Final Deliverable
Team 3
Members: Dallin Gordon, Shuo Zhang, Sonu Kumar, Rakin Munim

Introduction

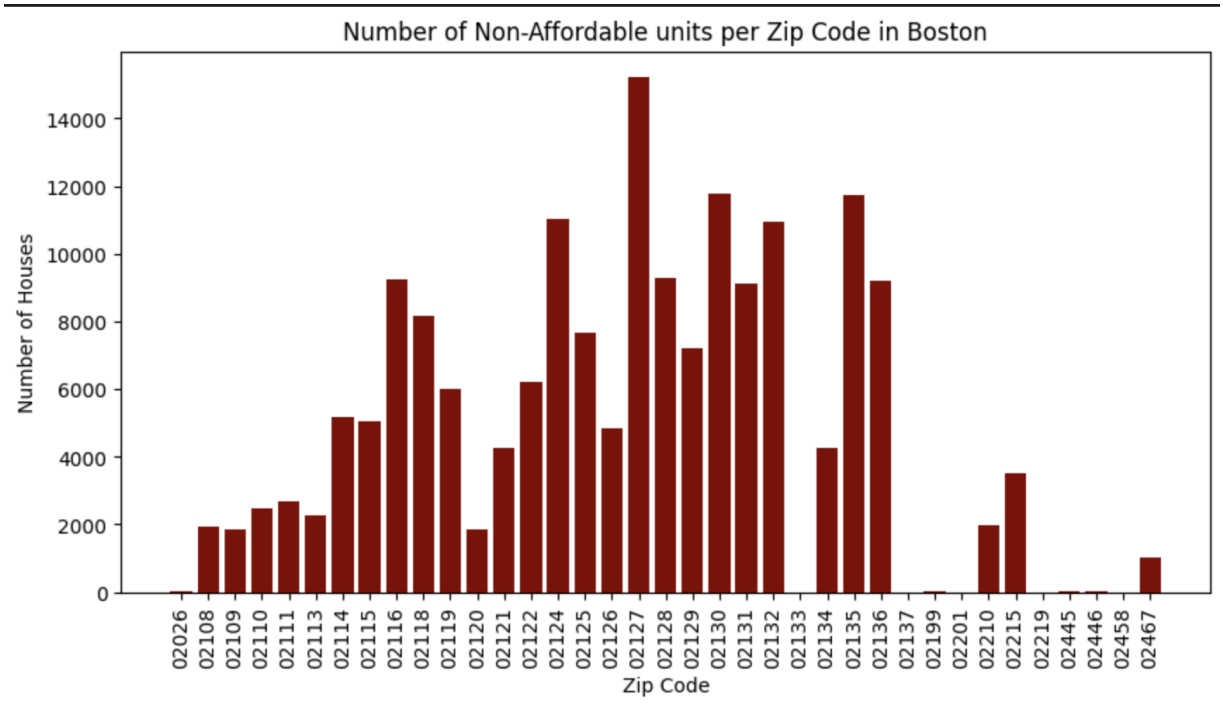
The main goal of this project is to get a better idea about the distribution of landlords not enrolled in different affordable housing programs. Important features to be considered: number of units, geographic location, and the demographic profile of the census block group. We will also look into the percentage of housing stock that is owned by the owner occupied and the small landlords, and at % affordable.

In addition to exploring features of affordable housing, we will dig into the details on whether housing has the potential to become affordable housing, we also merged with voter data to see the identity of the owners.

Basic Project

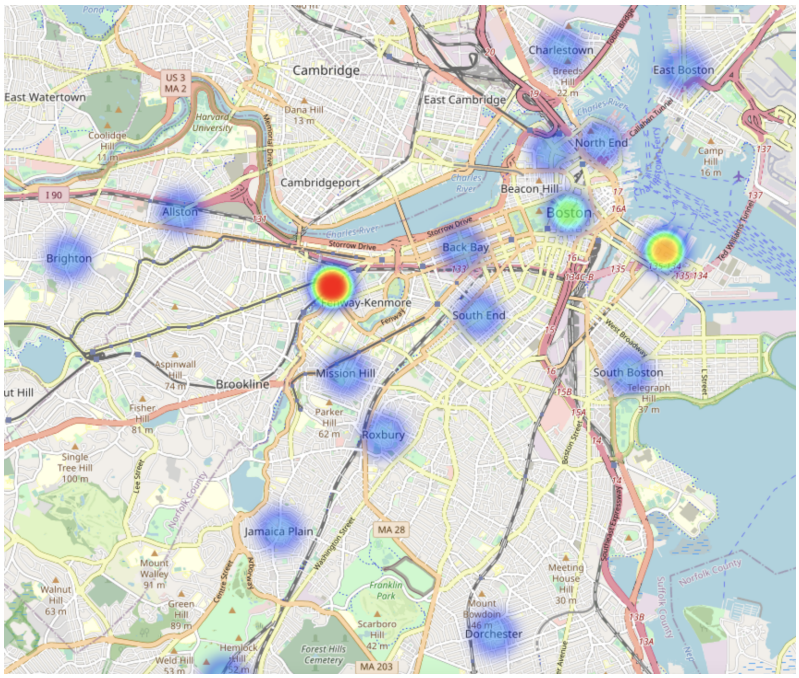
Q1: What is the current distribution of landlords NOT currently enrolled in different affordable housing programs?

