Coursera Capstone Project

IBM Data Science Professional Certificate

Best location for opening a gym accessory retail outlet in Singapore

By: Yang Lin

Date: January 2021

Content Page

A. Introduction and backgrounds

With the concepts of wellness and mindfulness becoming ever more important for consumers, fitness has become a more relevant goal for many people (Forbes, 2019). Nowadays, more people are going to health clubs and gyms in order to curb the side-effects that come with the hectic urban lifestyle. Holding a gym membership and spending money on fitness used to be seen as a luxury, but today it has become a part of people's lifestyle. According to Entrepreneur, "growing urbanization, rising middle class, and increasing disposable incomes are boosting the need for gyms (F45, 2020).

The high demand for gym and fitness also boom the retails business for the gym peripheral products. Both the gym operators and their consumers can easily access the gym accessory shop around the gym centers. Therefore, beside more and more gym and fitness centers opened island-wide in Singapore, the gym accessory retailers and potential investors also are more interested to know the numbers of gym centers located in a different neighborhood in Singapore. Then they can make a wise investment decision that opening their new retail outlets in a place with more gym centers.

1. Objective and Business Problems

The objective of this project is to help the retailers and potential investors to find the answers for the question: where are the best neighborhood locations to open new gym accessory retail outlets in Singapore? This study will be done by using the data science methodologies and machine learning techniques and concluded in a business solution and advice.

2. Target Audience of this project

This research project is particularly useful for gym accessory retailers who want to grow their business by opening more gym accessory retail outlets and the potential new investors in these industry sectors. As Singapore is quite a small country, the limited land space caused very crowd in the neighborhood areas. Therefore, high rental and labor cost is a big burden to run an offline

business. Although there are increasing numbers of gym and fitness centers over the recent years, choose a location with low potential customers surrounding may easily result in business failure.

B. Data

To find the solution for the problems, the following data will be required for this project:

- List of neighborhoods in Singapore. This defines the scope of this project that only within the Country of Singapore.
- Latitude and longitude coordinates of the neighborhoods. This is required in order to plot on the map as well as getting the venue data.
- Venue data, particularly data related to gym and fitness centers. The venue data will be used in the machine learning process to cluster the labels on neighborhoods.

1. Data sources and how it helps to solve the problem

A list of neighborhoods in Singapore can be obtained from the Wikipedia web page at https://en.wikipedia.org/wiki/Category:Places_in_Singapore. The data contains 191 neighborhoods in total over the island-wide in Singapore. Then we can retrieve the geographical coordinates for each neighborhood through the python package called Geocoder. With the latitude and longitude of each neighborhood, it can plot the neighborhood as circle markers on the map. It can be visualized where the neighborhoods are located.

Next, the most important data in this project is the venue data. It can be obtained by using the Foursquare API in python which can help to provide many categories of venue data by giving the neighborhood geographical coordinates. But we will only filter out the Gym and Fitness category from the venue data and then use it to train the k-means model for clustering. With the final clustering labels and map visualization, we are able to advise the best location for opening new gym accessory outlets.

C. Methodology

The overall methodology is to get the required data and transform it into the DataFrame format for easy analysis and then use several python packages and API to obtain further relevant data like venues and plot them on the map for visualization. Once the data is ready for machine learning, we will use the k-means model to cluster the targeted venue " Gym / Fitness Center". The labels of clustering will help us to find a solution to the problem.

First, we download the neighborhood list from Wikipedia in HTML format. The BeautifulSoup package could help only retrieve the neighborhood name and then the pandas package can transfer it into a DataFrame format for easy processing. Another required data is the geographical coordinates which are used for obtaining the venue data later. The geocoder package will help to find and map the latitude and longitude based on the name of the neighborhood.

Second, the Foursquare API is very helpful to get the venue data. We did not use the premium version as the free version is enough to provide our requirements in this project. Since Singapore is quite small and there have been 191 neighborhoods already selected, we set 1,000 meters for the radius and limit to 100 explore results. This could avoid too many duplicate results from overlap exploration. The venue data will be returned in JSON format. We will filter the data and merge it with our previous neighborhood data frame.

Third, we apply one-hot encoding and get the venue data in means based on each neighborhood group for the preparation of machine learning. There are 355 unique venues in total. However, only the "Gym / Fitness Center" will be used in the k-means clustering in our project. 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 is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters as more clusters may result in too small a size of the cluster that is not presentable to answer the question. Each cluster is based on the level of frequency of occurrence for "Gym / Fitness Centers".

