Applied Data Science Capstone Project — Strategic Clustering for Opening a Cafe in Hong Kong

Cara Lo

1. Introduction

Hong Kong is famous for being a 'busy city'. To escape from the hustle and bustle of the city life, sometimes a cup of good coffee, great service and beautiful interiors are all you want. These can all be found in a café. There are limitless choices on café in Hong Kong. Therefore, it is not easy for a new business to survive in the market. A strategic location would be vital for a great start of the cafe. In this project, geospatial analysis and data science methodology and techniques like K-means clustering will be performed to determine the optimal location to open a cafe in Hong Kong.

2. Data

The following data is used to reach the solution:

1. List of districts in Hong Kong (Source: HK GeoSpatial Open Data)
   This defines the scope of this project. There are 18 districts in Hong Kong, Central and Western, Eastern, Southern, Wan Chai, Sham Shui Po, Kowloon City, Kwun Tong, Wong Tai Sin, Yau Tsim Mong, Islands, Kwai Tsing, North, Sai Kung, Sha Tin, Tai Po, Tsuen Wan, Tuen Mun and Yuen Long.
2. Population of districts (Source: Census and Statistics Department)
   One of the important factors to decide the optimal location
3. Latitude and longitude of districts (Source: HK GeoSpatial Open Data)
   For map visualization using Folium
4. Median monthly household income of districts (Source: Census and Statistics Department)
   One of the important factors to decide the optimal location
5. Venue data (Source: Foursquare API)
   Data related to cafe to perform clustering on the districts

List of districts, coordinates, population and median monthly household income data are combined into one dataframe. Each district is assigned with one object ID. Missing data are dropped and the columns are rearranged. The merged table is shown below:

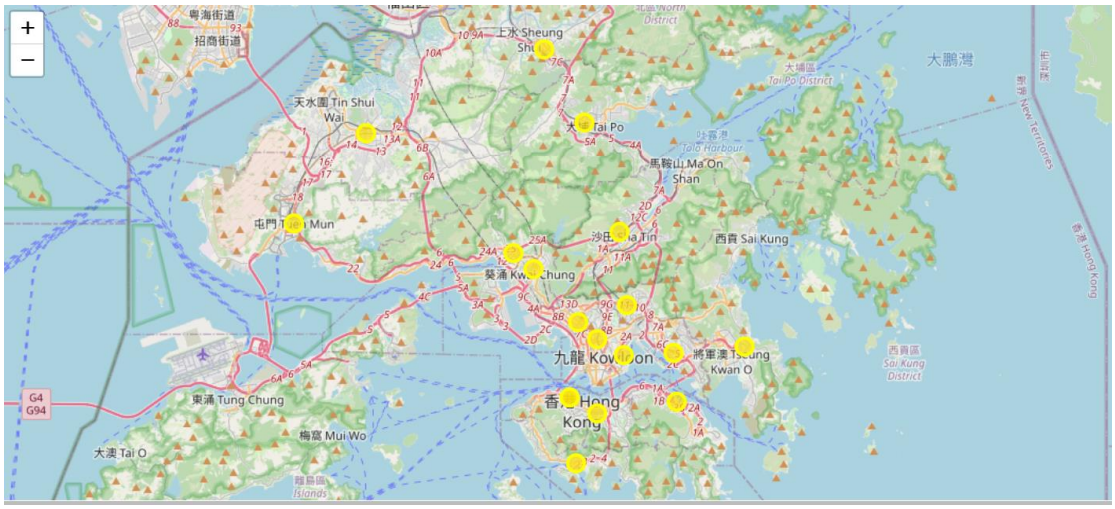
	ObjectID	District	Latitude	Longitude	Population	Median monthly household income (HK$)
0	1	Central And Western	22.28674	114.15488	243266	41400.0
1	2	Eastern	22.28418	114.22429	555034	34300.0
2	3	Southern	22.24734	114.15916	274994	32800.0
3	4	Wan Chai	22.27726	114.17283	180123	44100.0
4	5	Sham Shui Po	22.33191	114.16041	405869	24300.0

Figure 1. Merged Dataframe

3. Methodology

3.1 Folium Map

After gathering the data, the locations of the districts are visualized using Folium map. Each district is marked with a yellow dot.



3.2 Foursquare API

Foursquare API allows us to explore different venues, including restaurants, shopping malls and hotels etc. in each district. We are able to find out the number of cafés in each district by using Foursquare API.

3.3 Bar Graph

Then, a bar graph is plotted to show the number of cafes in a descending order so that we could tell which district is more saturated with café. It is to be believed that the higher the number, the more the competitive in the district.

