Predicting promising locations for opening a bakery in Toronto

Amy Jursz

December 18,2020

1. Introduction/Business Problem

1.1 Background

According to a 2019 study from Ryerson University, Toronto was the second fastest growing Metropolitan Area and the top growing city in all of the United States and Canada. The population of Toronto increased by 77,435 persons, far out pacing the next city – Phoenix with 25,288.

1.2 Business Problem

With this large influx of people, more business will be required to support them including gyms, retail stores, restaurants, etc. One of the main questions, and one of the main reasons for success, regarding opening a business is - as they say “location, location, location”.

In this exercise, we will analyze current venue locations in Toronto to determine possible new locations for opening a Bakery.

2. Data

2.1 Data Sources

1. Wikipedia.org:

   This will be used to obtain postal codes of Toronto, Canada which will be used in conjunction with Concl.us data to map the greater Toronto area.

   Example:

Postal Code	Borough	Neighborhood
M1S	Scarborough	Agincourt
M8W	Etobicoke	Alderwood, Long Branch

2. Cocl.us:

Contains longitude and latitude of Toronto, Canada postal codes which will be used in conjunction with Wikipedia data to map the greater Toronto area.

Example:

Postal Code	Latitude	Longitude
M1B	43.80669	-79.1944
M1C	43.78454	-79.1605

3. Foursquare.com:

This website contains business data for a multitude of businesses worldwide. This data includes location, business category and can include ratings and reviews of businesses. We will use this website to identify current bakery locations.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.655330	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653240	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

4. Ryerson.ca:

Contains growth data regarding Toronto, Canada referenced in the introduction. Full link provided:

https://www.ryerson.ca/cur/Blog/blogentry35/

2.2 Data cleaning

Data downloaded from Wikipedia contained many postal codes that were not assigned a borough or neighborhood. These data points were scrubbed from the dataset. Then the updated dataset from Wikipedia was combined with the dataset from Cocl.us to form a new dataset. This new dataset encompassed all location data, postal codes, boroughs, and neighborhoods of the greater Toronto area.

3. Methodology

3.1 Data analysis

After downloading and cleaning the data from Wikipedia and Conc.us, we need to locate the latitude and longitude for Toronto. This was done with geopy geolocator.

Next was using the appropriate latitude and longitude to get venue data using the Foursquare API. All venues within 500 metes from the designated latitude and longitude were pulled. The information on the venues pulled included:

- Neighborhood
- Neighborhood Latitude
- Neighborhood Longitude
- Venue
- Venue Latitude
- Venue Longitude
- Venue Category

A new data frame was created with the above information – containing 851 rows and 7 columns. Below is a sample of the list:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

The next step was to convert the data into categories and reindex it using one hot encoding.

	Neighborhood	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	...	Theater	Theme Restaurant	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Wine Bar	Yoga Studio
0	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 196 columns

Then the data was grouped by neighborhood and the average occurance (mean) was calculated for each venue category for that neighborhood.

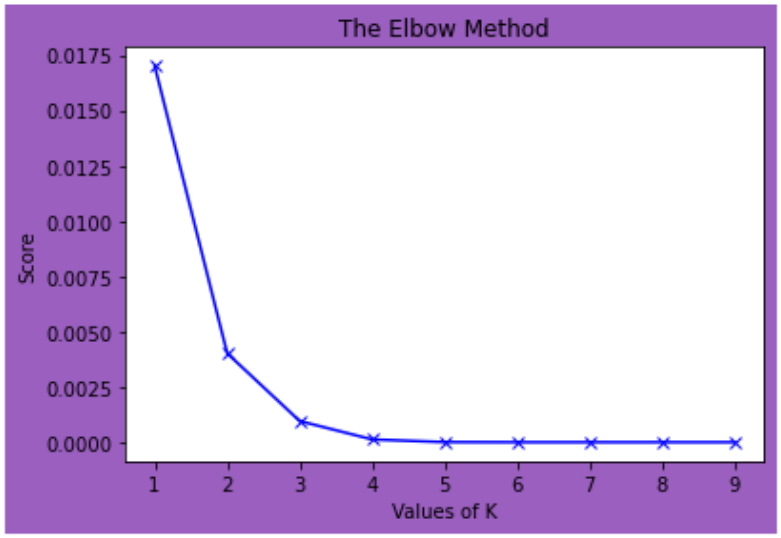
	Neighborhood	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	...	Theater	Theme Restaurant	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Wine Bar	Yoga Studio
0	Berczy Park	0.0000	0.0000	0.0000	0.000	0.000	0.0000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.033333	0.0	0.0	0.0	0.000000
1	Brockton, Parkdale Village, Exhibition Place	0.0000	0.0000	0.0000	0.000	0.000	0.0000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
2	Business reply mail Processing Centre, South C...	0.0000	0.0000	0.0000	0.000	0.000	0.0000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.0625	0.0625	0.0625	0.125	0.125	0.0625	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
4	Central Bay Street	0.0000	0.0000	0.0000	0.000	0.000	0.0000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.033333	0.0	0.0	0.0	0.033333

5 rows × 196 columns

Since we are only interested in Bakeries, the dataset was filtered to only show that venue. A sample of the data is shown below:

