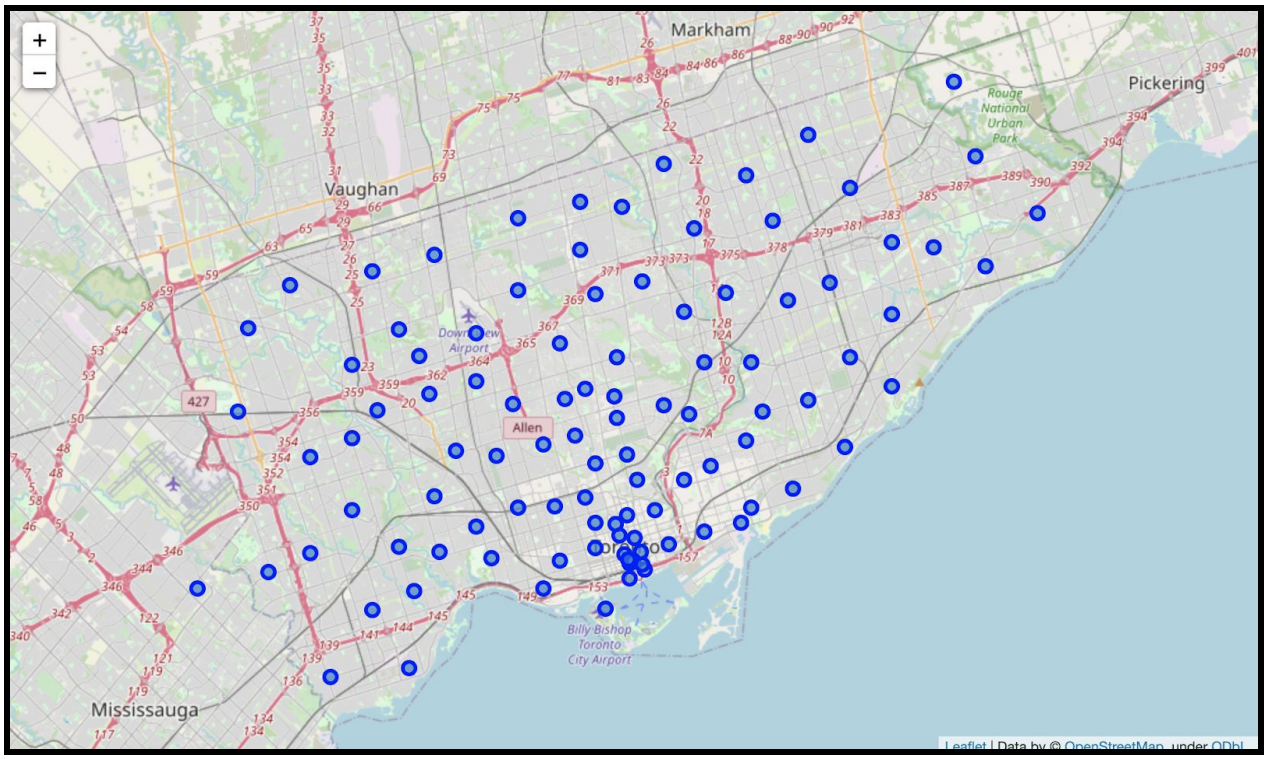
IBM Data Science Professional Certificate offered by Coursera
Capstone Project

Battle of the Neighborhoods
Data Science using Foursquare API

By Eshita Goel



Introduction

This report is part of the Capstone Project for IBM's Applied Data Science Professional Certificate offered by Coursera. This is part of the final course in this 9-course series.

We will be using several data visualization techniques, in particular, we will be making use of Foursquare API to retrieve location data for the state of Toronto in Canada and use this data to perform data analysis.

1. Introducing the Project

1.1 Background

Toronto is a big state with numerous neighbourhoods. Each neighbourhood has its share of shops, restaurants, cafes, beaches etc. The different places add to the vibrance of the state and make it a great place to live and travel for others. The place also provides many opportunities for entrepreneurs, especially those that want to start afresh. Being so close to the capital of Canada, it receives its fair share of foot traffic from foreigners. In this project, we aim to analyze the various places in Toronto using location data imported from the Foursquare API.