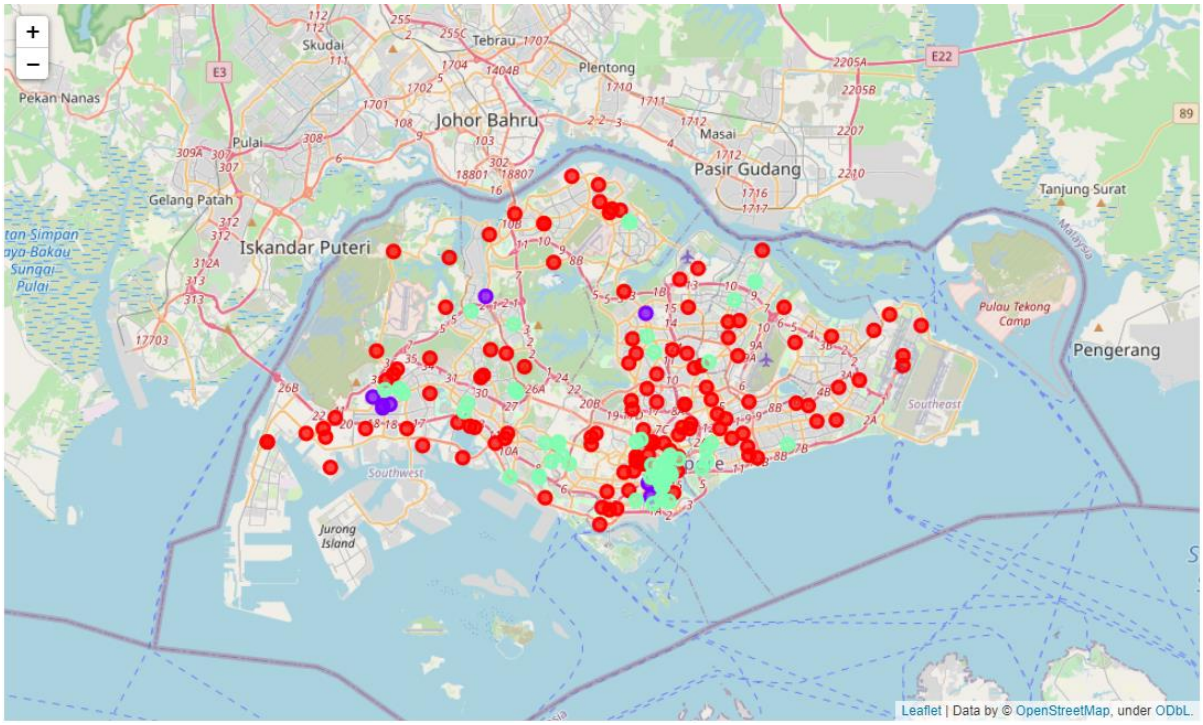
Lastly, we will use the Foursquare API again to plot each cluster in different colors on the map. It will provide a straight view of the distribution of clusters. We could call out the data by filtering each cluster and from there, the answers can be observed.

D. Results

The results show the 3 clusters with labeling 0, 1 and 2 for the "Gym / Fitness Center". Where:

- Cluster 0 (red): Neighborhoods with low frequency of occurrence and numbers of Gym and Fitness Centers.
- Cluster 1 (purple): Neighborhoods with high frequency of occurrence and numbers of Gym and Fitness Centers.
- Cluster 2 (green): Neighborhoods with moderate frequency of occurrence and numbers of Gym and Fitness Centers.

Below map shows how the clusters of neighborhoods are distributed.



E. Observations and Discussion

There are 123 out of the total 191 neighborhoods that fall into cluster 0 which showing as red circle on the map. Cluster 0 provides a category of very low frequency of occurrence of gym and fitness center in the neighborhoods. In fact, there is zero existence of it in these areas. In other words, most neighborhood locations in Singapore are not suitable for opening a gym accessory outlet since the potential customers are seen to be very low from these locations.

