IBM DATA SCIENCE CAPSTONE PROJECT REPORT

Opening a Gym/Fitness Centre in Ho Chi Minh City, Viet Nam

August 26, 2019

Cuong H. Nguyen-Dinh

Contents

Introduction

Ho Chi Minh (HCM) city (its former name is Sai Gon) is known as the most populous city in Viet Nam (with a population of over 10 million people). Moreover, HCM city is also the financial centre of Viet Nam. In this dynamic city, there are many opportunities for business activities. Therefore, it is good to think of doing business in this southern beautiful city of Viet Nam.

As the most attractive city of young people in Viet Nam, the gym/fitness service is highly required. Because this sports provides the youth not only the good health but also the well-form body. Hence, opening gym/fitness centre in HCM city promises a beneficial business.

Business problem

In this report, we try to answer the question of "Can we figure out areas in HCM city where a gym/fitness centre can be launched?" Our solution to the above business question mainly depends on data science methodology and cluster analysis of machine learning technique.

Data

We use the following data to solve the aforementioned business problem:

- Data about neighborhoods in HCM city. These are starting points to navigate companies or entrepreneurs around each neighborhood.
- Coordinate data of neighborhoods including latitudes and longitudes. This kind of data will be used to search venues and to draw map.
- Venue data is served clustering method which divides neighborhoods to groups. By analyzing these groups, we can locate areas where we can set up the gym/fitness centre.

The data used in this study comes from two sources:

- Wikipedia page[1];
- Foursquare service[2].

The data collection and processing include two steps:

1. In order to build the list of neighborhoods in HCM city, we crawl data at this URL `https://en.wikipedia.org/wiki/Category:Districts_of_Ho_Chi_Minh_City`. After having analyzed the html content, the list of 24 neighborhoods is generated. Then this data is used as the input for the next step.
2. Foursquare service is used to get coordinates of each neighborhood. Then this service is repeatedly used to get nearby venues of each neighborhood. The venue data are vectorized in order to become input for cluster analysis.

In the next section, we will describe our analytic method using these data.

[1] `https://en.wikipedia.org/wiki/Category:Districts_of_Ho_Chi_Minh_City`
[2] `https://foursquare.com/`

Methodology

In this study, we use Folium library to call Foursquare API in order to collect:

1. coordinates of neighborhoods;

2. venue data including venue category and venue coordinates.

In the first step, based on the list of neighborhoods collected at Wiki page (as mentioned in the previous section), we find coordinates of those neighborhoods by using Foursquare API. The neighborhood data including names and relevant coordinates is depicted in Fig 3.1.

```
(24, 3)
[55]:
```

	Neighborhood	Latitude	Longitude
0	Bình Chánh District	10.690115	106.582677
1	Bình Tân District	10.749809	106.605664
2	Bình Thạnh District	10.804659	106.707848
3	Cần Giờ District	10.398254	106.921951
4	Củ Chi District	10.975609	106.499711

Figure 3.1: Neighborhood data

The location of these neighborhoods is plotted in map as shown in Fig 3.2.

In the second step, with every neighborhood, we search for near by venues within the radius of 1 kilometer and the maximum number of collected venues is 100. An example of result is shown in Fig 3.3.

The venue data composes 123 unique categories and 662 records in total.

In the third step, we analyze these neighborhood - venue data with the aim at providing data for clustering step. Hence, we vectorize category data by using one-hot encoding method. The excerpt of transformed data is depicted in Fig 3.4.

