PREDICT POTENTIAL BUSINESS SERVICES IN VIETNAM

Le Huy Hoang

July 05, 2021

Contents

1. Introduction

1.1 Background

Vietnam is a country in Southeast Asia with a population of over 96 millions. In recent decades, thanks to the reforms in economy & politics, the country has transformed from one of the poorest nations into a lower middle-income country.

Rapid urbanization and the emerging of middle class make the country become a very attractive investment destination, especially its major cities like Hanoi, or Ho Chi Minh.

1.2 Problems

To make investment decision, investors will have to figure out that in the near future, which types of business will have more room to grow, hence, be more profitable.

Due to the fact that countries in East and Southeast Asia that were heavily influenced by Chinese culture, their development directions may share a lot of similarities. This aspect could be utilized to resolve the business problem.

In the project, modern cities in China, and Singapore will be analyzed, then common characteristics in their structures will be used to predict the most potential business services in Vietnam.

1.3 Interest

The work would be valuable for investors who are looking for business opportunities in Vietnam, and urban planners to optimize the effectiveness of land use and infrastructure.

2. Data

We are going to analyze the two cities:

- Shanghai, China
- Singapore

To extract the common structure in modern cities in East and Southeast Asia. The results will then be compared to the status of Hanoi, Vietnam to assess the accuracy.

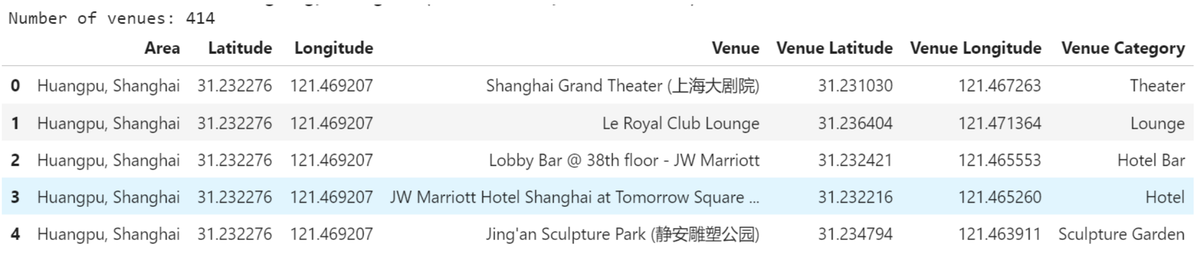
To extract the structure of business services in a city, we could use Foursquare Database, but first, we will need to have information about administrative divisions of the city, and their locations.

2.1 Shanghai, China

Administrative divisions of Shanghai could be obtained from Wikipedia, and via Geopy Library, we could query their locations from OpenStreetMap:

Area	Latitude	Longitude
Huangpu, Shanghai	31.2322758	121.4692071
Xuhui, Shanghai	31.163698	121.4279938
Changning, Shanghai	31.2092762	121.3899859
Jing'an, Shanghai	31.2297756	121.44306
Putuo, Shanghai	31.2513263	121.3912291
Hongkou, Shanghai	31.266703	121.501751
Yangpu, Shanghai	31.2620112	121.5214305
Pudong, Shanghai	31.2217826	121.5387401
Baoshan, Shanghai	31.4066338	121.4851577
Minhang, Shanghai	31.1147666	121.3769429
Jiading, Shanghai	31.377756	121.2606119
Jinshan, Shanghai	30.744817	121.3372571
Songjiang, Shanghai	31.0344052	121.2232077
Qingpu, Shanghai	31.1521636	121.1195519
Fengxian, Shanghai	30.9204487	121.4693834
Chongming, Shanghai	31.6313393	121.5337768

Using Foursquare Explore Request, with latitude, and longitude as parameters, we could the list of recommended venues around the location. Besides location information, I also passed the radius (2000m), and the upper limit (50) of results to return. Below is a portion of venues' data around Shanghai districts.



```
Number of venues: 414
```

