Report on Capstone Project -- Singapore Private Property Analysis (Price, Locations, Nearby Venues, Primary Schools)

Zeng Yunjia

1. Introduction

1.1 Business Problem

Singapore, the city state and island country located in Southeast Asia, is one of the original Four Asian Tigers. Having surpassed its peers in terms of GDP per capita, Singapore has one of the highest stand of living in Asia-Pacific [1]. With its "forward-thinking policies", Singapore has delivered on world-class infrastructures, and has become renowned for its airports, housing, safety, as well as the advanced infocomm networks, making Singapore one of the most attractive cities to live in Asia [2].

Singapore's private properties, therefore, continue to be highly sought-after by not only Singaporeans but also foreign investors. Buying properties is never a trivial decision to make, which becomes particularly true if the context is in Singapore with the high prices of its private properties. However, there are way too many factors that need to be considered in the decision making process of buying properties. Data analyses therefore can be very useful in this aspect thanks to its capability to help in decision-making processes.

1.2 Target Audience

This project aims to collect data about the private properties in Singapore, and to perform data analyses on the collected data. Potential buyers of Singapore's private properties as well as property agents can benefit from this project in terms of better understanding of the available properties in market, cross-comparison of the properties, as well as information about near-by venues that the future property owners could enjoy. One special bonus section will be dedicated to home buyers with preschool kids, with the primary schools that are nearby properties be sorted out and analyzed for them. Potential home buyers or investors can therefore make more informed decisions based on the results of this project, and property agents could also provide better recommendations based on the report to their customers.

2. Data

Based on the definition of the business problem, essential data for the analysis include:

* Private properties for sale in the market

* Recorded sale prices of the properties

* Locations of the properties

* Total units for each properties

* Venues nearby the properties

* Primary schools that are located within 1km of the properties

Due to the data size and availability constraints, the analyses will focus on the properties directly sold by all developers in a six-month time frame (from 2019 October to 2020 March).

Following data sources will be utilized to generate the required data:

* The private property sale records from Singapore URA website [3]

* The location data of each property extracted from **geopy.geocoders**

* A number of venues with their categories within a specified distance (1km) of every single property obtained via **Foursquare API**

* All the primary schools within a distance of 1km of every single property obtained via **Foursquare API**

2.1 Private Property Sale Data

The Singapore private property sale data from 2019 October till 202 March was downloaded from URA website [3], and saved as csv files. The data are retrieved and put into dataframes. For the purpose of this analysis, only the columns with property name, street address, developer information, type and locality of the properties, total units of the property, median price of the property are kept. The median price is calculated as the average over the six-month period. The dataframe to be used is ‘cfa_df’. The sale data are based on the properties sold directly from developers.

2.2 Private Property Location Data

The location data with latitude and longitude for every property is retrieved from geocoder, and included in the 'cfa_df' dataframe.

2.3 Private Property Nearby Venue Data

The top 100 venues within a distance of 1 kilometer from the property are retrieved from Foursquare. The 'property_venue' dataframe includes the venue name, venue location, and venue category besides the property information for each property.

2.4 Private Property Nearby School Data

To retrieve the nearby school information, the 'property_venue' dataframe was first explored. But it turns out very few schools are included in the top 100 venues for the properties. Another query was made via Foursquare to retrieve the top 100 schools within a distance of 1 kilometer from the properties. Among the retrieved schools, primary schools are picked and stored in the dataframe 'ps_df'.

2.5 List of All the Dataframes to Use and Initial Visualization

```
1 cfa_df.head()
```

	Total Units	Median Price	Property	Street	Developer	Type	Locality	Latitude	Longitude
0	56.0	2807.0	10 EVELYN	EVELYN ROAD	Creative Investments Pte Ltd	Non-Landed	CCR	1.31674	103.84
1	56.0	3234.0	120 GRANGE	GRANGE ROAD	RH Orchard Pte Ltd	Non-Landed	CCR	1.29967	103.825
2	101.0	3351.0	19 NASSIM	NASSIM HILL	Parksville Development Pte Ltd	Non-Landed	CCR	1.30665	103.821
3	58.0	1855.0	1953	TESSENSOHN ROAD	Oxley Amethyst Pte Ltd	Non-Landed	RCR	1.31523	103.856
4	96.0	3551.6	3 CUSCADEN	CUSCADEN WALK	SL Capital (2) Pte Ltd	Non-Landed	CCR	1.30387	103.829

