Coursera Capstone

IBM Applied Data Science Capstone

Finding a good location for opening a new Asian restaurant in Kaula Lumpur, Malaysia

By: Vasanta Prayaga

June 2020

Introduction

For many food lovers, people with limited cooking skills, and critics alike, restaurants are a great way to give in to those cravings, needs and search for the better respectively. With fresh taste, health and local demand in view, an Asian restaurant in Kuala Lumpur is an always a good natural choice made by the locals and mouthwatering menus will also surely lure in a lot of tourists, waiting to try the authentic cuisines of the city.

With more and more people working, eating out has become the new norm. A restaurant with a great choice of food and excellent customer service is sure to succeed. In addition to this, if it is placed in a good location which will save the customer commute, it is more likely to succeed and rake in profits from the start.

All this being said, opening a new restaurant and running it and sustaining in the business, and then earn profits at the earliest, is not an easy feat. As with any business decision, it needs a lot of serious planning and ideas along with timely decisions. One major decision for a successful restaurant is its location. It plays a very important role in the success or failure of the restaurant.

Business Problem

The objective of my capstone project is to analyze and select the best locations in the city of Kuala Lumpur, Malaysia to open a new Asian restaurant. Using Data Science methodology and Machine learning techniques like clustering, my project aims to provide options and solutions to answer the business question: **In the city of Kuala Lumpur, Malaysia, if an entrepreneur is looking to open a new Asian restaurant, which would be the best location for it?**

Target Audience of this project

This project is particularly useful to small business entrepreneurs and investors looking open or invest in a new Asian restaurant in the capital city of Malaysia, Kuala Lumpur. This project is timely as commuting to far off places for good food is not a good option any more, keeping the current corona virus travel restrictions in perspective. The stay at home orders and people working from homes, are overburdened with the additional duty of cooking and cleaning, kids and other chores. So, a warm, homely meal nearby will always be welcomed now and even when things go back to normal.

Data

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

- List of neighborhoods in Kuala Lumpur. This defines the scope of this project which is confined to the city of Kuala Lumpur, the capital city of Malaysia in South East Asia.
- Latitude and Longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to Asian restaurants. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them:

This Wikipedia page (https://en.weikipedia.org/wiki/Category:Suburbs in Kuala Lumpur) contains a list of neighborhoods in Kuala Lumpur, with a total of 70 neighborhoods. We will be using web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API to get the venue data for those neighborhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Asian 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.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Kuala Lumpur. This list is available to everyone in the Wikipedia page (https://en.weikipedia.org/wiki/Category:Suburbs in Kuala Lumpur). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitudes and longitudes in oreder to be able to use Foursquare API. Next we use the Geocoder package that will allow us to convert addresses into geopraphical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into pandas dataframe and then visualize the neighborhoods 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 Kuala Lumpur.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 meters. We need to register to Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API call to Foursquare passing in the geographical coordinated of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category venue latitudes and longitude. With the data, we can check how many venues were returned for each

neighborhood and taking the mean for the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the "Asian Restaurants" data, we will filter the "Asian Restaurants" as venue category for the neighborhoods.

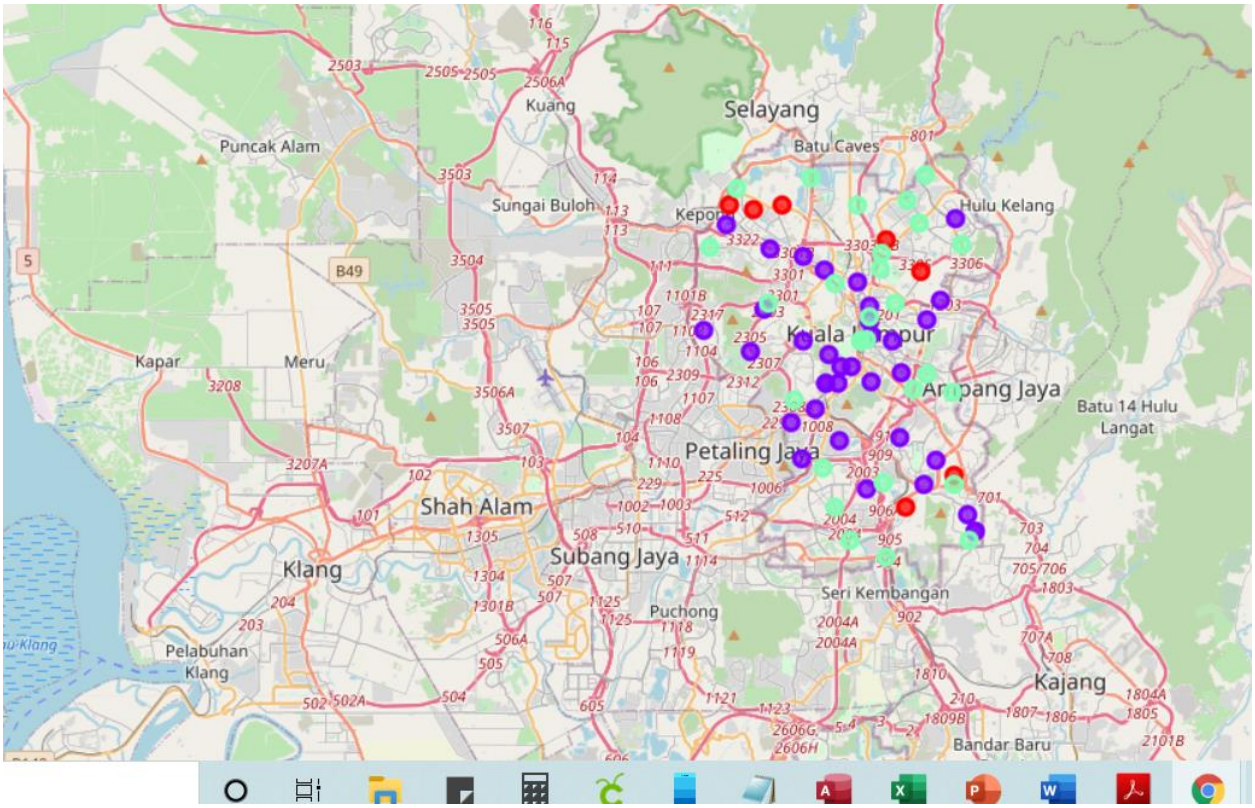
Lastly, we will perform clustering on the data by using K-means clustering algorithm, which identifies K the number of centroids, and allocates every data point ot the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into three clusters, based on their frequency of occurrence for "Asian restaurants". The results will allow us to identify which neighborhoods have higher concentration of Asian restaurants and which have fewer number of Asian restaurants. Based on the occurrence of the Asian restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Asian restaurants.

