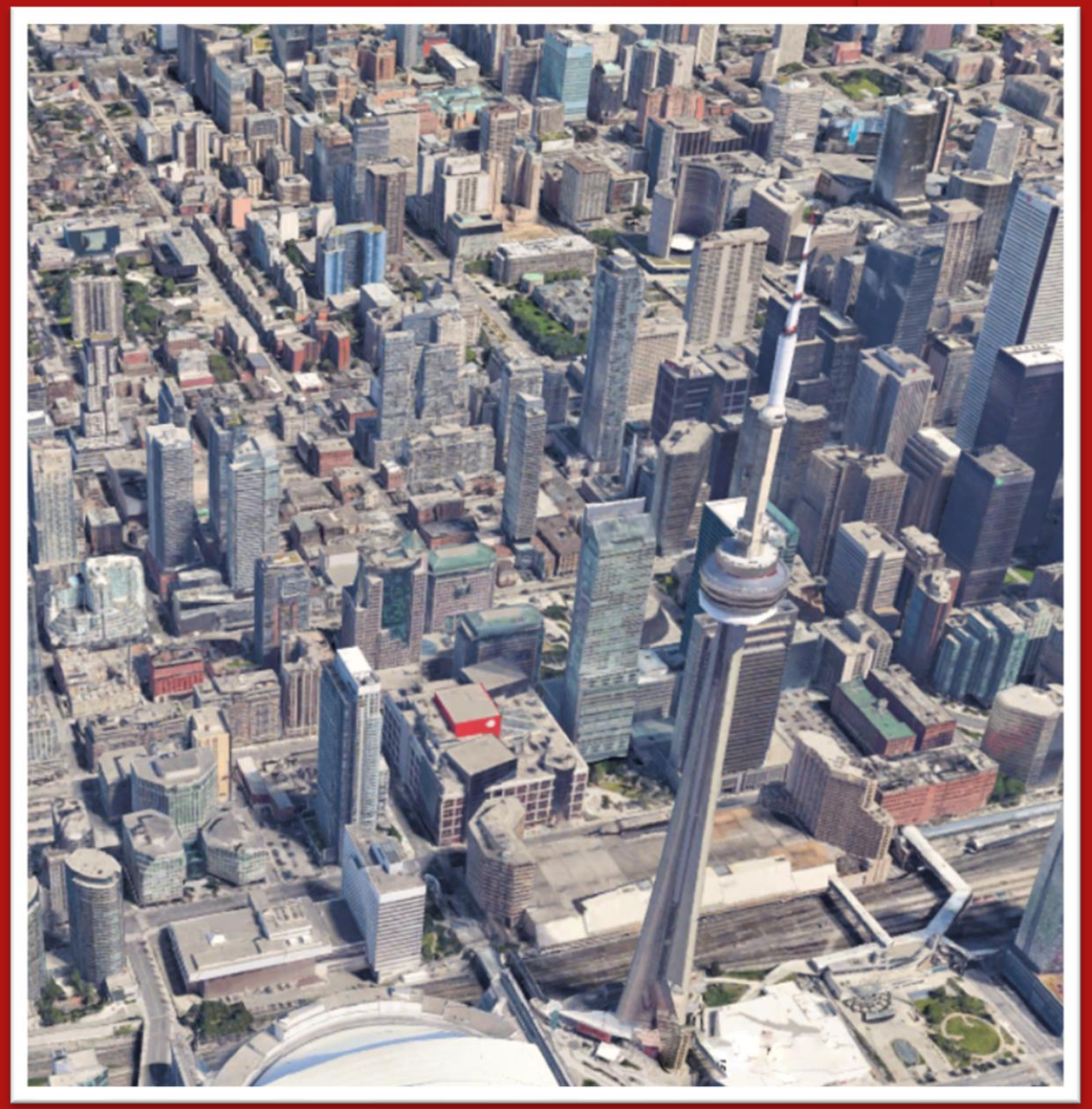
Clustering of Toronto neighborhoods for better business opportunities

