Battle of the SBK Line Metro Station Neighborhoods, Klang Valley, Malaysia.

Prepared by: Joe Alexander

Date: 29th September 2020

1. Introduction

1.1. Background

This is the report of the capstone project for the Coursera Data Science for Professionals Certificate. The remit is to frame a problem or need where Foursquare location data is utilized to explore and compare geographical locations.

In addition to attempting to meet the capstone project requirements, there are personal goals aligned. Being a retired petroleum engineer of 65 years old and having absolutely no previous coding experience but with time and resources at my disposal, I now have a curiosity for data science and wish to engage in it in my retirement activities, e.g. long-distance endurance cycling. In the near term, I wish to investigate and better understand the physiology demands of cycling with age. As I develop more skills, I hope to utilize it for other wellness activities.

To this end, I did not rush to complete my project and instead took a lot of time to educate myself at every opportunity during of the project completion with extended knowledge capturing in python and its libraries.

My goals were to meet the Coursera capstone project requirement and to gain a deeper understanding of the skills that I learnt during this course.

1.2. Opportunity

The Sungai Buloh- Kajang (SBK) Metro line spanning the Klang Valley in Malaysia is selected for this project. This Metro line runs through the heart of Kuala Lumpur, one of the largest metropolitan cities in South East Asia and by far the largest city in Malaysia. Being the economic and financial center of Malaysia, it continually attracts young professionals to the city seeking opportunities and enhancements in their careers. The Kuala Lumpur area has a population of nearly 2 million and the Greater Kuala Lumpur area, often called the Klang Valley, has a population of more than 7 million.

For young professionals starting their careers in Kuala Lumpur, selecting a residential neighborhood can be challenging. More likely than not, living in Kuala Lumpur city center will be unaffordable for young professionals. However, residential locations outside Kuala Lumpur city center but within the Klang Valley is an option that still allows a decent work play lifestyle.

A very good option is to select residential locations in the proximity of the Metro (MRT) stations along the SBK (Sg. Buloh – Kajang) line; the SBK line runs from Kajang in the South East of Kuala Lumpur, travers across the heart of Kuala Lumpur City Center (KLCC), and continues to Sg. Buloh to the North West of Kuala Lumpur.

This exercise set out to cluster and explore the neighborhoods within a kilometer radius of each the SBK Metro stations. The information could be informative to new incoming professionals working within the city of Kuala Lumpur or in the surrounding the Klang Valley. Neighborhoods are explored for different venue categories listed in Foursquare and used to select Metro Station neighborhoods based on their preferences.

2. Methodology

The following methodology was followed:

i. Data was mined from the Wikipedia site mentioned below using Beautiful Soup.
ii. A pandas data frame with only the key relevant features was created and cleaned.
iii. An initial visualization exploratory data analysis was performed on the data and confirmed by checking duplicates and uniqueness of the data in the data frame
iv. Alternate sources of geographical coordinates were used to replace incorrect and duplicated coordinates in the data frame.
v. Once the station names, station codes and station coordinates for all the SBK Metro stations were cleaned, a second visualization exploration data analysis was conducted using Folium map; to ascertain the neighborhoods of 1 km radius around the stations to be evaluated did not overlap.
vi. Selected stations were dropped from the data frame to avoid overlapping station neighborhoods.
vii. Foursquare was used to explore the highest-level venue categories and count for the station neighborhoods.
i. The venue categories count was then normalized for all the stations considered in the analysis.
ii. The station neighborhoods were then clustered and segmented according to the venue categories.
iii. The clusters were analyzed based on the category venue counts.
iv. Further categories were drilled down for understanding of the clusters.
v. The venue categories count was then normalized for all the stations considered in the analysis.
vi. The station neighborhoods were then clustered and segmented according to the venue categories.
vii. The results are discussed and the exercise concluded.

3. Data

3.1. Data sources

The map of the MRT route and train stations is obtain from the Wikipedia link shown below, https://en.wikipedia.org/wiki/Kajang_line# .

A map of the line with the 33 stations extracted from the above website is shown below.

Geopy Geocoders - Nominatim and Google Maps are also used to extract geographical coordinates when duplicated coordinates were detected in the data extracted form Wikipedia.

Foursquare is used for exploring venues around the selected Metro Stations.

All the information used in this exercise is publicly available.

