IBM Data Science Capstone Project
The Battle of Neighborhoods

Open a new Yoga center in Toronto city

Madhu Pottabathula

Business Problem

In this project a hypothetical business problem of a client from Ontario Canada will be addressed by presenting some viable options resulting from the data analysis of the available data using Machine Learning techniques.

One local businessman who lives in Ontario Canada (client) wants to invest and open a new Yoga center in Toronto city. Client has a concept in his mind and a fair idea about Toronto city setting, but would like to get some advice in choosing a best location from his perspective to open a new Yoga center and therefore needs help.

During an initial meeting, client has discussed the business goal that is in his mind and defined some criterion for selecting a spot for Yoga center like listed below.

Specifics	It should be located in urban setting like within city limits
	Within the 15 minutes walking distance from the city center
	As close as possible to other popular places in the area and it is as crowded as possible where we can expect heavy foot traffic
	There should be less or no competition
	He needs a place where labor is available easily at reasonable salaries.
	In what best location Neighborhood and/ or borough should he open a Yoga center to have the best chance of being successful

Data gathering and Data wrangling

Toronto city Neighborhood Data

- Toronto city data will be scrapped and collected from the following Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
- The information pulled above will be transformed into a pandas data frame

Geographical coordinates data for each Neighborhood in Toronto:

- The following csv file will give us the geographical coordinates (Longitude and Latitude) of each postal code: http://cocl.us/Geospatial_data

FourSquare location data

- FourSquare API will be utilized to obtain all the location data related to neighborhoods and popular venues in the surrounding area along with other details.
- Data will be obtained in Json file format and will be parsed using Python code to convert it into datasets.

Data Cleaning

- The data gathered will be cleaned to remove the duplicate data
- Data records with incomplete data will be discorded
- Some empty data will be replaced with other known values or common data values
- Only required data that is identified will be used for data analysis

Process and Methodology

Below is the step by step process followed to perform the data analysis and all other Machine learning techniques used to arrive at a solution to the defined business problem addressing all the client requirements.

- Data collection and wrangling to create required data sets for the data analysis (data sources and details discussed in the data section above)
- Utilize the Foursquare API to fetch location data such as venues in a given point, popularity, ratings, tips etc.
- Use One-hot encoding to find out most common venues and group the rows by neighborhood and by taking the mean of the frequency of occurrence of each category
- Execute K-means clustering algorithm to segment neighborhoods into required number of clusters
- Visualizing neighborhood clusters using Folium library
- Perform Data Analysis and derive insights from the results to arrive at a solution to the given Business Problem
- Finally recommend a viable solution to the client

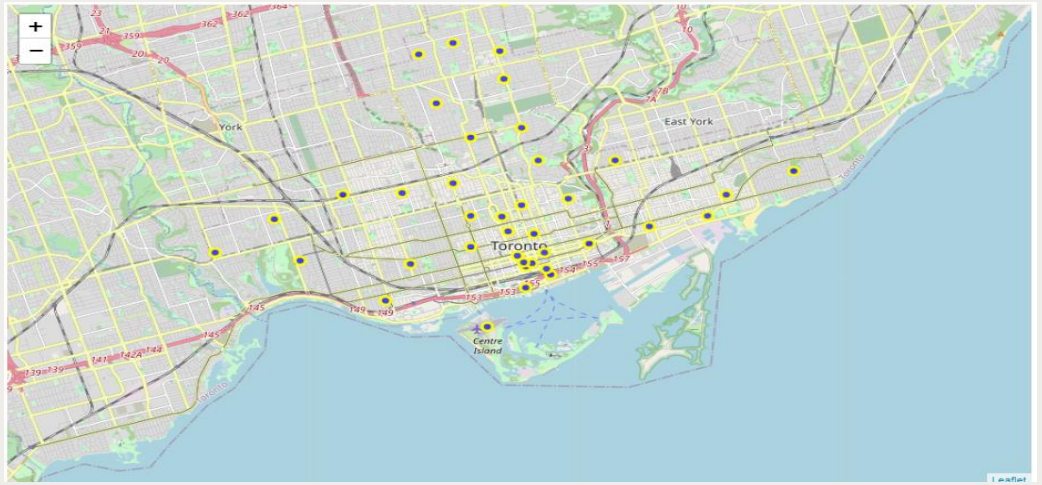
Results – Few Samples (Refer to Notebook for complete results)

