Applied Data Science Capstone

Capstone Project –
The Battle of Neighbourhoods

Selection of Potential Neighbourhoods in Kuala Lumpur for New GYM

By

Ter Boon Way

12th April 2020

Table of Contents

Table of Figures

1.Introduction

1.1. Business Problem

An investor is looking to open a new gym in Kuala Lumpur. Based on his previous experiences and marketing strategy, he would like to tap-into a mature neighbourhood with high traffic, but low competition. He has listed down the area selection criteria as shown below:

* Area that has hotels or shopping mall or residential (apartments or condo) in vicinity.

* Area that is not already crowded with gyms

To solve this problem, this project is initiated. Data science approach is utilized here to answer the investor's question, which is to locate a potential neighbourhood in Kuala Lumpur.

1.2. Stakeholders

Besides the investor who initiate this project, there are some other stakeholders who might be interested in this project:

- Those that are interested in knowing high traffic attraction areas in Kuala Lumpur
- Those that are interested in knowing the area with sports amenities in Kuala Lumpur
- Those that have interest in using the exploration result in this project

2.Data

Based on the business problem, we would require the following data in this project.

- Neighbourhood list of Kuala Lumpur
- Coordinates of each neighbourhood
- Venues in each neighbourhood

2.1. Neighbourhood list of Kuala Lumpur

First, we gather the list of neighbourhoods in Kuala Lumpur. This information is available at the link below. There are 71 neighbourhoods in Kuala Lumpur. Web scraping technique is applied here to capture these information.

https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur

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

Figure 1: First 5 rows of KL neighbourhood list

2.2. Coordinates of each neighbourhood

After getting the list of neighbourhoods in Kuala Lumpur, the next step will be to get the coordinates for each of the neighbourhood. This process is realized by geocoding library of Python.

	Neighborhood	Latitude	Longitude
0	Alam Damai	3.057690	101.743880
1	Ampang, Kuala Lumpur	3.148494	101.696729
2	Bandar Menjalara	3.190350	101.625450
3	Bandar Sri Permaisuri	3.103910	101.712260
4	Bandar Tasik Selatan	3.072750	101.714610

Figure 2: First 5 rows of KL neighbourhood list with geo coordinates

2.3. Venues in each neighbourhood

Foursquare location data is utilized to get the list of venues within 3km of radius in each neighbourhood. The information of geo-coordinates collected from previous step will be used to explore venues in each neighbourhood in Foursquare.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Main Category
0	Alam Damai	3.05769	101.74388	Pengedar Shaklee Kuala Lumpur	3.061235	101.740696	Supplement Shop	Supplement Shop
1	Alam Damai	3.05769	101.74388	Jc Deli 皆喜食坊	3.058397	101.748560	Food & Drink Shop	Food & Drink
2	Alam Damai	3.05769	101.74388	Machi Noodle 妈子面	3.057695	101.746635	Noodle House	Noodles
3	Alam Damai	3.05769	101.74388	628火焰蠡茶室	3.058442	101.747947	Chinese Restaurant	Chinese
4	Alam Damai	3.05769	101.74388	Minang Tomyam	3.057185	101.749812	Seafood Restaurant	Seafood

Figure 3: First 5 rows of venues list in each neighbourhood

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Main Category
Alam Damai	76	76	76	76	76	76	76
Ampang, Kuala Lumpur	100	100	100	100	100	100	100
Bandar Menjalara	100	100	100	100	100	100	100
Bandar Sri Permaisuri	100	100	100	100	100	100	100
Bandar Tasik Selatan	100	100	100	100	100	100	100
...	...	...	...	...	...	...	...
Taman Tun Dr Ismail	100	100	100	100	100	100	100
Taman U-Thant	100	100	100	100	100	100	100
Taman Wahyu	47	47	47	47	47	47	47
Titiwangsa	81	81	81	81	81	81	81
Wangsa Maju	100	100	100	100	100	100	100

71 rows × 7 columns

Figure 4: Count of Venues in Each Neighborhood

3. Methodology

Below are the methodologies used in this project:

- Pre-processing
- Clustering
- Cluster visualization on map
- Examine the cluster

3.1. Pre-processing

Firstly, the data collected is pre-processed.

