Coursera Data Science Capstone Project Report

Business Site Selection with Foursquare and Python
Problem Definition and Analytic Approach

Dec 19, 2018
Bryant Sheehy Jr.

Introduction

For my capstone project, I chose to use the Foursquare data set and Python data wrangling and mapping tools to help solve the problem of where to locate a new business in the city of Chicago. I live in the Chicagoland area, so I'm familiar with many of the neighborhoods of the City and the general layout. I also happen to have some experience in a previous life as a commercial real estate broker, so I know a bit about site selection. So I am going to attempt to create a tool that would allow an entrepreneur, the "user", the ability to search the City of Chicago to find potential locations for a new retail or service business.

When deciding where to open a new retail or service business, there are a number of considerations to think about. First of course, you want to think about who your potential new customers are and where they might be located. Then you want to look at similar businesses already located there to get a sense of what the potential competition might look like.

To keep things simple for the purpose of this project, we're going to address the first consideration by focusing the study within the city limits of Chicago. Chicago is a very diverse and relatively dense urban area. It is made up of seventy-seven distinct communities or neighborhoods with a wide variety of cultures and income, similar to boroughs in New York. So there should be potential customers of all types for just about any kind of retail or service business here.

The Problem

The problem we will try to solve is where to locate a new business such that the competition will be minimized and potential customer income will be adequate to support a new business, all else being equal.

The Analytic Approach

We will try to solve this problem by modeling the income distribution in the City, and then identifying communities within the City where there are fewer potential competing businesses

and therefore potentially unmet needs. Within the Jupyter notebook, a user looking for the right location to open their business would be able to enter one or more business categories to search, and see a list of how many businesses in this category are located in each community along with a couple of maps showing which communities have higher or lower concentrations of these types of businesses based on locations per capita and/or locations per square kilometers.

The Data

The Foursquare Search API provides a convenient method of searching any urban geographic area and pulling up a list of most businesses located in that area with each business assigned to one or more very granular business type categories. Bing provides an API that allows you to pull the geographic coordinates for any location in the world that can be described in terms of a place name. Wikipedia provides a good description of each community within the City of Chicago which includes the number of people living there and the size of the geographic area in square kilometers. So we will use data from these three sources.

From Wikipedia, we will pull the names of each Chicago community along with its population, geographic area and income statistics. You can see an example of where this information is located on Wikipedia at https://en.wikipedia.org/wiki/Community_areas_in_Chicago.

From Bing, we will pull the geographic coordinates for each community. You can see what MSN/Bing offers in this area at https://docs.microsoft.com/en-us/bingmaps/spatial-data-services/geodata-api. We're actually going to use the geocoder library to access this data.

With the data from Wikipedia and Bing, we can create some starting data sets that look something like these dataframes:

	Community	Area	Population	Income
0	Albany Park	5.00	52079	51969
1	Archer Heights	5.21	13266	43394
2	Armour Square	2.56	14068	24336
3	Ashburn	12.61	42752	63573
4	Auburn Gresham	9.76	45842	29389

	Community	Latitude	Longitude
0	Albany Park	41.968094	-87.721542
1	Riverdale	41.660000	-87.610001
2	Edgewater	41.985710	-87.663460
3	Archer Heights	41.811539	-87.725563
4	Armour Square	41.834579	-87.631889

With Foursquare, we will use the Search API to pull the names, categories, addresses and geographic coordinates for each business located in the Chicago city limits. We will then allow the user to enter in one or more Foursquare business categories to search for and display a couple of maps, choropleth and marker clusters, that indicate which communities have lower densities of the chosen business type, higher or lower median household income and where the potential competition is located. Here are a couple dataframe examples of what we can pull from Foursquare, then combined with Wikipedia and Bing:

	Community	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Category ID
0	Albany Park	41.968094	-87.721542	Lawrence Eye Care	41.968289	-87.721245	Accessories Store	4bf58dd8d48988d102951735
1	Albany Park	41.968094	-87.721542	El Gallo Bravo #6	41.968324	-87.721338	Mexican Restaurant	4bf58dd8d48988d1c1941735
2	Albany Park	41.968094	-87.721542	Cuenca's Family Hair Cut	41.968330	-87.722832	Salon / Barbershop	4bf58dd8d48988d110951735
3	Albany Park	41.968094	-87.721542	Chicago Canvas & Supply	41.968304	-87.721737	Building	4bf58dd8d48988d130941735
4	Albany Park	41.968094	-87.721542	Gallo El Bravo Number Six	41.968413	-87.721364	Food	4d4b7105d754a06374d81259

	Community	Area	Population	Income	Coffee Shops	Coffee Shops/SqKM	Coffee Shops per Capita	PerSqKM Norm	PerCap Norm
58	Riverdale	8.70	7090	14846	0.0	0.000000	0.000000	0.000000	0.000000
16	Clearing	6.63	24962	60624	0.0	0.000000	0.000000	0.000000	0.000000
20	East Side	7.25	23784	43421	1.0	0.137931	0.000042	0.011283	0.017820
4	Auburn Gresham	9.76	45842	29389	1.0	0.102459	0.000022	0.008381	0.009246
62	South Deering	23.03	15305	35056	1.0	0.043422	0.000065	0.003552	0.027692
30	Hegewisch	12.38	8985	50338	1.0	0.080775	0.000111	0.006607	0.047171
61	South Chicago	8.65	28095	28504	2.0	0.231214	0.000071	0.018913	0.030171

The idea is the help the user find areas in the City where customer income is adequate, competition might be lower, and where they might be able to find underserved geographic holes in the market.