Neighborhoods to build gyms in Toronto

Introduction:

statistics shows that the revenue in the Fitness segment is projected to reach US$350m in 2020 in Canada, and is expected to show an annual growth rate (CAGR 2020-2024) of 0.3%, resulting in a projected market volume of US$354m by 2024. [1] The industries related to health and fitness are growing in other developed countries too. Based on an article in *forbes*, according to International Health, Request& Sportsclub Association (IHRSA) the $30 billion health and fitness industry in the U.S. has been growing by at least 3 - 4% annually for the last ten years.[2] Consumer trends and the proliferation of public health campaigns have increased the strength of the Gym, Health and Fitness Clubs industry in Canada in recent years. Adult obesity rate is expected to rise. Therefore, the Public Health Agency of Canada (PHAC) has increasingly stressed the merits of fitness regimens and healthy lifestyle choices. Besides, health insurance companies and businesses seeking to improve workforce productivity and lower healthcare costs have provided incentives for health club memberships as a means of preventative care, buoying industry revenue growth.[3]

The growing industry of fitness has attracted many investors. Although there is a great increase in fitness apps that helps people to exercise on their own, buying each fitness or body building equipment is not affordable by most of the people and the best option is buying a gym membership and exercise under supervision of professionals. Therefore, investors are interested to invest in building gyms for this growing demand especially in cities like Toronto as financial capital of a developed country because there is a positive relationship between the income of individuals and their attending the gyms. People in wealthier states get more exercise as they have more disposable income to buy gym membership and sport clothes and running shoes.[4]

Deciding where to build a gym, in addition to other factors like the professional instructors and gym equipment, is important for its success. The population of people living in an area, the more population the more potential gym membership, and their income, due to positive relationship between income and gym membership, are two factors affecting the decision of choosing the location of a gym. Although there are other factors like real estate price, in this project the focus is on the two mentioned factors and suggestions for better location will be made based on the related information. In addition to these two factors, we must get the data about the number of existing gyms in each area because the more existing gyms in an area makes it less attractive for invest on building a new gym there.

Data:

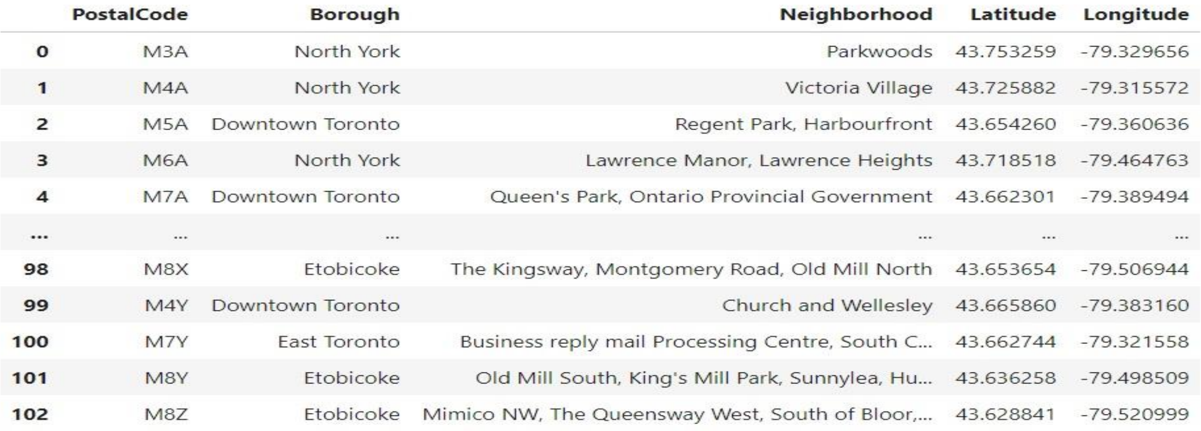
The data including information about postal codes of Canada, their latitude and longitude and also the population of neighborhoods in addition to the average income of the neighborhood households as a sign of neighborhood wealth were needed. Besides, the data showing the number of existing gyms in each neighborhood is needed as the more existing gyms, the less attractive the place for a new gym.

Four different sources were used to collect data. One, a *Wikipedia* page including the postal codes of Toronto, Boroughs and neighborhoods.[5] There are 180 postal codes with according boroughs and neighborhoods in this page but boroughs for some postal codes are not assigned. The second part of data were gotten from a downloaded CSV file that contained the latitudes and longitudes of 103 postal codes of Toronto.[6] Scraping the Canada statistics website[7] was used to get the population and average income of each neighborhood. We got only the related information for postal codes that were chosen after cleaning and merging the two data gathered from the first two sources. Then, the number of gyms in each neighborhood (in an area with defined radius around the geolocation of related postal codes) were gotten by using exploring the gyms in each neighborhood on *foursquare*.[8]

Methodology:

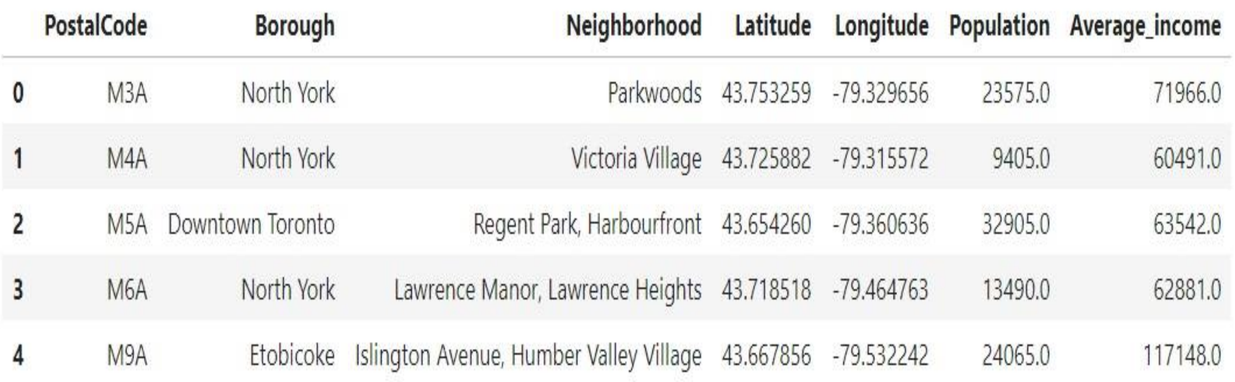
The data collected from W*ikipedia* included 180 rows and three columns of postal codes, boroughs and neighborhoods. Some boroughs were nor assigned. The rows containing not assigned borough were deleted. Besides, the extra elements like '\n' were also deleted to make the data cleaner.

