Capstone Project - Organic Bubble Tea Shop in HK

1. Introduction

1.1. Background

Hong Kong is one of the most densely populated and sophisticated cities in the world with a population of over 7.49 million. Being known as a 'Food Paradise', Hong Kong is famous for food and its wide-ranging cuisines have become part of the popular culture in Hong Kong. Recent years have seen a growing trend of 'foodies' in Hong Kong, showing that Hongkongers have increasing demand on food quality and variety. As a resident of this city, I decided to explore my hometown in my project, with the combination of one of the most popular food cravings in Hong Kong - Bubble Tea.

1.2. Problem and Interest

In the past year with the existence of COVID, people are more aware of their health as well as food qualities. Hence, there is a shift in customer preference from carbonised drinks to healthier drinks. Despite its popularity, bubble tea has always been labelled with health issues because of the cost-efficient ingredients (e.g. tapioca pearls and artificial flavourings). With society's growing health awareness, customers are seeking healthier options, but there is not much in Hong Kong's milk tea market.

As seen from this, I would like to explore the ideal location in Hong Kong for opening a Bubble Tea shop
that offers authentic Taiwanese organic milk tea to milk tea lovers, with the use of organic ingredients.

Hong Kong covers an area of 1,106 km^2 and has 3 regions in general: Hong Kong Island, Kowloon and New Territories. The city is divided into 18 districts, where diverse characteristics are observed in various districts: some are more commercial, some are more residential, and some are more touristic.

As the target and potential customers of the Bubble Tea shop are mostly teenagers and young adults, and the fact that organic drinks will be priced slightly higher than ordinary bubble teas, several factors will be taken into account in determining the location for the shop: median monthly household income, population and age groups. Considering all these factors, we can create a map of Hong Kong and information chart where each district is clustered with reference to the venue density.

2. Data acquisition and cleaning

2.1. Data sources

2.1.1. The first table extracted from the page on Wikipedia: https://en.wikipedia.org/wiki/Districts_of_Hong_Kong will be used as the starting point. It shows the names and regions of the 18 districts in Hong Kong, its respective population as well as population density. The name of the districts will be used to map with the geo data to create a map

2.1.2. Latitude and longitude of 18 Hong Kong districts can be found in hklatlon.csv, in which the coordinates are found from https://latitude.to/map/hk/hong-kong/cities/hong-kong

2.1.3. Foursquare API will be used to get the 10 most common venues of the 18 districts in Hong Kong, which will be further used for clustering

2.1.4. Analysis of social parameters (age groups and median monthly household income) will be referenced to HK_median income.csv and HK_age.csv. These datas are available in the Census and Statistics Department of Hong Kong - Population and Household Statistics Analysed by District Council District

2.2. Data cleaning and merging

2.2.1. The table on Wikipedia includes the subtotal of Population, Population Growth and Density/(km2) grouped by Region. These data are not useful for the analysis and I have removed the 3 rows (Hong Kong Island subtotal, Kowloon and New Kowloon subtotal and New Territories subtotal). Also, the last row 'Marine' does not constitute part of the 18 districts in Hong Kong, so I removed the row as well. The Chinese name of districts and Density(/km2) will be not the focus for our analyze, so I drop the columns

2.2.2. I combine the cleaned dataframe as mentioned above with hklatlon.csv to create a new dataframe, which now includes District, Population, Population Growth, Latitude, Longitude and Region as the columns. The newly merged dataframe is then used to create the map of Hong Kong

