Battle of the Neighborhoods: Analyzing Neighborhoods to Select Optimal Office Location For Startup

Kairn Brannon

July 4th, 2020

I. Introduction

Business Problem

For this project we will investigate neighborhoods in Chicago, New York City, and Toronto to determine which city and neighborhood a mid sized start-up company should choose for its headquarters. The selection process for a budding company's main office can be strenuous, and it's often hard to figure out the best location without physically being in the city. For business reasons they have narrowed their search to Chicago, New York (Manhattan or Brooklyn only), and Toronto, but have no way to decide which city is best. The company leaders have decided that they would like to put their office in a fun and vibrant neighborhood with plenty of good food and bars nearby for their employees. In addition, they require that the neighborhood has bars, parks, and if possible a gym nearby. The output of this project will be a recommendation of the ideal city and neighborhood for the start-up's office and the findings will be presented to the executive team.

II. Data Acquisition

Neighborhood Location

We will use neighborhood location data sets for Chicago, New York City, and Toronto. The name and coordinates of each neighborhood in each city are found at various locations online. For New York City, we find the data is publicly available here. For Chicago, we find the neighborhoods and their coordinates at this website. For Toronto, we will scrape neighborhood data from the wikipedia page and then find the coordinates of the neighborhoods here. An example of the resulting data for New York CIty is shown below.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

For New York, after importing our dataset for all 5 boroughs we drop the columns for Bronx, Staten Island, and Queens since our search is restricted to Manhattan and Brooklyn.

For Chicago, we parse the coordinates of the GeoJSON multipolygon and find the minimum and maximum longitude and latitude for each neighborhood. We then take the midpoint of both to and use the resulting coordinates as the center of the neighborhood.

For Toronto, we use the beautiful soup package to web crawl the wikipedia page for neighborhood names and then append the coordinates sourced from a different location.

Venue Data

I will leverage the Foursquare API and request venue data for each neighborhood by passing the longitude and latitude of each neighborhood. Calling the “explore” endpoint, Foursquare will return data including the venue name, location, and category, all of which we will save and store in a dataframe. We define a function that gets up to 100 venues within 500 meters of each neighborhood and stores the venue information in a large dataframe along with the city and neighborhood that the venue resides. The head of this dataframe is shown below.

	Neighborhood	Latitude	Longitude	City	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	New York	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	New York	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	New York	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	New York	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	New York	Dunkin'	40.877136	-73.906666	Donut Shop

III. Data Analysis

Clustering Analysis

The first operation we do with our combined dataset is to represent how frequently venues occur within each neighborhood. We do this to create a dataset to perform a k-means cluster analysis which will group our neighborhoods into clusters based on the venues returned by Foursquare. However, we have 290 neighborhoods to cluster so we should try to eliminate a few neighborhoods before performing the analysis to improve our results.

The start-up executives informed us that having lots of bars is necessary for the neighborhood location. To narrow down our neighborhood pool, we calculate the top 10 most common venues for each neighborhood. The resulting dataframe head is shown below.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt\n	Breakfast Spot	Lounge	Latin American Restaurant	Clothing Store	Skating Rink	Exhibit	Egyptian Restaurant	Electronics Store	Empanada Restaurant	English Restaurant
1	Albany Park	Mexican Restaurant	Bus Station	Bakery	Korean Restaurant	Chinese Restaurant	Grocery Store	Park	Asian Restaurant	Taco Place	Karaoke Bar
2	Alderwood, Long Branch\n	Pizza Place	Pub	Sandwich Place	Coffee Shop	Gym	Pool	Dance Studio	Ethiopian Restaurant	Entertainment Service	Empanada Restaurant
3	Archer Heights	Mexican Restaurant	Bakery	Hotel	Discount Store	Fast Food Restaurant	Nightclub	Electronics Store	Bar	Bank	Sandwich Place
4	Armour Square	Chinese Restaurant	Cosmetics Shop	Italian Restaurant	Sandwich Place	Asian Restaurant	Gas Station	Grocery Store	Indian Restaurant	Hot Dog Joint	Filipino Restaurant

Next we reduce the pool of neighborhoods to ones that have bars as one of their top 10 most common venues and create a new venue frequency dataframe with the resulting 57 neighborhoods. This is the dataframe we will use for our clustering analysis.