In the next step, the CSV file was downloaded, checked, cleaned and merged with the first data. The result is shown below:



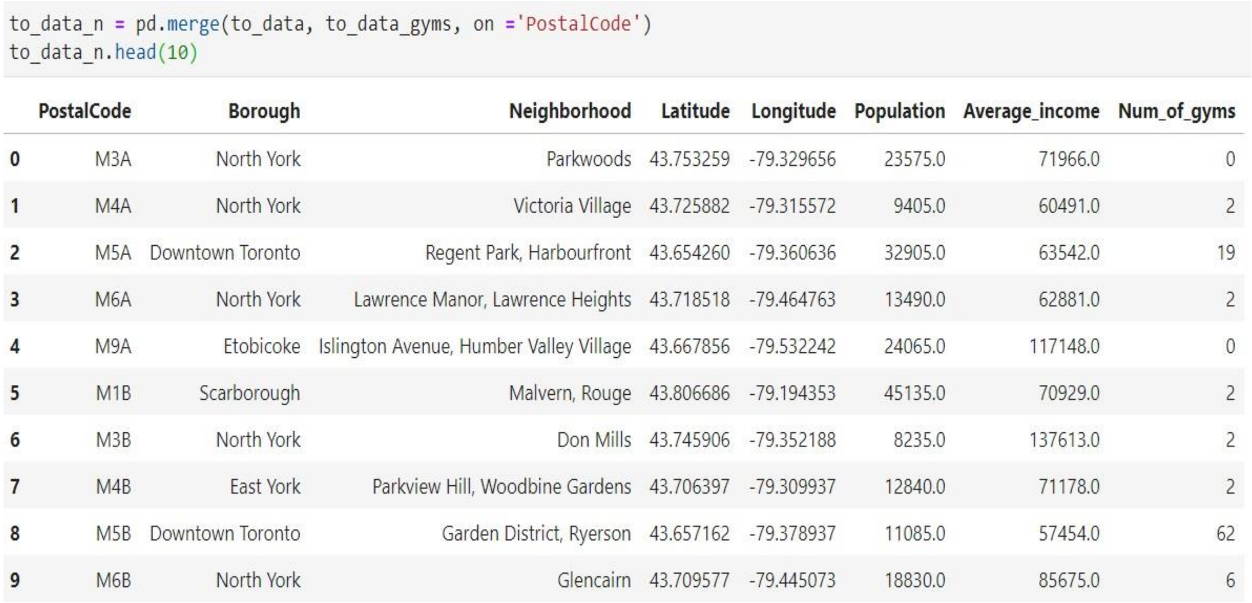
	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
...	...	...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

Getting the population (15 – 64 years old)[1] and household average income for each neighborhood in our dataframe was the next step. To do that, the Canada statistics website was scraped using ***Besutifulsoup***. The data contained some non-number values. Cleaning the data and merging the new information with our past data resulted in a new dataframe with 96 rows and 7 columns shown below: (the first five rows)



	PostalCode	Borough	Neighborhood	Latitude	Longitude	Population	Average_income
0	M3A	North York	Parkwoods	43.753259	-79.329656	23575.0	71966.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	9405.0	60491.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	32905.0	63542.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	13490.0	62881.0
4	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242	24065.0	117148.0

Bu using the ***Foursqaure API*** the number of gyms in an area with the radius of 1000 meters around location of each postal code were collected and saved. Then this new data were added to the previous data. The result is shown below: (the first ten rows)

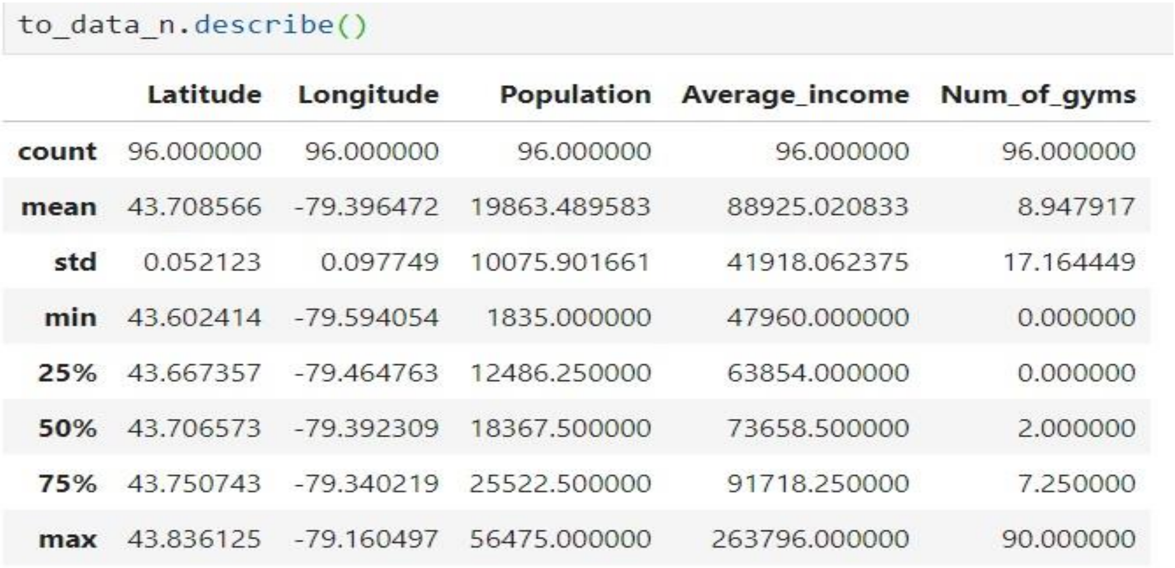


```
to_data_n = pd.merge(to_data, to_data_gyms, on ='PostalCode')
to_data_n.head(10)
```

