State Surplus Land Assessment: Final Report

Taylor Hazlett, Thuy Pham, Mengting Song

I. Introduction & Background

i. Introduction

The goal of this project is to identify Massachusetts' state-owned land most appropriate for low-income housing development for Representative Nika Elugardo's (15th Suffolk District in the Massachusetts House of Representatives) land equity bill.

A variety of metrics were considered in the determination of a land parcel's suitability of affordable housing construction. The following steps were taken to identify state-owned land and perform the analysis:

1. Filter out privately owned land from Massachusetts Land Parcel Database
2. Standardize owner_name column for state-owned land: consistent spellings for all unique agency names
3. Merge dataset with other State Surplus project team to yield the most comprehensive dataset
4. Identify **only** land parcels owned by state-operated housing and transportation-related agencies
5. Determine the number of public transit stations within walking distance (.5 miles) of each land parcel
6. Using Census 2010 data, determine the median household income of the Census tract that each land parcel belongs to
7. Analyze the relationship between income data and transit stops per land parcel, make final conclusions about findings

The rest of this paper follows the course of steps taken to complete the project and explains the tools used and findings at each stage. We conclude with our final results and suggestions as a result of our analysis.

ii. Assumptions and Limitations

The team faced several limitations that either slowed down the progress of analysis in an attempt to resolve or were left as is:

- **Lack of standardization:** the lack of standardization across various features is a limitation that consistently came up in all steps of the project. The inconsistent formatting of column values made the initial dataset difficult to parse. We describe the steps taken to resolve this issue in section **III.i**.
- **Unavailability of relevant datasets:** multiple steps of the project required the use of datasets external to the source database used throughout the project. However, these resources were only found available as information on a website or in a PDF file. The team has time to scrape Mass.gov's website for an A-Z list of state agencies (**IVi**). Additionally, the team attempted to mine a PDF maintaining the maximum quota for affordable housing development within each MA municipality, but the attempt was not fruitful and the team did not use this data in its work (**V**)
- **Missing address values for land parcels:** Approximately 50-60% of land parcels had unusable location information. This made the initial retrieval of census and transportation data unusable. **IV.ii** describes the process used to mitigate this issue.

The team also operated under the following assumption:

- **Local housing authorities:** According to David J. Hedison, Executive Director of Chelmsford Housing Authority, local housing authorities were created by a vote of the local government under Mass General Law Chapter 121. Occasionally these local authorities can receive funding from the state, and any land upon which state funding was used is subject to restrictions from the state. The team assumes that all local housing authority owned parcels in the dataset were not allocated state money and were removed from the dataset.

II. Data Filtering: Identifying All State-Owned Land

Source Database: Massachusetts Land Parcel Database (mapc.org)

i. Poly_typ*:* Only data entries with *poly_typ* equal to '*FEE'* or '*TAX'* were considered. This represented the type of the land parcel, with entries of '*FEE*' or '*TAX*' identifying properties that provide incomes to the state.

ii. Land Use Code: Parcels were also filtered by land-use code (columns: *luc_1, luc_2, luc_adj_1, luc_adj_2*). These are property type classification codes found on all of the associated assessors records, where *luc_1* and *luc_2* columns may include non-standard codes assigned by local staff. In *luc_adj_1* and *luc_adj_2,* all non-standard codes were assigned by MAPC to the best match of standard codes. Only records where any of these columns had values contained in the set {91*,92*,97*} were considered. Land use codes starting with 9* identify all property which is totally exempt from taxation under various provisions of the law, with each subgroup indicating separate entities, and further subgroups indicating branches within those entities. These codes identify state-owned land under the ownership of: Commonwealth of Massachusetts -- Reimbursable Land, Commonwealth of Massachusetts -- Non-Reimbursable Land, and Authorities, respectively.

iii. Building Value: One factor we did not use as a basis for inclusion/exclusion in this analysis but could be significant in determining where to build affordable housing is "building value." Based on criteria from the BU Spark! Team, the project client was both interested in literal building value (i.e. $ amount), as well as whether the land parcel already had existing buildings. The latter was an indication that the land may be more amenable for affordable housing development, as the presence of existing infrastructure indicates access to utilities. The following fields in the land parcel database are used to identify buildings existing on a land parcel (values > 0):

- *Bldg_value* -- sum of assessed building value in $ for all associated assessors records. For condominiums, generally includes land value.

