<u>BATTLE OF NEIGHBOURHOODS</u>

Abhishek Purohit

August 16, 2020

1. Introduction

1.1 Background

A chain of restaurant owners in Ontario, Canada want to expand their business in other cities. Currently they have their restaurants open in cities like Ottawa, Brampton and Hamilton. They figured out that they would make much more profit by opening up a restaurant in Toronto as Toronto is the largest city of Canada and has large population density. So they want to open up a new restaurant some place nice with good neighbourhood in Toronto.

1.2 Problem

As Toronto is a very large city, they are having trouble figuring out which place to choose within Toronto for their new restaurant. We have to help them figure out which place to choose where their business will be good, they have less competition and nice people live around. They want to know about 3-4 such places so that they can decide for themselves which one is the best for them according to the type of their restaurant.

1.3 Interest

Obviously, people in the business of restaurant chains, hotels, etc. who are willing to expand their business in new cities would be very interested in my project for competitive advantage and business values. Others who are new to this business and want to set up their business in a new city might also be interested.

2. Data Acquisition and cleaning

2.1 Data Sources

There were two main datasets that were used for this project.

First Dataset: List of all the neighbourhoods in Toronto

Firstly, I used data from a Wikipedia page which provides information about all the neighbourhoods of Toronto, Canada. Then I used a web scrapping tool named BeautifulSoup for extracting the data in the form of a csv table from this Wikipedia page. This table consisted of 3 columns: Postal Code, Borough and Neighbourhood. The link for this Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M . After importing this table into a data frame, pre-processing this data frame and adding two more columns of Latitude and Longitude of each Neighbourhood, this data frame was ready for use. Final data frame will have 5 columns: Postal Code, Borough, Neighbourhood, Latitude, Longitude. And it will contain 39 rows having 39 unique neighbourhoods of Toronto and 4 unique Boroughs. For example, below photo depicts first 5 rows of the dataset:

	Postal Code	Borough	Neighbourhood	latitude	longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031

Second Dataset: List of different venues in the neighbourhoods of Toronto:

This dataset will be formed using the Foursquare API. Foursquare is a website that provides any information about a particular venue. I used the Foursquare location data to explore different venues in each neighbourhood of Toronto.

These venues can be any place. For example: Parks, Coffee Shops, Hotels, Gyms, etc.

Using the Foursquare location data, information about these venues can be taken and the neighbourhoods of Toronto can be easily analysed based on this information.

I will use the geographical coordinates from above dataset to generate this Location dataset. This dataset is named **toronto_venues.**

	Neighbourhood	Neighbourhood Latitude	Neighbourhood 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

```
toronto_venues.groupby('Neighbourhood').count()
```

2.2 Data Pre-processing

After the 2 datasets were obtained, pre-processing of the second dataset was needed so that it can be used for clustering algorithm easily. I pre-processed **toronto_venues** data frame using **one-hot encoding** tool. The pre-processed data was stored in a data frame named **toronto_onehot**.

Now, we have a dataset named **toronto_onehot** that is pre-processed and through one-hot encoding, it is ready to be used for clustering technique. But this dataset contains information about all the nearby venues like Park, Gym, Shops, etc. which is not necessary. As we are only interested in venues in 'food' category, therefore venues like Park, Gym, Playground are discarded from the **toronto_onehot** data frame.

Also we are looking for only those venues that are proper restaurants. Hence venues such as coffee shops, pizza places, bakeries etc. are not direct competitors of the restaurant business, so we don't care about those. Hence we will include in our list only venues that have 'restaurant' in category name, and we'll make sure to detect and include all the subcategories of different restaurants in the neighbourhood. For example, Afghan restaurant, Italian restaurant, etc. For this, we locate venues from **toronto_onehot** data frame that are restaurants only and store this in a new data frame named **toronto_restaurants**. This new data frame will now be used for clustering algorithm.

