Battle of Neighbourhoods

Scarborough, Toronto

Table of Contents

Problem Statement

Identifying a suitable neighbourhood Toronto, to open an Indian Restaurant.

Background of the Problem

An Indian immigrant to Canada is our client and is planning to open an Indian restaurant. Though, the client knew the nuances of the restaurant business, he is new to Canada and must know the demography of Canada. Toronto, being the financial capital of Canada, is one of the widely preferred city for new immigrants to set up a business. But in order to choose the perfect neighbourhood in Toronto, the client has approached our Data Science team to come up with an analysis.

Objective

Data Science team has zeroed in on Scarborough, Toronto; because as per the 2016 statistics available on Wikipedia, South Asians are predominantly (around 25%) in Scarborough compared to the other Boroughs of Toronto which is less than 10%. To further narrow down the options, our team has decided to use the Battle of Neighbourhoods approach for Scarborough using K-Means Clustering algorithm and FourSquare API - to understand the top 10 venues from each neighbourhood of Scarborough.

Data

Data collection is an important part for any Data science project. We will not be getting any readymade data to work with. Hence, with the limited data that we have obtained from Wikipedia, we will be proceeding with our analysis.

Wikipedia: *List of Postal Code for Toronto*

Below is the sample of the postal codes that we will be using for this analysis. We will be scrapping the below table data from Wikipedia and will be doing the necessary data wrangling and data cleansing activity and will use the final data for our further analysis. One such example is that, we will be removing the Borough's that has the value 'Not assigned' and will work with the remaining data. From the formatted data, we will be taking the data of Scarborough alone to apply the Data science methodology.

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M5A	Downtown Toronto	Regent Park
M6A	North York	Lawrence Heights

Wikipedia: *Demography of Toronto*

Complete demography of Toronto has been provided in the above link. We have taken only the below section to select a Borough – Scarborough, upon which we will be applying the Battle of Neighbourhood analysis. We have taken Scarborough for analysis, because, as we can see from the below image that South Asians are around 25% of the population of Scarborough population. Whereas, the South Asian population is very less in other boroughs which is evident from the below image. Apart from this, we will not be using the below data for any other purpose.

Community Councils [edit]

The top visible-minority groups per **Community Council** (2016 Census) [33] are as follows:

- **Toronto & East York** (643,015): White: 64.7%, Chinese: 9.7%, South Asian: 7.1%, Black: 5.3%
- **North York** (635,260): White: 44.6%, Chinese: 13.3%, South Asian: 8.9%, Filipino: 8.0%, Black: 7.1%, West Asian: 5.1%
- **Scarborough** (602,645): White: 26.6%, South Asian: 25.4%, Chinese: 19.0%, Black: 10.8%, Filipino: 8.4%
- **Etobicoke York** (595,420): White: 55.2%, Black: 13.2%, South Asian: 10.4%, Latin American: 5.2%

Coursera: *Geospatial Data*

We will be using the Geospatial data - Latitude and Longitude details of all the postal code of Toronto that we have received through the capstone project as well. A sample of the data is provided in the below image

	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

Methodology

With the raw data, we will not be able to apply any ML algorithms or perform any visualization. In order to apply the data science methodology, we must first perform data wrangling and carry out the data analysis. Below are the step by step process of preparing the data.

Data Wrangling

a) Using beautiful soup package and lxml library, we must scrape the list of postal code of Toronto from Wikipedia.

b) Once the data is scrapped, we must format it by cleaning the 'Not assigned' values of Boroughs and also if any Neighbourhoods of the borough has 'Not assigned' value, we must update it with the borough value.

c) Also, we must make sure there is one unique row for each postal code. For this, we will be joining the Neighbourhood data for each postal code. Below is the image of the formatted data frame that we have obtained from the raw data of Wikipedia.