1.2 The Problem

If a person has to open a new restaurant in a neighbourhood in Toronto then what neighbourhood should he/she choose based on the restaurant type. And if they have a specific restaurant type in mind then what place would be best ensuring a good amount of customer traffic but also keeping in mind the amount of competition. We also cluster the neighbourhoods based on the most popular spots to visit so that we can make it easier for a new business person to choose the right neighbourhood for their restaurant/shop.

1.3 Interest

This report will be useful for those who want to start a new business in the state of Toronto. It will also be helpful for those who want to travel to Toronto and want to visit specif locations based on their interest. For example, if a tourist wants to visit multicultural restaurants in Toronto, then what neighbourhoods are best? They should ideally visit those neighbourhoods where multicultural restaurants are popular among the people. The clustering of the neighbourhoods based on the most visited spots allows people to decide where to travel and explore more. Clusters tell us what neighbourhoods are fairly similar to each other so the person can skip travelling to many of the same neighbourhoods.

2. Data Acquisition and Cleaning

2.1 Data Available

We make use of a few data sources to get the data required for this project. We get different kinds of Neighborhoods in Toronto along with the Boroughs and Postal codes from Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
We get the Latitudes and Longitudes for each postal code through the CSV file provided to us by the Coursera Applied Data Science Capstone Week 3 Module. We get the various location-related data, like the kinds of places in a particular neighbourhood, using Foursquare API. This data will include the type of shops, restaurants, cafes, beaches etc in each neighbourhood.

2.2 Acquiring the Data

We acquire the data about the various neighbourhoods, boroughs and postal codes from the Wikipedia page using Beautiful Soup. We put it into a data frame. The latitudes and longitudes are in a CSV file that can be read using pandas. We will make calls to Foursquare API using our credentials to acquire the location-related data.

	Postal Code	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

2.3 Cleaning the data

Once we have our data that includes the Postal Codes, Boroughs and Neighbourhoods in Canada, we drop all rows where the borough is unknown. We want to focus on the data that has been assigned a borough.

More than one neighbourhood can exist that has the same postal code, we combine such rows. We make a single row for each postal code and the subsequent neighbourhoods that are associated with that postal code would be put into the same row separated by commas.

Is a particular neighbourhood is not assigned but that row has an assigned borough, then the neighbourhood is considered to be the same as the borough.

We sort the data by the postal codes.

	index	Postal Code	Borough	Neighborhood
0	6	M1B	Scarborough	Malvern, Rouge
1	12	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	18	M1E	Scarborough	Guildwood, Morningside, West Hill
3	22	M1G	Scarborough	Woburn
4	26	M1H	Scarborough	Cedarbrae

Then we merge the data frame with the stored values of the latitudes and longitudes of each postal code.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Here we get our final data frame.

