IBM – Coursera

Data Science Specialization

Capstone project - Final report

Opening a New Coffee House 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 coffee house. Of course, as with any business decision, opening new coffee house requires serious consideration, the location of the coffee house is one of the most important decisions that will determine whether the coffee house 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 coffee house. 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 coffee house 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 coffee house 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 HanoiCSV1.txt from information in wiki (https://en.wikipedia.org/wiki/Hanoi). There are 30 district 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. With 2 data tables above, I merge them into 1 file (District_withGeo.csv) contains 5 fields (Dist., Area, Pop, Latitude, Longitude) (I renamed columns to easy understanding and processing).

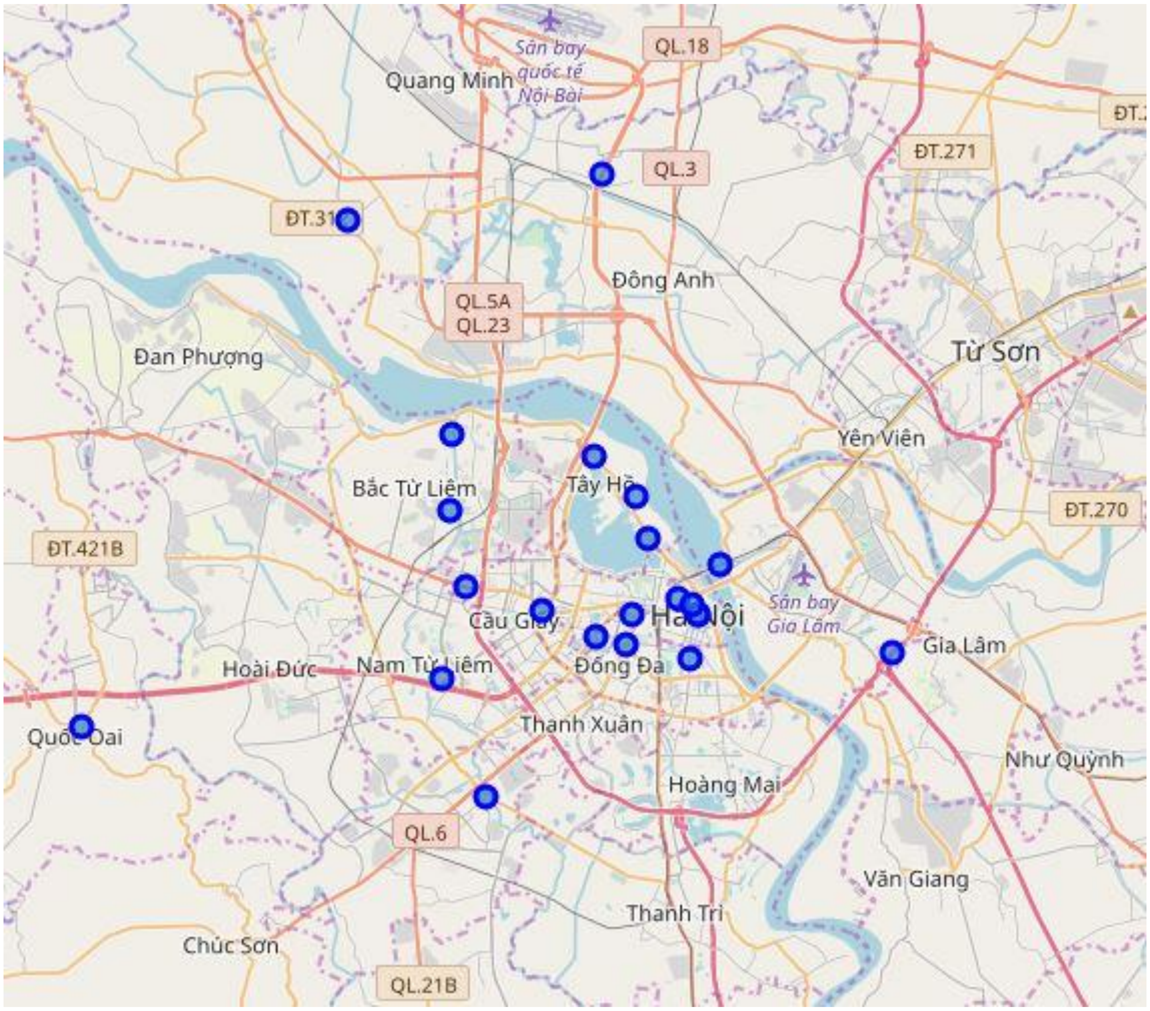
(https://raw.githubusercontent.com/nxtacctau/Coursera_Capstone/master/District_withGeo.csv)

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

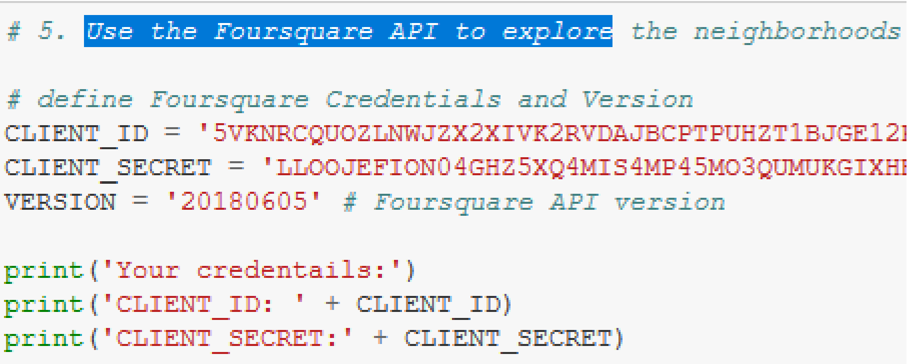
Next, I use Foursquare API to get the venue data for those Districts. I concentrated on Cafe category in order to solve the business problem mentioned above.

