Capstone Project – The Battle of Neighbourhoods

1. Introduction

1.1 Background

Singapore is a country with a very diverse culture. It has been recognized as a city with "best investment potential" and world's most competitive economy, as per Wikipedia, and is a home to 5.6 million residents. Singaporeans love to eat. Whenever a new food place would open up, a queue is expected that could last for hours.

1.2 Problem

A lot of businesses, especially food places, open up regularly. Unfortunately, some of them cannot sustain the business after the initial hype has mellowed down. It would be interesting to get a feel of the trending markets and how saturated the market is for a particular neighbourhood, in Singapore's case, a particular sub-zone. The aim of this project is identify new market opportunities/ potential businesses in Singapore using Four Square location data.

1.3 Interest/Target Audience

This project would utilize Foursquare data to identify the popular venues in Singapore, what businesses are already available and how saturated the market is. The target audience would be the potential investors.

2. Data

1.1 Data Source and description

I will be using Singapore boundary data to identify the regions and sub-zones in Singapore. The source of the data is from https://data.gov.sg/ I have downloaded the data for the master plan subzone boundary for 2014 to get the divisions of the zones. The files come in kml and shp formats; so will need to be converted first before I can use them in my analysis. I used https://mygeodata.cloud/converter/kml-to-csv and converted them to a CSV file. Below is the snapshot of the CSV files:

	X	Y	gid	Name	description	X_ADDR	SHAPE_Leng	FMEL_UPD_I	Region Name	Planning Are	Region Code	Subzone Nar	Central Area	SHAPE	Y_ADDR	Subzone Cod	INC_CRC	Subzone Nu
2	103.822322	1.28651192	4	HENDERSON HILL		26782.8261	3313.6249	##########	CENTRAL REGION	BUKIT MERA	CR	HENDERSON	N	Polygon	29933.7651	BMSZ08	3775D82C5D	
3	103.81718	1.28773673	5	REDHILL		26201.958	2825.5941	##########	CENTRAL REGION	BUKIT MERA	CR	REDHILL	N	Polygon	30005.7035	BMSZ03	85D9ABEF0A	
4	103.809556	1.28768285	6	ALEXANDRA HILL		25358.8208	4428.91344	##########	CENTRAL REGION	BUKIT MERA	CR	ALEXANDRA	N	Polygon	29991.3842	BMSZ07	9D286521EF	
5	103.844945	1.28975077	8	CLARKE QUAY		29253.2073	2208.6193	##########	CENTRAL REGION	SINGAPORE I	CR	CLARKE QUA	Y	Polygon	30222.8645	SRSZ02	48661DC0FB	
