Identifying a suitable location to open a restaurant in Toronto

Cesar A

May 08, 2020

1. Introduction

1.1. Background

Finding an optimal location to open a business is important for its success. Businesses in Toronto like in many other cities with dense population are negatively affected by crime. Also, there are neighborhoods where the competition among businesses are higher because they close to each other.

1.2. Business Problem

In this project, I identify selected neighborhoods in Toronto where a person can open a restaurant. The selection of the neighborhoods depends on the requirements of the person/client. The client's requirements are:

- Location must be a safe place where breaking and entering crimes are low.
- There should be a few or no restaurants in a 500-meter radius from location.
- There should be commercial buildings/offices nearby where employees go to work.
- Location must be near parks. Client likes parks and think people will gather there and eventually eat at his restaurant.

2. Data

To identify the neighborhoods with the lowest crime rate, I use the crime reports found on Toronto Police Website http://data.torontopolice.on.ca/pages/open-data

Crimes are grouped in different categories such as Homicide, Assault, Break & Enter, etc. To simplify the project, I use the crime report for breaking & entering crimes. Also, I focus on the crimes that happened between 2014 and 2019 because it is most current data available.

Luckily, the crime report table shows the latitudes and longitudes of the neighborhoods where breaking & entering crimes happened. Using those coordinates, I use https://developer.foursquare.com/ to retrieve all different businesses(venues) located 500 meters from each neighborhood's coordinates.

3. Methodology

3.1. Exploring Data

Loading the break & enter crime report into Jupyter notebook, I can see a table with many columns. One of the columns is "premisetype." This column classifies the Break & Enter crime into further subgroups such House, Apartment, Commercial. In this case, I select the Commercial subgroup because is relevant to businesses (i.e., restaurant.) Additionally, I select the following columns: Lat, Long, and Neighbourhood. Table 1 shows the clean data needed for further analysis.

Table 1. Table contains breaking and entering crimes in Toronto neighborhoods between 2014 and 2019

	premisetype	Lat	Long	Neighborhood
0	Commercial	43.670227	-79.386787	Rosedale-Moore Park (98)
1	Commercial	43.706944	-79.375648	Leaside-Bennington (56)
2	Commercial	43.773617	-79.261131	Bendale (127)
3	Commercial	43.630154	-79.485252	Stonegate-Queensway (16)
4	Commercial	43.790829	-79.445381	Westminster-Branson (35)
...	...	...	...	...
13677	Commercial	43.779770	-79.415573	Newtonbrook West (36)
13678	Commercial	43.660816	-79.385857	Bay Street Corridor (76)
13679	Commercial	43.760429	-79.570091	Humber Summit (21)
13680	Commercial	43.762623	-79.564423	Humber Summit (21)
13681	Commercial	43.657909	-79.381584	Bay Street Corridor (76)

Table 1 also shows that there are 13682 breaking & entering crimes to commercial businesses in all neighborhoods in Toronto between 2014 and 2019. To find the number of crimes for each neighborhood, I group the crimes by neighborhoods. Grouping the crimes shows that there are 140 neighborhoods in Toronto. Figure 1 shows the frequency of crimes for 10 out of 140 neighborhoods.

