



APPLIED DATA SCIENCE CAPSTONE

Final Report

Justin Yu

Introduction

The purpose of this report is to showcase the application of data science in the business world. More specifically, this report aims to illustrate how advanced data analytics create value to the business sector by conducting in depth analysis and generating actionable recommendations. Through understanding the implication of the analysis and the recommendation, stakeholders can make a better business decision.

Business Problem

The business problem that the analysis aims to solve is where in the Greater Toronto Area is the ideal location to open a new gym / fitness center. The trend of living a healthy lifestyle has transformed the daily lives of many, and the popularity of working out is growing exponentially. In specific, the percentage of people going to the gym has increased significantly in recent years in Ontario. As a result, there are many gyms in the Greater Toronto Area now and the intense competition is a headache to many gym owners. Therefore, this analysis will help gym investors and gym owners to identify neighborhoods in the Greater Toronto Area that has a low prevalence of gyms, so as to open up a new gym in a profitable location.

Target Audience

The target audience of the analysis are investors of Gym Investors. Gym Investors include owners of gym chains, owners of boutique gyms, and potential gym owners who are looking for an opportunity to start a gym. In this analysis, the owners of gym chains are the main target audience, as the analysis can help them expand their footprint in new neighborhoods. Among all the factors to be considered, the location of the gym is a critical element contributing to the success of the gym. As gym memberships are relatively long, meaning members are unlikely to switch gyms often, entering a neighborhood with less competition significantly enhances the customer loyalty and long term profitability. On the other hand, enter a saturated neighborhood will likely lead to the failure of an investment. In essence, the analysis can guide the Gym Investors to strategically open a new gym at a specific location that has compelling prospects.

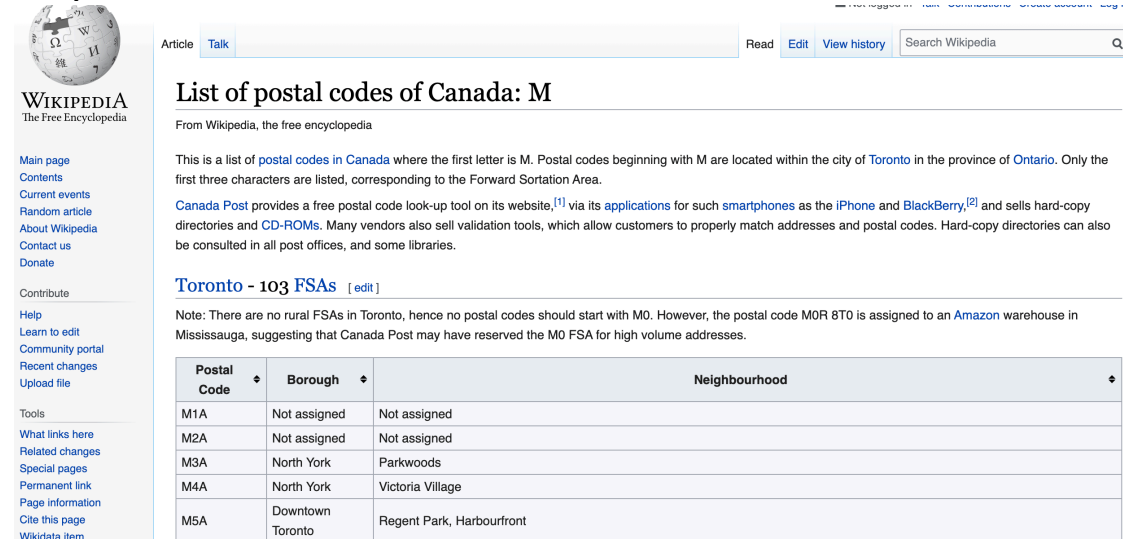
Data

The analysis will require two sets of data. The first set of data includes geographical information, latitude and longitude, of all the neighborhoods in the Greater Toronto Area. The second set of data is the Foursquare location data, which includes the venues of the neighborhoods.

Neighborhoods in Toronto

The first set of the data is obtained by utilizing Wikipedia and Geospatial data from a csv file provided, which are illustrated as follows:

Wikipedia :



The screenshot shows the Wikipedia article titled "List of postal codes of Canada: M". The article text states: "This is a list of [postal codes in Canada](#) where the first letter is M. Postal codes beginning with M are located within the city of [Toronto](#) in the province of [Ontario](#). Only the first three characters are listed, corresponding to the Forward Sortation Area. [Canada Post](#) provides a free postal code look-up tool on its website,^[1] via its applications for such [smartphones](#) as the [iPhone](#) and [BlackBerry](#),^[2] and sells hard-copy directories and [CD-ROMs](#). Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries."

