IBM APPLIED DATA SCIENCE

CAPSTONE PROJECT

SETTING UP A COFFEE SHOP IN HO CHI MINH CITY, VIETNAM

BY: NGUYEN TUAN LE GIANG

INTRODUCTION

In Ho Chi Minh City – the South of Vietnam, the people usually drink coffee in the morning and coffee shop is the common place for meeting with customer or hangout with friends.

There are many kinds of coffee shop in Vietnam. It can be the small shop on the pavement or luxury coffee shop with big garden for children playing or group gathering for big event.

Figure 1: Small coffee shop on pavement and luxury coffee shop

For opening the coffee shop, the investment is not so high and does not require the big team so there were many start-ups or franchising to setup this business. Therefore, the competitiveness is the big challenge before doing it. In the other hand, the location of coffee shop is playing the important role that will determine whether the coffee shop will be a success or a failure.

BUSINESS PROBLEM

The objective of this capstone project is to analyze and select the best locations in HO Chi Minh City, Vietnam to make the report and give the consultant to the customer. By using the data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the Ho Chi Minh city, Vietnam if an investor is looking to open a new coffee shop, where would you recommend?

DATA

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

- List of districts in Ho Chi Minh city, Vietnam. This defines the scope of this project which is finding the best place to open the new coffee shop.
- Latitude and longitude coordinates of those districts. This is required in order to plot the map and also get the venue data.
- Venue data, specific data related to coffee shops. We shall use this data to perform clustering on the districts.

SOURCE OF DATA AND MEDTHODOLOGY TO EXTRACT THE DATA

The Wikipedia page https://en.wikipedia.org/wiki/Category:Districts_of_Ho_Chi_Minh_City contains a list of districts in Ho Chi Minh city, with a total of 24 districts. We shall use the web scrapping techniques to extract the data from Wikipedia page, with the help of Python requests and Beautifulsoup packages. Then we will get the geographical coordinates of the districts using Python Geocoder package which will give us the latitude and longitude coordinates of the districts.

After that, we will use Foursquare API to get the venue data for those districts. Foursquare has one of the largest databases 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 Coffee Shop or Café category in order to help us to solve the business problem. 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).

HO CHI MINH CITY MAP

This map represents all the 24 districts of the Ho Chi Minh City. Then we try to make clusters by using the Machine learning technique (K-means clustering) and group all the districts into different clusters.

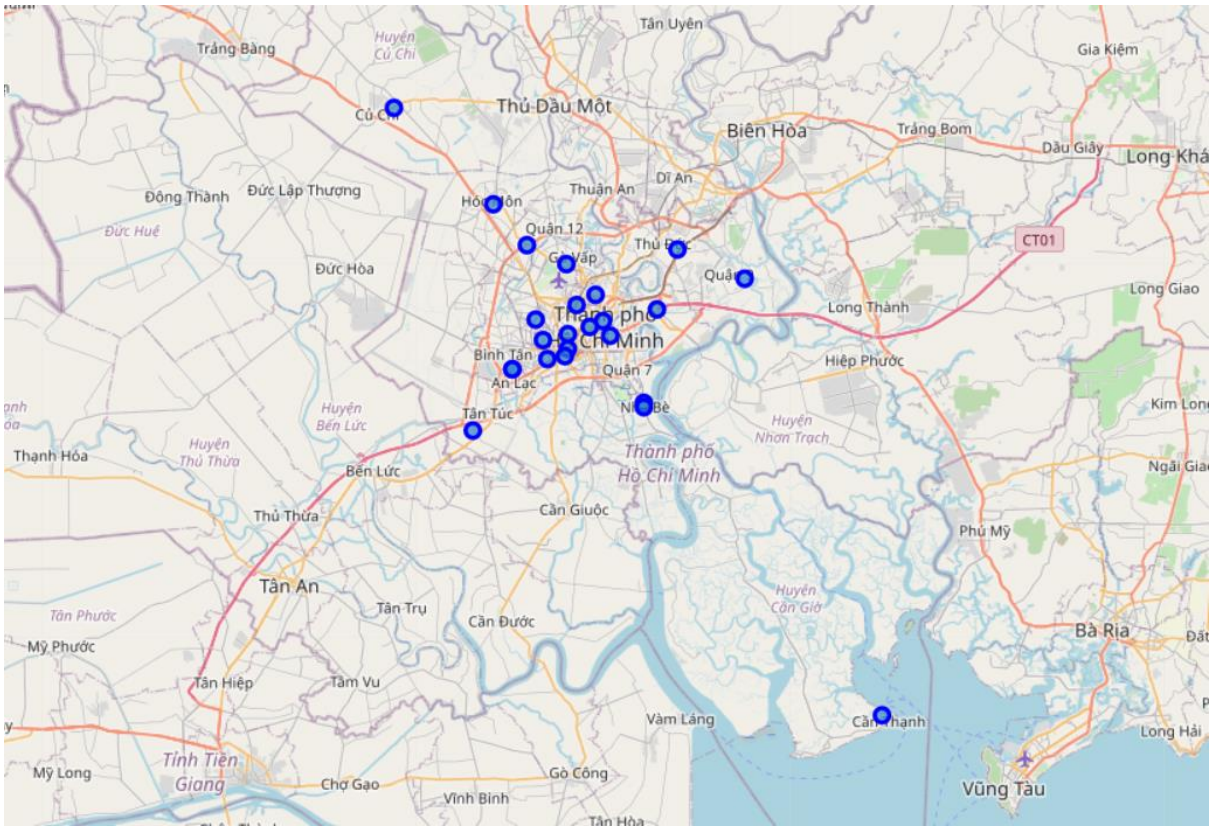

Figure 2: Map of Ho Chi Minh City

METHODOLOGY

Get the list of districts in Ho Chi Minh City. The list is available in the Wikipedia page https://en.wikipedia.org/wiki/Category:Districts_of_Ho_Chi_Minh_City

Do web scraping using Python requests and Beautifulsoup packages to extract the list of districts data. However, this is just a list of names.

