IBM Data Science Capstone Project - The Battle of Neighborhoods

Open a new Yoga center in Toronto city

Madhu Pottabathula

1 Introduction

The intent of this project is to help a small scale business owner who is planning to invest and establish a new Yoga center in the city of Toronto. Toronto, being the most popular city in Canada and getting ranked as an important global city continuously based on a high quality of living, is the prime choice for many investors to start or expand their business. However, with limited knowledge of the Toronto market, the client has approached to get assistance in the selection of a best location to open a new Yoga center. The efforts will begin with analyzing the neighborhood data of Toronto city, use Machine learning algorithms and perform data analysis to address the business problem of a client

Yoga has grown in popularity over the last few years, with passionate yogis stretching all around the world. Roughly 36% of the world's population practices yoga. It's no secret that the yoga industry is booming. In the last several years, yoga has exploded in the U.S and Canada showing no signs of slowing down. The practice of yoga makes the body strong and flexible, it also improves the functioning of the respiratory, circulatory, digestive, and hormonal systems. Yoga brings about emotional stability and clarity of mind.

2 Business Problem

In this project I would be working on a hypothetical business problem of a client from Ontario Canada and try to solve it by presenting some viable options resulting from the analysis of the available data.

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.

- ✓ 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

3 Data Requirements and preparation

The data requirement section provide the details of various sources of data, different processes to obtain the data, what kind of data cleaning would be performed to present a clean data for data analysis

3.1 Toronto city Neighborhood Data:

a. Toronto city data including the neighborhood places, borough details and postal codes etc. will be scrapped and collected from the following Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
b. The information pulled above will be transformed into a pandas data frame and will be used for further analysis after data clean-up.

3.2 Geographical coordinates data (Longitude and Latitude) for each Neighborhood in Toronto:

a. The following csv file will give us the geographical coordinates of each postal code: http://cocl.us/Geospatial_data

3.3 FourSquare location data

a. FourSquare API will be utilized (via the Request library in Python) to obtain all the data related to locations, popular venues in the surrounding area, category, ratings of the venues and other necessary details.

b. Data will be obtained in Json file format and will be parsed using Python code to convert it into datasets.

c. Data will be sorted based on rankings

d. Finally, the data be will be visually assessed using graphing from various Python libraries.

3.4 Data Cleaning

a. The data gathered will be cleaned to remove the duplicate data

b. Data records with incomplete data will be discorded

c. Some empty data will be replaced with other known values or common data values

d. Only required data that is identified will be used for data analysis

4 Methodology used.

The methodology section explains 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, the steps will include:

4.1 Data collection and wrangling to create required data sets for the data analysis (data sources and details discussed in the data section above)

The effort begins with scrapping the Wikipedia web page of Toronto city containing information about boroughs and neighborhoods present in each borough using the library BeautifulSoup. The raw data will be parsed into a data frame using Pandas methods and then geographical coordinates (latitude and longitude) values for each neighborhood would be appended to the data frame to construct a base data set. Further this data frame will be cleaned by removing the duplicates, replace the null values with appropriate values, dropping the unnecessary columns etc. and a final cleaned data frame will be presented for the data analysis in the subsequent steps.

4.2 Utilize the Foursquare API to fetch location data such as venues in a given point, popularity, ratings, tips etc.

Foursquare API will be used to explore the venues data in each neighborhood area. A Foursquare developer account has been created and obtained the account ID and API key to pull the data from Foursuare. Foursquare API request is configured to return 100 venues in each neighborhood with a radius of 1000 meters. The data returned will have all the details about venues such as venue name, coordinates, category, popularity ratings, tips etc. and all these data will be fetched into a Pandas data frame for easy accessibility. As per the client's requirement only boroughs within around the Toronto city center are considered, the filter used is the word 'Toronto' in the borough names to filter the required data.

4.3 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

One hot encoding will be performed on the resulting data frame for each neighborhood to represent the categorical variables as binary vectors (0s and 1s) for K-mean analysis and group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. All the data will be captured into a Pandas data frame

4.4 Execute K-means clustering algorithm to segment neighborhoods into required number of clusters

Neighborhoods, cities and venues are unlabeled data and they need to be categorized based on their features. Hence under the category of unsupervised Machine Learning Algorithms, K-Means Clustering algorithm (from the Scikit-learn Machine Learning library) would be best fit to cluster those venues based on popularity. K-means clustering will be performed on the data frame and the will be fit to get the venues segmented into 5 clusters. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. This information will be used to check the pattern for each neighborhood and get the information about the top ten common restaurants for each neighborhood.

4.5 Visualizing neighborhood clusters using Folium library

Folium library functions are used to visualize the actual map of Toronto city to visually look at ow the neighborhoods are spotted, and see how the cluster centroids are drawn with color coding to get a better feeling about the area.

4.6 Perform Data Analysis and derive insights from the results to arrive at a solution to the given Business Problem

Performed analysis on the data results which are the output from Python code components as explained in Results section below. Neighborhoods and Venues data is analyzed to figure out the most popular venues, number of venues in each neighborhood, number of venues segmented under each cluster after executing the K-means cluster logic, number of Yoga centers already existing in each cluster etc.

4.7 Finally recommend a viable solution to the client

As a result from data analysis performed in above step, several insights and observations noted as listed in the Discussion section below. The logic behind the neighborhood selection for establishing a new Yoga center is first to find out most popular areas in the selected boroughs based on the numbers of venues in each cluster. Secondly find out existing Yoga centers in each cluster to assess the competition. Once we have the information of total number of venues and existing Yoga centers in each cluster, do some analysis to pick a cluster having a good number of venues and less number of existing yoga centers so the area is equally popular as well as less competitive for opening a new Yoga center. Finally pull the list of neighborhoods falling under the

selected cluster and provide the list to client as an educated recommendation to establish a new Yoga center in Toronto city.

5 Results.

The results section provide the complete output from various python code components, information returned from Foresquare API and data frames constructed after each step of the process etc.

5.1 Created an initial Dataset after scraping the data from Toronto city Wikipedia website and required data cleaning done along with geographical coordinates of the each neighborhood added to utilize as base data.

In [10]:
```
# Load coordinate data from CSV file
Toronto_neighborhood = Toronto_neighborhood.rename({'Postal code':'Postal Code'}, axis=1)
lltemp_df = pd.read_csv('http://cocl.us/Geospatial_data')
geome_df = lltemp_df.rename({'Postal code':'PostalCode'}, axis=1)

# Merge coordinates into neighbourhood
Toronto_neighborhood = Toronto_neighborhood.merge(geome_df)
Toronto_neighborhood.shape
Toronto_neighborhood.head()
```

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

5.2 Client request was to focus only within Toronto City to explore a best location to open a new Yoga center. Considering only the Boroughs in the Toronto City (assumption: take only the Boroughs with word Toronto)

In [12]:
```
#Focusing only on Boroughs that contain the word Toronto
toronto_dfone = Toronto_neighborhood[Toronto_neighborhood['Borough'].str.contains('Toronto')].reset_index(drop=True)
toronto_dfone
```

Out[12]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park / Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031
5	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306

5.3 Create map of Toronto city using geographical latitude and longitude values to represent the Neighborhood locations with in the selected 4 Boroughs

Obtain the Geographical Coordinates using Geo Locator - latitude and longitude values of Toronto city

In [14]:
```
from geopy.geocoders import Nominatim
address = 'Toronto, ON'

geolocator = Nominatim(user_agent="toronto_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

Create map of Toronto city using latitude and longitude values obtained above

In [53]:
```
# create map of Toronto using latitude and longitude values
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=12)

for lat, lng, borough, neighborhood in zip(toronto_dfone['Latitude'], toronto_dfone['Longitude'], toronto_dfone['Borough'], toro
nto_dfone['Neighborhood']):
    label = '{}, {}'.format(toronto_dfone, borough)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='yellow',
        fill=True,
        fill_color='blue',
        fill_opacity=0.7,
        parse_html=False).add_to(map_toronto)

