Clustering strategy for finding potential optimal location of a Chinese restaurant in Toronto

Yilun Zhang

2021.05.21

1. Introduction

1.1 Background

Canadians who identify themselves as being of Chinese ethnic origin make up about 4.6% of the Canadian population, or about 1.57 million people according to the 2016 census. In addition, 11.1% of total population in Toronto are Chinese.

1.2 Business Problem

Under this background, opening a Chinese restaurant in Toronto seems to be a great business opportunity for our clients/stakeholders. However, according to my personal experience, not all Chinese restaurant can have a big success or live long in Toronto. Although there are many factors accounting for this situation, I think a cautious decision on restaurant location is the first thing needs to be considered for our clients/stakeholders. In this project we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening a Chinese restaurant in Toronto. By intuition, population, the number and types of existing Chinese restaurants should be taken into consideration when deciding a good location candidate.

2. Data

Based on the population of a neighborhood, the number of Chinese restaurants in a neighborhood and the types of Chinese restaurants in a neighborhood, we can use clustering strategy to segment these neighborhoods in Toronto and decide which neighborhoods our clients/stakeholders should choose when they want to start up a new Chinese restaurant in Toronto.

To be more specific, we will use web-crawling technique to get all neighborhoods in Toronto with the corresponding postal codes and boroughs. See example below:

	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	Queen's Park	Ontario Provincial Government

Then we will use the GeoSpatial Dataset given by Coursera to get the latitude and longitude coordinates of each neighborhood. See example below:

	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	Queen's Park	Ontario Provincial Government	43.662301	-79.389494

We will also use the census data to get the population information of each neighborhood downloaded from Canada open data portal. See example below:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Population
0	M3A	North York	Parkwoods	43.753259	-79.329656	34615
1	M4A	North York	Victoria Village	43.725882	-79.315572	14443
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	41078
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	21048
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494	10

Lastly, we will utilize Foursquare API to get the venue information of each neighborhood, or in other words, Chinese restaurant related information. See example below:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	The Bean Sprout	43.742229	-79.313577	Chinese Restaurant
1	Parkwoods	43.753259	-79.329656	Omni Palace Noodle House	43.771047	-79.331570	Chinese Restaurant
2	Parkwoods	43.753259	-79.329656	Dragon Pearl Buffet 龍珠	43.753693	-79.349730	Chinese Restaurant
3	Parkwoods	43.753259	-79.329656	China Gourmet	43.755189	-79.348382	Chinese Restaurant
4	Parkwoods	43.753259	-79.329656	Spicy Chicken House	43.760639	-79.325671	Chinese Restaurant
...	...	...	...	...	...	...	...
2044	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Ancom Chinese Restaurant	43.624481	-79.509448	Chinese Restaurant
2045	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Mandarin Buffet	43.621352	-79.523015	Chinese Restaurant
2046	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Li's Oriental Kitchen	43.637787	-79.539134	Chinese Restaurant
2047	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Lemongrass	43.645010	-79.522379	Thai Restaurant
2048	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Chinese Food Gallery	43.643905	-79.533167	Chinese Restaurant

3. Methodology

3.1 Data Cleaning

Firstly, we use web-crawling technique to get all neighborhoods in Toronto with the corresponding postal codes and boroughs. Note that, when we crawl the required information from Wikipedia, we only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned. More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 2 in the following table. If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

	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	Queen's Park	Ontario Provincial Government

After having the above information about neighborhoods, we will merge the coordinates and population information into the dataframe as following.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Population
0	M3A	North York	Parkwoods	43.753259	-79.329656	34615
1	M4A	North York	Victoria Village	43.725882	-79.315572	14443
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	41078
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	21048
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494	10

3.2 Exploratory data analysis

As we already have the coordinates information of each neighborhood, we can plot these neighborhood on a map of Toronto to check the distribution. In order to do this, we will use the Folium package in Python.