3.1.1. Data Filtering

There are a lot of venues categories returned from Foursquare location data. The interest here is to look for venues categories that contribute to traffic (which are shopping mall, hotel, residential) as well as competition (gym and sports amenities). Thus, the data is filtered to only these venue categories:

- Gym
- Sports
- Shopping mall
- Hotel
- Residential

3.1.2. One-hot encoding

One hot encoding is carried out to convert the categorical variables into a form that could be provided to machine learning algorithms for analysis.

	Neighborhood	Athletics & Sports	Boxing Gym	Gym	Gym / Fitness	Hotel	Mall	Residential	Sports Club
0	Ampang, Kuala Lumpur	0	0	0	0	1	0	0	0
1	Ampang, Kuala Lumpur	0	0	1	0	0	0	0	0
2	Ampang, Kuala Lumpur	0	0	1	0	0	0	0	0
3	Ampang, Kuala Lumpur	0	0	0	0	1	0	0	0
4	Ampang, Kuala Lumpur	0	0	0	0	1	0	0	0

Figure 5 One-hot encoding

3.1.3. Group-By Analysis

The data collected at this stage could be seen as a table consists of information for each venue. Each row is for one venue. There are total of 428 rows (428 venues).

Group-By analysis groups the row by neighbourhoods. Two tables are created, showing the sum of venues category and average of venues category in each neighbourhoods respectively.

	Neighborhood	Athletics & Sports	Boxing Gym	Gym	Gym / Fitness	Hotel	Mall	Residential	Sports Club
0	Ampang, Kuala Lumpur	0	0	2	0	9	0	0	0
1	Bandar Menjalara	0	0	3	0	0	1	0	0
2	Bandar Sri Permaisuri	1	0	1	0	0	0	2	0
3	Bandar Tasik Selatan	1	0	1	1	1	1	1	0
4	Bandar Tun Razak	0	0	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...
63	Taman Tun Dr Ismail	0	0	0	2	1	3	0	0
64	Taman U-Thant	0	0	0	2	12	2	0	0
65	Taman Wahyu	0	0	0	0	1	0	1	0
66	Titiwangsa	1	0	1	0	2	0	0	0
67	Wangsa Maju	0	0	2	3	0	1	0	0

Figure 6: Sum of venues categories in each neighbourhood

	Neighborhood	Athletics & Sports	Boxing Gym	Gym	Gym / Fitness	Hotel	Mall	Residential	Sports Club
0	Ampang, Kuala Lumpur	0.000000	0.0	0.181818	0.000000	0.818182	0.000000	0.000000	0.0
1	Bandar Menjalara	0.000000	0.0	0.750000	0.000000	0.000000	0.250000	0.000000	0.0
2	Bandar Sri Permaisuri	0.250000	0.0	0.250000	0.000000	0.000000	0.000000	0.500000	0.0
3	Bandar Tasik Selatan	0.166667	0.0	0.166667	0.166667	0.166667	0.166667	0.166667	0.0
4	Bandar Tun Razak	0.000000	0.0	0.500000	0.500000	0.000000	0.000000	0.000000	0.0
...	...	...	...	...	...	...	...	...	...
63	Taman Tun Dr Ismail	0.000000	0.0	0.000000	0.333333	0.166667	0.500000	0.000000	0.0
64	Taman U-Thant	0.000000	0.0	0.000000	0.125000	0.750000	0.125000	0.000000	0.0
65	Taman Wahyu	0.000000	0.0	0.000000	0.000000	0.500000	0.000000	0.500000	0.0
66	Titiwangsa	0.250000	0.0	0.250000	0.000000	0.500000	0.000000	0.000000	0.0
67	Wangsa Maju	0.000000	0.0	0.333333	0.500000	0.000000	0.166667	0.000000	0.0

Figure 7: Average of venue categories in each neighbourhood

3.2. Clustering

Machine learning algorithm (KMeans clustering) will be utilized to segment neighbourhoods into clusters. The dataset of average venues in each neighbourhoods as illustrated in figure 7 are used for clustering. Clustering enables effective and efficient results filtering. The clusters are examined individually to locate potential neighbourhoods.

