Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Restaurant in Hanoi, Vietnam

By: Duy Quang Bui

January 2021

Introduction

Most of us eat out at least once a week, and this is why the restaurant industry is a lucrative business and is thriving perpetually in Vietnam. And of course, we all love to eat and will continue to enjoy food prepared outside of the home. This is also another reason starting a restaurant business in Vietnam is among the most amazing business ideas.

Although the COVID-19 epidemic is causing certain difficulties in relocation decisions and activities, Vietnam is still considered an attractive investment destination for many multinational companies in the world. Currently, doing catering business is a potential investment sector which is growing rapidly. Tastes and trend of Vietnamese diners are diverse and variable. They are eager to try new dishes. In large cities, it is not hard to find restaurants owned by foreigners which provides their traditional dishes of many nations such as: France, Italy, Japan, Korea, Singapore, Thailand, China, India… No matters how the economy changes, our food demand is an everyday issue.

But, you need to know that opening a restaurant requires a lot of effort. Running a medium and above restaurant requires a huge capital and year-long pay pack period as well as an effective marketing plan in order to compete with many competitors. Many foreigners have been living in Vietnam for a long time wish to open restaurants here to serve their traditional dishes. It is not only passion in culinary arts but also their pride helping them through homesick. Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurant is one of the most important decisions that will determine whether it will be a success or a failure.

Business Problem

Suppose you want to open a restaurant in Hanoi, and you are wondering where is a suitable location to open a restaurant. You don't want to open a restaurant in a place that already has too many restaurants and want to have an overview of the restaurants that have opened in Hanoi.

The objective of this capstone project is to analyze and select the best locations in Hanoi, capital of Vietnam, to open a new restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Hanoi, Vietnam, if someone is looking to open a restaurant, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new restaurants in Hanoi, the capital city of Vietnam.

Data

To solve the problem, we will need the following data:

- List of neighborhoods in Hanoi, Vietnam. This defines the scope of this project which is confined to Hanoi, the capital city of Vietnam.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

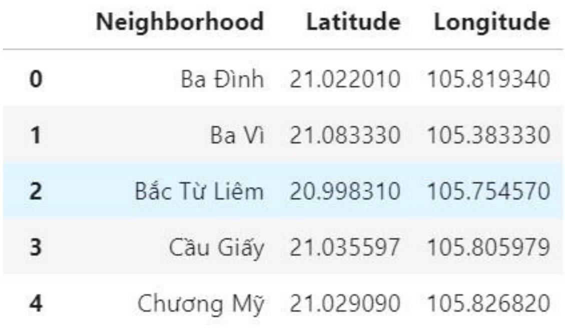
Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Districts_of_Hanoi) contains a list of neighborhoods in Kuala Lumpur, with a total of 70 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and BeautifulSoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.
Foursquare API will provide many categories of the venue data, we are particularly interested in the Restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