Also, a data frame named **venues_sorted** was also created which listed all the neighbourhoods of Toronto along with their respective 10 most common venues. This dataset would eventually help in visualising the solution. First 5 rows of this data frame is depicted in figure below:

	Neighbourhood	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	Cocktail Bar	Cheese Shop	Bakery	Farmers Market	Seafood Restaurant	Restaurant	Café	Beer Bar	Park
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Coffee Shop	Furniture / Home Store	Burrito Place	Restaurant	Italian Restaurant	Stadium	Intersection	Bar
2	Business reply mail Processing Centre	Yoga Studio	Spa	Auto Workshop	Brewery	Burrito Place	Butcher	Comic Shop	Farmers Market	Fast Food Restaurant	Garden
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Airport Terminal	Boutique	Sculpture Garden	Rental Car Location	Coffee Shop	Boat or Ferry	Bar	Harbor / Marina
4	Central Bay Street	Coffee Shop	Café	Sandwich Place	Italian Restaurant	Salad Place	Department Store	Japanese Restaurant	Burger Joint	Bubble Tea Shop	Poke Place

3. Methodology and Analysis

In **toronto_restaurants** data frame, I also added a column containing total number of restaurants in that neighbourhood. This will help us in making good clusters using K-Means clustering algorithm.

Now I use K-Means clustering algorithm to make clusters of dataset so that our analysis of the neighbourhoods is easy. For this I set number of clusters to be 5. The input for this clustering algorithm was **toronto_restaurants** data frame.
After the clusters were made, I merged the first dataset and the **venues_sorted** data frame and inserted cluster labels also. The result data frame was named **toronto_merged** which looked like this:

	Postal Code	Borough	Neighbourhood	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 Com V
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	1	Coffee Shop	Bakery	Pub	Café	Park	Breakfast Spot	Theater	S
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	3	Coffee Shop	College Cafeteria	Beer Bar	Smoothie Shop	Sandwich Place	Café	Portuguese Restaurant	Pe Resta
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	2	Clothing Store	Coffee Shop	Café	Cosmetics Shop	Bubble Tea Shop	Japanese Restaurant	Bookstore	I I
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	2	Coffee Shop	Café	American Restaurant	Cosmetics Shop	Clothing Store	Restaurant	Cocktail Bar	Fa M
4	M4E	East Toronto	The Beaches	43.676357	-79.293031	1	Trail	Neighborhood	Health Food Store	Pub	Doner Restaurant	Diner	Discount Store	Distrib C
5	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	0	Coffee Shop	Cocktail Bar	Cheese Shop	Bakery	Farmers Market	Seafood Restaurant	Restaurant	
6	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	0	Coffee Shop	Café	Sandwich Place	Italian Restaurant	Salad Place	Department Store	Japanese Restaurant	B
7	M6G	Downtown Toronto	Christie	43.669542	-79.422564	1	Grocery Store	Café	Park	Nightclub	Baby Store	Restaurant	Diner	Athlet S
8	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568	2	Coffee Shop	Café	Clothing Store	Restaurant	Hotel	Gym	Steakhouse	Resta
9	M6H	West Toronto	Dufferin, Dovercourt Village	43.669005	-79.442259	1	Bakery	Pharmacy	Gym / Fitness Center	Middle Eastern Restaurant	Music Venue	Park	Café	Bre

Next part was Analysis of each cluster to get the correct neighbourhood. I calculated total number of neighbourhoods and total number of restaurants for each cluster. Then I calculated Restaurant/Neighbourhood ratio and found that this ratio was lowest for cluster with cluster label=3. Hence this cluster was chosen for further analysis.

Cluster 4 consisted of total 7 neighbourhoods. Out of these, 2 had very high total number of restaurants, therefore these 2 neighbourhoods were discarded. Out of the remaining 5 neighbourhoods, 1 more were discarded because they had Restaurant as their most common venue more than once in the **toronto_merged** data frame and hence these neighbourhoods were not suitable for Restaurant business and hence discarded.

