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 xxxxxxx. 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:

Khwaeng of Bangkok [edit]

District (*Khet*)			Subdistrict (*Khwaeng*)			Notes
Code	Name	Name (Thai Alphabet)	Code	Name	Name (Thai Alphabet)	
01	Phra Nakhon	พระนคร	01	Phra Borom Maha Ratchawang	พระบรมมหาราชวัง	
			02	Wang Burapha Phirom	วังบูรพาภิรมย์	
			03	Wat Ratchabophit	วัดราชบพิธ	
			04	Samran Rat	สำราญราษฎร์	
			05	San Chaopho Suea	ศาลเจ้าพ่อเสือ	
			06	Sao Chingcha	เสาชิงช้า	Seat of BMA Office
			07	Bowon Niwet	บวรนิเวศ	
			08	Talat Yot	ตลาดยอด	
			09	Chana Songkhram	ชนะสงคราม	
			10	Ban Phan Thom	บ้านพานถม	
			11	Bang Khun Phrom	บางขุนพรหม	
			12	Wat Sam Phraya	วัดสามพระยา	Seat of the District Office
02	Dusit	ดุสิต	01	Dusit	ดุสิต	Seat of the District Office
			02	Wachiraphayaban	วชิรพยาบาล	
			03	Suan Chitlada	สวนจิตรลดา	
			04	Si Yaek Maha Nak	สี่แยกมหานาค	
			06	Thanon Nakhon Chai Si	ถนนนครไชยศรี	
03	Nong Chok	หนองจอก	01	Krathum Rai	กระทุ่มราย	Seat of the District Office
			02	Nong Chok	หนองจอก	
			03	Khlong Sip	คลองสิบ	

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.)

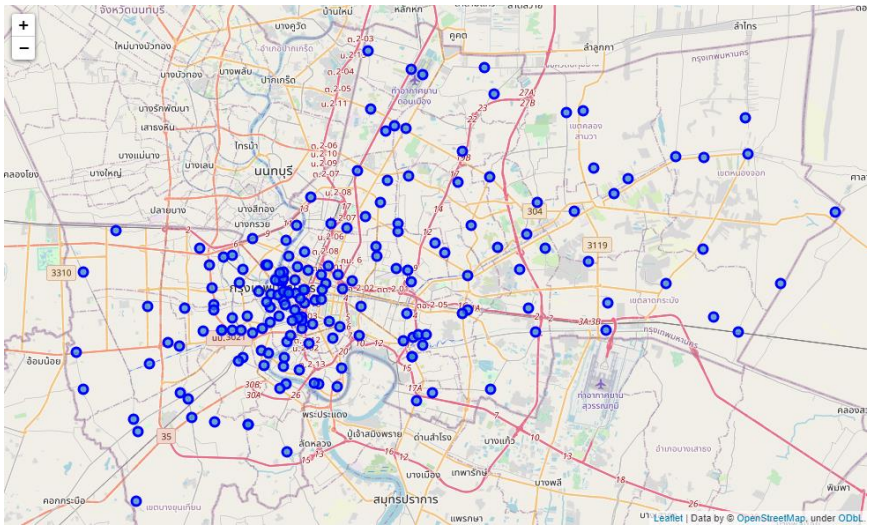
	Subdistrict	Subdistrict_in_Thai	District	District_in_Thai	Latitude	Longitude
0	Phra Borom Maha Ratchawang	พระบรมมหาราชวัง	Phra Nakhon	พระนคร	None	None
1	Wang Burapha Phirom	วังบูรพาภิรมย์	Phra Nakhon	พระนคร	None	None
2	Wat Ratchabophit	วัดราชบพิธ	Phra Nakhon	พระนคร	None	None
3	Samran Rat	สำราญราษฎร์	Phra Nakhon	พระนคร	None	None
4	San Chaopho Suea	ศาลเจ้าพ่อเสือ	Phra Nakhon	พระนคร	None	None
...	...	...	...	...	...	...
175	Thung Khru	ทุ่งครุ	Thung Khru	ทุ่งครุ	None	None
176	Bang Bon Nuea	บางบอนเหนือ	Bang Bon	บางบอน	None	None
177	Bang Bon Tai	บางบอนใต้	Bang Bon	บางบอน	None	None
178	Khlong Bang Phran	คลองบางพราน	Bang Bon	บางบอน	None	None
179	Khlong Bang Bon	คลองบางบอน	Bang Bon	บางบอน	None	None

180 rows × 6 columns

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.

