The Battle of Neighborhood

Best location to open GYM/Fitness Centre in Toronto

Laxmikantha Herle

Introduction

This project aims to utilize all Data Science concepts learned in the Data Science Professional Course. We define a Business Problem, the data that will be utilized and using that data, we are able to analyze it using Machine Learning tools. In this project, we will go through all the processes in a step by step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

Table to Contents

- Problem Description
- Data Description
- Methodology
- Machine Learning
- Data Analysis
- Discussion and Conclusion
- References

Problem Description

The objective of this project is to determine “what might be the ‘best’ Neighborhood in *Toronto* to open a GYM/Fitness Centre”. Will use *foursquare* location data and regional clustering of venue information to determine the ‘best’ Neighborhood in *Toronto* to open a GYM/Fitness Centre. We will find the most suitable location for an entrepreneur to open a new GYM/Fitness Centre in *Toronto*, *Canada*.

Target Audience

Information provided by this report would be useful for People who wants open GYM/Fitness Centre in *Toronto*, *Canada*. The Objective is to locate and recommend to People which neighborhood of *Toronto* will be the best choice to open GYM/Fitness Centre.

Data Description

Below Data Sources used for the Analysis

- *Toronto* Neighborhood Data: The following Wikipedia page was scraped to pull out the necessary information:

  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- Coordinate data for each Neighborhood in *Toronto*: The following csv file gave us the geographical coordinates of each postal code:

  http://cocl.us/Geospatial_data

- Venue Data in *Toronto, Canada*. Geographical Coordinates data will be utilized as input for the *Foursquare* API that will be leveraged to provision venues information for each Neighborhood. We will use *Foursquare* API to explore Neighborhood in *Toronto, Canada*.

Methodology

Toronto Neighborhood Data

- Used the *BeautifulSoup* package to transform the data in the table on the *Wikipedia* page into the pandas data frame.
- Cleansing and Merging of the data was required to start the process of analysis.
- There were Boroughs that were not assigned to any Neighborhood therefore, the following assumptions were made;
  - Only the cells that have an assigned borough will be processed,
  - Borough's that were not assigned get ignored,
  - More than one Neighborhood can exist in one postal code area, rows will be combined into one row with the Neighborhood separated with a comma,
  - If a cell has a borough but a not assigned Neighborhood, then the Neighborhood will be the same as the borough.

Methodology

Toronto Neighborhood Data

	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

Methodology

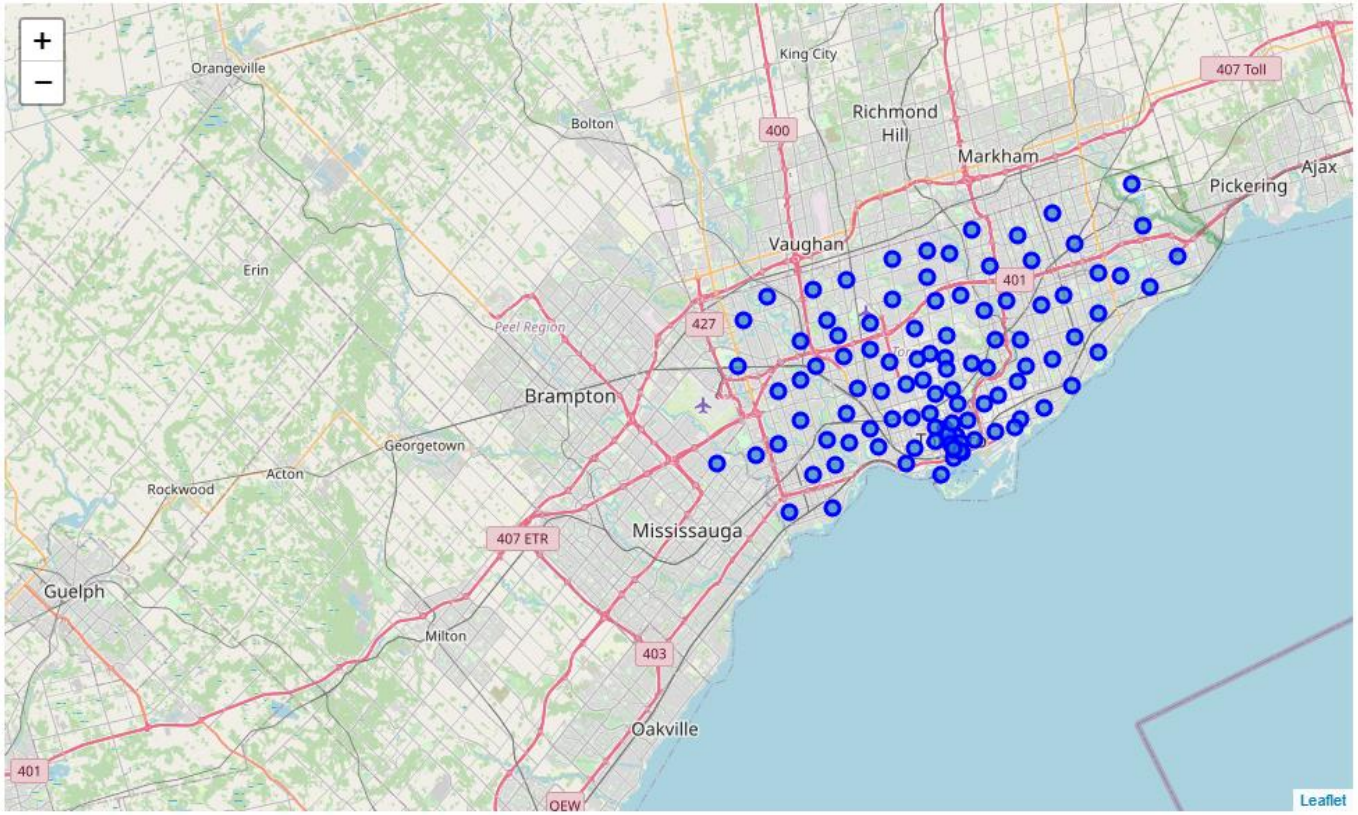
Coordinate data for each Neighborhood in Toronto

