Project Stage I

Data Collection & Defining the Data Science Problem

Questions

Our provisional questions are:

1. Is there a correlation between earnings and home values?
2. Which areas of the US have seen the greatest fluctuations in home values before and after The Great Recession?
3. Are denser areas more expensive than more rural areas?
4. Which areas have the greatest disparity between female and male earnings?
5. Which areas have the best earnings : home value ratio (i.e. which areas are most affordable given the median total earnings)?
6. Is there a correlation between home value and month? Some speculate that selling prices are higher in the summer months due to increased demand--is this observable in our data?

Data Sources

Zillow's Home Value Per Sq Ft ($)
This source provides the median home value per square foot of 12,315 US zip codes. There are an estimated 43,000 zip codes in total, so not all areas of the US are represented. Home values are listed month by month from April 1996 to November 2016. The values are based on estimates and the methodology is explained in further detail here.

American FactFinder 2015 Estimated Earnings (2015 inflation-adjusted dollars)
These are estimates of 2015 earnings from the Census Bureau's Population Estimates Program. It includes total mean and median earnings and earning distributions based on gender, earning amounts, and educational attainment. We chose the 2015 dataset because it is the most recent. It includes 33,120 tuples.

Classifieds Real Estate Listings
The 300 text documents are property listings that were scraped from free-classifieds-usa.com using WebHarvy. They include homes of various sizes and from multiple regions of the US. The WebHarvy data was stored in *.csv format, which we extracted to *.txt using this script.

Structured Data Extraction Methods

Zillow Home Value Data
The Zillow home value estimates were already available in *.csv format and we downloaded the data directly from their site.

Earnings Data

The earnings data came from the American FactFinder site, which is run by the US Census Bureau. We customized our table to include the zip code and downloaded the *.csv file from the site.

Text Document Extraction Proposal

Some attributes we are interested in extracting from the text documents are:

1. List Price
2. Square Footage
3. Number of Bedrooms/Bathrooms
4. City
5. Type (single-family, condo, duplex)

Tools Used

WebHarvy

WebHarvy is a web scraper that has a relatively user friendly point-and-click GUI. It requires no programming as it detects patterns (ostensibly using HTML DOM elements). We chose to use this tool as it was faster than writing a custom crawler to harvest the 300 text documents from Craigslist. It is not open-source but it is freely available for download with some locked features.