	Subdistrict	Subdistrict_in_Thai	District	District_in_Thai	Latitude	Longitude
13	Wachiraphayaban	วชิรพยาบาล	Dusit	ดุสิต	None	None
117	Chomphon	จอมพล	Chatuchak	จตุจักร	None	None
137	Sanambin	สนามบิน	Don Mueang	ดอนเมือง	None	None
163	Khlong Chaokhun Sing	คลองเจ้าคุณสิงห์	Wang Thonglang	วังทองหลาง	None	None

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.



3.2 Venue Data from FourSquare API

The FourSquare API was called for each Seattle neighborhood zip code. 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”.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Phra Borom Maha Ratchawang	13.749994	100.49175	Ama Art & Eatery	13.746449	100.490782	Thai Restaurant
1	Phra Borom Maha Ratchawang	13.749994	100.49175	Home Café Tha Tien (โฮม คาเฟ่ ท่าเตียน)	13.745696	100.491260	Café
2	Phra Borom Maha Ratchawang	13.749994	100.49175	A Pink Rabbit + Bob	13.745032	100.492030	Café
3	Phra Borom Maha Ratchawang	13.749994	100.49175	Supanniga Eating Room x Roots Coffee (ห้องทานข...	13.744162	100.491700	Thai Restaurant
4	Phra Borom Maha Ratchawang	13.749994	100.49175	Tonkin - Annam (ตงกิง - อันนัม)	13.745103	100.491297	Vietnamese Restaurant

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

```
Number of unique categories: 88
['Thai Restaurant' 'Café' 'Vietnamese Restaurant' 'Japanese Restaurant'
 'Restaurant' 'Noodle House' 'Bakery' 'Soup Place' 'Chinese Restaurant'
 'Wings Joint' 'Asian Restaurant' 'BBQ Joint' 'Dumpling Restaurant'
 'Burger Joint' 'Fast Food Restaurant' 'Fried Chicken Joint' 'Food Truck'
 'Italian Restaurant' 'Som Tum Restaurant' 'Diner' 'Breakfast Spot'
 'Indian Restaurant' 'Seafood Restaurant' 'Snack Place'
 'Hotpot Restaurant' 'Food Court' 'Dim Sum Restaurant' 'Cafeteria'
 'Halal Restaurant' 'Sandwich Place' 'Steakhouse' 'Food Stand'
 'American Restaurant' 'Hainan Restaurant' 'Food'
 'Modern European Restaurant' 'French Restaurant' 'Buffet'
 'German Restaurant' 'Tapas Restaurant' 'Deli / Bodega' 'Pizza Place'
 'Scandinavian Restaurant' 'Vegetarian / Vegan Restaurant' 'Irish Pub'
 'Sushi Restaurant' 'Israeli Restaurant' 'Korean Restaurant' 'Salad Place'
 'Bistro' 'Japanese Curry Restaurant' 'Satay Restaurant' 'Creperie'
 'Shabu-Shabu Restaurant' 'Comfort Food Restaurant' 'Donut Shop'
 'Yoshoku Restaurant' 'Ramen Restaurant' 'Donburi Restaurant'
 'Tonkatsu Restaurant' 'Greek Restaurant' 'Middle Eastern Restaurant'
 'Shanghai Restaurant' 'Malay Restaurant' 'Taiwanese Restaurant'
 'Gastropub' 'South Indian Restaurant' 'Lebanese Restaurant'
 'New American Restaurant' 'Turkish Restaurant' 'Pet Café'
 'Mediterranean Restaurant' 'Cantonese Restaurant' 'Taco Place'
 'Molecular Gastronomy Restaurant' 'Mexican Restaurant'
 'Fish & Chips Shop' 'Bagel Shop' 'Yakitori Restaurant'
 'Spanish Restaurant' 'Portuguese Restaurant'
 'Southern / Soul Food Restaurant' 'Sri Lankan Restaurant'
 'Udon Restaurant' 'African Restaurant' 'Cajun / Creole Restaurant'
 'Szechuan Restaurant' 'Australian Restaurant']
```