- *Bldg_area* -- sum of assessor's records of finished building area for each parcel (measured in square feet). Assessors' methods vary widely but gross building area may include garages, stairwells, basements, and other uninhabitable areas.
- *Bldg_psf* -- building value per square foot (*bldg_value* divided by the *bldg_area)*.
- *Sqm_bldg* -- estimated land parcel area covered by a building(s) (in square meters).
- *Pct_bldg* -- estimated percent of land parcel area covered by a building(s).

Although some of these fields are redundant, we maintained all of them as we could not be sure which columns were automated in their generation versus manually entered.

After applying all of these filters, we found **1875** state-owned land parcels within the original dataset. When excluding the building value filter, we found **7765** state-owned land parcels within the original dataset.

III. Data Standardization and Cleaning

i. Name Standardization & Identifying Relevant Agencies

One major complication with the original dataset was the lack of standardization throughout features. This issue was most prevalent in the *owner_name* field. The *owner_name* column indicates which agency owns the particular land parcel listed. Agencies that should map to one unique spelling of each name could map to more than one dozen unique spellings. For example, "MBTA" could show up as "MBTA," "M.B.T.A.," "Massachusetts Bay Transportation Authority," "Mass. Bay Trans Authority," etc.

To tackle this problem, the BU Spark! team asked us to use fuzzy string matching. The *fuzzywuzzy* Python package achieves this by calculating the Levenshtein distance between two strings (sequences). Essentially, the Levenshtein distance is the minimum number of single-character edits required to change one sequence into the other. The formulation is as follows:

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

https://en.wikipedia.org/wiki/Levenshtein_distance

where *a,b* are the strings/sequences being compared and *i, j* = |*a*|, |*b*| (the length of strings *a, b*, respectively). The *fuzzywuzzy* package utilizes this distance metric but outputs a score, where a higher score indicates that the two input strings are more similar.

Our overall approach was to first retrieve a master list of agency names via scraping (as described in section **IV.i**). Then for each land parcel, we would use the *fuzzywuzzy* package to calculate scores between the land parcel's *owner_name* and every standard agency name in our list. The standard agency name with the highest score, i.e. the most similar to the *owner_name*, was considered the most appropriate name to match *owner_name* to.

After taking a look at some examples, we recognized that acronyms/abbreviations of complete agency names would have a bad score, indicating that they were not in fact the same. For example, "MBTA" versus "Massachusetts Bay Transportation Authority" would have a significantly low score. With this in mind, we added an additional preprocessing step where we:

- Changed strings to all upper-case letters
- Removed extra white spacing at the front and end of the strings
- Removed special characters & punctuation
- Expanded any abbreviations in *owner_name* to the full-length agency name

For the last task above, we created an intermediate .csv file that contained full-length

standard agency names and their corresponding abbreviations by using the information scraped from the web. Since this information had been stored in a PDF file, we used the package *PyPDF2* to mine the data. *Fuzzywuzzy*'s *token_sort_ratio* method was used to ignore different orderings of words in the two input strings. This seemed to make the most logical sense for the given dataset because one of the worst offenders was all variations of the agency name "Commonwealth of Massachusetts," which would often appear as "Mass Commonwealth," etc. Unfortunately, although the approach appeared sound, we obtained undesirable results. In the table below, you can see some examples:

owner_name	***std_owner_name***
BROCKTON REDEVELOPMENT	DEPARTMENT OF CORRECTION
BROCKTON REDEVELOPMENT AUTHORI	DEPARTMENT OF CORRECTION
CITY OF BOSTON	DEPARTMENT OF CORRECTION
CITY OF BROCKTON	DEPARTMENT OF CORRECTION
COMM OF MASS MET DIST COMM	DEPARTMENT OF CORRECTION
COMM OF MASSDEPT OF ENV MGMT	DEPARTMENT OF CORRECTION
BROCKTON AREA MULTI SERVICES	DEPARTMENT OF FIRE SERVICES
COMM OF MASSHIGHWAY DEPT	DEPARTMENT OF FISH AND GAME

Table 1. Displays original *owner_name* field and its corresponding match in the generated *std_owner_name* column. Results are nonsensical, e.g. "City of Boston" should not be mapped to the "Department of Correction."