3.2.1. Elbow Method

The suitable number of clusters is determined by using elbow method.

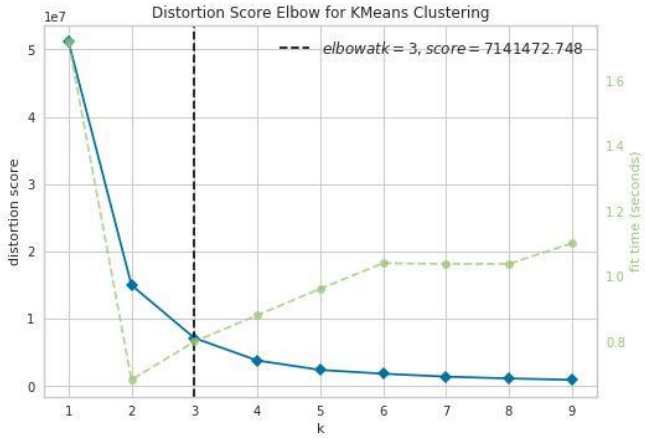

Figure 8: Elbow Method

3.2.2. Clustering

KMeans clustering will be performed by setting the number of cluster to the value found in previous step (elbow method).

After clustering the neighbourhoods, the cluster number of each respective neighbourhood is updated to the table.

	Cluster Label	Neighborhood	Athletics & Sports	Boxing Gym	Gym	Gym / Fitness	Hotel	Mall	Residential	Sports Club	Latitude	Longitude
0	1	Ampang, Kuala Lumpur	0	0	2	0	9	0	0	0	3.148494	101.696729
1	2	Bandar Menjalara	0	0	3	0	0	1	0	0	3.190350	101.625450
2	0	Bandar Sri Permaisuri	1	0	1	0	0	0	2	0	3.103910	101.712260
3	0	Bandar Tasik Selatan	1	0	1	1	1	1	1	0	3.072750	101.714610
4	2	Bandar Tun Razak	0	0	1	1	0	0	0	0	3.082800	101.722810
...	...	...	...	...	...	...	...	...	...	...	...	...
63	0	Taman Tun Dr Ismail	0	0	0	2	1	3	0	0	3.152830	101.622710
64	1	Taman U-Thant	0	0	0	2	12	2	0	0	3.157700	101.724520
65	1	Taman Wahyu	0	0	0	0	1	0	1	0	3.222400	101.671730
66	1	Titiwangsa	1	0	1	0	2	0	0	0	3.180730	101.703210
67	0	Wangsa Maju	0	0	2	3	0	1	0	0	3.203910	101.737190

Figure 9: labelling the cluster number of each neighbourhood

3.3. Cluster Visualization on Map

The cluster distribution is visualized on map, with Python library: Folium.

