

Capstone Project:

Neighborhood Recommendation for Asia Fine Dining Restaurant Chain Expansion Project

1 Introduction

1.1 Background

Client is an Asian restaurant group specialized in Chinese cuisine. Originated from Mainland China, it already has extended footprints around East Asia and earned its reputation on its artistic and fine dining experience. With a lot of tourists visiting them from overseas, it is decided to further expand their chain to North America. Considered New York and Toronto are one of the most alive and populous cities in United States and Canada respectively, Client decided to pick a location among their neighborhoods to open their flag ship restaurant in order to boost up their reputation further.

1.2 Problem

We are engaged by the Client for this important task for picking the right neighborhood. The sponsor of this project from this Asian Restaurant Group would be the Target Audience.

By adopting an appropriate Data Science methodology, we are going to recommend the best neighborhood among New York and Toronto for establishing this flag ship restaurant based on the data analysis conducted.

With the business goal established, client has further illustrated their criteria on what's "the best neighborhood":

- It should be a prime location, where it is popular to the crowd.
- It should have many good restaurants surrounding, so that people could recognize it is a good dining area.
- It is preferred to have similar kind of restaurant (i.e. Asian fine dining) around the area.

Therefore, we are going to focus on these criteria to develop our recommendation.

2 Data

In this project, we would apply the following data sources:

#	Data Acquired	Details of Data Source			
1	Neighborhood Information	<p>For New York's, it is based on the Wikipedia Page: Neighborhoods in New York City. (URL: https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City)</p> <p>For Toronto's, it is based on the Toronto City Government Website: Neighborhood Profiles. (URL: https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/)</p> <p>Data Sample:</p> <table><tr><th>#</th><th>City</th><th>Neighborhood</th></tr></table>	#	City	Neighborhood
#	City	Neighborhood			

		<table><tr><td>1</td><td>NY</td><td>Bay Terrace</td></tr><tr><td>2</td><td>TO</td><td>North Riverdale</td></tr><tr><td>3</td><td>...</td><td>...</td></tr></table>	1	NY	Bay Terrace	2	TO	North Riverdale	3																																																																															
1	NY	Bay Terrace																																																																																								
2	TO	North Riverdale																																																																																								
3																																																																																								
2	Geometry Information	<p>Because there is no geometry information provided from the sources above, we need to further acquire the geometry information (i.e. latitude and longitude) of each neighborhoods from the OpenCage Geocoding API. By inputting the neighborhood name, it would returns the required information.</p> <p>Data Sample:</p> <table><tr><th>#</th><th>City</th><th>Neighborhood</th><th>Lat</th><th>Lng</th></tr><tr><td>1</td><td>NY</td><td>Bay Terrace</td><td>40.555278</td><td>-74.134167</td></tr><tr><td>2</td><td>TO</td><td>North Riverdale</td><td>43.66547</td><td>-79.352594</td></tr><tr><td>3</td><td>...</td><td>...</td><td>...</td><td>...</td></tr></table>	#	City	Neighborhood	Lat	Lng	1	NY	Bay Terrace	40.555278	-74.134167	2	TO	North Riverdale	43.66547	-79.352594	3																																																																				
#	City	Neighborhood	Lat	Lng																																																																																						
1	NY	Bay Terrace	40.555278	-74.134167																																																																																						
2	TO	North Riverdale	43.66547	-79.352594																																																																																						
3																																																																																						
3	Location Information of Venues around Neighborhood	<p>Venue information around each neighborhood is the key elements for the analysis. By adopting the FourSquare API, the following set of location information are obtained:</p> <p>a. List of Venue around each neighborhood</p> <p>Data Sample:</p> <table><tr><th>#</th><th>id</th><th>Venue Name</th><th>Lat</th><th>Lng</th><th>dist</th><th>Category</th><th>Cat ID</th></tr><tr><td>1</td><td>4b92...</td><td>Metro North...</td><td>40.825...</td><td>-73.915...</td><td>22</td><td>Train Station</td><td>4bf8...</td></tr><tr><td>2</td><td>4dd9...</td><td>Tender Tot...</td><td>40.826...</td><td>-73.915...</td><td>49</td><td>Daycare</td><td>4d95...</td></tr><tr><td>3</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr></table> <p>Remark: “dist” is the distance between the venue from the central point of neighborhood in meter</p> <p>b. Detailed information for each venue and that could be divided into two sets (Standard and Premium)</p> <p>Data Sample (Standard):</p> <table><tr><th>#</th><th>id</th><th>Venue Name</th><th>Like Counts</th><th>Related List</th><th>List count</th></tr><tr><td>1</td><td>4b92...</td><td>Metro North...</td><td>161</td><td>...</td><td>4</td></tr><tr><td>2</td><td>4dd9...</td><td>Tender Tot...</td><td>154</td><td>...</td><td>1</td></tr><tr><td>3</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr></table> <p>Remark:</p> <p>(1) Four Square user can like a venue, “Like Counts” is the number of user who likes the venue;</p> <p>(2) Four Square user can create Venue List for sharing, “Related List” is all lists with such venue, “List count” is number of list with such venue.</p> <p>Data Sample (Premium):</p> <table><tr><th>#</th><th>id</th><th>Venue Name</th><th>url</th><th>tipcount</th><th>Price tier</th><th>Rating</th><th>RatingSignal</th></tr><tr><td>1</td><td>4b92...</td><td>Metro North...</td><td>http://...</td><td>13.0</td><td>2</td><td>8.0</td><td>255</td></tr><tr><td>2</td><td>4dd9...</td><td>Tender Tot...</td><td>http://...</td><td>10.0</td><td>4</td><td>5.0</td><td>12</td></tr><tr><td>3</td><td>...</td><td>...</td><td></td><td></td><td></td><td></td><td></td></tr></table> <p>Remark:</p> <p>(1) For Premium details, FourSquare only provides limited access and we can’t obtain premium data for all venues. Thus, we need to select a group of venues for accessing premium details.</p>	#	id	Venue Name	Lat	Lng	dist	Category	Cat ID	1	4b92...	Metro North...	40.825...	-73.915...	22	Train Station	4bf8...	2	4dd9...	Tender Tot...	40.826...	-73.915...	49	Daycare	4d95...	3	#	id	Venue Name	Like Counts	Related List	List count	1	4b92...	Metro North...	161	...	4	2	4dd9...	Tender Tot...	154	...	1	3	#	id	Venue Name	url	tipcount	Price tier	Rating	RatingSignal	1	4b92...	Metro North...	http://...	13.0	2	8.0	255	2	4dd9...	Tender Tot...	http://...	10.0	4	5.0	12	3					
#	id	Venue Name	Lat	Lng	dist	Category	Cat ID																																																																																			
1	4b92...	Metro North...	40.825...	-73.915...	22	Train Station	4bf8...																																																																																			
2	4dd9...	Tender Tot...	40.826...	-73.915...	49	Daycare	4d95...																																																																																			
3																																																																																			
#	id	Venue Name	Like Counts	Related List	List count																																																																																					
1	4b92...	Metro North...	161	...	4																																																																																					
2	4dd9...	Tender Tot...	154	...	1																																																																																					
3																																																																																					
#	id	Venue Name	url	tipcount	Price tier	Rating	RatingSignal																																																																																			
1	4b92...	Metro North...	http://...	13.0	2	8.0	255																																																																																			
2	4dd9...	Tender Tot...	http://...	10.0	4	5.0	12																																																																																			
3																																																																																								

