APPLIED DATA SCIENCE CAPSTONE

CAPSTONE PROJECT

February 2020

Week 5 - Assignment

Selection of Your Dream House in Hong Kong

I. Introduction

According to the latest Annual Demographia International Housing Affordability Survey [1], housing in Hong Kong is rated the most expensive and least affordable in the world. Buying a property of your own is probably one of the most important decisions people make in their lives, hence careful considerations must be made as you surely will not want to buy a property that is unsuitable to you.

In Hong Kong, though a very small city, it is geographically and administratively divided into 18 areas named as the 18 districts of Hong Kong. Each district has its own characteristics e.g. the most famous district is probably Central and Western District which is a central business district area surrounded by skyscrapers, however what about the other districts' characteristics?

The target audience of this project is for the people who would like to buy their dream house in Hong Kong. Though these homebuyers most likely have their wish list of the criteria and features of the dream house they want e.g. preferred facilities available, target property price they can afford etc., there is no specific info available which not only centralize but also visualize them.

It is very easy especially for the first-time homebuyer to only focus in the district area they are currently living with their parents and miss out the opportunities in other district areas where they can find their desirable house meeting the criteria in their wish list. The objective of this project is to help these homebuyers to easily understand the characteristics of these 18 district areas by classifying these areas and visualizing them on the map with its average pricing. Essentially, it helps to answer the homebuyer's question: Which district area's characteristics meet most of my criteria of my dream house?

II. Data

Following data will be used for this project:

- Hong Kong districts info: There are 18 geographically divided districts in Hong Kong which were used to separate the areas in Hong Kong for segmentation
    - GeoJSON file for the 18 districts from Esri China (HK) Ltd [2]
    - Latitude and longitude of the 18 districts from Hong Kong GIS Resources [3]
- Hong Kong property statistics of the 18 districts from Centaline Property Agency Limited including the below: [4]
    - Stock distribution by the age of private domestic
    - Latest transaction records in Jan-2020 including price per square feet
- Most common venues for the 18 districts in Hong Kong to be found from the data gathered from Foursquare which were used for segmentation of the 18 districts into groups [5]

III. Methodology

Historical property transaction and building age data are some of the key purchasing info for homebuyer, hence

a) Building Age Label: Each district was first assigned with a label according to the age of the buildings located in the district;
b) Choropleth Map on Sales Unit Price: Average unit price from the transactional data of the districts were used to create the choropleth map

The above labels and map were then combined with our clustering results based on the venues located in the districts to visualize the characteristics of the 18 districts on a single map.

A. Building Age Label

From the Hong Kong property statistics data [4], it included the stock distribution by the age of private properties in percentage as shown in Table 3.1.

Table 3.1. Portion of the Hong Kong property statistics with sales transaction in Jan-2020 [4]

District	Total_stock-units	Total_saleable_area thousand-s.f.	Average_building_age-years	Average_saleable_area_of_units-s.f.	Pre-1970 Num of Building	1970-79 Num of Building	1980-89 Num of Building	1990-99 Num of Building	2000-09 Num of Building	Post-2009 Num of Building	Pre-1970 %	1970-79 %	1980-89 %	1990-99 %	2000-09 %	Post-2009 %	Sales-Average unit price-HKpers.f.
CENTRAL & WESTERN	98,448	6,718.80	35	683	15848	21334	23909	20736	10482	6139	0.161	0.217	0.243	0.211	0.106	0.062	18,981
WAN CHAI	66,928	4,676.60	41	699	23446	16576	12942	5894	4136	3934	0.35	0.248	0.193	0.088	0.062	0.059	17,590
EASTERN	124,437	7,257	37	583	19964	30501	51082	8279	9063	5548	0.16	0.245	0.41	0.067	0.073	0.045	16,799
SOUTHERN	42,434	3,969.20	32	935	1945	10088	10378	13360	4837	1826	0.046	0.238	0.245	0.315	0.114	0.043	16,400
YAU TSIM MONG	113,002	6,171	39	546	45150	20175	9390	4274	26931	7082	0.4	0.179	0.083	0.038	0.238	0.063	15,906

And these data were further grouped into 3 categories as listed in Table 3.2. and the group with the highest percentage were identified per district as shown in Table 3.3. This info were to be used for map labelling purpose.

