IBM Data Science Final Capstone Project: Opening A Bubble Tea Shop in Singapore

Report By Esther Tan
December 2019

1. Introduction

Bubble Tea, also known as pearl milk tea or boba tea is a growing popular drink in Asian countries. The pearls/ bobas are made of tapioca with freshly brewed tea as the base drink. Users can choose from different variations of teas like black tea, green tea or oolong tea, or from a wide range of flavors such as earl grey, passion fruit, taro, plum etc. There is an increasing number of bubble tea shops opening over the years and food establishments offering bubble tea on their menu. The global bubble tea market was valued at US$1.96 billion in 2016 and is projected to reach US$3.21 billion by 2023 (Sharma, 2018).

1.1 Business Problem

Many established Taiwan bubble tea brands have started selling bubble tea in Singapore more than 10 years ago. An establish brand is KOI which started selling back in 2007 and now has a strong presence of 57 outlets across Singapore. Singapore being an island state is a small country, the current seemingly saturated bubble tea market means there is potentially more than one bubble tea shop operating at each area country wide.

For a new entrant that wants to open a store in Singapore, we want to locate the areas where competition is less stiff. I will be using K-Means, an unsupervised learning clustering algorithm to cluster the shops island wide and see if we can find any patterns among to determine the suitability of the area.

2. Dataset

Data required:

- Singapore Planning Areas as demarcated by the government[1]
- Latitude and Longitude coordinates taken using OpenCage Geocoder API [2]
- Corresponding household population size residing in each Planning Area. Statistics taken from the Singapore government portal, latest figures available for download is 2019.

[1] https://data.gov.sg/dataset
[2] https://opencagedata.com/

- List of bubble tea shops and shops selling bubble tea within the planning areas. List will be retrieved using the Foursquare API[3]

3. Methodology

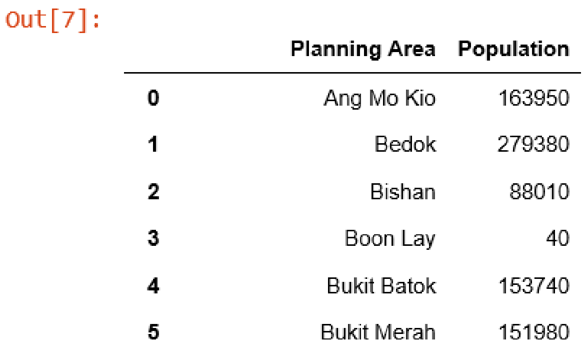
3.1 Data acquisition and data cleaning

First we import the csv file containing the planning areas and population statistics for each planning area into a new dataframe. We will only need the planning areas and 2019 population figures. Replace empty population cell values with 0 and keep planning areas that are for residential or commercial uses. Areas for example the Central Water Catchment are mainly reservoirs and the Western Islands and Tengah is a secluded island, hence they are excluded from the dataset. Notice that some planning areas have very low population figures signaling incomplete figures that will create outliers in the data when in reality the actual population figures are much higher. Hence for the following planning areas:

- Paya Lebar and Geylang
- Straits View, Marina South and Downtown Core
- Boon Lay, Pioneer and Jurong West
- Orchard and Newton
- Changi and Changi Bay

these planning areas are beside or in overlaps with each other, are combined together to decrease the disparity in the population figures for the entire dataset.

After combining the areas, the resulting dataframe is ready to use for retrieving the geo coordinates.



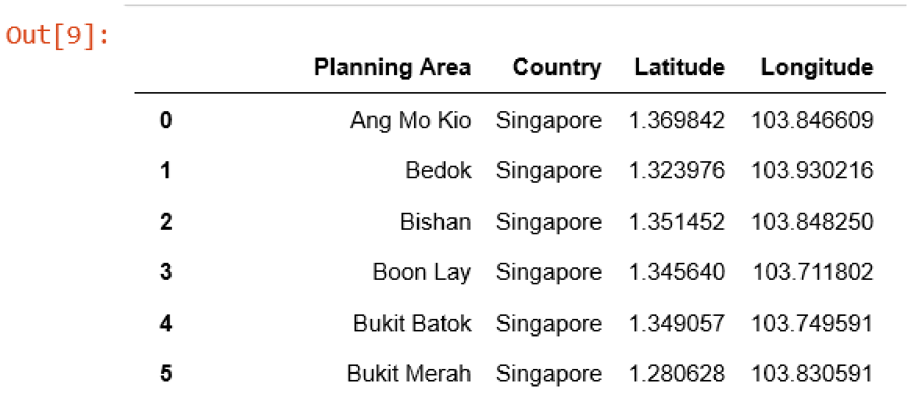
Out[7]:

	Planning Area	Population
0	Ang Mo Kio	163950
1	Bedok	279380
2	Bishan	88010
3	Boon Lay	40
4	Bukit Batok	153740
5	Bukit Merah	151980

[3] https://developer.foursquare.com/

3.2 Retrieving Geocoordinates

Connect to OpenCage API, retrieve the latitude and longitutde coordinates, clean the json results and store into the dataframe. Check through the coordinates to ensure a good estimate of the coordinates are retrieved and update the rows if necessary.



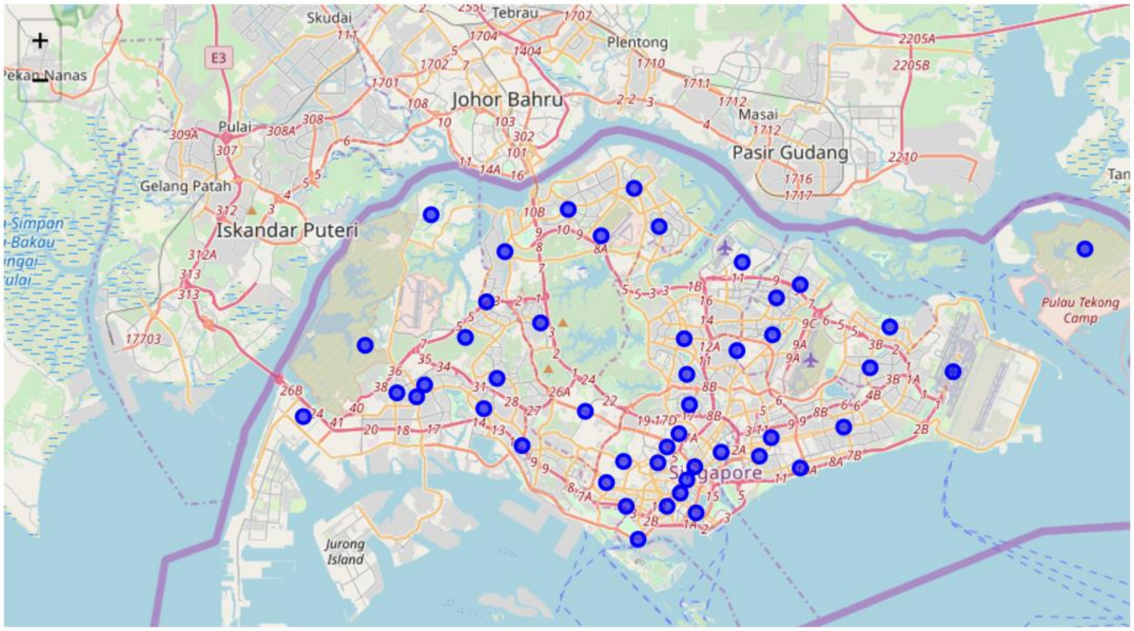
Out[9]:

	Planning Area	Country	Latitude	Longitude
0	Ang Mo Kio	Singapore	1.369842	103.846609
1	Bedok	Singapore	1.323976	103.930216
2	Bishan	Singapore	1.351452	103.848250
3	Boon Lay	Singapore	1.345640	103.711802
4	Bukit Batok	Singapore	1.349057	103.749591
5	Bukit Merah	Singapore	1.280628	103.830591

3.3 Plotting Planning Area Markers on Folium Map

Resulting planning areas are plotted on the map using Folium:

Out[14]:



3.4 Getting list of shops selling bubble tea using Foursquare

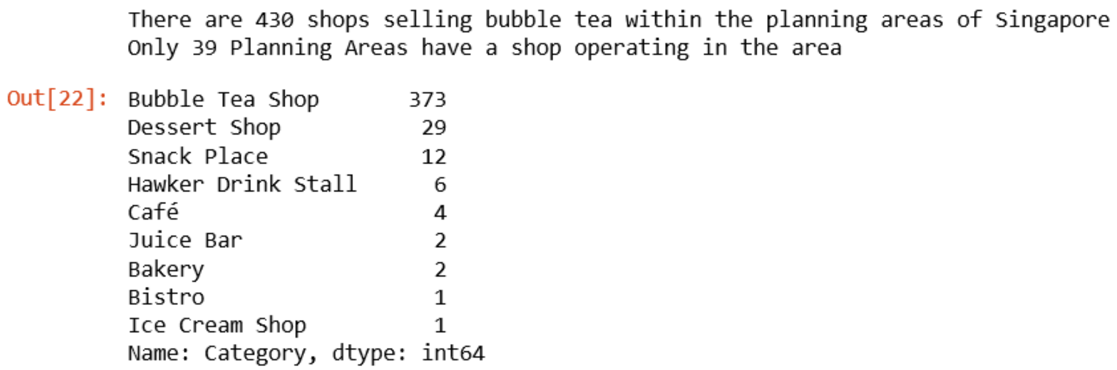
Use Foursquare API to get all the bubble tea shops in the respective planning areas. Limit the query to 100 bubble tea shops within 1500m radius. Clean the json results and store into a dataframe.

Some Planning Areas tend to be closer to each other hence the shops may be sorted under more than one Planning Area even under 1500 radius. Find duplicate shop IDs and assign them to their respective Planning Area base on the minimum distance. Drop duplicates.

Export the dataframe as a csv file and check through the results. Validate the shops existence by googling and searching on the brand's website if available. Shops that are no longer in operation are removed from the list, variations in shop name are edited for consistency. Near-similar addresses that appeared under two different shop IDs are also removed. Two additional dessert shops- BlackBall and Nine Fresh, whom also sells bubble tea but is not found in Foursquare's database were added in. Check through the data and load the csv file back in as a dataframe.