49	South Side	0.000000	0.0	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.000000	0.000000	0.0
50	University of Toronto, Harbord\n	0.000000	0.0	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.000000	0.000000	0.0
51	Upper West Side	0.000000	0.0	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.000000	0.000000	0.0
52	Uptown	0.000000	0.0	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.000000	0.012346	0.0
53	West Pullman	0.000000	0.0	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.000000	0.000000	0.0
54	West Town	0.000000	0.0	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.000000	0.000000	0.0
55	Williamsburg	0.000000	0.0	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.000000	0.000000	0.0
56	Yorkville	0.000000	0.0	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.000000	0.000000	0.0

57 rows × 445 columns

Now that we have refined our pool, we perform a clustering analysis on the remaining neighborhoods in an effort to group them based on venue similarities. We run a k-means clustering analysis and group the neighborhoods into 10 clusters. Of the 10 clusters, clusters 2 and 5 stand out as having venues suitable for the office location. We create a new dataframe with neighborhoods from cluster 2 and 5. A park was the 2nd most desired venue for the office location, so we reduce the remaining neighborhood pool to neighborhood with parks also in their top 10 venues:

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	1	Avondale	Chinese Restaurant	Park	Food Truck	Donut Shop	Supermarket	Bar	Light Rail Station	Electronics Store	Sandwich Place	Grocery Store
1	4	Financial District	Coffee Shop	American Restaurant	Pizza Place	Bar	Sandwich Place	Cocktail Bar	Hotel	Park	Falafel Restaurant	Italian Restaurant
2	1	Gerritsen Beach	Ice Cream Shop	Pizza Place	Bar	Bagel Shop	Harbor / Marina	Liquor Store	Gas Station	Event Space	Park	Department Store
3	1	Irving Park	Bar	Breakfast Spot	Farmers Market	Latin American Restaurant	Park	Asian Restaurant	Donut Shop	Martial Arts Dojo	Café	Thai Restaurant
4	1	Red Hook	Seafood Restaurant	Art Gallery	Park	Bar	American Restaurant	Bagel Shop	Flower Shop	Farm	Café	Ice Cream Shop

This operation resulted in 5 neighborhoods, which are the finalists in our battle of the neighborhoods competition.

IV. Conclusion

For final considerations among the five remaining neighborhoods, we take into account that the executives also would like to have a cafe or coffee shop near the office. The Financial District in New York City, Gerritsen Beach in New York City, Irving Park in Chicago, and Red Hook in New York City all feature this, eliminating Avondale from the conversation. Of the remaining neighborhoods we see that Red Hook doesn't have many restaurants nearby besides seafood, instead featuring venues like art galleries, farms, and flower shops. The same argument can be had for Gerritsen beach, leaving the Financial District in New York City and Irving Park in Chicago as the winners of our competition. In conclusion of this project, we recommend the Financial District or Irving Park as destinations for their new headquarters.

V. Future Considerations

If I were to attempt this project again in the future I would expand the number of venues for each neighborhood. Upon investigation a lot of my Foursquare API calls returned 100 venues and I imagine a lot of valuable neighborhood venues were left out of our analysis. In addition, I would choose a different form of clustering because through the k-means cluster analysis a lot of the clusters only had one or two venues. There must be some other form of clustering of data processing to mitigate this. Last but not least, I originally attempted to include population demographics for each neighborhood but found it incredibly hard to math up the population with each neighborhood. Perhaps different data sets or more clever data cleaning would have been useful to achieve this goal.