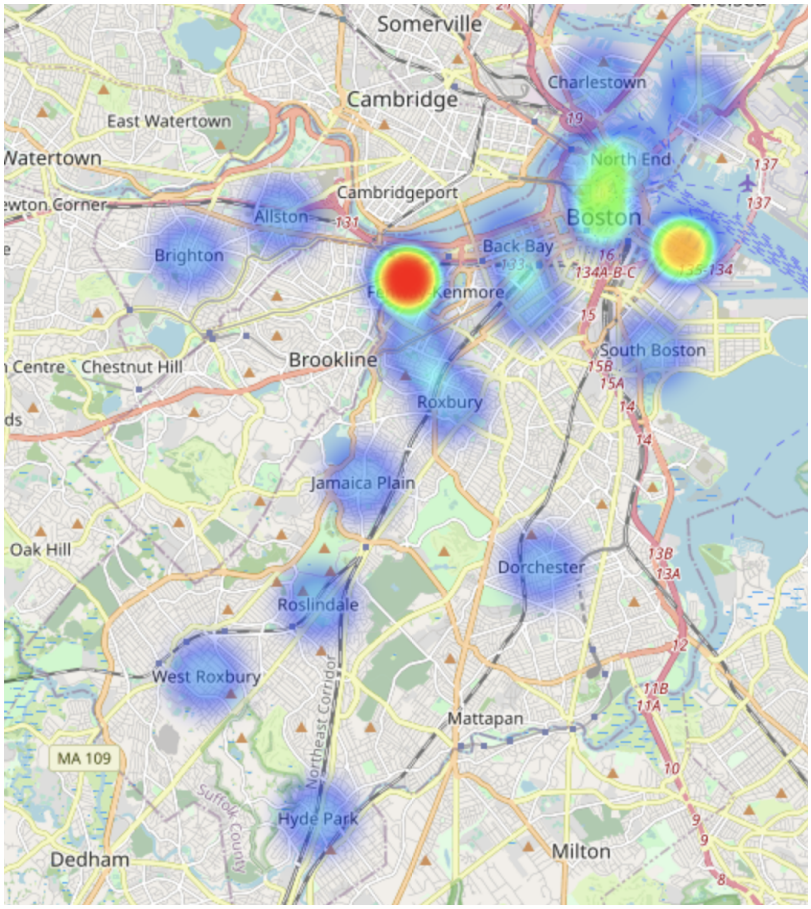
1. # of units
    a. 99 % of landlords are not enrolled in affordable housing programs:
2. Geographic distribution (by zip code)



Q2: What is the current distribution of landlords and housing listed in current affordable housing programs?

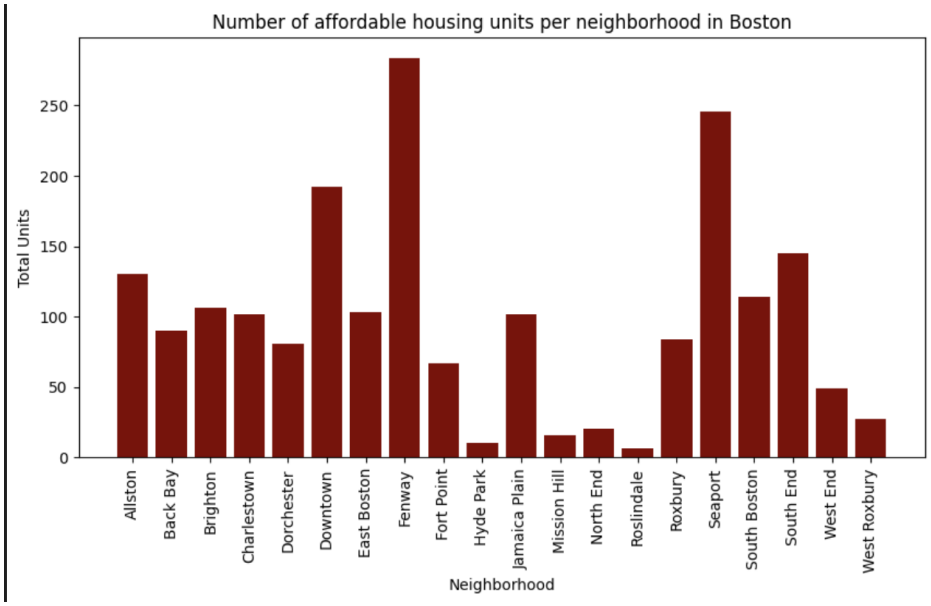
Here are the results for the Affordable Housing Stock dataset of Boston
Link to data: https://www.bostonplans.org/housing/finding-housing/property-listings





The two images above showcase the distribution of affordable housing in Boston by neighborhood. Note **this means that the heat maps you see are more distributed in real life but our heat map points are all aggregated to a single point for the whole neighborhood**.

