The Battle of Neighborhood

Introduction

Kuala Lumpur and Johor Bahru are two major cities in Malaysia. Both cities become a center of attention for residential, job employment, tourism, education, shopping and sports activity. Both cities are well known in Malaysia, and become the top choice for local and foreign communities.

Brief information about both cities:

> **Kuala Lumpur: is the national capital of Malaysia as well as its largest city. The only global city in Malaysia, it covers an area of 243 km2 (94 sq mi) and has an estimated population of 1.73 million as of 2016. Greater Kuala Lumpur, also known as the Klang Valley, is an urban agglomeration of 7.25 million people as of 2017.It is among the fastest growing metropolitan regions in South-East Asia, in both population and economic development. (source: https://en.wikipedia.org/wiki/Kuala_Lumpur)**
>
> **Johor Bahru: formerly known as Tanjung Puteri or Iskandar Puteri, is the capital of the state of Johor, Malaysia. It is situated along the Straits of Johor at the southern end of Peninsular Malaysia. Johor Bahru has a population of 497,097, while its metropolitan area, with a population of 1,638,219, is the third largest in the country. (source: https://en.wikipedia.org/wiki/Johor_Bahru)**

Objective

In this project, we will study in details the area classification using Foursquare data and machine learning segmentation and clustering. The aim of this project is to segment areas of Kuala Lumpur and Johor Bahru based on the most common places captured from Foursquare.

Using segmentation and clustering, we hope we can determine:

1. **the similarity or dissimilarirty of both cities**
2. **classification of area located inside the city whether it is residential, tourism places, or others**

Data

The data acquired from wikipedia pages and restructure to csv file for easier manipulation and reading. Another aspect to consider for this project is the Foursquare data. I believe that the data as good as provided, meaning although we are using Foursquare data for segmentation and clustering, the amount and accuracy of data captured can't 100% determine correct classification in real world.

To start, let's get and look at the data. I've already downloaded it, so let's read it (from local drive) and load it to dataframe.

Methodology

In this project, I will use the basic methodology as taught in Week 3 lab.

Above, we have done convert addresses into their equivalent latitude and longitude values.
Then we will use the Foursquare API to explore neighborhoods in both cities, Kuala Lumpur and Johor Bahru After that, explore function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters

K-means clustering algorithm will be use to complete this task. And also, the Folium library to visualize the neighborhoods in Kuala Lumpur and Johor Bahru and their emerging clusters.

Based on dataframe analysis above, we found out that Bukit Bintang area in Kuala Lumpur and Johor Bahru area in Johor Bahru are both have the highest number of area within it those district.

Discussion

Based on cluster for each cities above, we believe that classification for each cluster can be done better with calculation of venues categories (most common) in each cities. Refering to each clsuter, we can't deterimine clearly what represent in each cluster by using Foursquare - Most Common Venue data.

However, for the sae of this project we assumed each cluster as follow:

Cluster 1: Kuala Lumpur: Tourism
Cluster 2: Kuala Lumpur: Residental
Cluster 3: Kuala Lumpur: Mix
Cluster 1: Johor Bahru: Residental
Cluster 2: Johor Bahru: Tourism
Cluster 3: Johor Bahru: Sport

Conclusion

Using Foursquare API, we can captured data of common places all around the world. Using it, we refer back to our main objectives, which is to determine;

the similarity or dissimilarirty of both cities
classification of area located inside the city whether it is residential, tourism places, or others

In conclusion, both cities Kuala Lumpur and Johor Bahru are the center of attraction among Malaysian. However, to declare both cities are similar or dissimilar base on common venues visited is quite difficult. Both cities is similar in some venues also dissimilar in certain venues. And for classitification based on common venues, again we must have more systematic or quantitative way to identify and declare this. Comparison can be made, but no such method or quantitative data to determine this. We hope in the future, a method to determine it can be establish and explore for references.