Segmenting and Clustering Neighborhoods of Kuala Lumpur and Johor Bahru

Christopher Rodriguez
2020

1. 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.

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

2. Data

The data acquired from wikipedia pages and restructure to csv file for easier manipulation and reading. Both files uploaded to my github for references.

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.

3. 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.

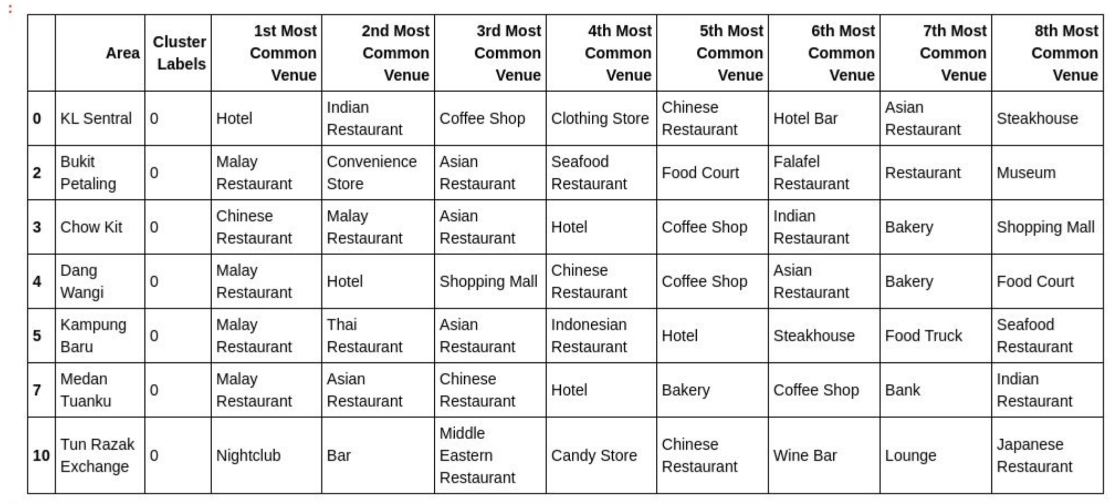
Kepong
Hulu Kelang
Kuala Lumpur
Ampang Jaya
Petaling Jaya

E22
Ulu Tiram
Tebrau
Plentong
Johor Bahru

4. Results

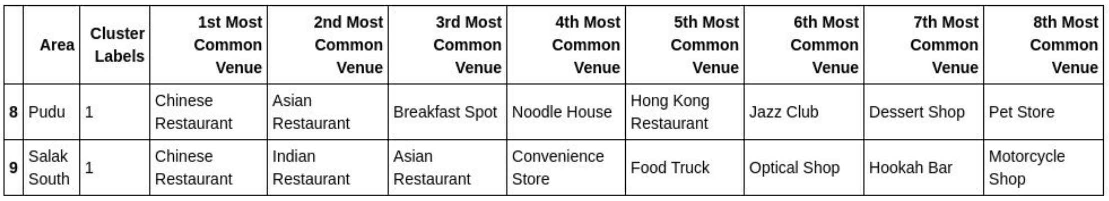
We found the next clusters for Kuala Lumpur:

Cluster 1:



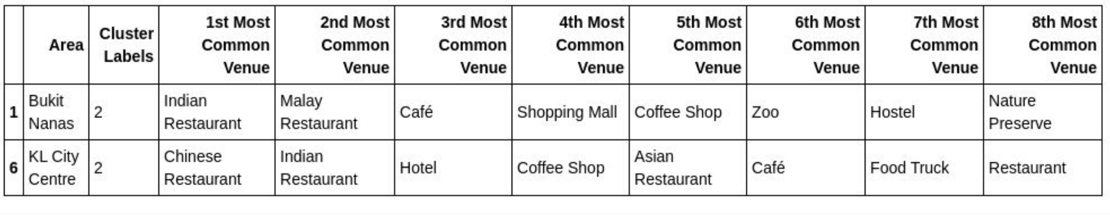
	Area	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	KL Sentral	0	Hotel	Indian Restaurant	Coffee Shop	Clothing Store	Chinese Restaurant	Hotel Bar	Asian Restaurant	Steakhouse
2	Bukit Petaling	0	Malay Restaurant	Convenience Store	Asian Restaurant	Seafood Restaurant	Food Court	Falafel Restaurant	Restaurant	Museum
3	Chow Kit	0	Chinese Restaurant	Malay Restaurant	Asian Restaurant	Hotel	Coffee Shop	Indian Restaurant	Bakery	Shopping Mall
4	Dang Wangi	0	Malay Restaurant	Hotel	Shopping Mall	Chinese Restaurant	Coffee Shop	Asian Restaurant	Bakery	Food Court
5	Kampung Baru	0	Malay Restaurant	Thai Restaurant	Asian Restaurant	Indonesian Restaurant	Hotel	Steakhouse	Food Truck	Seafood Restaurant
7	Medan Tuanku	0	Malay Restaurant	Asian Restaurant	Chinese Restaurant	Hotel	Bakery	Coffee Shop	Bank	Indian Restaurant
10	Tun Razak Exchange	0	Nightclub	Bar	Middle Eastern Restaurant	Candy Store	Chinese Restaurant	Wine Bar	Lounge	Japanese Restaurant

Cluster 2:



	Area	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
8	Pudu	1	Chinese Restaurant	Asian Restaurant	Breakfast Spot	Noodle House	Hong Kong Restaurant	Jazz Club	Dessert Shop	Pet Store
9	Salak South	1	Chinese Restaurant	Indian Restaurant	Asian Restaurant	Convenience Store	Food Truck	Optical Shop	Hookah Bar	Motorcycle Shop

