Recommending a Location for a New Business

IBM Applied Data Science Capstone

Brittany Easton

July 2020

Table of Contents

[Business Problem 3](#_Toc48496638)

[Client 3](#_Toc48496639)

[Background 3](#_Toc48496640)

[Business Problem 3](#_Toc48496641)

[Data 3](#_Toc48496642)

[Data Sources 3](#_Toc48496643)

[Data Cleaning & Feature Selection 5](#_Toc48496644)

[Foursquare 6](#_Toc48496645)

[Methodology 6](#_Toc48496646)

[Exploratory Analysis 6](#_Toc48496647)

[Results 6](#_Toc48496648)

[Discussion 6](#_Toc48496649)

[Conclusion 6](#_Toc48496650)

# Business Problem

## Client

A black belt in Brazilian Jiujitsu is looking to open up her own martial arts studio in the Greater Toronto Area (GTA). She does not have any investors, and needs her business to become profitable quickly in order for it to be sustainable. Her vision is to provide Brazilian Jiujitsu training to an underserviced community where she can really make an impact. She is also concerned about affordability since Brazilian Jiujitsu can be an expensive hobby. Through research she has found that similarly priced forms of recreation have more participants whose households make $100,000 in income per year or more. She has asked for recommendations on where to start her business. She would like a shortlist of 3 locations provided to her so that she can explore them as her primary options.

## Background

The location of a business is a big factor in its success, particularly for businesses that customers attend for services, and require a bricks-and-mortar environment. As an educational, exercise-based service, location will influence the number of potential customers the studio has, and the affluence of those customers. Location will therefore play a big factor in how accessible the studio is, and what percentage of the population has enough disposable income to afford a Brazilian Jiujitsu membership. Location will also determine how many other martial arts businesses the studio will be competing with directly, which will influence the speed of the business’s growth to profitability.

## Business Problem

The analysis will focus on creating a recommendation of 3 locations within the Greater Toronto Area (GTA) for the studio to be opened. Optimization of neighbourhoods will be focused through the maximization of the population, population density, and proportion of affluent households within the neighbourhood, and the minimization of competition.

# Data

## Data Sources

In order to answer the business problem and create recommendations for the location of this new martial arts business, data from several sources will be aggregated in order to determine population, density, competition, and affluence of the neighbourhoods within the Greater Toronto Area (GTA).

|  |  |  |
| --- | --- | --- |
| Foursquare | Data Contents | Location & Venue data |
| What insights does this data provide? | The level of competition in a particular region. |
| What data will be used? | The count of, and the location of martial arts studios in the GTA. |
| How will the data be used? | In determining population density per studio, and the number of high income households per studio in each region. |
| Example Data | Action & Reaction MMA  485 McNicoll Ave, North York, M2H 3H9 |
| Data Source | Foursquare API |
| Canada Revenue Agency | Data Contents | Household Income by FSA |
| What insights does this data provide? | The ability of the households in the GTA to afford a recreational hobby in the martial arts. |
| What data will be used? | The number of households per FSA which make over $100,000. |
| How will the data be used? | Proportion of households making over $100,000 will inform a conversion rate. |
| Example Data | |  |  | | --- | --- | | **FSA** | **$100K to $149K** | | A0A | 1900 | |
| Data Source | <https://www.canada.ca/en/revenue-agency/programs/about-canada-revenue-agency-cra/income-statistics-gst-hst-statistics/individual-tax-statistics-fsa.html> |
| Statistics Canada | Data Contents | Population by FSA |
| What insights does this data provide? | How much room for membership growth exists for the business. |
| What data will be used? | Geographic Name, Population (2016), Total Private Dwellings (2016). |
| How will the data be used? | Population size and density will inform metrics for growth. |
| Example Data | |  |  |  |  | | --- | --- | --- | --- | | **FSA** | **POP (2016)** | **DWELL (2016)** | | | A0A | 46,587 | 26,155 | |
| Data Source | <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&S=22&O=A> |
| Postal Codes in Canada | Data Contents | FSAs by County |
| What insights does this data provide? | Which FSAs lay within which geographical region. |
| What data will be used? | FSA, Type (indicates region, city or county name). |
| How will the data be used? | To define the geographical region of each FSA. |
| Example Data | |  |  | | --- | --- | | **FSA** | **TYPE** | | K1C | OTTAWA | |
| Data Source | <https://www.postalcodesincanada.com/province-ontario/> |
| AggData | Data Contents | Coordinates by FSA |
| What insights does this data provide? | The geographic coordinates of each FSA within the Greater Toronto Area. |
| What data will be used? | FSA, Latitude, Longitude, Province, Place Name. |
| How will the data be used? | Map the Greater Toronto Area regions. |
| Example Data | L3R: Markham, ON  43.8479, -79.3288 |
| Data Source | <https://www.aggdata.com/free/canada-postal-codes> |