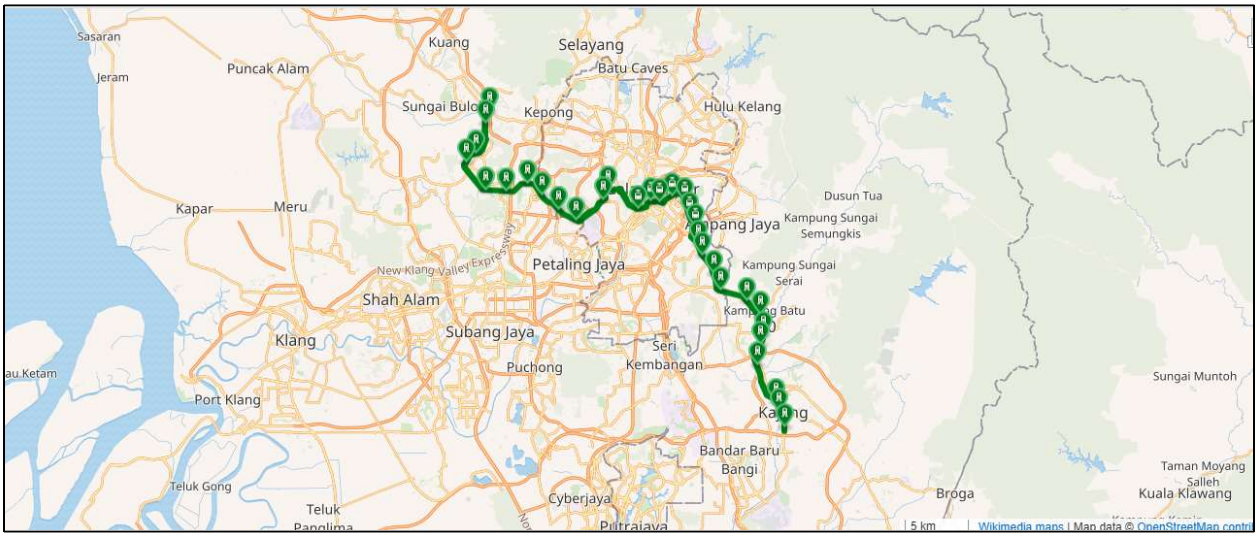


Figure 1 – Map of SBK Line, Wikipedia.

3.2. Data cleaning

The information scaped from the website into pandas data frame contained several non-relevant features. The database was cleaned to contain only the following features – ‘Station Names’, ‘Station Codes’ and ‘Coordinates’.

The Metro Stations were then mapped using Folium. Only 26 stations were displayed on the map.

A check on the number of stations in the scraped data showed 33 stations, which is correct. However, when the database ‘Coordinates” were checked for the length of unique ‘Coords’, only 26 stations are shown. Furthermore, when ‘Coords’ were checked for duplicity, seven stations (SBK24, SBK 25, SBK 26, SBK28, SBK 29, SBK30 and SBK 31) were found to have duplicate coordinates.

The coordinates for these seven stations were then sourced from Geopy Geocoders – Nominatim and plotted again using Folium. On this map, the coordinates of 2 of the stations looked incorrect because of a severe zigzag pattern between adjacent stations. The incorrect coordinates then replaced with coordinates from Google Maps.

The information in the data frame containing the 33 station was viewed again using Folium (refer Figure 2 below) and used for Exploratory Data Analysis.

3.3. Visual Exploratory Data Analysis

For this exercise, only visualization exploratory data analysis was used. The Metro Line station neighborhoods of 1 km radii is shown in Figure 2 below. The neighborhood boundaries of 1 km radius around each station are to scale.

From the map below, it is evident that the stations are closely located and if 1 km radius boundaries are used, then several of the neighborhoods will overlap. This will skew the Foursquare search results.

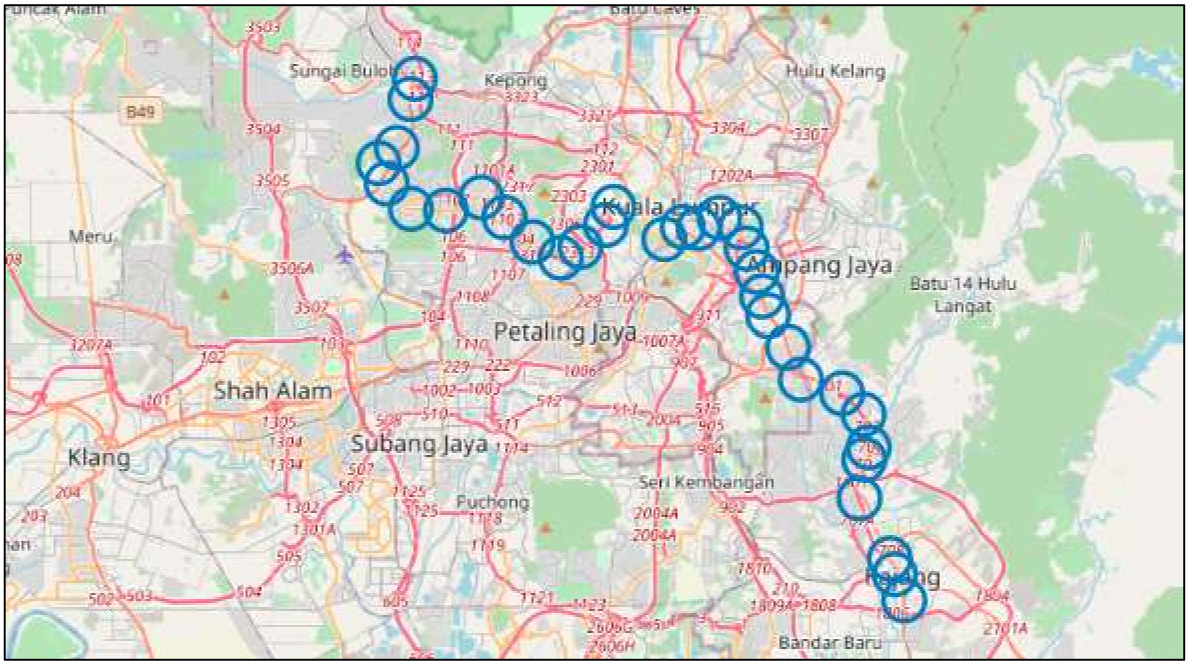


Figure 2 – SBK Line Stations (33) with 1 km Neighborhoods

To overcome the overlapping issue, only alternate station neighborhoods along the route were selected for Foursquare exploration. Only 17 station neighborhoods are used for the exploration.

