Coursera Capstone

IBM Applied Data Science Capstone – Week 2

Opening a new business opportunity in Toronto

By: Harsh Vaishnav

April 2021

Introduction

With increase in demand in many areas having increase in population and the COVID-19 pandemic all over the globe, there is an increment in chances of opening up small business venture especially after the tremendous effort of Canadian government to support local businesses. It would be great support to the city's tourism sector as well. As the places will open up eventually to local as well as foreign people, the local business opportunities will start gaining profit. Hence the idea is to find a place suitable to open up any kind of business venture (small sector). Businesses like healthcare, bakeries, restaurants, groceries will not only provide easy accessibility of resources but will also support local population by providing employment opportunities to some of them. There are so many benefits of opening up a small business venture to some areas. There are lots of factors involving setting up a business especially a small-scale business since many resources are at stake and the hopes and dreams of people setting up them. Location plays a very important role creating an environment for a profitable business venture. It is quite understandable that lot of other factors involving choosing a place to set up business venture. But for the sake of understanding and simplicity we will consider only location. There are several key business challenges in small-scale industry also such as lack of service orientation, lack of facilities, safety concern, competition with nearby vendors and difficulty in maintain optimum levels of profitability. All these factors play key role in determining proper setup of business location. Since we are doing something much simpler version, we are going to ignore these real-life factors and solely focus on Data analysis of cluster-based location search in a neighborhood of city of Toronto. I hope that although this analysis is a very limited in its sense but it will be able to present a rough idea of how to set up a business venture in neighborhood of Toronto city.

Business Problem

The objective of this capstone project is to analyze and then determine the best possible locations for opening a small business venture in a big city like Toronto. By using machine learning algorithm and data science methods like clustering, this project aims to provide an answer to the following question: If a person is looking to open a small-scale business venture in Toronto, in which location would you recommend them to open one?

Target Audience of This Project

This project is going to be helpful for the persons and aspiring businessmen in Toronto city who are looking to open new small-scale setup of their business venture or small business startup idea. A recent study by Loco BC, which looked at the impacts of local business on the economy, showed that if consumers made a 10% shift from big-store buying to local businesses, it would create $4.4 billion for the BC economy and 14,000 jobs. One news source also mentioned in their article that “The government is creating a $100 million Tourism and Hospitality Small Business Support Grant, giving those businesses one-time payments of up to $20,000 in 2021.” As part of the 2021-22 budget, Ontario Finance Minister Peter Bethlenfalvy revealed small businesses will get a second grant of up to $20,000, replicating a $1.7 billion program the government was forced to introduce in December amid a province-wide shut down. This will attract a large amount of population and small-scale industry have a chance to succeed really well, provided they can be built in locations where their chances of profitability are higher. Currently they are more focused on Tourism sector and tech setups involving work from home as well as delivery services.

Data

To do this project we will need the following data:

1. List of Postal Codes of city of Toronto. This will be done web scraping Wikipedia page of postal codes of Toronto.
2. The Latitude and Longitude co-ordinates of the neighborhoods. These will be helping in setting up map and visualize clusters.

How to extract Data

The Wikipedia page with the list of postal codes of neighborhoods in Toronto. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains the list of the neighborhoods in Toronto, nearly 103 of them. Then we will have latitudes and longitudes. Then we will use the geolocator to locate the city and clusters.

After that we will use the Foursquare API to get the data of the neighborhoods. The Foursquare Places API provides location-based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag. With the enhanced database and new enterprise-grade API, Foursquare further cements its status as the #1 independent provider of point-of-interest (POI) data for enterprises and developers. This project will require data science skills like working with Foursquare API, data cleaning, data visualization, data wrangling, data analysis and machine learning with help of K-Means Clustering.

Work Methodology

First, I installed packages like NumPy, pandas, matplotlib, etc and then I installed some packages like beautifulsoup, geopy and folium. Then I needed to create database on the list of postal codes in Toronto which I found on a Wikipedia page:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Now, extracting data became quite simple as the data table was available. The data looks quite clean but since users can Wikipedia content, sometimes data may not be reliable. But in our case, it will be as with very negligible margin of error as we will view and confirm by map later. (Few points after decimal may differ). After that I have made the table contents little bit clearer by using beautifulsoup library package. The data looks like a json data with headers namely Postal Code, Borough and Neighborhood respectively. The size of data frame is then confirmed to be 103 rows × 3 columns. Then I have added latitudes and longitudes from http://cocl.us/Geospatial_data . After that I used geolocator to visualize map of Toronto city along with clusters. Then the resulting data frame is to be used for further analysis.