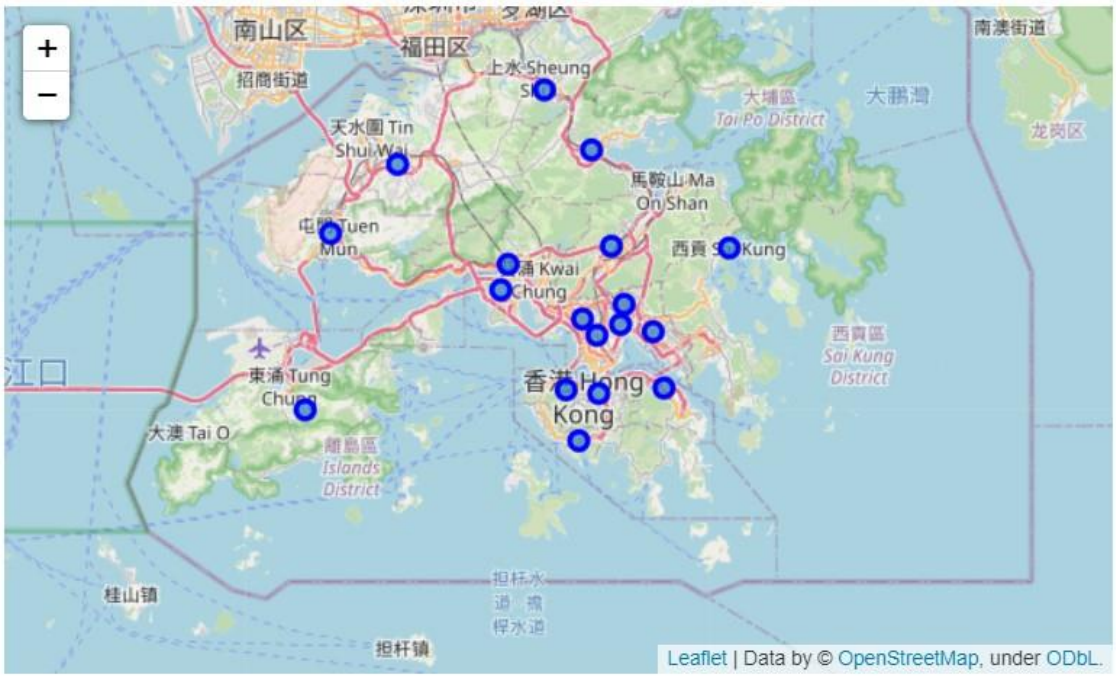
2.2.3. Before clustering, I removed the rows that returned NaN after running Foursquare API, including districts Yuen Long and Islands. It is likely that no values are returned because these two districts are rural, huge in area and scattered.

2.2.4. HK_age.csv is combined with the first dataframe (the cleaned version of the table extracted from Wikipedia). Computation is then performed to obtain the population of the target age group of 16 districts

3. Exploratory Data Analysis and Predictive Modelling

3.1. Common venues of districts

3.1.1. My master data (the cleaned and merged dataframe with columns District, Population, Population Growth, Latitude, Longitude and Region) is used to visualize geographic details of Hong Kong and the 18 districts with the use of python Folium library. A map of Hong Kong with districts superimposed on top is then created for the first glance of the city. Each blue dot represents one district in Hong Kong.



3.1.2. I explore the first district in the dataframe (i.e. Central and Western) to make sure that the latitude and longitude values are extracted properly. We can see that the latitude and longitude values of Central and Western are 22.2833, 114.15, which matches with the data. We can then proceed to get the top 100 venues that are located in Central and Western District with a radius of 500 meters. The results obtained are in json format. I define a function to get the category of the venue, clean the json and structure it into a pandas dataframe

	name	categories	lat	lng
0	PMQ (元創方)	Arts & Crafts Store	22.283298	114.151971
1	The Old Man	Bar	22.282770	114.151774
2	Craftissimo	Beer Store	22.284589	114.148293
3	Teakha (茶·家)	Tea Room	22.284506	114.148407
4	Blue Supreme	Bar	22.285148	114.149385

A total of 73 venues were returned by Foursquare.

3.1.3. I define another function getNearbyVenues to explore and cluster the districts in Hong Kong. We create another dataframe hk_venues which stores the district latitude, district longitude, venue latitude, venue longitude, venue (name), and venue category.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central and Western	22.2833	114.15	PMQ (元創方)	22.283298	114.151971	Arts & Crafts Store
1	Central and Western	22.2833	114.15	The Old Man	22.282770	114.151774	Bar
2	Central and Western	22.2833	114.15	Craftissimo	22.284589	114.148293	Beer Store
3	Central and Western	22.2833	114.15	Teakha (茶·家)	22.284506	114.148407	Tea Room
4	Central and Western	22.2833	114.15	Blue Supreme	22.285148	114.149385	Bar

I also checked the number of venues returned for each district as well as the number of unique categories to check if the results are representative. The result is satisfactory.