	Area	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Huangpu, Shanghai	31.232276	121.469207	Shanghai Grand Theater (上海大剧院)	31.231030	121.467263	Theater
1	Huangpu, Shanghai	31.232276	121.469207	Le Royal Club Lounge	31.236404	121.471364	Lounge
2	Huangpu, Shanghai	31.232276	121.469207	Lobby Bar @ 38th floor - JW Marriott	31.232421	121.465553	Hotel Bar
3	Huangpu, Shanghai	31.232276	121.469207	JW Marriott Hotel Shanghai at Tomorrow Square ...	31.232216	121.465260	Hotel
4	Huangpu, Shanghai	31.232276	121.469207	Jing'an Sculpture Park (静安雕塑公园)	31.234794	121.463911	Sculpture Garden

2.2 Singapore

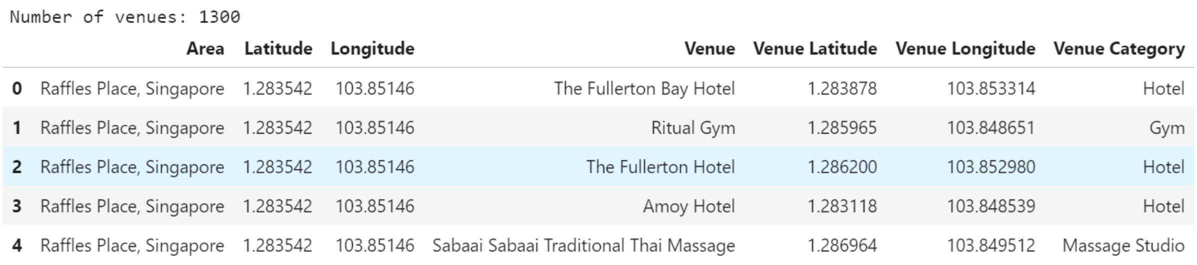
Unlike Shanghai 's Wikipedia page, the structure of Singapore 's page is quite complicated, and it's difficult to directly extract the districts from the page. Therefore, I used District Code and District Map of Singapore instead.

Repeat the same procedures with Shanghai, we could achieve all needed data.

Districts with locations:

Area	Latitude	Longitude
Raffles Place, Singapore	1.2835417	103.851460232669
Anson, Singapore	1.2737957	103.843473
Queenstown, Singapore	1.2946226	103.8060366
Telok Blangah, Singapore	1.27102005	103.809694766557
Pasir Panjang, Singapore	1.27620135	103.79147582342
High Street, Singapore	1.290383	103.8497316
Middle Road, Singapore	1.2986301	103.8538615
Little India, Singapore	1.30684265	103.849273551709
Orchard, Singapore	1.3034266	103.831341955707
Ardmore, Singapore	1.3089844	103.8288974
Watten Estate, Singapore	1.3282039	103.8092589
Balestier, Singapore	1.326226	103.8473149
Macpherson, Singapore	1.32620695	103.889506953939
Geylang, Singapore	1.3181862	103.8870563
Katong, Singapore	1.3016238	103.9045983
Bedok, Singapore	1.3239765	103.930216
Loyang, Singapore	1.3753678	103.9772924
Tampines, Singapore	1.3546528	103.9435712
Serangoon Garden, Singapore	1.3624579	103.8660127
Bishan, Singapore	1.3509859	103.848255074929
Upper Bukit Timah, Singapore	1.3466479	103.7719139
Jurong, Singapore	1.2596166	103.670471298337
Hillview, Singapore	1.3624042	103.767427293728
Lim Chu Kang, Singapore	1.4342172	103.7149872
Kranji, Singapore	1.4252189	103.7619891
Upper Thomson, Singapore	1.3507896	103.8359951
Yishun, Singapore	1.4293839	103.8350282
Seletar, Singapore	1.4098488	103.8773789

Venues:



```
Number of venues: 1300
```

	Area	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Raffles Place, Singapore	1.283542	103.85146	The Fullerton Bay Hotel	1.283878	103.853314	Hotel
1	Raffles Place, Singapore	1.283542	103.85146	Ritual Gym	1.285965	103.848651	Gym
2	Raffles Place, Singapore	1.283542	103.85146	The Fullerton Hotel	1.286200	103.852980	Hotel
3	Raffles Place, Singapore	1.283542	103.85146	Amoy Hotel	1.283118	103.848539	Hotel
4	Raffles Place, Singapore	1.283542	103.85146	Sabaai Sabaai Traditional Thai Massage	1.286964	103.849512	Massage Studio

