The Battle of Neighborhoods week-V Report

Project Title: Analyzing and Recommending the best locations for opening hotels in the city of Kuala Lumpur, Malaysia.

Introduction

For many tourists, booking good hotel is very tedious job for holidays. There are many websites available for hotel booking. The Hotels should provide good residential facility, nearer to airport and railways, easy access to transportation services and nearer to tourist's spots. These provide the hotels to provide their services in a better way. Property developers are also taking advantage of this trend to build more hotels to cater to the demand. As a result, there are hotels in the city of Kuala Lumpur and many more are being built. Opening hotels allows property developers to earn consistent rental income. Of course, as with any business decision, opening a hotel requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the hotel is one of the most important decisions that will determine whether the hotel will be a success or a failure.

Business Problem

The objective of this capstone project is to analyses and select the best locations in the city of Kuala Lumpur, Malaysia to hotel. 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 Lumpur, Malaysia, if a property developer is looking to open a hotel, where would you recommend that they open it?

Target Audience of this project

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

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 Hotel. 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/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 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 hotel 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 neighborhoods 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 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 coordinate's 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 analyses 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 "Hotel" data, we will filter the "Hotel" as venue category for the neighborhoods.

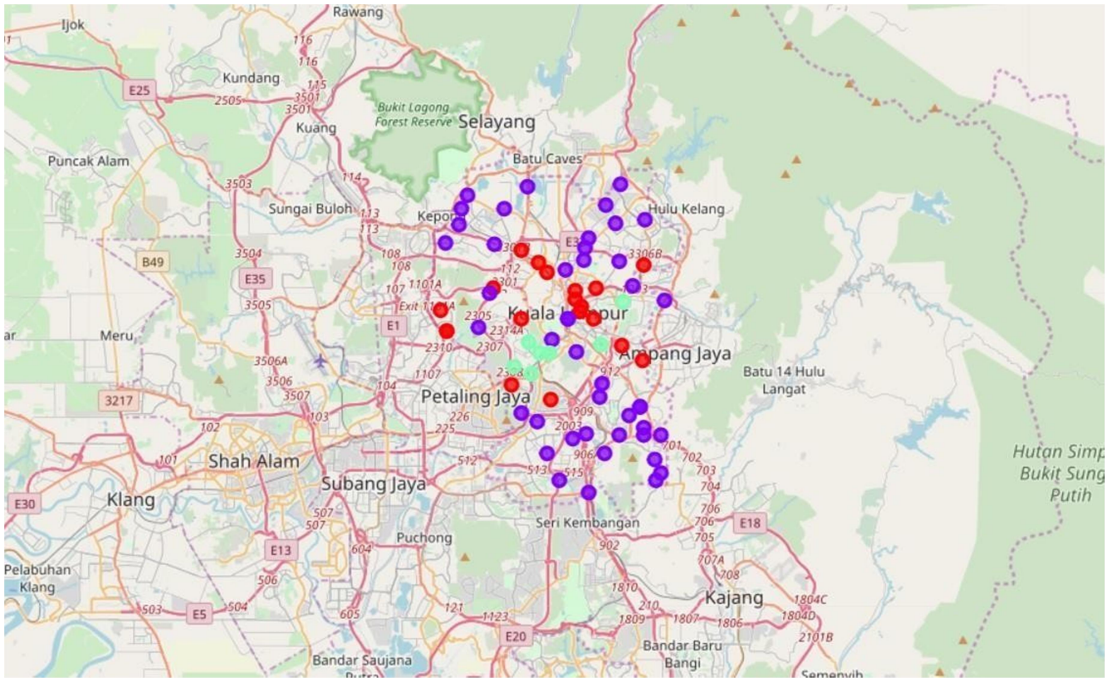
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 "Hotel". The results will allow us to identify which neighborhoods have higher concentration of hotel while which neighborhoods have fewer number of Hotel. Based on the occurrence of hotel in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new hotel.

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 “Shopping Mall”:

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

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



Discussion

As observations noted from the map in the Results section, most of the Hotel 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 to no shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new Hotel as there is very little to no competition from existing Hotel. Meanwhile, Hotel in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Hotel. From another perspective, the results also show that the oversupply of Hotel mostly happened in the central area of the city, with the suburb area still have very few Hotel. Therefore, this project recommends property developers to capitalize on these findings to open new Hotel in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new Hotel 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 Hotel 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 Hotel, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. 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 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 i.e. 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 neighborhoods in cluster 1 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

References

Category: Suburbs in Kuala Lumpur. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur

Foursquare Developers Documentation. *Foursquare*. Retrieved from https://developer.foursquare.com/docs

Appendix

Cluster 0

- Bangsar South
- Bukit Bintang
- Bukit Nanas
- Bukit Tunku
- Chow Kit
- Damansara Heights
- Damansara Town Centre
- Damansara, Kuala Lumpur
- Dang Wangi
- Jalan Cochrane, Kuala Lumpur
- Jalan Duta
- Kampung Baru, Kuala Lumpur
- Medan Tuanku
- Mont Kiara
- Segambut
- Setiawangsa
- Shamelin
- Taman Desa
- Taman Tun Dr Ismail

Cluster 1

- Alam Damai
- Ampang, Kuala Lumpur
- Bandar Menjalara
- Bandar Sri Permaisuri
- Bandar Tasik Selatan
- Bandar Tun Razak
- Batu 11 Cheras
- Batu, Kuala Lumpur
- Bukit Jalil
- Bukit Kiara
- Bukit Petaling
- Cheras, Kuala Lumpur
- Desa Petaling
- Federal Hill, Kuala Lumpur
- Happy Garden
- Jinjang
- Kampung Datuk Keramat
- Kepong
- Kuchai Lama
- Maluri
- Miharja
- Pantai Dalam
- Putrajaya
- Salak South
- Semarak
- Sentul Raya
- Setapak
- Sri Hartamas
- Sri Petaling
- Sungai Besi
- Taman Bukit Maluri
- Taman Cheras Hartamas
- Taman Connaught
- Taman Ibukota
- Taman Len Seng
- Taman Melati
- Taman Midah
- Taman OUG
- Taman P. Ramlee
- Taman Sri Sinar
- Taman Taynton View
- Taman Wahyu
- Titiwangsa
- Wangsa Maju

Cluster 2

- Bangsar
- Bangsar Park
- Brickfields
- KL Eco City
- Lembah Pantai
- Pudu, Kuala Lumpur
- Taman U-Thant