After reconvening with the BU Spark! team, we ultimately decided that this had to be done manually to ensure appropriate handling of semantics. One of our team members went through column *owner_name*, alphabetically sorted it, and then collapsed variations of spellings for a particular agency name into a single spelling. Since the client was primarily interested in housing authority- and transportation authority-related agencies, our team then went through the manually standardized *owner_name* column and mapped any housing authority- and transportation authority-related agencies to the corresponding

state-level agencies in our master list. In total, there were **7** state-level housing authorities and transportation authorities prevalent in the dataset; **1253** of the land parcel records had *owner_name* corresponding to these.

ii. Address Fields Standardization

a. Street number

We simply removed all white spaces. Unfortunately, some entries in the *addr_num* column contained non-numeric characters. The client manager did not have a recommendation for how to standardize these, so they were left as-is.

b. Street name

Extra white spaces at the beginning and ends of the string, as well as between words were removed. Special characters were also removed. Alphabetical characters potentially designating unit numbers were left as-is, despite suspicion that certain characters, e.g. "R," might be representing some other type of information. We did not develop a method to standardize or check for correct spelling of street names. The client manager did not have a recommendation for further standardization.

c. Zip code

Another inconsistency prevalent in the dataset was the inaccurate formatting of zip code entries (i.e 2138.0 vs 02138). To resolve this issue, the data type of addr_zip and owner_zip were converted to type string. This removed the occurrences of float values. Finally, entries with length less than 5 were edited to include a '0' at the start of the string, to compensate for the digit lost converting from float to string.

IV. Data Transformation and Augmentation

i. Scrape website for standardize list of agency names & addresses

Standardization of agency names was a vital step towards effective analysis of the dataset. After cleaning agency name spelling, we wanted to map transportation and housing related agencies to their official name on Mass Gov's website. There was no official dataset recording these names, so we had to scrape the website to obtain a list of agency names and their addresses. This was completed using the tool WebScraper.io. Due to inconsistent formatting of the website, only about 80% of addresses were recorded. The other team working on the State Surplus project had working code to recover missing addresses, so this dataset was passed off to them for completion.

ii. Convert coordinate information from one spatial reference to another

The need for (longitude, latitude) coordinate representation of each land parcel arose in the process of trying to augment the dataset with Census and transportation proximity data. In both cases, geographical information had to be provided, e.g. an address is needed to identify the Census tract number associated with a particular land parcel. Due to a lack of standardization in data entry, approximately 50-60% of the records (after filtering for parcel type and land use code) had incorrect, incomplete, or missing address fields (*addr_num, addr_str, addr_zip*). This was about 4400 of approximately 7700 rows, rendering it difficult to obtain meaningful data augmentation results.

Through discussion with the BU Spark! team and an additional resource, we identified that the *parloc_id* in the land parcel database represents the *loc_id* from MassGIS Level 3 Parcel assessor's database tables. This is a unique id assignment by MassGIS based on the (longitude, latitude) coordinates of the parcel center. MassGIS documentation stated that the spatial reference used is the North American Datum 1983 (NAD83). The data is registered to the Massachusetts State Plane Coordinate System, Mainland Zone, where units are in meters (at the federal level) or feet. The target spatial reference system for interpretable (longitude, latitude) coordinates is the World Geographic System (WGS) 84, which is used by the Global Positioning System.

We were able to use an open-source Python package called *pyproj* in order to transform coordinates in one space to the other. This first required identifying a common coordinate system to convert between the two, their equivalent European Petroleum Survey Group (EPSG) codes. The EPSG organization maintains the Geodetic Parameter Dataset, which is a collection of definitions of coordinate reference systems and coordinate transformations which may be global, regional, national or local in application. EPSG 4326 is the equivalent of WGS 84, while EPSG 3586 and EPSG 26986 are the equivalent s for NAD83 in feet and meters, respectively. The resulting implementation successfully augments the dataset with two columns for (longitude, latitude) coordinates for every land parcel record.

iii. Add Census median household income data

One feature that the client specified as important to the classification of whether a land parcel should be chosen for affordable housing development versus selling for funds was the socio-economic status of its neighborhood. The intuition being that communities with lower socio-economic status would welcome and benefit more greatly from affordable housing development. At the direction of the BU Spark! team, this feature was to be numerically represented by Census household income data.

We decided to use the median household income, rather than the mean household income primarily to mitigate the influence of outliers. Census 2010 data was used as it is the most recent at this time. The Census uses a hierarchical designation to group/represent physical land in geographic units; the three levels are:

- Census Blocks
- Census Block Groups
- Census Tracts

where tracts are composed of block groups, and block groups are composed of blocks, the smallest unit. At the direction of the BU Spark! team, we were to use Census tract numbers as a link between the land parcel data and Census 2010 median household

income data. Census tracts are small, relatively permanent statistical subdivisions of a county or equivalent, whose boundaries are specified with the intention of being maintained for long time so that statistical comparisons can be made across censuses.