Get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we use the wonderful Geocoder package that will allow us to convert the address into geographical coordinates in the form of latitude and longitude. After collecting the data, we will populate the data into a pandas DataFrame and then visualize the districts in a map by using Folium package. This

allows us to perform a sanity check to make sure that the geographical coordinate data returned by Geocoder are correctly plotted in Ho Chi Minh City map.

Use the Foursquare API to get the top 1000 venues that are within a radius of 20 kilometers. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. Then we make API calls to Foursquare passing in the geographical coordinates of the districts 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 district and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each district by grouping the rows by district 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 “Coffee shop” and “Café” data, we will filter the “Coffee shop” and “Café” as venue 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 districts into 4 clusters based on their frequency of occurrence for ““Coffee shop” and “Café”. The results will allow us to identify which districts have a higher concentration of coffee shop while which districts have a fewer number of shopping malls. Based on the occurrence of coffee shop in different districts, it will help us to answer the question which districts are most suitable to open new coffee shop. Therefore, this project recommends the investor to capitalize on these findings to open new coffee shop in districts in cluster 0 with less competition.

RESULT

The result from K-Means clustering showing:

- Cluster 0: These districts with a very less concentration of coffee shop
- Cluster 1: These districts with a high concentration of coffee shop
- Cluster 2: These districts with a high concentration of coffee shop

The result of clustering are visualized in the map below which cluster 0 is in red color; cluster 1 in pink color, cluster 3 is in green color.

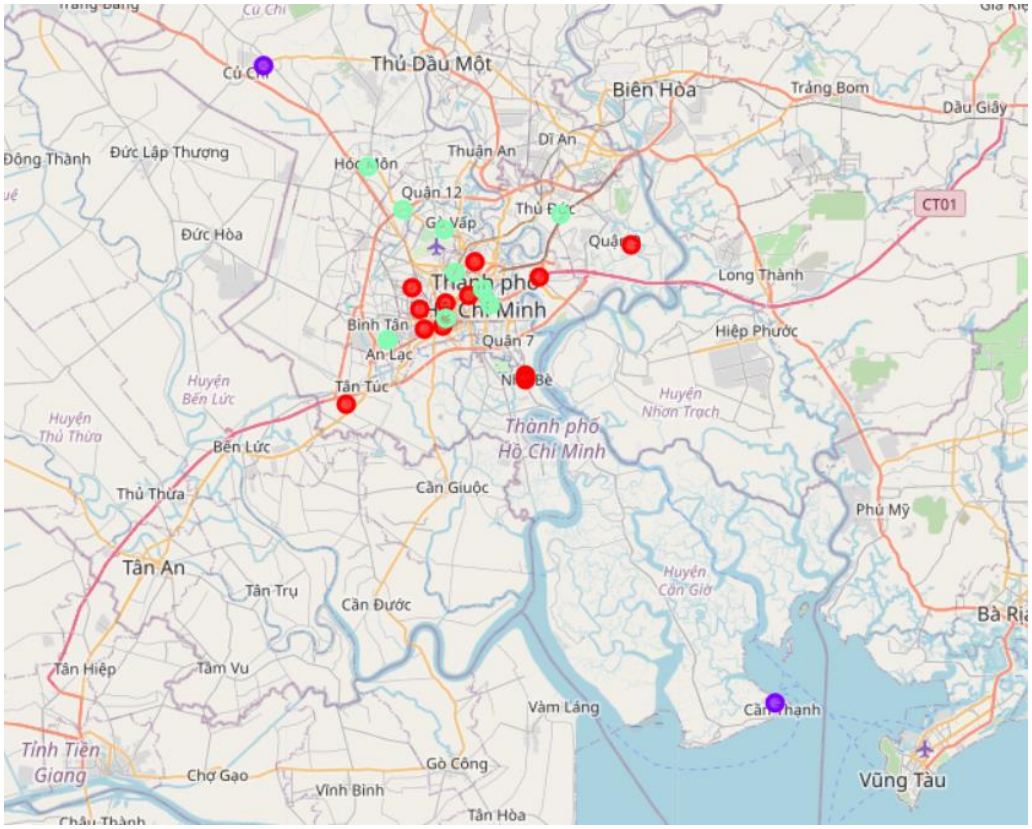

Figure 3: Clustering of coffee shop in Ho Chi Minh City

CONCLUSION

As the observation noted in Figure 3, most of coffee shop are concentrated in central area of Ho Chi Minh City, with the highest number in cluster 1 and cluster 2. It is showing that coffee time is the culture of Vietnamese people in the South of country. Therefore, the coffee shop are distributed everywhere in Ho Chi Minh City. Only cluster 0 has a very low number (19). It means that Cu Chi and Can Gio has the great opportunity and high potential to open the new coffee shop. However, these areas have less population and lower income, therefore, to open the luxury coffee shop is

impossible. The investor can consider to setup the normal coffee shop with reasonable price to attract the customer.

REFERENCE

- https://codekarim.com/node/55
- https://realpython.com/beautiful-soup-web-scraper-python/
- https://realpython.com/python-web-scraping-practical-introduction/