Battle of Neighborhoods

Gnyana Teja Samudrala

Problem description

The business activity within any local market tends to be unevenly distributed, some neighborhoods simply attract more activity than others. Economic development within a neighborhood also tends to have a direct impact on several measures of well-being among its residents.

So, for this reason a fitness expert who is trying to open an own Yoga studio in Toronto, Canada wants to do analysis of the neighborhood in deciding the place to establish the business.

Data description

The list of neighborhoods in Toronto are fetched from the Wikipedia page. Also, the latitude and longitude of these locations are added by using the geocoder package. By using the foursquare API, we can get a list of various venues that are present in each of the neighborhood. The steps to obtain data are as follows:

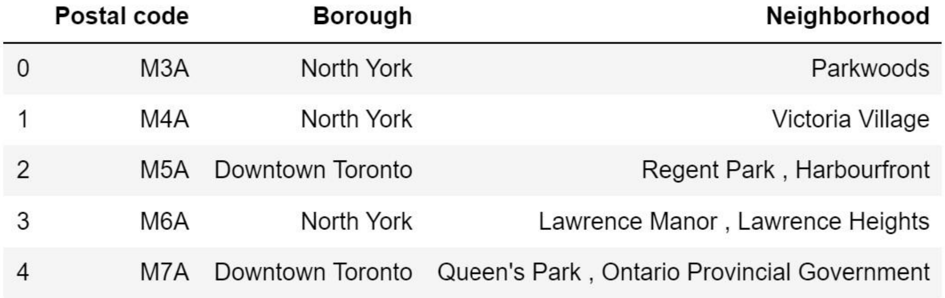
- We obtain top 100 venues of each neighborhood in 1000m radius.
  - Using this analyze most common top 10 places of each neighborhood.
  - Then cluster the neighborhoods.
  - From this clustering we can decide on a subset of neighborhoods for consideration.
- Once we have the first subset of neighborhoods, then we fetch the details of venues which offer physical fitness such as gyms, yoga studios, spas, pools etc.
  - From this data we can analyze and cluster each neighborhood with what type of venues they have and give out a well data backed judgement of the locality to open a yoga studio.

Methodology

Once the data is obtained, from different sources as stated above such as Wikipedia, foursquare API, and geocoder package, we proceed to the next step of analysis. The steps followed are as follows:

Scraping neighborhood data

This is the first step, in which we obtain the neighborhood data of Toronto from the Wikipedia page. Once we have the data, the not assigned borough rows are removed from the data frame. So, the imported data looks like shown below in fig.1



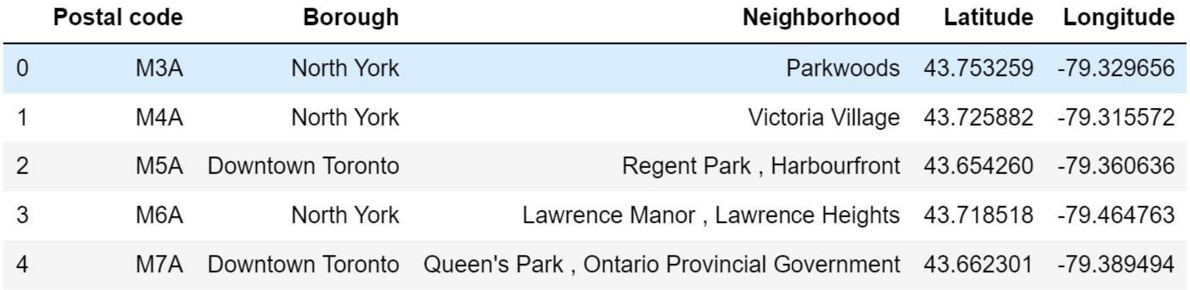
	Postal code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park , Harbourfront
3	M6A	North York	Lawrence Manor , Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park , Ontario Provincial Government

Fig.1: Showing the neighborhood data frame

This data frame has 103 entries, which means we have that many number of options to choose from.

