Coursea Capstone 2021

Singapore: a global city of opportunity

Authored by: Scarlett GENDREY

Table of content

Table of figures

Introduction

Background: Singapore is located off the southern edge of the Malay Peninsula, between Malaysia and Indonesia and has the 2nd highest population density in the world [1]. It has a population of 5.7 million with a population density of 7,894/km2 [2].

Problem: The business problem of this project is to analyze and provide insights into discovering ideal locations for people who are migrating such as housing prices and schools. Offering them a look at which neighborhood might be the best fit for easy access to cafes, supermarkets, grocery shops, malls, theaters, etc. It will help people make smart and efficient decisions about which area and neighborhood to choose before moving to a new city, state, country.

Target audience: The target audiences are foreigners who intend to move to Singapore or Singaporeans who simply want to know better their neighborhood.

Data description

The different data gathered for this project are

- The postal districts, sectors and general location of Singapore's neighborhoods. Singapore's postal codes are in 6 digits format and the first 2 digits represent the postal sector [3].
- The latitude and longitude of the neighborhoods for visualization and plotting of data which can be obtained from the Python Geocode package.
- Private Property Transaction Data API (Feb 2019 to Feb-2021) by postal district [4].

- Private education institution by region [5].
- FourSquare API to get the venues by district [6].

Methodology

Data preparation

Data downloaded or scraped from multiple sources were combined into one table.

- The postal districts data was web scraped from Wikipedia [3]. Pandas library was used to extract the table of neighborhood data. There are a total of 28 districts in Singapore. The postal districts (colum1 figure1) were first implemented in 1950 and then replaced by the postal sectors in 1955 (colum2 figure1). Although the old districts are not used by the Singapore Post anymore, they are still widely used for buying and selling of properties. That's why for this project postal districts are more relevant than postal sectors. Also, due to the land size of the country, one postal district can refer to a few locations in the vicinity.
- In order to use FourSquare API, geographical coordinates are needed. The Geocoder package was used to allow the conversion of "general location" data (addresses) obtained via web scraping into geographical coordinates.
- The data retrieved from URA API [4] which consist of private property transactions in the last 2 years by postal district. The data was downloaded in 6 batches in pandas and were concatenated. The average price per district in then calculated and extracted.

- The Private education institution data was downloaded from Data.gov.sg [5]. The geojson file contains, name, address and geographical coordinates of each school. From this data we gather the number of schools per postal district.

After gathering and transforming each different data source information, they all gathered into one DataFrame.

	Postal district	Postal Code	General location	Latitude	Longitude	Private schools	Price ($)
0	1	01, 02, 03, 04, 05, 06	Raffles Place, Cecil, Marina, People's Park	1.281890	103.849120	NaN	3.648721e+06
1	2	07, 08	Anson, Tanjong Pagar	1.278890	103.845390	6.0	1.802246e+06
2	3	14, 15, 16	Bukit Merah, Queenstown, Tiong Bahru	1.293048	103.806248	24.0	1.547536e+06
3	4	09, 10	Telok Blangah, Harbourfront	1.265331	103.818861	6.0	1.627808e+06
4	5	11, 12, 13	Pasir Panjang, Hong Leong Garden, Clementi New...	1.314730	103.756799	2.0	1.299735e+06
5	6	17	High Street, Beach Road (part)	1.290619	103.849451	2.0	NaN
6	7	18, 19	Middle Road, Golden Mile	1.299462	103.852847	NaN	1.457149e+06
7	8	20, 21	Little India, Farrer Park, Jalan Besar, Lavender	1.307100	103.858420	NaN	1.262430e+06
8	9	22, 23	Orchard, Cairnhill, River Valley	1.306540	103.839410	20.0	2.375163e+06
9	10	24, 25, 26, 27	Ardmore, Bukit Timah, Holland Road, Tanglin	1.323305	103.784985	2.0	3.122330e+06
10	11	28, 29, 30	Watten Estate, Novena, Thomson	1.326670	103.811390	NaN	1.841279e+06
11	12	31, 32, 33	Balestier, Toa Payoh, Serangoon	1.355540	103.876600	NaN	1.404034e+06
12	13	34, 35, 36, 37	Macpherson, Braddell, Potong Pasir, Bidadari	1.290410	103.852110	100.0	1.408717e+06
13	14	38, 39, 40, 41	Geylang, Eunos, Aljunied	1.313990	103.881970	23.0	1.302171e+06
14	15	42, 43, 44, 45	Katong, Joo Chiat, Amber Road	1.300876	103.901634	12.0	2.213112e+06
15	16	46, 47, 48	Bedok, Upper East Coast, Eastwood, Kew Drive	1.320397	103.950729	1.0	2.095990e+06
16	17	49, 50, 81	Loyang, Changi	1.373017	103.968394	NaN	1.056976e+06
17	18	51, 52	Simei, Tampines, Pasir Ris	1.371940	103.949940	2.0	1.117437e+06
18	19	53, 54, 55, 82	Serangoon Garden, Hougang, Punggol	1.364027	103.860205	8.0	1.144315e+06
19	20	56, 57	Bishan, Ang Mo Kio	1.364470	103.835060	5.0	1.605939e+06
20	21	58, 59	Upper Bukit Timah, Clementi Park, Ulu Pandan	1.328381	103.766423	3.0	1.543815e+06
21	22	60, 61, 62, 63, 64	Penjuru, Jurong, Pioneer, Tuas	1.320880	103.745320	19.0	NaN
22	23	65, 66, 67, 68	Hillview, Dairy Farm, Bukit Panjang, Choa Chu ...	1.365453	103.769995	2.0	1.291487e+06
23	24	69, 70, 71	Lim Chu Kang, Tengah	1.419670	103.702320	NaN	NaN
24	25	72, 73	Kranji, Woodgrove, Woodlands	1.429190	103.781140	1.0	NaN
25	26	77, 78	Upper Thomson, Springleaf	1.401268	103.817369	1.0	1.073759e+06
26	27	75, 76	Yishun, Sembawang, Senoko	1.447940	103.818910	2.0	1.038677e+06
27	28	79, 80	Seletar	1.410000	103.874170	2.0	1.162021e+06

