Capstone Project - The Battle of Neighborhoods

Recommendation of Opening a new shop in Hong Kong for ABC Outdoor Supply Store

Introduction

Hong Kong is one of the world's most important global cities. Hong Kong is one of the famous tourist destinations in the world and ranked as one of the most visited city in the world. Also, Hong Kong became home to many immigrants, as one of the most diverse cities worldwide. There are not only local citizens but also tourists visits Hong Kong to do outing, hiking, camping…such outdoor events. Also Due to COVID 19, many local citizens could not go oversea travel and turn to travel in Hong Kong, the increase needs of outdoor supply related products took place from 2020. There are several outdoor supply shops in Hong Kong. For any new players, they need to have business research to select a location to open the shop.

Problem Description

One European Outdoor Supply Store Group (ie. ABC Outdoor Supply Store), would like to consider open a new shop in the major Hong Kong districts. They did not open shop in Hong Kong before. They want to choose a district first for their 1st shop to test the water. Then they will replicate to other districts if successful. So comes to them is the question, how to select a district for the 1st shop?

Target Audience

The ABC Outdoor Supply Store has formed a project team “GO Hong Kong”, with members from managements, business team, marketing team. And they request their IT data science team to assist on this project team. The IT data science team needs to find out the best choice to open the 1st shop to the project team with reasons supporting on its recommendations.

Success Criteria

There are at least 2 factors of the shop location recommendations in Hong Kong must be found out for the ABC Outdoor Supply Store project team:
-Less competition: Lack of Outdoor supply stores
-High demand: Higher population

Data Description

The list of data to put into structured format should contain:

-List of major Hong Kong district, contain the district name

-Latitude and Longtitude coordinates of these districts. As this is used to plot the map for explore and get venue data

-Venue data for each district, contains the venue name, venue latitude, venue longitude, venue category

-District population

How to get the data?

First, we need to do web scrapping from internet. Although there is not easy to find structured format of the data, we could find it in Wikipedia page:

https://en.wikipedia.org/wiki/Districts_of_Hong_Kong

We could get the data from this page and put into a dataframe to make it as structured data for our analysis.

And, this page lists there are major 18 districts in Hong Kong, we need to select only the data we need.

We need to find out the major 18 Hong Kong districts data and explore, segment and do clustering on it.

District	Chinese	Population of 2016	Population Growth from 2006	Density
Central and Western	中西區	243,266	-2.7%	19,391
Eastern	東區	555,034	-2.8%	30,861
Southern	南區	274,994	-0.6%	7,080
Wan Chai	灣仔區	180,123	-0.1%	17,137
Sham Shui Po	深水埗區	405,869	+11.0%	43,381
Kowloon City	九龍城區	418,732	+15.5%	41,802
Kwun Tong	觀塘區	648,541	+10.4%	57,530
Wong Tai Sin	黃大仙區	425,235	+0.4%	45,711
Yau Tsim Mong	油尖旺區	342,970	+22.3%	49,046
Islands	離島區	156,801	+14.4%	886
Kwai Tsing	葵青區	520,572	-0.5%	22,307
North	北區	315,270	+12.3%	2,310
Sai Kung	西貢區	461,864	+13.6%	3,563
Sha Tin	沙田區	659,794	+8.6%	9,602
Tai Po	大埔區	303,926	+3.5%	2,233
Tsuen Wan	荃灣區	318,916	+10.5%	5,149
Tuen Mun	屯門區	489,299	-2.5%	5,894
Yuen Long	元朗區	607,200	+15.0%	4,435

We also get the geographical coordinates of the 18 districts to have the latitude and longitude coordinates for each district.

District	Chinese	Population of 2016	Population Growth from 2006	Density	Latitude	Longitude
Central and Western	中西區	243,266	-2.7%	19,391	22.282190	114.144860
Eastern	東區	555,034	-2.8%	30,861	22.272090	114.221396
Southern	南區	274,994	-0.6%	7,080	22.258010	114.153080
Wan Chai	灣仔區	180,123	-0.1%	17,137	22.277101	114.173837

We also get the venue data for the 18 districts, from reliable location information provider to explore the various types of venues and its categories available in each district.

We get the information such as below and process data cleansing into structured format.

District	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Central and Western	22.282190	114.144860	Morning Trail, The Peak (山頂晨運徑)	22.278008	114.144432	Trail
Central and Western	22.282190	114.144860	Frantzén's Kitchen by Björn Frantzén	22.284808	114.148220	Scandinavian Restaurant
Central and Western	22.282190	114.144860	Okra Hong Kong	22.286108	114.146104	Japanese Restaurant
Central and Western	22.282190	114.144860	Craftissimo	22.284589	114.148293	Beer Store
Central and Western	22.282190	114.144860	Brutl	22.286156	114.143600	Tapas Restaurant
Eastern	22.272090	114.221396	Master Low-key Food Shop (低調高手大街小食)	22.279212	114.229606	Snack Place
Eastern	22.272090	114.221396	Quarry Gap (大風坳)	22.266827	114.213317	Other Great Outdoors

We should particularly find out if any “Outdoor Supply Store” category as this is our project related for the business analysis.

District	Outdoor Supply Store
Central and Western	0.00
Eastern	0.00
Islands	0.00
Kowloon City	0.00
Kwai Tsing	0.00
Kwun Tong	0.00
North	0.00
Sai Kung	0.00
Sha Tin	0.00
Sham Shui Po	0.01

We will provide details step in the “Methodology” section include data process, analysis and if any machine learning method taken.

Methodology

We found the data from internet (Wikimedia), to get the current Hong Kong major districts and related information such population which is one of the key factor in this study. We use Python programming to scrap the website information and put the data collected into data frame.

Then, we need to find the geographical coordinates such us the latitude and longitude information of each district. We use geocode package to convert the district location address

into latitude, longitude. This is because we need to input latitude and longitude to use information provider such as Foursquare.com, to get each district's venue information.

We will visualize the major districts into a map using folium package. The result shown on the map could be used to check again if the geographic coordinates are correct.

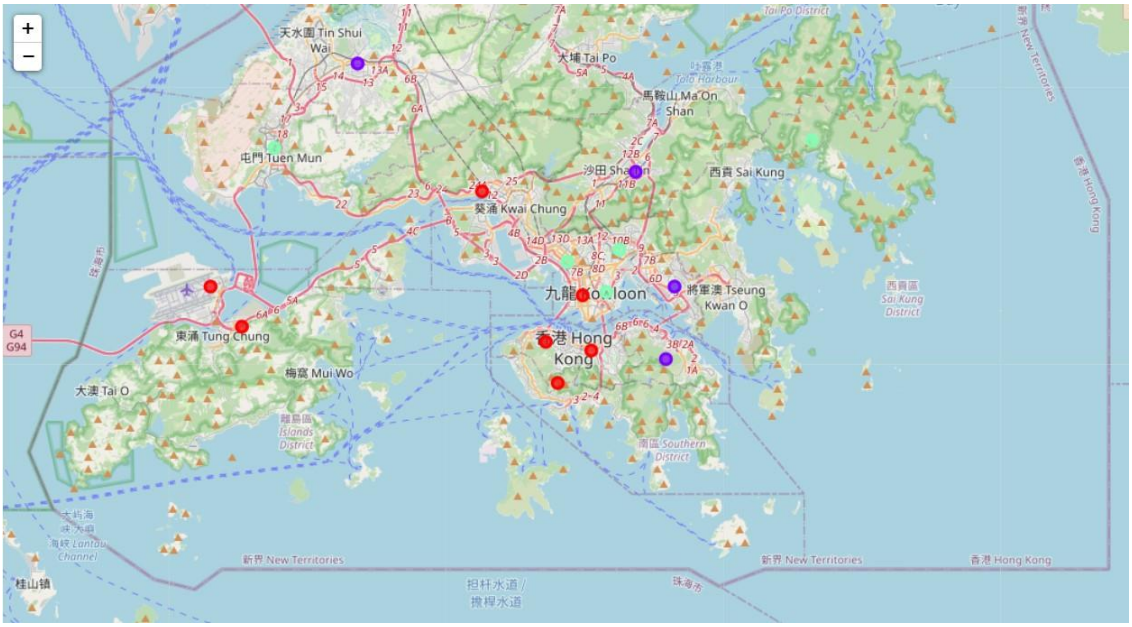
Then we get the venue information through Foursquare API, each district we get 100 venues information within 2 km area. We register a developer account in Foursquare.com to have Foursquare ID, Foursquare secret key and Foursquare access token. We use these to access Foursquare service by API calls with input district latitude and longitude information. And we could get the returned venue information in JSON such as venue name, venue category, venue latitude, venue longitude. We store the data into data frame. And we study the data by each district taking the mean of frequency occurrence of venue category. Since we are study for "Outdoor Supply Store", we filter the district by this venue category for clustering data.

