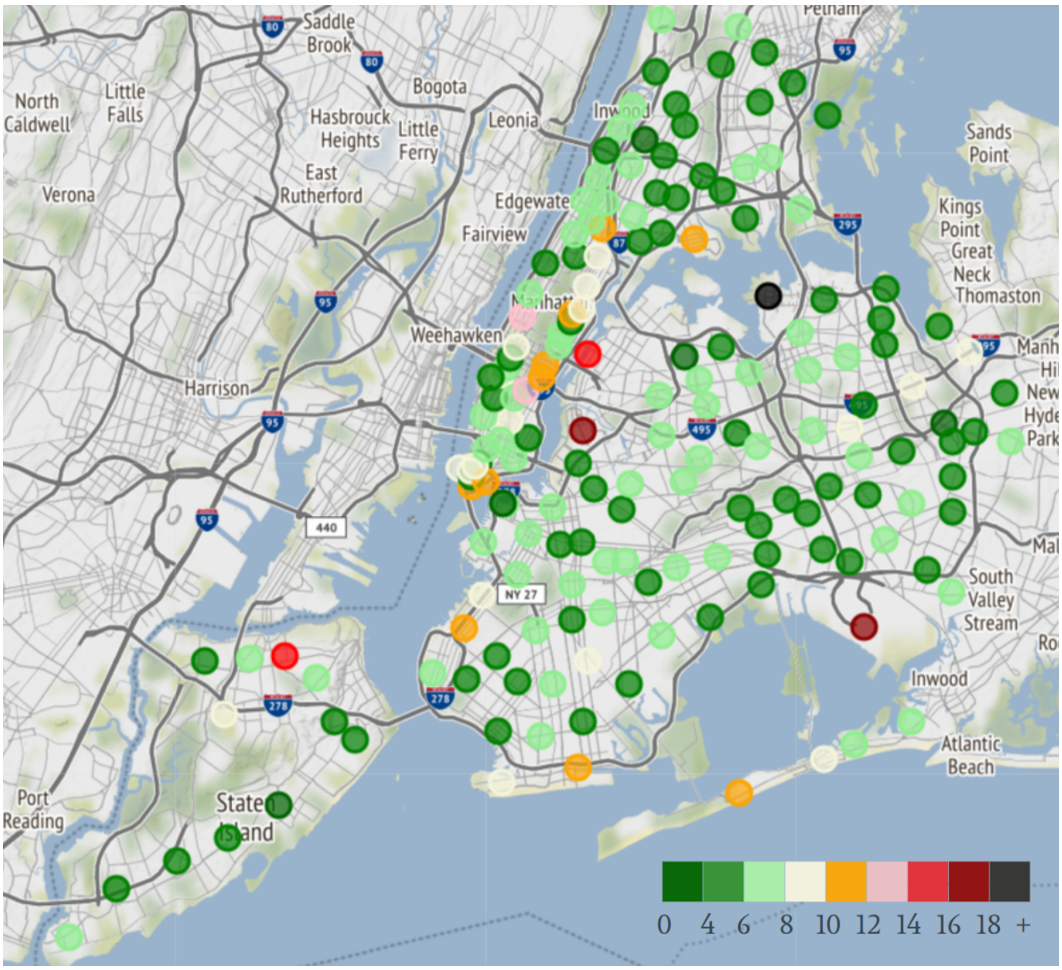
The largest data set is the first one, it contains data about all buildings in New York City over all 5 boroughs. The data was provided individually for each borough and we merged them all together when needed.

For each building, we have information about the location as well as characteristics of that building. More detailed presentation of the features included can be found on the Kaggle page of this data set. The data was last updated in 2017 and contains almost a million points.

2.3 Energy Database

This database is much smaller (around 3000 points) and contains data about all municipal buildings over all five boroughs. For each we have location information that will enable us to link it with other databases. We then have the greenhouse gas emissions intensities for year 2013 for each building. The lower the intensity the more efficient the building is.

