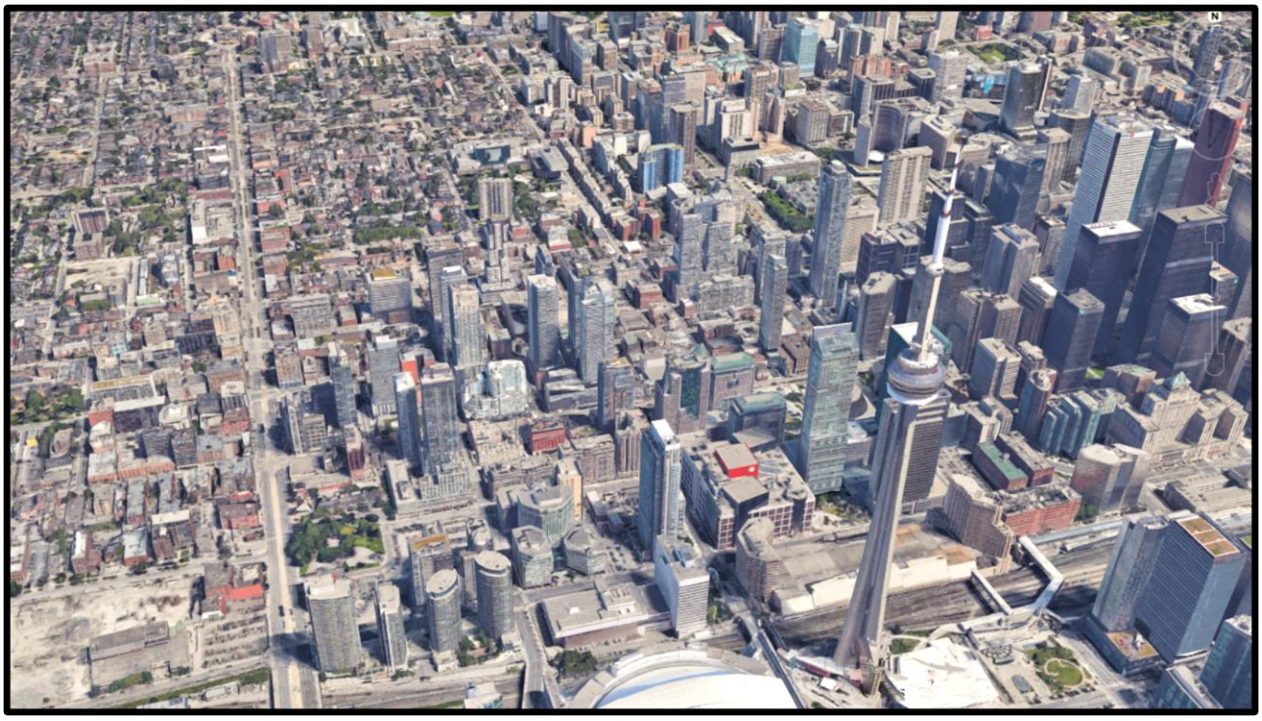
Clustering of Toronto Neighborhoods for better business opportunities: Opening a Yoga Studio



1. Introduction:

Toronto, the provincial capital of Ontario is the most populous city in Canada with a population of over 2.7 million. It is a dynamic metropolis which is a fascinating mix of neighborhoods ranging from the busy commercial business districts of Downtown to the leafy suburbs of Old Toronto. The neighborhoods of Toronto are a myriad mix of residential areas, public places, healthcare facilities, educational institutions and business enterprises of various scales. This provides each neighborhood with its distinct character. However, it is also possible to find similarities among neighborhoods and categorize them into groups, depending upon their profiles. Such categorizations in the past would be very time-consuming and wrought with subjective errors and lack of reproducibility. However, with access to a vast array of resources which allows the collection of vast volumes of relevant data and the use of smart machine learning algorithms like clustering, it possible to categorize the neighborhoods into clusters. Such an endeavor would be extremely useful to a wide range of stakeholders ranging from policy makers trying to decide upon allocation of public resources, a newbie who has just moved into Toronto, or an entrepreneur looking to start a new business. This project is an attempt to create an algorithm that will help an entrepreneur to decide upon a neighborhood to begin a business enterprise.

