The Battle of Neighborhoods Project

1. Introduction

Background:

The Project is the fulfillment of the Coursera IBM Data Science certification course. The project requirements are to leverage the "Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data and other related data to solve."

My initial intention is to design this project as a consulting investigation for the catering industry investors by data exploration technologies. I select New York City as the city which we would explore and investigate. Since New York City is one of the most popular travel destinations in the world, and also it has very large kinds of different cuisines in the city.

Problem Definition:

Suppose a Chinese chain restaurant (SWJ) wish to expand its business to North America. SWJ is a very popular brand in China and it has more than 1000 brand restaurants in China. We select Queens, New York City as the first landing location for SWJ in North America. According to its brand awareness in China, so SWJ's target customers are the tourists visiting New York City in China. And the location of new-open restaurant would be near tourism attractions, shopping stores and Chinese-like cuisine restaurants. Because for tourists from China, visiting attraction venues, shopping and dining are the most parts of the whole trip.

So, the mission can be the statement:

To find a good location to open a Chinese restaurant in Queens, New York City, which the neighborhood has the most of attractions, shops and Chinese-like cuisine restaurants.

2. Data Sources

Data Source 1 - Neighborhood Data

Queens, which is the target boroughs of this investigation within New York City, has 80 neighborhoods. We first need to obtain a list of all the locations of the neighborhoods in Queens. This information is available on the following web address: https://geo.nyu.edu/catalog/nyu_2451_34572

The data is from a JSON File, the structure is as the following example including neighborhood name, borough name and other information.


```
{'geometry': {'coordinates': [-73.84720052054902, 40.89470517661],
  'type': 'Point'},
 'geometry_name': 'geom',
 'id': 'nyu_2451_34572.1',
 'properties': {'annoangle': 0.0,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661],
  'borough': 'Bronx',
  'name': 'Wakefield',
  'stacked': 1},
 'type': 'Feature'}
```


Data Source 2 - Geographical Coordinates

Geographical coordinates for each neighborhood will be obtained with the aid of GEOPY Library. Each neighborhood will be assigned a latitude and longitude coordinate.

Combined with neighborhood data and geographical data, the data source should be structured like the following example:

	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

Data Source 3 - Venue categories

We will use the Foursquare API to retrieve venues, using the coordinates obtained in Data Source 2 above. We shall further obtain a list using Foursquare API for related venues such as attractions, shops and restaurants in Queens.

The example Foursquare API which I will use to collect this venue data is as the following:

```
url =
'https://api.foursquare.com/v2/venues/explore?&client_id={}&cli
ent_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(

    CLIENT_ID, CLIENT_SECRET,VERSION,

    neighborhood_latitude,

    neighborhood_longitude,

    radius,

    LIMIT)
```

After retrieving data from Foursquare and transferred into dataframe, the data source should be structured as the following example:

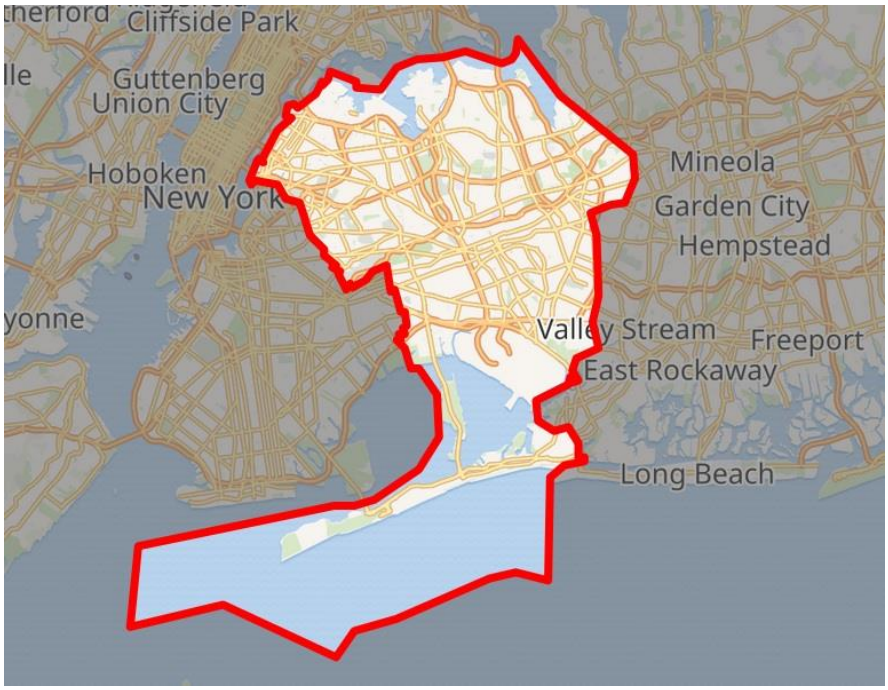
	name	categories	lat	lng
0	Favela Grill	Brazilian Restaurant	40.767348	-73.917897
1	Orange Blossom	Gourmet Shop	40.769856	-73.917012
2	Titan Foods Inc.	Gourmet Shop	40.769198	-73.919253
3	CrossFit Queens	Gym	40.769404	-73.918977
4	Simply Fit Astoria	Gym	40.769114	-73.912403

3. Methodology and Approach

Based on the business problem and Data Sources described above, I decide to leverage data exploration, data preparation, data visualization and machine learning technologies and tools to investigate the neighborhood in Queens, New York City. Our target is to find a good location to open a Chinese chain restaurant.

