Capstone Project
The Battle of Neighborhoods(week 2)

Explore neighborhoods in the cities of Greater Kuala Lumpur to recommend a location for opening a coffee Shop.

Introduction/Business Problem:

A friend just moved back to Malaysia from Australia recently and is planning to open a coffee shop in Greater Kuala Lumpur. However, he is not sure about the best location to start up his coffee shop. So he ask for my help to recommend the location. With my knowledge in Data Science and skills using location, I will explore cities and neighborhoods in Greater Kuala Lumpur to identify and recommend him the best location for his new coffee shop. This approach is more time-efficient compare to conventional approach(market research, online surveying, etc).

Greater Kuala Lumpur, also refer as Klang Valley is an area in Malaysia which is centered in Kuala Lumpur, and includes its adjoining cities and towns in the state of Selangor. It comprises 3 cities(Kuala Lumpur, Petaling Jaya, Shah Alam) & 6 Municipality(Kajang, Klang, Subang Jaya, Selayang, Ampang Jaya, Sepang). To begin with, I pick the top 2 most populated cities in Greater Kuala Lumpur which is Kuala Lumpur(1,588,750 inhabitants) and Petaling Jaya(613,977 inhabitants), detail refer to here. Both Kuala Lumpur(KL) and Petaling Jaya are the popular choice amongst city dwellers and visitors alike with its abundance of shopping and gastronomic delights – with thousands of hawker stalls, cafes, and restaurants.(Read more here) and here.

There is few critical factors to be considered when choosing a new coffee shop location in order to be profitable and successful such as population, crime rates, visibility & accessibility, surrounding environment and affordability(rental/property pricing).

Data:

Below are the data needed to solve this problem:

- **List of suburbs and neighborhoods of KL and Petaling Jaya with Postal code:**
  - Data source from web here and here for suburbs and postal code from google.
  - Will use the postal code to get geo-coordinate data using geopy library which can be useful when I use foursquare location API to explore neighborhoods later.

- **Foursquare location API:**
  - to explore the neighbourhoods(venues & categorical data) in Kuala Lumpur and Petaling Jaya to research surrounding business(is the area affluent?, What types of restaurants/shops in the area? Are they any coffee shop around the neighborhoods, what are trending places in the neighborhoods).

- **Population, Avg Rental Pricing, Crime Level of all suburbs and neighborhoods.**
  - Data source from web here and store in IBM DB2 server. To be mapped to the location data above to help identify the potential/selected location if it has enough population, low/average crime level and reasonable rental.

Methodology:

1. I collected the data(list of suburbs and neighborhoods with postal code) and demographic(population, Average Rental pricing, crime levels) from web and create a datatables and store it in IBM DB2 server.

2. Load the datasets of suburbs, neighborhoods and demographic from server, transform them into pandas dataframe then merge the 2 datasets. There are 12 suburbs and 107 neighborhoods in Kuala Lumpur and Petaling Jaya as below.

	SUBURBS	NEIGHBORHOODS	POSTALCODE	OVERALLPOPULATION	MALAYPOPULATION	OTHERBUMIPOPUL
0	Kepong	Jinjang	52100	54946.0	13229.0	807.0
1	Kepong	Taman Bukit Maluri	52100	NaN	NaN	NaN
2	Segambut	Bandar Manjalara	52200	10438.0	2041.0	56.0
3	Segambut	Bukit Kiara	60000	NaN	NaN	NaN
4	Segambut	Bukit Tunku	50480	NaN	NaN	NaN

```
print('The dataframe has {} suburbs and {} neighborhoods.'.format(
        len(GKLDF['SUBURBS'].unique()),
        GKLDF.shape[0]
    )
)
```

```
The dataframe has 12 suburbs and 107 neighborhoods.
```

3a. Perform data wrangling/data cleaning. There were a lot of missing values from demographic datasets for most of the neighborhoods, because of lack of record keeping. I decided to only use whatever available data. Hence after clean up the missing values, the datasets contain 12 suburbs and 37 neighborhoods.

```
GKLDF = GKLDF.dropna()
GKLDF.reset_index(drop=True, inplace=True)
print(GKLDF.shape)
GKLDF.head()
```

```
(37, 13)
```

