TACC CRAIGSLIST SCRAPER DOCUMENTATION

VERSION 2

LAST UPDATED: MARCH 5, 2018

INTRODUCTION

One of the challenges that we faced during our research projects about affordable housing was the lack of information that would provide a more comprehensive understanding of the current housing market in the state. In order to make up that challenge, starting in July 2017, The American City Coalition has been working collecting rental housing information from Craigslist.com for the State of Massachusetts. The first months were dedicated to create the initial tool and to test for 60 zip codes in the Boston area. Since October 2017, the tool has been collecting for all the extent of the state.

Versions log:

Version	**Date**	**Purpose/Change**
0	July 2017	Initial scraper and test with 60 zip codes for the City of Boston and the surrounding area.
1	October 2017	Expansion to include all 288 zip codes across the State of Massachusetts.
2	February 2018	Addition to collect more information for each individual listing.

SCRAPER

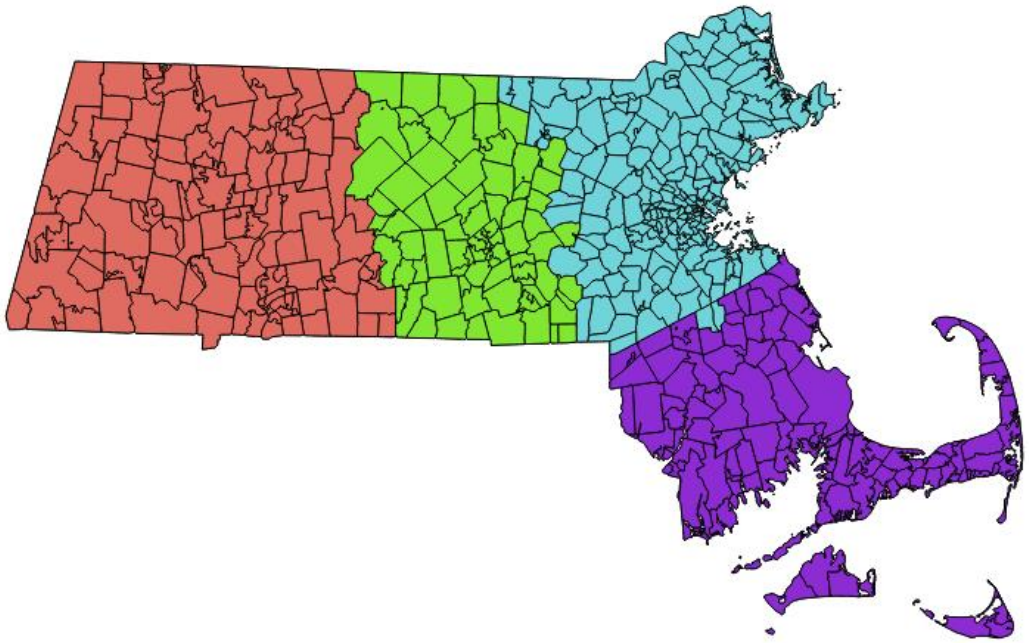
REGIONAL SITES

Craigslist has organized its site to cover different areas of the state with different regional sites, as it can be seen in Figure 1. The four regional sites are[1]:

- boston.craigslist.org
- southcoast.craigslist.org
- westernmass.craigslist.org
- worcester.craigslist.org

[1] The list of all the zip codes can be found in Appendix 1.

