**Community Profiles**

**A methodology for producing automated, accessible information to drive better local conversations about housing production**

**Tom Hopper**

**DA5020**

**Summer 2017**

**Introduction**

There are 351 municipalities in the Commonwealth of Massachusetts. Each of these municipalities has its own set of zoning bylaws, which govern what can be built and where within the community. While this allows an individual community to control its own destiny, it makes it very difficult to plan for regional and state housing needs. Furthermore, a large proportion of municipalities adopt zoning that purposefully restricts the development of new housing units, and in particular, multifamily housing units. There are even indications that the lack of supply of housing across the Boston metro region might have an adverse impact on the area’s ability to maintain a strong economy.1

In addition to potential impact on the regional economy, local regulation has had an impact on the geography of demographics, as restricted supply and lack of multifamily and rental housing have kept certain populations out of certain communities.2 This includes non-white racial groups as well as millennials. As a housing policy practitioner and an active community member, I can attest to the fact that discussions around development at the local level are rarely so blatantly exclusionary. Opponents of new housing development often cite concerns such as increased school costs from new residents, or increased traffic and tighter parking. Other times the discussion centers on maintaining the “character of the neighborhood,” which is often a thinly veiled allusion to keeping outsiders (or people who are different) out of a community. Whatever the anti-housing narrative, local debate often takes place without an abundance of facts and information. Community members should know how the demographics of their town are impacted by the housing they choose to allow.

With readily accessible data in an easily interpretable format, proponents of new housing development might be able to educate community members about what the demographics and housing situation really are in the town. With better information, a community can have more fruitful conversations, which might help drive better local outcomes around housing production. There are countless kinds of information that would be useful for a community to access when considering new housing development. This project aims to create a starting point for future development of a more expansive and thorough resource. This project will use data collection, storage, and retrieval techniques to create a module that generates a one-page dashboard of charts and tables related to the existing demographics and housing stock of a user-chosen community. In order to demonstrate how readily available the data can be made through the chosen data techniques, the module will be developed so that a user is able to simply input the name of any town in the state and generate the one-page fact sheet.

1. Oddleifson, C. and Ziegler, C. “Moment of reckoning for housing and state economy.” Commonwealth Magazine. March 30, 2016. <https://commonwealthmagazine.org/economy/moment-of-reckoning-for-housing-and-state-economy/>
2. Florida, R. “The Segregation that Zoning Inflicts on Cities.” The Atlantic. January 5, 2016. <https://www.theatlantic.com/business/archive/2016/01/zoning-land-use-segregation/422595/>

**Data sources**

The single most comprehensive resource for community-level information on demographics and housing stock is the U.S. Census Bureau’s American Community Survey (ACS) dataset. This is the primary data source for the project. The ACS collects information from a sample of citizens on an annual basis on a wide variety of topics, including demographics and housing. Although information is collected annually, sample size prevents the Census Bureau to release one-year data on small geographic scales for privacy reasons. So, they aggregate information across multiple years in order to provide a suitable sample size for smaller geographic areas. In order to obtain information for every one of the state’s 351 communities, I am using the 5-year ACS product for the years 2011 through 2015. The following tables are being accessed for this project:

*B25077 Median Value (Dollars)*

This data was not actually get used in the final product, since I am also using market information from trulia.com (as detailed below). However, it might provide a useful benchmark to measure against the results obtained from Trulia in the future. Median price is being included to provide a sense of how expensive homes are in the community.

*B25004 Vacancy Status*

Vacancy rate will be calculated from the information in this table. A vacancy rate will provide a proxy for how tight the market is. A high vacancy rate might indicate that there are not huge market pressures on housing in the town. A low vacancy rate might indicate that there is insufficient housing supply to meet housing demand. Vacancy rate might also be influenced by the volume of vacation homes in a town.

*B25070 Gross Rent as a Percentage of Household Income in the Past 12 Months*

When a household spend a large proportion of its income on housing expenses, there are less funds available for other necessities. One industry standard for weighing housing affordability is that households that are paying over 30% of income on housing expenses are “housing cost burdened,” while those paying more than 50% of income on household expenses are considered “extremely cost burdened.”

*B25095 Household Income by Selected Monthly Owner Costs as a Percentage of Household Income in the Past 12 Months*

Same as above, but for homeowners instead of renters. Pulling this data separately will allow users to see any differences in proportions of rent burdened populations for renters versus home owners.

