Restaurant Location Selection

----The Battle of Neighborhoods

Gary Liang

October 30, 2019

A. Introduction

A.1. Background & Problem Description

New York City, the most populous city in the United States, one of the greatest metropolises over the world, is a dream place for gourmet to seek delicious cuisine. Here, you may find all types of restaurant from each corner of the world. Its food culture includes an array of international cuisines influenced by the city's immigrant history. Central and Eastern European immigrants, especially Jewish immigrants from those regions, brought bagels, cheesecake, hot dogs, knishes, and delicatessens (or delis) to the city. Italian immigrants brought New York-style pizza and Italian cuisine into the city, while Jewish immigrants and Irish immigrants brought pastrami and corned beef, respectively. Chinese and other Asian restaurants, sandwich joints, trattorias, diners, and coffeehouses are ubiquitous throughout the city. Some 4,000 mobile food vendors licensed by the city, many immigrant-owned, have made Middle Eastern foods such as falafel and kebabs examples of modern New York street food. The city is home to "nearly one thousand of the finest and most diverse haute cuisine restaurants in the world", according to Michelin. As of 2019, there were 27,043 restaurants in the city, up from 24,865 in 2017[1].

As the figures tells, New York City attracts many to start their business in food industry. Before they take action, they need to find out where they would open it? What would they consider when selecting a location? I hope to explore regional characteristics of these restaurants and figure out is the neighborhood of restaurant an important for factor for success of a restaurant with sound analysis.

As I mentioned above, there are hundreds types of restaurants, it is impractical for me to run analysis for each type of restaurant. Based on the maximum total numbers among these restaurants, I choose Pizza Place for the following analysis. The analysis of other types of restaurants can be conducted with the same method. The conclusion on relationship between success and neighborhood of a restaurant can be a useful finding for those who plan to operate a restaurant.

A.2. Data Preparation

Data used in analysis are listed as below:

- Neighborhoods in New York City from Wikipedia. https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City I cleaned the data and reduced it to boroughs of NYC so that I can use it to find geological location for further venues analysis.
- Using **Geopy** to get geological location by address name
- Using **Forsquare API** to get the most common venues of given Borough of New York city.
- Using **Forsquare API** to get the venues record of given venues of New York city.

B. Methodology

I used **BeautifulSoup** to scrape boroughs from Wikipedia, and organize a table containing *Community Board*, *Area*, *Pop.Census*, *Neighborhoods* information of New York City.

And used Geopy to get geological location of each community board. (Because Geopy cannot recognize the address like 'Bronx CB 1', I use the first address in the list of Neighborhood of each community board. If it is still not found, the second address will be used.)

	Community Board	Area/km2	Pop.Census	Pop/km2	Neighborhoods	Latitude	Longitude
0	Bronx CB 1	7.17	91,497	12,761	Melrose, Mott Haven, Port Morris	40.824545	-73.910414
1	Bronx CB 2	5.54	52,246	9,792	Hunts Point, Longwood	40.812601	-73.884025
2	Bronx CB 3	4.07	79,762	19,598	Claremont, Concourse Village, Crotona Park, Mo...	40.839876	-73.907328
3	Bronx CB 4	5.28	146,441	27,735	Concourse, Highbridge	40.874217	-73.890410
4	Bronx CB 5	3.55	128,200	36,145	Fordham, Morris Heights, Mount Hope, Universit...	40.859267	-73.898469
5	Bronx CB 6	4.01	83,268	20,765	Bathgate, Belmont, East Tremont, West Farms	40.855278	-73.886389

I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as 100 venue and the radius 500 meter for each borough from their given latitude and longitude informations. Here is a head of the result, adding venue id, venue name, category, latitude and longitude informations from Forsquare API.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	id	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bronx CB 1	40.824545	-73.910414	5956be26123a195de6701c2b	Porto Salvo	40.823887	-73.912910	Italian Restaurant
1	Bronx CB 1	40.824545	-73.910414	4fa534cee4b0fed4819dc7d4	Perry Coffee Shop.	40.823433	-73.910940	Diner
2	Bronx CB 1	40.824545	-73.910414	5591837d498ee4167d48bb73	Cinco de Mayo	40.822674	-73.911592	Mexican Restaurant
3	Bronx CB 1	40.824545	-73.910414	553980e0498e68099069f65c	Old Bronx Courthouse	40.822894	-73.909565	Art Gallery
4	Bronx CB 1	40.824545	-73.910414	4d38c0949ca8236a1b12aee8	Popeyes Louisiana Kitchen	40.824605	-73.909819	Fried Chicken Joint
5	Bronx CB 1	40.824545	-73.910414	4c6d654f65eda093577b4ed0	McDonald's	40.825183	-73.908625	Fast Food Restaurant
6	Bronx CB 1	40.824545	-73.910414	4ca84cbc14c337040abcd23b	CTown Supermarkets	40.823888	-73.909905	Supermarket
7	Bronx CB 1	40.824545	-73.910414	4e011a3a45dd1e4999dceb4e	Duane Reade	40.823881	-73.909203	Pharmacy

It returns with 2555 records. I summarize venues by category. Among these 2555 records, Pizza Place counts 117 with the maximum total number. Therefore, I choose Pizza Place as an example of restaurants for further analysis.

