APPLIED DATA SCIENCE CAPSTONE (WEEK 5)

suitable location in Singapore for a new Food Court

By: Lee Kwong Yee

1. Introduction

For many people living in Singapore, there is a wealth of local food that can satisfy the palate of even the most fastidious. Food courts and coffee shops are the most popular locations for people to have their three meals as unlike restaurants, they are more affordable. In addition, food courts and coffee shops offer a variety of dishes, both local and regional cuisines under one roof.

Food courts and coffee shops offers great business opportunities in Singapore due to its metropolitan environment where most people prefer to eat out rather than cook. My client is interested to setup a food court as he believes that food court provides a cosier and more comfortable eating place as it is larger and air-conditioned, in comparison to a coffee shop which is smaller and open-air, albeit at a higher capital outlay. This to him is especially important, given Singapore's hot and humid climate.

My client has considered the food Court's prevalence, but he believes that as more and more people are eating out because of convenience, there is still considerable demand in this market segment. As with any other business decision, opening a food court requires serious consideration and is a lot more complicated than it seems. In particular, the location of the food court is one of the most important factors that will determine the profitability and sustainability of the business.

2. Business Problem

The objective of this project is to analyse and assist my client to select the best locations in the Singapore to open a new food court. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the following business question:

In Singapore, which is small, densely populated and with a metropolitan vibe, what is the best location to set up a food court, amidst the prevalence of eating places in the republic?

3. Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Singapore. For the listing, I have decided to look at the district codes in Singapore and use the listed neighbourhoods of the respective district codes to provide a holistic coverage of all the residential and commercial areas of Singapore.
- Latitude and longitude coordinates of these neighbourhoods. This is required in order to plot the map and to acquire the venue data.
- Venue data, which will be used to perform K-means clustering on the neighbourhoods.

Sources of data and methods to extract them:

This web page (https://www.mingproperty.sg/singapore-district-code/) contains a list of locations that are tied to the district codes of Singapore with a total of 68 neighbourhoods. I will use web scraping techniques to extract the data from web page with the help of Python Pandas requests. The geographical coordinates of the neighbourhoods will then be acquired either through Python Geocoder package or manually through the website https://www.maps.ie/coordinates.html.

After that, I will use Foursquare API to get the venue data for these neighbourhoods. Foursquare API will provide many categories of the venue data: I am particularly interested in the following categories:

- 'Food Court' and 'Coffee Shop', which are competing categories that will be assigned negative scores; and
- 'Shopping Mall' and 'Metro Station' which are complementing categories because of the crowd that they generate, in addition to being potential locations for setting up a food court; positive scores will be assigned to these categories.

Data science skills used for the project as follow:

- web scraping (Wikipedia),
- working with API (Foursquare),
- data cleaning and data wrangling (Pandas),
- machine learning (K-means clustering) and
- map visualization (Folium)

4. Methodology

Firstly, I need to get the list of neighbourhoods in Singapore based on the republic's district code. I leverage on web scraping through Python requests to extract the listing from the web page "https://www.mingproperty.sg/singapore-district-code/" into the following data frame:

	Postal District	Postal Sector(1st 2 digits of postal codes)	General Location
0	01	01, 02, 03, 04, 05, 06	Raffles Place, Cecil, Marina, People's Park
1	02	07, 08	Anson, Tanjong Pagar
2	03	14, 15, 16	Queenstown, Tiong Bahru
3	04	09, 10	Telok Blangah, Harbourfront
:			
25	26	77, 78	Upper Thomson, Springleaf
26	27	75, 76	Yishun, Sembawang
27	28	79, 80	Seletar

The column 'General Location' is used to provide the list of 68 neighbourhoods needed for the project.

Next step is to acquire the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, I try using the Geocoder package to convert the General Locations into the form of latitude and longitude but most of results return the same coordinates, in part attributable to the small size of Singapore.

As such, I decide to use the website "https://www.maps.ie/coordinates.html" to manually acquire the coordinates as it is able to suggest the relevant street names that are associated with the neighbourhoods and hence provides more precise coordinates.

The final listing of neighbourhoods and their respective coordinates as follow:

	Neighborhood	Latitude	Longitude
0	Raffles Place	1.283595	103.851568
1	Cecil	1.276564	103.846958
2	Marina	1.291274	103.857184
3	People's Park	1.284139	103.842557
	:		
64	Springleaf	1.396743	103.818899
65	Yishun	1.428136	103.833694
66	Sembawang	1.448065	103.820760
67	Seletar	1.409849	103.877379

I perform visualization of the neighbourhoods in a map using Folium package as a sanity check to make sure that the geographical coordinates are correctly plotted. The plot reveals that

the neighbourhoods do indeed cover most of the residential and commercial districts in Singapore.

Next, I use Foursquare API to get the top 50 venues that are within a radius of 500 meters.

I make API calls to Foursquare by passing the geographical coordinates of the neighbourhoods. Foursquare returns the venue data in JSON format from which I proceed to extract the venue name, venue category, venue latitude and venue longitude.