First 5 results of cluster 0

	Neighborhood	Gym / Fitness Center	Cluster Labels	No. of Gym/Fitness Center	Latitude	Longitude
0	Admiralty, Singapore	0.0	0	0.0	1.457351	103.818184
91	Kembangan, Singapore	0.0	0	0.0	1.444033	103.825493
90	Kebun Baru	0.0	0	0.0	1.370710	103.837080
89	Katong	0.0	0	0.0	1.304570	103.902880
88	Kampong Ubi	0.0	0	0.0	1.316670	103.900000

There are only 8 neighborhoods in cluster 1 as purple show the highest frequency of occurrence of gym and fitness centers in the neighborhoods. It represents the highest concentration areas such as Telok Ayer Street, Central Area Singapore, Boon Lay and Yew Tee. High number of gym and fitness centers are located in the central CBD areas as we can see there are 9 out of the total 17 gym and fitness centers in Telok Ayer Street and Central Area, Singapore. Therefore, we strongly recommend focusing on this cluster for the investment decision.

Results of cluster 1

	Neighborhood	Gym / Fitness Center	Cluster Labels	No. of Gym/Fitness Center	Latitude	Longitude
182	Western Islands, Singapore	0.062500	1	1.0	1.331100	103.695510
166	Telok Ayer Street	0.050000	1	5.0	1.280960	103.847752
186	Yew Tee	0.071429	1	2.0	1.394440	103.753890
190	Yunnan, Singapore	0.043478	1	1.0	1.337348	103.690630
93	Kian Teck	0.055556	1	1.0	1.332837	103.695545
31	Central Area, Singapore	0.040000	1	4.0	1.288830	103.846250
4	Ang Mo Kio Police Division	0.037037	1	1.0	1.385010	103.845100
14	Boon Lay	0.042553	1	2.0	1.333330	103.700000

The balance of 60 neighborhoods in green in cluster 2 is considered as a moderate level of frequency of occurrence. They are located quite widely over the country but we can see from the map that majority of them also squeezed in the CBD area of Singapore. This again supports our recommendation that considering opening the new gym accessory outlets in these neighborhoods. However, we might be hinting that some areas also have more than one gym and fitness centers existence. For example, there are 3 centers with only 0.03 frequency of occurrence rate in the Central Police Division. More centers is good sign but the low frequency of occurrence rate implies a high number of other venues existence in this neighborhood. In other words, this neighborhood is a more popular place where the rental cost might be higher than in other neighborhoods. This should be brought to consideration when making investment decisions.

First 5 results of cluster 2

	Neighborhood	Gym / Fitness Center	Cluster Labels	No. of Gym/Fitness Center	Latitude	Longitude
127	Pasir Panjang	0.029412	2	1.0	1.292290	103.768190
32	Central Police Division	0.030000	2	3.0	1.278930	103.839090
169	Template:Punggol	0.016393	2	1.0	1.402460	103.906860
168	Template:Places in Singapore	0.020000	2	2.0	1.290410	103.852110
36	Choa Chu Kang	0.027778	2	1.0	1.386160	103.746180

F. Limitations

This project only takes two factors the frequency of occurrence and the number of Gym / Fitness Center for analysis. However, there are many other factors such as the population, lifestyle, shopping mode (on-line/off-line), etc. not taken into this research and analysis. Further analysis is advisable for a more comprehensive study. For instance, the number of people visit the gym center in the CBD area at the weekend, the popularity of on-line purchases that all impair the location as the only factor for decision making. However, this is not in this project scope.

Another limitation is from the use of the free version of Foursquare API. The limited number of calls and unstable call results might restrict sufficient venue data for analysis. In this project, 19,100 (191 x 100) venues were supposed to be retrieved but there were only 12,306 returned.

G. Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. gym accessory retailers and potential investors regarding the best locations to open a new gym accessory outlet. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations and certain neighborhoods in cluster 2 are next tier to consider for opening a new outlet. The findings of this project will help the relevant stakeholders to capitalize on the opportunities in high potential locations while avoiding the low potential locations for business growth. At the end of the analysis, the report also highlights the limitations of other factors to be considered in the decision making for the stakeholder.

H. References

Wikipedia, Category: Places in Singapore

https://en.wikipedia.org/wiki/Category:Places_in_Singapore

Forbes, 2019. The Biggest Trends In Gyms And The Fitness Industry

https://www.forbes.com/sites/richardkestenbaum/2019/11/20/the-biggest-trends-in-gyms-and-the-fitness-industry/?sh=277f4fc37465

F45, 2020. Why the fitness industry is booming

https://www.f45invest.com/blog/why-the-fitness-industry-is-booming

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs