Exploring the Hong Kong Districts

A. Introduction

1. Introduction

Hong Kong is one of the most populous cities in Asia. It is diverse and is the financial capital of China. It is multicultural with business opportunities provided and a business-friendly environment. It has attracted many different players into the market as a global hub of business and commerce. The city is a major center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance in Asia.

In this project, I will go through the following step:

- Obtain the data;
- Data visualization and some simple statistical analysis;
- Analysis using **Cluster,** specially K-Means Clustering;
- Compare the District to find the best place to start up a restaurant;
- Inference from these results and related conclusion.

2. Preparation for Data

I Use Wikipedia and scrap the names of **18** Districts.

Process the information properly to retrieve the necessary details and create a data frame.

Use Geopy to get the coordinates of all the districts.

Data 1: The data of 18 districts of Hong Kong. https://en.wikipedia.org/wiki/Districts_of_Hong_Kong

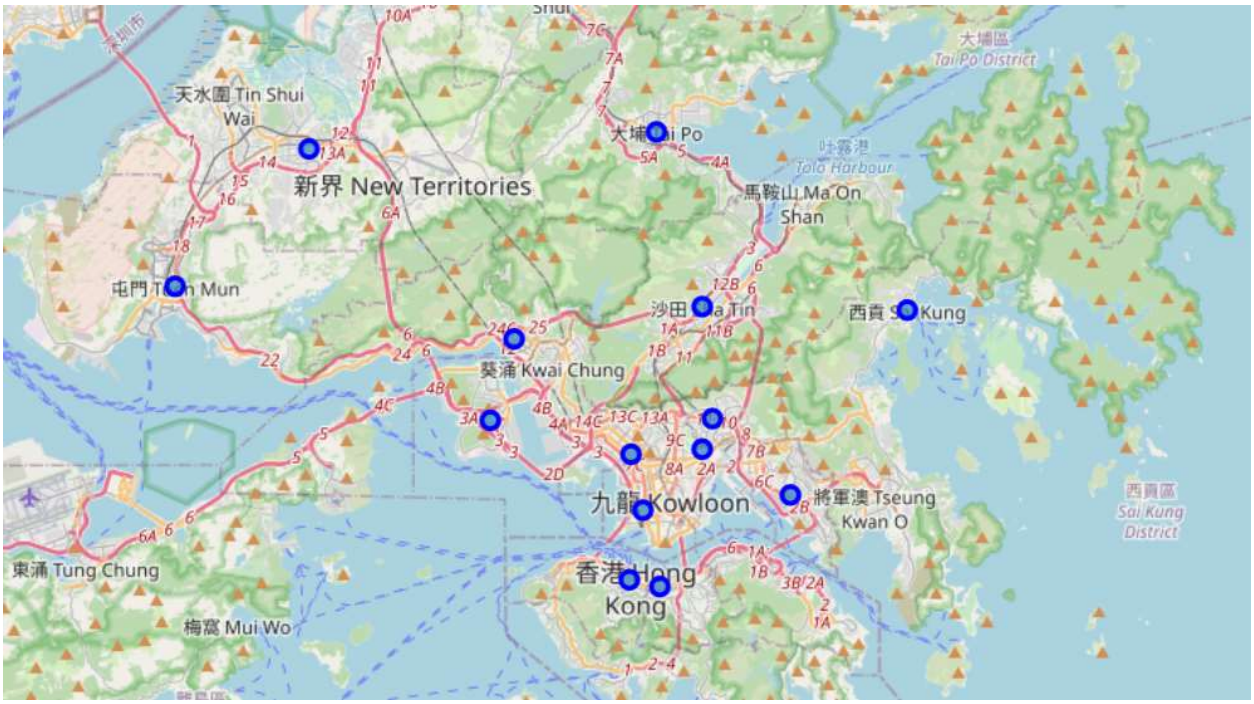
Data 2: Foursquare API to collect information on the venues/competitors in the neighborhoods of Hong Kong.

B. Methodology

I use GitHub repository in my study. My master data which has the main components *District, Area, Density, Region, Longitude and Latitude.*

	District	Population	Area/km^2	Density/km^2	Region	Latitude	Longitude
0	Central and Western	244600	12.44	19983.92	Hong Kong Island	22.281322	114.160258
1	Eastern	574500	18.56	31217.67	Hong Kong Island	1.178718	38.602580
2	Southern	269200	38.85	6962.68	Hong Kong Island	31.990978	-102.071297
3	Wan Chai	150900	9.83	15300.10	Hong Kong Island	22.279015	114.172483
4	Sham Shui Po	390600	9.35	41529.41	Kowloon	22.328190	114.160854

I used python **folium** library to visualize geographic details of Hong Kong and its region and I created a map of Hong Kong with regions superimposed on top. I used latitude and longitude values to get the visual as below:



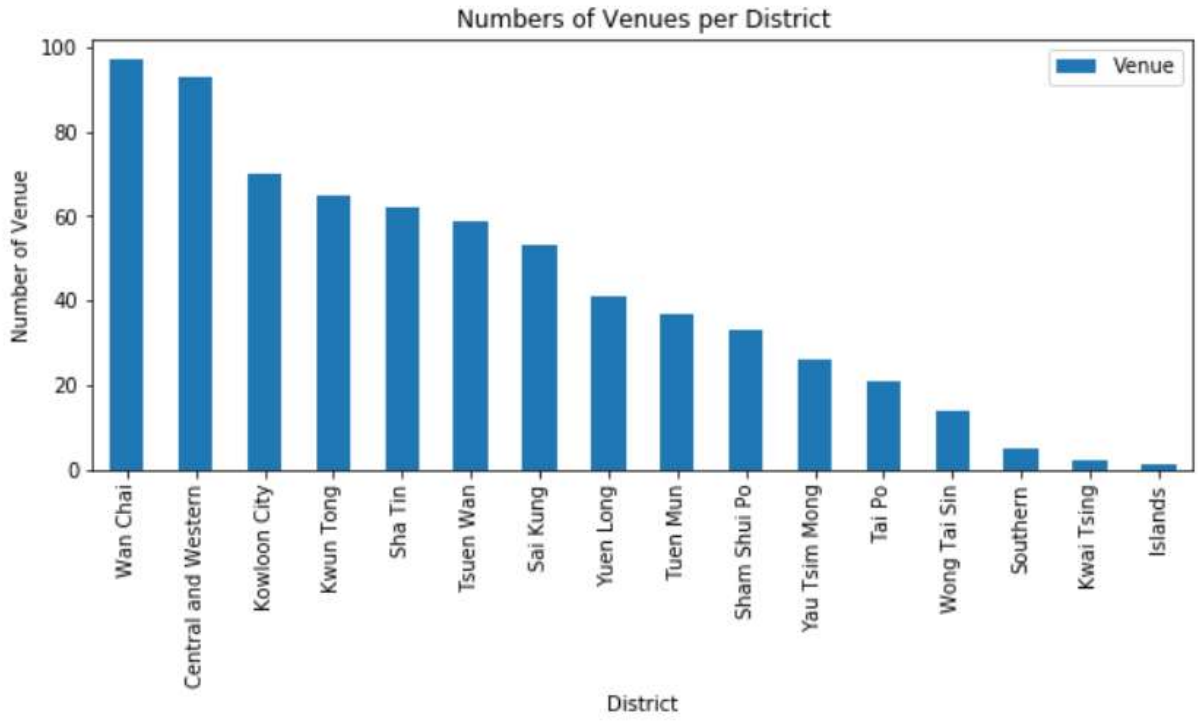
I utilized the **Foursquare API** to explore the regions and segment them. I designed the limit as **100** venue and the radius **500 meter** for each region from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

	name	categories	lat	lng
0	Sports Rehab Physiotherapy (HK) centre	Physical Therapist	22.279498	114.161780
1	Bank Negara Indonesia, Hong Kong Branch	Bank	22.279864	114.164397
2	Pure Fitness	Gym	22.279925	114.163022
3	MTR Admiralty Station (港鐵金鐘站)	Metro Station	22.279412	114.164559
4	Lippo Chiuchow Restaurant (力寶軒)	Chinese Restaurant	22.279242	114.163212

In summary of this data **679 venues** were returned by Foursquare. Here is a merged table of districts and venues.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central and Western	22.281322	114.160258	Mandarin Grill + Bar (文華扒房＋酒吧)	22.281928	114.159408	Steakhouse
1	Central and Western	22.281322	114.160258	Mandarin Oriental Hong Kong (香港文華東方酒店)	22.281857	114.159382	Hotel
2	Central and Western	22.281322	114.160258	Mott 32 (卅二公館)	22.280286	114.159080	Dim Sum Restaurant
3	Central and Western	22.281322	114.160258	The Mandarin Cake Shop	22.281959	114.159416	Bakery
4	Central and Western	22.281322	114.160258	Dr. Fern's Gin Parlour	22.280985	114.158391	Lounge

We can see that **none** of the districts reached the 100 limit of the venues., while *Island, Kwai Tsing, Southern, Wong Tai Sin and Tai Po* are below 20 venues in our coordinates with Latitude and Longitude, shown as in below.



