Coursera Data Science Project Report

Finding an ideal location to open a Chinese Restaurant in NYC

By: Qadir Ali Danish

1. Introduction

1.1 Problem Description and the Background

Description: New York City is the most populous city in the US, which has been described as the cultural, financial, and media capital 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. The New York City Department of Health and Mental Hygiene assigns letter grades to the city's restaurants based upon their inspection results. As of 2019, there were 27,043 restaurants in the city, up from 24,865 in 2017[1].

Problem: As per the study, New York is one of the ideal places for gourmet to seek delicious cuisine and also a good place for those who want to start their food business. As we all know New York is home to the largest ethnic Chinese population outside of Asia, with multiple distinct Chinatowns across the city [1]. So, NYC is a good place for people who want run a Chinese Restaurant. Before people take action, they need to know where they open it? By exploring the characteristics of Chinese restaurants, I hope to figure out whether there is a pattern in the distribution of the restaurants, so that we can give some advices for those who want to start a new Chinese Restaurant.

[1]. https://en.wikipedia.org/wiki/New_York_City#Cuisine

1.2 Data and Solution

Data: NYC Dataset Link: https://geo.nyu.edu/catalog/nyu_2451_34572

	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

Fig 1.2 – NYC Dataset

How to use the data to solve the problem?

- Using Geopy to get the geological location by address name.
- Using Foursquare API to get Chinese restaurants' category list of New York City
- Using Foursquare API to get Chinese restaurants' records of New York City
- Using Heatmap to show the density of Chinese restaurants in each neighborhood
- Using Pandas corr() function to get the standard correlation coefficient of different variables in restaurant detail records
- Using K-means Machine Learning algorithm to cluster the restaurant data

And the Data get by Foursquare API is as following:

	Venue Id	Name	Rating	Price	Likes	Photos	Tips
0	4dabc3dc93a04642f09ccabd	Xing Lung Chinese Restaurant	7.0	1.0	6	0	0
1	4b9d6b45f964a52078ab36e3	Mr. Q's Chinese Restaurant	7.7	1.0	9	11	10
2	4e4d0387bd413c4cc66dfd72	Hung Hing Chinese Restaurant	6.0	1.0	5	1	2
3	4e2e08021838f1c552b6b8eb	Choi Yuan - Chinese Restaurant	6.3	1.0	6	4	8
4	566f33e7498e44c2501bda81	Panda Express	7.6	1.0	12	14	2
5	4bbe0e0407809521db5bdb91	Green Dragon	7.6	2.0	10	16	7
6	4b89b62df964a520ff4c32e3	Sabor Latino Seafood Restaurant	7.0	2.0	5	12	8
7	4c28cca7ed0ac9b6b2c160aa	New China	6.5	1.0	2	3	4
8	4ca4fda4d971b1f77da5f2e0	Golden Phoenix Chinese Restaurant	5.7	1.0	4	1	1
9	4d9a695ee5fd6ea8e3096df5	Lucky House Chinese Restaurant	7.4	1.0	5	0	2

Fig 1.3 – Chinese Restaurants Details Record in NYC

Venue Id	Rating	Price	Likes	Photos	Tips	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
bc3dc93a04642f09ccabd	7.0	1.0	6	0	0	Bronx	Eastchester	40.887556	-73.827806	Xing Lung Chinese Restaurant	40.888785	-73.831226	Chinese Restaurant
d6b45f964a52078ab36e3	7.7	1.0	9	11	10	Bronx	Pelham Parkway	40.857413	-73.854756	Mr. Q's Chinese Restaurant	40.855790	-73.855455	Chinese Restaurant
d0387bd413c4cc66dfd72	6.0	1.0	5	1	2	Bronx	Bedford Park	40.870185	-73.885512	Hung Hing Chinese Restaurant	40.871181	-73.886759	Chinese Restaurant
e08021838f1c552b6b8eb	6.3	1.0	6	4	8	Bronx	Bedford Park	40.870185	-73.885512	Choi Yuan - Chinese Restaurant	40.873078	-73.889086	Chinese Restaurant
f33e7498e44c2501bda81	7.6	1.0	12	14	2	Bronx	Fordham	40.860997	-73.896427	Panda Express	40.863001	-73.900894	Chinese Restaurant

Fig 1.4 – Chinese Restaurants grouped by Neighborhood

2. Methodology

2.1 Download and Explore New York City Dataset

In order to segment the neighborhoods of New York City, a dataset is required that contains the 5 boroughs and the neighborhoods, that exist in each borough, with respective latitude and longitude coordinates. This dataset is downloaded using the mentioned URL.

Once the .json file is downloaded, it is analyzed to understand the structure of the file. A python dictionary is returned by the URL and all the relevant data is found to be in the features key, which is basically a list of the neighborhoods. The dictionary is transformed, into a panda Dataframe, by looping through the data and filling the Dataframe rows one at a time.

