Applying K-Means Machine Learning Algorithm to choose the best location for coffee shop start-up in Ho Chi Minh City, Vietnam

Phan Bao Dung

January 07, 2021

1. Introduction

1.1 Background

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. 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.

1.2 Problem

The objective of this article is to report the methodology to analyze and select the best locations in HO Chi Minh City, Vietnam for coffee shop start-up. By using the data science and machine learning techniques, 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?

2. Data acquisition and cleaning

2.1 Data sources

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. I shall use the web

scrapping techniques to extract the data from Wikipedia page, with the help of Python requests and Beautifulsoup packages.

	Neighborhood
0	Bình Chánh District
1	Bình Tân District, Ho Chi Minh City
2	Bình Thạnh District
3	Cần Giờ District
4	Củ Chi District
5	District 1, Ho Chi Minh City
6	District 3, Ho Chi Minh City
7	District 4, Ho Chi Minh City
8	District 5, Ho Chi Minh City
9	District 6, Ho Chi Minh City
10	District 7, Ho Chi Minh City

Figure 1. List of district in Ho Chi Minh city, Vietnam

Then I will get the geographical coordinates of the districts using Python Geocoder package which will give us the latitude and longitude coordinates of the districts.

```
[[10.679220000000043, 106.57654000000008],
 [10.73684000000003, 106.61448000000007],
 [10.806080000000065, 106.69297000000006],
 [10.41566000000006, 106.96130000000005],
 [10.977340000000027, 106.50223000000005],
 [10.78096000000005, 106.69911000000008],
 [10.775650000000041, 106.68672000000004],
 [10.766700000000071, 106.70647000000008],
 [10.755690000000072, 106.66637000000009],
 [10.745970000000057, 106.64769000000007],
 [10.70515000000006, 106.73748000000006],
 [10.74771000000004, 106.66334000000006],
 [10.768830000000037, 106.66599000000008],
 [10.763160000000028, 106.64314000000007],
 [10.850440000000049, 106.62731000000008],
 [10.833790000000022, 106.66556000000008],
 [10.888360000000034, 106.59640000000007],
 [10.701530000000048, 106.73818000000006],
 [10.795650000000023, 106.67464000000007],
 [10.73684000000003, 106.61448000000007],
 [10.782320000000027, 106.63667000000004],
 [10.861779986287589, 106.79610692772711]]
```

Figure 2. List of coordinator of districts in Ho Chi Minh city, Vietnam

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. I merge the list of district into its coordinator (latitude and longitude).

	Neighborhood	Latitude	Longitude
0	Bình Chánh District	10.67922	106.576540
1	Bình Tân District, Ho Chi Minh City	10.73684	106.614480
2	Bình Thạnh District	10.80608	106.692970
3	Cần Giờ District	10.41566	106.961300
4	Củ Chi District	10.97734	106.502230
5	District 1, Ho Chi Minh City	10.78096	106.699110
6	District 3, Ho Chi Minh City	10.77565	106.686720
7	District 4, Ho Chi Minh City	10.76670	106.706470
8	District 5, Ho Chi Minh City	10.75569	106.666370
9	District 6, Ho Chi Minh City	10.74597	106.647690
10	District 7, Ho Chi Minh City	10.70515	106.737480

Figure 3. List of district and its coordinator

3. Exploratory Data Analysis

3.1 Visualize the location

I used Folium to show the location of each district on the map. Then I use the Foursquare API to explore the district data in Ho Chi Minh city, Vietnam. 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.