Fig. 1. First five rows of the dataframe 'cfa_df'

```
1 property_venues.head()
```

	Property	Property Latitude	Property Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	10 EVELYN	1.316736	103.83962	Ah Chew Desserts 阿秋甜品 (Ah Chew Desserts)	1.318411	103.843714	Dessert Shop
1	10 EVELYN	1.316736	103.83962	Hai Yan BBQ Seafood	1.312084	103.839500	Seafood Restaurant
2	10 EVELYN	1.316736	103.83962	Udders	1.318253	103.843948	Ice Cream Shop
3	10 EVELYN	1.316736	103.83962	Chui Huay Lim Teochew Cuisine	1.313970	103.841545	Chinese Restaurant
4	10 EVELYN	1.316736	103.83962	Starbucks Reserve	1.317673	103.844021	Coffee Shop

Fig. 2. First five rows of the dataframe 'property_venue'

All the dataframes are listed in the figure below with the first five records shown in Figs. 1-4.

```
1 ps_df.head()
```

	Property	Property Latitude	Property Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	120 GRANGE	1.299670	103.825410	Alexandra Primary School	1.291213	103.823912	Elementary School
1	ARTRA	1.290528	103.816572	Gan Eng Seng Primary School	1.285765	103.815187	Elementary School
2	ARTRA	1.290528	103.816572	Alexandra Primary School	1.291213	103.823912	Elementary School
3	AVENUE SOUTH RESIDENCE	1.276497	103.830543	Radin Mas Primary School	1.274728	103.823972	Elementary School
4	AVENUE SOUTH RESIDENCE	1.276497	103.830543	Zhangde Primary School	1.284212	103.826825	Elementary School

Fig. 3. First five rows of the dataframe ‘ps_df’

```
7 allps_df.head()
```

```
Data downloaded!
(190, 7)
```

	Name	Funding	Type	Area[3]	Notes	Website	School Code
0	Admiralty Primary School	Government	Mixed	Woodlands	NaN	[1]	1744.0
1	Ahmad Ibrahim Primary School	Government	Mixed	Yishun	NaN	[2]	1738.0
2	Ai Tong School	Government-aided, SAP	Mixed	Bishan	Affiliated to Singapore Hokkien Huay Kuan[4]	[3]	5625.0
3	Alexandra Primary School	Government	Mixed	Bukit Merah	NaN	[4]	1266.0
4	Anchor Green Primary School	Government	Mixed	Sengkang	NaN	[5]	1254.0

Fig. 4. First five rows of the dataframe ‘allps_df’

The list of all the Singapore primary school information were retrieved from the website [4] and stored in ‘allps_df’ dataframe.

```
17 map_singapore
```

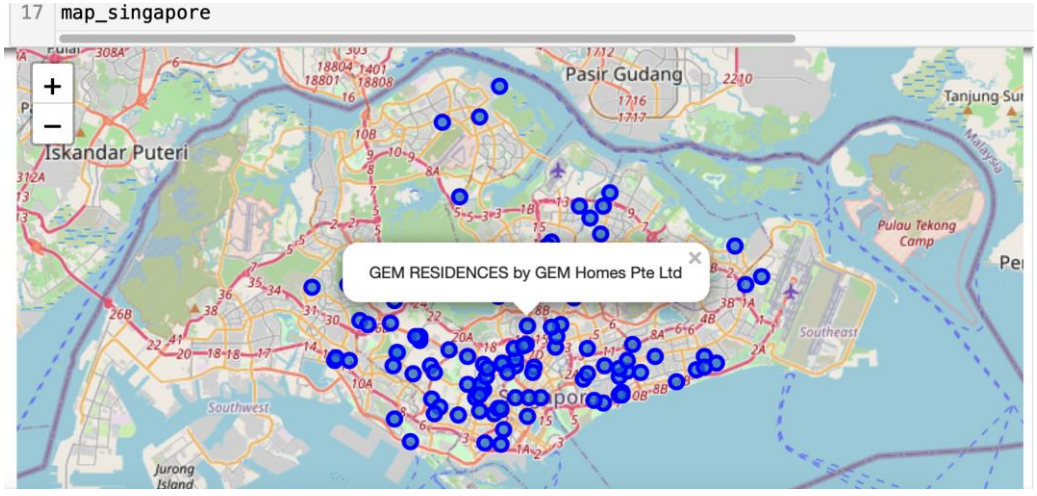


Fig. 5. Visualization of the properties on Singapore map.