We utilized the Census geocoder REST API to obtain a census tract number for each land parcel. It supports two different types of searches: by address or by (longitude, latitude) coordinates. When searching by address, the minimum requirements are the street number, street name, and either the zipcode or the town plus two-letter state abbreviation. The first attempt was made using the address fields described in section **III.ii**. The approach was as follows:

1. Check that a record has data for each column (*addr_num, addr_str, addr_zip)*
2. Try to make a request to the API
3. If error or empty response, replace zip code information with town and two-letter state abbreviation
4. Populate tract numbers as new columns in the table with NaN values when information unavailable

As mentioned in section **IV.ii**, only about 40-50% of the records had "complete" address information. However, only half of those (20% of the 7719 rows) had "good" data that was able to get a meaningful response from the Census geocoder API. Some examples of "bad" data include decimals in the street number field, a value of 0 in the street number field, and misspelled street names. As a result, address information was deemed unreliable for this purpose.

In the second iteration, we first obtained (longitude, latitude) coordinates for each land parcel as described in section **IV.ii**, then used this as input to the Census geocoder API. We were able to obtain tract numbers for every record. We then used a different Census API to access the American Community Survey (ACS) 2018 Census data, specifically the ACS 2018 5-year detailed table "B19013_001E." This is the most recent table with median household income information. We pulled the median household income data in

MA grouped by Census tract number. We then were able to augment the dataset with median household income information for every record by relating tract numbers across the tables. In some cases, a tract number was associated with multiple median household income entries. In these cases, we took the average of the values.

The table below provides a breakdown of the number of land parcels grouped by median household income brackets (in increments of $25k):

Median HH income brackets	# land parcels
0-25k	169
25k-50k	3252
50k-75k	749
75-100k	1115
100k+	2390
Grand Total	**7675**

Table 2. Shows the number of land parcels for different brackets of median household income. The total number of records in the table, 7675, is less than the 7723 rows of the complete table due to erroneous values for 48 rows of the data. These rows had negative median household income when matching by tract, possibly indicating that the ACS 2018 database did not have data available for these tracts.

The following table further breaks down this information by providing grouping information by municipality:

Municipality	0-25k	25k-50k	50k-75k	75-100k	100k+	Total
122409	1					1
167206	1					1
169214	1					1
Abington		3			5	8
Acton		27				27
Amesbury				13	4	17

Andover		44			22	66
Arlington		25			20	45
Ashland		101			36	137
Attleboro		1	6		1	8
Avon				5		5
Ayer					3	3
Bedford		25				25
Bellingham		6			4	10
Belmont		71		13		84
Berlin		52				52
Beverly		7		13	15	35
Billerica		13			12	25
Blackstone					26	26
Bolton		5				5
Boxborough		2				2
Boxford		21				21
Braintree		6			2	8
Bridgewater		55	25	5	135	220
Brockton	47		58	45	9	159
Brookline		4				4
Burlington		8			10	18
Cambridge		71	7	41	31	150
Canton		75			40	115
Carlisle		14				14
Carver				5	31	36
Chelmsford		46			25	71

Chelsea			6	38		44
Clinton				5	14	19
Cohasset		17				17
Concord		31				31
Danvers		5		5	30	40
Dedham		28			97	125
Dover		3				3
Dracut				13	17	30
Dunstable		16				16
Duxbury		27				27
East Bridgewater					11	11
Easton		11			166	177
Essex		4				4
Everett			31	7		38
Foxborough		18			42	60
Framingham		21	42	22	40	125
Franklin		42			7	49
Georgetown		49				49
Gloucester		1	20	5	7	33
Groton		56		1		57
Groveland					30	30
Halifax					20	20
Hamilton		4			7	11
Hanover		5				5
Hanson		3			20	23

Harvard		11				11
Haverhill		5	13	29	6	53
Hingham		37				37
Holbrook				6		6
Holliston		19				19
Hopedale		11				11
Hopkinton	1	24				25
Hudson		4		6	7	17
Hull				8	39	47
Ipswich		36		9	31	76
Kingston		24			19	43
Lakeville		3		4		7
Lancaster		34				34
Lawrence	8		6			14
Lexington		6				6
Lincoln		31			2	33
Littleton		24				24
Lowell	88		156	163	3	410
Lynn	19	13	66	12	20	130
Lynnfield		1				1
Malden				3		3
Manchester		16				16
Mansfield		38			13	51
Marblehead		28			4	32
Marlborough				20	32	52
Marshfield	1	2			1	4

