Week 5

1. Introduction

1.1 Background

In the past decade, the lifestyle of urban people has changed with the trend and habit of drinking coffee. Coffee, which is ancient, is the same drink commonly used by older men, and now women and men of all ages are used to drinking coffee. Not only to enjoy coffee, many people are looking for places to drink coffee. This coffee shop has finally become a cool place with internet connection and enjoy all kinds of coffee beans.

This coffee drinking trend will become a huge business opportunity. Businesses are starting to work where specialty coffee is available. With the trend in Hong Kong, coffee shops are likely to make good profits. However, it is not as easy to enter the business world as people think, especially in Hong Kong, where coffee shops are very common.

If you already have the capital to open a coffee shop, then you have to have the courage to start designing strategies and looking at the market. If you like coffee for a long time and have a habit of drinking coffee, that means you can start your business with the right passion. Therefore, I try to practice my learning in coursera to answer the related questions, that is, design strategies to determine which areas are suitable for opening coffee shops.

1.2 Problem

Finding data in Hong Kong is a challenge that must be addressed because Hong Kong does not divide the region into communities like some countries. Therefore, this project will use Wikipedia's region list to define this region. Renting a place to determine the exact location of the coffee shop is also one of the problems that must be solved.

1.3 Interest

I believe it's a related challenge and an effective one for anyone who wants to open a coffee shop and determine the right location. The same method can be applied as needed. This case also applies to anyone who is interested in exploring and starting a business or looking for new business in any city. Finally, it can also be a good practice for developing data science skills.

2. Data Acquisition and Cleaning

2.1. Data Acquisition

The data acquired for this project is a combination of data from two sources. The first data source of data is scraped from a Wikipedia page that contains the list of districts in Hong Kong https://en.wikipedia.org/wiki/Districts_of_Hong_Kong.

The following are the columns:

	Districts	Regions
0	Central and Western	Hong Kong Island
1	Eastern	Hong Kong Island
2	Southern	Hong Kong Island
3	Wan Chai	Hong Kong Island
4	Sham Shui Po	Kowloon

District : Name of the district Region: Name of the region

The Second data source is the list of Longitude & Latitude from website latlong.net, the following are columns:

	Districts	Latitude	Longitude
0	Tsuen Wan	22.374630	114.115100
1	Sha Tin	22.383381	114.198517
2	Tuen Mun	22.396910	113.974411
3	Tai Po	22.445400	114.167709
4	Yuen Long	22.445570	114.022290

District : Name of the district Latitude : Latitude of the town Longitude : Longitude of the town.

2.2.Data Cleaning

The data is preprocessed separately. The Districts information of Hong Kong is scraped from Wikipedia using the Beautiful Soup library in Python. This library can help us extract data in the tabular format on the website. After extracting data, a panda data frame (as shown in Fig 2.1) is created using string manipulation

	Districts	Regions
0	Central and Western	Hong Kong Island
1	Eastern	Hong Kong Island
2	Southern	Hong Kong Island
3	Wan Chai	Hong Kong Island
4	Sham Shui Po	Kowloon

Fig 2.1 Hong Kong 18-District Data after preprocessing

The second data is a list of coordinates for the 18 districts which we get the data from latlong.net and store them in a csv file. A panda dataframe(as shown in Fig 2.2) is then created in order to store the data

	Districts	Latitude	Longitude
0	Tsuen Wan	22.374630	114.115100
1	Sha Tin	22.383381	114.198517
2	Tuen Mun	22.396910	113.974411
3	Tai Po	22.445400	114.167709
4	Yuen Long	22.445570	114.022290

Fig 2.2 Coordinates for the 18 districts in Hong Kong

3. Methodology

After creating a dataframe storing the data of 18 districts and their coordinates, by using the Foursquare API, we can find different venues for different districts within a 500-meter radius. It then return a JSON file which turn into a dataframe containing venues within districts(as shown in Fig 3.1) with further processes.