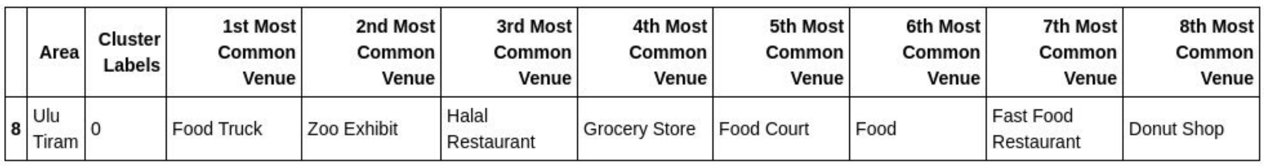
Cluster 3:



	Area	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
1	Bukit Nanas	2	Indian Restaurant	Malay Restaurant	Café	Shopping Mall	Coffee Shop	Zoo	Hostel	Nature Preserve
6	KL City Centre	2	Chinese Restaurant	Indian Restaurant	Hotel	Coffee Shop	Asian Restaurant	Café	Food Truck	Restaurant

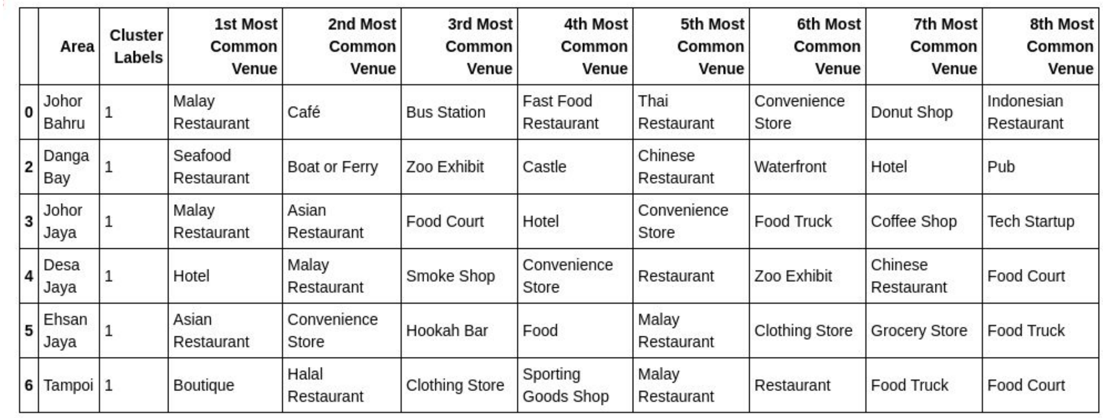
We found the next clusters for Johor Bahru:

Cluster 1:



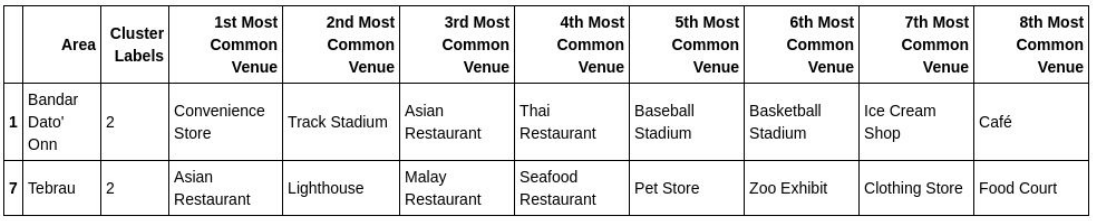
	Area	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
8	Ulu Tiram	0	Food Truck	Zoo Exhibit	Halal Restaurant	Grocery Store	Food Court	Food	Fast Food Restaurant	Donut Shop

Cluster 2:



	Area	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	Johor Bahru	1	Malay Restaurant	Café	Bus Station	Fast Food Restaurant	Thai Restaurant	Convenience Store	Donut Shop	Indonesian Restaurant
2	Danga Bay	1	Seafood Restaurant	Boat or Ferry	Zoo Exhibit	Castle	Chinese Restaurant	Waterfront	Hotel	Pub
3	Johor Jaya	1	Malay Restaurant	Asian Restaurant	Food Court	Hotel	Convenience Store	Food Truck	Coffee Shop	Tech Startup
4	Desa Jaya	1	Hotel	Malay Restaurant	Smoke Shop	Convenience Store	Restaurant	Zoo Exhibit	Chinese Restaurant	Food Court
5	Ehsan Jaya	1	Asian Restaurant	Convenience Store	Hookah Bar	Food	Malay Restaurant	Clothing Store	Grocery Store	Food Truck
6	Tampoi	1	Boutique	Halal Restaurant	Clothing Store	Sporting Goods Shop	Malay Restaurant	Restaurant	Food Truck	Food Court

Cluster 3:



	Area	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
1	Bandar Dato' Onn	2	Convenience Store	Track Stadium	Asian Restaurant	Thai Restaurant	Baseball Stadium	Basketball Stadium	Ice Cream Shop	Café
7	Tebrau	2	Asian Restaurant	Lighthouse	Malay Restaurant	Seafood Restaurant	Pet Store	Zoo Exhibit	Clothing Store	Food Court

5. 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

What is lacking at this point is a systematic, quantitative way to identify and distinguish different district and to describe the correlation most common venues as recorded in Foursquare. The reality is however more complex: similar cities might have or might not have similar common venues. A further step in this classification would be to find a method to extract these common venues and integrate the spatial correlations between different of areas or district.

We believe that the classification we propose is an encouraging step towards a quantitative and systematic comparison of the different cities. Further studies are indeed needed in order to relate the data acquired, then observe it to more meaningful and objective results.

6. 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 classification 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.