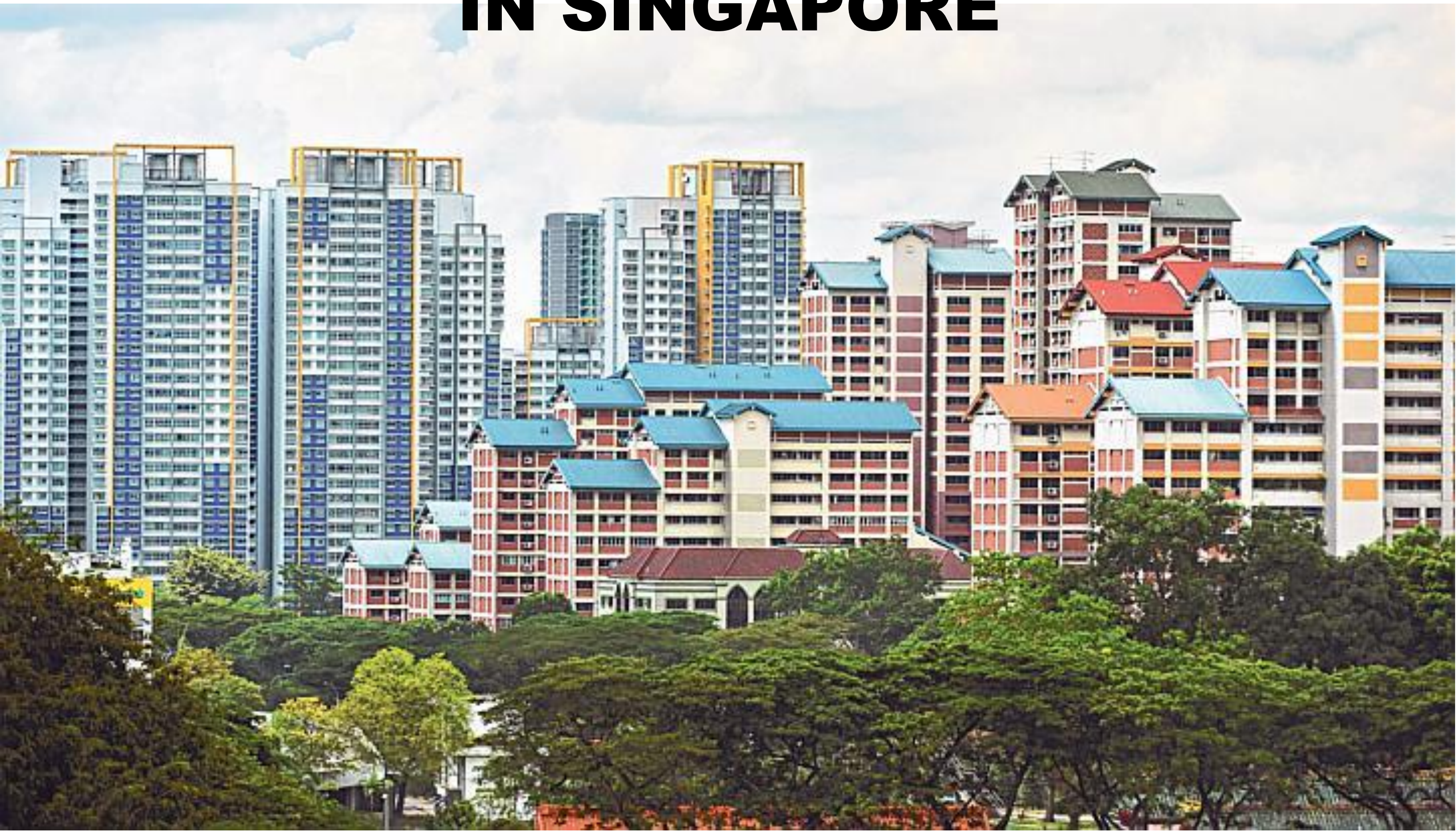
FINDING THE PERFECT HDB FLAT IN SINGAPORE



NAME: GENG ZHAOXIN

EMAIL: dorigeng.lg@gmail.com

DATE: 8th April 2019

CONTENTS

Abstract

In this report, I built a new HDB flats recommendation system for local property seekers and agents by using Python. This recommendation system stressed that customers' different needs on lifestyles is beyond location preferences, and it contains two parts: (1) a dataset that include six major variables as filter to find the perfect HDB, (2) two clustering maps which can be used as a reference in finding blocks with similar venue features. This kind of work has never been done before, and has a big implication for local property market.