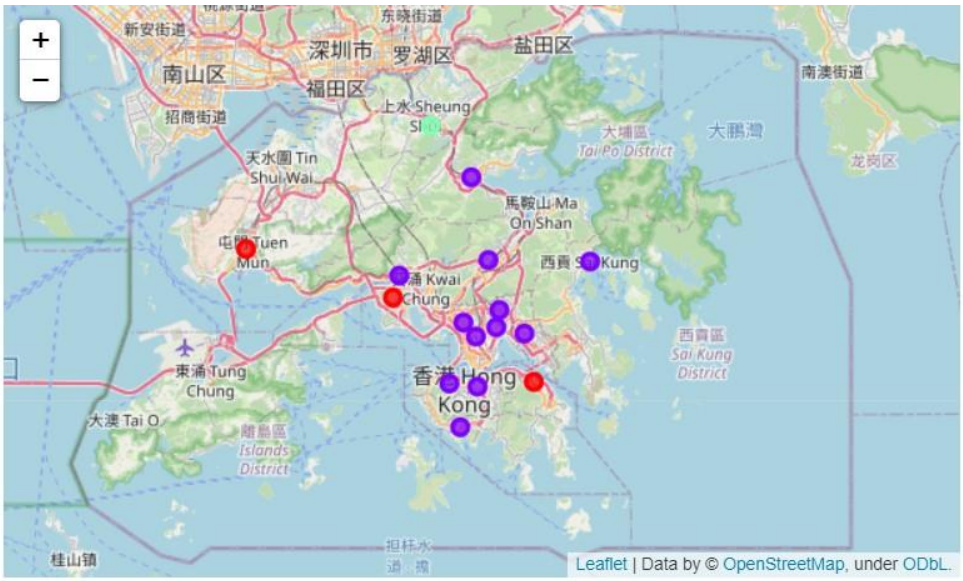
3.1.4. For the purpose of clustering, I performed one hot encoding for venue category (in hk_onehot dataframe) and found that the shape of the data was (546, 129), the result is also satisfactory. I created another dataframe named hk_grouped to group the results from hk_onehot by district, with the shape of the dataframe (16, 129).

3.1.5. I displayed the top 10 venues for each district with the help of the self-defined function return_most_common_venues, and store the resulting data in the new dataframe districts_venues_sorted.

3.2. Clustering of districts

3.2.1. K-means, one of the most common clustering method of unsupervised learning, were used for clustering the districts of Hong Kong in this project. I set the number of clusters as 3 and run the k-means clustering. Clustering labels were added to the dataframe hk_merged and visualization was performed, using Folium once again.

The new map with 3 clusters are shown below: Cluster 0 in Red, Cluster 1 in Purple, and Cluster 2 in Green.