After that I have used Foursquare API to retrieve the top 50 venues that are within a radius of 300m. In order to use Foursquare API, I had to register a developer account and then I had to make an app which gave me a 'Client ID' and 'Client Secret' which helped me in making the API call using the postal codes. I made a loop in Python notebook and passed the postal codes of neighborhoods until there Data Frame had no neighborhoods left. The Foursquare API returned the data in the form of a JSON file. By using this JSON file I extracted, I got to know the name of the venue, its category and its co-ordinates. Now with this data, we can check the number of venues. We can also check different type of categories each of these returned venues can be categorized. Then, I created a Data Frame table with each of those neighborhoods, neighborhood latitudes, neighborhood longitudes, venue latitude, venue longitudes along with their category for that neighborhood.

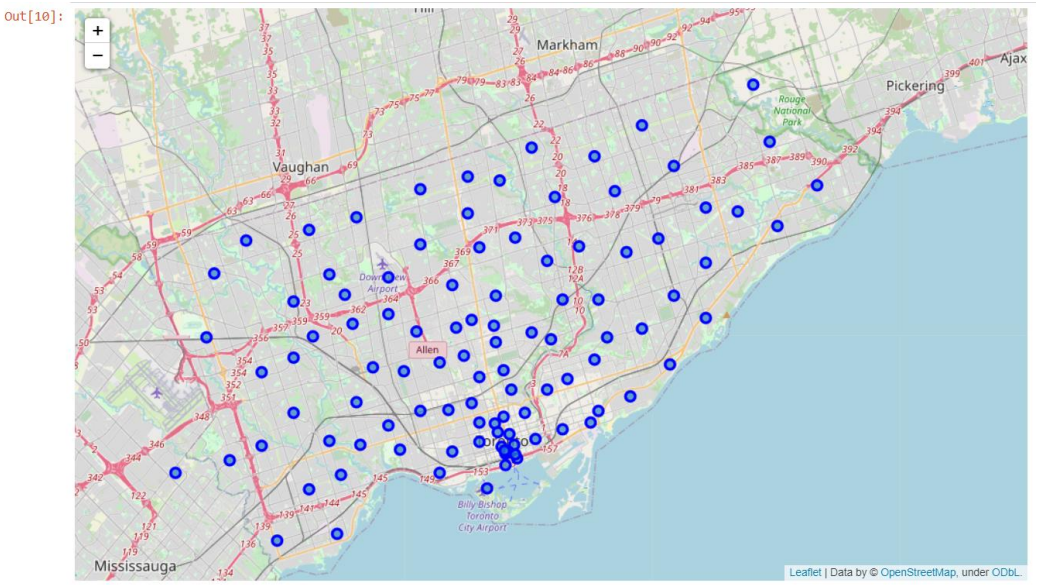
After that the frequencies of each of categories in those particular neighborhoods were calculated and hence a new data frame was obtained. The

priority for this project was “venues” since I had to analyze all small-scale business venue types in a particular cluster in a particular neighborhood. Hence, I created a data table for analyzing these conditions. Then I grouped them according to the neighborhoods. In order to further analyze I observed top 5 venues along with their frequencies in their neighborhood. Now I sorted out the most common venues like 1st common,2nd common… so that data obtained would tell us that which neighborhood lacks what kind of business venues hence one can easily identify which small scale business is lacking in which neighborhood. The reason is to obtain a neighborhood which would give optimum result to our business venture and for that demand in that area is very necessary.

