Battle of Neighbourhoods in Toronto

Submitted to IBM (Applied Data Science Capstone)

By Zequn Wang

1. Introduction

Modern cities are composed of neighbourhoods with different demographics such as population, income level, education background, cultural diversity, etc. Such difference lead towards business opportunities that are dependent on locations. Take food for example, places represented by white collars are probably more appealed to delicate restaurants. This article aims to compare 140 neighbourhoods in Toronto Figure 1) and explore their chances of opening a generic restaurant.

The city of Toronto is one of the largest and most populated multicultural metropolises in North America. Where there are people, there are needs for gourmet food. Toronto is no different. By comparing 140 neighbourhoods by their current types of restaurants, this article will discuss how location influences food characteristics across the city. The discussions could serve as a solid reference to business developers or restaurant owners who are looking forward to opening a new restaurant in Toronto.

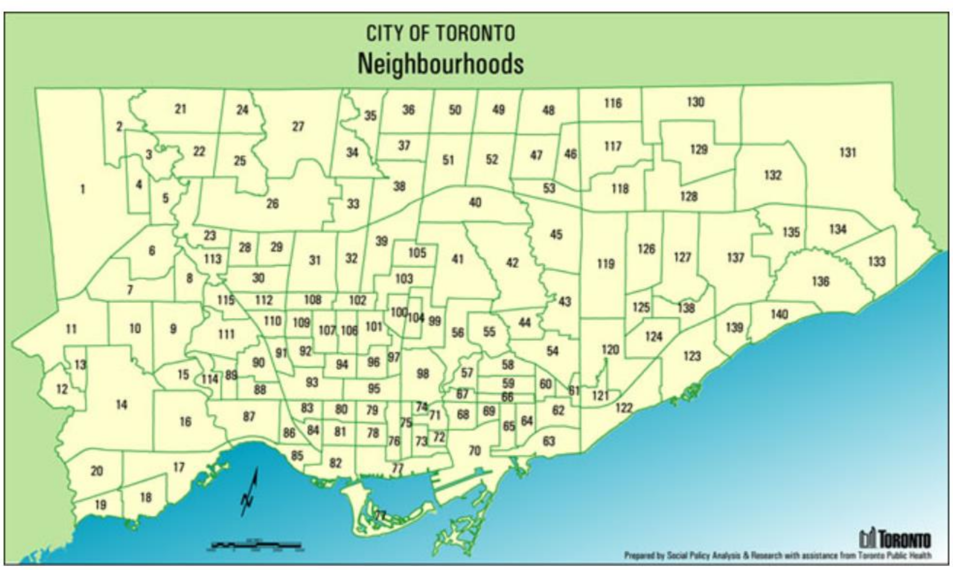

Figure 1 City of Toronto Neighbourhoods and Codes (source: https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/)

2. Data Description

Data used in the analysis come from two sources: City of Toronto Open Data Portal and Foursquare API Services. The key features of imported data include:

City of Toronto Open Data Portal:

- List of Neighbourhoods in Toronto and Codes
- Population in 2016
- Average individual income in 2015
- Geometry coordinates of each neighbourhood

Foursquare API Services:

- List of food venues around centroid of each neighbourhood
- Venue category
- Geometry coordinates of each venue

The food venues are retrieved from Foursquare at the centroid of each neighbourhood with a radius proportional to its area. Retrieved venues are carefully checked to ensure that their coordinates are located within the boundary of its associated neighbourhood. The same venue that is assigned to multiple neighbourhoods is scrutinized to remove duplicates.

With proper preprocessing, data from two sources are combined into one large dataset. The following Table 1 shows a selection of the dataset including some key features used for analysis.

