Coursera Capstone

IBM Applied Data Science Capstone

Opening a café near train stations (MRT) and shopping malls in Singapore

By: Lee Jun Yi

February 2020

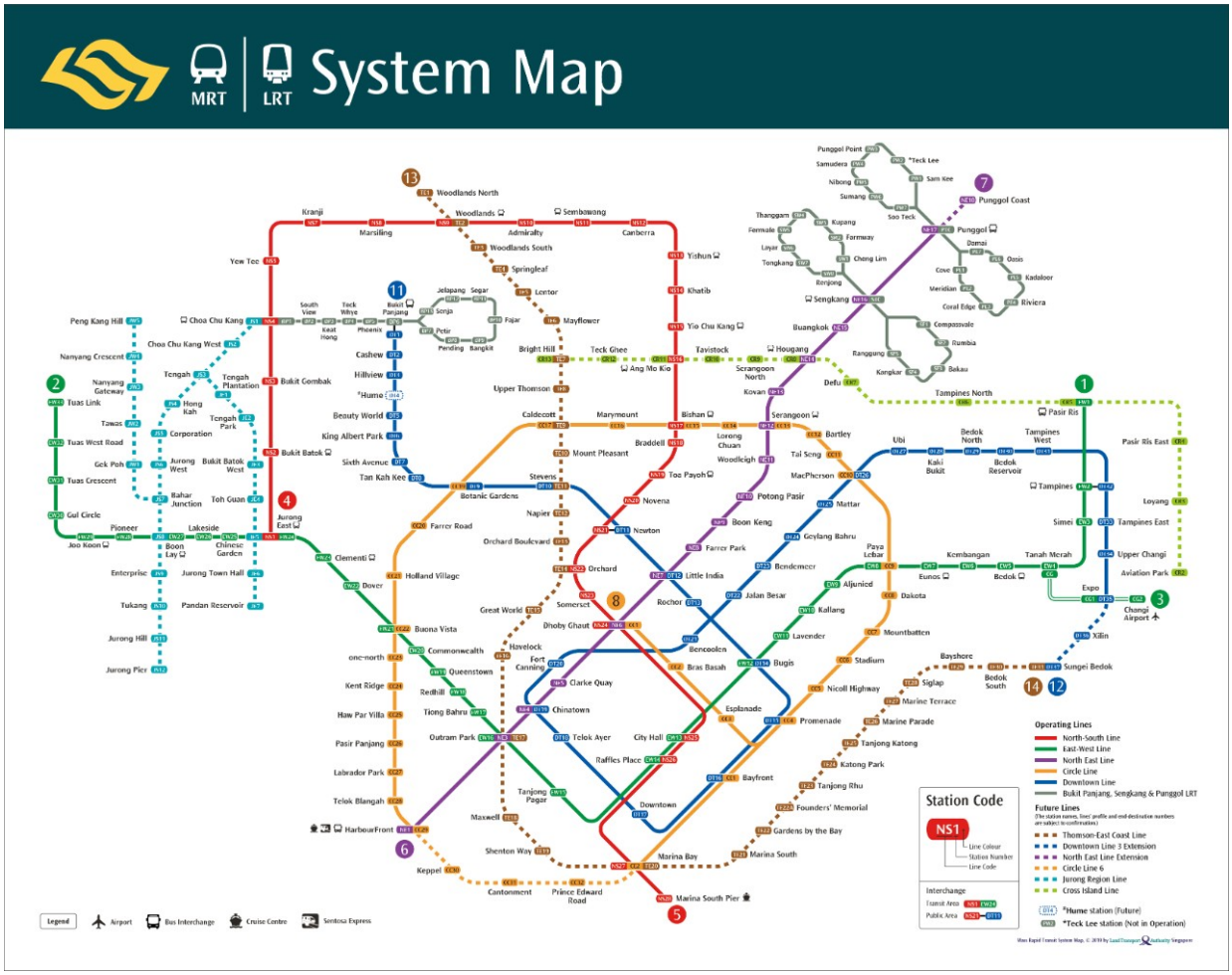


Image source : https://www.lta.gov.sg/content/ltagov/en/getting_around/public_transport/rail_network.html

Introduction

For many business owners, opening a shop is a costly decision and many of them depend on Real Estate salesperson to facilitate the search for a suitable location. However, this introduces a biased selection based on the limited number of properties that the salesperson can sell or rent.