At the end, I used clustering the data points using ‘k-means clustering’. I already explained in notebook, 'k-means clustering' is a machine learning technique used for clustering data. I have selected the number of centroids for clustering, i.e., the number of clusters I wanted to group the data into, or the value for particular 'k'. The data points were grouped into the cluster of the closest centroid (Expectation). Then the new centroid/mean of the cluster is calculated (Maximization). The Expectation Step and Maximization step are in an iteration and continue until none of the data points are left that need to be grouped, or the centroid for each cluster does not change. Then I used this technique for clustering the neighborhoods for sake of this project. I kept k = 5 so that it will be helpful in understanding which neighborhoods have a higher frequency of a particular business venue, which have a moderate frequency and which have a lower frequency. This will help us in identifying the neighborhoods where we can set up a small-scale business venture.

Next, I examined the clusters by varying the value of cluster labels as 0,1,2,3 and 4 to find out the nth common venue in that particular neighborhood. The results were quite satisfying and are discussed in Result section.

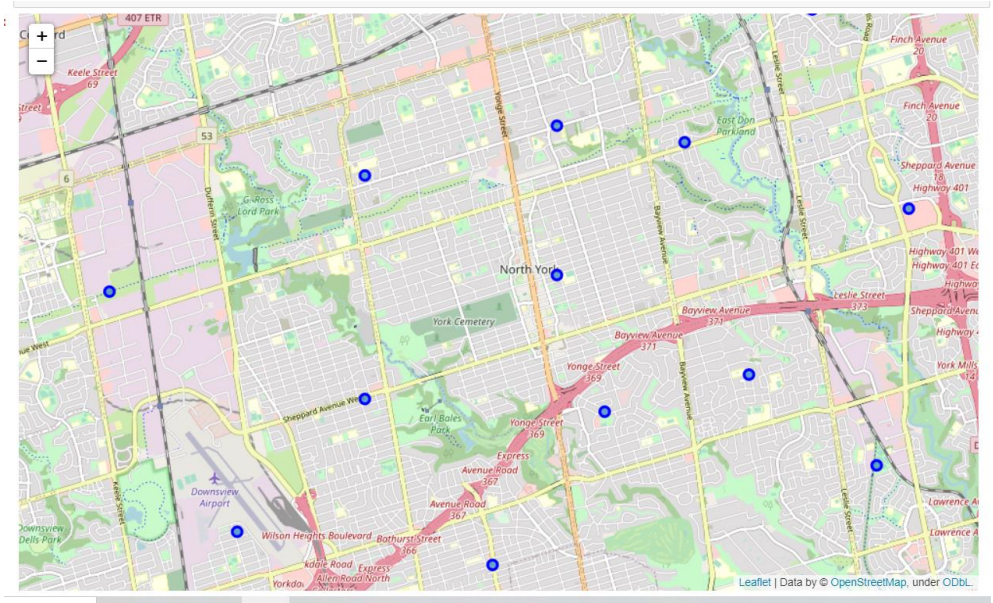
Results



Map showing total clusters in Toronto City

The Results obtained by Analyzing Map of Toronto city's neighborhood have been described here. The map in figure 1 shows all clusters in city in blue dots. Now we will observe each cluster labels individually.

