IBM Applied Data Science Capstone

April 2020

Opening a Bubble Tea Chain in Kuala Lumpur, Malaysia

Coursera Capstone Project:
The Battle of Neighbourhoods

EUGENE WONG WEE LUNN

TABLE OF CONTENTS

INTRODUCTION

1.1 Background

Bubble tea is a tea-based drink originated from Taiwan in the 1980's (Martin, Laura C., 2007), it is served typically with tapioca balls called "pearls" or "boba" which adds a unique texture to the drink. There are many varieties of the drink with a wide range of flavours and toppings, with the two most popular varieties being: "black pearl milk tea" and "green pearl milk tea" (Chang, Derrick, 2012). The sweet, creamy milk tea paired with chewy "boba" has proved to be an addictive combination amongst the people.

Taiwan has an established market for bubble tea, having many successful brands throughout the country. In recent years, the popularity of bubble tea has skyrocketed and made its way internationally with many Taiwanese bubble tea brands opening their franchises to the global market, including Malaysia. The people of Malaysia tend to favour bubble tea places as a staple 'go-to' hangout spot, instead of the usual cafes. This overwhelming trend continues to lure more bubble tea brands into Malaysia.

1.2 Business Problem

Throughout the year of 2019, Malaysia has seen a high influx of bubble tea chains entering its market. This project aims to utilise the data science methodology and techniques to analyse, if a bubble tea brand intends to open a chain in Kuala Lumpur, Malaysia, where would be the ideal location to set up a store?

1.3 Target Audience

New bubble tea brands looking to venture into the Malaysian market or even existing bubble tea chains in Malaysia looking to expand their store locations, would be the target audience of this project. The analysis provided by this project will give an idea to interested stakeholders, of the level of market saturation in a particular area in Kuala Lumpur, Malaysia.

DATA

2.1 Data Sources

The following sets of data are needed to carry out this project:

- *List of neighbourhoods in Kuala Lumpur, Malaysia*

  A list of neighbourhood names in Kuala Lumpur is essential to the scope of the project. The source of this data can be found in a Wikipedia page regarding Kuala Lumpur (https://en.wikipedia.org/wiki/Kuala_Lumpur), a sub-page located at the bottom, called 'Kuala Lumpur metropolitan area', contains a list of neighbourhood names that are present in Kuala Lumpur, arranged in table format.

- *Geographical coordinates of the neighbourhoods*

  This project will be utilizing Foursquare location data which requires the latitude and longitude coordinates of each neighbourhood to be used as inputs. The geographical coordinate data can be extracted using Geopy's Geocoder library in Python.

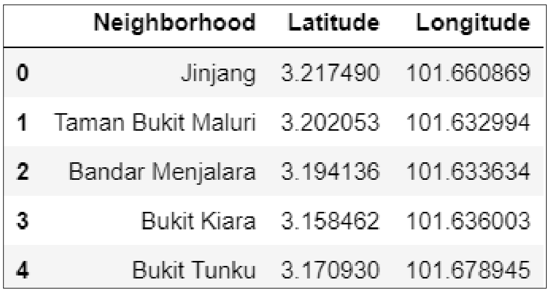
- *Bubble tea shops in each neighbourhood*

  To answer the business problem, analysing the level of market saturation would be the focus of this analysis. Foursquare API will be called to search a venue (neighbourhood) for existing bubble tea shops in the neighbourhood's vicinity. This enables us to collect data on the total number of bubble tea shops in a neighbourhood, as well as number of unique bubble tea shops. Following that, K-means clustering will be done to segregate the neighbourhood into different cluster labels, which would give us a rough idea on the level of market saturation per neighbourhood. The resulting clusters will be superimposed onto Kuala Lumpur's map, as part of the data visualization process.

