Battle of Neighborhood - report

Kenneth Nguyen

March 2020

1 Introduction

1.1 Problem Background

Toronto is one of the most important city in Canada. It can be considered as the financial capital of Canada. Toronto contains also the head office of one of the top five biggest banks of Canada. It contains also the stock exchange of Toronto. Then it is not surprising that Toronto contains banks and financial industries.

Since Toronto contains a lot of industries, it contains also cultural venues such as park, zoo or museum. There are public services such as restaurants, shops, supermarket and gym which is the venue that we focus on for this project.

1.2 Business Plan

As mentioned in the previous section, we focus on opening a new gym or fitness center in Toronto. A gym or fitness center allows people to practice and have physical activities. In this project, we suppose that the company Gold's gym, one of the biggest fitness center chain, would like to open one of its gym in Toronto. If the company wants to make profit with its fitness center, there are three criterion to take into account:

1. The location: we need to find an appropriate place to open the gym to be sure that people go to this gym.

2. The target audience: which type of person can be a potential client for this fitness center.

3. Is there already a gym in the neighborhood: people have already subscription to this gym and might not change.

2 Data acquisition

In order to solve our problem, we need to have data about Toronto city. We use two datasets on Toronto.

Dataset 1: The first dataset contains geographical data on Toronto which consists of the following:

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights
4	M6A	North York	Lawrence Manor

Figure 1: Dataset 1

This dataset can be provided for free with the following website: `https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M`.

Dataset 2: The second dataset is provided by the data science course from coursera and contains the following:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Figure 2: Dataset 2

This dataset allows to segment and explore the neighborhood of Toronto if we merge this Dataset 2 with Dataset 1.

Dataset 3: The third dataset contains all the venues in Toronto and their geographical coordinate. The dataset can be find on the Foursquare API `https://developer.foursquare.com/docs/data`.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Parkwoods	43.753259	-79.329656	Donalda Golf & Country Club	43.752816	-79.342741	Golf Course
2	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
3	Parkwoods	43.753259	-79.329656	Graydon Hall Manor	43.763923	-79.342961	Event Space
4	Parkwoods	43.753259	-79.329656	Galleria Supermarket	43.753520	-79.349518	Supermarket

Figure 3: Dataset 3

The information on the gym or fitness center are contained in the third dataset. From these three datasets, we need to build a new datasets and create new data to solve our business problem.

3 Methodology

In this section, we solve our business problem by using our data given in the previous section. Our goal is to segment the neighborhood in Toronto and highlight each venue in order to analysis the most frequented venue in each neighborhood. First, we create a new dataset build from the initial dataset, then we use a machine learning method to determine cluster on the map of Toronto and finally, we display the map of Toronto with the clusters build with the machine learning method.

3.1 Dataset cleaning and processing

This is the first step of our analysis. We need to find the number of venues per neighborhood if we want to segment each neighborhood. In Dataset 1, for some Postcode, the borough and neighborhood are not assigned, then we remove these rows with missing values. The Dataset 1 contains duplicate since one borough contains more than one neighborhood, then we remove these duplicates by combining rows with common postcode as one postcode with one borough and a list of neighborhoods assigned to the post code and the borough. We observe that Dataset 1 and Dataset 2 have a common feature which is 'Postcode', then the first step is to merge the two datasets on the feature postcode. The new dataset is Dataset 4 with the following features:

	Postal Code	Borough	Neighbourhood	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	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights,Lawrence Manor,Lawrence Heigh...	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494

Figure 4: Dataset 4

From the Foursquare API, we get the data of venues in Toronto. We convert each venue in a neighborhood as a categorical values. We only consider the ten most common venues in each neighborhood and we compute the mean amount of each most common venues by neighborhood.

Finally, we get two datasets: the dataset 5 containing the neighborhoods and the name of the ten most common venues and the datasets 6 containing the mean amount of each venues per neighborhood.

	Neighborhood	1 most common venue	2 most common venue	3 most common venue	4 most common venue	5 most common venue	6 most common venue	7 most common venue	8 most common venue
0	Adelaide,King,Richmond	Coffee Shop	Restaurant	Hotel	Theater	Beer Bar	Pizza Place	Cosmetics Shop	Japanese Restaurant
1	Agincourt	Chinese Restaurant	Coffee Shop	Pharmacy	Restaurant	Sandwich Place	Indian Restaurant	Sushi Restaurant	Bank
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	Chinese Restaurant	Coffee Shop	Park	Gas Station	Pizza Place	Dessert Shop	Bakery	Vietnamese Restaurant
3	Albion Gardens,Beaumond Heights,Humbergate,Jam...	Coffee Shop	Pizza Place	Indian Restaurant	Fast Food Restaurant	Grocery Store	Convenience Store	Park	Skating Rink
4	Alderwood,Long Branch	Coffee Shop	Fast Food Restaurant	Department Store	Pizza Place	Clothing Store	Burger Joint	Pharmacy	Restaurant