## Data Cleaning & Feature Selection

The four data tables will be loaded and cleaned, and then joined together to create a single table which will serve as the main dataframe for analysis outside of Foursquare data. The final dataframe will include only postal codes in Ontario beginning with postal codes L, or M, which encompass the Greater Toronto Area for consideration.

Household Income by FSA (Canada Revenue Agency)

This table reports annual household income per FSA within Canada. The table is formatted to break income ranges into ranges of $5,000 - $10,000. For simplicity’s sake, this analysis requires only the number of households that are making over $100,000, and the ratio of those homes to the total number of homes per FSA (affluence ratio). Any FSAs that do not begin with the letters L or M can be dropped, as these postal codes encompass the Greater Toronto Area’s geography.

Population by FSA (Statistics Canada)

This table reports population, as well as how many private dwellings, Indian reserves there are per FSA within Canada. The analysis requires only that we keep the population per FSA. Population will be used to compute household size when combined with the CRA data on number of total households within an FSA. Any FSAs that do not begin with the letters L or M can be dropped, as these postal codes encompass the Greater Toronto Area’s geography.

Coordinates by FSA (AggData)

This table reports Place Name, Province, Latitude & Longitude for each FSA in Canada. All data other than latitude, longitude and FSA will be dropped. Over two phases the number of records was reduced to include only those within the province of Ontario, and then again in joining with the amalgamated table containing each of these four datasets, to keep only the FSAs within the Greater Toronto Area.

FSAs by County (Postal Codes in Canada)

This table reports the place name for each county or region per FSA within Canada. The only fields are the FSA and the Place Name. Both fields will be kept.

## Foursquare

The Foursquare API is being used for real-time data provision of martial arts studios within the city of Toronto. This data will inform the rate of competition in different areas of Toronto, a key metric for analyzing the optimal location to start a new business. Those areas