METHODOLOGY

3.1 Data Acquisition & Cleaning

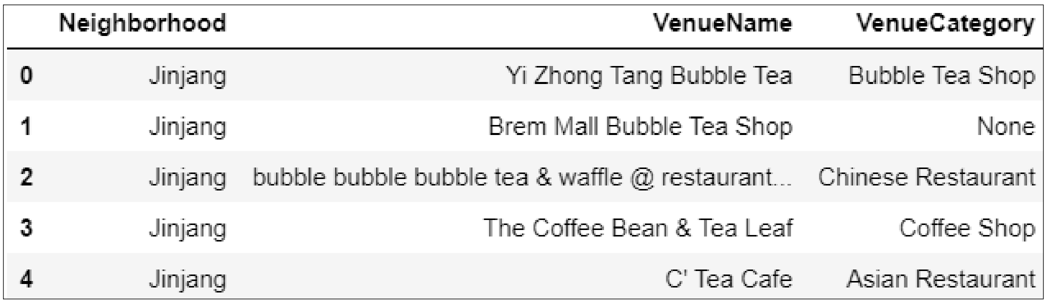
Begin with acquiring data of a list of neighbourhood names in Kuala Lumpur, Malaysia. The data source is located in a Wikipedia page, hence, web scraping was done using Beautiful Soup to extract the list of names of neighbourhoods. Beautiful Soup is a Python package for parsing HTML and XML documents. An empty pandas data frame was created to store the data and any duplicate data was dropped. According to the data source, there are 70 metropolitan areas or major neighbourhoods in Kuala Lumpur.



	Neighborhood	Latitude	Longitude
0	Jinjang	3.217490	101.660869
1	Taman Bukit Maluri	3.202053	101.632994
2	Bandar Menjalara	3.194136	101.633634
3	Bukit Kiara	3.158462	101.636003
4	Bukit Tunku	3.170930	101.678945

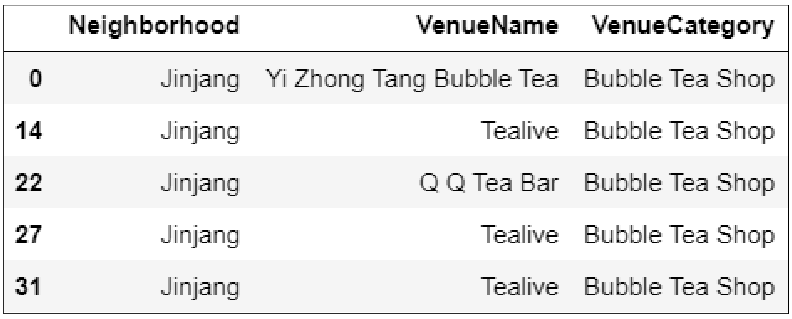
Figure 1: First 5 neighbourhoods with its corresponding latitude and longitude.

Geopy’s geocoder library was used to convert the address of the neighbourhoods into geographical coordinates of latitude and longitude. This was done with a for-loop, iterating each neighbourhood’s name as inputs to the geocoder. Unfortunately, some neighbourhoods’ geographical coordinates were unable to be detected by the geocoder. These neighbourhoods were omitted as data points for the analysis (a total of 9 neighbourhoods were removed; 61 neighbourhoods remain as workable data). The latitude and longitude coordinate data were merged together with neighbourhood names, to display each neighbourhood with its corresponding geographical coordinates, as shown in Figure 1.



	Neighborhood	VenueName	VenueCategory
0	Jinjang	Yi Zhong Tang Bubble Tea	Bubble Tea Shop
1	Jinjang	Brem Mall Bubble Tea Shop	None
2	Jinjang	bubble bubble bubble tea & waffle @ restaurant...	Chinese Restaurant
3	Jinjang	The Coffee Bean & Tea Leaf	Coffee Shop
4	Jinjang	C' Tea Cafe	Asian Restaurant

Figure 2: First 5 rows of results using Foursquare location data.