I utilized the Foursquare API again by pizza places ID to explore detailed record of these pizza places. Select out Rating, Price, Likes, Photos, Tips into a dataframe. And drop those places without a rating.

	id	Name	Rating	Price	Likes	Photos	Tips
0	4e4cf9c1bd413c4cc66dae95	Linda's Pizzeria	0.0	1	0	4	0
1	4df7be37aeb7f7c3b5436f7a	Little Caesars Pizza	5.7	1	1	3	4
2	4ca3e96c5720b1f78d0936ef	Fratellis Pizza Cafe	6.8	1	3	3	5
3	4afac66af964a520a11822e3	Jerome's Pizza	7.5	1	6	7	8
4	4bf329ef2d62952165ec5f58	Best Italian Pizza	7.5	1	15	10	13
5	4cb0f72d75ebb60cba60c9ad	Domino's Pizza	6.3	1	2	3	3
6	4ca680d1b7106dcba0d35ea5	Susie's Pizza	0.0	1	3	2	0
7	513fd3c3e4b0a8eca5f8b719	Joey Pepperoni's Pizza	0.0	1	0	1	0
8	5a272952fe37404aed37cf4e	Bella Pizza	0.0	1	0	0	0
9	4aabd3e6f964a5204a5a20e3	Zero Otto Nove	9.1	3	231	144	76
10	4aee1220f964a520ced121e3	Full Moon Pizzeria	8.7	2	114	95	42

Then I tried to find correlation among these variables:

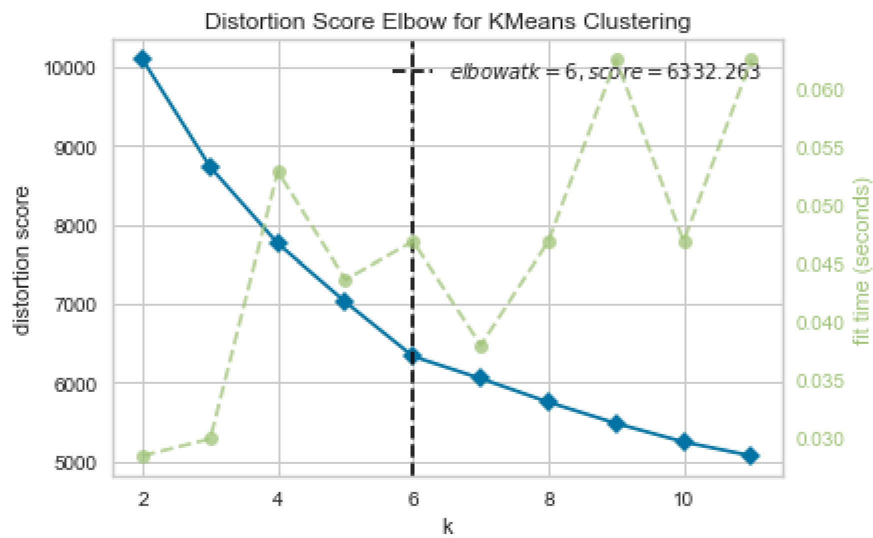
	Rating	Price	Likes	Photos	Tips
Rating	1.000000	0.468022	0.493203	0.450074	0.510075
Price	0.468022	1.000000	0.624174	0.528872	0.713245
Likes	0.493203	0.624174	1.000000	0.868937	0.892418
Photos	0.450074	0.528872	0.868937	1.000000	0.829932
Tips	0.510075	0.713245	0.892418	0.829932	1.000000

Showing from the correlation matrix, Likes, Photos and Tips are highly correlated to each other. But Likes is not highly related to Rating. Customers who click Likes for some specific reasons but give lower ratings to the general performance might cause this low correlation. Therefore I choose Rating to represent the restaurant. Rating is somewhat correlated to Price, which indicates that the price might not affect impressions of customers on that place significantly.

