New York Real Estate Transactions

Alex Chaffers, Robbie Zielinski

5/2/2018

Motivation

The Great Recession of 2008 caused housing prices to plummet across the nation. As time passed different areas have recovered at different rates. New York State makes this disparity apparent. While areas around New York City have largely recovered, in Upstate New York large areas have still not reached pre recession economic levels. We wanted to explore this dynamic through looking at real estate transactions. Theoretically, fewer jobs created should lead to lower population growth, fewer real estate transactions, and lower prices. We used population and economic data provided by the Census ACS survey to real estate transaction data and population and economic data to explore these trends.

Data

American Community Survey

The American Community Survey or ACS, is carried out every year by the Census Bureau to provide government agencies and businesses with up to date information about population, educational attainment and various economic indicators. It is designed to help local government agencies and businesses gather data about their communities. To acquire this data we used the `acs` package in R. The package allowed us to easily specify particular areas of interest, using the `geo.make` function. Then, the `acs.fetch` function allowed us to import a specific acs table for our area of interests. We used the ACS package to find population and median income data from Albany, Rensselaer, Saratoga, Orange, Dutchess, Erie and Monroe Counties in New York.

Real Estate Transactions

We accessed real estate transactions for many of the counties in New York through nydatabases.com. In order to pull them from the website, we wrote a web scraper in Python and created a MySQL server to store the data. When we gathered the data from the website, it included the street address and the information of the location's town and county, but not latitude and longitude coordinates, which would allow us to generate a map of the transactions. In order to get this data, we used the `geocode` function in the `ggmap` package. Because we were dealing with over 1.8 million observations, geocoding each address would take far too long, so we used the doParallel package to be able to cut down the run time.

Shiny App

To present our data, we developed a Shiny app using `leaflet` to make an interactive map of the transactions. The user is able to filter by year and county, choosing the year of data they want to examine and a reference year to compare it to. On the map, we show points of all real estate transactions in the chosen counties and year, colored by their price, and in a separate tab, we provide tables giving an indication of the change in average real estate prices, population, and average salary in those counties in the time between the chosen years.

! [Alt text] ("~/STAT231-S18-GroupM/Shiny_screenshot.png")

Next Steps

First and foremost, the next steps that we would take for this project involve gathering more data. Because the real estate database's website crashed, we were only able to collect data real estate data from 28 of the 57 counties included in the dataset. We would certainly finish gathering the real estate data from 1993-2017 for all the counties included in the database. Additionally, because the database is updated weekly, we would be interested in developing a program that automatically pull this new data into our MySQL server. The ACS API only included population and income estimates from 2009 to 2016, although there is data on the Census Bureau's website dating back to 2000. We would like to find a way to incorporate those estimates into our project. Once we have more complete data, we would be interested in more closely examining the relationships between the population, income, and real estate price variables. Potential questions include:

- What order does causation take? Does income influence population, then house prices, or do we see that population drives house prices, which then changes incomes?
- How much of a lag do we see between changes in variable values? How long does it take for a change in population to translate to a change in house prices? Are these values different? If so, why?