Figure 1: Craigslist regional sites



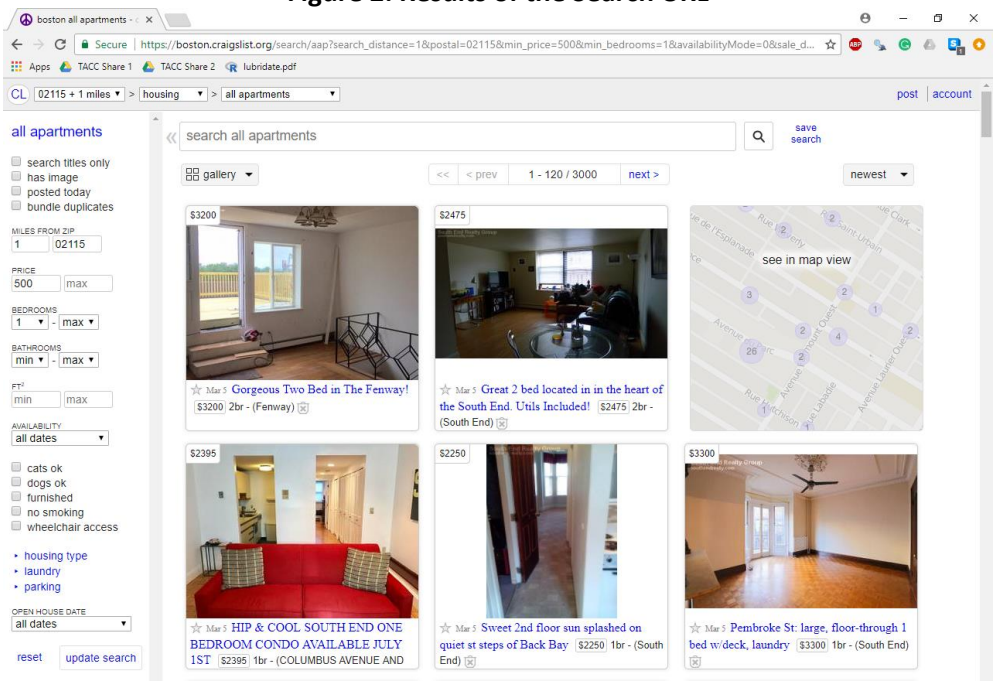
SEARCH PARAMETERS

The search parameters used to create the search URL are:

- Postal = Zip code of interest.
- Distance radius from zip code = 1 mile
- Page number = 0, 120, 240 and 360. Craigslist shows by default 120 listings per landing page. Since the objective of the tool is to collect the largest amount of listings, the first four pages are scraped.
- Availability = 0. This parameter will include all available listings at the time that the program is executed. Minimum Price = 500 US dollars.
- Maximum price = 3333 US dollars. (This parameter was discontinued on February 12, 2018).
- Minimum number of bedrooms = 1. (This parameter was discontinued on February 12, 2018).

Once the search URL is ready, the tool goes online and uses it. Figure 2 shows the results of the URL in the browser. As it can be seen, several listings match the search parameters. Craigslist.com shows up to 3000 listings. However, since listings usually last 30 days and the tool runs on weekly basis, only the first 4 pages are scraped, that can be translated in up to 480 listings.

Figure 2: Results of the Search URL



DATA COLLECTED

From the landing page of the URL seen in Figure 2, the following general data for each listing is collected:

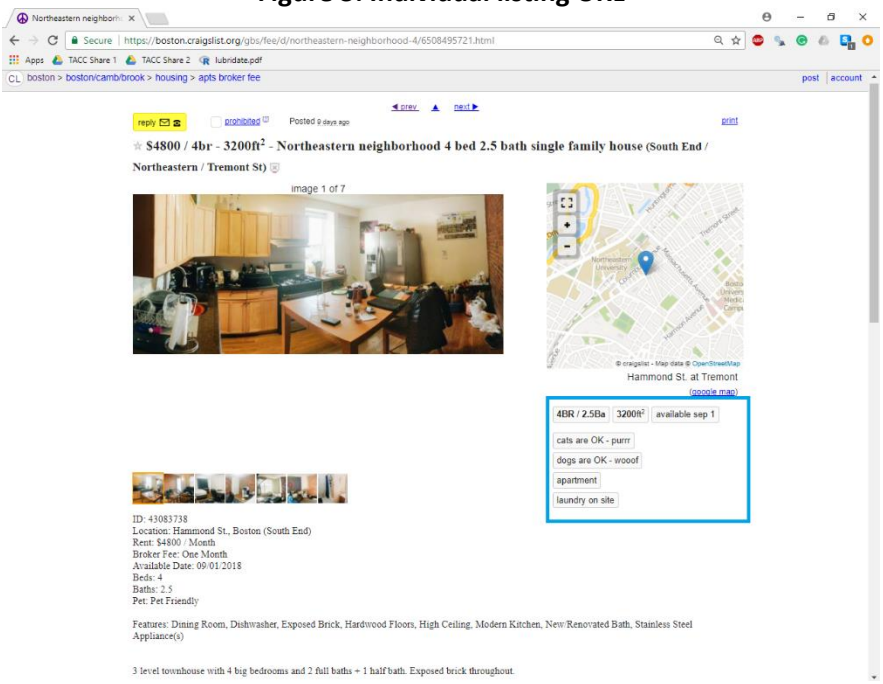
- **Title:** Given by the poster of the listing.
- **Date:** Posting date on Craigslist
- **Price:** Price for the listing in US$.
- **Bedrooms:** Number of bedrooms in the unit
- **SqFt:** Number of bedrooms in the unit
- **Location:** Given by the poster of the listing. This is an open field when posting in Craigslist.com, so this location is not necessarily the true location. In addition, it can be a City or Neighborhood name.
- **Latitude and Longitude:** If the poster of the listing assigns a specific address, Craigslist geocodes it and assigns a latitude and longitude.
- **Listing URL:** URL for each individual listing.