Figure 1: DataFrame obtained after data wrangling.

With the coordinates retrieved through the Geocoder package, we can visualize the neighborhoods on a folium map.

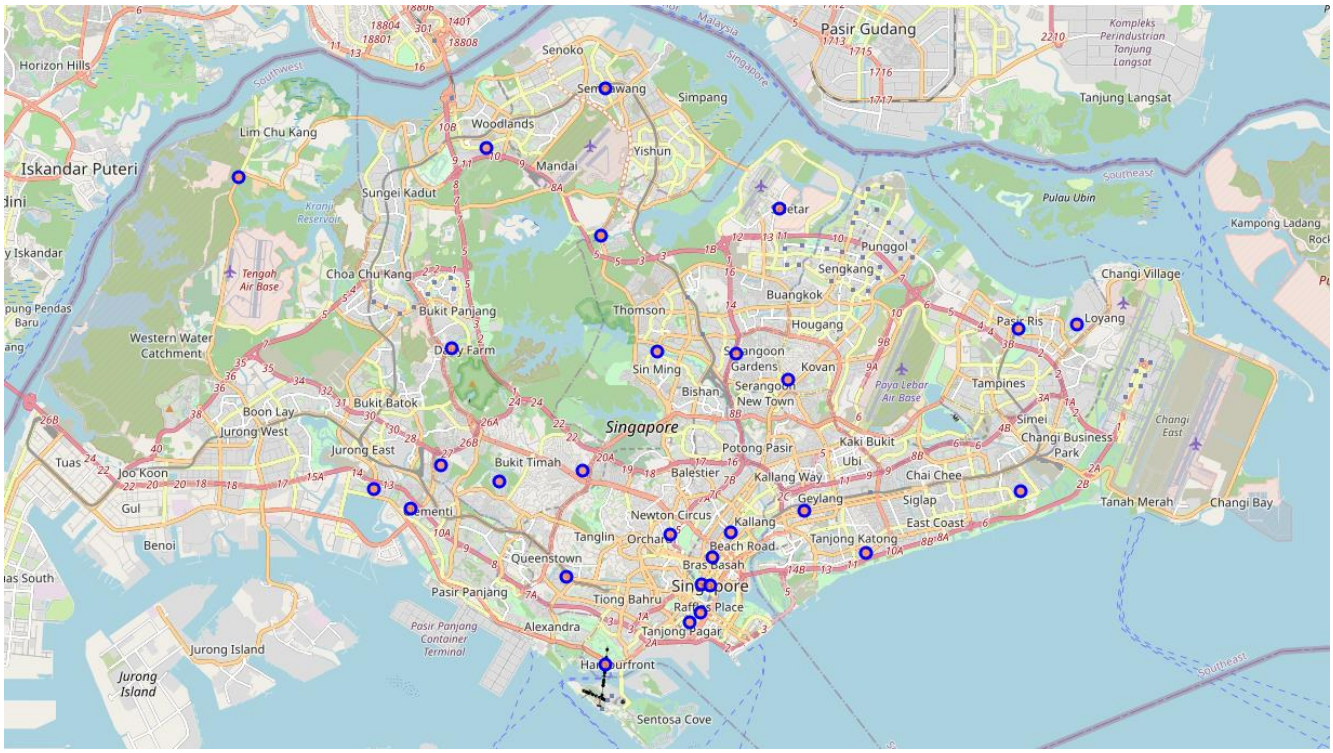

Figure 2: Singapore's neighborhoods on a folium map

The first thing to be noted is that most if the districts are located around the central and north south region.

Foursquare

Using Foursquare, we can now explore each district. With the Foursquare developer account and our credentials, we can make an API call to retrieve 100 venues in a radius of 1 km around each district. The data are returned in json file from which we can extract the venue name, category, latitude and longitude.

In the next step of the analysis, we will explore the Foursquare API and understand more about the districts in Singapore. Using the get request, the 100 venues of a district

in a 1 km radius from Foursquare were obtained. The get request will then return a list of recommended venues near the current location. Below shows some of the venues that are retrieved for a district.

	Postal district	Postal Code	General location	Latitude	Longitude	Private schools	Price ($)	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
0	1	01, 02, 03, 04, 05, 06	Raffles Place, Cecil, Marina, People's Park	1.281890	103.849120	NaN	3.648721e+06	2	Hotel	Coffee Shop	Japanese Restaurant	Cocktail Bar	Food Court	Café	Gym / Fitness Center	Bar	Korean Restaurant	Salad Place
1	2	07, 08	Anson, Tanjong Pagar	1.278890	103.845390	6.0	1.802246e+06	2	Coffee Shop	Japanese Restaurant	Hotel	Italian Restaurant	Bakery	Cocktail Bar	Korean Restaurant	Ramen Restaurant	Seafood Restaurant	Spanish Restaurant
2	3	14, 15, 16	Bukit Merah, Queenstown, Tiong Bahru	1.293048	103.806248	24.0	1.547536e+06	0	Chinese Restaurant	Café	Food Court	Coffee Shop	Supermarket	Indian Restaurant	Asian Restaurant	Park	Pizza Place	Noodle House
3	4	09, 10	Telok Blangah, Harbourfront	1.265331	103.818861	6.0	1.627808e+06	0	Japanese Restaurant	Fast Food Restaurant	Chinese Restaurant	Clothing Store	Toy / Game Store	Spa	Multiplex	Coffee Shop	Asian Restaurant	Shopping Mall
4	5	11, 12, 13	Pasir Panjang, Hong Leong Garden, Clementi New...	1.314730	103.756799	2.0	1.299735e+06	0	Food Court	Japanese Restaurant	Bus Station	Indian Restaurant	Dessert Shop	Asian Restaurant	Coffee Shop	Noodle House	Fast Food Restaurant	Shopping Mall

Figure 3: Top venues per district

The data are then grouped by district and show the top 10 most common venues per district.

Data modeling

We perform a machine learning algorithm, K-means clustering, in order to identify k number of centroids and allocates every data point to the nearest cluster, while keeping the centroids as small as possible. Without any labeled data available, using this unsupervised learning model is the most suited to this project.

The districts will be clustered into 4 clusters. Then we will be able to identify which cluster has most infrastructures, how many private schools and what the average sale price in each cluster is. With this information we will be able to identify the best neighborhoods/districts to explore or move into.