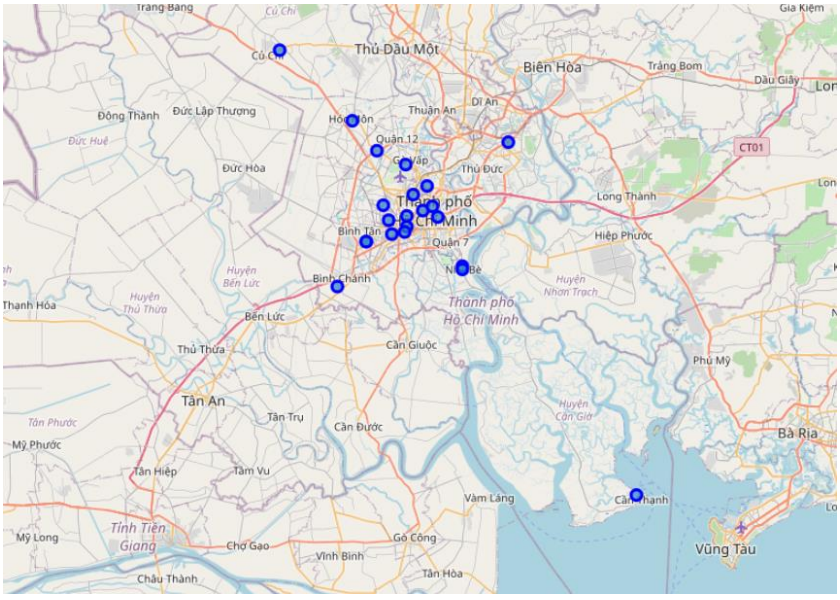

Figure 4. Map of Ho Chi Minh city, Vietnam

3.2 Group the data

I group the venues into categories and map them to the district name accordingly.

```
[22] #check out how many unique categories can be curated from all the returned values
     print('There are {} uniques categories.'.format(len(venues_df['VenueCategory'].unique())))

     There are 83 uniques categories.
```

```
[23] # displaying the first 50 Venue Category names
     venues_df['VenueCategory'].unique()[:50]

     array(['Shopping Mall', 'Multiplex', 'Bed & Breakfast', 'Dessert Shop',
            'Hotel', 'Coffee Shop', 'BBQ Joint', 'Pizza Place', 'Park',
            'Deli / Bodega', 'Sushi Restaurant', 'Vietnamese Restaurant',
            'Café', 'Whisky Bar', 'Noodle House', 'Indian Restaurant',
            'Vegetarian / Vegan Restaurant', 'Flower Shop', 'Beer Bar',
            'Asian Restaurant', 'Supermarket', 'Department Store',
            'Sandwich Place', 'Brewery', 'Hotel Bar', 'Tapas Restaurant',
            'Bar', 'Steakhouse', 'Massage Studio', 'Seafood Restaurant',
            'French Restaurant', 'German Restaurant', 'Italian Restaurant',
            'Spa', 'Burger Joint', 'Japanese Restaurant', 'Bookstore',
            'Nightclub', 'Hotpot Restaurant', 'Bistro', 'Clothing Store',
            'Ramen Restaurant', 'Middle Eastern Restaurant',
            'Korean Restaurant', 'Lounge', 'Golf Course', 'Mexican Restaurant',
            'Chinese Restaurant', 'Public Art', 'Health & Beauty Service'],
           dtype=object)
```

Figure 5. Data categories in Ho Chi Minh City, Vietnam

After that, I filter the categories from dataset; choose the “Coffee shop” or “Café” and create the data frame.

	Neighborhoods	Coffee Shop	Café
0	Bình Chánh District	4	7
1	Bình Thạnh District	2	7
2	Bình Tân District, Ho Chi Minh City	4	7
3	Cần Giờ District	2	8
4	Củ Chi District	0	9
5	District 1, Ho Chi Minh City	2	6
6	District 10, Ho Chi Minh City	3	6
7	District 11, Ho Chi Minh City	3	7
8	District 12, Ho Chi Minh City	3	7
9	District 3, Ho Chi Minh City	3	6
10	District 4, Ho Chi Minh City	2	6

Figure 6. List of district with detail of “Coffee shop” and “Café”

4. Classification modeling

I used the Clustering algorithm K-Means to classify the dataset into 3 clusters.

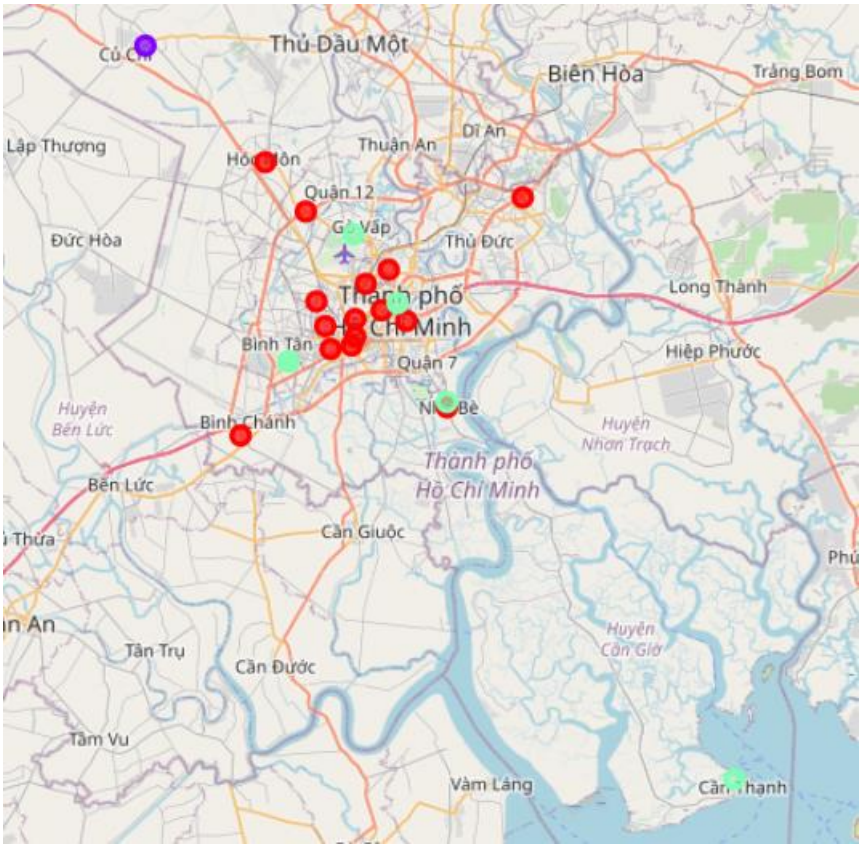

Figure 7. District clustering

5. Conclusion

Our analysis shows that there is a great number of Coffee shops in the center of Ho Chi Minh City which are located in cluster 0 and 2. It is not good if we setup the new business here because we might get the competitive from competitors.

At Cu Chi district, cluster 1, only 10 coffee shops are operating. Hence, this is the right location for our customers to open the business.

Purpose of this project was to identify which location has less coffee shop in order to aid customer in narrowing down the search for optimal location for a new coffee shop. By collecting the number of coffee shop density distribution from Foursquare data we have first identified general boroughs that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby coffee shops. Clustering of coffee shops based on location was then performed in order to show the insight about distribution of major zones which are containing greatest number of coffee shops. It helps the customer to take the decision.

Final decision on optimal coffee shop location will be made by customer based on specific characteristics of neighborhoods and locations in every recommended zone.