IBM Applied Data Science
Capstone Project

Opening a new restaurant in Kuala Lumpur, Malaysia

Introduction

Business Problem:

Kuala Lumpur is the cultural, financial and economic centre of Malaysia. It is among the fastest growing metropolitan regions in Southeast Asia, in both population and economic development. Tourism plays an important role in the city's service driven economy. It is driven by the city's cultural diversity, relatively low costs, and wide gastronomic and shopping variety. Businessmen are looking for opening new restaurants in Kuala Lumpur to take maximum benefit out of city's tourism.

The objective of this report is to analyse and select the best locations in Kuala Lumpur, Malaysia to open a new restaurant. This report is intended for investors and businessmen looking to open restaurants in a strategic location. In this project, we will use foursquare location data and regional clustering of venue information to determine what must be the best neighborhood in Kuala Lumpur to open a restaurant. The objective of performing analysis is to select a cluster in such a way that the investor will face very little competition from others.

Data

We considered following data to build the analytical model:

- List of neighborhoods in Kuala Lumpur. This also defines the scope of project, which is confined to the city of Kuala Lumpur, capital of Malaysia.
- Latitude and Longitude coordinates of those neighborhoods. This will be required to plot the map and get the venue data.
- The data obtained will be used to perform clustering on the neighborhoods.

The mentioned Wikipedia page: https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur contains a list of neighborhoods in Kuala Lumpur, with total of 70 neighborhoods. Web scrapping technique was used to extract the data from Wikipedia page with the help of Python requests and

BeautifulSoup packages. Geographical coordinates (latitudes and longitudes) of the neighborhoods were obtained using Python Geocoder package. Next, Foursquare API was used to obtain venue data for those neighborhoods. It provided many categories of the venue data but only restaurant category was used in this model to solve the business problem.

This project puts into implementation combination of skills learnt in data science course ranging from web scrapping (Wikipedia), working with API Foursquare, data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

1. Firstly, the list of neighbourhoods in the city of Kuala Lumpur is obtained from the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur).
2. List is obtained by web scrapping using Python requests and beautifulsoup packages.

```
# create a new DataFrame from the list
kl_df = pd.DataFrame({"Neighborhood": neighborhood_List})

kl_df.head()
```

	Neighborhood
0	Alam Damai
1	Ampang, Kuala Lumpur
2	Bandar Menjalara
3	Bandar Sri Permaisuri
4	Bandar Tasik Selatan

3. However, we obtain only list of names through web scrapping. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that converts address into geographical coordinates in the form of latitude and longitude.

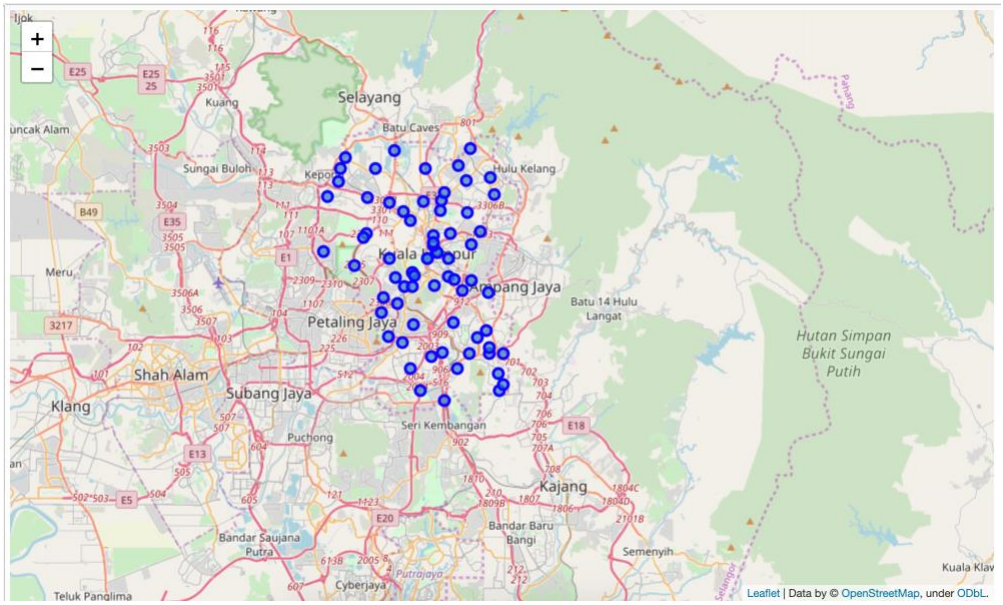
```
In [39]: # check the neighborhoods and the coordinates
         print(kl_df.shape)
         kl_df
```

```
(71, 3)
```