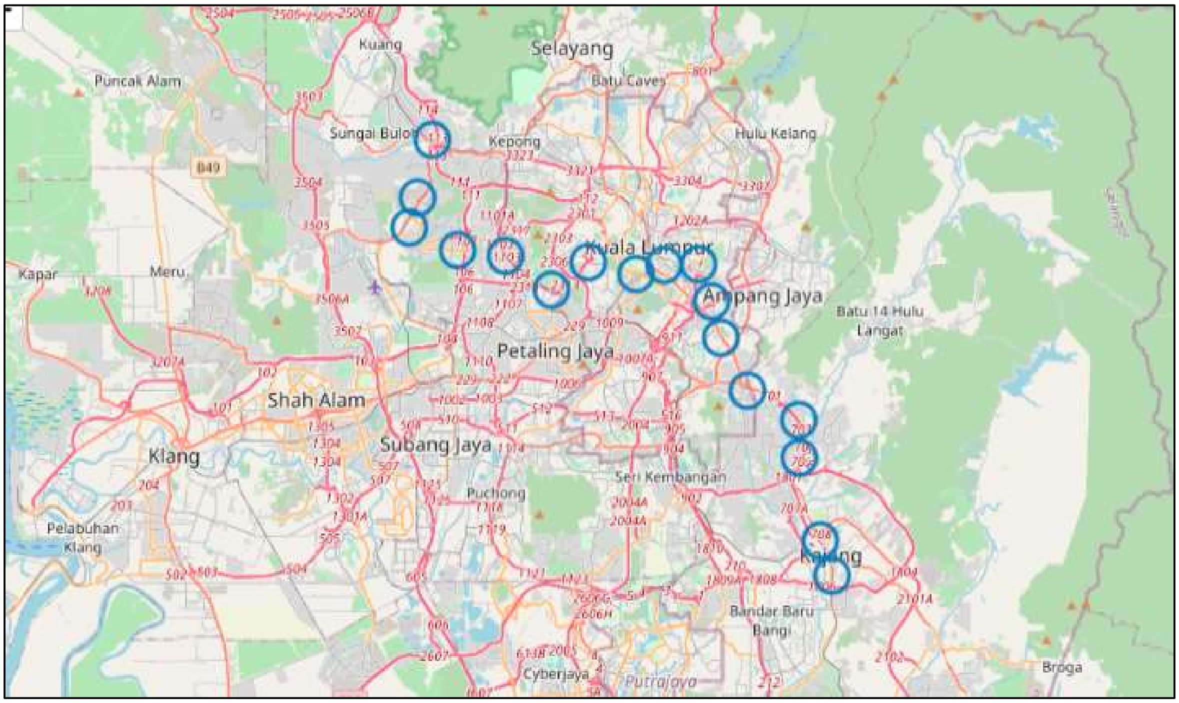
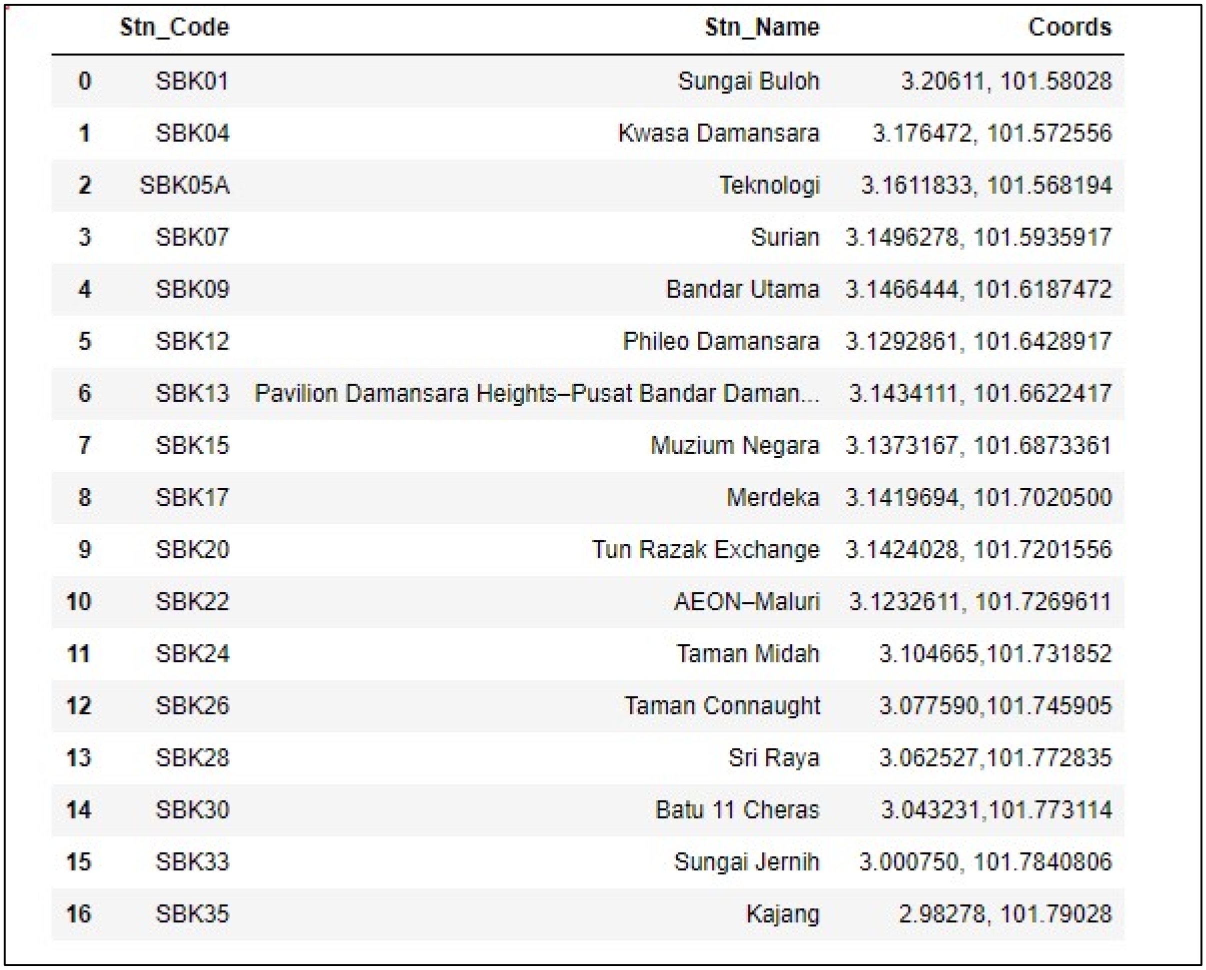