# Methodology

## Exploratory Analysis

Table 1: Household Size Histogram

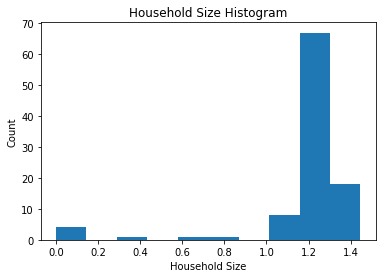


Table 2: Total Households Histogram

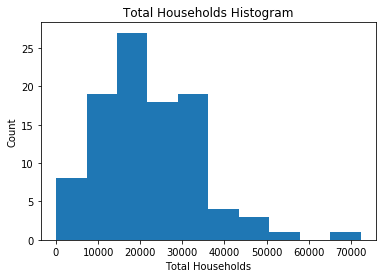


Table 3: Affluence Ratio Histogram

# 

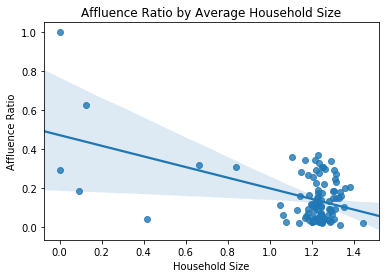
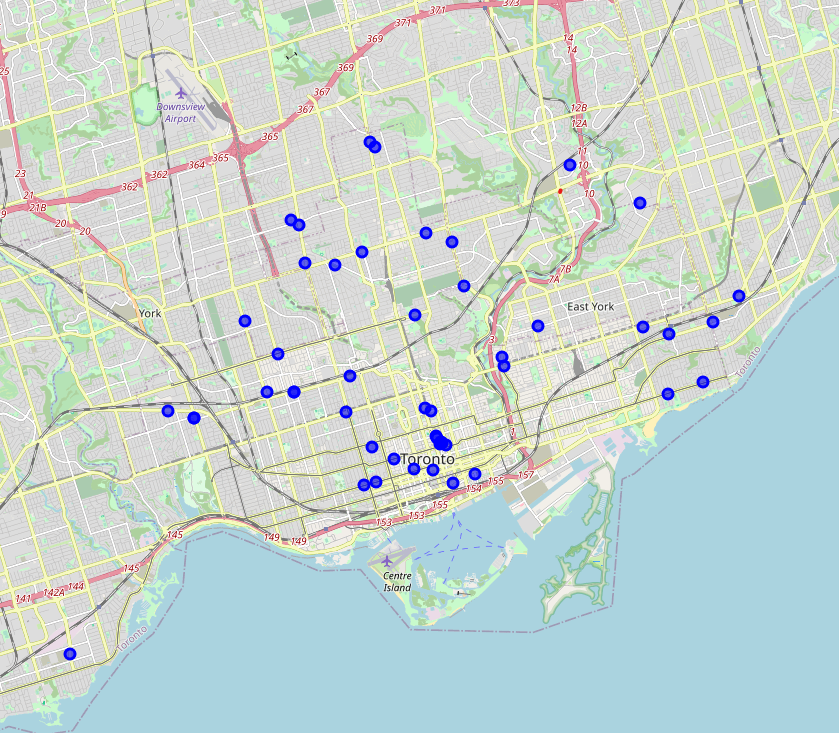
Table 4: Affluence Ratio by Average Household Size

Table 5: Map of Competing Martial Arts Studios in Toronto



## K-Means Clustering

Using K-means clustering on the features Competitors, Population, Total Households, Household Size, Affluence Ratio, Latitude & Longitude allowed like FSAs to be clustered together. These clusters were then mapped in order to determine proximity to competitors.

Most clusters were shown to have little competition. A larger number of total households within a cluster was negatively associated with the affluence ratio of the cluster. A recommend ratio was determined using the product of each cluster’s Population and Affluence Ratio over the number of Competitors per cluster plus one.

