Clustering in Singapore based nearby venues and supplemented with median home prices

Glen Teoh

June 13, 2020

Introduction

Issue and background

Food is a popular past time, and also business in Singapore. Business owners, especially new ones, often have a hard time knowing where to launch their restaurant. Ideally, they would want to come up with **some areas that may be good to enter the market** and potentially extend in. Although the shopping and business district are obvious starting locations, the rents and availability of space there is also hard to come by. Therefore, there is a need to look at residential areas for potential business since rents are lower, but number of people is not necessarily lower. As such, there is a lot of attention focused on the shopping and working districts but not much on the 'heartlands' or residential areas. Ideally, we would like to know which areas have **many other attractions nearby, with food being a priority, and the property prices of those locations**, since property prices are actually a reliable indicator of the affluency of the local population.

Data Acquisition

Since we are interested in nearby venues, the Foursquare API comes to mind as a good data source to request venues based on a list of latitudes and longitude coordinates. This would help us to identify the most common venues in an area, and give us an idea of the people who might visit there.

In order to obtain these coordinates, we are also going to use the Geocoder library in python which makes further API calls to Open Street Map in order to get geocodes for the names of the neighbourhoods.

Lastly, we will need a list of the towns that are in Singapore, along with the median property price of each of those towns. This data can be easily obtained from Gov.sg, which gives us many different statistics as it pertains to Singapore.

Feature Selection

The Foursquare API, aside from nearby venues, is also able to provide us tips, photos, coordinates etc. of the venues. However, we only really require the *category* of the venues, for example: restaurant, club, bar, gym etc. Although reviews and photos could give us an idea of the popularity and thus footfall of the area, the amount of work that would have to go into that would not be meaningful for the value we will be able to obtain, which is why all of those were excluded in this study.

In terms of home property prices, I opted for 4-room flats under HDB (the housing development board of Singapore, which manages public housing), since 78.7% of the population lives in these apartments, and 4 room flat types are also the most common type of apartment. The prices for these flats differ by town, and that is really what we need – a determinant of household prices weighted against other towns. Therefore, the *absolute* prices aren't very important, more their

relative price to other towns with 4-room flats. As such, other housing types, such as condominiums, and 2 or 3 room flats, were excluded.

Methodology

Since we would like to find neighbourhoods that contain certain type of venues along with some median home prices as supplementary information, cluster analysis is the choice of method for grouping towns based on nearby venues. I opted to use K-means clustering as it is a very common and popular method of clustering. The median home prices would not be considered in the k-means clustering, but rather used as additional information to be visualized on top of the map.

