Coursera Data Science Capstone Project Report

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

Dec 20, 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.

Methodology

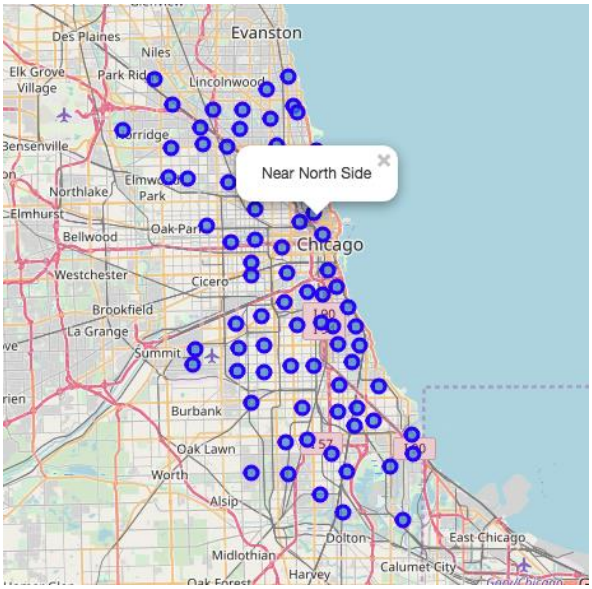
In terms of methodology, here is how set up the processes in the Jupyter Notebook and the steps that I executed.

Fetching the Data

I started with the list of Chicago neighborhoods from Wikipedia, using a Python library called BeautifulSoup to parse the list table into a Pandas dataframe. Initially, Wikipedia yielded a list of communities which were then subdivided into smaller neighborhoods. Knowing that I would be uploading this data into one or two choropleth maps, I found a GeoJSON file for Chicago and examined it to see how the geographic descriptions matched up with the Wikipedia list. I found that the GeoJSON file appeared match the list of communities more closely than the more granular list of neighborhoods, so I chose to focus on the communities as the primary geographic unit of measure. Even then, I had to make manual adjustments to the place names in the GeoJSON file to get it to match up close to 100%.

While examining the community data and setting up the scraping routine, I noticed that they had population, geographic area and median household income for each community. That data was buried on 77 separate pages in a rather unstructured way. So I decided to just manually type that data into a CSV file in about 30 minutes, instead of spending a couple days trying to figure out how to scrape it. If I were to expand this project to multiple cities, I would take the time to figure out a scraping routine at that point.

The next step was to get the corresponding latitudes and longitudes from Bing, via Geocoder, for each community and pull them into a dataframe. Once that was done, I used the Folium library to create a basic circle marker plot map with labels just to make sure everything laid out like it was supposed to. From this screenshot, it looks like I got the job done so far.



Next, I set out to get the venue data from Foursquare using their Search API. I used a very similar code module as the ones we used to get the Foursquare venue data for New York and

Toronto. My objective was to get data for every commercial venue recorded by Foursquare in the City of Chicago. Using the geographic size data in square kilometers for each community that I copied from Wikipedia, I figured out that if each community was a perfect circle, the average radius would be about 1.6 kilometers.

So for the Foursquare API parameters, I used a radius of 2 kilometers and limit of 10,000 businesses per community just make sure I captured as many venues as possible. For each venue, I captured its name, ID code, latitude, longitude, category and category ID code. I ended up with 9,743 venues after deduping. From this data set, I sliced out the list of venue categories and exported it to a CSV file so I could use it to pick out a few business categories to investigate further.

Choosing a Business Category to Focus On

Looking over the category list, I picked out a few at random to play with – Dentists, Irish Pubs, Coffee Shops, Clothing Boutiques and others. For the purpose if demonstrating the functionality of this tool set, I chose to use Coffee Shops because there are a lot of them all around Chicago. In the notebook, I set up a couple of input cells where the user can enter their choice of category name and ID, which are then stored in variables to be used to pull data for analysis in the rest of the notebook.