As a result, a Dataframe is created with Borough, Neighborhood, Latitude and Longitude details of the New York City's neighborhoods.

	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

Fig 2.1 – NYC Dataset

Upon analysis, it is found that the Dataframe consists of 5 boroughs and 306 neighborhoods. And Geopy library is used to get the latitude and longitude values of New York City, which was returned to be Latitude: 40.7127281, Longitude: -74.0060152. Then neighborhoods Dataframe is used to visualize by using folium lib, the result is as following.

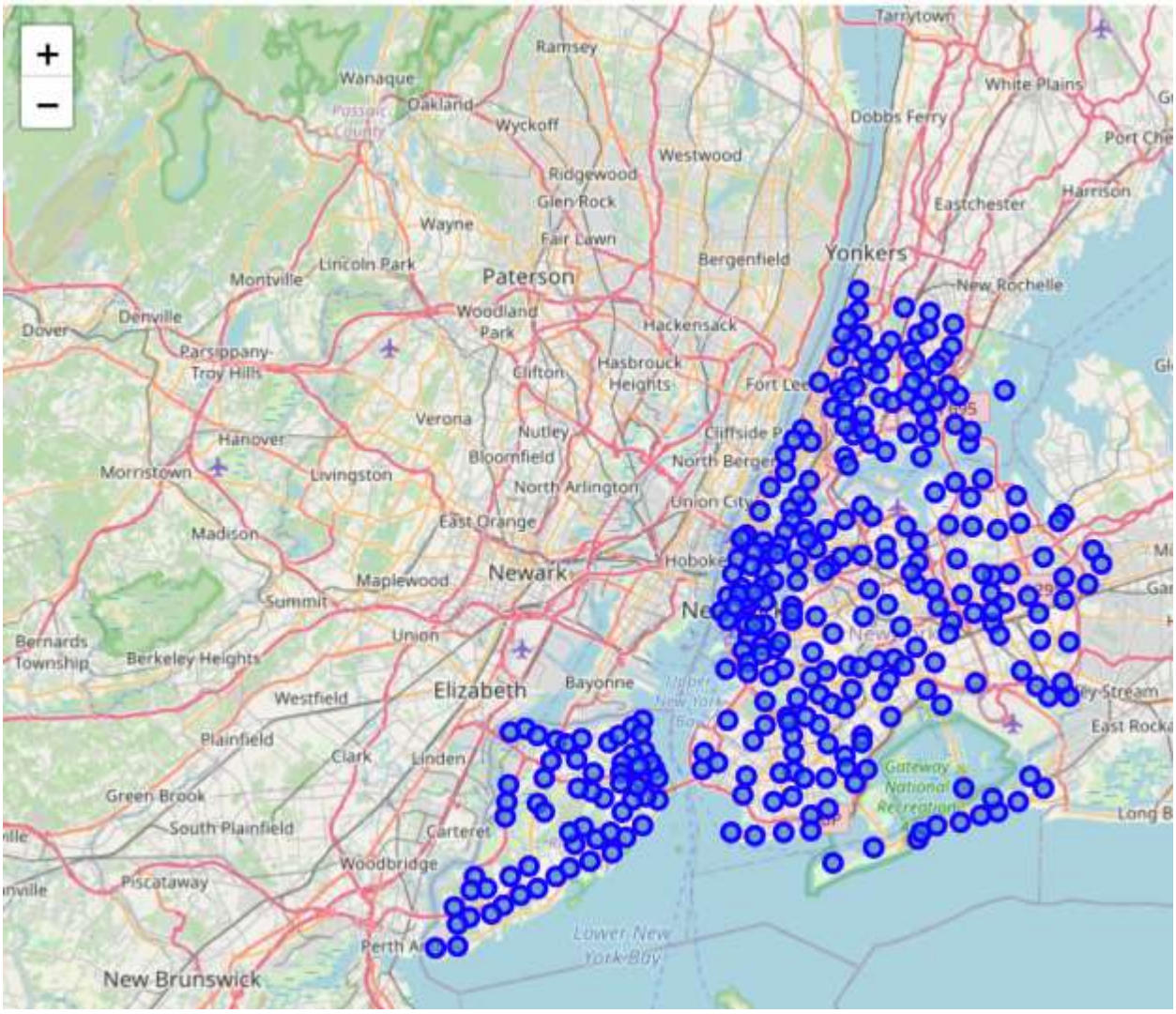

Fig 2.2 – Restaurants' locations

2.2 RESTful API Calls to Foursquare

The Foursquare API is used to explore the neighborhoods and segment them. To access the API, CLIENT_ID', 'CLIENT_SECRET' and 'VERSION' is defined. There are many endpoints available on Foursquare for various GET requests. But, to explore the Chinese restaurants in each neighborhood, it is required that all the venues extracted are from 'Chinese Restaurant' category. Foursquare Venue Category Hierarchy is retrieved and returned request is further analyzed. At first we need get all 'Chinese restaurant' categories, so I create a function call getCNRestaurantIds() to get all the 'Chinese Restaurant' categories.

```
# This func is using to get the detail records of each Chinese Restaurant
def getDetailRecords(venue_ids):
    valid =0
    count = 0
    like_missing = 0
    erro = 0
    records = []
    print('Obtaining Restaurants detail datas: ', end='')
    for id in venue_ids:
        # Create the API request URL
        url = 'https://api.foursquare.com/v2/venues/{}?client_id={}&client_secret={}&v={}'.format(
            id,
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION)

        # make the GET request
        try:
            result = requests.get(url).json()['response']['venue']
        except:
                erro +=1
                #print("erro {} data".format(erro))
                continue
        # choosing the useful info
        # Note: Since some records don't have some attributes, we have to handle this exception
        try:
            rating =  result['rating']
        except:
            # Drop no rating data
```

Fig 2.3 – Restful Function

Following is the result:

	id	name
0	52af3a5e3cf9994f4e043bea	Anhui Restaurant
1	52af3a723cf9994f4e043bec	Beijing Restaurant
2	52af3a7c3cf9994f4e043bed	Cantonese Restaurant