Key words: HDB flats, recommendation system, similar venue features

1. Introduction

Singapore is a very small but well-developed country in Southeast Asia. It has only one city, named Singapore, which has gained a reputation for holding first Trump-Kim Summit. Long before that, Singapore is always being a very international city where most regional headquarters for top 500 fortune companies are located. According to the report of Department of Statistics Singapore, the proportion of Singapore citizens in the total population is only 61.56% in 2018, and for Permanent Residents and foreigners, it is 9.26% and 29.16% respectively. For people moving to Singapore, settling down and finding the perfect place to live would be the first thing to be considered.

However, the current property recommendation system in Singapore is not perfect as it should be. Generally speaking, if someone decides to live in a new place, before contacting a specific property agent, he or she should clearly know that which location is a preference, as different districts usually have different agents. This question is actually very difficult to be answered for residents in Singapore, especially those newcomer foreigners. Singapore has no rural areas, and the mature communities always have convenient transportation and life facilities. If your work location is in the CBD like Raffles Place, even if you live in the westmost or east most areas of Singapore, it may only take an hour from your home to work place. If you can bear the distance for 30 minutes' public transportation, you still have too many choices about where to live. As a result, region factor cannot be in the first place to classify costumers in local property market.

In this report, I try to build a new property recommendation system for local agents. I believe customers' different needs on life-styles is beyond location preference in most cases in Singapore. For example, parents with young kids would take the distance to elite primary schools on as their priority, while foreign students and young employees maybe enjoy a rapid and modern life-style, and they might choose somewhere close to MRT, coffee shops and gyms. So, I suggest we should firstly cluster properties types in the whole city (which is equal to the whole country) by differences of customers' needs, then add that customers' detailed preferences to decide which regions and what blocks are the best choice for them. Over eighty percent of Singapore residents choose to live in HDB (Housing and Development Board) blocks, and the data about those HDB blocks is very easy to get from government supported websites. As a result, I narrowed down my research scope, and focus on HDB only.

2. Methodology

2.1 Method

Data analysis method is commonly used to solve recommendation problem. In this report, I choose python as main data analysis tool to: (1) collect relevant data from websites, (2) clean and process data, and (3) create a dataset and cluster data.

Dataset is the key to make recommendations by filtering some conditions. Besides that, I intent to make an overall category result as a guideline for HDB seeker. Since at least 5 variables are included as filter conditions, if we classify the HDB blocks by applying the filter, there will be at least 32 categories (2^5) in total. As a result, it will be difficult to show a simplified visual classification result. So, I decide to use two kinds of clustering algorithms to reduce the category scale. They are k-means clustering and hierarchical clustering.

K-means clustering is more commonly used to analyze data, but hierarchical clustering is more intuitive and can be applied to small and medium size dataset. It is said in some cases, the result of these two kinds of clustering is similar. We will include the clustering results in the result section.

2.2 Data and resources

HDB blocks data: I got the basic HDB data from website *https://data.gov.sg/* , which is run by local government. This dataset includes the basic information of all the 12132 HDB blocks, such as the name of every HDB block, which street it belongs to, how many rooms it has, which year it is built and so on. However, the location information is not available in that dataset. So, I decided to use python geocoder package to get all the related location information.

Firstly, to make our analysis more efficient, I decided to put the HDB blocks that located in the same street and built in the same year together and labeled them as similar blocks. As I believe those similar blocks could be located in the same place and share the same venue features and living standard. As a result, I got 1570 similar block groups, and use the features of the first block in that group as a representative of that group. Then I got the latitude and longitude of that 1570 representatives by using python geocoder package.

Besides location data, I intend to put the completed year of the building into our dataset as a variable. Because the living condition of HDB flats is directly related with which year it is built. For example, the HDB blocks that build after year of 2000 has higher ceilings, bigger windows and of course better living standard.

Price data: In most cases of renting and buying properties, price is still the first thing to be considered. The rentals or selling price per square inch(psf) for Singapore HDB blocks varies a lot from each other. The communities where rich people clustered have relative higher psf, even if the venue features and living conditions is the same.

We collect the price data from a property website *https://www.srx.com.sg/*, which is run by SRX company. By inputting the street and block number of some block, we can get a web page showing a list of selling price details. I use the mean price from that list of blocks as our price index.

Figure 2.1 shows the mean psf of the HDB blocks. We can see that HDB flats in south is much more expensive.