Medfield		15				15
Medford		3		9	44	56
Medway		5				5
Melrose		16		13	1	30
Mendon		7				7
Merrimac					4	4
Methuen		7	1	8		16
Middleborough		68		31	55	154
Middleton		44				44
Milford				7	19	26
Millis		1				1
Millville					6	6
Milton		29				29
Nahant					8	8
Natick		219			4	223
Needham	1	89				90
Newbury					52	52
Newburyport		5				5
Newton		152			13	165
Norfolk		38				38
North Attleborough		52		11	8	71
North Reading		6				6
Northborough		79				79
Northbridge				3	55	58
Norton		40			3	43

Norwell		4				4
Norwood		1		11	20	32
Peabody		26	4	20	14	64
Pembroke		12			4	16
Pepperell		18			14	32
Plainville					6	6
Plymouth		80		29	30	139
Plympton					2	2
Quincy		6	23	106	62	197
Randolph				8	9	17
Raynham		16			4	20
Reading		16			11	27
Revere			188	56		244
Rockland				1	5	6
Rockport					14	14
Rowley		49				49
Salem			4	44	42	90
Salisbury				76	28	104
Saugus		21		5	53	79
Scituate		19				19
Sharon		47				47
Sherborn		6				6
Shirley				55	6	61
Somerville		34		26	39	99
Southborough		45				45
Stoneham		3				3

Stoughton		11		16	14	41
Stow		6				6
Sudbury		31				31
Swampscott		2			7	9
Taunton			92	63	402	557
Tewksbury		10			20	30
Topsfield		21				21
Tyngsborough		24				24
Upton		14				14
Uxbridge		31			18	49
Wakefield					4	4
Walpole		63				63
Waltham		11		4	8	23
Watertown		19			29	48
Wayland		21				21
Wellesley		81				81
West Bridgewater		6			8	14
West Newbury		16				16
Westborough		76			78	154
Westford		61				61
Weston		14				14
Westwood		24				24
Weymouth		2		12	23	37
Whitman				4	8	12
Wilmington		35				35

Winthrop			1	1	2	4
Woburn					3	3
Wrentham		5			15	20
(blank)	1					1
Grand Total	**169**	**3252**	**749**	**1115**	**2390**	**7675**

Table 3. Shows the number of land parcels for different brackets of median household income, broken down by municipality. You will notice some of the issues with standardization mentioned throughout the paper. For example, some municipalities have numerical data for entries, rather than town/city names. In addition, there was 1 blank record. The total number of records in this table is also 7675 because of the 48 records that the ACS 2018 Census data lacked median income information for.

iv. Add transportation proximity data

Another relevant feature in the determination of a surplus land parcel for housing development is proximity to public transportation. It is necessary that affordable housing developments are within walking distance to transit stations, as people seeking subsidized housing are less likely to possess a car or more expensive means of transportation. After consulting the team's project managers at BU Spark, 0.5 mile was decided as the maximum threshold for walking distance.

We used the Google Places API to determine the number of stations close to each land parcel. Google Places API offers a *nearbysearch* feature. The feature takes in:

- Latitude & longitude (of central data point)
- Place type
- Search Radius (in meters)

And returns a JSON object containing all places of "place type" within the specified radius of the central data point. Place type was set as: **transit_station.** Transit station includes all bus stations, train stations, and subway stations. Search radius was set to **806 m**, which is equivalent to .5 miles. We created a **numTransitStops** column and updated

the rows with the length of the respective parcel's JSON object, to represent the number of stations within ½ mile.

Similar to the process described in **IV.iii,** this process took two iterations. Nearly half of the dataset was missing addresses for land parcels, and only 20% of that data was usable in the dataset. We used Python's **geocode** library to find location coordinates for each parcel just by using street address, city, state, and zip code. However, this data proved to be inaccurate and still only 20% of the data yielded useful results. The remaining 80% in the dataset were held with a -1 value for **numTransitStops**.