Merging Latitude and Longitude

Now using the geocoder package, we get the latitude and longitude of each neighborhood and put it in the data frame. This data frame is shown in fig.2

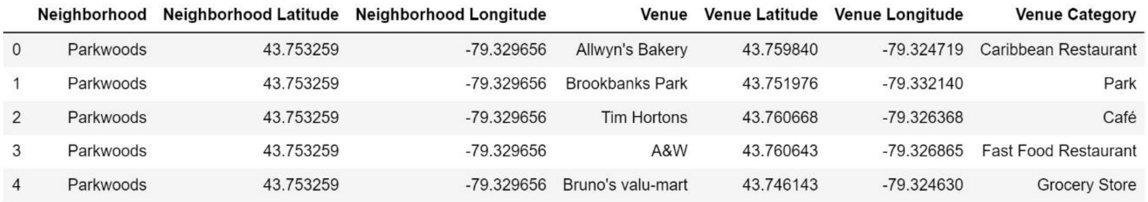


	Postal code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park , Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor , Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park , Ontario Provincial Government	43.662301	-79.389494

Fig.2: Showing the neighborhood data frame with Latitude and Longitude merged

Fetching the top 100 venues

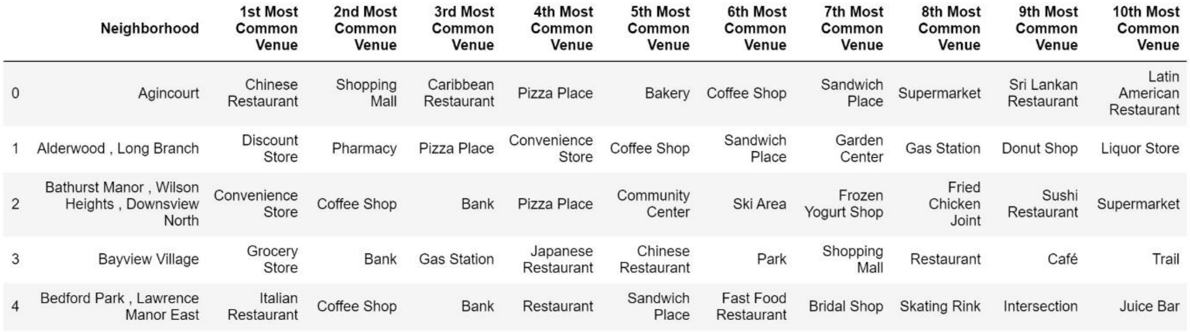
Using the foursquare API, we collect top 100 venues around 1000m radius of each neighborhood. These venues belong to various categories, in this case we got a total number of 327 unique categories of venues. The data frame is shown in fig.3



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
2	Parkwoods	43.753259	-79.329656	Tim Hortons	43.760668	-79.326368	Café
3	Parkwoods	43.753259	-79.329656	A&W	43.760643	-79.326865	Fast Food Restaurant
4	Parkwoods	43.753259	-79.329656	Bruno's valu-mart	43.746143	-79.324630	Grocery Store

Fig.3: The data frame with 100 top venues from 1000m radius of each neighborhood

From this we analyze the occurrence of each category type and sort them out as 10 most common places for each neighborhood, to get an idea of each type of neighborhood. We form a data frame as shown in fig.4



	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	Agincourt	Chinese Restaurant	Shopping Mall	Caribbean Restaurant	Pizza Place	Bakery	Coffee Shop	Sandwich Place	Supermarket	Sri Lankan Restaurant	Latin American Restaurant
1	Alderwood , Long Branch	Discount Store	Pharmacy	Pizza Place	Convenience Store	Coffee Shop	Sandwich Place	Garden Center	Gas Station	Donut Shop	Liquor Store
2	Bathurst Manor , Wilson Heights , Downsview North	Convenience Store	Coffee Shop	Bank	Pizza Place	Community Center	Ski Area	Frozen Yogurt Shop	Fried Chicken Joint	Sushi Restaurant	Supermarket
3	Bayview Village	Grocery Store	Bank	Gas Station	Japanese Restaurant	Chinese Restaurant	Park	Shopping Mall	Restaurant	Café	Trail
4	Bedford Park , Lawrence Manor East	Italian Restaurant	Coffee Shop	Bank	Restaurant	Sandwich Place	Fast Food Restaurant	Bridal Shop	Skating Rink	Intersection	Juice Bar