Results

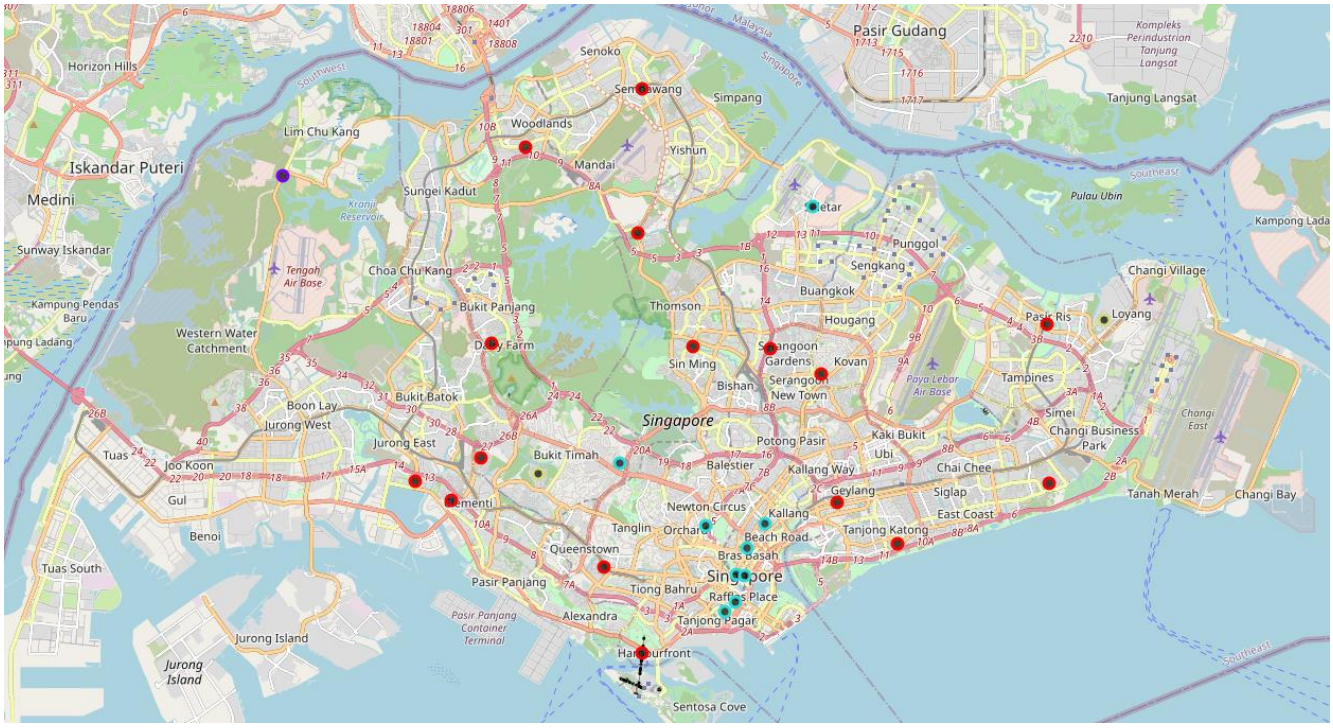

Figure 4: Singapore's neighborhoods after k = 4 K-means clustering

The table below show the number of districts for each cluster.

Cluster	Number of districts
0	16
2	9
3	2
1	1

Table 1: Number of districts by cluster

- Cluster 0 is the densest. It includes 16 of the 28 districts in Singapore. It's composed of very attractive venues such as food courts, supermarkets, hotels, coffee shop, dessert shop, parks, gyms, trails, shopping malls, restaurants (Asian, Chinese, Japanese), bakeries, etc. Cluster0 presents a lot a different type of venue seems to be a very attractive area.
- Cluster 2 is composed of 9 districts. It also presents attractive venues. One of the most common venues is hotels, boutiques, playground, waterfront, the airport, etc.
- Cluster 3 is composed of 2 districts. Its characterized by bus stations, government building, shopping malls, bar, restaurants, bookstore, etc.
- Cluster 1 is postal district 24

Cluster 0 and 2 are the most interested ones. They have respectively 16 and 9 districts. In the most common venue we can find restaurants, coffee shops, hotels, bus stations, trails, golf courses, supermarkets, cosmetic shops, bakeries, cocktail bars, the airport, etc. These clusters are very dense and have many different types of venues that are very attractive.

Number of private schools

The number of private schools by postal district is represented below on a bar graph. We can clearly observe that most of them are in one postal district with almost 100 private schools in this area.