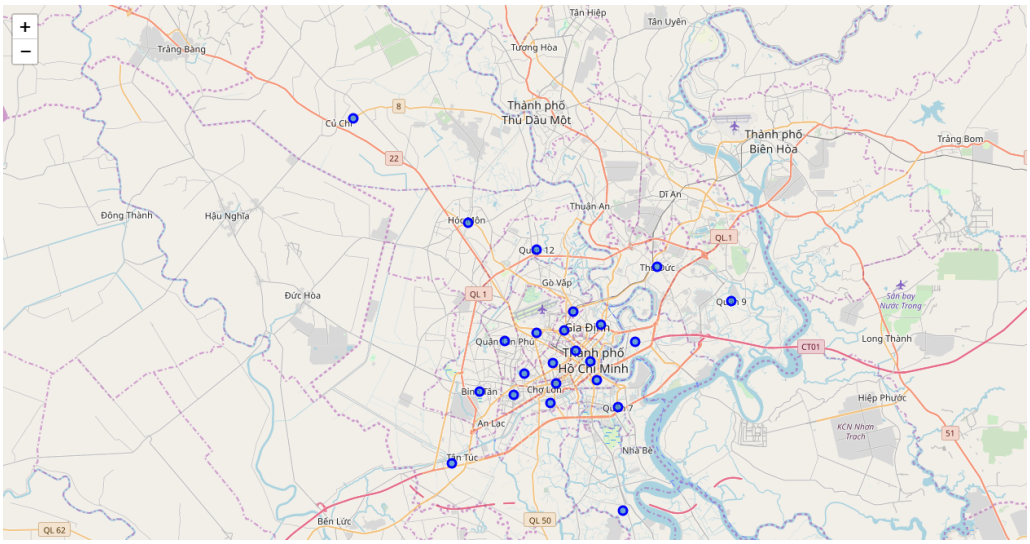

Figure 3.2: Location of neighborhoods in map

```
(662, 7)
```

[65]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bình Chánh District	10.690115	106.582677	Van Thinh Rest Stop	10.687641	106.584976	Vietnamese Restaurant
1	Bình Chánh District	10.690115	106.582677	Bánh Bao Phong Lan	10.692805	106.577061	Bakery
2	Bình Chánh District	10.690115	106.582677	Binh Dien Quan Restaurant	10.695333	106.587668	Vietnamese Restaurant
3	Bình Chánh District	10.690115	106.582677	Kedai Sarah	10.688974	106.574965	Women's Store
4	Bình Tân District	10.749809	106.605664	Coffee Bui Van Ngo	10.751648	106.612534	Café

Figure 3.3: Venue data

```
(662, 124)
```

	Neighborhood	Airport Service	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	...	Trail	Travel Agency	Vegetarian / Vegan Restaurant	Vide Gam Stor
0	Bình Chánh District	0	0	0	0	0	0	0	0	0	...	0	0	0	
1	Bình Chánh District	0	0	0	0	0	0	0	0	1	...	0	0	0	
2	Bình Chánh District	0	0	0	0	0	0	0	0	0	...	0	0	0	
3	Bình Chánh District	0	0	0	0	0	0	0	0	0	...	0	0	0	
4	Bình Tân District	0	0	0	0	0	0	0	0	0	...	0	0	0	

Figure 3.4: An excerpt of transformed data

In the fourth step, we group these transformed data by neighborhood column and calculate the mean weight for every category. Fig 3.5 shows an excerpt of the result.

(23, 117)

	Neighborhood	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	Bar	...	Track Stadium	Trail	Travel Agency	Vegetar / Veg Restaur
0	Bình Chánh District	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.250000	0.0	...	0.0	0.0	0.000000	
1	Bình Thạnh District	0.0	0.034483	0.0	0.068966	0.0	0.0	0.0	0.034483	0.0	...	0.0	0.0	0.034483	
2	Bình Tân District	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.000000	
3	Cần Giờ District	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.000000	
4	Củ Chi District	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.000000	

5 rows × 117 columns

Figure 3.5: Grouped data

In the fifth step, we try to find the top 10 venue categories and the result is shown in Fig 3.6

[71]: (23, 11)

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bình Chánh District	Vietnamese Restaurant	Women's Store	Bakery	Electronics Store	Food & Drink Shop	Food	Flower Shop	Flea Market	Fast Food Restaurant	Eastern European Restaurant
1	Bình Thạnh District	Café	Coffee Shop	Vietnamese Restaurant	Soup Place	Seafood Restaurant	Road	French Restaurant	Bookstore	Art Gallery	Asian Restaurant
2	Bình Tân District	Café	Shopping Mall	Vietnamese Restaurant	Yoga Studio	Food & Drink Shop	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant
3	Cần Giờ District	Café	Yoga Studio	Food Court	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant	Eastern European Restaurant	Electronics Store
4	Củ Chi District	Café	Coffee Shop	History Museum	Vietnamese Restaurant	Bakery	Yoga Studio	Food	Flower Shop	Flea Market	Fast Food Restaurant

Figure 3.6: The top 10 venue category of each neighborhood

As for cluster analysis, we apply k-means model. Hence, before entering the final model, we apply elbow analysis to find out the optimum k value for k-means model. As shown in Fig 3.7, we see that the optimum k value is 3.