Figure 5: Dataset 5

Neighborhood	Zoo Exhibit	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	American Restaurant	Amphitheater	Animal Shelte
Adelaide,King,Richmond	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.0
Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0
Agincourt North,L'Amoreaux East,Milliken,Steel...	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0
Albion Gardens,Beaumond Heights,Humbergate,Jam...	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0
Alderwood,Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0

Figure 6: Dataset 6

3.2 Clustering method

In this section, we segment the neighborhood of Toronto. We use the k-means clustering method for this segmentation. We decide to choose three different clusters that are dominated by the most common venues. With the k-means cluster method, we get the label for each cluster denoted by $0, 1$ and 2 by training the k-means on the dataset 6. Since we get the the label from the kmeans clustering method, we add the label features to the dataset 5 to display the venues on the Toronto map. The final dataset has the following features:

1. Postal Code
2. Borough
3. Neighborhood
4. Latitude
5. Longitude
6. 1 to 10 most common venues
7. Cluster labels

The neighborhoods will be display on the map with their latitude and longitude and the venues will be display according to the cluster labels.

4 Results

In this section, we display our features on the world map. We use the latitude and the longitude of the city of Toronto.

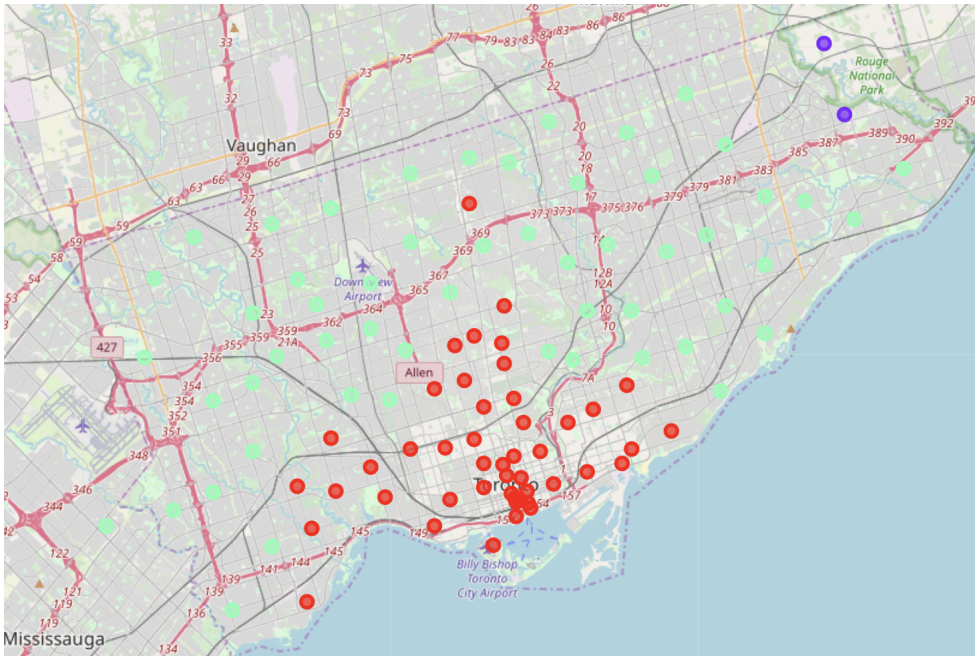

Figure 7: Toronto map

We observe three clusters on the Toronto map. We can clearly distinguish three specific locations:

1. The red cluster situated near the sea.
2. The blue cluster near to the National Park.
3. The light green cluster in the city center.

The next step is to analyze each cluster. We observe which are the most common venues in each clusters.