Out[39]:

	Neighborhood	Latitude	Longitude
0	Alam Damai	3.057690	101.743880
1	Ampang, Kuala Lumpur	3.153153	101.700413
2	Bandar Menjalara	3.190350	101.625450
3	Bandar Sri Permaisuri	3.103910	101.712260
4	Bandar Tasik Selatan	3.072620	101.714710
5	Bandar Tun Razak	3.082800	101.722810
6	Bangsar	3.129200	101.678440
7	Bangsar Park	3.134780	101.672620
8	Bangsar South	3.111020	101.662830
9	Batu 11 Cheras	3.098980	101.734990

4. After gathering the data, it is populated into a pandas data frame and the neighborhoods are visualized in a map using Folium package.



5. Next, Foursquare API was used to get the top 100 venues that are within a radius of 2000 meters. Foursquare ID and Foursquare secret key are obtained by registering a Foursquare developer account. API calls then can be made to Foursquare by passing in the geographical coordinates of the neighborhoods in a Python loop.
6. Foursquare will return the venue data in JSON format and the venue name, venue category, venue latitude and longitude will be extracted from it. With the data, we can

check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Alam Damai	3.05769	101.74388	Pengedar Shaklee Kuala Lumpur	3.061235	101.740696	Supplement Shop
1	Alam Damai	3.05769	101.74388	Machi Noodle 妈子面	3.057695	101.746635	Noodle House
2	Alam Damai	3.05769	101.74388	628火焰鑫茶室	3.058442	101.747947	Chinese Restaurant
3	Alam Damai	3.05769	101.74388	Restoran Ikbal	3.061134	101.750220	Restaurant
4	Alam Damai	3.05769	101.74388	沙巴生肉面	3.057715	101.749096	Chinese Restaurant

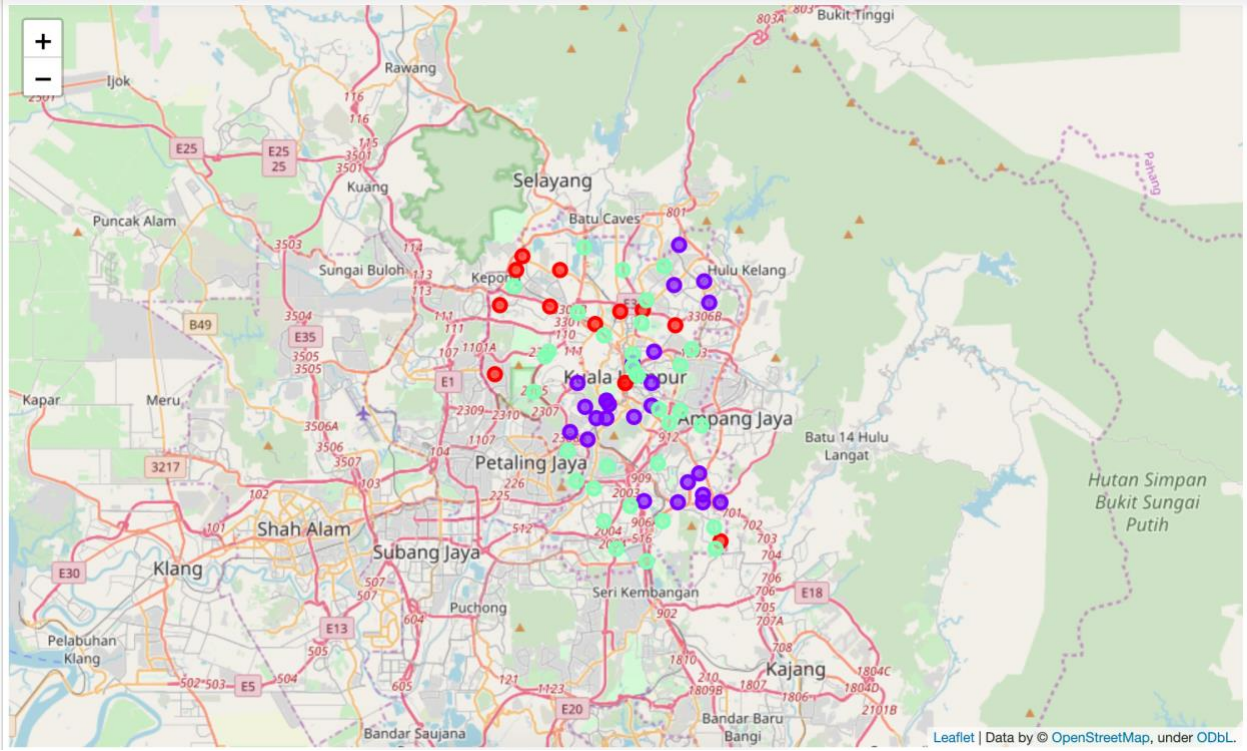
7. Then, each neighborhood is analyzed by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for the use in clustering.
8. Since we are analyzing the “Restaurants” data, we filter the “Restaurants” as venue category for the neighborhoods.
9. Lastly, clustering is performed on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.
10. Neighborhoods are clustered into 3 clusters based on their frequency of occurrence for “Restaurants”. The results allow us to identify which neighborhoods have higher concentration of restaurants and which have fewer number.
11. Based on the occurrence of restaurants in different neighborhoods, it helps us to answer the question as to which neighborhoods are most suitable to open new restaurants.