6	103.798388	1.28860707	10	QUEENSWAY		24168.3106	3454.23928	##########	CENTRAL REGION	QUEENSTOW	CR	QUEENSWAY	N	Polygon	30104.1768	QTSZ07	3580D2AFFB	
7	103.819522	1.29249624	12	ALEXANDRA NORTH		26548.2507	2907.05146	##########	CENTRAL REGION	BUKIT MERA	CR	ALEXANDRA	N	Polygon	30519.3922	BMSZ06	4DC4BF8D86	
8	103.843342	1.2741159	22	ANSON		29145.35	1586.10006	##########	CENTRAL REGION	DOWNTOWN	CR	ANSON	Y	Polygon	28466.7783	DTSZ10	4893F02202	1
9	103.837635	1.29356851	14	INSTITUTION HILL		28465.3999	2842.52551	##########	CENTRAL REGION	RIVER VALLE	CR	INSTITUTION	Y	Polygon	30711.2179	RVSZ05	C3C22D1EE3	
10	103.670222	1.25997756	16	JURONG ISLAND AND BUK		13012.8805	68083.9365	##########	WEST REGION	WESTERN ISL	WR	JURONG ISLA	N	Polygon	27225.8674	WISZ01	699F7210FBI	
11	103.71921	1.20801575	17	SUDONG		15931.7586	24759.0658	##########	WEST REGION	WESTERN ISL	WR	SUDONG	N	Polygon	19579.069	WISZ03	F718C723E0	
12	103.772538	1.20284639	18	SEMAKAU		21206.3338	18703.6811	##########	WEST REGION	WESTERN ISL	WR	SEMAKAU	N	Polygon	20465.8075	WISZ02	E69207D4F7	
13	103.851586	1.24158367	19	SOUTHERN GROUP		29815.09	25626.9772	##########	CENTRAL REGION	SOUTHERN I	CR	SOUTHERN G	N	Polygon	23412.5866	SISZ02	5809FC5472	
14	103.834323	1.24892189	20	SENTOSA		27593.9393	17496.1937	##########	CENTRAL REGION	SOUTHERN I	CR	SENTOSA	N	Polygon	25813.3546	SISZ01	A6FCDC9C44	
15	103.859296	1.27135716	23	STRAITS VIEW		30832.9017	5277.76083	##########	CENTRAL REGION	STRAITS VIEV	CR	STRAITS VIEV	Y	Polygon	28194.0843	SVSZ01	21451799CA	
16	103.812406	1.26848725	24	MARITIME SQUARE		25805.7867	13737.1162	##########	CENTRAL REGION	BUKIT MERA	CR	MARITIME S	N	Polygon	27911.4193	BMSZ01	C1AC31ABF9	
17	103.81935	1.28198445	35	BUKIT MERAH		26360.799	3074.96324	##########	CENTRAL REGION	BUKIT MERA	CR	BUKIT MERA	N	Polygon	29384.1429	BMSZ02	1C51019439	
18	103.848562	1.27849651	27	CECIL		29730.1983	2116.09467	##########	CENTRAL REGION	DOWNTOWN	CR	CECIL	Y	Polygon	29011.3284	DTSZ08	65AA82AF6F	
19	103.82721	1.27840723	28	KAMPONG TIONG BAHRU		27335.8305	3304.06047	##########	CENTRAL REGION	BUKIT MERA	CR	KAMPONG T	N	Polygon	28979.7731	BMSZ11	CF28DD4E78	1
20	103.85732	1.34075605	159	LORONG 8 TOA PAYOH		30646.8429	3271.27119	##########	CENTRAL REGION	TOA PAYOH	CR	LORONG 8 T	N	Polygon	35893.1126	TPSZ05	4A20B4AB7C	
21	103.795491	1.27753463	30	PASIR PANJANG 2		23692.6385	5797.23132	##########	CENTRAL REGION	QUEENSTOW	CR	PASIR PANJA	N	Polygon	28894.4099	QTSZ14	C6E665749C	1

