Coursera Capstone

Opening a New Asian Restaurant in Ho Chi Minh City, VietNam

By: Tran Hong Vinh

Jun 2020

Introduction

Today, Vietnamese often eat at the restaurant. Life is so busy so it makes cooking at home and cleaning becomes a burden for many people, especially in big cities. So breakfast which is easy to buy at convenience stores or restaurants such as bread, sticky rice, noodles, wet bread ... is preferred. The Officers usually have more time so that they choose to have breakfast with friends at the cafe. Lunch time is short, only about 1 hour, so most people choose to order office lunch or invite each other to the restaurants. Only dinner or weekends left to spend time cooking family meals.

The trend of eating out has led to the increase of a variety of dining types such as restaurants and bars from luxurious to popular to serve different audiences. Korean and Japanese restaurants have poured into Vietnam market and despite the rise for nearly ten years, until now, it has attracted a lot of customers. Well-known brands are always in the right place at all times.

Opening asian restaurant allows to earn income. Of course, as with any business decision, opening a new asian restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the asian restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the Ho Chí Minh city, VietNam to open a new asian reataurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the Ho Chí Minh city, VietNam, if the businessman is looking to open a new asian restaurant, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to the businessman and investors looking to open or invest in new asian restaurant in the Ho Chí Minh city, VietNam. This project is timely as the city is currently suffering from oversupply of asian restaurant. According to the General Statistics Office just published, in the first month of 2020, accommodation and catering services revenue is estimated at over 45,000 billion VND, up 0.5% from the previous month and up 14.7%. over the same period in 2019.

Data

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

• List of neighbourhoods in Ho Chi Minh city. This defines the scope of this project which is confined to the Ho Chi Minh city.

• Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to asian restaurant. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_districts_of_Vietnam#H%E1%BB%93_Ch%C3%AD_Minh_City)contains a list of neighbourhoods in Ho Chi Minh city, with a total of 24 neighbourhoods. 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 neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. 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 Shopping Mall 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 neighbourhoods in the Ho Chi Minh city. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_districts_of_Vietnam#H%E1%BB%93_Ch%C3%AD_Minh_City). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods 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 wonderful 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 neighbourhoods 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 the Ho Chi Minh city.

