

# **BATTLE OF NEIGHBOURHOODS**

Kushal Deswal

July 12, 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](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 103 rows having 103 unique neighbourhoods of Toronto and 11 unique Boroughs. For example, below photo depicts first 5 rows of the dataset:

	Postcode	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, Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494

#### 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	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
5	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
6	Victoria Village	43.725882	-79.315572	Eglinton Ave E & Sloane Ave/Bermondsey Rd	43.726086	-79.313620	Intersection
7	Harbourfront, Regent Park	43.654260	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
8	Harbourfront, Regent Park	43.654260	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
9	Harbourfront, Regent Park	43.654260	-79.360636	Toronto Cooper Koo Family Cherry St YMCA Centre	43.653191	-79.357947	Gym / Fitness Center
10	Harbourfront, Regent Park	43.654260	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
11	Harbourfront, Regent Park	43.654260	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
12	Harbourfront, Regent Park	43.654260	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant
13	Harbourfront, Regent Park	43.654260	-79.360636	Figs Breakfast & Lunch	43.655675	-79.364503	Breakfast Spot

For example, the neighbourhood named Parkwoods in Toronto contains 3 nearby venues depicted by first 3 rows of above dataset. Information about these venues is also provided in this dataset.

## 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 5 most common venues. This dataset would eventually help in visualising the solution. First 10 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
0	Adelaide, King, Richmond	Coffee Shop	Café	American Restaurant	Bar	Steakhouse
1	Agincourt	Lounge	Sandwich Place	Breakfast Spot	Chinese Restaurant	Yoga Studio
2	Agincourt North, L'Amoreaux East, Milliken, St...	Park	Playground	Asian Restaurant	Yoga Studio	Drugstore
3	Albion Gardens, Beaumont Heights, Humbergate, ...	Grocery Store	Fast Food Restaurant	Pizza Place	Sandwich Place	Beer Store
4	Alderwood, Long Branch	Pizza Place	Coffee Shop	Skating Rink	Dance Studio	Pharmacy
5	Bathurst Manor, Downsview North, Wilson Heights	Coffee Shop	Delii / Bodega	Fast Food Restaurant	Bank	Supermarket
6	Bayview Village	Café	Japanese Restaurant	Bank	Chinese Restaurant	Yoga Studio
7	Bedford Park, Lawrence Manor East	Fast Food Restaurant	Coffee Shop	Italian Restaurant	Sandwich Place	Sushi Restaurant
8	Berczy Park	Coffee Shop	Cocktail Bar	Bakery	Cheese Shop	Café
9	Birch Cliff, Cliffside West	College Stadium	General Entertainment	Skating Rink	Café	Drugstore

### 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:

	Postcode	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
0	M3A	North York	Parkwoods	43.753259	-79.329656	0	Fast Food Restaurant	Park	Food & Drink Shop	Dumpling Restaurant	Diner
1	M4A	North York	Victoria Village	43.725882	-79.315572	0	Intersection	Coffee Shop	Hockey Arena	Portuguese Restaurant	Drugstore
2	M5A	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636	3	Coffee Shop	Pub	Bakery	Park	Theater
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763	0	Clothing Store	Furniture / Home Store	Women's Store	Coffee Shop	Fraternity House
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494	1	Coffee Shop	Sushi Restaurant	Gym	Japanese Restaurant	Park
5	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	0	Fast Food Restaurant	Print Shop	Dessert Shop	Diner	Discount Store

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=4. Hence this cluster was chosen for further analysis.

Cluster 4 consisted of total 8 neighbourhoods. Out of these, 2 had very high total number of restaurants, therefore these 2 neighbourhoods were discarded. Out of the remaining 6 neighbourhoods, 2 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:

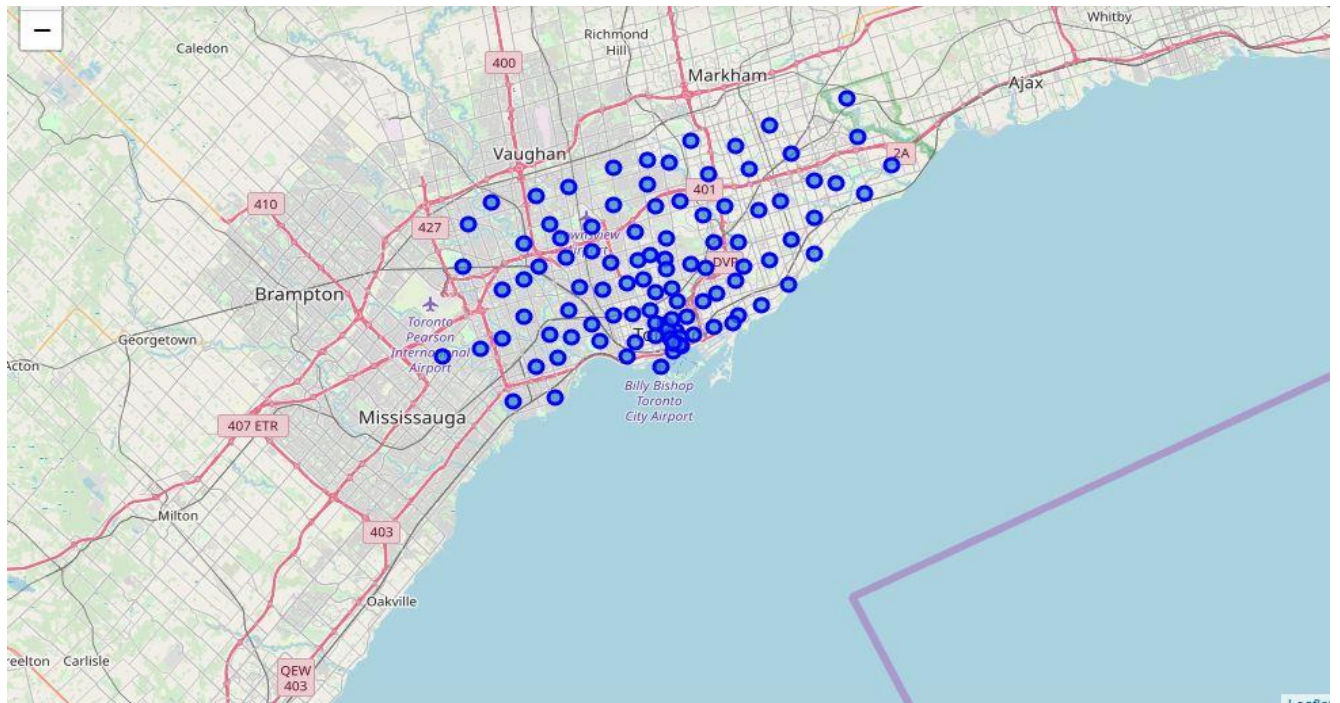
	Postcode	Borough	Neighbourhood	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
14	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	4	Café	Coffee Shop	Restaurant	Hotel	Bakery
29	M5H	Downtown Toronto	Adelaide, King, Richmond	43.650571	-79.384568	4	Coffee Shop	Café	American Restaurant	Bar	Steakhouse
41	M5K	Downtown Toronto	Design Exchange, Toronto Dominion Centre	43.647177	-79.381576	4	Coffee Shop	Hotel	Café	Restaurant	Gastropub
90	M5W	Downtown Toronto	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	4	Coffee Shop	Restaurant	Café	Hotel	Beer Bar

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.

