**IBM Applied Data Science Capstone Project  
The competition of restaurants by district in Bangkok**

Kitithat Pansang  
March 31, 2021

**1. Introduction**

1.1 Background

Thailand is a paradise for tourism that is popular with people from all over the world. One of the key factors is the diversity of food and the culture of eating. Thai cooking places emphasis on lightly prepared dishes with strong aromatic components and a spicy edge. And there are also many nationalities of food in Thailand such as Japan, Korea, Europe, Asia, the Middle East.

1.2 Business Problem

Longyang is a name of Thai restaurant that is mean “Have you tried?”. It is a restaurant that sells Thai and Chinese food. At present, the second generation of heirs is inheriting the business. The owners are deciding to expand their food sales category between Japan and Traditional Thai food, which will open next to the original store at Bang Wa district. She is working on various factors that help make decisions, and one of them is a competition in the neighborhood. In additional, it is also a good idea to look for another location that will open new restaurants in the future in the new area.

To address this business problem, research will be done using data science methodology and machine learning techniques such as clustering to find the answers.

**2. Data acquisition and cleaning**

2.1 Data sources

* The **5o** **Districts and 180 Subdistricts of Bangkok** which can be obtained from Wiki page “Khwaeng”.
* **Geospatial** of each location can use GeoPy package to get latitude and longitude base on subdistricts name.
* The **Foursquare API** was used to add venue data for the neighborhoods such as number of restaurants, category, etc.

1.2 Data cleaning

The **5o** **Districts and 180 Subdistricts of Bangkok** table included name of districts and subdistricts that use to get **Geospatial** of each location and combine into one table. But some names may be specific, which cannot be obtained from GeoPy package. Therefore, need to explore the deficiency and find another source such as google map.

The data that obtained form **Foursquare API** will be in the form of JSON file and in order to be able to work, it must be adjusted to homogeneous dataset in the form of a table which will later on be used for the k-means clustering technique.

**3. Methodology**

3.1 Data loading

First, Scrapping from the Wikipedia that contains Districts and Subdistricts of Bangkok. The table use attribute “rowspan” to merge the table like pivot table for easy to explore. Show example below:

Table

Description automatically generated

Source: <https://en.wikipedia.org/wiki/Khwaeng>

After scraping the data, we will obtain 5o districts and 180 subdistricts of Bangkok. Packing it into dataframe. (Note: I intend to get name in Thai language. To be used in the future.)

Table

Description automatically generated

Next, use name of subdistricts for search latitude and longitude by using GeoPy python package and check whether the information has been completed or not. Found that the dat had not been completed in all 4 subdistricts. Therefore need to find more from Google map.

Graphical user interface, text

Description automatically generated with medium confidence

After that, display all location by using folium python package in map to see the distribution of the area and check the data set is correct by seeing that it covers all over Bangkok.

Map

Description automatically generated

3.2 Venue Data from FourSquare API

The FourSquare API was called for each rastuarant. Our FourSquare request was limited to just the top 100 most common venues per request because that is more than sufficient for our analysis, and to avoid cap on daily request size. Our FourSquare request was also limited to 800 meters and set query category to “Food”.

Graphical user interface, text, application, chat or text message

Description automatically generated

A quick sanity check reveals that we start with 88 unique venue categories.

Text, letter

Description automatically generated

3.3 Cleaning Venue Data

First, venue data has several categories of restaurant. So, we will be grouping category that same nationality such as sushi restaurant, Japanese restaurant or Udon Restausrant to Japanese, so it can reduce to 65 categories.



Create a temporary analysis DataFrame with just the subdistript along with Food Venue Categories. Each of the rows represents a Venue we are interested in (restaurant related). Within each row is the borough and neighborhood for that restaurant as well as a “1” put in the Venue Category column applicable to that restaurant.

A screenshot of a computer

Description automatically generated with medium confidence

next step where the DataFrame of all Food Venue is collapsed down into amount of subdistrict, one row each. The Venue Category columns all remain, but the individual 1’s are all rolled up int means (percent frequency) per row.



**Japanese restaurants**

**Thai restaurants**

3.4 Rank Order Venue Categories

In this section, we will create a table indicating the order of the number of 10 descending stores in each region. Making it possible to compare between different categories of restaurants in the same subdistrict.

Table

Description automatically generated

3.5 Use K-Means to Generate 5 Clusters

Next, the matrix of Venue Categories was fed into the K-means data model engine to try and find any patterns or relationships between neighborhoods based on the types of venues operating within. In this paper, we use scikit-learn python package to train model, the K-Means returns an array of cluster labels (numbers 0 thru 4) corresponding to the neighborhoods passed in.

Graphical user interface, text, application

Description automatically generated

**4.** Results

3.4 Visual frequency Japanese and Thai restaurants

A picture containing background pattern

Description automatically generatedBackground pattern

Description automatically generatedAfter creating the frequency table, we can draw a graph in descending order to compare between Thai food and Japanese food. We compared to all over Bangkok, Japanese restaurants look a little more interesting in our subdistrict, Bang Wa. This is because it has less frequency proportions and a lower order than Thai food. is less than Thai food.

**Japanese restaurants**

**Thai restaurants**

Form rank order venue categories table, we will see our subdistrict, Thai restaurant is the most common venue in Bang Wa and Japanese restaurants falling at sixth. This supports our previous opinion that in this area the competition of Thai restaurants is more intense than Japanese restaurants.

Table

Description automatically generated with low confidence

Finally display the results of the clustering on the map. Which was found to have dispersed of each group throughout Bangkok. Therefore, in the future, when there is a need to expand our branches, we can come to the area that is in the same group as ours first so as not to waste time looking around Bangkok.

Chart, map

Description automatically generated