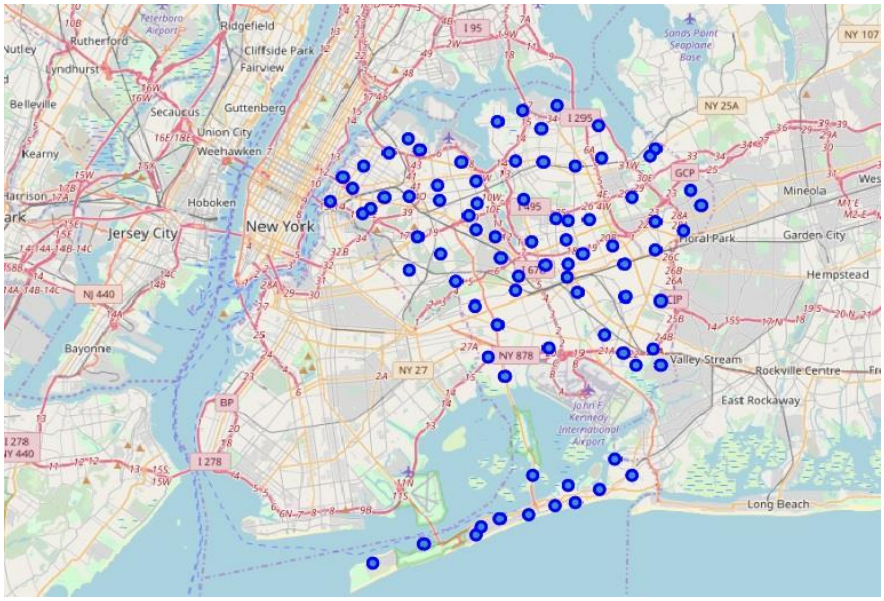
About Queens, New York City:

Queens is the easternmost of the five boroughs of New York City. It is the largest borough geographically. the borough of Queens is the second largest in population (after Brooklyn), with an estimated 2,358,582 residents in 2017. Queens County also is the second most populous county in the U.S. state of New York, behind Brooklyn.



From https://en.wikipedia.org/wiki/Queens

There’re totally 80 neighborhoods in Queens. We can visualize the distribution of the neighborhoods in the map, as well get the geographical data from public data source.



Also, I leverage Foursquare API to retrieve venue data in Queens and transfer the data into a dataframe, which include neighborhood, latitude, longitude, venue, venue category.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Astoria	40.768509	-73.915654	Favela Grill	40.767348	-73.917897	Brazilian Restaurant
1	Astoria	40.768509	-73.915654	Orange Blossom	40.769856	-73.917012	Gourmet Shop
2	Astoria	40.768509	-73.915654	Titan Foods Inc.	40.769198	-73.919253	Gourmet Shop
3	Astoria	40.768509	-73.915654	CrossFit Queens	40.769404	-73.918977	Gym
4	Astoria	40.768509	-73.915654	Simply Fit Astoria	40.769114	-73.912403	Gym

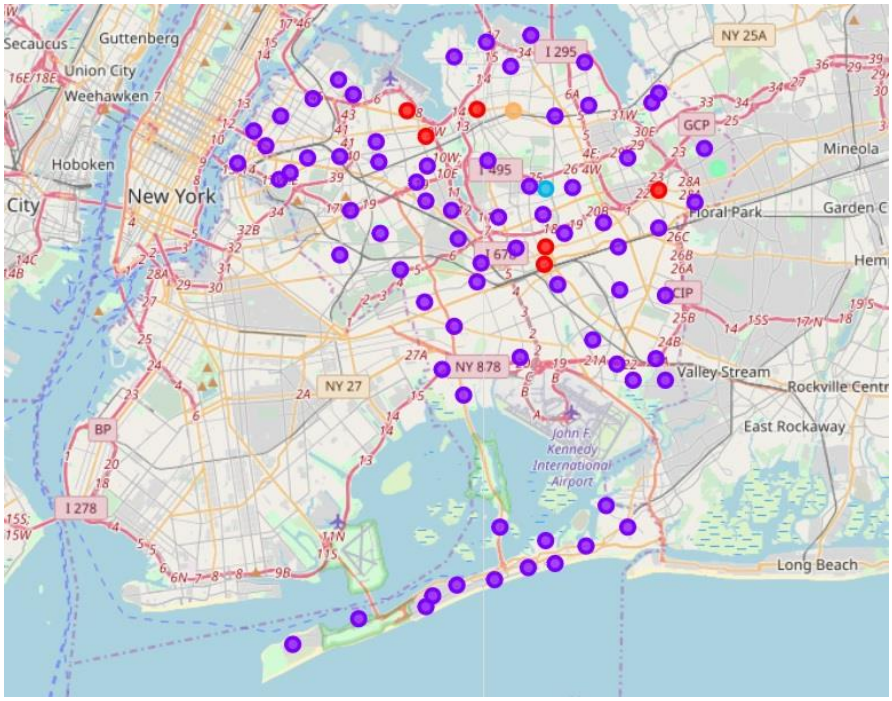
Since the new-open restaurant mainly focus on tourists from China, I selected venue categories related to tourism, attractions, shopping and restaurants. The following is the venue category list:

'Art Museum', 'Arts & Entertainment', 'Asian Restaurant', 'Beach', 'Cantonese Restaurant', 'Chinese Restaurant', 'Clothing Store', 'Concert Hall', 'Department Store', 'Dim Sum Restaurant', 'Diner', 'Fast Food Restaurant', 'General Entertainment', 'Gift Shop', 'Hotel', 'Japanese Restaurant', 'Korean Restaurant', "Men's Store", 'Monument / Landmark', 'Museum', 'Park', 'Restaurant', 'Seafood Restaurant', 'Shanghai Restaurant', 'Shoe Store', 'Shopping Mall', 'Sushi Restaurant', 'Taiwanese Restaurant', 'Theater'

So in the next step, I summarized and normalized the venue quantity grouped by neighborhoods with the important venue category list.

	Neighborhood	Art Museum	Arts & Entertainment	Asian Restaurant	Beach	Cantonese Restaurant	Chinese Restaurant	Clothing Store	Concert Hall	Department Store	Dim Sum Restaurant	Diner	Fast Food Restaurant	General Entertainment	Gift Shop	Hotel
0	Arverne	0.000000	0.000000	0.000000	0.058824	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
1	Astoria	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.010000	0.000000	0.000000	0.000000	0.000000
2	Astoria Heights	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
3	Auburndale	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.052632	0.000000	0.000000	0.000000
4	Bay Terrace	0.000000	0.000000	0.024390	0.000000	0.000000	0.000000	0.121951	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.024390	0.000000
5	Bayside	0.000000	0.000000	0.014706	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
6	Bayswater	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
7	Beechhurst	0.000000	0.000000	0.000000	0.000000	0.000000	0.058824	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
8	Bellaire	0.000000	0.000000	0.000000	0.000000	0.000000	0.166667	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000

In the machine learning phase, I leveraged clustering(K-means) to cluster the neighborhoods into 5 clusters based on the attributes of venues. The following is the visualization of the 5 clusters:



4. Results

As a result, it is found that cluster 1 is the cluster in which the neighborhoods are good locations to open a Chinese chain restaurant.

	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
9	Flushing	Chinese Restaurant	Korean Restaurant	Asian Restaurant	Hotel	Sushi Restaurant	Seafood Restaurant	Cantonese Restaurant	Taiwanese Restaurant	Fast Food Restaurant	Arts & Entertainment
12	East Elmhurst	Chinese Restaurant	Theater	Gift Shop	Arts & Entertainment	Asian Restaurant	Beach	Cantonese Restaurant	Clothing Store	Concert Hall	Department Store
31	Jamaica Center	Clothing Store	Chinese Restaurant	Concert Hall	Department Store	Men's Store	Japanese Restaurant	Theater	General Entertainment	Arts & Entertainment	Asian Restaurant
65	Bellaire	Chinese Restaurant	Theater	Gift Shop	Arts & Entertainment	Asian Restaurant	Beach	Cantonese Restaurant	Clothing Store	Concert Hall	Department Store
66	North Corona	Hotel	Museum	Fast Food Restaurant	Gift Shop	Arts & Entertainment	Asian Restaurant	Beach	Cantonese Restaurant	Chinese Restaurant	Clothing Store
68	Jamaica Hills	Fast Food Restaurant	Shopping Mall	Asian Restaurant	Seafood Restaurant	Chinese Restaurant	Theater	General Entertainment	Arts & Entertainment	Beach	Cantonese Restaurant

This cluster contains 6 neighborhoods: **Flushing, East Elmhurst, Jamaica Center, Bellaire, North Corona, Jamaica Hills.** We can find that all of these 6 neighborhoods have very similar popular venues like Chinese Restaurants, Clothing Store, Theater, Museum, Asian Restaurants.

So, our result of this investigation is to suggest SWJ to open its new restaurant in these 6 neighborhoods.

5. Discussion

- The clustering result reflect all cluster 3,4,5 have only one neighborhood, different K parameter may lead to different result
- In cluster 1, the first three common venues reflect that the neighborhoods have a good circumstance combined with Asian cuisine, attractions and shops.
- There’re two Jamaica culture neighborhoods in cluster 1: Jamaica Center and Jamaica Hills. We can see Asian food is welcome in this area.

6. Conclusion

As a conclusion, I investigated the neighborhoods data in Queens, New York City, with Foursquare venue categories. By using data preparation, exploration, visualization and machine learning, I finally find a group of neighborhoods as the good location to open a Chinese chain restaurant (SWJ). That’s a fulfillment and practice of Data Science course.