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Population	Average_income	Num_of_gyms
0	M3A	North York	Parkwoods	43.753259	-79.329656	23575.0	71966.0	0
1	M4A	North York	Victoria Village	43.725882	-79.315572	9405.0	60491.0	2
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	32905.0	63542.0	19
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	13490.0	62881.0	2
4	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242	24065.0	117148.0	0
5	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	45135.0	70929.0	2
6	M3B	North York	Don Mills	43.745906	-79.352188	8235.0	137613.0	2
7	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937	12840.0	71178.0	2
8	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	11085.0	57454.0	62
9	M6B	North York	Glencairn	43.709577	-79.445073	18830.0	85675.0	6

[1] The population with age between 15 to 64 was gathered as almost all of the people attending the gyms are between 15 and 64.

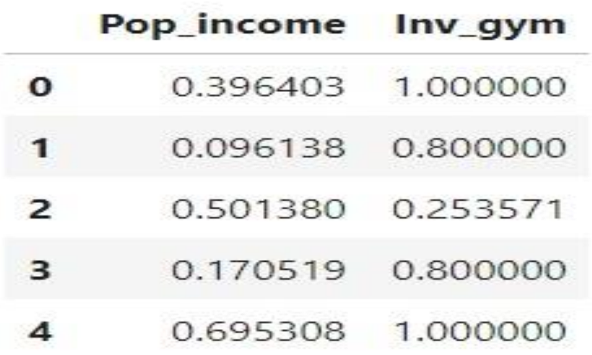
Preparing the data and clustering the neighborhoods was the next step. Before preparing the data, it was checked again.



```
to_data_n.describe()
```

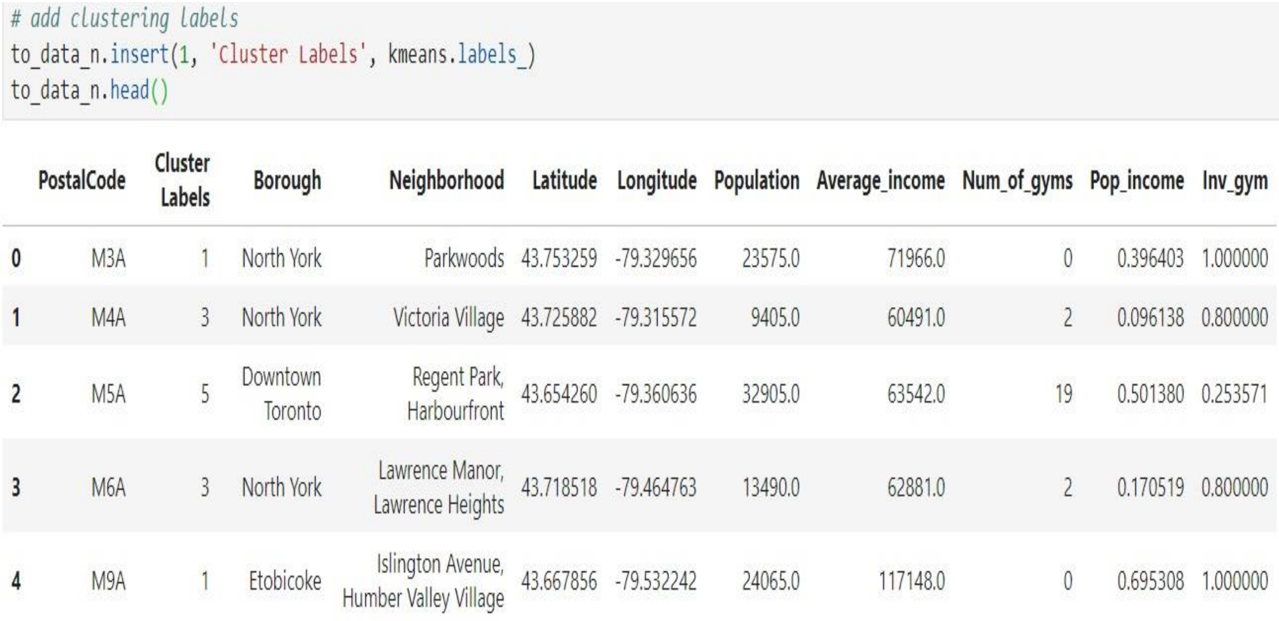
	Latitude	Longitude	Population	Average_income	Num_of_gyms
count	96.000000	96.000000	96.000000	96.000000	96.000000
mean	43.708566	-79.396472	19863.489583	88925.020833	8.947917
std	0.052123	0.097749	10075.901661	41918.062375	17.164449
min	43.602414	-79.594054	1835.000000	47960.000000	0.000000
25%	43.667357	-79.464763	12486.250000	63854.000000	0.000000
50%	43.706573	-79.392309	18367.500000	73658.500000	2.000000
75%	43.750743	-79.340219	25522.500000	91718.250000	7.250000
max	43.836125	-79.160497	56475.000000	263796.000000	90.000000

As mentioned in the introduction part, the population of people living in a neighborhood and also the average income of its households can be factors affecting the suitability of the neighborhood for building a gym. Besides, the number of existing gyms in the neighborhood has a negative effect on the intention for an investor to provide budget for a new gym. Therefore, two new variables were made, first one, the product of population and average income (**Pop_income**) and the second one, inverse of the number of gyms (**Inv_gym**) in the neighborhood. To reduce the effect of having zero or one gym, I added the average of number of gyms to each number of gym and then inversed it. These two new variables were normalized and then used for clustering the neighborhoods.



