Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Chinese Restaurant in Singapore

By: Junli Wu
September 2019

1 Introduction

1.1 Background

According to WiKi Demographics of Singapore, Singapore is a multiracial and multicultural country with ethnic Chinese (76.2% of the citizen population), Malays (15.0%), and ethnic Indians (7.4%) making up the majority of the population. The 76.2% Chinese population decides that there is a great potential demanding for Chinese Restaurant. For investors, the location of the new restaurant is one of the most important decisions that will determine whether the restaurant will be a success of a failure.

1.2 Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Singapore to open a new Chinese Restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Singapore, if a property developer is looking to open a new Chinese Restaurant, where would you recommend that they open it?

2 Data Acquisition

In this project, we will need the following data:

• List of neighbourhoods in Singapore. This defines the scope of this project which is confined to the city of Singapore

• Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to Chinese Restautant. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore) contains a list of neighbourhoods in Singapore, with a total of 55 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and Beautiful Soup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Chinese Restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

3 Methodology

Firstly, we need to get the list of neighbourhoods in the city of Singapore. It is in the Wikipedia page https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore. We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Singapore.

Import the necessary libraries:

```
import numpy as np # library to handle data in a vectorized manner

import pandas as pd # library for data analsysis
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", None)

import json # library to handle JSON files

from geopy.geocoders import Nominatim # convert an address into latitude and longitude values
import geocoder # to get coordinates

import requests # library to handle requests
from bs4 import BeautifulSoup # library to parse HTML and XML documents

from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors

# import k-means from clustering stage
from sklearn.cluster import KMeans

import folium # map rendering library

print("Libraries imported.")
```

After performing the web scraping of the Wikipedia page, the data is retrieved and saved in a Pandas dataframe as below.

```
import pandas as pd
df=parse_html_table(My_table)
df.head()
```

	Name (English)	Malay	Chinese	Pinyin	Tamil	Region	Area (km2)	Population[6]	Density (/km2)
0	Ang Mo Kio		宏茂桥	Hóng mào qiáo	ஆங் மோ கியோ	North-East	13.94	165,710	12,000
1	Bedok	*	勿洛	Wù luò	பிடோக்	East	21.69	281,300	13,000
2	Bishan		碧山	Bì shān	பீஷான்	Central	7.62	88,490	12,000
3	Boon Lay		文礼	Wén lǐ	பூன் லே	West	8.23	30	3.6
4	Bukit Batok	*	武吉巴督	Wǔjí bā dū	புக்கிட் பாத்தோக்	West	11.13	144,410	13,000

Get the geo-coordinates of London city using the geopy library of Python.

	Neighborhood	Region	Latitude	Longitude
0	Ang Mo Kio	North-East	1.371610	103.845460
1	Bedok	East	1.324260	103.952960
2	Bishan	Central	1.350790	103.851100
3	Boon Lay	West	1.333330	103.700000
4	Bukit Batok	West	1.349520	103.752770
5	Bukit Merah	Central	1.283220	103.816760

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. 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 use in clustering. Since we are analysing the "Chinese Restaurant" data, we will filter the "Chinese Restaurant" as venue category for the neighbourhoods.

	Neighborhoods	ATM	Accessories Store	Airport	Airport Food Court	Airport Lounge	American Restaurant	Amphitheater	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Aust Resta
0	Ang Mo Kio	0.00	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.030000	0.000000	
1	Bedok	0.00	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.040000	0.000000	
2	Bishan	0.00	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.040000	0.000000	
3	Boon Lay	0.00	0.00	0.000000	0.000000	0.0	0.010000	0.000000	0.000000	0.00	0.00	0.00	0.060000	0.000000	
4	Bukit Batok	0.00	0.01	0.000000	0.000000	0.0	0.010000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.000000	
5	Bukit Merah	0.00	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.01	0.00	0.00	0.010000	0.000000	

Lastly, we will perform clustering on the data 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 simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Chinese Restaurant". The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of Chinese Restaurants. Based on the occurrence of Chinese Restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Chinese Restaurants.