Table 6: Clusters defined by Population & Competitors

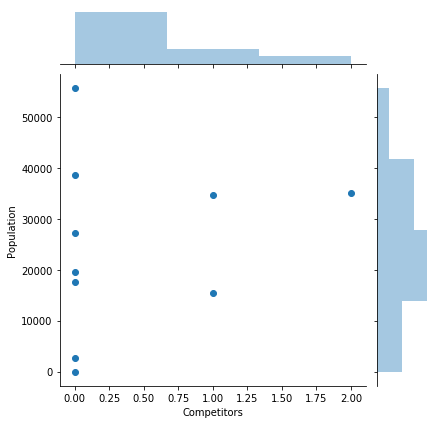


Table 7: Clusters defined by Affluence Ratio & Total Households

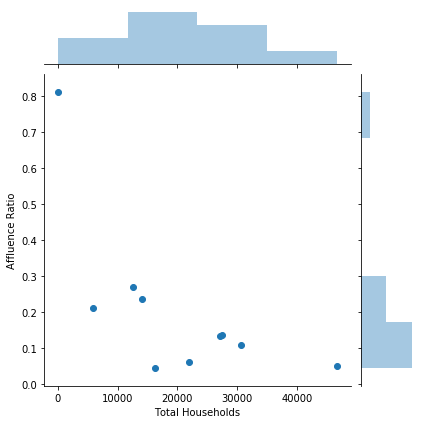
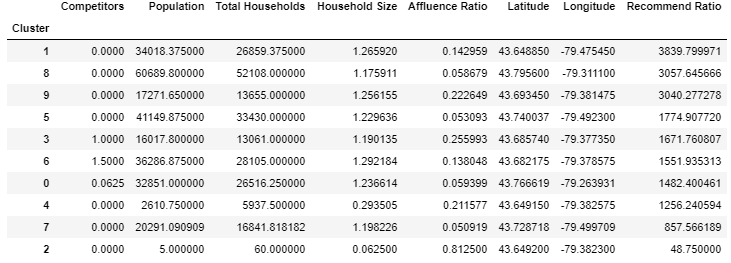
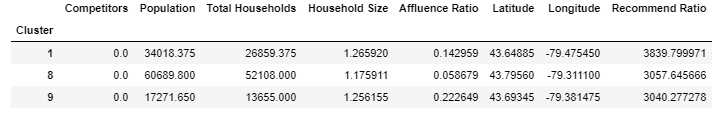


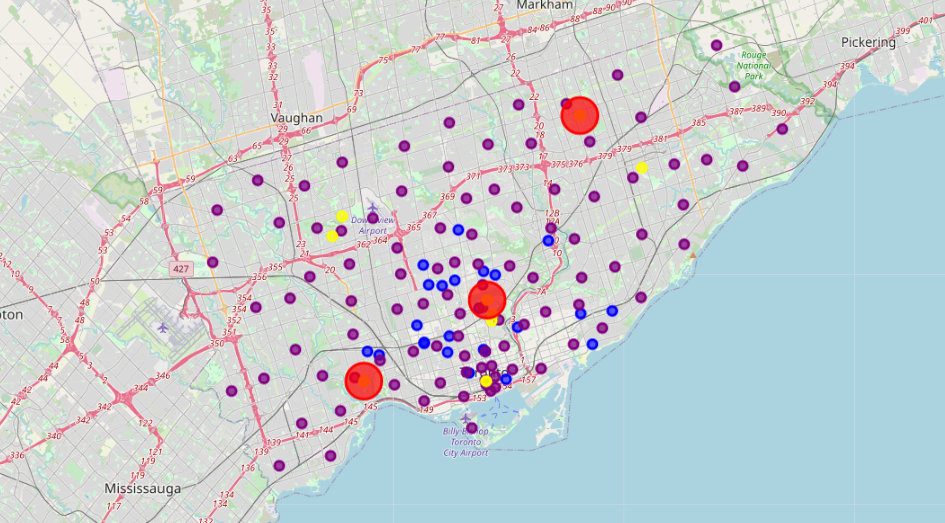
Table 8: Cluster Dataframe Ranked by Recommend Ratio



# Results

A recommend ratio was determined using the product of each cluster’s Population and Affluence Ratio over the number of Competitors per cluster plus one. The three clusters with the largest Recommend Ratio are recommended as possible locations for the client’s new studio.





# Discussion

The analysis of Toronto's martial arts industry has shown that Downtown and Central Toronto are heavy in competition compared with East York, Scarborough, and Etobicoke. The most affluent FSAs within the less competitive neighbourhoods are in the regions of Downtown Toronto, however this area also showcases the smallest average Household Size ratio.

The Recommend Ratio was created to maximize the Population & Affluence Ratio of a cluster compared to the level of competition present there. The top three clusters were selected as recommendations, so that the client may decide based on other factors, such as availability & pricing of commercial real estate.

Suggestions for future improvements of this analysis include:

* taking more location data features into account
* using membership data for competitive studios to show areas with less penetration, and therefore more opportunity
* using commercial real estate pricing to show areas with higher rental prices
* using a choropleth map to show competition, population, and affluence
* using a clustermap to differentiate relative similarity between clusters
* collecting publicly available revenue data for competitive studios for predictive modeling.

# Conclusion

The recommendation made to the client for where to start her new business in Toronto are:

1. Etobicoke
2. Scarborough
3. Central Toronto