I utilized the Foursquare API centering these pizza places to explore their neighborhoods with a 500-meter radius.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	id	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Linda's Pizzeria	40.823803	-73.909060	5956be26123a195de6701c2b	Porto Salvo	40.823887	-73.912910	Italian Restaurant
1	Linda's Pizzeria	40.823803	-73.909060	4fa534cee4b0fed4819dc7d4	Perry Coffee Shop.	40.823433	-73.910940	Diner
2	Linda's Pizzeria	40.823803	-73.909060	4f0f4fe14fc6e3adf353a950	Blink Fitness	40.819543	-73.910554	Gym / Fitness Center
3	Linda's Pizzeria	40.823803	-73.909060	5591837d498ee4167d48bb73	Cinco de Mayo	40.822674	-73.911592	Mexican Restaurant
4	Linda's Pizzeria	40.823803	-73.909060	553980e0498e68099069f65c	Old Bronx Courthouse	40.822894	-73.909565	Art Gallery
5	Linda's Pizzeria	40.823803	-73.909060	542c31e1498e76106c273492	Blink Fitness St Ann's	40.819470	-73.910522	Gym
6	Linda's Pizzeria	40.823803	-73.909060	4d38c0949ca8236a1b12aee8	Popeyes Louisiana Kitchen	40.824605	-73.909819	Fried Chicken Joint

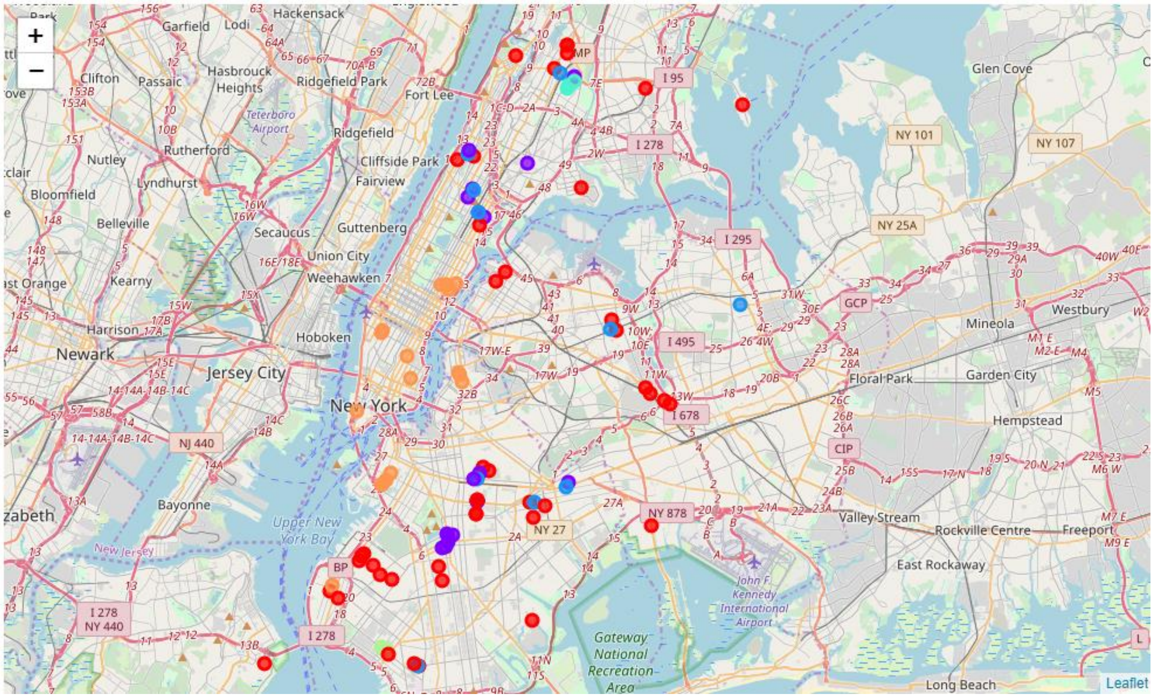
According to venues categories and numbers surrounding each pizza house, I use k-means to cluster pizza house into several groups. I use “elbow” method to help select the optimal number of clusters by fitting the model with a range of values for k. The “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. In the visualizer “elbow” k=6 is annotated with a dashed line.



I merged cluster labels of each pizza place with its geological location.

	PizzaPlace	ClusterLabels	id	Rating	Price	Likes	Photos	Tips	Venue Latitude	Venue Longitude
90	Sottocasa Pizzeria - Harlem	2	56e86a57cd1017cb53f3e8f9	9.4	1	87	39	29	40.805550	-73.947435
88	Saraghina	1	4a593de0f964a52015b91fe3	9.2	2	736	432	204	40.683590	-73.935340
63	Nick's Pizza	1	45ac11b0f964a5205b411fe3	9.1	2	191	143	94	40.718180	-73.840692
98	Zero Otto Nove	4	4aabd3e6f964a5204a5a20e3	9.1	3	231	144	76	40.854714	-73.888388
79	Posto	0	43c4bb46f964a5205a2d1fe3	9.0	2	346	146	121	40.734737	-73.983049
100	babbalucci	2	55a4456a498e29417f5c71a2	9.0	2	92	73	30	40.808875	-73.944796
47	La Villa Pizzeria	1	4b193722f964a52056d923e3	8.9	2	75	42	40	40.616867	-73.909858
73	Peppino's	1	4b58f8f2f964a5203f7628e3	8.8	1	32	24	16	40.629851	-74.028481
93	Vesuvio Pizzeria & Restaurant	0	4b3eaefff964a520bfa025e3	8.8	2	105	80	149	40.632580	-74.027096
23	F&F Pizzeria	0	5d56bcea6ac2400008de14d9	8.8	1	15	13	0	40.677342	-73.998162
27	Full Moon Pizzeria	4	4aee1220f964a520ced121e3	8.7	2	114	95	42	40.855506	-73.887557

