IBM Applied Data Science Project

Opening shopping mall at Kuala Lumpur, Malaysia

June 9th, 2020

Introduction

Shopping malls are a good investment as they are one-stop destinations for entertainment, shopping and food. This is a win-win for the retailers as they get a good distribution channel, and also for property developers who get to take advantage of it. Hence, location of a shopping mall is very important which will determine if the same will be a success or failure.

Business Problem

The objective of this project is to analyze and select the best place to open a new shopping mall in Kuala Lumpur. Using data science and machine learning, need to recommend where the shopping mall can be opened.

Data acquisition and cleaning

Data source:

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

- List of neighbourhoods in Kuala Lumpur
- Latitude and Longitude of the neighbourhoods in Kuala Lumpur.
- Venue data which gives data regarding the shopping malls

Data cleaning:

- Neighbourhoods data can be obtained using the below wiki page and can be extracted using the Beautiful soup packages in python
  https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur
- Geocoder package can be used to get the latitude and longitude information
- Foursquare website can be used to get the venue details regarding the shopping malls.

Analysis and resolution

1. Neighbourhoods data needs to be extracted from the below wiki 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 neighbourhood data.

   Using geocoder we will get the coordinates of these neighbourhoods. It can be converted to a dataframe and populated the same in maps using Folium.

Foursquare API is used to get the 100 venues within 2000 meters of radius using the foursquare id and key. We can make calls to foursquare api in a loop for getting details for various neighbourhood latitude and longitude. This data is again converted to a dataframe and populated in maps using Folium.

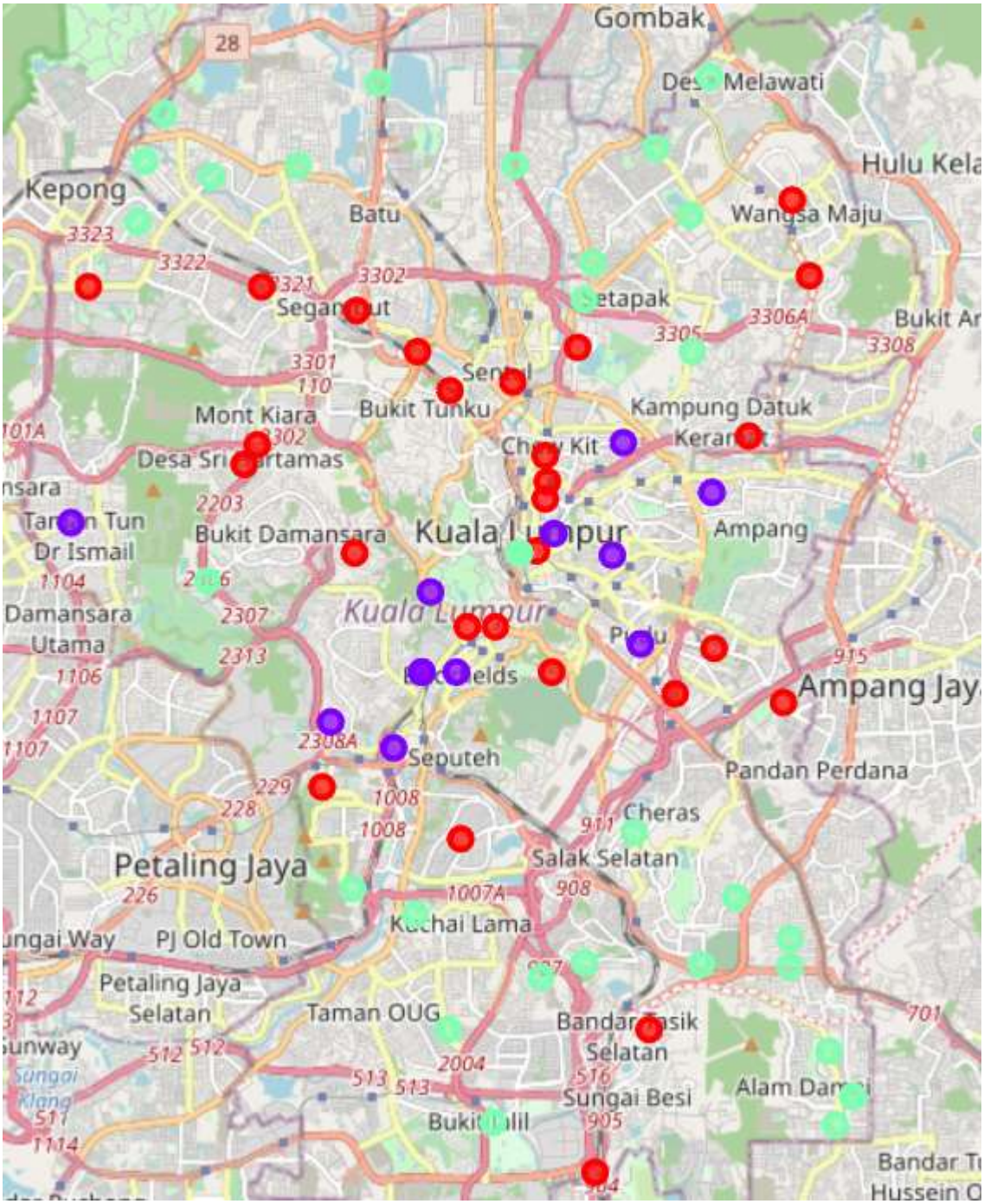
Since we are analyzing shopping malls, we will need to filter the shopping mall data. We can use K-means clustering for performing clustering on the data. It is the simplest and popular unsupervised ML algorithm and this method suits the problem in hand. Let's keep the cluster count at 3 and we will cluster the neighbourhoods based on their frequency of occurrence of shopping malls. The results will allow us to identify which neighbourhood has fewer number of shopping malls and which one has more. Based on these results, we will be able to come to a conclusion regarding the most suitable neighbourhood to open a shopping mall.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on frequency of occurrence for shopping malls.

- 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 that the cluster 0 in red colour, 1 in purple and 2 in mint green color.



Discussion

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 a very low number of totally no shopping malls 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, this also shows that the oversupply of shopping malls mostly happened

in the central area of the city, with the suburb area still having very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls 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 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.

Limitations and suggestions

At present, the results are only dependent on location and frequency of shopping malls. We can further enhance it to include parameters like population density, income of residents in that neighbourhood, etc.

Conclusion

In this project we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing data, performing machine learning by clustering of data based on similarities, and lastly providing recommendation for relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall.

Answer to the business question is:

The neighbourhood in cluster 1 is the most preferred location to open a shopping mall.