Capstone Project Report

A business location recommending system

Introduction/Business Problem

The purpose of the project is to recommend neighborhoods in New York city to start a business of a given type.

Selection of location is essential to start a business. There are lots of approaches to identify locations best fitted for a type of store/venue and one of which is referring to the best rated venues in the city and selecting similar locations. Ideally, if a business of a given type is successful in neighborhood A, such business type will most likely be successful in similar neighborhoods B or C. The success of a store/venue is measured by its rating and count of likes. Thus, given a business type, the first several best rated stores of the same type will be searched in a city. Each of these stores will have its borough and nearest neighborhood calculated, referred to as target neighborhood. Within the same borough, all neighborhoods will be clustered and assigned with cluster labels. Only neighborhoods with the same cluster labels as the target neighborhood will be returned and a brief summary will be generated to describe the neighborhoods.

Data

Data to be used for this project include location and venue data from Foursquare and neighborhood/borough data of New York retrieved from Wikipedia.

- The Foursquare location data, combined with New York neighborhood data, will be used to describe the physical locations and calculate the nearest neighborhood of venues. Neighborhoods will be clustered based on the number of each venue type within a given radius.

- The venue data from Foursquare will be used to measure the quality of a venue given its rating and count of likes. Venues of a given type in a city will be ranked based on ratings and then count of likes. That being said, if two venues have the same ratings, the one with more count of likes will be ranked over the other one.

Methodology

The project aims to automatically generate analysis results and visualize the results with a Folium map. Steps are shown as follow:

1. Find the top best rated venues of a given type in a city.
   1). Use Geopy library, identify the latitude and longitude of the city.
   2). Use Foursquare venue search API, search for venues of the given type within radius.
   3). Sort the returned venue lists on ratings and then likes_count.

2. List the neighborhoods these venues locate
   1). Retrieve New York neighborhood data from Wikipedia and use Geopy to find latitudes of longitudes of all neighborhoods.
   2). Calculate the distance from venues in step 1 to each NY neighborhood. Assign a neighborhood with minimum distance to the venue.
   3). List the top neighborhoods as the target neighborhoods and corresponding boroughs as the target borough.

3. Use K-means clustering, find similar neighborhoods
   1). For each target borough, count the number of venues of each category.
   2). Use K-means, cluster neighborhoods with similar categories of venues in each target borough.
   3). Filter the clustering results with target neighborhoods. Only neighborhoods with the same cluster labels of target neighborhoods are listed.

4. Generate descriptive summary of these similar neighborhoods
   1). For each of the listed neighborhoods,a Foursquare venue search will be performed to collect information of venues with the type designated at the beginning of the analysis.
   2). All listed neighborhoods will be marked on the Folium map. Colors are used to denote cluster labels.

3). In the pop-up of each neighborhood market, following information will be displayed as a brief summary:

a. Name: Borough.Neighborhood

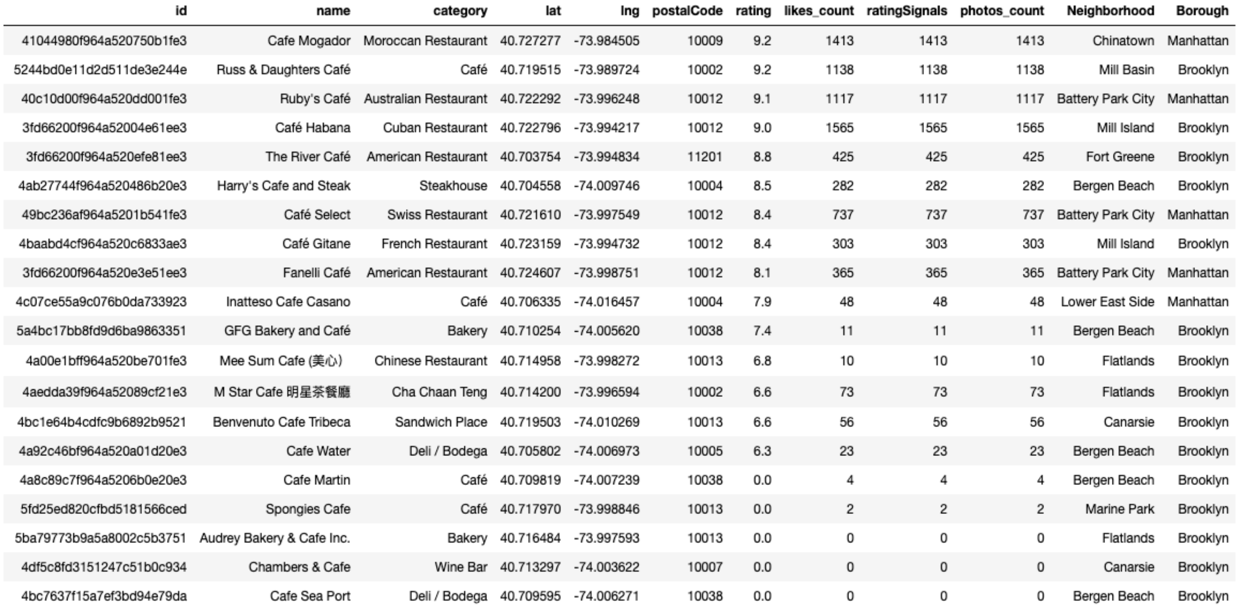
b. Top 3 venue types

c. Number of designated venues

d. Average Ratings / Average likes_count

Results

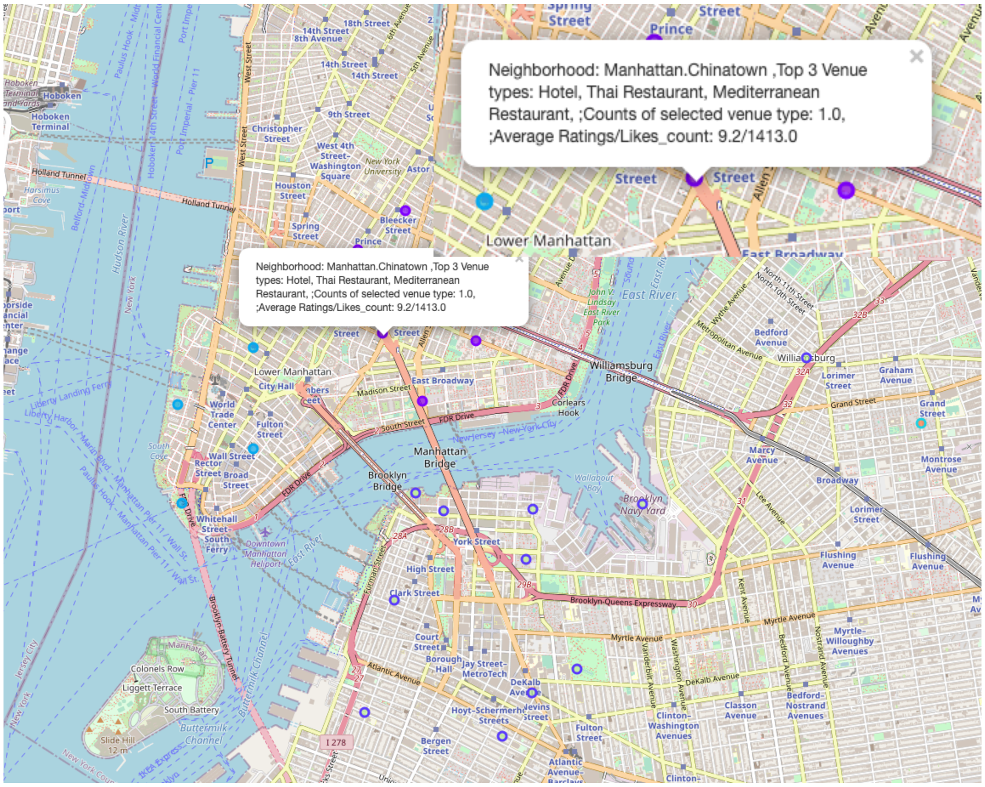
Results are from an illustration of using "cafe" as the designated venue type. After searching for venues in New York of given category and sorting by ratings, the returned results show as follows:



id	name	category	lat	lng	postalCode	rating	likes_count	ratingSignals	photos_count	Neighborhood	Borough
41044980f964a520750b1fe3	Cafe Mogador	Moroccan Restaurant	40.727277	-73.984505	10009	9.2	1413	1413	1413	Chinatown	Manhattan
5244bd0e11d2d511de3e244e	Russ & Daughters Café	Café	40.719515	-73.989724	10002	9.2	1138	1138	1138	Mill Basin	Brooklyn
40c10d00f964a520dd001fe3	Ruby's Café	Australian Restaurant	40.722292	-73.996248	10012	9.1	1117	1117	1117	Battery Park City	Manhattan
3fd66200f964a52004e61ee3	Café Habana	Cuban Restaurant	40.722796	-73.994217	10012	9.0	1565	1565	1565	Mill Island	Brooklyn
3fd66200f964a520efe81ee3	The River Café	American Restaurant	40.703754	-73.994834	11201	8.8	425	425	425	Fort Greene	Brooklyn
4ab27744f964a520486b20e3	Harry's Cafe and Steak	Steakhouse	40.704558	-74.009746	10004	8.5	282	282	282	Bergen Beach	Brooklyn
49bc236af964a5201b541fe3	Café Select	Swiss Restaurant	40.721610	-73.997549	10012	8.4	737	737	737	Battery Park City	Manhattan
4baabd4cf964a520c6833ae3	Café Gitane	French Restaurant	40.723159	-73.994732	10012	8.4	303	303	303	Mill Island	Brooklyn
3fd66200f964a520e3e51ee3	Fanelli Café	American Restaurant	40.724607	-73.998751	10012	8.1	365	365	365	Battery Park City	Manhattan
4c07ce55a9c076b0da733923	Inatteso Cafe Casano	Café	40.706335	-74.016457	10004	7.9	48	48	48	Lower East Side	Manhattan
5a4bc17bb8fd9d6ba9863351	GFG Bakery and Café	Bakery	40.710254	-74.005620	10038	7.4	11	11	11	Bergen Beach	Brooklyn
4a00e1bff964a520be701fe3	Mee Sum Cafe (美心)	Chinese Restaurant	40.714958	-73.998272	10013	6.8	10	10	10	Flatlands	Brooklyn
4aedda39f964a52089cf21e3	M Star Cafe 明星茶餐廳	Cha Chaan Teng	40.714200	-73.996594	10002	6.6	73	73	73	Flatlands	Brooklyn
4bc1e64b4cdfc9b6892b9521	Benvenuto Cafe Tribeca	Sandwich Place	40.719503	-74.010269	10013	6.6	56	56	56	Canarsie	Brooklyn
4a92c46bf964a520a01d20e3	Cafe Water	Deli / Bodega	40.705802	-74.006973	10005	6.3	23	23	23	Bergen Beach	Brooklyn
4a8c89c7f964a5206b0e20e3	Cafe Martin	Café	40.709819	-74.007239	10038	0.0	4	4	4	Bergen Beach	Brooklyn
5fd25ed820cfbd5181566ced	Spongies Cafe	Café	40.717970	-73.998846	10013	0.0	2	2	2	Marine Park	Brooklyn
5ba79773b9a5a8002c5b3751	Audrey Bakery & Cafe Inc.	Bakery	40.716484	-73.997593	10013	0.0	0	0	0	Flatlands	Brooklyn
4df5c8fd3151247c51b0c934	Chambers & Cafe	Wine Bar	40.713297	-74.003622	10007	0.0	0	0	0	Canarsie	Brooklyn
4bc7637f15a7ef3bd94e79da	Cafe Sea Port	Deli / Bodega	40.709595	-74.006271	10038	0.0	0	0	0	Bergen Beach	Brooklyn

Five of the best rating venues are located in Manhattan and 15 are located in Brooklyn. The last five venues located in Brooklyn lack ratings and have limited likes. To select target neighborhoods, we only use the top 4 distinct neighborhoods presented in the sorted results, which are Chinatown and Battery Park City in Manhattan along with Mill Basin and Fort Greene in Brooklyn. A K-means clustering is performed on neighborhoods in Manhattan as well as Brooklyn, separately. After clustering, only neighborhoods with the same cluster labels as the target neighborhoods are marked on the map as follows.

The popup label of each marker shows information as expected which includes the name and borough of the neighborhood, top 3 venue types around this neighborhood as well as the amount, average ratings and average likes_count of the designated venue type within range of this neighborhood. Neighborhoods with the same cluster labels show similar top 3 venue types as expected. The situation where counts and ratings are 0 will be further discussed in the next section. However, neighborhoods with 0 counts can be filtered out as an alternative presentation but such a



Neighborhood: Manhattan.Chinatown ,Top 3 Venue types: Hotel, Thai Restaurant, Mediterranean Restaurant, ;Counts of selected venue type: 1.0, ;Average Ratings/Likes_count: 9.2/1413.0

Neighborhood: Manhattan.Chinatown ,Top 3 Venue types: Hotel, Thai Restaurant, Mediterranean Restaurant, ;Counts of selected venue type: 1.0, ;Average Ratings/Likes_count: 9.2/1413.0

solution is not recommended in terms of business. Further implications of the result will be discussed in the next section.

Conclusion and Discussion

In this section, the following topics will be discussed.

1. Business implication of the results.
2. Situations with 0 counts.
3. Parameters that can be tuned for better results.

Business implication of the results.

As shown in the map, the program recommends numerous neighborhoods, in the form of 4 clusters located at 2 boroughs, to start a cafe business. Such recommendation is still considered to be general and a set of new factors need to be considered to finally make decisions. Such factors should include population, traffic, rent, etc. In the case of starting a cafe business, the style of the store and target market should also be considered as part of the decision process. That being said, instead of using the program as a "where to start" recommendation, the clustering analysis could serve better as a "where not to start" recommendation.

Situations with 0 counts

Another issue with the analysis is that certain neighborhoods show a count of 0 for the designated venue type, i.e. cafe. A few factors could contribute to the results. First, "cafe" is a narrower term compared to "restaurant". Thus, searching for a restaurant could decrease the chances of having a 0 count result. Another factor is the search radius. Currently, the illustration uses a radius of 5000 meters to search for venues of designated type near a neighborhood. The optimal radius to search varies borough by borough. 5000 meters could be a relatively large radius in Manhattan but not enough in Brooklyn. Thus, optimizing the radius according to boroughs can be a next step for this project. However, results with 0 count can still play an important role as such neighborhoods could serve as a potential opportunity in starting the business. Therefore, neighborhoods with 0 counts are kept in the results.

Parameters to be tuned

A set of parameters can be tuned in this analysis. One that has been mentioned above is the venue type. Switching venue categories between synonyms can potentially yield different results. Two radii are being used in the analysis. One is city radius, which represents the radius to be used to search for best rated venues within the city. The default is 10,000 meters for New York city. Another radius is neighborhood radius which is being used to search for venues of designated type within a range of neighborhoods. The next parameter can be tuned is top_n which represents the top n venues within the city that should be chosen to list the target neighborhoods and boroughs. In addition, the k clusters parameter can also be tuned to define the number of clusters. Tuning such parameters will yield different results to fit in various needs in using the program.