Opening a new Yoga Studio



MRINMAY KUMAR MALLIK

Introduction

Introduction

The background

- **Toronto** : a dynamic metropolis and home to 2.7 million people.
- **Its neighborhoods** : A myriad mix of features including residential areas, public facilities and commercial enterprises.
- **The need** : A way to find similarities and differences among neighborhoods according to the above features and categorize them into groups.
- **In the past** catering to such a need was tedious , error prone and non-reproducible
- **Today** with access to huge accessible databases and smart machine learning algorithms the process is easier, more objective and less error-prone

Introduction
General Problem and stakeholders

- **Problem ?** To categorize neighborhoods into categories according to public facilities, commercial venues , educational institutions and demographic data to better understand their character.
- **How to solve such a problem ?** Use machine unsupervised machine learning algorithms like Kmeans clustering on relevant data that caters to such a need to categorize neighborhoods into clusters.
- **How would be the general stakeholders ?** Government policy makers in deciding the suitable areas to allocate public resources , a person looking for a good neighborhood to move into or an entrepreneur in the process of beginning a new business or expanding one.

Introduction : Finding a suitable location for a business venture

- **Most important determinant for any business enterprise** : Location
- **What is one looking for ?** : Less competitors and more demand
- **How can you help ?** Find a solution by coming up with a list of neighborhoods which satisfies the above criteria.
- **Who would be interested stakeholders** : Anyone willing to start a new business or expand it.

The Current Project

FINDING A SUITABLE LOCATION TO OPEN A YOGA STUDIO.

Introduction : Finding a neighborhoods suitable for yoga studio

- **Strategy** : To create neighborhood clusters , find the frequency of different kinds of venues among these clusters and further filter the clusters for neighborhoods where Yoga studios are not frequent. In addition based on the premise of a possible greater demand for Yoga studios amongst people who are like to stay more connected with nature, find neighborhoods amongst the above which have more of parks as compared to other
- **Audience** : Anyone willing to open a new Yoga studio.

The Datasets

- 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

The Python libraries used

- 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

Methodology

The General Methodology Roadmap

- 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) Finding neighborhoods with low frequency of existing Yoga studios and high frequency of parks

Creating the neighborhood dataframe

- **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 was **scraped** using "BeautifulSoup" and "Requests"
- iii) **Cleaning of data** by dropping boroughs with unassigned neighborhoods and assigning neighborhoods with no names to the names of their boroughs
- iv) This created a dataframe with Postalcodes, boroughs and neighborhoods
- v) A second dataframe was created with postal code with latitude and longitides from http://cocl.us/Geospatial_data
- vi) The above two dataframes were merged