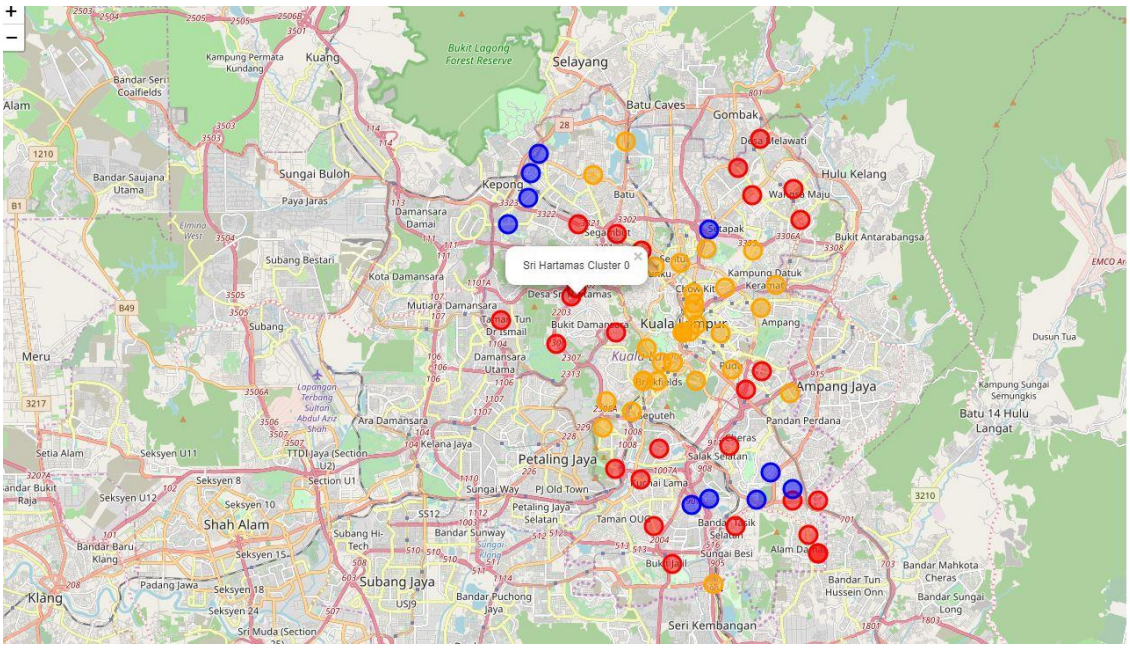

Figure 10: Custer Visualization on Map

3.4. Examine the Cluster

The clusters are examined by plotting them on bar chart.

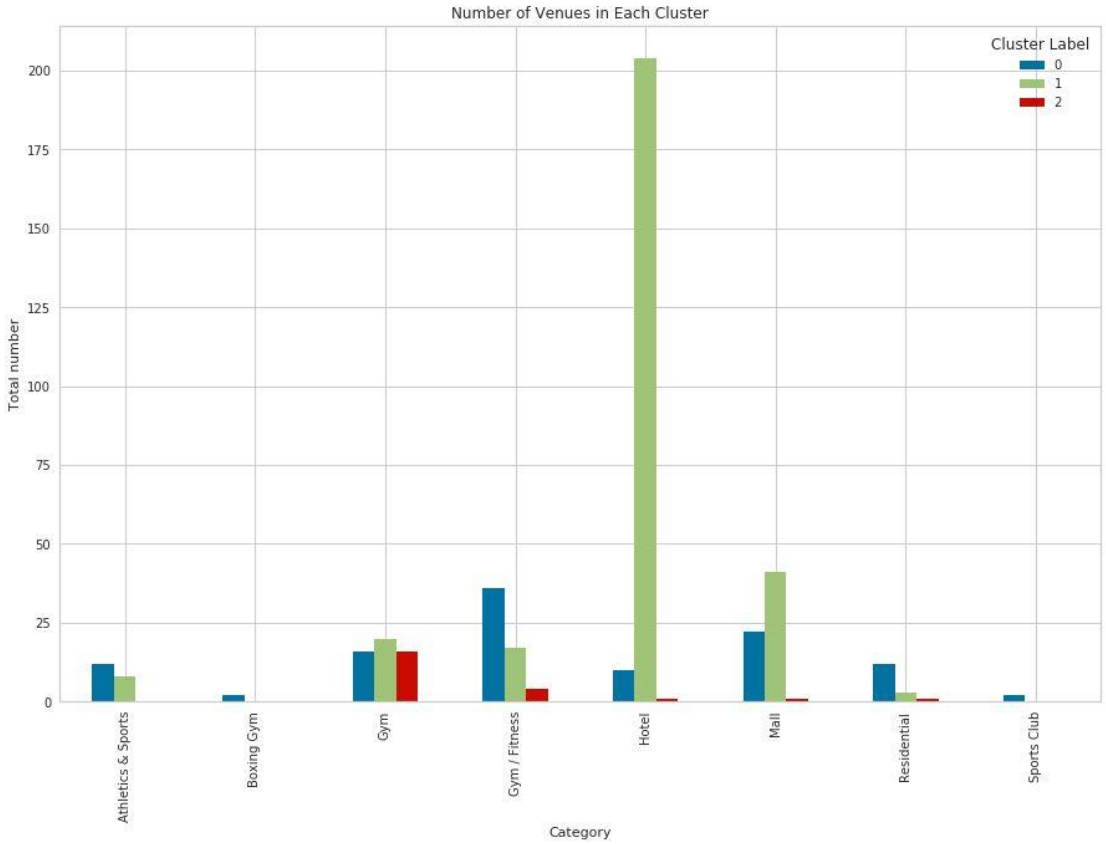

Figure 11: Examine the cluster

4. Results, Discussions & Limitations

4.1. Results & Discussions

There are total of 71 neighbourhoods in Kuala Lumpur. After filtering out the data to only neighbourhoods that consists of target venues, 68 neighbourhoods remain. 428 venues in total are recorded.