Table 3.2. Groups of building age

Building Age Group	Building Year
Over 40 years [Old]	• Pre-1970 • 1970-79
20 – 40 years [Middle]	• 1980-89 • 1990-99
Less than 20 years [Young]	• 2000-09 • Post-2009

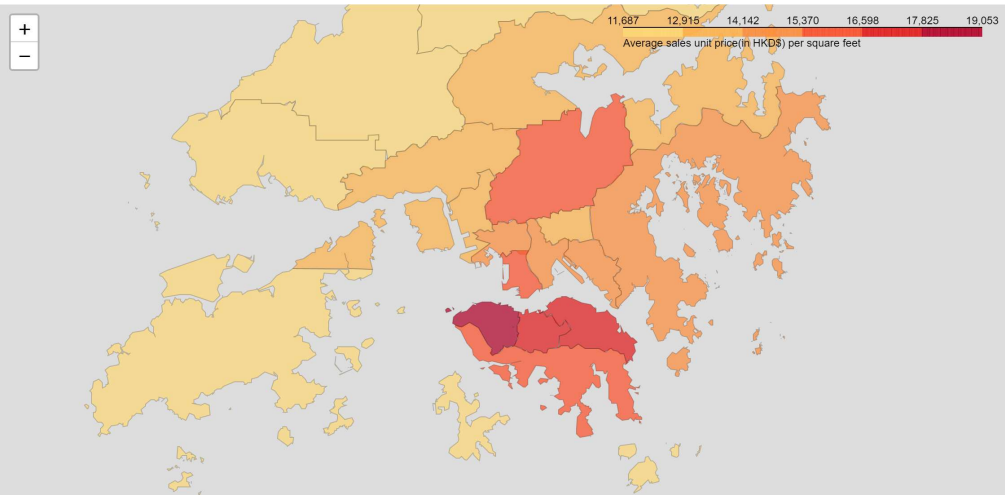
Table 3.3. Portion of stock distribution in the 3 groups mentioned in Table 3.2. and its highest age group

	District	Sales-Average unit price-HKpers.f.	Average_building_age-years	Over 40 years	20 - 40 years	Below 20 years	Top building age group
0	CENTRAL & WESTERN	18981	35	0.378	0.454	0.168	20 - 40 years
1	WAN CHAI	17590	41	0.598	0.281	0.121	Over 40 years
2	EASTERN	16799	37	0.405	0.477	0.118	20 - 40 years
3	SOUTHERN	16400	32	0.284	0.560	0.157	20 - 40 years
4	YAU TSIM MONG	15906	39	0.579	0.121	0.301	Over 40 years

B. Choropleth map based on the average sales unit price of the 18 districts

In order to create the choropleth map as shown in Fig. 3.1. , GeoJSON file retrieved from Esri China (HK) Ltd [2] was used to visualized the region of space for the 18 geographically divided districts in Hong Kong and the areas were shaded in proportion to the property sales unit price as of Jan-2020 in Table 3.1. [4]

Fig 3.1. Choropleth map showing the sales price per square feet in HK dollar for the 18 dsitricts



C. Frequency of occurrence of venues categories per district using Foursquare API

Foursquare API [5] is used to explore the venues of the 18 districts by passing the coordinates of each district to it. The latitude and longitude of the 18 districts was downloaded from the Hong Kong GIS Resources [3] for this purpose as shown in Table 3.4.

Table 3.4. Portion of the coordinate retrieved from Hong Kong GIS Resource [3]

ID	District	Lat	Long
1	WONG TAI SIN	22.333611	114.196944
2	KWAI TSING	22.353611	114.106389
3	SHAM SHUI PO	22.330833	114.162222
4	YAU TSIM MONG	22.321389	114.172500
5	KOWLOON CITY	22.328333	114.191667

And below parameters were passed to the Foursquare API to search for the nearby venues of the given coordinates and Table 3.5. shows a sample of the venues data returned for the API.

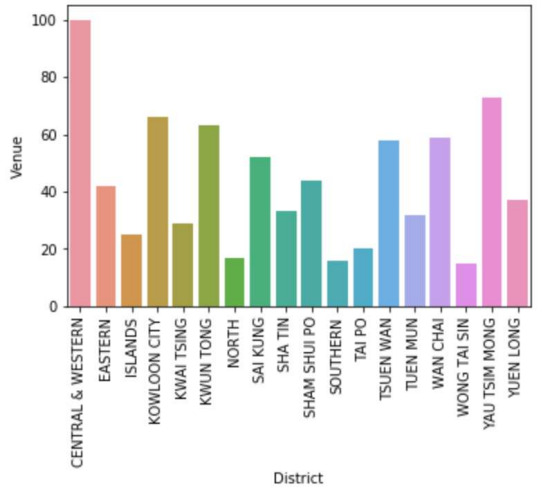
- Radius: 800 meters
- Maximum number of venues: 100 venues

