Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Restaurant in Kula Lumpur, Malaysia

By: Sachin Hulekal

Introduction

Restaurant is one of the successful business one good restaurant can attract many people all around the place, people travel around the world some people try to prefer regional food and some try to have the different taste. The southern part of India is famous for its various spices and also spicy foods. Having variety of taste and culture in food attracts many food lovers. South Indian cuisine includes the cuisines of the five southern state of India. South Indian dishes have variety for both Vegetarian and Non-Vegetarian so the people can enjoy both type of dishes. Main aim of this project to help people to open a restaurant within the city where people visit more and less competitors allows the restaurants to grow faster. Through this project allow client to know which locality suits best to open a restaurant where they can attract more people and allow them to enjoy the meal.

Business Problem

The objective of Capstone project is to analyze and select the best location in the city of Kuala Lumpur, Malaysia to open a new restaurant. Using data science and machine learning techniques like clustering, this project aims to provide solution to answer the business question: In the city of Kuala Lumpur, Malaysia, If a property developer is looking to open a new restaurant, where would you recommend that they open it?

Data

List of following data required.

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

Source of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur) contains list of neighborhoods in Kula Lumpur, with the total of 70 neighborhoods. 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 geographical coordinates of the neighborhoods using Python Geocoder packages which will give us the Latitude and Longitude coordinates of the neighborhoods.

After that, we will use the Foursquare API to get the venue data of the neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers, Foursquare API provide many categories of the venue data, we are particularly interested in the 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 where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that we used.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Kuala Lumpur. Fortunately, this list isavailable in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur)We will do web scraping using Python requests and beautiful- soup packages to extract the list of the neighborhoods data. However, this is the just a list of names. We need to get the geographical coordinates in the form of Longitude and Latitude in order to be able to use the Foursquare API. To do so we need to use Geocoder package that will allow us to convert address into geographical coordinates in the form of longitude and latitude. After gathering the data, we will populate the data into a pandas Data Frame and then visualize the neighborhoods in a map using the folium package. This allows 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.

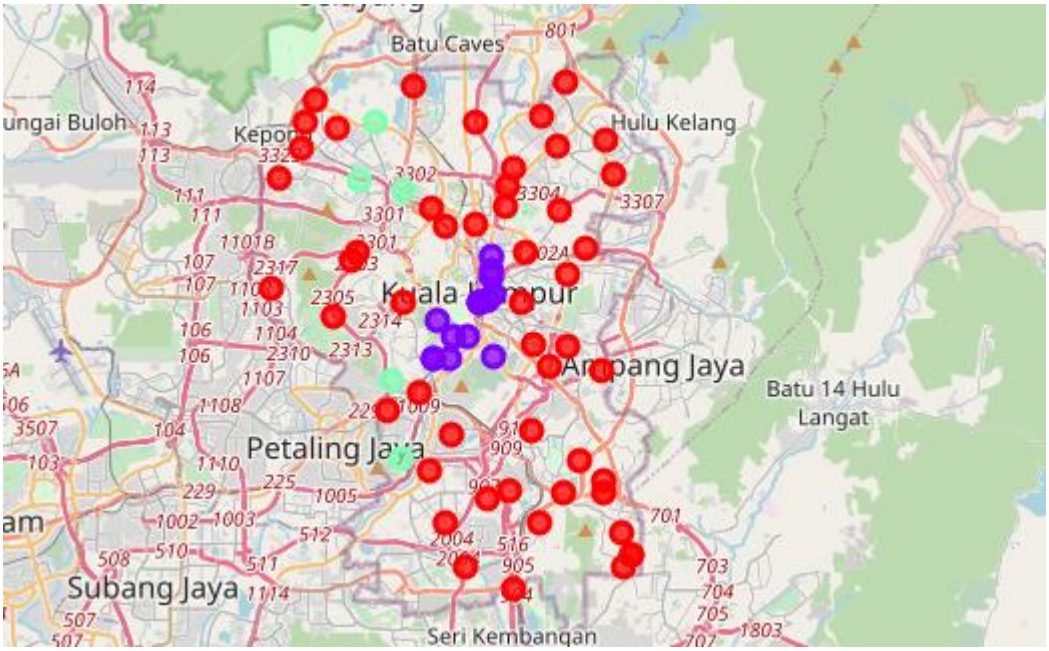
Next, we will us the Foursquare API to get the top 100 venues that the are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare Id and Foursquare secrete key. We then make an API calls to Foursquare passing the geographical coordinates of the neighborhood in a Python loop. Foursquare will return the venue data to the 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 neighborhood and examine how many unique categories can be curated from all these returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data of use in clustering. Since we are analyzing the “Restaurant” data, we will filter the “Restaurant” as venue category for the neighborhoods.

Lastly, we will perform clustering on the data by using 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 particularly suited to solve the problem for this project. We will cluster the neighborhoods in 3 cluster based on this frequency of occurrence for “Restaurant”. The result will allow us to identify which neighborhood has fewer number of Restaurant. Based on the occurrence of Restaurant in different neighborhood are most suitable to open new Restaurants.

Results

The results from the k-means clustering shows that we categorize the neighborhood into 3 clusters based on the frequency of occurrence for “Restaurants”:

- Cluster 0: Neighborhoods with the with low number of Restaurants.
- Cluster 1: Neighborhoods with the high number of Restaurants.
- Cluster 2: Neighborhoods with the moderate number of Restaurants.



The results of the clustering are visualized in the map below with the cluster 0 in red color, cluster

Discussion

An observation noted from the map in the result section, most of the restaurants are concentrated in the central area of Kula Lumpur city, with the highest number in cluster 1 and moderate number in cluster 2, On the other hand, cluster 0 has very low number of restaurants in the neighborhood. This represent a great opportunity and high potential area to open new restaurants as there is very little competition from existing restaurants, Meanwhile, restaurants in cluster 1 are likely suffering from the of intense competition due to oversupply and high concentration of restaurants. From another perspective, the result also shows the oversupply of restaurant mostly happened in the central area of the city, with the suburb area still have the few restaurant. Therefore, this project recommends property developers to capitalize on these new finding’s tom open a new restaurant in cluster 0 with the little to no competition. Property developers with the unique selling proposition to stand out from the competition can also open new restaurant in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhood in cluster 1 which already have the high concentration of restaurant and suffering the intense competition.

Limitation and Suggestion for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of restaurant, there are other factors such as population and income of residents that could influence the location decision of a new restaurant. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in clustering algorithm to determine the preferred location to open a new restaurant. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with the limitation 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 result.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing and preparing the data, performing the machine learning by clustering the data into 3 cluster based on their similarities, and lastly providing recommendation to the relevant stakeholders i.e. property developers and investors regarding the best location to open a restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhood in cluster 0 are the most prefer location to open a restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential location while avoiding overcrowded areas in their decision to open a new restaurant.