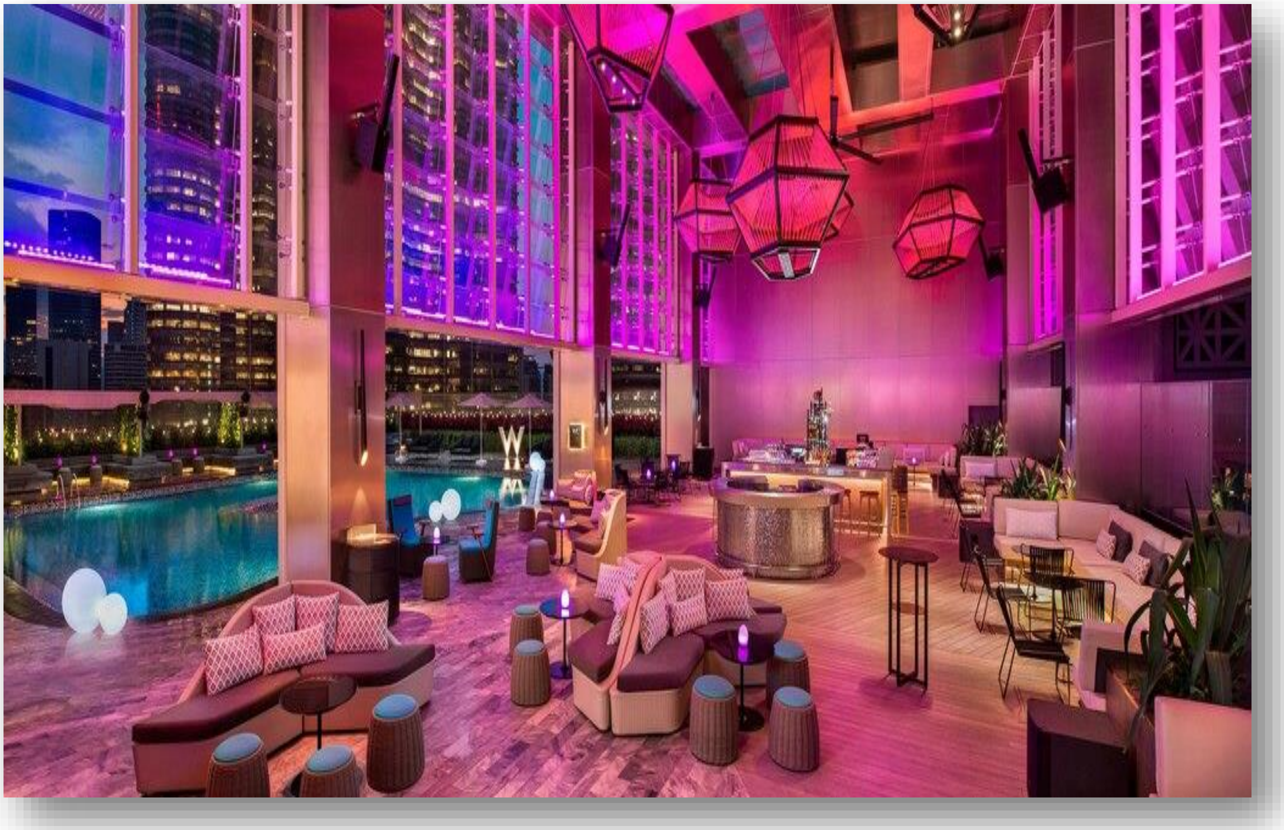
Data Science Capstone Project

Opening an New Restaurant in
Kuala Lumpur, Malaysia



By: Vishal Anand Rao

Introduction

Malaysia is a hub of tourist places with an annual footfall of about 26,100,800 in 2019. Whereas its capital city Kuala Lumpur welcomes around 11.4 million according to Mastercard in 2019. With this local businesses are growing at a rate of 6.1% in the city. With growing economy and number of tourists the restaurant business in the city is also booming and with this the search for best locations to open a new restaurant in the city and its neighborhood is also increasing rapidly. The choice of correct location plays a crucial role in success of a restaurant. A place is termed as "correct location" depending on many factors in which one of the factor is number of restaurants already there in the neighborhood of that location. A square already overcrowded with restaurants offer little scope of success of a new restaurant. Hence businessmen or investors usually prefer locations with fewer crowds of restaurants.

Business Problem

The objective of this Capstone Project is to analyze the data and select some best locations to open a new restaurant. Using Data Science methodology and Machine Learning techniques this project aims at providing solutions to answer the business problem: If a businessmen or investor is looking for some best locations to open a new restaurant in city Kuala Lumpur or its neighborhood, which places you will be recommending?

Target Audience

This project is particularly useful to businessmen and investors looking to open or invest in new restaurants in the capital city of Malaysia i.e. Kuala Lumpur. This project is timely as the city is currently suffering from oversupply of restaurants. Data from the National Property Information Centre (NAPIC) released last year showed that an additional 15 per cent will be added to existing space, and the agency predicted that total occupancy may dip below 86 per cent. The local newspaper The Malay Mail also reported in March last year that the true occupancy rates in restaurants may be as low as 40 per cent in some areas, quoting a Financial Times (FT) article cataloguing the country's continued obsession with building more restaurants despite chronic oversupply.