Table 3.5. Portion of the data returned from Foursquare API [5] for Wong Tai Sin District

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	WONG TAI SIN	22.333611	114.196944	英記美點小食	22.334636	114.196827	Asian Restaurant
1	WONG TAI SIN	22.333611	114.196944	Just Climb Bouldering Gym	22.334145	114.199200	Climbing Gym
2	WONG TAI SIN	22.333611	114.196944	Starbucks (星巴克)	22.334026	114.197196	Coffee Shop
3	WONG TAI SIN	22.333611	114.196944	Pentahotel Hong Kong, Kowloon (香港九龍貝爾特酒店)	22.337191	114.199625	Hotel
4	WONG TAI SIN	22.333611	114.196944	Yata Supermarket (一田超市)	22.333312	114.196586	Supermarket

Below figure shows the number of venue returned from the Foursquare per district and there were a few districts – North, Southern and Wong Tai Sin districts were with number of venues less than 20. Increasing the radius and adjusting the coordinates within the districts were tested but still with no help in increasing the number of venues found.

Fig 3.2. Bar chart showing the number of venue returned from Foursquare API per district



Using Foursquare data, the frequency of occurrence of each category per district as shown in Table 3.6 were prepared which were used as the data for clustering.

Table 3.6. Portion of the frequency of the occurrence of each category per district using returned data from Foursquare

	District	Airport Service	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bank	Bar	Beer Bar	Beer Garden	Beer Store	Beijing Restaurant	Bettin Sho
0	CENTRAL & WESTERN	0.01	0.0	0.01000	0.01	0.020000	0.0	0.010000	0.010000	0.0	0.02	0.0	0.01	0.01	0.0	0.
1	EASTERN	0.00	0.0	0.02381	0.00	0.000000	0.0	0.000000	0.000000	0.0	0.00	0.0	0.00	0.00	0.0	0.
2	ISLANDS	0.00	0.0	0.00000	0.00	0.000000	0.0	0.000000	0.000000	0.0	0.00	0.0	0.00	0.00	0.0	0.
3	KOWLOON CITY	0.00	0.0	0.00000	0.00	0.030303	0.0	0.000000	0.030303	0.0	0.00	0.0	0.00	0.00	0.0	0.
4	KWAI TSING	0.00	0.0	0.00000	0.00	0.000000	0.0	0.034483	0.034483	0.0	0.00	0.0	0.00	0.00	0.0	0.

D. k-means clustering and determination of k

k-means clustering was used for segmentation in this case as this is one of the simplest algorithm and can be used to find groups which have not been labeled in the data.

And to determine the best k (the number of clusters), Silhouette Score was used. This score is a measure of how similar an object is to its own cluster compared to other clusters. 'The silhouette ranges from -1 to +1, where a high value indicates the object is well matched to its own cluster and poorly matched to neighboring clusters.' [6] If the score is close to -1, it means that the value is assigned to the wrong cluster.

IV. Results

According to Part III Section D, seven was the chosen k in our k-means clustering as it was with the highest silhouette score as shown in Fig 4.1.

Fig. 4.1. Silhouette score against the number of cluster (k)

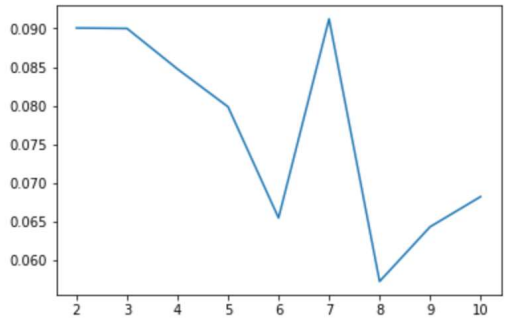


Table 4.1 is the master table showing the top 10 common venues of each district and its cluster label. Naming of these 7 clusters were derived based on the top venues within the same clusters as show in Table 4.2.