*B25008 Total Population in Occupied Housing Units by Tenure*

This table will provide a break down of renter households versus owner households in the community. This will provide insight into the relative stock of each kind of tenured housing.

*B02001 Race*

For reasons of social justice, it is important to understand to what extent there is racial diversity in a community. Years of red-lining and housing discrimination, combined with exclusionary housing practices and restrictive local zoning have led to spatial segregation by race. Comparing a municipality’s diversity (or lack thereof) might be eye-opening for some community members

*B01001 Sex by Age*

A simpler table for just age could not be found, so this table it being downloaded and reformatted to aggregate age demographics for the entire town population. The data will be organized by cohort and shown in a histogram that will be compared to the distribution for the entire state. This comparison will allow a community to see what age groups are over- or under-represented relative to the state population.

*B25024 Units in Structure*

Many local restrictions do not allow much, if any, multi-family housing. By providing detailed information on how many housing units are in each kind of building, it might shed some light on how many communities have almost completely relied on detached single-family homes.

The other data source I will be using is market information scraped from trulia.com. Trulia provides a ‘Real Estate Overview’ report on the town level. There are three useful pieces of information right at the top of this report:

* + - * Median sales price
      * Price per square foot
      * Median rent

These pieces of information will provide a rough estimate of the current market prices within the town, and might place some of the census information that is being collected into a little bit of context. For example, if a high proportion of renters are cost-burdened, then does the median rent provide us a little bit of insight into what kind/price of rental units exist in the community. Places with high rents with high levels of cost-burdened renters might indicate that although cost-burdened, these renters are higher income. A community with high renter cost burden but low rents might indicate that many renters are lower-income. Median home prices might offer some insight into the demographics of the community. For example, a town with very high home prices might not be able to attract younger adults/millennials because there might not be many starter homes in the town’s inventory.

**Data techniques and platform choices**

I have decided to split my code into two separate RMarkdown files. The first document collects the census data through an API-based package, reformats and tidies the collected data, and then stores the resulting data frames into database tables using SQLite. The second document retrieves and displays information dynamically, based on a town name parameter defined by the user, and uses an R-to-html package to deliver a one-page fact sheet.

Phase One - Collect, Tidy, Store

Census information is retrieved through use of the ‘acs’ package in R. After registering for a code through the U.S. Census Bureau’s API, the package allows you to specify the year, geography, and table number of the data you would like to access. Once specified, the functions return the requested information in an ‘acs-class’ object, which is a set of arrays. The array format makes it a little more difficult to gather, tidy and select the desired data fields, but since the architecture of the acs-class object is virtually the same across all tables, the same data processing methodology can be used for every table mentioned above. A function was briefly considered for this, but because the desired fields varied between the tables, with slightly different naming conventions, a function seemed like it might sacrifice data integrity for programming efficiency.

Once each data frame was created from the census tables, I write the table to a SQLite database. The tables have been designed to link on a common GEO\_ID field that I have included in every table. There is a reference table as well, with one record for every GEO\_ID and Community name. This will be the primary table that I use later on when writing joins into my SQL statements.

I have annotated my RMarkdown document, which provides a little more step-by-step guidance and insight into the coding choices that were made. when formatting. A couple key points:

- While the architecture of the database is simple, the formatting of the raw census information into a clean table was fairly involved.

- As the primary key for the Community table, the GEO\_ID field is the most crucial field in the database, as it connects a community’s record to all of the tables of information. Additionally, the formatting of the GEO\_ID is crucial, because it needs to match up with the GEO\_ID that is used by the Census Bureau’s mapping layers. In the downloaded tables, codes for state, county and county subdivision must be aggregated in order to properly join with these shapefiles, which use a single composite ID field.

- I am not using SQLite to store information from Trulia. Census information is updated once a year, and this tool might be used frequently between updates. Therefore, it makes sense to store this information. Trulia’s market information get calculated based on the properties on the market at any given time. In order to make sure the end report has the freshest information, this information will be retrieved on the fly as the report is generated.

Phase 2 - Retrieve and present data

After some research, I chose to use flexdashboard as the vehicle to display the data I wish to present. My goal was to have a user type the name of a community in Massachusetts, run the program, and produce a dashboard of charts and tables specific to that community, with some comparisons to statewide data. I like that flexdashboard allows me to use html widgets and some css styling for the components. Flexdashboard allows for multiple kinds of presentation techniques, including standard ggplots, stylized html tables, and even interactive content like clickable and scrollable maps. Flexdashboard is available for download through CRAN.