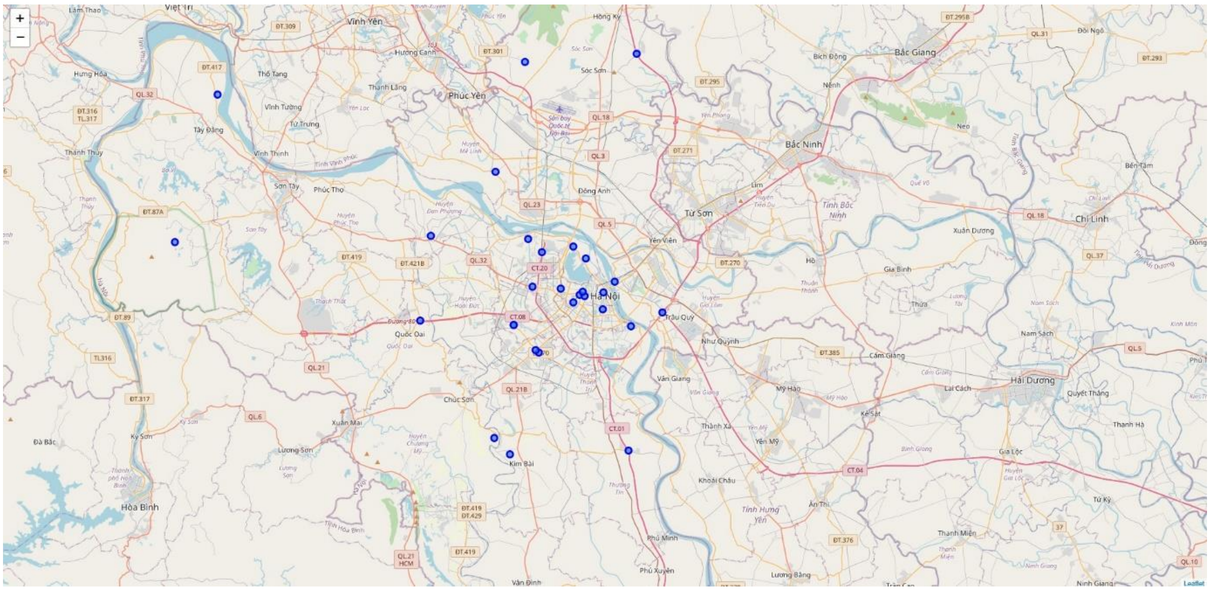
Methodology

Firstly, we need to get the list of neighborhood in the Hanoi. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Districts_of_Hanoi). We will do web scraping using Python requests and BeautifulSoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in Hanoi.



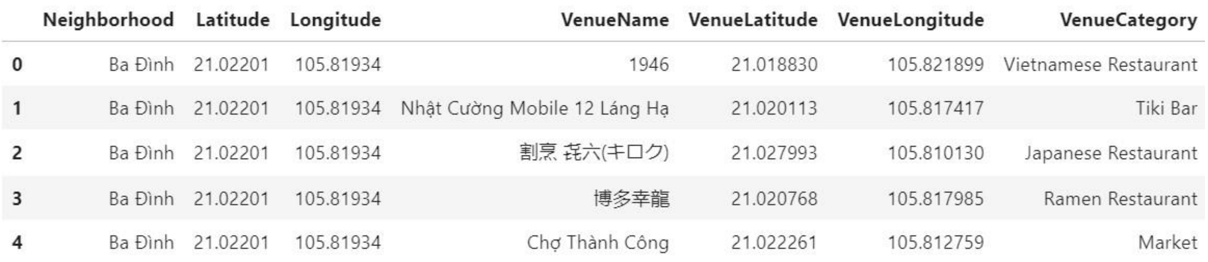
	Neighborhood	Latitude	Longitude
0	Ba Đình	21.022010	105.819340
1	Ba Vì	21.083330	105.383330
2	Bắc Từ Liêm	20.998310	105.754570
3	Cầu Giấy	21.035597	105.805979
4	Chương Mỹ	21.029090	105.826820

Then we use python folium library to visualize geographic details of Hanoi and its neighborhood as below.



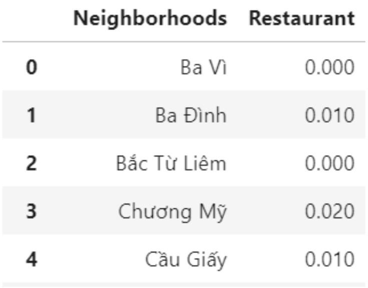
Next, we will use Foursquare API to get the top **100 venues** that are within a radius of **2000 meters.**

We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude.



	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Ba Đình	21.02201	105.81934	1946	21.018830	105.821899	Vietnamese Restaurant
1	Ba Đình	21.02201	105.81934	Nhật Cường Mobile 12 Láng Hạ	21.020113	105.817417	Tiki Bar
2	Ba Đình	21.02201	105.81934	割烹 㐂六(キロク)	21.027993	105.810130	Japanese Restaurant
3	Ba Đình	21.02201	105.81934	博多幸龍	21.020768	105.817985	Ramen Restaurant
4	Ba Đình	21.02201	105.81934	Chợ Thành Công	21.022261	105.812759	Market

With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the "Restaurant" data, we will filter the "Restaurant" as venue category for the neighborhoods.



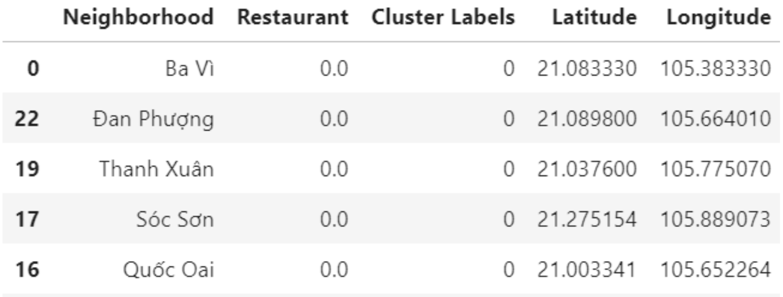
	Neighborhoods	Restaurant
0	Ba Vì	0.000
1	Ba Đình	0.010
2	Bắc Từ Liêm	0.000
3	Chương Mỹ	0.020
4	Cầu Giấy	0.010