Table 1 Dataset used for analysis (with selected features)

Neighbourhood	Venue	Venue Category	Population 2016	Individual Income 2015
Agincourt North	Saravanaa Bhavan South Indian Restaurant	Indian Restaurant	29113	25005
Agincourt North	Fahmee Bakery & Jamaican Foods	Caribbean Restaurant	29113	25005
Agincourt North	Grandeur Palace 華丽宫 (Grandeur Palace 華麗宫)	Chinese Restaurant	29113	25005
Agincourt North	Congee Town 太皇名粥	Chinese Restaurant	29113	25005
Agincourt North	Subway	Sandwich Place	29113	25005
Agincourt North	Pizza Pizza	Pizza Place	29113	25005
Agincourt North	Popeyes Louisiana Kitchen	Fried Chicken Joint	29113	25005
Agincourt North	Sushi Legend	Sushi Restaurant	29113	25005
Agincourt North	Subway	Sandwich Place	29113	25005
Agincourt North	Aromaz Cake and Pastry 龍騰閣	Bakery	29113	25005
Agincourt North	Pizza Pizza	Pizza Place	29113	25005
Agincourt North	Subway	Sandwich Place	29113	25005
Agincourt North	Lac Vien Vietnamese Restaurant- Scarborough	Vietnamese Restaurant	29113	25005
Agincourt South-Malvern West	The Roti Hut	Caribbean Restaurant	23757	20400
Agincourt South-Malvern West	Yummy Cantonese Restaurant 老西關腸粉	Cantonese Restaurant	23757	20400
Agincourt South-Malvern West	Mona's Roti	Caribbean Restaurant	23757	20400
Agincourt South-Malvern West	Congee Me 小米粥鋪	Chinese Restaurant	23757	20400
Agincourt South-Malvern West	Panagio's Breakfast & Lunch	Breakfast Spot	23757	20400
Agincourt South-Malvern West	Wonton Chai Noodle 雲吞仔	Noodle House	23757	20400
Agincourt South-Malvern West	One2 Snacks	Asian Restaurant	23757	20400

3. Methodology

To identify which neighbourhood is the most suitable for opening a restaurant, we must first determine factors that boost the demand for dining services. Intuitively, population, income, transportation, age distribution, race, gender could all contribute the need for restaurants in a certain area. Given the data availability and capacity of project, the author has chosen the two leading factors, i.e., population and income level, to differentiate different neighbourhoods.

Secondly, current restaurants in each neighbourhood should be explored to measure the level of competition across the city. Foursquare offers many metrics to evaluate a restaurant such as its type, price, rating, etc. Since the goal of this article is to identify a location for opening a generic restaurant, it is necessary to investigate the diversity and quantity of restaurants in each neighbourhood. Analysis of restaurants' details is out of scope for this article but can be performed in the future to determine what kind of restaurant to open.

Based on available data, we ought to compare neighbourhoods by their demand for restaurants and the current level of competition. As such, the analysis is designed to follow several steps:

a. Aggregate food venues (i.e., restaurants) for each neighbourhood
b. Cluster 140 neighbourhoods into five groups based on the category and frequency of different food venues located within the boundary of each neighbourhood.

c. Investigate the clustering of neighbourhoods based on food venues against population, individual income, and population income (individual income multiplied by population).
d. Identify neighbourhoods that have potential for opening more restaurants.

The clustering of neighbourhoods based on food venues will be accomplished using k-Means algorithm. k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster (https://en.wikipedia.org/wiki/K-means_clustering).

Another measure in additional to population and income is employed called population income, which is equal the multiplication of income by population. This metric can be interpreted as the accumulated annual income of a neighbourhood determined by both population and individual income. It incorporates population and individual income into one factor to evaluate different clusters of neighbourhoods.

4. Results & Discussion

Based on the categories of food venues, the 140 neighbourhoods are grouped into five clusters using k-Means method. A screenshot of the clustered neighbourhoods along with their cluster labels are shown in Table 2. The total number of neighbourhoods in each cluster is given in Table 3 along with the average no. of restaurants per neighbourhood, and first and second most popular food venues in each cluster. The results indicate that two main clusters (cluster 0 and 1) and three 1-sample clusters. Since the most neighbourhoods are grouped into cluster 0 and 1, the following discussion will focus on comparison of these two clusters.