	SUBURBS	NEIGHBORHOODS	POSTALCODE	OVERALLPOPULATION	MALAYPOPULATION	OTHERBUMIPOPULATION	CHINESEPOPULATION
0	Kepong	Jinjang	52100	54946.0	13229.0	807.0	33574.0
1	Segambut	Bandar Manjalara	52200	10438.0	2041.0	56.0	7353.0
2	Segambut	Damansara Heights	50490	12335.0	4111.0	256.0	5098.0
3	Segambut	Jalan Duta	50480	9885.0	5212.0	42.0	3279.0
4	Segambut	Mont Kiara	50480	13477.0	830.0	56.0	4465.0

```
print('The dataframe has {} suburbs and {} neighborhoods.'.format(
        len(GKLDF['SUBURBS'].unique()),
        GKLDF.shape[0]
    ))
```

```
The dataframe has 12 suburbs and 37 neighborhoods.
```

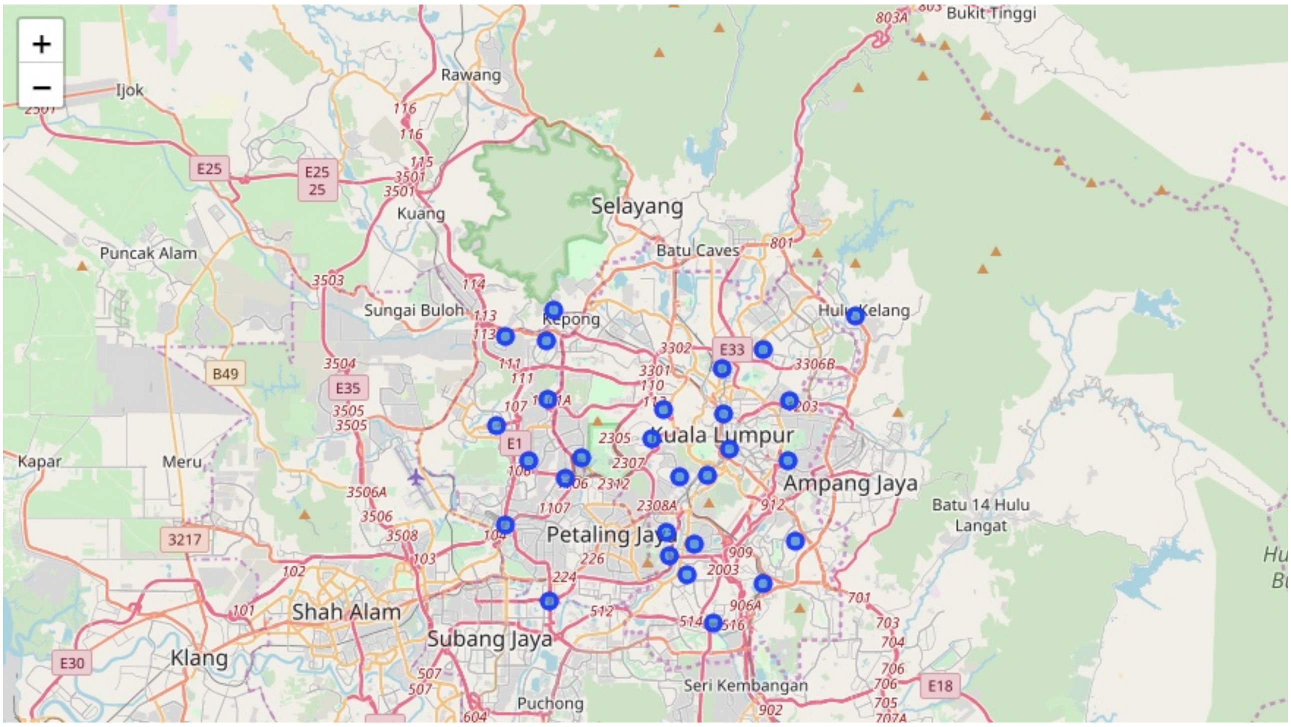
3b. Group the data by postal code and get the sum of population and Avg for rent listing price. The grouped datasets now having 28 unique postal code which will be used to get the latitude & latitude using geopy library.

```
# Generating latitude and longitude for each postalcode of Greater KL(Kuala Lumpure & Petaling Jaya)neighbourhoods.
pbar = ProgressBar()
geolocator = Nominatim()
for index in range(0, GKLNeighF['CountryZip'].shape[0]):
    address = GKLNeighF.loc[index, 'CountryZip']
    location = geolocator.geocode(address, timeout = None)
    if (location != None):
        GKLNeighF.loc[index, 'Latitude'] = location.latitude
        GKLNeighF.loc[index, 'Longitude'] = location.longitude
    sleep(1)
print(GKLNeighF.shape)
GKLNeighF.head()
```

```
(28, 16)
```

