The Battle of Neighborhoods

YW
6/12/2019

1. Introduction

1.1. Background

Food-In Pro, Inc. is a well-known supplier in the manufacturing of powdered food ingredients (e.g. salt, sugar, flour and yeast) and food additives (e.g. sweetener, artificial color, creamer and softener) in Canada as well as worldwide. Due to the growth of business, the company decides to open a small distribution warehouse in the city of Toronto for serving the local customers effectively. Before the financial budget of investment is evaluated, the company would first like to know if there is an optimal location/neighborhood in Toronto for the warehouse based on the needs from target and/or potential clients.

According to the sales and marketing strategies, the most clients in a city are roughly summarized and rated into five categories depending on the business scale and stability as below:

Table 1. Categories of clients with business description

No.	Client's Category	Business description	Score rate
I	Supermarket and grocery	Large and stable business	5
II	Restaurant	Large business but varied stability	4
III	Food spot, court, place, house, pub, joint, diner...	Medium business and varied stability	3
IV	Ice cream, bakery, dessert, chocolate, donut, smoothie and cafe	Small business and varied stability	2
V	Other food stores	Small and unstable business	1

1.2. Problem Description

The company assigns this project to the data analyst to explore the distribution of the above categorical clients against the neighborhood in the city of Toronto. The analyst is expected to give a recommendation to the facility address searching group about the best area of neighborhood(s) to locate the warehouse in the city, according to the distribution of clients. The principles of priority for the recommendation of location are: 1) more amount of high rate clients (especially supermarkets and groceries), then 2) more amount of total clients.

2. Data Acquisition and Cleaning

2.1 Data Sources

The data of neighborhoods in the city of Toronto are acquired from Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) and scraped by the BeautifulSoup package. The geographical coordinates are acquired from the course material. The location data of those categorical venues are acquired by querying Foursquare API. The

keyboards sent for querying are “Supermarket”, “Restaurant”, and “Ice cream”, where the other categories can also be returned. The venue lists are acquired and combined into one dataframe.

2.2 Data Selection and Cleaning

Only “Toronto boroughs” (Downtown Toronto, East Toronto, West Toronto and Central Toronto) and their neighborhoods are selected for evaluation in this project. “Not assigned” neighborhoods under the assigned postcodes are also dropped. Venue categories are examined to exclude unreasonable ones (e.g. Pharmacy). Then, the venues are identified by venue categories as shown in **Table 2**. This is the initial dataframe for the analysis.

Table 2. Neighborhood venues in client categories (partial view)

Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Ryerson	43.657162	-79.378937	Metro	43.658404	-79.376748	Supermarket
Garden District	43.657162	-79.378937	Metro	43.658404	-79.376748	Supermarket
St. James Town	43.651494	-79.375418	Metro	43.649027	-79.373313	Supermarket
Berczy Park	43.644771	-79.373306	Loblaws	43.644462	-79.369486	Supermarket
Berczy Park	43.644771	-79.373306	Metro	43.649027	-79.373313	Supermarket
Central Bay Street	43.657952	-79.387383	The Market by Longo's Elizabeth	43.655357	-79.385115	Supermarket
Central Bay Street	43.657952	-79.387383	Metro	43.660569	-79.383768	Supermarket
Christie	43.669542	-79.422564	Fiesta Farms	43.668471	-79.420485	Supermarket
Christie	43.669542	-79.422564	Loblaws	43.671807	-79.421102	Supermarket
Adelaide	43.650571	-79.384568	Rabba Marché	43.649216	-79.386908	Supermarket

2.3 Feature Identification

There are 2802 venues provided at those neighborhoods. Venue category is the principle feature that contributes to score the rates of neighborhoods which gives us the candidate(s). Neighborhood coordinates are used to map the neighborhood variables, and may contribute to search the compromised location if the candidates are not adjacent. Although there is another type of coordinates for venues which gives a more accurate information, the tiny variance gives difficulty to mapping and observing since we are targeting the neighborhoods. They are therefore dropped in the following analysis.