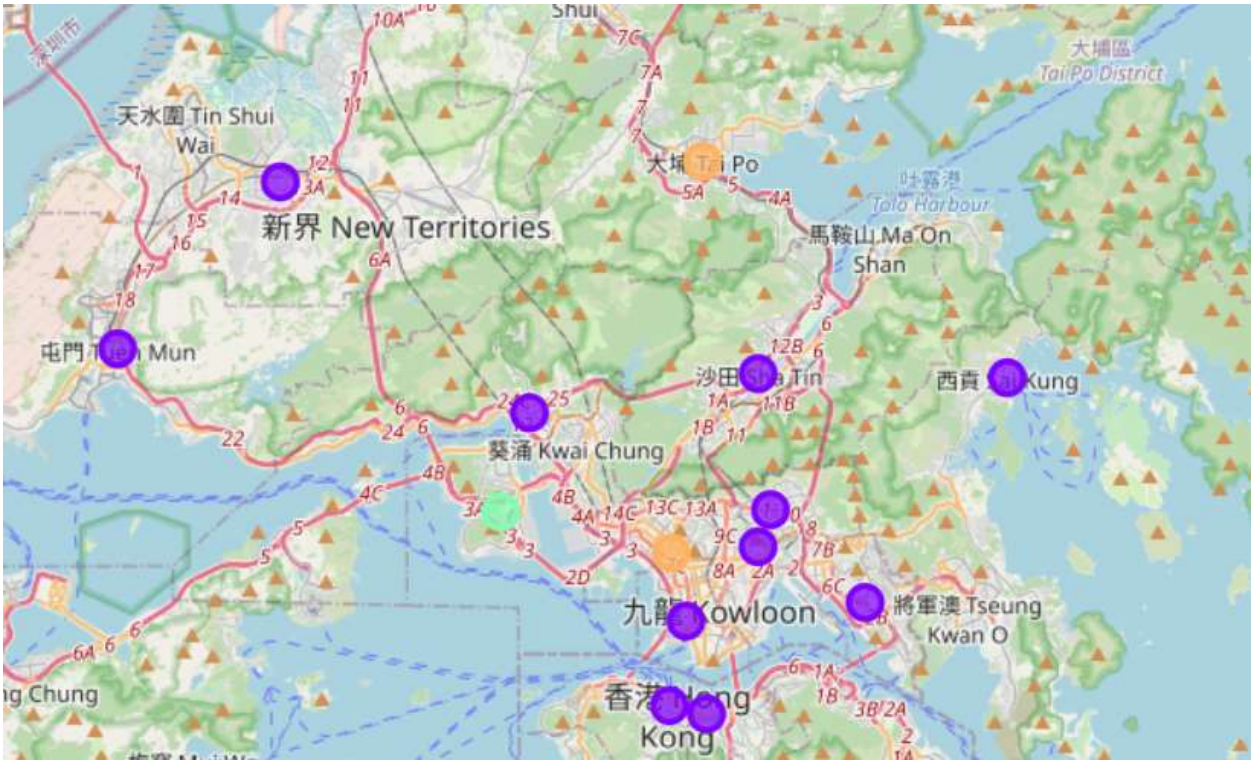
In summary of this graph **143** unique categories were returned by Foursquare, then I created a table which shows list of top 5 venue category for each borough in below table.

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Central and Western	Steakhouse	French Restaurant	Chinese Restaurant	Japanese Restaurant	Social Club
1	Islands	Beach	General Travel	Furniture / Home Store	Fujian Restaurant	French Restaurant
2	Kowloon City	Thai Restaurant	Dessert Shop	Chinese Restaurant	Coffee Shop	Café
3	Kwai Tsing	Tunnel	Chinese Restaurant	Zoo	Fast Food Restaurant	Furniture / Home Store
4	Kwun Tong	Chinese Restaurant	Café	Coffee Shop	Cha Chaan Teng	Fast Food Restaurant

We have some common venue categories in boroughs. In this reason I used unsupervised learning K-means algorithm to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning. And below is my merged table with cluster labels for each District.

	District	Population	Area/km^2	Density/km^2	Region	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Central and Western	244600	12.44	19983.92	Hong Kong Island	22.281322	114.160258	0.0	Steakhouse	French Restaurant	Chinese Restaurant	Japanese Restaurant	Social Club
2	Southern	269200	38.85	6962.68	Hong Kong Island	31.990978	-102.071297	2.0	Mexican Restaurant	American Restaurant	Breakfast Spot	Zoo	Gastropub
3	Wan Chai	150900	9.83	15300.10	Hong Kong Island	22.279015	114.172483	0.0	Japanese Restaurant	Coffee Shop	Café	Italian Restaurant	Hong Kong Restaurant
4	Sham Shui Po	390600	9.35	41529.41	Kowloon	22.328190	114.160854	0.0	Noodle House	Shopping Mall	Café	Chinese Restaurant	Italian Restaurant
5	Kowloon City	405400	10.02	40194.70	Kowloon	22.330160	114.189937	0.0	Thai Restaurant	Dessert Shop	Chinese Restaurant	Coffee Shop	Café

You can also see a clustered map districts of Hong Kong in the below.



C. Results and Conclusion

In this project, I took a brief look at different districts in Hong Kong. I was able to recognize which district is the best choice for restaurant to start up, and to give the restaurant owner an idea of what is the most competitive if one wants to start his business in the same district.

I used the Kmeans algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 5. However, only 18 district coordinates were used. For more detailed and accurate guidance, the data set can be expanded, and the details of the neighborhood or street can also be drilled.

I hope my project could provide some useful data analysis to people who wants to turn to Hong Kong and the city managers who manage the city.