3	58daa1558bbb0b01f18ec1d3	Cha Chaan Teng
4	52af3a673cf9994f4e043beb	Chinese Aristocrat Restaurant
5	52af3a903cf9994f4e043bee	Chinese Breakfast Place
6	4bf58dd8d48988d1f5931735	Dim Sum Restaurant
7	52af3a9f3cf9994f4e043bef	Dongbei Restaurant

Fig 2.4 – Chinese Restaurant's Categories

Then I define another function to and pass 'Chinese Restaurant' categories List, radius = 500 meters, and every neighborhood' latitude and longitude values to get the total Chinese restaurants

in New York. Since the radius = 500 so there must be some duplicate data, we must drop these data. We got 1370 restaurants The results are as following.

Venue Id	Rating	Price	Likes	Photos	Tips	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
ibc3dc93a04642f09ccabd	7.0	1.0	6	0	0	Bronx	Eastchester	40.887556	-73.827806	Xing Lung Chinese Restaurant	40.888785	-73.831226	Chinese Restaurant
d6b45f964a52078ab36e3	7.7	1.0	9	11	10	Bronx	Pelham Parkway	40.857413	-73.854756	Mr. Q's Chinese Restaurant	40.855790	-73.855455	Chinese Restaurant
d0387bd413c4cc66dfd72	6.0	1.0	5	1	2	Bronx	Bedford Park	40.870185	-73.885512	Hung Hing Chinese Restaurant	40.871181	-73.886759	Chinese Restaurant
e08021838f1c552b6b8eb	6.3	1.0	6	4	8	Bronx	Bedford Park	40.870185	-73.885512	Choi Yuan - Chinese Restaurant	40.873078	-73.889086	Chinese Restaurant
f33e7498e44c2501bda81	7.6	1.0	12	14	2	Bronx	Fordham	40.860997	-73.896427	Panda Express	40.863001	-73.900894	Chinese Restaurant

Fig 2.5 – Chinese Restaurants

Now, we got all the Chinese restaurants, the next step we have to get the detail records of each restaurants including Rating, Price, Likes, Photos, Tips. Then putting these data into Dataframe and dropping those data without a rating. Finally, I get 515 data, the result is as following.

	Venue Id	Name	Rating	Price	Likes	Photos	Tips
0	4dabc3dc93a04642f09ccabd	Xing Lung Chinese Restaurant	7.0	1.0	6	0	0
1	4b9d6b45f964a52078ab36e3	Mr. Q's Chinese Restaurant	7.7	1.0	9	11	10
2	4e4d0387bd413c4cc66dfd72	Hung Hing Chinese Restaurant	6.0	1.0	5	1	2
3	4e2e08021838f1c552b6b8eb	Choi Yuan - Chinese Restaurant	6.3	1.0	6	4	8
4	566f33e7498e44c2501bda81	Panda Express	7.6	1.0	12	14	2
5	4bbe0e0407809521db5bdb91	Green Dragon	7.6	2.0	10	16	7
6	4b89b62df964a520ff4c32e3	Sabor Latino Seafood Restaurant	7.0	2.0	5	12	8
7	4c28cca7ed0ac9b6b2c160aa	New China	6.5	1.0	2	3	4
8	4ca4fda4d971b1f77da5f2e0	Golden Phoenix Chinese Restaurant	5.7	1.0	4	1	1
9	4d9a695ee5fd6ea8e3096df5	Lucky House Chinese Restaurant	7.4	1.0	5	0	2