1. **Cluster label '0':**

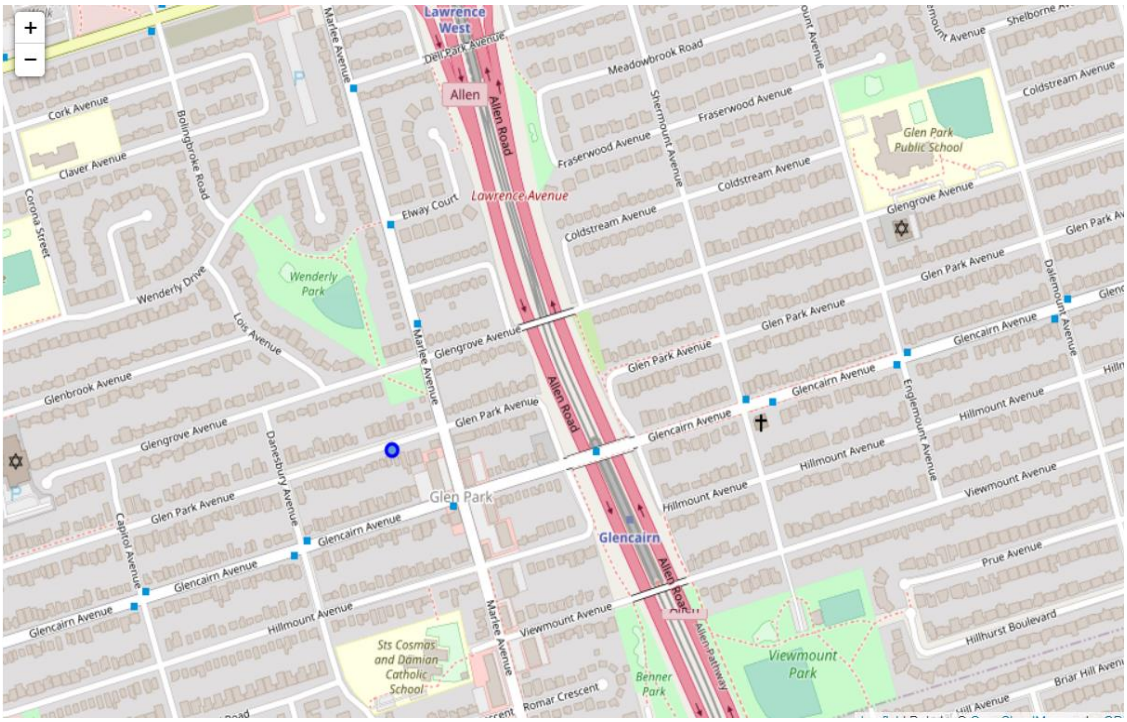


Clusters across North York Region

Points of interest: (top 5)

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	North York	0.0	Coffee Shop	Pizza Place	Hockey Arena	Portuguese Restaurant	Women's Store
3	North York	0.0	Japanese Restaurant	Caribbean Restaurant	Café	Gym	Women's Store
4	North York	0.0	Pizza Place	Park	Bakery	Japanese Restaurant	Department Store
5	North York	0.0	Gym	Coffee Shop	Restaurant	Sandwich Place	Clothing Store
6	North York	0.0	Golf Course	Mediterranean Restaurant	Pool	Fast Food Restaurant	Dog Run

2.Cluster label '1':



Clusters across Lawrence Heights

Points of interest: (all)

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Lawrence Manor, Lawrence Heights	1.0	Clothing Store	Accessories Store	Boutique	Coffee Shop	Vietnamese Restaurant

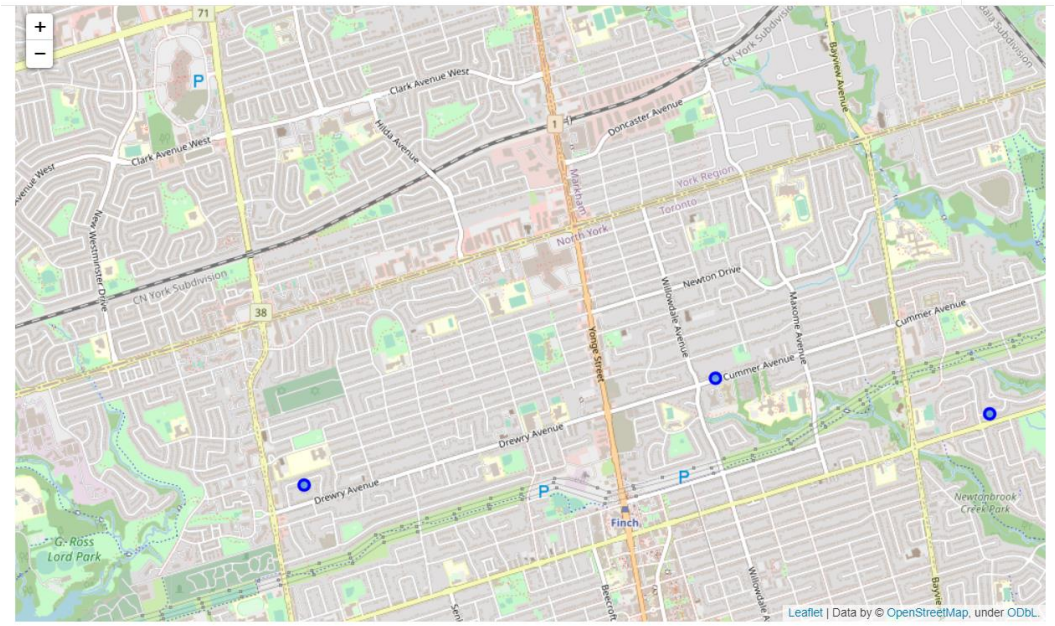
3.Cluster label '2':