Table 4.1. Table of the 18 districts with its cluster labels and top 10 common venues

District	Lat	Long	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
YAU TSIM MONG	22.321389	114.1725	0	Coffee Shop	Noodle House	Cha Chaan Teng	Dessert Shop	Chinese Restaurant	Market	Cantonese Restaurant	Steakhouse	Beer Bar	Flower Shop
KWUN TONG	22.313333	114.225833	0	Chinese Restaurant	Coffee Shop	Café	Japanese Restaurant	Cantonese Restaurant	Hong Kong Restaurant	Department Store	Sushi Restaurant	Fast Food Restaurant	Restaurant
SAI KUNG	22.381389	114.270556	0	Seafood Restaurant	Café	Thai Restaurant	Pub	Pizza Place	Coffee Shop	Restaurant	Italian Restaurant	Burger Joint	Dessert Shop
CENTRAL & WESTERN	22.286667	114.155	0	Coffee Shop	Japanese Restaurant	Chinese Restaurant	Café	Wine Bar	Hotel	French Restaurant	Cocktail Bar	Gym / Fitness Center	Yoga Studio
WAN CHAI	22.279722	114.171667	0	Coffee Shop	Hotel	Café	Japanese Restaurant	Cantonese Restaurant	Chinese Restaurant	Clothing Store	Italian Restaurant	Lounge	Bakery
EASTERN	22.286842	114.210768	0	Coffee Shop	Japanese Restaurant	Chinese Restaurant	Café	French Restaurant	Park	Sushi Restaurant	Fried Chicken Joint	Ramen Restaurant	Building
SHAM SHUI PO	22.330833	114.162222	1	Noodle House	Chinese Restaurant	Dessert Shop	Dumpling Restaurant	Hong Kong Restaurant	Italian Restaurant	Snack Place	Shopping Mall	Fast Food Restaurant	Cha Chaan Teng
TAI PO	22.447587	114.164341	1	Noodle House	Cha Chaan Teng	Chinese Restaurant	Pub	Dumpling Restaurant	Dessert Shop	Snack Place	Fast Food Restaurant	Market	Coffee Shop
YUEN LONG	22.444291	114.030362	1	Chinese Restaurant	Noodle House	Dessert Shop	Fast Food Restaurant	Coffee Shop	Hong Kong Restaurant	Market	Bookstore	Shopping Mall	Café
TSUEN WAN	22.370556	114.119722	1	Chinese Restaurant	Dessert Shop	Shopping Mall	Hong Kong Restaurant	Cantonese Restaurant	Noodle House	Fast Food Restaurant	Taiwanese Restaurant	Thai Restaurant	Japanese Restaurant
WONG TAI SIN	22.333611	114.196944	2	Pizza Place	Thai Restaurant	Cha Chaan Teng	Diner	Supermarket	Noodle House	Chinese Restaurant	Climbing Gym	Paintball Field	Coffee Shop
KOWLOON CITY	22.328333	114.191667	2	Thai Restaurant	Dessert Shop	Chinese Restaurant	Coffee Shop	Café	Halal Restaurant	Noodle House	Cha Chaan Teng	Seafood Restaurant	Fast Food Restaurant
KWAI TSING	22.353611	114.106389	3	Bus Station	Chinese Restaurant	Convenience Store	Dessert Shop	Hong Kong Restaurant	Taiwanese Restaurant	Rest Area	Fast Food Restaurant	Cha Chaan Teng	Cantonese Restaurant
SHA TIN	22.373419	114.182831	3	Bus Stop	Hong Kong Restaurant	Chinese Restaurant	Cantonese Restaurant	Bus Station	Korean Restaurant	Fast Food Restaurant	Italian Restaurant	Asian Restaurant	Restaurant
NORTH	22.504498	114.126779	4	Athletics & Sports	Chinese Restaurant	Fast Food Restaurant	Food Court	Seafood Restaurant	Betting Shop	Gym / Fitness Center	Shopping Mall	Cantonese Restaurant	Coffee Shop
TUEN MUN	22.391667	113.977222	4	Coffee Shop	Cantonese Restaurant	Shopping Mall	Fast Food Restaurant	Zoo	Shanghai Restaurant	Dim Sum Restaurant	Multiplex	Chinese Restaurant	Shabu-Shabu Restaurant
ISLANDS	22.253985	113.904984	5	Dessert Shop	Chinese Restaurant	Vegetarian / Vegan Restaurant	Tea Room	Cable Car	Middle Eastern Restaurant	Bus Station	Noodle House	Market	Sandwich Place
SOUTHERN	22.247222	114.158889	6	Sushi Restaurant	Thai Restaurant	Jewelry Store	Furniture / Home Store	Noodle House	Supermarket	Market	Seafood Restaurant	Coffee Shop	Park