Figure 3 – Alternate SBK Line Stations (17) with 1 km Neighborhoods

As can be observed from Figure 3 above, when alternate station neighborhoods are considered, there is barely any overlap of neighborhoods.

4. Exploration of venues with Foursquare

The station data used for the Foursquare exploration is given in Table 1 below.

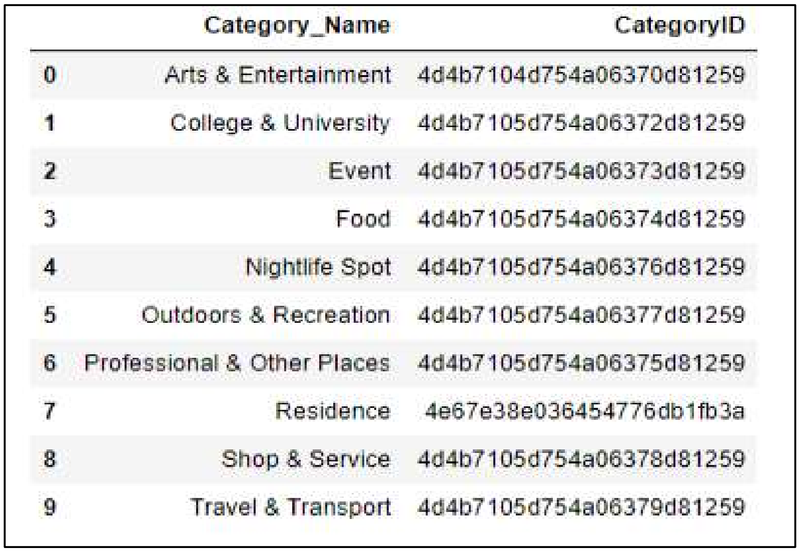


	Stn_Code	Stn_Name	Coords
0	SBK01	Sungai Buloh	3.20611, 101.58028
1	SBK04	Kwasa Damansara	3.176472, 101.572556
2	SBK05A	Teknologi	3.1611833, 101.568194
3	SBK07	Surian	3.1496278, 101.5935917
4	SBK09	Bandar Utama	3.1466444, 101.6187472
5	SBK12	Phileo Damansara	3.1292861, 101.6428917
6	SBK13	Pavilion Damansara Heights–Pusat Bandar Daman...	3.1434111, 101.6622417
7	SBK15	Muzium Negara	3.1373167, 101.6873361
8	SBK17	Merdeka	3.1419694, 101.7020500
9	SBK20	Tun Razak Exchange	3.1424028, 101.7201556
10	SBK22	AEON–Maluri	3.1232611, 101.7269611
11	SBK24	Taman Midah	3.104665,101.731852
12	SBK26	Taman Connaught	3.077590,101.745905
13	SBK28	Sri Raya	3.062527,101.772835
14	SBK30	Batu 11 Cheras	3.043231,101.773114
15	SBK33	Sungai Jernih	3.000750, 101.7840806
16	SBK35	Kajang	2.98278, 101.79028

Table 1 – Alternate SBK Line Stations (17)

4.1. Exploring venues in station neighborhoods with Foursquare.

One of the main goals of this exercise was to explore venues categorized at the highest level in Foursquare. The categories (and their IDs) explored are listed in Table 2 below.



	Category_Name	CategoryID
0	Arts & Entertainment	4d4b7104d754a06370d81259
1	College & University	4d4b7105d754a06372d81259
2	Event	4d4b7105d754a06373d81259
3	Food	4d4b7105d754a06374d81259
4	Nightlife Spot	4d4b7105d754a06376d81259
5	Outdoors & Recreation	4d4b7105d754a06377d81259
6	Professional & Other Places	4d4b7105d754a06375d81259
7	Residence	4e67e38e036454776db1fb3a
8	Shop & Service	4d4b7105d754a06378d81259
9	Travel & Transport	4d4b7105d754a06379d81259

Table 2 – Highest Level Foursquare Category IDs

Once the value counts for the different categories count was obtained for all the 17 SBK Metro stations' neighborhoods, they were normalized using the MinMaxScaler from SKLearning.

The normalized venue counts for the 17 stations were then boxplotted using seaborn. The box plots are shown below in Figure 4. From the box plots, all the features i.e. Category IDs have sufficient value counts to be included for the clustering exercise. Travel and Transport cold have been omitted due to its low count and spread but I decided not to omit it.