		<p>(2) "tipcount" is the number of users provided tips for such venue.</p> <p>(3) "Price_tier" indicate how expensive the restaurant is in the range of 1(cheap) to 4(expensive).</p> <p>(4) "Rating" is the average rate provided by users. "Rating Signal" is the no. of users provided such rating.</p>
--	--	--

3 Methodology

The analysis is mainly relied on the location data provided by Four Square. By submitting the latitude and longitude of each neighbourhood in scope (i.e. within Toronto and New York), Four Square returned a list of venues (specifically, we are interested in restaurants only in this study) around such neighbourhood with relevant attributes.

We would first explore these attributes to access their distribution, their relationship with each other in order to establish their connection with the analysis. Then, based on comparing & aggregating these attributes of venues within the neighbourhood, we would further establish our understanding to the neighbourhood and further conclude our recommendation to the problem above.

4 Exploratory Data Analysis

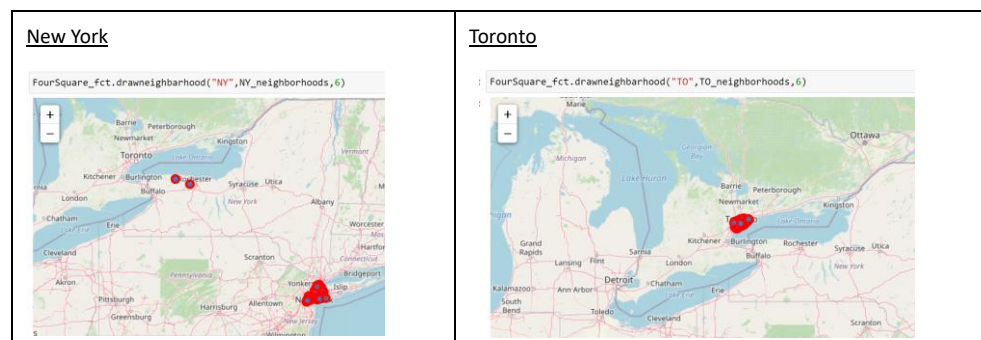
4.1 Geometry Information of neighbourhoods

There are 318 neighbourhoods in New York and 140 neighbourhoods in Toronto. And we rely on Open Cage to obtain the geometry information. We validate them based on the descriptive statistics, as below >>

New York			Toronto		
	lat	lng		lng	lat
count	318.000000	318.000000	count	140.000000	140.000000
mean	40.802827	-74.101653	mean	-79.396461	43.706296
std	0.492997	0.793809	std	0.089970	0.047625
min	40.511217	-78.210376	min	-79.597457	43.592005
25%	40.633368	-74.005166	25%	-79.437317	43.671664
50%	40.713689	-73.949132	50%	-79.416300	43.700110
75%	40.787462	-73.875184	75%	-79.354615	43.730064
max	43.280191	-73.660795	max	-79.150768	43.823174

The geometry information of Toronto neighbourhoods seems consistent. However, that of New York neighbourhoods may be incorrect. For latitude, the maximum is 43.28 and for longitude, the minimum is -78.21, which are far away from New York.