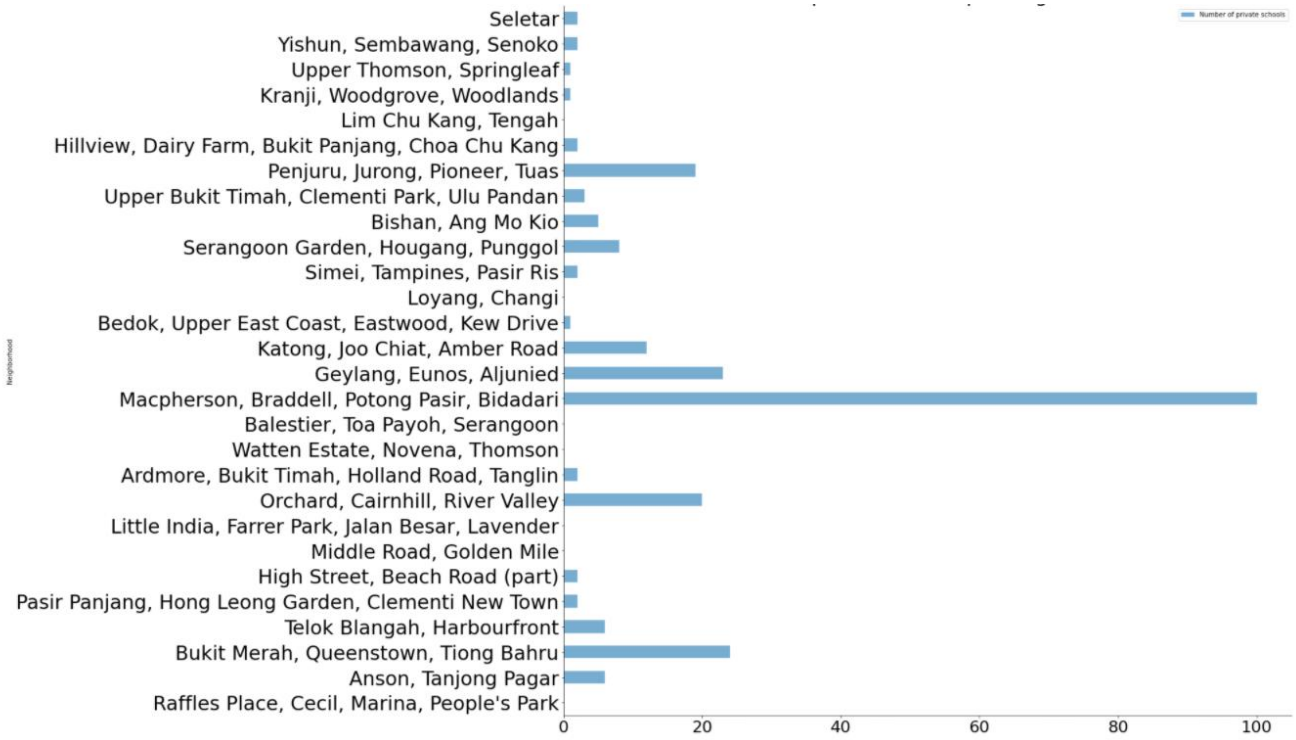

Figure 5: Number of private schools by postal districts

The graph below shows the number of private schools per cluster.