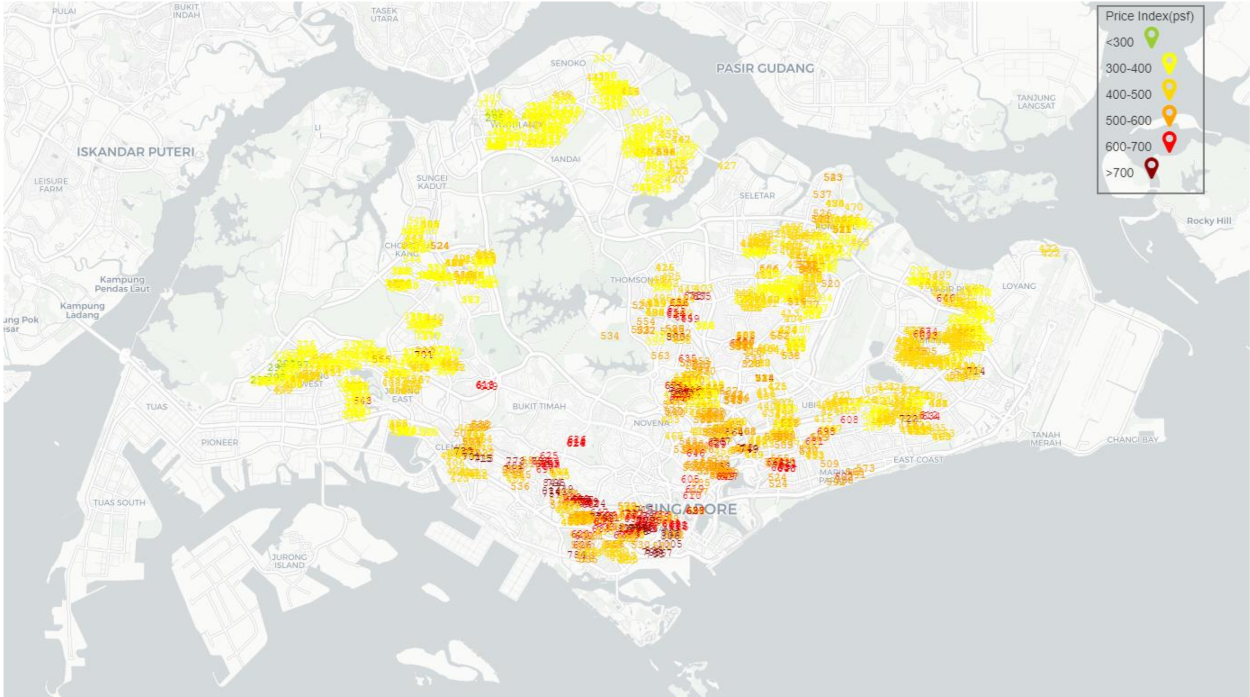
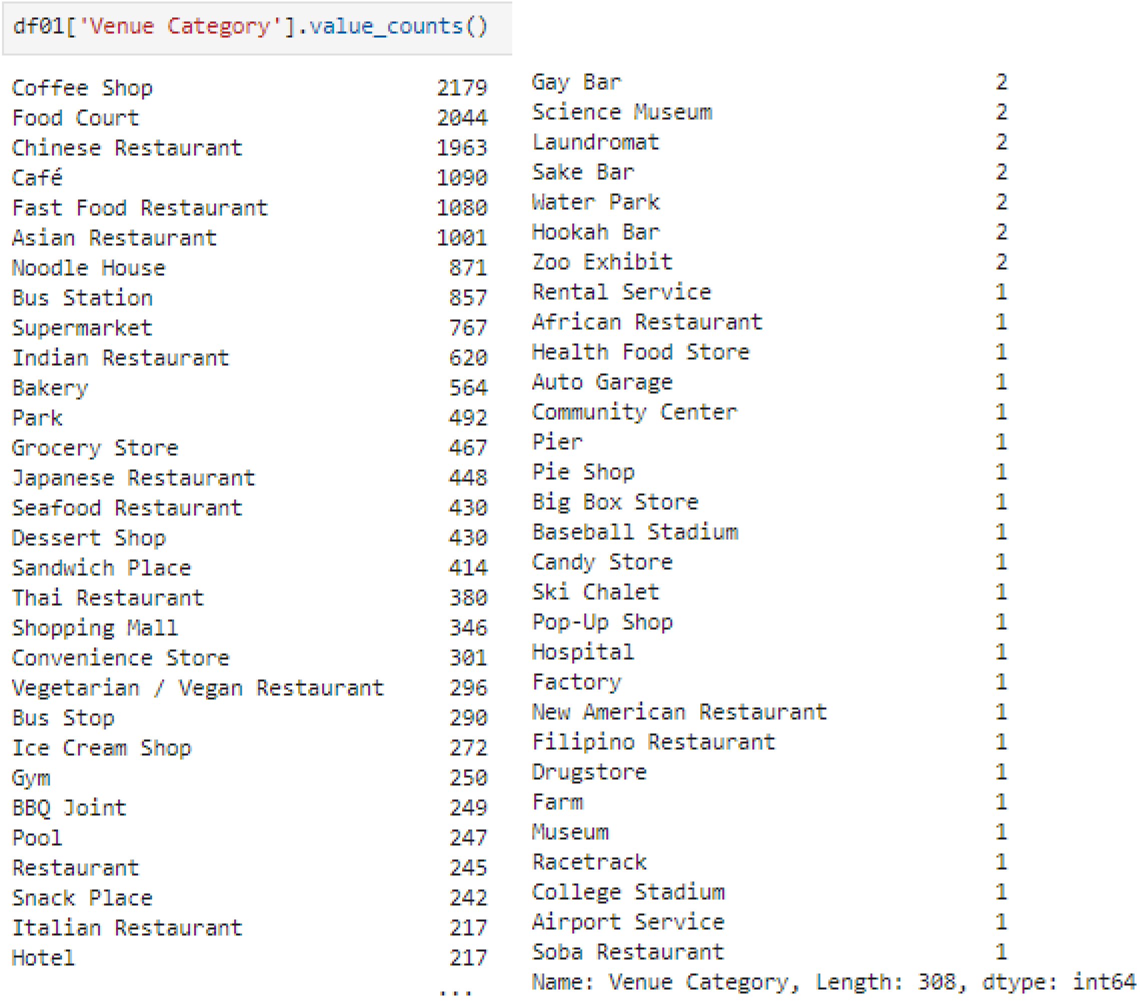


