Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Supermarket in Kuala Lampur, Malaysia

By: Sahil Aggarwal

May 2020

Introduction

A supermarket is self-service shop offering a wide variety of food, beverages and household products, organized into sections. It is larger and has a wider selection than earlier grocery stores, but is smaller and more limited in the range of merchandise than a hypermarket or big-box market.

The supermarket typically has aisles for meat, fresh produce, dairy, and baked goods. Shelf space is also reserved for canned and packaged goods and for various non-food items such as kitchenware, household cleaners, pharmacy products and pet supplies. Some supermarkets also sell other household products that are consumed regularly, such as alcohol (where permitted), medicine, and clothes, and some sell a much wider range of non-food products: DVDs, sporting equipment, board games, and seasonal items.

There has been a rapid transformation of the food sector in developing countries, beginning in the 1990s. This applies particularly to Latin America, South-East Asia, Malaysia, China and South Africa. However, growth is being witnessed in nearly all countries. With growth, has come considerable competition and some amount of consolidation. The growth has been driven by increasing affluence and the rise of a middle class; the entry of women into the workforce; with a consequent incentive to seek out easy-to-prepare foods; the growth in the use of refrigerators, making it possible to shop weekly instead of daily; and the growth in car ownership, facilitating journeys to distant stores and purchases of large quantities of goods.

The opportunities presented by this potential have encouraged several European companies to invest in these markets (mainly in Asia) and American companies to invest in Latin America and China. Local companies also entered the market.

They are taking advantage of this trend to build supermarkets to cater to the demand. As a result, there are many supermarkets in the capital city of Malaysia and many more are being built. Opening supermarkets allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new supermarket requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the supermarkets is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Kuala Lampur, Malaysia to open a new Supermarket. 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 Kuala Lampur, Malaysia, if an organization is looking to open a new supermarket, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to investors looking to open or invest in new supermarkets in the capital city of Malaysia i.e. Kuala Lampur. This project is timely as the city is currently suffering from oversupply of supermarkets. There are places where there are a huge number of supermarkets within 2-4 Kilo meters from residential places and then there are some residential places where one has to travel a lot to reach a good supermarket.

Data

To solve the problem, we will need the following data:
• List of neighbourhoods in Kuala Lampur. This defines the scope of this project which is confined to the city of Kuala Lampur, the capital city of the country of Malaysia in South East Asia.
• 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 supermarkets. 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/Neighbourhoods_of_Kuala Lampur) contains a list of neighbourhoods in Kuala Lampur. 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 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 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 Supermarket 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 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 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 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 "Supermarket" data, we will filter the "Supermarket" as venue category for the neighbourhoods.

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 "Supermarket". The results will allow us to identify which neighbourhoods have higher concentration of Supermarkets while which neighbourhoods have fewer number of supermarkets. Based on the occurrence of Supermarkets in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Supermarkets.

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 "Supermarket":

- Cluster 0: Neighbourhoods with moderate number of Supermarkets
- Cluster 1: Neighbourhoods with low number to no existence of Supermarkets
- Cluster 2: Neighbourhoods with high concentration of Supermarkets

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.