Assignment on IBM Applied Data Science Capstone

Project Report of Data Analysis on Plan of Opening A New Shopping Mall in Kuala Lumpur, Malaysia

Prepared By: Winter Melon
Date: Dec 2020

Introduction Section

For most people, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can go to supermarket for daily supplies, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop termination 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, there are many shopping malls in the city of Kuala Lumpur and many 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 serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem Section

The objective of this capstone project is to analyze and select the best locations in the city of Kuala Lumpur, Malaysia to open a new shopping mall. Using data science methodology and machine learning techniques such as clustering, this project aims to provide solutions to answer the business question: In the city of Kuala Lumpur, Malaysia, if a property developer is planning to open a new shopping mall, where would you recommend that they open it?

This project is particularly useful for property developers and investors looking to open or invest in new shopping malls in the capital city of Malaysia i.e. Kuala Lumpur. This project is timely as the city is currently suffering from oversupply of shopping malls. Data from the National Property Information Centre (NAPIC) released last year showed that an additional 15 per cent will be added to existing mall 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 malls 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 shopping space despite chronic oversupply.

Data Section

To solve the problem, we will need the following data such as 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 shopping malls. We will use this data to perform clustering on the neighborhoods.
This Wikipedia page with the link showing as below:
(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 beautiful soup 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 neighbourhoods.

Foursquare has one of the largest databases 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 Shopping Mall 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 Section

Firstly, we need to get the list of neighborhoods in the city of Kuala Lumpur. Fortunately, the list is available in the Wikipedia page. We will do web scraping using Python requests and beautiful soup 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 Data Frame 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 "Shopping Mall" data, we will filter the "Shopping Mall" 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 "Shopping Mall". The results will allow us to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results Section

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Shopping Mall", which are shown as below:

Cluster 0: Neighborhood's with moderate number of shopping malls.

Cluster 1: Neighborhood's with low number to no existence of shopping malls.

Cluster 2: Neighborhood's with high concentration of shopping malls.

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

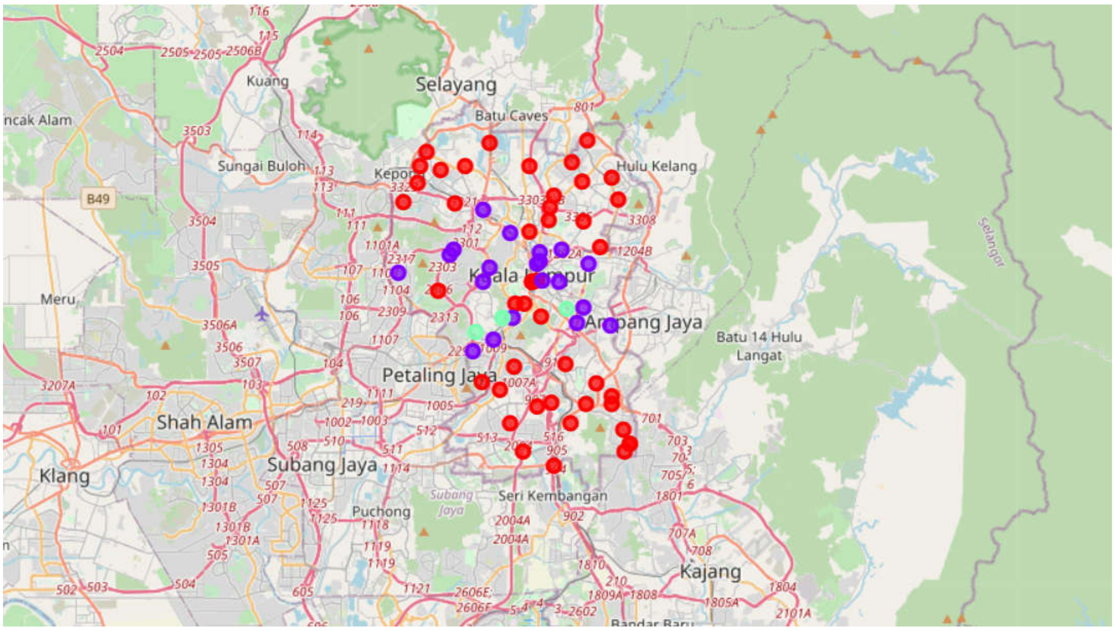

Figure 1. Results of the data analysis

Discussions Section

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Kuala Lumpur city, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number of shopping mall in the neighborhoods. 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.

Conclusions Section

Accordingly, this project recommends property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 1 with little competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

It is a good option to open a shopping mall with main business on outlets shopping as well as large entertainment place in the suburb area, owing to the lower land price for the property developer and rental for the commercial dealers. Considering more parking lots than the central area, more people will enjoy the new shopping mall built in the suburb area.