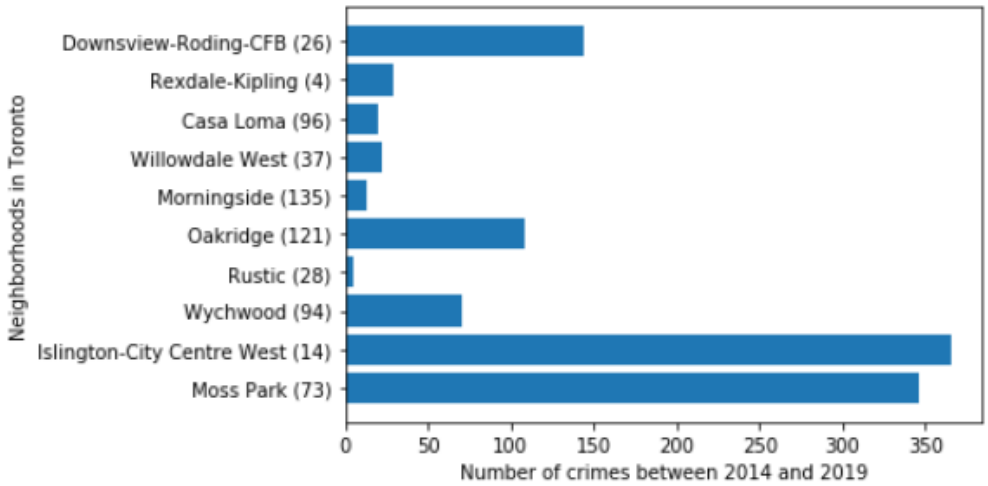

Figure 1. Number of crimes of some neighborhoods in Toronto from 2014 to 2019

There are neighborhoods like Moss Park that accumulated over 300 crimes in the 5-year period. There are other neighborhoods such as Casa Loma with less 50 crimes in the same period. I arbitrarily select neighborhoods with less than 30 crimes in that period for further analysis.

There are 36 neighborhoods with less than 30 crimes. I call these neighborhoods 'safe neighborhoods.' I show the location of these neighborhoods in Figure 2.

Figure 2. Map with locations of the 36 safe neighborhoods

Using the coordinates of the 36 neighborhoods and FOURSQUARE, I retrieve the commercial venues located 500 meters from each neighborhood coordinate. Table 2 shows that there are 261 venues in the 36 neighborhoods and 102 unique categories of venues.

Table 2. Table contains venues located within 500-meter radius of each neighborhood coordinates.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bayview Woods-Steeles	43.796005	-79.383685	Bestview Hiking Trails	43.797368	-79.382227	Trail
1	Bayview Woods-Steeles	43.796005	-79.383685	Baseball Fields/Park	43.795644	-79.385214	Dog Run
2	Bayview Woods-Steeles	43.796005	-79.383685	Don Valley Parklands	43.792490	-79.380347	Park
3	Black Creek	43.766580	-79.520558	Petro-Canada	43.766586	-79.519870	Gas Station
4	Black Creek	43.766580	-79.520558	Rexall	43.766590	-79.519852	Pharmacy
...	...	...	...	...	...	...	...
256	Woodbine-Lumsden	43.693936	-79.312704	The Beer Store	43.693731	-79.316759	Beer Store
257	Woodbine-Lumsden	43.693936	-79.312704	Luxy Nails	43.692856	-79.315849	Spa
258	Woodbine-Lumsden	43.693936	-79.312704	Stan Wadlow Park	43.697836	-79.314303	Park
259	Woodbine-Lumsden	43.693936	-79.312704	New Star Video	43.692565	-79.315937	Video Store
260	Woodbine-Lumsden	43.693936	-79.312704	Dance Kids Canada	43.696563	-79.317385	Dance Studio

3.2. Analyzing Data

Next, I use k-means clustering technique to group the neighborhoods that share similar venue categories. First, I count the number of unique venue categories for each neighborhood. Then, I normalize these numbers before I can apply K-means to group similar neighborhoods. Technically, one could identity the number of clusters to group the neighborhoods by using the elbow method. This method identifies the cluster number where the distortion (sum of the square distances from each point of a cluster to its assigned center) flattens out.

In my case, figure 3 does not show a flatten curve. Nonetheless, the figure shows that the distortion decreases rapidly from cluster number 1 to cluster number 3. For that reason, I choose 3 for the number of clusters to group the neighborhoods

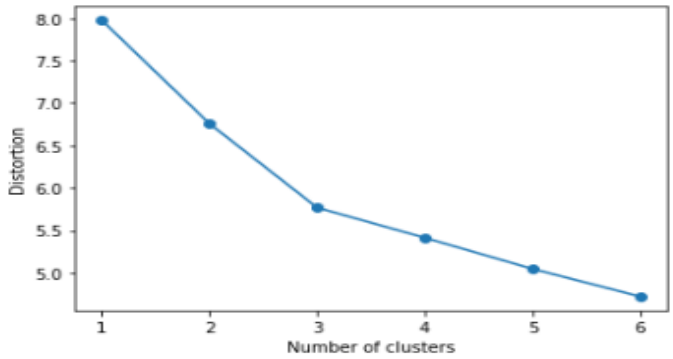

Figure 3. Elbow method to determine k-means number of clusters

After applying K-means clustering technique, I generate figure 4. This figure shows a map containing the neighborhoods grouped into 3 clusters. The clusters are numbered and colored as follows:

Cluster number: 0, color: red
Cluster number: 1, color: purple
Cluster number: 2, color: teal

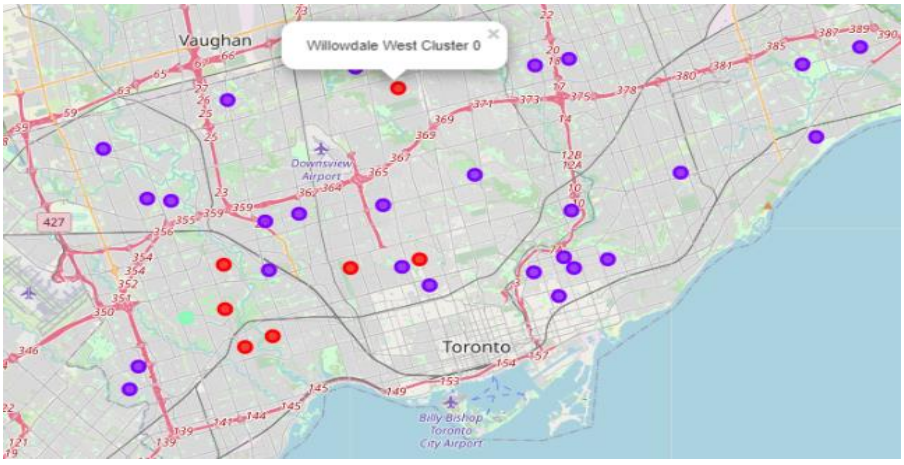

Figure 4. Map contains the location of 3 clusters

Table 3 contains information about cluster 0 red. It shows the neighborhoods where the most common venue category is Park. Additionally, in the following neighborhoods there are not many nearby restaurants:

-Bayview Woods-Steeles
-Caledonia-Fairbank
-Willowdale West

Table 3. Table contains information about cluster 0

	Neighborhood	Lat	Long	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Bayview Woods-Steeles	43.796005	-79.383685	0	Park	Dog Run	Women's Store	Construction & Landscaping
5	Caledonia-Fairbank	43.689843	-79.453226	0	Park	Women's Store	Bakery	Dog Run
9	Edenbridge-Humber Valley	43.670956	-79.521711	0	Dog Run	Park	Fast Food Restaurant	Women's Store
14	Forest Hill South	43.693905	-79.415058	0	Playground	Park	Furniture / Home Store	Discount Store
17	Humber Heights-Westmount	43.691419	-79.522331	0	Pizza Place	Park	Grocery Store	Coffee Shop
20	Kingsway South	43.653247	-79.510498	0	Lounge	Park	Pool	Garden
21	Lambton Baby Point	43.658649	-79.495432	0	Garden	Park	Mini Golf	Women's Store
34	Willowdale West	43.771972	-79.426922	0	Park	Mobile Phone Shop	Women's Store	Grocery Store

These three neighborhoods meet the following criteria:

- They are safe (less than 30 crimes in 5-year period.)
- There are parks around them.
- There are a few or no restaurants around them.

The last criterion that these neighborhoods must satisfy is that there must be commercial buildings/offices nearby where employees go to work.

I obtain the number offices around those three neighborhoods using their coordinates and FOURSQUARE. Figure 5 shows the number of offices around neighborhoods.