We further draw the neighbourhoods on map by using Folium as below, where each red dot represents one neighbourhood >>



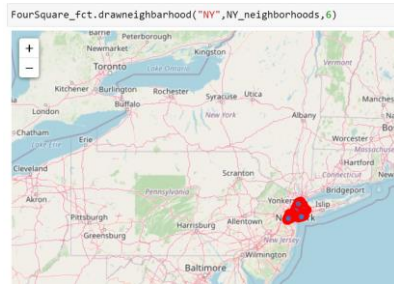
For New York's, some neighbourhoods are with incorrect geometry information and outside New York; while, for

Toronto's, all the red dots are within the Toronto's area.

Data Cleaning

Based on Google Map, the range of latitude and longitude of New York are (40.495788, 40.915845) and (-74.253752, -73.726185). 13 neighbourhoods are identified outside the ranges. As the amount is manageable, we manually check in the google map and updated the data.

Redraw the neighbourhoods again as below with updated data.



4.2 Location Information of Venues around Neighborhood

Based on the geometry information of each neighbourhood, we obtained the venues list related to each of them from Four Square. In total, there are over 47,000 venues. Similarly, we quickly check on their descriptive statistics and will review the attributes as below.

	Neighborhood Latitude	Neighborhood Longitude	Venue Latitude	Venue Longitude	distance
count	47237.000000	47237.000000	47237.000000	47237.000000	4.723700e+04
mean	41.573449	-75.516692	41.572665	-75.518434	9.481881e+02
std	1.361775	2.469393	1.365899	2.537447	4.910421e+04
min	40.511217	-79.597457	27.852840	-121.272707	0.000000e+00
25%	40.666770	-79.276515	40.666202	-79.278550	1.120000e+02
50%	40.755906	-73.988504	40.756625	-73.987956	2.370000e+02
75%	43.654644	-73.902746	43.653699	-73.902261	4.280000e+02
max	43.823174	-73.684722	46.601696	28.742530	8.050293e+06

4.2.1 Distance

The field "distance" is the distance between each venue and the neighbourhood. From the descriptive statistics table above, while majority (i.e. around ~75% of venues) are within the range 428m. However, the maximum distance is over 8000km away, which shouldn't be relevant to the neighbourhood anymore. To make the venue data more relevant, we have filtered out those venues with distance over 500m away. That makes the venue list down to ~38000 venues.

4.2.2 Venue Category

The field "Venue Category" indicates the category of venue. Four Square has its own categories.

Handling of Missing Data

First, we checked if there is any missing data. From the table below, there are only 35183 non-null entries over 37985 entries. (i.e. there are ~3000 entries without Venue Category). As this field is key to determine if a venue is relevant, we would drop those data and results 35183 entries left.