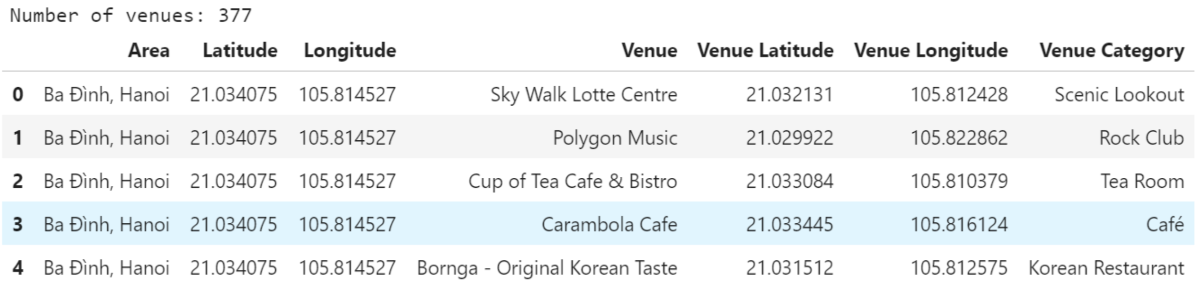
2.3 Hanoi, Vietnam

All information of Hanoi, Vietnam could be obtained from its Wikipedia page, OpenStreetMap, and Foursquare Database.

Districts with locations:

Area	Latitude	Longitude
Ba Đình, Hanoi	21.0340746	105.8145271
Bắc Từ Liêm, Hanoi	21.0698605	105.7573392
Cầu Giấy, Hanoi	21.02916475	105.803437672196
Đống Đa, Hanoi	21.0129196	105.8271961
Hà Đông, Hanoi	20.97026	105.775000621979
Hai Bà Trưng, Hanoi	21.0059701	105.8574845
Hoàn Kiếm, Hanoi	21.0285237	105.8507155
Hoàng Mai, Hanoi	20.9745977	105.8637067
Long Biên, Hanoi	21.0393411	105.8922453
Nam Từ Liêm, Hanoi	21.0128458	105.7608745
Tây Hồ, Hanoi	21.079042	105.8154324
Thanh Xuân, Hanoi	20.9936873	105.8143014

Venues:



```
Number of venues: 377
```

	Area	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ba Đình, Hanoi	21.034075	105.814527	Sky Walk Lotte Centre	21.032131	105.812428	Scenic Lookout
1	Ba Đình, Hanoi	21.034075	105.814527	Polygon Music	21.029922	105.822862	Rock Club
2	Ba Đình, Hanoi	21.034075	105.814527	Cup of Tea Cafe & Bistro	21.033084	105.810379	Tea Room
3	Ba Đình, Hanoi	21.034075	105.814527	Carambola Cafe	21.033445	105.816124	Café
4	Ba Đình, Hanoi	21.034075	105.814527	Bornga - Original Korean Taste	21.031512	105.812575	Korean Restaurant

3. Methodology

To predict potential business services in Vietnam, we could refer from the economic models of modern cities which have similar culture characteristics and locate in the same or nearby regions. Therefore, I choose Shanghai, China, and Singapore as the role models.

Firstly, I collect information about popular venues around the administrative divisions of the two cities. With the purpose of finding out development trends in service sector, I removed districts/divisions which had less than 10 venues around.

Due to the fact that Shanghai and Singapore are ones of the most developed cities these days, the numbers of venues are very big, and their categories are various. To be able to get the insights about the two cities, we will need to segment the divisions, then analyze the most common venues in each cluster.

In this project, K-means Clustering Algorithm will be used for the task, with the inputs are encoded venues' categories. To find the best number of clusters, and evaluate the models' performance, I used Elbow Method in combination with Silhouette Score.

Based on the results of clustering process, we will extract a set of business services which are common in both two cities, and we could expect that those services will be also promising in Vietnamese cities.