Much of the information for the report is being retrieved from the SQLite database that stores the census information. SQL statements are being used, but they are parameterized by the name of the community that has been inputted at the top of the document by the user. Once information is pulled from the SQLite database, it is formatted for easy reporting, either through a table or a chart.

The second retrieval method is web scraping using the rvest and XML packages. First, the url that is being used to access the webpage is parameterized based on the town designated by the user, and based on the url convention of the Trulia website. After using ‘Inspect’ in the browser to identify the node and css wrapper that the desired information is in, the XPath was copied and inserted into an xPathApply argument on a parsed html document, which extracts the three data points: median price, price per square foot, and median rent. Of course, the accessed data is not in a proper numeric format, so some tidying has to occur to make the data workable. Once tidied, a table is compiled using these fields, stylized using css statements.

The third retrieval technique is through the census bureau, again, but this time I am downloading a shape file that I can use to connect the data from my database and combine them both on an interactive html map. This download is made easy through the ‘tigris’ package. Once the shape file is downloaded, it is joined to a table generated on vacancy data through a SQL statement. In order to display the joined information, the ‘leaflet’ package is used, which provides a basemap and a platform for displaying geographic data.

The RMarkdown for a flex dashboard document knits to an html dashboard output. Flexdashboard resizes content based on the column size definitions provided by the user. The html-ready format allows interactive content to shine. The map included in the report is clickable and zoomable, allowing a user to explore how the community they chose compares to others in relation to vacancy rate.

**Issues and troubleshooting**

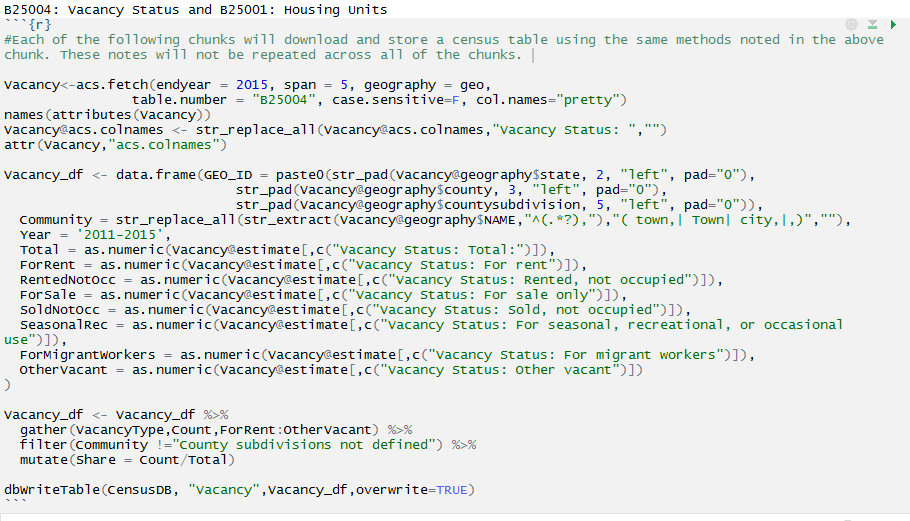
One of the more complex parts of the data retrieval was working with the array-style format of the download census tables. I was able to find some guidance through several threads on stack overflow to better understand how to use the proper syntax to obtain the data I needed. Once I learned how to use the @ and $ figures to drill down into the second level of arrays, the problem was resolved. Luckily, the process was very similar across all of the census tables, so the pain was only experienced once.

Another issue I had was in obtaining market data through web scraping. I had initially aimed to get information by scraping multiple pages of Zillow search results. However, the parsed html document I obtained through rvest did not match the structure I saw when using ‘Inspect’ in my browser. Instead the nodes that were parsed included captcha code and seemed to be a roadblock to obtaining the site data. After spending way too much time trying to figure this out, I decided to see what other sites might have to offer. The neat and accessible format of Trulia’s Real Estate Overview page seemed to be a much simpler way to access information that is just as useful. Zillow was abandoned as a data source.