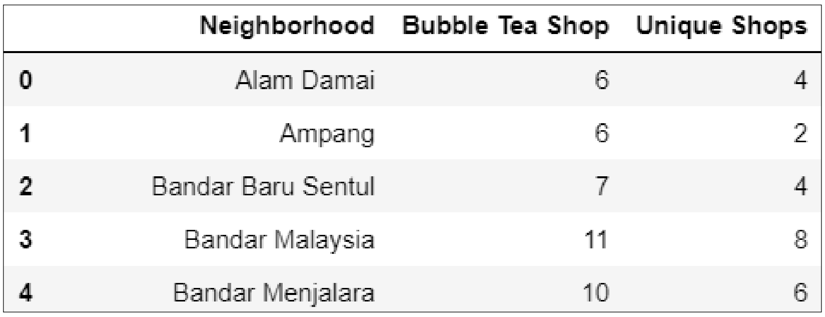


	Neighborhood	VenueName	VenueCategory
0	Jinjang	Yi Zhong Tang Bubble Tea	Bubble Tea Shop
14	Jinjang	Tealive	Bubble Tea Shop
22	Jinjang	Q Q Tea Bar	Bubble Tea Shop
27	Jinjang	Tealive	Bubble Tea Shop
31	Jinjang	Tealive	Bubble Tea Shop

Figure 3: First 5 rows of results after filtering venue category for ‘Bubble Tea Shop’.

Next, we will need to search for bubble tea shops in each neighbourhood’s vicinity. Foursquare API was called, iterating over each neighbourhood’s geographical coordinates as inputs to search for bubble tea shops located in respective neighbourhoods. A radius limit of 3,000 metres and results limit of 100 was set when calling the Foursquare API, and the output results will be in JSON format. Relevant data such as the venue name and venue category were extracted and appended into an empty pandas data frame along with its corresponding neighbourhood (Figure 2). Then, all venue categories that are not bubble tea shops were removed from the data frame (Figure 3).

3.2 Data Analysis



	Neighborhood	Bubble Tea Shop	Unique Shops
0	Alam Damai	6	4
1	Ampang	6	2
2	Bandar Baru Sentul	7	4
3	Bandar Malaysia	11	8
4	Bandar Menjalara	10	6

Figure 4: First 5 neighbourhoods with its number of bubble tea shops and unique bubble tea shops.

According to Foursquare location data, there are a total of 410 bubble tea shops in Kuala Lumpur and 48 unique bubble tea shops. To further analyse the dataset, bubble tea shops of each neighbourhood were aggregated and grouped according to their respective neighbourhoods, showing the total number of bubble tea shops in each neighbourhood. This step was replicated to also find the total number of unique bubble tea shops for each neighbourhood. Now, we have a data frame which consists of the total number of bubble tea shops, as well as total number of unique bubble tea shops in each neighbourhood, as shown in Figure 4.

3.3 Clustering

The data set was grouped into clusters of similar characteristics using K-Means clustering. K-Means clustering is an unsupervised machine learning algorithm which works by identifying k number of centroids, then it allocates every data point to the nearest centroid forming a cluster, while minimizing intra-cluster distance and maximizing inter-cluster distance. It is one of the simplest unsupervised machine learning algorithm to implement and works well with multidimensional data sets. Since the goal is to determine market saturation levels based on two variables (1. Number of bubble tea shops; 2. Number of unique bubble tea shops), this method will provide a good idea on which group of neighbourhoods would be the best to open a bubble tea shop.

3.4 Data Visualization

Finally, using a library in Python for map-rendering called Folium, a map was created, centred on Kuala Lumpur, Malaysia. The resulting clusters were superimposed onto Kuala Lumpur's map with markers, showing the neighbourhood's name and its corresponding cluster label. A colour scheme was also added to provide better visualization for different clusters.