3. Methodology

I. Rank the neighborhoods by the number of venues

II. Rank the neighborhoods by scoring the rates based on **Table 2**. Explore the difference between the two ranks.

III. Rank the neighborhoods by the number of supermarkets and groceries.

IV. Run k-mean to cluster the neighborhoods to identify the top common venues.

V. Examine the ranks and clusters, and recommend the candidate(s) according to the principles in Section 1.2.

Those procedures are achieved by Python 3 in Jupyter Notebook. The codes are saved in the link below:

https://github.com/Nuercom/Forlearn/blob/master/Final.ipynb

4. Results and Discussion

4.1. Neighborhoods ranking by the number of venues

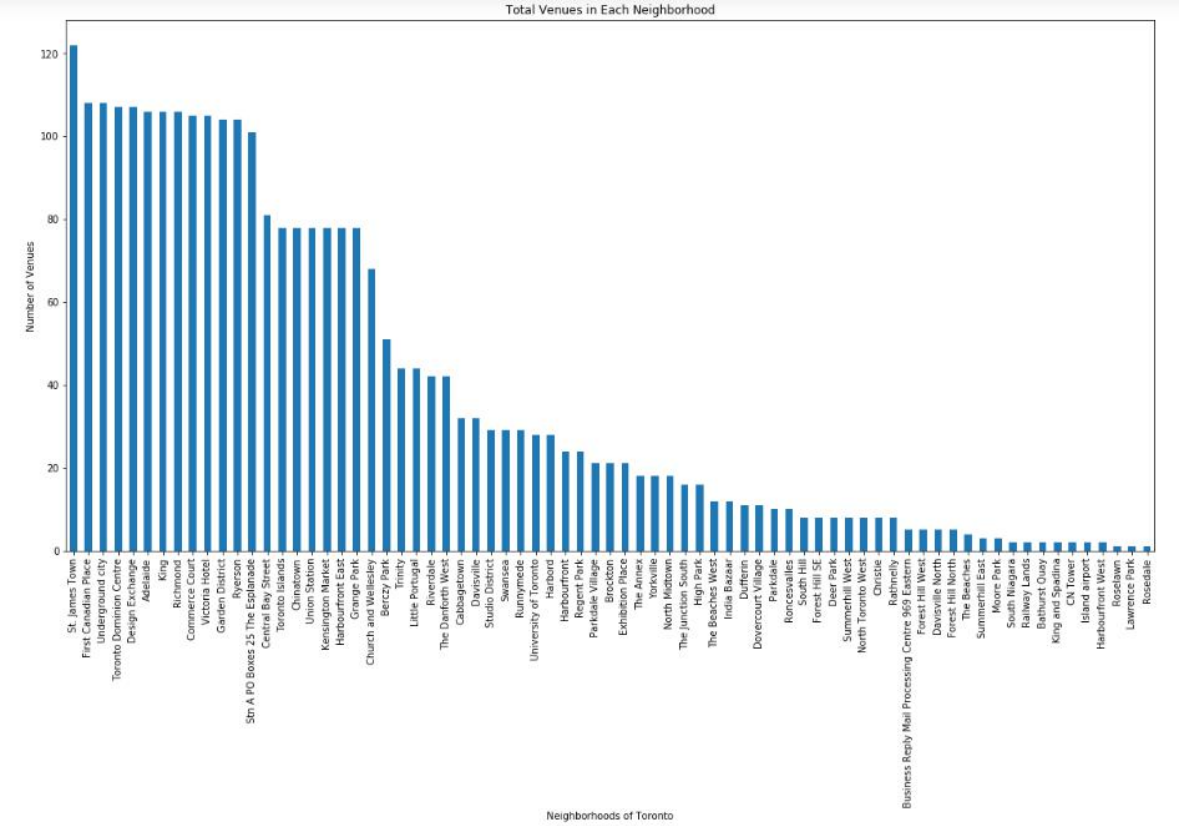

Figure 1. Bar chart of venue numbers in neighborhoods

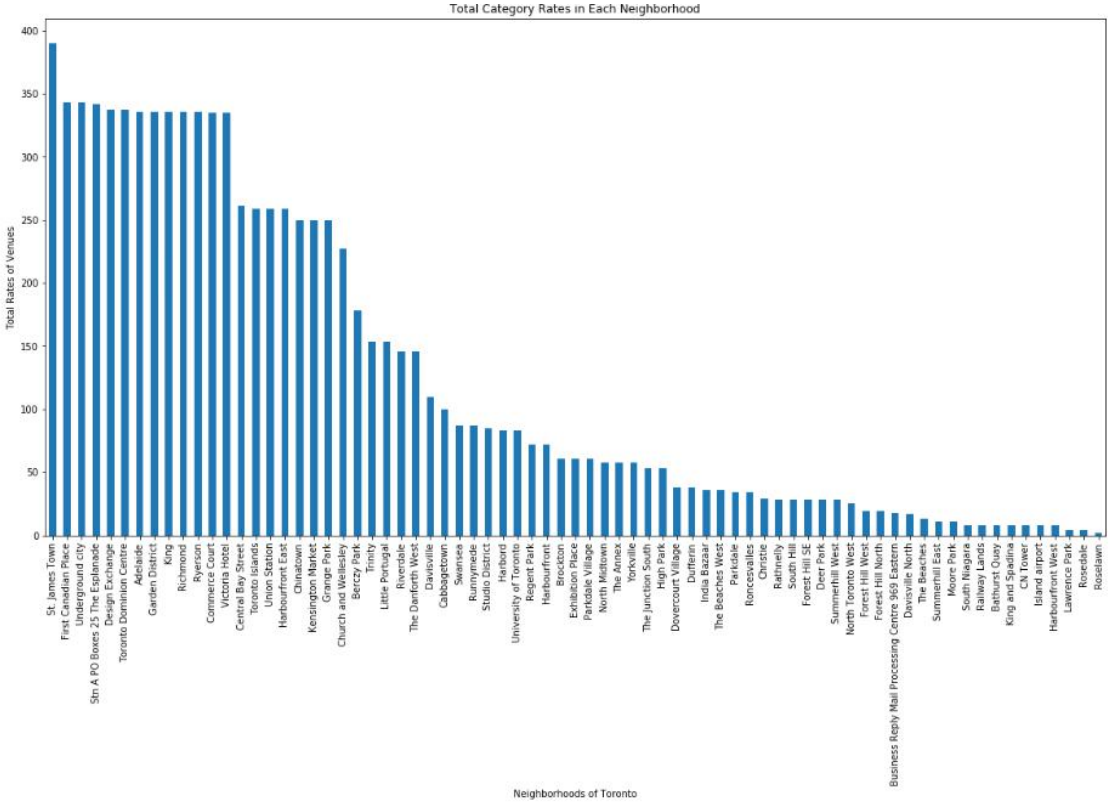

Figure 2. Bar chart of category rate in neighborhoods

The sum of venues in each neighborhood is visualized in a bar chart as shown in **Figure 1**. There are twelve neighborhoods which have over 100 venues of our target clients. We keep these twelve in comparison to the bar chart of venue rates as discussed in Section 4.2, which gives us a better observation toward the categorical clients.