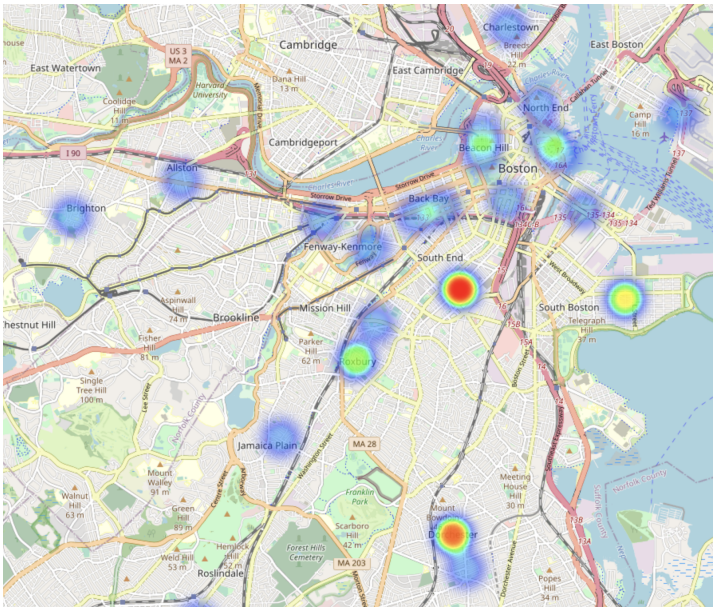
Here is a histogram for this data with neighborhood:

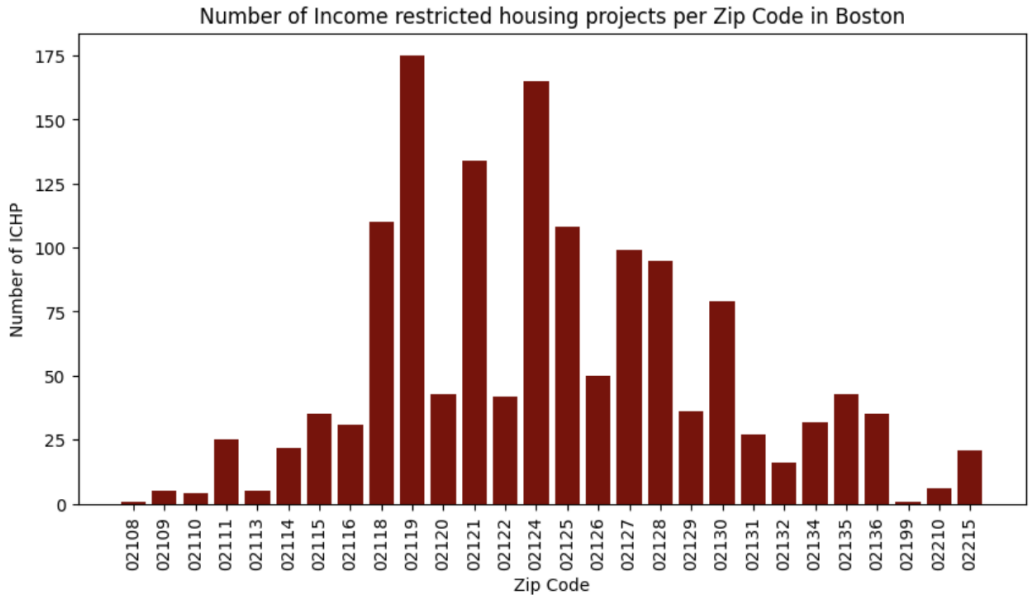


Our interpretation: From this analysis we can see that the distribution is bimodal and the affordable housing stock seems to be concentrated around the downtown area between Fenway and Seaport. There also does seem to be a roughly uniform distribution, if a few neighborhoods such as Hyde Park, Roslindale, North End, and Mission Hill, are ignored.

Distribution of housing units for the Income-Restricted Housing Units Program:
Dataset:
https://data.boston.gov/dataset/income-restricted-housing/resource/464bd32f-ebac-49e4-884a-01c4549d3cd3