INDIANPOPULATION	OTHERSPOPULATION	NONMALAYSIAN	AVGFORRENTLISTINGPRICE_RM	AVGFORRENTLISTINGPRICE_PSF_RM	CRIMELEVEL	Country	CountryZip	Latitude	Longitude
300.0	27.0	2326.0	1634.0	8.640	High	, Malaysia	46150, Malaysia	3.073520	101.612935
281.0	38.0	307.0	7433.0	2.740	Low	, Malaysia	47301, Malaysia	3.108895	101.592445
892.0	158.0	910.0	9633.5	3.720	High, Average	, Malaysia	47400, Malaysia	3.130599	101.620482
978.0	94.0	1634.0	7722.0	4.840	Average	, Malaysia	47410, Malaysia	3.138934	101.603415
5165.0	445.0	4292.0	11709.5	6.025	Average, Average	, Malaysia	47810, Malaysia	3.155758	101.588617

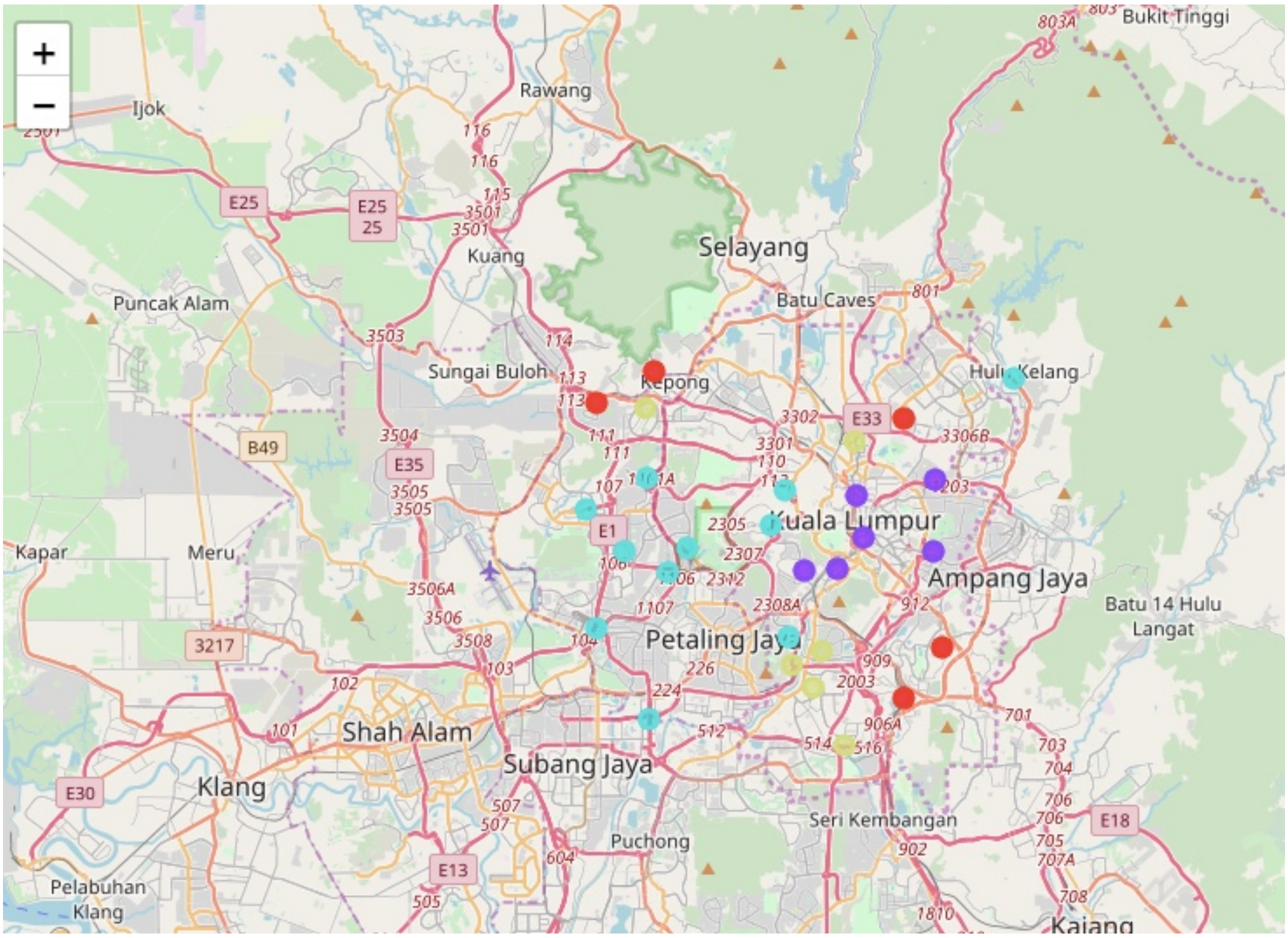
3c. Visualize the neighborhoods in Kuala Lumpur and Petaling Jaya using folium map.