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 37985 entries, 0 to 13620
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Neighborhood           37985 non-null  object
1   Neighborhood Latitude  37985 non-null  float64
2   Neighborhood Longitude 37985 non-null  float64
3   id                     37985 non-null  object
4   Venue                  37985 non-null  object
5   Venue Latitude         37985 non-null  float64
6   Venue Longitude        37985 non-null  float64
7   distance               37985 non-null  int64
8   Venue Category         35183 non-null  object
9   Venue Category ID      35183 non-null  object
10  city                   37985 non-null  object
dtypes: float64(4), int64(1), object(6)
memory usage: 3.5+ MB

```

To further explore, we listed out the top 15 categories in term of counts and % of total.

	Venue Category	cnt	%
0	Residential Building (Apartment / Condo)	1569	4.46%
1	Office	1470	4.18%
2	Building	1172	3.33%
3	Salon / Barbershop	1142	3.25%
4	Doctor's Office	1009	2.87%
5	Deli / Bodega	777	2.21%
6	Bus Line	562	1.60%
7	Church	528	1.50%
8	Laundry Service	524	1.49%
9	Dentist's Office	506	1.44%
10	Pizza Place	502	1.43%
11	Automotive Shop	475	1.35%
12	Medical Center	443	1.26%
13	Chinese Restaurant	421	1.20%
14	Bank	388	1.10%
15	Subtotal	11488	32.65%

Among these top 15 categories, only 2 are related to dining (i.e. Pizza Place and Chinese Restaurant).

Thus, we need to further narrow down the list for item related to dining only. Luckily, Four Square defined its Venue Category in a hierarchy manner and we could locate all relevant categories under the key categories "Food". With that, the list further reduced to 6848 venues as below:

	Venue Category	cnt	%
0	Deli / Bodega	777	11.35%
1	Pizza Place	502	7.33%
2	Chinese Restaurant	421	6.15%
3	Bakery	326	4.76%
4	Coffee Shop	294	4.29%
...
141	Soba Restaurant	1	0.01%
142	Colombian Restaurant	1	0.01%
143	Souvlaki Shop	1	0.01%
144	Japanese Curry Restaurant	1	0.01%
145	Subtotal	6848	100.00%

The most popular category from the list above is "Deli / Bodega". However, it is kind of convenient stores instead of a place serving food and drinks. Thus, venues under this category is dropped and the list further reduced to 6071 venues.

Composition of venue categories in New York and Toronto

New York	Toronto
----------	---------

NY % NY			TO % TO		
Venue Category			Venue Category		
Pizza Place	9.09%	383.0	Coffee Shop	7.44%	139.0
Chinese Restaurant	7.88%	332.0	Pizza Place	6.64%	124.0
Bakery	5.44%	229.0	Bakery	5.03%	94.0
Food Truck	5.27%	222.0	Restaurant	4.82%	90.0
Food	4.18%	176.0	Café	4.60%	86.0
Italian Restaurant	3.89%	164.0	Italian Restaurant	4.55%	85.0
Coffee Shop	3.75%	158.0	Chinese Restaurant	4.50%	84.0
Café	3.39%	143.0	Fast Food Restaurant	4.18%	78.0
Mexican Restaurant	3.25%	137.0	Sandwich Place	3.91%	73.0
Caribbean Restaurant	2.87%	121.0	Caribbean Restaurant	2.62%	49.0

The two tables above show the top 10 venue categories in New York and Toronto. It is observed that Chinese Restaurant is 2nd most popular category in New York (vs. 7th most popular in Toronto). Also, there are more restaurants (for all categories) in New York. It is likely to be related to the size of the city, where New York have over 300 neighbourhoods while Toronto only have 140

4.2.3 Like count

“Like count” is an important attribute related to each restaurant. It indicates how many users clicked “like” button to a restaurant. It is a key indicator to show how popular and how good one is.

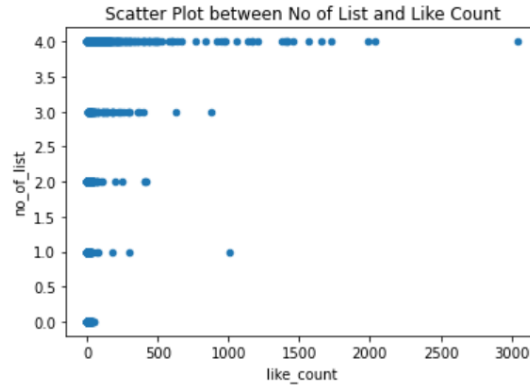
Range	cnt	cnt %	NY	NY %	TO	TO %
0	2640	43.41%	1759	41.75%	881	47.16%
1	924	15.19%	661	15.69%	263	14.08%
2	439	7.22%	302	7.17%	137	7.33%
3	266	4.37%	174	4.13%	92	4.93%
4 - 10	677	11.13%	437	10.37%	240	12.85%
11 - 100	864	14.21%	630	14.95%	234	12.53%
101 - 500	177	2.91%	166	3.94%	11	0.59%
501 or above	34	0.56%	34	0.81%	0	0.00%
Total	6021	100%	4163	100%	1858	100%

We have bucketized those 6021 restaurants in various “like count” range and presented as the table above. It is observed that

- ✧ Almost half of restaurant receive 0 “like”;
- ✧ Over 80% of restaurants receive 10 likes or less.
- ✧ Only 211 restaurants receive 101 likes or above, which majority are in New York. Among those, only 34 (0.56%) received 501 likes or above.

4.2.4 No of list