Table 4.2 Cluster name based on the features of each custer

Cluster Label	Dot color on Fig 4.1	Features (top venues)	Cluster Name
0		Coffee shops/Café/Hotel	***Commercial and residential mix area with Coffee shops/Café***
1		Noodle house/Cha Chaan Teng/Chinese restaurant	***Local foods area***
2		Thai restaurant/ Pizza Place / Dessert shop	***Non-local foods area***
3		Bus station/Bus stop	***Massive bus transportation network area***
4		Athletics&sports/Coffee Shop/Shopping Mall/Zoo	***Recreation facilities area***
5		Dessert shop/Cable car	***Tourist area***
6		Sushi restaurant/Thai restaurant/Jewelry store/Furniture store	***Others***

Fig 4.1 is the finalized choropleth map combined with the cluster results and 16 out of 18 districts were able to be grouped together with other districts except for Islands (Cluster 5) and Southern (Cluster 6) which are the sole district within the cluster. It's no surprising that Islands district (circled in Fig 4.1) where the Hong Kong International Airport and the Big Buddha Temple located was grouped on its own. The Southern district case is to be discussed under Part VI.

Fig 4.1 Choropleth map based on average sales price per square feet in HK dollars with cluster labels included

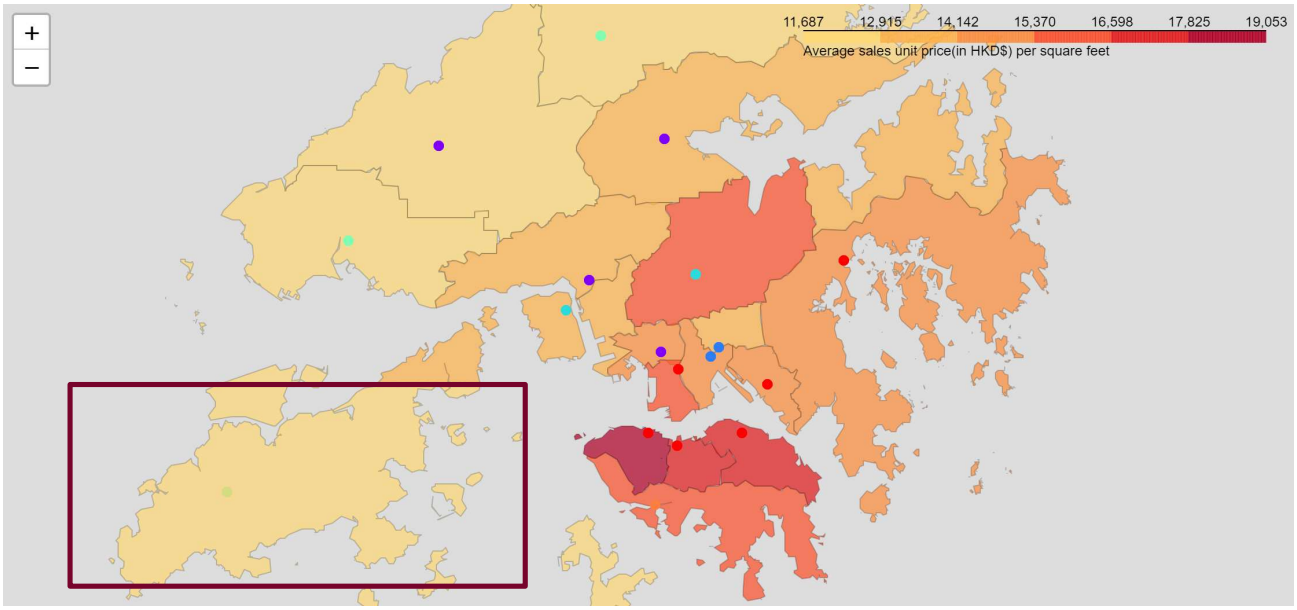
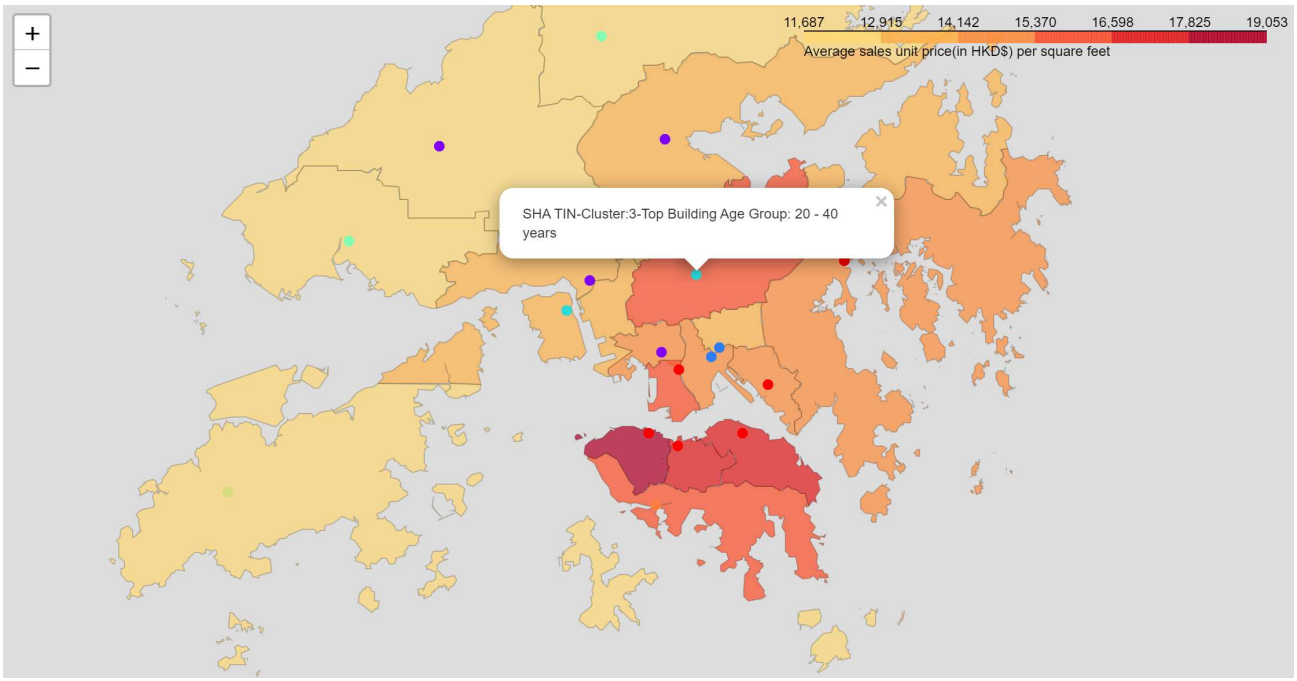