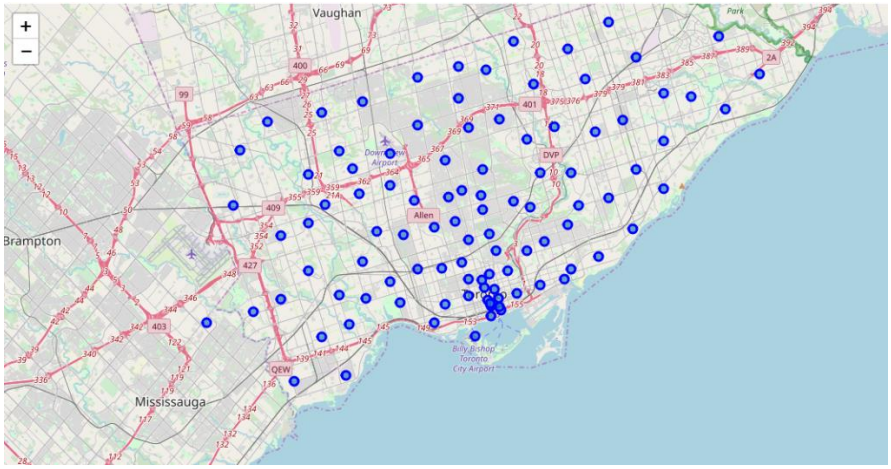
	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

d) Now, we must append the coordinates for each borough with the above data frame to proceed with our analysis. Coordinates details are obtained from Geospatial Data csv file. After appending, a sample of the final data frame looks like below,

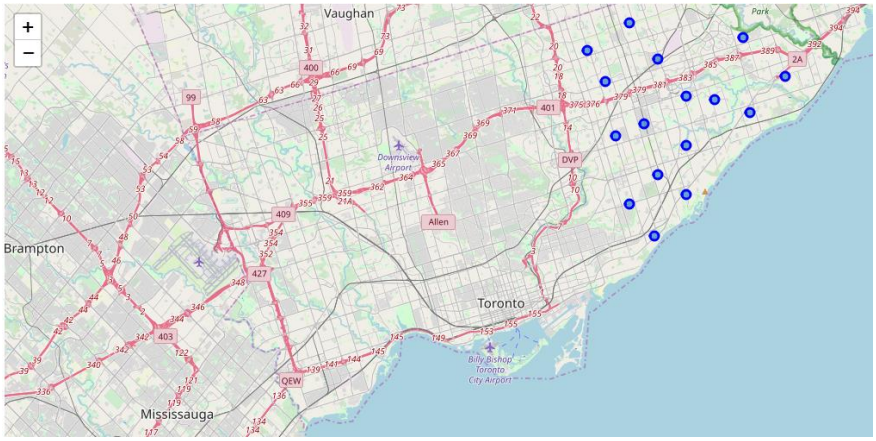
	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917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

Data Visualization

Once the data analysis is over and the data is prepared, we will be proceeding to visualization part to understand in a better way. Below is the map of Toronto with all the postal codes marked.



a) Now, for analysis, we are taking only Scarborough's neighbourhoods as the South Asian population is more here. Below is the map of the Scarborough neighbourhoods.



b) Since, we have decided the neighbourhood, we must analyse the neighbourhoods by knowing the most common venues around the each neighbourhoods. For this, we will be using the **FourSquare** API, where providing the neighbourhood coordinates along with the radius, we will be getting the Venue details for all the neighbourhoods in Scarborough. Below is the same of the data frame after the venue extraction and appending it to our neighbourhood data frame.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Guildwood, Morningside, West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	Marina Spa	43.766000	-79.191000	Spa

c) Though, we have received the necessary data, but this is not in the expected format to apply ML algorithms. We must further format it. As we are dealing with categorical data

(Venue categories), we will have to group these data based on OneHot encoding method for each neighbourhoods. After grouping, the sample of the data frame looks like below.

	Neighbourhood	Accessories Store	American Restaurant	Athletics & Sports	Auto Garage	Bakery	Bank	Bar	Breakfast Spot	Bus Line	Bus Station	Café	Caribbean Restaurant	Chinese Restaurant	Coffee Shop	College Stadium	Department Store	Disc
0	Agincourt	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.2	0.000000	0.0	0.00	0.000000	0.2	0.000000	0.00	0.0	
1	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.0	0.00	0.000000	0.0	0.333333	0.00	0.0	
2	Birch Cliff, Cliffside West	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.0	0.25	0.000000	0.0	0.000000	0.25	0.0	
3	Cedarbrae	0.0	0.0	0.142857	0.0	0.142857	0.142857	0.0	0.0	0.000000	0.0	0.00	0.142857	0.0	0.000000	0.00	0.0	
4	Clairlea, Golden Mile, Oakridge	0.0	0.0	0.000000	0.0	0.222222	0.000000	0.0	0.0	0.222222	0.0	0.00	0.000000	0.0	0.000000	0.00	0.0	