Figure 2.1 The mean psf of the HDB block in Singapore

Venue data: venue data can be used to depict the different living styles of each block. We can get the venue data from **Foursquare API**. By inputting the latitude and longitude information of each block into the Foursquare research API, we got all the venue information within 500 m of these blocks. The unique values of those venues are shown in figure 2.2



```
df01['Venue Category'].value_counts()
```

```
Coffee Shop                          2179
Food Court                           2044
Chinese Restaurant                   1963
Café                                 1090
Fast Food Restaurant                 1080
Asian Restaurant                     1001
Noodle House                          871
Bus Station                           857
Supermarket                           767
Indian Restaurant                     620
Bakery                                564
Park                                  492
Grocery Store                         467
Japanese Restaurant                   448
Seafood Restaurant                    430
Dessert Shop                          430
Sandwich Place                        414
Thai Restaurant                       380
Shopping Mall                         346
Convenience Store                     301
Vegetarian / Vegan Restaurant         296
Bus Stop                              290
Ice Cream Shop                        272
Gym                                   250
BBQ Joint                             249
Pool                                  247
Restaurant                            245
Snack Place                           242
Italian Restaurant                    217
Hotel                                 217
                                     ...
Gay Bar                                 2
Science Museum                          2
Laundromat                              2
Sake Bar                                2
Water Park                              2
Hookah Bar                              2
Zoo Exhibit                             2
Rental Service                          1
African Restaurant                      1
Health Food Store                       1
Auto Garage                             1
Community Center                        1
Pier                                    1
Pie Shop                                1
Big Box Store                           1
Baseball Stadium                        1
Candy Store                             1
Ski Chalet                              1
Pop-Up Shop                             1
Hospital                                1
Factory                                 1
New American Restaurant                 1
Filipino Restaurant                     1
Drugstore                               1
Farm                                    1
Museum                                  1
Racetrack                               1
College Stadium                         1
Airport Service                         1
Soba Restaurant                         1
Name: Venue Category, Length: 308, dtype: int64
```

Figure 2.2 Python codes that show all the names of venue categories

The next step is to refine the venue data, select those most import categories and put them in our model as variables. After asking some consulting expert and interviewing some HDB seekers, I abstract the following features as key features, which are: MRT data, Fairprice data, Primary school data, swimming complex data.

I have to say, the Singapore venue data we get from Foursquare API is not very accurate. For example, in some cases, they classify seafood restaurants into coffee shop, which is obviously wrong. So, I decide to collect those feature related data we mentioned above from other resources, which are more accurate and with more details.

MRT data: MRT is the most convenient public transportation in Singapore. A lot of people put the distance to MRT station as their priority in choosing where to live. I get the accurate MRT exit location data from the website where we got the HDB data before: *https://data.gov.sg/*.

The MRT exit location data we download was in kml format, so I used python to change it into csv format. We got xx exit data for xx MRT. By using python and the location data both from HDB blocks and MRT exits, we can easily calculate the distance between each HDB block and its nearest MRX exit.

Fairprice data: Fairprice supermarket stands for a lower living expenses in Singapore. It is a chain supermarket run by NTUC(the National Trade Union Congress), and the fruit and vegetables there is usually lower in price but fresher, compared to other chain supermarkets like Cold storage.

I downloaded Fairprice location data from its official website *https://www.fairprice.com.sg/*. It is a pdf file includes all the street information of all Fairprice stores, so I used python to abstract those text and input them in Geocoder package to get the exact latitude and longitude data.

