Battle of the Neighborhoods

Potential location for new sushi restaurant in Toronto

1.Introduction

As an international city, Toronto is greatly populated and open-minded. And Japanese food is one of the most well-known meal categories all over the world. Sushi is a traditional Japanese dish and always be treated as high-level cuisine, which stands for being healthy, aesthetical and cultural, becoming increasingly popular. Sushi restaurant could be a business with large profitable potential. However, running a restaurant depends on a lot of conditions and location is one of the keys for its success.

The object of this project is to use Foursquare location data, getting venues distribution information. We aim to figure out the most ideal location for starting a sushi-restaurant business in Toronto, providing investors some useful suggestions.

2.Target Audience

This project is aimed towards investors or business owners who want to open a new sushi restaurant in Toronto. Our analysis would be a good reference and provide some fundamental information for our target audience.

3.Data Overview

Firstly, we grab information from Wikipedia website. The website is about of list of postal codes of Canada: M, and we scrape the list table, clean and organize and get the information of boroughs and neighborhoods of Toronto; We got according latitude and longitude information of each neighborhood from https://cocl.us/Geospatial_data. The neighborhood location information would be used to get venues information from foursquare website. Then, we kept only sushi information from foursquare and use them to calculate the condition of sushi restaurant in Toronto and make the final business recommendation and decision.

1.Toronto Neighborhoods Information Source https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

From Wikipedia website, the postal code, borough and name of neighborhood information of Toronto would be grabbed. And the data would be clean and stored in data frame as follow:

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

2.Geographical Location Information Source https://cocl.us/Geospatial_data

We got the location information of each neighborhood from geospatial data provided by IBM cognitive class. The data was stored in csv file and we use pandas library to read it and stored in data frame. The data in the file includes postal code, latitude and longitude.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

We merged the data frame with borough and neighborhood data frame based on postal code. The results were as follows:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031

3.Venue Data Information source –Foursquare

We queried venues information from foursquare based on geographical location of each neighborhood in Toronto. And we stored the data in data frame as follow:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Corktown Common	43.655618	-79.356211	Park

Then, we prepared data by using one hot coding to analyze venue categories, grouping rows by neighborhood and taking the mean of the frequency of occurrence of each category. And we keep venue category 'sushi restaurant' information as our target analyzing object for later data clustering work.

4.Methodology

4.1 Data Cleaning

We scraped data of Toronto neighborhoods from Wikipedia and stored them in pandas data frame. We only process the cells which the borough is assigned. The boroughs ‘Not assigned’ were ignored. For the neighborhoods cell which is ‘Not assigned’, we kept them be same as the borough they belong to. The processed data frame was as follow:

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

We added the geographical location data to each neighborhood, and we only kept the cells which the borough name contains ‘Toronto’.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031

We used folium library to visualize the distribution of the neighborhoods in Toronto

We used Foursquare API to query the venues located in each neighborhood of Toronto. The results were shown as follow:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Corktown Common	43.655618	-79.356211	Park

4.2 Machine Learning

4.2.1 Data prepared

The venue data was processed using one hot coding method to transform 'venue category' into numeric data.

Neighborhood	New American Restaurant	Nightclub	Noodle House	Office	Opera House	Optical Shop	Organic Grocery	Other Great Outdoors	Park	Performing Arts Venue	Pet Store	Pharmacy	Pizza Place	Playground	Plaza	Poke Place	Poutine Place	Pu
Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Regent Park, Harbourfront	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	

Next, we grouped rows by neighborhood and took the mean of the frequency of occurrence of each category.

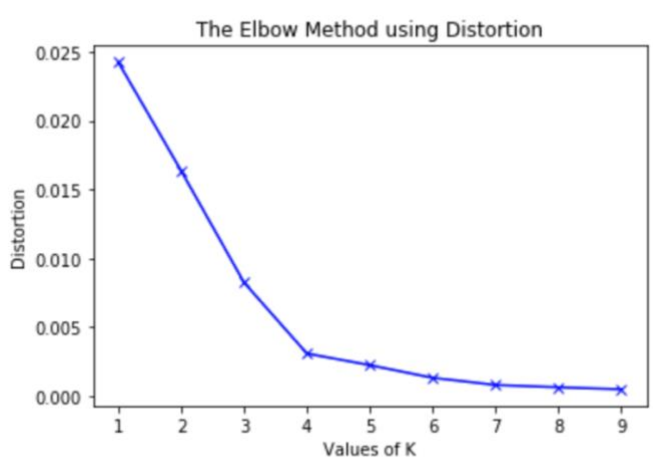
	Neighborhood	Yoga Studio	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Workshop	
0	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.017544	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
1	Brockton, Parkdale Village, Exhibition Place	0.041667	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	0.000000	0.0
2	Business reply mail Processing Centre, South C...	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	0.066667	0.0
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.058824	0.058824	0.058824	0.117647	0.176471	0.117647	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0

After, we extracted the neighborhood and sushi restaurant data into a new data frame.