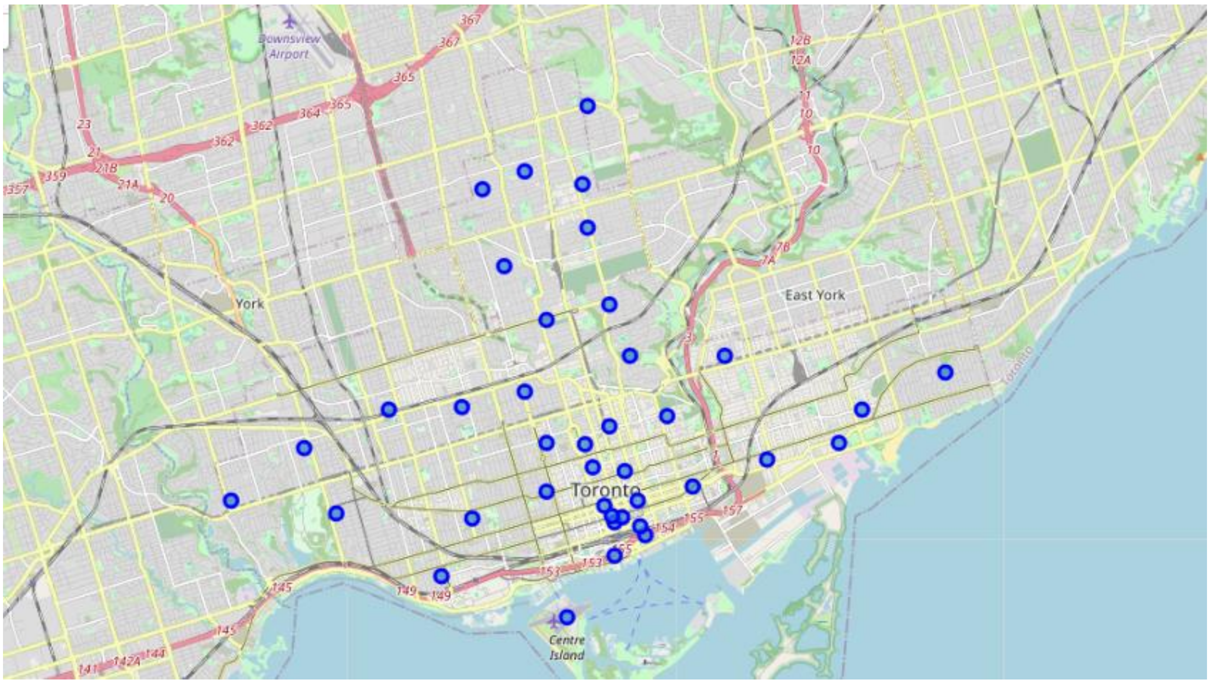
The final dataset contains all the information about these remaining 4 neighbourhoods:

	Postal Code	Borough	Neighbourhood	latitude	longitude	Cluster Labels	1st Most Common Venue	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
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	3	Coffee Shop	College Cafeteria	Beer Bar	Smoothie Shop	Sandwich Place	Café	Portuguese Restaurant	Persian Restaurant
26	M4S	Central Toronto	Davisville	43.704324	-79.388790	3	Pizza Place	Sandwich Place	Dessert Shop	Sushi Restaurant	Coffee Shop	Gym	Italian Restaurant	Café
27	M5S	Downtown Toronto	University of Toronto, Harbord	43.662696	-79.400049	3	Café	Bar	Japanese Restaurant	Bookstore	Sandwich Place	Restaurant	Bakery	Yoga Studio
35	M4X	Downtown Toronto	St. James Town, Cabbagetown	43.667967	-79.367675	3	Coffee Shop	Pub	Pizza Place	Italian Restaurant	Bakery	Chinese Restaurant	Café	Restaurant