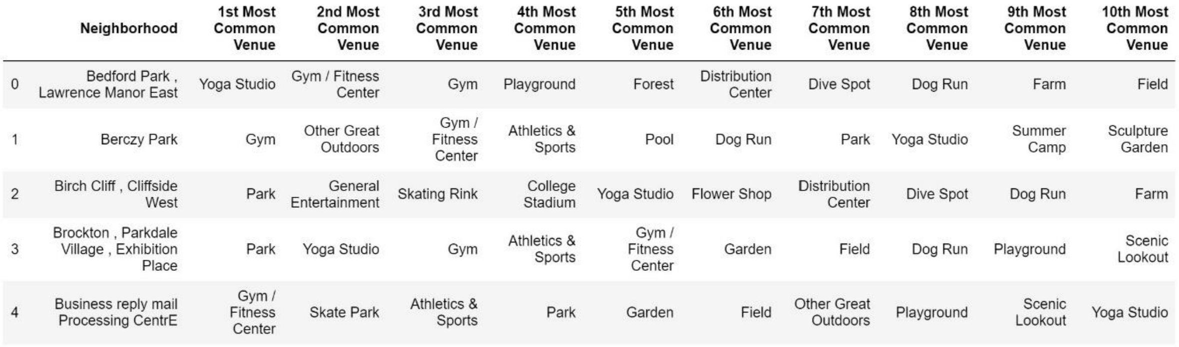
Fig.4: The data frame with 10 most common venue categories for each neighborhood

Clustering

The K-Means clustering algorithm is used to cluster the neighborhoods, depending on the 10 most common venue categories. We cluster them into 5 groups. By looking at each of the cluster we can get an idea of what cluster to proceed to for establishing a yoga studio. Once a cluster is decided we can move ahead with further refining the cluster to decide upon a specific neighborhood. The results of the clustering are discussed in the results section.

Fetching venues of Outdoor and Recreation category

Once the cluster is decided from a above method, we fetch the venues belonging to the outdoor and recreation category from the foursquare API. Once we have the venues, we repeat the same process of analyzing by getting the 10 most common venues of each neighborhood and clustering. This time clustering is done into 3 groups. The results are discussed in detail in the results section below.



	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	Bedford Park , Lawrence Manor East	Yoga Studio	Gym / Fitness Center	Gym	Playground	Forest	Distribution Center	Dive Spot	Dog Run	Farm	Field
1	Berczy Park	Gym	Other Great Outdoors	Gym / Fitness Center	Athletics & Sports	Pool	Dog Run	Park	Yoga Studio	Summer Camp	Sculpture Garden
2	Birch Cliff , Cliffside West	Park	General Entertainment	Skating Rink	College Stadium	Yoga Studio	Flower Shop	Distribution Center	Dive Spot	Dog Run	Farm
3	Brockton , Parkdale Village , Exhibition Place	Park	Yoga Studio	Gym	Athletics & Sports	Gym / Fitness Center	Garden	Field	Dog Run	Playground	Scenic Lookout
4	Business reply mail Processing CentrE	Gym / Fitness Center	Skate Park	Athletics & Sports	Park	Garden	Field	Other Great Outdoors	Playground	Scenic Lookout	Yoga Studio

Fig.5: The data frame with 10 most common venue categories in outdoor and recreation category for each neighborhood

Results

From the first clustering of top 100 venues of all categories, we get to know the neighborhoods based on the frequency of occurrence of each category in top 3 most common places. From the bar graph below in fig.6 corresponding to each cluster we can observe cluster 1 would be the idea spot to establish a yoga studio as gym is the third most common category in this cluster and it occurred more frequently than any other cluster. All the other cluster are ideal places where we can have food joints, shopping malls, and tourism spots, which are not a place we want to establish a yoga studio.