	Pop_income	Inv_gym
0	0.396403	1.000000
1	0.096138	0.800000
2	0.501380	0.253571
3	0.170519	0.800000
4	0.695308	1.000000

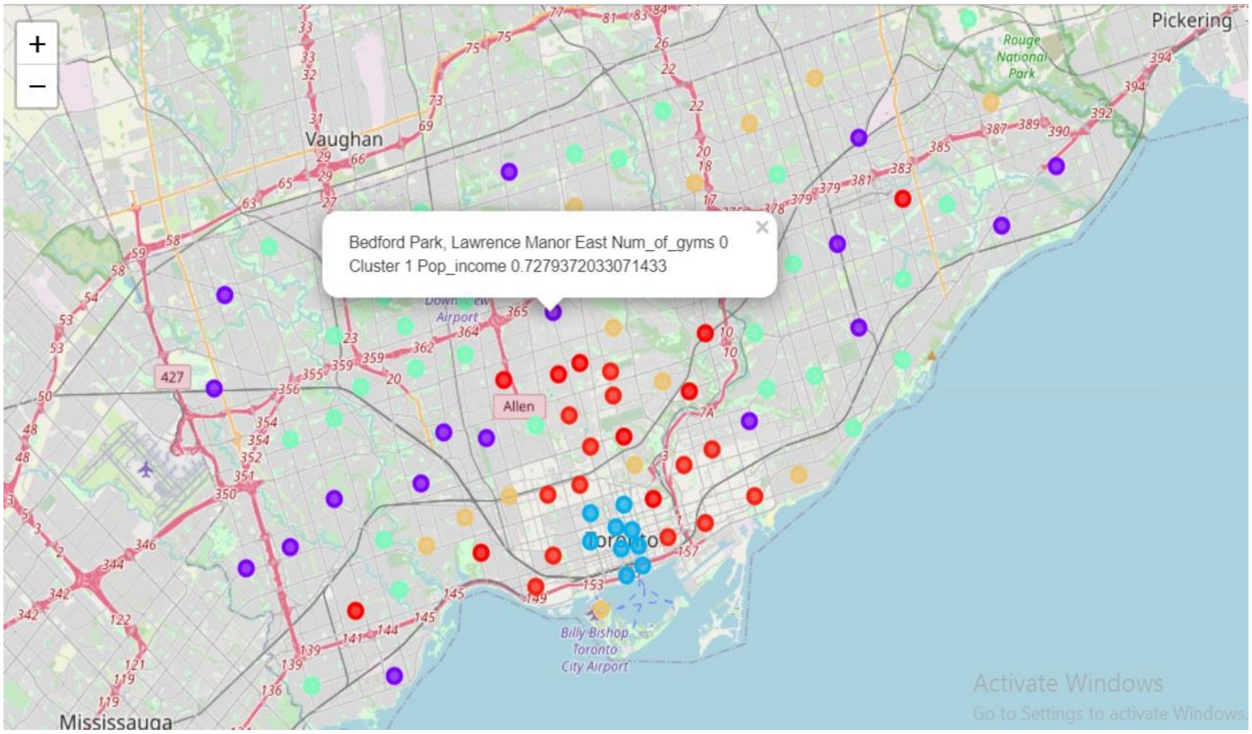
Before using the ***K-means*** algorithm to cluster neighborhoods, checking for the elbow point was tried and finally the K = 6 were chosen for the number of clusters. Then, the cluster labels were added to the dataframe. The result is shown below:



```
# add clustering labels
to_data_n.insert(1, 'Cluster Labels', kmeans.labels_)
to_data_n.head()
```

	PostalCode	Cluster Labels	Borough	Neighborhood	Latitude	Longitude	Population	Average_income	Num_of_gyms	Pop_income	Inv_gym
0	M3A	1	North York	Parkwoods	43.753259	-79.329656	23575.0	71966.0	0	0.396403	1.000000
1	M4A	3	North York	Victoria Village	43.725882	-79.315572	9405.0	60491.0	2	0.096138	0.800000
2	M5A	5	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	32905.0	63542.0	19	0.501380	0.253571
3	M6A	3	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	13490.0	62881.0	2	0.170519	0.800000
4	M9A	1	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242	24065.0	117148.0	0	0.695308	1.000000

Then, the results were visualized using ***folium*** library.