Data Cleaning & Transformation

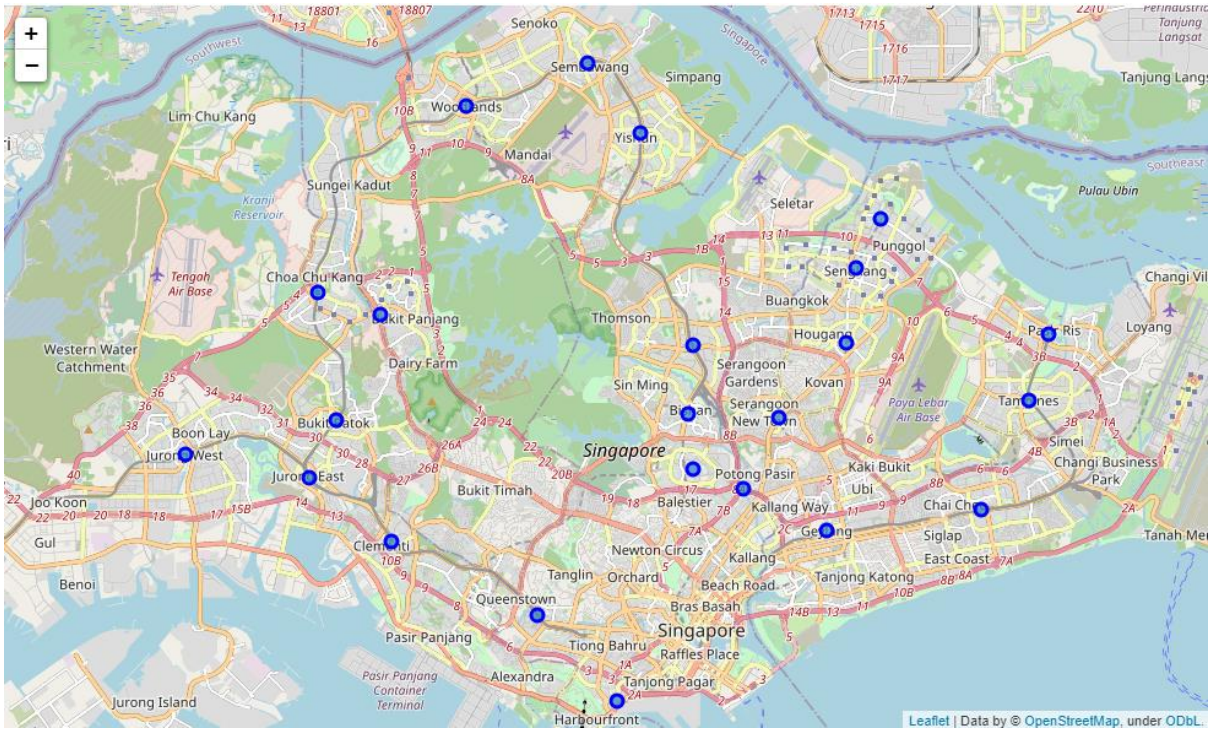
Home resale prices were obtained from gov.sg, but contained information stretching back more than 10 years, split by quarter. The data also included home prices for several other types of apartments. I removed the unnecessary data, and only used the latest (2020 Q1) median prices for 4 room flats:

	quarter	town	flat_type	price
0	2020-Q1	ANG MO KIO	4-ROOM	397500
1	2020-Q1	BEDOK	4-ROOM	387300
2	2020-Q1	BISHAN	4-ROOM	521500
3	2020-Q1	BUKIT BATOK	4-ROOM	350000
4	2020-Q1	BUKIT MERAH	4-ROOM	640000
5	2020-Q1	BUKIT PANJANG	4-ROOM	411000
6	2020-Q1	CHOA CHU KANG	4-ROOM	339000
7	2020-Q1	CLEMENTI	4-ROOM	644000
8	2020-Q1	GEYLANG	4-ROOM	450000
9	2020-Q1	HOUGANG	4-ROOM	400000
10	2020-Q1	JURONG EAST	4-ROOM	375000
11	2020-Q1	JURONG WEST	4-ROOM	360000
12	2020-Q1	KALLANG/WHAMPOA	4-ROOM	500000
13	2020-Q1	PASIR RIS	4-ROOM	435500
14	2020-Q1	PUNGGOL	4-ROOM	460000
15	2020-Q1	QUEENSTOWN	4-ROOM	728000
16	2020-Q1	SEMBAWANG	4-ROOM	351000
17	2020-Q1	SENGKANG	4-ROOM	425000
18	2020-Q1	SERANGOON	4-ROOM	429000
19	2020-Q1	TAMPINES	4-ROOM	422500
20	2020-Q1	TOA PAYOH	4-ROOM	539000
21	2020-Q1	WOODLANDS	4-ROOM	340000
22	2020-Q1	YISHUN	4-ROOM	360000

Using the geocoder API, I then queried and obtained a list of geocodes for the respective towns, which I then added into the table:

	quarter	town	flat_type	price	Latitude	Longitude
0	2020-Q1	ANG MO KIO	4-ROOM	397500	1.370073	103.849516
1	2020-Q1	BEDOK	4-ROOM	387300	1.323976	103.930216
2	2020-Q1	BISHAN	4-ROOM	521500	1.350986	103.848255
3	2020-Q1	BUKIT BATOK	4-ROOM	350000	1.349057	103.749591
4	2020-Q1	BUKIT MERAH	4-ROOM	640000	1.270439	103.828318
5	2020-Q1	BUKIT PANJANG	4-ROOM	411000	1.378629	103.762136
6	2020-Q1	CHOA CHU KANG	4-ROOM	339000	1.384749	103.744534
7	2020-Q1	CLEMENTI	4-ROOM	644000	1.315100	103.765231
8	2020-Q1	GEYLANG	4-ROOM	450000	1.318186	103.887056
9	2020-Q1	HOUGANG	4-ROOM	400000	1.370682	103.892545
10	2020-Q1	JURONG EAST	4-ROOM	375000	1.333115	103.742297
11	2020-Q1	JURONG WEST	4-ROOM	360000	1.339636	103.707339
12	2020-Q1	KALLANG/WHAMPOA	4-ROOM	500000	1.329750	103.863840
13	2020-Q1	PASIR RIS	4-ROOM	435500	1.373031	103.949255
14	2020-Q1	PUNGGOL	4-ROOM	460000	1.405258	103.902330
15	2020-Q1	QUEENSTOWN	4-ROOM	728000	1.294623	103.806045
16	2020-Q1	SEMBAWANG	4-ROOM	351000	1.449093	103.820055
17	2020-Q1	SENGKANG	4-ROOM	425000	1.391654	103.895364
18	2020-Q1	SERANGOON	4-ROOM	429000	1.349862	103.873729
19	2020-Q1	TAMPINES	4-ROOM	422500	1.354653	103.943571
20	2020-Q1	TOA PAYOH	4-ROOM	539000	1.335391	103.849741
21	2020-Q1	WOODLANDS	4-ROOM	340000	1.436897	103.786216
22	2020-Q1	YISHUN	4-ROOM	360000	1.429384	103.835028

Using the Folium library, I placed these coordinates on the map to ensure that they were correct:



Now, with household prices already in hand, the next thing I needed to do was to get the nearby venues for each of the neighbourhoods. In order to reduce the data sizes (and also because of the real-world limitations of using a free account on IBM Watson and Foursquare), I limited the radius for the search to 1 kilometre and placed a limit of 100 venues before querying the Foursquare API.

	Town	Town Latitude	Town Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ANG MO KIO	1.370073	103.849516	FairPrice Xtra	1.369279	103.848886	Supermarket
1	ANG MO KIO	1.370073	103.849516	Old Chang Kee	1.369094	103.848389	Snack Place
2	ANG MO KIO	1.370073	103.849516	Face Ban Mian 非板面 (Ang Mo Kio)	1.372031	103.847504	Noodle House
3	ANG MO KIO	1.370073	103.849516	NTUC FairPrice	1.371507	103.847082	Supermarket
4	ANG MO KIO	1.370073	103.849516	MOS Burger	1.369170	103.847831	Burger Joint

Armed with a list of venues for each of those towns, I had to convert the venue categories into a numerical type in order for the clustering to work:

	Town	ATM	Accessories Store	American Restaurant	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	...
0	ANG MO KIO	0	0	0	0	0	0	0	0	0	...
1	ANG MO KIO	0	0	0	0	0	0	0	0	0	...
2	ANG MO KIO	0	0	0	0	0	0	0	0	0	...
3	ANG MO KIO	0	0	0	0	0	0	0	0	0	...
4	ANG MO KIO	0	0	0	0	0	0	0	0	0	...

As you can see, each category became one column, and if there was 1 instance of the category in the town, the value would be 1; 2 if there were 2 and so on.

Since I wanted to create only 1 record for each town, I grouped the table above by towns, and calculated the mean of the frequency of occurrence for each category. This gives us an idea of how likely a type of category appears in a town, and also takes into account the number of venues returned for that town, so that they can be compared against each other. This gives us a table like the below:

	Town	ATM	Accessories Store	American Restaurant	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	...
0	ANG MO KIO	0.000000	0.000000	0.011364	0.00	0.000000	0.000000	0.034091	0.000000	0.000000	...
1	BEDOK	0.000000	0.010417	0.010417	0.00	0.000000	0.000000	0.031250	0.000000	0.000000	...
2	BISHAN	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.042254	0.000000	0.000000	...
3	BUKIT BATOK	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	...
4	BUKIT MERAH	0.000000	0.000000	0.016129	0.00	0.016129	0.000000	0.048387	0.016129	0.000000	...
5	BUKIT PANJANG	0.000000	0.000000	0.019231	0.00	0.000000	0.000000	0.057692	0.000000	0.000000	...