The second source of data provided with the Geographical coordinates of the neighborhood with the respective Postal Codes, Merged the two tables together based on Postal Code.

	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

Methodology

Use the *Python Folium* library to visualize geographic details of *Toronto* Neighborhoods



Methodology

Venue Data in Toronto, Canada through Geographical Coordinates

Utilize the *Foursquare* API to explore the Neighborhood venues and segment them.

Merged the *Foursquare* Venue data with the Neighborhood data.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	North York	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	North York	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	North York	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	North York	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Machine Learning

Perform ***One Hot Coding*** on Neighborhood venue Data, Convert Categorical Data into Numerical Data. For each of the Neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each Neighborhood.

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	C
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.04	0.0	0.0	0.0	0.0	

Machine Learning

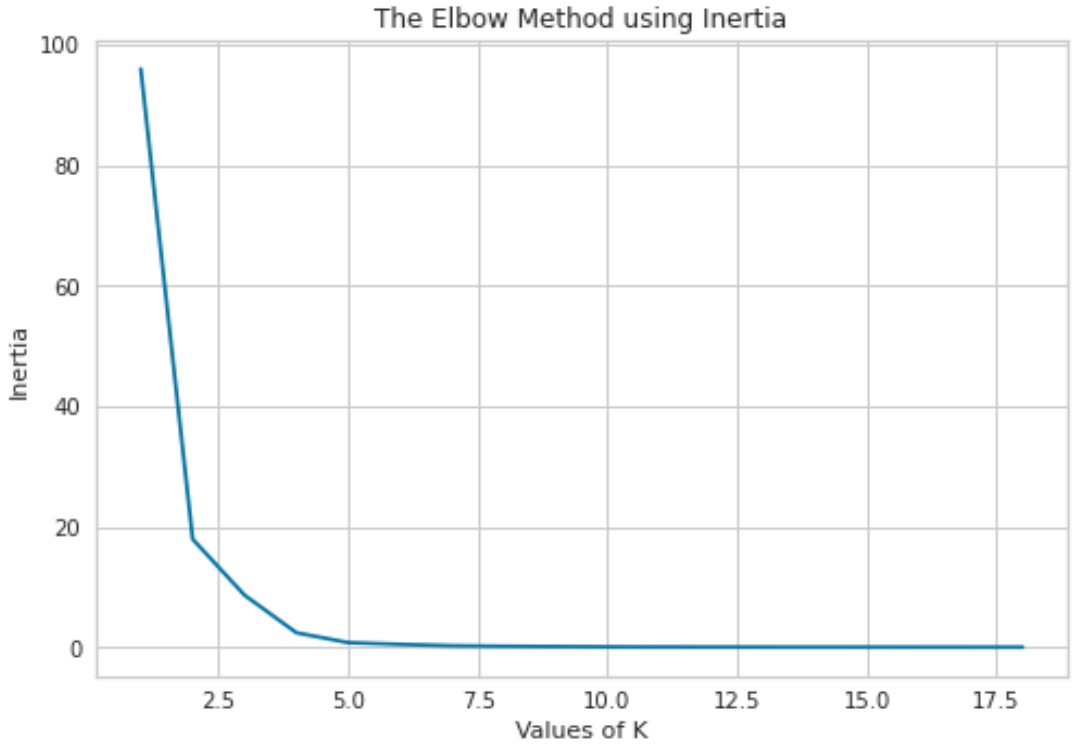
Create a new data frame that only stored the Neighborhood names as well as the mean frequency of GYM/Fitness Centre in that Neighborhood and combine Both GYM & Fitness Centre Venues for better identification and Analysis.