Fig 2.6 – Chinese Restaurant's Details

Next, because there are some missing values in column 'Price', I use the mean price of the same category to replace the missing values. Then I try to find correlations among these variables:

	Rating	Price	Likes	Photos	Tips
Rating	1.000000	0.137413	0.462308	0.401613	0.413512
Price	0.137413	1.000000	0.218063	0.255197	0.200156
Likes	0.462308	0.218063	1.000000	0.922552	0.960418
Photos	0.401613	0.255197	0.922552	1.000000	0.917763
Tips	0.413512	0.200156	0.960418	0.917763	1.000000

Fig 2.7 – Correlation Coefficient Matrix

From the correlation coefficient Matrix, we can see Likes, Photos, Tips have a strong relation. This means they are redundant features; we can just select one. But Likes, Photos, Tips are not highly related to Rating. We can explain this in two ways, in the one hand, Rating can be a feature, in another hand, Customers who click Likes for some specific reasons but give lower ratings to the general performance might cause this low correlation. And Rating has a weak relation with Price, which means that the price might not affect impressions of customers in that place significantly and Price is a good feature.

2.3 Explore Restaurants data

First, we need to calculate the total restaurant in each neighborhood then draw the restaurant density of each neighbor by using folium lib:

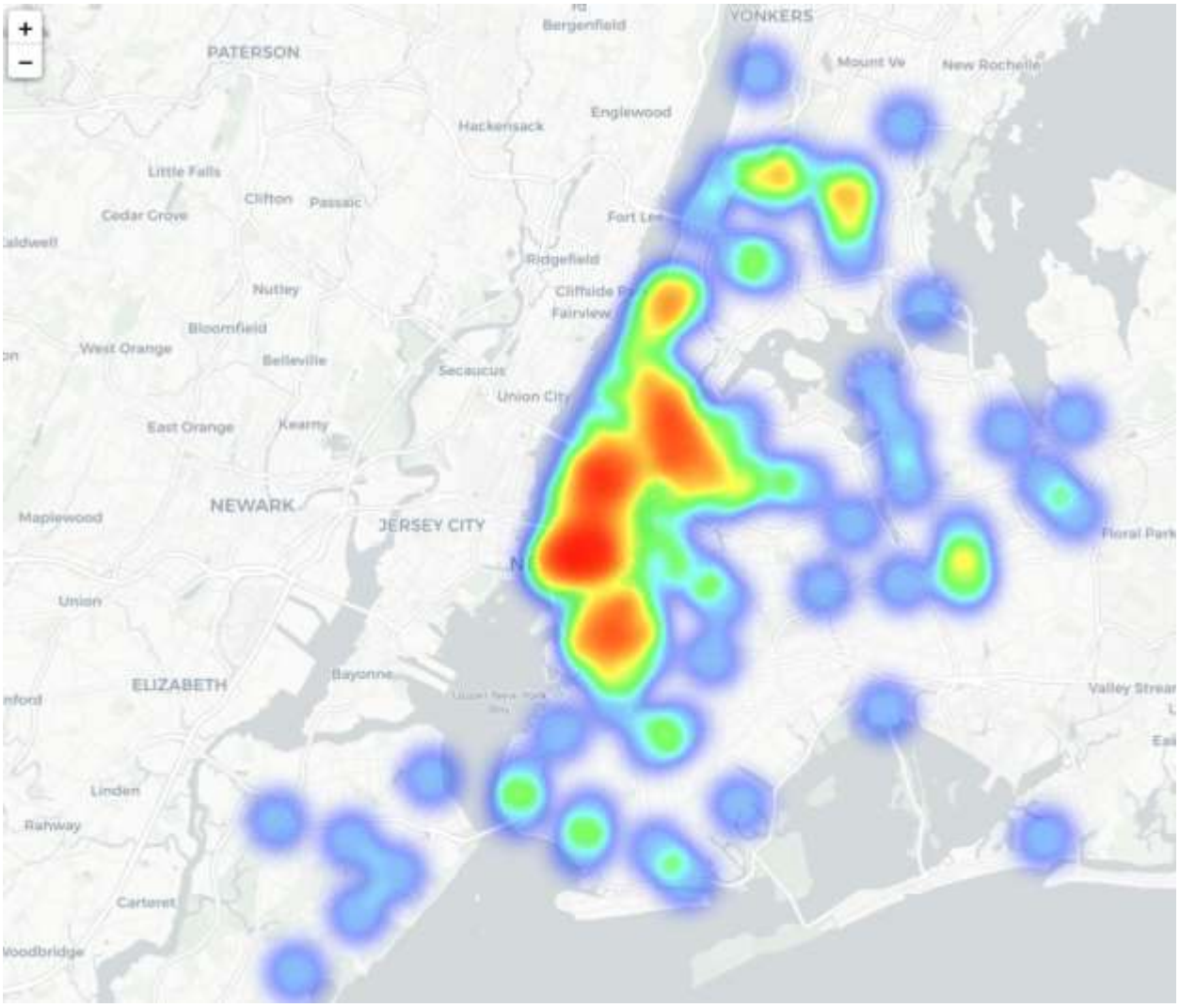

Fig 2.8 – Restaurants Density

From the Heatmap, we can see that most of restaurants are concentrated in Manhattan. Second, I use Horizontal Bar Plot to Visualize the distribution of Chinese restaurants, the results are as following:

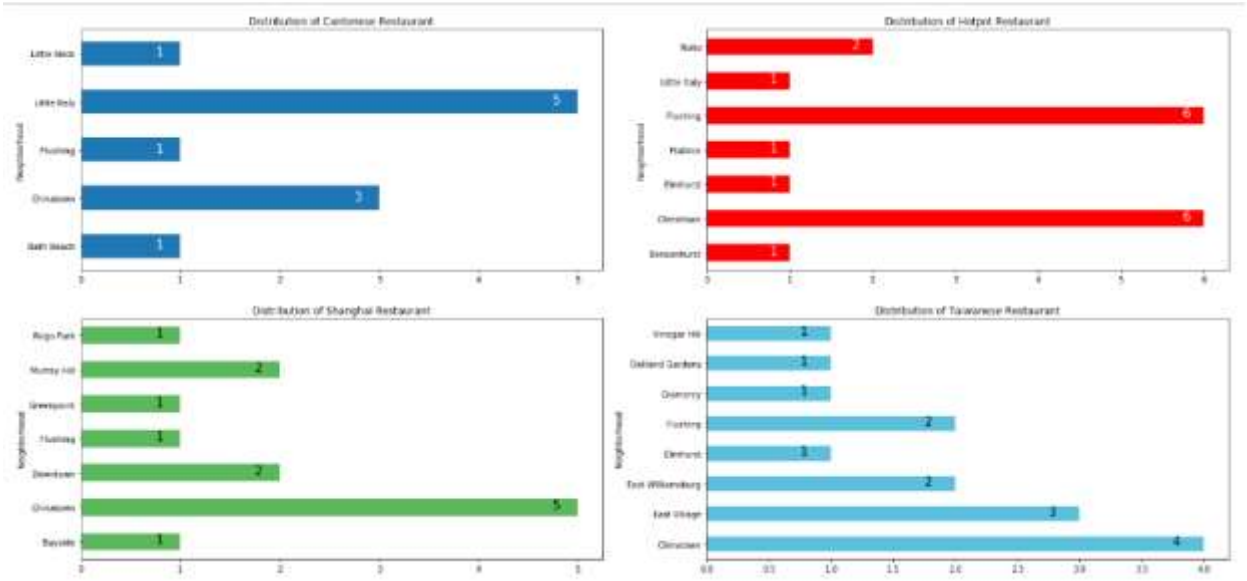

Fig 2.9 – Distribution of Chinese Restaurants

Third, I use the K-means algorithm to classify the Chinese Restaurant into several groups. I use the "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=3 is annotated with a dashed line.