As there are lots and lots of different types of categories, in order to reduce dimensionality, I decided to use only the top 5 most common venues. This actually gave me an interesting insight:

	Town	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	ANG MO KIO	Food Court	Coffee Shop	Asian Restaurant	Japanese Restaurant	Chinese Restaurant
1	BEDOK	Coffee Shop	Chinese Restaurant	Food Court	Café	Supermarket
2	BISHAN	Food Court	Coffee Shop	Chinese Restaurant	Seafood Restaurant	Japanese Restaurant
3	BUKIT BATOK	Food Court	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Malay Restaurant
4	BUKIT MERAH	Coffee Shop	Asian Restaurant	Clothing Store	Food Court	Bus Stop

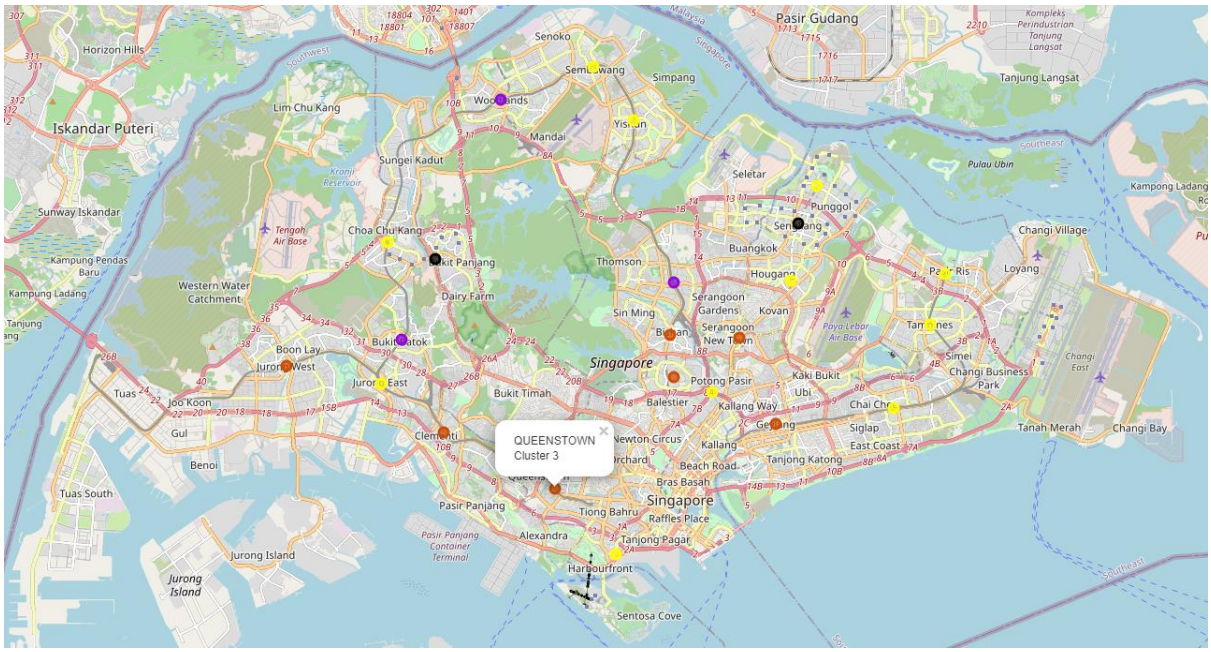
Being a nation that loves its food, I was actually confirming my own suspicions that the top 5 most common venues for many of the neighbourhoods were actually related to food. Here I began to have some suspicions that the clustering would end up grouping towns by the type of food establishments in the towns.

Clustering

With the above data, now we are ready to do clustering analysis, based on the top 5 most common venues across the towns. Running the K-means clustering method with sci-kit learn with K=4, 4 clusters were created based on the data given. Please note that clustering was **not** done with the median home prices – that is supplementary data that we are using to enhance our insights.