4.2. Neighborhoods ranking by the rate of venues

The score rates of venue categories based on **Table 1** is added to **Table 2** as shown in below.

Table 3. Neighborhood venues with the score rates in client categories (partial view)

Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Category Rate
Ryerson	43.657162	-79.378937	Metro	43.658404	-79.376748	Supermarket	5
Garden District	43.657162	-79.378937	Metro	43.658404	-79.376748	Supermarket	5
St. James Town	43.651494	-79.375418	Metro	43.649027	-79.373313	Supermarket	5
Berczy Park	43.644771	-79.373306	Loblaws	43.644462	-79.369486	Supermarket	5
Berczy Park	43.644771	-79.373306	Metro	43.649027	-79.373313	Supermarket	5
Central Bay Street	43.657952	-79.387383	The Market by Longo's Elizabeth	43.655357	-79.385115	Supermarket	5
Central Bay Street	43.657952	-79.387383	Metro	43.660569	-79.383768	Supermarket	5
Christie	43.669542	-79.422564	Fiesta Farms	43.668471	-79.420485	Supermarket	5
Christie	43.669542	-79.422564	Loblaws	43.671807	-79.421102	Supermarket	5
Adelaide	43.650571	-79.384568	Rabba Marché	43.649216	-79.386908	Supermarket	5

The rank of rates in neighborhood is visualized in **Figure 2**. The neighborhoods of the top 12 rates are slightly varied. A comparison table is generated for review in below.