	Districts	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central and Western	22.28666	114.15497	Four Seasons Hotel Hong Kong (香港四季酒店)	22.286554	114.156929	Hotel
1	Central and Western	22.28666	114.15497	Galerie Perrotin	22.285455	114.156215	Art Gallery
2	Central and Western	22.28666	114.15497	Central Indian Restaurant	22.285622	114.153839	Indian Restaurant
3	Central and Western	22.28666	114.15497	The Spa at Four Seasons	22.286279	114.157623	Spa
4	Central and Western	22.28666	114.15497	忠記粥品	22.285031	114.154474	Chinese Breakfast Place

Fig 3.1 Dataframe containing all venues for different districts

The data is further processed using one hot encoding(one hot encoding is commonly used to turn categorial data to numerical data in order to help the machine do a better job in prediction when we provide it to ML algorithm). The venue data is then grouped by districts and then the mean for each category is calculated. Therefore, we can find the most common category of venues for each district(as shown in Fig 3.2).

	Districts	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Central and Western	Coffee Shop	Chinese Restaurant	Japanese Restaurant	Wine Bar	French Restaurant	Cocktail Bar	Hotel	Yoga Studio	Sushi Restaurant	Modern European Restaurant
1	Eastern	Chinese Restaurant	Park	Coffee Shop	Cantonese Restaurant	Indian Restaurant	Hong Kong Restaurant	Japanese Restaurant	Restaurant	French Restaurant	Harbor / Marina
2	Islands	Clothing Store	Sporting Goods Shop	Coffee Shop	Sushi Restaurant	Café	Korean Restaurant	Chinese Restaurant	Cha Chaan Teng	Accessories Store	Pharmacy
3	Kowloon City	Thai Restaurant	Dessert Shop	Chinese Restaurant	Café	Coffee Shop	Fast Food Restaurant	Cha Chaan Teng	Noodle House	Cantonese Restaurant	Bakery
4	Kwai Tsing	Mobile Phone Shop	Bus Station	Trail	Scenic Lookout	Dive Bar	Flea Market	Fast Food Restaurant	English Restaurant	Electronics Store	Dumpling Restaurant

Fig 3.2 The most common category of venues for each district

After getting the top 10 categories of venues for each district, we cluster the districts into 5 clusters using k-means clustering which is a form of unsupervised machine learning algorithm that clusters data to predefined cluster size. In this project, we will cluster the districts into 5 group(as shown in Fig 3.3). The reason for using k-means clustering is to group districts with similar venues so that people can shortlist the area of their interest based on the venues for each district.

	Cluster Labels	Districts	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	4	Central and Western	Coffee Shop	Chinese Restaurant	Japanese Restaurant	Wine Bar	French Restaurant	Cocktail Bar	Hotel	Yoga Studio	Sushi Restaurant	Modern European Restaurant
1	3	Eastern	Chinese Restaurant	Park	Coffee Shop	Cantonese Restaurant	Indian Restaurant	Hong Kong Restaurant	Japanese Restaurant	Restaurant	French Restaurant	Harbor / Marina
2	4	Islands	Clothing Store	Sporting Goods Shop	Coffee Shop	Sushi Restaurant	Café	Korean Restaurant	Chinese Restaurant	Cha Chaan Teng	Accessories Store	Pharmacy
3	0	Kowloon City	Thai Restaurant	Dessert Shop	Chinese Restaurant	Café	Coffee Shop	Fast Food Restaurant	Cha Chaan Teng	Noodle House	Cantonese Restaurant	Bakery
4	1	Kwai Tsing	Mobile Phone Shop	Bus Station	Trail	Scenic Lookout	Dive Bar	Flea Market	Fast Food Restaurant	English Restaurant	Electronics Store	Dumpling Restaurant

Fig 3.3 Districts with their cluster labels

Plotting the clusters on the map of Hong Kong using Folium can better visualize the clusters(as

shown in Fig 3.4)