Four Square user can select venues they like and create a venue list for sharing. “No of list” is how many lists have referenced a restaurant. Thus, similar to “like count”, if a restaurant is referenced by more lists, it should be relatively more popular.



The above is scattered chart between “no of list” and “like count”. That shows:

- ✧ There are only 5 buckets for “no of list”. (i.e. At most, a venue is only referenced by 4 lists)
- ✧ Restaurants with high “like count” (i.e. those with > 500 like count) are likely to be referenced in more list (i.e. no of list = 4)
- ✧ Restaurant with low “like_count” may not referenced by fewer lists.

Not sure why the range of “no of list” is quite small (only between 0 and 4). Maybe it is because the venue list creation is not as widely adopted as like count. However, comparing with “like count”, that make it less effective for identifying which restaurants is more popular.

The above are all standard data available from Four Square related to our analysis. We would further utilize premium data provided by Four Square. However, Four Square imposed restriction on Premium Data, in which we could only access limited restaurants for such. Therefore, we would need to be selective and refine our list for accessing premium data.

From the above, it is observed that “like count” is a better indicator. Thus, we aggregated the like count for all the restaurants from neighbourhoods and obtained the “total like count”. With that, we selected the top 10 neighbourhood based on the “total like count” as below:

Neighborhood	like_count	cnt	no_of_list
East Village(Manhattan)	8078.0	28	82.0
Mount Hope(Bronx)	6522.0	41	94.0
Greenpoint(Brooklyn)	4485.0	39	89.0
Prospect Heights(Brooklyn)	3578.0	28	79.0
SoHo(Manhattan)	3428.0	16	18.0
Upper West Side(Manhattan)	3245.0	21	48.0
West Village(Manhattan)	3020.0	14	39.0
Tribeca(Manhattan)	2940.0	29	49.0
Fish Bay(Bronx)	2940.0	20	44.0
Dumbo(Brooklyn)	2898.0	14	35.0

In the following premium data analysis, we would focus in the 250 restaurants as above.

The premium data attributes for the top 10 neighbourhoods are summarized as the table below; there are in total 277 restaurants.

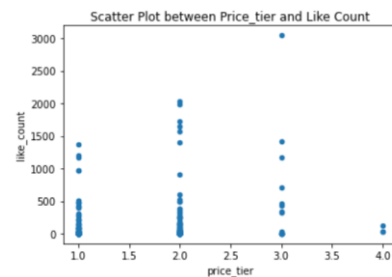
	Neighborhood	like_count	cnt	no_of_list	tipCount	rating	ratingSignal	No pricetier	pricetier 1	pricetier 2	pricetier 3	pricetier 4
0	East Village(Manhattan)	8078	28	82	2717	7.97778	11253	0.0	13.0	14.0	1.0	0.0
1	Mount Hope(Bronx)	6522	41	94	2473	7.70833	9201	2.0	24.0	13.0	2.0	0.0
2	Greenpoint(Brooklyn)	4485	39	89	1376	7.412	6136	4.0	16.0	19.0	0.0	0.0
3	Prospect Heights(Brooklyn)	3578	28	79	1152	7.95789	4824	0.0	13.0	11.0	3.0	1.0
4	SoHo(Manhattan)	3428	16	18	1146	7.9	4765	1.0	10.0	4.0	1.0	0.0
5	Upper West Side(Manhattan)	3245	21	48	1141	7.34167	4611	3.0	9.0	9.0	0.0	0.0
6	West Village(Manhattan)	3020	14	39	1078	7.85	4290	0.0	6.0	4.0	4.0	0.0
7	Fish Bay(Bronx)	2940	20	44	1455	7.67692	4496	0.0	9.0	8.0	3.0	0.0
8	Tribeca(Manhattan)	2940	29	49	899	6.97143	4271	0.0	15.0	11.0	1.0	2.0
9	Dumbo(Brooklyn)	2898	14	35	817	7.88889	4078	0.0	12.0	1.0	1.0	0.0
10	Total		250					10.0	127.0	94.0	16.0	3.0

4.2.5 Price tier

“Price tier” is an indicator of how expensive a restaurant is. It is from range of 1 to 4 (1: Cheapest, 4: Most Expensive). Reference to the table above,

- ✧ Only 262 restaurants provided “Price tier” indicator
- ✧ There are only 20 restaurants under the top two tiers

Further, we explore if there is any correlation between “Price tier” and “Like count”. From the scattered graph below, there is no observable correlation between them.

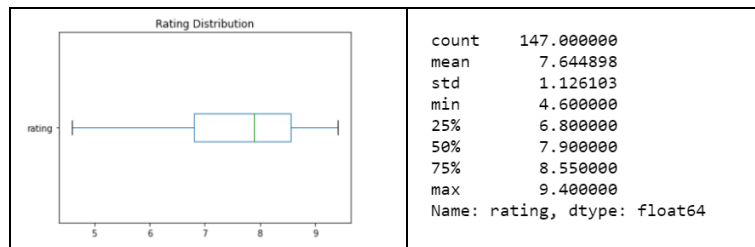


4.2.6 Rating and Rating Signal

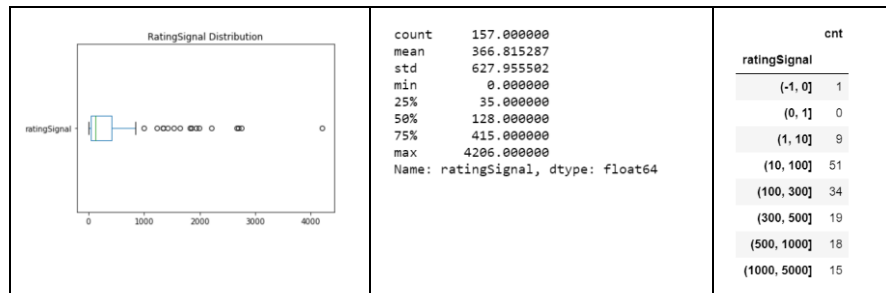
Four Square users could rate a restaurant accordingly from 0 to 10, with 10 to be the best rating. “Rating” is the average from all rates received; “Rating Signal” is the number of users provided such rate. In general, “Rating” may be an useful attribute for accessing a restaurant.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 250 entries, 0 to 249
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  -
0   id          250 non-null   object
1   rating      147 non-null   float64
2   ratingSignal 147 non-null   float64
dtypes: float64(2), object(1)
memory usage: 17.8+ KB
```

