**IBM Data Science Certification**

**Capstone Project Report**

**Recommend Toronto Neighborhood to open a Yoga Studio**

**Shailendra Dungarwal**

**Dec 24, 2020**

1. **Introduction**
   1. **Background**

Toronto is the most populous city of Canda and is recognized as one of the most multicultural and cosmopolitan cities in the world. The current metro area population of Toronto in 2020 is around 6 million and trend is getting stronger with many people moving to Toronto because of immigration policies and opportunities.

With recent pandemic, there is increased awareness and focus on individual health. It is not just about physical health but overall wellbeing including mental health. Regular Yoga practice creates mental clarity and calmness, increases body awareness, relieves chronic stress patterns, relaxes the mind and sharpens concentration. Yoga Studio is the best way to connect yoga teachers and students.

This increased awareness about Yoga makes a business case for opening a Yoga Studio. However, Yoga Studio has become so popular that such studios are everywhere. Starting a new Yoga Studio is great opportunity, however we need to find a suitable location for long term sustainability of the business. This project explores best locations to open a new Yoga Studio in Toronto, Canada.

* 1. **Business Problem**

Taking care of physical and mental health is a big part of the healthy lifestyle, however, due to heavy competition, the location of the Yoga Studio can make or break the business. The purpose of the project is to explore the most suitable neighborhood for to open a new Yoga Studio in Toronto, Canada. The project plans to use data science tools, methodology and machine learning algorithms to recommend most appropriate neighborhoods for opening a new Yoga Studio.

* 1. **Target Audience**

The entrepreneur who wants answer to question – Which Toronto neighborhood I should consider to open a Yoga Studio.

1. **Data and Packages/Resources**

**2.1 Data**

To perform my analysis, I need geographical location data for Toronto as well as Yoga Studio data in Toronto

1. List of Toronto neighborhoods:

Data Source: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

Description: Above page has list of the names of the different neighborhoods of Toronto that will be scrapped from above Wikipedia page. I might need to append this data set with additional information about coordinates using the Nominatim module of the Python Geopy library.

1. Yoga Studios in Toronto:

Data Source: Foursquare API

Description: By using this API we will get all the venues in the Toronto area. We can further filter data to get Yoga Studios only.

**2.2 Packages**

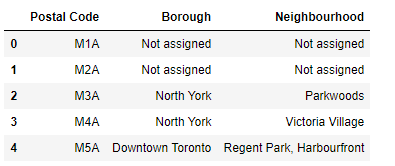
* BeautifulSoup – a python package to scrap a webpage
* Nominatim – Module from python Geopy library to get additional geographical data (longitude and latitude) of Toronto neighborhoods
* Folium - for visualization
* Foursquare API - to search various venues
* KMeans module from sklearn.cluster library – for clustering algorithm

1. **Methodology**
   1. **Step 1 – Get Canada neighborhood and Postal Codes**

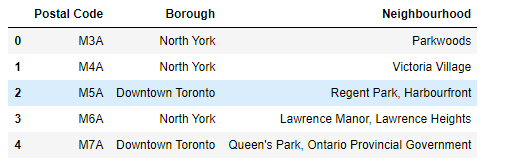
As a starting point, I need to get list of all neighborhoods of Toronto, Canada. The data extracted by scrapping from following Wikipedia page using Python BeautifulSoup package -

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

A Pandas data frame was created from scrapped data. First 5 rows are shown below -



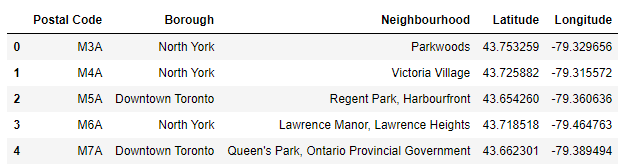
Then, I performed a series of data cleaning steps (e.g. ignore cells with not assigned boroughs, combine cells with same postal code) to prepare data for further analysis.



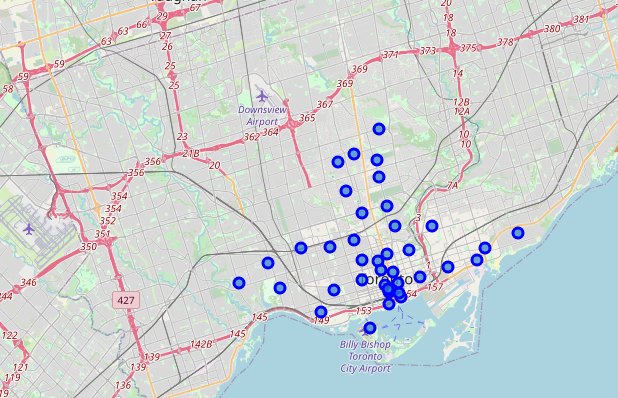
* 1. **Step 2 – Append Latitude and Longitude to neighborhood data**

Wikipedia link above only provided list of neighborhood names and postal code. To utilize Foursquare API, I need to append the data frame with latitude and longitude information.

I used CSV (link - <https://cocl.us/Geospatial_data> ) provided in earlier IBM certification modules to get latitude and longitude for Toronto Postal Codes.



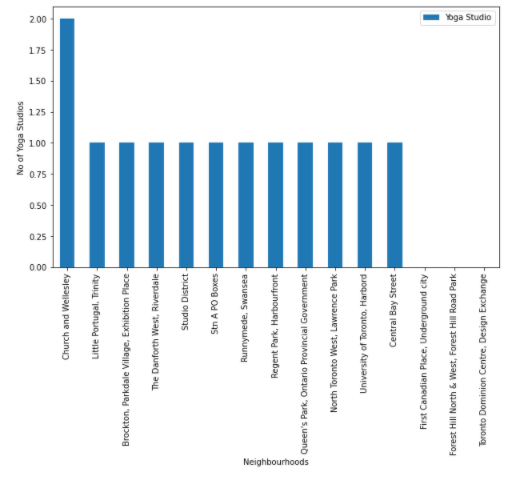
After having all relevant geographical information, I used Folium package to generate maps to visualize neighborhoods.