RESULTS

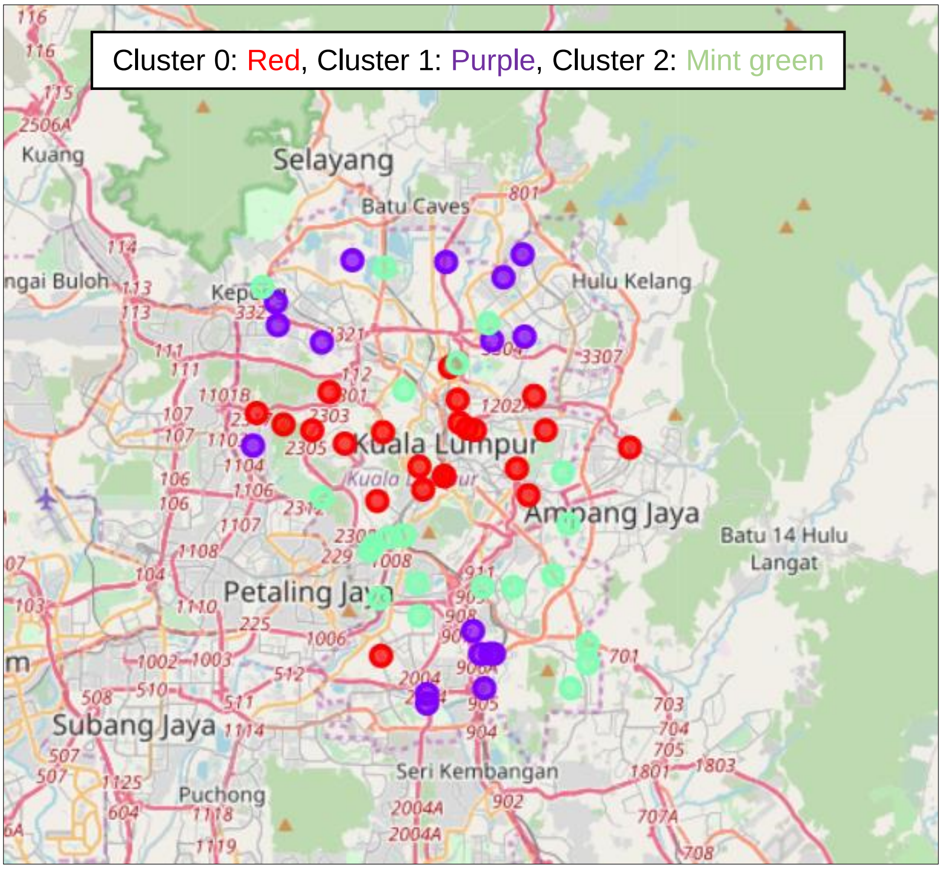


Figure 5: Neighbourhood clusters superimposed onto Kuala Lumpur's map.

Table 1: Neighbourhoods in Cluster 0 (23 neighbourhoods).

Neighbourhood	Bubble Tea Shops	Unique Shops	Cluster Label
Ampang	6	2	0
Bangsar	6	2	0
Brickfields	2	1	0
Bukit Kiara	3	2	0
Bukit Nanas	5	2	0
Chow Kit	5	2	0
Damansara Heights	4	1	0
Dang Wangi	4	2	0
Federal Hill	2	1	0
Jalan Cochrane	3	3	0
Jalan Duta	5	1	0
Kampung Datuk Keramat	2	2	0
Kampung Sungai Penchala	4	2	0
Kuala Lumpur City Centre	2	1	0
Kuala Lumpur Sentral	3	2	0
Medan Tuanku	4	1	0
Mont Kiara	5	3	0
Perdana Botanical Gardens	3	1	0
Sentul	6	2	0

Sri Hartamas	5	3	0
Taman OUG	5	3	0
Taman U-Thant	3	3	0
Tun Razak Exchange	4	3	0

Table 2: Neighbourhoods in Cluster 1 (17 neighbourhoods).

Neighbourhood	Bubble Tea Shops	Unique Shops	Cluster Label
Bandar Malaysia	11	8	1
Bandar Menjalara	10	6	1
Bandar Tasik Selatan	11	8	1
Bukit Jalil	10	9	1
Desa Petaling	8	7	1
Jinjang	10	5	1
Kampung Malaysia	10	8	1
Padang Balang	8	7	1
Semarak	12	6	1
Setapak	9	6	1
Sri Petaling	10	9	1
Sungai Besi	9	8	1
Taman Bukit Maluri	9	6	1
Taman Ibukota	12	8	1
Taman Melati	13	9	1
Taman Sri Sinar	11	6	1
Taman Tun Dr Ismail	10	6	1

Table 3: Neighbourhoods in Cluster 2 (21 neighbourhoods).

