Introduction

Business Problem

The objective of this capstone project is to analyse and select the best location in Singapore to open a new bubble tea store. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Singapore, if franchise owners and investors are looking to open a new outlet, where would you recommend that they open it?

Target Audience

This project is particularly useful to franchise owners and investors who are looking for high potential locations to open a bubble tea store. Despite the large supply of outlets around the country, there is still a demand for the sugary beverage. As a precautionary measure due to the COVID-19 outbreak, the Ministry of Trade and Industry announced that all standalone food and beverage outlets have to be temporarily shut. Following the sudden announcement, long queues were spotted at bubble tea shops across many parts of the island as Singaporeans tried to get their last bubble tea fix.

As of June 2020, bubble tea shops in Singapore can only operate from a limited number of central kitchens. Consequently, this reduced supply has led to an even higher demand for the drink. GrabFood - a delivery service provider popular in South East Asia, offers a subscription plan exclusively for bubble tea lovers. Despite these unprecedented times, the amount of love Singaporeans have for the popular Taiwanese drink remains unwavered, indicating a strong demand for bubble tea even after we tide over the pandemic.

Data

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

- List of MRT (mass rapid transit) stations in Singapore. Each MRT station is associated with its surrounding area. This defines the scope of the project which is confined to the city of Singapore.
- Latitude and longitude coordinates of those MRT stations. This is required to plot the map and also to get the venue data.
- Venue data, particularly data related to bubble tea shops. We will use this data to perform clustering on the MRT stations.

Sources of data and methods to extract them

We will read a CSV file which contains a list of 122 MRT stations in Singapore using the pandas library. Then we will get the geographical coordinates of the train stations using the Geocoder package which will give us the latitude and longitude coordinates of the surrounding area. After that, we will use the Foursquare API to get the venue data for those train stations. The API provides many categories of the venue data, we are particularly interested in the 'Bubble Tea Shop' category in order to solve the business problem. This is a project that will make use of many data science skills, from extracting information from a CSV file, working with Foursquare API, data cleaning, data wrangling, to machine learning (k-means clustering) and map visualization (Folium).

Methodology

As Singapore is mostly accessible by public transport, we obtain the list of neighbourhoods by the train station associated with a particular area. The list of 122 MRT stations is then manually compiled and extracted. Given a list of names, we need to get the geographical coordinates in the form of latitude and longitude in order to effectively use the Foursquare API. To do so, we use the Geocoder package which converts the address of train stations to geographical coordinates. After gathering the data, we will populate the data into a pandas dataframe and then visualize the neighbourhoods in a map using the Folium package. This allows us to perform a sanity check to make sure the geographical coordinates data returned by Geocoder are correctly plotted.

Next, we will use Foursquare API to get the top 150 venues that are within a radius of 1km. Calls are then made by passing in the geographical coordinates of the MRT stations in a loop. Foursquare will return the venue data in JSON format, which is then extracted to obtain the venue name, category, latitude and

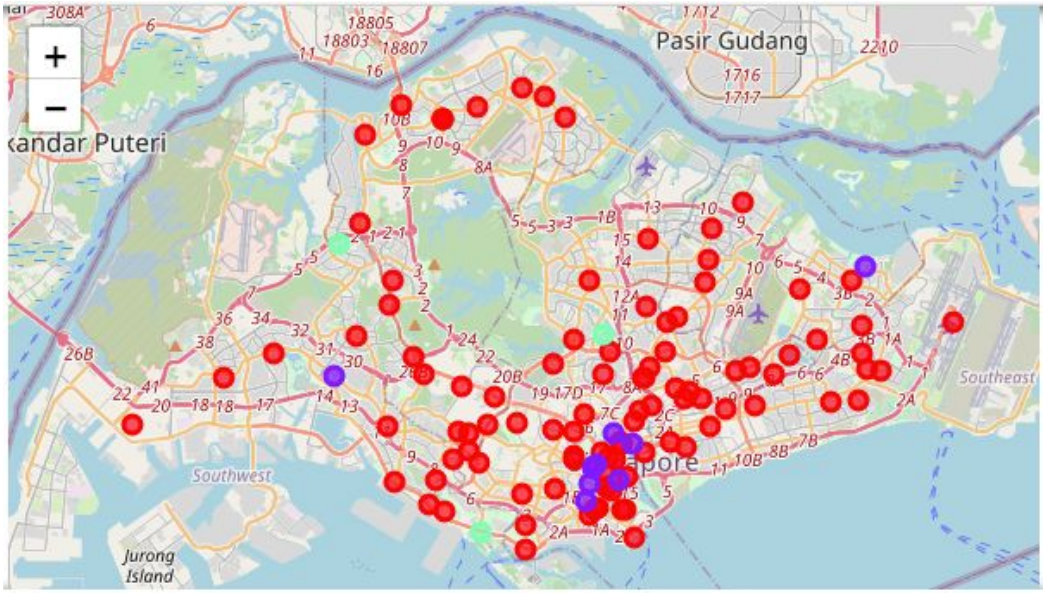
longitude. With the data, we can check how many values were returned for each neighbourhood and examine how many unique categories can be curated from all the returned values. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for clustering. Since we are analysing the 'Bubble Tea Shop' data, we will filter the 'Bubble Tea Shop' as a venue category for the neighbourhoods.

Lastly, we will perform clustering on the data by using k-means clustering. It is an algorithm which identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. This clustering method is particularly suited to solve the problem for this project. The neighbourhoods are grouped into 2 clusters based on their frequency of occurrence for 'Bubble Tea Shop'. The results will allow us to identify which neighbourhoods have higher as well as lower concentration of bubble tea outlets. Based on occurrence of stores in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new outlets.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for 'Bubble Tea Shop'

- Cluster 0 (Red): Neighbourhoods with high concentration of bubble tea outlets
- Cluster 1 (Purple): Neighbourhoods with low concentration of bubble tea outlets
- Cluster 2 (Mint): Neighbourhoods with very low concentration of bubble tea outlets



Results

As observed from the map above, most of the bubble tea stores are evenly scattered throughout Singapore, with the highest number in cluster 0. On the other hand, clusters 1 and 2 have very low to low stores in the neighbourhoods, This represents a great opportunity and high potential areas to open new outlets as there is little to no competition from existing stores. Meanwhile, outlets in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of stores. From another perspective, the results also show that the oversupply of bubble tea stores mostly happen in the central area of the city, with residential areas still having very few stores. Therefore, this project recommends franchise owners and investors to capitalize these findings and open new outlets particularly in the 3 neighbourhoods in cluster 2. Those with unique selling propositions to stand out from the competition may consider to open outlets in cluster 1. Ideally, franchise owners are advised to avoid neighbourhoods in cluster 0 which already have high concentration of bubble tea stores and suffering from intense competition.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. franchise owners and investors regarding the best locations to open a new bubble tea outlet. The neighbourhoods in cluster 2 are the most preferred locations i.e. Labrador Park, Choa Chu Kang and Bishan. The findings of this project will help stakeholders to seize opportunities on high potential locations with significantly lesser competition.