	Neighborhood	GYM_T
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.111111
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.000000

Machine Learning

K-Means Clustering

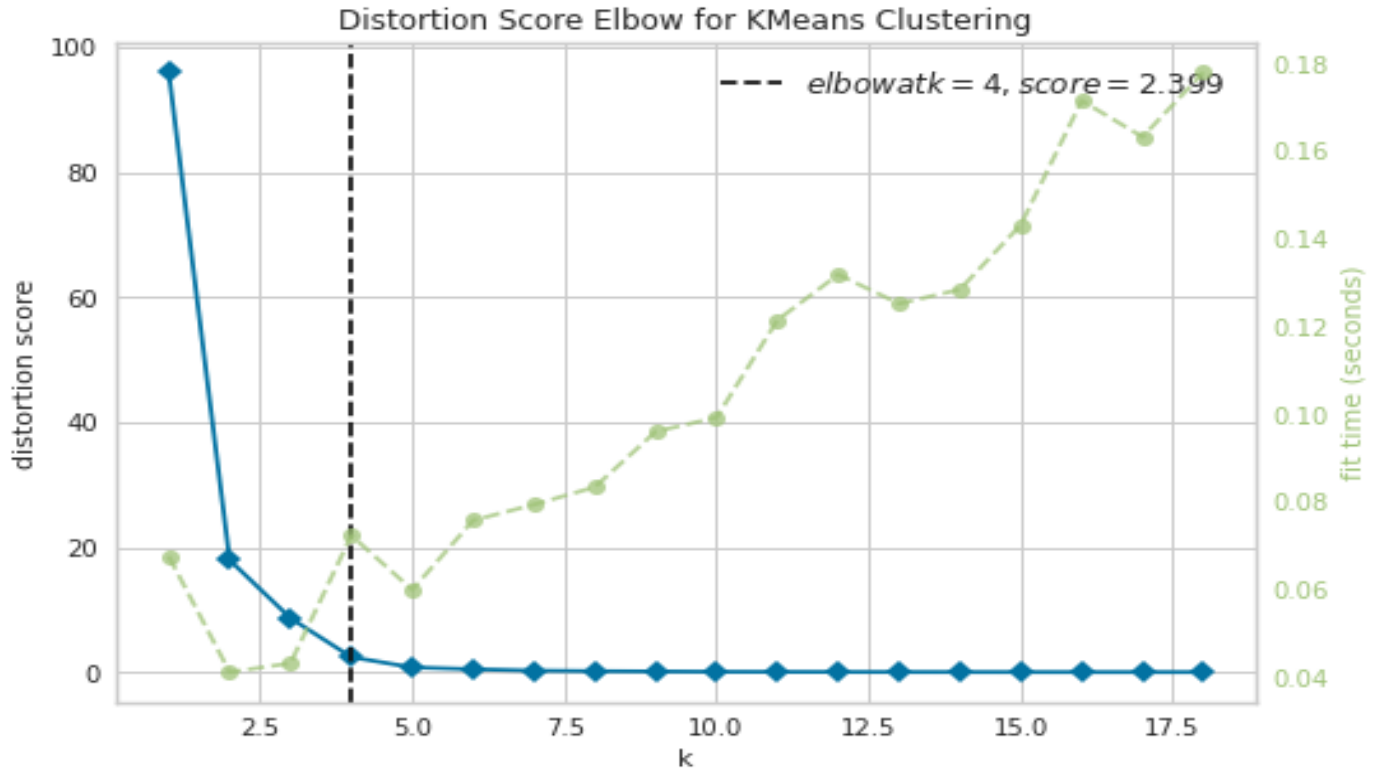
Use *Elbow Point* Technique to get our optimum K value that was neither overfitting nor under fitting the model