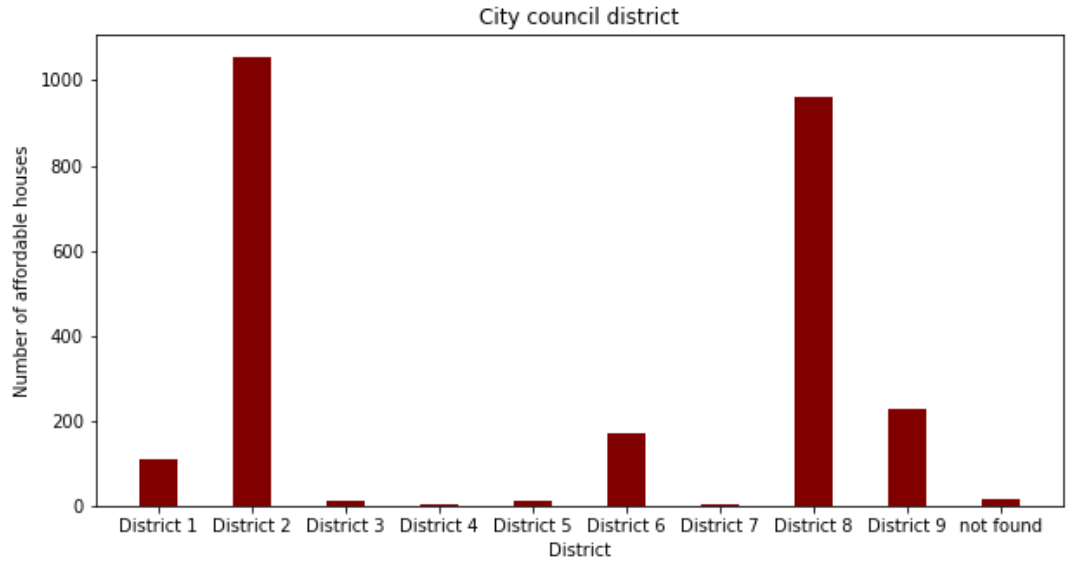
Associated Histograms

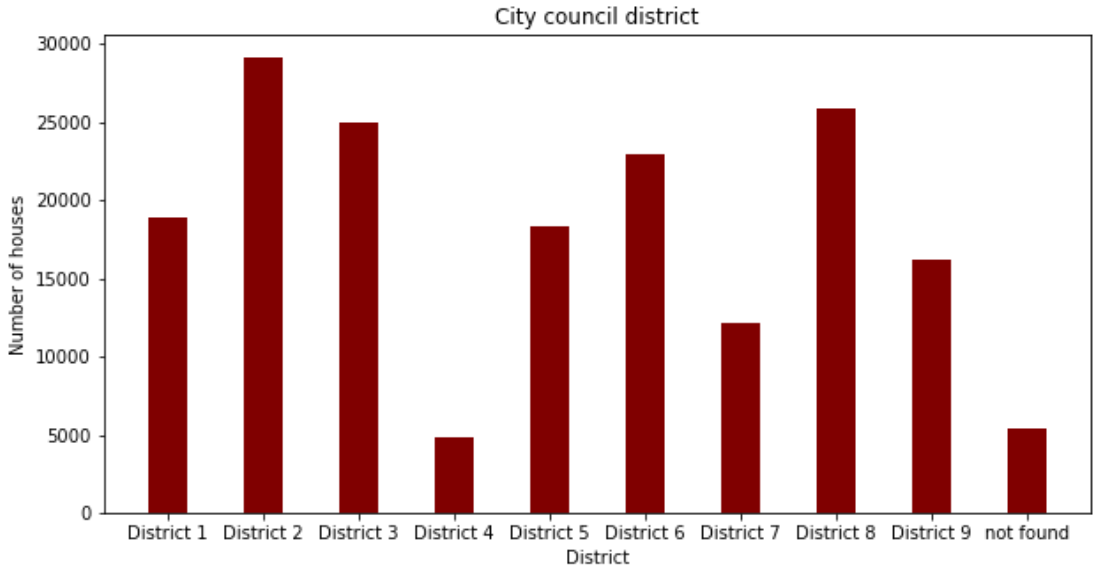




Our Interpretation: In contrast to the uniformness of the affordable housing stock data, the income restricted housing seem to be more random and we see larger disparities between zip codes. We see much of the income restricted houses in the neighborhoods of Roxbury and Dorchester. These two concentrations open interest to us as to why such a strong concentration exists. We also want to find out why areas like Allston and Brookline seem to have very little income restricted housing.

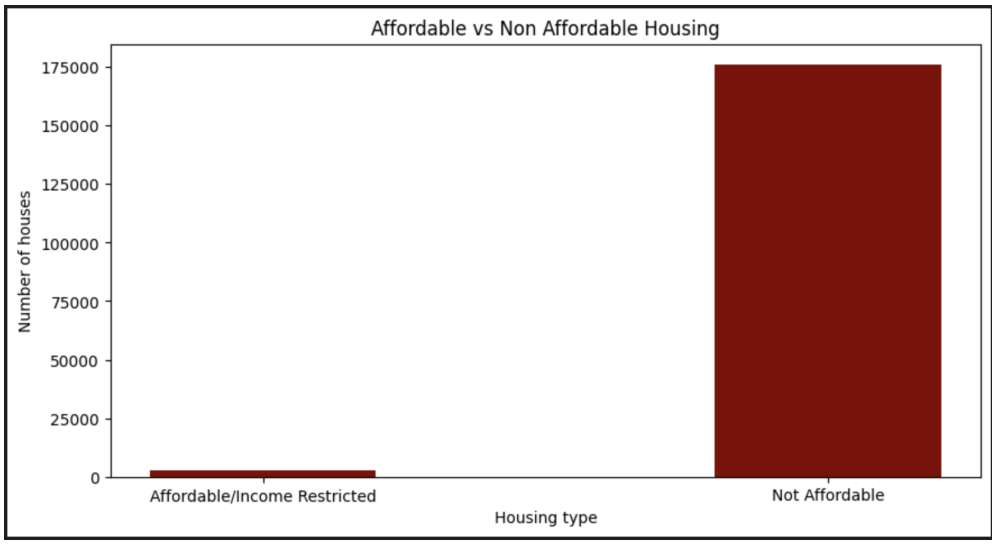
Q3: What is the geographic distribution of these landlords by city council district?
Housing per district