3.3 Feature engineering

After Data Cleaning section, there are still two more required features which are not included in our dataframe, namely, the number of Chinese restaurants in a neighborhood and the type of Chinese restaurants in a neighborhood. For getting these two features, we need to use Foursquare API. To be more specific, we search top 100 Chinese restaurant related venues within 2000 meters of the center of a neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	The Bean Sprout	43.742229	-79.313577	Chinese Restaurant
1	Parkwoods	43.753259	-79.329656	Omni Palace Noodle House	43.771047	-79.331570	Chinese Restaurant
2	Parkwoods	43.753259	-79.329656	Dragon Pearl Buffet 龍珠	43.753693	-79.349730	Chinese Restaurant
3	Parkwoods	43.753259	-79.329656	China Gourmet	43.755189	-79.348382	Chinese Restaurant
4	Parkwoods	43.753259	-79.329656	Spicy Chicken House	43.760639	-79.325671	Chinese Restaurant
...	...	...	...	...	...	...	...
2042	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509	Ancom Chinese Restaurant	43.624481	-79.509448	Chinese Restaurant
2044	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Ancom Chinese Restaurant	43.624481	-79.509448	Chinese Restaurant
2045	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Mandarin Buffet	43.621352	-79.523015	Chinese Restaurant
2046	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Li's Oriental Kitchen	43.637787	-79.539134	Chinese Restaurant
2048	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Chinese Food Gallery	43.643905	-79.533167	Chinese Restaurant

Using the table above, we can firstly calculate the number of Chinese restaurants in a neighborhood and add this feature into our dataframe with a variable name ‘Number of CN Restaurant’.

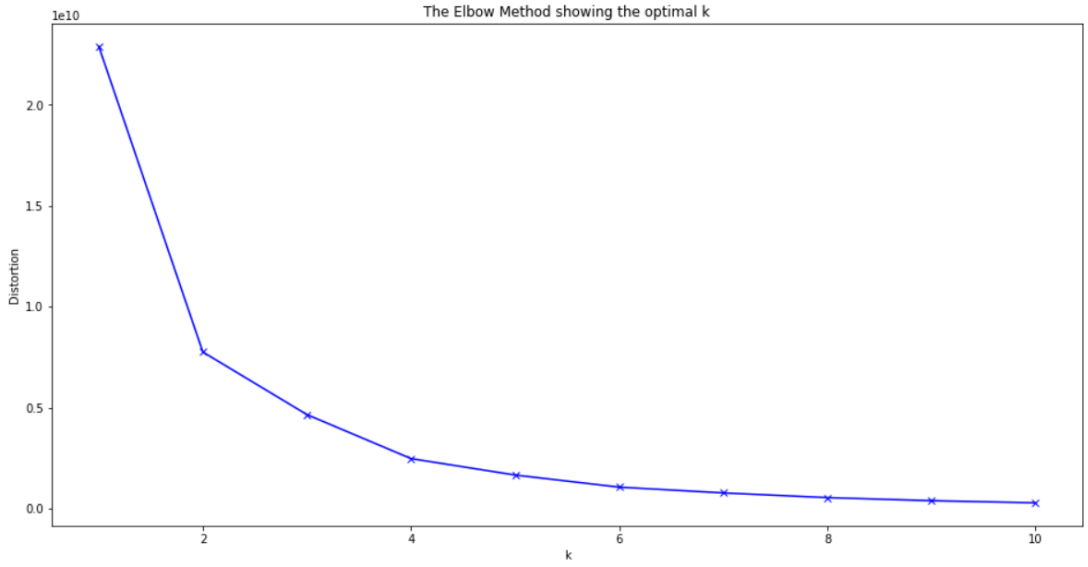
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Population	Distance from center	Number of CN Restaurant
0	M3A	North York	Parkwoods	43.753259	-79.329656	34615	11917.949753	7
1	M4A	North York	Victoria Village	43.725882	-79.315572	14443	9751.235551	5
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	41078	1881.525138	39
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	21048	9730.670477	5
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494	10	1077.658778	98
...	...	...	...	...	...	...	...	...
97	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944	10787	9923.318339	5
98	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	30472	1376.706656	95
99	M7Y	East Toronto Business	Enclave of M4L	43.662744	-79.321558	10	5135.793455	3
100	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509	21299	9440.207261	3
101	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	17038	11393.296323	4