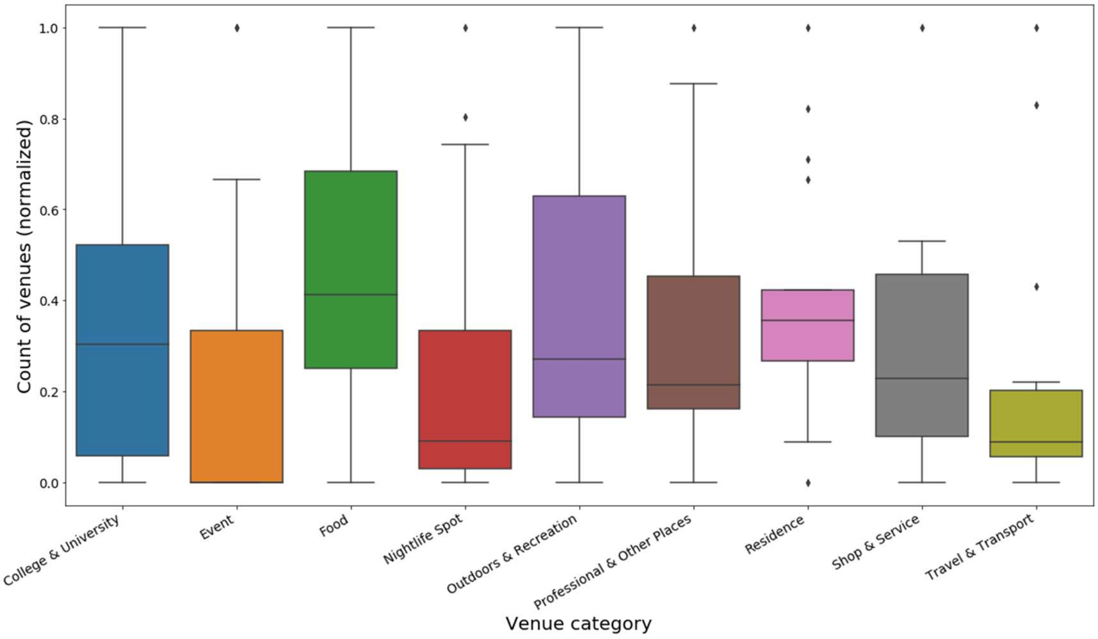


Figure 4 – Normalized Venue Counts for the 17 station

4.2. Clustering

K Means is utilized to cluster the stations based on their category venue counts. The k-elbow method suggests that the optimal number of clusters in this evaluation is 3. Refer Figure 5.