Then I used **folium** to visualize distribution of these pizza places in NYC as below:



C. Results

Let's see if there is different in performance of these pizza places with different clustering labels:

```
ClusterLabels
0    8.30
1    7.45
2    7.25
3    6.40
4    8.40
5    7.80
Name: Rating, dtype: float64
```

Based on median, it seems there is difference. But is it significant? I ran one-way ANOVA analysis:

```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower   upper  reject
-----------------------------------------------------
     0      1   -0.871 0.0044 -1.5518 -0.1901   True
     0      2  -0.7924 0.0472 -1.5789 -0.0058   True
     0      3  -1.7924  0.001 -2.7425 -0.8422   True
     0      4   0.0176    0.9 -1.3073  1.3426  False
     0      5  -0.4157   0.89 -1.5478  0.7165  False
     1      2   0.0786    0.9 -0.5643  0.7216  False
     1      3  -0.9214 0.0218 -1.7566 -0.0861   True
     1      4   0.8886 0.3087 -0.3565  2.1337  False
     1      5   0.4553 0.7714 -0.5823  1.4929  False
     2      3     -1.0 0.0258 -1.9234 -0.0766   True
     2      4     0.81 0.4693 -0.4959  2.1159  False
     2      5   0.3767    0.9 -0.7331  1.4864  False
     3      4     1.81 0.0043  0.3995  3.2205   True
     3      5   1.3767 0.0192  0.1455  2.6079   True
     4      5  -0.4333    0.9 -1.9723  1.1057  False
-----------------------------------------------------
[0 1 2 3 4 5]
```

Showing in the table, the performances of these 6 types pizza places are significantly different. Places with cluster label 0 and 4 perform best while those with label 1,2 and 3 perform worst. Hence, different neighborhood might affect the impression of the customers on this restaurant. Restaurant owners may search for a similar location to start their business. So what is special in these locations around pizza places?

I select out pizza places with label 0 and 4, and drop categories with empty value, sort venue types by average numbers around pizza places in a descending way.

ClusterLabels	0	4
Coffee Shop	5.111111	0.2
Pizza Place	3.666667	5.8
Italian Restaurant	3.500000	16.2
Bar	2.388889	1.4
Café	2.333333	1.0
Sushi Restaurant	2.055556	0.0
Cocktail Bar	1.833333	0.0
Gym / Fitness Center	1.777778	0.4
Ice Cream Shop	1.777778	0.0
Bakery	1.722222	5.2

Pizza place as a classic light meal, normally is opened at a location where different light meals gather, like Cafe, bakery or bar. It explains why places with higher ratings gather around Manhattan. Does the pizza there truly taste better than other places'? Short time spent on tasting could be a reason that render the quality of ratings. Another interesting thing is that Italian restaurant numbers is apparently more than other types of restaurant, especially for pizza places with label 1. These places are in Bronx which is happen to be an Italian area. Pizza as a representative of Italian food seems meet the correct market there.

D. Conclusion

As a result, for those types of pizza place where providing quick services and moderate flavor with moderate price, they should consider locations in busy area and close to other light meals restaurant.

For those aiming to provide delicious pizza to most picky gourmet, they should open their pizza place at places close to their target customers or customers with certain background.

E. Discussion

As a recommendation to those who plan to operate restaurant, location selection is only one basic problem to think over. The analysis of this report assumes the type of restaurant is selected, for example, a pizza house. It can not solve the problem about whether a type restaurant is the most popular type and how much customer will visit every day. And as location suggestion, it offer a opportunity analysis but lack risk analysis, like the cost of the location and competition in that area.

Although in this report, it demonstrates the relations between location and ratings, but ratings might not reflect the operation status of the restaurant. Restaurant with a high rating could still be unprofitable, which is unsuccessful from business perspective. So the suggestion is relatively narrow. In order to suggest with more practical and profitable ideas, the relationship between customer reactions and financial performance should be evaluated.

With all these analyses done, the report finally become constructive for a restaurant owner in real business world.