In the second iteration, we obtained location coordinates for each parcel (**see IV.ii**) instead of geocode, and ran the Google Places API using these values. This final iteration was a success, and we were able to find the number of transit stops relative to each parcel in the dataset.

muni	owner_name	longitude	latitude	numTransitStops
Amesbury	COMM OF MASS FLOOD CONTROL	-71.3099	42.6392	60
Cambridge	COMM OF MASS	-71.0337	42.3956	60
Middleborough	COMM OF MASS	-71.3008	42.6441	58
Lowell	MASS DEPT OF PUBLIC WORKS	-71.0828	42.3847	53
Uxbridge	COMM OF MASS	-71.3181	42.6358	50

Table 4. Shows the 5 unique parcels closest to the most transit stations. This indicates that public transportation is not only generally accessible, but also that these parcels are likely near a variety of different transit stations and lines.

muni	numTransitStops
Taunton	4148
Lowell	3090

Middleborough	2506
Revere	2251
Saugus	1601

Table 5. Shows the 5 municipalities with the largest total number of transit stops. This indicates that these towns potentially have many lucrative spots for building affordable housing and that there may be a need for affordable housing in the municipality.

V. Results and Recommendations

Additional data that can make analysis more robust:

- Each town in Massachusetts has a maximum quota for affordable housing development within their own municipality. This data was recorded in a PDF that we attempted to mine with little success. We manually generated a usable .csv file, but ultimately did not move forward with using this dataset. This data could be incredibly useful in determining municipalities in need of more affordable housing, and could help legislators decide where to allocate funds.
- MA income level limits (i.e. threshold for low-income) vary by municipality, as well as household size. The addition of this information can provide more meaningful interpretation of the median household income data. (Average) household size should be available in the Census data and only requires determining which table to pull from. The state has publicly available information about poverty levels grouped by municipality. However, this may only be available in the form of a pdf, which will require another attempt at PDF mining.

The most significant recommendation we have for improvement of the dataset is the requirement of standardization. As previously mentioned throughout the report, the inconsistent labeling of agency names and street addresses can make it difficult to search through the dataset on these parameters and consequently can skew analysis of the data. Guidelines for entering accurate and consistent data will make the dataset easier to maintain moving forward and will yield more meaningful results upon a query or further analysis.

- Recommendations:
    - Use a structured web form for database insertions:
        - Fixed datatypes behind form fields
        - Certain fields should only contain numbers or alphanumeric chars or alpha chars
        - Removal of extra white spaces in fields
        - Drop-down menus to limit spelling variations, especially for agency names
    - State provides clearer guidelines on what code should be used (i.e. 91_, 92_, 97_) in the luc_1, luc_2, luc_adj_1, luc_adj_2 fields in land parcel database:
        - Automatic assignment to land use code
        - Reducing number of codes, if possible

After a closer examination of census data and transportation data, we found results that suggest which state land could potentially be attractive for affordable housing development.

In **IV.iv**, the municipalities of Middleborough and Lowell appeared in **Table 4** and **Table 5**, suggesting that both towns contain (multiple) land parcels in convenient locations for individuals seeking affordable housing. Additionally, **Table 3** of **IV.iii** shows that both Lowell and Middlebourough contain between 10.7% and 22.% of available state parcels within the 0-50k income census tracts, respectively. It should also be noted that these two municipalities also have among the five largest raw numbers of parcels within these census tracts when compared with other municipalities.

When discussing the project and ideal locations for affordable housing, BU Spark! and the team concluded that areas with a lower income demographic are likely to be closer to public transportation, as families and individuals without access to a car might live in an area where they can access other means of travel.

Middleborough and Lowell are two examples of municipalities that exhibited these traits, and could be two areas to further look into when making a decision on state housing construction.

Municipality	0-25k	25k-50k	50k-75k	75-100k	100k+	Total
Lowell	88		156	163	3	410
Middleborough		68		31	55	154

Excerpt from **Table 3**

One final feature to consider in looking for attractive land of development is whether or not there is pre-existing construction on the parcel site. Approximately 24.15% of land parcels in the dataset have buildings on the land. While parcels with construction are more likely to already have utilities on site, the cost of demolition of unusable buildings may offset the money saved from the pre-existing utilities setup. Additionally, there are more empty lots available for construction. Overall, it is important to weigh the pros and cons of each building site before constricting the scope of the search.

While we were able to produce meaningful results to point towards lucrative affordable housing parcels, we did not get a chance to understand the complete flow of data. This can play a role in our recommendations for standardization of data entry. Standardization will likely be a long-term project and require participation across various municipalities/entities. Given more time, the team would like to explore different means of classification and analysis that could be used to recommend land within the dataset for housing development vs selling and to classify land as either category upon insertion into the database.