The Best Housing Estate in Singapore

Applied Data Science Capstone Project (Week 2)

Steven Sie

Table of Contents

Introduction

As the one of the most advanced country in South East Asia, Singapore has attracted talents from around the region. There is a large number of foreigners from the region stay and work in Singapore. Many have settled as Permanent Resident (PR) in Singapore.

Up to 80% of Singapore's population live in Housing and Development Board (HDB) flats. HDB flats are the mainstream public housing flats in Singapore. The HDB flats have an advantage over the private housing for being relatively more affordably. PRs are not permitted to buy new HDB flats in Singapore, as the HDB flats are reserved for the citizens only. PRs could only buy the HDB flats from the resale market.

Residential properties close to the Mass Rapid Transit (MRT) are popular among home buyers as MRT is the main form of transportation in the island city. Living near to MRT also means more choices of entertainment, retails, eateries and other amenities. There are HDB flats near to most of the MRT stations. PR home buyers typically look at businesses and amenities around the station when deciding the choice of housing.

This project sets out to assist the many PRs from the region to look for ideal housing estate near to the MRT stations. The main considerations in the selection of housing estate are the proximity to the MRT station, types of business near the station and the price of housing flats.

Data

The data used in the project was taken from three sources, i.e. the MRT stations information from the land transport authority of Singapore, the Foursquare location data, and the resale market price data from the housing authority of the Singapore. The analysis was conducted on Jupyter notebook using Python 3.

MRT station data

In order to analyse the businesses and amenities around the MRT stations, the location data of the stations was sourced from the official data from the authority. The train station data was downloaded from the Singapore Land Transport Authority Datamall (1) webpage. It is in Esri shapefile (SHP) format. It contains the name and the geographical information of all the MRT stations.

The .shp and .dbf files were read into the notebook. See the Diagram 1.

	stn_names	stn_codes	coordinates
0	EUNOS MRT STATION	EW7	[35782.955299999565, 33560.077600000426]
1	CHINESE GARDEN MRT STATION	EW25	[16790.746600000188, 36056.30189999938]
2	KHATIB MRT STATION	NS14	[27962.310800000094, 44352.56799999997]
3	KRANJI MRT STATION	NS7	[20081.697399999946, 45214.54790000059]
4	REDHILL MRT STATION	EW18	[26163.47800000012, 30218.819599999115]
...	...	...	...
180	TELOK BLANGAH MRT STATION	CC28	[25376.846900000237, 28138.96670000069]
181	TELOK AYER MRT STATION	DT18	[29667.621600000188, 29414.58669999987]
182	NEWTON MRT STATION	NS21	[28519.36160000041, 32735.207200000063]
183	CANBERRA MRT STATION	NS12	[27598.621771013826, 47200.24193103178]
184	WOODLANDS NORTH MRT STATION	TE1	[22699.99319899935, 47770.399814307006]

185 rows × 3 columns

Diagram 1

The data showed coordinates that looked abnormal. It was because Singapore has a special coordinate system. Fortunately, there was a code shared in github called SVY21 that could convert the data into the normal latitude and longitude coordinates.

After converting into the normal coordinate system, the coordinates column was split into latitude and longitude for easy of handling. The final dataframe is as shown in Diagram 2.

	stn_names	stn_codes	latitude	longitude
0	EUNOS MRT STATION	EW7	1.319778	103.903252
1	CHINESE GARDEN MRT STATION	EW25	1.342352	103.732596
2	KHATIB MRT STATION	NS14	1.417383	103.832980
3	KRANJI MRT STATION	NS7	1.425177	103.762165
4	REDHILL MRT STATION	EW18	1.289562	103.816816

Diagram 2

Foursquare location data

The next source of data was retrieved via the Foursquare API. The API was used to call the top 50 popular venues at within 400m radius of each MRT station. This was done using the 'explore' function of requesting URL. The data extracted is as shown in Diagram 3.

	stn_names	stn_names latitude	stn_names longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	EUNOS MRT STATION	1.319778	103.903252	福春叻沙卤面 Laksa & Lor Mee	1.320254	103.904132	Noodle House
1	EUNOS MRT STATION	1.319778	103.903252	GYMM BOXX	1.317912	103.901094	Gym
2	EUNOS MRT STATION	1.319778	103.903252	Mr Bean	1.319703	103.902479	Breakfast Spot
3	EUNOS MRT STATION	1.319778	103.903252	香港街珍達記	1.316703	103.901415	Chinese Restaurant
4	EUNOS MRT STATION	1.319778	103.903252	LiHO	1.319729	103.902512	Bubble Tea Shop

Diagram 3

The dataframe contains the station names and geographical coordinates, the extracted venue names, their categories and coordinates. After grouping them by station names, there were 319 unique categories from all the stations in the dataframe.