Y and X columns give the latitude and longitudes respectively. The sub-zones fall under 5 regions, namely: Central Region, West Region, North-East Region, East Region and North Region. There is also a field called Central Area Indicator, which indicates whether the sub-zone can be found in the central area. This would be useful if we are only interested in the sub-zones that belong to the central districts.

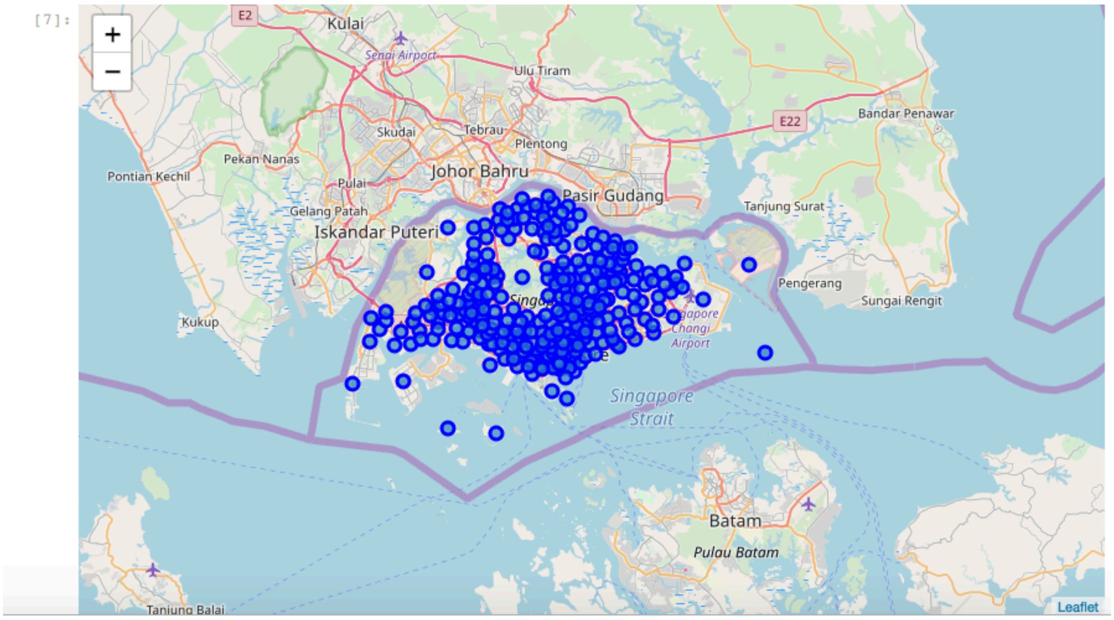
Given the location of each sub-zone, we can now do queries using Foursquare API to get the businesses or venues for each sub-zone and determine which is the most popular type of establishments.

3. Methodology

The raw data has been uploaded to Github repository. I removed any unnecessary columns from the dataset. Since I only needed the region names and the sub-zones, together with the Latitude and Longitude and the Central Area Indicator, I retained those columns. I also renamed the columns to reflect the content such as X to longitude, Y to latitude and Name to Subzone. The first five rows of the final data are shown below:

	Longitude	Latitude	Subzone	Region Name	Central Area Indicator
0	103.822322	1.286512	HENDERSON HILL	CENTRAL REGION	N
1	103.817180	1.287737	REDHILL	CENTRAL REGION	N
2	103.809556	1.287683	ALEXANDRA HILL	CENTRAL REGION	N
3	103.844945	1.289751	CLARKE QUAY	CENTRAL REGION	Y
4	103.798388	1.288607	QUEENSWAY	CENTRAL REGION	N

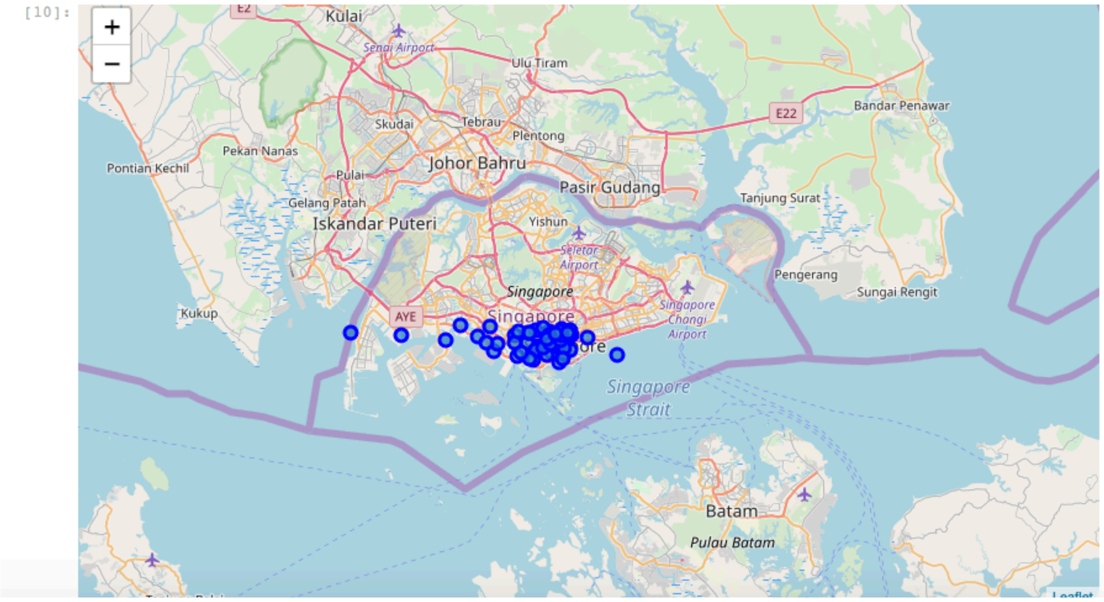
For the initial analysis of all the regions in Singapore, I used Folium to create a map of Singapore with the sub-zones superimposed and labeled. The geopy library was used to get the Latitude and Longitude of Singapore.