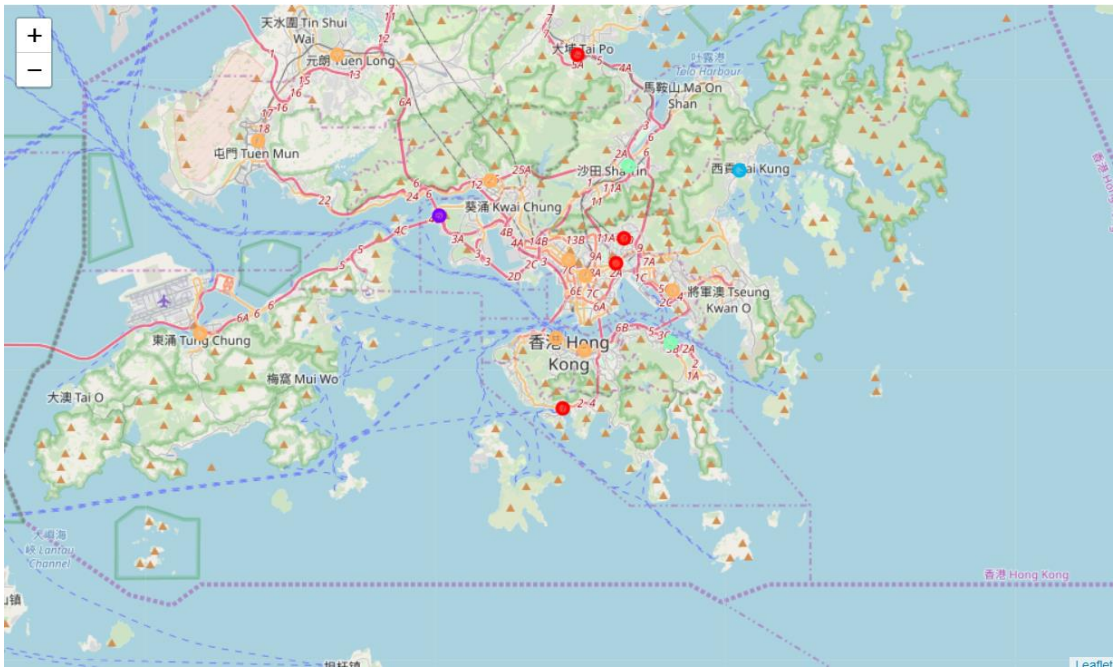

Fig 3.4 Cluster result on a map with red = cluster 0, purple = cluster 1, blue = cluster 2, green = cluster 3, orange = cluster 4

4. Results

As in this project, the objective is to find a proper district to run a coffee shop, in this case, we want to lower our risk by choosing districts with fewer competitors. Therefore, we drop the districts with café being the top 10 most common venues (as shown in Fig 4.1).

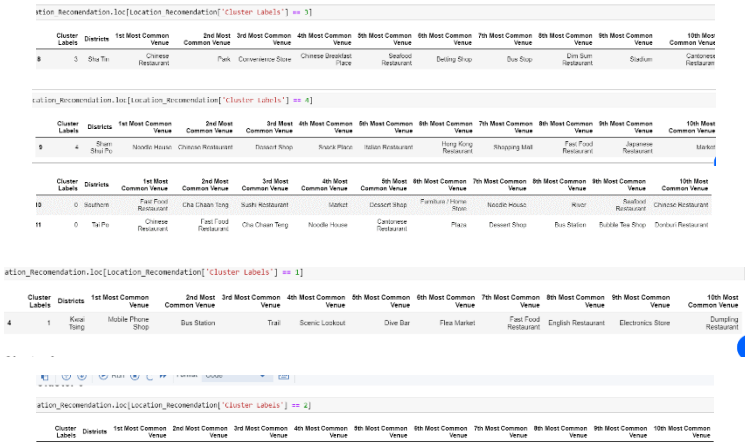

Fig 4.1 Districts without café being the top 10 most common venue

From the result, we see that there are actually a large competition in Hong Kong, we see that

out of the 18 districts in Hong Kong, we only have 5 districts where coffee shop is not in the top 10 common venue. Therefore, running a coffee shop in Hong Kong now may not be the best option. In case you really want to run a coffee shop, area in Cluster 0 may be the best option you have as there are some indirect competition in the area of other Cluster, like Dessert Shop, Bubble Tea Shop etc.

5. Conclusion

This project helps one get a better understanding of the environment in relation to the most suitable place to open coffee shops. The future of this project includes considering other factors such as the cost of renting a place, the price of land to open a new coffee shop or even the work and salaries of each person in the area to be able to more accurately determine the price of coffee to be sold.