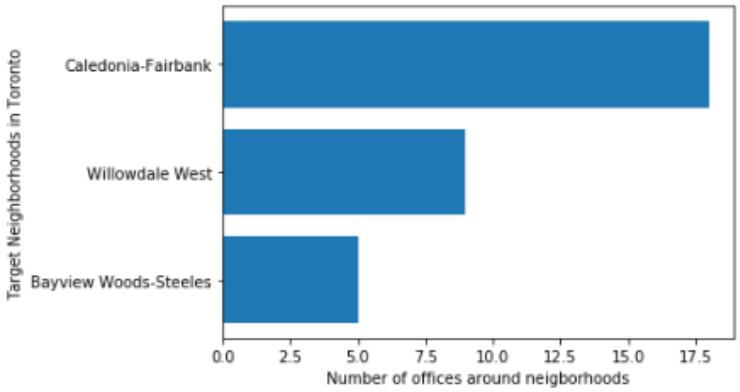

Figure 5. Number of commercial businesses near target neighborhoods

Additionally, the figure shows that neighborhood Caledonia-Fairbank meets all the client's requirements because not only this neighborhood is safe, has parks and a few or no restaurants but it also has the greatest number of offices around it.

Finally, figure 6 shows a map locating this neighborhood in Toronto

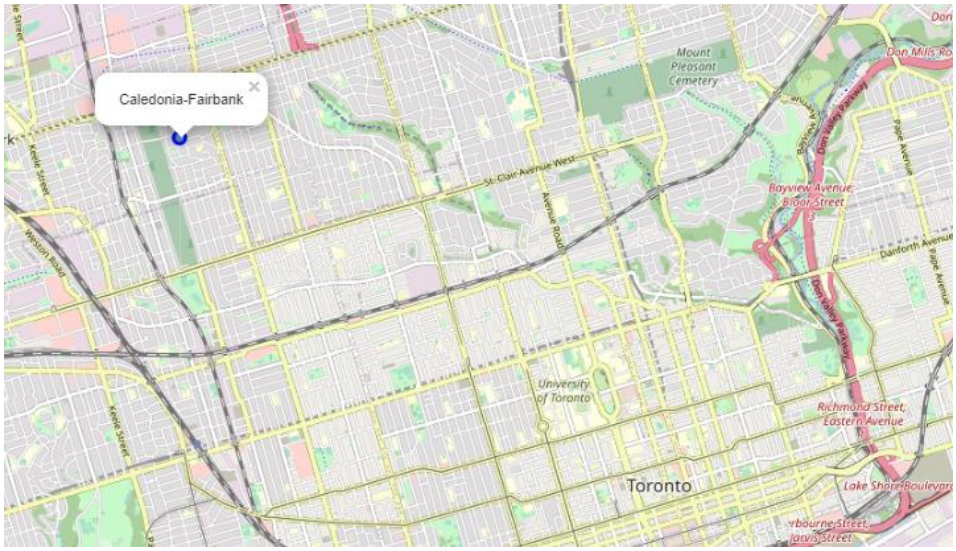

Figure 6. Location of neighborhood that meets all the client's requirements

4. Results

The analysis performed in the methodology section allowed me to discover the following:

- From the 140 neighborhoods in Toronto only 36 are considered safe.
- Using the elbow method, I identify three clusters to group the 36 neighborhoods.
- Cluster 0 contains the neighborhoods where parks is the most common venue category.
- From the three neighborhoods that contain the most parks around them, neighborhood Caledonia-Fairbank also contains the greatest number of commercial buildings where employees go to work.
-

5. Discussions and Recommendations

There are some arbitrary choices I make in this project that impact the selection of neighborhoods that meet the client's requirements. First, I only use data for one type of crime (break & enter) to analyze the neighborhoods. Adding other types of crimes such as theft can impact the final result. Second, I only select the neighborhoods with less than 30 crimes in a 5-year period. Increasing the cutoff from 30 to 50 crimes can change the results.

The datasets used in the analysis are constantly changing (i.e., new crimes, new venues.) Different datasets can also be added such data on rent prices if cost was a client's requirement. These changes can alter the results.

I use k-means method to group the neighborhoods based on their venue types. It is recommended that you insert a seed in your code to reproduce the same results. Every time you run k-means, results change because the positions of the center of the clusters also change.

One of the clusters contain the desired most common venue type, parks. If for some reason, none of the clusters contains the desired most common venue type, one can manually select the neighborhoods with the desired venue type from the table containing all the venue types.

6. Conclusion

I used crime data in Toronto and foursquare venue data to identify a neighborhood that meets all the client's requirements to open a restaurant. The analysis made in this project can be easily modified to satisfy requirements of another client to open a different business.