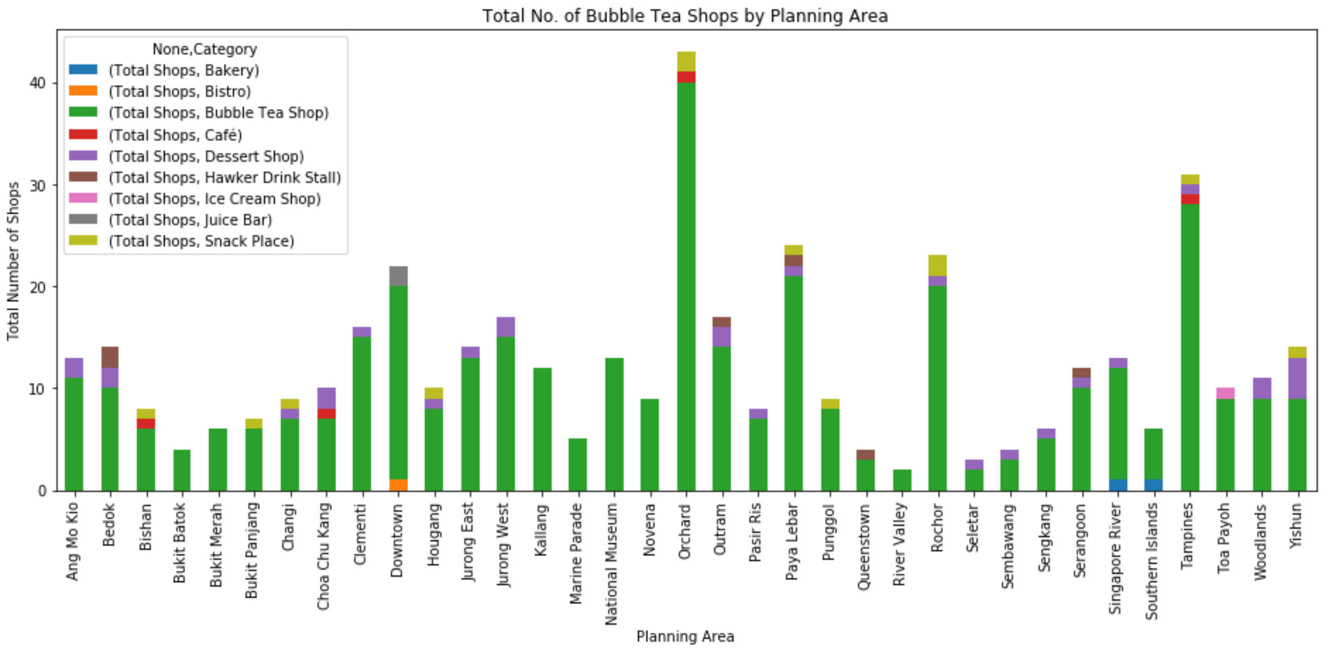
4. Exploratory Data Analysis

After cleaning the data, there are 430 shops selling bubble tea within 39 planning areas of Singapore as retrieved by the Foursquare API. The following shows the different category of shops selling bubble tea:

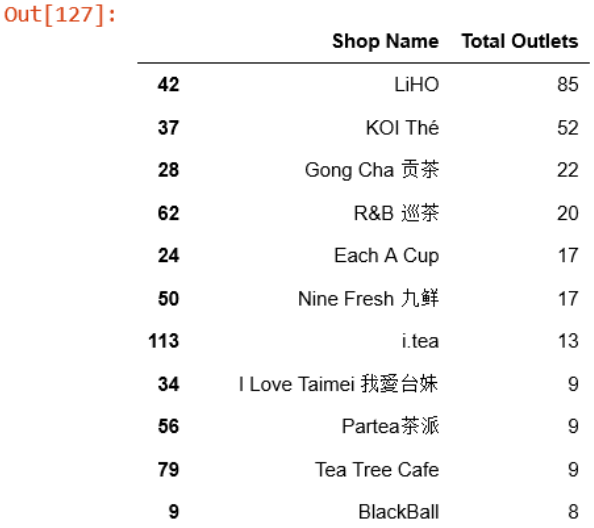


```
         There are 430 shops selling bubble tea within the planning areas of Singapore
         Only 39 Planning Areas have a shop operating in the area

Out[22]: Bubble Tea Shop       373
         Dessert Shop           29
         Snack Place            12
         Hawker Drink Stall      6
         Café                    4
         Juice Bar               2
         Bakery                  2
         Bistro                  1
         Ice Cream Shop          1
         Name: Category, dtype: int64
```



The main competing bubble tea brands that have more than the average 4 outlets per brand:



Out[127]:

	Shop Name	Total Outlets
42	LiHO	85
37	KOI Thé	52
28	Gong Cha 贡茶	22
62	R&B 巡茶	20
24	Each A Cup	17
50	Nine Fresh 九鲜	17
113	i.tea	13
34	I Love Taimei 我愛台妹	9
56	Partea茶派	9
79	Tea Tree Cafe	9
9	BlackBall	8

5. K-Means Clustering

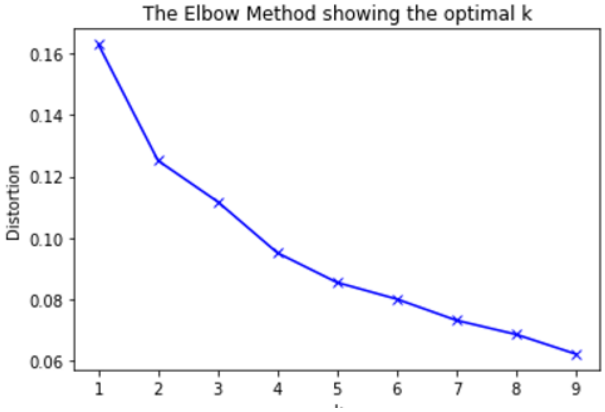
K-Means is an unsupervised learning algorithm that tries to find interesting patterns among the data and splits them into clusters based on its attributes. To prepare the data for K-Means clustering, first encode the bubble tea shops through one hot encoding and take the mean of frequency occurrence for each shop category.