1.1 The problem, the stakeholders and the audience

It is often said that the three key ingredients for a successful business enterprise are location, location and location. Before starting a business, any entrepreneur would spend a considerable amount of time and energy to decide where to "set up shop". So, the problem which this project is going to address is the decision of finding an ideal location for a business and the stakeholders or audience in this would be the entrepreneurs.

By clustering the neighborhoods of a city like Toronto spread over an area of more than 630 square kilometer, it should be possible to assist this process by narrowing down the areas which the entrepreneur wants to explore.

Yoga studios have become extremely popular over the last decade and most major cities in the developed world have witnessed a spurt, mainly driven by the human desire to connect to spirituality and good health in an increasingly stressful environment. Anyone who wishes to start a Yoga studio would be interested in neighborhoods where Yoga studios are relatively less as compared to others.

Thus, although the project would interest a wide range of business entrepreneurs, I shall look at one kind of business enterprise i.e. Yoga studios to help them in deciding which would be the ideal neighborhoods to open such an enterprise.

So, anyone who wishes to open a Yoga studio would be the most important part of the profile of an audience or a stakeholder in this project.

The aim of this project is to provide someone who wishes to open a Yoga studio with a list of locations among the neighborhoods of Toronto to start a Yoga studio

2a Data which has been used

For this project I shall be using the following data

1) The page on Wikipedia which provides a list of postal codes of Toronto
   https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2) Toronto geospatial data to get the corresponding latitude and longitude
   http://cocl.us/Geospatial_data
3) The data with the venues corresponding to the neighborhoods from Foursquare API

2b. The Python Libraries imported

Pandas
Numpy
BeautifulSoup for webscraping of html and xml documents
Requests for http requests on Python
Matplotlib for plotting
Folium for rendering maps
Kmeans from sklearn.cluster for kmeans clustering

3. Methodology

The methodology comprises of five major steps

1) Creating a dataframe comprising of the neighborhoods and boroughs of the city of Toronto along with their corresponding latitude and longitude.

2) Creating the neighborhood map

3) Getting the venues of the neighborhood

4) Clustering the neighborhoods according to the venues.

5) Refining the clusters further for ideal locations for Yoga studios.

3.1 Creating a dataframe comprising of the neighborhoods and boroughs of the city of Toronto along with their corresponding latitude and longitude

i) The following libraries were imported

a) Pandas b) BeautifulSoup and c) requests

ii). The "url" https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M from the Wikipedia page on Toronto postal code was scraped using "BeautifulSoup" and "Requests" to create a pandas dataframe

iii) The data was cleaned wherein the boroughs with unassigned neighborhoods were dropped and neighborhoods which had no names assigned to them were assigned the names of the corresponding boroughs.

iv) This led to the creation of the following dataframe.

	Postalcode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

v) A second dataframe was created from the following url http://cocl.us/Geospatial_data which comprised of the latitude and longitude of the postal shown below.

	Postalcode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

vi) The above to dataframes were merged to create the following dataframe

	Postalcode	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

3.2 creating the neighborhood map.

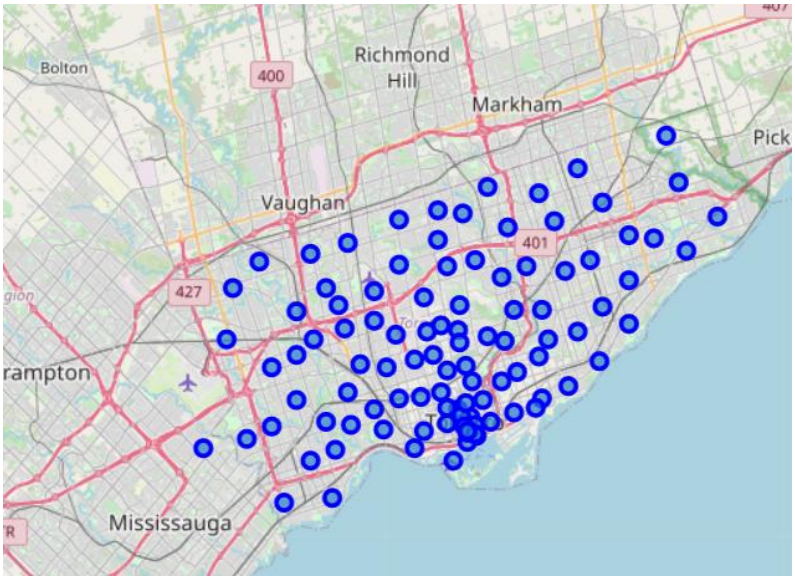
i) The following libraries were imported

a) Nominatim from geopy.geocoders to get city, state and country of a particular latitude and longitude.

b)matplotlib for plotting

c) folium for visualization of data created by Python on a map

ii) The above was used to create the neighborhood map of Toronto. See figure below

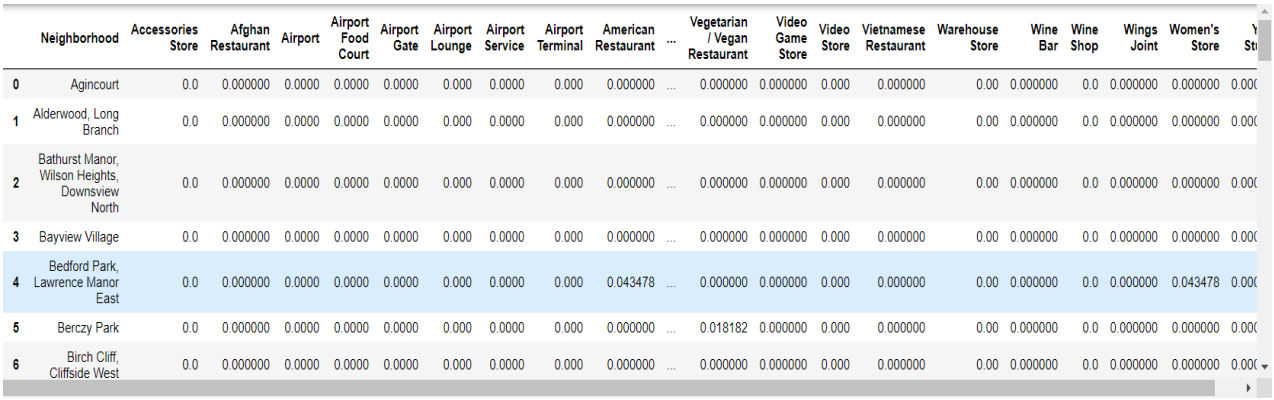


3.3 Getting the venues in the neighborhood

i) The Foursquare API credentials were defined.

ii) A function was created to get the nearby venues

iii) A dataframe was created using onehot encoding for the venues to get the following dataframe.



	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Wine Bar	Wine Shop	Wings Joint	Women's Store	Y St
0	Agincourt	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000
1	Alderwood, Long Branch	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000
3	Bayview Village	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000
4	Bedford Park, Lawrence Manor East	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.043478	...	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.0	0.000000	0.043478	0.000
5	Berczy Park	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	...	0.018182	0.000000	0.000	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000
6	Birch Cliff, Cliffside West	0.0	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000

iv) The frequency top five venues in each neighborhood and then the top 10 most common venues in each neighborhood were found and a pandas dataframe was created as follows

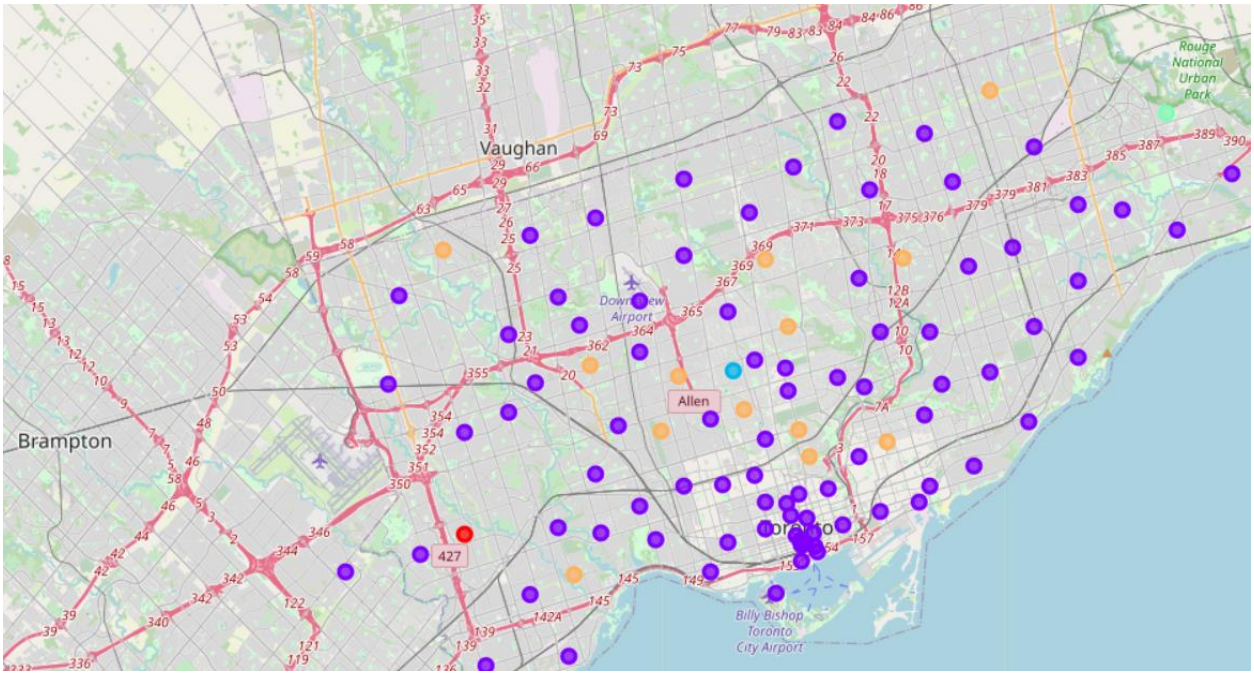
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Latin American Restaurant	Breakfast Spot	Lounge	Chinese Restaurant	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant
1	Alderwood, Long Branch	Pizza Place	Pharmacy	Coffee Shop	Sandwich Place	Athletics & Sports	Pub	Pool	Gym	Dumpling Restaurant	Drugstore
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Fried Chicken Joint	Ice Cream Shop	Shopping Mall	Sandwich Place	Diner	Middle Eastern Restaurant	Supermarket	Restaurant
3	Bayview Village	Café	Japanese Restaurant	Bank	Chinese Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Yoga Studio
4	Bedford Park, Lawrence Manor East	Coffee Shop	Restaurant	Sandwich Place	Italian Restaurant	Greek Restaurant	Thai Restaurant	Grocery Store	Pharmacy	Pizza Place	Pub

3.4: Clustering the neighborhood according to the frequency of venues

The KMeans clustering algorithm was imported to cluster the neighborhoods

	Postalcode	Borough	Neighborhood	Latitude	Longitude	Clusterlabels	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 M Com Ve
0	M3A	North York	Parkwoods	43.753259	-79.329656	2.0	Park	Construction & Landscaping	Food & Drink Shop	Yoga Studio	Doner Restaurant	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog
	click to unscroll output; double click to hide		bria Village	43.725882	-79.315572	1.0	Intersection	Pizza Place	French Restaurant	Coffee Shop	Portuguese Restaurant	Hockey Arena	Discount Store	Deli / Bodega	Department Store	Dessert S
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	1.0	Coffee Shop	Park	Pub	Bakery	Theater	Breakfast Spot	Café	Performing Arts Venue	Beer Store	
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	1.0	Clothing Store	Accessories Store	Furniture / Home Store	Event Space	Boutique	Vietnamese Restaurant	Coffee Shop	Gift Shop	Women's Store	Miscellan S
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	1.0	Coffee Shop	Sushi Restaurant	Park	Bar	Beer Bar	Smoothie Shop	Sandwich Place	Burrito Place	Café	Col Audito
			Islington Avenue													

A map of the resulting clusters was created on folium



3.5: Refining clusters further for ideal location for Yoga studios

The resulting dataframes were analyzed to find neighborhood where Yoga Studios did not figure amongst the top 10 most common venues. This was further refined to neighborhoods which feature parks amongst the top 10 venues on the assumption that such parks would be more frequent in areas where people might like to stay more connected to nature and might be more likely to subscribe to a Yoga studio

4. Results:

There were 5 clusters as follows

Cluster name	Number of neighborhoods
Cluster 1	1
Cluster 2	83
Cluster 3	1
Cluster 4	0
Cluster 5	13
TOTAL	98

Clusters 2 and Cluster 5 were the major clusters although cluster 2 was much larger than cluster 5.

Cluster 2

```
In [40]: cluster2 = toronto_merged.loc[toronto_merged['Clusterlabels'] == 1]
         cluster2
```

M1G	Scarborough	Woburn	43.770992	-79.216917	1.0	1	Coffee Shop	Soccer Field	Korean Restaurant	Dumpling Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Electronics Store
M4G	East York	Leaside	43.709060	-79.363452	1.0	1	Coffee Shop	Bank	Burger Joint	Furniture / Home Store	Sporting Goods Shop	Liquor Store	Sandwich Place	Supermarket	Fish & Chips Shop	Sports Bar
M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	1.0	1	Coffee Shop	Café	Italian Restaurant	Sandwich Place	Burger Joint	Japanese Restaurant	Department Store	Salad Place	Bubble Tea Shop	Yoga Studio
M6G	Downtown Toronto	Christie	43.669542	-79.422564	1.0	1	Grocery Store	Café	Park	Restaurant	Diner	Baby Store	Nightclub	Italian Restaurant	Athletics & Sports	Coffee Shop
M1H	Scarborough	Cedarbrae	43.773136	-79.239476	1.0	1	Hakka Restaurant	Thai Restaurant	Fried Chicken Joint	Bank	Bakery	Athletics & Sports	Gas Station	Caribbean Restaurant	Cuban Restaurant	Cupcake Shop
M2H	North York	Hillcrest Village	43.803762	-79.363452	1.0	1	Pool	Athletics & Sports	Dog Run	Golf Course	Mediterranean Restaurant	Drugstore	Discount Store	Distribution Center	Doner Restaurant	Donut Shop
M3H	North York	Bathurst Manor, Wilson Heights, Downsview North	43.754328	-79.442259	1.0	1	Coffee Shop	Bank	Grocery Store	Mobile Phone Shop	Bridal Shop	Sandwich Place	Diner	Restaurant	Deli / Bodega	Supermarket

```
In [64]: cluster2.shape
Out[64]: (83, 17)
```

Cluster 5

```
In [66]: cluster5 = toronto_merged.loc[toronto_merged['Clusterlabels'] == 4,]
         cluster5
Out[66]:
```

	Postalcode	Borough	Neighborhood	Latitude	Longitude	Cluster_Labels	Clusterlabels	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 M Comm Ven
0	M3A	North York	Parkwoods	43.753259	-79.329656	4.0	4	Park	Construction & Landscaping	Food & Drink Shop	Fabric Shop	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Diner	East Europe Restaur
10	M6B	North York	Glencairn	43.709577	-79.445073	4.0	4	Park	Pizza Place	Japanese Restaurant	Pub	Doner Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Do St
21	M6E	York	Caledonia-Fairbanks	43.689026	-79.453512	4.0	4	Park	Women's Store	Pool	Falafel Restaurant	Fabric Shop	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dim S Restaur
35	M4J	East York	East Toronto, Broadview North (Old East York)	43.685347	-79.338106	4.0	4	Park	Pizza Place	Convenience Store	Drugstore	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Do St
49	M6L	North York	North Park, Maple Leaf Park, Upwood Park	43.713756	-79.490074	4.0	4	Park	Basketball Court	Construction & Landscaping	Bakery	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	East Europe Restaur
50	M9L	North York	Humber Summit	43.756303	-79.565963	4.0	4	Pizza Place	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	East Europe Restaur
61	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	4.0	4	Park	Bus Line	Swim School	Drugstore	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	D

Cluster 2 was further analyzed by first creating a dataframe comprising of neighborhoods with no Yoga studios amongst the top 10 venues. This reduced the number of neighborhoods from 83 to 61. Following this, the neighborhoods from the latter dataframe was filtered for those which had parks included amongst the top ten venues. This created a new dataframe which comprised of all neighborhoods in cluster 2 that satisfied our required criteria. This comprised of 12 neighborhoods. Therefore, among the 83 neighborhoods in cluster 2, there were 12 neighborhoods which were considered to be suitable for starting a Yoga studio.

THERE ARE 12 SUCH NEIGHBORHOODS IN CLUSTER 2 WHICH MAY BE RECOMMENDED TO START A YOGA STUDIO

```
In [81]: YogaMinusParkPlus2['Neighborhood']
Out[81]: 2                    Regent Park, Harbourfront
         25                                 Christie
         31               Dufferin, Dovercourt Village
         40                                Downsview
         44            Golden Mile, Clairlea, Oakridge
         46                                Downsview
         47             India Bazaar, The Beaches West
         53                                Downsview
         60                                Downsview
         67                          Davisville North
         74        The Annex, North Midtown, Yorkville
         84    Kensington Market, Chinatown, Grange Park
         Name: Neighborhood, dtype: object
```

Cluster 5 was similarly analyzed. From the point of view of our filtering it was found that cluster 5 returned a much higher proportion of neighborhoods which were suitable for opening a Yoga studio. After using a similar filtering mechanism comprising of neighborhoods with no yoga studios among the first 10 venues but with parks among such venues it was found that 11 of the 23 neighborhoods fulfilled the criteria.