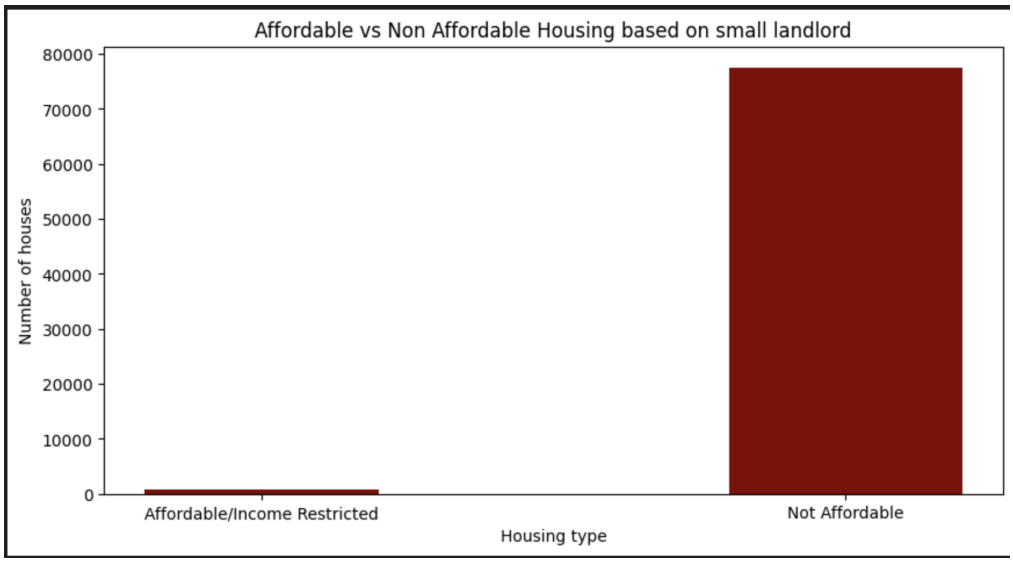




Interpretation: Most affordable housing is concentrated in district 2 and district 8; this matches the trend of affordable housing and the zip code graph. Such a finding shows that affordable housing is not evenly distributed in Boston. Several possible situations may lead to this result. First, districts that include neighborhoods such as Newbury may have many commercial housings. Second, the area size of each district may vary. Third, housing owners of districts 3, 4, 5, and 7 have not participated in affordable housing programs.

Q4: What percentage of housing stock is owned by owner occupied and small landlords, and at what % affordable





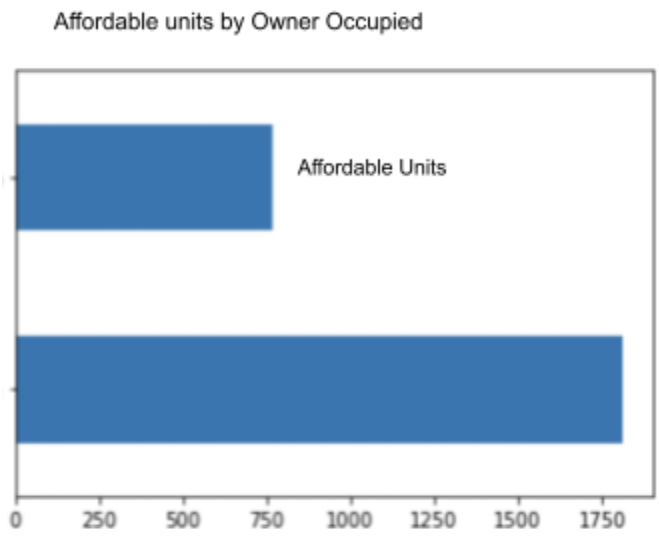
Only about 1.4% of total housing stock is actually affordable

Interpretation: Roughly 1% of owner occupied housing is actually deemed affordable which is about the same percentage as the overall amount(~1% as well) which suggests to us that owner occupied housing isn't any more or less likely to be affordable housing based on the charts above.