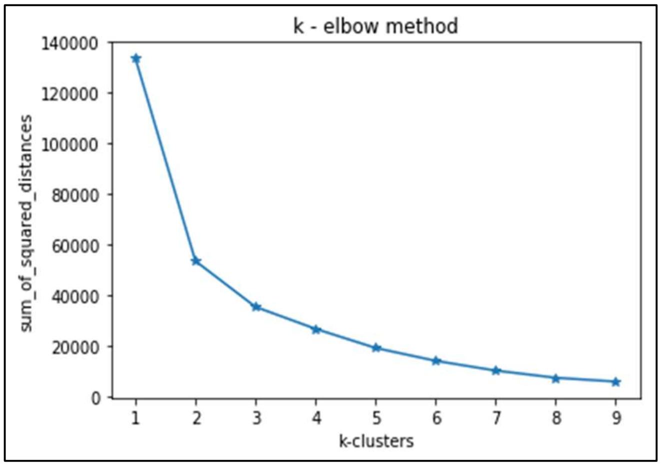


Figure 5 – Optimal clusters

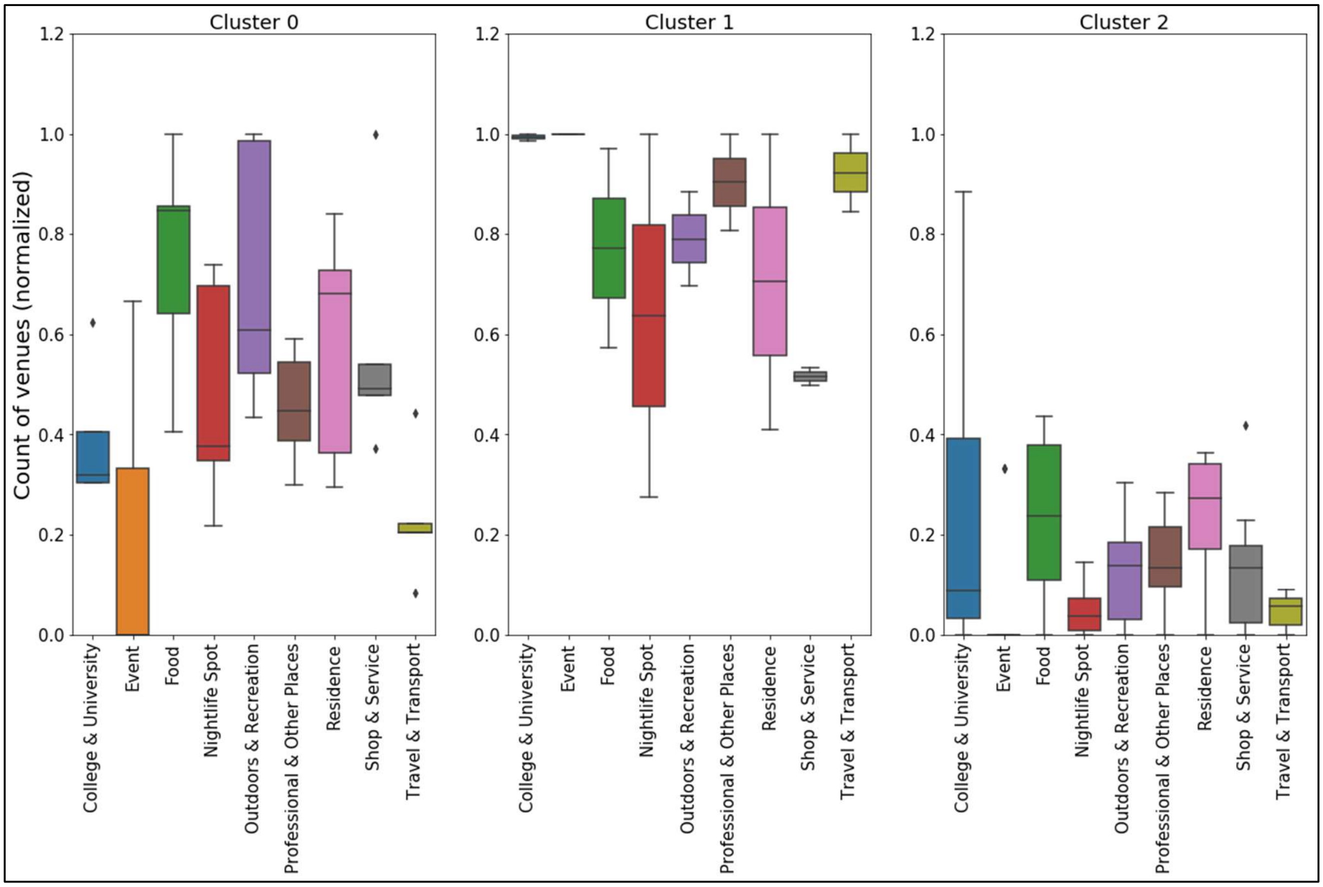


Figure 6 – Venue counts for all Categories in Clusters 0, 1 and 2.

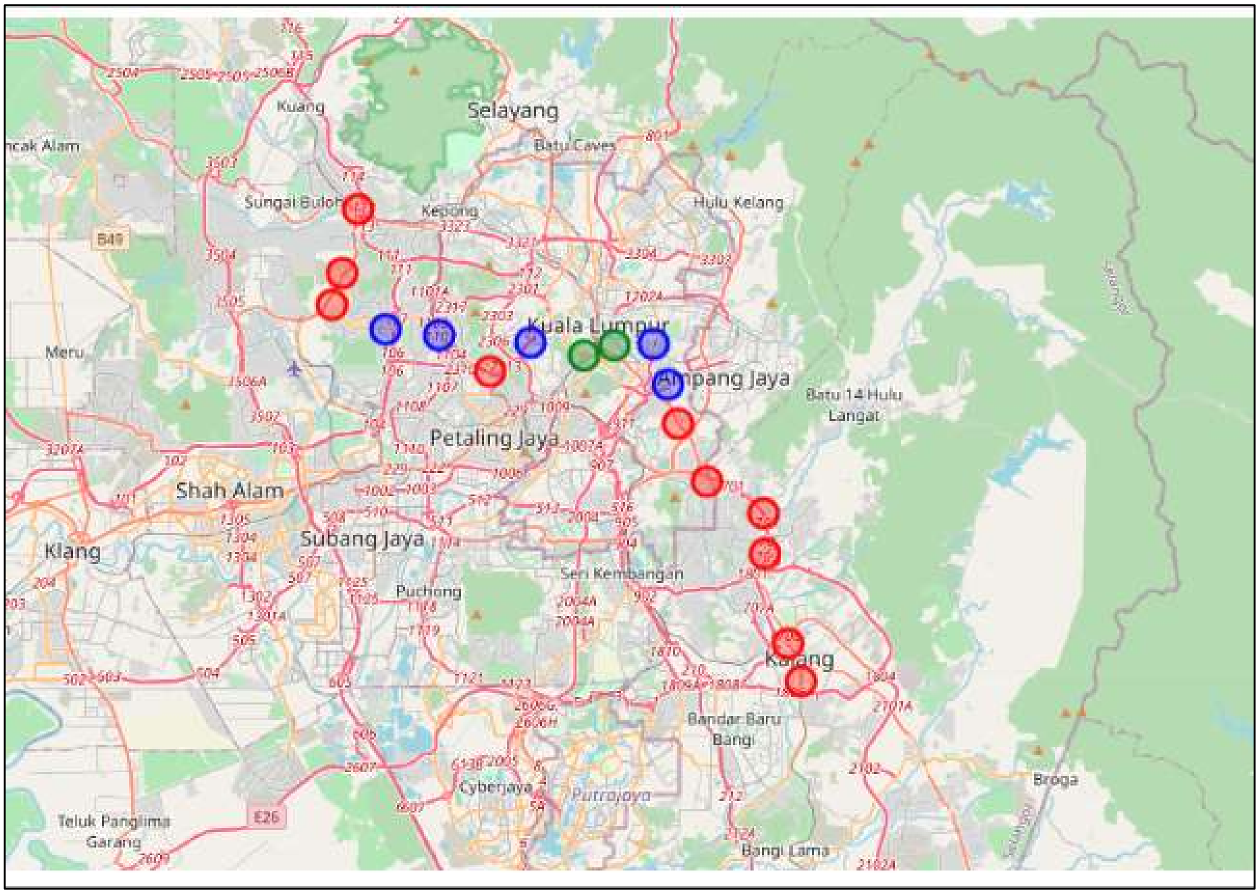


Figure 7 – Station Clusters

Cluster	Station Code	Station Name	Color
0	SBK07	Surian	Blue
0	SBK09	Bandar Utama	Blue
0	SBK13	Pavilion Damansara	Blue
0	SBK20	Tun Razak Exchange	Blue
0	SBK22	AEON-Malun	Blue
1	SBK15	Muzium Negara	Green
1	SBK17	Merdeka	Green
2	SBK01	Sungai Buloh	Red
2	SBK 04	Kwasa Damansara	Red
2	SBK05A	Teknologi	Red
2	SBK12	Phileo Damansara	Red
2	SBK24	Taman Midah	Red
2	SBK26	Taman Connaught	Red
2	SBK28	Sri Raya	Red
2	SBK30	Batu 11 Cheras	Red
2	SBK33	Sungai Jernih	Red
2	SBK35	Kajang	Red

Table 3 – Metro Stations by Clusters

5. Analysis and Discussion

5.1. Discussion of Clustered Neighborhoods

The clusters shown in the map above can be used for an initial high-level screening for newcomer professionals seeking their preferential neighborhood in the Klang Valley.