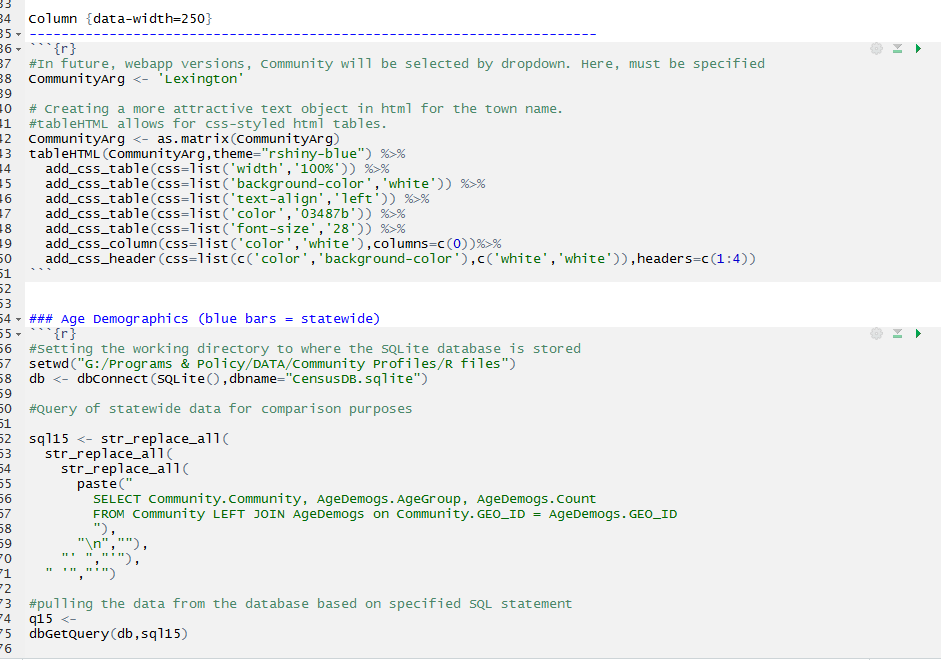
One issue that remains unresolved has to do with the tigris package download. I developed the two RMarkdown documents over time on two different machines: a PC and a Mac. While I had no issues running the code for the download of the tigris shapefile on the PC (both knitting and running the chunk), the Mac will not allow me to access the URL and download/unzip the shapefile when knitting. Oddly, the code runs fine on the Mac when just running the chunk. I am not sure if this has to do with my Mac’s browser settings or not. This could be an issue in trying to unroll this program to other users (or instructors).