Table 4. Comparison Table of venue numbers versus category rates (top 20 shown) by neighborhoods

Neighbourhood	Latitude	Longitude	Venue_Number	Category_Rate
St. James Town	43.651494	-79.375418	122	390
First Canadian Place	43.648429	-79.382280	108	343
Underground city	43.648429	-79.382280	108	343
Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	101	342
Toronto Dominion Centre	43.647177	-79.381576	107	337
Design Exchange	43.647177	-79.381576	107	337
Richmond	43.650571	-79.384568	106	336
Garden District	43.657162	-79.378937	104	336
Ryerson	43.657162	-79.378937	104	336
King	43.650571	-79.384568	106	336
Adelaide	43.650571	-79.384568	106	336
Victoria Hotel	43.648198	-79.379817	105	335
Commerce Court	43.648198	-79.379817	105	335
Central Bay Street	43.657952	-79.387383	81	261
Harbourfront East	43.640816	-79.381752	78	259
Toronto Islands	43.640816	-79.381752	78	259
Union Station	43.640816	-79.381752	78	259
Chinatown	43.653206	-79.400049	78	250
Kensington Market	43.653206	-79.400049	78	250
Grange Park	43.653206	-79.400049	78	250

Figure 3. Map of grouped neighborhoods by category rates

A feature from the comparison is indicated that, even though the neighborhood rank is varied in these two manipulations, the groups of neighborhoods is almost kept the same. For example, the neighborhoods which have over 100 venues do have over 300 of category rate. So the scoring of clients gives a matched reflection of venue number in neighborhood, as well as gives a more accurate pathway to identify the candidate neighborhoods toward our target clients.

The neighborhoods are mapped (**Figure 3**) by dividing them into four intervals >300 (purple), 200~300 (blue), 100~200 (grey) and <100 (red). Two features are observed. First, each level is routinely surrounding one by one without obvious intersection. Second, they are concentrated to the center of Toronto from low level to high. This step-like distribution supports that we could be serving more target clients as we are getting close to the center of Toronto. The 20 neighborhoods from **Table 4** are selected as the preliminary candidates. The exploratory analysis of each venue category will be discussed in the next two sections, which will contribute to narrow the candidates.

4.3. Neighborhoods categories

Table 5 lists how many venues of each category in all selected neighborhoods. This distribution of amounts indicates that only 2.03% of them are supermarkets/groceries. This is a very different feature from our previous records in rural cities. Since the supermarkets/groceries are in the company's top priority, the distribution of them in neighborhoods is explored.

The neighborhoods with two of more supermarkets/groceries are listed in **Table 6**. Seven of the ten (marked in yellow) can be found in the preliminary list as discussed in Section 4.2. It is shown that the rank of supermarket is differed from the rate rank of total venues. For example, St. James Town has an outstanding rate of venues, while it is not the outstanding neighborhood of supermarket. Kensington Market, Chinatwon and Grange Park are the top neighborhoods with supermarkets, but their venue category rates are still in medium group.

So from this section the data analyst realizes the difference of category occupations between Toronto and previous rural cities. To target the clients based on the priority principle, the popularity of each category in those neighborhoods is examined. It will be useful if some neighborhoods have top common venue as supermarket/grocery.

Table 5. Amounts of venues in each category

Client's Category	Score rate	Venues
Supermarket and grocery	5	57
Restaurant	4	1333
Food spot, court, place, house, pub, joint, diner...	3	723
Ice cream, bakery, dessert, chocolate, donut, smoothie and cafe	2	584
Other food stores	1	105

Table 6. List of neighborhoods which have 2 or more supermarkets/groceries

Neighborhood	**Supermarket/grocery**
Kensington Market	5
Chinatown	5
Grange Park	5
Dufferin	2
Christie	2
Dovercourt Village	2
Stn A PO Boxes 25 The Esplanade	2
St. James Town	2
Berczy Park	2
Central Bay Street	2

4.4. Common venues in neighborhoods

This section examines the common venues of the five categories in the neighborhoods. Those neighborhoods are divided into 5 clusters to run k-means clustering (**Figure 4**). The clusters are displayed in the Jupyter Notebook. Matched to **Table 5**, II, III and IV category occupies the top 2 common venue in those neighborhoods. "Supermarket" neighborhood could not be identified. However, even though the supermarket/grocery is the smallest group, in some neighborhoods it is not the last common venue (**Table 7**). Neighborhoods which have supermarket/grocery in 3rd and 4th common venue are extracted and shown in **Table 8** with the category rates.

