Opening a Chinese restaurant in Toronto

1. Introduction

1.1 Problem Description and Background

In this project, we will focus on the discussion of starting a business in a city, more specifically, opening a Chinese restaurant in Toronto. It is not an easy decision for a business man who is not very familiar with the city. Usually, we need to consider the following factors at least when opening a restaurant:

- Location: where to open the restaurant? We have to investigate the distribution of restaurants in a city. Usually, it is good place to open a new restaurant close to a group of restaurants, as there are already a large flow of people coming to eat there every day, but the competition may also be a challenge, so we need to have differentiation in restaurant style.
- Style of restaurant: it depends on what kind of restaurant we are good at. Let's choose Chinese restaurant in this project.
- Return of investment: this is a quite complicated problem, we need to know the investment of renting a house, hiring people, etc. as well as a prediction of revenue/profit based on the revenue/profit data of other restaurants. As we don't have the data, we won't cover this part in this project.

In a nutshell, the problem we will address in this project is where is a good place to open a Chinese restaurant in Toronto. I think most people who are planning to start a restaurant business in Toronto would have interest to this analysis.

1.2 Data Description

The data I will use for the analysis includes:

- a Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M that has all the information we need to explore and cluster the postal codes, boroughs and neighborhoods in Toronto. We will scrape the Wikipedia page and wrangle the data, clean it, and then read it into a *pandas* dataframe.
- a csv file that has the geographical coordinates of each postal code of Toronto city: http://cocl.us/Geospatial_data
- Forsquare API to get the most common venues of given neighborhoods of Toronto city.

Based on the venue data, I can pull out the restaurant information in each neighborhood of Toronto, then make statistics how many restaurants in each neighborhood and the style of

each restaurant. I think the good place to open Chinese restaurant is the neighborhood with most restaurants but very few Chinese restaurant.

2. Methodology

Firstly, I got the neighborhood info of Toronto city from the Wikipedia page, cleaned the data and read it into a *pandas* dataframe. Then read the geographical coordinates of each postal code from a csv file, and combined both info into one dataframe as follows:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	East Toronto,Riverdale	43.679557	-79.352188
2	M4L	East Toronto	East Toronto,India Bazaar	43.668999	-79.315572
3	M4M	East Toronto	East Toronto	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
5	M4P	Central Toronto	Central Toronto	43.712751	-79.390197
6	M4R	Central Toronto	Central Toronto	43.715383	-79.405678
7	M4S	Central Toronto	Central Toronto	43.704324	-79.388790
8	M4T	Central Toronto	Moore Park,Central Toronto	43.689574	-79.383160
9	M4V	Central Toronto	Deer Park,Central Toronto,Rathnelly,South Hill...	43.686412	-79.400049

After that, I utilized the Foursquare API to explore the neighorhoods and segment them. I designed the limit as **100 venue** and the radius **500 meter** for each neighborhood from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude info from Forsquare API.

	PostalCode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M4E	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	M4E	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	M4E	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	M4E	43.676357	-79.293031	Glen Stewart Ravine	43.676300	-79.294784	Other Great Outdoors
4	M4E	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood

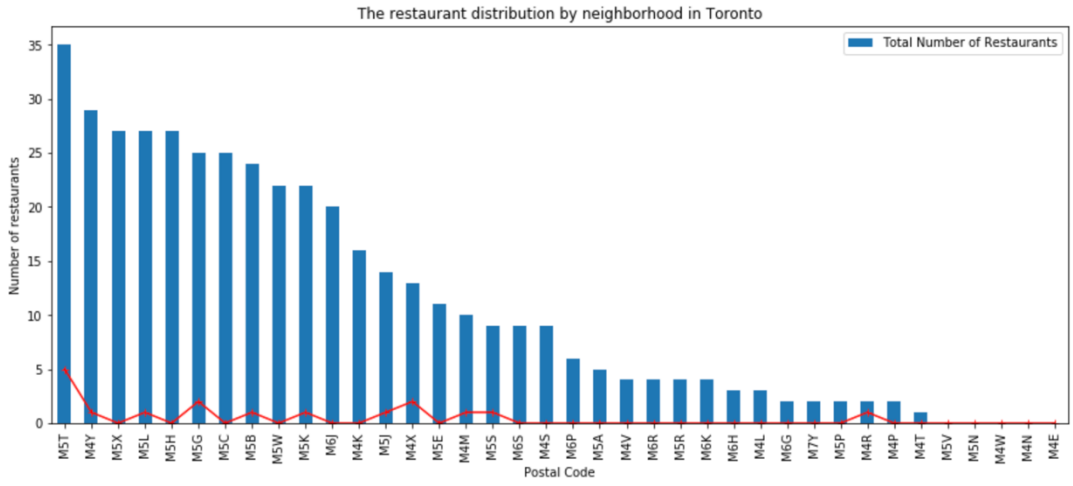
As we care about only the restaurant info in the venue category, I created a dataframe to calculate the number of different kinds of restaurants in each neighborhood by postal code. Here is a head of the dataframe:

	PostalCode	Afghan Restaurant	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Colombian Restaurant	...	Restaurant	Seafood Restaurant	Southern / Soul Food Restaurant	S Restau
0	M4E	0	0	0	0	0	0	0	0	0	...	0	0	0	
1	M4K	0	1	0	0	0	0	1	0	0	...	1	0	0	
2	M4L	0	0	0	0	0	0	0	0	0	...	0	0	0	
3	M4M	0	2	0	0	0	0	0	1	0	...	0	1	0	
4	M4N	0	0	0	0	0	0	0	0	0	...	0	0	0	

Next step, I summarized the total number of restaurants to compare with the number of Chinese restaurants in each neighborhood, listed by descending of total number of restaurants.

	PostalCode	Neighborhood	Chinese Restaurant	Total Number of Restaurants
0	M5T	Chinatown,Grange Park,Kensington Market	5	35
1	M4Y	Church and Wellesley	1	29
2	M5X	First Canadian Place,Underground city	0	27
3	M5L	Commerce Court,Downtown Toronto	1	27
4	M5H	Downtown Toronto,Downtown Toronto,Downtown Tor...	0	27
5	M5G	Downtown Toronto	2	25
6	M5C	St. James Town	0	25
7	M5B	Downtown Toronto,Downtown Toronto	1	24
8	M5W	Downtown Toronto	0	22
9	M5K	Design Exchange,Toronto Dominion Centre	1	22

In order to demonstrate the info in visualization mode, I used a bar chart to show the total number of restaurants, and a line to show the number of Chinese restaurants by postal code in the following:

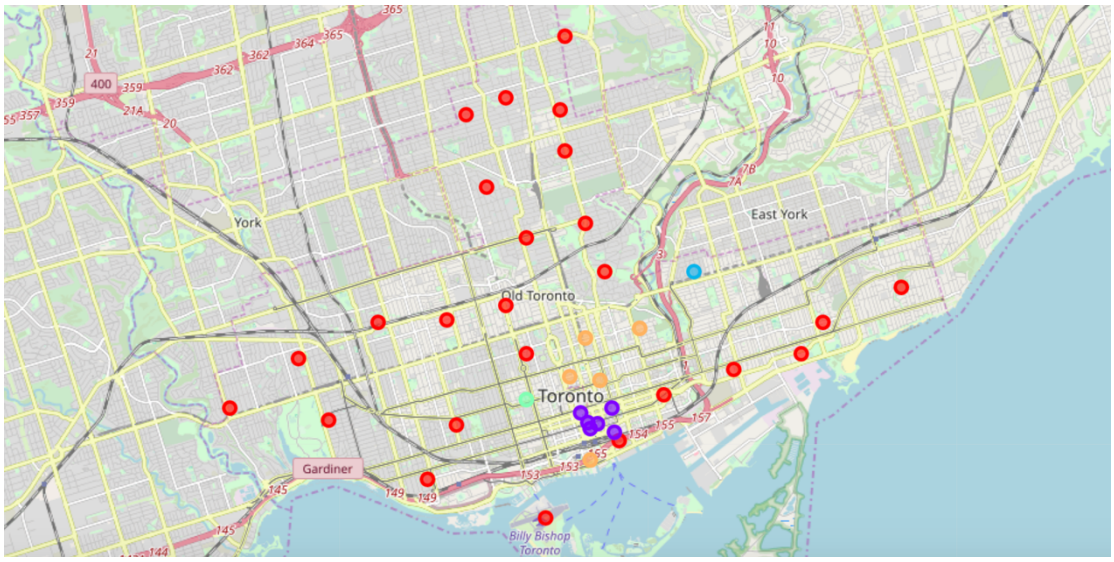


According to our investment philosophy, we tend to open a Chinese restaurant in a place where has more restaurants but with very few Chinese restaurant, so from the picture above, we can find some good places like M5X - First Canadian Place, Underground city, M5H – Downtown Toronto, etc.

To find out all the good places for opening a Chinese restaurant, I used unsupervised learning **K-Means algorithm** to cluster the neighborhoods. Based on the restaurant data in each neighborhood, I run K-Means to cluster the neighborhoods into 5 clusters. I also found out the top 5 most common restaurant for each neighborhood. Here is a list of the neighborhoods by clustering with most common restaurant:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant
0	M4E	East Toronto	The Beaches	43.676357	-79.293031	0	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Greek Restaurant	Gluten-free Restaurant	German Restaurant
1	M4K	East Toronto	East Toronto,Riverdale	43.679557	-79.352188	2	Greek Restaurant	Italian Restaurant	Caribbean Restaurant	Restaurant	Indian Restaurant
2	M4L	East Toronto	East Toronto,India Bazaar	43.668999	-79.315572	0	Italian Restaurant	Sushi Restaurant	Fast Food Restaurant	Vietnamese Restaurant	Doner Restaurant
3	M4M	East Toronto	East Toronto	43.659526	-79.340923	0	American Restaurant	Italian Restaurant	Thai Restaurant	Seafood Restaurant	Chinese Restaurant
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	0	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Greek Restaurant	Gluten-free Restaurant	German Restaurant
5	M4P	Central Toronto	Central Toronto	43.712751	-79.390197	0	Asian Restaurant	Restaurant	Vietnamese Restaurant	Doner Restaurant	Gluten-free Restaurant
6	M4R	Central Toronto	Central Toronto	43.715383	-79.405678	0	Chinese Restaurant	Mexican Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Gluten-free Restaurant
7	M4S	Central Toronto	Central Toronto	43.704324	-79.388790	0	Sushi Restaurant	Italian Restaurant	Indian Restaurant	Restaurant	Seafood Restaurant
8	M4T	Central Toronto	Moore Park,Central Toronto	43.689574	-79.383160	0	Restaurant	Vietnamese Restaurant	Doner Restaurant	Gluten-free Restaurant	German Restaurant
9	M4V	Central Toronto	Deer Park,Central Toronto,Rathnelly,South Hill...	43.686412	-79.400049	0	Vietnamese Restaurant	American Restaurant	Sushi Restaurant	Restaurant	Doner Restaurant

Then I created a map to visualize the neighborhood clustering in the following. From the picture, we can see both M5H and M5X are in cluster 1 (purple mark), so we can think all the neighborhoods in cluster 1 are our recommendation for opening a Chinese restaurant.



3. Results

According to our analysis, the first recommendation for opening a Chinese restaurant is the neighborhoods in cluster 1. It includes 6 neighborhoods listed below, all those neighborhoods have god number of restaurants but without or with only one Chinese restaurants.