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 the country 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 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.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur) contains a list of neighborhoods in Kuala Lumpur, with a 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 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 database of 105+ million places and is used by over 125,000 developers. Foursquare API will 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, 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 neighbourhoods in the city of Kuala Lumpur. Fortunately, the list is available in the Wikipedia page - (https://en.wikipedia.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 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 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 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 neighborhoods 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 neighborhood and examine how many unique categories can be curated from all the 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 for use in clustering.

Since we are analyzing the “Restaurants” data, we will filter the “Restaurants” as venue category for the neighborhoods. 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 neighborhoods into 3 clusters based on their frequency of occurrence for “Restaurants”. The results will allow us to identify which neighborhoods have higher concentration of restaurants while which neighborhoods have fewer number of restaurants. Based on the occurrence of restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new restaurants.

Results

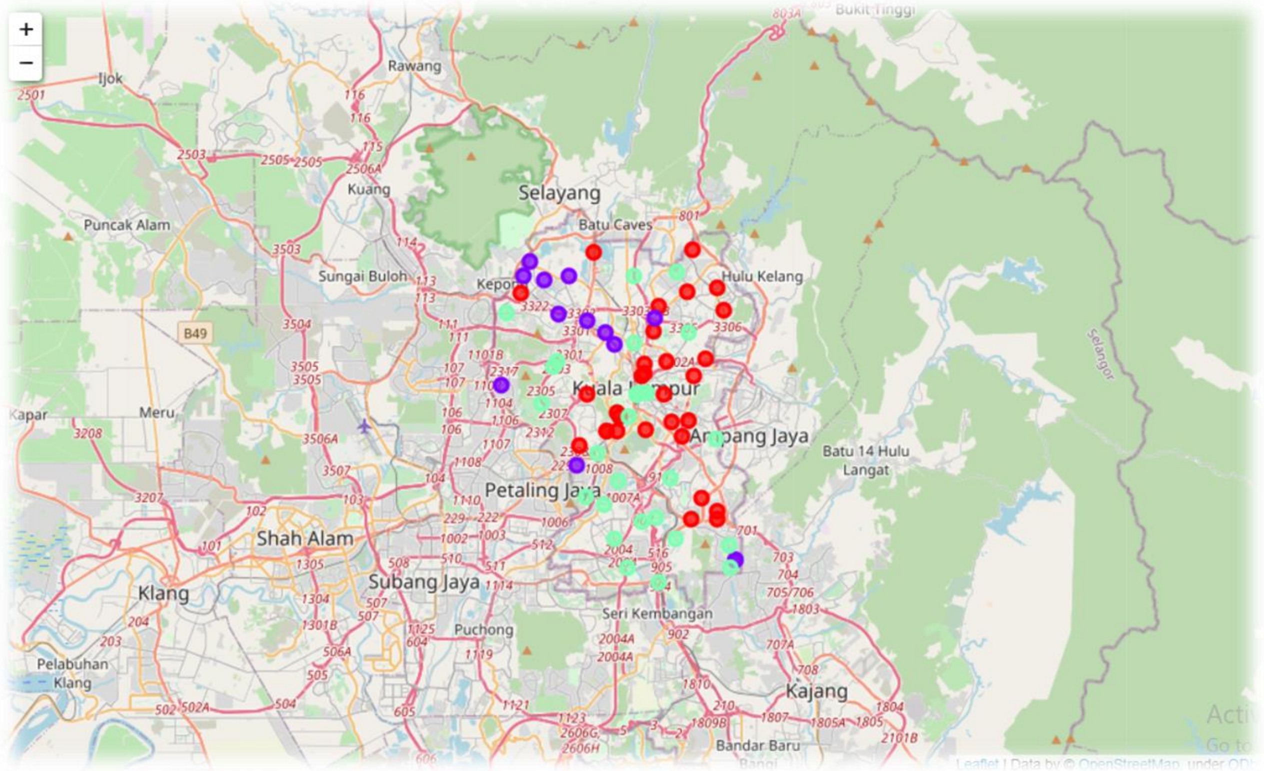
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Restaurant”:

- Cluster 0: Neighborhoods with less number of restaurants

- Cluster 1: Neighborhoods with high concentration of restaurants

- Cluster 2: Neighborhoods with moderate of 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.



Discussion

As observations noted from the map in the Results section, most of the restaurants are concentrated in the central area of Kuala 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 to no restaurants in the neighborhoods. This represents a great opportunity and high potential areas to open new restaurants as there is very little to no competition from existing restaurants.

Meanwhile, restaurants in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of restaurants. From another perspective, the results also show that the oversupply of restaurants mostly happened in the central area of the city, with the suburb area still have very few restaurants. Therefore, this project recommends property developers to capitalize on these findings to open new restaurants in neighborhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new restaurants in neighborhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 1 which already have high concentration of restaurants and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of restaurants, 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 research 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 the clustering algorithm to determine the preferred locations to open a new 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.