2.4 Feature Selection

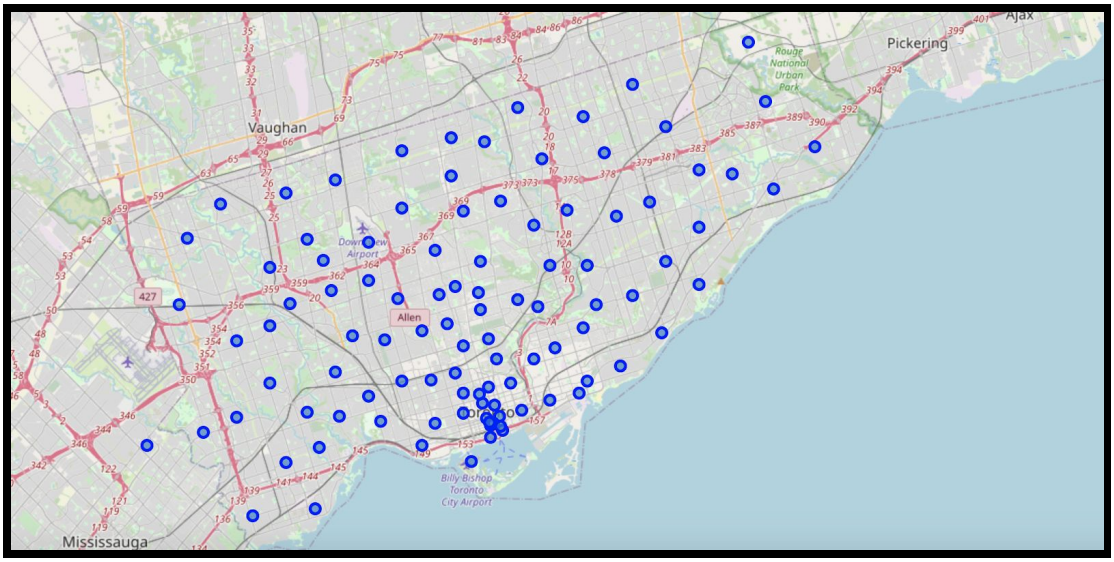
We want to focus on neighbourhoods in Toronto. So we drop all rows that have Boroughs outside of Toronto. We keep the data from Toronto. Finally, our data frame is ready to use.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
2	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