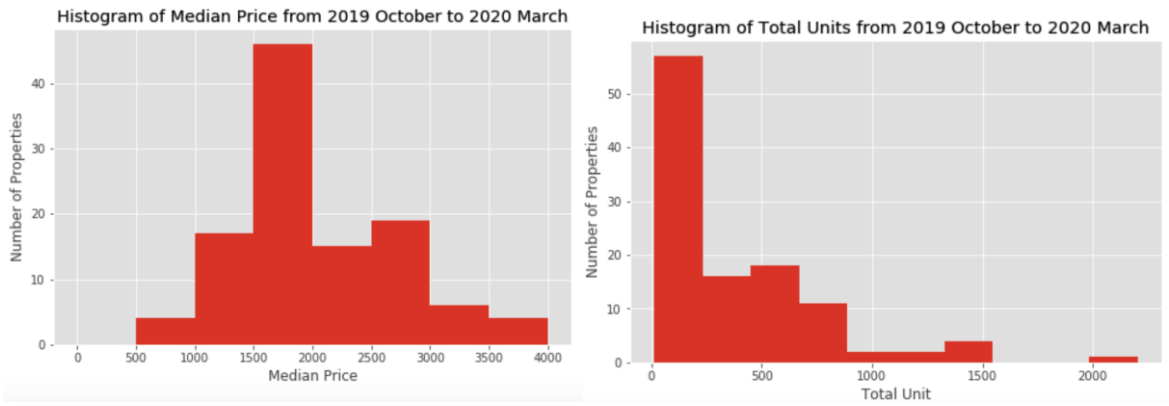
All the properties considered are plotted as markers on the Singapore map as show in Figure 5. Each pop-up shows on property and its developer.

3. Methodology

In this section, exploratory analysis will first be performed on the data. Different types of diagrams are used to visualize the property data. The dataframes are also cross checked to find any abnormal entry, which are then corrected. K-means is used to cluster all the properties considered by the types and density of their nearby venues. The generated different clustered are examined in details to find their individual characteristics. The school information is then apply to sort out properties that are suitable for different customers.

3.1 Exploratory Data Analysis

First, the median prices and total units in property data are visualized through histogram.



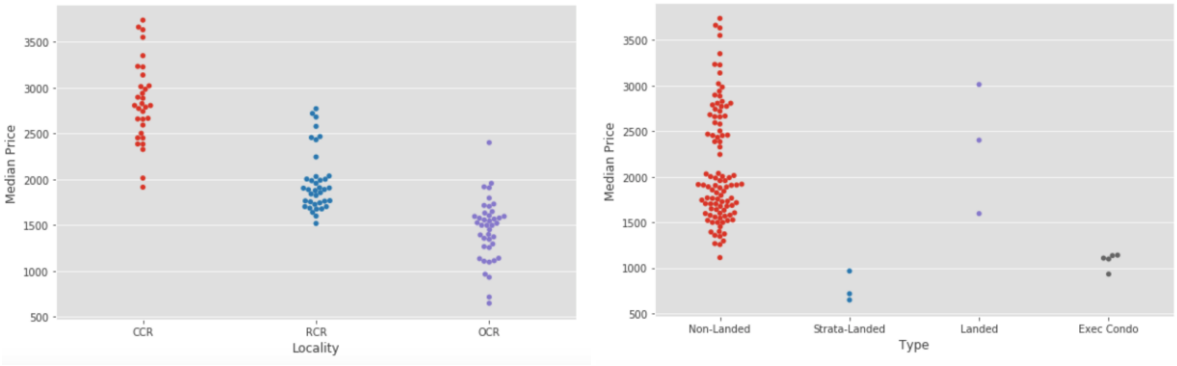
(a) (b)

Fig. 6. Histograms of (a) median price and (b) total units of the properties in Singapore.

As shown in Figure 6, the median price mostly ranges from 1500 to 2000 SGD per square feet, while majority of the total units in properties are less than 500.

As for the category data locality and type of the properties, swarm plots are used to visualize their distribution together with median price. Such plots adjust the points along the categorical axis and prevent them from overlapping.

As can be seen from Figure 7(a), the median price varies significantly across different locality type, with CCR type has the highest property price, followed by RCR, and the OCR being the category with the lowest price.



(a) (b)

Fig. 7. Swarm plots of (a) Locality and (b) Type versus median prices.