From the table above, out of 250 entries, only 147 entries (58.8%) with rating.

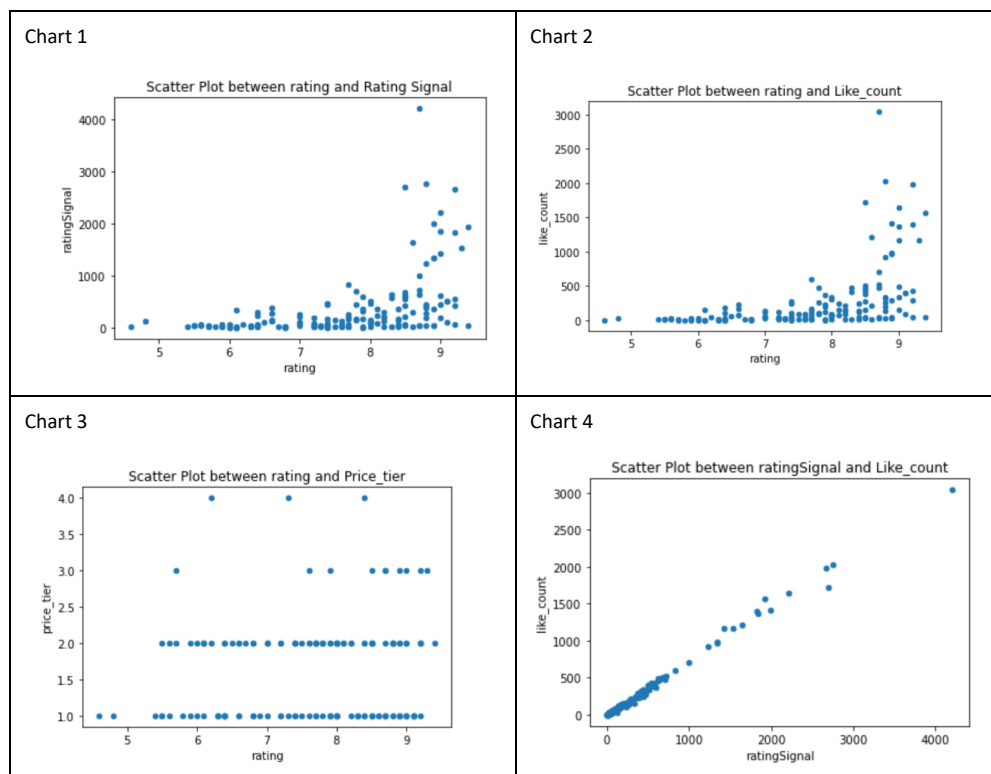


Refer to the Rating Distribution box graph and table above, the average of rating is 7.64 with standard derivation to be 1.12.



For "RatingSignal", the distribution is as the box graph and tables above. The variance of "RatingSignal" is quite big and there are quite a lot of outliers. Thus, it is meaning to reference the bucket count (from the table on right-hand side). Among of the 147 restaurants with rating, most of them are in the range of (10, 100).

Then, we explore if there is any correlation among the attributes: (1) Rating, (2) RatingSignal, (3) like count and (4) Price tier.

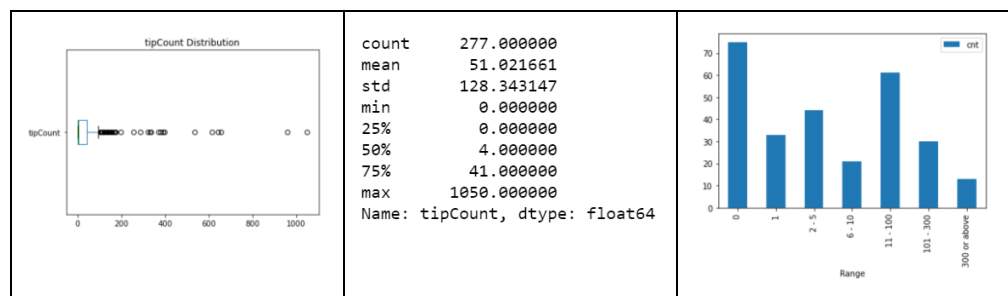


- ✧ In Chart 1, it is observed that, for data with “ratingSignal” lower than 500 (i.e. less than 500 users have rated), they are quite widely distributed. It could be interpreted that the same rating (high or low) could be rated by small group of people or large group of people. Nevertheless, the correlation between them is weak.
- ✧ In Chart 2, it is found that the pattern is very similar to Chart 1. While it shows the correlation between “Rating” and “Like count” is weak, it indicates the correlation between “ratingSignal” and “Like count” could be strong. That’s proved by Chart 4.
- ✧ In Chart 3, it shows the average rating in Price Tier 3 are the best and that of Price Tier 1 is the worst.

Overall, it shows that, while “Rating” is designed to rate a restaurant, direct adoption of that to interpret the popularity may be misleading. For example, with two restaurants of rating of 9.0, the one with only 3 rates and the one with 1000 rates shouldn’t consider with the same popularity. Thus, “like count” is considered a better indicator for that.