Machine Learning

Used a model that accurately pointed out the optimum K value, fit K-Means model above to the *Elbow Visualizer.*

From the dotted line, we see that the Elbow is at K=4. Moreover, in K-Means clustering, objects that are similar based on a certain variable are put into the same cluster

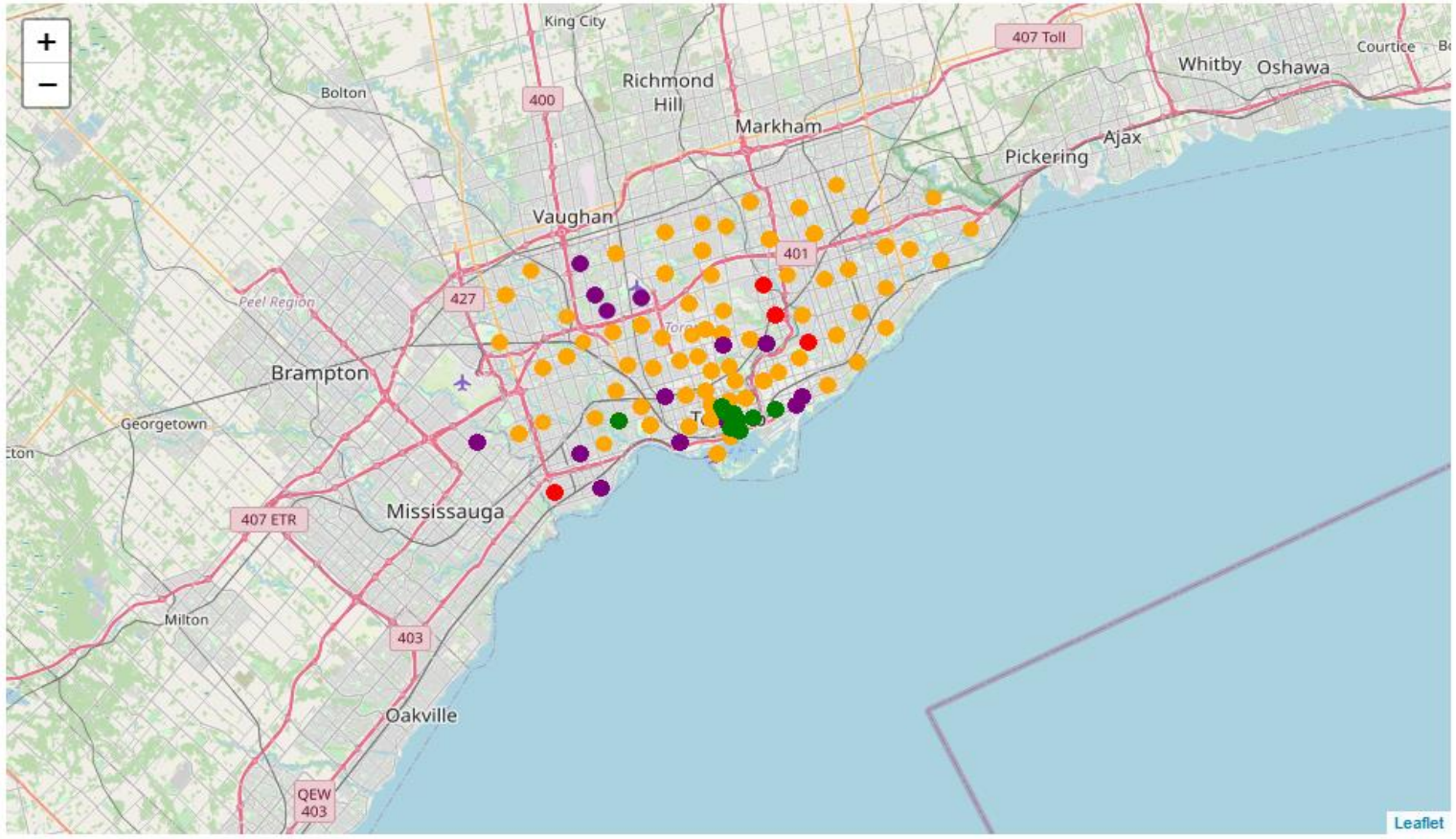


Machine Learning

Clusters in Map

Created a map using the Folium package in Python and each Neighborhood was colored based on the cluster label.

