Best Locations for new mall in Kuala Lumpur, Malaysia
IBM Applied Science Capstone Project Report

Name: Surbhi Chaturvedi **Date: 8/20/2020**

Introduction

The shopping mall is a one stop shop for recreation, gastronomy, retail and everything in between. Malls have become a important part of city lives and in lot of places they provide a great deal of convenience in the immensely populated cities.

For retailers, they are a central location which is 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

The objective of this project is to analyze geographical data and select the best locations for opening a new mall in the city of Kuala Lumpur, Malaysia. 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 new shopping mall, 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 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.

Data Analysis

We will use the following data for the data analysis:

- List of neighborhoods in Kuala Lumpur.
- Latitude and longitude coordinates of those neighborhoods to plot geographical maps and coordinates of venues
- Venue data, particularly data related to shopping malls, to perform clustering on the neighborhoods.

Sources of data and methods of extraction:

This Wikipedia page (***https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur***) contains a list of neighborhoods in Kuala Lumpur, a total of 70. We will use previously learnt techniques of web scraping to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. We will get the geographical coordinates, both latitudes and longitudes of the neighborhoods using Python Geocoder package.

Foursquare API will be used to get the venue data for those neighborhoods. 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 our business problem. The project will use of many data science skills, like web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, machine learning (K-means clustering) and map visualization (Folium). In the next section, we will discuss the steps for the data analysis and machine learning technique that were used for this project.

Methodology

- The list of neighborhoods of Kuala Lumpur which is readily available through Wikipedia will be used for web scraping using python request. For this we use the beautifulsoup package and extract the list of neighborhoods in the city.
- We analyze the data to transition this list of neighborhoods into geographical coordinates. For this purpose, we use the Foursquare API and gather the latitude and longitude data. Python's Geocoder package will be used to convert the list of venues into the geographical coordinates.
- Next, we will create a dataframe using Pandas library and perform data visualization of the neighborhoods using Folium package. This will perform a sanity check and will ensure that the coordinates returned by Geocoder are correctly plotted.
- Foursquare API will be used to identify the top 100 venues that are within the radius of 2000 meters. Similar to the lab exercises, we make an API call to Foursquare and pass the geographical coordinates of the neighborhoods. Foursquare will then return a JSON file with the venue data. We will use this file to extract venue name, category, latitude and longitude. We checked how many venues are returned and then examining the number of unique categories that can be curated from the returned venues.
- Further we analyze each neighborhood by grouping them and taking the mean frequency of occurrence of each venue category. This preparation is done for the data to be used for clustering. We then filter the venue data based on category as Shopping mall for each of the neighborhoods.
- Finally, we perform clustering by using k-means algorithm. 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's one of the simplest and popular unsupervised machine learning algorithms and is well suited for this project. We will create 3 clusters of neighborhoods based on their frequency of occurrence for shopping mall. The results will help identify the neighborhoods with both higher and lower concentration of shopping malls. And this will determine which neighborhoods are best suited to open a new shopping mall.

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 shopping malls
- Cluster 1: Neighborhoods with low number to zero shopping malls
- Cluster 2: Neighborhoods with high concentration of shopping malls

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