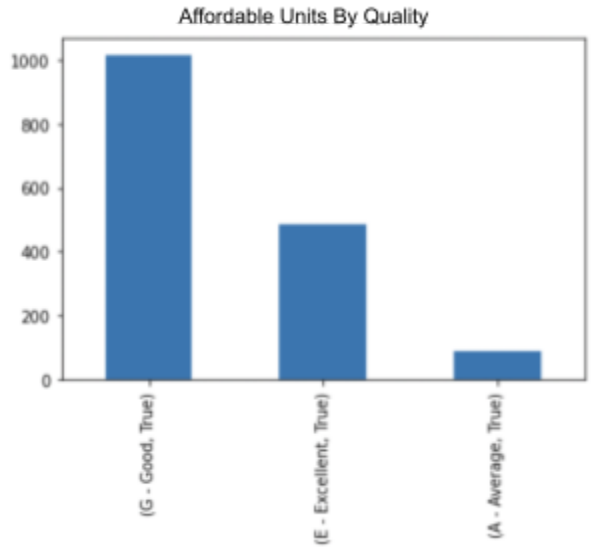
Extension Project

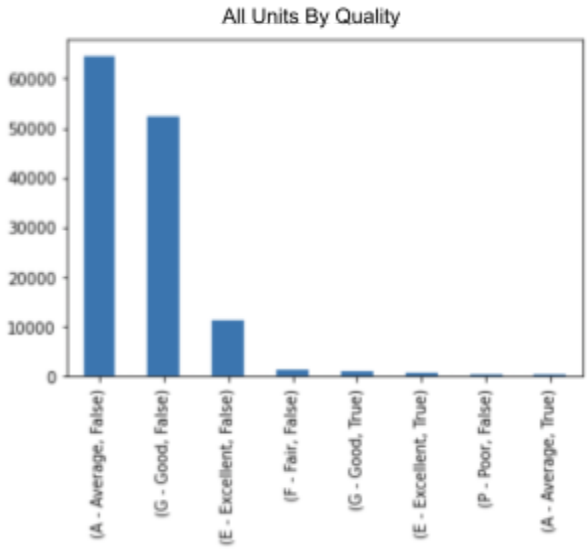
Data Collection.

Collecting the data was not difficult, we acquired one external dataset needed to facilitate joining. The actual assembling was a challenge. We needed a list of neighborhoods and area codes. It was not difficult to find. We did some manual work to make the names of the neighborhoods identical in each dataset. The biggest challenge was cleaning and joining. The Inventory dataset was rich. It had address and area code, and ultimately neighborhood. The affordable housing datasets both lacked these. We used building name and partial addresses found there, passed those through google places api, and were able to join them to the Inventory dataset that way. The api calls to google were done in the location_api_and_dataset_ensebling notebook. You will need a google api account and key to run that. The joining was performed in the cleaning_and_join notebook. It generates the Is_Affordable_Master_Dataset.csv. The voter data we had access to was aggregated at voter district in one file, and by neighborhood in the other. We examined both, but used the neighborhood level data for the join.



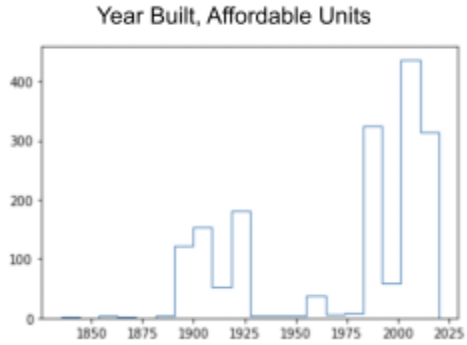
After collecting the data and doing preliminary analysis, we started looking for well correlating variables. And our first major challenge became clear to us. The client made us aware that the affordable units we had in our dataset was a subsection, and likely a small subsection of the actual affordable unit stock. We had just over 2000 units listed as affordable. Most of those were not owner occupied. We also think that lower quality units are likely left out. This is when it became clear to us of the major bias issue. Even after the joins that we had to do, which did inject a degree of uncertainty, we can only be reasonably confident in them. The





negatives cases could be existing affordable units outside our datasets, potential units for an affordable program, or true negatives.

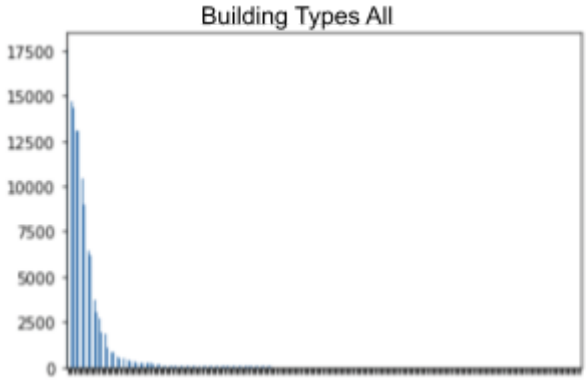
There were similar biases with the year the building was built. We were interested in seeing if there was a correlation between the age of the building and its presence on the list. We thought either older buildings would be more likely used for affordable housing programs, or that it would benefit us to look for new units with a higher probability given the year. In stead we found another anomaly with the data. (found in Extension notebook)

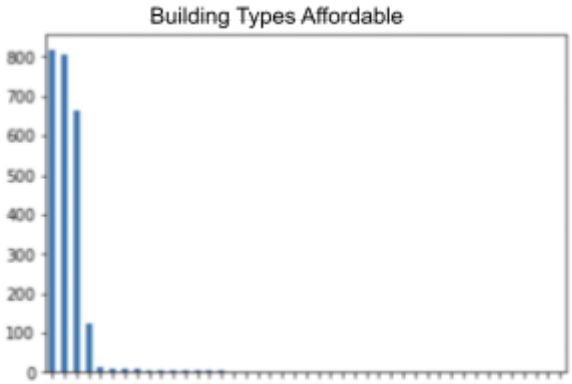


The multi modal nature of the Buildings in the Affordable set is very visible. We did make attempts to incorporate it into further analysis, but it did not seem explanatory or useful. It appears that there is some considerable selection bias in the Affordable data we had access to. We have access to the higher quality, newer buildings.