Table 2 A screenshot of clustered neighbourhoods

	Neighbourhood	Code	Population	Income	Cluster Labels	Latitude	Longitude
0	Agincourt North	129	29113	25005	0	43.805441	-79.266712
1	Agincourt South-Malvern West	128	23757	20400	1	43.788657	-79.265611
2	Alderwood	20	12054	10265	0	43.604937	-79.541611
3	Annex	95	30526	26295	1	43.671585	-79.404000
4	Banbury-Don Mills	42	27695	23410	1	43.737658	-79.349720

Table 3 Cluster of 140 Neighbourhoods based on food venues

Cluster	No. of Neighbourhoods	Average no. of venues per neighbourhood	The first popular venue	The second popular venue
0	99	12	Pizza Place	Pizza Place
1	38	46	Restaurant	Sandwich Place
2	1	67	Café	Vegetarian / Vegan Restaurant
3	1	59	Bakery	Café
4	1	57	Pizza Place	Italian Restaurant

Table 3 suggests that cluster 0 is most popularized with pizza places while cluster 1 has more mixed type restaurants and sandwich places. Also, cluster 1 has more restaurants than cluster 0 on average per neighbourhood. When plotting the number of neighbourhoods in each of these two clusters against population and individual income, two histograms are obtained as given in Figure 2 and Figure 3. The relationship between neighbourhood clustering and population/income is rather straightforward. Namely, cluster 0 peaks at neighbourhoods with relatively low population and low income while cluster 1 has more neighbourhood of medium population and income.

Figure 2 and Figure 3 enhanced results in Table 3. First, cluster 1 represents neighbourhoods with more restaurants than cluster 0, which can be explained by the fact that neighbourhoods in cluster 1 have more population and higher income than those in cluster 0. Second, cluster 0 neighbourhoods are popularized with food venues like pizza places which are more affordable than the most popular food venues in cluster 1 neighbourhoods (mixed type restaurants). Therefore, the best restaurant business opportunities should occur in cluster 0 neighbourhoods with high population/income or cluster 1 neighbourhoods with low population/income. For cluster 0 neighbourhoods with high population/income, the investors should consider opening medium to high end restaurants. For cluster 1 neighbourhoods with low population/income, the investors should consider affordable food chains such as pizza places.