Finally, we will assess the current status of those services in Hanoi, Vietnam at the moment to evaluate the accuracy.

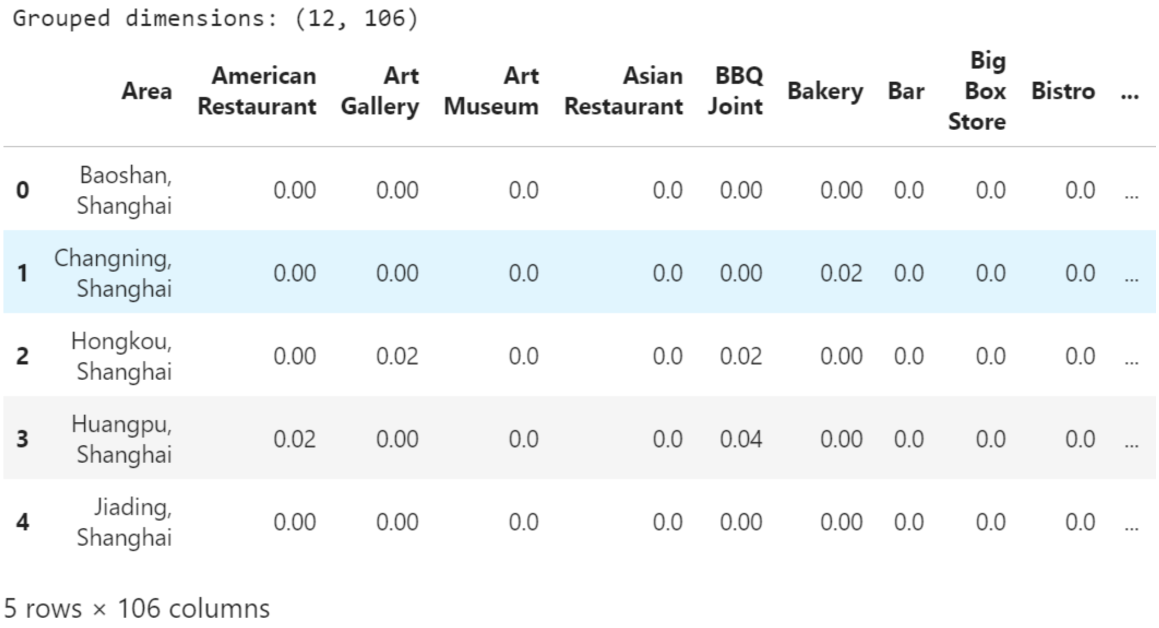
4. Analysis

In the section, we will process venues' data in each city, then use them as inputs of K-means Clustering Model.

4.1 Shanghai, China

Venues' categories from Foursquare Database are in text format, to use it for clustering, we will need to encode the categories. The encode technique will be One-Hot Encoding. After this phase, each venue will be associated with a category vector.

Next step, I group the venues by their central division, and calculate the mean value of each category in each division. Below is a screenshot of the grouped table:

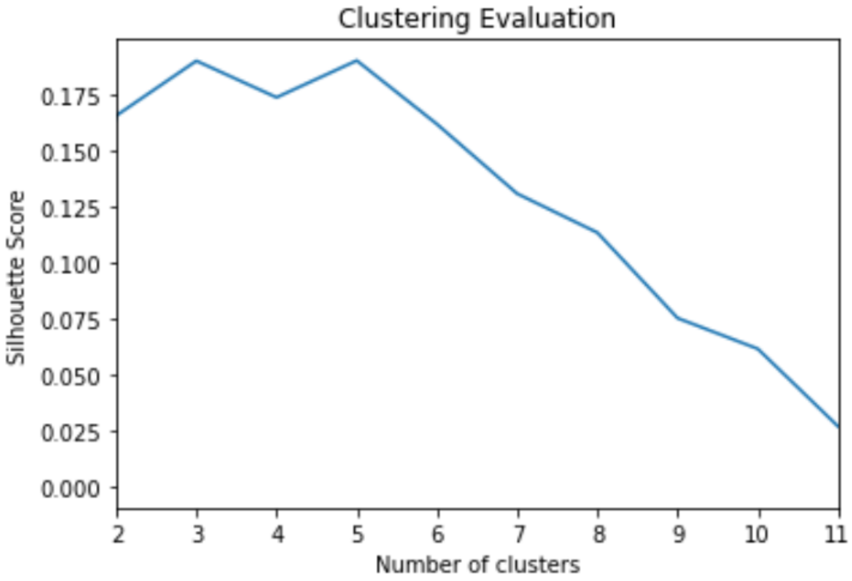


```
Grouped dimensions: (12, 106)
```

	Area	American Restaurant	Art Gallery	Art Museum	Asian Restaurant	BBQ Joint	Bakery	Bar	Big Box Store	Bistro	...
0	Baoshan, Shanghai	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0	...
1	Changning, Shanghai	0.00	0.00	0.0	0.0	0.00	0.02	0.0	0.0	0.0	...
2	Hongkou, Shanghai	0.00	0.02	0.0	0.0	0.02	0.00	0.0	0.0	0.0	...
3	Huangpu, Shanghai	0.02	0.00	0.0	0.0	0.04	0.00	0.0	0.0	0.0	...
4	Jiading, Shanghai	0.00	0.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0	...

5 rows × 106 columns

One more thing we need to do before clustering the data is to find the best number of clusters (K). To do it, I used Elbow method in combination with Silhouette Score. The chart below display the relation between number of clusters with Silhouette Score:



Based on this chart, we could assume the best K is 5.

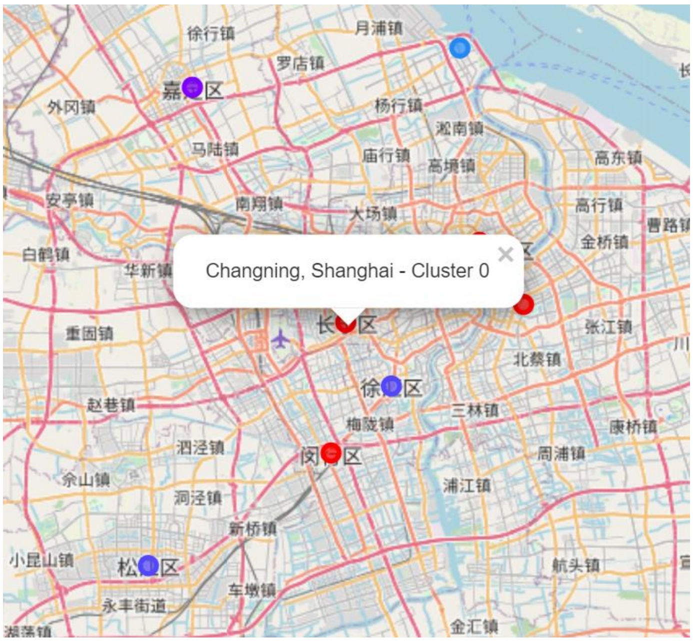
By performing K-means Clustering on Shanghai venues' data with **K = 5**, we got cluster label for each division in Shanghai. Below is the summary of clusters:

Cluster	Most Common Venues		
	1st	2nd	3rd
0	{'Coffee Shop'}	{'Japanese Restaurant', 'Hotel', 'Fast Food Restaurant'}	{'Italian Restaurant', 'Hotel', 'Fast Food Restaurant', 'Shopping Mall'}
1	{'Hotel'}	{'Garden'}	{'Food'}
2	{'Hotel', 'Fast Food Restaurant'}	{'Hotel', 'Fast Food Restaurant', 'Coffee Shop'}	{'Metro Station', 'Hotel', 'Coffee Shop'}
3	{'Shopping Mall'}	{'Port'}	{'Boat or Ferry'}
4	{'Hotel'}	{'Dumpling Restaurant', 'French Restaurant'}	{'Chinese Restaurant', 'Hotpot Restaurant'}

From the cluster table, we could see that the most popular business services in Shanghai are related to Restaurant (especially Fast Food), Coffee Shop, and Hotel.

The differences between clusters mostly come from the order of those three groups.