Potential Affordable Units.

This initiative was born out of necessity, our exploration, and client interest. We continued to look at the inputs from the inventory dataset to see what would make for a good input to a predictive model. Due to the extremely disproportionate and biased differences in the positive and negative cases,





We used a subsampling scheme. We trained a logistic regression on some features that we thought would be useful and descriptive. We had some issues with collinearity that caused us to leave out some inputs. We also binned the categorical inputs by necessity (for example, we only included Building Type where there were at least 10k instances, many of those with fewer were not represented in the Affordable datasets, those included are shown in our regression output below). We looked at a number of different

potential inputs. We left out some inputs (like Building Quality) because of perceived bias. We ultimately decided to use Building Value, Owner Occupied, Residence Number of Units, and Building Type. Included in our results is the average score from this Regression performed 3 times on a sample of all the Affordable units and 5000 randomly selected samples from the units not in our Affordable set. (found in the Prediction Models notebook, outputs of the individual models are stored as reg1.png, reg2.png, reg3.png)

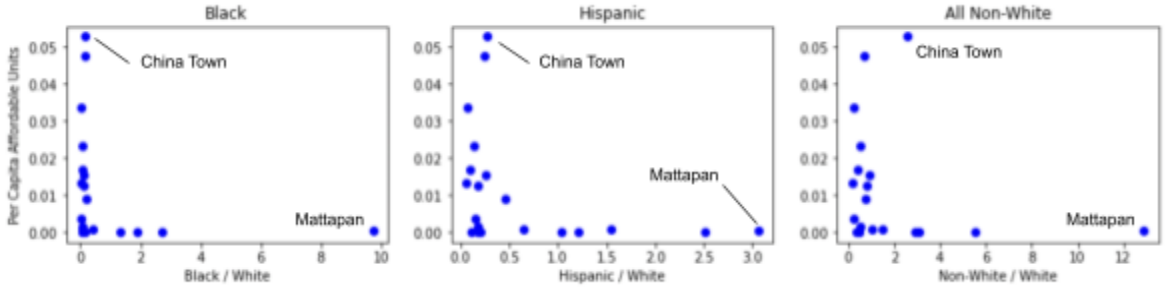
Regression 2 output

Intercept	-4.9910
C(RES_UNITS_INPUT)[T.1]	1.6784
C(RES_UNITS_INPUT)[T.2]	-16.3484
C(RES_UNITS_INPUT)[T.3]	-1.2718
C(RES_UNITS_INPUT)[T.4]	0.2497
C(RES_UNITS_INPUT)[T.5]	-1.1769
C(RES_UNITS_INPUT)[T.6]	-30.5638
C(RES_UNITS_INPUT)[T.Other]	0.9208
C(BLDG_TYPE_INPUT)[T.CV - Conventional]	0.1253
C(BLDG_TYPE_INPUT)[T.DK - Decker]	-0.0689
C(BLDG_TYPE_INPUT)[T.FS - Free Standing]	-19.5089
C(BLDG_TYPE_INPUT)[T.LR - Low Rise]	2.1645
C(BLDG_TYPE_INPUT)[T.MR - Mid Rise]	4.1522
C(BLDG_TYPE_INPUT)[T.NoBld -]	6.1152
C(BLDG_TYPE_INPUT)[T.Other_type]	3.4469
C(BLDG_TYPE_INPUT)[T.RM - Row Middle]	-0.5638
BLDG_VALUE	0.4547
OWN_OCC	-0.0049

From our regression analysis, there were some notable coefficients. 1 unit availability had a positive correlation with being affordable through all 3 regressions, the rest of the units we separate out (2-6) were negative. The highest correlation was actually above 6 units (T.Other). This went against client expectations. There were 2 building codes that were notable. Mid-Rise and No Building Code units both had high correlations with affordability. Building Value had a positive contribution to the affordability probability. We are not sure if that is due to the data bias, but it was consistent across all three regressions, and we like the idea of higher quality units being added to the list. Being owner occupied is very close to zero in all three cases. Given the seemingly cherry-picked nature of the data, we do think it would be worth additional effort to add owner occupied units to affordable housing programs. We found 13675 units that were owner occupied and we predicted as affordable. This list can be further focussed by using the probability we include in the reg_avg column of the master_with_regression_output.csv file.