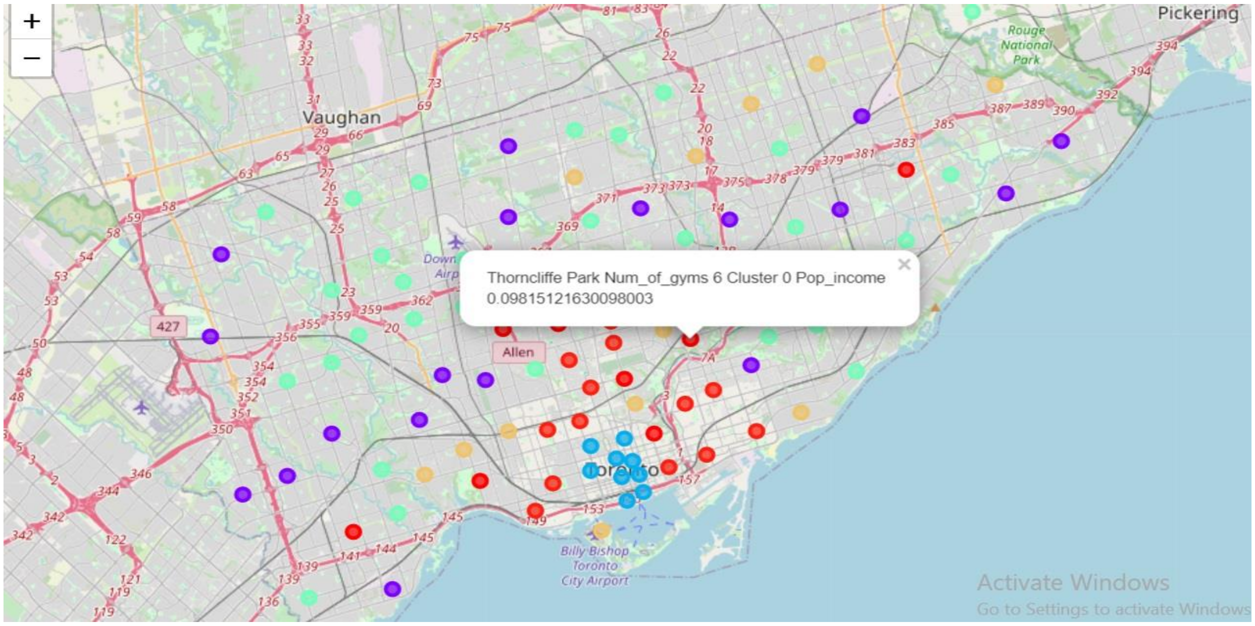


Bedford Park, Lawrence Manor East Num_of_gyms 0 Cluster 1 Pop_income 0.7279372033071433

Results:

Examples of a neighborhoods within each cluster are shown in the following pictures.

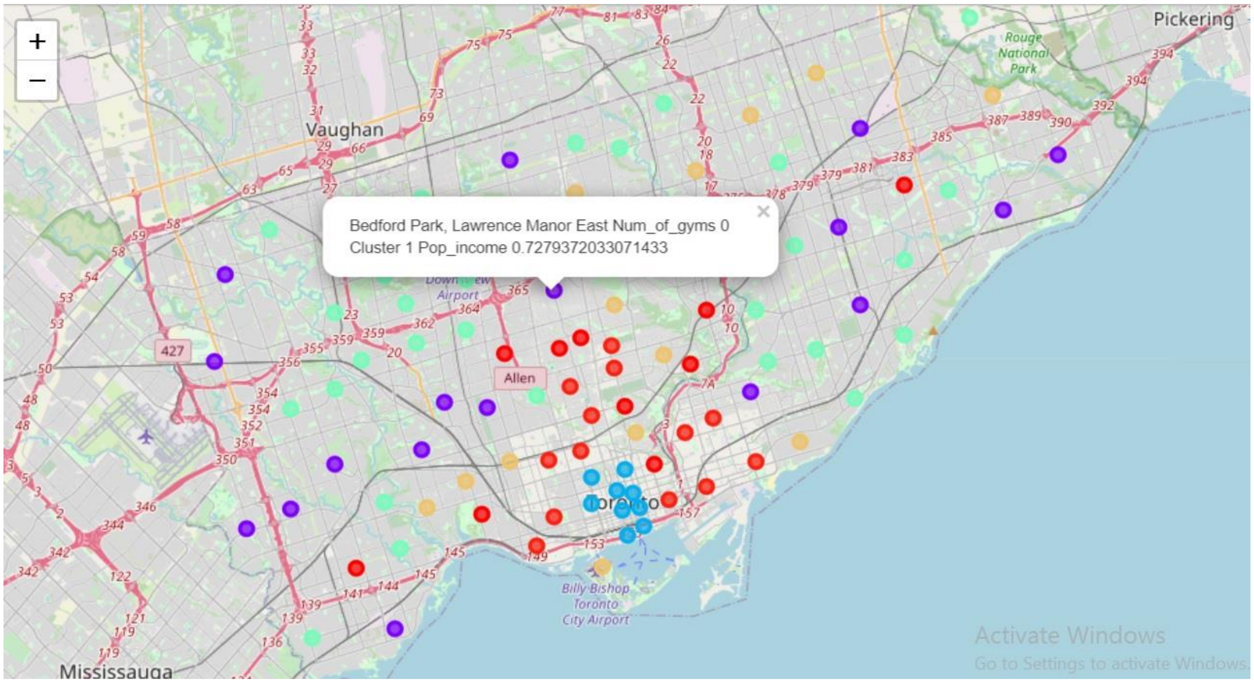
Cluster label 0 – moderate number of gyms -- low pop_income value



Thorncliffe Park Num_of_gyms 6 Cluster 0 Pop_income 0.09815121630098003

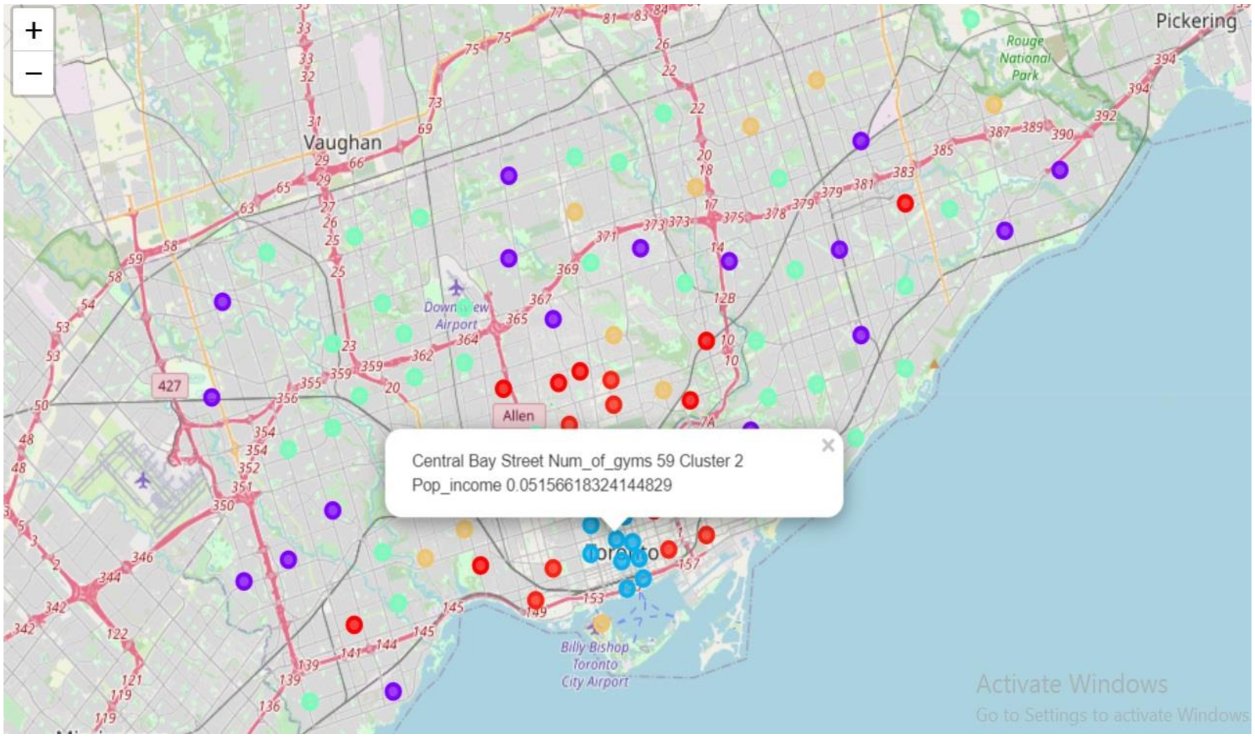
+
−
Pickering
Rouge National Park
Vaughan
Allen
Toronto
Billy Bishop Toronto City Airport
Activate Windows
Go to Settings to activate Windows