Selecting Singapore for this analysis, it is a densely populated city with most area accessible by train and a shopping mall is located at a stone's throw every few stations away. While shopping malls provide the towns and neighborhoods a centralized location for shopping, dining, and entertainment, it naturally attracts the crowd to there to get products or services. The heavy traffic near shopping malls and stations will benefit any business daily with the large number of potential customers.

To make an informed decision by first understanding the competitors and businesses around shopping malls close to every station, business owners can save the time of researching each area and focus on the locations that satisfy their criteria. While this solution is designed to find out the businesses situated at the most convenient locations, it is highly adaptable for other uses e.g. finding eateries in all neighborhoods, concentration of cafes in every town, schools available at each area, etc.

Business Problem

The objective of this Capstone project is to enable users to select the best locations in the city of Singapore for opening a shop in the vicinity of a shopping mall beside the train station. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to the business need of finding the most suitable location with high human traffic in Singapore. In essence, the question is: where should a business owner in Singapore open a shop?

Target Audience of this project

This project is particularly useful for business owners or investors who intends to understand the business landscape and concentration around shopping malls conveniently located beside train stations. While there is no lack of stations or shopping malls, businesses struggle to demonstrate a Unique Selling Point to consumers since shopping malls can already provide a one-stop location for a wide variety of products and services. Hence, business owners or investors can utilize this project to assess whether a shop is strategically located near to (or away from) the similar businesses.

Data

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

- List of train stations in Singapore — This defines the scope of this project and the area is confined to the country Singapore in South East Asia.
- Latitude and longitude coordinates of every station — This is required in order to plot the map and also get the venue data.
- Venue data, especially data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_Singapore_MRT_stations) contains a list of train stations in Singapore, with a total of 226 stations. I will be using web scraping techniques to extract the data table from the Wikipedia page, with the help of Python's "Request" and "BeautifulSoup" packages. Then we will get the geographical coordinates of the stations using Python Geocoder package which will give us the latitude and longitude coordinates of the stations.

After that, I will be using Foursquare API to get the venue data for those stations. As we have learnt from this course, Foursquare has one of the largest database of places and is used by most developers.

Foursquare API will provide many categories of the venue data, we are particularly interested in the "Shopping Mall" category in order to help us to solve the business problem listed. This is a project that will make use of many data science skills, from web scraping (using BeautifulSoup on Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning modeling, and map visualization (Folium).

Methodology

Firstly, we need to get the list of train stations in Singapore from the Wikipedia page (https://en.wikipedia.org/wiki/List_of_Singapore_MRT_stations). We will do web scraping using Python requests and BeautifulSoup packages to extract the list of train stations data. Next, we need to get the geographical coordinates (latitude and longitude) in order to use Foursquare API. Combining with the Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude, we gather the data and populate the data into a Pandas DataFrame and visualize the businesses around each station using the Folium package. This allows us to check that the geographical coordinates data returned by Geocoder are correctly plotted on the map of Singapore.

Next, we will use Foursquare API to get the top 50 venues that are within a radius of 300 meters.

Using the Foursquare ID and Foursquare secret key from our registered Foursquare Developer Account, we make API calls to Foursquare using the geographical coordinates in a Python loop. Foursquare will return the venue data in JSON format to extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and filter unique categories.

Next, we will analyze each area near the train stations by grouping the rows by station and taking the mean of the frequency of occurrence under each venue category. By doing so, this forms the pre-processing of data for use in Clustering. Since we are analyzing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the stations.