Below the text is a table titled "Toronto - 103 FSAs" with the following data:

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront

Source: [https://en.wikipedia.org/wiki/List of postal codes of Canada: M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

CSV File from Geocoder Package:

Postal Code	Latitude	Longitude
M1B	43.806686	-79.194353
M1C	43.784535	-79.160497
M1E	43.763573	-79.188711
M1G	43.770992	-79.216917
M1H	43.773136	-79.239476
M1J	43.744734	-79.239476
M1K	43.727929	-79.262029
M1L	43.711112	-79.284577
M1M	43.716316	-79.239476
M1N	43.692657	-79.264848
M1P	43.757410	-79.273304

Source: http://cocl.us/Geospatial_data

Venues in Different Neighborhoods

Foursquare location data is leveraged in this project in order to obtain all the information required to conduct the analysis. In specific, the scope covers the venues in different neighborhoods in order to better understand the neighborhood, and the data is illustrated as follows:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Parkwoods	43.753259	-79.329656	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
Parkwoods	43.753259	-79.329656	Tim Hortons	43.760668	-79.326368	Café
Parkwoods	43.753259	-79.329656	A&W	43.760643	-79.326865	Fast Food Restaurant
Parkwoods	43.753259	-79.329656	Bruno's valu-mart	43.746143	-79.324630	Grocery Store

Data Explanation

Combining all three sources of data, a comprehensive dataset can be created. In particular, the combined dataset will include all the venues in different neighborhoods in Toronto, including geospatial data, the name of the venue, and the category of the venue. Hence, the dataset can facilitate the process of identifying the most ideal location to start a gym, which is the purpose of the analysis.

Methodology

Web Scraping Data & Foursquare API

The first step of the analysis is to retrieve the data using the techniques of web scraping, data engineering, and leverage the Foursquare API in Python.

To obtain a list of neighborhoods in Toronto, the BeautifulSoup and Request packages in Python are utilized. By extracting the information from the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), a list of the names of all neighborhoods are extracted. Moreover, using the Geocoder package streamlines the process of obtaining the latitude and longitude information of each neighborhood, which are essential in extracting the information of different venues using Foursquare API.

Moreover, Foursquare API allows information of all venues in a neighborhood to be extracted, and the top 100 venues of each neighborhoods are included. In specific, by doing so, the category each venue can be obtained and be used to cluster neighborhood in order to identify neighborhoods with low prevalence of gym, which illustrates compelling investment prospects.

Exploratory Data Analysis

Having all the data required to conduct clustering, exploratory data analysis is essential to ensure creditability of the data set. In specific, there are different types of gym, which is shown as follows:

```
1 for i in toronto_venues_lst:
2     if 'gym' in i.lower():
3         print (i)
```

Boxing Gym
Climbing Gym
College Gym
Gym
Gym / Fitness Center
Gym Pool

It is apparent that Gym and Gym / Fitness Center are essentially the category that the target audience will be the most interested, and hence should be considered as the same category in the analysis.

Statistical Analysis

The distribution of the prevalence of gym of different neighborhoods is summarized as follows:

Gym	
count	99.000000
mean	0.020659
std	0.026258
min	0.000000
25%	0.000000
50%	0.010000
75%	0.030000
max	0.125000

Clustering: K - means

The final step of the analysis is to cluster neighborhoods based on the prevalence of gym in different neighborhoods. In specific, k - means clustering requires a fixed number of clusters, which is 5 in this analysis, and all the neighborhoods is clustered into 5 groups based on the occurrence of gym in these neighborhoods. This unsupervised machine learning technique is exceptionally suitable for the scope of the analysis, as the target audience can easily identify a list of neighborhoods that have the potential for the gym investors to open a new gym.

Moreover, for gym chains that re undergoing expansion, essentially the entire cluster, which contains a number of neighborhoods, can be a new market to expand to.