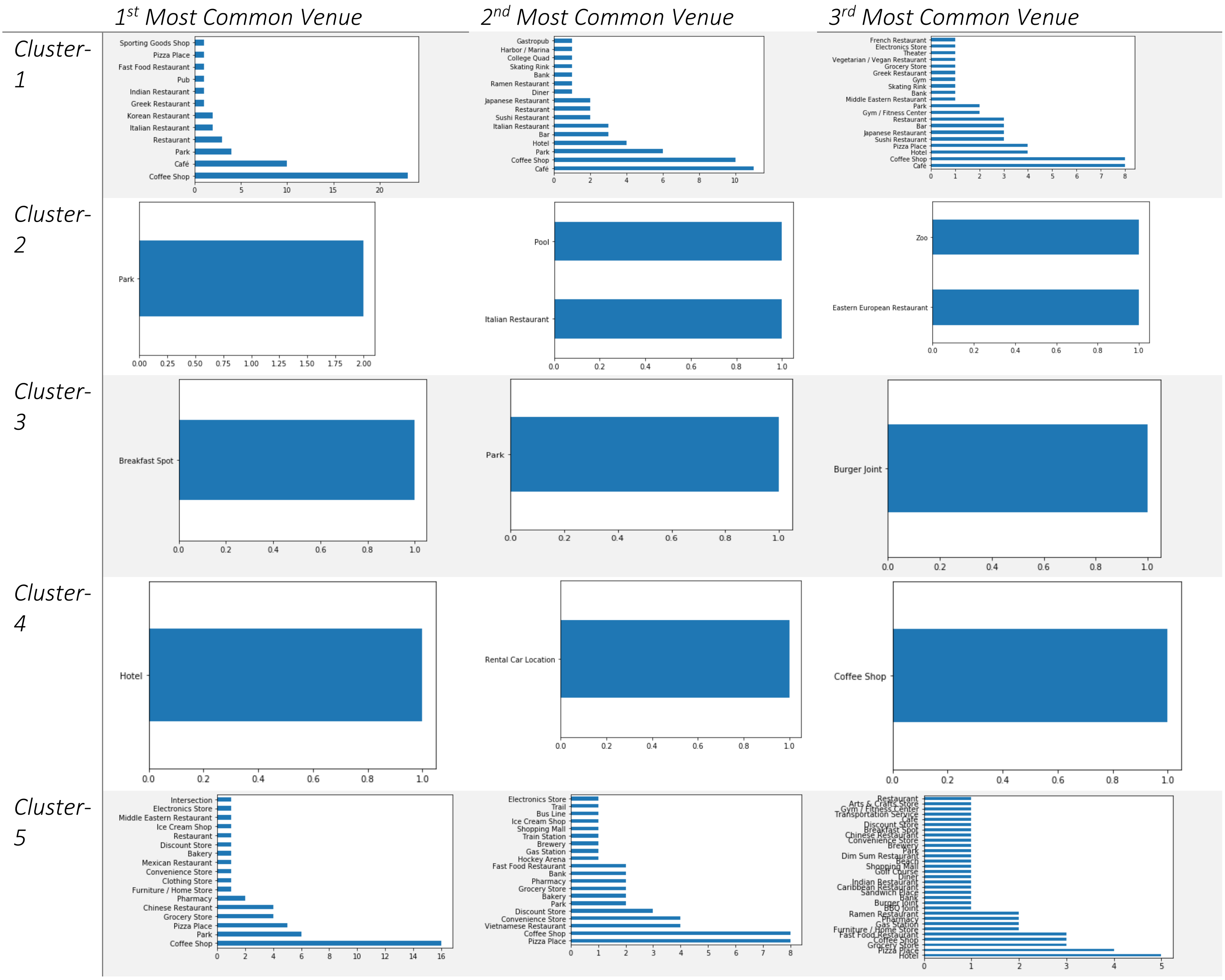


Fig.6: The bar graph of frequency of occurrence of a category in each cluster column

Further analysis on Cluster-1

So, in cluster 1 we have a total of 50 neighborhood options to decide. From each of the neighborhood we get the list of venues belonging to the outdoor and recreation category. We repeat a similar kind of most common venue categories among them and cluster them into 3 groups. The clustering of this yields the following result shown in the fig.7. We can remove cluster 2 & 3 from the option as we have occurrences of a yoga studio in all 3 most common venues and on the other hand cluster 1 frequency for yoga studio is low. So cluster 1 is where we need to look closely for a potential neighborhood.