In the next step, I went back to Foursquare to pull the name, ID code and geographic coordinates for every coffee shop in Chicago. Then I examined the other Foursquare API endpoints to see if I could find more qualitative data about each coffee shop to try and distinguish them from each other. When I was looking at many of the venue locations in detail, I noticed many of them had what looked like bogus residential addresses. One good example was “Tito and Courtney’s Strip Club” located in the middle of a residential block in Albany Park! So I thought it would be good to find a way to focus on the more legitimate venue locations. ;-)

Unfortunately, Foursquare does not provide a lot of qualitative data for free, but they do provide the number of “likes” for each venue, that is the number of people who tagged each venue as one of their favorites. So I used the Likes API to pull this data into the notebook for each coffee shop. Out of a total of ~1,100 coffee shop locations, 855 were “liked”. I figured that if I was looking at each one as a potential competitor, I would want to focus on the ones that people like. After pulling the number of likes for each coffee shop, I calculated the mean and median Likes values per venue as well and printed them out. This give the user the option of filtering out the venues with zero likes or go further and filter out those below the median or below average. I chose to be less restrictive and only filter out the coffee shops with zero likes.

Here’s what that looks like:

```
(1113, 3)
The average number of likes is: 36
The median number of likes is: 10
The number of competitors with any likes is: 855 out of 1113 Coffee Shops in total.
```

	Community	Community Latitude	Community Longitude	Coffee Shops	Competitor ID	Competitor Latitude	Competitor Longitude	Likes Count
0	Albany Park	41.968094	-87.721542	Starbucks	4b1bdfabf964a5206efe23e3	41.964527	-87.708840	58
1	Albany Park	41.968094	-87.721542	Nighthawk	579964a8cd10bd6545689b88	41.967974	-87.713415	19
2	Albany Park	41.968094	-87.721542	Backlot Coffee	5992ee341c675b5afdb4b94c	41.953261	-87.731976	15
3	Albany Park	41.968094	-87.721542	Starbucks	4b5afeaaf964a520e7dd28e3	41.968911	-87.728817	49
4	Albany Park	41.968094	-87.721542	Café Descartes	4c4082e6cc410f4794b4a961	41.981552	-87.718156	5

Next, I grouped the coffee shop venues by community, uploaded the community population and income stats file from Wikipedia and appended it to the grouped venues. This give us a choice of calculating the density of competition either by the number of venues per square kilometer or per capita. Since we are focusing on targeting customers more than geographic areas per se, I chose to use the number of venues per capital as the primary density evaluation metric, but I wanted to give a user the ability to make that choice either way. For the purpose of showing the differences between communities on a choropleth map, I also added new columns to normalize this data to a uniform 0 to 1 scale. Here's what that looks like:

	Community	Area	Population	Income	Coffee Shops	Coffee Shops/SqKM	Coffee Shops per Capita	PerSqKM Norm	PerCap Norm
16	Clearing	6.63	24962	60624	0.0	0.000000	0.000000	0.000000	0.000000
30	Hegewisch	12.38	8985	50338	0.0	0.000000	0.000000	0.000000	0.000000
58	Riverdale	8.70	7090	14846	0.0	0.000000	0.000000	0.000000	0.000000
62	South Deering	23.03	15305	35056	0.0	0.000000	0.000000	0.000000	0.000000
20	East Side	7.25	23784	43421	1.0	0.137931	0.000042	0.011513	0.026730
13	Calumet Heights	4.58	13732	49923	1.0	0.218341	0.000073	0.018225	0.046297
61	South Chicago	8.65	28095	28504	1.0	0.115607	0.000036	0.009650	0.022629
4	Auburn Gresham	9.76	45842	29389	1.0	0.102459	0.000022	0.008552	0.013868