Results

The results from the k-means clustering shows that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Restaurants”:

- Cluster 0: Neighborhoods with low number of restaurants.
- Cluster 1: Neighborhoods with moderate number of restaurants.
- Cluster 2: Neighborhoods with high concentration of restaurants.

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



Cluster 0

In [67]:
```
kl_merged.loc[kl_merged['Cluster Labels'] == 0]
```

Out[67]:

	Neighborhood	Restaurant	Cluster Labels	Latitude	Longitude
48	Setapak	0.04000	0	3.188160	101.704150
18	Cheras, Kuala Lumpur	0.04000	0	3.061870	101.746750
28	Jalan Duta	0.04000	0	3.180163	101.677880
29	Jinjang	0.05000	0	3.209500	101.658740
62	Taman OUG	0.05000	0	3.210050	101.634508
64	Taman Sri Sinar	0.05102	0	3.190070	101.652930
37	Maluri	0.04000	0	3.147890	101.694050
34	Kepong	0.05000	0	3.217500	101.637630
39	Miharja	0.04000	0	3.147890	101.694050
66	Taman Tun Dr Ismail	0.06000	0	3.152830	101.622710
2	Bandar Menjalara	0.04000	0	3.190350	101.625450
46	Semarak	0.04000	0	3.179927	101.721442
47	Sentul Raya	0.05000	0	3.187431	101.691453

In [68]:
```
kl_merged.loc[kl_merged['Cluster Labels'] == 1]
```

Out[68]:

	Neighborhood	Restaurant	Cluster Labels	Latitude	Longitude
36	Lembah Pantai	0.01	1	3.121202	101.663899
55	Taman Cheras Hartamas	0.00	1	3.082630	101.746710
25	Federal Hill, Kuala Lumpur	0.00	1	3.136370	101.685640
56	Taman Connaught	0.00	1	3.082690	101.736890
31	Kampung Baru, Kuala Lumpur	0.01	1	3.165460	101.710280
30	KL Eco City	0.01	1	3.117140	101.673890
49	Setiawangsa	0.01	1	3.191803	101.740070
26	Happy Garden	0.00	1	3.201630	101.721070
38	Medan Tuanku	0.01	1	3.159260	101.698340
70	Wangsa Maju	0.01	1	3.203910	101.737190
9	Batu 11 Cheras	0.01	1	3.098980	101.734990
5	Bandar Tun Razak	0.01	1	3.082800	101.722810
6	Bangsar	0.00	1	3.129200	101.678440
7	Bangsar Park	0.00	1	3.134780	101.672620
65	Taman Taynton View	0.01	1	3.087070	101.736810
22	Damansara, Kuala Lumpur	0.01	1	3.138759	101.684046
10	Batu, Kuala Lumpur	0.01	1	3.135760	101.708370
11	Brickfields	0.00	1	3.129160	101.684060
24	Desa Petaling	0.01	1	3.083310	101.704380
16	Bukit Petaling	0.01	1	3.129290	101.698920
61	Taman Midah	0.01	1	3.093590	101.728370
60	Taman Melati	0.01	1	3.223570	101.723990
20	Damansara Heights	0.00	1	3.147980	101.667980
21	Damansara Town Centre	0.01	1	3.138759	101.684046
12	Bukit Bintang	0.01	1	3.147770	101.708550