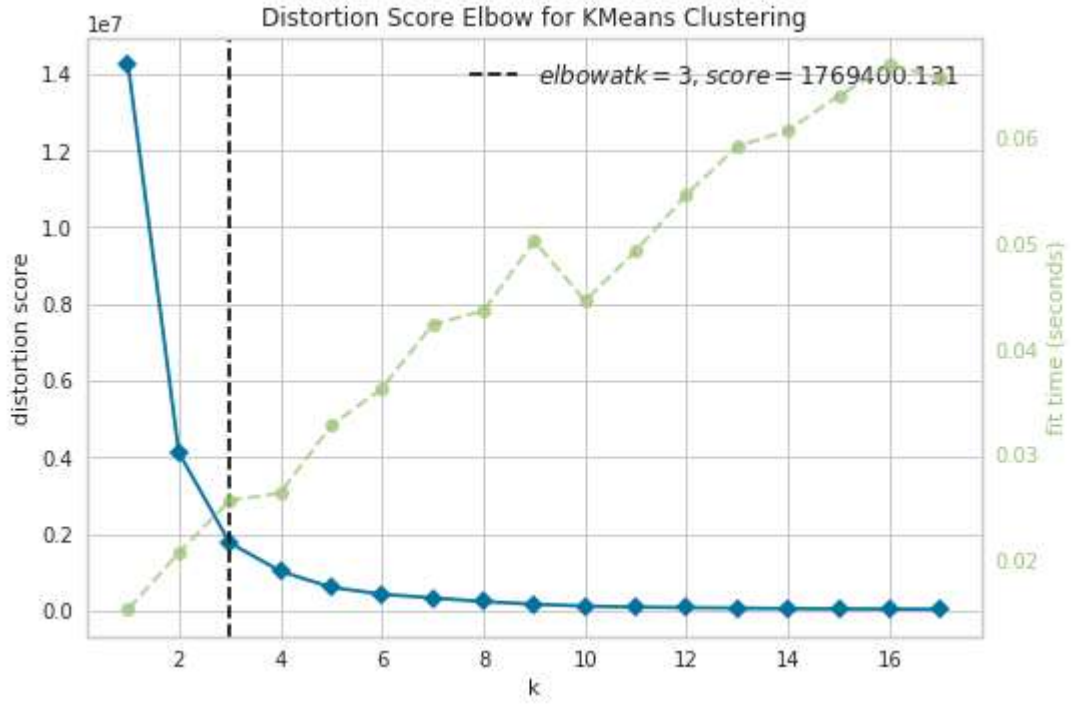

Fig 2.10 – Elbow for KMeans Clustering

I merged cluster labels of Chinese restaurants with its geological location.

	Venue Id	Rating	Price	Likes	Photos	Tips	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels
0	4dabc3dc93a04942f09ccabd	7.0	1.0	6	0	0	Bronx	Eastchester	40.887556	-73.827806	Xing Lung Chinese Restaurant	40.889785	-73.831226	Chinese Restaurant	0
1	4b9d9b45f964a52070ab36e3	7.7	1.0	9	11	10	Bronx	Pelham Parkway	40.857413	-73.854756	Mr. Q's Chinese Restaurant	40.855790	-73.855455	Chinese Restaurant	0
2	4e4d0287bd413c4ca96d6d72	6.6	1.0	5	1	2	Bronx	Bedford Park	40.870185	-73.885512	Hung Hing Chinese Restaurant	40.871181	-73.890759	Chinese Restaurant	0
3	[illegible]	6.3	1.0	6	4	8	Bronx	Bedford Park	40.870185	-73.885512	Choi Yuan - Chinese Restaurant	40.873079	-73.888086	Chinese Restaurant	0
4	[illegible]	7.6	1.0	12	14	2	Bronx	Fordham	40.860997	-73.890427	Panda Express	40.863001	-73.900894	Chinese Restaurant	0

Fig 2.11 – Restaurants Information

Then, I used folium to visualize the distribution of these Chinese restaurants in NYC as below:

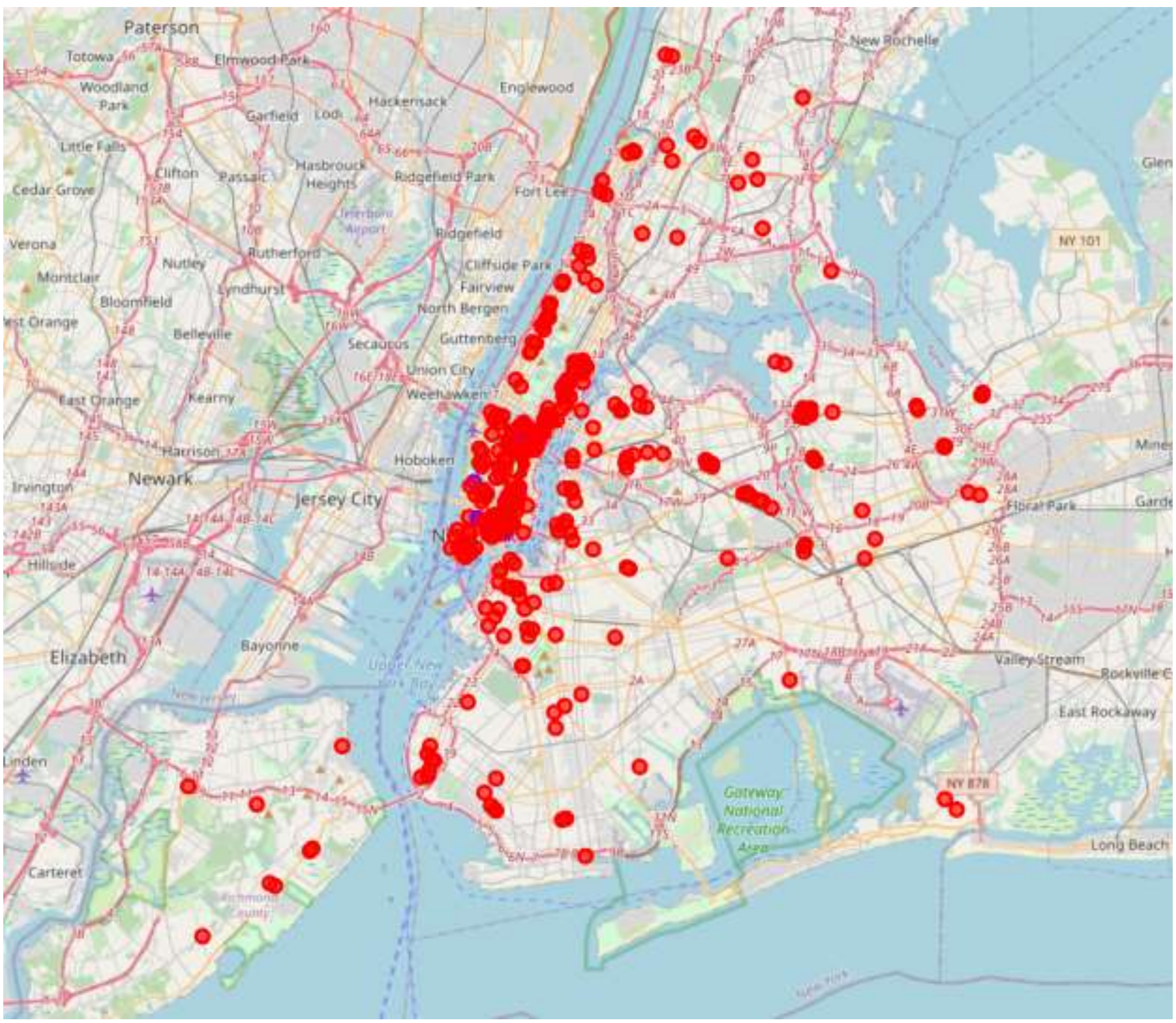

Fig 2.12 – Distribution of Clustering

3. Results

3.1 Cluster 1

Following are the results of the Cluster — 1 analysis:

	Borough	Neighborhood	Venue	Venue Category	Rating	Price	Likes	Photos	Tips	Cluster Labels
0	Bronx	Eastchester	Xing Lung Chinese Restaurant	Chinese Restaurant	7.0	1.0	6	0	0	0
1	Bronx	Pelham Parkway	Mr. Q's Chinese Restaurant	Chinese Restaurant	7.7	1.0	9	11	10	0
2	Bronx	Bedford Park	Hung Hing Chinese Restaurant	Chinese Restaurant	6.0	1.0	5	1	2	0
3	Bronx	Bedford Park	Choi Yuan - Chinese Restaurant	Chinese Restaurant	6.3	1.0	6	4	8	0
4	Bronx	Fordham	Panda Express	Chinese Restaurant	7.6	1.0	12	14	2	0
5	Bronx	Throgs Neck	Green Dragon	Chinese Restaurant	7.6	2.0	10	16	7	0
6	Bronx	Parkchester	Sabor Latino Seafood Restaurant	Chinese Restaurant	7.0	2.0	5	12	8	0
7	Bronx	North Riverdale	New China	Chinese Restaurant	6.5	1.0	2	3	4	0
8	Bronx	North Riverdale	Golden Phoenix Chinese Restaurant	Chinese Restaurant	5.7	1.0	4	1	1	0
9	Bronx	Concourse	Lucky House Chinese Restaurant	Chinese Restaurant	7.4	1.0	5	0	2	0
10	Brooklyn	Bay Ridge	XIN	Chinese Restaurant	7.3	3.0	6	2	4	0

Fig 3.1 – Cluster 1

	Neighborhood	Counts
18	Chinatown	70
62	Little Italy	29
38	Flushing	23
29	East Village	18
69	Midtown	17
73	Murray Hill	10
32	Elmhurst	10
28	Downtown	9
45	Gramercy	9
95	Sutton Place	9
21	Clinton	9
48	Greenwich Village	9
35	Financial District	8
37	Flatiron	8
75	Noho	7

Fig 3.2 – Neighborhood Counts

This is the biggest cluster of all which means that majority of the neighborhoods are clustered in it. Then I calculate the mean of Rating, Price and Likes in Cluster 1, we get: Rating = 6.754762, Price = 1.387446, Likes = 28.640693. I find that Cluster1 is the cluster of very ordinary Chinese restaurants which Rating, Price and Likes are not high. And Top 5 places are Chinatown, Little Italy, Flushing, East Village, Midtown. At the same time, I check the boroughs of these neighborhoods, I find they are all in Manhattan and Queens, which fits the HeatMap we draw before.

3.2 Cluster 2

Following are the results of the Cluster — 2 analysis:

	Borough	Neighborhood	Venue	Venue Category	Rating	Price	Likes	Photos	Tips	Cluster Labels
81	Manhattan	Chinatown	Jing Fong Restaurant 金豐大酒樓	Dim Sum Restaurant	7.9	2.0	1000	1635	290	1
87	Manhattan	Chinatown	Nom Wah Tea Parlor	Dim Sum Restaurant	8.5	1.0	1256	1446	375	1
91	Manhattan	Chinatown	Shanghai 21	Shanghai Restaurant	8.9	2.0	744	663	263	1
100	Manhattan	Chinatown	Vanessa's Dumpling House	Chinese Restaurant	8.4	1.0	1185	688	415	1
102	Manhattan	Chinatown	Shanghai Café Deluxe	Chinese Restaurant	7.9	1.0	895	859	369	1
103	Manhattan	Chinatown	Golden Unicorn Restaurant 麒麟金閣	Dim Sum Restaurant	7.9	2.0	810	1042	228	1
109	Manhattan	Chinatown	Mission Chinese Food	Chinese Restaurant	8.2	3.0	945	1091	406	1
240	Manhattan	Murray Hill	Café China	Chinese Restaurant	8.8	2.0	1046	568	273	1
250	Manhattan	Chelsea	Buddakan	Chinese Restaurant	9.0	3.0	1486	1718	521	1