3.3 Cleaning Venue Data

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

```
group_catagory = { 'American': ['American Restaurant','New American Restaurant'],
                   'Bakery': ['Bagel Shop','Bakery','Donut Shop'],
                   'Cafe': ['Cafeteria','Café','Creperie','Pet Café'],
                   'Chinese': ['Cantonese Restaurant','Chinese Restaurant','Dim Sum Restaurant','Dumpling Restaurant',
                               'Shanghai Restaurant','Taiwanese Restaurant'],
                   'Japan': ['Donburi Restaurant','Yoshoku Restaurant','Yakitori Restaurant','Udon Restaurant',
                             'Tonkatsu Restaurant','Japanese Curry Restaurant','Japanese Restaurant','Ramen Restaurant',
                             'Sushi Restaurant'],
                   'Bar': ['Gastropub','Irish Pub','Bistro'],
                   'Thai': ['Som Tum Restaurant','Thai Restaurant'],
                   'Indian': ['Indian Restaurant' ,'South Indian Restaurant']
                  }
```

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.

	Neighborhood	African Restaurant	American	Asian Restaurant	Australian Restaurant	BBQ Joint	Bakery	Bar	Breakfast Spot	Buffet	...	Sri Lankan Restaurant	Steakhouse	Szechuan Restaurant	Taco Place
0	Phra Borom Maha Ratchawang	0	0	0	0	0	0	0	0	0	...	0	0	0	0
1	Phra Borom Maha Ratchawang	0	0	0	0	0	0	0	0	0	...	0	0	0	0
2	Phra Borom Maha Ratchawang	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	Phra Borom Maha Ratchawang	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	Phra Borom Maha Ratchawang	0	0	0	0	0	0	0	0	0	...	0	0	0	0

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.

	Neighborhood	African Restaurant	American	Asian Restaurant	Australian Restaurant	BBQ Joint	Bakery	Bar	Breakfast Spot	Buffet	...	Sri Lankan Restaurant	Steakhouse	Szechuan Restaurant
0	Anusawari	0.0	0.000000	0.057143	0.0	0.028571	0.057143	0.000000	0.000000	0.000000	...	0.0	0.028571	0.0
1	Arun Ammarin	0.0	0.000000	0.108696	0.0	0.021739	0.000000	0.000000	0.000000	0.000000	...	0.0	0.021739	0.0
2	Ban Bat	0.0	0.013158	0.157895	0.0	0.000000	0.026316	0.000000	0.013158	0.000000	...	0.0	0.000000	0.0
3	Ban Chang Lo	0.0	0.000000	0.052632	0.0	0.000000	0.039474	0.000000	0.000000	0.000000	...	0.0	0.052632	0.0
4	Ban Phan Thom	0.0	0.000000	0.096774	0.0	0.010753	0.043011	0.010753	0.010753	0.000000	...	0.0	0.010753	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
170	Wat Sommanat	0.0	0.015385	0.061538	0.0	0.015385	0.015385	0.000000	0.015385	0.000000	...	0.0	0.046154	0.0
171	Wat Tha Phra	0.0	0.000000	0.142857	0.0	0.071429	0.000000	0.000000	0.000000	0.000000	...	0.0	0.000000	0.0
172	Wat Thep Sirin	0.0	0.000000	0.156863	0.0	0.019608	0.000000	0.000000	0.000000	0.019608	...	0.0	0.000000	0.0
173	Wong Sawang	0.0	0.000000	0.041667	0.0	0.041667	0.125000	0.000000	0.000000	0.000000	...	0.0	0.083333	0.0
174	Yan Nawa	0.0	0.000000	0.066667	0.0	0.066667	0.066667	0.000000	0.000000	0.000000	...	0.0	0.000000	0.0

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.

	Neighborhood	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	Anusawari	Noodle House	Fast Food Restaurant	Thai	Japan	Hotpot Restaurant	Food Court	Asian Restaurant	Bakery	Steakhouse	Restaurant
1	Arun Ammarin	Noodle House	Thai	Asian Restaurant	Chinese	Japan	Food Truck	Shabu-Shabu Restaurant	Hotpot Restaurant	Seafood Restaurant	Pizza Place
2	Ban Bat	Noodle House	Asian Restaurant	Thai	Chinese	Fast Food Restaurant	Seafood Restaurant	Bakery	Cafe	Japan	American
3	Ban Chang Lo	Thai	Noodle House	Cafe	Japan	Asian Restaurant	Steakhouse	Bakery	Fried Chicken Joint	Food Court	Vietnamese Restaurant
4	Ban Phan Thom	Thai	Noodle House	Cafe	Asian Restaurant	Vegetarian / Vegan Restaurant	Bakery	Chinese	Italian Restaurant	Indian	Israeli Restaurant

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.