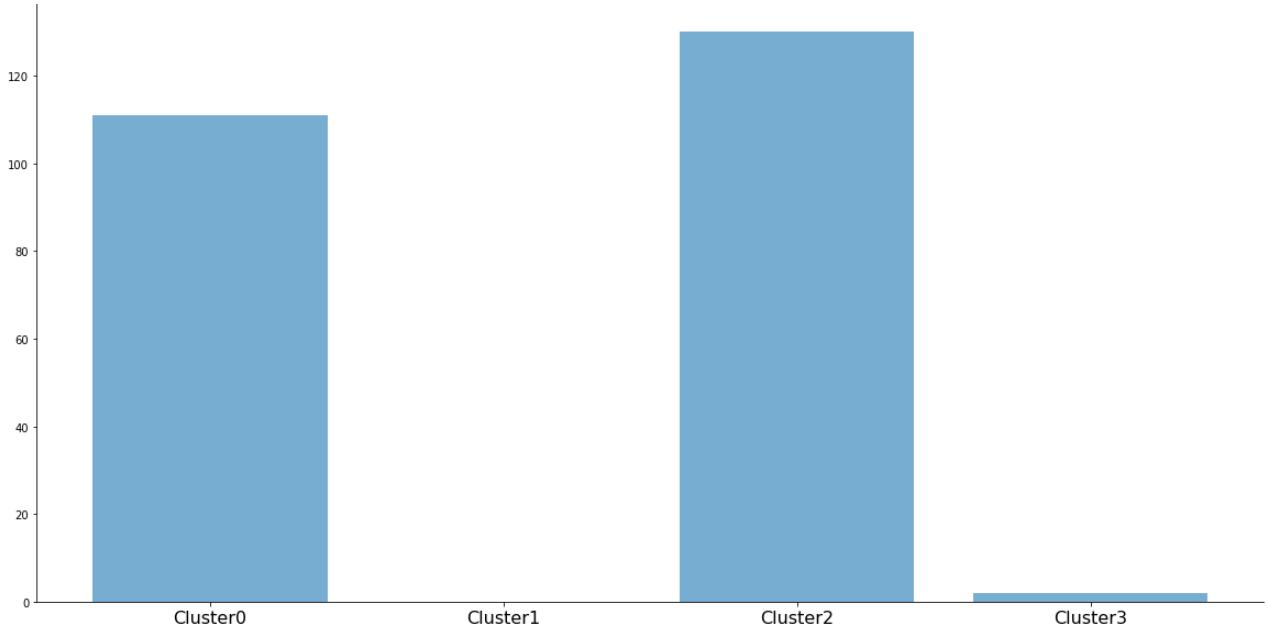

Figure 6: Number of private schools per cluster

Cluster 0 and 2 have most of the private schools, and as shown before also have most of the interested common venues. Cluster has no private school.

Average housing sales prices

The average housing sales prices by postal districts are represented below on a bar graph. The prices seem to be in the same order of magnitude except for 2 districts where the prices are clearly higher.