On the map here we can visualize the average GHG emissions intensity by zip code. We can see in particular that there is the most variation in Manhattan and the other boroughs are all quite efficient. Indeed, the greener the better and the red/black dots are worse in terms of energy performance.



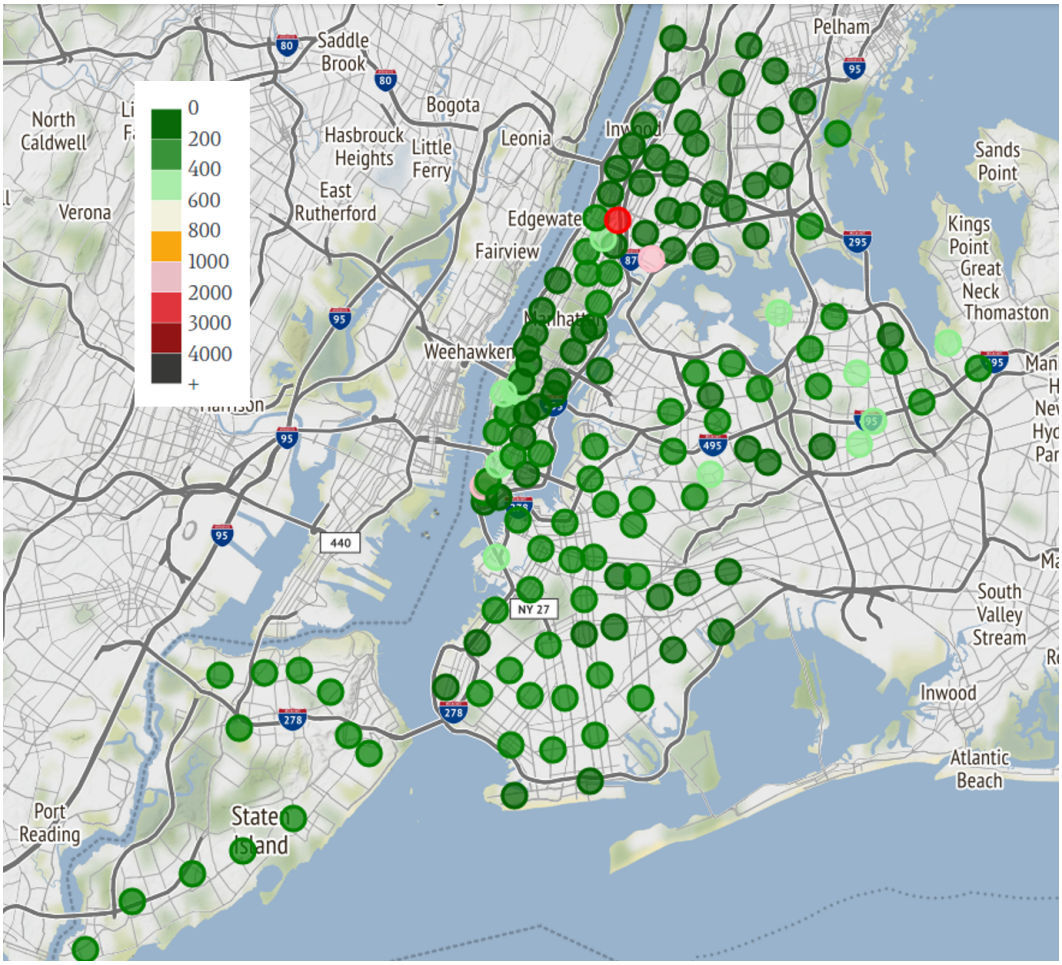
Average GHG Emission Intensity ($kgCO2/ft^2$) per Zip Code

On top of this visualization, we also printed out in our notebook, the efficiency grouped by other features than zip codes to understand for example how lot area or number of building on the lot can influence the GHG emission intensity. We also listed the energy efficiency by owner and the result was interesting. In particular, one of the most efficient buildings in NYC turned out to be the Lincoln Center for Performing Arts. When looking into this result, we found out that the Lincoln Center is particularly concerned with sustainability and efficiency, and they have spent a lot of time and money in improving their carbon footprint. It is therefore not surprising that their buildings rank very well in energy efficiency.