Fig 3.3 – Cluster 2

	Neighborhood	Counts
1	Chinatown	7
0	Chelsea	1
2	Murray Hill	1
3	Soho	1
4	West Village	1

Fig 3.4 – Cluster 2 Counts

I calculate the mean of Rating, Price and Likes in Cluster 1, we get: Rating = 8.427273, Price = 2.090909, Likes = 983.454545. I find that Cluster 2 is the cluster of high-end Chinese restaurants which Rating, Price, Likes are high. And Top5 places are Chinatown, Chelsea, Murray Hill, Soho, West Village. And these neighborhoods are in Manhattan and Queens.

3.3 Cluster 3

Following are the results of the Cluster — 3 analysis:

	Borough	Neighborhood	Venue	Venue Category	Rating	Price	Likes	Photos	Tips	Cluster Labels
49	Brooklyn	Downtown	Yaso Tangbao	Shanghai Restaurant	7.2	2.0	222	160	69	2
57	Brooklyn	East Williamsburg	Win Son	Taiwanese Restaurant	9.1	2.0	436	262	101	2
59	Brooklyn	North Side	Birds of a Feather	Chinese Restaurant	9.0	1.0	344	189	72	2
60	Brooklyn	North Side	M Shanghai Bistro	Chinese Restaurant	7.7	2.0	230	122	129	2
61	Brooklyn	North Side	Vanessa's Dumpling House	Chinese Restaurant	8.1	1.0	516	248	144	2
72	Manhattan	Chinatown	Spicy Village	Chinese Restaurant	8.8	1.0	501	282	179	2
73	Manhattan	Chinatown	Wah Fung Number 1 Fast Food 華豐快餐店	Chinese Restaurant	8.5	1.0	194	194	96	2
80	Manhattan	Chinatown	Great N.Y. Noodletown	Chinese Restaurant	8.0	1.0	547	769	288	2
82	Manhattan	Chinatown	Taiwan Pork Chop House 臺灣正昌好味道	Taiwanese Restaurant	8.4	1.0	226	329	84	2
85	Manhattan	Chinatown	99 Favor Taste 99號餐廳	Hotpot Restaurant	7.9	2.0	466	412	102	2

Fig 3.5 – Cluster 3

	Neighborhood	Counts
1	Chinatown	23
7	Flushing	3
9	North Side	3
0	Chelsea	2
4	East Village	2
10	Upper West Side	2
2	Clinton	1
3	Downtown	1
5	East Williamsburg	1
6	Flatiron	1
8	Murray Hill	1
11	West Village	1

Fig 3.6 – Cluster 3 Counts

I calculate the mean of Rating, Price and Likes in Cluster 1, we get: Rating = 8.185366, Price = 2.090909, Likes = 316.731707. I find that Cluster 3 is cluster of Chinese restaurants with high cost performance. They have high Rating and Likes but with a low price. And Top5 places are Chinatown, Flushing, North Side, Chelsea, East Village. These neighborhoods are in Manhattan, Queens and Brooklyn,

4. Conclusion

As a result, there exists a pattern in the distribution of Chinese restaurants in NYC, and most of the restaurants are located in areas where a large number of Chinese live. And there are 3 types of Chinese restaurants in NYC, so people who want to open a corresponding type of restaurant should consider starting with the corresponding location where has less restaurants. For example, for those who want to run a high-end Chinese restaurant could consider starting their preliminary investigation on Murray Hill, Soho or West Village. Others are similar.

5. Discussion

As a recommendation to those who plan to operate a restaurant, location selection is only one fundamental problem to think over. It can't solve the problem of whether a type of restaurant is the most popular type and how many customers will visit every day. And as for location suggestion, it offers an opportunity analysis but lacks risk analysis, like the cost of the location and competition in that area.

Although in this report, it demonstrates the relations between location, ratings, price and likes, but ratings might not reflect the operation status of the restaurant. A restaurant with a high rating could still be unprofitable, which is not desirable from a business perspective. So the suggestion is relatively narrow. To suggest more practical and profitable ideas, the relationship between customer reactions and financial performance should be evaluated, so that the report can become constructive for a restaurant owner in the real business world.

6. Reference

[1]. https://en.wikipedia.org/wiki/New_York_City#Cuisine