Primary school data: For young parents who cares about kids' education, the distance to primary school might be very important for them. Singapore government claims that the primary school enrollment policy strictly obeys 'within 1 km' and 'within 2 km' regulations, while middle school enrollment only depends on students' PSLE scores. As a result, the phenomenon that parents move to a new place to get a higher chance in primary school enrollment is very common in Singapore.

I got the list of all the primary schools from website: *https://www.moe.gov.sg/admissions/primary-one-registration/information-on-primary-schools/listing-by-planning-area*. Later, I use python geocoder package again to get their location information.

Swimming complex data: In Singapore, there are 26 public swimming complexes that run by local government, and will cost only 1 dollar for each entrance. For the people living in HDB that is close to public swimming complex, the life style may be more comfortable. Swimming is a very popular in Singapore as it is a tropical country. Therefore, I put the venue feature swimming complex into our consideration.

I got the swimming complex information from website
https://www.myactivesg.com/Facilities/Swimming-Pools .

2.3 variables

Based on the data we discussed above, we can refine 6 variables, shown as table 2.1.

Table 2.1 variable names and definitions

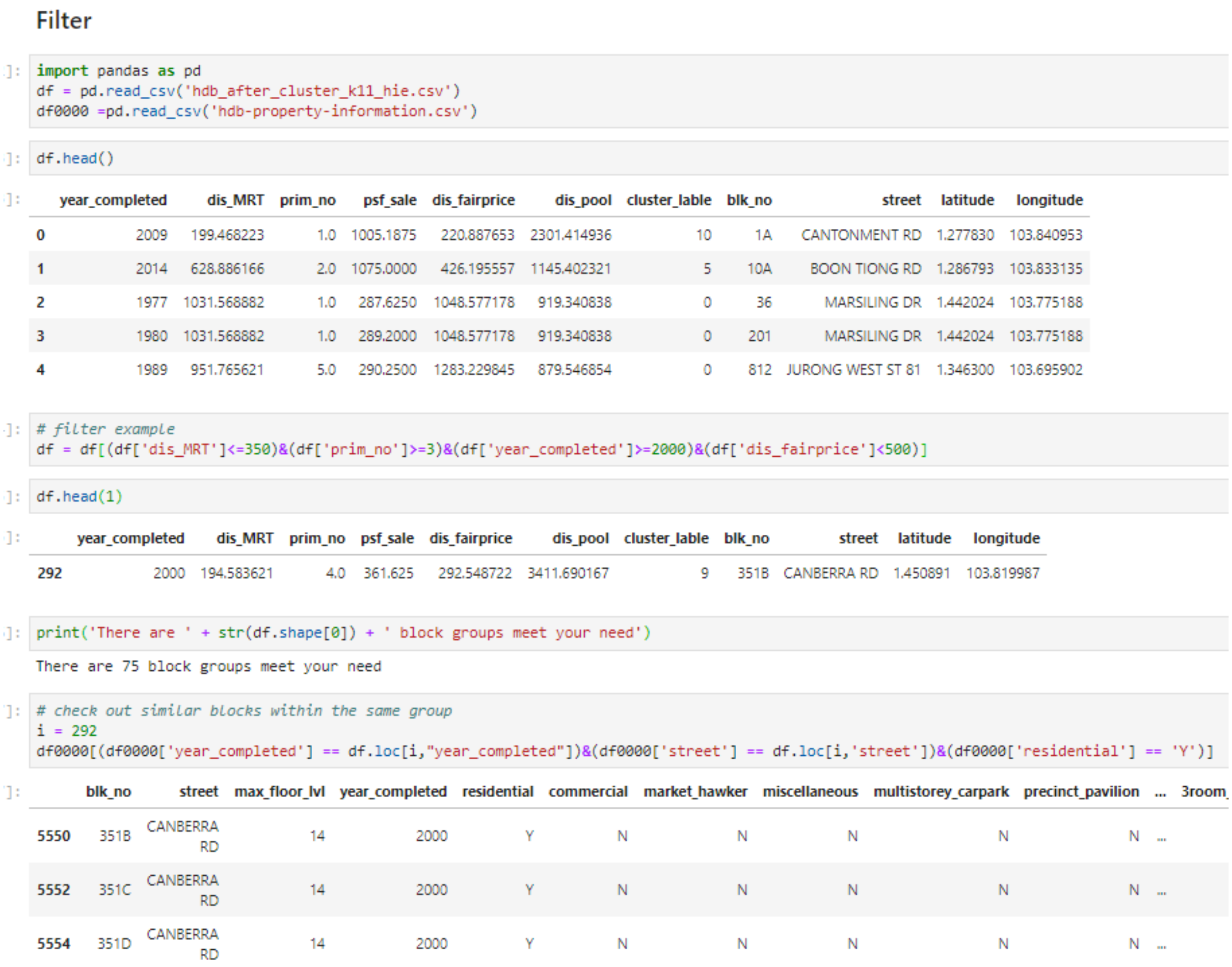
Name	Definition
dis_MRT	Distance to the nearest MRT exit
dis_fairprice	Distance to the nearest Fairprice supermarket
prim_no	Numbers of primary schools within 1km
dis_pool	Distance to the nearest swimming complex
psf_sale	Selling price per square inch
year_completed	The year when the HDB block is completed

