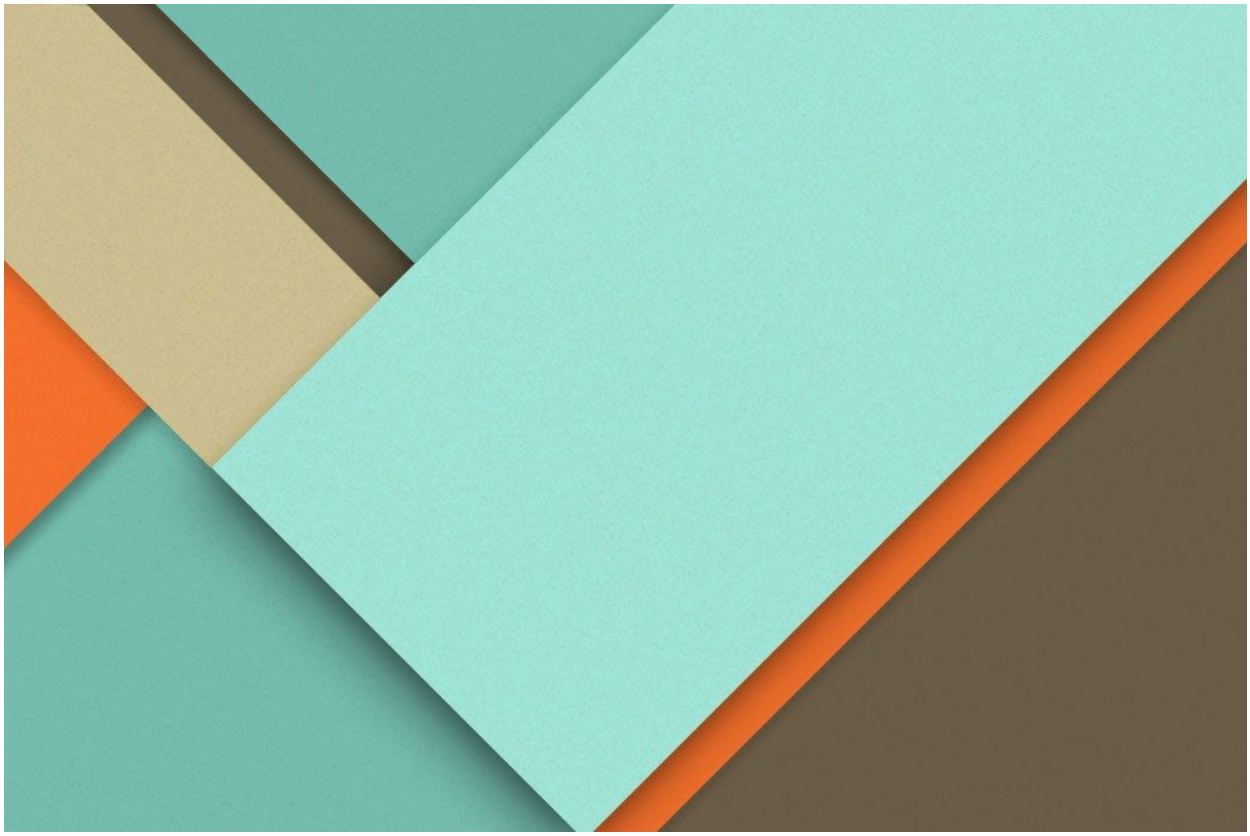


IBM Capstone Project

Open a Chinese Restaurant in Toronto, Canada

Ruoyu Zhang
email: zoezhang41@gmail.com

Table of Contents

1. Introduction

1.1 Background

Toronto is one of the most diverse cities in the world. The local residents come from more than 100 nationalities and speak more than 140 different languages. Toronto has the largest Chinese ethnic group outside of Asia, and the Chinese restaurant here gathers Sichuan, Hunan, Cantonese and even northern dishes. These Chinese restaurants are located in Chinatown in the city center, as well as Scarborough, Richmond Hill, Markham and Mississauga in the north. Chinese cuisine attracts many tourists and local residents in Toronto.

1.2 Business Understanding and Target Audience

The objective of this project is to find the most suitable location for the client to open a Chinese restaurant in Toronto. I will use machine learning skills such as K-means to analyse data extracted from various sources to find the best solution for my client. My target audience is entrepreneur who wants to open a Chinese restaurant in Toronto, and need suggestions on the best location to operate the restaurants.

2. Data

2.1 Neighborhoods data

Scarpping list of neighborhoods in Toronto from Wikipedia.

Neighbourhood data link

2.2 Latitude and Longitude data

Getting the geographical coordinates data of each postal code from a csv file.

Geographical coordinates data

2.3 Venue data

Getting venue data from Foursquare API.