In order to collect more information, the tool also scrapes information from each individual listing using the Listing URL collected from the landing page. Figure 3 shows an individual listing: it has the same information already collected, but also the body of the listing, the pictures that the poster adds to the listing and the information in the blue box that provides deeper insights about the renter conditions such as if the unit has a washer and dryer or if pets are allowed.

New general data for collected:

- **Body:** Description of each listing that each poster writes. This field is open to as many characters and details the poster thinks that will help him renting the place.

Figure 3: Individual listing URL



Given the nature of the information in the blue box, dummy variables were created to capture the information from each of the information boxes. For each one of the variables, 1 means yes, 0 means no.

Type of housing variable:

- **Apartment**
- **Condo**
- **Cottage**
- **Duplex**
- **Flat**
- **House**
- **In Law**
- **Loft**
- **Townhouse**
- **Manufactured**
- **Assisted Living**
- **Land**

Information about laundry:

- **No Laundry**
- **In-Unit Laundry**
- **Laundry Hookup**
- **In-Building Laundry**
- **Laundry-On-Site**

Parking:

- **Carport**
- **Attached Garage**
- **Detached Garage**
- **Off-Street Parking**
- **On-Street Parking**
- **Valet**
- **No Parking**

Another general:

- **No Smoking**
- **Furnished**
- **Wheelchair Accessible**
- **Cat**
- **Dog**

DATA MANAGEMENT

Once all the information has been collected and stored, several additional steps are taken before using the data for the different purposes that we have.

REMOVING DUPLICATES AND SHARED APARTMENT LISTINGS

Craigslist has an anti-spam policy: publishers are not supposed to publish several times the same add. However, reality shows that that policy is not very efficiently adopted, which leads to having multiple listings for the same unit. In addition, each listing has a Unique Listing ID that is valid while the listing is valid (has not expired). If the add expires and is reposted, it will receive a new Unique Listing ID.

The first challenge was to start analyzing the data to identify those duplicate listings to have a better representation of the housing market. For that reason, the scraper has two instances where duplicated are removed:

- Listings by Unique Listing ID: if two or more listings have the same Unique Listing ID, only one will remain in the data set.
- Tile + bedrooms + price: if two or more listings share the exact same title, the exact same number of bedrooms and the same rent price, then the listing is classified as duplicate and only one listing will remain in the data set.

Since the main objective of this infrastructure is to analyze the housing market for renters, all those units that refer to shared apartments or roommate requests were filtered out if their titles matched any of these patterns (case insensitive):

- Beginning with Room
- Beginning with Shared
- Containing roommate anywhere
- Containing private room anywhere
- Containing room for rent anywhere

- Containing sub lease anywhere

However, these two instances are not exhaustive in the task of removing duplicates. Further research in new methodologies are required in order to keep reducing the duplicate listings.

CENSUS TRACT ID

Most of the information available from different sources (Children Opportunity Index, ACS data, etc.) are published at Census Tract level[2]. Hence, for us to be able to link each listing to other information, it was necessary to assign the Census Tract ID to each listing. For that purpose, we used the Federal Communications Commission API, a tool available to the public and free.

In order to assign the Census Tract ID to each listing, the tool requires the Latitude and Longitude for each listing. Once the query has completed, the result is the Census Tract ID, only for those listings that have valid Latitude/Longitude pair.

REVERSE GEOCODING

Even though Craigslist.com has the option for the add poster to insert the full address of the listing to show it on the map, the address is not published; rather it is transformed into Latitude and Longitude. In order to link the data to other data sources for the analysis (e.g. Payment Standards at zip code level, American Community Surveys at Census Tract level), we needed to have the corresponding full address for each listing. The process for obtaining a full address from the Latitude and Longitude is called Reverse Geocoding. Using Google Maps API, we were able to obtain the full address for each listing, only for those listings that have valid Latitude/Longitude pair. This process allows us to identify true location of each listing and compare it with the Location given by Craigslist.com.

METHODOLOGY

The entire infrastructure was built using the software R and the following packages: ggmap, placement, rvest.

[2] The Census Tract is a geographical unit established by the Bureau of Census.