2.4 Sales Database

This database has 84 thousand data points and contains information about the selling price of all buildings or building units sold between 2016 and 2017 in NYC. Once again we have the location information necessary to enable us to link this database with the exhaustive buildings database.

On the map below, we can visualize the average selling price by zip code. We can see that most neighborhoods are in the same selling price range, but a few areas stand out, especially in Manhattan.



Average Selling Price ($/ft^2) per Zip Code

2.5 Zip Codes

For all of our visualizations, we needed to use zip codes to plot the data. To do so, we used the Area Connect Zip Code Search Tool to find the latitude and longitude of all zip codes in the 5 boroughs of NYC.

With the information collected, we generated a zip code data set that we included in our repository and used for our visualizations.

2.6 Additional Databases

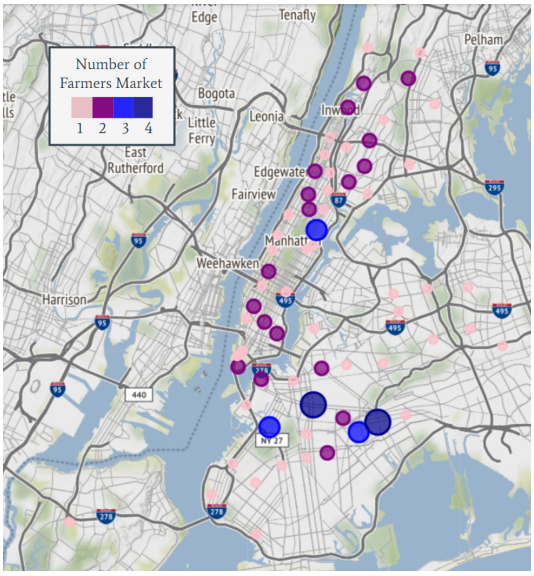
For the second step of our analysis, we want to use more data sets to visualize the distribution of different points of interests over the city.

2.6.1 Farmers Market

The first we can look at is Farmers Markets. We found this database on Kaggle at the following link: https://www.kaggle.com/new-york-city/new-york-city-farmers-markets

This dataset contains information about all farmers market over all 5 boroughs. Each data point is characterized, amongst other, by the zip code it is located in. Therefore, we were able to produce the visualization below that shows the number of farmers markets by zip code.

With the map below, we can see that the highest density of farmers markets per zip code is in Brooklyn, in Flatbush and Prospect Park. If we want to look only at Manhattan, we have a high density in the Upper East Side. Dure to the proximity of the zip codes in manhattan, we can consider that the density of farmers market is also quite high in Chelsea, the Lower East Side and Harlem.

