# Coursera Capstone

**IBM Applied Data Science Capstone**

# Opening a Thai Restaurant in Kuala Lumpur, Malaysia

**By: Arif Zahari**

**August 2020**

## Introduction: Business Problem

Malaysia is known as a food paradise with great diverse culinary experience. Thai cuisine is one of the most popular cuisines amongst Malaysians and foreigners in Kuala Lumpur, Malaysia.

In this project we will try to find an optimal location for a Thai restaurant. Specifically, this report will be targeted to stakeholders interested in opening a Thai restaurant in Kuala Lumpur, Malaysia. Since there are lots of restaurants in Kuala Lumpur, we will try to detect locations that are not already crowded with restaurants**.** We are also particularly interested in areas with no Thai restaurants in vicinity. We would also prefer locations as close to city center as possible**,** assuming that first two conditions are met.

We will use our data science powers to generate a few most promising neighborhoods based on these criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders

## Data

To solve the problem, we will need the following data:

• List of neighborhoods in Kuala Lumpur. This defines the scope of this project which is confined to the city of Kuala Lumpur, the capital city of the country of Malaysia in South East Asia.

• Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to Thai restaurants. We will use this data to perform clustering on the neighborhoods.

## Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur) contains a list of neighborhoods in Kuala Lumpur, with a total of 70 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the
neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.

Foursquare API will provide many categories of the venue data, we are particularly interested in the Thai restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

## Methodology

Firstly, we need to get the list of neighborhoods in the city of Kuala Lumpur. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur).
We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. 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 wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Kuala Lumpur.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood 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 use in clustering. Since we are analyzing the “Thai Restaurant” data, we will filter the “Thai Restaurant” as venue category for the neighborhoods.

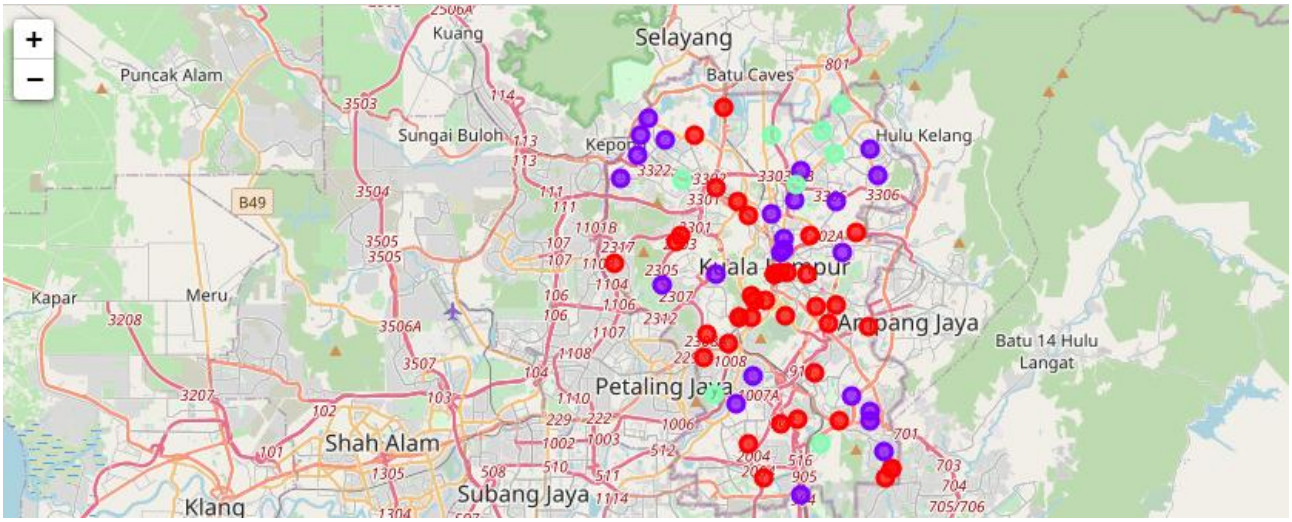
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Thai Restaurant”. The results will allow us to identify which neighborhoods have higher concentration of Thai restaurants while which neighborhoods have fewer number of Thai restaurants. Based on the occurrence of Thai restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Thai restaurants.

## Results

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

• Cluster 0: Neighborhoods with low to no existence number of Thai restaurants
• Cluster 1: Neighborhoods with moderate number of Thai restaurants
• Cluster 2: Neighborhoods with high concentration of Thai 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.