3. Methodology

3.1 Data Preparation

Firstly, I used Beautifulsoup package in Python to compile the raw neighborhood data from Wikipedia. Then, I extracted geographical coordinates from csv file, and merge with the neighborhood data. Now, the primary table look like this:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

After that, I used Folium package to visualize the map of Toronto along with the marked coordinates of its neighborhoods.

3.2 Foursquare API

I used Foursquare API to pull the list of top 100 venues within 500 meters around each coordinates. A new dataframe including venues characteristics were retreived.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

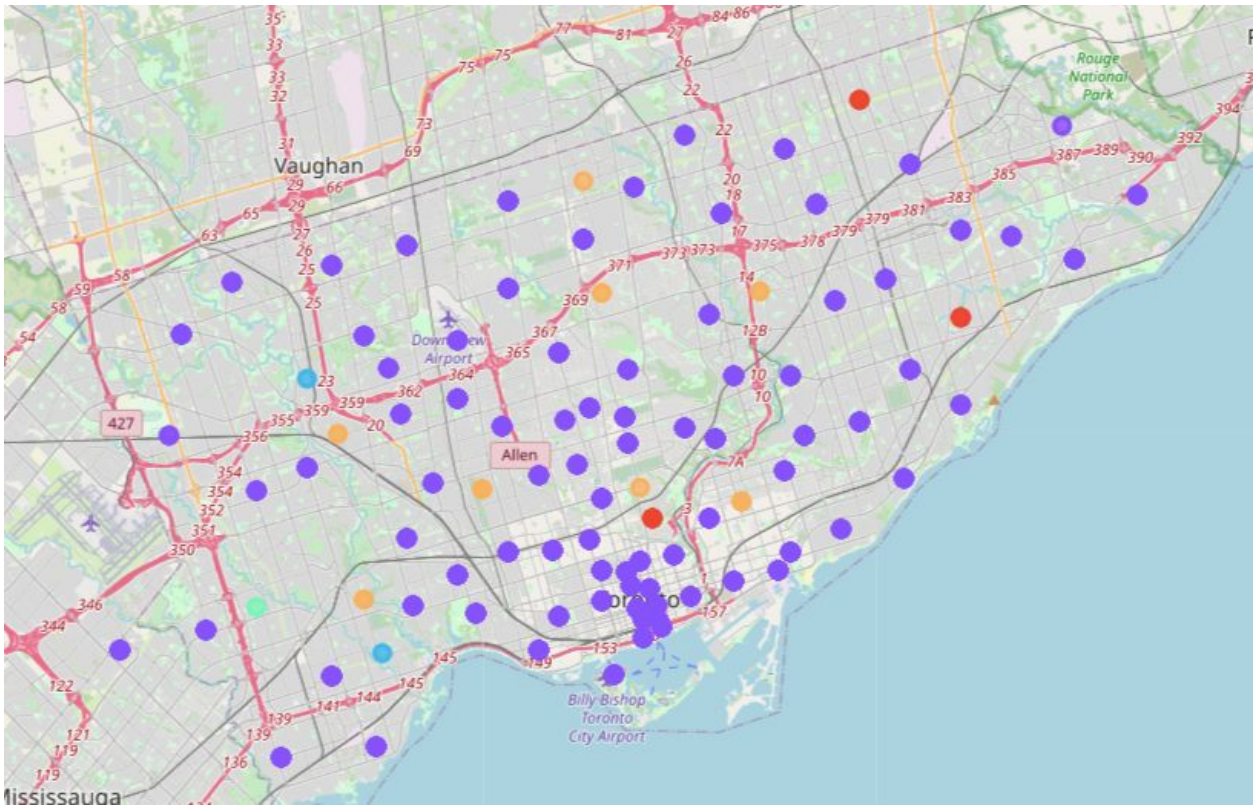
3.3 One-hot Encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to machine learning algorithms to do a better job in prediction.

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery
0	Agincourt	0.000000	0.000000	0.000000	0.000000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.000000
1	Alderwood, Long Branch	0.000000	0.000000	0.000000	0.000000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.000000
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000	0.000000	0.000000	0.000000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.000000
3	Bayview Village	0.000000	0.000000	0.000000	0.000000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.000000
4	Bedford Park, Lawrence Manor East	0.000000	0.000000	0.000000	0.000000	0.0000	0.000	0.0000	0.000	0.041667	0.000000	0.00	0.000000

3.4 Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. I used k-means to cluster the neighborhood into 5 clusters. With Folium, I created map of Toronto with five clusters.