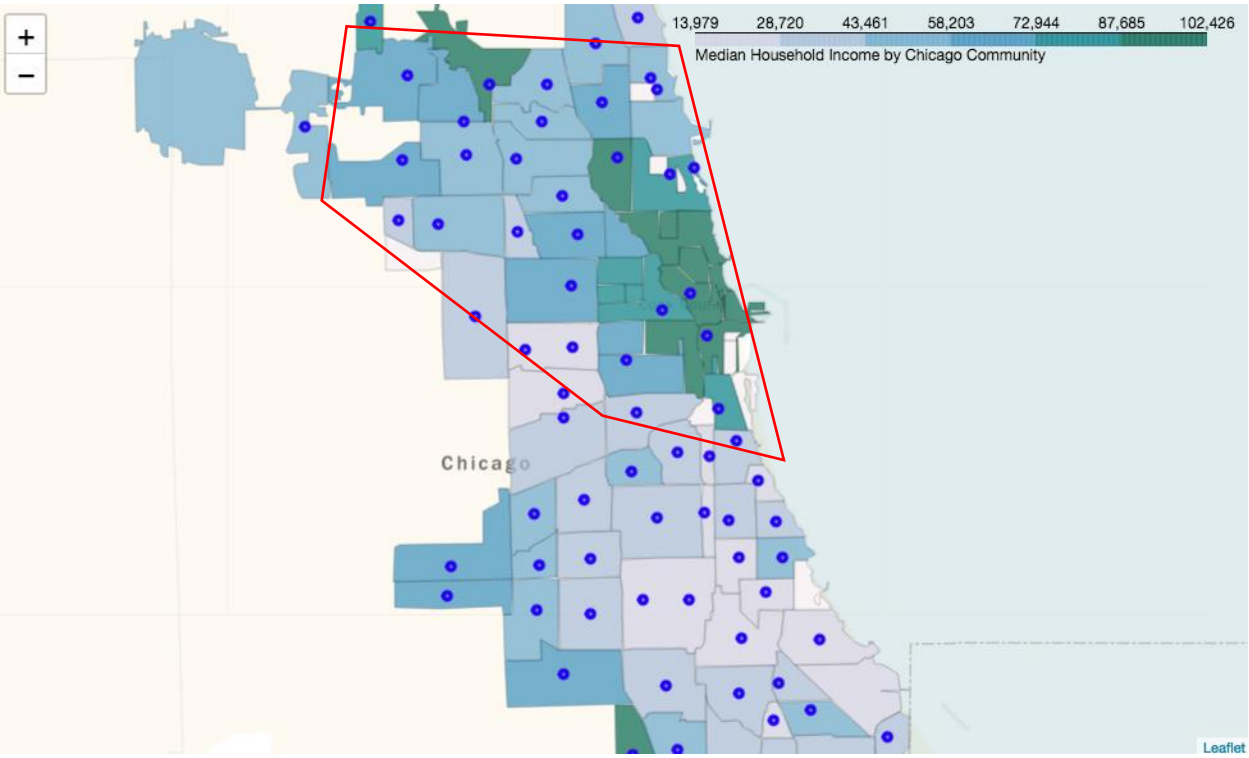
Sorting by the number of coffee shops with fewer at the top gives us a few initial prospective communities to investigate further. The first four communities have no coffee shops according to Foursquare. One question we might want to ask is Why? Where are these communities located?

The first community on the list, Clearing, is located on the southwest side, next to Midway Airport. This is a working class area of very well kept mid-century bungalow style homes, sitting on top of a large industrial park in the neighboring town of Bedford Park. The media income in Clearing is solidly in the middle-class range, so this place might be worth further investigation.

The other three of communities, Hegewisch, Riverdale and South Deering, are located at the far south end of the City, a run-down older industrial area with income levels on the lower end of the range. Putting myself in the user's shoes, its been my anecdotal experience that coffee shops tend to appeal to people who can afford to pay $3-5 for a cup of joe. So I'd make a note to investigate these locations further, but I would not put them at the top of the list.