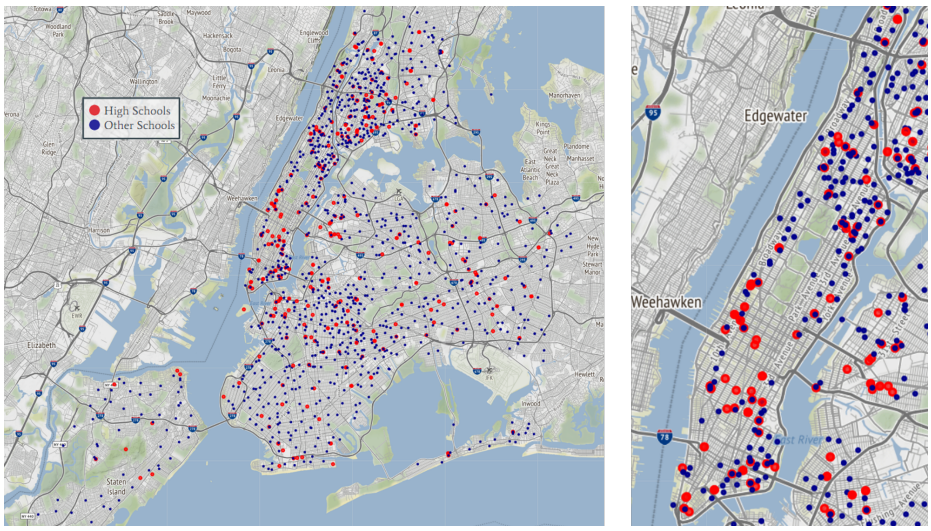


Number of Farmers Market per Zip Code

2.6.2 Schools

Next, we looked at schools. We found two datasets on Kaggle for this part. The first one contains most schools in NYC and we combined that with the second data set containing all high schools in NYC. Both of these data sets were updated quite recently and contain a quite exhaustive directory of schools and high schools in all 5 boroughs of New York City.

- Schools : https://www.kaggle.com/adamschroeder/nyc-schools
- High Schools: https://www.kaggle.com/new-york-city/nyc-high-school-directory



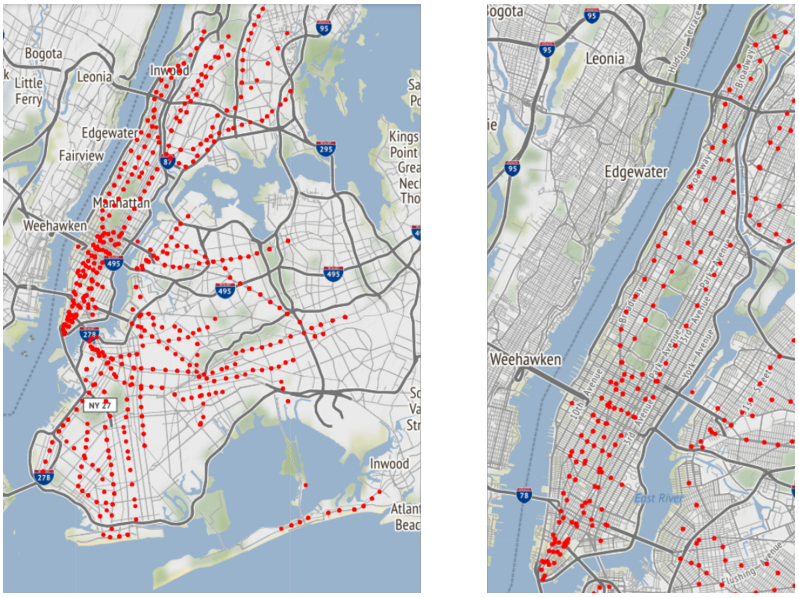
Distribution of Schools over NYC and in Manhattan

The map visualizing the distribution of schools in the city can be a very useful tool for someone with kids trying to buy an apartment. Indeed, we can easily see the locations of schools and see that the highest density is in the Bronx and Brooklyn. If we focus on Manhattan only by zooming in to the map, we can see that the highest density of schools is on the west side along the Hudson river, in the lower east side and Harlem.

2.7 Subway Stations

Then, we looked at subway stations using another kaggle dataset https://www.kaggle.com/new-york-state/nys-nyc-transit-subway-entrance-and-exit-data.

This data set contains data about all subway station across the entire city except the stops on Staten Island. For each, we only focused on the longitude and latitude and were able to visualize station locations on the map below.



Distribution of Subway Stations over NYC and in Manhattan

We can see that while some neighborhoods are very well accessible in terms of subway station density, a lot of neighborhoods in Queens, Brooklyn and the Bronx are very far from any subway stops.

Focusing only on Manhattan, we can see that the areas of high density of school locations we had earlier are locations of low subway station density. The places of highest accessibility are in the southern part of Manhattan, in the central axis of the island.

3 Objective, Method and Models

3.1 Objective

Our objective is the following:
Given someone's preferences, guide them in which neighborhood to look into for an apartment to live in.

