Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Shopping Mall in Kuala Lumpur, Malaysia

Introduction

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shops at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination 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 this 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. As with any business decision, opening a new shopping mall requires serious consideration and is much more complicated that it seems. In particular, the location of the shopping mall remains one of the most important decisions in determining the success or failure of the shopping mall.

Business Problem

The objective of this capstone project is to analyze the locations around the city of Kuala Lumpur with the goal of opening a new shopping mall. To accomplish this data science was used along with machine learning techniques such as clustering. This project aims to provide solutions for the business question: In the city of Kuala Lumpur, Malaysia, where would we recommend a property developer open a new shopping mall.

Target Audience

This project was done for the specific use of investors and property developers looking to invest in the city of Kuala Lumpur. The city is currently facing an oversupply of shopping malls, so this study should be particularly useful. The National Property Information Center (NAPIC) released data showing existing mall space will increase 15 percent, with total occupancy dipping below 86 percent, with local press reporting true occupancy rates as low as 40 percent. Clearly the city is facing an issue of chronic oversupply, with development showing no signs of slowing.

Data

To solve the problem, the following data is needed:

- List of neighborhoods in the city of Kuala Lumpur, which will define the scope of the project.
- Latitude and Longitude coordinates of each neighborhood, to plot the map and retrieve venue data.
- Venue data, specifically data for shopping malls, which will be used in clustering.

Sources of data and method to extract them

The Wikipedia page contains a list of the 70 neighborhoods in the city of Kuala Lumpur, Malaysia. Using web scraping through Python (beautiful soup package) we will extract the name and Geocoder to obtain the geographical coordinates (latitude and longitude). After the data is gathered, we will populate a pandas DataFrame and the visualize the neighborhoods in a map using the Folium package. This

will provide a sanity check to make sure that the Geocoder coordinates are correctly plotted in the city of Kuala Lumpur.

Next, Foursuare API will be used in the retrieval of the top 100 venues in a radius of 2,000 meters. API calls will be used to pass the geographical data, using a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, category and latitude/longitude. With this data we can check to see the venues within each neighborhood and examine the unique categories. Next, 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 clustering. Finally, we will filter the data looking specifically for shopping mall data within each neighborhood.

Clustering will b
e performed on the data by using k-means clustering, with the goal of keeping the centroids as small as possible. We will cluster the neighborhoods into three clusters based on frequency of occurrence for shopping malls. The results will allow us to identify concentrations of shopping malls across all neighborhoods, which will help us to identify which is most suitable for new investment.