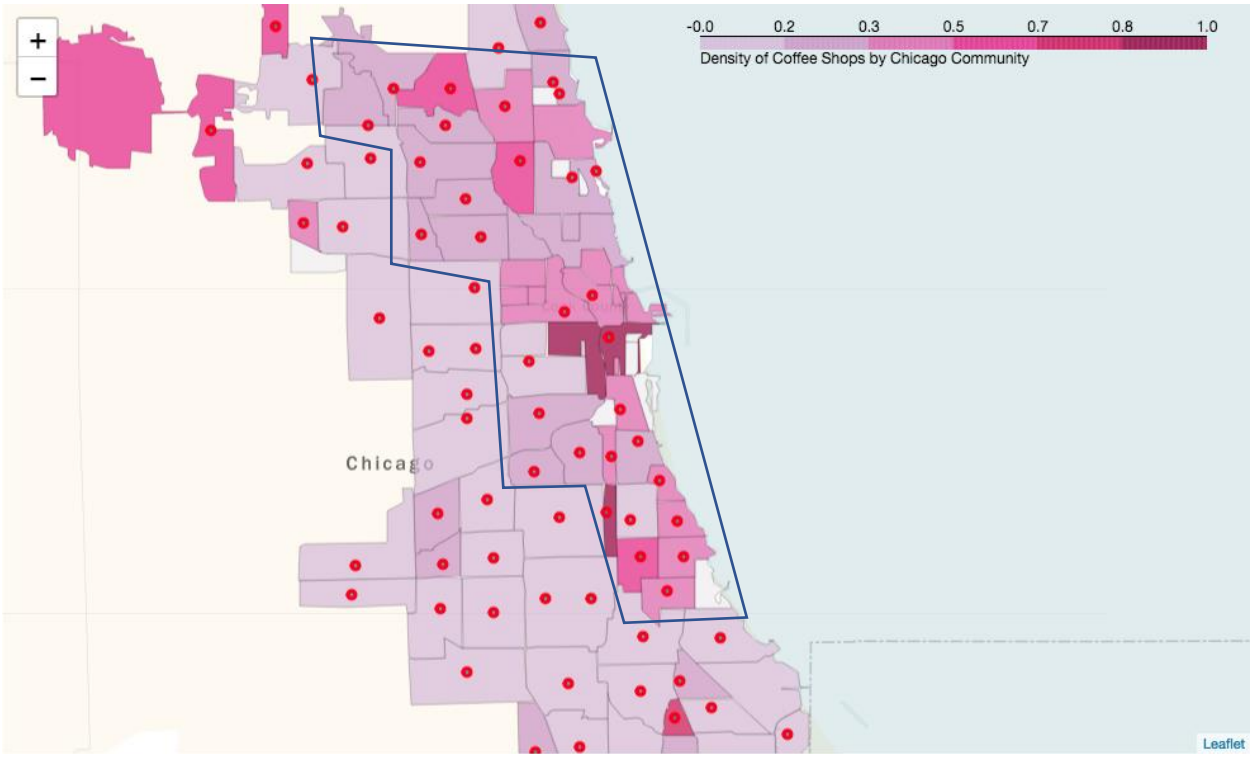
Now that we have pulled in all the data, sorted, cleaned and compiled it into a series of useful Pandas dataframes, the next step is to create a series of maps showing income distribution and competitor venue density. For the first two maps, I used the choropleth map type to show the differences between communities on these metrics. The third map uses marker clusters to give the user the ability to drill in geographically and see how the potential competition might be geographically distributed in a more close-up granular fashion. Here's what they look like:

Income Distribution



The darker green areas represent higher median income levels. As you can see, most of the money is located on the north side of the City, as indicated with the red polygon.

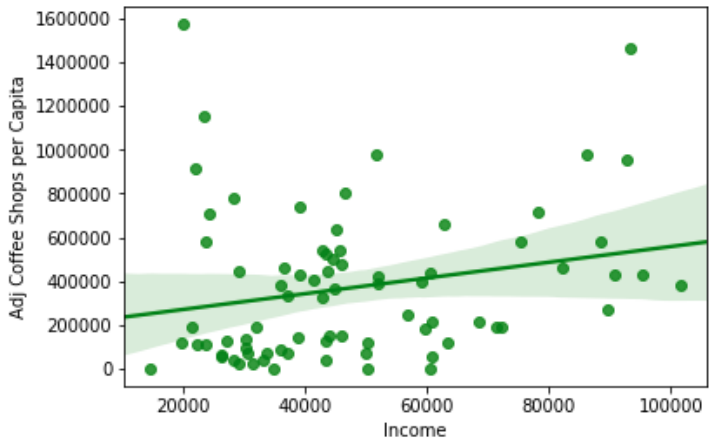
Competitor Density (Venues per Capita)



The darker red areas on this map represent more coffee shops per capita. It is interesting how the locations of these higher density areas appear to correspond to a noticeable degree to the higher income areas. I suspect there is a reason for that as I alluded to above.