Focusing first on finding the neighborhoods that are the cheapest and the most concerned with energy efficiency, we aim to add other features to help people seamlessly make informed decisions about where to buy an apartment in NYC.

3.2 Method

To do so, we used the following four steps method:

1/ Build a prediction model on our sales and energy data
2/ Use the model to predict the price and energy efficiency of the buildings for which we are missing the data
3/ Visualize on a map the average efficiency and sales price per neighborhood
4/ Visualize the distribution of the other points of interest by neighborhood (Farmers Market, Schools ...)

3.3 Data Weaknesses and Assumptions

We know that our data had weaknesses and understanding them led us to make a few assumptions.

First of all, the energy efficiency data was given only for municipal buildings. Therefore, we assumed that it made sense to use the model built on municipal buildings to predict the efficiency of the other buildings (residential for example).

The second main weakness is the number of data points we had to build our models compared to the number of data points for which we have to predict values. We believe that the predictions we are making would be best made separately for each borough. However, because we had few data points per borough, we made two assumptions. The first one is that it makes sense to build an efficiency prediction model over all 5 boroughs. The second concerns the sales predictions. We assume that it makes sense to build a model by borough, even if that model will be built on few data points.

3.4 Models

To build the best models, we iterated and tried different options. Eventually, we chose the models that procured the highest regression score. More specifically, we chose the following:

	Model	Score
Energy Efficiency	Random Forest Regression	0.69
Manhattan Selling Price	Random Forest Regression	0.92
Brooklyn Selling Price	Random Forest Regression	0.09
Queens Selling Price	Random Forest Regression	0.26
Bronx Selling Price	Ridge Regression	0.06
Staten Island Selling Price	Random Forest Regression	0.67

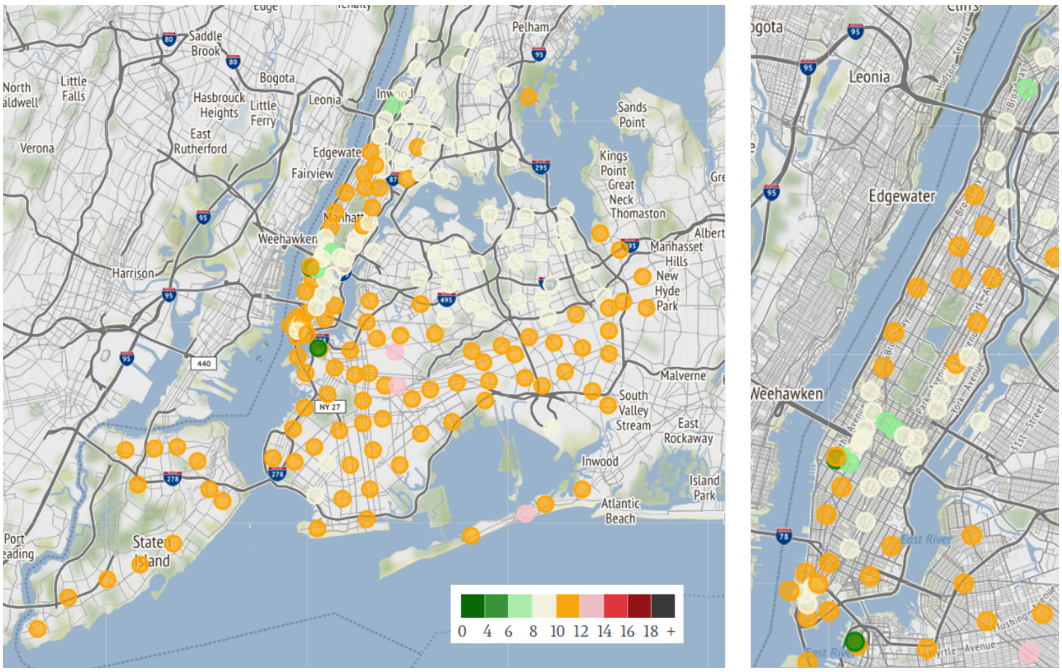
We can see that some of the models perform pretty badly. We tried different regression models to try to improve these scores and yet these were the best we got. As mentioned in the weaknesses and assumptions, this comes from the fact that we do not have enough data points in our sales database.