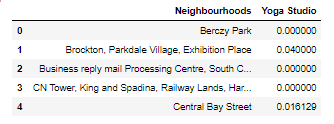
* 1. **Step 3 – Get Venue details of Toronto**

I created a Foursquare developer account to get my client id and client secret. Foursquare API search is used to find the Toronto Venues. Here I’ve chosen 150 popular spots for each neighborhood with a radius of 500 meters. API provided names, categories, latitude and longitude of the venues. There are 236 unique categories in which Yoga Studio is one of them. I also ensured that ‘Yoga Studio’ is one of the venue category.

I visualize max number of Yoga Studio and here is what it looks like below. In the image below, Church and Wellesley neighborhood has highest number of Yoga Studios.

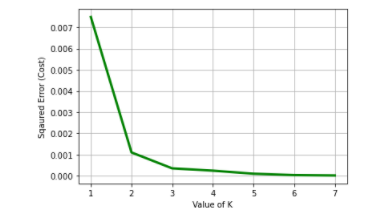


Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. I made a justification to specifically look for “Yoga Studio”.



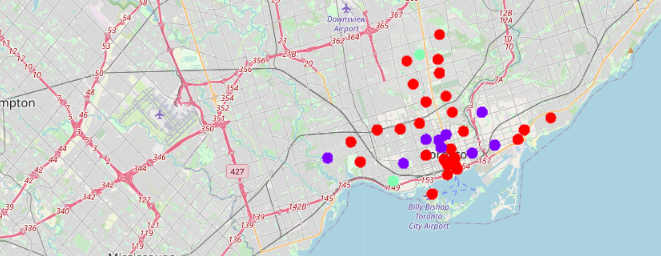
* 1. **Step 4 – Clustering using K-Means**

We will extract Yoga Studio data from the above table and fit this into the code for finding the best value of K.



From the above image, we see that the best value of K will be 3 according to the Elbow method.

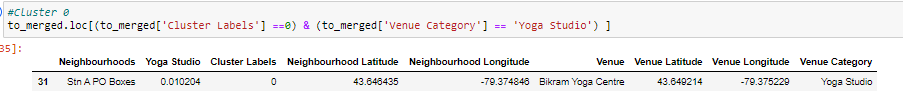
Then, I performed the clustering algorithm by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for Yoga Studios. The results were visualized in below map.



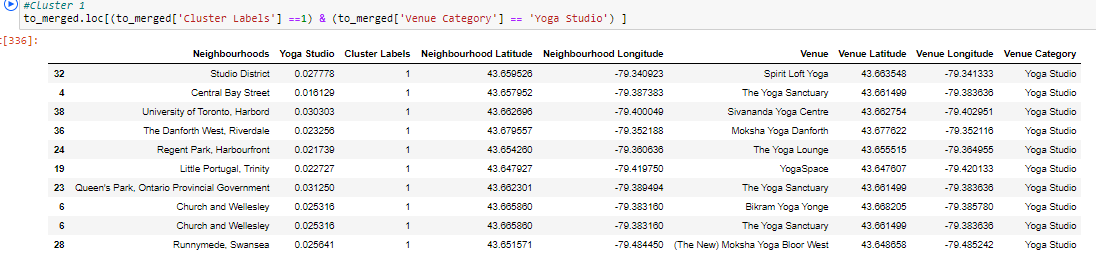
# ****Results****

K-Means clustering demonstrate that Toronto neighborhoods can be categorized into 3 clusters based on how many Yoga Studios are in each neighborhood:

* **Cluster 0** has Toronto neighborhoods with the lesser number of Yoga Studios (Red color)



* **Cluster 1** has Toronto neighborhoods with medium dense Yoga Studios (Indigo color)



* **Cluster 2** has Toronto neighborhoods with highly dense Yoga Studios (Green color)

# 

# ****Results Summary –****

* Church and Wellesley neighborhood has highest number of Yoga Studios.
* Cluster 0 neighborhoods have the least number of Yoga Studios.
* I will recommend opening Yoga Studio in Central Bay Street or Stn A PO Boxes as both these neighborhoods have least number of yoga studios currently.

# ****Discussion****

According to the analysis, Central Bay Street or Stn A PO Boxes neighborhoods will provide the least competition for an upcoming Yoga Studio. Church and Wellesley neighborhood has highest number of Yoga Studios so analysis does not recommend opening a new Yoga studio there. Also, analysis is not recommending Brookton, Parkdale Village, Exhibition Place, North Toronto West or Lawrence Park locations because of already dense presence of Yoga Studios.

Here are some of the drawback of the analysis performed - clustering is completely based on the data provided by Foursquare API. Since land price/rental price, the number of potential customers and median income of the neighborhood can play a major role so this analysis is definitely not conclusive. However, it provides a preliminary recommendation for locations where one can open a new Yoga Studio in Toronto.

# ****Conclusion****

# With this project, I got exposure to how a real-life Data science project would looks like. I have used common Python libraries, visualize data by plotting graphs and exploratory data analysis. Foursquare API provided important information of Toronto venues. The analysis results depends on the accuracy of FourSquare data.

# As a final conclusion, this project recommends the entrepreneur to open a Yoga Studio in Central Bay Street or Stn A PO Boxes neighborhoods of Toronto.