Resale market price data

To understand the latest price of HDB housing, the median resale prices of flats was sourced from the official data from the authority. The csv file was downloaded from the Data.gov.sg (2) webpage. Generally, the town and the MRT train station have the similar name, which enable the analysis of price at various MRT stations.

The into the notebook. See the Diagram 4. It registered the resale median resale price of different types of flat in various towns and the period of sales.

	quarter	town	flat_type	price
0	2007-Q2	Ang Mo Kio	1-room	na
1	2007-Q2	Ang Mo Kio	2-room	-
2	2007-Q2	Ang Mo Kio	3-room	172000
3	2007-Q2	Ang Mo Kio	4-room	260000
4	2007-Q2	Ang Mo Kio	5-room	372000
...	...	...	...	...
7951	2019-Q4	YISHUN	2-ROOM	-
7952	2019-Q4	YISHUN	3-ROOM	265000
7953	2019-Q4	YISHUN	4-ROOM	370000
7954	2019-Q4	YISHUN	5-ROOM	470000
7955	2019-Q4	YISHUN	EXEC	565000

7956 rows × 4 columns

Diagram 4

The data contains the prices of flats from 2007 to 2019. There are close to 8000 rows in the dataframe. As we are interest in the latest prices, only the data from the fourth quarter of 2019 is extracted to be used in the project. There were also some rows with no price indicated. It could be due to no such transaction data available, or simply no such flat type in the location. The rows with no price information was removed from the dataframe.

After the cleaning the data, it is as shown in Diagram 5. There are 74 entries in the dataframe, showing the resale price in the fourth quarter of 2019, according to the town and flat type. The index number was not edited as it is irrelevant and would be removed in the subsequent analysis.

|	town	flat_type	price
7802	ANG MO KIO	3-ROOM	270000.0
7803	ANG MO KIO	4-ROOM	397500.0
7804	ANG MO KIO	5-ROOM	614000.0
7808	BEDOK	3-ROOM	275000.0
7809	BEDOK	4-ROOM	380000.0
...	...	...	...
7949	WOODLANDS	EXEC	550000.0
7952	YISHUN	3-ROOM	265000.0
7953	YISHUN	4-ROOM	370000.0
7954	YISHUN	5-ROOM	470000.0
7955	YISHUN	EXEC	565000.0