4 Predictions Results

We then used these models to predict the selling price and energy efficiency of all buildings. More specifically, we now included the buildings that were not used to build the models.

4.1 Energy Efficiency Prediction Results

On the NYC map below we can see the results of our energy efficiency predictions.



Average GHG Emissions Intensity Predictions by Zip Code (kgCO2/ft^2)

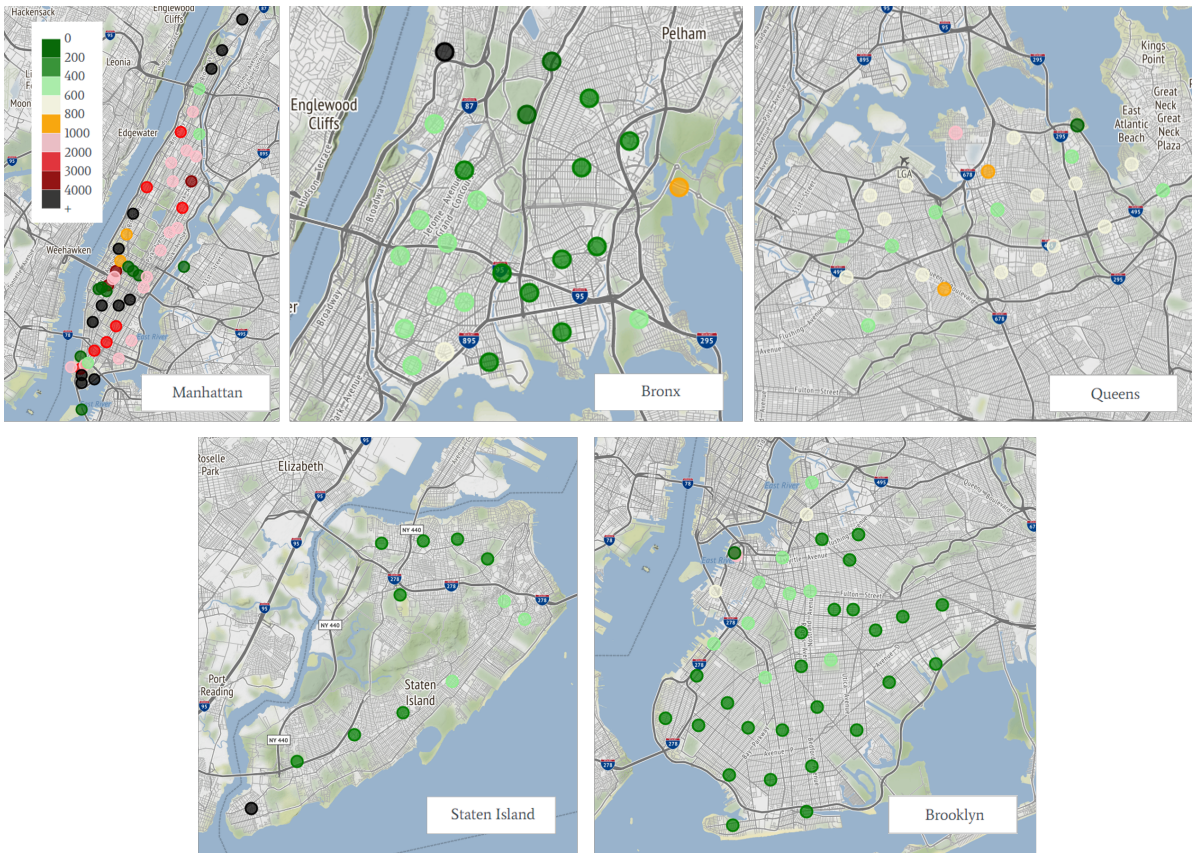
We can see that the general trend is that all Zip Codes have higher GHG emission intensity than on the visualization of the data (2.3). This can be explained by the environmental policies of NYC. In the early 2000s, NYC passed a policy to reduce GHG emissions, they targeted buildings in particular because they saw that it was the largest source of GHG emissions. Although a lot of improvements have been measured on government buildings the same cannot be said for residential buildings which explains why the map with all buildings has higher emission levels. Indeed, when considering all building, the non-municipal buildings bring the average up. The link below describes the energy efficiency policies and the measured results.