4a. Then, utilizing Foursquare location API to explore neighborhoods in Greater Kuala Lumpur using explore function to get the most common venue categories in each neighborhood and use this feature to group the neighborhoods into 4 clusters. I will use the Machine-Learning K-means clustering algorithm to complete this task.

	Neigh	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	9th Most Common Venue	10th Most Common Venue
0	Alam Damai, Bandar Sri Permaisuri	Chinese Restaurant	Malay Restaurant	Asian Restaurant	Seafood Restaurant	Café	Breakfast Spot	Pizza Place	Noodle House	Indonesian Restaurant	Food Truck
1	Bukit Jalil, Sri Petaling, B...	Café	Chinese Restaurant	Asian Restaurant	Japanese Restaurant	Coffee Shop	Restaurant	Vegetarian / Vegan Restaurant	Massage Studio	Bubble Tea Shop	Burger Joint
2	Bandar Manjalara, Bandar Sri Damansara	Chinese Restaurant	Café	Coffee Shop	Vegetarian / Vegan Restaurant	Asian Restaurant	Indian Restaurant	Food Truck	Pizza Place	Restaurant	Park
3	Kota Damansara, Mutiara Damansara	Café	Coffee Shop	Malay Restaurant	Bakery	Golf Course	Massage Studio	Bubble Tea Shop	Thai Restaurant	Burger Joint	Spa
4	Jalan Duta, Mont Kiara, Sri Hartamas	Japanese Restaurant	Café	Asian Restaurant	Korean Restaurant	Ice Cream Shop	Italian Restaurant	Bakery	Gym / Fitness Center	Seafood Restaurant	Spa

4b. use the Folium library to visualize the neighborhoods in Greater KL(Kuala Lumpur and Petaling Jaya) and their emerging clusters.



Result:

Then examine each cluster and determine the discriminating venue categories that distinguish each cluster.

Cluster 1

```
GKL_combine.loc[GKL_combine['Cluster Labels'] == 0, GKL_combine.columns[[1] + list(range(2, GKL_combine.shape[1]))]]
```

SF_RM	CRIMELEVEL	Country	CountryZip	Latitude	Longitude	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	9th Most Common Venue	10th Most Common Venue
1.34	Average	Malaysia	47830, Malaysia	3.197427	101.592788	0	Malay Restaurant	Chinese Restaurant	Asian Restaurant	Indian Restaurant	Café	Fast Food Restaurant	Burger Joint	Ice Cream Shop	Pizza Place	Flower Shop
1.79	High	Malaysia	52100, Malaysia	3.209695	101.615605	0	Chinese Restaurant	Noodle House	Asian Restaurant	Café	Indian Restaurant	Burger Joint	Cantonese Restaurant	Seafood Restaurant	Malay Restaurant	Trail
2.60	High	Malaysia	53200, Malaysia	3.191315	101.713538	0	Chinese Restaurant	Malay Restaurant	Asian Restaurant	Café	Coffee Shop	Food Court	Thai Restaurant	Fast Food Restaurant	Dance Studio	Halal Restaurant
5.68	Average, Average	Malaysia	56000, Malaysia	3.101620	101.728337	0	Chinese Restaurant	Malay Restaurant	Asian Restaurant	Seafood Restaurant	Café	Breakfast Spot	Pizza Place	Noodle House	Indonesian Restaurant	Food Truck
0.00	High	Malaysia	57100, Malaysia	3.081433	101.713298	0	Chinese Restaurant	Malay Restaurant	Asian Restaurant	Indian Restaurant	Burger Joint	Convenience Store	Flea Market	Pool Hall	Bakery	Department Store