4. Results

The five clusters are shown below:

Cluster 0

	Neighborhood	Chinese Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
50	Milliken, Agincourt North, Steeles East, L'Amo...	0.0	0	43.815252	-79.284577	Port Royal Park	43.815477	-79.289773	Park
50	Milliken, Agincourt North, Steeles East, L'Amo...	0.0	0	43.815252	-79.284577	Milliken Public School Playground	43.815383	-79.289867	Playground
65	Rosedale	0.0	0	43.679563	-79.377529	Rosedale Park	43.682328	-79.378934	Playground
65	Rosedale	0.0	0	43.679563	-79.377529	Whitney Park	43.682036	-79.373788	Park
65	Rosedale	0.0	0	43.679563	-79.377529	Alex Murray Parkette	43.678300	-79.382773	Park
65	Rosedale	0.0	0	43.679563	-79.377529	Milkman's Lane	43.676352	-79.373842	Trail
70	Scarborough Village	0.0	0	43.744734	-79.239476	McCowan Park	43.745089	-79.239336	Playground
70	Scarborough Village	0.0	0	43.744734	-79.239476	Tumbe Cafe	43.744058	-79.244021	Grocery Store

Cluster 1

	Neighborhood	Chinese Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agincourt	0.000000	1	43.794200	-79.262029	Panagio's Breakfast & Lunch	43.792370	-79.260203	Breakfast Spot
0	Agincourt	0.000000	1	43.794200	-79.262029	Twilight	43.791999	-79.258584	Lounge
0	Agincourt	0.000000	1	43.794200	-79.262029	El Pulgarcito	43.792648	-79.259208	Latin American Restaurant
0	Agincourt	0.000000	1	43.794200	-79.262029	Mark's	43.791179	-79.259714	Clothing Store
0	Agincourt	0.000000	1	43.794200	-79.262029	Commander Arena	43.794867	-79.267989	Skating Rink
1	Alderwood, Long Branch	0.000000	1	43.602414	-79.543484	Il Paesano Pizzeria & Restaurant	43.601280	-79.545028	Pizza Place
1	Alderwood, Long Branch	0.000000	1	43.602414	-79.543484	Timothy's Pub	43.600165	-79.544699	Pub
1	Alderwood, Long Branch	0.000000	1	43.602414	-79.543484	Toronto Gymnastics International	43.599832	-79.542924	Gym

Cluster 2

Cluster 2

```
trt_merged.loc[trt_merged['Cluster Labels'] == 2]
```

	Neighborhood	Chinese Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
39	Humberlea, Emery	0.0	2	43.724766	-79.532242	Strathburn Park	43.721765	-79.532854	Baseball Field
58	Old Mill South, King's Mill Park, Sunnylea, Hu...	0.0	2	43.636258	-79.498509	Woodford Park	43.633152	-79.496266	Baseball Field

Cluster 3

Cluster 3

```
trt_merged.loc[trt_merged['Cluster Labels'] == 3]
```

	Neighborhood	Chinese Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
86	West Deane Park, Princess Gardens, Martin Grov...	0.0	3	43.650943	-79.554724	Seaforth Golf Club	43.651183	-79.556107	Golf Course

Cluster 4

Cluster 4

```
trt_merged.loc[trt_merged['Cluster Labels'] == 4]
```

	Neighborhood	Chinese Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
10	Caledonia-Fairbanks	0.0	4	43.689026	-79.453512	SNSD's home	43.690405	-79.455011	Café
10	Caledonia-Fairbanks	0.0	4	43.689026	-79.453512	Nairn Park	43.690654	-79.456300	Park
10	Caledonia-Fairbanks	0.0	4	43.689026	-79.453512	Maximum Woman	43.690651	-79.456333	Women's Store
10	Caledonia-Fairbanks	0.0	4	43.689026	-79.453512	Fairbank Memorial Park	43.692028	-79.448924	Park
26	East Toronto, Broadview North (Old East York)	0.0	4	43.685347	-79.338106	Aldwych Park	43.684901	-79.341091	Park
26	East Toronto, Broadview North (Old East York)	0.0	4	43.685347	-79.338106	The Path	43.683923	-79.335007	Park
26	East Toronto, Broadview North (Old East York)	0.0	4	43.685347	-79.338106	Sammon Convenience	43.686951	-79.335007	Convenience Store
26	East Toronto, Broadview North (Old East York)	0.0	4	43.685347	-79.338106	Donlands Subway Station	43.680960	-79.337759	Metro Station

5. Discussion

After examining five clusters, cluster 1 has the most Chinese restaurants such as Bayview Village, North Toronto Wes, etc. The cluster results suggests that the client can open a Chinese restaurant in potential neighborhoods with less Chinese restaurant. Cluster 4 has several good potential neighborhoods because they are close to cluster but have no Chinese restaurant. This way, client can gain profit while minimize competition.

6. Conclusion

In this project, I discussed the potential location for an entrepreneur to open a Chinese restaurant in Toronto, Canada. I used neighborhood data, geographical corrdinates data, and Foursquare API to create a map of Toronto. Then used machine learning algorthm to predict the pontential location for the client. One improvement could be include population density to my model, which can increase the overall prediction accuracy.