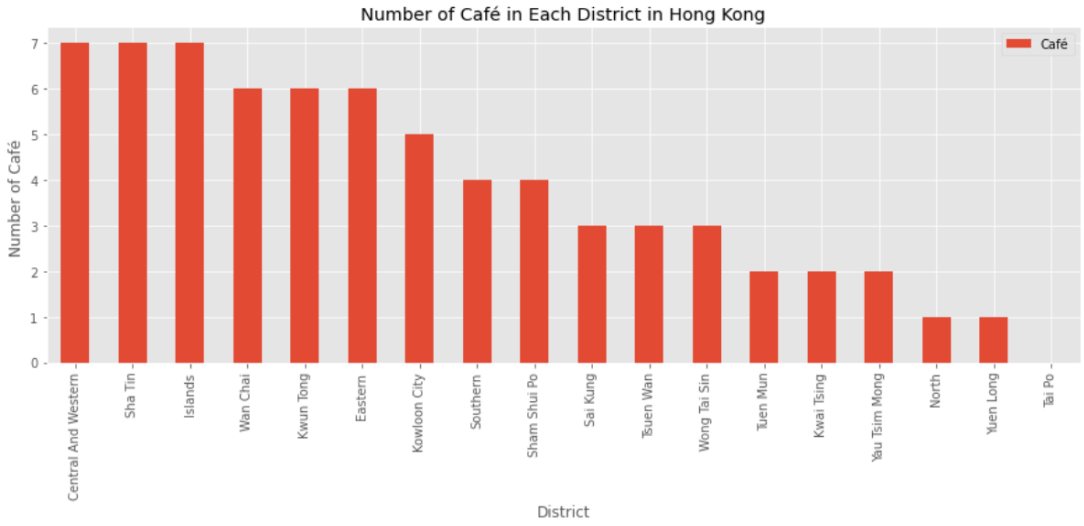


Figure 2. Number Of Café In Each District In Hong Kong

From Figure 2, Central and Western, Sha Tin and Islands have the highest number of 7 cafés.

The population and the median monthly household income are also important factors to take into consideration for the café location. They indicate the size of potential customers and their spending power.

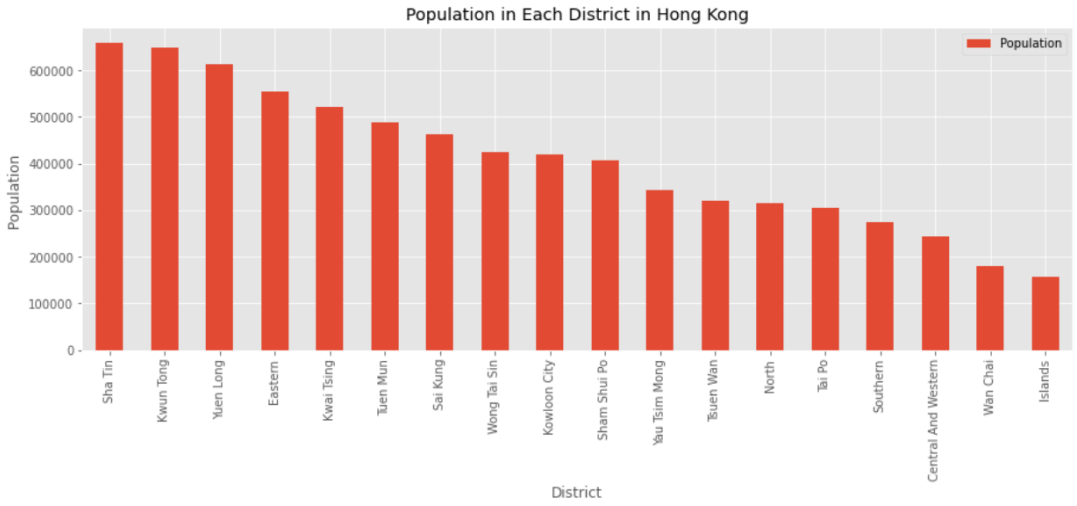


Figure 3. Population In Each District In Hong Kong

From Figure 3, Sha Tin, Kwun Tong and Yuen Long have a relatively large population among all districts.