Neighbourhood	Bubble Tea Shops	Unique Shops	Cluster Label
Alam Damai	6	4	2
Bandar Baru Sentul	7	4	2
Bandar Sri Permaisuri	6	5	2
Bangsar South	9	4	2
Bukit Tunku	8	3	2
Damansara	8	5	2
Damansara Town Centre	6	3	2
KL Eco City	8	4	2
Kampung Pandan	5	4	2
Kerinchi	8	4	2
Kuchai Lama	5	4	2
Mid Valley City	8	4	2
Pantai Dalam	7	5	2
Salak South	5	5	2
Shamelin Perkasa	6	5	2
Taman Connaught	7	4	2
Taman Desa	7	5	2
Taman Len Seng	7	4	2
Taman Midah	6	4	2
Taman P. Ramlee	8	5	2
Taman Wahyu	9	5	2

Figure 5 and Table 1, 2 and 3 above shows that Cluster 0 (RED) consist of neighbourhoods with generally low number of bubble tea shops and low number of unique bubble tea shops. In contrast, neighbourhoods in Cluster 1 (PURPLE) demonstrated opposing characteristics with a seemingly high number of bubble tea shops, as well as a high number of unique bubble tea shops. Cluster 2 (MINT GREEN) contain neighbourhoods with moderately high bubble tea shops and unique bubble tea shops.

DISCUSSION

Neighbourhoods in Cluster 0 have great potential for a bubble tea brand to open a chain as they would face low levels of competition in contrast to the other clusters. There are a few neighbourhoods in Cluster 0, such as Ampang, Bangsar and Sentul which has 6 bubble tea shops, a comparatively high value among other neighbourhoods of the same cluster. However, there are only 2 unique bubble tea shops in these neighbourhoods, which could enable a bubble tea chain of a different brand to still be effective in setting up a chain here as it would act as a direct competitor in an overall low brand competition area.

Cluster 1 has the highest level of competition among the clusters, it has a combination of high product and brand competition. Opening a bubble tea chain here would likely cause the brand to suffer from intense competition in a highly saturated market. Meanwhile, Cluster 2 exhibits some considerable potential to open a chain. Although, having moderately high bubble tea shops in the area, the market is not as saturated as Cluster 1. Moreover, if positioned strategically, the brand could challenge and benefit from existing customer traffic from the other bubble tea shops.

CONCLUSION

This project was conducted by utilising the data science methodology to first, identify the business problem of which would be the best location to open a bubble tea chain in Kuala Lumpur, Malaysia, as well as identifying relevant stakeholders. Following the scope of the project, we have sourced, acquired and cleaned the data needed using various techniques and tools. Next, we clustered the data with machine learning algorithm into different groups with varying characteristics, to analyse the results and finally visualizing it on a map.

Ultimately, for a bubble tea brand looking to open a chain in Kuala Lumpur, it is recommended to do it in Cluster 0, as it has minimal competition but it would require the bubble tea brand to generate an initial customer traffic . Cluster 2 can be considered, as many bubble tea shops are present to provide existing customer traffic in the area, whilst the market is not at the point of saturation yet. Lastly, Cluster 3 should be avoided due to its nature of a highly saturated market.

REFERENCES

Martin, Laura C. (2007). Tea: The drink that changed the world. Rutland: Tuttle Publishing. p. 219. ISBN 9780804837248. (https://en.wikipedia.org/wiki/Bubble_tea)

Chang, Derrick (2012). "Is this the inventor of bubble tea?" International Edition. CNN (https://en.wikipedia.org/wiki/Bubble_tea)

Kuala Lumpur metropolitan areas. Wikipedia. (https://en.wikipedia.org/wiki/Kuala_Lumpur)

Foursquare Developers Documentation. (https://developer.foursquare.com/docs)