The data prepared are clustered by machine learning algorithm (KMeans clustering). Firstly, the suitable number of clusters are found by elbow method. Refer to figure 8 above. It's suggested from the elbow method that 3 clusters of neighbourhoods exist in the data.

The venues distribution are visualized on bar chart, as illustrated in figure 11. The result is also tabulated as shown below.

Cluster Label	0	1	2
Athletics & Sports	12	8	0
Boxing Gym	2	0	0
Gym	16	20	16
Gym / Fitness	36	17	4
Hotel	10	204	1
Mall	22	41	1
Residential	12	3	1
Sports Club	2	0	0

Figure 12: Clustering Result

From the bar chart and table above, it's noticed that these clusters could be described as below:

	Description
Cluster 0	Moderate number of traffic attraction venues (malls/hotel/apartments), high amount of sports amenities
Cluster 1	High amount of traffic attraction venues , moderate amount of sports amenities
Cluster 2	Low amount of traffic attraction venues, low amount of sports amenities

Our target neighbourhoods are those with high traffic attraction venues (such as malls, hotel, apartments), with low competition (sports amenities). Thus, it's suggested from the result above that cluster 1 fits into the requirements.

The cluster 1 is further analysed. There are 30 neighbourhoods that fall under cluster 1. To visualize the cluster better, the venues are divided into two categories, which are:

- Traffic (shopping malls, hotels, residential etc)
- Sports (Gym, boxing gym, sports club etc)

To further narrow down the potential neighbourhood list, cluster 1 is sub-segmented into 3 sub-clusters. The average number of venues in each sub-cluster could be visualized in the bar chart below.