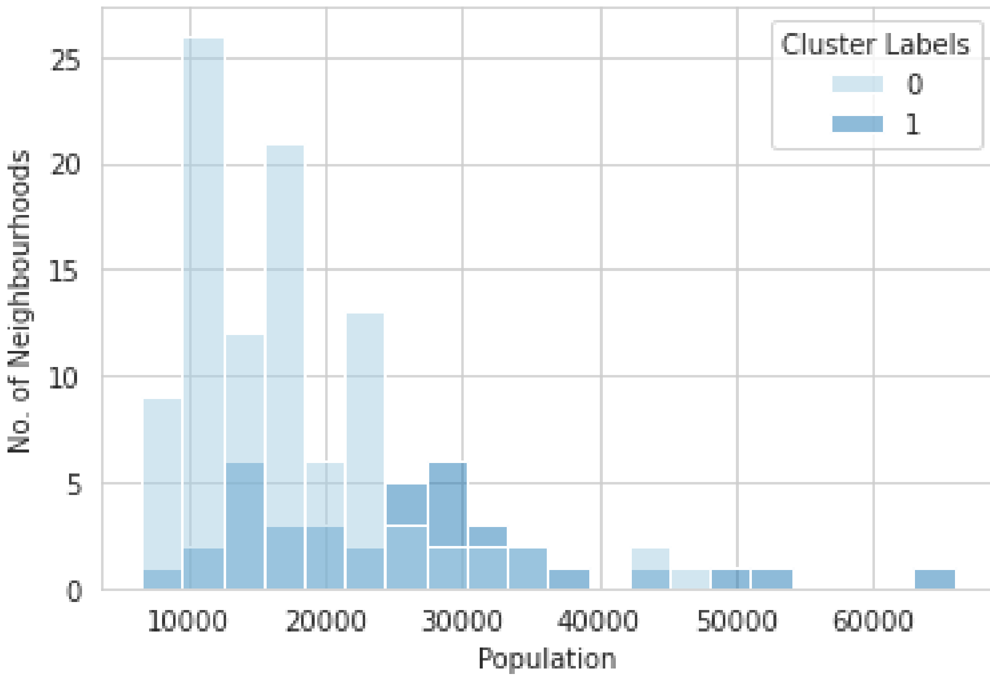

Figure 2 Number of neighbourhoods in cluster 0 and 1 against neighbourhood population

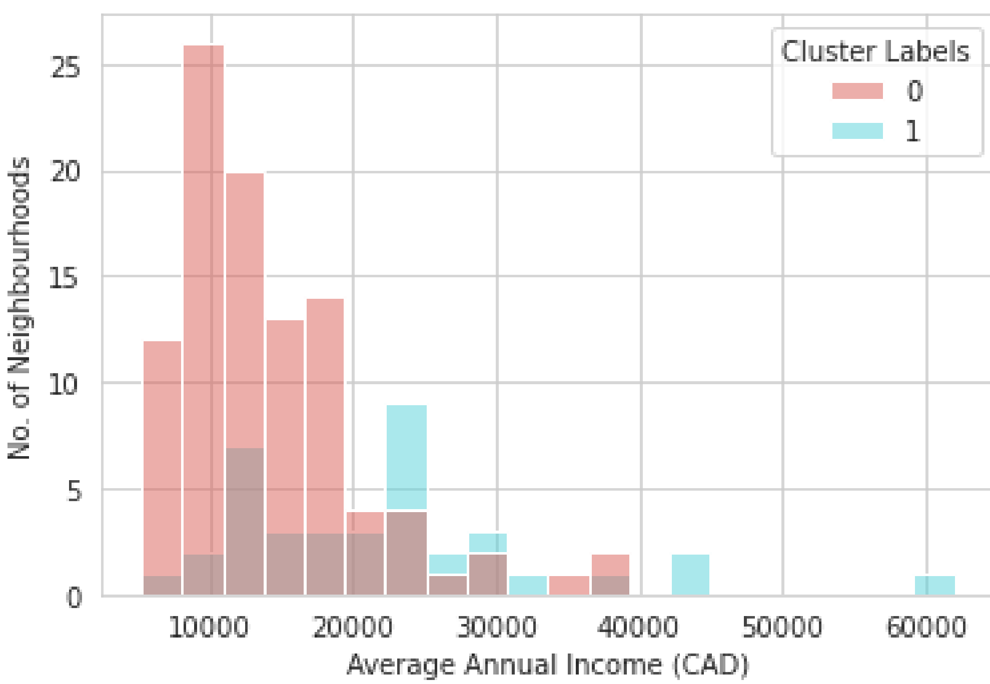

Figure 3 Number of neighbourhoods in cluster 0 and 1 against neighbourhood individual income

The following map (Figure 4) integrated population income and cluster of neighbourhoods for the 140 neighbourhoods in Toronto. The map visually displays our discussion of the relationship of population-income and food venues in different neighbourhood clusters. It is easily seen that most red circle dots (cluster 0) are labelled on light green neighbourhoods (low population-income) and most purple circle dots (cluster 1) are labelled on darker green neighbourhoods (medium to high population-income). Those neighbourhoods that do not follow such rules such as Rouge (131) are places that have better potential to open new restaurants.

