The Battle of the Neighborhoods

June 26, 2019

Xueyue Xiao

1. Introduction: Business Problem

New York is one of my favorite cities in the world. There is so many kinds of restaurants, like Chinese food, American food, Portuguese cuisine, and etc. If you get tired of one kind, you can always find another brilliant one that suits your taste. In addition to a variety of types of food, restaurants of all levels can be found in NYC, from street food or breakfast truck to Michelin star restaurants. If you are a foodie, you will never get tired of this city. Therefore, I would like to choose NYC as the city to explore.

In this project I will try to help Mr. Panini to find an optimal location or block for his business. Mr. Panini is an Italian restaurant owner in California, and recently he would like to expand his business to NYC. He would like to open a new restaurant in Manhattan, but he doesn't know where should him to set up his new business. Following is my strategy for Mr. Panini:

1) Since Mr. Panini is new to NYC, I will first create a map of NYC and Manhattan to him. Given the map, he is supposed to have a big picture about what does NYC and Manhattan look like, how the venues distribute.
2) Mr.Panini is going to open an Italian restaurant. His main competitors are existed Italian restaurant. Therefore, I will list all the Italian restaurant in Manhattan with their name, coordinates, rating.
3) In order to show a clear picture, I will create a map to show how Italian restaurants are distributed in each neighborhood. 4) I will cluster the restaurants to narrow down the location and calculate the average rating.
4) By choosing the lowest average rating to be our target cluster, I will choose the neighborhood with most Italian restaurants. Comparison will make Mr.Panini's restaurant stand out.
5) Finally, the location will be determined according to the Italian restaurants in this neighborhood.

2. Data description

The source of data is from New York data json file given by the course and from Foursquare API.

1) **New York data json** file provides me with the borough, neighborhood, latitude and longitude of each neighborhood in NYC. This data source allows me to create the map of NYC and Manhattan.
2) **Foursquare API** allows me to get the Venue in each neighborhood, its latitude, longitude, category, ID and rating. This data source allows me to obtain the list of Italian restaurants, conduct calculation of the rating, and create the map with Italian restaurants.

