IBM Data Science Course Capstone Project

May 2020

(Source: https://www.timeout.com/tokyo/restaurants/best-brunch-in-tokyo)

1. Introduction

I have been acquiring skills related to data science by taking the IBM Data Science Professional Course on Coursera. The last course contains a capstone project. This project is about applying data science toolset and obtained skills to analyze a problem in reality and creating value. My project's theme concerns a topic that I have been really interested in: Food & Beverage industry. My analysis was performed in Python. The details are pushed to Github, containing detailed report and the Jupyter notebook. I will link them at the end of the report.

2. Business Problem

My client decide to open a new restaurant in Ho Chi Minh City, Vietnam. In recent years, there is a big boom in all-day brunch&bakery. She is keen on opening a new unit, which will focus on the American and Asian fusion vegetarian kitchen. Taking into account the financial plan at which the restaurant will operate, the intent is to find an optimal location in an area, where vegetarian all-day brunch is booming. The following criteria should be considered:

- Accessibility for tourists and local citizens (transportation)
- Nearby competitors
- Metropolitan area

The assumption behind the analysis is that we can use unsupervised machine learning to create clusters of districts that will provide us with a list of areas for potential locations for

the restaurant. The purpose is that the restaurant to be situated close to one of high populated areas and touristic hotspots, with less competition, and easily accessed to.

3. Data

To perform this analysis, we will need the following data:

- List of the districts of Ho Chi Minh City
- Geo-coordinates of the districts in Ho Chi Minh City
- Top venues of districts

List of districts will be obtained from https://www.gso.gov.vn/dmhc2015/Default.aspx

Geo-coordinates of districts will be obtained with the help of the geocoder tool in the notebook.

Top venues data will be obtained from Foursquare through an API.

4. Methodology

As a database, I used GitHub repository in my study. My master data has the main components Districts and their Latitude/Longitude.

	District Code	District	Latitude	Longitude
0	760	District 1	10.774540	106.699184
1	761	District 12	10.747343	106.666116
2	762	Thu Duc District	10.822023	106.718302
3	763	District 9	10.747343	106.666116
4	764	Go Vap District	10.840150	106.671083
5	765	Binh Thanh District	10.804659	106.707848
6	766	Tan Binh District	10.797979	106.653805
7	767	Tan Phu District	10.791640	106.627302
8	768	Phu Nhuan District	10.800118	106.677042
9	769	District 2	10.804963	106.747470
10	770	District 3	10.775844	106.701756
11	771	District 10	10.773198	106.667833
12	772	District 11	10.764208	106.643282
13	773	District 4	10.759243	106.704890
14	774	District 5	10.756129	106.670376
15	775	District 6	10.746928	106.634495
16	776	District 8	10.747343	106.666116
17	777	Binh Tan District	10.749809	106.605664
18	778	District 7	10.736573	106.722432
19	783	Cu Chi District	10.971846	106.487258
20	784	Hoc Mon District	10.891609	106.594945
21	785	Binh Chanh District	10.696676	106.593789
22	786	Nha Be District	10.701211	106.739009

Data frame with coordinates of districts in Ho Chi Minh City

As you can notice, we have only 23 districts here and that is due to the error with geocode that failed to retrieve latitude of Can Gio District. As a result, I decided to proceed with the data I had obtained. Also, the district is outskirts and unlikely to have a material impact on the analysis.

Map of Ho Chi Minh City with showing the districts

In the next step, I obtain venue information for each district to see which venues are the most common. In this step, the venue data was collected from Foursquare via API. After collecting the data and organizing into a pandas dataframe, I have a table of top 10 common venues of each district that looks like this (this is only a portion of the whole table).