https://www.dec.ny.gov/docs/administration_pdf/nycghg.pdf

If we focus on Manhattan only, we can see that their are a few neighborhood with relatively better GHG Emission Intensities. Similarly we can see that the neighborhood of Dumbo in Brooklyn stands out on this map.

4.2 Selling Price Predictions Results

We trained different models for each borough to predict sales prices within that region. We then visualize the predictions on the maps below.



Selling Price Predictions by Zip Code ($/ft^2)

The colors from dark green to black imply increasing prices. The prices are shown as dollar value per square foot. It is clear that some of the most expensive buildings are in Manhattan and the cheaper ones are located in the other four boroughs. For Manhattan, some of the cheaper areas are in Chelsea and Mid-Town areas.

Though not perfect, our models and visualizations of the predictions of the sales prices can help one gain a basic understanding of sales prices for the locations in their mind. One can combine the sales price prediction result with other criteria such as energy efficiency, number of train stations, and other preferences in order to pick the best place under their budget.

5 Recommendations

5.1 Results

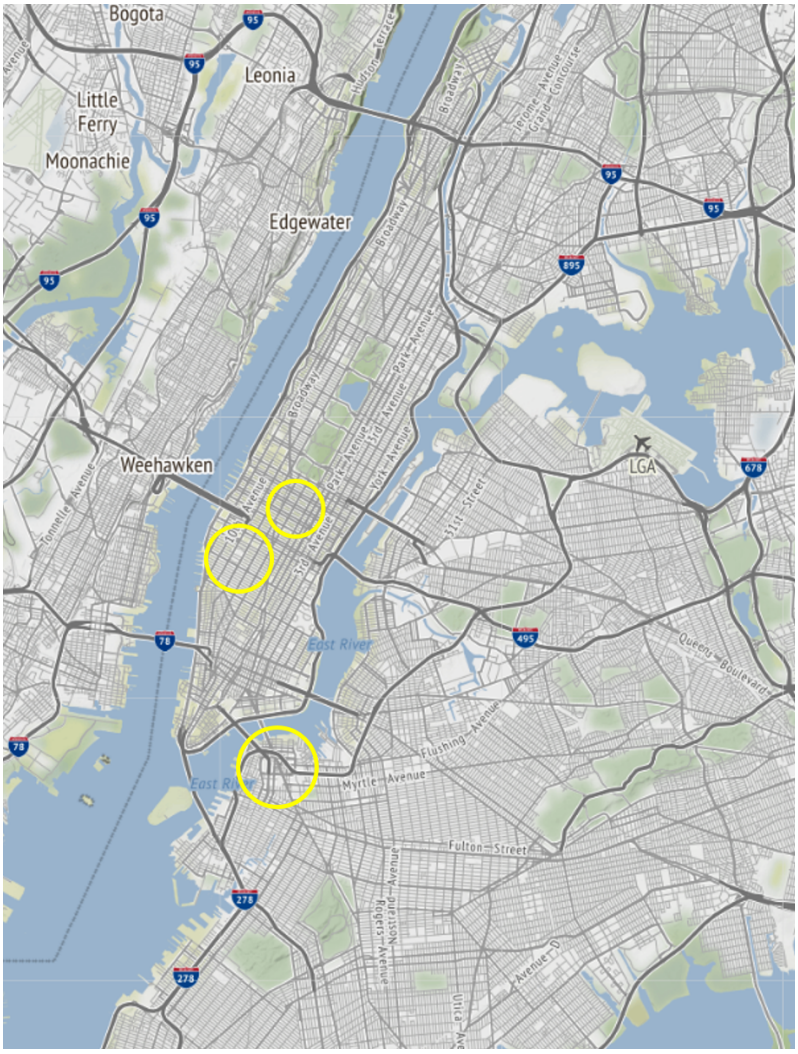
Given the visualization seen previously here would be our recommendations for users. If you are concerned with low selling price and low GHG emission intensity, and if you want to be close to schools, farmers markets and subway stops, you should buy an apartment in