For this particular study, I focused on the sub-zones from the central area, that is those sub-zones with Central Area indicator = “Y”. I created a new dataframe containing only those sub-zones. Below are the first five rows of the new dataframe:

	Longitude	Latitude	Subzone	Region Name	Central Area Indicator
0	103.844945	1.289751	CLARKE QUAY	CENTRAL REGION	Y
1	103.843342	1.274116	ANSON	CENTRAL REGION	Y
2	103.837635	1.293569	INSTITUTION HILL	CENTRAL REGION	Y
3	103.859296	1.271357	STRAITS VIEW	CENTRAL REGION	Y
4	103.848562	1.278497	CECIL	CENTRAL REGION	Y

Using Folium again to visualize the Central Area superimposed in Singapore Map provided the following map:



I utilized the Foursquare API to explore the Central Area and all the associated sub-zones. I opted to extract 100 venues within a 500 meter radius since Singapore is not a big city. First five rows of venues returned as per below:

	Subzone	Subzone Latitude	Subzone Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	CLARKE QUAY	1.289751	103.844945	Clarke Quay Riverside	1.289661	103.846464	Waterfront
1	CLARKE QUAY	1.289751	103.844945	Swissôtel Merchant Court	1.288434	103.845674	Hotel
2	CLARKE QUAY	1.289751	103.844945	Nirai Kanai Okinawan Restaurant	1.291493	103.845279	Japanese Restaurant
3	CLARKE QUAY	1.289751	103.844945	Zouk	1.290995	103.845947	Nightclub
4	CLARKE QUAY	1.289751	103.844945	Jumbo Seafood Restaurant	1.288983	103.844812	Seafood Restaurant

To find out how many venues were returned for each subzone, I used the groupby() function:

Subzone	Subzone Latitude	Subzone Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
ANSON	72	72	72	72	72	72
BAYFRONT SUBZONE	62	62	62	62	62	62
BENCOOLEN	65	65	65	65	65	65
BOAT QUAY	100	100	100	100	100	100
BOULEVARD	100	100	100	100	100	100
BRAS BASAH	73	73	73	73	73	73
BUGIS	100	100	100	100	100	100
CAIRNHILL	72	72	72	72	72	72
CECIL	100	100	100	100	100	100
CENTRAL SUBZONE	20	20	20	20	20	20
CHINA SQUARE	100	100	100	100	100	100
CHINATOWN	100	100	100	100	100	100
CITY HALL	83	83	83	83	83	83
CLARKE QUAY	100	100	100	100	100	100
CLIFFORD PIER	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

I then proceeded to get the top 10 most common venues using the frequency count for each subzone and assigned this to a dataframe. This dataframe became the input to the clustering of the sub-zones.

