IBM – Coursera

Data Science Specialization

Capstone project - Final report

Opening a New Spa in Ha Noi, Vietnam

Nguyễn Xuân Toản – 2019

I. Introduction:

This report is for the final course of the Data Science Specialization. A 9-courses series created by IBM, hosted on Coursera platform. The problem and the analysis approach are left for the learner to decide, with a requirement of leveraging the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

Hanoi is the capital and one of the five municipalities of Vietnam. Covering an area of 3,328.9 square kilometres (1,285 sq mi), it is the largest city in Vietnam by area. With an estimated population of 7.7 million as of 2018, it is the second largest city in Vietnam by population. The metropolitan area, encompassing nine additional neighboring provinces, has an estimated population of 16 million. Located in the central area of the Red River Delta, Hanoi is the commercial, cultural, and educational centre of Northern Vietnam. Having an estimated nominal GDP of US$32.8 billion, it is the second most productive economic centre of Vietnam, following Ho Chi Minh City. *(https://en.wikipedia.org/wiki/Hanoi).*

With 7.7 million of population, the need of place for leisure is very high, especially Spa. Of course, as with any business decision, opening new Spa requires serious consideration, the location of the Spa is one of the most important decisions that will determine whether the Spa will be success or failure.

II. Business Problem:

The objective of this final assignment is to analyze and select best locations Ha Noi to open new Spa. I use data science methodology and machine learning techniques that have learned through 9 course in IBM Data Science Professional Certificate such as data collection, data cleaning, Foursquare

location data, clustering to answer the business question: where to open new coffee how in Hanoi. Of course, In the scope of capstone, I just mention some factors that effect the result. In order to launch this in real, we have to improve more. With this report, I show that I can use Python data science tools to solve business problem.

III. The target audience for this capstone project:

The target audience for this report are:

- Developers and investors looking to open new Spa in Ha Noi.
- And of course, to this course's instructors and learners who will grade this project. Or to anyone who catch this shared on the social media showing that I can use Python data science tools.

IV. Data description:

To solve this question, I need data:

- List of District (neighborhoods) in Ha Noi.
- Latitude and longitude coordinates of those neighborhoods to plot map and to get the venue data.
- Venue data include Spa data to perform clustering on the neighborhoods.

Source of Data and Methods to extract:

Because, there is not detail data about districts of Ha Noi, so I make CSV file HanoiCSV.txt from information in wiki (https://en.wikipedia.org/wiki/Hanoi). In this project I used only inner districts. There are 13 districts in Hanoi with fields (name of the district, Area of the district and population).

After that, I use Python Geocoder Package to get latitude and longitude coordinates of the neighborhoods.

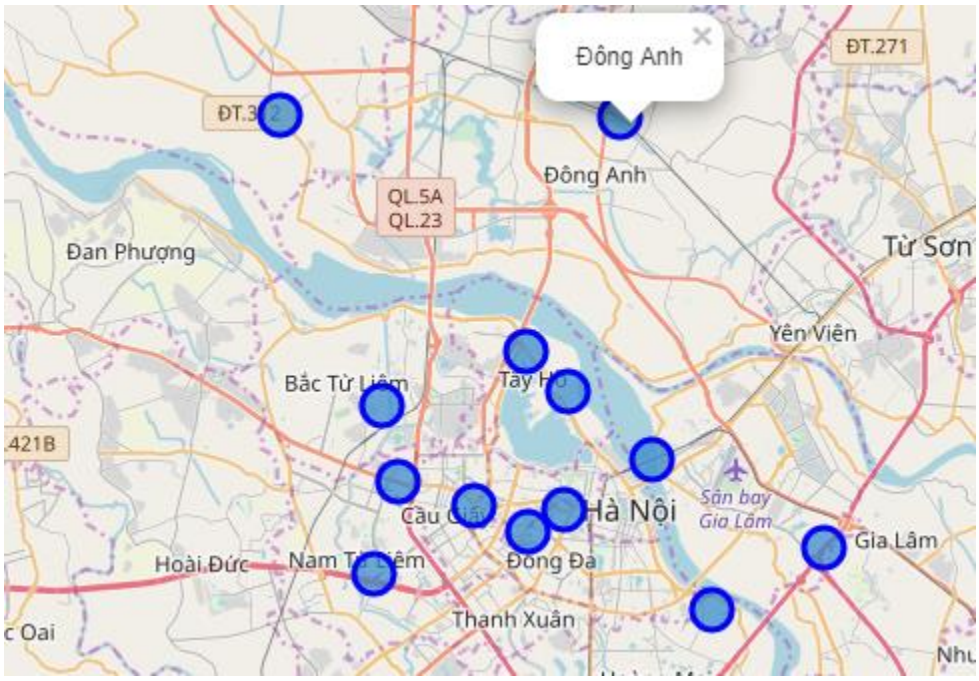
(https://raw.githubusercontent.com/nxtacctau/Coursera_Capstone/master/District_withGeo.csv)