Table 7. Example of common venues in neighborhoods

Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Harbourfront	2	II	IV	III	V	I
Regent Park	2	II	IV	III	V	I
Christie	2	II	IV	I	III	V
Dovercourt Village	2	II	IV	III	I	V
Dufferin	2	II	IV	III	I	V
Brockton	2	IV	II	III	V	I
Exhibition Place	2	IV	II	III	V	I
Parkdale Village	2	IV	II	III	V	I

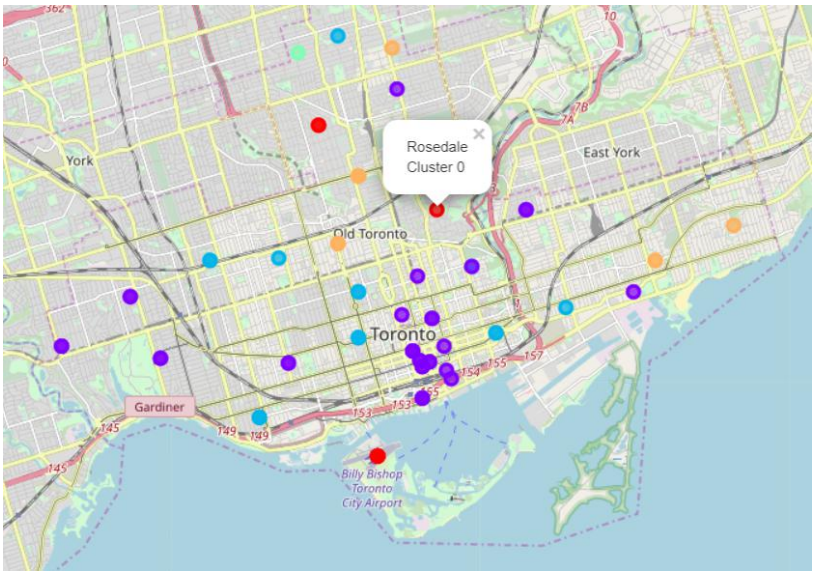

Figure 4. Map of clusters by clustering common venues

Table 8. Selected neighborhoods of non-bottom venues of supermarket/grocery

Neighbourhood	Latitude	Longitude	Category_Rate
Berczy Park	43.644771	-79.373306	178
The Danforth West	43.679557	-79.352188	146
Riverdale	43.679557	-79.352188	146
Parkdale	43.648960	-79.456325	34
Roncesvalles	43.648960	-79.456325	34
Christie	43.669542	-79.422564	29
Dovercourt Village	43.669005	-79.442259	38
Dufferin	43.669005	-79.442259	38
Chinatown	43.653206	-79.400049	250
Grange Park	43.653206	-79.400049	250
Kensington Market	43.653206	-79.400049	250

5. Recommendation

Comprehensively review the results in **Table 4**, **Table 6** and **Table 8** against the principles in Section 1.2. Chinatown, Grange Park and Kensington Market win the principle 1), and have the medium level against principle 2). They are the recommended neighborhoods therefore. Since those are adjacent neighborhoods, there is no need to figure out a center point to cover them. So that area is recommended to open the distribution warehouse to fulfill the company's need.

6. Conclusion

The analysis performed in this project can be modelized to predict another city or area. More cities can be selected for the training and testing to improve the model. There are also some tips to improve the methodology. For example, according to the distribution feature of venues in big cities, can the analysis be improved with a more detailed rating system for small business units of food stores, or if the distribution of competitors can be added to the current rating system (e.g. as negative score?).