## Coursera Capstone

**IBM Applied Data Science Capstone**

# Setting up a Bubble Tea Shop in a Singapore Shopping Mall


By: Cheng Seng Tan

## Introduction

Bubble tea, which started in Taiwan, has won fans worldwide, with an industry that is worth $60 billion.

A Taiwanese bubble tea brand owner is eyeing the Singapore market. He is unsure if the bubble for bubble tea in Singapore is nearing the bursting point or if there is still room for him to bring his unique bubble tea concoction to this sunny island. He is aware that major Taiwanese brands such as Koi and Gong Cha has made their forays into the Singapore market. He wants to know if there are still untapped markets in Singapore and where they are located. The investor prefers to setup his shop in an existing shopping mall where high traffic can be expected and most setup and infrastructure costs are taken care of in the shop rental.

## Business Problem

The objective of this capstone project is to analyze and select the best districts in Singapore to setup a bubble tea shop. within an existing shopping mall. It aims to provide answers to the Taiwanese investor as to the best locations for his bubble tea business. He wants to ensure the recommended location has the best chance for success and without the fierce heat from existing competitors.

This study explores all the districts in Singapore and locates the occurrences of bubble tea shops and shopping malls. Next, cluster analysis is applied to group districts with similar characteristics. Through visualizing the resulting clusters on the map, we are able to gain a clearer understanding of how bubble tea shops and shopping malls are currently distributed in Singapore.

## Target Audience

Although the target audience of this project is for the Taiwanese bubble tea investor, the study can be re-used for other ventures like restaurant business in business districts, hair salons in public housing districts, etc

## Data

We web-scrapped the list of districts (called Planning Areas) in Singapore from the Wikipedia page (https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore) The resultant table also contain area, population and population density information.

Next, we use Python Geocoder package, to obtain the longitudes and latitudes of the district. The geographical co-ordinates are needed to map the districts and also obtain venues within a determined radius.

Venue data for each district is gathered using Foursquare API. We filtered the venue categories to “Bubble Tea Shop” and “Shopping Mall” as these are the ones we are interested to analyse in this study.

# Methodology

The Wikipedia page (https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore) contains the table of Planning Areas within the broader Regions in Singapore. Planning Area is a term used by the Urban Renewal Authority (URA). I have chosen to rename it as District for clearer understanding.

```
(35, 7)
```

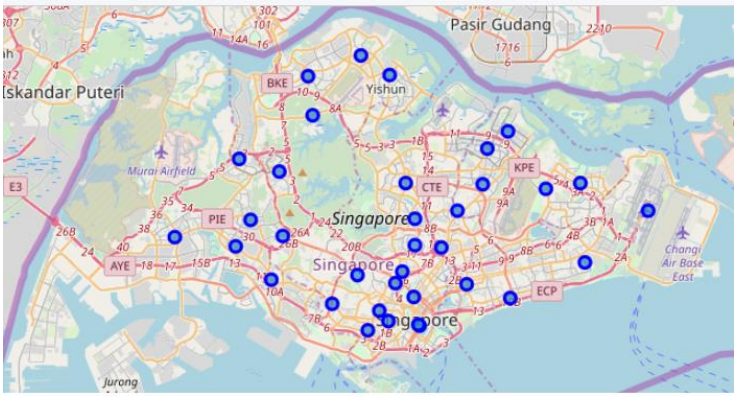
Out[7]:

| | District | Region | Area | Population | Density | Latitude | Longitude |
|---|---|---|---|---|---|---|---|
| 0 | Ang Mo Kio | North-East | 13.94 | 165710.0 | 12000.0 | 1.371610 | 103.845460 |
| 1 | Bedok | East | 21.69 | 281300.0 | 13000.0 | 1.324260 | 103.952960 |
| 2 | Bishan | Central | 7.62 | 88490.0 | 12000.0 | 1.350790 | 103.851100 |
| 3 | Bukit Batok | West | 11.13 | 144410.0 | 13000.0 | 1.349520 | 103.752770 |
| 4 | Bukit Merah | Central | 14.34 | 151870.0 | 11000.0 | 1.284170 | 103.823060 |
| 5 | Bukit Panjang | West | 8.99 | 140820.0 | 16000.0 | 1.378770 | 103.769770 |
| 6 | Bukit Timah | Central | 17.53 | 77280.0 | 4400.0 | 1.340410 | 103.772210 |
| 7 | Changi | East | 40.61 | 2080.0 | 62.3 | 1.355140 | 103.990060 |
| 8 | Choa Chu Kang | West | 6.11 | 187510.0 | 31000.0 | 1.386160 | 103.746180 |
| 9 | Clementi | West | 9.49 | 93000.0 | 9800.0 | 1.314380 | 103.765370 |
| 10 | Downtown Core | Central | 4.34 | 2510.0 | 580.0 | 1.286667 | 103.853611 |
| 11 | Geylang | Central | 9.64 | 111610.0 | 12129.0 | 1.311470 | 103.882180 |
| 12 | Hougang | North-East | 13.93 | 223010.0 | 16000.0 | 1.371140 | 103.891440 |
| 13 | Jurong East | West | 17.83 | 81180.0 | 4600.0 | 1.334370 | 103.743670 |
| 14 | Jurong West | West | 14.69 | 266720.0 | 27000.0 | 1.339490 | 103.707390 |

We used Pandas read.html to parse the data. Next, we remove unwanted columns. We also remove districts that have population less than 2000. These districts are unlikely to have the traffic required for the bubble tea business.

We next obtain the geographical coordinates in the form of latitude and longitude to map the districts and get venue information from Foursquare API. The Geocoder package is used for this purpose. On inspecting the mapped co-ordinates, one or two locations were wrongly geo-coded so I corrected them using values obtained from the internet. We now have a total of 37 districts to consider.

A map of Singapore is visualized using the Folium package with the 37 districts superimposed on top.



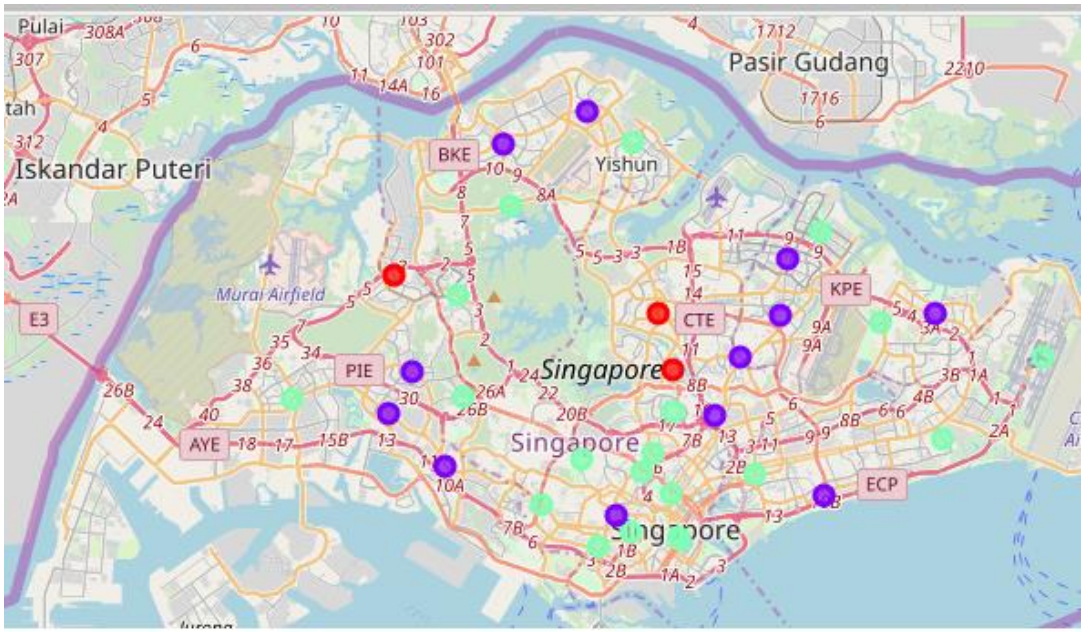

Next, we called the Foursquare API to obtain the top 100 nearby venues for each district. explore the districts. This analysis yields a total of 1230 venues with 193 unique venue categories.

To prepare for the cluster analysis, we used one hot encoding to transform the data and grouped rows by district d by taking the mean of the frequency of occurrence of each venue category. We then filtered the results to contain only “Bubble Tea Shop” and “Shopping Mall” venues. We now have 35 districts.

Finally, we ran k-means clustering to cluster the 37 districts into 3 clusters based on the frequency of occurrence of the two venues of interest. The clusters are mapped onto the Singapore district map and the visual allow us to discover underlying patterns. Details of the 3 clusters are listed for analysis of the pattern and inherent characteristics. The cluster results will allow us to identify which district have higher concentration of bubble tea shops and shopping malls while which districts have fewer number of them.