Voter Data

We wanted to get a measure of under serving; what segments of the population, if any, were being inadequately provided with low income housing. Due to the anonymous nature of the voter data, we could not join on anything more granular than neighborhood. Those challenges have been addressed previously. At the neighborhood level we looked at the number of units, and graphed it compared to the relative percent of black to white populations. Number of units alone didn't take into consideration the entire population of the neighborhood. So we graphed the per capita affordable units by the ratio of minority to white population of a neighborhood. (found in the Extension notebook)



Each point represents a neighborhood. The y axis is the per capita affordable units in the neighborhood. The higher a point is, the more "coverage" the neighborhood has. The x axis is the ratio of the minority group to the white population of the area. The farther a point is to the right the less "white" the neighborhood is. The disparity is shocking. Even with the bias toward high quality units in the affordable datasets, The lack of coverage for areas with higher minority populations is a major social failure. China Town appears to have a little better coverage for the smallest minority groups. It is however the best served area. The data used in the minority separated graphs is in the affordability_by_Neighborhood.csv file, or can be generated from the Extension notebook. Filtering by the aff_per_cap column in ascending order will show you the most underserved neighborhoods first. Filtering by the ratio columns in descending order will show you the areas with highest minority population concentrations. Mattapan is tragically underserved. Our regression identified 73 units in Mattapan as being affordable candidates. These predictions are available in the master_with_regression_output.csv as well as in the Predict Models notebook.

Master_with_regression_output.csv.

This is the output file that contains everything. It is over 100 MB so we left it off of git. To generate this file, run the Prediction Models notebook. Running the last cell will save it. It contains all the headers from all the datasets. The Is_Affordable_Master column is True if the unit was in either of the Affordable datasets. The reg_avg is our composite regression score. It is a probability between 0 and 1, 1 being the most likely to be an affordable candidate, zero the least. We suggest looking for potential new affordable units using this score. The NEIGHBORHOOD column will be useful as well, to narrow down the results to a particular area. OWN_OCC is a binary column where Y is owner occupied and N is not.

How To Use our Codebase:
Files and Instructions

File: landPDExploration.ipynb
Usage Instructions: This notebook is to be used to analyze the bostonParcelsData.csv. This csv file was generated from the dataset found at: https://datacommon.mapc.org/browser/datasets/360. The bostonParcelsData.csv holds all properties relevant to Boston and can be used directly in a data pipeline without cleaning.

File: incomeRestricted.ipynb
Data Source:https://data.boston.gov/dataset/income-restricted-housing
Usage Instructions: This notebook allows for analysis of Boston's income restricted datasets. The notebook is divided into two parts, one dedicated to the former and the other for the latter. Usage of these notebooks allows for matplotlib visualization generations as well as geocode map generation. The only requirements for this notebook is the csv file :

1. Income-restricted-inventory-2021.csv

File: data explore.ipynb
Data Sources:

1. ***https://www.bostonplans.org/housing/finding-housing/property-listings***
2. ***https://data.boston.gov/dataset/property-assessment***

Usage Instructions: This notebook is used for analysis of affordable housing dataset and the 2020 Property assessment dataset. This notebook is a very useful since these two datasets are part of the the code datasets for this project. The notebooks, when run, load the datasets into dataframe and provide some analyses. Afterwords the user can manipulate these data frames as needed to actually produce more analysis.

File: masterDataAnalysis.ipynb
Data Source: Is_Affordable_Master_Dataset.csv (IAM) (Team generated)
Usage Instructions: This notebook uses the joined dataset we created from the affordable housing dataset and income restricted data set. The IAM dataset performed an inner join on both datasets giving us a complete single dataset without repeats. Additionally full addresses have been added for ease of use and mapping. The notebook creates a dataframe and outputs some basic analysis and from there the user can manipulate and extract data from the dataframe.

File: cleaning_and_join.ipynb
Usage Instructions: This notebook was used to merge the affordable housing and income restricted housing datasets together. If new data is generated for these files, and Is_Affordable_Master_Dataset.csv needs to be recomputed this notebook is to be used.

File: city_council_district.ipynb
Usage Instructions: This notebook was used to group housings into city council district. It will generate a column based on Is_Affordable_Master_Dataset.csv called DISTRICT and store the new data set into master_with_district.csv

CSV File Descriptions:
Dataset: neighborhoods_and_codes.csv
Usage: Contains a mapping from Boston Neighborhoods to Area Codes.

Dataset: bostonParcelsData.csv
Usage: Contains a cleaned version of the full eastern MA parcels dataset. This dataset particularly includes only Boston parcels, making it ready for data pipelining.

DataSet: Is_Affordable_Master_Dataset.csv
Usage: Contains an exclusive join between the affordable housing dataset and income restricted housing dataset. The key columns to keep in mind are:

1. Income_restricted_inv
2. Existing_affordable_inv
3. Is_Affordable_Master

All three columns are of the type boolean. Column 1 tells whether or not the unit is in the income restricted housing inventory. 2 tells us whether or not the unit is in the affordable housing dataset. 3 Tells us whether the unit is in either. This will be the most valuable dataset to be used.

Dataset: boston_neighborhood.csv
Usage: This dataset is useful for racial data analysis and contains population information. Loading it into a dataframe would be the best way to use this dataset.

Dataset: boston_voting_district.csv
Usage: Useful for voting analysis, the titles contain information about the column values.

Dataset: BostonAssessorsDataCleaned.csv
Usage: This is a cleaned version of the Boston FY assessors dataset. This dataset is to be used to analyze holistic Boston property data.