For the property type distribution, it can be seen that most of the properties are of the type 'Non-landed', which also has the widest spread of price range. The results will be further discussed in Section 4.

When exploring the data of venues nearby properties, it is first noted that there are in total 300 unique categories of venues. The venue data are then encoded via one-hot of the 300 unique categories, and the data are further grouped by every single property considered in the analysis. It was then figured out that the number of properties in the venue data is one less than the original property data. The missing property is found out, and it turns out that the geographical data of this missing property is wrongly obtained from geocoder with a negative values in its latitude. The correct geo data is generated, and the dataframes are updated accordingly.

The densities of each venue categories nearby the every single properties are then calculated. And the top five venues of the properties are printed out in the notebook. In the next step, the top ten common venues of every property are summarized into the dataframe 'property_venues_sorted'.

3.2 Clustering with Venues

In this section, the venue data that has been preprocessed in the previous section is used to cluster all the properties. Grid search is used to figure out the best cluster value for the k-mean approach. It turns out the properties can be optimally categorized into seven clusters based on the density of venue types near the properties. The ‘singapore_merged’ dataframe contains all the information, including property information, top ten venue information, as well as the newly assigned cluster labels. As shown in Fig. 8, the clustered properties are then visualized in Singapore map, with different colors of the markers representing different clusters.

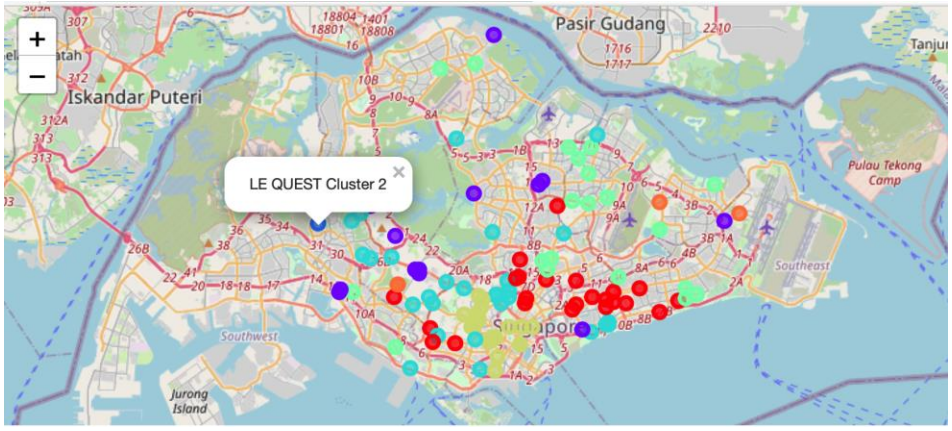


Fig. 8. Visualization of the clustered properties on Singapore map.

The seven clusters are further examined in details to better understand the characteristics of the property clusters, and to identify the discriminating venue categories that distinguish each cluster. And thus, the results can be used to make recommendation to our target audience. The detailed results and discussion will be in Section 4.

3.3 Primary School Sortout

The primary school data are grouped by the nearby property to find out the 41 properties that have primary school within one kilometer distance. The price distribution of these properties are plotted as in Fig. 9. It shows a similar distribution as that of the median price distribution of all the properties. Most of the properties has a median price between 1000 to 2000 SGD per square feet. The plot indicates that the primary school might not be very influential in determining the price of the properties.

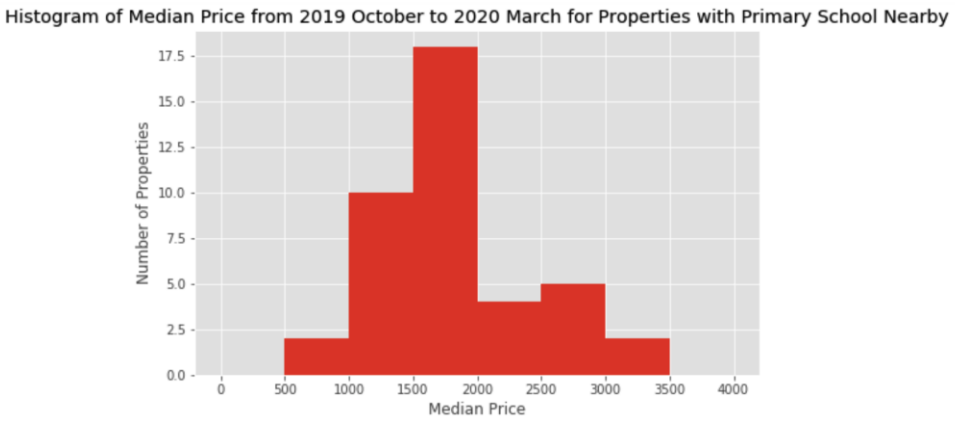


Fig. 9. Histogram of the median prices of properties with schools nearby.