Once the “best” cluster is chosen, we can narrow down the districts in the cluster. We use a horizontal bar plot of cluster to show the districts in descending order of population density.

## Results



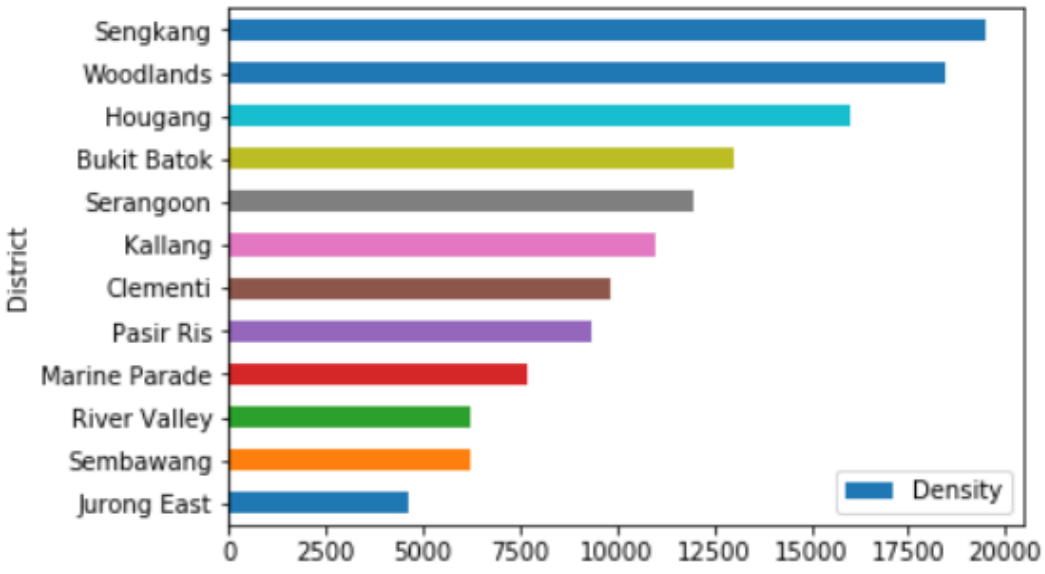

The resulting three clusters are as follows:

● Cluster 1 ● shows districts with high penetration rate for bubble tea shops and presence of shopping malls.

● Cluster 2 ● contains districts with presence of shopping malls and almost no (except for Jurong East) presence of bubble tea shops.

● Cluster 3 ● has the districts that have no or low presence of bubble tea shops. These districts are also characterised by no or few shopping malls.

We drill down the population density of each district within the preferred Cluster 2.



## Discussion

The clustering results revealed that:

Cluster 1 has the highest saturation of bubble tea shops housed within shopping malls. Intense competition will be expected in these districts.

Cluster 2 & 3 has low presence of bubble tea shop in the districts. The difference in the profiles is that Cluster 2 has the shopping malls whereas Cluster 3 has no or low presence of shopping malls. The client's preference is to setup a bubble tea shop in a high-traffic shopping mall so Cluster 3 will not be suitable.

Districts in Cluster 2 will be the most suitable area to setup a bubble tea shop within an existing shopping centre. The population density of the districts in Cluster 2 is used to infer the highest traffic flow.

I would recommend the client to setup a bubble tea shop at either Sengkang, Woodlands or Hougang. In fact, these districts are away from the Central Business District or main shopping belt and the rentals in the shopping centres will be more affordable.

The criteria for choosing the districts within the preferred cluster is based on population density in this project. This is resident population. Other factors that could influence traffic could be income level, age group of customers or location within or proximity to office buildings or train stations. Additional data is needed to factor in these criteria for a sharper conclusion.

Also, the input to the clustering algorithm is the venue information from Foursquare API which may not be complete. Some relevant locations may not be tagged in Foursquare. Perhaps data can be complemented from other review sites like Yelp and Google.

## Conclusion

This project has allowed me to use data science techniques in solving a realistic business problem.

- Web scraping of data from the internet
- Data Wrangling to prepare the data for visualisation
- Using Foursquare API to gather venue information
- Mapping using Folium
- Applying K-means clustering to group similar districts in terms of the frequency of occurrence of nearby Shopping Malls and Bubble Tea Shops.
- Plotting bar chart to visualise population density

Based on the clustering results, I recommend the districts of Sengkang, Woodlands or Hougang as ideal to open a Bubble Tea Shop in one of the existing shopping malls. The recommendation is based on evidence from data.