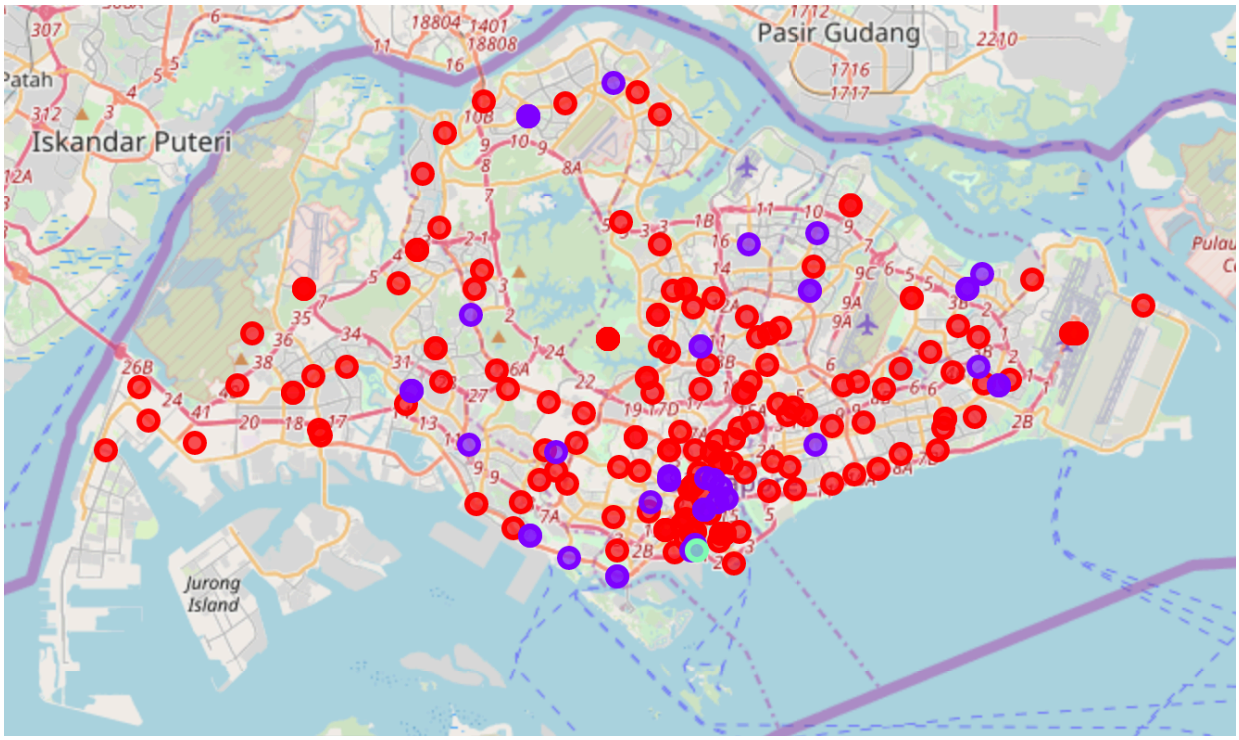
Lastly, we will perform Clustering by using k-means Clustering. K-means clustering algorithm identifies ‘k’ number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the clearest and popular unsupervised machine learning algorithms and is suitable for deriving the required data for this project to answer the business problem. We will form 3 clusters based on their frequency of occurrence for “Shopping Mall” near the stations. The results will allow us to identify the neighborhoods near train stations that have higher concentration of shopping malls, moderate concentration, or even no shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the business question for the most suitable location to open new cafes.

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 the Venue Category “Shopping Mall”:

- Cluster 0: very few to no shopping mall near the train stations
- Cluster 1: moderate number of shopping malls near the train stations
- Cluster 2: high concentration of shopping malls near the train stations

The results of the clustering are visualized in the map below with Cluster 0 in red colour, Cluster 1 in purple colour, and Cluster 2 in cyan colour.



Discussion

With reference to the map in the Results section, most of the shopping malls are concentrated in the city at the South area of Singapore, with the highest concentration of shopping malls near train stations in Cluster 2, followed by moderate number in Cluster 1. This represents a great opportunity and high potential areas to open new cafes near shopping malls, as there is heavy human traffic in the area, brought about by the convenience of being in the vicinity of the train stations and the existing malls.

However, the comparatively high concentration of shopping malls in Cluster 2 suggests that businesses located in the area may not be able to survive well due to the intense competition.

In contrast, Cluster 0 has few to no shopping mall near the train stations, signifying that the traffic is expected to be less active in the area even with the presence of a train station. However, café owners who are confident of having Unique Selling Points of their concept can consider opening their shop at these areas to establish first-mover advantage.

Limitations and opportunities for further research

In this project, we only considered the frequency of occurrence of shopping malls as the main factor. For an even more comprehensive analysis of the potential locations, there are other factors such as population and income of residents that could influence the location decision of a new cafe. Due to the limited access of data to such granularity currently, this is an opportunity for future research to be done in conjunction with surveys conducted to gather the required information. Furthermore, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned but future research could make use of paid account to overcome these limitations and obtain more extensive results for precise analysis.

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 café owners and investors to find suitable locations near train stations with shopping malls.

Answering the business problem: The neighborhoods near the train stations in Cluster 1 are the most preferred locations to open a new café.

The findings through this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shop. We can observe the higher concentration of shopping malls in the central to southern area of the country, which signifies the key office areas that coincides with the Central Business District having the heaviest traffic on average daily.

Lastly, this finding can also help business owners who have a shopfront experiencing intense competition to find alternatives for relocation in other area.