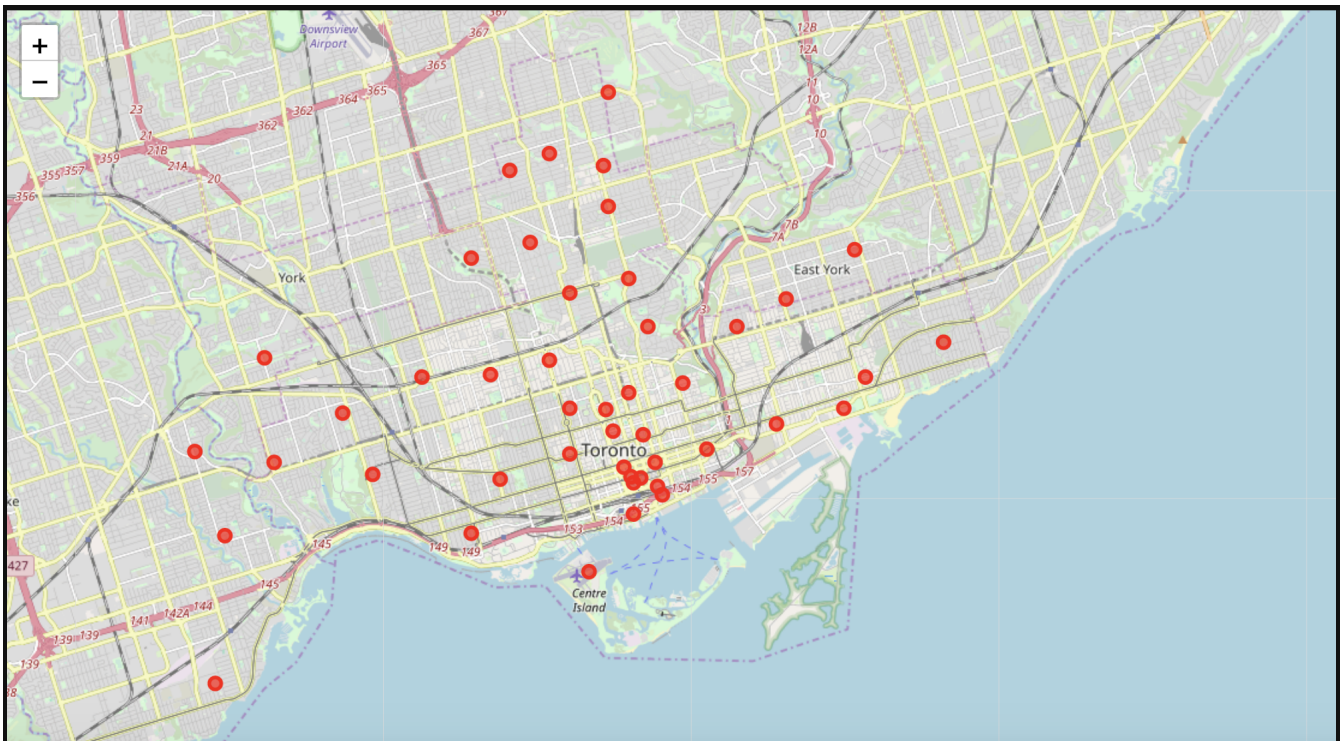

Figure 8: Toronto cluster 0 map

	Neighborhood	1 most common venue	2 most common venue	3 most common venue	4 most common venue	5 most common venue	6 most common venue	7 most common venue	8 most common venue
2	Harbourfront	Coffee Shop	Café	Park	Theater	Restaurant	Japanese Restaurant	Bakery	Gastropub
4	Queen's Park	Coffee Shop	Café	Japanese Restaurant	Park	Mexican Restaurant	Breakfast Spot	Gastropub	Restaurant
9	Ryerson,Garden District	Coffee Shop	Café	Restaurant	Gastropub	Japanese Restaurant	Park	Italian Restaurant	Pizza Place
14	Woodbine Heights	Park	Coffee Shop	Pizza Place	Thai Restaurant	Café	Gastropub	Ethiopian Restaurant	Pharmacy
15	St. James Town	Coffee Shop	Café	Restaurant	Hotel	Italian Restaurant	Farmers Market	Japanese Restaurant	Gastropub
16	Humewood-Cedarvale	Italian Restaurant	Coffee Shop	Bank	Café	Bakery	Park	Caribbean Restaurant	Indian Restaurant
19	The Beaches	Coffee Shop	Pub	Breakfast Spot	Beach	Bakery	Thai Restaurant	BBQ Joint	Bar
20	Berczy Park	Coffee Shop	Hotel	Café	Japanese Restaurant	Restaurant	Farmers Market	Beer Bar	Gastropub