**Input and output results from an example session of your R code and insightful screenshots of the project running.**

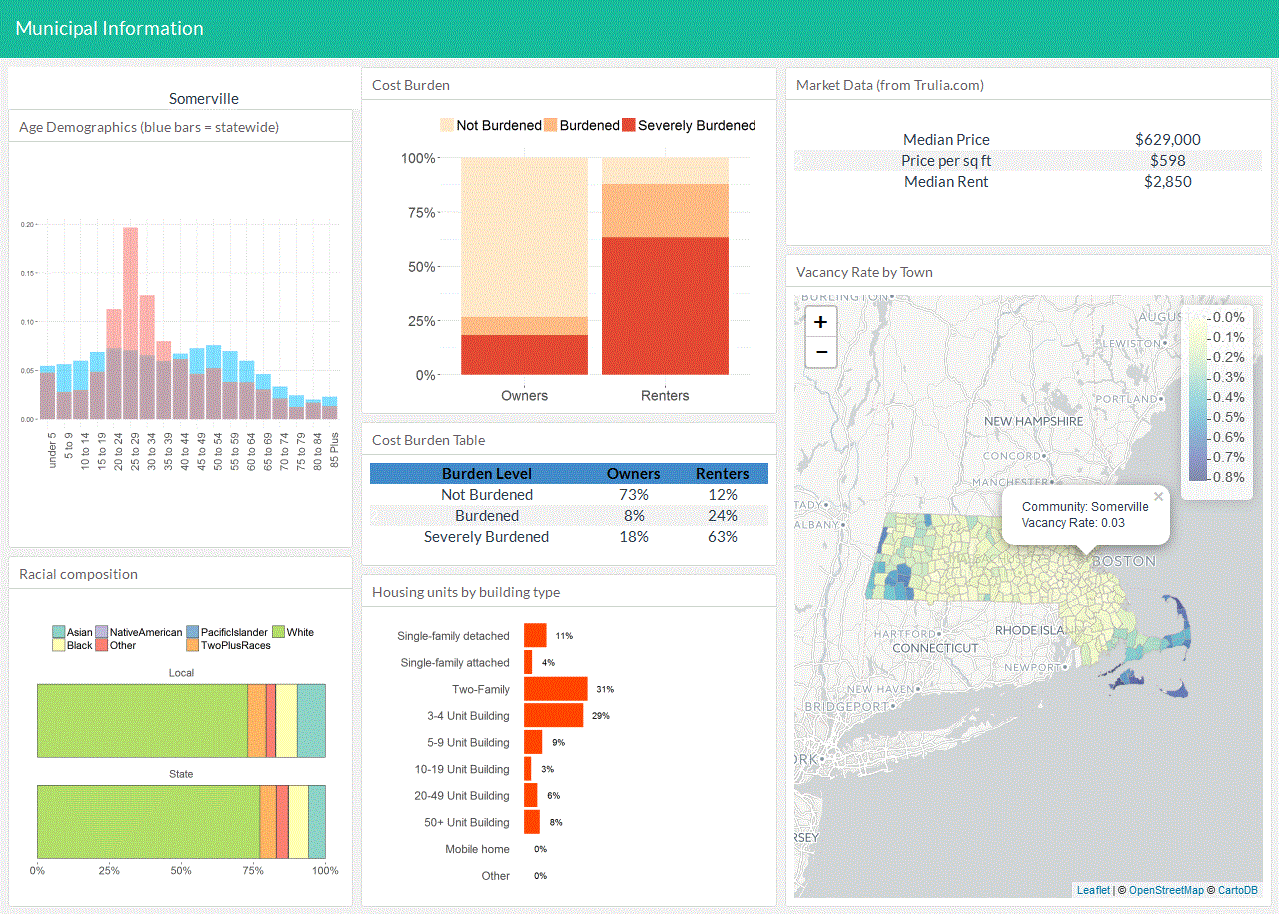