Results Analysis

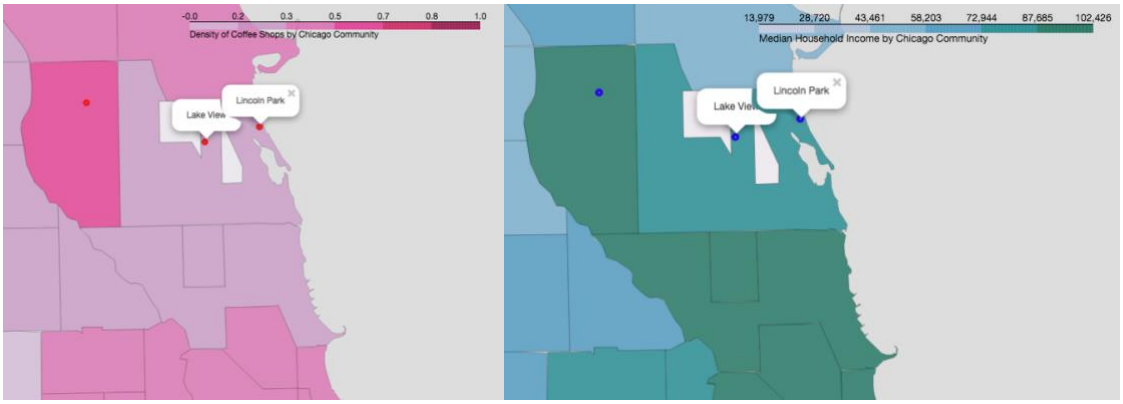
I began this study with the idea that one would look to geographic areas where the density of potential competition is lower as potentially good places to locate a new retail service business. But then looking at the apparent overlap between the higher venue density areas and the higher income areas, I decided to create a Seaborn regression chart to see if indeed, there might be a correlation. Here's what that chart shows:



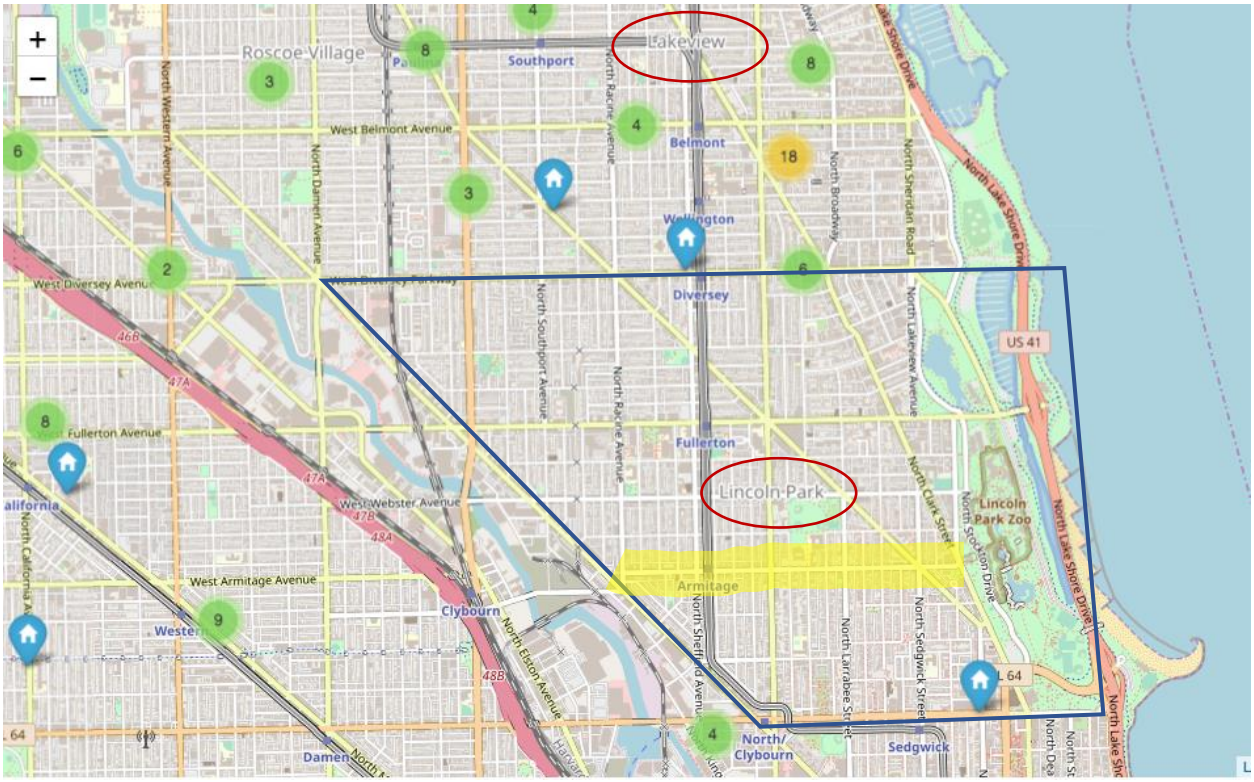
It looks to me like there is a positive correlation. Not a strong one, but it's there.

