The Battle of the Neighborhoods – Report

1. **Introduction and Business Problem**

   Shopping Malls are a great way for many shoppers to relax themselves and find enjoyment in their spare time. Whether it is grocery shopping, dining at restaurants or browsing various retail stores, they surely have a great time as it iis a one-in-all solution for all type of shoppers. For retailers, the central location and the large footfall at the malls provide great opportunities to promote and market their products and services. Property developers are also taking advantages in this trend to build more shopping malls. As a result more and more shopping malls are being built in Kuala Lumpur to increase the number of shopping malls in the city further. Opening Shopping Malls provides property developers with consistent income from lease agreements.
   To decide where to build a new Shopping Mall a proper analysis is necessary as the location will determine its success or failure.

   Business Problem

   The objective of this report is to analyse and select the best locations in the city of Kuala Lumpur to open a new shopping Mall. Using Data Science Techniques like K-Means clustering, this projects has the objective to provide an answer to the underlying business question where a property developer shall open its new Shopping Mall.

   Target Audience

   Target Audience of this report are property developers and investors looking to open a new Shopping Mall in Kuala Lumpur.

2. **Data**

   To answer the question, we need the following data:

   - List of neighborhoods in Kuala Lumpur
   - Latitude and Longitude of the above mentioned neighborhoods of Kuala Lumpur to be able to plot the map an retrieve the venue data
   - Venue data from the Foursquare API with filter on ‘Shopping Malls’ as a basis for applying the clustering algorithm.

   Data Sources:

   The Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur) contains a listof neighborhoods in Kuala Lumpur. We will use web scraping techniques to extract the datafrom the Wikipedia page, with the help of Python requests and beautifulsoup packages. Next, we will get the geographical coordinates of the neighbourhood using Python geocoder package, which will give use the latitude and longitude information of the neighborhoods.

   After that, we will use the Foursquare API to get the venue data for each of those neighborhoods. Foursquare will provide various categories of venues, but as we’re particular interested in Shopping Malls we will filter the data accordingly.

3. Methodology

At first, we need to get the list of neighborhoods in Kuala Lumpur from the above mentioned Wikipedia page. We utilise web scraping using Python request and the beautifulsoup packages to extract the list of neighbourhood data. As this is only a list of name, we have to enrich it with the longitude and latitude information to process the data further. Hence, we use the geocoder package to convert address information into geographical coordinates in form of latitude and longitude. After pulling the data, we will populate it into a pandas dataframe and then visualize the neighborhoods in a map using the Folium package to enable us to perform a sanity check of the data.

Next, we will use the Foursquare API to get the top 100 venues within a radius of 2km. We then make the API calls to Foursquare passing the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the data in a JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues there are for each neighbourhood and examine how many unique categories can be found. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of each venue category. This way we also prepare the data for usage in the clustering. As next step we filter the data to 'Shopping Mall' only.

As last step we will perform the clustering of the data using the K-Mean algorithm. We will cluster the neighbourhood into 3 clusters based on their frequency of occurrence for 'Shopping Mall' The result will allow us to identify which neighborhoods have fewer numbers of shopping malls and which have a higher concentration. Based on the occurrence it'll help us to answer the question which neighborhood is best suited to receive a new shopping mall.

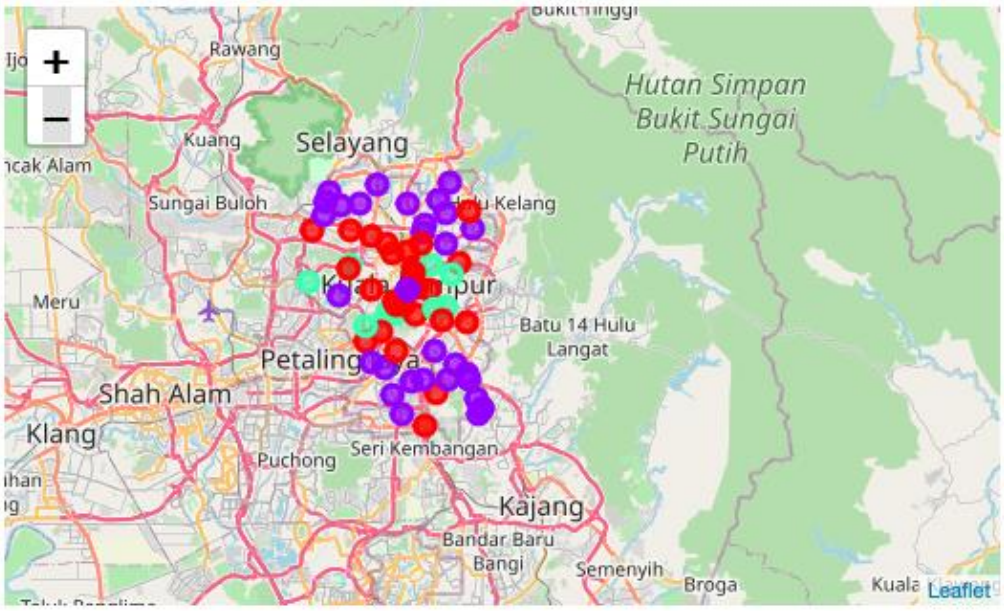
4. Result

The results from the k-means clustering show that we can cluster the neighbourhood in three parts.

- Cluster 0: Moderate number of already existing shopping malls
- Cluster 1: Low number to no existence of shopping malls
- Cluster 2: Already high concentration of shopping malls

The results are shown in the map below with cluster 0 in red, cluster 1 in purple and cluster 2 on green:

[39]:



5. **Discussion**

Most Shopping malls are already concentrated on the central area of Kuala Lumpur, with the highest number in Cluster 2 and a moderate number in cluster 0. Cluster 1 has, on the other hand, only a very low number of shopping malls in their neighbourhoods. This resembles a great opportunity for future new openings as there is almost no competition. Meanwhile shopping malls in cluster 2 are very likely to suffer from competition due to an oversupply and a high concentration.

Therefore, this project recommends to focus the neighborhoods on cluster 1 and avoid looking into cluster 2.

6. **Limitations**

This projects solely focusses on one variable: the frequency of occurrence of existing shopping malls. For further consideration of this report, it is advisable to look at more data e.g. population and income of the neighborhoods in cluster 1 whether the demographical structure of these neighborhoods confirm the findings that these areas are a great location for a new shopping mall. Also, if available, data from the existing shopping malls should be considered (e.g. revenue) to identify potential locations within cluster 2, if there are currently some struggling malls.

As this project was using the free sandbox tier account of the Foursquare API, the number of API calls was limited. Future research could switch to a premium account to bypass these limitations and obtain more results.

7. **Conclusion**

In this project we have gone through the process of identifying the business problem, specifying the needed data, extracting and preparing data, performing machine learning algorithm by clustering the data into 3 clusters, based on their similarities and lastly provided recommendations to the relevant stakeholders. The neighborhoods in cluster 1 are the most preferred locations to investigate further via adding more data points e.g. demographical data.