

*Rohith P R – 2nd May 2020*

|  |
| --- |
|  |
| The Battle of Neighborhoods |
|  |
| Opening new Yoga Studio in Toronto |
| *IBM Data Science Professional Certificate – Capstone Project* |

The Battle of Neighborhoods

Opening new Yoga Studio in Toronto

# 1. Introduction

**1.1 Background**

The basis of this study is to help a group of investors planning to open their first Indian based Yoga Studio expansion in Toronto. Being that Toronto is the most populated city in Canada, and continually ranks as an important global city based on a high quality of living, the choice to expand into the neighbor of the north market was an easy selection for the investing group. However, with limited knowledge of the Toronto market, the group of investors have selected me to assist in the selection of which areas of Toronto will facilitate a launch of their business.

They are more interested in locations that are densely populated and nothing else, because:

* There are no age limits for doing Yoga
* There are no education requirements
* There are no income requirements as their packages are very affordable

With the requirements given by the investing group, based on previous success in other markets, the objective is to locate and recommend to the investors, the target audience, which neighborhood(s) of Toronto will be the best choice to start their international growth plan. The information gained will assist in choosing the right location by providing data about the population of each neighborhood, in addition to other established venues present in these areas.

Additionally, this information could be of interest to other potential investors looking to open a new restaurant or entertainment venue in Toronto.

**1.2. Problem**

In this project, I will investigate that if there is any densely populated location in Toronto for opening a Yoga Studio and the business should also be far away from the existing Yoga Studios.

**1.3. Target Audience**

The increasing number of health-conscious North Americans over the last 20 years has helped drive the growth of the [yoga](http://karmayoga.ca/index.php/2016/08/27/what-is-yoga/)industry. The popularity of this market is not just a short-term boom either. Y[oga](http://www.yogajournal.com/category/beginners/)has become much more accessible and appeals to a wide variety of demographics for a variety of reasons. The market is expected to reach [$14.03 billion in 2020](http://www.technavio.com/blog/yoga-upswing-north-america).

Yoga's gentle movements are a big reason for why it is so popular. Yoga is good for people who have not been active in a while. It is good for people who have certain health conditions like arthritis or osteoporosis. But yoga is also great if you are already fit and want a challenging workout.

The general target audience for Yoga Studios are customers who prioritize health and well-being as one of their goals; and are in the age group of 19 to 65 (but not limited to only this age group); with an average income of USD 45,000 per annum and not afraid in expressing themselves in fashionable clothing designed to be worn for exercising and for general use. As it stands right now, the median household **income in Toronto**, based on a living **wage** of roughly $19 per hour, is $72,830 CAD.

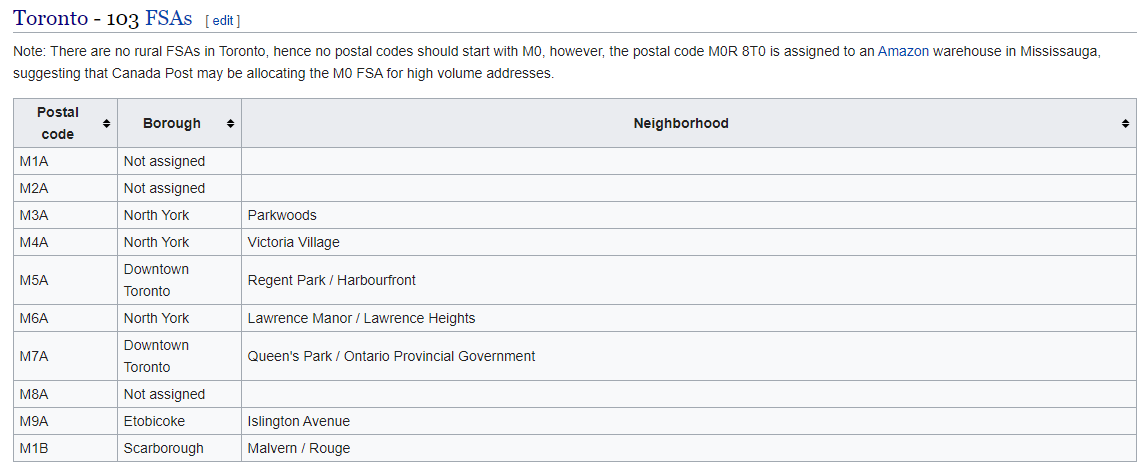
# 2. Data

* 1. **Data sources**

The information needed to the solve investors problem comes from the following sources:

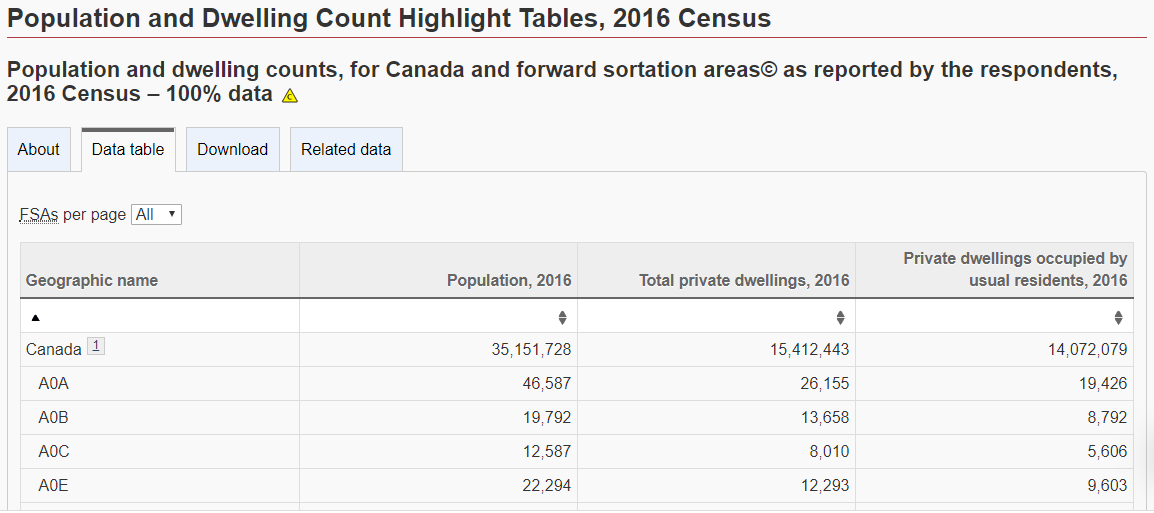
**Neighborhood data of Toronto from Wikipedia [1]:**

This is a list of [postal codes in Canada](https://en.wikipedia.org/wiki/Postal_codes_in_Canada) where the first letter is M. Postal codes beginning with M are located within the city of [Toronto](https://en.wikipedia.org/wiki/Toronto) in the province of [Ontario](https://en.wikipedia.org/wiki/Ontario). Only the first three characters are listed, corresponding to the Forward Sortation Area.



**Toronto Neighborhood population data from Statistics Canada [2]:**

The Census of Population is held across Canada every five years (the last being in 2016). Statistics Canada has collected all the information based on the postal codes. This information is useful to merge with the previous neighborhoods data to find the population of each neighborhood. It helps me to find the densely populated areas in Toronto.



**Latitude and Longitude data of Postal Codes [3]:**

This data collected from the coursera. It is a CSV file with the latitude and longitude information of the Toronto postal codes, which is extracted from the Geocoder Python package.

**Toronto Neighborhood Shapes for Choropleth Maps [4]:**

A GeoJSON file is downloaded to map the Toronto neighborhoods on the Choropleth map.

**Foursquare API [5]:**

[Foursquare API](https://developer.foursquare.com/) to collect information on other venues/competitors in the neighborhoods of Toronto.