	Neighborhood	Sushi Restaurant
0	Berczy Park	0.017544
1	Brockton, Parkdale Village, Exhibition Place	0.000000
2	Business reply mail Processing Centre, South C...	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
4	Central Bay Street	0.014493

4.2.2 K-means Clustering

We use machine learning to cluster the neighborhood into several clusters based on their sushi restaurant occurrence frequency by k-means clustering. To start the k- means clustering, we need to pick the optimal k value in case the model is underfitting or overfitting. The Elbow Point technique was used to test the best k.



From the figure, it is clear that the elbow point was at k = 4. In this case, we would separate the neighborhoods into four clusters.

We then built the model and assigned the neighborhoods into 4 different clusters, and we assigned the labels to the according rows.

	Neighborhood	Sushi Restaurant	Labels
0	Berczy Park	0.017544	3
1	Brockton, Parkdale Village, Exhibition Place	0.000000	1
2	Business reply mail Processing Centre, South C...	0.000000	1
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	1
4	Central Bay Street	0.014493	3

4.2.3 Cluster visualization

Cluster 1 (label0) --Purple
Cluster 2 (label1) --Bright cyan
Cluster 3 (labe2) --Yellow-green
Cluster 4 (label3) –Red

5.Results

At the end of the analysis, we examined each cluster and concluded the feature of each cluster.

Cluster 1

	Neighborhood	Sushi Restaurant	Labels	Latitude	Longitude
6	Church and Wellesley	0.054054	0	43.665860	-79.383160
8	Davisville	0.058824	0	43.704324	-79.388790
16	India Bazaar, The Beaches West	0.055556	0	43.668999	-79.315572
28	Runnymede, Swansea	0.085714	0	43.651571	-79.484450
33	Summerhill West, Rathnelly, South Hill, Forest...	0.062500	0	43.686412	-79.400049

There are 5 neighborhoods belongs to cluster 1. Based on the average frequency of sushi restaurant occurrence, cluster 1 can be defined as neighborhood cluster with high frequency sushi-restaurant occurrence, which is between 0.05-0.09. This data results indicates that the neighborhood assigned in cluster 1 are the neighborhood where sushi restaurants are densely distributed.

Cluster 2

	Neighborhood	Sushi Restaurant	Labels	Latitude	Longitude
1	Brockton, Parkdale Village, Exhibition Place	0.000000	1	43.636847	-79.428191
2	Business reply mail Processing Centre, South C...	0.000000	1	43.662744	-79.321558
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	1	43.628947	-79.394420
5	Christie	0.000000	1	43.669542	-79.422564
7	Commerce Court, Victoria Hotel	0.000000	1	43.648198	-79.379817
9	Davisville North	0.000000	1	43.712751	-79.390197
10	Dufferin, Dovercourt Village	0.000000	1	43.669005	-79.442259
13	Garden District, Ryerson	0.000000	1	43.657162	-79.378937
14	Harbourfront East, Union Station, Toronto Islands	0.010000	1	43.640816	-79.381752
15	High Park, The Junction South	0.000000	1	43.661608	-79.464763
17	Kensington Market, Chinatown, Grange Park	0.000000	1	43.653206	-79.400049
18	Lawrence Park	0.000000	1	43.728020	-79.388790
19	Little Portugal, Trinity	0.000000	1	43.647927	-79.419750
20	Moore Park, Summerhill East	0.000000	1	43.689574	-79.383160
21	North Toronto West, Lawrence Park	0.000000	1	43.715383	-79.405678
22	Parkdale, Roncesvalles	0.000000	1	43.648960	-79.456325
24	Regent Park, Harbourfront	0.000000	1	43.654260	-79.360636
26	Rosedale	0.000000	1	43.679563	-79.377529

27	Roselawn	0.000000	1	43.711695	-79.416936
29	St. James Town	0.000000	1	43.651494	-79.375418
31	Stn A PO Boxes	0.010309	1	43.646435	-79.374846
32	Studio District	0.000000	1	43.659526	-79.340923
34	The Annex, North Midtown, Yorkville	0.000000	1	43.672710	-79.405678
35	The Beaches	0.000000	1	43.676357	-79.293031

There are 24 neighborhoods assigned in cluster 2, and from the result of the average frequency of sushi restaurant occurrence, which showed 23 neighborhoods has value 0.0 and only 1 neighborhood has value around 0.01. The average frequency of sushi restaurant value in this cluster indicates that the neighborhoods assigned in clusters has almost 0 sushi restaurant in their area. The distribution of the sushi restaurant is extremely sparsely.