I used folium to create a map of Hanoi using Latitude and longitude.

Next, I use Foursquare API to get the venue data for those Districts. I concentrated on Spa category in order to solve the business problem mentioned above.

V. Methodology:

First, we need the list of neighborhoods in Hanoi. Then I got geographical coordinates in the form of longitude and latitude by using the wonderful Geocoder package that allow us to get geographical coordinates (longitude and latitude) from address. I use folium to create and mark the neighborhoods in map. *(Picture below).*



Next, I used Foursquare API to get the top 100 venues within a radius of 7kms. I used information from Foursquare account created in the previous assignment.

```
# 5. Use the Foursquare API to explore the neighborhoods

# define Foursquare Credentials and Version
CLIENT_ID = '5VKNRCQUOZLNWJZX2XIVK2RVDAJBCPTPUHZT1BJGE12I
CLIENT_SECRET = 'LLOOJEFION04GHZ5XQ4MIS4MP45MO3QUMUKGIXHI
VERSION = '20180605' # Foursquare API version

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

I got data from Foursquare, then I grouped the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, I was also preparing the data for use in clustering. Since I analyzed the "Spa" data, I filter the "Spa" as venue category for the neighborhoods.

```
In [54]: hn_mall = hn_grouped[["neighborhood","Spa"]]
         hn_mall.head()
```

Out[54]:

	neighborhood	Spa
0	Ba Đình	0.030000
1	Bắc Từ Liêm	0.000000
2	Cầu Giấy	0.030000
3	Gia Lâm	0.018868
4	Hai Bà Trưng	0.000000

Lastly, I performed clustering on the data by using K-means clustering. The results will allow us to identify which Districts have higher concentration of Spa. Based on the information I got, it helped me to answer the question as to which districts are most suitable for opening new Spa.

VI. Results:

The results from k-means clustering show that we can categorize the districts into 3 clusters based on the frequency of occurrence of "Spa".

Cluster 0: District with moderate number of Spas