We use machine learning techniques, K-Means, to segment and cluster the districts in order to group them and understand their similarities. This is needed as we need to find out the district which will be recommended to ABC Outdoor Supply Store as the first shop location.

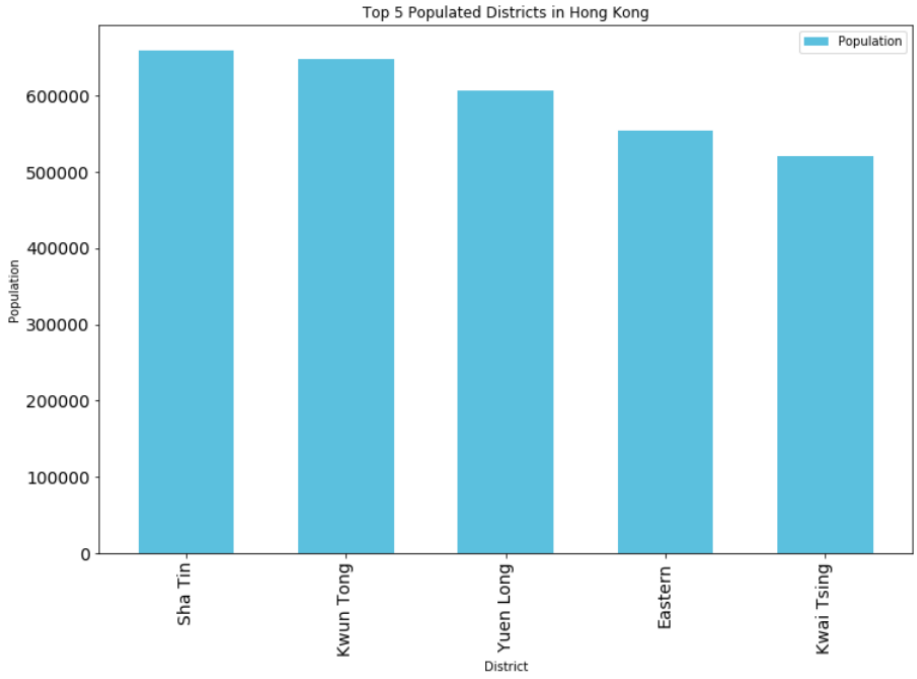
By all the methodology mentioned, we could find out the best suitable district for ABC Outdoor Supply Store to open the first shop base on the districts similarities, high demand and low competence.

Result and Discussion

With K-Means clustering, we could category the districts into 3 x clusters, based on the frequency of the occurrence of venues category as "Outdoor Supply Store" and the population.



And, we use Bart Chart and find out the districts with highest population. This is important because high population means high demand of the Outdoor supply, which we may consider as the recommendation of district to ABC Outdoor Supply Store. The top 5 districts with highest population is show below.



Base on the K-Means clustering, we found Cluster 1 has the highest population districts. Further, Cluster 1 also has low supply of "Outdoor Supply Store".

In Cluster 1, we checked out the district "Sha Tin", has the highest population in the cluster, and very few "Outdoor Supply Store".

Cluster 1

	District	Outdoor Supply Store	Chinese	Population	Population Growth	Density	Latitude	Longitude	Cluster Labels
	Sha Tin	0.0	沙田區	659794.0	+8.6%	9,602	22.382036	114.202102	1
5	Kwun Tong	0.0	觀塘區	648541.0	+10.4%	57,530	22.314236	114.226625	1
17	Yuen Long	0.0	元朗區	607200.0	+15.0%	4,435	22.445131	114.025732	1
1	Eastern	0.0	東區	555034.0	-2.8%	30,861	22.272090	114.221396	1

Further, the above Bart Chart of top 5 districts with highest population also show that the district “Sha Tin” is the district with highest popular among all the other districts.
Both the factor of “high demand” and “less competition” are fulfilled.

Conclusion

With the above findings, we will provide the recommendation as:

Cluster selection: Cluster 1

District : Sha Tin

That means we recommend District "Sha Tin" be selected as the first shop location for ABC Outdoor Supply Store, because the factor that High demand and Less competition are considered.

Further, we suggest ABC Outdoor Supply Store project team to run the program again to have updated results. The project team should consider the result for selecting 2nd shop for growth of business. This is because the result could be used to re-validate the findings from this project. This will be important for better decision making for the project team.