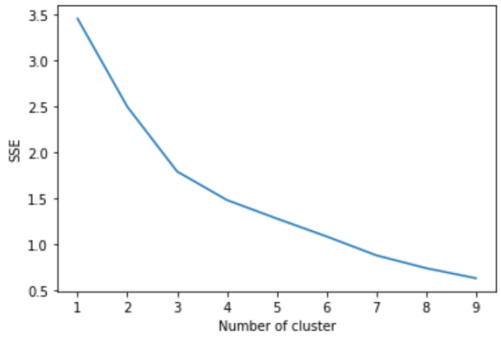

Figure 3.7: Elbow analysis

Based on the previous steps, we build the k-means clustering model with k = 3 and use the vectorized data for this model. The cluster column is then added to the data

frame in order to serve visualization and analysis purposes in the nest steps. Fig 3.8 shows an excerpt of the clustering result.

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Binh Chánh District	10.690115	106.582677	1	Vietnamese Restaurant	Women's Store	Bakery	Fried Chicken Joint	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant	Electronics Store	Fast Food Restaurant
1	Binh Tân District	10.749809	106.605664	1	Italian Restaurant	Shopping Mall	Chinese Restaurant	Café	Yoga Studio	Food Court	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant
2	Binh Thanh District	10.804659	106.707848	1	Café	Vietnamese Restaurant	Asian Restaurant	Coffee Shop	Seafood Restaurant	Steakhouse	Bakery	Soup Place	French Restaurant	Fried Chicken Joint
3	Cần Giờ District	10.398254	106.921951	2	Café	Yoga Studio	French Restaurant	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant	Electronics Store	Fast Food Restaurant	Flea Market
4	Củ Chi District	10.975609	106.499711	2	Café	Coffee Shop	Seafood Restaurant	Convention Center	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant	Electronics Store

Figure 3.8: An excerpt of clustering result

The clustering results are now plotted in map as in Fig 3.9.

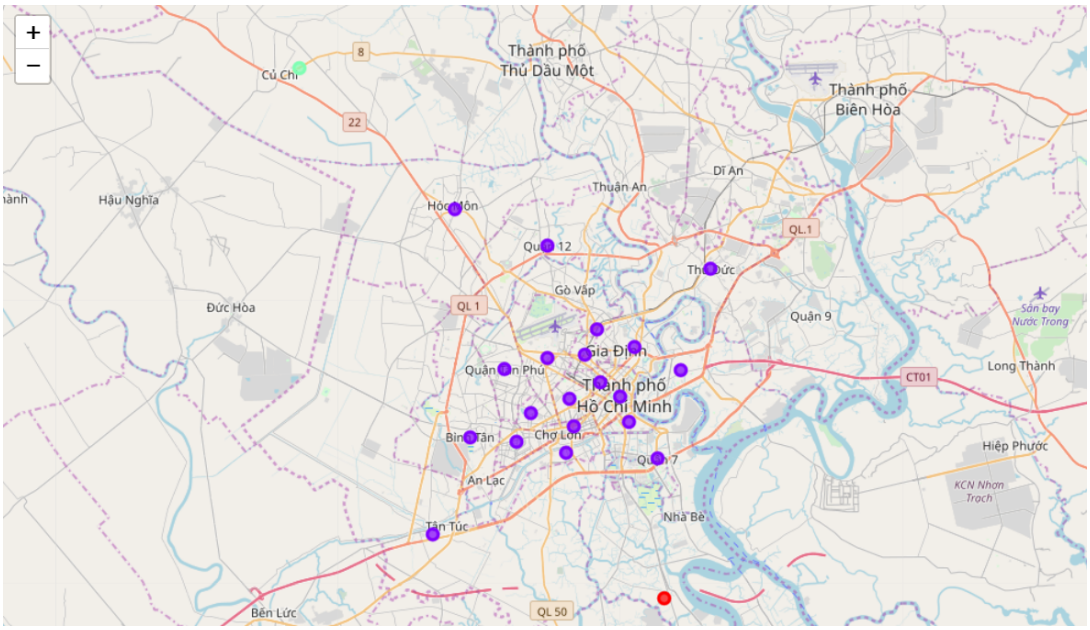

Figure 3.9: Visualizing clusters

In the next section, we analyze the clustering results.