Cluster 2

```
GKL_combine.loc[GKL_combine['Cluster Labels'] == 1, GKL_combine.columns[[1] + list(range(2, GKL_combine.shape[1]))]]
```

_RM	CRIMELEVEL	Country	CountryZip	Latitude	Longitude	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	9th Most Common Venue	10th Most Common Venue
9.21	High	Malaysia	50000, Malaysia	3.144790	101.697514	1	Hotel	Café	Spa	Food Truck	Indian Restaurant	Speakeasy	Malay Restaurant	Motel	Boutique	Juice Bar
0.00	High	Malaysia	50350, Malaysia	3.161159	101.694878	1	Hotel	Malay Restaurant	Coffee Shop	Thai Restaurant	Chinese Restaurant	Convenience Store	Clothing Store	Restaurant	Department Store	Soup Place
0.00	High	Malaysia	50470, Malaysia	3.132625	101.687314	1	Hotel	Indian Restaurant	Café	Ice Cream Shop	Hotel Bar	Coffee Shop	Steakhouse	Garden	Convenience Store	Malay Restaurant
1.87	Average	Malaysia	54000, Malaysia	3.167197	101.725897	1	Hotel	Coffee Shop	Juice Bar	Café	Malay Restaurant	Cosmetics Shop	Italian Restaurant	Cocktail Bar	Seafood Restaurant	Restaurant
0.00	NaN	Malaysia	55100, Malaysia	3.139113	101.725283	1	Chinese Restaurant	Hotel	Asian Restaurant	Nightclub	Shopping Mall	Coffee Shop	Noodle House	Middle Eastern Restaurant	Café	Bar
1.41	Average	Malaysia	59000, Malaysia	3.131593	101.674625	1	Indian Restaurant	Hotel	Shopping Mall	Café	Steakhouse	Ice Cream Shop	Juice Bar	Clothing Store	Bar	Coffee Shop