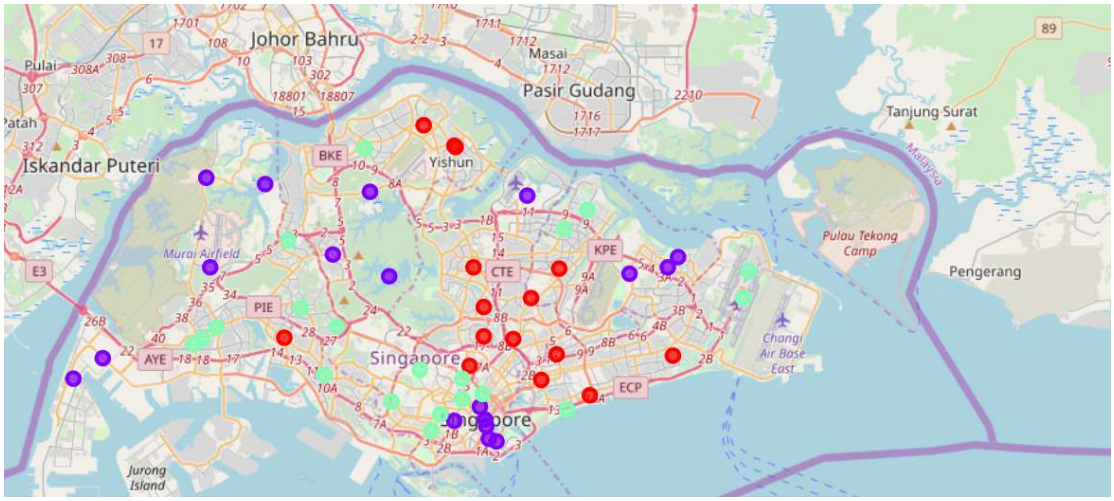
	Neighborhood	Chinese Restaurant	Cluster Labels	Region	Latitude	Longitude
0	Ang Mo Kio	0.13	0	North-East	1.37161	103.84546
1	Bedok	0.08	0	East	1.32426	103.95296
2	Bishan	0.11	0	Central	1.35079	103.85110
3	Boon Lay	0.04	2	West	1.33333	103.70000
4	Bukit Batok	0.07	2	West	1.34952	103.75277

4 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 “Chinese Restaurant”:

- Cluster 0: Neighbourhoods with high number of shopping malls
- Cluster 1: Neighbourhoods with low number to no existence of Chinese Restaurant
- Cluster 2: Neighbourhoods with moderate concentration of Chinese Restaurants

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 mint green colour.



5 Discussion

As observations noted from the map in the Results section, Chinese restaurant are scattered evenly in the different regions of Singapore city, with the highest number in cluster 0 and moderate number in cluster 2. On the other hand, cluster 1 has very low number to no Chinese Restaurant in the neighbourhoods. This represents a great opportunity and high potential areas to open new Chinese Restaurants as there is very little to no competition from existing malls. Meanwhile, Chinese Restaurants in cluster0 are likely suffering from intense competition due to oversupply and high concentration of Chinese Restaurants Therefore, this project recommends property developers to capitalize on these findings to open new Chinese Restaurants in neighbourhoods in cluster 1 with little to no competition.

Property developers with unique selling propositions to stand out from the competition can also open new Chinese Restaurants in neighbourhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 0 which already have high concentration of Chinese Restaurants and suffering from intense competition.

6 Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of Chinese Restaurants, there are other factors such as population and income of residents that could influence the location decision of a new Chinese Restaurants. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new Chinese Restaurant. In addition, 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. Future research could make use of paid account to bypass these limitations and obtain more results.

7 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. property developers and investors regarding the best locations to open a new Chinese Restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new shopping mall. The findings of 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 Chinese Restaurant.

8 References

Planning Areas of Singapore. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore

Foursquare Developers Documentation. *Foursquare*. Retrieved from https://developer.foursquare.com/docs