map_toronto
```

Out[53]:

5.4 Use the Foursquare API to get location data to explore the neighborhoods

In [21]:
```
CLIENT_ID = 'LQWEPZR310MZTHC1CKEPP2EPHM2FFONLHK1KWB1GSDN4ZVTS'
CLIENT_SECRET = 'YRZX0HIR31A5M3JTQPDD4GNI2S1ONDSUR1RUEGKW4UGFBMMK'
VERSION = '20180605'
LIMIT = 100
print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentails:
CLIENT_ID: LQWEPZR310MZTHC1CKEPP2EPHM2FFONLHK1KWB1GSDN4ZVTS
CLIENT_SECRET:YRZX0HIR31A5M3JTQPDD4GNI2S1ONDSUR1RUEGKW4UGFBMMK
```

5.5 Fetching all the venues in each neighborhood in the selected Boroughs (all Boroughs with word Toronto in the name) into a new data frame called toronto_venues to perform the data analysis

In [22]:
```
toronto_venues = getNearbyVenues(names=toronto_dfone['Neighborhood'],
                                 latitudes=toronto_dfone['Latitude'],
                                 longitudes=toronto_dfone['Longitude']
                                )
```

```
Regent Park / Harbourfront
Queen's Park / Ontario Provincial Government
Garden District, Ryerson
St. James Town
The Beaches
Berczy Park
Central Bay Street
Christie
Richmond / Adelaide / King
Dufferin / Dovercourt Village
Harbourfront East / Union Station / Toronto Islands
Little Portugal / Trinity
The Danforth West / Riverdale
Toronto Dominion Centre / Design Exchange
Brockton / Parkdale Village / Exhibition Place
India Bazaar / The Beaches West
Commerce Court / Victoria Hotel
Studio District
Lawrence Park
Roselawn
Davisville North
Forest Hill North & West
High Park / The Junction South
North Toronto West
The Annex / North Midtown / Yorkville
Parkdale / Roncesvalles
Davisville
University of Toronto / Harbord
Runnymede / Swansea
Moore Park / Summerhill East
Kensington Market / Chinatown / Grange Park
Summerhill West / Rathnelly / South Hill / Forest Hill SE / Deer Park
CN Tower / King and Spadina / Railway Lands / Harbourfront West / Bathurst Quay / South Niagara / Island airport
Rosedale
Stn A PO Boxes
St. James Town / Cabbagetown
First Canadian Place / Underground city
Church and Wellesley
Business reply mail Processing CentrE
```

5.6 Cluster the Neighborhoods using K-Means Algorithm into 5 Clusters

In [30]:
```
# set number of clusters
kclusters = 5

toronto_grouped_clustering = torontoarea_grouped.drop('Neighborhood', 1)
#toronto_grouped_clustering.head()
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