3. Results

Our results can be divided into two parts:

(1) dataset searching for preferred HDB blocks. (an example of searching results is shown in Figure 3.1)

(2) category clustering result. (the comparison of two clustering results is shown in Figure 3.2, the reason about choosing the best k is explained in **appendix** part)



Filter

```
import pandas as pd
df = pd.read_csv('hdb_after_cluster_k11_hie.csv')
df0000 =pd.read_csv('hdb-property-information.csv')
```

```
df.head()
```

	year_completed	dis_MRT	prim_no	psf_sale	dis_fairprice	dis_pool	cluster_lable	blk_no	street	latitude	longitude
0	2009	199.468223	1.0	1005.1875	220.887653	2301.414936	10	1A	CANTONMENT RD	1.277830	103.840953
1	2014	628.886166	2.0	1075.0000	426.195557	1145.402321	5	10A	BOON TIONG RD	1.286793	103.833135
2	1977	1031.568882	1.0	287.6250	1048.577178	919.340838	0	36	MARSILING DR	1.442024	103.775188
3	1980	1031.568882	1.0	289.2000	1048.577178	919.340838	0	201	MARSILING DR	1.442024	103.775188
4	1989	951.765621	5.0	290.2500	1283.229845	879.546854	0	812	JURONG WEST ST 81	1.346300	103.695902

```
# filter example
df = df[(df['dis_MRT']<=350)&(df['prim_no']>=3)&(df['year_completed']>=2000)&(df['dis_fairprice']<500)]
```

```
df.head(1)
```

	year_completed	dis_MRT	prim_no	psf_sale	dis_fairprice	dis_pool	cluster_lable	blk_no	street	latitude	longitude
292	2000	194.583621	4.0	361.625	292.548722	3411.690167	9	351B	CANBERRA RD	1.450891	103.819987

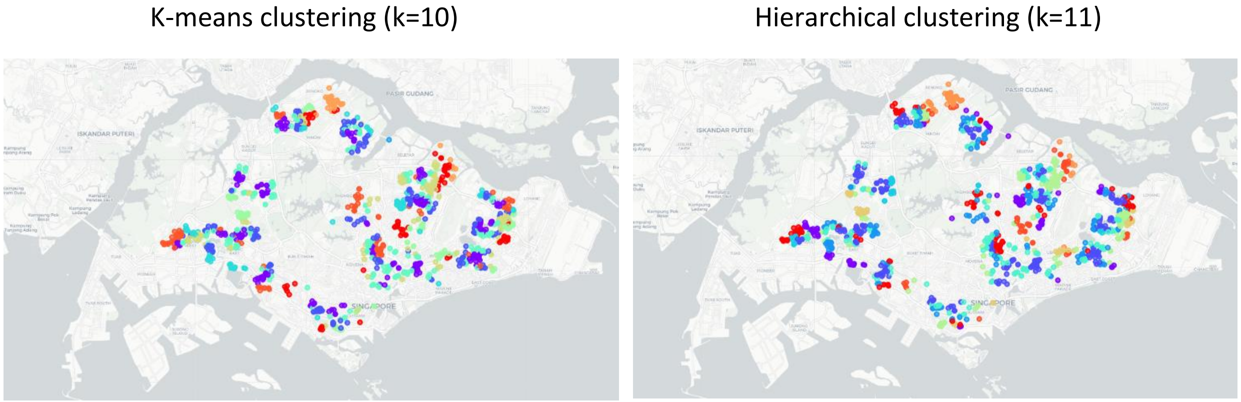
```
print('There are ' + str(df.shape[0]) + ' block groups meet your need')
```

```
There are 75 block groups meet your need
```

```
# check out similar blocks within the same group
i = 292
df0000[(df0000['year_completed'] == df.loc[i,"year_completed"])&(df0000['street'] == df.loc[i,'street'])&(df0000['residential'] == 'Y')]
```

	blk_no	street	max_floor_lvl	year_completed	residential	commercial	market_hawker	miscellaneous	multistorey_carpark	precinct_pavilion	...	3room_
5550	351B	CANBERRA RD	14	2000	Y	N	N	N	N	N	...	
5552	351C	CANBERRA RD	14	2000	Y	N	N	N	N	N	...	
5554	351D	CANBERRA RD	14	2000	Y	N	N	N	N	N	...	