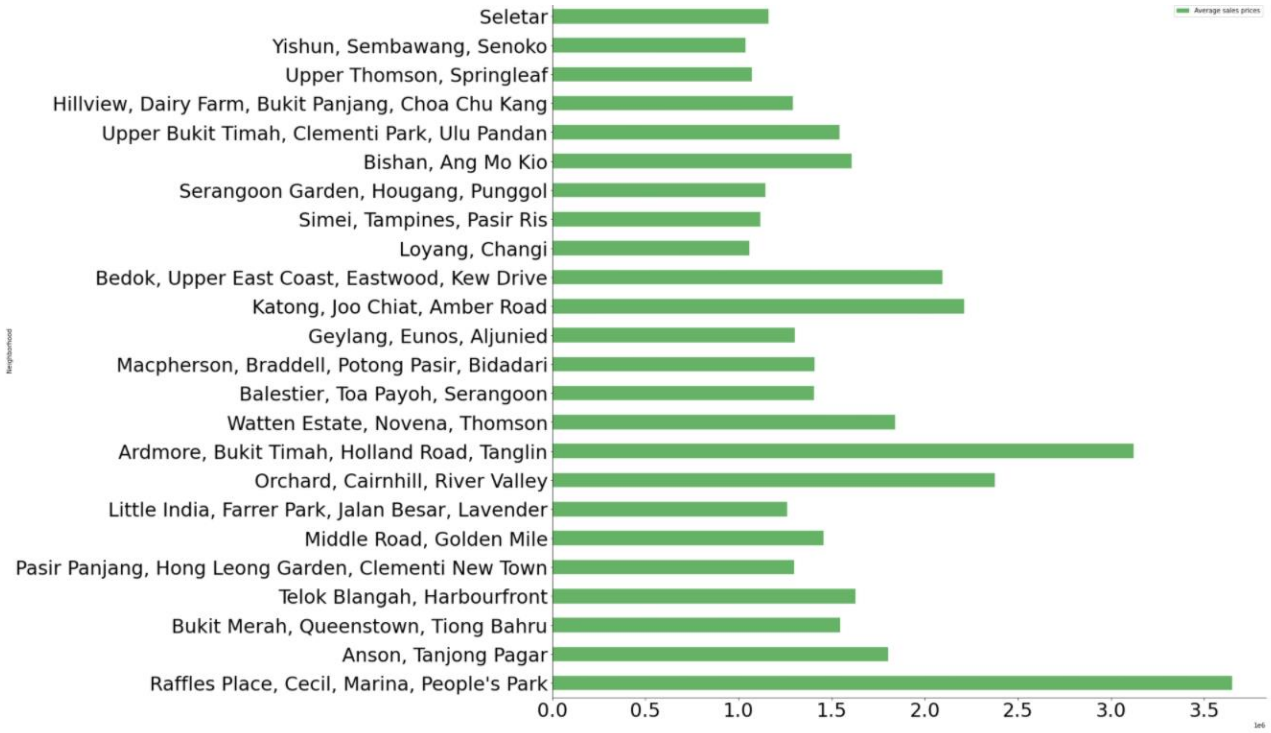

Figure 7: Average housing sales prices by postal districts

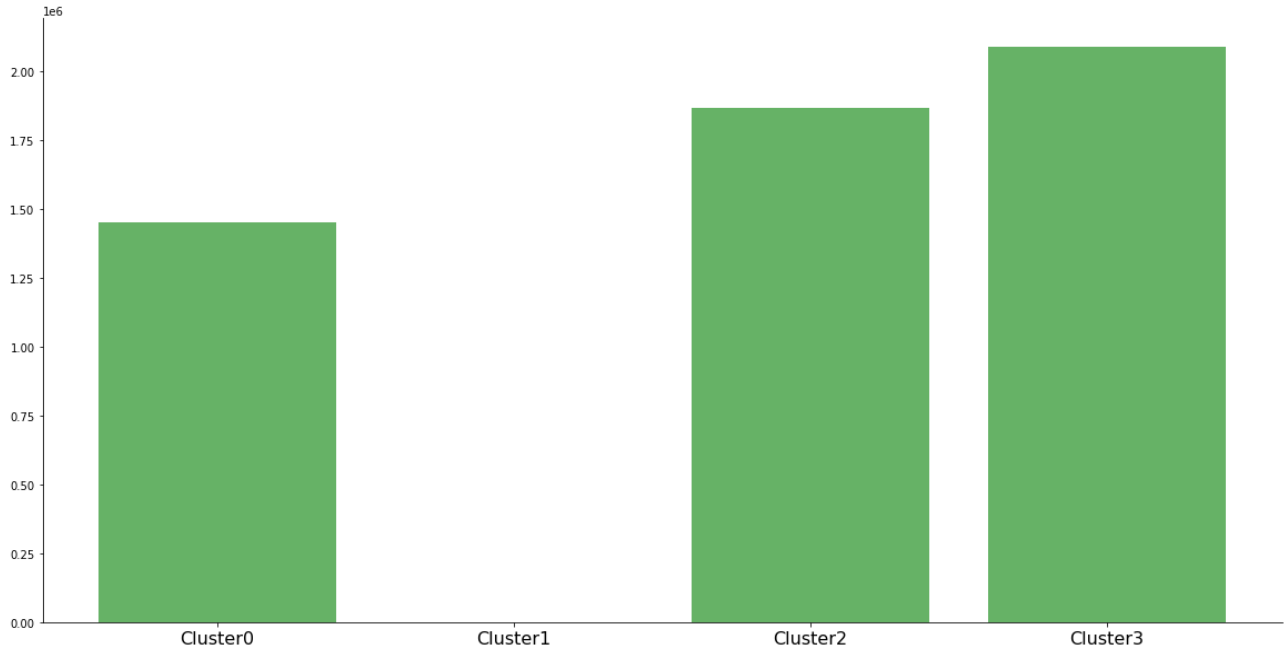

Figure 8: Average housing sales prices by cluster

Cluster 0 and 2 have more average prices then Cluster 3. Cluster 3 is far smaller than Clusters 0 and 2 but the average sale price is higher.

Discussion and Recommendations

Based on the clustering model and on the different data studied. It seems that clusters 0 and 2 are the most attractive one to explore or move into. With most of the private schools and average sales prices lower than cluster 3, they seem to be the clusters that have the most to offer. Nevertheless Cluster 2 presents most of the hotels and the airport, it might be a more touristic area. Cluster 0 might be the most appropriate to move into.

Conclusions

In this we have gone through the process of identifying a business problem, gathered the required data, extracting and preparing it. Visualizing the result and performing a machine learning algorithm by clustering the data into 4 clusters in order to reach a solution to the business problem.

This project also provides recommendations to visitors, potential immigrants and investors. However, there are clear limitations to this project and all the assumptions made. Notably because of the data gathered. Foursquare is a great source of information for general venues. There are other possible factors to account for like, income of the residents residing, rental rates, essential infrastructures (Pharmacy, Bank, Police station, etc.) and many more. Lastly, the assumption was made by taking the 100 venues within a 1 km radius only.

References

1. https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population_density
2. https://www.singstat.gov.sg/modules/infographics/population
3. https://en.wikipedia.org/wiki/Postal_codes_in_Singapore
4. https://www.ura.gov.sg/realEstateIIWeb/transaction/search.action
5. https://data.gov.sg/dataset/private-education-institutions
6. https://developer.foursquare.com