```
hn_merged_SM.loc[hn_merged_SM['Cluster Labels'] == 0]
```

	neighborhood	Spa	Cluster Labels	Latitude	Longitude
3	Gia Lâm	0.018868	0	21.016521	105.921027
8	Nam Từ Liêm	0.010000	0	21.008130	105.766500
12	Đống Đa	0.020000	0	20.996983	105.882695

Cluster 1: District with low number to no existence of Spas

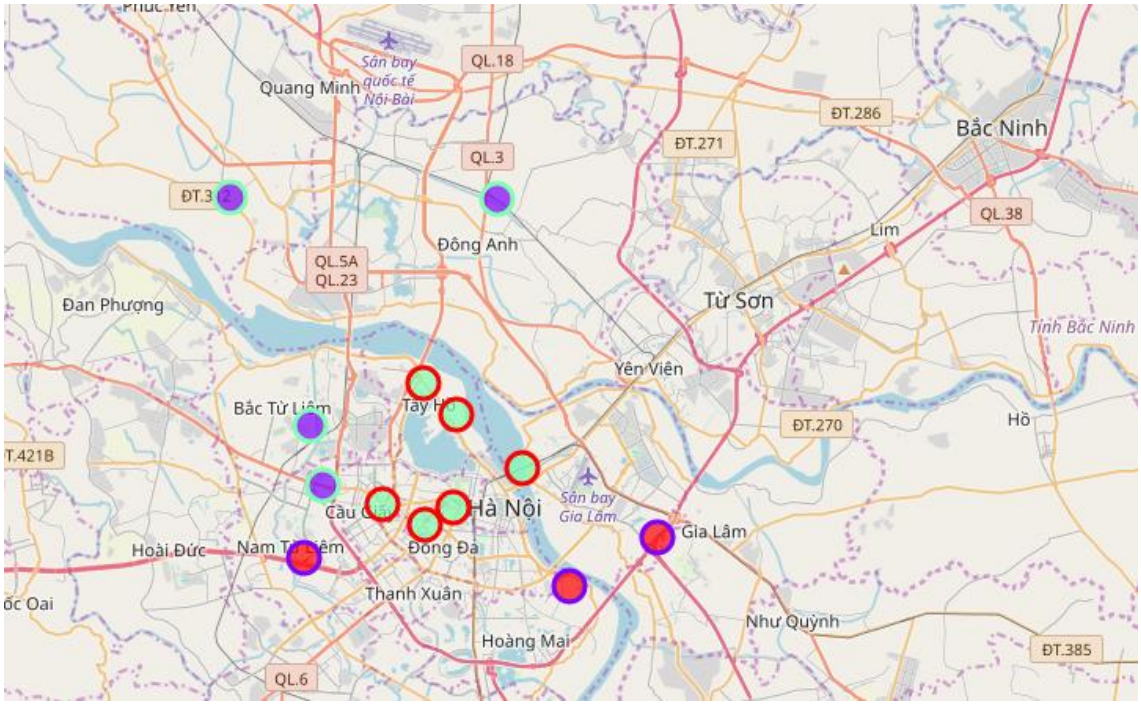
```
hn_merged_SM.loc[hn_merged_SM['Cluster Labels'] == 1]
```

	neighborhood	Spa	Cluster Labels	Latitude	Longitude
1	Bắc Từ Liêm	0.0	1	21.06217	105.76941
4	Hai Bà Trưng	0.0	1	21.15505	105.73429
9	Thanh Xuân	0.0	1	21.03760	105.77507
11	Đông Anh	0.0	1	21.15418	105.85131

Cluster 2: District with high number of Spas

```
hn_merged_SM.loc[hn_merged_SM['Cluster Labels'] == 2]
```

	neighborhood	Spa	Cluster Labels	Latitude	Longitude
0	Ba Đình	0.03	2	21.022010	105.819340
2	Cầu Giấy	0.03	2	21.030245	105.801359
5	Hoàn Kiếm	0.03	2	21.079070	105.819280
6	Hoàng Mai	0.03	2	21.028580	105.832070
7	Long Biên	0.04	2	21.044537	105.861879
10	Tây Hồ	0.03	2	21.066670	105.833330



VII. Discussion:

As we can see from the picture in Results section, most of the Spas are in the district that located in center of Ha noi.

So If you want to develop new Spa, I recommend you donot open in Center of Ha noi (cluster 2) already high concentration of Spa and intense competitions.

Open new Spa in neighborhoods in Cluster 1: No existence of Spa, so there isnt any competition.

VIII. Conclusion:

In this project, I only use 1 factor as frequency of occurrence of Spa to find out the answer for the question from business. There are other factors such as population, income, etc... that could influence the decision for investors to invest in Spa.

However, through this project I show that with data science methodology, toolbox we can use to solve real business problems, and of course we need invest more money to get data and register the paid account in Foursquare to get more information.

Thank you!