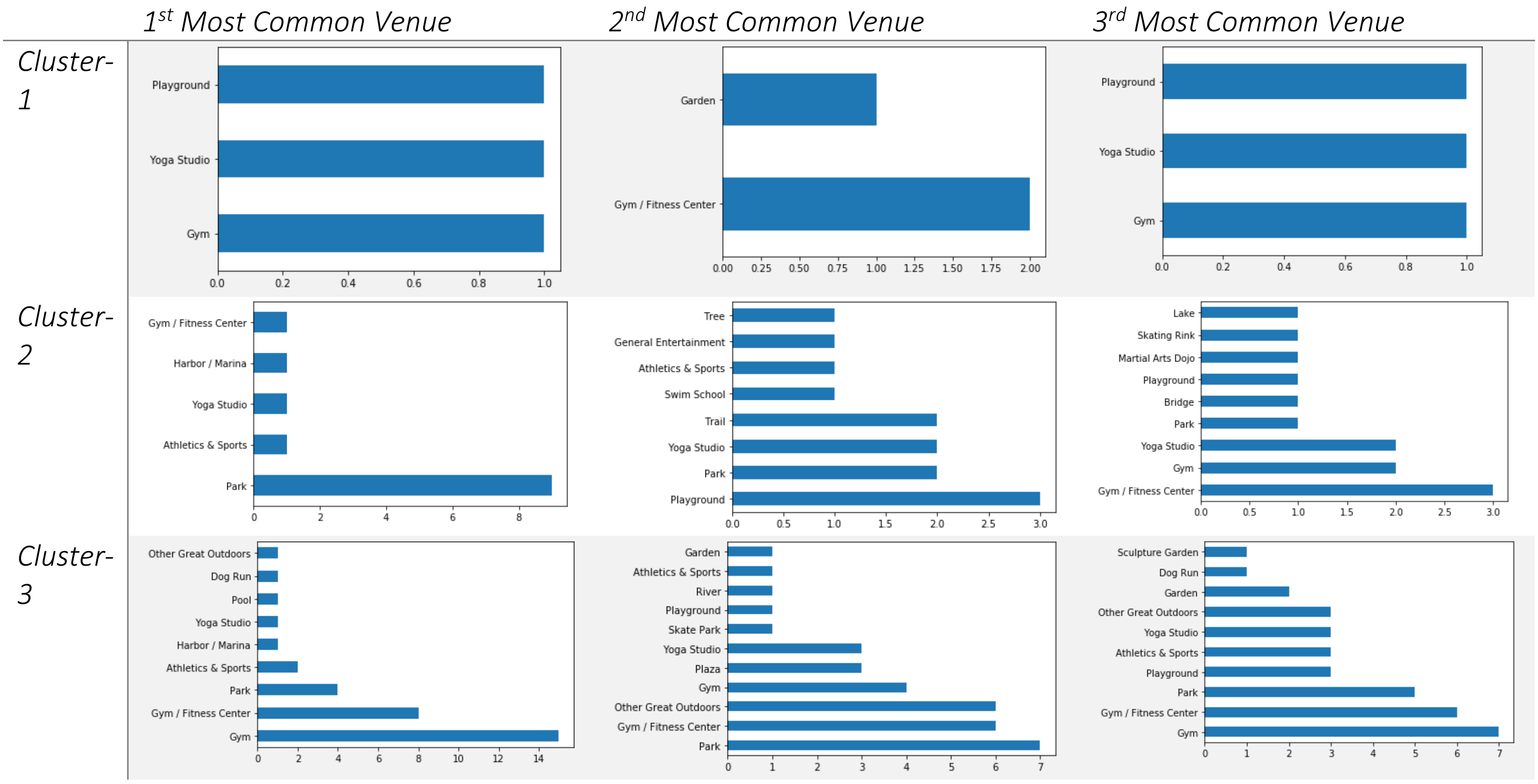
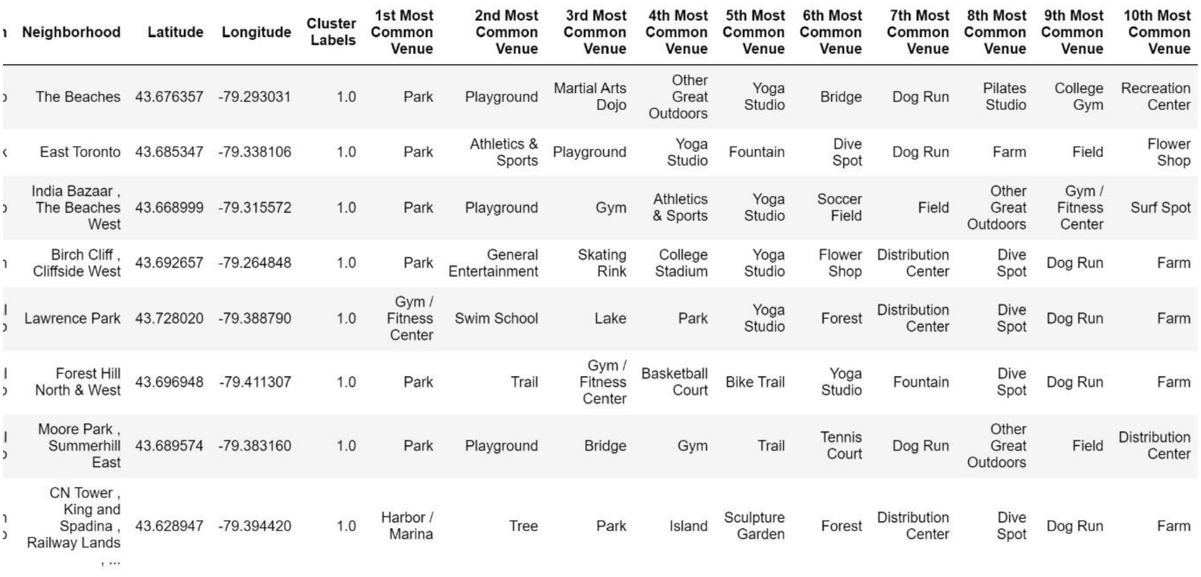