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).*

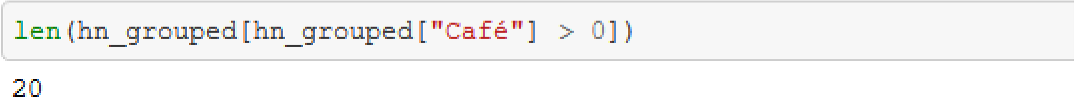


Next, I used Foursquare API to get the top 100 venues within a radius of 2kms. I used information from Foursquare account I created in the previous assignment.



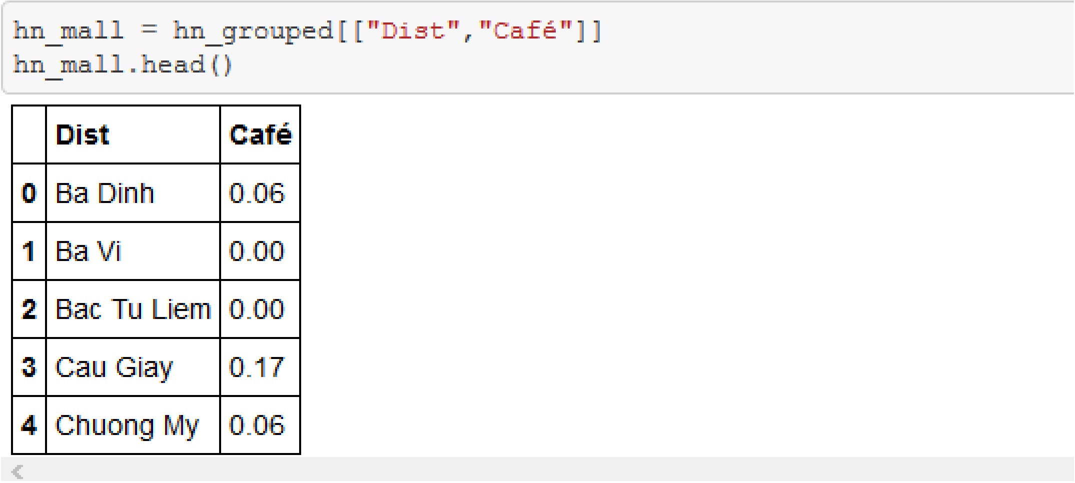
```
# 5. Use the Foursquare API to explore the neighborhoods

# define Foursquare Credentials and Version
CLIENT_ID = '5VKNRCQUOZLNWJZX2XIVK2RVDAJBCPTPUHZT1BJGE12
CLIENT_SECRET = 'LLOOJEFION04GHZ5XQ4MIS4MP45MO3QUMUKGIXH
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 "Cafe" data, I filter the "Cafe" as venue category for the neighborhoods.



```
len(hn_grouped[hn_grouped["Café"] > 0])
```

```
20
```



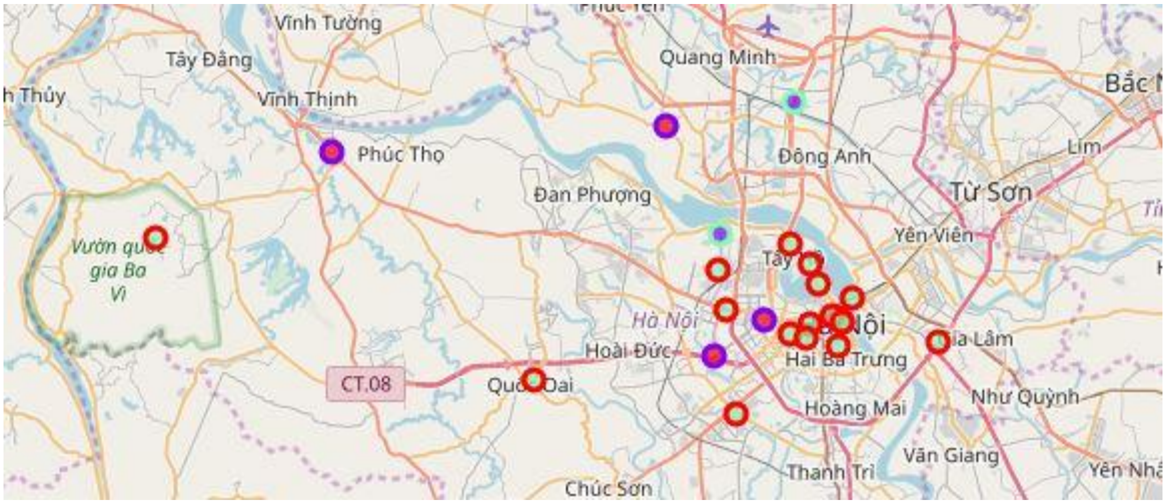
```
hn_mall = hn_grouped[["Dist","Café"]]
hn_mall.head()
```

	Dist	Café
0	Ba Dinh	0.06
1	Ba Vi	0.00
2	Bac Tu Liem	0.00
3	Cau Giay	0.17
4	Chuong My	0.06

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 coffee house. Based on the information we got, it helped us to answer the question as to which districts are most suitable for opening new coffee house.

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 "Coffee house".



VII. Discussion:

As we can see from the picture in Results section, most of the coffee house are in the district that not located in center of Hanoi. We can guess because of the rent price of properties in center Hanoi is very high so that investor did not invest in coffee house in the center of Ha Noi.

So for investors, it is not wise to invest in coffee house in Hanoi.

VIII. Conclusion:

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

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

Thank you!