74 rows × 3 columns

Diagram 5

Methodology

Visualisation of stations

The adjusted coordinate data of the MRT stations was first plotted on folium map. The plot is as shown in Diagram 6. By cross checking with Google map, they are reasonably similar. Hence the data could be applied to the Foursquare API to generate the locations database.
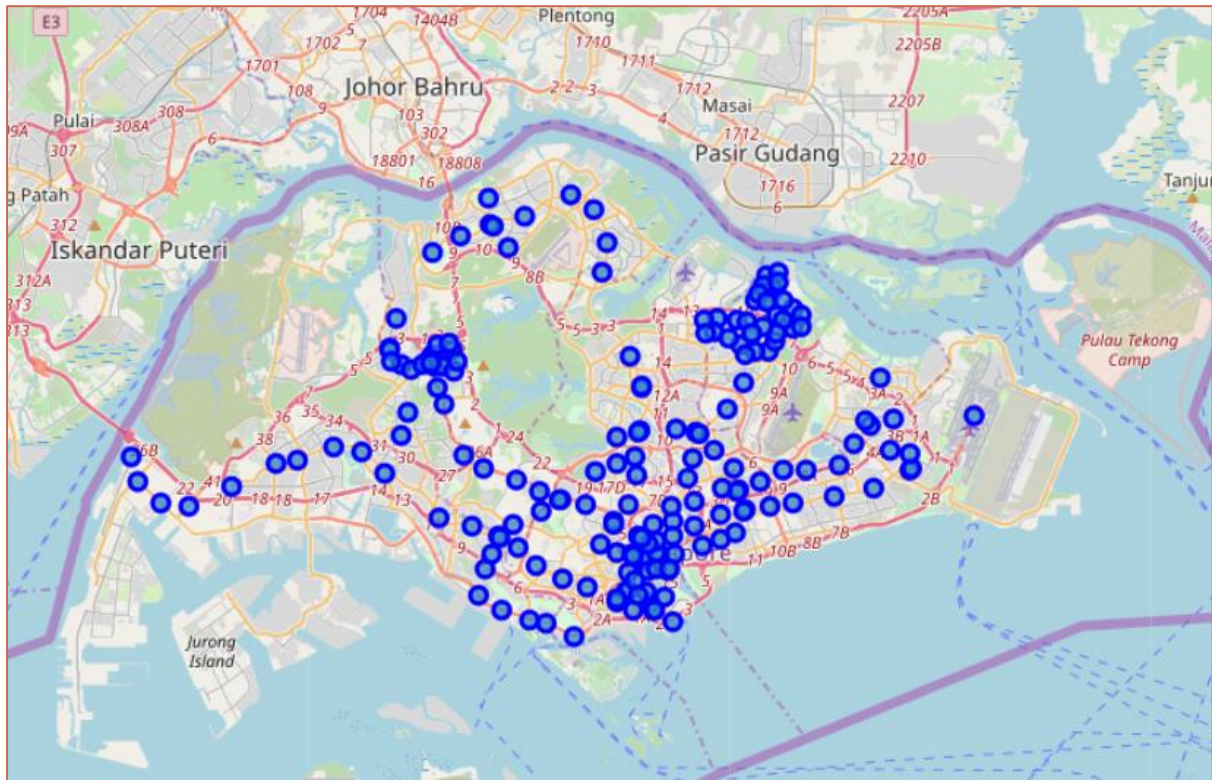
Diagram 6

Analyse stations by clustering

The location information from Foursquare was prepared before the analysis was carried out. First, one hot encoding was applied to the data. Then, the rows with identical station were grouped together and the mean of the frequency of occurrence of each location category was calculated. The data was further arranged into a new dataframe showing the ten most common venues at each station. The result is shown in Diagram 7.

	stn_names	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
0	ADMIRALTY MRT STATION	Bakery	Supermarket	Bus Station	Plaza	Food Court	Coffee Shop	Paper / Office Supplies Store	Food Truck	Food Stand	Food Service
1	ALJUNIED MRT STATION	Asian Restaurant	Chinese Restaurant	Noodle House	Dim Sum Restaurant	Coffee Shop	Food Court	Vegetarian / Vegan Restaurant	BBQ Joint	Seafood Restaurant	Farmers Market
2	ANG MO KIO MRT STATION	Coffee Shop	Dessert Shop	Food Court	Supermarket	Bubble Tea Shop	Japanese Restaurant	Snack Place	Miscellaneous Shop	Modern European Restaurant	Frozen Yogurt Shop
3	BAKAU LRT STATION	Basketball Court	Supermarket	Fast Food Restaurant	Trail	Sandwich Place	Malay Restaurant	Bus Station	Shoe Store	Japanese Restaurant	Dessert Shop
4	BANGKIT LRT STATION	Food Court	Fruit & Vegetable Store	Park	Noodle House	Bike Trail	Market	Miscellaneous Shop	Coffee Shop	Food & Drink Shop	Fishing Spot

```
stn_venues_sorted.shape
```

```
(165, 12)
```

Diagram 7

Next, clustering is applied to the data by k-means approach. The number of cluster was set to 5. After the passing the clustering function, the specific cluster number of each station was added into the data in Diagram 7. The result is shown in Diagram 8. The column 'Cluster Labels' was added to show the cluster where each station belongs to. Visualisation of the clusters was done by plotting into folium map. This is shown in Diagram 9.