All these properties together with the nearby schools are plotted in Figure 10 on Singapore map. The properties are represented as circle markers with colors representing the cluster labels. The schools are in green icon markers.

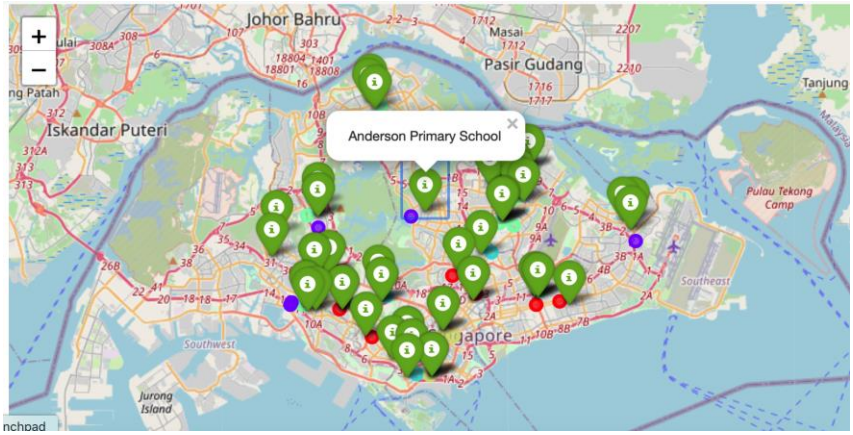


Fig. 10. Visualization of the properties and their nearby schools.

In the next step, the four primary schools that are affiliated with Hokkien Huay Kuan are sorted out. And all the special assistance plan (SAP) schools are also filtered out. These schools are included in the dataframe ‘sps_df’. Those properties that are within one kilometer distance from these schools are also sorted out in the dataframe ‘picked_merged’. These properties are visualized on Singapore map as shown in Fig. 11.

As can be observed from Fig. 11, only nine properties are within the one-kilometer distance from these schools, and they belong to three clusters “2, 4, 5”.

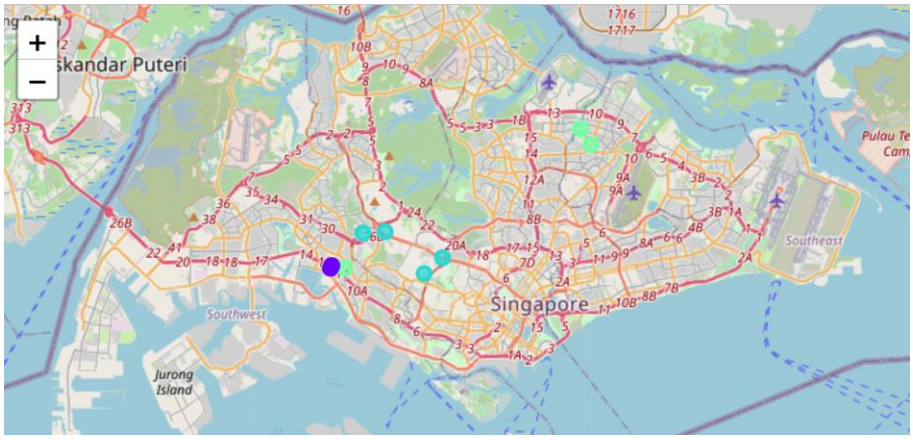
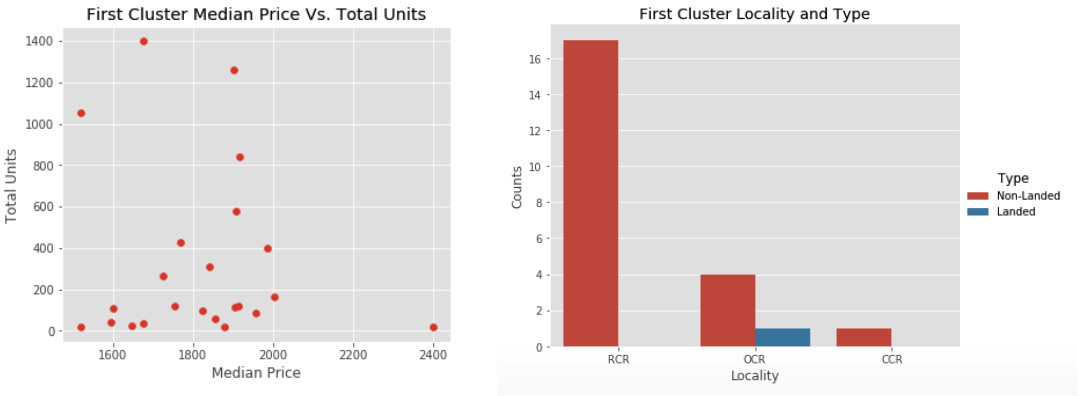