Results

The k-means cluster model divide the neighborhoods to three distinct groups which are numbered 1, 2 and 3. The explanatory data of these three groups are shown in the Fig 4.1, 4.2 and 4.3 respectively.

10	District 6, Ho Chi Minh City	Supermarket	Café	Flea Market	Movie Theater	Pizza Place	Vietnamese Restaurant	Asian Restaurant	Department Store	Bagel Shop	Fast Food Restaurant
11	District 7, Ho Chi Minh City	Vietnamese Restaurant	Café	Supermarket	Flea Market	Sushi Restaurant	Multiplex	Gym / Fitness Center	Residential Building (Apartment / Condo)	Shopping Mall	Scandinavian Restaurant
12	District 8, Ho Chi Minh City	Café	Coffee Shop	Trail	Vietnamese Restaurant	Flea Market	Dim Sum Restaurant	Fast Food Restaurant	Track Stadium	Dumpling Restaurant	Flower Shop
14	District 10, Ho Chi Minh City	Vietnamese Restaurant	Café	Coffee Shop	Seafood Restaurant	Vegetarian / Vegan Restaurant	Dessert Shop	Diner	Bookstore	Burger Joint	Convenience Store
15	District 11, Ho Chi Minh City	Café	Department Store	Pizza Place	Basketball Stadium	Theme Park	Cantonese Restaurant	Water Park	Diner	Dumpling Restaurant	Food Court
16	District 12, Ho Chi Minh City	Café	Vietnamese Restaurant	Department Store	Yoga Studio	Food Truck	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant	Electronics Store
17	Gò Vấp District	Café	Vietnamese Restaurant	Asian Restaurant	Market	Convention Center	Pizza Place	Brewery	Shopping Mall	Park	Flea Market
18	Hóc Môn District	Café	Market	Seafood Restaurant	Vietnamese Restaurant	Supermarket	Spa	Flower Shop	Flea Market	Fast Food Restaurant	Electronics Store
20	Phú Nhuận District	Café	Coffee Shop	Vietnamese Restaurant	Asian Restaurant	Hotel	BBQ Joint	Spa	Bookstore	Chinese Restaurant	Gym / Fitness Center

Figure 4.1: An excerpt of Group 1

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Cần Giờ District	Café	Yoga Studio	French Restaurant	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant	Electronics Store	Fast Food Restaurant	Flea Market
4	Củ Chi District	Café	Coffee Shop	Seafood Restaurant	Convention Center	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant	Electronics Store

Figure 4.2: Group 2

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
19	Nhà Bè District	Coffee Shop	Yoga Studio	French Restaurant	Dessert Shop	Dim Sum Restaurant	Diner	Dumpling Restaurant	Electronics Store	Fast Food Restaurant	Flea Market

Figure 4.3: Group 3

As we analyze the results, gym/fitness centres still have not been developed as many as restaurants or coffee shops in all three groups. However, in group 1, gym/fitness centre get the top 7 & 10 category in some neighborhoods, while in group 2 & 3, gym/fitness centre takes no place.

Discussion

As the observation results in the previous section, gym/fitness centre have not been developed as many as other kinds of business like restaurants or coffee shops in HCM city. Hence there are big opportunities for opening a gym/fitness centre in this city in general. In particular, we can select the areas of group 2 and/or group 3 to launch our centre. Because gym/fitness does not appear in the top 10 venue categories of these two groups. Eventually, we can think of launching our centre in many points of group 1, because the number of this kind of centre is small in comparison with other kinds of business.

Conclusion

In this study, we have defined the business problem, figured out data requirements, collected and merged data from multiple sources, transformed data, visualized data and analyzed clustered results. The findings of this study can be briefly reported as follows.

- There is big opportunity for opening gym/fitness centre in HCM city because this kind of business has not been developed;
- The opportunity is high at areas of group 2 & 3; and
- In the areas of group 1, there is still lot of opportunities because gym/fitness centres are just minority and low rank in the top 10 venue categories of this group.