The basic pandas dataframe of neighborhoods of Toronto

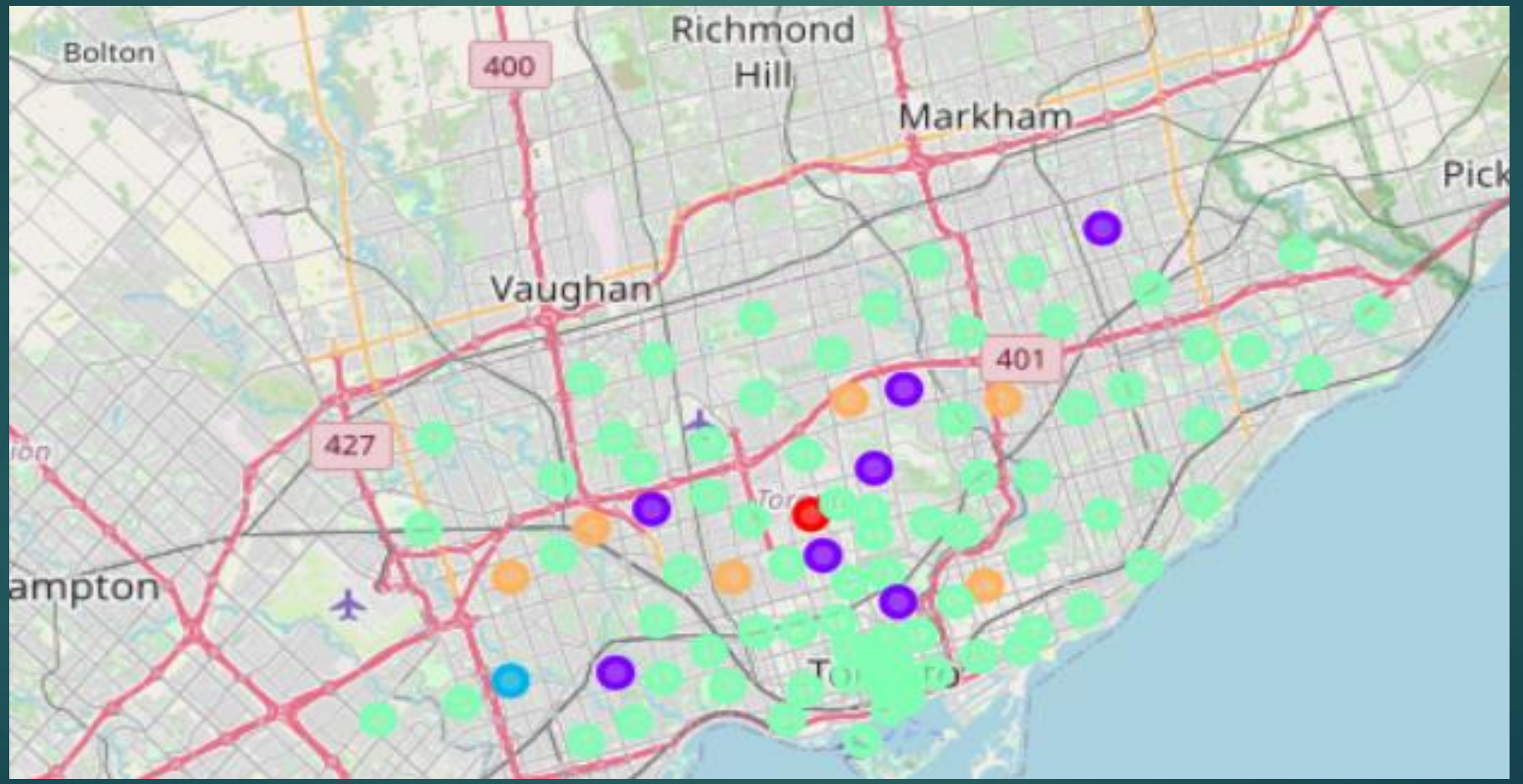
	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A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

The folium map was created for the neighborhoods

The 10 most frequent venues were obtained through Foursquare API and later one hot encoding

	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

K means clustering was performed to generate the clusters



Further filtering of data

- **Neighborhoods in 2 of the clusters were further filtered for neighborhoods with less frequent Yoga studios and more frequent parks.**

Results

The clusters.

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

Cluster 2 (Only the head out of 83 neighborhoods)

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 2 neighborhoods dataframe.head suitable for yoga studios

Finding the neighborhoods with no studios but with parks among the top ten venues

```
filt2 = (YogaMinus['1st Most Common Venue'] == 'Park') |(YogaMinus['2nd Most Common Venue'] == 'Park')|(YogaMinus['3rd Most Common Venue'] == 'Park')|(YogaMinus['4th Most Comm
YogaMinus[filt2]
YogaMinusParkPlus2 = YogaMinus[filt2]
YogaMinusParkPlus2   # Dataframe with neighborhoods in cluster 2 with parks in top ten venues but no Yoga studios
```

6]:

	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 Mo Commo Venu
2	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	Electroni Sto
25	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 Sh
31	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	Supermark
40	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	Discou Sto
44	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	Dumpli Restaura
46	M3L	North York	Downsview	43.739015	-79.506944	1.0	1	Grocery Store	Park	Bank	Airport	Snack Place	Gym / Fitness Center	Shopping Mall	Baseball Field	Liquor Store	Disc S

Cluster 2 list of neighborhoods suitable for 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

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 Ver
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 Sh
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 Sh
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 Europ Resta
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	Di

Cluster 5 neighborhoods dataframe.head suitable for yoga studios

```
YogaMinus5 = cluster5[-filt3]
YogaMinus5  # Neighborhoods in cluster 5 without yoga studios in first ten commonest venues
```

]:

	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 Ver
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 Sh
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 Sh
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
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	Di
66	M2P	North York	York Mills West	43.752758	-79.400049	4.0	4	Park	Construction & Landscaping	Convenience Store	Bar	Dumpling Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drug

Cluster 5 list of neighborhoods suitable for Yoga studio

All the neighbourhoods in Cluster 5 without Yoga studios have parks and therefore there is no need for further refinement since all have parks

```
In [59]: YogaMinus5['Neighborhood']
```

```
Out[59]: 0                                          Parkwoods
         10                                         Glencairn
         21                               Caledonia-Fairbanks
         35         East Toronto, Broadview North (Old East York)
         49              North Park, Maple Leaf Park, Upwood Park
         61                                     Lawrence Park
         66                                   York Mills West
         68       Forest Hill North & West, Forest Hill Road Park
         85     Milliken, Agincourt North, Steeles East, L'Amo...
         91                                         Rosedale
         101    Old Mill South, King's Mill Park, Sunnylea, Hu...
         Name: Neighborhood, dtype: object
```