To facilitate the analysis process, I convert all the venue categories from row data to columns using a Pandas Transpose function:

From:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Raffles Place	1.283595	103.851568	The Fullerton Bay Hotel	1.283878	103.853314	Hotel
1	Raffles Place	1.283595	103.851568	CITY Hot Pot Shabu shabu	1.284173	103.851585	Hotpot Restaurant
2	Raffles Place	1.283595	103.851568	Virgin Active	1.284608	103.850815	Gym / Fitness Center

To:

	Neighborhood	Accessories Store	Airport	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Video Game Store	Vietnar Restar
0	Raffles Place	0	0	0	0	0	0	0	0	0	...	0	
1	Raffles Place	0	0	0	0	0	0	0	0	0	...	0	
2	Raffles Place	0	0	0	0	0	0	0	0	0	...	0	

With the data, I then analyse each neighbourhood by grouping the venues by neighbourhood and taking the mean of the frequency of occurrence of each venue category (see below). By doing so, I am also preparing the data for use in clustering.

	Neighborhood	Accessories Store	Airport	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Vietnamese Restaurant
0	Amber Road	0.00	0.00	0.052632	0.00	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000
1	Ang Mo Kio	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.040816	0.000000	...	0.000000
2	Anson	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000
3	Ardmore	0.00	0.00	0.025641	0.00	0.025641	0.00	0.00	0.000000	0.000000	...	0.000000
4	Balestier	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.041667	0.000000	...	0.000000
5	Beach Road (part)	0.00	0.00	0.000000	0.00	0.022727	0.00	0.00	0.000000	0.000000	...	0.045455
6	Bedok	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.083333	0.000000	...	0.000000

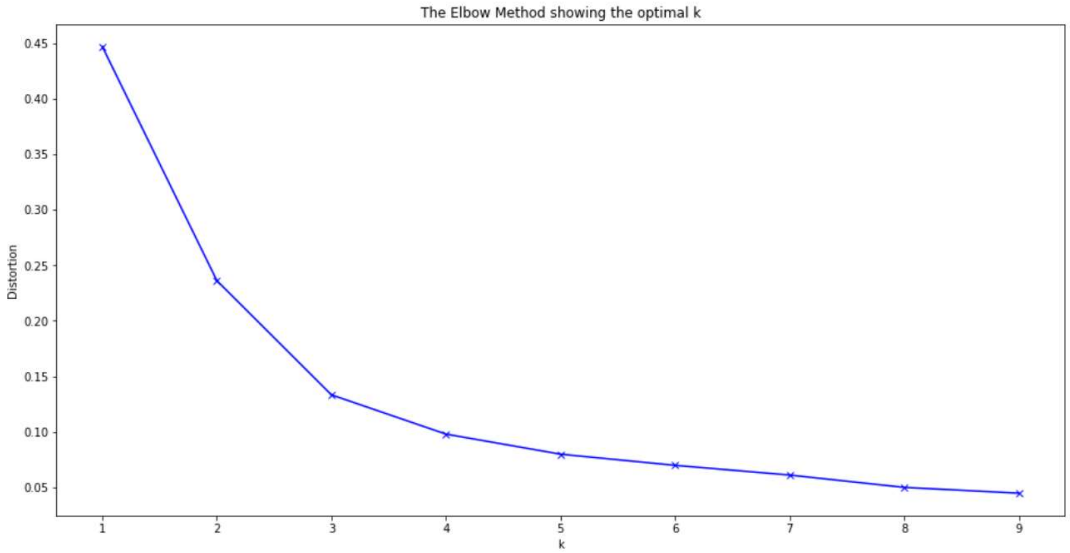
As mentioned in the data section, I am only interested in the following categories:

- Food Court' and 'Coffee Shop', which are competing categories, and
- 'Shopping Mall' and 'Metro Station' which are complementing categories because of the crowd that they generate, in addition to being potential locations for setting up a food court.

The categories of interest are extracted to form the following data frame for clustering:

	Neighborhood	Food Court	Coffee Shop	Shopping Mall	Metro Station
0	Amber Road	0.000000	0.000000	0.000000	0.0
1	Ang Mo Kio	0.081633	0.081633	0.020408	0.0
2	Anson	0.000000	0.080000	0.020000	0.0
3	Ardmore	0.000000	0.000000	0.000000	0.0
4	Balestier	0.062500	0.020833	0.020833	0.0

Lastly, I perform clustering on the data by using K-means clustering. K-means clustering algorithm identifies K number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. I determine that the number of clusters (k) to use by using the Elbow method (see following diagram) and the result indicates that k=5 is the optimal.