Results

The results from the K-means clustering show that we can categorize the neighborhoods into three clusters based on the frequency of occurrence for "Asian Restaurants":

- Cluster 0: Neighborhoods with least number of Asian restaurants
- Cluster 1: Neighborhoods with highest number of Asian restaurants
- Cluster 2: Neighborhoods with moderate number of Asian restaurants

The results of the clustering is visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.



Discussion

As seen from the map in the results section, most of the Asian restaurants are concentrated in the central area of Kuala Lumpur city, with the highest in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has a very low number of Asian restaurants in the neighborhoods. This represents a great opportunity and high potential area to open new Asian restaurants as there is very little competition of existing restaurants. From another perspective, the results also show that the oversupply of Asian resturants mostly happenend in the central area of the city, with the suburb area still have very few restaurants.

Therefore, this project recommends business investors to capitalize on these findings to open new Asian restaurants in neighborhoods of cluster 0 and with little competition. Business investors with unique ideas and those that are an established chain, can also venture into the neighborhoods or cluster 2. It would be wise to conclude to avoid cluster 1 neighborhoods, as it already has a high

concentration of Asian restaurants and so a demand as well as survival of more in the already intense competition is risky.

Limitations and Suggestions for Future Research

In this project we only consider one factor, which is the number of existing Asian restaurants in a neighborhood. There are other factors such as population, income of residents, interest in Asian cuisine, preference to eat outside rather than cooking at home etc. These other factors also could influence the success or failure of a restaurant.

As such data is not readily available for the neighborhood chosen for this project, there is scope for future research. Future research could come up with a methodology to identify such data, to be used in the clustering algorithm, to determine the best location for opening new Asian restaurants.

Also, as this project was implemented with the free Sandbox Tier account of Foursquare API, which has limitations to the number of API calls and results returned, there is scope in future to bypass these limitations and use more data to obtain more results.

Conclusions

In this project, the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into three clusters based on their similarities, and recommendations to different business investors and entrepreneurs was achieved.

To answer the business question that was raised in the introduction section, the answer concluded from the results achieved from the projects is: The neighborhoods in cluster 0 are the most in need of more Asian restaurants. These findings can be used by relevant business investors to make profitable investments in high demand locations, while avoiding highly over crowded locations for their business decisions for investing in opening new Asian restaurants.

References

Category: Suburbs in Kuala Lumpur. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs in Kuala Lumpur

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

Appendix

Cluster 0 Kepong Baru, Jinjang, Taman Taynton View, Bandar Tasik Selatan, Taman P.Ramlee , Taman OUG, Semarak
Cluster 1 KL Eco City, Kampung Datuk Keramat, Pantai Dalam, Pudu, Kuala Lumpur, Salak South, Segambut, Taman Bukit Maluri, Sri hartamas, Jalan Duta, Taman Desa, Taman Len Seng, Taman Midah, Taman Sri Sinar, Taman Tun Dr Ismail, Taman U-Thant, Sentul, Kuala Lumpur, Federal Hill, Kuala Lumpur, Wangsa Maju, Dang Wangi, Bandar Sri Permaisuri, Bandar Tun Razak, Bangsar, Bangsar Park, Bangsar South, Batu 11 Cheras, Brickfields, Bukit Bintang, Bukit Nanas, Bukit Petaling, Cheras, Kuala Lumpur, Bukit Kiara, Damansara, Kuala Lumpur, Damansara Town Centre, Damansara Heights, Chow kit
Cluster 2 Maluri, Taman Wahyu, Ampang, Kuala Lumpur, bandar Menjalara, Happy Garden, Jalan Cochrane, Kuala Lumpur, Taman Ibukota, Batu, Kuala Lumpur, Taman Connaught, Lembah Pantai, Sungai Besi, Kampung Padang Balang, Shamelin, Setiawangsa, Setapak, Bukit Jalil, Desa Petaling, Kepong, Titiwangsa, Putrajaya, Bukit Tunku, Kuchai Lama, Mont Kiara, Miharja, Medan Tuanku, Sri Petaling, Alam Damai