Figure 3.1 An example of how to use HDB filter dataset


K-means clustering (k=10)

Hierarchical clustering (k=11)

Figure 3.2 Comparison of two clustering method results

4. Discussion

Our HDB selection dataset works very well, as is shown in Figure 3.1. The conditions we input is: the distance between the targeted HDB blocks and the nearest MRT (in the dataset it is named as dis_MRT) is less than 350 m; the numbers of primary school within 1 km (prim_no) is more than 3; the year that HDB block is built and completed is after year 2000; the distance to the nearest fairprice supermarket is less than 500 m. Turns out there are 75 HDB block groups meet the condition.

What we get from the above dataset is not the final list of targeted HDB blocks, that is because we have combined the similar HDB blocks in the same group at the first place. The next step is to get the complete HDB list from the original csv file, which contains all the blocks in Singapore. We realize it by inputting Python code that search for rows that contain the same street name and same completed year with the targeted HDB blocks. The result of the example is shown at the bottom of Figure 3.1.

By using the dataset above, we can then create some visual results, as is shown in Figure 3.2. The clustering results have two meanings: Firstly, for new clients that just moved to Singapore, the clustering map can be used as an overall review of local property market, as we can see from the map that blocks in the same community can be divided into different categories since they have different venue features. Secondly, for those who already live in Singapore, and intend to move to somewhere else with a similar living style, the map can tell them other blocks that with similar venue features.

Figure 3.2 shows that the clustering results varies according to different clustering method. Which one can be chosen as the better method depends on checking on categories' details.

.

5. Conclusion and expansion

So far, what I did in this report is just a first step in building a complete recommendation system. Clustering those HDB blocks helps potential property demanders to have a basic review of local HDB market. But individuals’ demand differs from person to person. A customized filter system is also needed to target the perfect HDB. Therefore, my plan is to create an app (by using python GUI package PyQT) with the clustering results we already got, and the existed data is also needed to create the filter interface. In apple store, there is currently no such an app to give Singapore residents guidelines like this.

Besides HDB data, I also consider in the future build a whole property system by adding condo data, even though it only owns 20% of the market share. Because those condos are, in some cases, substitutions of HDB blocks that build in after year 2000 (both in better living condition and higher rentals than the average HDB).

6. References

[1] https://en.wikipedia.org/wiki/Hierarchical_clustering

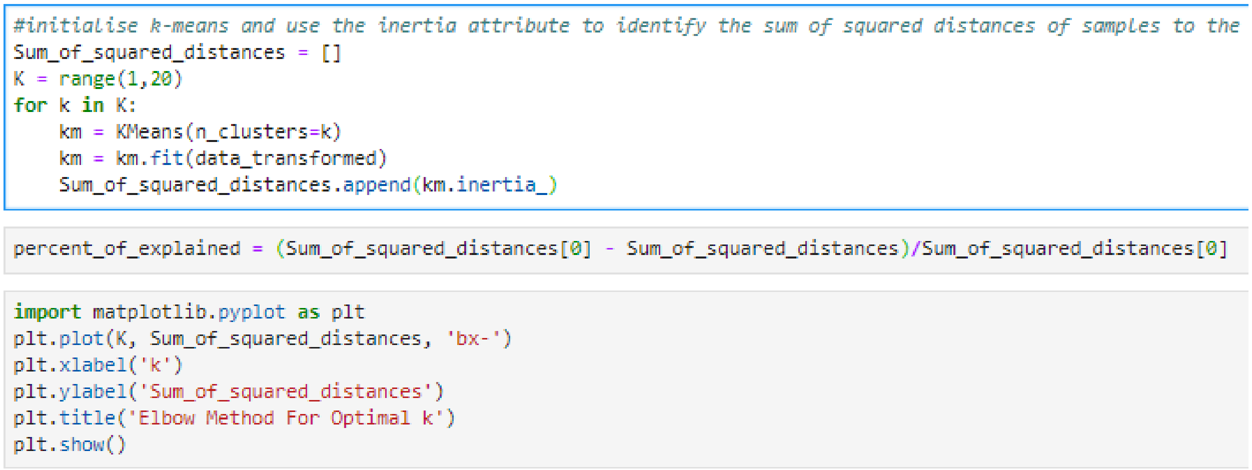
[2] https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/