Cluster label 1 – zero or very low number of gyms -- moderate or high pop_income value



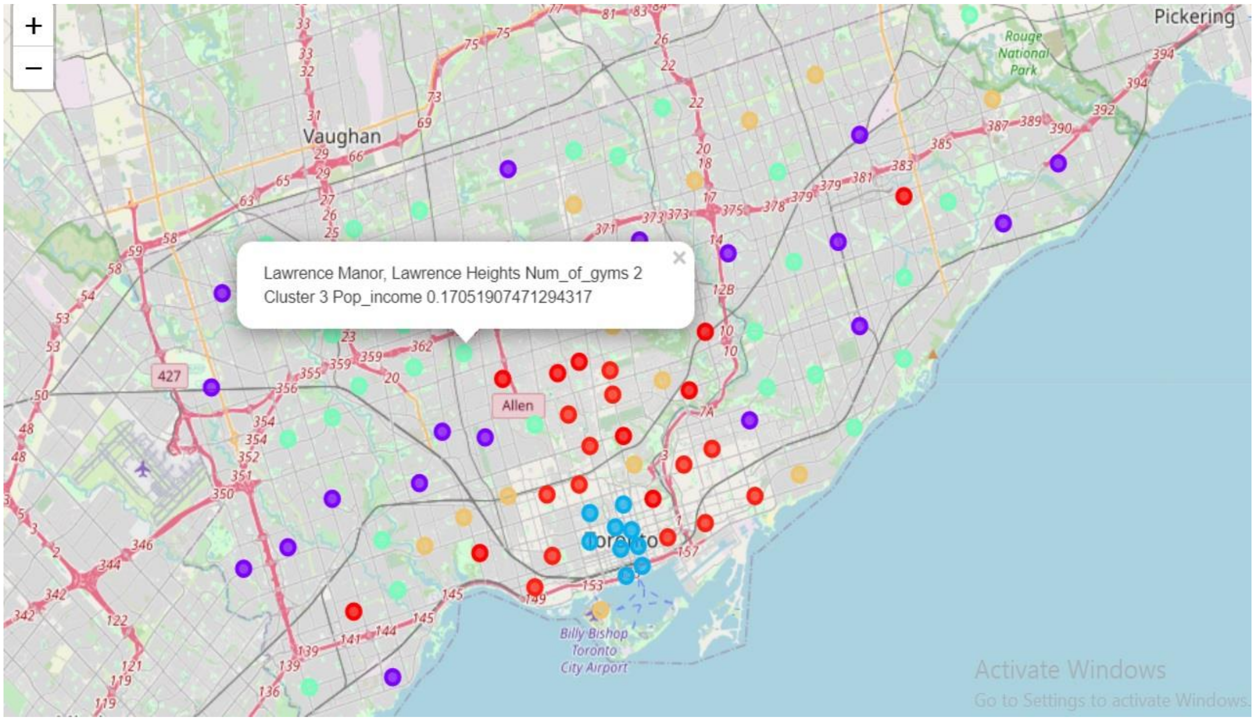
Bedford Park, Lawrence Manor East Num_of_gyms 0 Cluster 1 Pop_income 0.7279372033071433