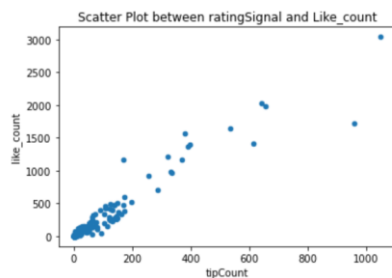
4.2.7 TipCount

Four Square user could leave comment on the restaurant called tips. While the tip could be good or bad, the high number of tips is likely to be related to its popularity.



Regarding “TipCount”, the distribution is shown as charts and table above. The variance is quite big and there are quite a lot of outliers. From the chart on the right-hand side, quite a lot of restaurants (>70) are without tipCount.

From the scattered chart below, it has strong correlation with like_count.



Overall, it could be summarized that a restaurant with more “Like count” would attract more users to rate and more users to provide tips to such place. However, its “Rating” is from the average of rates of all users are depend on its user group, which could be big or small.

4.3 Chinese Restaurant

As this analysis is about Chinese Restaurant, it is worth to explore the distribution of Chinese Restaurant among the two cities: New York and Toronto.

city	NY	TO
price_tier		
1.0	313.0	77.0
2.0	19.0	12.0
3.0	2.0	0.0
4.0	0.0	2.0

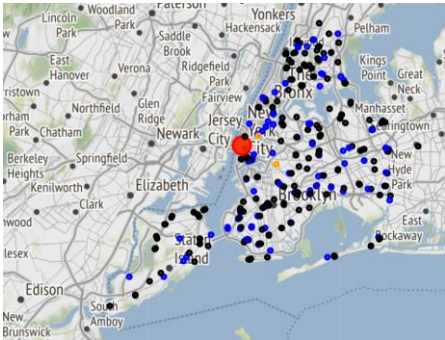
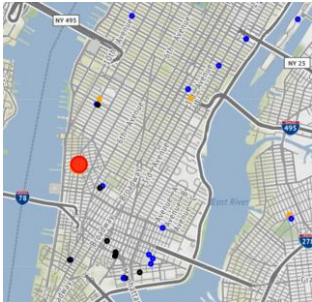
From the table and as illustrated in Section 5.2.2 above,

- ✧ Chinese Restaurant are more common in New York compared to Toronto.
- ✧ Only four of them (a very small number) are in high-tier price range.

Range	cnt	
	NY	TO
0	147.0	62.0
1	83.0	15.0
2	42.0	10.0
3 - 10	52.0	8.0
11 - 100	12.0	8.0
101 - 500	7.0	0.0
500 or above	1.0	0.0

From the table above in which Chinese restaurants are grouped by “Like count” range, all restaurants with high “Like count” are all in New York.

We further visualized the restaurants in the map:

<p>New York</p>  	<ol style="list-style-type: none">1) Chinese Restaurants are blooming everywhere in the city.2) Many of them are with low like count. However, those with high “like count” (indicated by Red / Orange) and higher price-tier (indicated with large circle), are all located in the neighbourhood of West Village in Manhattan.
<p>Toronto</p>	<ol style="list-style-type: none">1) Chinese Restaurants are only sparsely spread across the city.2) None of them are with Red / Orange circle (i.e. none of them have high “like count”).



5 Result and Discussion

From Section 4 above, there are a few attributes by Four Square, which are useful to access how popular a restaurant is. They are “Like count”, “No of List”, “Rating”, “RatingSignal” and “Tip Count”. However, we choose to use “Like count” as the key attribute for the assessment because:

- (1) It is standard data instead of premium data so that we don’t have limitation on data availability.
- (2) There are no missing data.
- (3) It has a wide range of value available which could differential the restaurants in a granular level.
- (4) It also shows correlation with other attributes.

Based on aggregation of like count, 10 neighborhoods are shortlisted as below:

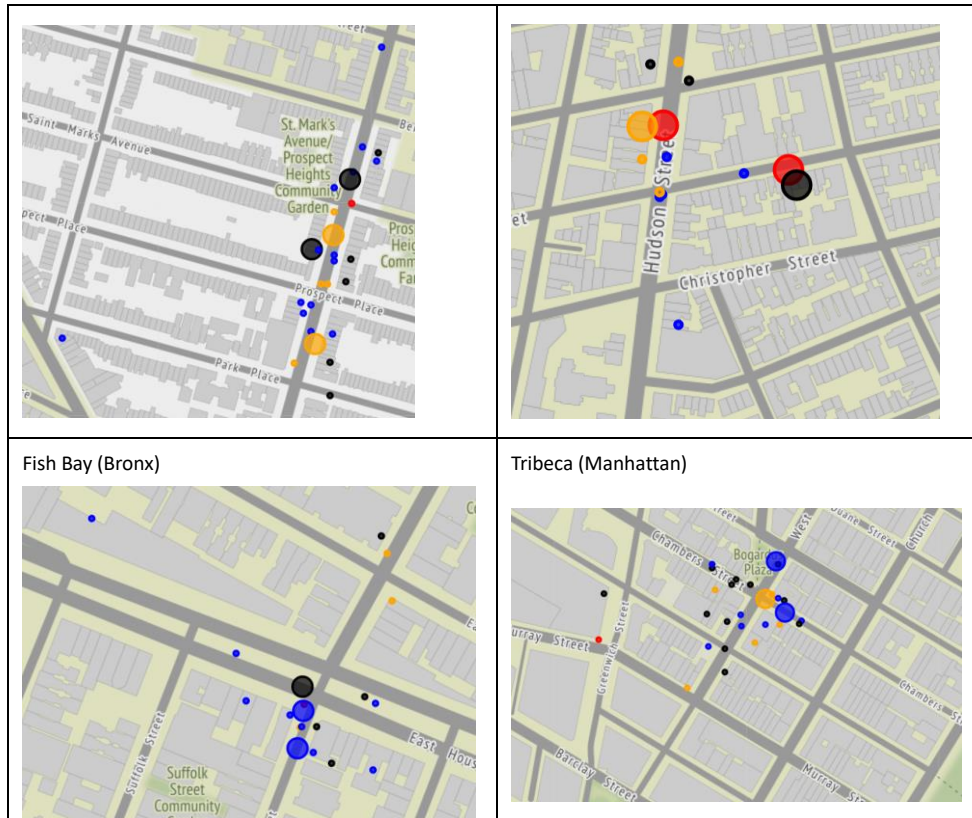
	Neighborhood	like_count	cnt	no_of_list	tipCount	rating	ratingSignal	No pricetier	pricetier 1	pricetier 2	pricetier 3	pricetier 4
0	East Village(Manhattan)	8078	28	82	2717	7.97778	11253	0.0	13.0	14.0	1.0	0.0
1	Mount Hope(Bronx)	6522	41	94	2473	7.70833	9201	2.0	24.0	13.0	2.0	0.0
2	Greenpoint(Brooklyn)	4485	39	89	1376	7.412	6136	4.0	16.0	19.0	0.0	0.0
3	Prospect Heights(Brooklyn)	3578	28	79	1152	7.95789	4824	0.0	13.0	11.0	3.0	1.0
4	SoHo(Manhattan)	3428	16	18	1146	7.9	4765	1.0	10.0	4.0	1.0	0.0
5	Upper West Side(Manhattan)	3245	21	48	1141	7.34167	4611	3.0	9.0	9.0	0.0	0.0
6	West Village(Manhattan)	3020	14	39	1078	7.85	4290	0.0	6.0	4.0	4.0	0.0
7	Fish Bay(Bronx)	2940	20	44	1455	7.67692	4496	0.0	9.0	8.0	3.0	0.0
8	Tribeca(Manhattan)	2940	29	49	899	6.97143	4271	0.0	15.0	11.0	1.0	2.0
9	Dumbo(Brooklyn)	2898	14	35	817	7.88889	4078	0.0	12.0	1.0	1.0	0.0
10	Total		250					10.0	127.0	94.0	16.0	3.0

At the same time, these 10 neighborhoods also covered 49 out of top 100 restaurants in term of like count across both cities, New York and Toronto.

As the goal is to open an upper tier dining place, it is preferable to have a neighbourhood which have similar restaurant arounds. Therefore, the 4 neighbourhoods, which are with more price tier 3 and 4 restaurants, are preferable. They are “Prospect Heights (Brooklyn)”, “West Village (Manhattan)”, “Fish Bay (Bronx)”, “Tribeca (Manhattan)”.

We further visualized the restaurant distribution for these neighbourhoods as below:

Prospect Heights (Brooklyn)	West Village (Manhattan)
-----------------------------	--------------------------



In the visualization above,

- For restaurants with price tier 3 or 4, they are in large circle. Otherwise, they are in small one.
- For restaurants with like_count > 500, it appears in "Red"; like_count > 100, it appears in "Orange"; like_count > 1, it appears in "Blue"; like_count > 100, it appears in "Black";

In view of the result above, it is recommended that the "West Village (Manhattan)" is preferable neighborhood because

- It is the top10 neighborhood in term of like-count.
- It is a high-end dining area. It has 4 restaurants with price tier 3 or above.
- Among those 4 restaurants, three have received quite a lot of like counts and two of them are Chinese Restaurants. It could help to establish a Chinese Fine Dining area to draw potential customers.

6 Conclusion

In the analysis above, we have compared the neighborhoods among New York & Toronto and recommended the best neighborhood to open a new the high-end Chinese Restaurant. Throughout the project, while there is no complicated algorithm and data modeling involved, most of the effort was spent are on data collecting, understanding & cleaning. It reflected what this course said before: 80% of effort is on those area. After all the data are tidy up, the analysis is almost completed.

At the same time, it is important to note that the analysis result is highly relied to the quality of Four Square location data and its availability of data. We would need to assume that Four Square would have a good coverage and representation for the venues in both cities. At the same time, the users from both cities are equally active and familiar with the tools from Four Square.