Figure 9: Toronto venues in cluster 0

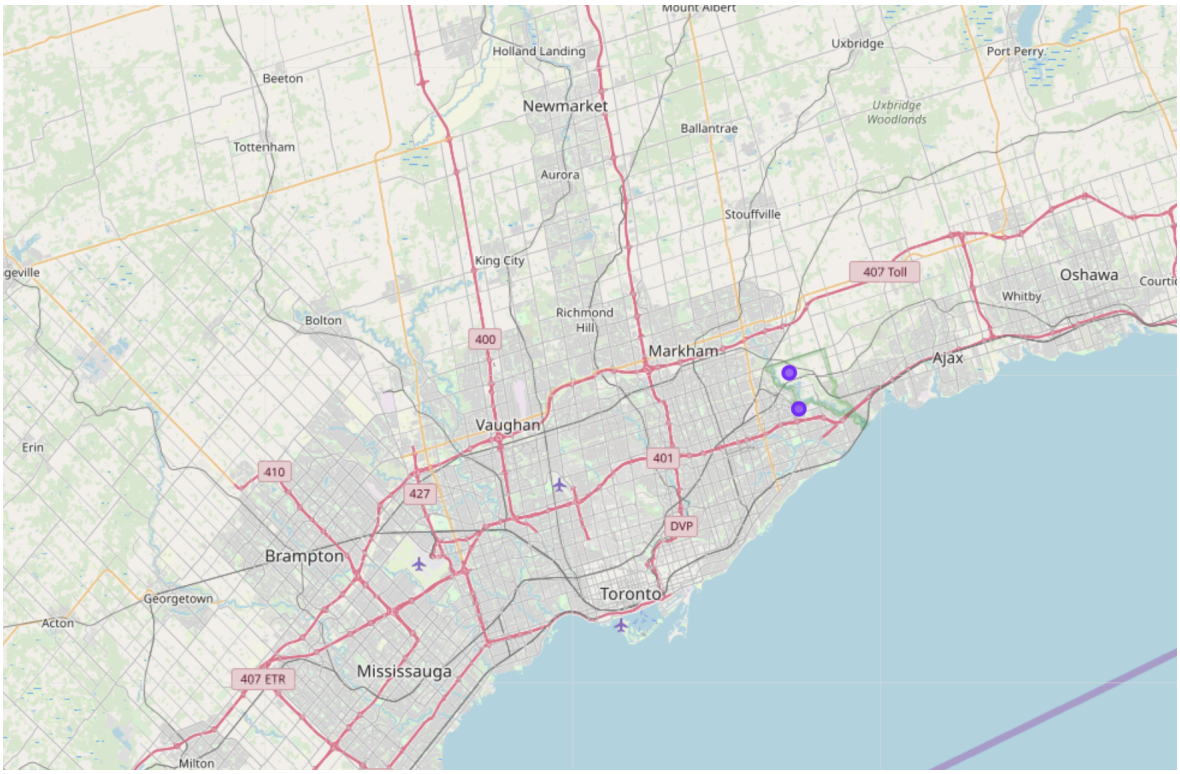

Figure 10: Toronto cluster 1 map

	Neighborhood	1 most common venue	2 most common venue	3 most common venue	4 most common venue	5 most common venue	6 most common venue	7 most common venue	8 most common venue
6	Rouge,Malvern	Zoo Exhibit	Fast Food Restaurant	Restaurant	Gas Station	Other Great Outdoors	Pizza Place	Zoo	Mediterranean Restaurant
95	Upper Rouge	Sculpture Garden	Golf Course	Grocery Store	Trail	Playground	Farm	Empanada Restaurant	Doner Restaurant