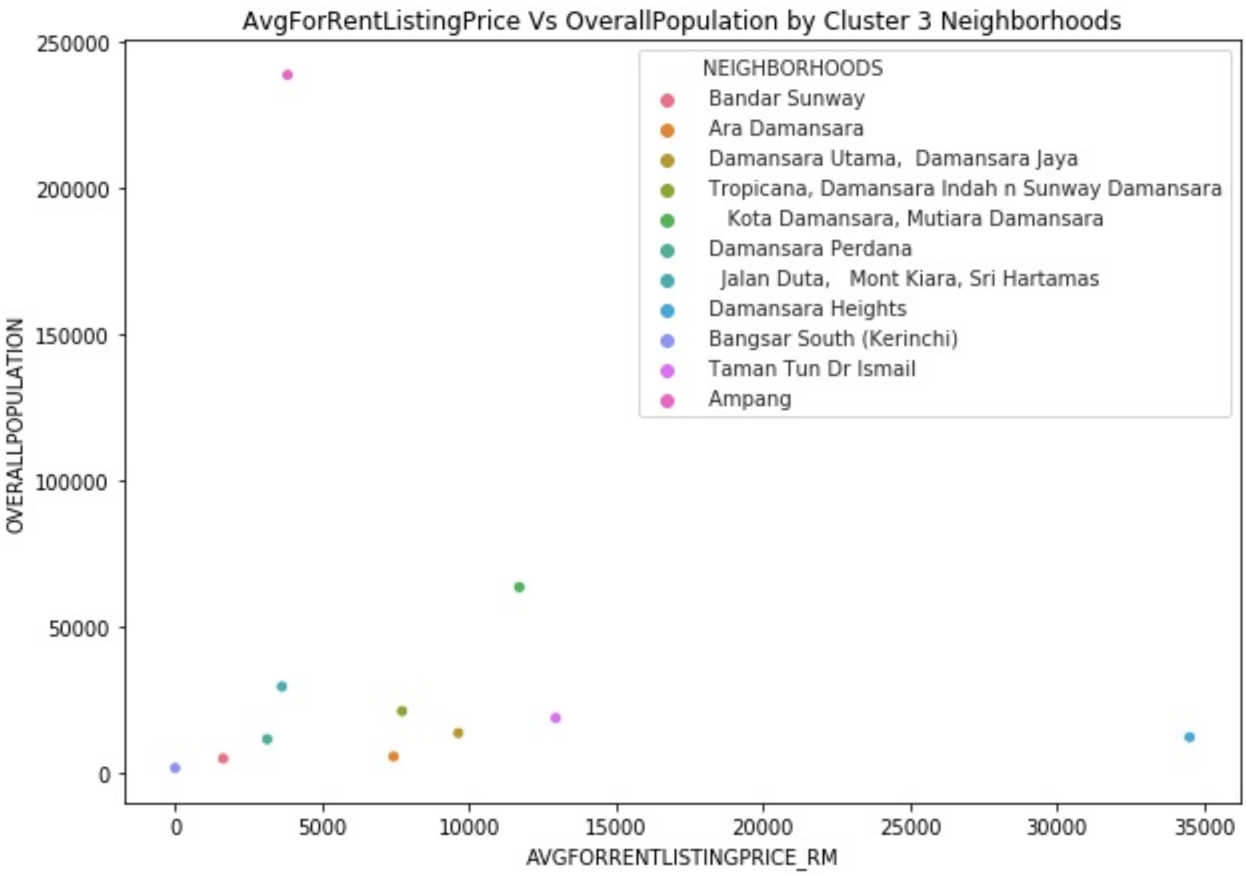
						Labels	Venue	Venue	Venue	Venue	Venue	Venue	Venue	Venue	Venue	Venue
8.640	High	Malaysia	46150, Malaysia	3.073520	101.612935	2	Chinese Restaurant	Coffee Shop	Café	Malay Restaurant	Restaurant	Smoke Shop	Ice Cream Shop	Indonesian Restaurant	Cosmetics Shop	Clothing Store
2.740	Low	Malaysia	47301, Malaysia	3.108895	101.592445	2	Café	Chinese Restaurant	Italian Restaurant	Malay Restaurant	Clothing Store	Steakhouse	Burger Joint	Thai Restaurant	Resort	Sushi Restaurant
3.720	High, Average	Malaysia	47400, Malaysia	3.130599	101.620482	2	Chinese Restaurant	Café	Korean Restaurant	Ice Cream Shop	Bakery	Burger Joint	Dessert Shop	Malay Restaurant	Convenience Store	Massage Studio
4.840	Average	Malaysia	47410, Malaysia	3.138934	101.603415	2	Multiplex	Juice Bar	Spa	Shopping Mall	Ice Cream Shop	Indonesian Restaurant	Café	Flea Market	Bubble Tea Shop	Bar
6.025	Average, Average	Malaysia	47810, Malaysia	3.155758	101.588617	2	Café	Coffee Shop	Malay Restaurant	Bakery	Golf Course	Massage Studio	Bubble Tea Shop	Thai Restaurant	Burger Joint	Spa
6.730	Low	Malaysia	47820, Malaysia	3.167831	101.612787	2	Coffee Shop	Malay Restaurant	Restaurant	Café	Japanese Restaurant	Shopping Mall	Burger Joint	Clothing Store	Snack Place	Gym / Fitness Center
8.460	Low, Average, Low	Malaysia	50480, Malaysia	3.162921	101.666975	2	Japanese Restaurant	Café	Asian Restaurant	Korean Restaurant	Ice Cream Shop	Italian Restaurant	Bakery	Gym / Fitness Center	Seafood Restaurant	Spa
4.530	Low	Malaysia	50490, Malaysia	3.149460	101.661541	2	Café	Japanese Restaurant	Italian Restaurant	Wine Bar	Yoga Studio	Food Truck	Thai Restaurant	Korean Restaurant	Chinese Restaurant	Bar
0.000	Average	Malaysia	59200, Malaysia	3.105585	101.667909	2	Japanese Restaurant	Coffee Shop	Ice Cream Shop	Café	Hotel	Asian Restaurant	Restaurant	Chinese Restaurant	Bakery	Breakfast Spot
1.930	Average	Malaysia	60000, Malaysia	3.140501	101.628224	2	Bakery	Dessert Shop	Burger Joint	Café	Ice Cream Shop	Indian Restaurant	Restaurant	Malay Restaurant	Chinese Restaurant	Supermarket
4.690	Average	Malaysia	68000, Malaysia	3.207350	101.756821	2	Malay Restaurant	Asian Restaurant	Coffee Shop	Restaurant	Burger Joint	Café	Gym / Fitness Center	Chinese Restaurant	Boutique	Food Truck

CLuster 4

```
GKL_combine.loc[GKL_combine['Cluster Labels'] == 3, GKL_combine.columns[[1] + list(range(2, GKL_combine.shape[1]))]]
```

