Mass Housing Partnership
April 11, 2021
CS 506: Deliverable 3

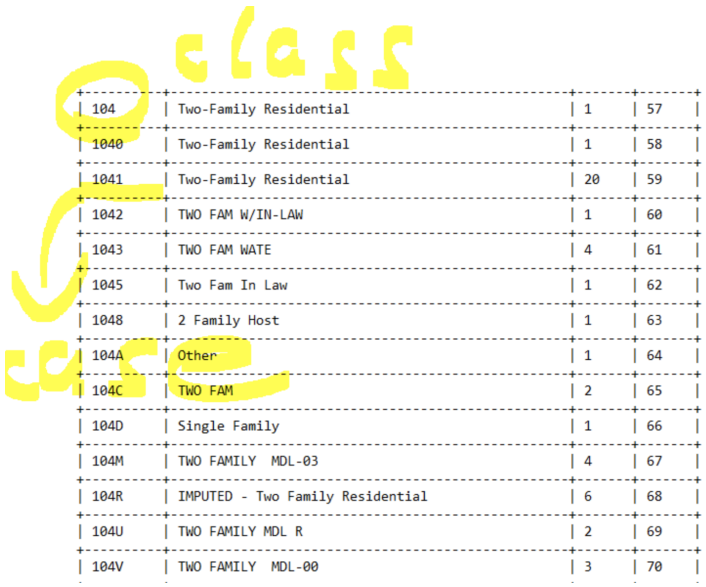
Refine the preliminary analysis of the data performed in PD1&2

We successfully merged the state-wide address data and state-wide parcel by using shapely's spatial join on the geometry of both data sets. This produced a comprehensive and combined dataset by connecting the addresses to the corresponding parcels. We noticed when we performed this step that there were significantly more duplicates than we had expected. After meeting with the client, this was due to commercial and other non-residential buildings being included in both the datasets we were working with. We first scaled down the state-wide data set by just analyzing the Quincy data specifically. There is a feature called "USE_CODE" in the parcel data set that categorizes them by alpha numeric codes. Codes starting with zero, one, and nine are considered residential properties. We then proceeded to filter out the results that were not categorized by these residential Use Codes. This gave us a better idea, and significantly fewer duplicates, of the final data set we needed. However, we realized some of the Use Codes had descriptions like "store" and "office." After meeting with the client again, he explained that there are many mixed-use properties that are commercial units on the first floor, but have residential units above. We then went back into the data processing steps and are utilizing the Use Code descriptions instead of the Styles of the properties. The Styles only describe the specific style of the property and not that it contains both residential and commercial units. We need to account for these anomalies. The address data set does not have a Use Code feature, so we need to filter out the non-residential and develop a process to analyze mixed-use properties before we merge the data sets. An additional challenge we faced is when there were discrepancies regarding the number of units on a property. The number of units reported by the address data and assessor data differed in many occasions, so a process will need to be developed to decide which data set to use for each property. For example, the parcel data reported one unit for an entire apartment building due to tax records, but state-wide address data reported a reasonable number of units. So in this case, we would use the address data.

Attempt to answer overarching project question

The overarching question for this project will provide information on an accurate residential density. The challenges with developing a process for the mixed-use properties affects the accuracy of the residential question. We plan on overcoming this challenge by utilizing the Use Code descriptions to be able to calculate the residential units only in mixed-use properties. There are special cases in the Use Code descriptions that need to be identified to properly filter out commercial units. For example in the image below, you can see that Use Code 104 has different descriptions. In order to find a pattern in these special cases, we will perform an

anomaly detection. This will help us better understand mixed-use properties. Once this is done, we would then merge the data sets and calculate the residential density for each LOC_ID.



104	Two-Family Residential	1	57
1040	Two-Family Residential	1	58
1041	Two-Family Residential	20	59
1042	TWO FAM W/IN-LAW	1	60
1043	TWO FAM WATE	4	61
1045	Two Fam In Law	1	62
1048	2 Family Host	1	63
104A	Other	1	64
104C	TWO FAM	2	65
104D	Single Family	1	66
104M	TWO FAMILY MDL-03	4	67
104R	IMPUTED - Two Family Residential	6	68
104U	TWO FAMILY MDL R	2	69
104V	TWO FAMILY MDL-00	3	70

Draft of Final Report

1. **Project Description**
2. **Data Description**
3. **Data Processing/Download**
4. **Key Question 1: How to merge state-wide address data and state-wide parcel data?**
5. **Key Question 2: What is the residential data of each LOC_ID?**
6. **Key Question 3: What is the process to account for discrepancies between the two datasets for individual properties? (Address and Assessor Data)**
7. **Final Dataset Description**
8. **Limitation and Challenges**
9. **Refined Project Scope**

Depending on the amount of time we have, we may not be able to predict the units for the mixed-use properties as accurately as we want. So far, we have just applied this data processing workflow to the city of Quincy. The city of Boston has significantly more mixed-use properties, which could cause more discrepancies between the data sets and more challenging to apply our process. However, we will still be able to calculate the residential density for each LOC_ID and generate a final data set for the client, regardless of how accurate the calculation will be. We will also be able to document our code, so the client can use it for future use, which is one of the main goals of the project.