Figure 11: Toronto venues in cluster 1

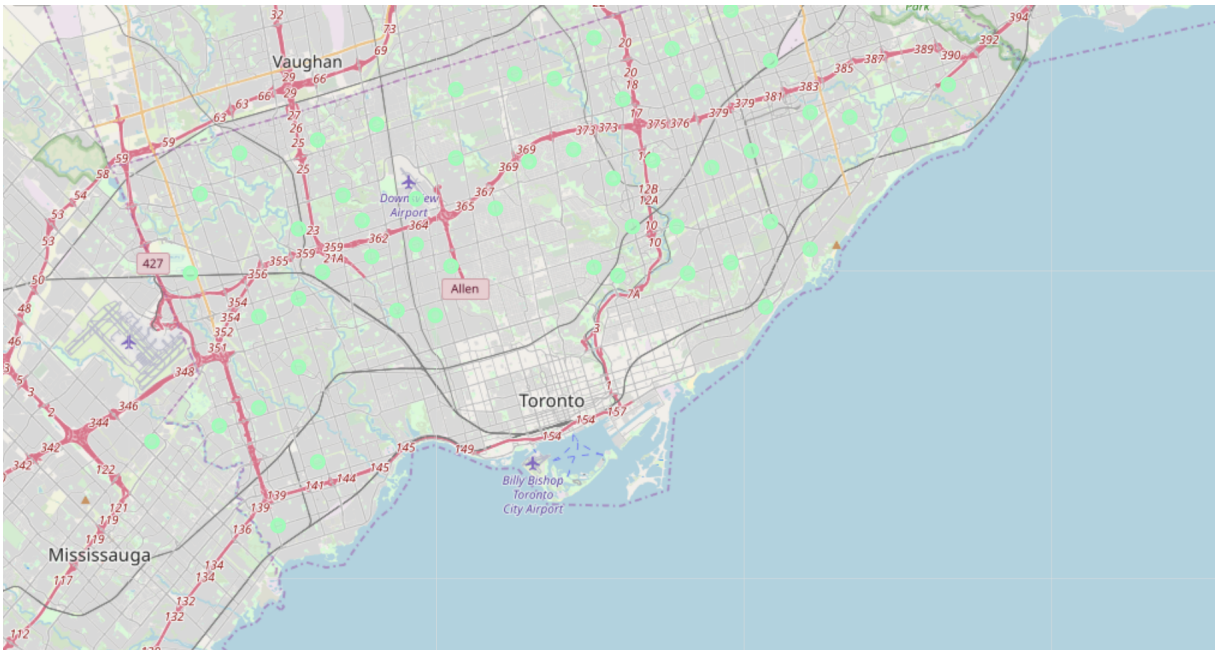

Figure 12: Toronto cluster 2 map

Neighborhood	1 most common venue	2 most common venue	3 most common venue	4 most common venue	5 most common venue	6 most common venue	7 most common venue
Parkwoods	Coffee Shop	Japanese Restaurant	Gas Station	Pizza Place	Supermarket	Sandwich Place	Chinese Restaurant
Victoria Village	Coffee Shop	Fast Food Restaurant	Gym	Sandwich Place	Gym / Fitness Center	Japanese Restaurant	Grocery Store
Lawrence Heights,Lawrence Manor	Clothing Store	Coffee Shop	Furniture / Home Store	Restaurant	Fast Food Restaurant	Dessert Shop	Grocery Store
Islington Avenue	Pharmacy	Coffee Shop	Bank	Park	Shopping Mall	Golf Course	Liquor Store
Don Mills North	Coffee Shop	Japanese Restaurant	Restaurant	Pizza Place	Italian Restaurant	Supermarket	Bank
Woodbine Gardens,Parkview Hill	Pizza Place	Sandwich Place	Park	Coffee Shop	Skating Rink	Pharmacy	Fast Food Restaurant
Glencairn	Clothing Store	Coffee Shop	Furniture / Home Store	Restaurant	Sushi Restaurant	Fast Food Restaurant	Fried Chicken Joint
Cloverdale,Islington,Martin Grove,Princess Gar...	Coffee Shop	Convenience Store	Fish & Chips Shop	Sandwich Place	Restaurant	Pizza Place	Pharmacy
Highland Creek,Rouge Hill,Port Union	Coffee Shop	Breakfast Spot	Pizza Place	Pharmacy	Pet Store	Sandwich Place	Hotel

Figure 13: Toronto venues in cluster 2

According to Figure 9 and 8, the most common venues near to the see are mostly restaurants, Café or pub. In Figure 11, we have only two neighborhoods that appear. The venues are Zoo, Garden and other outdoors venues. The Figure 13 represents the neighborhoods that are near to the city center and we observe that we still have restaurants, but we have also bank, supermarkets, stores and also Gyms and Fitness centers which are our target.

We observe that the gym and fitness centers are not the most frequented venues in each of the cluster. The gyms are only classified as the third most common venue in the neighborhood of Victoria Village. The gyms and fitness centers are only located in the cluster 0 and cluster 2 since cluster 1 is dedicated to parks, outdoor venues and restaurants due to its location.

5 Discussion

According to our analysis, Gold's Gym should better open its fitness center in the neighborhoods of cluster 0 or cluster 2, where there are restaurants, stores, supermarkets or other services such as hotel or bank. The gyms are popular in these neighborhoods since clients or employees of these services can go directly to the gym after works, but it should avoid neighborhood where there is already a gym, where concurrency can occur. However, it is not a good idea to build a fitness center in the neighborhoods of cluster 1 due to its proximity with the National Park. People prefer to walk or run in the park to workout or to visit the zoo in the neighborhood of the park.

6 Conclusion

To conclude, based on our assumptions and analysis, Gold's gym would better open its center in the fitness in the city center and avoid places where there is already a gym. The company should choose the neighborhood located in the cluster 0 or 2 where all services are located.