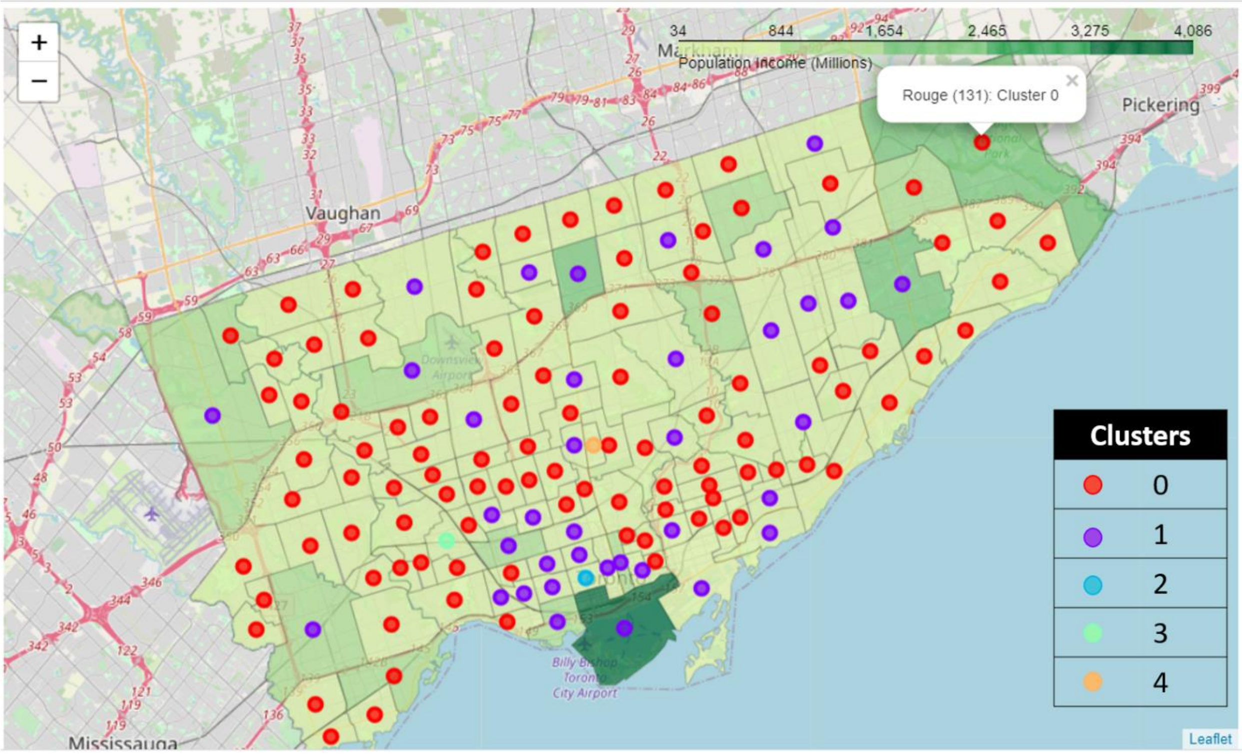

Figure 4 Map of all neighbourhoods colored by population income, and circle marked by clusters

5. Conclusion

This article has compared 140 neighbourhoods in Toronto and explore their chances of opening a generic restaurant. Two data sources (City of Toronto Open Data Portal and Foursquare Location API Services) are used in retrieving location, population, and income information and different food venues in each of the neighbourhoods. Based on the categories of food venues (i.e., restaurants), the neighbourhoods are grouped into five clusters with two main clusters that are studied further.

It is noticed that cluster 0, corresponding to neighbourhoods with relatively low population and income, has a smaller number of food venues per neighbourhood and the venues are mostly affordable restaurants like pizza places. In comparison, cluster 1 neighbourhoods have more food venues and popularized with mixed type restaurants. Cluster 1 corresponds to medium to high population and income neighbourhoods.

As such, the best restaurant business opportunities should occur in cluster 0 neighbourhoods with high population/income or cluster 1 neighbourhoods with low population/income. For cluster 0 neighbourhoods with high population/income, the investors should consider opening medium to high end restaurants. For cluster 1 neighbourhoods with low population/income, the investors should consider affordable food chains such as pizza places.