Fig 4.2 Popup label showing the name of district, cluster label and top building age group



V. Discussion

Though Hong Kong being a very small city, the k-means clustering has grouped districts into 7 clusters each with its own characteristics. However, based on the results, there are still some recommendation as listed below:

Inadequate venue data

As mentioned in Part III, three districts North, Southern and Wong Tai Sin districts were with number of venues less than 20. Testing of using different coordinate within the district and increasing the radius have been performed, but there were no improvement in the number of venues returned from the Foursquare API.

Though Southern district itself was segmented as its own cluster while North and Wong Tai Sin districts were successful grouped together with other districts into cluster 4 and 2 respectively, the low number of venue available from these districts made the clustering result of these districts in question. More data points gather from other channel e.g. Places API from Google Maps Platform [7] should improve the accuracy and reliability of the grouping.

Other data

Apart from the inadequate venue data in this study, only the venue category was used for clustering with top building age group and average unit price info included in the map. Other type of data e.g. number of primary/secondary schools with high ranking available within the districts which is one of the main consideration factor in choosing the location of the house for parents as their child can only apply for the primary/secondary schools within their living district areas, can be included in future study to enrich the info available for the homebuyer's consideration.

VI. Conclusion

Combining the property transaction data, statistics data and venue occurrence found within the district via Foursquare API [4], homebuyer can now view all these info in one single map. This should help answer the homebuyer's question - Which district area's characteristics meet most of my criteria of my dream house?

REFERENCE

[1]: **15th Annual Demographia International Housing Affordability Survey: 2019**
http://www.demographia.com/dhi.pdf

[2]: **Hong Kong 18 districts GeoJSON file**
http://opendata.esrichina.hk/datasets/eea8ff2f12b145f7b33c4eef4f045513_0?geometry=113.136%2C21.852%2C114.992%2C22.298

[3]: **Hong Kong 18 districts coordinates**
http://www.hkgisa.org.hk/hong-kong-gis-resources

[4]: **Hong Kong property statistics Jan-2020 from one of the largest property agency - Centaline Property Agency Limited**
http://hkdata.centanet.com/BigData/hong-kong?field=T_POP&sort=default

[5]: **Foursquare**
https://foursquare.com/city-guide

[6]: **Silhouette Score**
https://en.wikipedia.org/wiki/Silhouette_(clustering)

[7]: **Google Map Platform**
https://developers.google.com/places/web-service/intro