	PostalCode	Borough	Neighborhood	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant
15	M5C	Downtown Toronto	St. James Town	1	Restaurant	Italian Restaurant	American Restaurant	Japanese Restaurant	Seafood Restaurant
18	M5H	Downtown Toronto	Downtown Toronto,Downtown Toronto,Downtown Tor...	1	Thai Restaurant	American Restaurant	Asian Restaurant	Restaurant	Sushi Restaurant
20	M5K	Downtown Toronto	Design Exchange,Toronto Dominion Centre	1	Italian Restaurant	Restaurant	American Restaurant	Japanese Restaurant	Seafood Restaurant
21	M5L	Downtown Toronto	Commerce Court,Downtown Toronto	1	Restaurant	American Restaurant	Seafood Restaurant	Thai Restaurant	Vegetarian / Vegan Restaurant
28	M5W	Downtown Toronto	Downtown Toronto	1	Restaurant	Fast Food Restaurant	Seafood Restaurant	Italian Restaurant	Japanese Restaurant
29	M5X	Downtown Toronto	First Canadian Place,Underground city	1	Restaurant	American Restaurant	Asian Restaurant	Seafood Restaurant	Thai Restaurant

When looking at the data and map, I think the neighborhoods in cluster 4 (yellow mark) also are good place to be considered for opening a Chinese restaurant, for example, M4Y – Church and Wellesley has 29 restaurants but only one Chinese restaurant. So the neighborhoods in cluster 4 will be our second recommendation for opening a Chinese restaurant, it includes 5 neighborhoods listed below:

	PostalCode	Borough	Neighborhood	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant
11	M4X	Downtown Toronto	Cabbagetown,St. James Town	4	Restaurant	Italian Restaurant	Chinese Restaurant	Indian Restaurant	American Restaurant
12	M4Y	Downtown Toronto	Church and Wellesley	4	Japanese Restaurant	Sushi Restaurant	Restaurant	Fast Food Restaurant	Mediterranean Restaurant
14	M5B	Downtown Toronto	Downtown Toronto,Downtown Toronto	4	Fast Food Restaurant	Middle Eastern Restaurant	Ramen Restaurant	Italian Restaurant	Japanese Restaurant
17	M5G	Downtown Toronto	Downtown Toronto	4	Italian Restaurant	Indian Restaurant	Sushi Restaurant	Japanese Restaurant	Chinese Restaurant
19	M5J	Downtown Toronto	Downtown Toronto,Toronto Islands,Union Station	4	Italian Restaurant	Restaurant	Indian Restaurant	Japanese Restaurant	New American Restaurant

4. Discussion

Looks like the K-Means algorithm provides good clustering for the neighborhoods based on the restaurant data to support our purpose, but I also found some neighborhoods, such as M6J - Little Portugal, Trinity (has 20 restaurants in total without Chinese restaurant), M4K – East Toronto, Riverdale (has 16 restaurants in total without Chinese restaurant), should also be good place for opening a Chinese restaurant, which has not been included into either cluster 1 or cluster 4. In this project, the restaurant data we used for K-Means algorithm is just the number of different style of restaurants in each neighborhood, that means we didn't consider too much the number of Chinese restaurants, so I think it would be more accurate if we could increase the weight of number of Chinese restaurants by multiply a factor when using the K-Means for clustering.

Whatever, our model is still too simple in this project. As I mentioned in the introduction, in addition to the location, we also need to consider a lot of other factors when opening a restaurant, so it is a very complicated problem in a real world.

5. Conclusion

In this capstone project, I used the data from Wikipedia page and Forsquare API to implement a solution for a real business problem, which demonstrate the power of data analysis, data visualization, machine learning in Python. It provides good recommendation for the business problem although the model is still needed to be enhanced.

I think this is a very good exercise for a learner to build skills on data science.