d) To understand the most occurring frequency of a particular venue of a Neighbourhood, we can generate top 5 venues for a neighbourhood.

```
----Agincourt----
                venue  freq
0              Lounge   0.2
1        Skating Rink   0.2
2      Breakfast Spot   0.2
3      Sandwich Place   0.2
4  Chinese Restaurant   0.2


----Agincourt North, L'Amoreaux East, Milliken, Steeles East----
                       venue  freq
0                Coffee Shop  0.33
1                       Park  0.33
2                 Playground  0.33
3  Latin American Restaurant  0.00
4          Light Rail Station  0.00
```

Similarly, we can get the Top 10 venues for each neighbourhoods and can for a dataframe, which will be used for applying ML algorithms.

	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	Agincourt	Skating Rink	Sandwich Place	Lounge	Breakfast Spot	Chinese Restaurant	Vietnamese Restaurant	Coffee Shop	Hakka Restaurant	Grocery Store	General Entertainment
1	Agincourt North, L'Amoreaux East, Milliken, St...	Playground	Park	Coffee Shop	Vietnamese Restaurant	Chinese Restaurant	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Fast Food Restaurant
2	Birch Cliff, Cliffside West	Skating Rink	General Entertainment	Café	College Stadium	Vietnamese Restaurant	Coffee Shop	Indian Restaurant	Hakka Restaurant	Grocery Store	Fried Chicken Joint
3	Cedarbrae	Caribbean Restaurant	Thai Restaurant	Athletics & Sports	Hakka Restaurant	Bakery	Bank	Fried Chicken Joint	College Stadium	Indian Restaurant	Grocery Store
4	Clairlea, Golden Mile, Oakridge	Bakery	Bus Line	Soccer Field	Fast Food Restaurant	Park	Metro Station	Intersection	Vietnamese Restaurant	College Stadium	Hakka Restaurant
5	Clarks Corners, Sullivan, Tam O'Shanter	Pizza Place	Chinese Restaurant	Thai Restaurant	Italian Restaurant	Fried Chicken Joint	Fast Food Restaurant	Pharmacy	Noodle House	Vietnamese Restaurant	Grocery Store
6	Cliffcrest, Cliffside, Scarborough Village West	American Restaurant	Motel	Movie Theater	Vietnamese Restaurant	Coffee Shop	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Fast Food Restaurant
7	Dorset Park, Scarborough Town Centre, Wexford ...	Indian Restaurant	Pet Store	Latin American Restaurant	Light Rail Station	Vietnamese Restaurant	Chinese Restaurant	Department Store	Hakka Restaurant	Grocery Store	General Entertainment
8	East Birchmount Park, Ionview, Kennedy Park	Discount Store	Train Station	Coffee Shop	Bus Station	Department Store	Intersection	Indian Restaurant	Hakka Restaurant	Grocery Store	General Entertainment
9	Guildwood, Morningside,		Rental Car	Electronics				Mexican	Vietnamese		

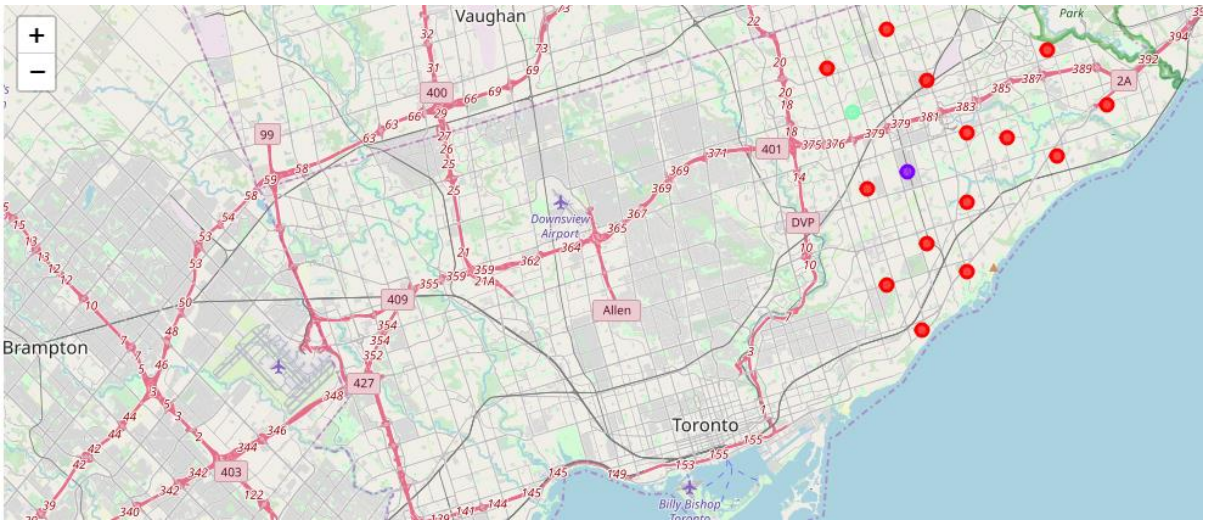
With the help of data visualization, we have further developed our dataframe by appending the venue details for each category. With this data, we will be applying ML algorithms to model and fit the data.