Cluster 2

```
In [69]: kl_merged.loc[kl_merged['Cluster Labels'] == 2]
```

	Neighborhood	Restaurant	Cluster Labels	Latitude	Longitude
57	Taman Desa	0.020000	2	3.102970	101.684710
68	Taman Wahyu	0.030000	2	3.222400	101.671730
67	Taman U-Thant	0.020000	2	3.157700	101.724520
50	Shamelin	0.020000	2	3.124580	101.735970
51	Sri Hartamas	0.030000	2	3.162200	101.650360
54	Taman Bukit Maluri	0.020000	2	3.200660	101.633370
53	Sungai Besi	0.020000	2	3.050640	101.706130
58	Taman Ibukota	0.030000	2	3.212160	101.715400
63	Taman P. Ramlee	0.020833	2	3.193600	101.705980
52	Sri Petaling	0.030000	2	3.072600	101.682520
59	Taman Len Seng	0.030000	2	3.069080	101.742870
0	Alam Damai	0.030000	2	3.057690	101.743880
44	Salak South	0.020000	2	3.081020	101.697240
1	Ampang, Kuala Lumpur	0.020000	2	3.153153	101.700413
3	Bandar Sri Permaisuri	0.020000	2	3.103910	101.712260
4	Bandar Tasik Selatan	0.020000	2	3.072620	101.714710
8	Bangsar South	0.030000	2	3.111020	101.662830
13	Bukit Jalil	0.030000	2	3.057800	101.689650
14	Bukit Kiara	0.030000	2	3.143480	101.644330
15	Bukit Nanas	0.020000	2	3.152017	101.701028
17	Bukit Tunku	0.030000	2	3.173810	101.682760
45	Segambut	0.030000	2	3.186390	101.668100
19	Chow Kit	0.020000	2	3.163590	101.698110
27	Jalan Cochrane, Kuala Lumpur	0.020000	2	3.132977	101.724669
32	Kampung Datuk Keramat	0.030000	2	3.166400	101.730460
33	Kampung Padang Balang	0.020619	2	3.209430	101.693180
69	Titiwangsa	0.030000	2	3.180670	101.703220
40	Mont Kiara	0.030000	2	3.165320	101.652430
41	Pantai Dalam	0.030000	2	3.094760	101.667470
42	Pudu, Kuala Lumpur	0.020000	2	3.133540	101.713070
43	Putrajaya	0.030000	2	3.125862	101.718624
23	Dang Wangi	0.020000	2	3.156685	101.698076
35	Kuchai Lama	0.030000	2	3.090740	101.677330

Observations

Most of the restaurants are concentrated in the central area of Kuala Lumpur city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has comparatively very less number of restaurants in the neighborhoods. This represents a great opportunity and high potential area to open new restaurant as there is very little competition from the existing ones.

Meanwhile, restaurants in cluster 2 are most likely suffering from intense competition due to high concentration of restaurant options available. Therefore, this project recommends an investor to capitalize on these findings to open new restaurant in neighborhoods of cluster 0 with less competition. Investor, if having unique cuisine propositions which stands out among the competitors can also open new restaurant in neighborhoods of cluster 1 with moderate competition. Lastly, investors are advised to avoid neighborhoods in cluster 2 which already have high concentration of restaurants and suffering from intense competition.

Limitations

In this project, only one factor was considered i.e. frequency of occurrence of restaurants. There are other factors such as population and income of residents that could influence the location decision of a new restaurant which were not considered in this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new restaurant. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities of high potential locations while avoiding the overcrowded areas.