	stn_names	stn_codes	latitude	longitude	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 Mos Commo Venu
0	EUNOS MRT STATION	EW7	1.319778	103.903252	0.0	Chinese Restaurant	Noodle House	Coffee Shop	Seafood Restaurant	Gym	Bubble Tea Shop	Train Station	Vegetarian / Vegan Restaurant	Asia Restaurar
1	CHINESE GARDEN MRT STATION	EW25	1.342352	103.732596	0.0	Chinese Restaurant	Coffee Shop	Asian Restaurant	Indian Restaurant	Train Station	Café	Food Court	Food	Filipin Restaurar
2	KHATIB MRT STATION	NS14	1.417383	103.832980	1.0	Coffee Shop	Asian Restaurant	Supermarket	Food Court	Grocery Store	Park	Shopping Mall	Middle Eastern Restaurant	Trai Statio
3	KRANJI MRT STATION	NS7	1.425177	103.762165	3.0	Stadium	Racetrack	Night Market	Bakery	Lighthouse	Bus Line	Go Kart Track	Food & Drink Shop	Fish Chip Sho
4	REDHILL MRT STATION	EW18	1.289562	103.816816	0.0	Chinese Restaurant	Coffee Shop	Gym	Pool	Park	Hainan Restaurant	Train Station	Vegetarian / Vegan Restaurant	Mala Restaurar

Diagram 8

Diagram 9

Exploratory analysis flats prices

The flat prices data consists of prices of different type of flats. In order to find the town with the lowest mean resale price, the data was broken down into five dataframes by different type of flats. An example of the dataframe of Executive type flats is as shown in Diagram 10. It is then plotted into bar chart and sorted for easy comparison of the prices across all towns. See Diagram 11. The town with lowest re-sale price for each type of flat could then be easily spotted.

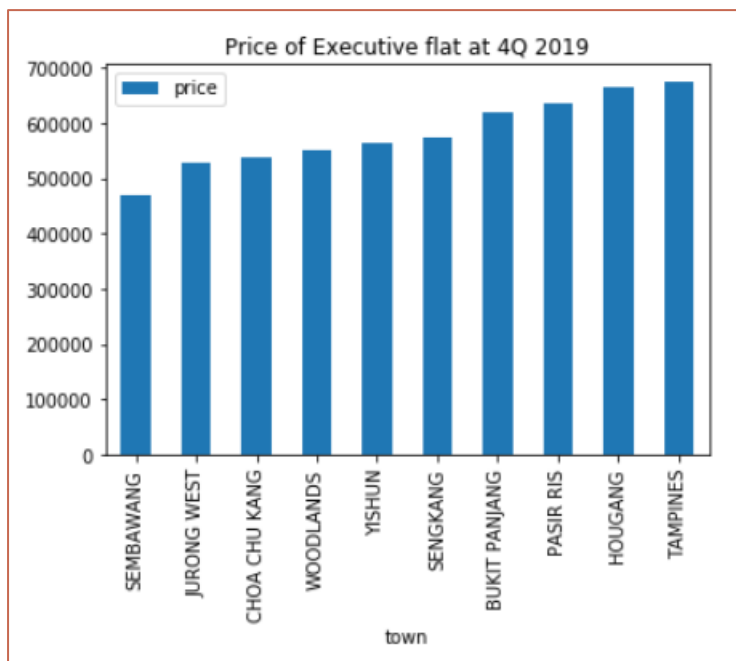| town	price
SEMBAWANG	470000.0
JURONG WEST	530000.0
CHOA CHU KANG	540000.0
WOODLANDS	550000.0
YISHUN	565000.0
SENGKANG	575000.0
BUKIT PANJANG	621500.0
PASIR RIS	635000.0
HOUGANG	665000.0
TAMPINES	675000.0

Diagram 10

Diagram 11

Result

The cluster with the most suitable venues for housing

By looking at the type of venues in the each cluster, the most common venues in each cluster could be identified from the clustered data.

Referring the plot in Diagram 9, the common venues in the red cluster are food courts, coffee shops, chinese restaurant. They are more eateries in this cluster.

The common venues of the blue cluster are rest area, indian restaurant and field. A further check online revealed that the MRT station is new, and the area is not as densely populated as the other cluster.

For the green cluster, the common venues are sports, hotel, Japanese restaurant, café and bakery. These seem venues of high expenses.

For the brown cluster, the common venues are playground, coffee shop, book store and sculpture garden. These venues are more recreational.