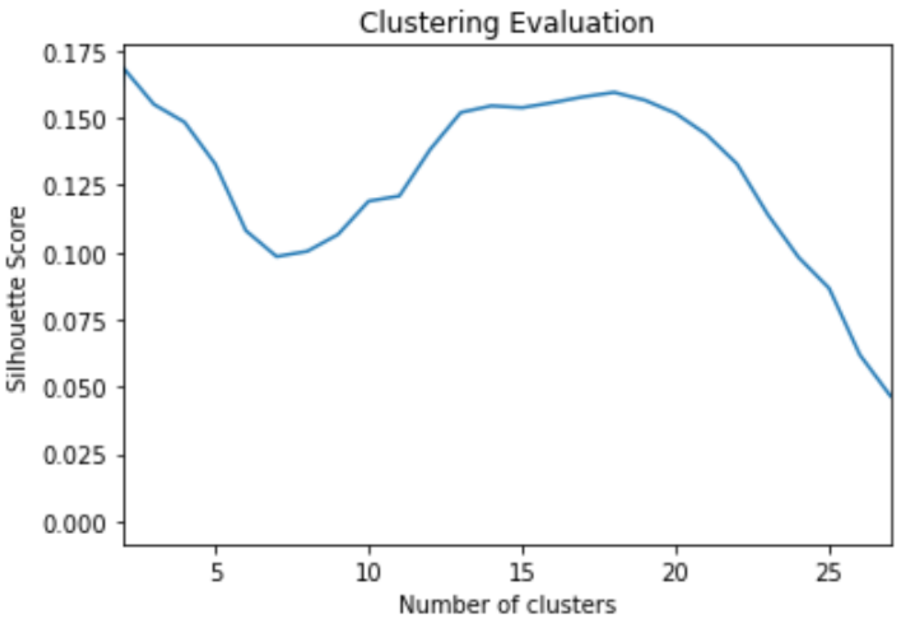
Now, let's visualize the clusters on Shanghai map:



Repeat the same workflow for Singapore, and Hanoi, I got their clustered venues data as below.

4.2 Singapore

Relation between K and Silhouette Score



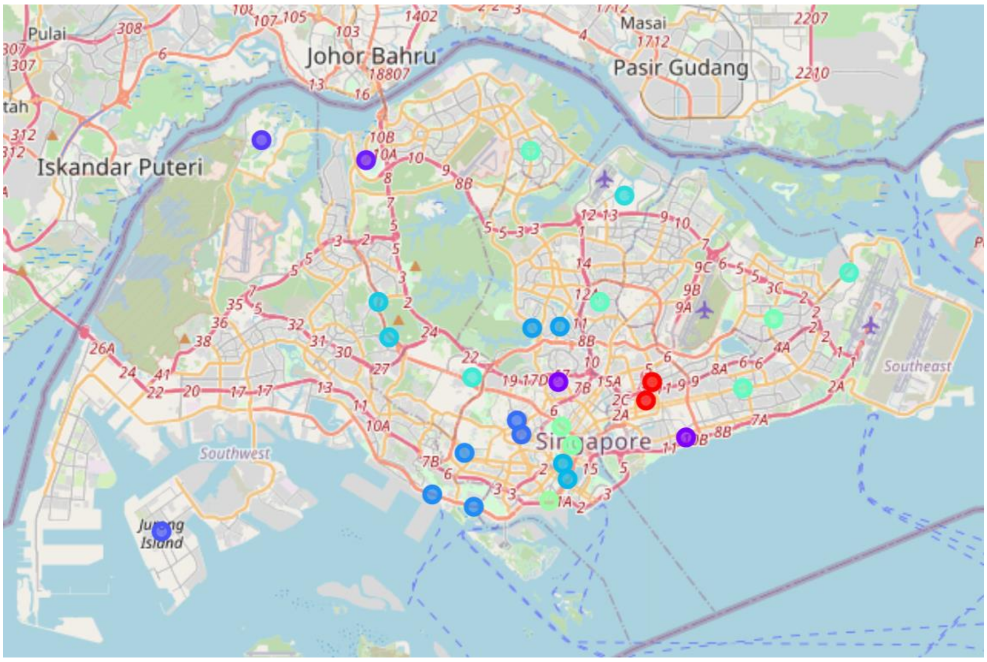
We could use **13 clusters** to segment Singapore districts:

Cluster	Most Common Venues		
	1st	2nd	3rd
0	{'Resort', 'Park'}	{'Café', 'Trail'}	{'Scenic Lookout', 'Indian Restaurant'}
1	{'Chinese Restaurant', 'Food Court', 'Thai Restaurant', 'Coffee Shop'}	{'Chinese Restaurant', 'Thai Restaurant', 'Food Court', 'Coffee Shop'}	{'Indian Restaurant', 'Chinese Restaurant', 'Food Court', 'Coffee Shop'}
2	{'Hotel'}	{'Bakery'}	{'Japanese Restaurant', 'Clothing Store'}
3	{'Harbor / Marina'}	{'Beach'}	{'Resort'}
4	{'Hotel'}	{'Japanese Restaurant', 'Waterfront'}	{'Performing Arts Venue', 'Waterfront'}
5	{'Farm'}	{'Campground'}	{'Harbor / Marina'}
6	{'Chinese Restaurant', 'Café', 'Airport'}	{'Chinese Restaurant', 'Café', 'Park'}	{'Italian Restaurant', 'Café', 'Coffee Shop'}
7	{'Korean Restaurant', 'Coffee Shop'}	{'Indian Restaurant', 'Nature Preserve'}	{'Café', 'Food Court'}
8	{'Grocery Store'}	{'Food Court'}	{'Pizza Place'}
9	{'Japanese Restaurant', 'Bakery', 'Coffee Shop'}	{'Asian Restaurant', 'Italian Restaurant', 'Coffee Shop'}	{'Supermarket', 'Chinese Restaurant', 'Spanish Restaurant'}
10	{'Asian Restaurant', 'Chinese Restaurant'}	{'Asian Restaurant', 'Chinese Restaurant'}	{'BBQ Joint', 'Noodle House'}

11	{'Coffee Shop'}	{'Park'}	{'Supermarket'}
12	{'Hotel'}	{'Café', 'Indian Restaurant'}	{'Japanese Restaurant', 'Café'}

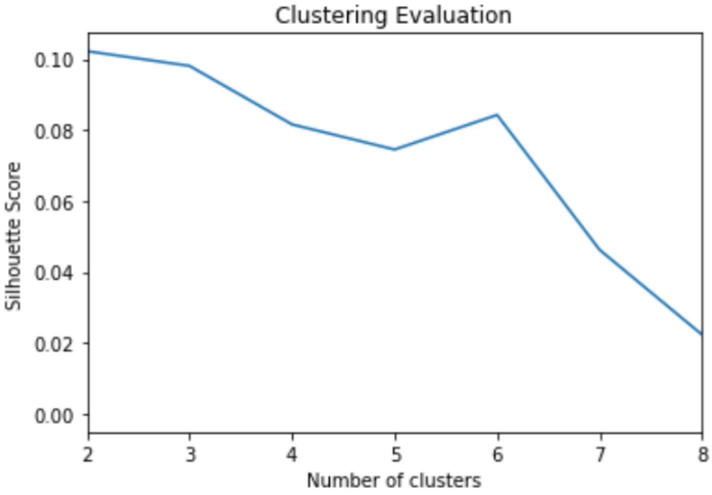
Although there are 13 clusters, business services in Singapore are also mostly related to Restaurant, Coffee Shop, and Hotel.

Divisions' clusters on Singapore 's Map:



4.3 Hanoi, Vietnam

Relation between K and Silhouette Score:

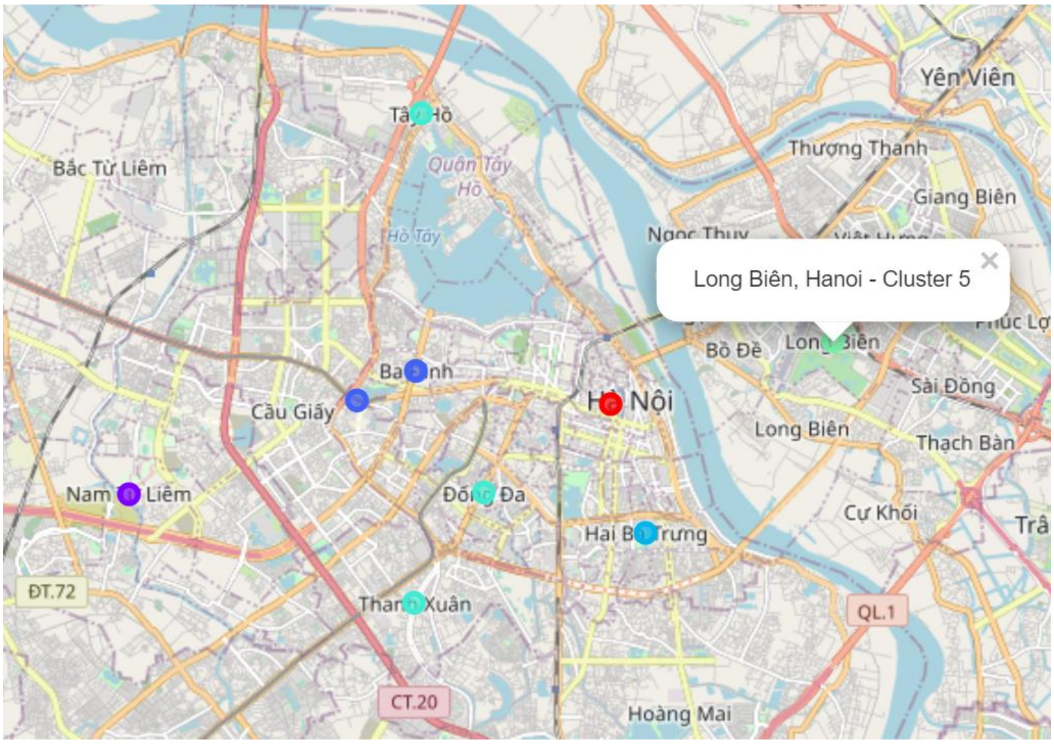


Results after clustering Hanoi divisions with **K = 6**:

Cluster	Most Common Venues		
	1st	2nd	3rd
0	{'Hotel'}	{'Vietnamese Restaurant'}	{'Coffee Shop'}
1	{'Café'}	{'Coffee Shop'}	{'Bakery'}
2	{'Café', 'Coffee Shop'}	{'Japanese Restaurant', 'Hotel'}	{'Vietnamese Restaurant'}
3	{'Vietnamese Restaurant'}	{'Café'}	{'Japanese Restaurant'}
4	{'Vietnamese Restaurant', 'Noodle House', 'Coffee Shop'}	{'Multiplex', 'Vietnamese Restaurant', 'Coffee Shop'}	{'Café', 'Bakery'}
5	{'Vietnamese Restaurant'}	{'Auto Garage'}	{'Shopping Mall'}

From the cluster table, we could see that the most popular business services in Hanoi are also Hotel, Coffee Shop, and Restaurant.

Divisions' clusters on Hanoi 's Map:



5. Results & Discussions

From the clustering results of Shanghai, and Singapore, we see that business service structures in the two cities are very similar. In both of cities, Restaurant, Coffee Shop, and Hotel are the most popular services.

One noticeable difference is that in Shanghai, Fast Food is the most popular type of restaurants, while in Singapore, people seem to prefer Foreign Cuisines.

From the similarities, we could predict that Restaurant, Coffee Shop, and Hotel will be the most promising business services in other cities in the nearby regions.

To evaluate the accuracy of the prediction, let's take a look at the data of Hanoi, one of the biggest cities in Vietnam. We could see that the business structure in Hanoi is closely matched with the prediction.

6. Conclusion

In this project, we have analyzed the business structures in Shanghai, China, and Singapore, to predict potential business services in Vietnam.

Although, the accuracy of the prediction has been proved by comparing to the data of Hanoi, one of the biggest cities in Vietnam, however, there are still many rooms for improvements. For example, we could dig further into the densities of venues in each city to find out saturation thresholds; or the historical aspects could be utilized to customize the prediction.