So in the interest of time management efficiency, I would start looking at the northern communities first even though the average density of competition appears to be somewhat higher. They say misery loves company. Well maybe success loves company too, at least with coffee shops?

There are two communities that catch my attention on the maps, Lake View and Lincoln Park. Both are on the lower side of the coffee shop density range and on the higher side of the income range.



So we will focus next on Lincoln Park to start with. That's where the third map comes in, the Marker Cluster map. I loaded all the filtered coffee shop locations, the ones with "Likes", into a marker cluster map that allows a user to drill into a specific area and see how the competition might be situated geographically. Here's what it looks like for Lincoln Park.



Each green and yellow circle represents several coffee shop locations that will open up and turn into blue markers like the markers showing at the top and bottom of the polygon as you zoom in. The polygon itself is Lincoln Park. Lake View sits right above it.

Notice that there are almost no coffee shops shown in Lincoln Park. This is one of the highest income areas in Chicagoland. The yellow highlighted street in the bottom half of the polygon is Armitage Avenue, one of the most trendy shopping areas in the City, chock full of eclectic boutiques and restaurants popular with tourists and locals alike. One would think this would be a prime location for a high-end upscale coffee shop catering to the local elite.

Let's dig in further. In the last few cells of the Jupyter Notebook, I created a procedure where the user can enter the name of a single community and the Notebook will look of the coordinates of that location and find all the coffee shops located there, and then see which ones, if any, are "liked". Here's what I found:

(30, 5)