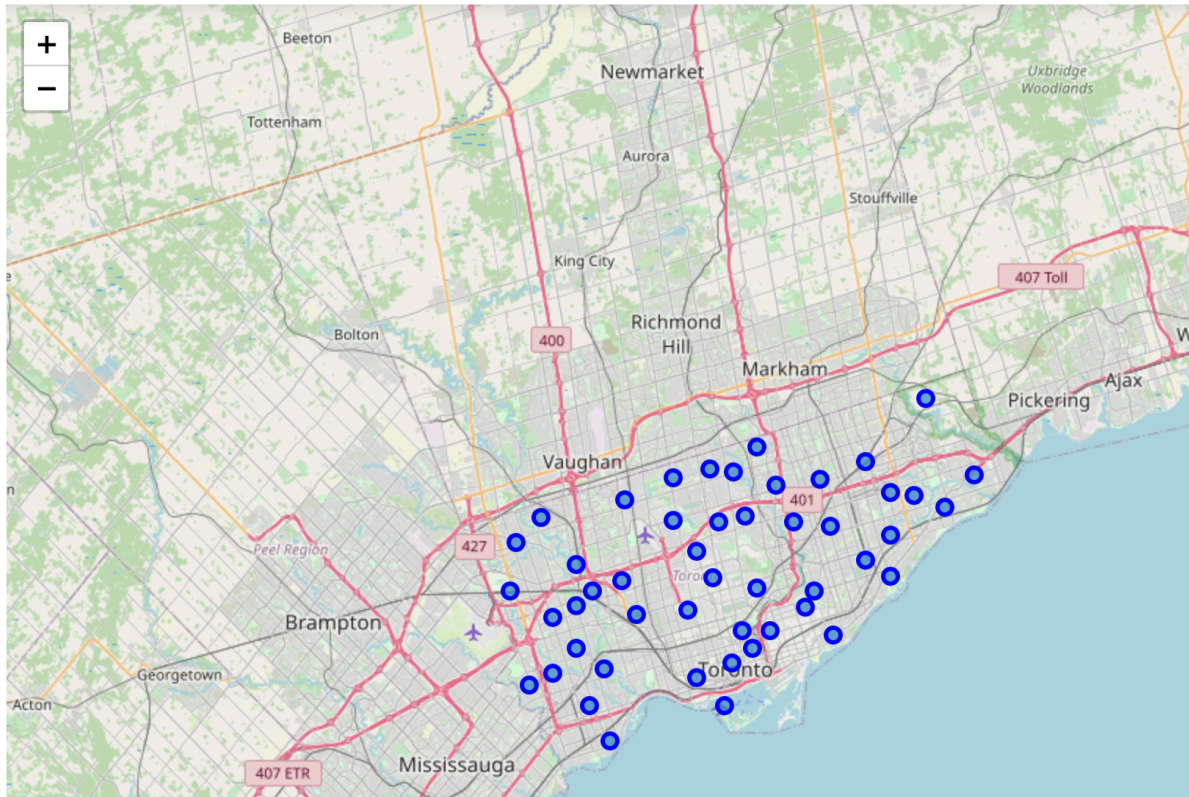
Results

After performing k - means clustering, 'Gym Mean' is the indicator of the prevalence of gym in the cluster. A low value means the prevalence of gym is low in the cluster, and a high value means the prevalence of gym is high in the cluster. In other words, the cluster of neighborhoods with the highest potential for investment is **Cluster 3**, and the cluster with the lowest potential for investment is **Cluster 2**.

The summary of the results is as follows:

Gym Mean	
Cluster Labels	
3	0.006849
1	0.022275
4	0.029426
0	0.035258
2	0.067376

A map illustrating the locations of all neighborhoods in cluster 3 is shown below:



Discussion & Recommendation

From the map above, it is apparent that Cluster 3 have the lowest gym prevalence. Compared to the cluster of the second gym prevalence (Cluster 1) , the gym mean value of Cluster 3 is approximately 70% lower, which is an important indicate that the competition of investing in a new gym in neighborhoods in Cluster 3 is significantly lower than investing in a new gym in neighborhoods in Cluster. Furthermore, comparing the value of gym mean in Cluster 3 (0.006849) with the mean value of gym mean in all other clusters (0.03858375), it is evident that opening a new gym in Cluster 3 will encounter much lower competition than in any neighborhoods that are not in Cluster 3. Therefore, there are many opportunities for gym investors to further expand their presence in the neighborhoods in Cluster 3 with a minimized competition.

Moreover, these neighborhoods scatter in different boroughs in Toronto and are not concentrated in a particular part of Toronto. This provides autonomy for investors to choose a neighborhood which they are more familiar with to open a new gym.

Cluster 3 presents the best potential to open a new gym in terms of the foreseeable level of competition. This analysis recommends gym investors to open a new gym in the neighborhoods in Cluster 3 and to avoid opening a new gym in the neighborhoods in Cluster 2.

Conclusion

In conclusion, this project identifies neighborhoods with great investment opportunity to start a new gym. The business problem is addressed by obtaining essential data, exploratory data analysis, and using machine learning to cluster neighborhoods.

The neighborhoods identified (Cluster 3) are neighborhoods that have a low number of gyms. For the investors, this significantly narrows down the neighborhoods to be considered as starting a gym in these neighborhoods is likely to have a low level of competition.

As a result, other than looking at all neighborhoods one by one, in – depth analysis can be performed on neighborhoods in cluster 3 to identify the best location to open a new gym. The final decision of where to open a gym should be made by the investors by considering factors other than prevalence of gym that will impact their success, such as average household income level, population of the neighborhood, demographic information of the neighborhood, etc. Further research and analysis can be done strategically on the neighborhoods to address factors that are not considered in this analysis in order to identify the ideal location to open a new gym.