Initial Data Frame for data analysis

Out[10]:

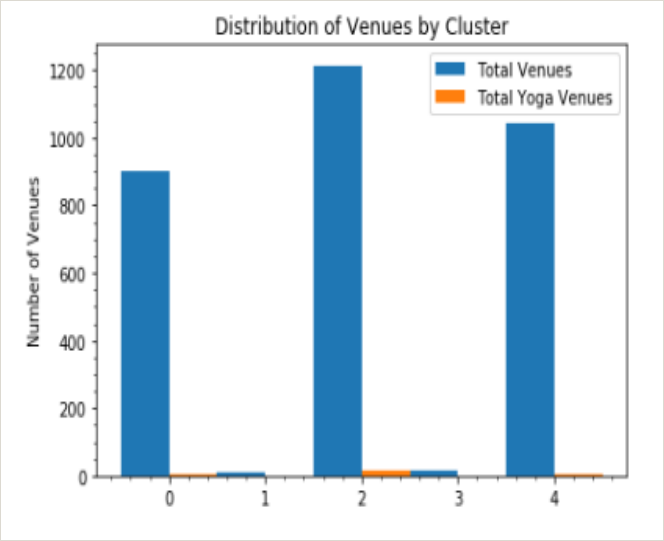
	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park / Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor / Lawrence Heights	43.718518	-79.464763
4	M7A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

Neighborhood locations represented on Toronto city map

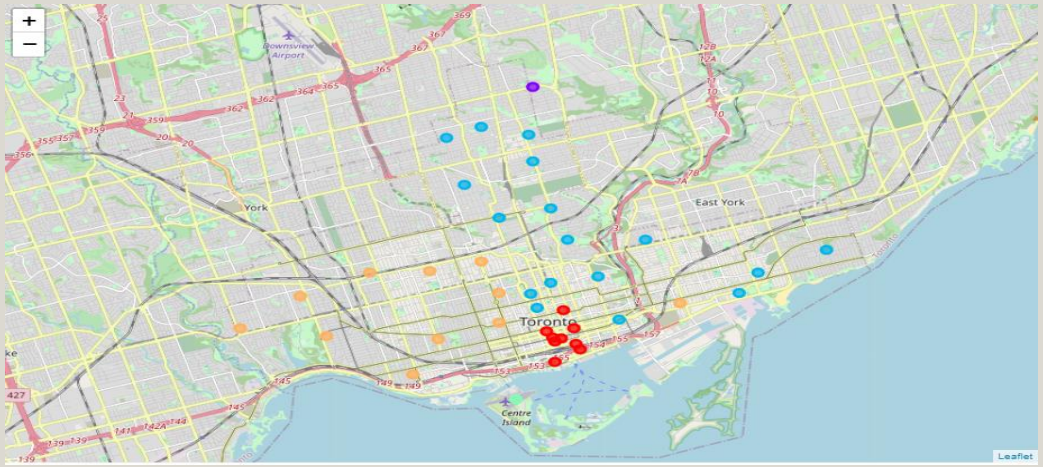


Data Frame constructed for best location determination for new Yoga Center

	Cluster Labels	Total Venues	Total Yoga Venues
0	0	900	3
1	1	8	0
2	2	1214	16
3	3	15	0
4	4	1042	7



Clusters represented after performing K-Means analysis



Recommendations

Insights

- Clusters 0, 2 and 4 are appear to be very busy areas having many existing venues
- Clusters 0, 2 and 4 are more or less equally popular based on the number of Venues that each of these Clusters has
- Clusters 1 and 3 appear to be less popular considering the number of Venues each has, we can eliminate these two clusters from consideration
- clusters 2 and 4 have more number of Yoga centers established already and the cluster 0 has less numbers of Yoga centers
- Competition in Cluster number 0 is less compared to other two Clusters 2 and 4

Recommendations

After careful consideration of the data and the insights after data analysis, Cluster 0 found to be the suitable area for establishing a new Yoga center in Toronto city being the popular area with less competition. Finally the Client has been advised and recommended that a new Yoga Center can be established in any of the following 9 Neighborhoods in the Downtown Borough:

- Garden District, Ryerson
- St. James Town
- Berczy Park
- Richmond / Adelaide / King
- Harbourfront East / Union Station / Toronto Islands
- Toronto Dominion Centre / Design Exchange
- Commerce Court / Victoria Hotel
- Stn A PO Boxes
- First Canadian Place / Underground city