_PSF_RM	CRIMELEVEL	Country	CountryZip	Latitude	Longitude	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	9th Most Common Venue	10th Most Common Venue
3.130	High	Malaysia	51000, Malaysia	3.182488	101.694217	3	Chinese Restaurant	Asian Restaurant	Café	Indian Restaurant	Motorcycle Shop	Coffee Shop	Malay Restaurant	Park	Thai Restaurant	Hotel
1.885	Average, Low	Malaysia	52200, Malaysia	3.195448	101.611645	3	Chinese Restaurant	Café	Coffee Shop	Vegetarian / Vegan Restaurant	Asian Restaurant	Indian Restaurant	Food Truck	Pizza Place	Restaurant	Park
2.625	Average, Average, High, Average	Malaysia	57000, Malaysia	3.062847	101.689961	3	Café	Chinese Restaurant	Asian Restaurant	Japanese Restaurant	Coffee Shop	Restaurant	Vegetarian / Vegan Restaurant	Massage Studio	Bubble Tea Shop	Burger Joint
0.000	Average	Malaysia	58000, Malaysia	3.094503	101.669320	3	Chinese Restaurant	Café	Indian Restaurant	Asian Restaurant	Japanese Restaurant	Korean Restaurant	Pet Store	Coffee Shop	Restaurant	Residential Building (Apartment / Condo)
3.430	High	Malaysia	58100, Malaysia	3.100055	101.681501	3	Chinese Restaurant	Café	Asian Restaurant	Japanese Restaurant	Coffee Shop	Thai Restaurant	Restaurant	Korean Restaurant	Malay Restaurant	Pet Store
2.630	High	Malaysia	58200, Malaysia	3.085899	101.677512	3	Chinese Restaurant	Café	Asian Restaurant	Vegetarian / Vegan Restaurant	Coffee Shop	Restaurant	Japanese Restaurant	Korean Restaurant	Bakery	BBQ Joint

Then visualize the overall population and avg for rent listing price of selected cluster of neighborhoods(with complimentary nearby business) using seaborn scatterplot to identify the best location based on high population, low/average crime level and reasonable rent price.



Discussion:

Based on the defining categories, each cluster can be categorized as below:
Cluster 1: Crime level is average to high.
Mostly with Local ethnic/Asian Restaurants, Fast foods, Cafe/coffee shops.
Cluster 2: Crime level is average to high.
Mostly with Hotels, Bars, local ethnic/Asian Restaurants, Cafe/Coffee shops, Dessert shops.
Cluster 3: Crime level is low to average.
Mostly with Cafe/Coffee shops, local ethnic/Asian Restaurants, Dessert/Ice-cream shops, Yoga/ Gym & Fitness Center, Bakery, Spa/massage, Retails.
Cluster 4: Crime level is average to high.
Mostly with local Chinese/Indian/Japanese Restaurant, Cafe/Coffee shops.

From the segmented clusters above, the neighborhoods in cluster 3 is the most suitable as it has most of the surrounding business that can compliment to coffee business which are Yoga/fitness centre, retails, bakery & restaurants compare to other clusters as well as crime level is low to average too.

The scatter plot(Avg For Rent Listing Price(RM) Vs Overall Population of Neighborhoods in Cluster 3) shown neighborhood 'Ampang' has the highest population with moderate Rent price and average crime level, follow by 'Damansara Perdana' among the neighborhoods in cluster 3.

Based on the clustering neighborhood data and scatterplot on Avg Rent price vs Overall population, I recommend neighborhood 'Ampang' from cluster 3 as the best location to open new coffee shop. ¶

Conclusion:

In summary, using machine learning- Kmean clustering and data analysis with python on location data using foursquare API can be useful in segmenting/clustering the neighborhoods to identify the best location for opening a new coffee shop in Greater Kuala Lumpur. The result show neighborhood 'Ampang' is the best location as it has the highest population with reasonable rent price and average crime level as well as complimentary business among the neighborhoods in cluster 3. In near future, can include foot/vehicle traffic as one of the factor in deciding the location for opening new restaurant or coffee shop.

This project not only can help my friend to solve the location problem for opening his new coffee shop in Greater Kuala Lumpur, but it can be beneficial to those who is interested in opening new restaurant or exploring neighborhoods in Greater Kuala Lumpur(Kuala Lumpur & Petaling Jaya).