Combined dataframe of suitable neighborhoods from both clusters 2 and 5

Concatenating the two dataframes from cluster 2 and cluster 5

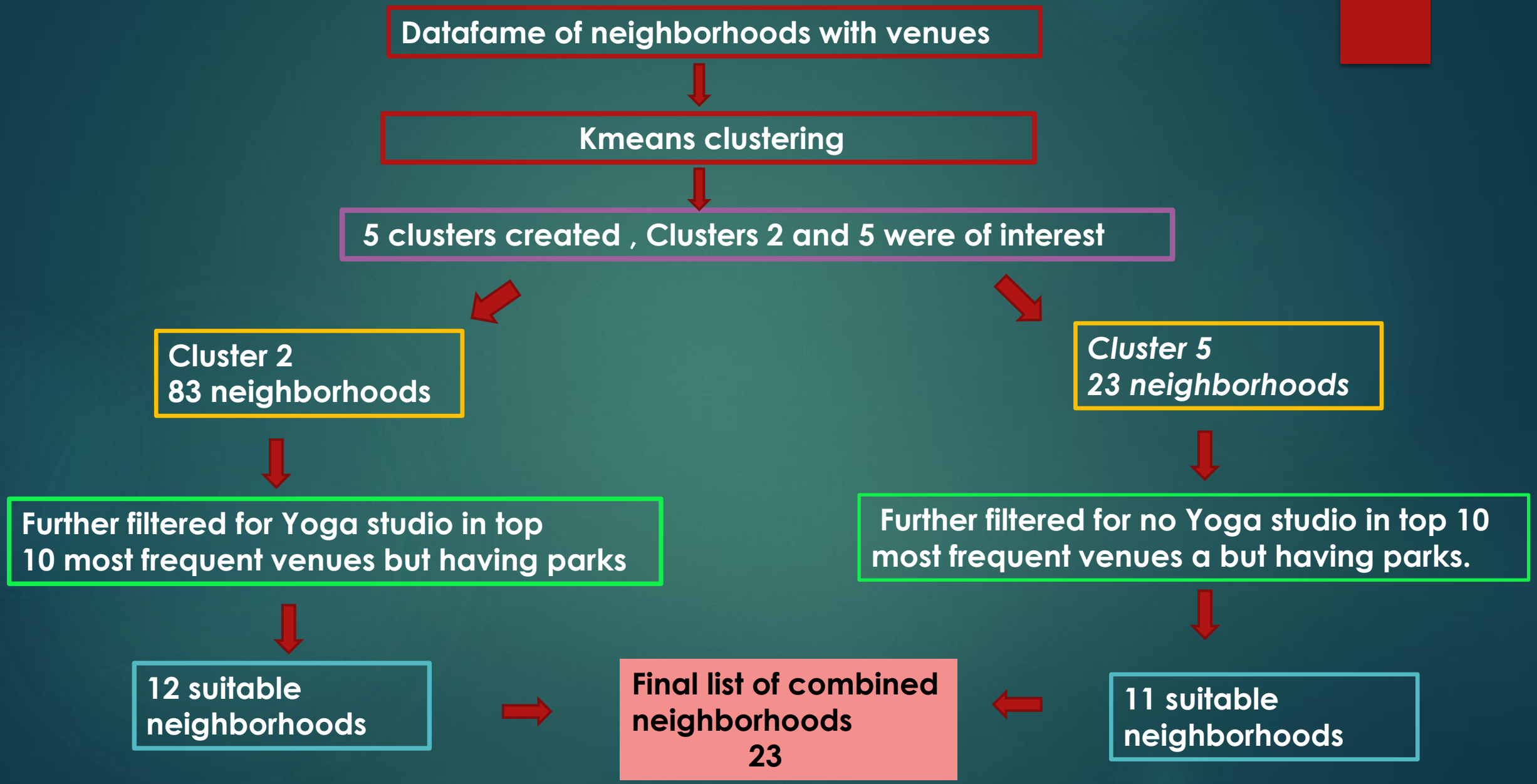
```
84]: YogaMinusParkPlus2.append(YogaMinus5,ignore_index = True, sort = False)
     YogaStudio_recommendation = YogaMinusParkPlus2.append(YogaMinus5,ignore_index = True, sort = False)

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 entire list of 23 neighborhoods suitable for Yoga studios from both clusters

```
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
```

Result Summary


Datafame of neighborhoods with venues
Kmeans clustering
5 clusters created , Clusters 2 and 5 were of interest
Cluster 2
83 neighborhoods
Cluster 5
23 neighborhoods
Further filtered for Yoga studio in top
10 most frequent venues but having parks
Further filtered for no Yoga studio in top 10
most frequent venues a but having parks.
12 suitable
neighborhoods
Final list of combined
neighborhoods
23
11 suitable
neighborhoods

Discussion and conclusions

Discussion and conclusions

- 12 of the 83 (14.45 %) neighborhoods in cluster 2 and 11 (84.61 %) of the 13 neighborhoods in cluster 5 were found to be suitable for opening a new Yoga studio.
- Thus although cluster size of 5 is much smaller than 2, it has a greater proportion of neighborhoods suitable for Yoga studios.
- Overall, the process was able to find that 23 out of the 98 neighborhoods examined were suitable for Yoga studios.

Discussion and conclusions

- Cluster sizes of 3 of the 5 were very small although on manually testing for clusters numbers between 3 and 20, 5 was found to be most suitable.
- Since a very recent version of Foursquare API with live data has been used , it is possible that the ongoing Corona virus pandemic might have altered the profile of active business enterprises.

Discussion and conclusions

- The project has been able to create a list of neighborhoods by filtering data according to the frequency of available business enterprises and a feature which is non commercial like a public park.
- It might be possible to find out interesting correlation with other features present in the combined dataframe of neighborhoods suitable for opening Yoga studios.
- Other features like demographic data including age of the population and socio-economic status may be merged within this dataframe to gather useful insights
- The stakeholders using the dataframe should provide the data scientist with crucial follow up data for further refinements and error detection.

Thank You