Fig. 11. Visualization of the properties that are nearby SAP and affiliated schools.

4. Results and Discussion

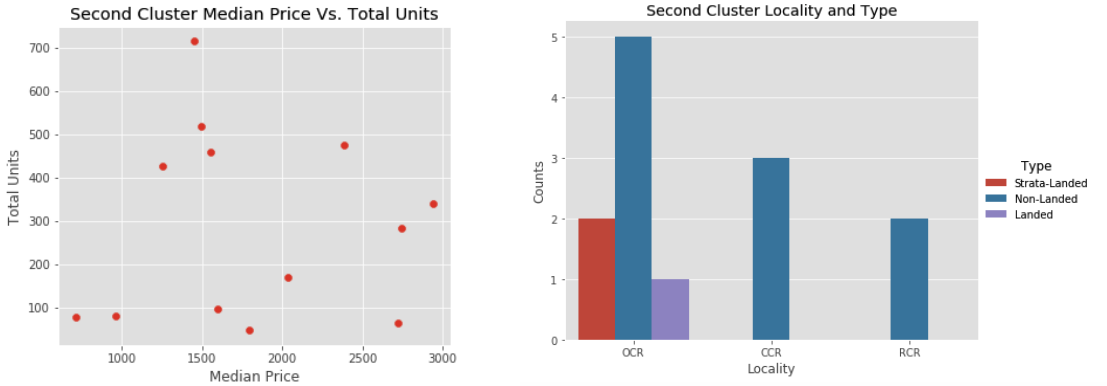
First of all, based on the property sale data analysis, it is found that the median price mostly ranges from 1500 to 2000 SGD per square feet, while majority of the total units in properties are less than 500. The median price varies significantly across different locality type, with CCR type has the highest property price, followed by RCR, and the OCR being the category with the lowest price. This is in accordance with expectation, as location plays a significant role in pricing properties. These three regions are Core Central Region (CCR), Rest of Central Region (RCR) and the Outside Central Region (OCR). As the name suggests, the CCR and RCR are the central regions of Singapore, and therefore the property prices for these two regions are higher than those in OCR. CCR has properties with the highest price because of its premier location. In view of the property type, most properties considered are non-landed, and their price ranges from 1000 to 4000 SGD per square feet; while the strate-landed properties have median prices between 500 to 1000 SGD per square feet; landed properties are with per-square-feet price above 1500 SGD; and the last type of executive condo are with median price around 1000 SGD per square feet.

Next, we will analyze every single clusters of the properties. For each cluster, the scatter plot of the property median price and total units is shown in part (a) of the figures, and the histograms of locality and property type is also included in part (b). The detailed listing of the venues in each clusters are printed out in the notebook.



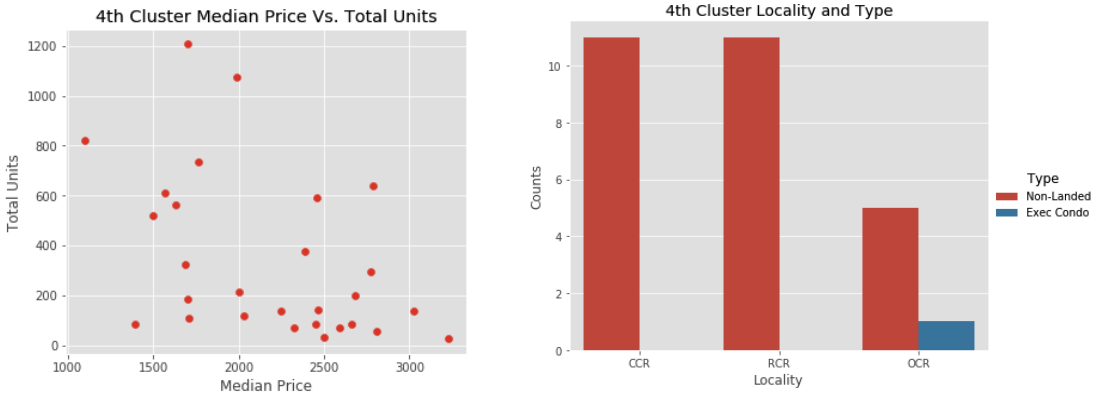
(a) (b)

Fig. 12. (a) Scatter plot and (b) bar plot of first cluster.