And finally, the purple cluster which has common venues such as fast food, dessert shops, supermarket, coffee shops and grocery. This seems to be ideal venues for residential purpose.

The town with the lowest resale flat price

The resale price of each type of flats are as shown in Diagram 12. The three towns with the lowest resale price were selected. They are Woodlands, Sembawang, Jurong West, Bukit Batok, and Choa Chu Kang.
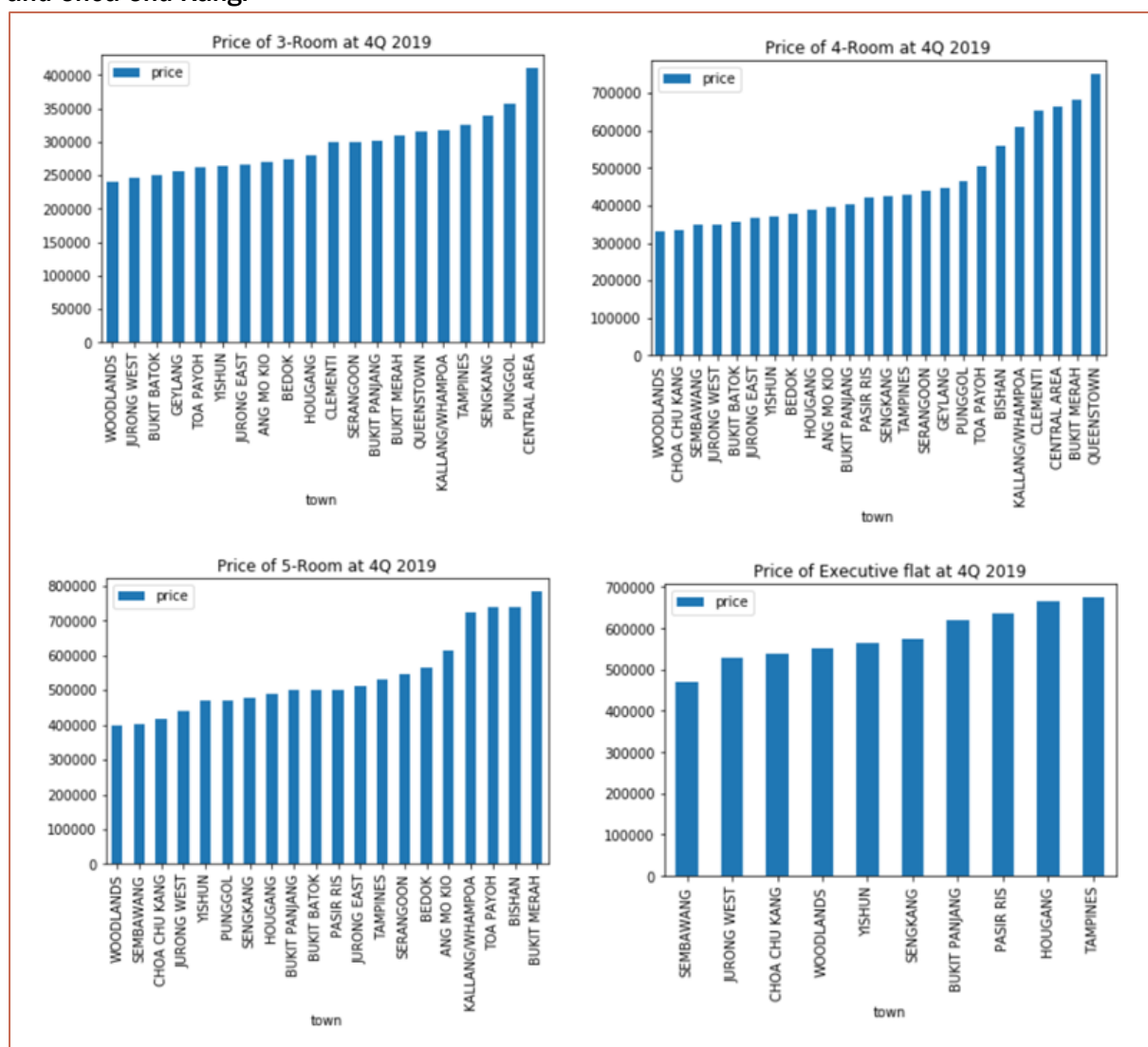

Diagram 12

Among the towns with the lowest resale flat price, Woodlands belongs to the red cluster, which is in the vicinity of venues with higher expenses. The other four towns, i.e. Sembawang, Jurong

West, Bukit Batok, and Choa Chu Kang belong to the purple cluster, which is ideal cluster for residential purpose.