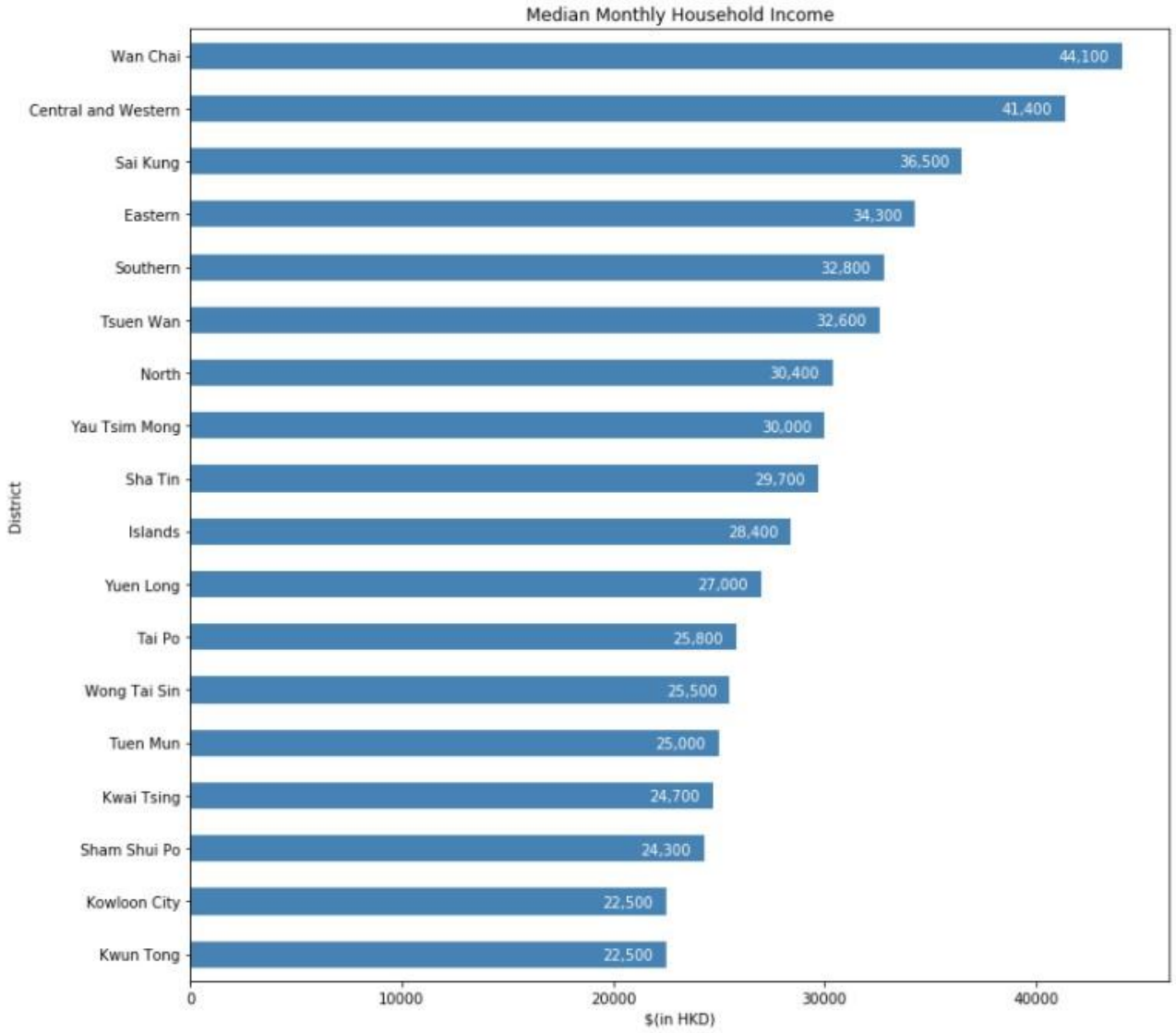


3.2.2. As observed, Cluster 1 dominated the map. I further explored the common venues in Cluster 1 and found that most of the venues are restaurants, bars and coffee shop. Districts in Cluster 1 could be ideal locations for the opening of the bubble tea shop.

3.3. Social parameters

3.3.1. Median monthly household income, population and age groups are selected as the main parameters for the analysis. As the mission and positioning of the organic bubble tea shop is to maximize profit, the prestige pricing strategy and the even-odd pricing strategy will be implemented. The price of the organic bubble tea has to be slightly higher than our competitors' average selling price, so that an indicator of wealth (median monthly household income) is important for the analysis. Also, bubble tea is generally not too preferable for children or elderly, age

groups are therefore taken into consideration instead of simply looking at the population of the districts in general. Last but not least, bubble tea is not a daily necessity, it is more beneficial to open the shop in a district with a larger population as a larger customer base is essential.