[3] https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

[4] https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f

7. Appendix

7.1 Best k choice in k-means cluster analysis



```
#initialise k-means and use the inertia attribute to identify the sum of squared distances of samples to the
Sum_of_squared_distances = []
K = range(1,20)
for k in K:
    km = KMeans(n_clusters=k)
    km = km.fit(data_transformed)
    Sum_of_squared_distances.append(km.inertia_)
```

```
percent_of_explained = (Sum_of_squared_distances[0] - Sum_of_squared_distances)/Sum_of_squared_distances[0]
```

```
import matplotlib.pyplot as plt
plt.plot(K, Sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```

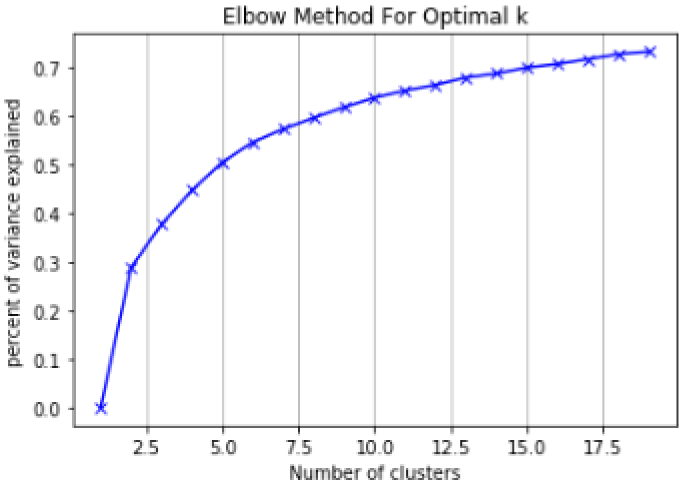
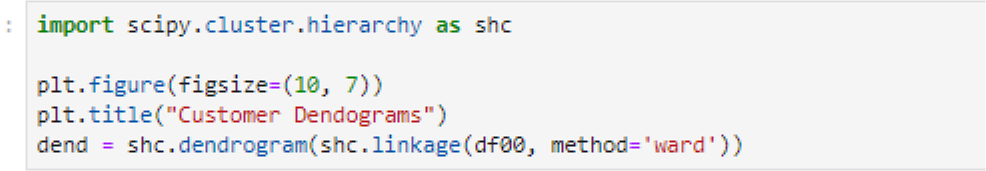


Figure 7.1 best k choice in k-means cluster analysis

7.2 Best k choice in Hierarchy cluster analysis

clustering -hierarchy class



```
import scipy.cluster.hierarchy as shc

plt.figure(figsize=(10, 7))
plt.title("Customer Dendograms")
dend = shc.dendrogram(shc.linkage(df00, method='ward'))
```

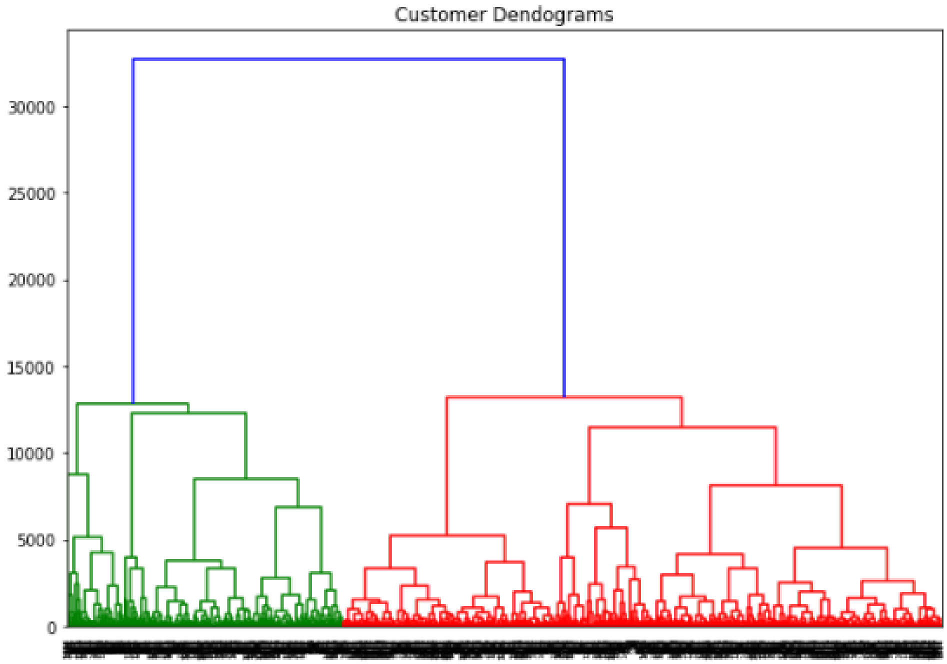


Figure 7.2 best k choice in Hierarchy cluster analysis