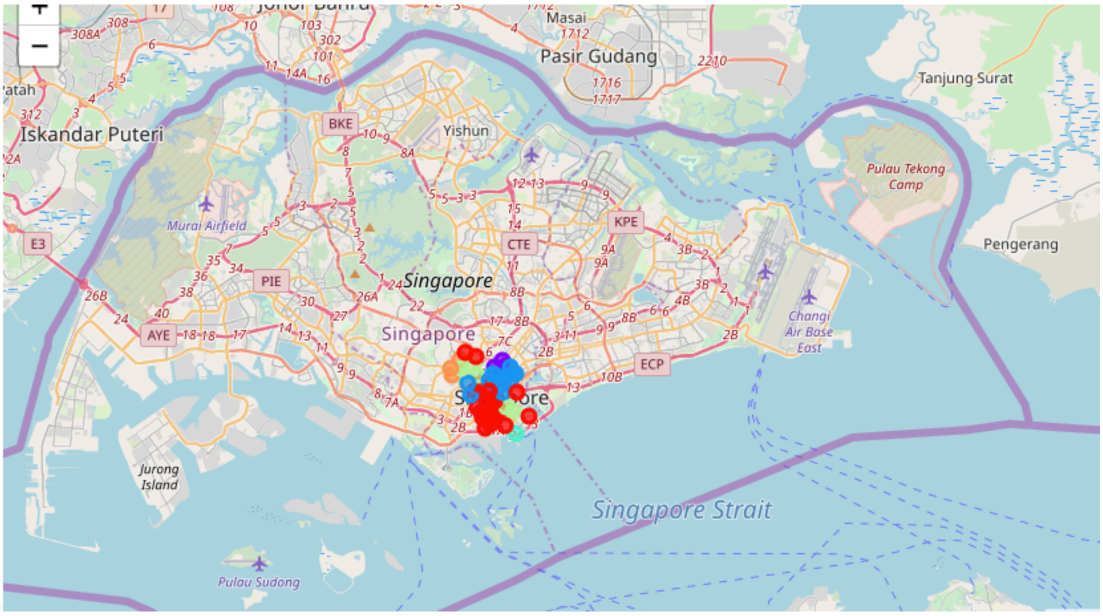
	Subzone	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	1 (
0	ANSON	Coffee Shop	Japanese Restaurant	Ramen Restaurant	Café	Hotel	Bakery	Italian Restaurant	Gym / Fitness Center	Korean Restaurant	R
1	BAYFRONT SUBZONE	Boutique	Hotel	Theater	Waterfront	Noodle House	Roof Deck	Casino	Tea Room	Bridge	R
2	BENCOOLEN	Café	Hotel	Indian Restaurant	Food Court	Sports Bar	Bakery	Chinese Restaurant	Dessert Shop	Art Gallery	I(
3	BOAT QUAY	Bar	Café	Nightclub	Gym / Fitness Center	Japanese Restaurant	Seafood Restaurant	Yoga Studio	Lounge	Food Court	
4	BOULEVARD	Boutique	Japanese Restaurant	Hotel	Café	Sushi Restaurant	Cosmetics Shop	Department Store	French Restaurant	Chinese Restaurant	

Looking at the initial output for the 1st 5 sub-zones, we can see that Cafe is a quite popular venue with 4 out 5 sub-zone:

Using the elbow curve to determine the number of cluster to use for K-means analysis, I used 6 clusters as the input. After getting the cluster labels I then merged the resulting dataframe with the original dataframe to get the Latitude and Longitude for the sub-zones:

	Longitude	Latitude	Subzone	Region Name	Central Area Indicator	Cluster name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Com V
0	103.844945	1.289751	CLARKE QUAY	CENTRAL REGION	Y	0	Japanese Restaurant	Hotel	Bar	Nightclub	Seafood Restaurant	
1	103.843342	1.274116	ANSON	CENTRAL REGION	Y	0	Coffee Shop	Japanese Restaurant	Ramen Restaurant	Café	Hotel	B
2	103.837635	1.293569	INSTITUTION HILL	CENTRAL REGION	Y	0	Japanese Restaurant	Hotel	Café	Coffee Shop	Wine Bar	Resta
3	103.859296	1.271357	STRAITS VIEW	CENTRAL REGION	Y	3	Boat or Ferry	Pier	Road	Bar	Seafood Restaurant	Govern Bu
4	103.848562	1.278497	CECIL	CENTRAL REGION	Y	0	Coffee Shop	Japanese Restaurant	Korean Restaurant	Café	Hotel	Ch Resta

And then used the Folium again to visualize the Map with the resulting clusters:



4. Results

From the clustering results, taking Cluster 1 as an example, we could describe Cluster 1 as Asian Restaurant given the number of Japanese Restaurants and Chinese Restaurants.