	Community	Coffee Shops	Address	Zip	Venue ID
0	Lincoln Park	Bittersweet Pastry Shop & Cafe	1114 W Belmont Ave	60657	4a14ef13f964a5208c781fe3
1	Lincoln Park	Osmium Coffee Bar	1117 W Belmont Ave	60657	53c159bd498e27a169be54bc
2	Lincoln Park	Starbucks	2754 N Clark St	NaN	58cdbddd375c4a6ccf6f1712
3	Lincoln Park	Stan's Donuts & Coffee	2800 N Clark St	60657	5618f770498ee790946d31ff
4	Lincoln Park	Hero Coffee Bar	2950 N Sheridan Rd	60657	5bb93e8447f876002cf49542
5	Lincoln Park	Bobtail Ice Cream Company	2951 N Broadway St	60657	4a25b16af964a5207f7e1fe3
6	Lincoln Park	Peet's Coffee & Tea	3025 Clark St.,	60657	5425c2f7498ef7818825c1fe
7	Lincoln Park	Starbucks	3030 N. Broadway	60657	57e155c0498e784cdf75cdf8
8	Lincoln Park	Intelligentsia Coffee	3123 N Broadway St	60657	4234d400f964a5200f201fe3
9	Lincoln Park	Starbucks	3184 N Clark St	60657	4a15b903f964a520c0781fe3
10	Lincoln Park	The Alley Chicago	3221 N Clark St	60657	5a0d81858ad62e5b0572e13c
11	Lincoln Park	Stan's Donuts & Coffee	3300 N Broadway St	60657	574059ef498e6a948ca72579
12	Lincoln Park	The Coffee & Tea Exchange	3311 N Broadway St	60657	4a58d9ddf964a52008b81fe3
13	Lincoln Park	Yefseis Café	3344 N Halsted St	60657	54a9810d498e959269e02ef4
14	Lincoln Park	Starbucks	3358 N Broadway St	60657	4b40f526f964a5203dbe25e3
15	Lincoln Park	Pick Me Up Café	3408 N Clark St	60657	40b68100f964a5206d001fe3
16	Lincoln Park	Lakeview Rewired Cafe	3508 N Broadway St	60657	596e6713d48ec155ebd81b18
17	Lincoln Park	Starbucks	3549 N. Sheffield Ave.	60657	4a8da3b1f964a5205a1020e3
18	Lincoln Park	Starbucks Reserve	3649 N Clark Street	60613	58fef9fb6fd626300700921c
19	Lincoln Park	Coffee Tree & Tea Leaves	3752 N Broadway St	60613	4ba64610f964a5204b4139e3
20	Lincoln Park	Uncommon Ground	3800 N Clark St	60613	49e676b0f964a5204e641fe3
21	Lincoln Park	Emerald City Coffee	3938 N Sheridan Rd	60613	4a01ecb1f964a5200d711fe3
22	Lincoln Park	Dunkin' Donuts	3949 N Broadway St	60613	4b5b6256f964a520aef928e3
23	Lincoln Park	Dollop Coffee & Tea Co.	4181 N Clarendon Ave	60613	422f8e00f964a520f91f1fe3
24	Lincoln Park	Starbucks	4446 N Broadway St	60640	4c51f1ed9d642d7f401b6ade
25	Lincoln Park	11 Degrees North	824 W Belmont Ave	60657	591a0228b546182516881644
26	Lincoln Park	Coronas Coffee Shop	909 W Irving Park Rd	60613	4b135fdbf964a520b49623e3
27	Lincoln Park	The Satellite Cafe	942 W Montrose Ave	60613	58dc0dd245005e6e8a125ecf
28	Lincoln Park	Dunkin' Donuts	949 W Addison St	60613	532488dc498e89b38c11821c
29	Lincoln Park	Dunkin Donuts	Belmont CTA	60657	5262bfa911d2701752411016

It turns out that there are 30 coffee shops in Lincoln Park after all. But most are located around the edges of the community, and none are located on Armitage Avenue. And check this out:

	Coffee Shops	Competitor ID	Likes Count
8	Intelligentsia Coffee	4234d400f964a5200f201fe3	404
15	Pick Me Up Café	40b68100f964a5206d001fe3	266
20	Uncommon Ground	49e676b0f964a5204e641fe3	246
5	Bobtail Ice Cream Company	4a25b16af964a5207f7e1fe3	211
1	Osmium Coffee Bar	53c159bd498e27a169be54bc	187
9	Starbucks	4a15b903f964a520c0781fe3	180
23	Dollop Coffee & Tea Co.	422f8e00f964a520f91f1fe3	152
3	Stan's Donuts & Coffee	5618f770498ee790946d31ff	152
14	Starbucks	4b40f526f964a5203dbe25e3	89

Only a handful of them have more than 100 Foursquare “Likes”! And only one of those is a recognizable national chain, Starbucks. Obviously, the folks in Lincoln Park like to have their coffee served with local flavor.

Conclusion

I think the next step is to hop in my car and go drive through Lincoln Park to see if I can find any vacancies in or around Armitage Avenue that look like good coffee shop locations. BTW, I’ve eaten at Stan’s Donuts (#4 on the list). Their donuts are really, really good!!

Before I hop in the car and drive off to Lincoln Park, I would repeat this drill-down exercise for several other communities on the north side. Perhaps Lake View, Uptown, Edgewater to the north and Logan Square and Avondale to the west. Maybe even Clearing too, over by Midway Airport.

In conclusion, I believe I have created a very logical data driven process for beginning a site selection search for a new business location within the City of Chicago. I hope you agree.

Bryant Sheehy