Coursera Capstone

IBM Applied Data Science Capstone

Opening a new Shopping Mall in Kuala Lumpur

By, Navin Krishna

June 2020

Introduction

Visiting shopping mall is a relaxation activity for most of family. Also wide variety of age groups should hangout in shopping malls. There are many features such as grocery shopping, dine at restaurant and shop at various fashion outlets. For retailers, the central location and large crowd at shopping malls provide great distribution channel to market their products and services.

Bussiness Problem

The objective of this capstone project is to analyze and select the best locations in Kuala Lumpur, Malaysia to open a new shopping mall. Using Data science methodology and machine learning techniques like clustering, this project aim to provide solution to answer the business question: In the city of Kula Lumpur, Malaysia, if a property developer is looking to open a new shopping mall, where would you recommend to open it?

Target audience of this project

This project is useful for developers and investors looking to open or invest new shoppping mall in kula Lumpur.

Data

- List of neighbourhood data from wikipedia(https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur)
- Latitude and Longitude coordinates of neighbourhood from geocoder
- Venue data is extracted from Foursquare using foursquare API

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 neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longtitude in order to be able 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 data, we will populate the data into pandas DataFrame 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 ploted in the city of Kula Lumpur.

Next, we will use Foursquare API to get the top 100 venues that are with in 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 ina Python loop. Foursquare passing in the geographical coordinates of the neighbourhoods ina a python loop. Foursquare will return the venue data data in JSON format and we will extract the venue name, venue category, venue latititude and longitude. With the data, we can check how many venues were returned for each neighborhood and taking the mean of frequency of occurance 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 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 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 result 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 while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of

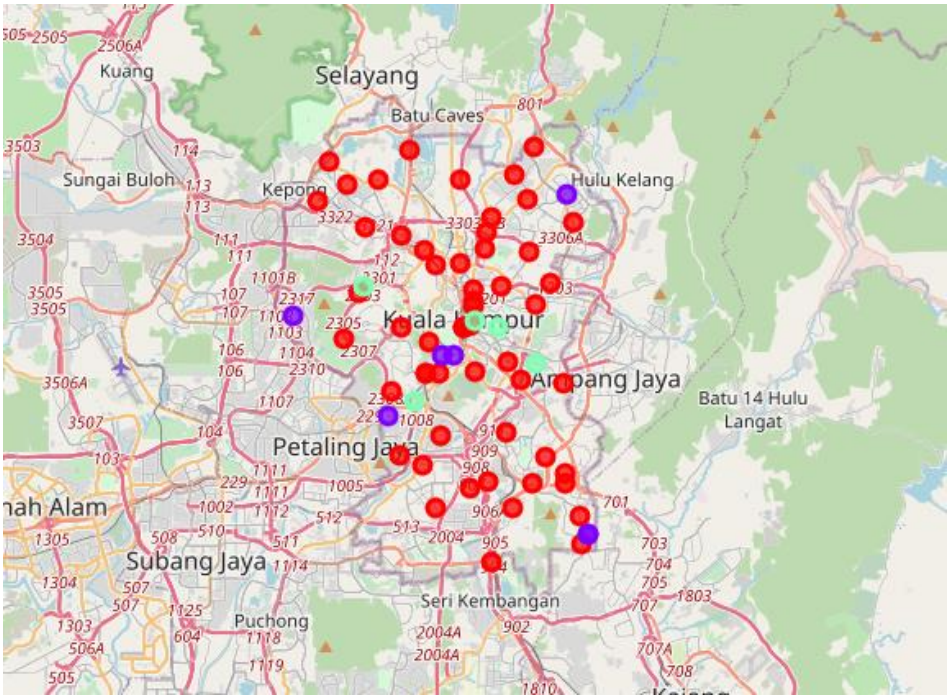
shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

Results

The results from the k-Means clustering show that we can categorize 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 of shopping malls
- Cluster 2: Neighbourhoods with high number of shopping malls

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



Disscusion

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 height number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no shipping mall in the neighbourhoods. This represents a great opportunity and high potential areas to open new shopping malls in cluster 2 are likely suffering from intense 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 captilize on these findings to open new shopping malls in neighbourhood in cluster 1 with little to no competition. Property developers will unique selling propositions to stand out from the competition. Property developers with unique selling propositions to stand out from the 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.

Conclusion

In this project, we have gone through the process of identifying the bussiness 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 neighbourhoods in cluster 1 are most preferred locations to open a new shopping mall. The finding of this project will help relevant stake holders to capitalize on the opportunities on high potential locations while avoiding overvrowed areas in their decisions to open a new shopping mall.