Cluster 0 (shown in blue on the map) consist of five stations, and is the commercial hub and tourist hub of Kuala Lumpur; where embassies, corporate headquarters of Malaysian companies like the oil giant Petronas, several regional bank headquarters, shopping malls, etc. are located. Eateries, bars and places of entertainment are plentiful besides the nightlife scene. Commercial and residential property rental is at a premium but much sought after. A popular residential neighborhood for expats, diplomats, oil and gas executives, bank executives and wealthy Malaysians and a very popular stop for local and foreign tourists. A popular place for events like trade shows, exhibitions, conferences and also has a few private colleges.

Cluster 1 (shown in green on the map) are two station neighborhoods adjacent to Cluster 0. It is a more congested neighborhood, with Professional places, Residences, and is being developed as a Transport Hub for the Klang Valley. These are hardly any Colleges, Event or spots for Outdoor & Recreation activities. This does not look suitable for families. Rental is premium at these neighborhoods.

Cluster 2 (shown in red on the map) are the station neighborhoods in the Klang Valley outside the Kuala Lumpur city station neighborhoods. The station neighborhoods in this cluster have much lower venue counts in all categories than the other clusters mainly because these neighborhoods are in less densely populated areas and venues are spread out away from the stations rather than concentrated around the stations. Residents often live further out than a 1000m radius and are more likely to use the 'park and ride scheme' where car are parked at the stations and car owners use the SBK Metro line to travel or commute to Kuala Lumpur or when e-hailing services are used between residences and Metro stations. These locations offer more affordable residences. Professionals who would need to work both in the Kuala Lumpur commercial hub and also need frequent air travel could select a residential neighborhood along the southern stations along the SBK line. Similarly, professionals who prefer a more rural and affordable setting besides easier access out of Kuala Lumpur to the northern states could select a neighborhood along the northern stations on the SBK line. Several of the stations in Cluster 2 are located close to several institutions of higher learning and student residential neighborhoods.

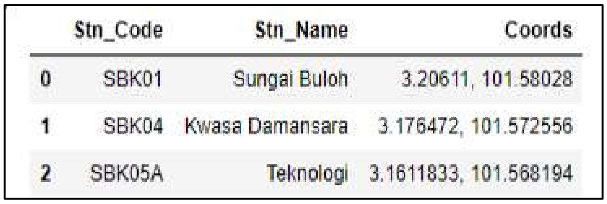
5.2. Exploration of venues of lower level categories at specific station neighborhoods

Once an initial screening has yielded some preferred neighborhood/s or preferred cluster, then venue categories and venues of specific interests can be further explored. For example, for a professional who has an office in Cluster 0 or 1, who drives back fortnightly to the northern city of George Town during weekends could explore Cluster 2 neighborhoods along the northern end of the SBK line. If he/she and their spouse play tennis and /or golf, have young children and a pet dog, then proximity to a Tennis Club, a Golf Club, a Veterinarian, a Daycare center and a Music School may be explored.

Let's do a search for such a station neighborhood. In this case, we shall assume that the spouse will have a vehicle, so the radius of the venues will be extended up to 3 km and the other main breadwinner commutes daily by Metro to Kuala Lumpur city

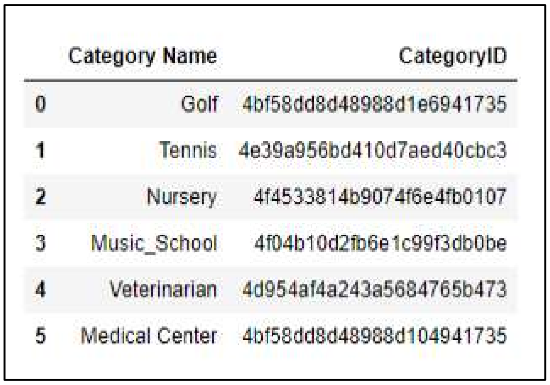
Venues for tennis, golf, music, nursery, medical center and veterinary services are explored for both, a 1km radius neighborhood and a 3km radius neighborhood around three of the stations

Table 3 – Metro Stations by Clusters



	Stn_Code	Stn_Name	Coords
0	SBK01	Sungai Buloh	3.20611, 101.58028
1	SBK04	Kwasa Damansara	3.176472, 101.572556
2	SBK05A	Teknologi	3.1611833, 101.568194

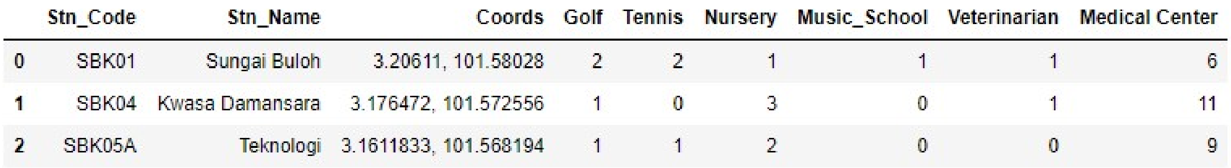
Table 4 – North-West line stations



	Category Name	CategoryID
0	Golf	4bf58dd8d48988d1e6941735
1	Tennis	4e39a956bd410d7aed40cbc3
2	Nursery	4f4533814b9074f6e4fb0107
3	Music_School	4f04b10d2fb6e1c99f3db0be
4	Veterinarian	4d954af4a243a5684765b473
5	Medical Center	4bf58dd8d48988d104941735