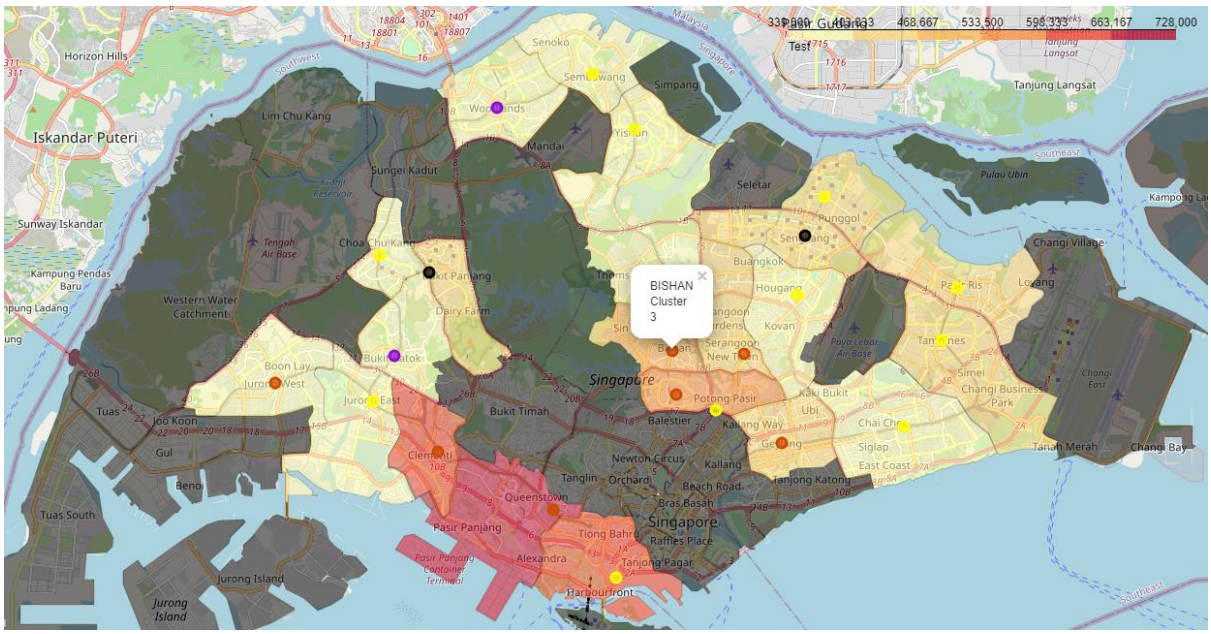
	town	price	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	ANG MO KIO	397500	1.370073	103.849516	2	Food Court	Coffee Shop	Asian Restaurant	Japanese Restaurant	Chinese Restaurant
1	BEDOK	387300	1.323976	103.930216	0	Coffee Shop	Chinese Restaurant	Food Court	Café	Supermarket
2	BISHAN	521500	1.350986	103.848255	3	Food Court	Coffee Shop	Chinese Restaurant	Seafood Restaurant	Japanese Restaurant
3	BUKIT BATOK	350000	1.349057	103.749591	2	Food Court	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Malay Restaurant
4	BUKIT MERAH	640000	1.270439	103.828318	0	Coffee Shop	Asian Restaurant	Clothing Store	Food Court	Bus Stop

Let's take a look at the visualisation of the clusters on the map:



Since these clusters are generated by the nearby venues, as somebody who lives in Singapore I can already start to see the clusters making sense - many of the dark orange clusters (cluster 3) are known as "mature estates", meaning that they have been around a long time, with lots of development and infrastructure already in place. Meanwhile, you can also see that many of the yellow clusters are located in areas that locally we know are pretty difficult to get to (as many of them are near the edges of the country). Let's take a look at whether the property prices can give us more insights.

Adding median home price data to the map



You'll notice that there are many 'blacked out' areas in the map - these are either industrial areas, or the central business and shopping districts where there are no public housing. Of

course, private housing is available there, but since we are using public housing data for this study, those were not included in the analysis.