(a) (b)

Fig. 13. (a) Scatter plot and (b) bar plot of second cluster.

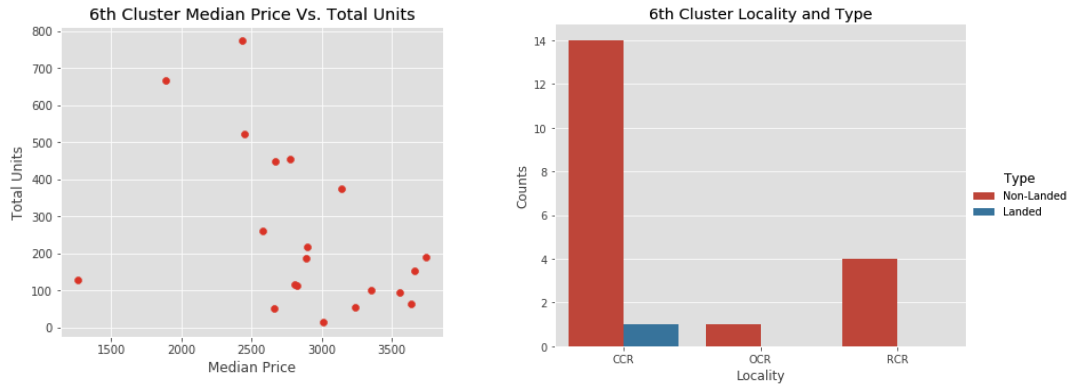


(a) (b)

Fig. 14. (a) Scatter plot and (b) bar plot of fouth cluster.

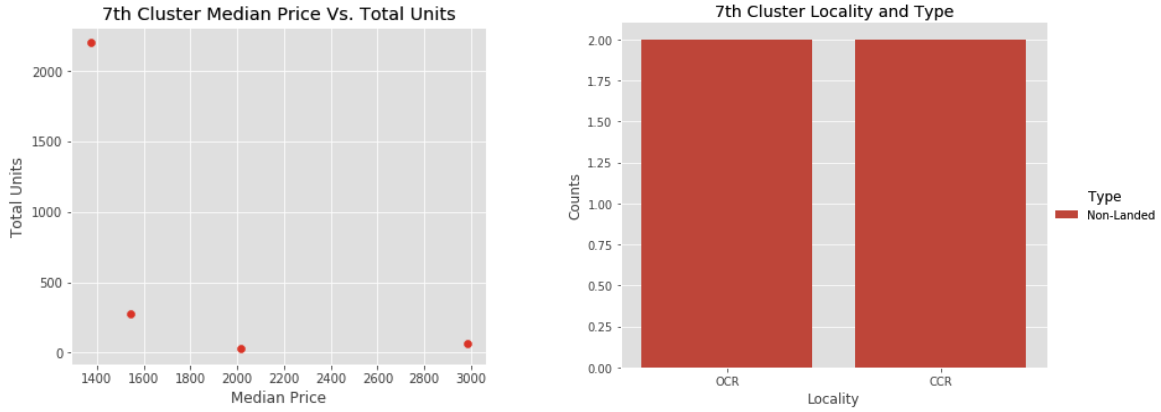
(a) (b)

Fig. 15. (a) Scatter plot and (b) bar plot of fifth cluster.



(a) (b)

Fig. 16. (a) Scatter plot and (b) bar plot of sixth cluster.



(a) (b)

Fig. 17. (a) Scatter plot and (b) bar plot of seventh cluster.

As shown in Fig. 12, the majority properties in the first cluster have a median price ranging from 1500 to 2000 SGD per square feet, and they are mainly located in the RCR region. Examining the detailed venue listing in the notebook, it can be found that the first cluster has food courts, various restaurants and cafes as the most common venues.