Cluster 3

	Neighborhood	Sushi Restaurant	Labels	Latitude	Longitude
12	Forest Hill North & West, Forest Hill Road Park	0.25	2	43.696948	-79.411307

Cluster 3 has only 1 neighborhood included, which the value of average frequency of sushi restaurant occurrence is highest among the four clusters. The value is 0.25, much higher than any of the value of neighborhood in the other three cluster.

Cluster 4

	Neighborhood	Sushi Restaurant	Labels	Latitude	Longitude
0	Berczy Park	0.017544	3	43.644771	-79.373306
4	Central Bay Street	0.014493	3	43.657952	-79.387383
11	First Canadian Place, Underground city	0.020000	3	43.648429	-79.382280
23	Queen's Park, Ontario Provincial Government	0.033333	3	43.662301	-79.389494
25	Richmond, Adelaide, King	0.021053	3	43.650571	-79.384568
30	St. James Town, Cabbagetown	0.021277	3	43.667967	-79.367675
36	The Danforth West, Riverdale	0.023810	3	43.679557	-79.352188
37	Toronto Dominion Centre, Design Exchange	0.020000	3	43.647177	-79.381576
38	University of Toronto, Harbord	0.028571	3	43.662696	-79.400049

The cluster 4 includes 9 neighborhoods. The mean frequency of the sushi restaurant is around 0.1-0.4, which means the value is in a middle level according to other three clusters information. This indicates that the neighborhoods assigned in cluster 4, their sushi restaurant distribution is not dense or sparse.

6.Discussion

According to the results, it is obvious that the distribution of sushi restaurant in different neighborhood has different feature. The feature could be separated into four clusters. Very few(sparsely), middle level (not dense nor sparse), densely distributed, very densely distributed. We noted that the neighborhoods which the sushi restaurants are densely distributed (cluster 1) and extremely densely distributed (cluster 3), have a smaller number of neighborhoods, which is 5 and 1, respectively. In contrast, the number of neighborhoods which the sushi restaurants are sparsely distributed (cluster 1) are great – 24, which indicates that there are a lot of neighborhoods which do not have a sushi restaurant or need more sushi restaurant. There are a great number of neighborhoods which have a potential need on sushi restaurant in another word. In addition, there are 9 neighborhoods which the mean frequency of sushi restaurant occurrence is middle level, which means these neighborhoods have a distribution of sushi restaurant which is not sparse nor dense.

The data we used for analyzing only including sushi restaurant. There are still a great many of information to be consider for a new sushi restaurant business, such as population of the neighborhood, age group, culture, education, income, etc. As our present result, the neighborhood assigned in cluster 2 has an extreme sparsely sushi restaurant distribution, and they are more recommended to be ideal location to start a sushi restaurant business for lack of competition. Also, the neighborhoods assigned in cluster 4 can be the second recommendation for a sushi restaurant business, for there maybe have more demand for sushi restaurants. It is because firstly, the neighborhoods in cluster 2, they already have some sushi restaurants in there, which means people living in these areas have interest and demand for this kind of restaurant. In the same time, the distribution of this kind of restaurant is not very dense at this time, so it is supposed to be not much competition. More research needed to be processed to refine our results.

7.Conclusion

In the study, we collected data of neighborhood in Toronto, their geographical location, venues information from Foursquare API. We cleaned the data use k-means clustering machine learning method to cluster the average frequency of occurrence of sushi restaurant in each neighborhood, in order to figure out the distribution of sushi restaurant in each neighborhood. Based on the result, we found the neighborhoods which has less density of sushi restaurant distribution which is more suitable for a new sushi restaurant business in Toronto. The analysis method can be used in a more extensive range, such as starting a Italian restaurant business, or even open a new grocery store, etc. This study can be an instruction for people to tackle with business location problem or explore for more complex realistic ideas.