Discussion

Observations

The Diagram 9 provided an overview of the clustering result. From the cluster analysis, the purple cluster was identified as the ideal cluster for residential purpose. The clustering result provided a good representation of the residential estate distribution in Singapore. The purple clusters generally match well with the actual public housing estates.
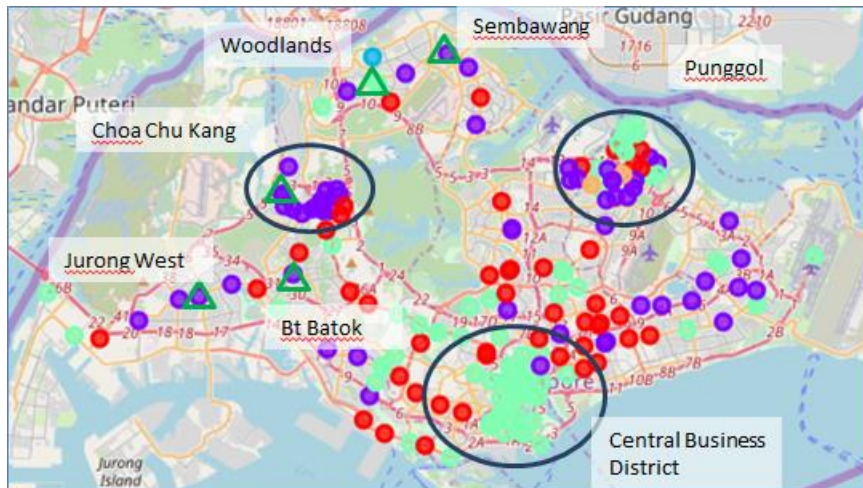

Diagram 13

Also from Diagram 9, from the density of the MRT stations distribution, one could see that the density at the South of the country is higher than the density at the North, East and West. The South side is actually the Central Business District. See Diagram 13, thus the majority of green cluster (of high expenses venues) stations are in the South.

There is a dense purple cluster at the North-West which correspond to Choa Chu Kang town. It was developed in the 80s. Majority of the residential properties in the town are public housing flats, hence the purple cluster and the price of resale price is relatively low. Compare this to the other dense mixture of cluster in the North-East, i.e. Punggol town. It is one of the newer town developed around 1998. Due to the greater effort of planning by the authority, there are a variety of venues in this new town, hence the different clusters from our analysis.

The towns with the lowest resale flat price are marked with green triangle in Diagram 13. Generally they are further away from the Central Business District and they fall under the purple cluster, except Woodlands. Woodlands is the largest town in the North, and it has rail and road link to the neighbouring Malaysia, hence the upmarket venues in the town.

Recommendations

The clustering in this project was done by assigning five clusters. The clustering could be done with less (or more) number of cluster to analyse whether a better representation could be achieved.

The project had embarked on the objective to assist PRs coming from the South East Asia region to look for ideal housing estate with relatively lower price. The analysis could also be

extended to the other group of buyers, such as the locals to identify housing estate that suit their style of living.

Conclusion

This project sets out to assist the many PRs from the South East Asia region to look for ideal housing estate near to the MRT stations. The main considerations in the selection of housing estate are the proximity to the MRT station, types of business near the station and the price of housing flats. The data of stations and the Foursquare venues were analysed and the result showed that four towns in the North and West, i.e. Sembawang, Choa Chu Kang, Bukit Batok and Jurong West, are most suitable for buyers looking for affordable properties with a good mix of venues for residential purpose.

Bibliography

1.   Land   Transport   Authority   Datamall.   [Online] https://www.mytransport.sg/content/dam/datamall/datasets/Geospatial/TrainStation.zip.

2.   Data.gov.sg.   [Online]   https://data.gov.sg/dataset/44b852a9-e7f8-4381-b896-e7c809da0f9c/download.