Cluster label 2 – very high number of gyms –low pop_income value



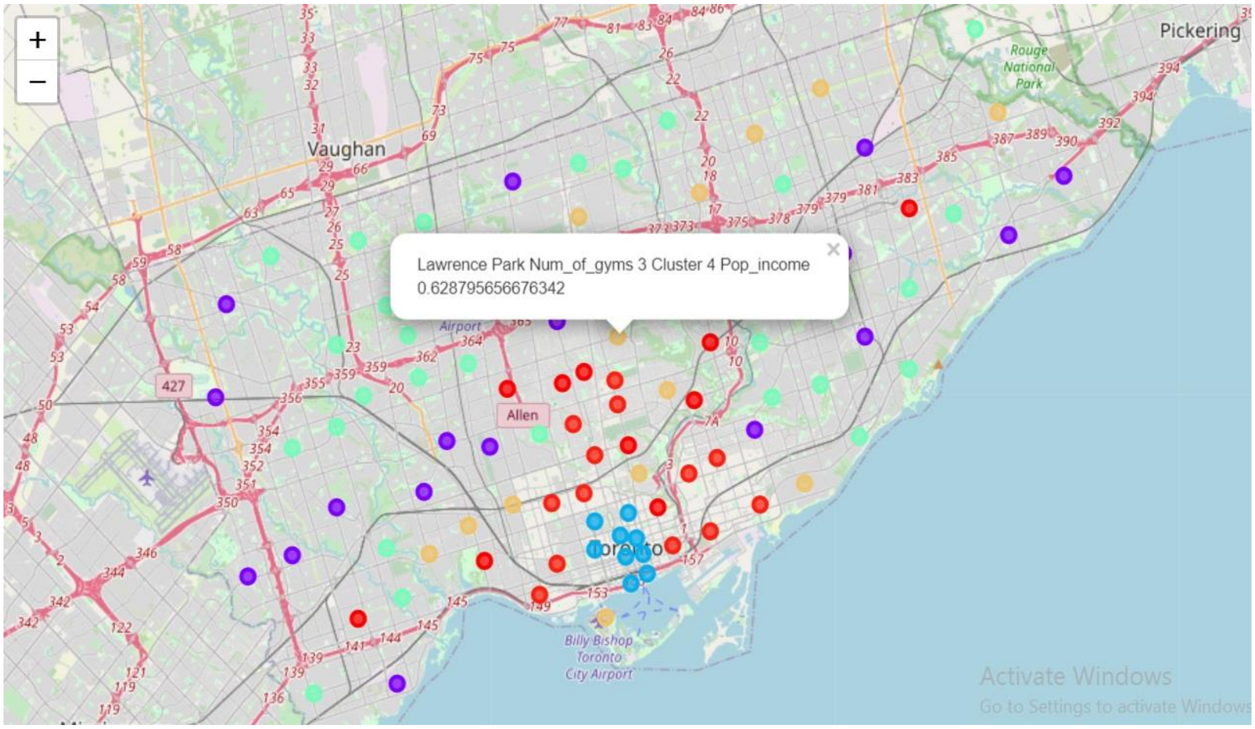
Central Bay Street Num_of_gyms 59 Cluster 2 Pop_income 0.05156618324144829

Cluster label 3 – low number of gyms – low pop_income value

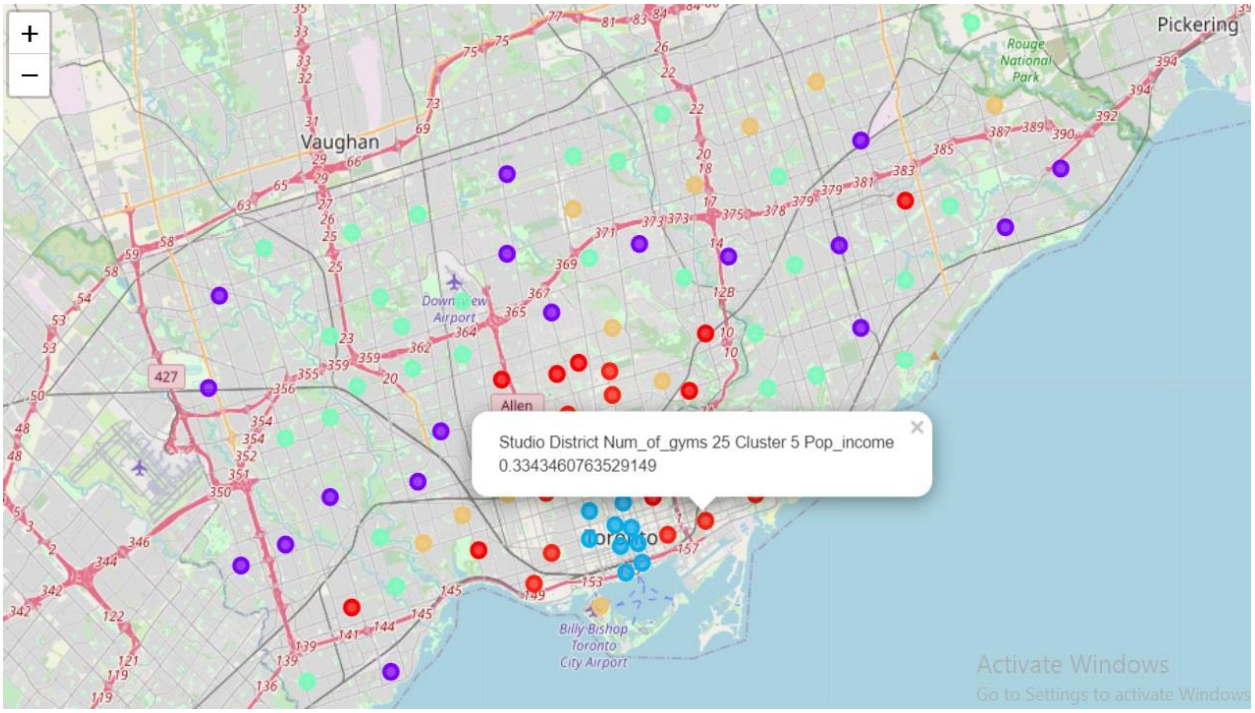

+
−
Vaughan
Pickering
Rouge National Park
427
Allen
Toronto
Billy Bishop Toronto City Airport
Activate Windows
Go to Settings to activate Windows.