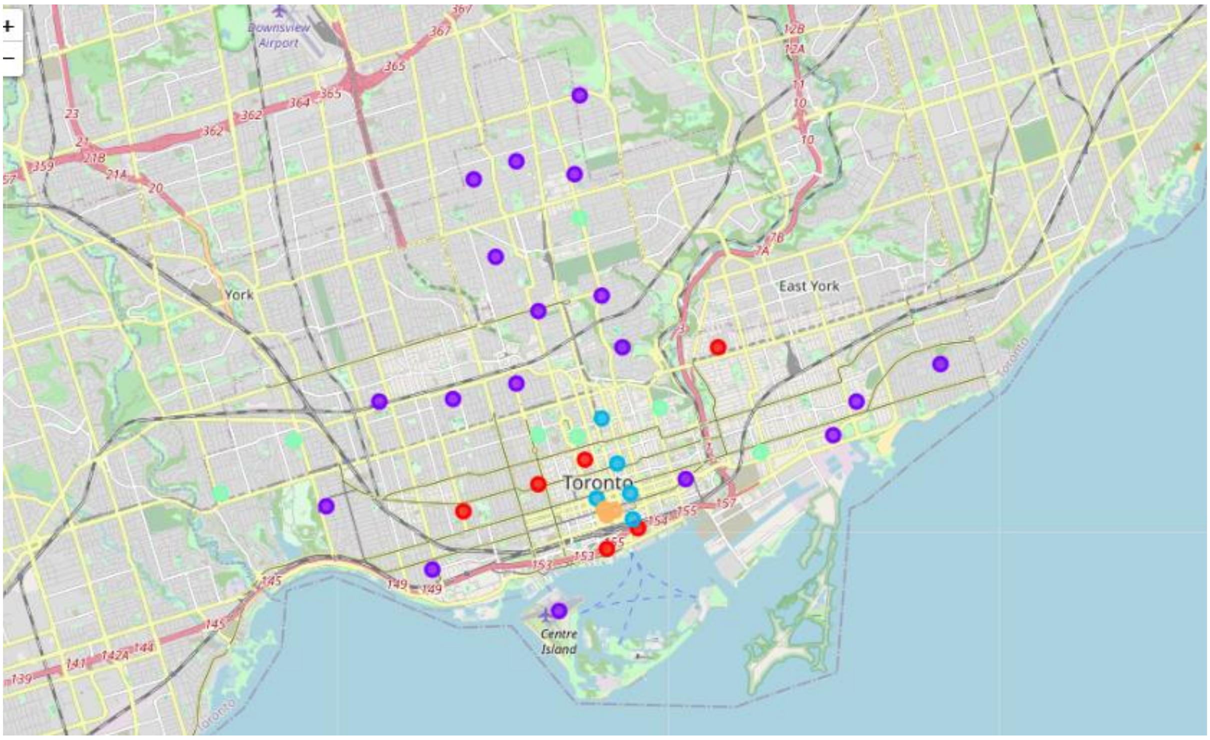
The owners can further choose from these 4 locations which will be the best according to the type of restaurant they are trying to open.

4. Data Visualisation

A map of Toronto city was generated using a great visualisation library named **Folium**. All the 39 neighbourhoods of Toronto were also marked with blue circles on the map with help of first dataset. The map looked like this:

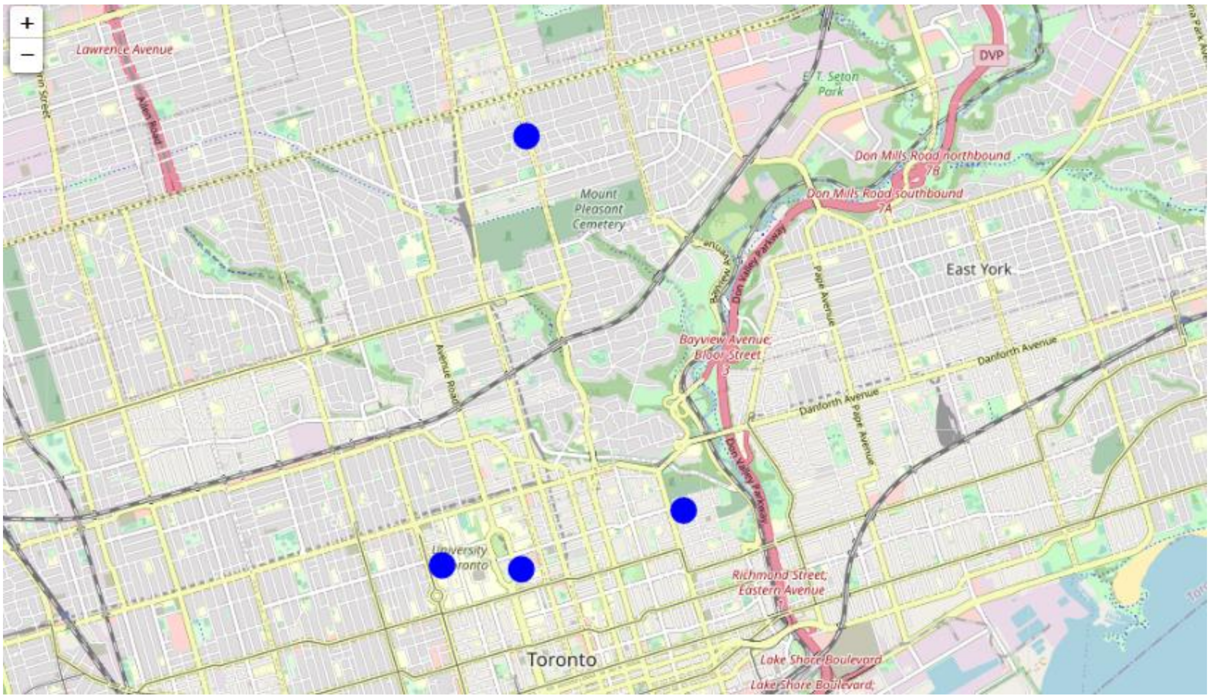


After using the clustering algorithm and creating 5 different clusters where each neighbourhood belong to one of these clusters, the new map of Toronto looked like this:



In the above map, 5 different colours, one for each cluster are used for representing each neighbourhood in Toronto.

The final 4 neighbourhoods were also presented on a map:



The 4 neighbourhoods are depicted by 4 blue dots in the above map.

5. Result and Discussion

Our analysis shows that although there is a great number of restaurants in Toronto, there are pockets of low restaurant density spread across the **Toronto** city. To identify these pockets, I used clustering algorithm and segmented our neighbourhood dataset accordingly.

I used K-means clustering algorithm for making 5 clusters each containing some neighbourhoods based on number of restaurants they have in their vicinity. Then I analysed each cluster by calculating the Restaurant/Neighbourhood ratio of each cluster. I saw that cluster 4 had the lowest ratio, which means very few restaurants are present within vicinity of each neighbourhood that belonged to that cluster. There were a total of 8 neighbourhoods belonging to cluster 4. Then upon further analysis, I found that 4 among those were not good for opening up a new restaurant. Hence, only 4 neighbourhoods were left.

According to my analysis, I got a total of 4 neighbourhoods where restaurant business will be good. There are two reasons for that. First reason is that we saw that these neighbourhoods does not contain much restaurants around their vicinity which will lower the competition in the restaurant business and give them a competitive advantage. Second reason is that, as we can see in the above map that these 4 neighbourhoods lie nearly in the centre of Toronto city which means these neighbourhoods must have high population density which means more customers and hence more profit.

The final 4 neighbourhoods that are perfect for opening a new restaurant are stored in a data frame named final which contains information about latitude, longitude and borough of these neighbourhoods.

The owners can further choose from these 4 locations which will be the best according to the type of restaurant they are trying to open.

6. Conclusion

Purpose of this project was to identify neighbourhoods in **Toronto** which have low number of restaurants in order to aid stakeholders in narrowing down the search for optimal location for a new restaurant. By calculating restaurant density distribution from Foursquare data we have first identified the most common nearby venues of each neighbourhood. Then with the help of clustering techniques and further analysis we were able to narrow down our analysis to 4 neighbourhoods which were good for opening up a new restaurant. This concludes this project of **Battle of Neighbourhoods.**