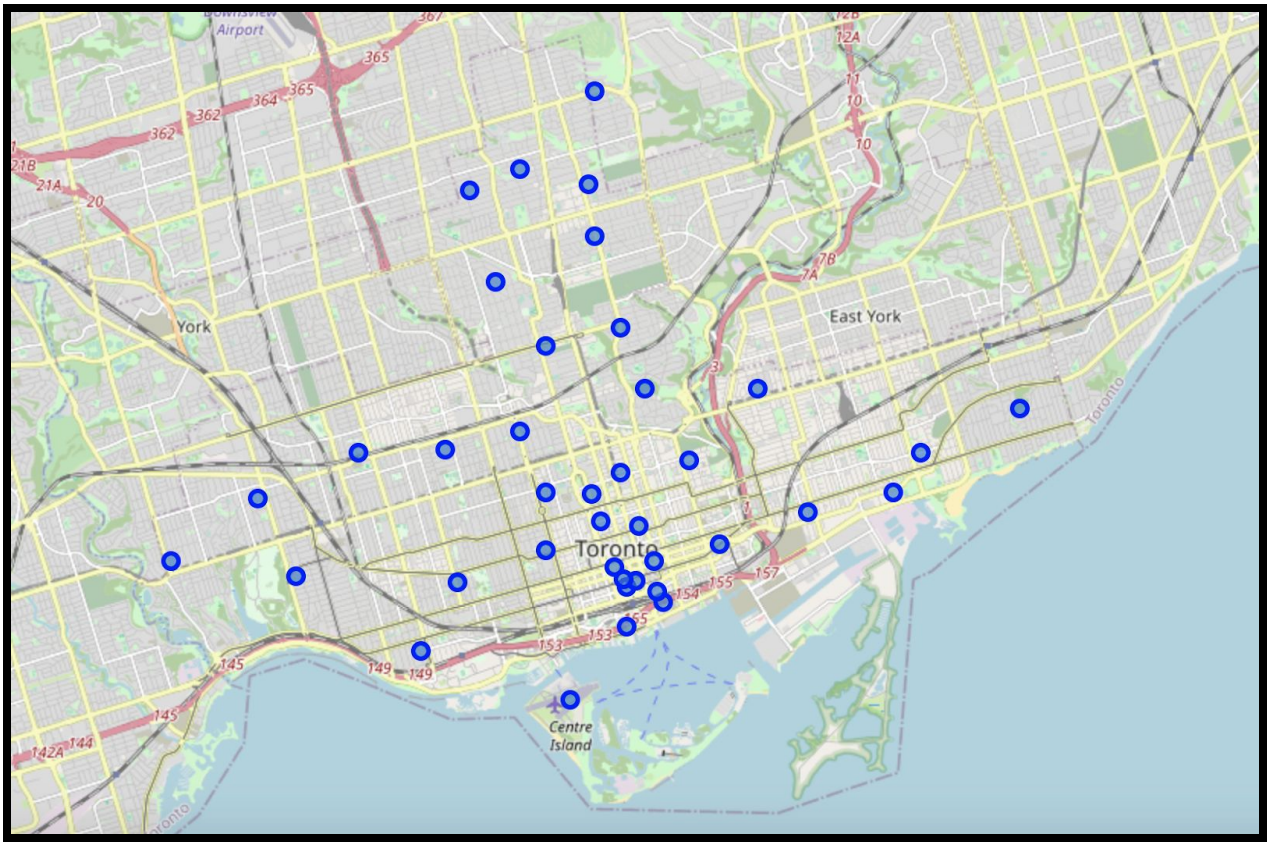
3. Exploratory Data Analysis

3.1 Visualizing the data

We can visualize the various neighbourhoods in Canada by drawing a map and plotting the neighbourhoods on top. This allows us to see what we are dealing with and how the neighbourhoods are scattered.



We need to focus on the neighbourhoods in Toronto. We have already made a separate data frame with data from Toronto (i.e. East, West, North and Downtown Toronto). We visualize this data using a map centred in Toronto but only plotting the neighbourhoods in Toronto.



3.2 Exploring the Neighborhoods

We now see how we can find out what venues are there in each neighbourhood in Toronto. We can do this by exploring any one neighbourhood.

We find what is the first neighbourhood in our list of neighbourhood sin Toronto. It is “The Beaches”. We use Foursquare API to get the top 100 venues that are in The Beaches within a radius of 700 metres. This gives us an idea of the types of locations that can be present in a neighbourhood.

	name	categories	lat	lng
0	Glen Manor Ravine	Trail	43.676821	-79.293942
1	The Beech Tree	Gastropub	43.680493	-79.288846
2	Beaches Bake Shop	Bakery	43.680363	-79.289692
3	Tori's Bakeshop	Vegetarian / Vegan Restaurant	43.672114	-79.290331
4	Ed's Real Scoop	Ice Cream Shop	43.672630	-79.287993

We can see that around “The Beaches”, few of the locations include a Trail, Gastropub, Bakery, Vegan Restaurant and Ice Cream shop.

Similarly, we can explore the other neighbourhoods and this information can be very useful for a potential business owner wanting to start a new business in any of the neighbourhoods.

4. Methodology

4.1 Exploring the various places with different categories

Our first step is making a dataset with all the neighbourhoods along with the different venues near that neighbourhood. This dataset will allow us to group the neighbourhoods together according to the similarity in the type of venues in each neighbourhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Danforth West, Riverdale	43.679557	-79.352188	MenEssentials	43.677820	-79.351265	Cosmetics Shop

Now we have a dataset of all the neighbourhoods and their corresponding venues along with the categories of the venues.

4.2 Grouping the neighbourhoods based on the topmost common venues

Using the above dataset, we can start to group the neighbourhoods based on the similarity of their topmost venues. If two neighbourhoods have the same top few venues then they can be groups together in the same row. We make use of the mean of the frequency of the occurrence of each category and combine the neighbourhoods with similar venues.

The top 5 rows of this data frame will look like :

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Coffee Shop	Seafood Restaurant	Cheese Shop	Farmers Market	Beer Bar	Bakery	Restaurant	Café	Cocktail Bar	Indian Restaurant
1	Brockton, Parkdale Village, Exhibition Place	Café	Coffee Shop	Nightclub	Breakfast Spot	Stadium	Restaurant	Bakery	Intersection	Italian Restaurant	Climbing Gym
2	Business reply mail Processing Centre, South C...	Park	Garden	Brewery	Farmers Market	Fast Food Restaurant	Burrito Place	Restaurant	Auto Workshop	Pizza Place	Skate Park
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Airport Terminal	Coffee Shop	Harbor / Marina	Boat or Ferry	Sculpture Garden	Rental Car Location	Boutique	Airport Food Court
4	Central Bay Street	Coffee Shop	Sandwich Place	Café	Italian Restaurant	Bubble Tea Shop	Japanese Restaurant	Department Store	Salad Place	Burger Joint	Discount Store

This clearly shows the similar neighbourhoods in the same row along with the top 10 most popular places in those neighbourhoods.