Then we can use the same table to calculate mean of the frequency of occurrence of each category, i.e. different types of Chinese restaurants. The results are also required features as mentioned before. Note that, there are 4 neighborhoods which do not have Chinese restaurant within 2000 meters radius of the center of the neighborhood so they will be excluded in this step. This will be further discussed in Section 5.

	Neighborhood	Asian Restaurant	BBQ Joint	Cantonese Restaurant	Chinese Restaurant	Comfort Food Restaurant	Dim Sum Restaurant	Dongbei Restaurant	Dumpling Restaurant	Fried Chicken Joint	Hakka Restaurant	Hong Kong Restaurant	Hotpot Restaurant	Indian Chinese Restaurant	Noodle House	Peking Duck Restaurant
0	Agincourt	0.025641	0.0	0.102564	0.692308	0.0	0.051282	0.0	0.0	0.0	0.0	0.076923	0.0	0.0	0.0	0.025641
1	Alderwood, Long Branch	0.000000	0.0	0.000000	1.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
2	Bathurst Manor, Wilson Heights, Downsview North	0.500000	0.0	0.000000	0.500000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
3	Bayview Village	0.000000	0.0	0.000000	0.750000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
4	Bedford Park, Lawrence Manor East	0.200000	0.0	0.000000	0.800000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000

3.4 Model

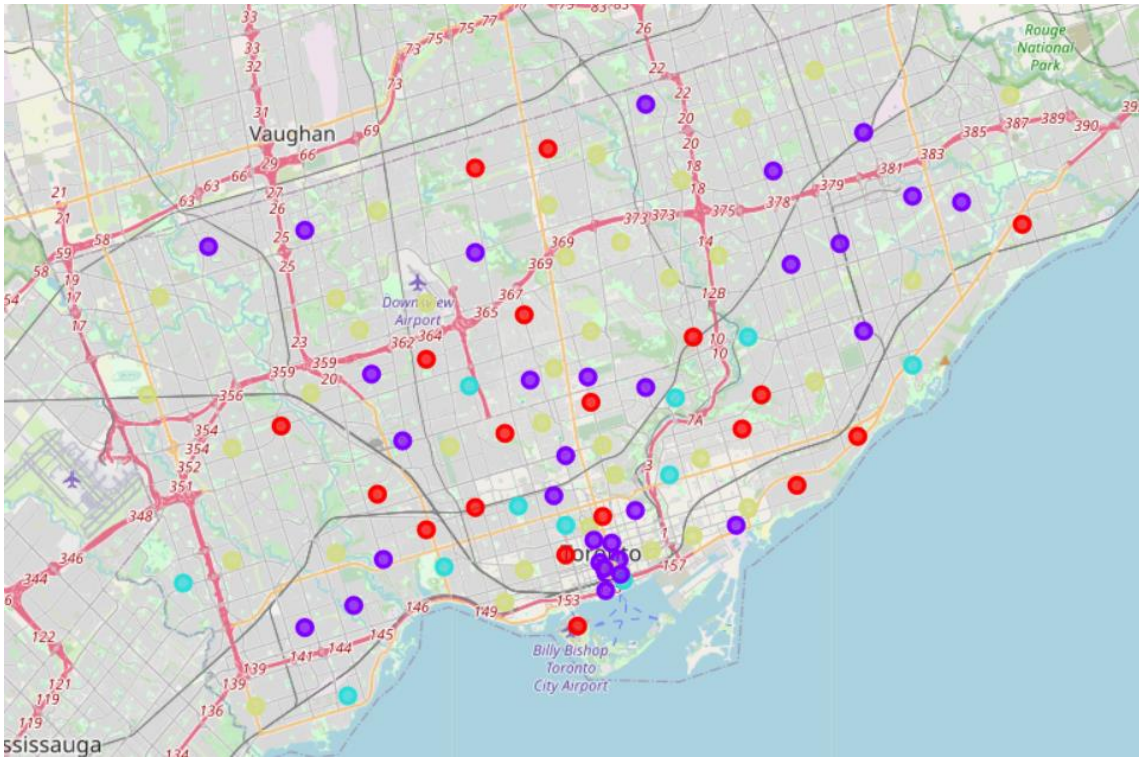
Now, our dataframe contains all features we need for analysis. In this section, we will use K-Means clustering algorithm to divide neighborhoods into clusters in order to find potential location for a new restaurant. Since K-Means clustering algorithm needs us to predefine the number of clusters K, we need to use the elbow method to find the optimal K first.