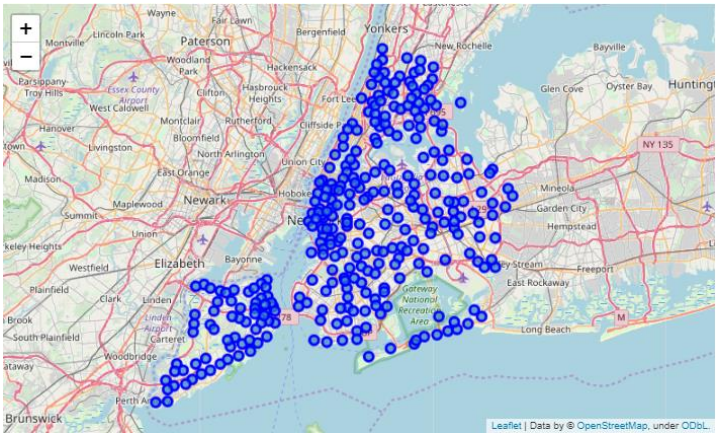
3. Methodology

1) Data Cleaning

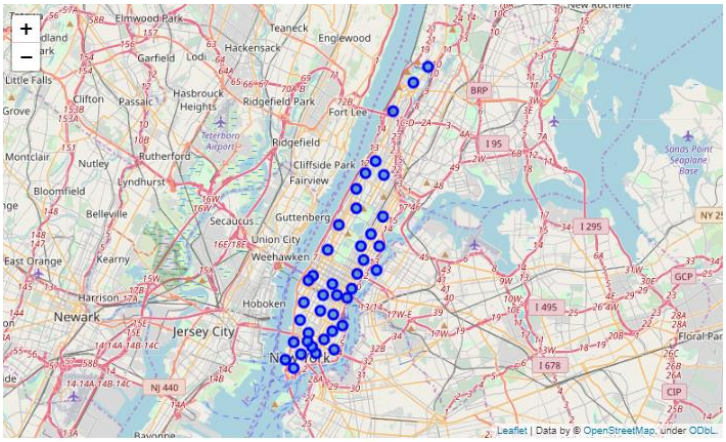
The original data is a json file which cannot be used to conduct data analysis and data visualization. I extracted the “features” part which include coordinates, properties name, borough and etc. Then, a DataFrame named *neighborhoods* is created with data from the json file. The columns includes **Borough, Neighborhood, Latitude, Longitude**. There are 305 rows and 4 columns, including 5 boroughs. Then, I filtered *neighborhoods* DataFrame to get a new DataFrame *manhattan_data* which only has data of Manhattan.

2) Data Visulization

I used python folium library to create maps of NYC and Manhattan with neighborhoods superimposed on top. By using the latitude and longitude of each neighborhood, I show the map of NYC:



The map of Manhattan:



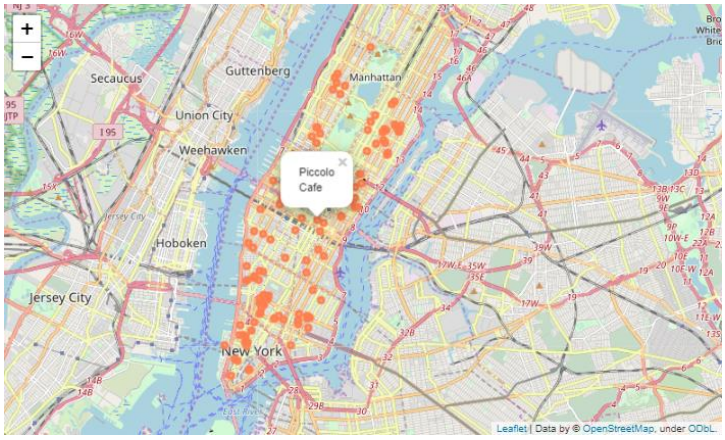
3) Foursquare API

Foursquare API allows me to explore all the venues in Manhattan. I utilized it to obtain the **venue name, latitude, longitude, venue category, venue ID** and **rating** in each neighborhood and transformed them into a DataFrame called *nearby_venues*. I filtered the DataFrame to select the venue category with Italian Restaurant. There are **125** Italian restaurants in Manhattan, all of them have ratings. A part of table is shown below:

	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID	Rating
0	Chinatown	Bacaro	40.714468	-73.991589	Italian Restaurant	472a027af964a520ea4b1fe3	8.2
1	Washington Heights	Saggio Restaurant	40.851423	-73.939761	Italian Restaurant	4d21107c6e8c37042b58ff9f	8.6
2	Hamilton Heights	Fumo	40.821412	-73.950499	Italian Restaurant	56d8e01d498ef1500ae7fbfe	8.9
3	Manhattanville	Pisticci Ristorante	40.814015	-73.960266	Italian Restaurant	457f1183f964a5204b3f1fe3	9.1
4	Manhattanville	Bettolona	40.814084	-73.959574	Italian Restaurant	4c956003f7cfa1cd2e2ebd15	8.0
5	Upper East Side	Sant Ambroeus Madison Ave	40.775328	-73.962819	Italian Restaurant	4a22d7f9f964a520977d1fe3	8.7
6	Upper East Side	Antonucci	40.775711	-73.956607	Italian Restaurant	4b3bedfaf964a5209e7e25e3	8.5

After I have a DataFrame of all Italian restaurants, I utilized Foursquare API again to get the rating of each Italian restaurant in Manhattan. Transforming and combining data from Foursquare API, I have a new DataFrame named *Italian_merged.*

By using **data visualization** in step 2) again, I created a map of Manhattan with Italian restaurants superimposed on top:

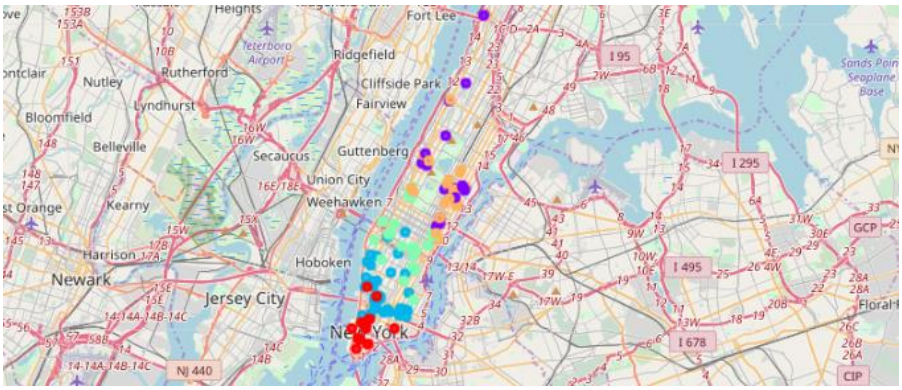


4) *k*-means algorithm

According to the venue coordinates and rating, I used unsupervised learning l-means algorithm to cluster the venues. k-means is one of the most common clustering algorithm which is very easy to understand. I set k as 5 to get appropriate amount of clusters. Merging the label list to the existed DataFram *Italian_merged.* The new DataFrame is shown below:

	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID	Rating	Label
0	Chinatown	Bacaro	40.714468	-73.991589	Italian Restaurant	472a027af964a520ea4b1fe3	8.2	0
1	Washington Heights	Saggio Restaurant	40.851423	-73.939761	Italian Restaurant	4d21107c6e8c37042b58ff9f	8.6	1
2	Hamilton Heights	Fumo	40.821412	-73.950499	Italian Restaurant	56d8e01d498ef1500ae7fbfe	8.9	1
3	Manhattanville	Pisticci Ristorante	40.814015	-73.960266	Italian Restaurant	457f1183f964a5204b3f1fe3	9.1	1
4	Manhattanville	Bettolona	40.814084	-73.959574	Italian Restaurant	4c956003f7cfa1cd2e2ebd15	8.0	4
5	Upper East Side	Sant Ambroeus Madison Ave	40.775328	-73.962819	Italian Restaurant	4a22d7f9f964a520977d1fe3	8.7	1
6	Upper East Side	Antonucci	40.775711	-73.956607	Italian Restaurant	4b3bedfaf964a5209e7e25e3	8.5	1

By using **data visualization** again, I created a map of Manhattan with five clusters superimposed on top:



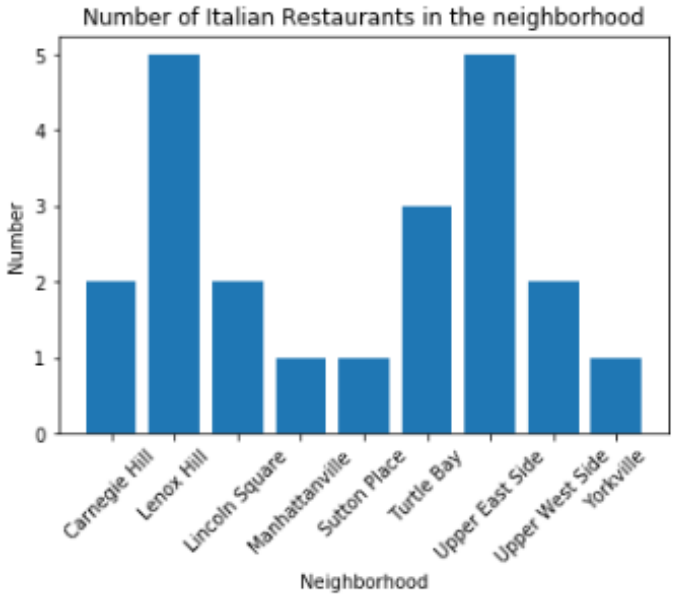
The first three clusters is highly like formed by location, while the rest two clusters looks to be formed according to rating. I extract five clusters and calculate average rating for each of them. It turns out that cluster five has the lowest rating 7.81, which is shown in purple in the map above.