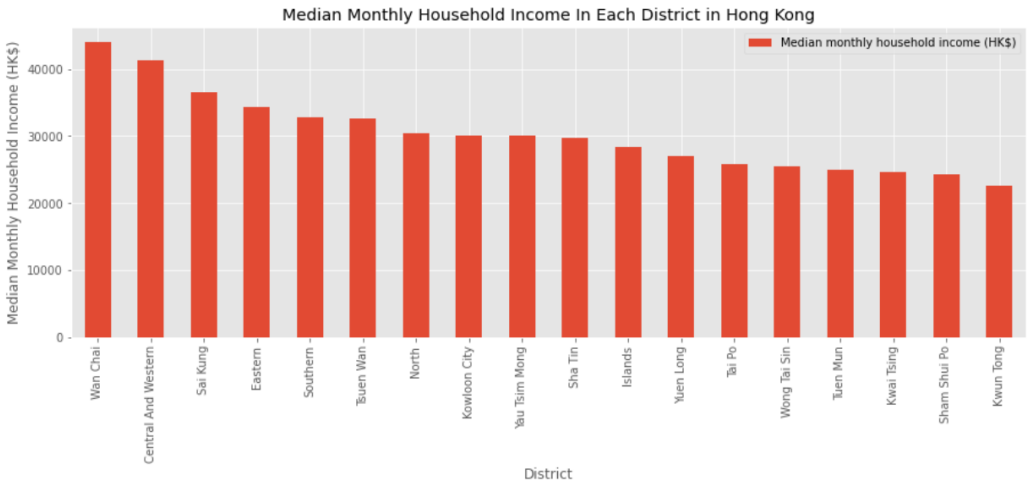


Figure 4. Median Monthly Household Income In Each District In Hong Kong

From Figure 4, Wan Chai, Central and Western and Sai Kung have the highest median monthly household income.

3.4 K-Means Clustering

K-means clustering, a machine learning algorithm used to create K clusters of data points based on feature similarity. In this project, the number of clusters is set to 3. Each district is then assigned with a cluster Label representing the cluster number (either 0 or 1 or 2).

ObjectID	District	Cluster Label	Population	Median monthly household income (HK$)	Café
1	Central And Western	1	243266	41400	7
2	Eastern	2	555034	34300	6
3	Southern	1	274994	32800	4
4	Wan Chai	1	180123	44100	6
5	Sham Shui Po	0	405869	24300	4

Figure 4. Dataframe After Clustering

4. Results

Cluster Label 0

ObjectID	District	Cluster Label	Population	Median monthly household income (HK$)	Café
5	Sham Shui Po	0	405869	24300	4
6	Kowloon City	0	418732	30000	5
8	Wong Tai Sin	0	425235	25500	3
11	Kwai Tsing	0	520572	24700	2
13	Sai Kung	0	461864	36500	3
17	Tuen Mun	0	489299	25000	2

Figure 5. Cluster Label 0

- Med-Low Population
- Low Household Income
- Low Number Of Cafes

Cluster Label 1

ObjectID	District	Cluster Label	Population	Median monthly household income (HK$)	Café
1	Central And Western	1	243266	41400	7
3	Southern	1	274994	32800	4
4	Wan Chai	1	180123	44100	6
9	Yau Tsim Mong	1	342970	30000	2
10	Islands	1	156801	28400	7
12	North	1	315270	30400	1
15	Tai Po	1	303926	25800	0
16	Tsuen Wan	1	318916	32600	3

Figure 6. Cluster Label 1

- Med-Low Population
- High-Med Household Income
- High-Med Number Of Cafes

Cluster Label 2

ObjectID	District	Cluster Label	Population	Median monthly household income (HK$)	Café
2	Eastern	2	555034	34300	6
7	Kwun Tong	2	648541	22500	6
14	Sha Tin	2	659794	29700	7
18	Yuen Long	2	614178	27000	1

Figure 7. Cluster Label 2

- High Population
- Med-Low Household Income
- Med Number Of Cafes

5. Recommendation & Conclusion

In order to find the optimal location for opening a café, several factors including size of customer base, spending power and number of competitors are taken into accounts. By using various data science methodologies and clustering, 3 cluster groups with different characteristics are found. After considering all the factors, districts in cluster label 2 are recommended, including Eastern, Kwun Tong, Sha Tin and Yuen Long. These districts have a large population, a moderate income and not many competitors which are suitable for a new business to start.

However, starting a new business is always not easy and location is not the only factor that needed to be considered. In real-life, much complex situation and more advance data science technology will be required. This project is just an example on how to use data science to solve business problem. Hope it can provide some insights.

Reference

https://en.wikipedia.org/wiki/Districts_of_Hong_Kong
https://opendata.esrichina.hk/datasets/district-offices-and-sub-offices-in-hong-kong-1?geometry=113.250%2C22.135%2C114.985%2C22.579
https://pandas.pydata.org/pandas-docs/stable/index.html