In the second cluster, the properties has a wider range of prices, and they include three different property types and locations as in Fig. 13. After examining the venues listing, it can be found that the most common venues include bus stations, cafes, restaurants, and recreation centers.

The third cluster only has one property with most common venues as coffee shop, restaurant, and gas station. Further examining the location of the property as in Fig. 8, it can be seen that it is the west most region of Singapore. The distinct location may present significantly different venue categories as compared with its counterparts.

In Fig. 14, it can be observed that the fourth cluster has a significant amount of properties that are in the price range between 2000 to 3000 SGD per square feet, and these properties are typically with a size smaller than 500 units. Examining the location and type of these properties, we can find that the properties in these cluster mainly located in CCR and RCR. The most common venues in this cluster include cafes, restaurants, bus stations, and hotels. It is also interesting to note that venues in more varieties are present in this cluster, including spa, gym, trail, massage studio, and etc.

In the fifth cluster, it is found that the properties are all priced below 2000 SGD per square feet, and they are mainly located in OCR, which may explain their relatively lower price as the location is outside city center. Looking at the venue list in notebook, it can be found that the most common venues in this clusters are coffee shops, food courts, bus stops and gyms.

The sixth cluster has properties priced mainly above 2500 SGD per square feet, and they are mainly located in CCR and RCR. The most common venues in this category includes hotels, boutiques, and Japanese restaurants. Shopping malls, spas, bars, and lounges are also commonly present near these properties.

In the last category, there are only four properties, and their common venues include bus station, cafes, and restaurants. It is noted that the most common venue in this category is bus station.

In summary, it is interesting to find that although the clustering is performed based on the density and type of venues nearby properties, the resulted clusters also exhibit distinct characteristics in terms of location and price. It is particularly obvious in clusters 1, 4, 5, and 6: in which cluster 1 is priced between SGD 1500 to 2000, and mainly located in RCR region; cluster 4 with average price between SGD 2000 to 3000 and include quite a significant amount of properties in CCR; cluster 5, on the other hand, are all priced below SGD 2000, and mainly located in OCR regions; cluster 6 has the highest median price above 2500 SGD, and they are mostly in central regions CCR and RCR. The categories of nearby venues are also quite distinct, like in cluster 6, hotels, boutiques, and Japanese restaurant are most common, while in cluster 5, the most common types are coffee shops, food courts and bus stops.

As for the schools data analysis, there are 41 properties that have primary school within one kilometer distance. The price distribution in Fig. 9 exhibits a similar trend as that of the median price distribution of all the properties. Moreover, among these properties, nine of them are located within 1 kilometer distance of the schools that are either affiliated or have special assistance plan. And these properties belong to cluster 2, 4, and 5. Parents with preschool kids may prefer these properties taking into account their children's admission to primary schools.

5. Conclusion

In this project, the Singapore properties have been studied in terms of price, total units, locality, type, nearby venues, and primary school. The property data includes all the properties sold by developers directly over the time period of 2019 October to 2020 March. The exact location data of the properties were then generated via Geocoder, and used to obtain nearby venue information including primary schools through Foursquare. Exploratory data analysis was first performed on the property data to generate an initial understanding of the Singapore properties. In examining the venue data, it was found that the location data was wrongly generated, and the datasets are all corrected and modified. Clustering of the properties was then performed based on the types and density of nearby venues. K-means and grid search were applied to find the best clustering results, which results in 7 clusters. Each of the seven clusters were then examined in terms of detailed listing of common nearby venues and plots of property price, number of units, locality, and type. The cluster characteristics were further analyzed and discussed. Last but not

least, data on all the primary schools in Singapore were also processed, and the schools that are nearby properties were sorted out. It will be of particular interest to parents with preschool kids, as the distance to primary school serves as one of the important criteria in children's admission to school. The schools that are with affiliation and special assistance plan were also sorted out. Based on the analysis results and discussion, investors and home buyers who intend to purchase Singapore properties, as well as property agents, could have a more systematic view on the property information in Singapore. They could narrow down their search based on the clusters. Each cluster has their characteristic price, type and locality, as well as the most common near-by venues, which also indicates features that could be essential for customers' life styles. The school information could be very useful for home buyers with preschool kids.

6. References

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

[2]. https://www.reuters.com/brandfeatures/infrastructure2030/singapore-hub

[3]. https://www.ura.gov.sg/realEstateIIWeb/price/search.action

[4]. https://en.wikipedia.org/wiki/List_of_primary_schools_in_Singapore