Out[30]:
```
array([0, 4, 2, 3, 2, 4, 2, 0, 2, 2], dtype=int32)
```

5.7 Create Toronto City map to visualize the resulting clusters

Create Toronto City map to visualize the resulting clusters

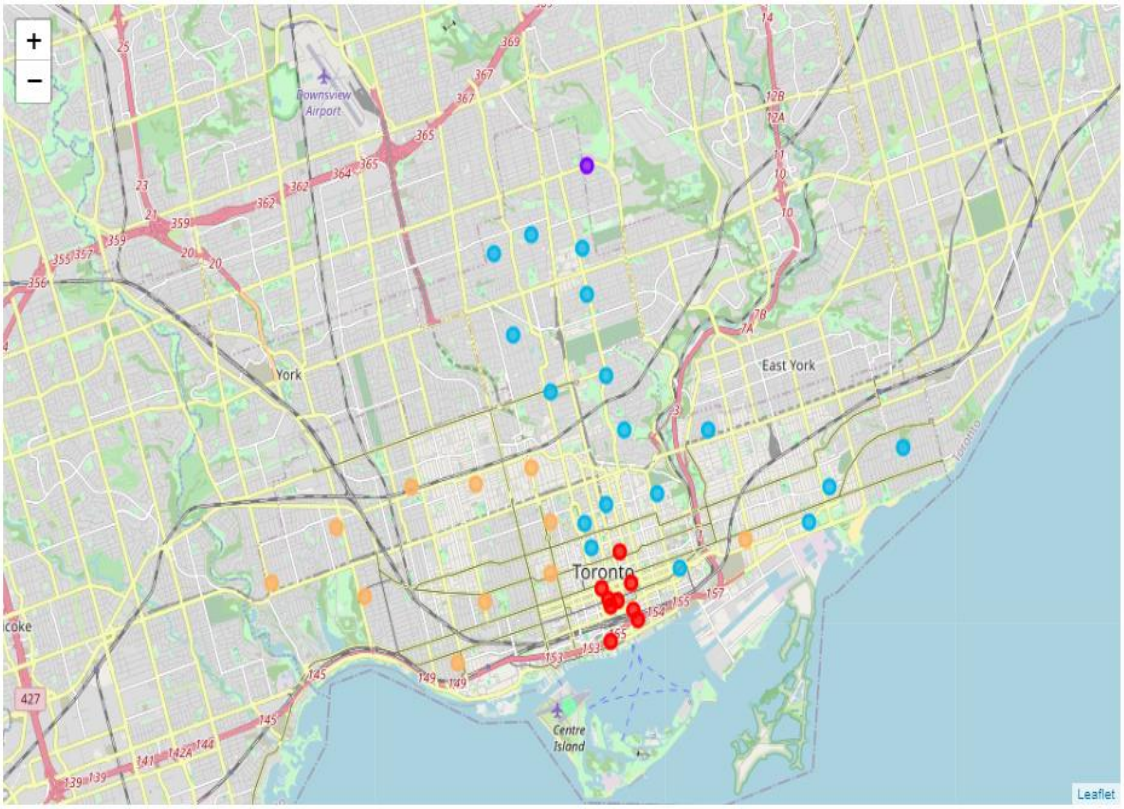
In [36]:
```
# create map
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=12)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(torontoarea_merged['Latitude'], torontoarea_merged['Longitude'], torontoarea_merged['Neighborh
ood'], torontoarea_merged['Cluster Labels']):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

Out[36]:

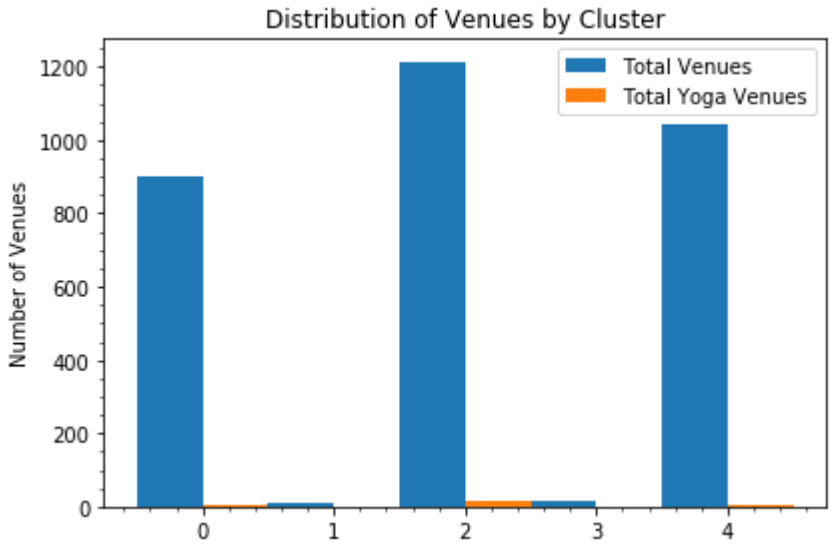


5.8 The total number Yoga centers verses total number of venues in each cluster

Out[45]:

	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

5.9 Visually look at the Venues and Yoga centers distribution across all the Clusters using bar charts



6 Discussion.

In the discussion section all the insights and observations based on the results will be explained and the method who the recommendation is arrived will be discussed as well. Finally the results and recommendation will be submitted to client

6.1 Few Insights noted based on the analysis as listed below

- 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

6.2 Discussion and final Recommendation

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 Neighborhoods in the Downtown Borough:

1. Garden District, Ryerson
2. St. James Town
3. Berczy Park
4. Richmond / Adelaide / King
5. Harbourfront East / Union Station / Toronto Islands
6. Toronto Dominion Centre / Design Exchange
7. Commerce Court / Victoria Hotel
8. Stn A PO Boxes
9. First Canadian Place / Underground city

7 Conclusion

During the course of the project, have gone through the process of identifying the business problem, specifying the data requirements and sources, extracting and preparing the data, perform a bit of the machine learning by utilizing k-means clustering, analysis the data results and draw insights and finally providing a viable recommendation to the stakeholders. This analysis was performed on limited data and basic operations were performed. This can be taken as a base version and can be scaled up to apply for a bigger business problems collecting more data and obtain the better results.

Here is the link to the Notebook with all the Python code work done and other project files published on the Github. Thank you for your interest and time.