Machine Learning Algorithm (*K-Means Clustering*)

We will be using the K-Means clustering ML algorithm to cluster the neighbourhoods of Scarborough which will be helpful to infer the suitable neighbourhood to open a restaurant. Below is the sample of the dataframe with the Cluster Labels after applying the ML algorithm.

	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	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th M Comm Ve
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	0	Fast Food Restaurant	Vietnamese Restaurant	Italian Restaurant	Indian Restaurant	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Electro S
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	0	Bar	Vietnamese Restaurant	Coffee Shop	Indian Restaurant	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Fast F Restau
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	0	Spa	Rental Car Location	Electronics Store	Pizza Place	Breakfast Spot	Medical Center	Mexican Restaurant	Vietnamese Restaurant	C S
3	M1G	Scarborough	Woburn	43.770992	-79.216917	0	Coffee Shop	Korean Restaurant	Vietnamese Restaurant	Indian Restaurant	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Fast F Restau
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	0	Caribbean Restaurant	Thai Restaurant	Athletics & Sports	Hakka Restaurant	Bakery	Bank	Fried Chicken Joint	College Stadium	In Restau
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476	0	Jewelry Store	Playground	Vietnamese Restaurant	Chinese Restaurant	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Fast F Restau
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029	0	Discount Store	Train Station	Coffee Shop	Bus Station	Department Store	Intersection	Indian Restaurant	Hakka Restaurant	Gro S
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577	0	Bakery	Bus Line	Soccer Field	Fast Food Restaurant	Park	Metro Station	Intersection	Vietnamese Restaurant	Col Sta
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	43.716316	-79.239476	0	American Restaurant	Motel	Movie Theater	Vietnamese Restaurant	Coffee Shop	Hakka Restaurant	Grocery Store	General Entertainment	F Chi
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848	0	Skating Rink	General Entertainment	Café	College Stadium	Vietnamese Restaurant	Coffee Shop	Indian Restaurant	Hakka Restaurant	Gro S

Plotting the clustered neighbourhood in the Toronto map, we will be getting the below. Each colour represents each cluster. Here, we have used 3 Clusters as the neighbourhoods are very less.



Result

Below are the results of each cluster with the top 10 venues of the neighbourhoods.

First Cluster

As per the K-Means clustering algorithm, first cluster is the red markers in the above map and most of the neighbourhoods (around 14) fall under this cluster. From the cluster result, we can infer that only for couple of neighbourhoods (Rogue, Malvern, Highland Creek, Rouge Hill, Port Union, Woburn), the Indian Restaurant venues comes in the 4th common venue of that neighbourhoods. In all other neighbourhoods, Indian Restaurants are either less common and or doesn't fall under the top 10 venues bracket.

Top 10 venues of first cluster

```
Scarborough_merged.loc[Scarborough_merged['Cluster Labels'] == 0, Scarborough_merged.columns[[2] + list(range(5, Scarborough_merged.shape[1]))]]
```