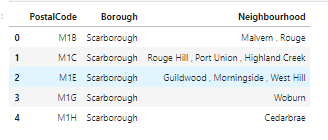
**References:**

* [1] [Toronto Postal Codes — Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
* [2] [Toronto Neighborhood Population from Statistics Canada](https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&SR=1&S=22&O=A&RPP=9999&PR=0)
* [3] [Latitude and Longitude data of Postal Codes](http://cocl.us/Geospatial_data)
* [4] [Toronto Neighborhood GeoJSON](https://raw.githubusercontent.com/ag2816/Visualizations/master/data/Toronto2.geojson)
* [5] [Foursquare API](https://api.foursquare.com/v2/venues/explore?&client_id=%7b%7d&client_secret=%7b%7d&v=%7b%7d&ll=%7b%7d,%7b%7d&radius=%7b%7d&limit=%7b%7d)

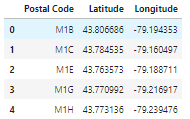
# 3. Data Cleaning and Transformation

# In order to find the target neighborhood(s), we will explore the demographics of the neighborhoods in the city of Toronto by segmenting the data and conducting descriptive analysis using Panda. Additional data will be gleaned by web scraping and API will be used to generate data. The below steps explain the process how data is retrieved, cleaned, and transformed for further analysis and visualization.

1. Toronto neighborhood data is pulled from the Wikipedia which has a table consisting of Postal Code, Borough, Neighborhood.
2. The scraped data is transformed into a data frame “**df1**”, which is then cleaned to eliminate the postal code that are not assigned to any borough.
3. The Neighborhood column with missing data is assigned with the respective borough name
4. And finally, all the neighborhoods are grouped by the postal code as shown below.



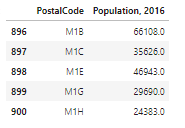
1. The geospatial data from the coursera csv file is copied into a new dataframe “**df2**”



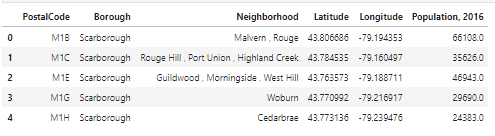
1. Both the dataframe df1 and df2 are merged on postal code to create a new dataframe df\_toronto



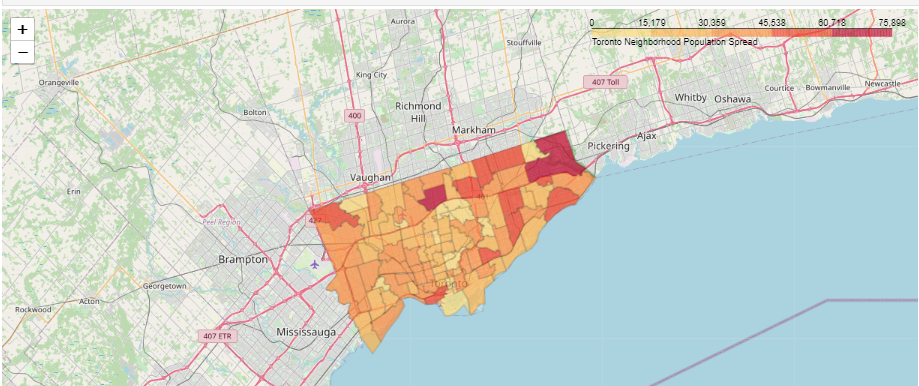
1. Then I pulled the Toronto neighborhood population data from Statistics Canada and cleaned the cleaned the data frame “dfp” to retain the information of Toronto postal codes



1. Lastly, the df\_toronto and dfp are merged on postal code to create the df\_toronto\_pop.



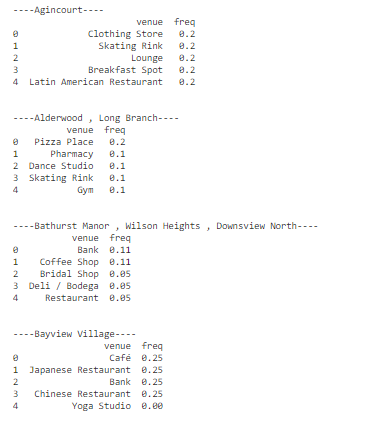
1. This dataframe is important in plotting the below choropleth map and highlighting the densely populated areas.