From the plot above we can see that, 4 probably is a good choice for our optimal K. Therefore, we use K-Means clustering algorithm to divide neighborhoods into 4 different clusters.

4. Results

Again, we can use Folium package to visualize the clustering results on a map of Toronto.



Legends of the map are following:
Red dots: Cluster 0
Purple dots: Cluster 1
Blue dots: Cluster 2
Yellow dots: Cluster 3

The following table shows some summary information of the clustering results.

	Number of neighborhoods	Mean population	Mean number of Chinese restaurants
Cluster 0	19	32343.37	14.47
Cluster 1	31	21400.74	31.03
Cluster 2	11	24531.64	17.36
Cluster 3	37	29074.70	14.38

5. Discussion

As mentioned in Section 3.3, there are 4 neighborhoods which do not have Chinese restaurant within 2000 meters radius of the center of the neighborhood. These 4

neighborhoods are ‘Islington Avenue’, ‘Rouge Hill, Port Union, Highland Creek’, ‘Humberlea, Emery’ and ‘Upper Rouge’. These 4 neighborhoods are all relatively remote areas of Toronto and this is probably one reason for no Chinese restaurants in these areas. We may need further analysis to decide whether these areas are qualified candidates. In terms of the clustering results, although we can not see a clear clustering pattern from the map alone, we can still get some insightful results with some additional background information about Toronto. The most interesting result should be Cluster 1 which has the highest mean number of Chinese restaurants and the lowest mean population. From the distribution of purple dots in the map we can see that, neighborhoods in Cluster 1 mainly gather in 4 areas, which are Downtown, North York, Etobicoke and Scarborough. This result makes sense since there are lots of universities and colleges in these 4 areas where many Chinese students are attending. For example, University of Toronto (St. George Campus and Scarborough Campus), York University and Humber College. Thus, even though the mean population is low in these areas, the number of Chinese restaurants are very high. This kind of popular regions should be a good choice for starting up a Chinese restaurants. In addition, both mean population and mean number of Chinese restaurants are medium for Cluster 2. And from the map we can see that neighborhoods in Cluster 2 are quite close to neighborhoods in Cluster 1 which explain the medium results. As far as I am concerned, neighborhoods in Cluster 2 might also be good choices due to their potential. Neighborhoods in the rest two cluster are relatively remote areas of Toronto so even though the market sizes are good thy may not be good choices.

6. Conclusion and future improvement

In conclusion, in this report, we find some potential optimal locations for starting up a new Chinese restaurant in Toronto using K-Means clustering algorithm. Popular regions near universities where many Chinese students are attending and regions near popular regions are some good choices. However, these results are still quite rough since location selection for a restaurant involves many more factors than what we used in this report. For example, rent and surrounding infrastructure. So this report is more like a demo. In order to make the results more promising, we need to collect more data to cover more factors and do further analysis.