Clusters across Downsview Central and Emery

Points of interest: (all)

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
17	Downsview Central	2.0	Food Truck	Home Service	Baseball Field	Women's Store	Dim Sum Restaurant
19	Humberlea, Emery	2.0	Construction & Landscaping	Baseball Field	Women's Store	Diner	Clothing Store

4.Cluster label ‘3’:

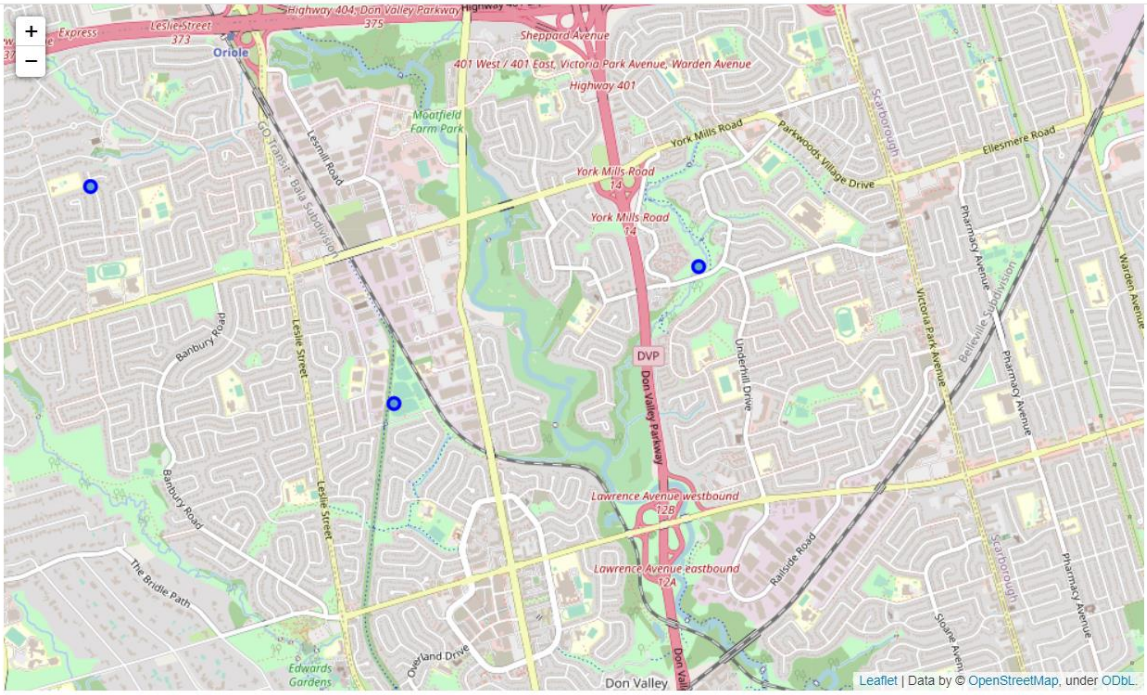


Clusters across Newtonbrook

Points of interest: (all)

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
16	Willowdale, Newtonbrook	3.0	Piano Bar	Women's Store	Caribbean Restaurant	Chocolate Shop	Clothing Store

5.Cluster label ‘4’:



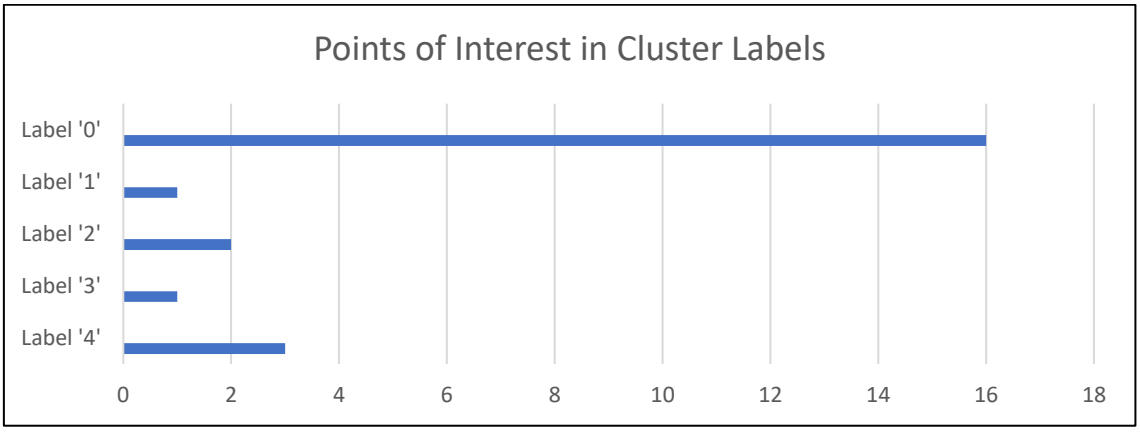
Clusters across York Mills West, Downsview East and Parkwoods

Points of interest: (all)

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Parkwoods	4.0	Food & Drink Shop	Park	Women's Store	Department Store	Chocolate Shop
11	Downsview East	4.0	Park	Airport	Business Service	Dim Sum Restaurant	Chocolate Shop
22	York Mills West	4.0	Convenience Store	Park	Women's Store	Dim Sum Restaurant	Chocolate Shop

The following is the summary of applying ‘k-means clustering’ based on the frequency value for Points of interest (presence of already existing business ventures):

Cluster Label	Total points of interest	Remarks (Presence of existing business ventures)
0	16	High presence
1	1	Low Presence
2	2	Low Presence
3	1	Low Presence
4	3	Moderate Presence



The above provided results (analyzing map and clusters) show that in cluster label 1,2 and 3 there is low presence of business ventures. Hence, they are ideal places to start a new business. Whereas the points of interest are comparatively high in clusters 0 and 4. Hence these areas can be avoided to set up a business venture.

Discussion

The cluster areas concentrated with Business venues in the city of Toronto are scattered across the city. We can see that Clusters 1, 2 and 3 show the low presence of business venues in the neighborhood while Cluster 0 and 4 shows high presence of business venues. Let's discuss about the clusters in 1,2 and 3, in our case low presence of business ventures. This neighborhood can be recommended to start up any business venture owing to less presence and setting up any business idea will be profitable as long as there is no other interference from other competitor in future. We can check out the list of most common venues for our reference and cancel out the ideas having most common venues. The idea is to maximize profit with less competition. Next, is the cluster label 4 which is not recommended but can be considered if other factors are suitable i.e., location of venue (must be on a major roadway), demand of local population, presence of tourist spots, distance of other competing venues across neighborhoods. The Cluster 1 is not a recommended to build a start a business because of presence of so many other business venues as they are already well established there and until and unless there is very good reason to setup, the business can experience loss there. I have kept this project simple by considering only the number of already established business venues in the neighborhood, I won't discuss about the other factors here. Hence the idea is clear here.

Conclusion

In this Project, first the business problem was identified. Then a through investigation was done in terms of research work in order to identify condition of setting up business ventures across the neighborhoods of Toronto city. Then the relevant data was extracted and cleaned for the purpose of data visualization. Then coordinates were obtained using Foursquare API tool. The resulting database was analyzed and examined using an unsupervised method of clustering 'k-means clustering' to cluster the data depending on their similarities. Finally, the results were obtained in terms of clusters which were further analyzed to obtain most profitable spot in order to start a business. In order to summarize the answer for the business problem the following lines would be conclusive:
Neighborhoods in Cluster 1,2 and 3 would be highly recommended to start a business venture because of less presence of existing business venue which would provide less competition and more profit.

References

Neighborhoods in Toronto:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare Developer Documents:

https://developer.foursquare.com/docs

Loco BC:

https://www.locobc.ca/cpages/resources

Ontario small business and tourism sector to receive support payments:

"https://toronto.ctvnews.ca/ontario-small-business-and-tourism-sector-to-receive-support-payments-1.5360598"