Table 5 – Lower-level categories explored

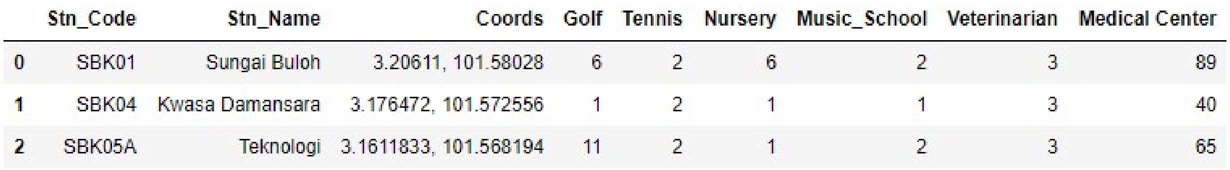
The value count for Golf, Tennis, Nursery, Music School, Veterinarian and Medical Center within 1 km radius neighborhoods around each of the north western stations (SBK01, SBK04 and SBK05A) are shown in Table 6.



	Stn_Code	Stn_Name	Coords	Golf	Tennis	Nursery	Music_School	Veterinarian	Medical Center
0	SBK01	Sungai Buloh	3.20611, 101.58028	2	2	1	1	1	6
1	SBK04	Kwasa Damansara	3.176472, 101.572556	1	0	3	0	1	11
2	SBK05A	Teknologi	3.1611833, 101.568194	1	1	2	0	0	9

Table 6 – Value count for specific requirements, 1 km radius neighborhoods

The value count for Golf, Tennis, Nursery, Music School, Veterinarian and Medical Center within 3 km radius neighborhoods around each of the north western stations (SBK01, SBK04 and SBK05A) are shown in Table 6.



	Stn_Code	Stn_Name	Coords	Golf	Tennis	Nursery	Music_School	Veterinarian	Medical Center
0	SBK01	Sungai Buloh	3.20611, 101.58028	6	2	6	2	3	89
1	SBK04	Kwasa Damansara	3.176472, 101.572556	1	2	1	1	3	40
2	SBK05A	Teknologi	3.1611833, 101.568194	11	2	1	2	3	65

Table 7 – Value count for specific requirements, 3 km radius neighborhoods

The above two searches for specific categories show results must be examined closely and well understood. For example, a search for golf courses in a 3 km radius neighborhood centered at SBK 01 should 6 venues; it is impossible to have 6 golf courses in such a small area. The results actually show any venue related or associated with golf clubs i.e. cafetaria at a golf club, golf driving range, golf accessories store, etc. And one of the golf courses listed is at least 200 km away.

Likewise, it is impossible to have 91 Medical Centers in a 3 km radius neighborhood; the search includes multiple departments at a hospital, daycare medical facility,

5.3. Discussion summary

In this exercise, information has been scraped from a public website which was subsequently cleaned and normalized prior to clustering and segmenting. Folium is used to map the clustered and segmented SBK Metro Station neighborhoods using the highest-level category IDs. This first part of the analysis yielded results that were expected. The categories venue count does give an insight of the neighborhoods in the three clusters

The second part of the analysis, where specific categories like 'golf', 'tennis’, 'nursery', 'medical’, ’center', 'music school' and 'veterinarian' were analyzed for specific SBK Metro station neighborhoods, yielded inaccurate results. The results from these searches must be fully understood before the results are used. For example, a golf course, Clearwater Sanctuary Resort, which is about 200 km away is shown to be in a 3 km radius of station SBK01. The three entries for the same club, Valencia, and a total count of 6 for golf clubs within a 3 km radius of SBK 01 shows the result is not correct. When delving into specific categories, the results must be closely examined and understood. The inaccuracy is probably due to the lack of verifications of the inputs submitted into Foursquare by users.

Last but not the least, Foursquare is an application that is perhaps not commonly used in this part of the world.

Finally, Foursquare is a useful application but must be understood well to ensure inaccurate results are not used in any analysis.

Going forward, a second Metro line is being constructed. I intend to include the second line into a side personal project for practice, but will also try to include rental rates and property prices per sq.ft in the neighborhoods considered. This may require some creative neighborhood boundaries to utilize difficult to obtain rental and property rates along the Metro lines.

6. Conclusion

In this Coursera Capstone project, the brief was to frame a project to use Foursquare API to explore neighborhoods. The neighborhoods in the vicinity of the SBK Metro Stations in the Klang Valley, Malaysia was selected for this exercise. The analysis is aimed at assisting young professionals relocating in the Klang Valley to select neighborhoods for their residences, especially for professionals who prefer not spend hours in the grid lock peak hour traffic of the city or professionals who may wish a greener living by using the Metros in the city.

The objective for doing the above project is to display the skills I have learnt during this Coursera for Professional Certificate course. The coding presented here may be long winded and not the most efficient but it is an attempt by a newbie in data science and to this end, I think I have gained insights into basic Python programming and data analysis.

A secondary but personal goal was to use every opportunity along the way “to learn as much as possible” rather than just completing the capstone project. To this end, I believe I achieved my goal so far.

I look forward to further educate myself in this discipline.