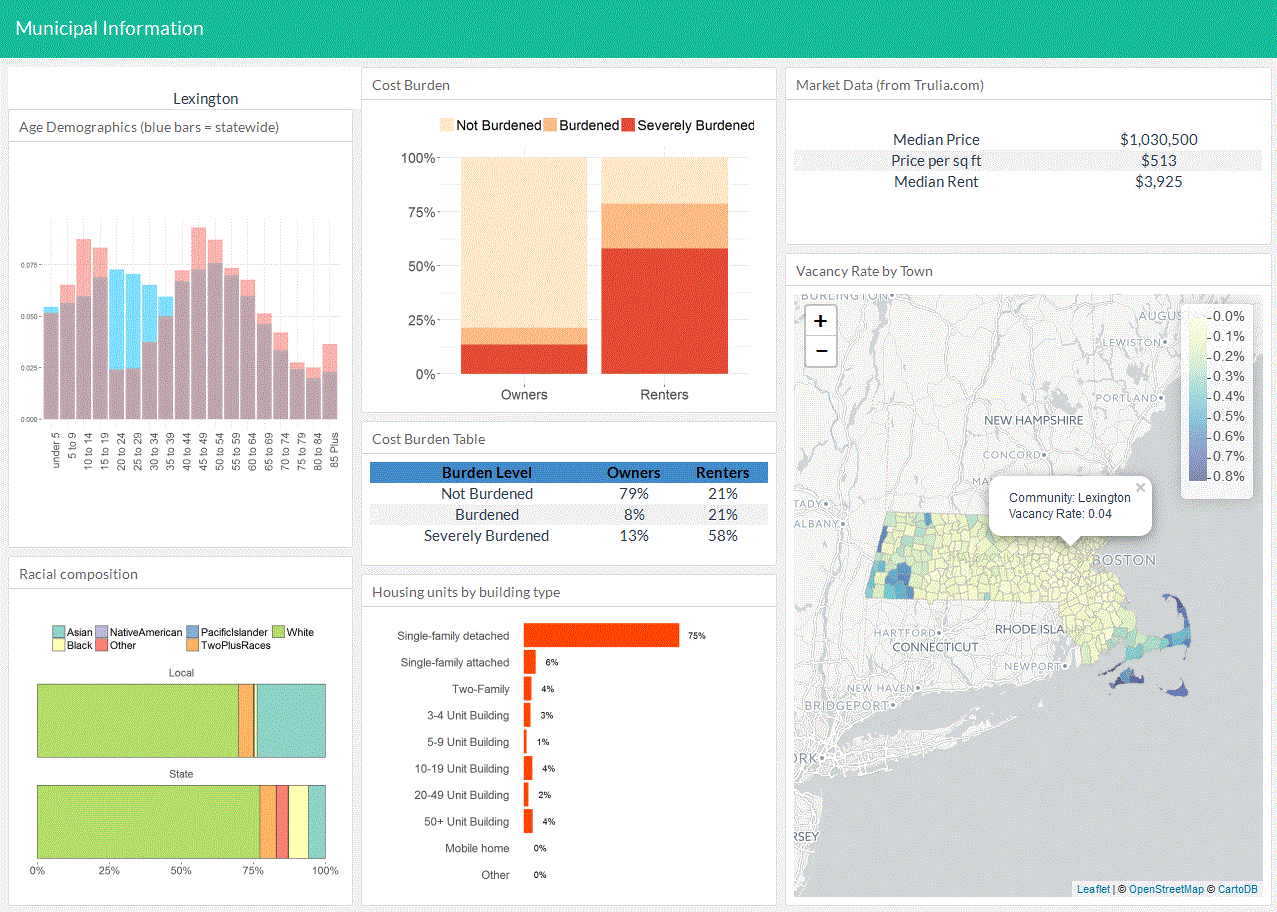
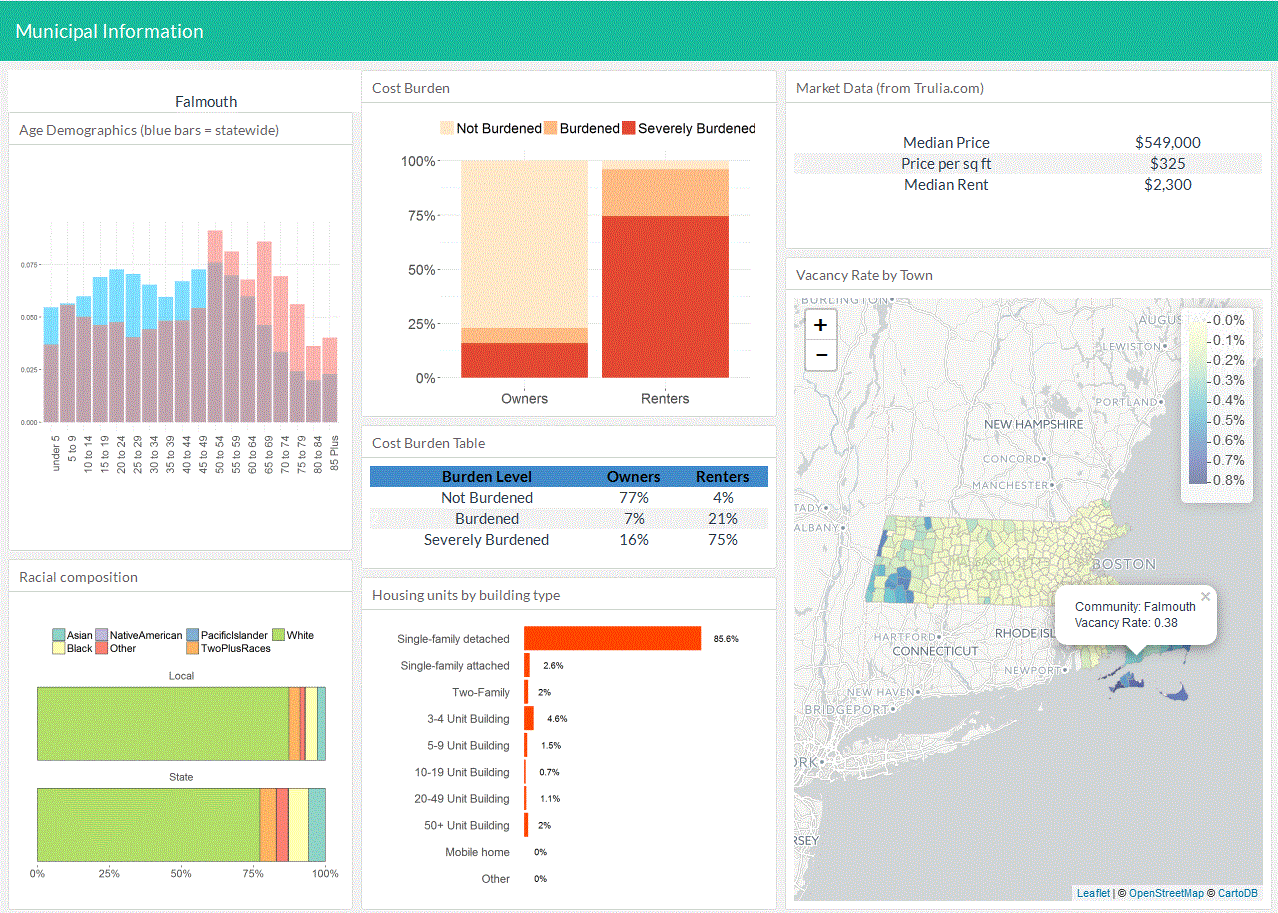
* Example of the table-making code for census table output from the acs package:



 - Column definition within flexdashboard document. User define column widths and add each square of report content in a separate code chunk:



* Three sample reports:



**Insights and next steps**

The information I’ve decided to include is not fully inclusive of all the information I think would be useful to have in a profile of a community. I wound up focusing heavily on demographic information, but there are certainly additional sources and datasets that would help to round out the picture of a community.

The mapping element of the dashboard was included primarily to show some diversity in the kinds of data and visualizations that are possible through this methodology and presentation method (flex dashboard). A more formal, final product would definitely include a map that used data more connected to the rest of the material. Additionally, I would like to explore how to set the map to zoom automatically to the selected community based on latitude and longitude from the shapefile, and perhaps highlight or outline the selected community so that it is easier to spot on the statewide map.

The SQL database schema I created was also very simple. Since all the data that I stored was from the U.S. Census Bureau, and all from the same census program (5-year ACS) the SQL statements used to retrieve the data were not particularly complex. Introducing additional datasets from other sources would undoubtedly make these SQL statements more complex. I would need to connect any other datasets via either a GEO\_ID key or, possibly, community name. I could possibly do this through a linking table that provided an intermediary step between the census acs data and the other sources.

Flexdashboard was deliberately chosen because of its relationship to html content. A long term goal of this project would be to build a web app or some other kind of user interface from the starting point I’ve developed in flexdashboard, allowing users to input or select a community name and generate the report themselves.

Another future goal is to be able to extrapolate this process to other states and regions. However, because Massachusetts is comprised of 351 cities and towns, with no unincorporated areas, the geographical level of the data differs from what might be appropriate for another state. For example, in most states there are active county governments, and decisions about housing and development are often made on the regional or state scale. Extrapolating this methodology for another state might not be terribly difficult, however, since there are only a few lines of code that are being used to identify the geographical level for which to pull the census information. For Massachusetts, town level data is pulled by selecting ‘County subdivisions,’ where in another state, the code might be altered to pull county information.

Overall, I hope the code and methodology I’ve developed for this project can be a starting point for something that can be very useful to those focused on housing issues on the local level. While it certainly does not include all of the information needed to have an informed conversation, it as least demonstrates a way to make this information accessible and organized.

**Accompanying documents:**

1. Two Rmarkdown documents:
   1. Scrape, tidy, store
      1. Contains the download, formatting and storage of census information
   2. Dashboard
      1. Must be knit and requires flexdashboard to be installed to run.
2. SQLite database – CensusDB
   1. While the first Rmarkdown document will set up a SQLite file, I am attaching the already-complete database.
3. Three sample reports generated from the dashboard code, for what I feel is a good range of community types. They are the same examples as in the screen shots above:
   1. Falmouth: demographics shift to older generations, with low levels of racial diversity, high proportions of single-family housing stock, and high vacancy rate (likely due to the preponderence of vacation homes and seasonal renters in the town. Renters, here and everywhere, are more cost burdened than owners.
   2. Lexington: a more affluent suburb, with disproportionate shares of children and baby boomers, but a demographic trough around millennials. High proportion of single-family homes and low vacancy. High cost burden for renters.
   3. Somerville: demographically opposite to Lexington, with high proportions of millennials, low levels of children and baby boomers. Low vacancy, high cost burden for renters.

Data sources:

Trulia.com. “Real Estate Market Overview” <https://www.trulia.com/real_estate/Boston-Massachusetts/>

U.S. Census Bureau, American Community Survey, 5-year estimates 2011-2015. Data accessed through API. <https://www.census.gov/programs-surveys/acs/>