Fig.7: The bar graph of frequency of occurrence of a category in each cluster column

The cluster 1 data has 13 neighborhoods to choose from, but by removing the neighborhoods with yoga studio among 3 most common venues we are down to eight venues to choose from. The data frame with 8 neighborhoods as shown in fig.8



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
The Beaches	43.676357	-79.293031	1.0	Park	Playground	Martial Arts Dojo	Other Great Outdoors	Yoga Studio	Bridge	Dog Run	Pilates Studio	College Gym	Recreation Center
East Toronto	43.685347	-79.338106	1.0	Park	Athletics & Sports	Playground	Yoga Studio	Fountain	Dive Spot	Dog Run	Farm	Field	Flower Shop
India Bazaar , The Beaches West	43.668999	-79.315572	1.0	Park	Playground	Gym	Athletics & Sports	Yoga Studio	Soccer Field	Field	Other Great Outdoors	Gym / Fitness Center	Surf Spot
Birch Cliff , Cliffside West	43.692657	-79.264848	1.0	Park	General Entertainment	Skating Rink	College Stadium	Yoga Studio	Flower Shop	Distribution Center	Dive Spot	Dog Run	Farm
Lawrence Park	43.728020	-79.388790	1.0	Gym / Fitness Center	Swim School	Lake	Park	Yoga Studio	Forest	Distribution Center	Dive Spot	Dog Run	Farm
Forest Hill North & West	43.696948	-79.411307	1.0	Park	Trail	Gym / Fitness Center	Basketball Court	Bike Trail	Yoga Studio	Fountain	Dive Spot	Dog Run	Farm
Moore Park , Summerhill East	43.689574	-79.383160	1.0	Park	Playground	Bridge	Gym	Trail	Tennis Court	Dog Run	Other Great Outdoors	Field	Distribution Center
CN Tower , King and Spadina , Railway Lands , ...	43.628947	-79.394420	1.0	Harbor / Marina	Tree	Park	Island	Sculpture Garden	Forest	Distribution Center	Dive Spot	Dog Run	Farm