In Brooklyn, the best neighborhood would be **Dumbo** as it satisfies all the preference criterias. There are neighborhoods with more farmers markets and more schools but if you want to bring together all preferences, Dumbo would be your best place to look.

In Manhattan, the best place to satisfy all these criterias would be in **central Midtown or the lower Chelsea area**. They are area of great variation in terms of price as we have some of the best and worst neighborhood. However, they really stand out in terms of energy efficiency on Manhattan. Also, they have a good amount of Farmers Markets and schools (both K12 and high schools). To add to this, these Manhattan locations benefit from very good subway accessibility.

Areas in the Bronx, Queens and Staten Island will satisfy some criteria but it will be hard to satisfy all of them as these boroughs all have great weaknesses in one of these fields. Queens and Staten Island both have accessibility issues and relatively less schools and farmers markets than other boroughs. For the Bronx, additional information (for example from other projects from this course) can show that this borough suffers from important criminal issues. Therefore, someone looking to settle down with children would not be recommended to buy an apartment in the Bronx. However, looking only at our analysis, all criterias are met in the Bronx.

The map below shows what we consider to be the 3 best locations to satisfy these criterias, the area circled in yellow being the area one should look into.



Selling Price Predictions by Zip Code ($/ft^2)

5.2 Next Steps

We believe that this project is the first step to a larger project that could really make a change and help people make better informed decision in a seamless way. More specifically, we think that their are a few steps left that an organization could take to make this into a real usable product.

For the purpose of this study, we decided to view our results on different visualization maps. However, we did not bring the all together. The first next step would be to create a scoring system that would give a score to each neighborhood given their value in each of the different features (price efficiency, farmers markets, schools...). To build this scoring system, we would first need to standardize the values of each of these features. Then, we would give an importance to each of them to give them a weight in the final score. Adding the weighted features together for each neighborhood would then give us a score for each neighborhood and we can return as an output the first few neighborhoods with the highest scores.

Finally, the last step would be to develop a user friendly interface where users would be able to give their preferences in terms of different factors such as price, energy efficiency... and the first few best neighborhoods would be returned to them on a map.

Another idea to build up on this project would be to add in more features that users could choose with to give more traction to this idea.

6 Guide to our Code

Our project is accessible on Github and the guide below to run our code is also detailed in this repo. *https://github.com/romanedoncieu/Where-to-live-in-NYC*

The first Notebook we used was to visualize energy efficiency for municipal buildings provided by the database. This same database also presents the model search and prediction of energy efficiency over all buildings in the Buildings Database. The 3 visualization notebooks contain the code to download the data and visualize the distribution of points of interests over a NYC map.

Most of our code was implemented in Google Colaboratory, you can view it using the following links:

- Energy predictions: *https://colab.research.google.com/drive/1uhzLjmmStZZ36L_dyNyAKLQlCFC9XJic*
- Farmers Markets visualization: *https://colab.research.google.com/drive/10gcDcwSnz6iww00VpoX9MeS4mu5V4nA8*
- Schools Visualization: *https://colab.research.google.com/drive/1vcTIEHlPzwPGqVNUjxjCKfcRG5P2xXoV*
- Subway Station Visualization: *https://colab.research.google.com/drive/1YtEaHI8YxCVkDTNY8FWx6Gp5_FJrfmIg*

All the data to run these notebooks are accessible in the following Google Drive folder: *Final-DataScience*

Our sales predictions however were run using iPython Notebook, you can download the code (Sales Predictions.ipynb) and the data required to run it in our github repository.