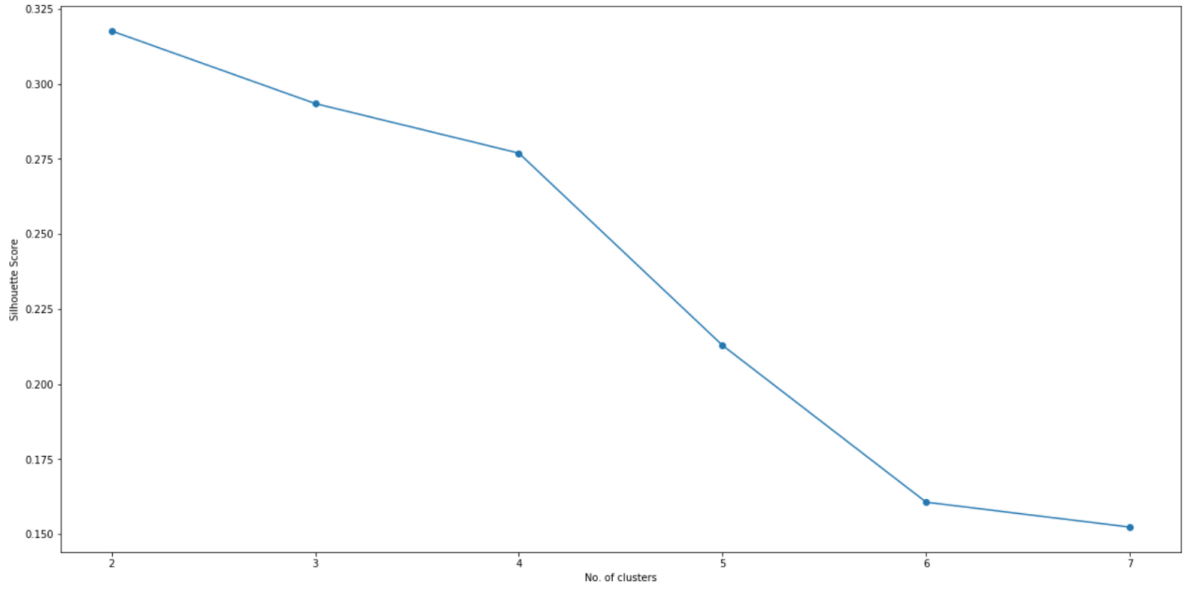
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Binh Chanh District	Vietnamese Restaurant	Brewery	Whisky Bar	Flower Shop	Convention Center	Cupcake Shop	Department Store	Dessert Shop	Dim Sum Restaurant
1	Binh Tan District	Café	Chinese Restaurant	Shopping Mall	Whisky Bar	Flea Market	Cupcake Shop	Department Store	Dessert Shop	Dim Sum Restaurant
2	Binh Thanh District	Café	Coffee Shop	Seafood Restaurant	Vietnamese Restaurant	Diner	Road	Bubble Tea Shop	Convenience Store	Soup Place
3	Cu Chi District	Hotel	Karaoke Bar	Bus Station	Market	Pharmacy	Vietnamese Restaurant	Golf Course	Film Studio	Convention Center
4	District 1	Vietnamese Restaurant	Hotel	Coffee Shop	Café	Japanese Restaurant	Vegetarian / Vegan Restaurant	Spa	Massage Studio	Hotel Bar
5	District 10	Vietnamese Restaurant	Café	Coffee Shop	Vegetarian / Vegan Restaurant	Soup Place	Dessert Shop	Gym / Fitness Center	Korean Restaurant	Frozen Yogurt Shop
6	District 11	Café	Water Park	Shopping Mall	Asian Restaurant	Cantonese Restaurant	Theme Park	Basketball Stadium	Electronics Store	Film Studio
7	District 12	Vietnamese Restaurant	Dim Sum Restaurant	Chinese Restaurant	Coffee Shop	Food	Bakery	Dessert Shop	Gym / Fitness Center	Dumpling Restaurant

One-hot coding is a required step before we can run the clustering algorithm. It converts the categorical values into dummies so they can be used for machine learning.

	Neighborhood	American Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bar	Basketball Stadium
0	District 1	0	0	0	0	0	0	0	0	0	0	0	0
1	District 1	0	0	0	0	0	0	0	0	0	0	0	0
2	District 1	0	0	0	0	0	0	0	0	0	0	0	0
3	District 1	0	0	0	0	0	0	0	0	0	0	0	0
4	District 1	0	0	0	0	0	0	0	0	0	0	0	0

One-hot coding result (extraction)