Lawrence Manor, Lawrence Heights Num_of_gyms 2 Cluster 3 Pop_income 0.17051907471294317

Cluster label 4 – low or moderate number of gyms -- high or very high pop_income value

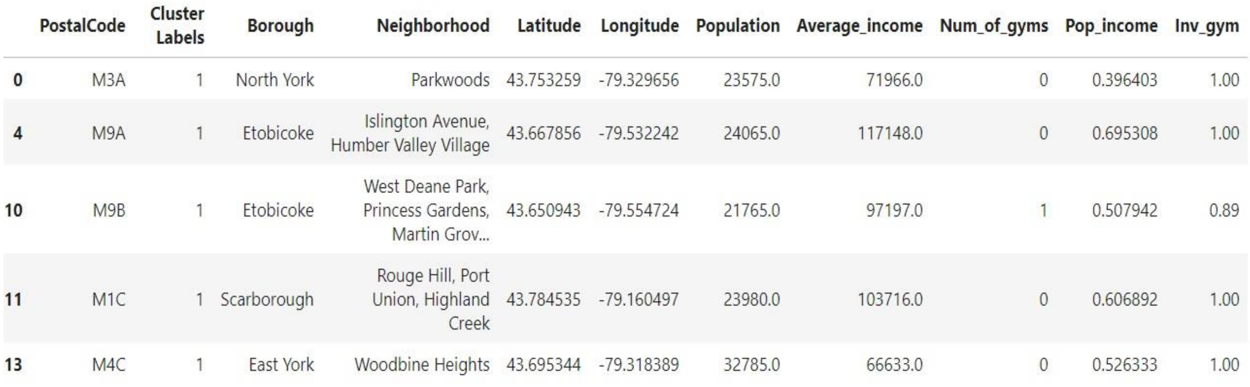

Lawrence Park Num_of_gyms 3 Cluster 4 Pop_income
0.628795656676342

Cluster 5 – above moderate number of gyms – moderate or high pop_income value


Studio District Num_of_gyms 25 Cluster 5 Pop_income
0.3343460763529149

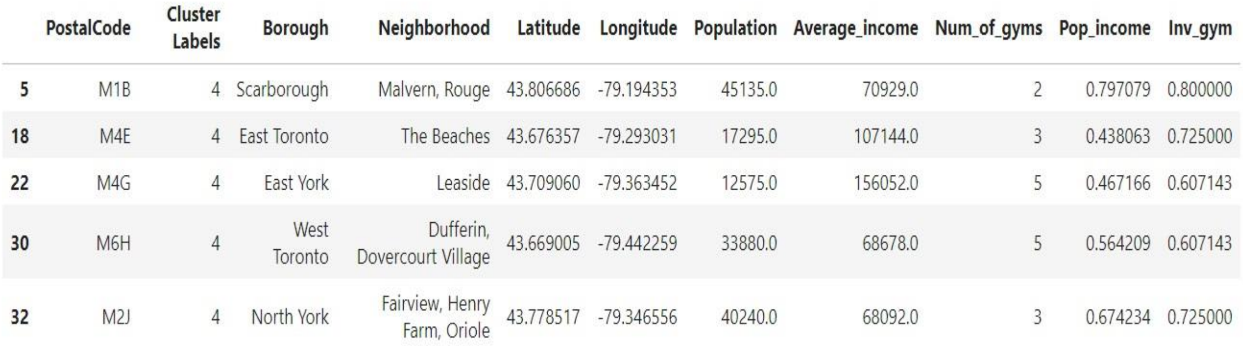
Discussion:

Checking the final dataframe and the map made by folium, it be can be seen that the neighborhoods with the clusters label of **1** (dark blue circles) are more appropriate. Neighborhoods in cluster **1** have zero or very low number of gyms and moderate or high *pop_income* value which is the normalized product of population and household average income. A sample of cluster 1 neighborhoods are shown below:



	PostalCode	Cluster Labels	Borough	Neighborhood	Latitude	Longitude	Population	Average_income	Num_of_gyms	Pop_income	Inv_gym
0	M3A	1	North York	Parkwoods	43.753259	-79.329656	23575.0	71966.0	0	0.396403	1.00
4	M9A	1	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242	24065.0	117148.0	0	0.695308	1.00
10	M9B	1	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...	43.650943	-79.554724	21765.0	97197.0	1	0.507942	0.89
11	M1C	1	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	23980.0	103716.0	0	0.606892	1.00
13	M4C	1	East York	Woodbine Heights	43.695344	-79.318389	32785.0	66633.0	0	0.526333	1.00

The next choice can be from the neighborhoods with cluster label of **4**. Neighborhoods in cluster **4** have low or moderate number of gyms, and high or very high *pop_income* value. A sample of cluster 4 neighborhoods are shown below:



	PostalCode	Cluster Labels	Borough	Neighborhood	Latitude	Longitude	Population	Average_income	Num_of_gyms	Pop_income	Inv_gym
5	M1B	4	Scarborough	Malvern, Rouge	43.806686	-79.194353	45135.0	70929.0	2	0.797079	0.800000
18	M4E	4	East Toronto	The Beaches	43.676357	-79.293031	17295.0	107144.0	3	0.438063	0.725000
22	M4G	4	East York	Leaside	43.709060	-79.363452	12575.0	156052.0	5	0.467166	0.607143
30	M6H	4	West Toronto	Dufferin, Dovercourt Village	43.669005	-79.442259	33880.0	68678.0	5	0.564209	0.607143
32	M2J	4	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556	40240.0	68092.0	3	0.674234	0.725000

Neighborhoods in these two cluster are more suitable for building a new gym. Although there are other factors like real estate price that can affect the decision.

Conclusion:

In this report, information about postal codes, their latitudes and longitudes, population and households average income of neighborhoods alongside with the data taken from *foursaure.com* about the gyms in each neighborhood helped us to choose more suitable locations for investing on building new gyms. This report shows that by using just online resources without need to buy any data accompanied by using data science techniques we can extract much useful information that can help us make better decisions.

Thanks for your attention.

Hamid Zaeimi

Reference:

[1] https://www.statista.com/outlook/313/108/fitness/canada

[2] https://www.forbes.com/sites/benmidgley/2018/09/26/the-six-reasons-the-fitness-industry-is-booming/#fcf8f01506db

[3] https://www.marketresearch.com/IBISWorld-v2487/Gym-Health-Fitness-Clubs-Canada-12808267/

[4] https://www.washingtonpost.com/news/wonk/wp/2018/07/03/more-money-more-fitness-why-people-in-the-wealthiest-states-get-more-exercise/

[5] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

[6] http://cocl.us/Geospatial_data

[7] https://www12.statcan.gc.ca/census-recensement/index-eng.cfm

[8] https://foursquare.com/