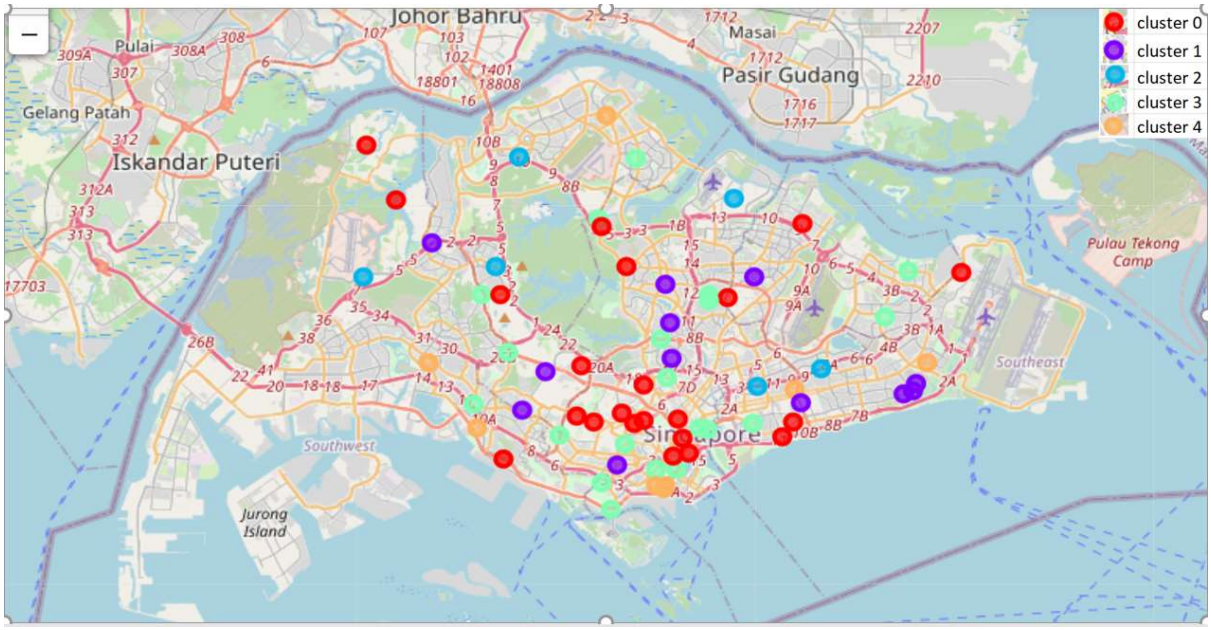


With the results, I further calculate the mean value of each category for the respective clusters to derive the following data frame. The mean values for “Food Court” and Coffee Shop” are multiplied by -1 to derive negative values to indicate that they are competing factors, while the mean values for “Shopping Mall” and “Metro Station” remain as positive to indicate that they are complementing factors.

	Cluster	Food Court(-)	Coffee Shop(-)	Shopping Mall(+)	Metro Station(+)
0	c0	-0.000000	-0.005455	0.007273	0.000000
1	c1	-0.107615	-0.082940	0.003595	0.002874
2	c2	-0.227062	-0.020833	0.015152	0.006944
3	c3	-0.051452	-0.042914	0.018459	0.003125
4	c4	-0.033705	-0.104152	0.023884	0.002500

5. Results

The folium map for the cluster distribution is as follow; for the two largest clusters, cluster 0 comprises of mostly residential districts whilst cluster 3 has a good mix of both residential and commercial districts.



The “Score” column of each cluster is tabulated as a summation of the row data based on the following formula:

*Cluster score = (‘Food Court’ + ‘Coffee Shop’) *(-1) + (‘Shopping Mall’+ ‘Metro Station’)*

The ranking of the clusters in terms of score is as follow after sorting:

Rank	Cluster	Food Court(-)	Coffee Shop(-)	Shopping Mall(+)	Metro Station(+)	Score
1st	c0	-0.000000	-0.005455	0.007273	0.000000	0.001818
2nd	c3	-0.051452	-0.042914	0.018459	0.003125	-0.072782
3rd	c4	-0.033705	-0.104152	0.023884	0.002500	-0.111473
4th	c1	-0.107615	-0.082940	0.003595	0.002874	-0.184087
5th	c2	-0.227062	-0.020833	0.015152	0.006944	-0.225800

From the ranking, it seems that cluster 0 is the best location to set up a food court. However, a holistic analysis would still be required as the rankings serve basically as a reference rather than a determinant.

6. Discussion

From the clustering exercise and based on the rankings above, cluster 0 is ranked the highest based on the score. However, we can observe that the high score is mainly attributable to the lack of Food Court and Coffee Shop in the vicinity of the neighbourhood. Although there is less competition, the mean value of Shopping Mall also indicates that the neighbourhood has a comparatively low number of shopping malls compared to other clusters. Cluster 0 also has a low mean value for Metro Station. All these suggest that the crowd volume might not be ideal to derive the necessary demand.

I would instead recommend cluster 3 (which is ranked second) as the cluster is moderate in terms of Food Court and Coffee Shop, and at the same time dense in terms of Shopping Mall and Metro Station. All these factors suggest that cluster 3 would be a safer bet than cluster 0, albeit its lower ranking.

7. Conclusion

Based on holistic analysis of the clustering results, it is deduced that cluster 3 would be an ideal location to set up a Food Court for my client.

However, other factors such as population and income of residents could influence the location decision of the Food Court and such data is not available at neighbourhood level required by this project.

Nevertheless, site survey would be able to bridge this gap to affirm and select the most ideal location amongst all neighbourhoods listed within cluster 3.