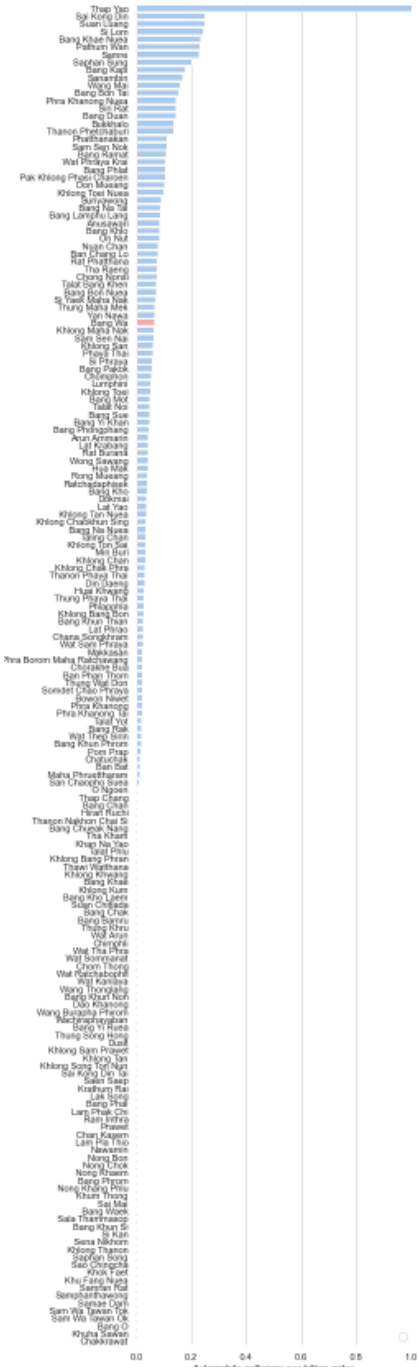
```
# Run k-means to cluster the neighborhood into 5 clusters
kclusters = 5  # set number of clusters

bkk_grouped_clustering = bkk_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(bkk_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([2, 0, 0, 3, 2, 1, 1, 0, 1, 1])
```

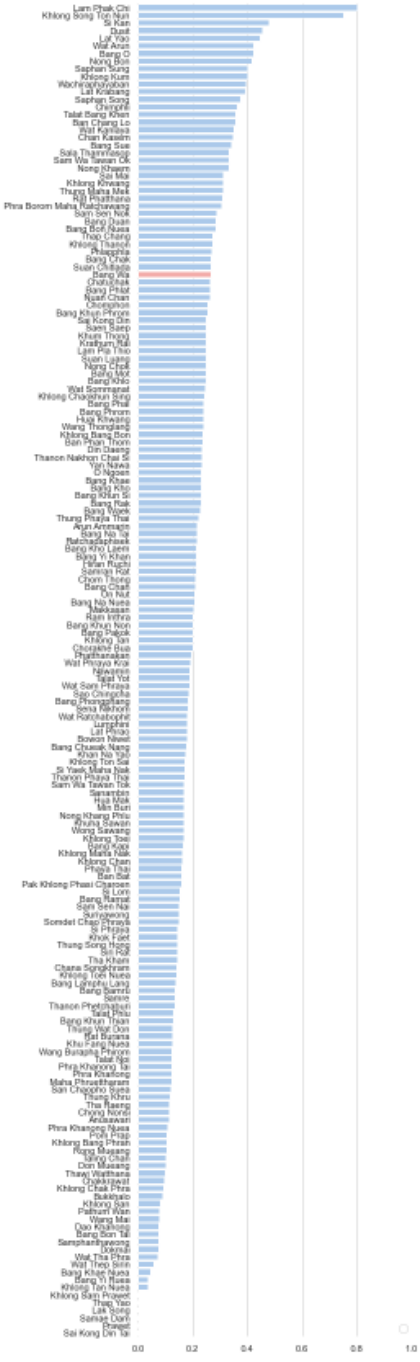
4. Results

3.4 Visual frequency Japanese and Thai restaurants

After 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.

	Neighborhood	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
35	Bang Wa	Thai	Asian Restaurant	Noodle House	Chinese	Steakhouse	Japan	Bakery	Seafood Restaurant	Middle Eastern Restaurant	Modern European Restaurant

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.