## Discussion

As observations noted from the map in the Results section, most of the Thai restaurants are concentrated in the suburb of Kuala Lumpur city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no Thai restaurants in the neighborhoods. This represents a great opportunity and high potential areas to open new Thai restaurants as there is very little to no competition from existing Thai restaurants. Meanwhile, shopping malls in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Thai restaurants. From another perspective, the results also show that the oversupply of Thai restaurants mostly happens in the suburb of the city, with the central city still with low number of few Thai restaurants. Therefore, this project recommends future Thai restaurant owners to capitalize on these findings to open new Thai restaurants in neighborhoods in cluster 0 with little to no competition. Thai restaurant developers with unique selling propositions to stand out from the competition can also open Thai restaurants in neighborhoods in cluster 1 with moderate competition. Lastly, Thai restaurant developers are advised to avoid neighborhoods in cluster 2 with high concentration of Thai restaurants and suffering from intense competition.

## Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of Thai restaurants, there are other factors such as population and income of residents that could influence the location decision of a new Thai restaurant. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by 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 Thai 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. Thai restaurant developers and investors regarding the best locations to open a new Thai 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 Thai restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Thai restaurant.

## References

Category:Suburbs in Kuala Lumpur. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs

# Appendix

### Cluster 0

| | Neighborhood | Thai Restaurant | Cluster Labels | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | Alam Damai | 0.00 | 0 | 3.057690 | 101.743880 |
| 51 | Shamelin | 0.01 | 0 | 3.124570 | 101.735970 |
| 25 | Federal Hill, Kuala Lumpur | 0.00 | 0 | 3.136370 | 101.685640 |
| 27 | Jalan Cochrane, Kuala Lumpur | 0.00 | 0 | 3.134630 | 101.721690 |
| 28 | Jalan Duta | 0.01 | 0 | 3.180025 | 101.677833 |
| 29 | Jinjang | 0.01 | 0 | 3.209500 | 101.658740 |
| 30 | KL Eco City | 0.01 | 0 | 3.117130 | 101.673840 |
| 31 | Kampung Baru, Kuala Lumpur | 0.01 | 0 | 3.165460 | 101.710280 |
| 32 | Kampung Datuk Keramat | 0.01 | 0 | 3.166400 | 101.730460 |
| 22 | Damansara, Kuala Lumpur | 0.00 | 0 | 3.138766 | 101.684015 |
| 37 | Lembah Pantai | 0.00 | 0 | 3.121189 | 101.663889 |
| 40 | Miharja | 0.01 | 0 | 3.147890 | 101.694050 |
| 41 | Mont Kiara | 0.00 | 0 | 3.165320 | 101.652430 |
| 53 | Sri Petaling | 0.01 | 0 | 3.072600 | 101.682520 |
| 43 | Pudu, Kuala Lumpur | 0.00 | 0 | 3.133540 | 101.713070 |
| 44 | Putrajaya | 0.01 | 0 | 3.125851 | 101.718513 |
| 45 | Salak South | 0.00 | 0 | 3.081540 | 101.696890 |
| 46 | Segambut | 0.01 | 0 | 3.186390 | 101.668100 |
| 52 | Sri Hartamas | 0.01 | 0 | 3.162200 | 101.650360 |
| 38 | Maluri | 0.01 | 0 | 3.147890 | 101.694050 |
| 21 | Damansara Town Centre | 0.00 | 0 | 3.136444 | 101.690294 |
| 24 | Desa Petaling | 0.00 | 0 | 3.083300 | 101.704380 |
| 12 | Bukit Bintang | 0.01 | 0 | 3.147770 | 101.708550 |
| 9 | Batu 11 Cheras | 0.01 | 0 | 3.061870 | 101.746750 |
| 10 | Batu, Kuala Lumpur | 0.01 | 0 | 3.147890 | 101.694050 |
| 6 | Bangsar | 0.00 | 0 | 3.129200 | 101.678440 |
| 5 | Bandar Tun Razak | 0.00 | 0 | 3.082800 | 101.722810 |
| 3 | Bandar Sri Permaisuri | 0.01 | 0 | 3.103910 | 101.712260 |
| 11 | Brickfields | 0.00 | 0 | 3.129160 | 101.684060 |
| 7 | Bangsar Park | 0.00 | 0 | 3.129200 | 101.678440 |
| 8 | Bangsar South | 0.01 | 0 | 3.111020 | 101.662830 |
| 66 | Taman Tun Dr Ismail | 0.01 | 0 | 3.152830 | 101.622710 |
| 15 | Bukit Nanas | 0.00 | 0 | 3.148609 | 101.699854 |
| 16 | Bukit Petaling | 0.00 | 0 | 3.129290 | 101.698960 |
| 68 | Taman Wahyu | 0.01 | 0 | 3.222400 | 101.671730 |
| 17 | Bukit Tunku | 0.00 | 0 | 3.173810 | 101.682760 |
| 18 | Cheras, Kuala Lumpur | 0.01 | 0 | 3.061870 | 101.746750 |
| 1 | Ampang, Kuala Lumpur | 0.01 | 0 | 3.148499 | 101.696728 |
| 13 | Bukit Jalil | 0.00 | 0 | 3.057810 | 101.689650 |