Household population is an attribute that will influence the decision of opening a bubble tea shop in a particular area as it's likely inferred that a higher population density will have more traffic flow. Before adding the population figures, use scikit learn preprocessing to normalize the population statistics for each Planning Area before adding back to the main dataframe. 5 bins are also created to segment the populations into very low, low, medium, high and very high for easy visualization of the data.

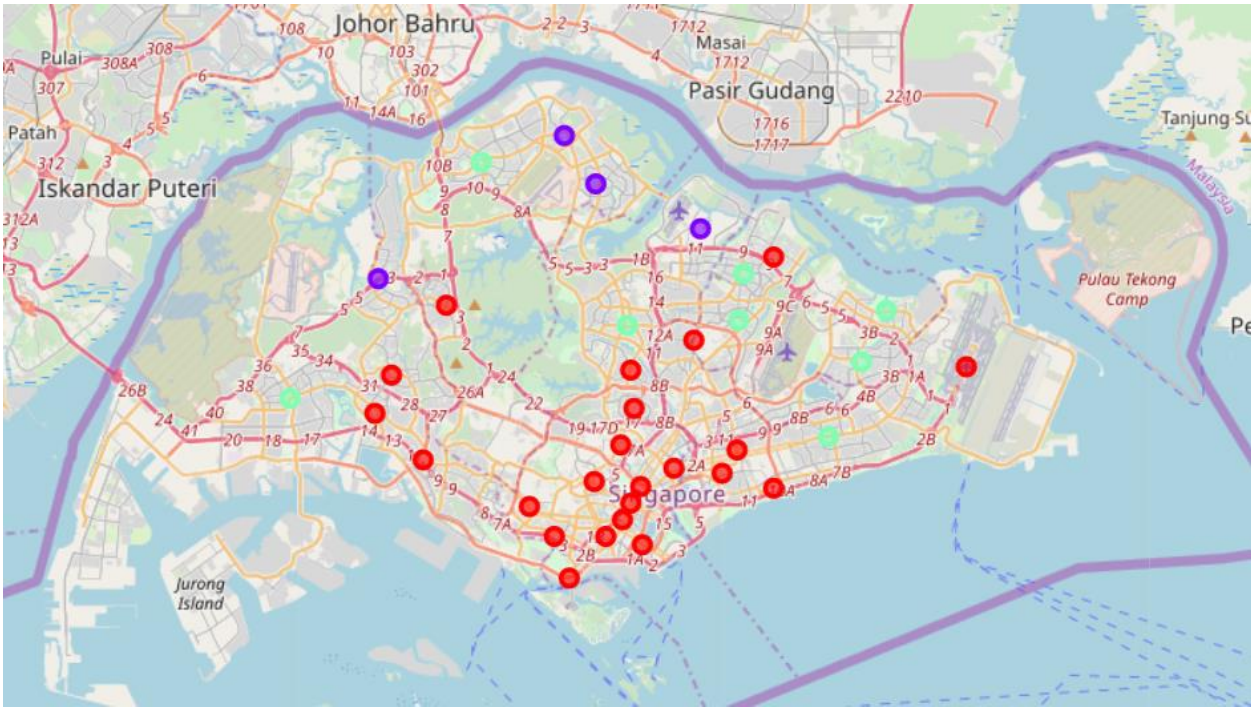
The final dataframe used for clustering consists of the Shop Categories represented by the columns, the last column being the population figures, and the planning areas represented by each row in the dataframe.

