# Capstone Project - The Battle of Neighborhoods

**Coursera IBM Data Science Certification**

**POORNA MOHANDAS**

July 2020

# Report Contents

1. Introduction
2. Data
3. Methodology
4. Result
5. Conclusion

# INTRODUCTION

This is a capstone project for IBM certification course on Data Science. In this project, I am using a hypothetical scenario where a yoga instructor currently residing in Canada wants to find the best place for starting a yoga studio.

In the middle of hustle and bustle of the modern life, emotional stability declines day by day. Yoga is the perfect solution to refurbish attitude towards life by maintaining a positive physical and mental health. With this social purpose in mind, it is also important to find the right location in order to reduce the competition as the yoga instructor is an entrepreneur as well. Hence I am designing this project to help this person to find a best location to start a yoga studio with data analysis skills acquired through the entire span of courses.

## BUSINESS PROBLEM

The objective of this capstone project is to find the most suitable location for the entrepreneur to start a new Yoga Studio in Toronto, Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open a Yoga Studio, where should they consider opening it?

## TARGET AUDIENCE

The entrepreneurs who wants to find the location to open an Exclusive Yoga Studio.

# DATA

To solve this problem, need below data:

- List of neighborhoods in Toronto, Canada
- Latitude and Longitude of these neighborhoods
- Venue data related to existing Yoga Studios. This will help us find neighborhoods that are more suitable to open a Yoga Studio.

### EXTRACTING THE DATA

- The scrapping of Toronto neighborhoods via Wikipedia.
- Getting Latitude and Longitude data of these neighborhoods using suitable method (via Geocoder package or csv method).
- Using Foursquare API to get venue data details related to these neighborhoods.

# METHODOLOGY

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia:https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. I did the web scraping by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a webpage into the data frame.

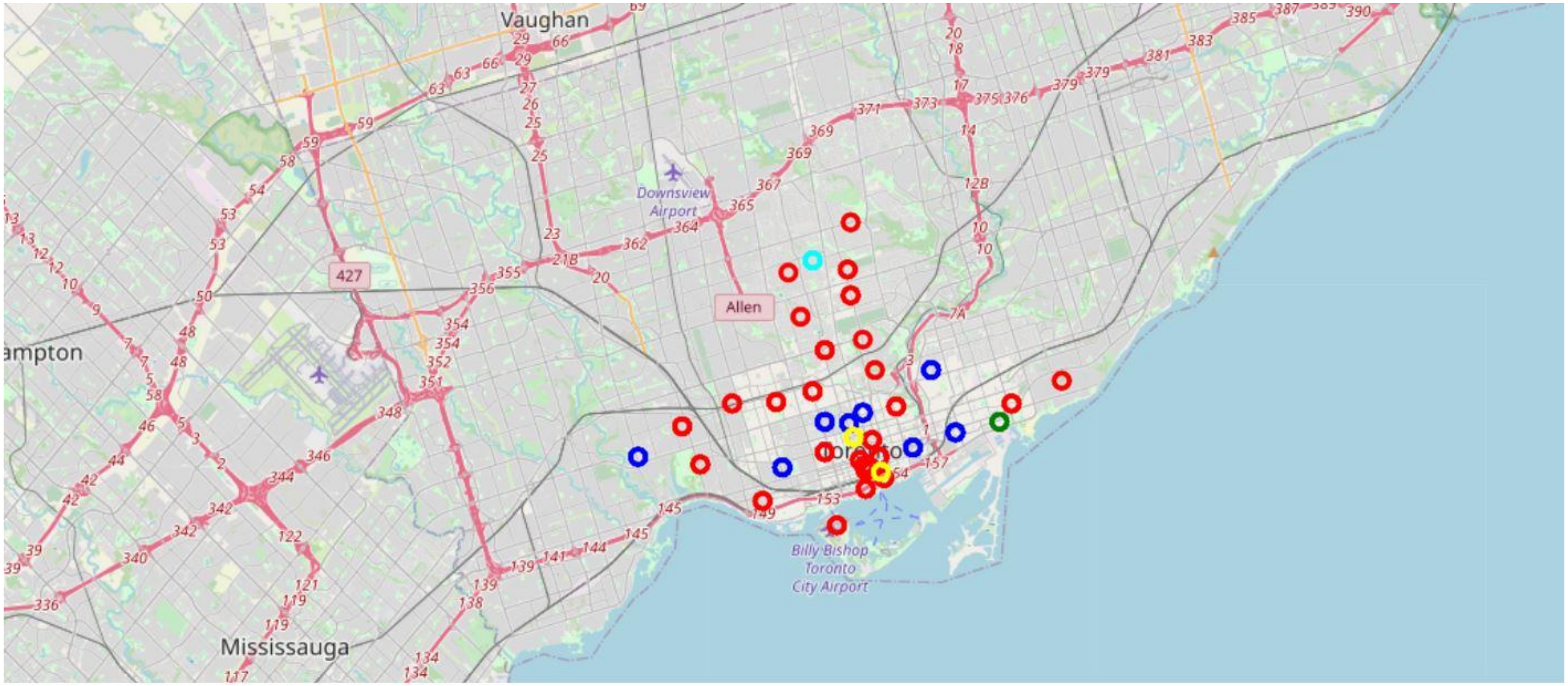
However, it is only a list of neighborhood names and postal codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates. I have used the CSV file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering these coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have Foursquare developer account and hence client ID and client key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for “Yoga Studio”. Lastly, I performed the clustering method 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. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 5 clusters based on their frequency of occurrence for “Yoga Studio”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the studio.

# RESULT

## CLUSTERS

The results from k-means clustering show that we can categorize Toronto neighborhoods into 5 clusters based on how many Yoga Studio in each neighborhood:



- Cluster 0: Neighborhoods with good no Yoga Studio.

- Cluster 1: Neighborhoods with more number of Yoga Studios.
- Cluster 2: Neighborhoods with less number of Yoga Studios.
- Cluster 3: Neighborhoods with less number of Yoga Studios.
- Cluster 4: Neighborhoods with less number of Yoga Studios.

The results are visualized in the above map with Cluster 0, Cluster 2, Cluster 3, Cluster 4.

# CONCLUSION

From the analysis found that no yoga studios are present in cluster 0. Also fewer are present in Clusters 3 and 4. More number of yoga studios are present in cluster 1.

Hence places such as Berczy Park, Brockton, Parkdale Village, Exhibition Place, Harbourfront East, Union Station, Toronto Islands (cluster 0) can be considered as the best places to start the studio. It is also ideal to start in places such as North Toronto West, Lawrence Park, Stn A PO Boxes, Business reply mail Processing Centre (clusters 2,3 $ 4). It is better to avoid places near to those in cluster 1 such as University of Toronto, Harbord, Church and Wellesley, Little Portugal, Trinity, etc.

Looking at nearby venues it seems cluster 0 might be a good location as there are no Yoga Studio in these areas. Therefore, this project recommends the entrepreneur to open a Yoga Studio in these locations which has minimum studios.