	Neighbourhood	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	Rouge, Malvern	0	Fast Food Restaurant	Vietnamese Restaurant	Italian Restaurant	Indian Restaurant	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Electronics Store	Discount Store
1	Highland Creek, Rouge Hill, Port Union	0	Bar	Vietnamese Restaurant	Coffee Shop	Indian Restaurant	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Fast Food Restaurant	Electronics Store
2	Guildwood, Morningside, West Hill	0	Spa	Rental Car Location	Electronics Store	Pizza Place	Breakfast Spot	Medical Center	Mexican Restaurant	Vietnamese Restaurant	Coffee Shop	Grocery Store
3	Woburn	0	Coffee Shop	Korean Restaurant	Vietnamese Restaurant	Indian Restaurant	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Fast Food Restaurant	Electronics Store
4	Cedarbrae	0	Caribbean Restaurant	Thai Restaurant	Athletics & Sports	Hakka Restaurant	Bakery	Bank	Fried Chicken Joint	College Stadium	Indian Restaurant	Grocery Store
5	Scarborough Village	0	Jewelry Store	Playground	Vietnamese Restaurant	Chinese Restaurant	Hakka Restaurant	Grocery Store	General Entertainment	Fried Chicken Joint	Fast Food Restaurant	Electronics Store
6	East Birchmount Park, Ionview, Kennedy Park	0	Discount Store	Train Station	Coffee Shop	Bus Station	Department Store	Intersection	Indian Restaurant	Hakka Restaurant	Grocery Store	General Entertainment
7	Clairlea, Golden Mile,	0	Bakery	Bus Line	Soccer Field	Fast Food	Park	Metro Station	Intersection	Vietnamese	College	Hakka

Second Cluster

Second cluster is the purple marker in the above map and only one of the neighbourhood fall under this cluster. From the cluster result, we can infer that the Indian Restaurant venues comes as the 1st common venue of that neighbourhood. This shows that there are lots of Indians residing in this neighbourhood

Top 10 venues of second cluster

```
Scarborough_merged.loc[Scarborough_merged['Cluster Labels'] == 1, Scarborough_merged.columns[[2] + list(range(5, Scarborough_merged.shape[1]))]]
```

	Neighbourhood	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
10	Dorset Park, Scarborough Town Centre, Wexford ...	1	Indian Restaurant	Pet Store	Latin American Restaurant	Light Rail Station	Vietnamese Restaurant	Chinese Restaurant	Department Store	Hakka Restaurant	Grocery Store	General Entertainment

Third Cluster

Third cluster is the cyan marker in the above map and only one of the neighbourhood fall under this cluster. From the cluster result, we can infer that the Indian Restaurant venue doesn't come in the top 10 most common venue of that neighbourhood. But from other venues of this neighbourhood, we can see that Asians are residing here.

Top 10 venues of third cluster

```
Scarborough_merged.loc[Scarborough_merged['Cluster Labels'] == 2, Scarborough_merged.columns[[2] + list(range(5, Scarborough_merged.shape[1]))]]
```

	Neighbourhood	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
13	Clarks Corners, Sullivan, Tam O'Shanter	2	Pizza Place	Chinese Restaurant	Thai Restaurant	Italian Restaurant	Fried Chicken Joint	Fast Food Restaurant	Pharmacy	Noodle House	Vietnamese Restaurant	Grocery Store

Discussion

We can observe from the clusters that Indian Restaurant tops the most common venue in 2nd Cluster, thus showing the higher probability of Indians residing in the neighbourhood - **Dorset Park, Scarborough Town Centre, Wexford.** Since, the more Indians are residing in this neighbourhood, we can recommend this neighbourhood to our client for opening an Indian restaurant. Having said that, there is a higher chance that there will be a stiff competition for our client as the Indian restaurants are already topping the chart in the top 10 most common venues.

Similarly, from the 1st cluster couple of neighbourhoods can be recommended to the client, since we can infer that there will be a considerable number of Indian population residing in these neighbourhoods. Hence, the Indian Restaurants are 4th most common venue in these neighbourhoods.

Alternatively, 3rd cluster contains most of the Asian type restaurants – Chinese, Thai, Vietnamese, confirming heavy Asian contingent residing here. Since, Indian restaurant didn't top the most common venue, the client will be less successful here. Or, he can open his restaurant by thinking out of box in this neighbourhood.

Conclusion

With this limited data that we have collected, we were able to recommend a neighbourhood in Scarborough, Toronto to our client. Final decision has to be taken by our client based on the neighbourhood that we have recommended and their strategy.

This, Battle of Neighbourhoods capstone project can be a prototype for many larger projects, especially dealing with bigger cities and choosing between cities for any purpose. When the project gets larger, we will have to collect a lot more data and process them, analyse them and build a model that helps our client.