- *Cluster 1 – Orange*
- *Cluster 2 – Purple*
- *Cluster 3 – Red*
- *Cluster 4 – Green*

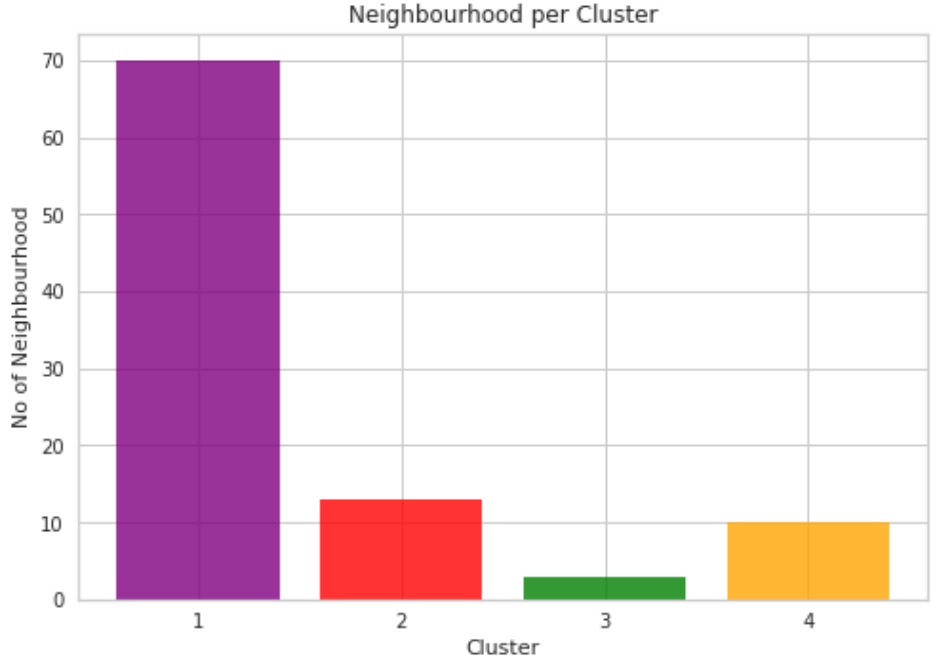


Data Analysis

Total No of Neighborhoods in each cluster and the average GYM/Fitness Centre in that cluster.

we can compare the number of Neighborhoods per Cluster. We see that Cluster 1 has the highest no Neighborhoods 70 while cluster 3 has least 3. Cluster 2 has 13 Neighborhoods and Cluster 4 has 10 Neighborhoods,

Cluster 3 is having highest Mean and Cluster 1 is having least.



Data Analysis

Analysis of each Cluster

Cluster 1

- Cluster-1 had 216 unique Venue Categories,
- Cluster 1 is not having any GYM or Fitness Center, But Cluster-1 is having highest no of Neighborhood in it,
- Cluster-1 having 70 Neighborhoods and 9 Boroughs,
- In that *North York* is having 17 neighborhoods and *Scarborough* having 16 Neighborhoods.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	GYM_T	Cluster Labels
0	North York	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park	0.0	0
2069	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	The Anndore House	43.668801	-79.385413	Hotel	0.0	0
2081	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	Cawthra Square Dog Park	43.666583	-79.380040	Dog Run	0.0	0
58	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	Suzy Shier	43.718846	-79.465906	Clothing Store	0.0	0
2080	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	NC Salon +	43.669406	-79.386748	Health & Beauty Service	0.0	0

Data Analysis

Cluster 2

- Cluster-2 had 13 Neighborhoods and 129 unique Venue Categories across 8 different Boroughs,
- 26 venue location in Cluster-2 is having GYM or Fitness Center, Which is highest among all 4 Clusters,
- *Downtown Toronto* is having highest of 15 GYM or Fitness Center in it,
- Cluster-2 had average GYM or Fitness Center rate as 0.05.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	GYM_T	Cluster Labels
1498	Mississauga	Canada Post Gateway Processing Centre	43.636966	-79.615819	Anoush	43.636769	-79.620840	Mediterranean Restaurant	0.076923	1
1490	Mississauga	Canada Post Gateway Processing Centre	43.636966	-79.615819	Hilton Garden Inn	43.638519	-79.618721	Hotel	0.076923	1
1745	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321	Hex-Mex	43.601261	-79.502284	Mexican Restaurant	0.076923	1
1743	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321	Sense Appeal	43.601729	-79.501063	Café	0.076923	1
1742	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321	Pet Valu	43.602431	-79.498653	Pet Store	0.076923	1