5) Determine the location

I focused on the cluster five and created a DataFrame *cluster* to show all the Italian restaurants in the cluster five:

	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID	Rating
4	Manhattanville	Bettolona	40.814084	-73.959574	Italian Restaurant	4c956003f7cfa1cd2e2ebd15	8.0
7	Upper East Side	Parma Restaurant	40.774930	-73.957210	Italian Restaurant	4ab972d6f964a520697f20e3	8.1
8	Upper East Side	Sistina	40.777716	-73.961668	Italian Restaurant	57d9ddfe498e8a66d21ce8bb	7.5
10	Upper East Side	Caffe Grazie	40.779538	-73.959986	Italian Restaurant	4b23ee1ef964a520255d24e3	8.2
11	Upper East Side	Quattro Gatti Ristorante	40.775449	-73.955977	Italian Restaurant	4b5260e2f964a520587a27e3	7.8

Then, I divided the restaurants based on the neighborhood which can help me narrow down the area. Using bar chart, I can show the number of Italian restaurants in each neighborhood:



Since my strategy is first locate the cluster with lowest rate, and then choose a neighborhood with the highest number of restaurants. Now, I focus on Lenox Hill and Upper East Side. Both of them have 5 Italian restaurants. Since they have the same number of Italian restaurant, then I chose the neighborhood with lower average rating, which turns out to be Upper East Side.

By using **data visualization** again, I show the five Italian restaurants on the map:

Upper East Side is going to be the neighborhood for Mr. Panini's new restaurant. The restaurant should not be too close to other restaurants but still not far from the others. Therefore, the center of five restaurants is a good choice.

4. Result

Through calculation, the latitude and longitude is (40.7765, -73.9584). I used google map to obtain address of this coordinates, which is 1187 Lexington Avenue, New York,

NY 10028, United States of America. This location is the optimal one through this project for Mr. Panini.

5. Discussion

The location I would like to provide with Mr. Panini is 1187 Lexington Avenue. Main reasons are:

1) Manhattan is the most popular borough in NYC, with the highest density and a variety of restaurants. Mr. Panini's restaurant can benefit from large population and demand.
2) I separate Manhattan into five clusters using k-mean method to figure out the average rating in each cluster. The restaurants at this Neighborhood have average lowest rating. When people are surrounded by low rating restaurants, people would tend to choose the new one. As long as the new restaurant can offer tasty dishes, it is highly likely that the new one can receive higher rating.
3) When I look inside the cluster, I have to decide to build in which neighborhood. The method I used is to compare the number of restaurants in each neighborhood. Therefore, I narrow down Lenox Hill and Upper East Side because they have more restaurants than other neighborhoods. However, Upper East Side has lower rating. When people conduct comparing, Mr. Panini's restaurant can stand out.

It is undoubtful that other factors, such as other type of restaurants, the price, population distribution and etc, should be considered. However, according to the methods I have learned from this course, 1187 Lexington Avenue would be the one I recommend. This address now is a La Marqueza Beauty Spa.

6. Conclusion

The purpose of this project is to help Mr. Panini find an optimal location to move his business from California to New York. Thanks for Cousera capstone project provides me with NYC data file so I can easily get the coordinates of neighborhood and borough in NYC. Through data from Foursquare API, I am able to cluster the Italian restaurants into five cluster and calculate the average rating. Clustering helps me to narrow down the area. This project determine the optimal location only based on the competitors' category, locations and ratings.

For Mr. Panini, he is supposed to consider more factors include price, other type of restaurants, food quality, menu, population density, real estate availability and etc.