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for "Restaurant". The results will allow us to identify which neighborhoods have higher concentration of restaurants while which neighborhoods have fewer number of restaurants. Based on the occurrence of restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results

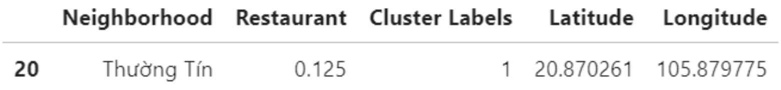
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Restaurant":

- Cluster 0: Neighborhoods with low number of restaurants



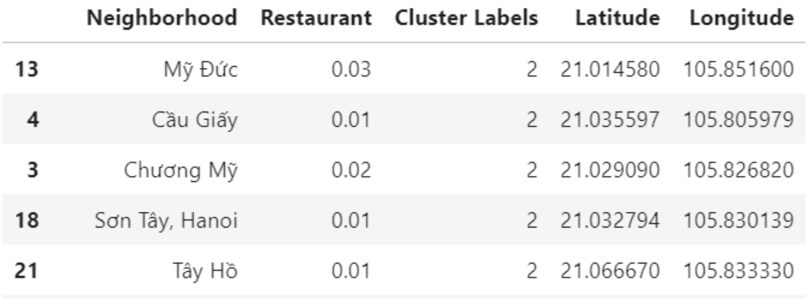
	Neighborhood	Restaurant	Cluster Labels	Latitude	Longitude
0	Ba Vì	0.0	0	21.083330	105.383330
22	Đan Phượng	0.0	0	21.089800	105.664010
19	Thanh Xuân	0.0	0	21.037600	105.775070
17	Sóc Sơn	0.0	0	21.275154	105.889073
16	Quốc Oai	0.0	0	21.003341	105.652264

- Cluster 1: Neighborhoods with moderate number of restaurants



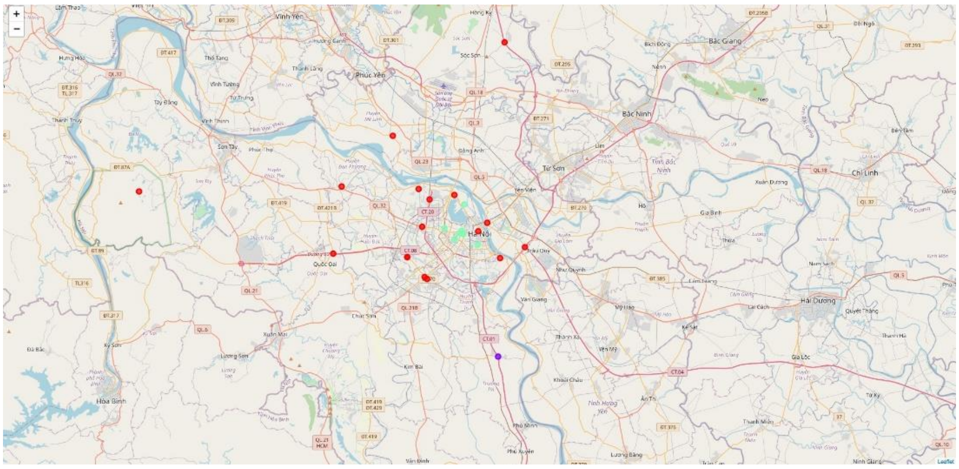
	Neighborhood	Restaurant	Cluster Labels	Latitude	Longitude
20	Thường Tín	0.125	1	20.870261	105.879775

- Cluster 2: Neighborhoods with high concentration of restaurants



	Neighborhood	Restaurant	Cluster Labels	Latitude	Longitude
13	Mỹ Đức	0.03	2	21.014580	105.851600
4	Cầu Giấy	0.01	2	21.035597	105.805979
3	Chương Mỹ	0.02	2	21.029090	105.826820
18	Sơn Tây, Hanoi	0.01	2	21.032794	105.830139
21	Tây Hồ	0.01	2	21.066670	105.833330

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in blue color, and cluster 2 in green color.



Discussion

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Hanoi, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new restaurants as there is very little to no competition from existing restaurants. Meanwhile, restaurants in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of restaurants.

From another perspective, the results also show that the oversupply of restaurants mostly happened in the central area of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open restaurants in neighborhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new restaurants in neighborhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor frequency of occurrence of restaurants, there are other factors such as population and income of residents that could influence the location decision of a new restaurant. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new restaurant. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders property developers and investors regarding the best locations to open a new restaurant.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant.

References

[1] District of Hanoi – Wikipedia

[2] Foursquare API