**Cluster 1**

| | Neighborhood | Thai Restaurant | Cluster Labels | Latitude | Longitude |
|---|---|---|---|---|---|
| 67 | Taman U-Thant | 0.020000 | 1 | 3.157700 | 101.724520 |
| 65 | Taman Taynton View | 0.020000 | 1 | 3.087070 | 101.736810 |
| 56 | Taman Connaught | 0.030000 | 1 | 3.082690 | 101.736890 |
| 55 | Taman Bukit Maluri | 0.020000 | 1 | 3.200660 | 101.633370 |
| 63 | Taman P. Ramlee | 0.030928 | 1 | 3.193940 | 101.705730 |
| 57 | Taman Desa | 0.030000 | 1 | 3.102970 | 101.684710 |
| 59 | Taman Len Seng | 0.020000 | 1 | 3.069080 | 101.742870 |
| 61 | Taman Midah | 0.020000 | 1 | 3.093590 | 101.728370 |
| 62 | Taman OUG | 0.020000 | 1 | 3.210051 | 101.634508 |
| 54 | Sungai Besi | 0.020000 | 1 | 3.049970 | 101.706030 |
| 50 | Setiawangsa | 0.030000 | 1 | 3.191802 | 101.740066 |
| 35 | Kepong Baru | 0.020000 | 1 | 3.207771 | 101.645173 |
| 48 | Sentul, Kuala Lumpur | 0.020000 | 1 | 3.175080 | 101.693050 |
| 2 | Bandar Menjalara | 0.030000 | 1 | 3.190350 | 101.625450 |
| 14 | Bukit Kiara | 0.030000 | 1 | 3.143480 | 101.644330 |
| 19 | Chow Kit | 0.020000 | 1 | 3.163590 | 101.698110 |
| 20 | Damansara Heights | 0.030000 | 1 | 3.147970 | 101.667950 |
| 23 | Dang Wangi | 0.020000 | 1 | 3.157825 | 101.697280 |
| 70 | Wangsa Maju | 0.030000 | 1 | 3.203870 | 101.737150 |
| 69 | Titiwangsa | 0.030000 | 1 | 3.180730 | 101.703210 |
| 36 | Kuchai Lama | 0.020000 | 1 | 3.090740 | 101.677330 |
| 39 | Medan Tuanku | 0.020000 | 1 | 3.159260 | 101.698340 |
| 47 | Semarak | 0.030000 | 1 | 3.179943 | 101.721449 |
| 34 | Kepong | 0.020000 | 1 | 3.217500 | 101.637630 |

**Cluster 2**

| | Neighborhood | Thai Restaurant | Cluster Labels | Latitude | Longitude |
|---|---|---|---|---|---|
| 58 | Taman Ibukota | 0.040000 | 2 | 3.21216 | 101.71540 |
| 33 | Kampung Padang Balang | 0.044444 | 2 | 3.20943 | 101.69318 |
| 49 | Setapak | 0.040000 | 2 | 3.18816 | 101.70415 |
| 26 | Happy Garden | 0.050000 | 2 | 3.20163 | 101.72107 |
| 64 | Taman Sri Sinar | 0.041667 | 2 | 3.19007 | 101.65293 |
| 42 | Pantai Dalam | 0.040000 | 2 | 3.09476 | 101.66747 |
| 4 | Bandar Tasik Selatan | 0.040404 | 2 | 3.07275 | 101.71461 |
| 60 | Taman Melati | 0.050000 | 2 | 3.22357 | 101.72399 |