IBM Applied Data Science Capstone Project

By LiuYuanxiu

July 2019

Opening a shopping mall in Kuala Lumpur, Malaysia

Introduction

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers.

For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, shopping malls are aplenty in Kuala Lumpur and more are being built.

Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires careful planning and is a complicated decision. Particularly, the location of the shopping is one of the key factors that will determine the success of the mall.

Business Problem

The objective of this capstone project is to analyse and select the best locations in Kuala Lumpur to open a new shopping mall. Using data science methodology and machine learning techniques, this project aims to provide solutions to answer the business question of "If a property developer is looking to open a new shopping mall, where would you recommend that they open it?"

Data

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

1. List of neighbourhoods in Kuala Lumpur – to define the scope of this project which is targeted for Kuala Lumpur, the capital city of Malaysia.
2. Latitude and longitude coordinates of these neighbourhoods – to plot the map and obtain the venue data.
3. Venue data related to shopping malls – to perform clustering on the neighbourhoods.

Sources of data and methods for extraction:

First, we pick a list of neighbourhoods in Kuala Lumpur from Wikipedia (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur), which contains 71 neighbourhoods.

We will use web scraping techniques to extract the data from Wikipedia via Python requests and Beautifulsoup packages. Next, we will obtain the geo coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

We will then use Foursquare API to obtain the venue data for these neighbourhoods. Foursquare API will provide categories of the venue data, especially categories on the shopping malls to help us solve the business problem mentioned in Part 1.

This project will make use of several data science skills from web scraping, Foursquare API, data cleaning, data wrangling, machine learning, and map visualisation.

Methodology

- First, we need to get the list of neighbourhoods in Kuala Lumpur, which is available in the Wikipedia page (please refer to above for the URL). 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 also get the geographical coordinates in the form of latitude and longitude in order to use Foursquare API. To do so, we will use the 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 visualise the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to ensure that the geographical coordinates data returned by Geocoder are correctly plotted in 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 "Shopping Mall" data, we will filter the "Shopping Mall" 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 "Shopping Mall". The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

Results

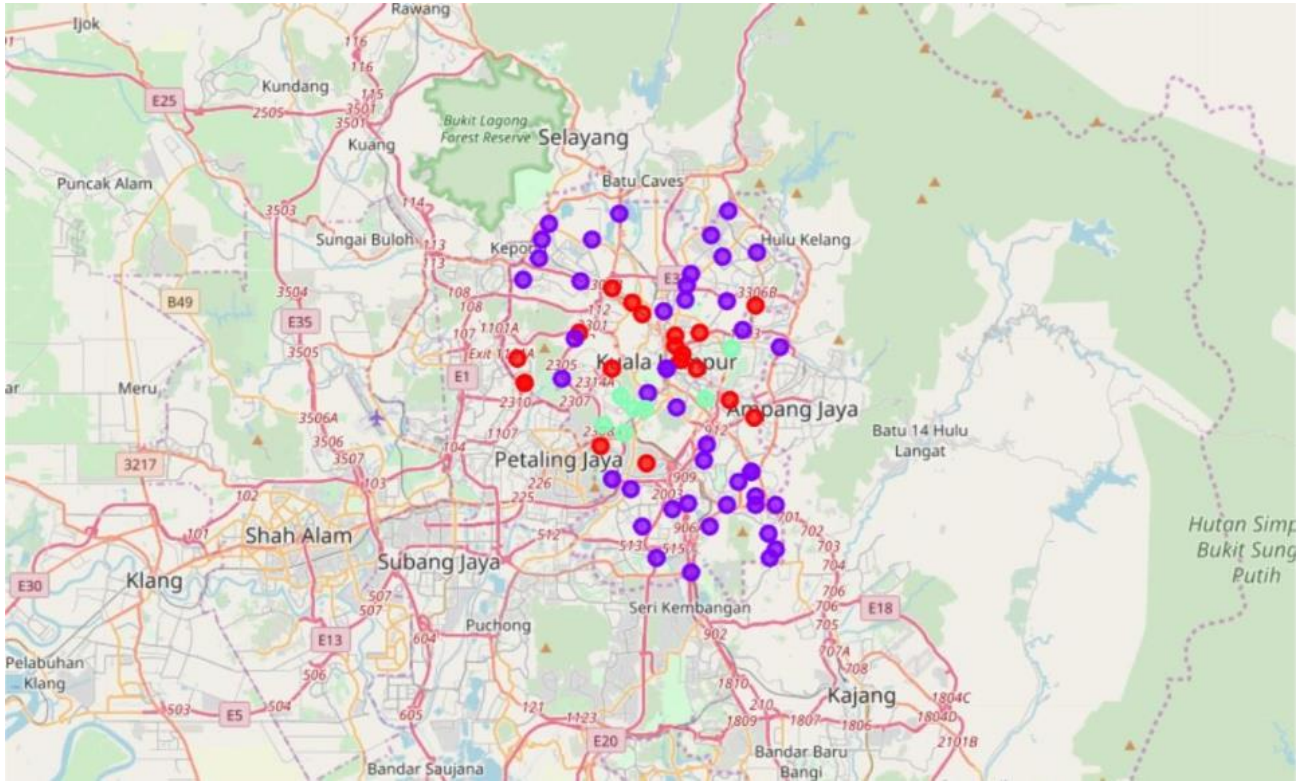
- The results from the k-means clustering show that we can categorise the neighbourhoods into 3 clusters based on the frequency of occurrence for "Shopping Mall":

✓ Cluster 0: Neighbourhoods with moderate number of shopping malls.

✓ Cluster 1: Neighbourhoods with low number to no existence of shopping malls.

✓ Cluster 2: Neighbourhoods with high concentration of shopping malls.

- The results of the clustering are visualised 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 shopping malls are concentrated in the central area of KL city, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no shopping mall in the neighbourhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few shopping malls.
- Therefore, this project recommends property developers to capitalise on these findings to open new shopping malls in neighbourhoods in cluster with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

Limitations and Suggestions for Future Research

- In this project, we only consider one factor (ie. frequency of occurrence of shopping malls). There ae other factors such as population and income of residents that could influence the location decision of a new shopping malls. However, to the best knowledge of this research 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 shopping mall.
- 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.

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 (ie. property developers and investors) regarding the best locations to open a new shopping mall. 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 most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalise on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

End of Report

Thank you.