5.1 Finding the optimal number of clusters and performing K-Means Clustering

The optimal k=3 clusters is found through the elbow method and silhouette score. After running the K-Means clustering algorithm, this is the clusters plotted on the map



5.2 Visualizing clusters on the map



5.3 Interpreting Cluster Results

We can label each cluster as:

- Cluster 1 (red markers): Planning area with high density of bubble tea shops and majority low household population
- Cluster 2 (green markers): Planning area with low density of bubble tea shops and medium size household population
- Cluster 3 (purple markers): Planning area with moderate density of bubble tea shops and high household population

Cluster 1

Cluster 1 (red markers) can be interpreted as areas that have a higher density of bubble tea shops compared to other clusters. It can also be inferred that the planning areas under cluster 1 have high customer traffic flow, since there are a great number of shops operating in those areas.

The clustering algorithm mainly segmented the central business district areas (Downtown, Orchard, Novena, Outram, Singapore River), areas that have a greater no. of commercial properties and areas that have very high residential population figures (Jurong West and Tampines). Although the housing population figures for the CBD area is low, that is because housing properties are fewer in the commercial district, it does not mean that customer traffic flow is poor. The daily office crowd would make up for the consumption figures.

Areas like Paya Lebar and Jurong East are subway interchange for passengers to change mrt trains to other parts of Singapore, hence it is a good location with decent crowd flow although population residing in those areas are binned as low.

In terms of crowd flow and possible higher customer spending power, cluster 1 is ideal if you have the confidence to operate and compete among the stronger brands that have already been operating there, especially in areas with high numbers of bubble tea stores. Or one may choose a planning area with fewer shops within cluster 1 to operate in.

Cluster 2

Cluster 2 (green markers) forms the far north planning areas of Singapore that have a mix of high and low number of residents living in the area. There are fewer number of bubble tea shops operating in cluster 2, with dessert shops being the second popular category of shops selling bubble tea. Although the size of Cluster 2 is the smallest compared to the other 2 clusters, it can be inferred the competition level is the lowest in Cluster 2 since these areas are distinctly separated from one another and from other clusters. A new entrant may benefit from the inferred low competition if it operates in cluster 2.

Cluster 3

Cluster 3 (purple markers) forms the heartland areas in Singapore that has a high number of residents compared to cluster 2. There is a mix of snacks and dessert shops as the second most popular category of shops that sells bubble tea. Interestingly, this cluster has a good combination of residential population and a healthy level of competition. Cluster 3 would be ideal for a new entrant that wants to test the market first. Residential population figures are high, indicating high and stable customer flow and one can benefit from the lower rental rates compared to majority of the areas in Cluster 1.

6. Discussion: Limitations

While we are using Foursquare to retrieve the shop information based on the demarcated Planning Areas, the number of shops collected hugely depends on Foursquare's database and their user submission of the shops.

While cleaning the data, there were outlets listed on the bubble tea brand websites but did not appear in the Foursquare API results, this potentially reduces the size of the dataset. To include more data into the dataset, I had to google for the missing individual brands websites and add them in.

Some areas of Singapore, for example One North which is about 4km away from Clementi is not marked as a Planning Area and hence not included, when in reality there are bubble tea shops operating in that area. By limiting the Foursquare query to a radius of 1.5km also leads to certain bubble tea shops that may not be included into the results. But a radius is specified to prevent to many overlapping results for the planning areas.

Planning areas like Downtown, Singapore River and National Museum are the central business district of Singapore where daily traffic is high but fewer residential properties since it is a commercial district. However, the people residing there have higher spending power since property prices in the CBD area are much more expensive compared to residents residing in the heartland areas. Hence if statistics of the average income of the residents of each Planning Area is available, the clustering results can be more robust.

7. Conclusion

For a moderately conservative new entrant, Cluster 3 will be ideal to open a shop in to test the market where there is a stable customer traffic flow and healthy level of competition.

Other factors to consider when choosing a store location includes store location accessibility and amenities around the area that can further increase crowd traffic. In addition to our cluster observations, on the brand competition level, one also has to consider competitor brands offering, their drink menu, price points and brand standing in Singapore, in evaluation to their own brand.

8. References

Sharma, Y. D. (2018, Feb). *Bubble Tea Market by Base Ingredient- Global Opportunity Analysis and Industry Forecast*. Retrieved from Allied Market Research: https://www.alliedmarketresearch.com/bubble-tea-market