	Neighborhood	Bakery
0	Berczy Park	0.033333
1	Brockton, Parkdale Village, Exhibition Place	0.043478
2	Business reply mail Processing Centre, South C...	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
4	Central Bay Street	0.000000
5	Christie	0.000000
6	Church and Wellesley	0.000000
7	Commerce Court, Victoria Hotel	0.033333
8	Davisville	0.000000
9	Davisville North	0.000000
10	Dufferin, Dovercourt Village	0.153846
11	First Canadian Place, Underground city	0.033333
12	Forest Hill North & West, Forest Hill Road Park	0.000000
13	Garden District, Ryerson	0.033333
14	Harbourfront East, Union Station, Toronto Islands	0.000000
15	High Park, The Junction South	0.040000
16	India Bazaar, The Beaches West	0.000000
17	Kensington Market, Chinatown, Grange Park	0.033333
18	Lawrence Park	0.000000
19	Little Portugal, Trinity	0.033333
20	Moore Park, Summerhill East	0.000000
21	North Toronto West, Lawrence Park	0.000000

K-Means clustering was performed on the data in order to group the neighborhoods into similar clusters so as to easily identify where there are the least existences of bakeries. Before this can be done though, we need to identify the optimal number of clusters for the data using the elbow method.



Using the elbow method, the optimal number of clusters is 3.

K-Means clustering was then performed on the data and clusters were assigned to each venue.

	Neighborhood	Bakery	Cluster Labels	Postal Code	Borough	Latitude	Longitude
0	Berczy Park	0.033333	2	M5E	Downtown Toronto	43.644771	-79.373306
1	Brockton, Parkdale Village, Exhibition Place	0.043478	2	M6K	West Toronto	43.636847	-79.428191
2	Business reply mail Processing Centre, South C...	0.000000	0	M7Y	East Toronto	43.662744	-79.321558
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0	M5V	Downtown Toronto	43.628947	-79.394420
4	Central Bay Street	0.000000	0	M5G	Downtown Toronto	43.657952	-79.387383

This data was then mapped using Folium where:

- Cluster 0 = Red
- Cluster 1 = Purple
- Cluster 2 = Green

4. Results

Based on K-Means clustering, the cluster with the least number of occurrences of a bakery is Cluster 0:

Neighborhood	Postal Code	Borough
The Beaches	M4E	East Toronto
India Bazaar, The Beaches West	M4L	East Toronto
Lawrence Park	M4N	Central Toronto
Davisville North	M4P	Central Toronto
North Toronto West, Lawrence Park	M4R	Central Toronto
Davisville	M4S	Central Toronto
Moore Park, Summerhill East	M4T	Central Toronto
Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park	M4V	Central Toronto
Rosedale	M4W	Downtown Toronto
Church and Wellesley	M4Y	Downtown Toronto

St. James Town	M5C	Downtown Toronto
Central Bay Street	M5G	Downtown Toronto
Harbourfront East, Union Station, Toronto Islands	M5J	Downtown Toronto
Roselawn	M5N	Central Toronto
Forest Hill North & West, Forest Hill Road Park	M5P	Central Toronto
The Annex, North Midtown, Yorkville	M5R	Central Toronto
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	M5V	Downtown Toronto
Christie	M6G	Downtown Toronto
Parkdale, Roncesvalles	M6R	West Toronto
Runnymede, Swansea	M6S	West Toronto
Queen's Park, Ontario Provincial Government	M7A	Downtown Toronto
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	M7Y	East Toronto

The cluster with the next least number of occurrences of a bakery is Cluster 2:

Neighborhood	**Postal Code**	**Borough**
The Danforth West, Riverdale	M4K	East Toronto
Studio District	M4M	East Toronto
St. James Town, Cabbagetown	M4X	Downtown Toronto
Garden District, Ryerson	M5B	Downtown Toronto
Berczy Park	M5E	Downtown Toronto
Richmond, Adelaide, King	M5H	Downtown Toronto
Toronto Dominion Centre, Design Exchange	M5K	Downtown Toronto
Commerce Court, Victoria Hotel	M5L	Downtown Toronto
University of Toronto, Harbord	M5S	Downtown Toronto
Kensington Market, Chinatown, Grange Park	M5T	Downtown Toronto
Stn A PO Boxes	M5W	Downtown Toronto
First Canadian Place, Underground city	M5X	Downtown Toronto

Little Portugal, Trinity	M6J	West Toronto
Brockton, Parkdale Village, Exhibition Place	M6K	West Toronto
High Park, The Junction South	M6P	West Toronto

The cluster with the greatest number of occurrences of a bakery is Cluster 1:

Neighborhood	**Postal Code**	**Borough**
Regent Park, Harbourfront	M5A	Downtown Toronto
Dufferin, Dovercourt Village	M6H	West Toronto

5. Observations and Recommendations

Observing the cluster map and the list of Boroughs above, there appears to be a focus of bakeries in the Downtown Toronto Borough, while the Central Toronto Borough there seems to be a deficit. This could be due to the fact Central Toronto appears to have more suburban boroughs (per Wikipedia).

Based on the data and observations, the recommendation is to look at the locations in Cluster 0 with emphasis on the Downton Toronto Borough – such as Christie, Central Bay Street, Rosedale or Queen's Park. This is due to location in Downtown and their lack of any bakery in that particular neighborhood. It is not recommended to look at locations in Cluster 1 as they have the greatest occurrence of bakeries and result in greater competition. Cluster 2 can be considered; however, the new bakery will have to contend with previously established businesses and that may already have a solid customer base.

6. Conclusion

In this study, I analyzed the boroughs and neighborhoods of Toronto, Canada to determine the optimal location to open a bakery. K-Means clustering, in tandem with the elbow method, was used to cluster the neighborhoods into similar groups so as to easily identify favorable locations to consider or avoid. These same models can be used to evaluate other business besides bakeries, such as gyms, retail, restaurants, etc.