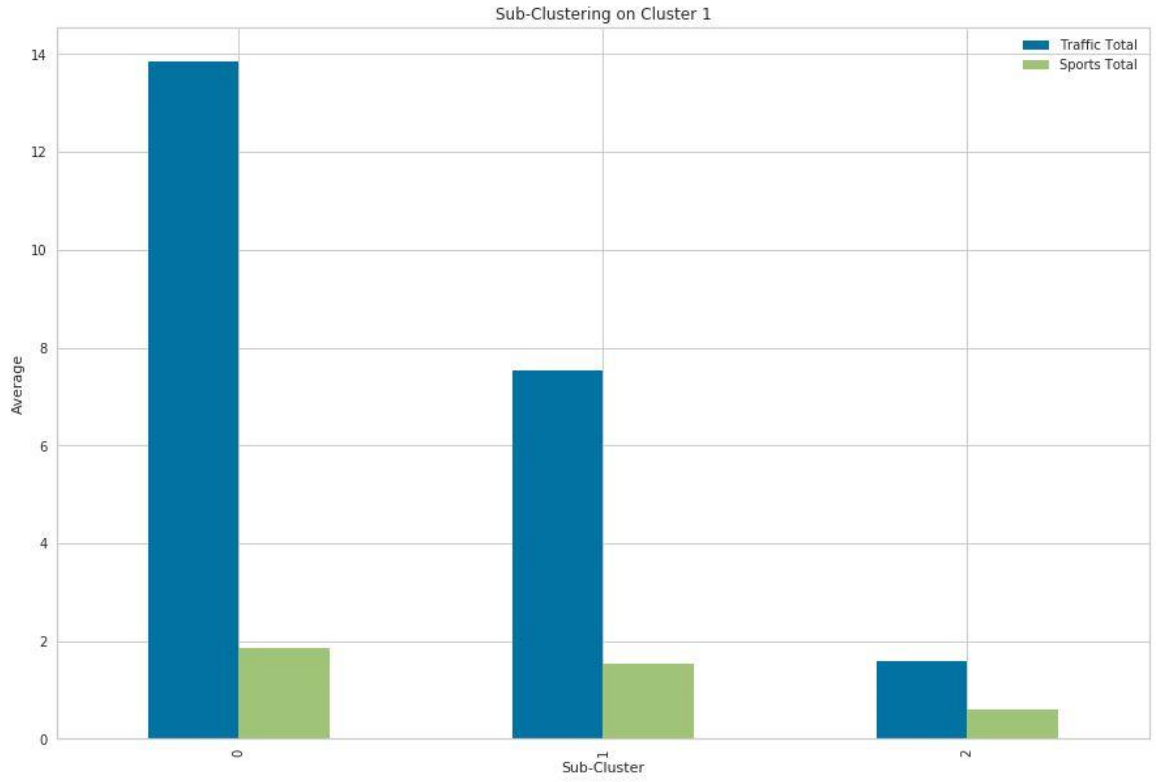

Figure 13: Average Venue Numbers of Cluster 1

The sub-clusters in cluster 1 could be described as below:

Sub-cluster	Number of Traffic Attraction Venues	Number of Sports Venues
0	highest	highest
1	2nd	2nd
2	lowest	lowest

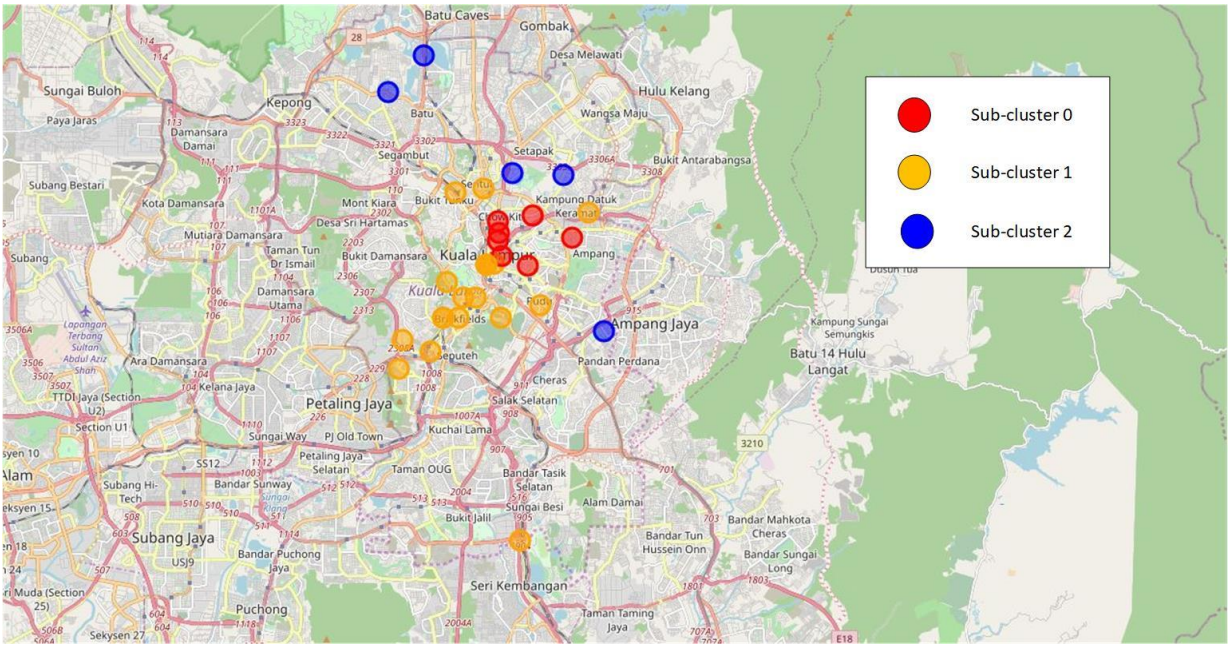

Figure 14: Distribution of sub-clusters of cluster 1 on map

The sub-clusters distribution are visualized on the map. It shows that sub-cluster 0 are located in the city centre of Kuala Lumpur, which explain the observation of high number of traffic attraction venues (shopping malls / hotels/ residential etc.).