Cluster 1

```
]: central_merged.loc[central_merged['Cluster name'] == 0, central_merged.columns[[2] + list(range(5, central_merged.shap
```

]:

	Subzone	Cluster name	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	CLARKE QUAY	0	Japanese Restaurant	Hotel	Bar	Nightclub	Seafood Restaurant	Café	Thai Restaurant	Hotpot Restaurant	Spa	Bakery
1	ANSON	0	Coffee Shop	Japanese Restaurant	Ramen Restaurant	Café	Hotel	Bakery	Italian Restaurant	Gym / Fitness Center	Korean Restaurant	Tapas Restaurant
2	INSTITUTION HILL	0	Japanese Restaurant	Hotel	Café	Coffee Shop	Wine Bar	Restaurant	Steakhouse	Yoga Studio	Burger Joint	Seafood Restaurant
4	CECIL	0	Coffee Shop	Japanese Restaurant	Korean Restaurant	Café	Hotel	Chinese Restaurant	Restaurant	Italian Restaurant	Food Court	Sandwich Place
5	PHILLIP	0	Chinese Restaurant	Japanese Restaurant	Café	Yoga Studio	Cocktail Bar	Italian Restaurant	Gym / Fitness Center	Korean Restaurant	Salad Place	Restaurant
6	RAFFLES PLACE	0	Chinese Restaurant	Italian Restaurant	Café	Japanese Restaurant	Coffee Shop	Hotel	Cocktail Bar	Salad Place	Gym / Fitness Center	Korean Restaurant
10	CHINA SQUARE	0	Japanese Restaurant	Cocktail Bar	Chinese Restaurant	Café	Hostel	Hotel	Wine Bar	Italian Restaurant	Korean Restaurant	Gym / Fitness

For Cluster 2, it could be defined as Indian and Vegan Restaurants due to the majority of these categories.

Cluster 3 could be described as Hotel and Cafe neighbourhood.

Cluster 4 could be described as water venue.

Cluster 5 could be described as Botique neighbourhood.

Cluster 6 could be described as Hotel neighbourhood.

	Cluster name	1st Most Common Venue	count
0	0	Japanese Restaurant	8
1	2	Café	7
2	0	Chinese Restaurant	5
3	2	Hotel	4
4	4	Boutique	4
5	5	Hotel	3
6	1	Indian Restaurant	3
7	0	Coffee Shop	2
8	2	Fried Chicken Joint	1
9	2	Indian Restaurant	1
10	4	Japanese Restaurant	1
11	0	Pub	1
12	3	Boat or Ferry	1
13	0	Italian Restaurant	1
14	0	Hotel	1
15	0	Garden	1
16	4	Hotel	1

Based on the clustering above, taking cluster 1 as example, we can see that there is an opportunity to open other kinds of restaurant aside from Japanese and Chinese restaurants. Fast food or other kinds of cuisines like Mexican or middle eastern cuisine might be a good alternative to give the people more options. A convenience store is also a good opportunity given that there are a number of bars/nightclubs and other tourist attractions in the cluster.

5. Discussion

There is a clear market opportunity for increasing the number of F&B outlets in the Central Area of Singapore.

Given the size of Singapore, and the number of Tourists arrivals in Singapore, which was at 18.5 Million last 2018 (as per the Singapore Tourism Board data), there are plenty of opportunities to open up food places/restaurants. The analysis done here can be expanded to include the number of Tourists arrival in Singapore and try to determine what cuisines are preferred mostly, given the country of origin of the tourists.

I only used the sub-zones from the Central Area based on Central Area Indicator. The clustering could be done differently using the other regions in Singapore, and see how that compares with the Central Area. Also the number of clusters that I derived using the elbow curve could be changed to see how that would affect the result using K-means.

More in depth analysis can be performed using other data such as tourist data or when we look at the trend of data, that is, given a specific period of time, we observe how often the top 10 most common venues are changing.

6. Conclusion

In conclusion, there is definitely a market for more food places in Singapore. More market to expand/diversify the cuisines to give people more options.

Potential business owners/investors would benefit from a similar analysis by looking at the trending venues and the number of venues available for a particular area.