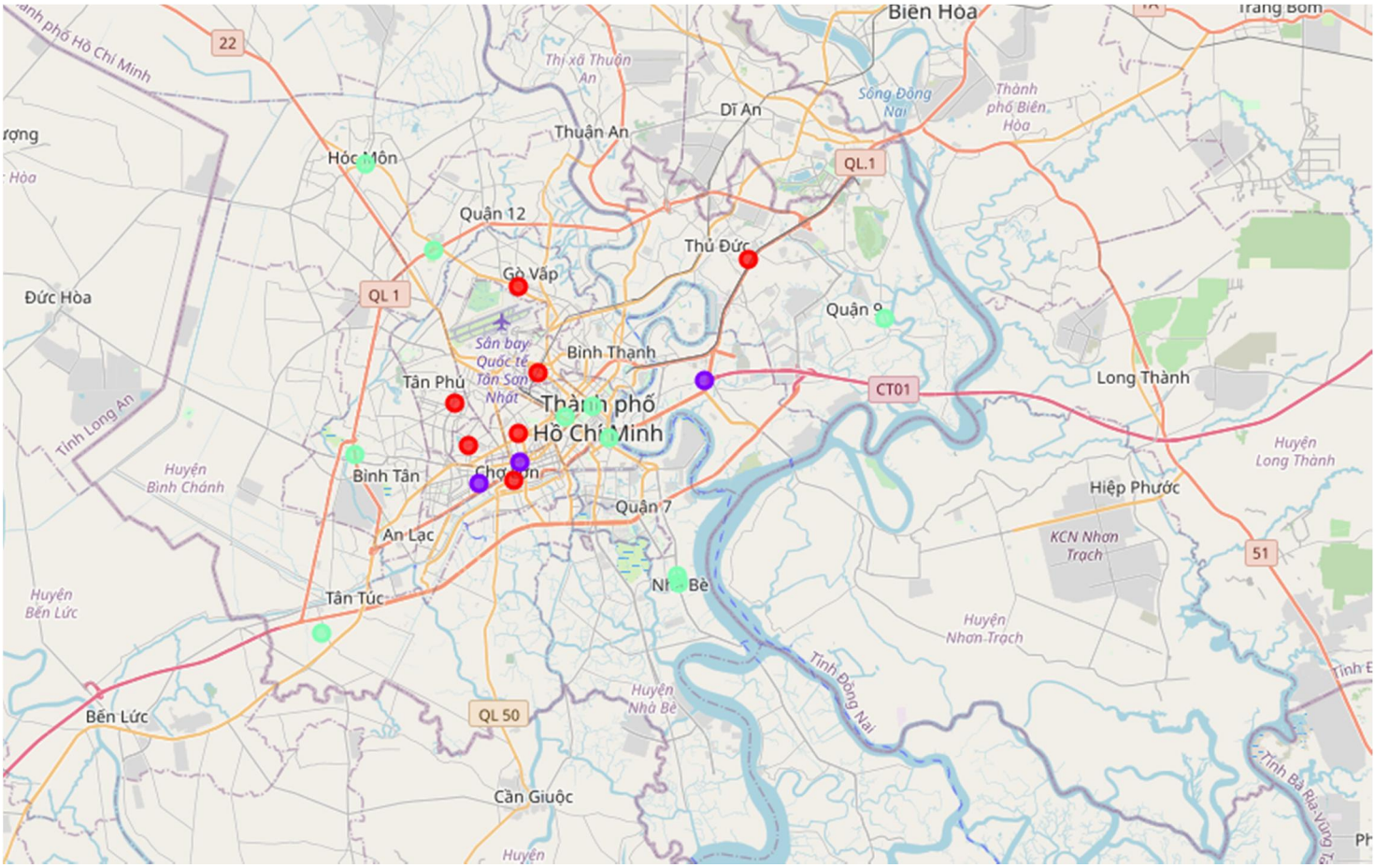
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods 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. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 analysing the “Asian Restaurant” data, we will filter the “Asian Restaurnat” as venue category for the neighbourhoods.
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 neighbourhoods into 3 clusters based on their frequency of occurrence for “Asian Restaurant”. The results will allow us to identify which neighbourhoods have higher concentration of asian restaurant while which neighbourhoods have fewer number of asian restaurant. Based on the occurrence of asian restaurant in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new asian restaurant

Results

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

- Cluster 0: Neighbourhoods with moderate number of Asian Restaurant
- Cluster 2: Neighbourhoods with low number to no existence of Asian Restaurant
- Cluster 1: Neighbourhoods with high concentration of Asian Restaurant

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



Discussion

As observations noted from the map in the Results section, most of the asian restaurant are concentrated in the District 6,5,2 with the highest number in cluster 1 and moderate number in cluster 0. On the other hand, cluster 2 has very low number to no asian restaurant in the neighbourhoods. This represents a great opportunity and high potential areas to open new asian restaurant as there is very little to no competition from existing malls. Meanwhile, asian restaurant in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of asian restaurant. From another perspective, the results also show that the oversupply of asian restaurant mostly happened in the center of the city, with the suburb area still have very few asian restaurant. Therefore, this project recommends property bussinessman to capitalize on these findings to open new asian restaurant in neighbourhoods in cluster 2 with little to no competition. Property bussinessman with unique selling propositions to stand out from the competition can also open new asian restaurant in neighbourhoods in cluster 0 with moderate competition. Lastly, property bussinessman are advised to avoid neighbourhoods in cluster 1 which already have high concentration of asian restaurant and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of asian restaurant, there are other factors such as population and income of residents that could influence the location decision of a new asian restaurant. However, to the best knowledge of this researcher such data are not available to the neighbourhood 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 asian 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 i.e. property developers and investors regarding the best locations to open a new asian restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to open a new asian restaurant. 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 asian restaurant.