Boston Area Research Initiative ● Northeastern & Harvard Universities
E-Mail: bari@northeastern.edu ● Web: www.bostonarearesearchinitiative.net

A guide to Craigslist data for the State of Massachusetts

Overview

This dataset contains information about 285,233 listings in Massachusetts that were listed on craigslist.com between February 13th, 2020 and December 31st, 2022. To collect data on property listings across the state, we ran a Python script once a day that scraped posts made to the "housing" sections of five Craigslist URLs: boston.craigslist.org; capecod.craigslist.org; southcoast.craigslist.org; westernmass.craigslist.org; and worcester.craigslist.org. Using the location data and timestamps embedded in listings, we have aggregated the data extracted from posts to generate neighborhood-level metrics about Craigslist activity.

- The main dataset (*CRAIGSLIST.Listings.csv*) is the base file derived from Craigslist property listings.
- *CRAIGSLIST.CT.csv* is a file containing aggregate measures (i.e. ecometrics) that describe neighborhoods (approximated with census tracts). It includes cross- sectional measures calculated by aggregating information from listings across the entire period of data collection and longitudinal measures calculated by aggregating information from listings in each month. These variables are provided in a spreadsheet format (*.csv*) and in mappable shapefiles (*.shp*).

This documentation contains a section for each of these groupings, describing the contents and variables.

Table of Contents

1. Listings 3
1.1. Summary of Contents 3
1.2. Description of Variables 3
1.2.1. Listings Characteristics 3
1.2.2. Geographical Information 4
2. Neighborhood Metrics 5
2.1. Summary of Contents 5
2.2. Description of Variables 5
2.2.1. Cross-Sectional Metrics 5
2.2.2. Longitudinal Metrics 5
2.2.3. Geographical Information 6

1. Listings

1.1. Summary of Contents

The listings dataset includes information on all unique postings that were scraped during the data collection period[1]. We processed the contents of 285,233 listings to extract 14 variables that provide information about the property listing and meta-information about the post.

1.2. Description of Variables

The Craigslist Listings data set includes variables that can be grouped into two categories: listing characteristics and geographical information. Listing characteristics include the basic characteristics and attributes of the property as drawn from the listing on Craigslist. Geographical information provides further detail on the location of the listed property drawn from semantic and geographic information embedded in listings.

1.2.1. Listings Characteristics

- *LISTING_ID* is a unique identifier for each listing that is generated from the listing's URL address.
- *LISTING_YEAR* indicates the year the listing was posted online
- *LISTING_MONTH* indicates the month the listing was posted online
- *LISTING_DATE* indicates the day of month the listing was posted online
- *LISTING_TIME* indicates the time the listing was posted online
- *RETRIEVED_ON* indicates the date and time the BARI scraper collected the listing information
- *BODY* is a string variable that includes the text from the main body section of the listing.
- *PRICE* is the price of the listed property in dollars.
- *AREA_SQFT* is the area of the listed property in square feet.

[1] To prevent duplicates from being created if a post is online for more than one day, we verify that the post's unique ID is not already present in our database prior to data ingestion. Code and explanation of scraping process can be found at: https://github.com/BARIBoston/craigslist2

- *ALLOWS_CATS* is a binary variable where "1" indicates that the listed property allows cats, and "0" indicates the property does not allow cats.
- *ALLOWS_DOGS* is a binary variable where "1" indicates that the listed property allows dogs, and "0" indicates the property does not allow dogs.

1.2.2. Geographical Information

- *ADDRESS* is the address posted for the listing.
- *LOCATION* is a string variable containing the location posted with the listing.
- *CT_ID_10* is the 2010 Census Tract ID number.
    - This is found by extract longitude and latitude from listed locations and spatially overlaying them onto BARI's Census Tracts shapefile.

 Note:
We have data lapses for the year 2021 and the data from July, August, October, and November in 2021 are compromised.

2. Neighborhood Metrics

2.1. Summary of Contents

A neighborhood-level dataset was created that describes aggregate features of neighborhood property listings. Neighborhoods are approximated with census tracts. Neighborhood-level measures are provided in both standard format (*.csv*) and as mappable shapefiles (*.shp*).

2.2. Description of Variables

Neighborhood-level variables are split into three categories: cross-sectional metrics, longitudinal metrics, and geographic information. Cross-sectional metrics have been generated by calculating statistics across all listings made during the data collection period. Longitudinal metrics have been generated by calculating statistics for the listings of a given month, with variables suffixes (MMM) indicating the month of measurement. Geographical information provides further detail on the location of the neighborhood.

2.2.1. Cross-Sectional Metrics

- *ListingsFreq_Cross* is the number of property listings in the neighborhood posted on Craigslist during the data collection period.
- *MedianRent_Cross* is the median rental price of property listings in the neighborhood posted on Craigslist during the data collection period.

2.2.2. Longitudinal Metrics

- *ListingsFreq_(MMM)* is the number of property listings in the neighborhood posted on Craigslist in the measured month.
    - *Note:* The February value of the year of 2020 is based only on listings from February 13th to February 29th .
- *MedianRent_(MMM)* is the median rental price of property listings in the neighborhood posted on Craigslist in the measured month
    - *Note:* Values are missing for tracts that had zero listings posted in the measured month.

2.2.3. Geographical Information

- *CT_ID_10* is the 2010 Census Tract ID number.
- *TOWN* is the town or city where the tract is located.
- *COUNTY* is the country where the tract is located.

 Note:
We have data lapses for the year 2021 and the data from July, August, October, and November in 2021 are compromised. Therefore, the neighborhood measurements do not include metrics from these months.