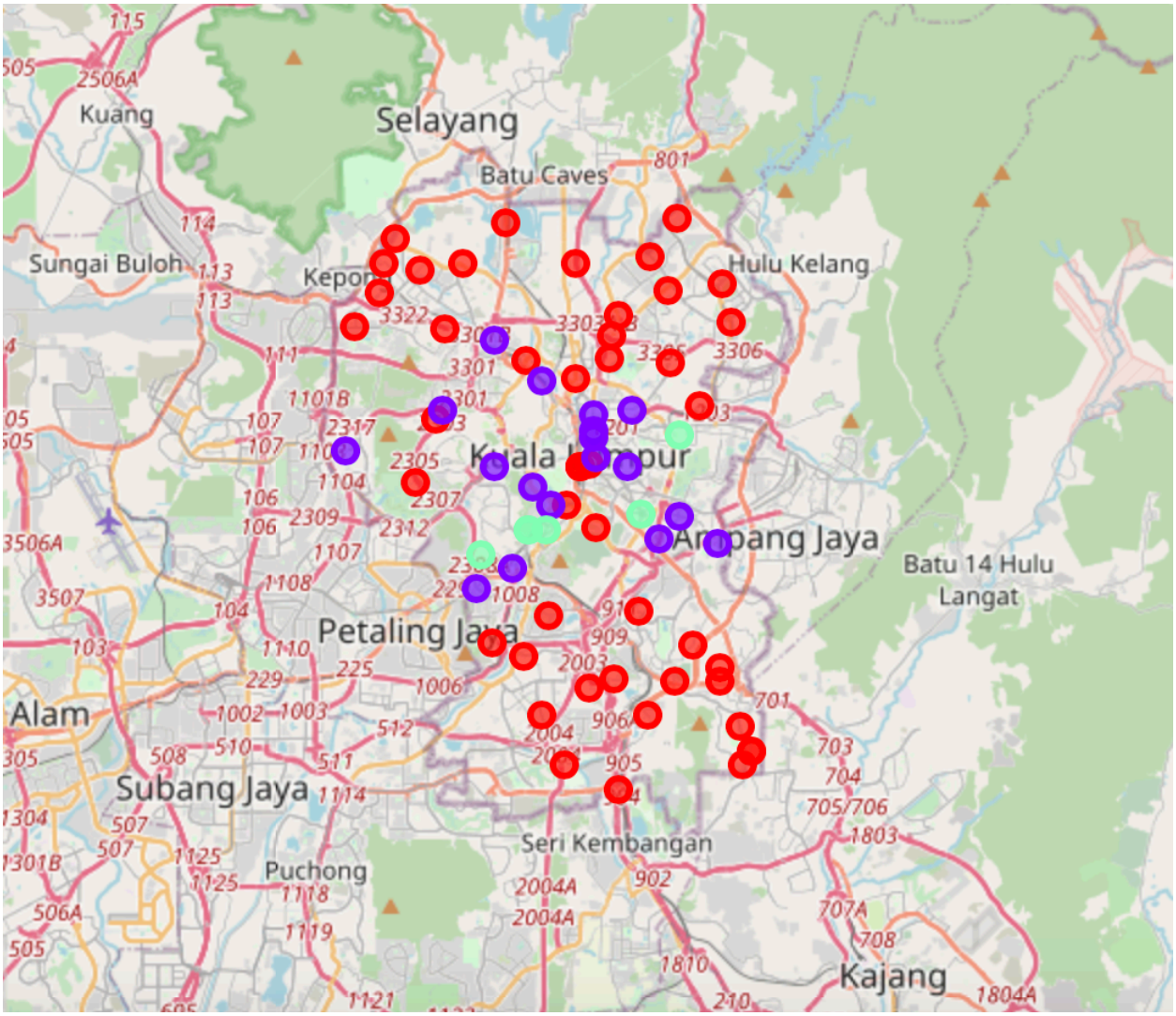
Results

The results from the k-means clustering show that we can categorize the neighborhoods of Kuala Lumpur into three clusters, based on the frequency of the occurrence of shopping malls:

- Cluster 0: Neighborhoods with moderate number of shopping malls

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

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



Results

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. A great opportunity exists within cluster 1, which has a low to no concentration meaning little to no competition. Cluster 2 is likely facing intense competition due to oversupply and high concentration. Alternatively, the results show the oversupply occurs in the center of the city, with the suburban areas having few shopping malls. Therefore, it is recommended that property developers capitalize on these findings to open new properties in cluster 1 to face little to no competition. Cluster 0 can be invested in only if the developer is offering unique opportunities for customers. Cluster 2 should be completely avoided due to the intense competition and concentration.

Limitation and Suggestions for Future Research

In this project the only factor considered was the occurrence of shopping malls. There exist other factors to be considered, such as population density and residential income. Unfortunately, this data could not be sourced on a neighborhood level. Future research could perhaps estimate this data to be used in the clustering algorithm to determine the preferred shopping mall locations.

Conclusion

In this project we have identified a business problem, specified the data required, extracted and prepared that data, performed machine learning through clustering based on similarities and finally, we provided recommendations to interested parties (property developers).

To answer the business question raised in the introduction section we determined that cluster one offers the preferred locations to open a new shopping mall. Stakeholders should be able to capitalize on the opportunities of these high potential locations while avoiding overcrowded areas in their decision to open a new shopping mall.