We can see that many of the orange clusters (cluster 3) also correspond to higher housing prices, except for one town in the west (Jurong West), which is something of an outlier here.

Breakdown of clusters

Cluster 0

	price	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	387300	Coffee Shop	Chinese Restaurant	Food Court	Café	Supermarket
4	640000	Coffee Shop	Asian Restaurant	Clothing Store	Food Court	Bus Stop
6	339000	Coffee Shop	Fast Food Restaurant	Food Court	Asian Restaurant	Gym
9	400000	Coffee Shop	Food Court	Chinese Restaurant	Fast Food Restaurant	Asian Restaurant
10	375000	Food Court	Coffee Shop	Japanese Restaurant	Chinese Restaurant	Café
12	500000	Coffee Shop	Chinese Restaurant	Convenience Store	Food Court	Bus Line
13	435500	Coffee Shop	Fast Food Restaurant	Food Court	Park	Supermarket
14	460000	Café	Fast Food Restaurant	Supermarket	Electronics Store	Japanese Restaurant
16	351000	Coffee Shop	Chinese Restaurant	Fast Food Restaurant	Italian Restaurant	Japanese Restaurant
19	422500	Coffee Shop	Café	Bakery	Supermarket	Bubble Tea Shop
22	360000	Coffee Shop	Food Court	Fast Food Restaurant	Chinese Restaurant	Asian Restaurant

Median home price of cluster: 424572

Cluster 1

	price	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	411000	Fast Food Restaurant	Café	Coffee Shop	Asian Restaurant	Supermarket
17	425000	Bus Station	Food Court	Fast Food Restaurant	Coffee Shop	Supermarket

Median home price of cluster: 418000

Cluster 2

	price	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	397500	Food Court	Coffee Shop	Asian Restaurant	Japanese Restaurant	Chinese Restaurant
3	350000	Food Court	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Malay Restaurant
21	340000	Food Court	Coffee Shop	Fast Food Restaurant	Café	Asian Restaurant

Median home price of cluster: 362500

Cluster 3

	price	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	521500	Food Court	Coffee Shop	Chinese Restaurant	Seafood Restaurant	Japanese Restaurant
7	644000	Food Court	Chinese Restaurant	Indian Restaurant	Asian Restaurant	Supermarket
8	450000	Chinese Restaurant	Asian Restaurant	Food Court	Vegetarian / Vegan Restaurant	Noodle House
11	360000	Japanese Restaurant	Asian Restaurant	Fast Food Restaurant	Chinese Restaurant	Coffee Shop
15	728000	Chinese Restaurant	Coffee Shop	Food Court	Café	Noodle House
18	429000	Asian Restaurant	Coffee Shop	Chinese Restaurant	Café	Bus Station
20	539000	Chinese Restaurant	Noodle House	Food Court	Asian Restaurant	Coffee Shop

Median home price of cluster: 524500

Based on the clusters above, we can start to find some meaningful insights, and even come up with some potential categories for the clusters:

- Cluster 0: Middle-High Property Price, Coffee shops & Restaurants
- Cluster 1: Medium Property Price, Outliers and less developed areas
- Cluster 2: Lower Property Price, Food courts and Coffee Shops
- Cluster 3: High Property Price, Restaurants

Based on these clusters, and also with my experience as a local, it looks like the above study was a good starting point for business owners to consider where they would like to start a "heartland" business targeting locals!

Conclusion

There are several limitations to this study, some financial (I can't do much with free accounts as I am already approaching the IBM Watson CPU runtime limits as I type this), some related to sample size (Singapore is a small country with not many towns), and foursquare also being one of them since it's not a very popular app in Singapore and therefore not as updated as other mapping providers such as Google. However, I believe the analysis gives a good overview of the 'type' of neighbourhood that business owners can make use of to consider where to launch their businesses. Given more time and resources, it would be interesting to make use of both private and public housing prices, and to dive down into further detail, such as clustering along individual or even blocks. This would require many more results from the Foursquare API, and a block-level pricing data which is already obtainable from Gov.sg. However, due to resource constraints, I am currently unable to proceed with such a study. I had fun doing this and I hope you had a good time reading through this too.