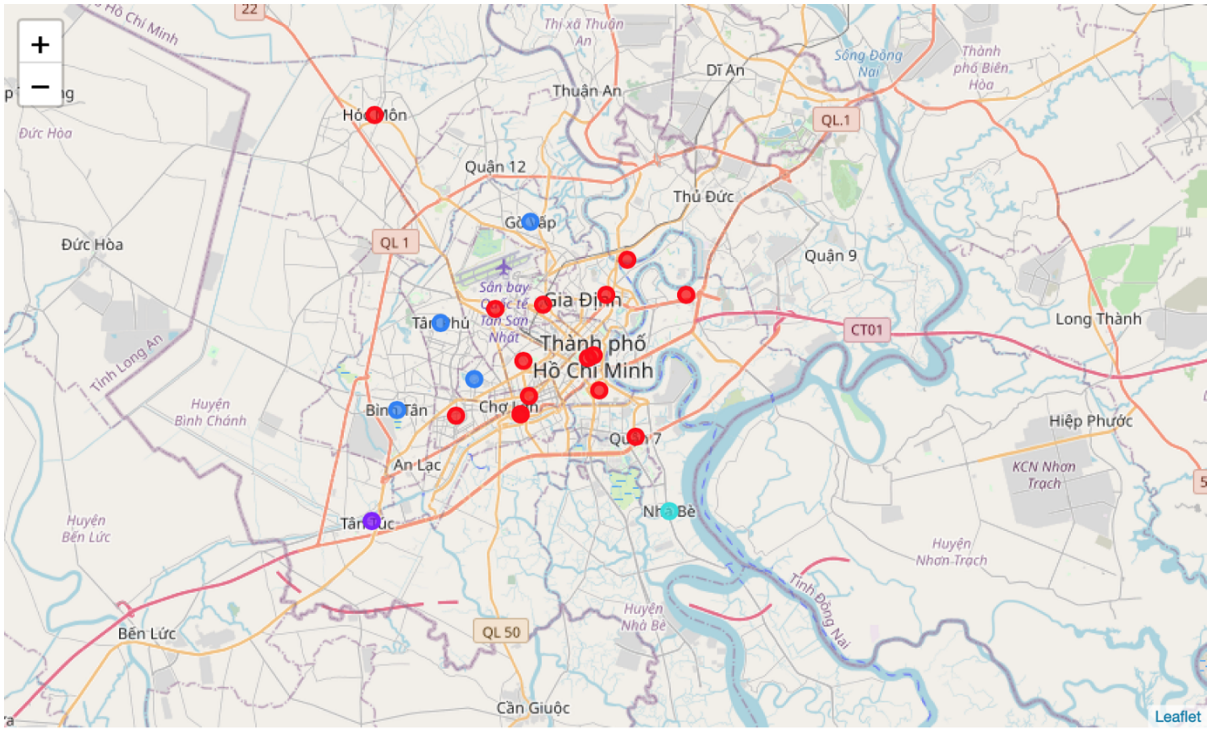
For the clustering process, the K-means approach was used, which is an unsupervised machine learning algorithm. This process also requires to set the parameter for the number of clusters. To be able to identify the optimal number for this parameter, the silhouette score was used. I used visualization of clusters and calculate silhouette score of each cluster quantity, which leads to a conclusion of optimal value. The optimal number of cluster is 2.



Visualization of clusters

```
For n_clusters = 2, silhouette score is 0.31766326058275424)
For n_clusters = 3, silhouette score is 0.2934693104030061)
For n_clusters = 4, silhouette score is 0.2769531214545777)
For n_clusters = 5, silhouette score is 0.21291435505245665)
For n_clusters = 6, silhouette score is 0.16068362755092586)
For n_clusters = 7, silhouette score is 0.1523873203372235)
```

I plot data with folio. Let's look at the following map



5. Results

By looking at the cluster data, we can see that cluster 1 is the one that we are the most interested in. Cluster 1 (Cluster label 0) is the biggest cluster. Recall the criteria that we set at the beginning of this analysis and evaluate:

- Accessibility for tourists and local citizens (transportation): Among the top common venue are hotels, which is ideal for attracting tourists. Bus stops are nearby, rendering easiness of access to our restaurant.
- Nearby competitors: Although there are a number of restaurants (Vietnamese restaurants, Dim Sum restaurants,…), there are not many café, especially all-day brunch with American and Asian fusion vegetarian style.
- Metropolitan area: There are flea markets, markets, parks, gym centers, multiplexes, convenience stores and so forth. It is a promising point of highly populated areas. It leads to opportunity of large customer number coming to our newly open restaurant.

The second cluster (Cluster label 1) is an outer district where café is not really represented (restaurants and bars are among the top).

6. Recommendation

Based on what we learned about the clusters, we can advise the restaurant owner to consider the districts from cluster 1 as a prospective location for the new restaurant. These are the districts where our planned business model is well represented and also hotels are frequent. These satisfy the three original criteria that the location should be in a metropolitan centre, with less competition and in a location that customers can easily find.

7. Conclusion

This report discussed the process of finding an answer for a hypothetical though real-life like business concern. The analysis was performed based mainly on the toolset of data science and the utilization of Python and Python libraries such as Pandas, Scikit, Folium, and so on. The output of the analysis offers recommendations for the business problem in question.

Final note

I hope you found my analysis interesting and it encourages you to dig deeper in this field too. I highly recommend the IBM Data Science course as I enjoyed learning from it greatly.

For more information of coding, I attached my Jupyter notebook here:

https://github.com/ngancrystal/github-example/blob/master/Courserafinalproject.ipynb

My report here:

https://github.com/ngancrystal/github-example/blob/master/IBM%20Data%20Science%20Course%20Capstone.pdf