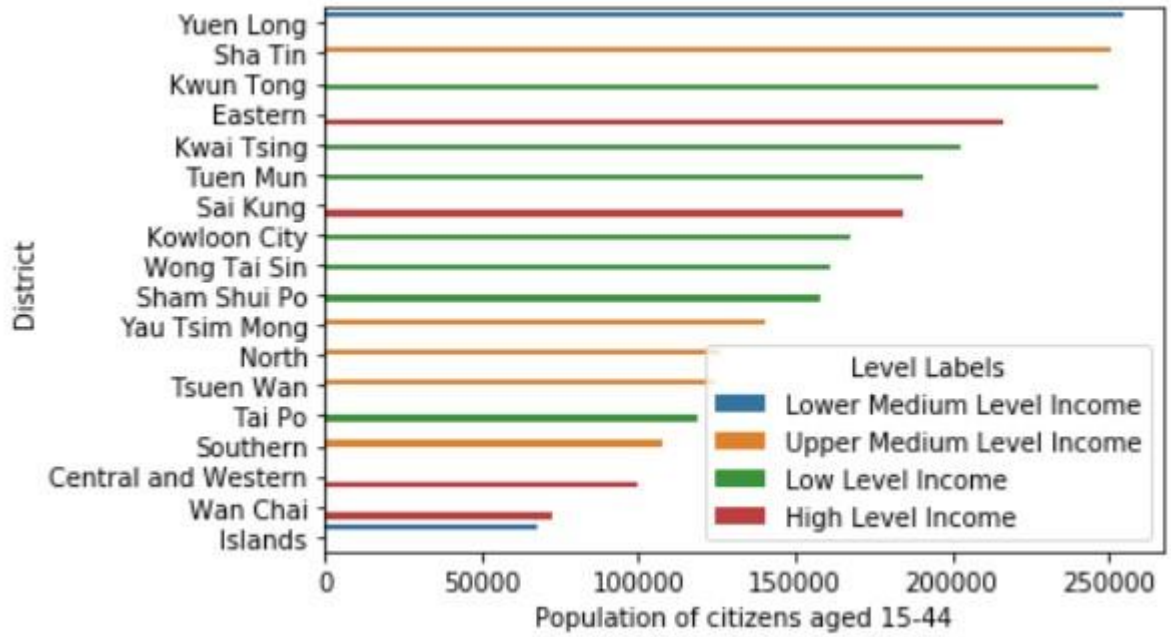


3.3.2. As seen from the bar chart, we can see that Wan Chai, Central and Western and Sai Kung are the top 3 districts in terms of Median Monthly Household Income, followed by Eastern and Southern districts.

3.3.3. I have divided the 18 districts into 4 income levels: High Level Income, Upper Medium Level Income, Lower Medium Level Income and Low. Level Income The table is shown below:

District	Region	Median monthly household income	Level Labels
Kwun Tong	Kowloon	22500	Low Level Income
Kowloon City	Kowloon	22500	Low Level Income
Sham Shui Po	Kowloon	24300	Low Level Income
Kwai Tsing	New Territories	24700	Low Level Income
Tuen Mun	New Territories	25000	Low Level Income
Wong Tai Sin	Kowloon	25500	Low Level Income
Tai Po	New Territories	25800	Low Level Income
Yuen Long	New Territories	27000	Lower Medium Level Income
Islands	New Territories	28400	Lower Medium Level Income
Sha Tin	New Territories	29700	Upper Medium Level Income
Yau Tsim Mong	Kowloon	30000	Upper Medium Level Income
North	New Territories	30400	Upper Medium Level Income
Tsuen Wan	New Territories	32600	Upper Medium Level Income
Southern	Hong Kong Island	32800	Upper Medium Level Income
Eastern	Hong Kong Island	34300	High Level Income
Sai Kung	New Territories	36500	High Level Income
Central and Western	Hong Kong Island	41400	High Level Income
Wan Chai	Hong Kong Island	44100	High Level Income

3.3.4. I have selected 2 age groups, 15-24 and 25-44, as the target customers of bubble tea shop. The age group % of 18 districts are multiplied by its respective population to generate the new column ‘Population of citizens aged 15-44’. From the barplot below, we can observe the population and the income level of 18 districts. Yuen Long district has the highest population aged 15-44 yet its income level is low, which is not ideal for the organic bubble tea shop. The second highest district in terms of population: Sha Tin, could be a better choice as the income level belongs to upper medium. Kwun Tong is likely to be rejected due to similar reasons as Yuen Long, and Eastern district is seemingly ideal as well. Further comparison is performed.