Data Analysis

Cluster 3

- Cluster-3 had 31 unique venue Categories in 3 neighborhoods,
- Had 5 venue location is having GYM or Fitness Centre in it,
- With an average rate of 0.111, which is highest among all 4 clusters,
- Which means Cluster-3 is having highest average rate of GYM or Fitness Center even though it had 3 Neighborhoods.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	GYM_T	Cluster Labels
233	North York	Don Mills	43.725900	-79.340923	Harvey's	43.726603	-79.341035	Restaurant	0.111111	2
218	North York	Don Mills	43.725900	-79.340923	Oomomo	43.726429	-79.343283	Discount Store	0.111111	2
1872	Etobicoke	Alderwood, Long Branch	43.602414	-79.543484	Toronto Gymnastics International	43.599832	-79.542924	Gym	0.111111	2
230	North York	Don Mills	43.725900	-79.340923	Genghis Khan Mongolian Grill	43.726906	-79.341216	Asian Restaurant	0.111111	2
237	North York	Don Mills	43.725900	-79.340923	Barber Greene Square	43.727654	-79.340810	Shopping Mall	0.111111	2

Data Analysis

Cluster 4

- Cluster-4 had 155 unique Venue Categories in that 15 venue location contains GYM or Fitness Center,
- Though Average Rate of GYM in Cluster-4 is 0.02 which is having 2nd highest numbers of GYM in it,
- *Downtown Toronto* is having 13 GYM and *East Toronto* and *West Toronto* contains one GYM or Fitness Center each.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	GYM_T	Cluster Labels
1785	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	Garrison Bespoke	43.648102	-79.376334	Tailor Shop	0.030928	3
1780	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	D.W. Alexander	43.648333	-79.373826	Cocktail Bar	0.030928	3
1778	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	Biff's Bistro	43.647085	-79.376342	French Restaurant	0.030928	3
1777	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	Berczy Park	43.648048	-79.375172	Park	0.030928	3
1776	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	Hockey Hall Of Fame (Hockey Hall of Fame)	43.646974	-79.377323	Museum	0.030928	3

Discussion and Conclusion

- Most of the GYMs are in Cluster-2, Neighborhood located in *Downtown Toronto* is having highest no of GYM in it.
- Neighborhood *Commerce Court* and *First Canadian Place* having 5 GYM these location.
- *East York*, *Mississauga* and *North York* is having less GYM in these locations, each contains 1 GYM.
- After *Downtown Toronto*, *North York* is having 19 Neighborhoods, but less No of GYM venues. So *North York* is suitable for opening New GYM.
- The second-best Borough that have a great opportunity would be *Scarborough*, which is not having any GYM Venues.
- Cluster-1 will be having more no of Neighborhoods in *North York* and *Scarborough* and Cluster-1 is not having GYM venues in it.
- So Cluster-1 Neighborhoods will be the best location to open a New GYM.

Discussion and Conclusion

- Some of the drawbacks of this analysis are
  - The clustering is completely based on data obtained from the Foursquare API.
  - Also, the analysis does not take into consideration of the population across neighborhoods as this can play a huge factor while choosing which place to open a new GYM or Fitness Center.

- Utilized numerous *Python* libraries to fetch the information, control the content and break down and visualize those datasets,
- Utilized *Foursquare* API to investigate the settings in neighborhoods of Toronto, get a great measure of data from Wikipedia which we scraped with the *Beautifulsoup* Web scraping Library,
- We can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Movie Theater and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data science.

References

- Wikipedia Content: https://en.wikipedia.org/wiki/Toronto
- CSV for Coordinate Data: http://cocl.us/Geospatial_data
- Foursquare API
- Statistics Data on GYM:
- https://www.ibisworld.com/canada/market-research-reports/gym-health-fitness-clubs-industry
- https://www.statista.com/outlook/313/108/fitness/canada#market-users
- https://www.cfa.ca/blog/new-year-new-franchise-opportunities-with-franchise-canada-2