1. By using Foursquare API I got venues **500 m** around center of each neighborhood with limit of **100 venue**. I merged data with data df\_toronto\_pop.



1. The venues are then groped by the neighborhood and the top 5 venues in each neighborhood are pulled as show below.

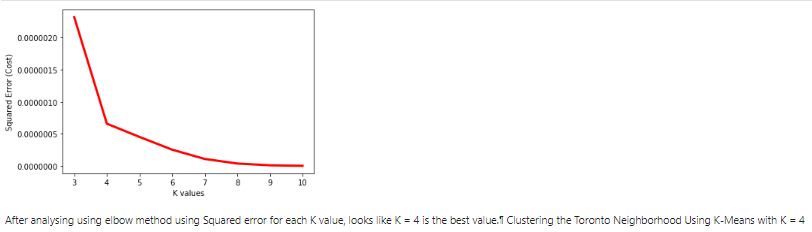


# 4. Data Visualization and Clustering

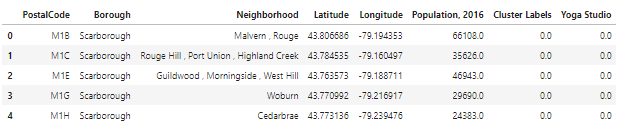
To find the neighborhoods with the “Yoga Studio”, a new data frame toronto\_part is created and visualized using the bar chart as show below. It shows that only quite a few neighborhoods have the “Yoga Studio”, but the data also reveals that in most of neighborhoods Yoga studios are in the top 5 venues list.

# 

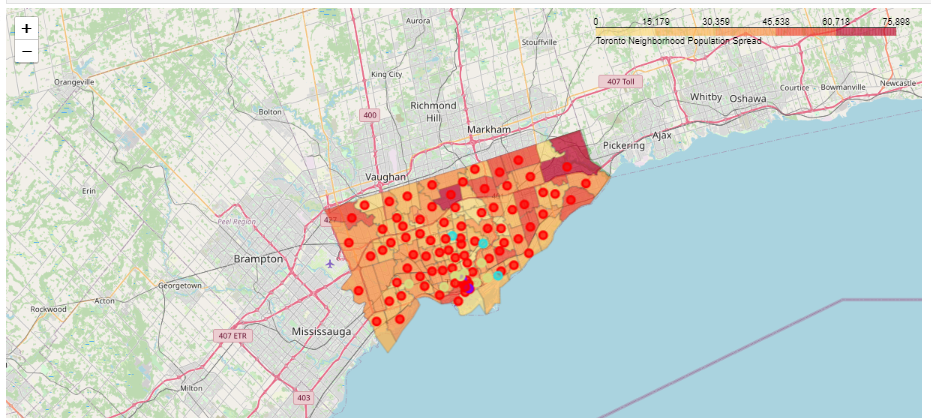
1. With the data now ready, we run k-means to cluster the neighborhoods. By using the Elbow method which is plot between squared error and an array K values we found that 4 cluster would be ideal to cluster the neighborhood.



1. All the neighborhoods are then labelled based on their respective cluster to create the below dataframe



1. The labelled neighborhoods are then plotted on the choropleth map, which is already designed earlier using the neighborhood population as a parameter.



# 5. Discussion

By examining the clusters, cluster-1,2,3 have neighborhoods with a yoga studio. Also, the established Yoga studios are in the top 5 venue list of the respective neighborhood.

However, cluster-0 neighborhoods do not have any Yoga studios in them and most od them are densely populated. Hence, these neighborhoods are better places for starting a new Yoga Studio.

# 6. Conclusion

After sorting the cluster-0 dataframe based on population of the neighborhood, I have identified the top 5 areas i.e. M2N, M1B, M2J, M9V, and M1V that are densely populated and good for starting a Yoga Studio by the investors.