4.3 Clustering the data

We will cluster the data using **KMeans Clustering.**

We already have our grouped data where similar neighbourhoods have been grouped together based on the top venues in these neighbourhoods. We can now perform KMeans to associate each group into a cluster. We cluster these neighbourhoods into 5 clusters and label them accordingly.

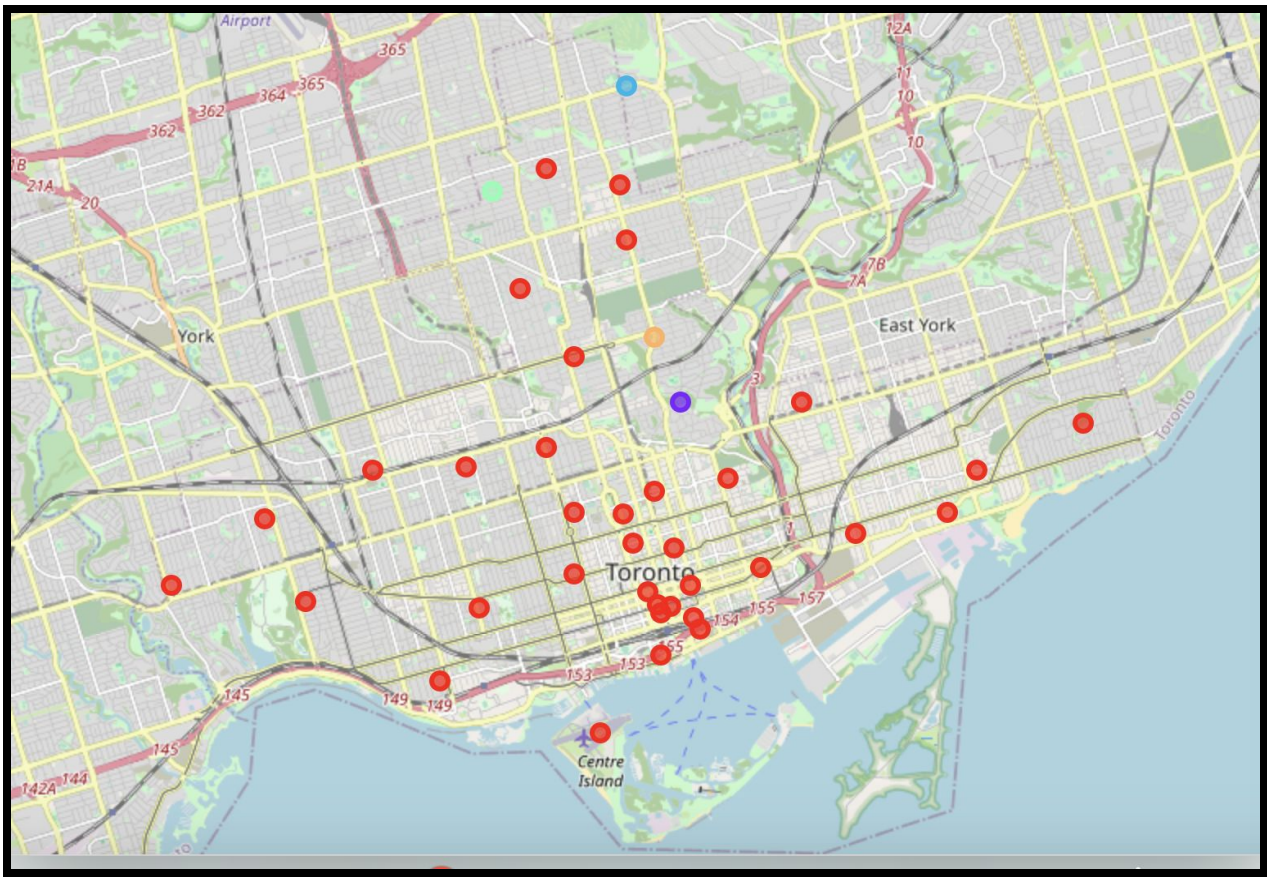
We will then analyze based on the clusters how similar the neighbourhoods are to each other. Neighbourhoods in the same cluster are likely to have similar categories of venues and thus opening a new branch for the business in the same cluster will not be ideal.