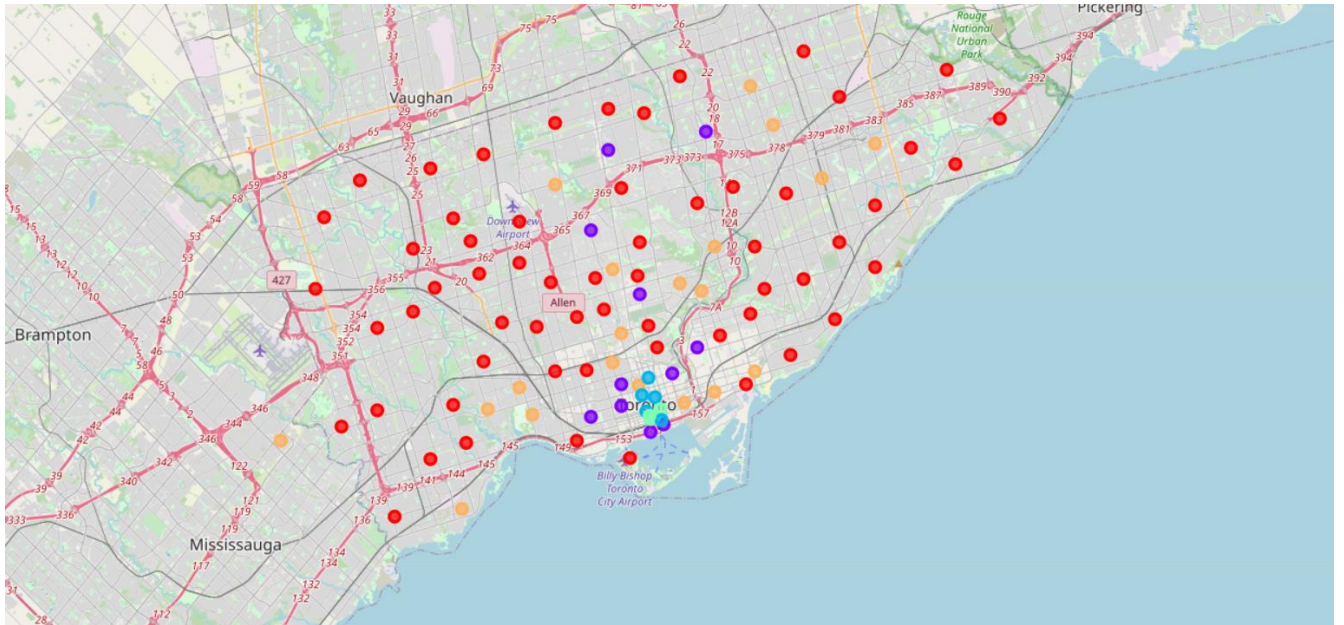


## 4. Data Visualisation

A map of Toronto city was generated using a great visualisation library named **Folium**. All the 103 neighbourhoods of Toronto were also marked with blue circles on the map with help of first dataset. The map looked like this:

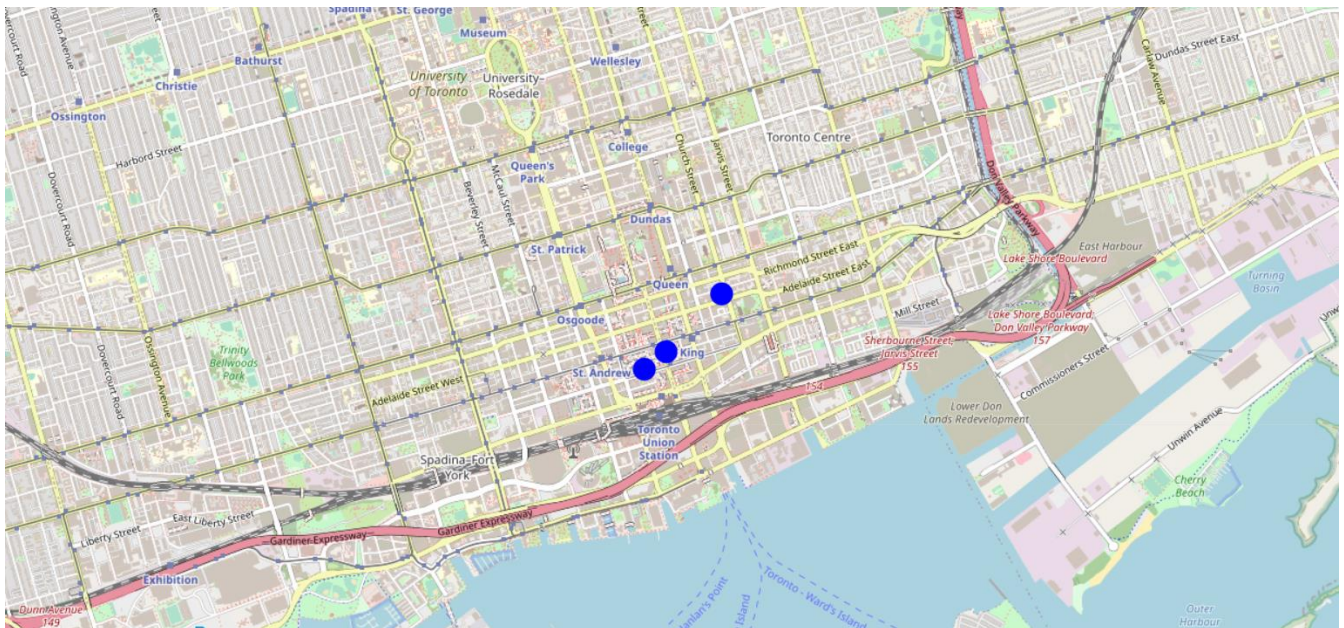


After using the clustering algorithm and creating 5 different clusters where each neighbourhood belong to one of these clusters, the new map of Toronto looked like this:



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

The final 3 neighbourhoods were also presented on a map:



The 3 neighbourhoods are depicted by 3 blue dots in the above map.

Here restaurants are most isolated among all clusters and within this, from the map we can conclude that neighbourhood with **Postal Code : M5C, Downtown Borough, St. James Town with Latitude, Longitude:{46.35,-79.37}** is the most isolated and hence the most ideal place for restaurant opening which can be most Profitable for the investors to put their funding into.

## 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 3 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 4 neighbourhoods belonging to cluster 3. Then upon further analysis, I found that 1 among those were not good for opening up a new restaurant. Hence, only 3 neighbourhoods were left.

According to my analysis, I got a total of 3 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 3 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 3 locations which will be the best according to the type of restaurant they are trying to open.

If we look closely the distance from each other in the map, we find neighbourhood with **Postal Code : M5C, Downtown Borough, St. James Town with Latitude, Longitude:{46.35,-79.37}** is most isolated of all and hence, the perfect location for restaurant opening



## 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**.