APPENDIX 1

Craigslist Regional Site	Zip codes
Boston Area	01432, 01434, 01450, 01460, 01463, 01464, 01469, 01474, 01701, 01702, 01718, 01719, 01720, 01721, 01730, 01731, 01741, 01742, 01746, 01748, 01749, 01752, 01754, 01760, 01770, 01773, 01775, 01776, 01778, 01801, 01803, 01810, 01821, 01824, 01826, 01827, 01830, 01832, 01833, 01834, 01835, 01840, 01841, 01843, 01844, 01845, 01850, 01851, 01852, 01854, 01860, 01862, 01863, 01864, 01867, 01876, 01879, 01880, 01886, 01887, 01890, 01902, 01904, 01905, 01906, 01907, 01908, 01913, 01915, 01921, 01922, 01923, 01929, 01930, 01938, 01940, 01944, 01945, 01949, 01950, 01951, 01952, 01960, 01966, 01969, 01970, 01982, 01983, 01984, 01985, 02019, 02021, 02025, 02026, 02030, 02032, 02035, 02038, 02043, 02045, 02052, 02053, 02054, 02056, 02062, 02067, 02071, 02072, 02081, 02090, 02093, 02108, 02109, 02110, 02111, 02113, 02114, 02115, 02116, 02118, 02119, 02120, 02121, 02122, 02124, 02125, 02126, 02127, 02128, 02129, 02130, 02131, 02132, 02134, 02135, 02136, 02138, 02139, 02140, 02141, 02142, 02143, 02144, 02145, 02148, 02149, 02150, 02151, 02152, 02155, 02169, 02170, 02171, 02176, 02180, 02184, 02186, 02188, 02189, 02190, 02191, 02210, 02215, 02301, 02302, 02322, 02343, 02368, 02420, 02421, 02445, 02446, 02451, 02452, 02453, 02458, 02459, 02460, 02461, 02462, 02464, 02465, 02466, 02467, 02468, 02472, 02474, 02476, 02478, 02481, 02482, 02492, 02493, 02494, 02762
Southcoast area	02048, 02050, 02061, 02066, 02301, 02302, 02324, 02330, 02332, 02333, 02338, 02339, 02341, 02346, 02347, 02351, 02356, 02359, 02360, 02364, 02367, 02370, 02375, 02379, 02382, 02532, 02534, 02535, 02536, 02537, 02538, 02539, 02540, 02542, 02543, 02553, 02554, 02556, 02558, 02559, 02561, 02562, 02563, 02564, 02568, 02571, 02576, 02601, 02630, 02631, 02632, 02633, 02635, 02637, 02638, 02639, 02641, 02642, 02644, 02645, 02646, 02647, 02648, 02649, 02650, 02652, 02653, 02655, 02657, 02659, 02660, 02663, 02664, 02666, 02667, 02668, 02670, 02671, 02672, 02673, 02675, 02702, 02703, 02713, 02713, 02715, 02717, 02718, 02719, 02720, 02721, 02723, 02724, 02725, 02726, 02738, 02739, 02740, 02743, 02744, 02745, 02746, 02747, 02748, 02760, 02763, 02764, 02766, 02767, 02769, 02770, 02771, 02777, 02779, 02780, 02790, 02791
Western Massachusetts area	01001, 01002, 01003, 01007, 01008, 01009, 01010, 01011, 01012, 01013, 01014, 01020, 01021, 01022, 01026, 01027, 01028, 01029, 01030, 01032, 01033, 01034, 01035, 01036, 01038, 01039, 01040, 01041, 01050, 01053, 01054, 01056, 01057, 01060, 01062, 01063, 01069, 01070, 01071, 01072, 01073, 01075, 01077, 01080, 01081, 01082, 01083, 01084, 01085, 01088, 01089, 01092, 01095, 01096, 01098, 01103, 01104, 01105, 01106, 01107, 01108, 01109, 01118, 01119, 01128, 01129, 01144, 01151, 01152, 01201, 01203, 01220, 01222, 01223, 01224, 01225, 01226, 01230, 01235, 01236, 01237, 01238, 01240, 01242, 01243, 01244, 01245, 01247, 01252, 01253, 01254, 01255, 01256, 01257, 01258, 01259, 01260, 01262, 01263, 01264, 01266, 01267, 01270, 01301, 01302, 01330, 01337, 01338, 01339, 01340, 01341, 01342, 01343, 01344, 01346, 01347, 01349, 01350, 01351, 01354, 01355, 01360, 01364, 01367, 01370, 01373, 01375, 01376, 01378, 01379, 01380, 01521

Worcester Area	01005, 01031, 01037, 01068, 01097, 01331, 01366, 01368, 01420, 01430, 01431, 01436, 01440, 01451, 01452, 01453, 01462, 01468, 01473, 01475, 01501, 01503, 01504, 01505, 01506, 01507, 01510, 01515, 01516, 01518, 01519, 01520, 01522, 01523, 01524, 01527, 01529, 01531, 01532, 01534, 01535, 01536, 01537, 01540, 01541, 01542, 01543, 01545, 01550, 01560, 01562, 01564, 01566, 01568, 01569, 01570, 01571, 01581, 01583, 01585, 01588, 01590, 01602, 01603, 01604, 01605, 01606, 01607, 01608, 01609, 01610, 01611, 01612, 01655, 01740, 01745, 01747, 01756, 01757, 01772