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	M4E	East Toronto	The Beaches	43.676357	-79.293031	0	Trail	Health Food Store	Pub	Women's Store	Cupcake Shop	Dumpling Restaurant	Donut Shop	Doner Restaurant	Dog Run	Distribution Center
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	0	Greek Restaurant	Coffee Shop	Italian Restaurant	Ice Cream Shop	Furniture / Home Store	Restaurant	Bubble Tea Shop	Bakery	Pub	Pizza Place
2	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572	0	Park	Board Shop	Food & Drink Shop	Sandwich Place	Light Rail Station	Italian Restaurant	Burrito Place	Liquor Store	Restaurant	Ice Cream Shop
3	M4M	East Toronto	Studio District	43.659526	-79.340923	0	Café	Coffee Shop	Gastropub	Bakery	Brewery	American Restaurant	Yoga Studio	Comfort Food Restaurant	Seafood Restaurant	Sandwich Place
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	2	Park	Swim School	Bus Line	Electronics Store	Dumpling Restaurant	Donut Shop	Doner Restaurant	Dog Run	Distribution Center	Discount Store

Neighborhood	Latitude	Longitude	Cluster Labels
The Beaches	43.676357	-79.293031	0
The Danforth West, Riverdale	43.679557	-79.352188	0
India Bazaar, The Beaches West	43.668999	-79.315572	0
Studio District	43.659526	-79.340923	0
Lawrence Park	43.728020	-79.388790	2

The Neighbourhoods have been assigned clusters.

Plotting clusters on a map centred in Toronto



5. Results

Most of the neighbourhoods in Toronto fall in the same cluster (Indicated in Red). Other than that there are 4 neighbourhoods each forming its own cluster.

Since the neighbourhoods were clustered based on the similarity in the categories of popular venues, it can be observed that most neighbourhoods have the same category of venues.

There are 4 neighbourhoods that don't fall in the majority category, suggesting that these have different than most neighbourhoods in Toronto.

6. Discussion

Our aim was to help potential business owners and tourists in picking our the right neighbourhood to travel or open a business in. For example, if a business owner wants to open a Vegan Cafe, he/she must choose the neighbourhood carefully making sure that there aren't any other popular Vegan Cafes in the same location. If so, he/she can face a lot of competition from an already established place.

But they should also keep in mind the interest of the people. People in a particular neighbourhood should be interested in the entrepreneur's business.

Also, when tourists visiting Toronto plan their holiday, they would want to visit different kinds of places. Visiting neighbourhoods that have almost the same characteristics would not be ideal.

This is why our aim was to cluster the neighbourhoods based on their similarity, i.e, their most popular locations. For example, if two neighbourhoods are popular for Indian restaurants, then both of them can be put into the same cluster. This alerts business owners that if they want to open a particular kind of restaurant and one neighbourhood is not ideal, then all neighbourhoods in that cluster are not ideal. This also gives them an idea about how they can scale profitably and open more branches in different clusters.

7. Conclusions

The clusters allow interested people to understand how similar neighbourhoods are in Toronto.

Using the data about the 10 most popular venues in each neighbourhood group allows people to choose the right neighbourhood for starting a new business or opening a new branch for their already existing business.

The similarity in neighbourhoods allows tourists to decide which places they should add on the to-visit list without making redundant choices. It also tells them what are the most popular places n each neighbourhood that they must visit.