THERE ARE 12 SUCH NEIGHBORHOODS IN CLUSTER 2 WHICH MAY BE RECOMMENDED TO START A YOGA STUDIO

```
In [81]: YogaMinusParkPlus2['Neighborhood']

Out[81]: 2                    Regent Park, Harbourfront
         25                                   Christie
         31                 Dufferin, Dovercourt Village
         40                                   Downsview
         44              Golden Mile, Clairlea, Oakridge
         46                                   Downsview
         47               India Bazaar, The Beaches West
         53                                   Downsview
         60                                   Downsview
         67                             Davisville North
         74          The Annex, North Midtown, Yorkville
         84     Kensington Market, Chinatown, Grange Park
         Name: Neighborhood, dtype: object
```

Finally, the two dataframes, one from the suitable neighborhoods of Cluster 2 and the other from the suitable neighborhoods of cluster 5 were merged into a single dataframe.

Concatenating the two dataframes from cluster 2 and cluster 5

```
84]: YogaMinusParkPlus2.append(YogaMinus5,ignore_index = True, sort = False)
     YogaStudio_recommendation = YogaMinusParkPlus2.append(YogaMinus5,ignore_index = True, sort = False)

85]: YogaStudio_recommendation
```

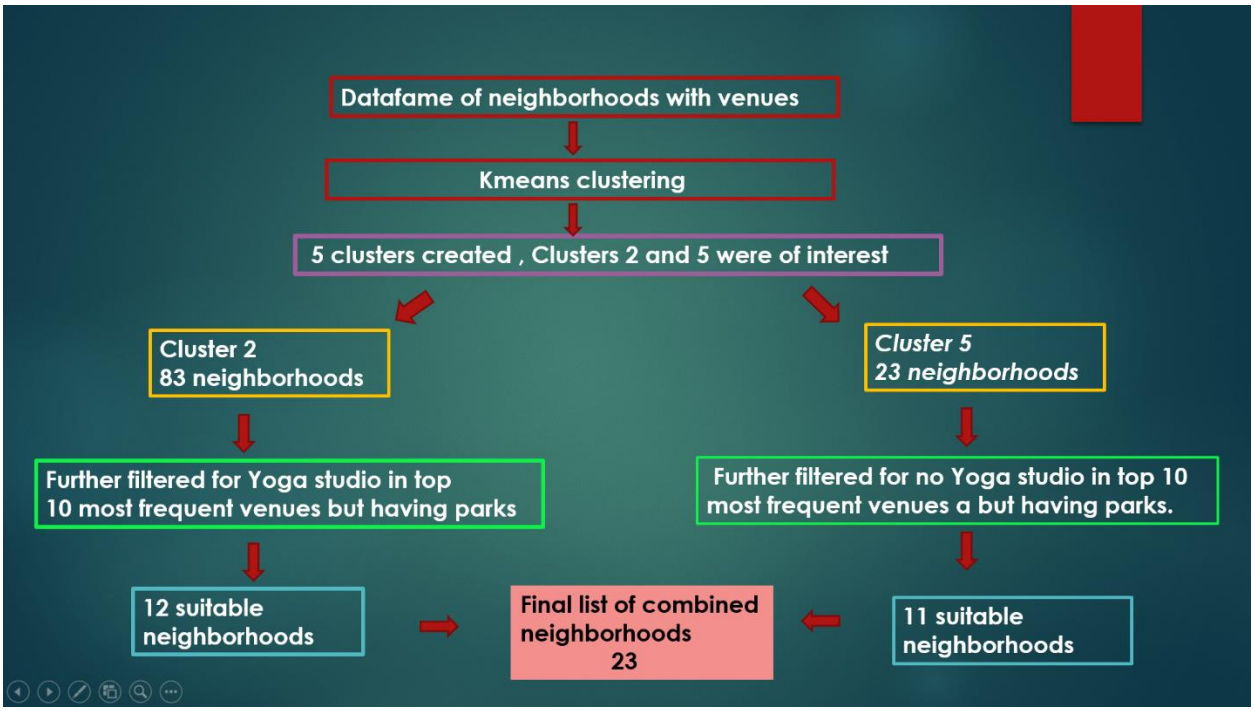
ut[85]:

	Postalcode	Borough	Neighborhood	Latitude	Longitude	Cluster_Labels	Clusterlabels	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
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	1.0	1	Coffee Shop	Bakery	Pub	Park	Breakfast Spot	Café	Restaurant	Theater	French Restaurant
1	M6G	Downtown Toronto	Christie	43.669542	-79.422564	1.0	1	Grocery Store	Café	Park	Restaurant	Diner	Baby Store	Nightclub	Italian Restaurant	Athletics & Sports
2	M6H	West Toronto	Dufferin, Dovercourt Village	43.669005	-79.442259	1.0	1	Bakery	Pharmacy	Park	Middle Eastern Restaurant	Music Venue	Pet Store	Pizza Place	Café	Brewery
3	M3K	North York	Downsview	43.737473	-79.464763	1.0	1	Grocery Store	Park	Bank	Airport	Snack Place	Gym / Fitness Center	Shopping Mall	Baseball Field	Liquor Store
4	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577	1.0	1	Bakery	Bus Line	Ice Cream Shop	Intersection	Metro Station	Bus Station	Soccer Field	Park	Eastern European Restaurant

The final list comprised of 23 neighborhoods which might be recommended for opening a Yoga studio out of 98 neighborhoods which were studied.

```
Out[95]: 0                                Regent Park, Harbourfront
         1                                                 Christie
         2                            Dufferin, Dovercourt Village
         3                                                Downsview
         4                          Golden Mile, Clairlea, Oakridge
         5                                                Downsview
         6                           India Bazaar, The Beaches West
         7                                                Downsview
         8                                                Downsview
         9                                          Davisville North
         10                     The Annex, North Midtown, Yorkville
         11                Kensington Market, Chinatown, Grange Park
         12                                               Parkwoods
         13                                               Glencairn
         14                                     Caledonia-Fairbanks
         15           East Toronto, Broadview North (Old East York)
         16                North Park, Maple Leaf Park, Upwood Park
         17                                         Lawrence Park
         18                                        York Mills West
         19          Forest Hill North & West, Forest Hill Road Park
         20        Milliken, Agincourt North, Steeles East, L'Amo...
         21                                                Rosedale
         22        Old Mill South, King's Mill Park, Sunnylea, Hu...
         Name: Neighborhood, dtype: object
```

Summary of Results



5. Discussion and conclusions

The project was undertaken to create a system of recommendation for locations for starting new business ventures in a huge city like Toronto. After acquiring the necessary data on the different neighborhoods and the venues which are present within the neighborhood, a KMeans clustering algorithm was used to cluster the neighborhood. It was found that when clustering the neighborhood into 5 clusters only two had a substantial number of neighborhoods in it. I used a manual method of trying out cluster numbers from 3 to 15 and 5 gave the maximum breakup of neighborhoods although we would like to see a better split. I used a very recent version of the Foursquare API (as recent as the current month of June 2020). It is possible that a number of business ventures and other commercial activities might have been affected due to the current pandemic and when using live data to check on venues there might be issues with data and the subsequent results.

After analyzing the two major clusters i.e. clusters 2 and 5, it was possible to filtering down the 98 neighborhoods in the city of Toronto to 23 neighborhoods. I used two features to filter; one commercial i.e. a less frequency of other Yoga studios and one non-commercial i.e. presence parks. This was created on the premise that people who opt to live in areas with many parks might be more inclined to remain connected with nature and thus explore their spirituality or good health through Yoga.

The neighborhood list was merged into a dataframe which might give the stakeholders an idea of other venues in that neighborhood and which might be used for further analysis based on demographic data like average age of the population and socio-economic status which might be obtained from other databases. In addition, one could get a snapshot of other businesses or

venues in these neighborhoods which might be relevant to those interested in joining a Yoga Studio like other gyms and fitness centers and clothing stores or even a dog run.
In conclusion, this project shows how one might put together a data science cum machine learning process to facilitate the process of decision making in business venture, policy making or other endeavors of a more personal nature