Fig.8: The 8 potential neighborhoods for opening a yoga studio

Conclusion

Among the 8 potential neighborhoods presented in fig.8, the first 6 neighborhoods have yoga studio among 10 most common venues. But the last two does not have a yoga studio among 10 most common venues, so these two are a better candidates for establishing a new yoga studio. But among these two going for Moore Park, Summerhill East neighborhood is a good option as it has Gym, tennis court and other outdoor physical activities, and opening a yoga studio there might be a useful as people in that neighborhood give a importance to physical fitness. But on the other hand, CN Tower has other kinds of activities in 10 most common venues which are sculpture garden, farm and forest which does not show their interest to a yoga studio. Hence Moore Park, Summerhill East the red marker in the map shown in fig.9 is the best potential neighborhood for opening a yoga studio.

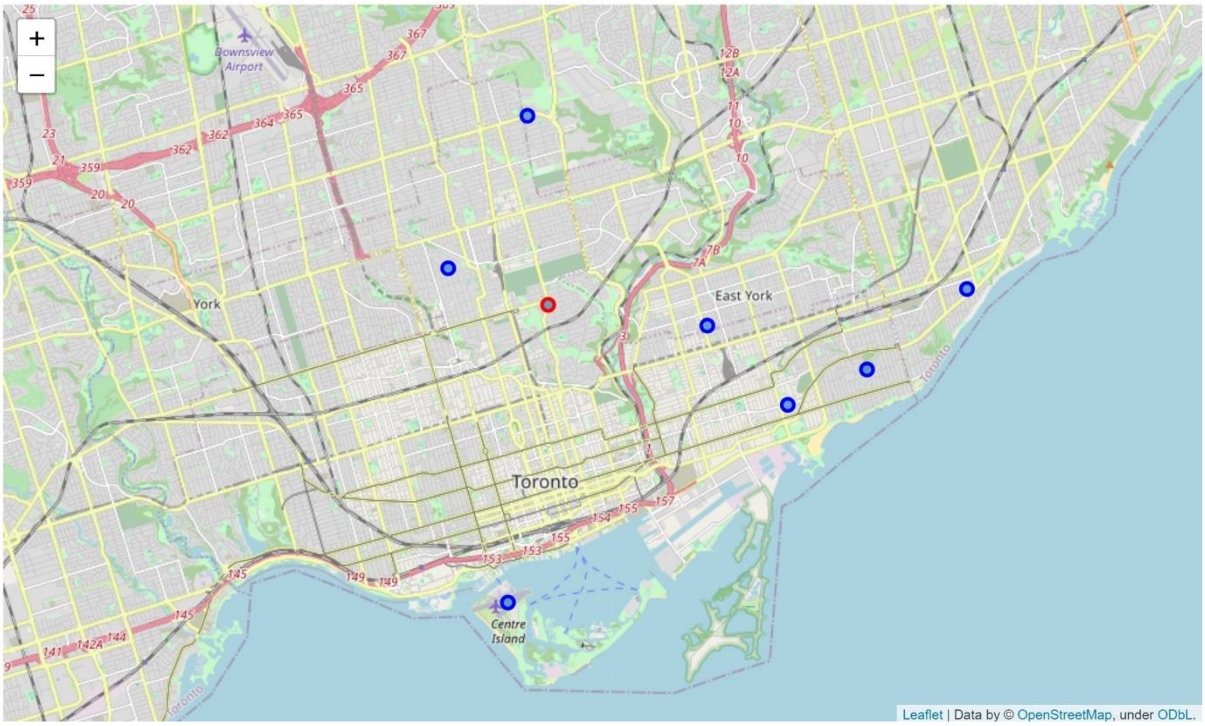


Fig.9: Final marker in red showing the neighborhood decided