Based on the results and selection criteria, it shows that neighbourhoods in sub-cluster 0 of cluster 1 are the potential neighbourhoods. The number of traffic venues are high, while the number of sports amenities are only slightly higher than that of sub-cluster 1.

Besides, sub-cluster 1 of cluster 1 could also be considered. Despite having lower traffic attraction venues, the number of competitors (sports amenities) are also lower.

The neighbourhood's lists of these sub-clusters are tabulated below.

Sub-cluster 0 of cluster 1. Seven (7) neighbourhoods in total.

	Neighborhood	Sub-Cluster label	Traffic Total	Sports Total	Latitude	Longitude
0	Bukit Bintang	0	15	1	3.147770	101.708550
1	Bukit Nanas	0	12	4	3.151142	101.699375
2	Chow Kit	0	14	1	3.163590	101.698110
3	Dang Wangi	0	15	3	3.156685	101.698077
4	Kampung Baru, Kuala Lumpur	0	12	0	3.165460	101.710280
5	Medan Tuanku	0	15	2	3.159260	101.698340
6	Taman U-Thant	0	14	2	3.157700	101.724520

Figure 15: Neighbourhoods in sub-cluster 0 of cluster 1

Sub-cluster 1 of cluster 1. Nineteen (19) neighbourhoods in total.

	Neighborhood	Sub-Cluster label	Traffic Total	Sports Total	Latitude	Longitude
0	Ampang, Kuala Lumpur	1	9	2	3.148494	101.696729
1	Bangsar	1	10	0	3.129200	101.678440
2	Bangsar Park	1	10	0	3.129200	101.678440
3	Bangsar South	1	6	2	3.111020	101.662830
4	Batu, Kuala Lumpur	1	7	2	3.147890	101.694050
5	Brickfields	1	10	2	3.129160	101.684060
6	Bukit Petaling	1	7	1	3.129290	101.698960
7	Bukit Tunku	1	6	2	3.173810	101.682760
8	Damansara Town Centre	1	8	2	3.136442	101.690296
9	Damansara, Kuala Lumpur	1	8	1	3.141906	101.679678
10	Federal Hill, Kuala Lumpur	1	9	2	3.136370	101.685640
11	KL Eco City	1	7	1	3.117130	101.673840
12	Kampung Datuk Keramat	1	6	2	3.166400	101.730460
13	Lembah Pantai	1	8	1	3.121202	101.663899
14	Maluri	1	7	2	3.147890	101.694050
15	Miharja	1	7	2	3.147890	101.694050
16	Pudu, Kuala Lumpur	1	6	0	3.133540	101.713070
17	Sentul, Kuala Lumpur	1	6	2	3.175080	101.693050
18	Sungai Besi	1	6	3	3.049970	101.706030

Figure 16: Neighbourhoods in sub-cluster 1 of cluster 1

4.2. Limitation

There are some limitations in this projects, which can be further improved should this project is extended. They are as listed below:

1. The traffic attraction is assumed by the number of venues (such as hotels, shopping malls etc.). Some factors which contribute to traffic attraction are not considered in this approach, such as the attractiveness of venues, capacity of these venues, transportation etc.

2. There are other venues with high traffics, such as corporate offices, which are not listed in Foursquare returned data.

3. The competitors are not studied in details (such as distances, pricing, rating etc.)

4. The neighbourhoods are not studied in details (such as transportation, area size, population etc.).

5. There are other determining factors besides the selection criteria listed, which is not in the scope of this project.

5.Conclusion

By utilizing foursquare location data complemented with neighbourhood list from Wikipedia, the data required for this project are collected.

By running through series of pre-processing and machine learning algorithm, the potential neighbourhoods are located. They are 2 classes of neighbourhoods recommended.

The first class are the best fit neighbourhoods, 7 neighbourhoods in total. They are labelled as sub-cluster 0, cluster 1. The neighbourhoods list are as shown in figure 15.

The second class consists of 19 neighbourhoods. The traffic attraction venues are lesser than that of first class, with slightly lower competitors in vicinity. The neighbourhoods list are as shown in figure 16.

This project has successfully answered the question raised by stakeholders. It's worth taking note that there are some limitations in this project, which could be seen as result improvement opportunities should this project is extend.