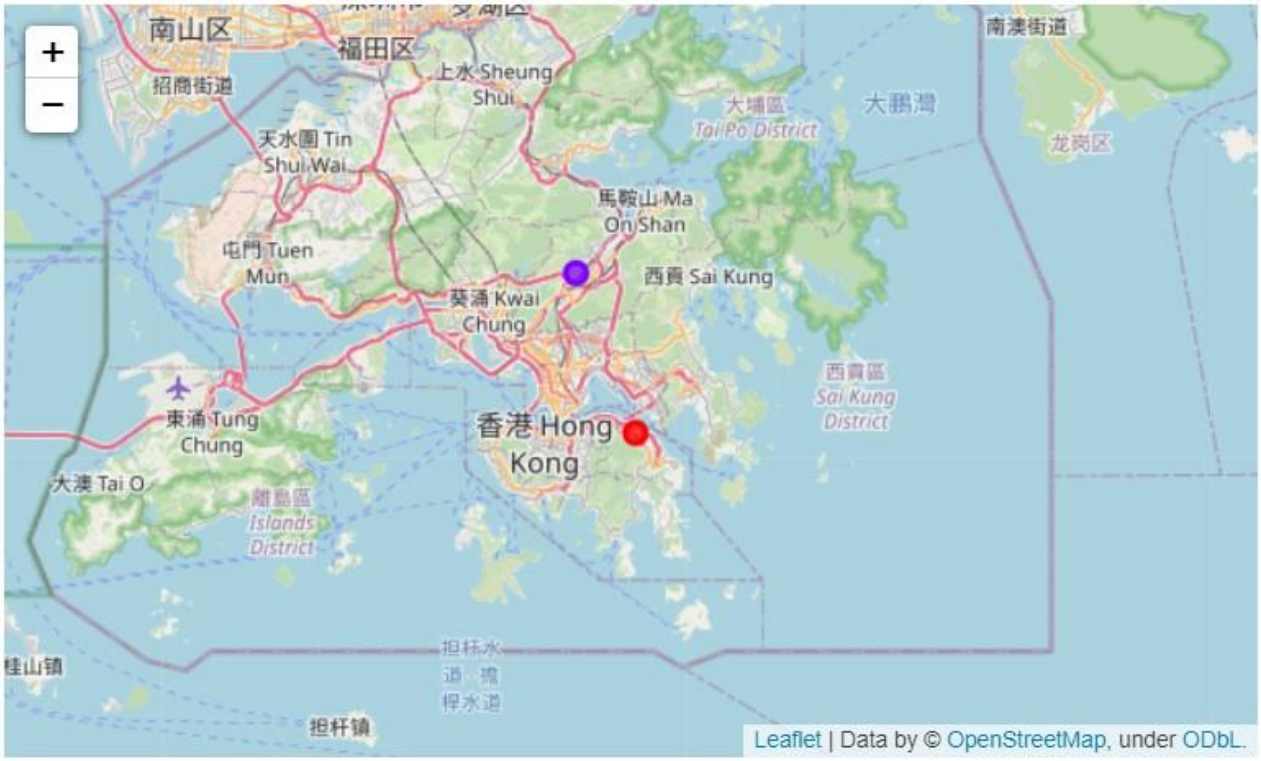
Comparison between Sha Tin and Eastern districts:

	District	Population	Population Growth	Latitude	Longitude	Region	Cluster Labels	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
1	Eastern	555034	-2.8%	22.28411	114.22414	Hong Kong Island	0	Chinese Restaurant	Park	Restaurant	Cantonese Restaurant	Japanese Restaurant	Hainan Restaurant	Indian Restaurant
12	Sha Tin	659794	+8.6%	22.3833	114.18330	New Territories	1	Campground	Vegetarian / Vegan Restaurant	Temple	Resort	Breakfast Spot	Café	Cantonese Restaurant

4. Conclusion

Sha Tin district (in purple dot) and Eastern district (in red dot) are two ideal districts for the opening of an organic bubble tea shop. Both districts have their pros and cons when we take different social parameters into account. Sha Tin district is located in the New Territories, and the top 10 list of venues of this district is not highly dominated by restaurants. With 8.6% of population growth, the organic bubble tea might be a potential and profitable new catering business in Hong Kong; Eastern district is located in Hong Kong Island. Keen competition among catering industries are likely to occur. Yet, the district is also close to Wan Chai/ Central & Western District, which is actually easily accessible by tram. Quarry Bay, which is located in the Eastern District is also developing to be a

second business district in HK (aside from the well-known Central Business District). Our organic bubble tea shop has its potential to deliver high quality milk tea by pricing relatively expensive in this district as citizens rendering around this area are generally wealthier.



5. Future directions

There are no doubt limitations in the analysis, in particular the demographic and social parameters in determining the ideal districts. For instance, the rent of different locations, the number of direct competitors within the same industry, etc. Also, as Hong Kong is a tiny city with numerous catering venues, clustering might not be so effective as most of the common venues in Hong Kong are restaurants, and the types of restaurants are pretty diverse in the city, making the result to be quite similar while observing the top 10 common venues of the districts in Hong Kong. Perhaps I will narrow down the scope of analysis to a particular district and a particular type of cuisine for further analysis in the future.