Finding the best district in Bangkok to open a foreign restaurant

Jeremy Kulcsar

June 10th, 2020

1. Introduction

1.1 Background

Bangkok, Thailand, has always been a city of wonders where anything could be found. There is everything for everyone: restaurants, street food, clubs, bars, temples, museums, malls, parks, rooftops, pools, sun... And it is what makes it the most attractive place in the world. With 22.7 million international tourists every year, Bangkok ranks among the most visited cities.

1.2 Problem

Opening a restaurant as an expat could be a good idea: other expats of the same nation would come to eat their native food, and thai locals would come to try our exotic food. However, with 8.3 million residents and a surface of 1569km², Bangkok is a huge and dense city, and each of the 50 districts has its own demographics and specialties.

1.3 Interest

The aim of this data science project is to help locate the best district to open our restaurant. This could help any expat who wants to open a new restaurant in Bangkok.

2. Data sources, visualization & preprocessing

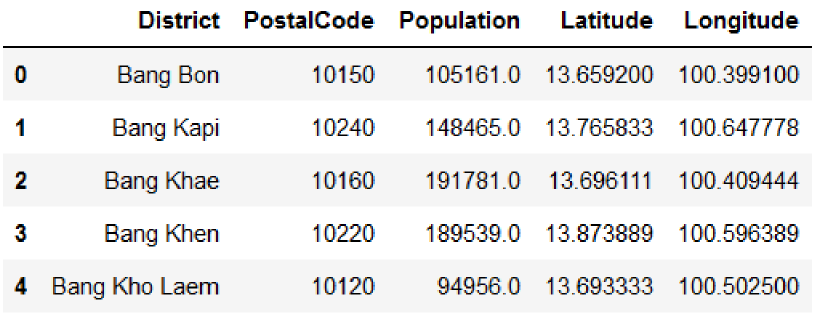
2.1 Data sources

2.1.1 Wikipedia

Wikipedia provides us a table containing all the necessary data related to the districts of Bangkok, namely:

- The Name
- The postal code
- The population
- The coordinates

As this data is displayed in an HTML table, some scraping will be required to gather this as a dataframe for our notebook (fig 1).



	District	PostalCode	Population	Latitude	Longitude
0	Bang Bon	10150	105161.0	13.659200	100.399100
1	Bang Kapi	10240	148465.0	13.765833	100.647778
2	Bang Khae	10160	191781.0	13.696111	100.409444
3	Bang Khen	10220	189539.0	13.873889	100.596389
4	Bang Kho Laem	10120	94956.0	13.693333	100.502500

Figure 1: Bangkok's districts dataframe

2.1.2 Foursquare

As Bangkok is a very metropolitan city with a very diversified population and choice of restaurants, we can make the hypothesis that the nature of the demographics in the district doesn't matter, and the nature of the venues is enough to classify the districts.

The Foursquare location data is therefore enough for this study: we will be able to locate which places are the most suited for opening a restaurant based on their activity there: do they already have many restaurants at that location? Is it rather a coffee/bar street, a restaurant one, or a club one? etc...

By exploring the different features accessed through the Foursquare API, we will be able to target and recommend locations that guarantee a profitable restaurant business.

2.2 Visualization

As we're trying to tackle a geographical problem here, our tool of choice will be folium for map visualization. We will use it to see the location of the districts on a map (fig 2), and then see where each district is and in which cluster it belongs.

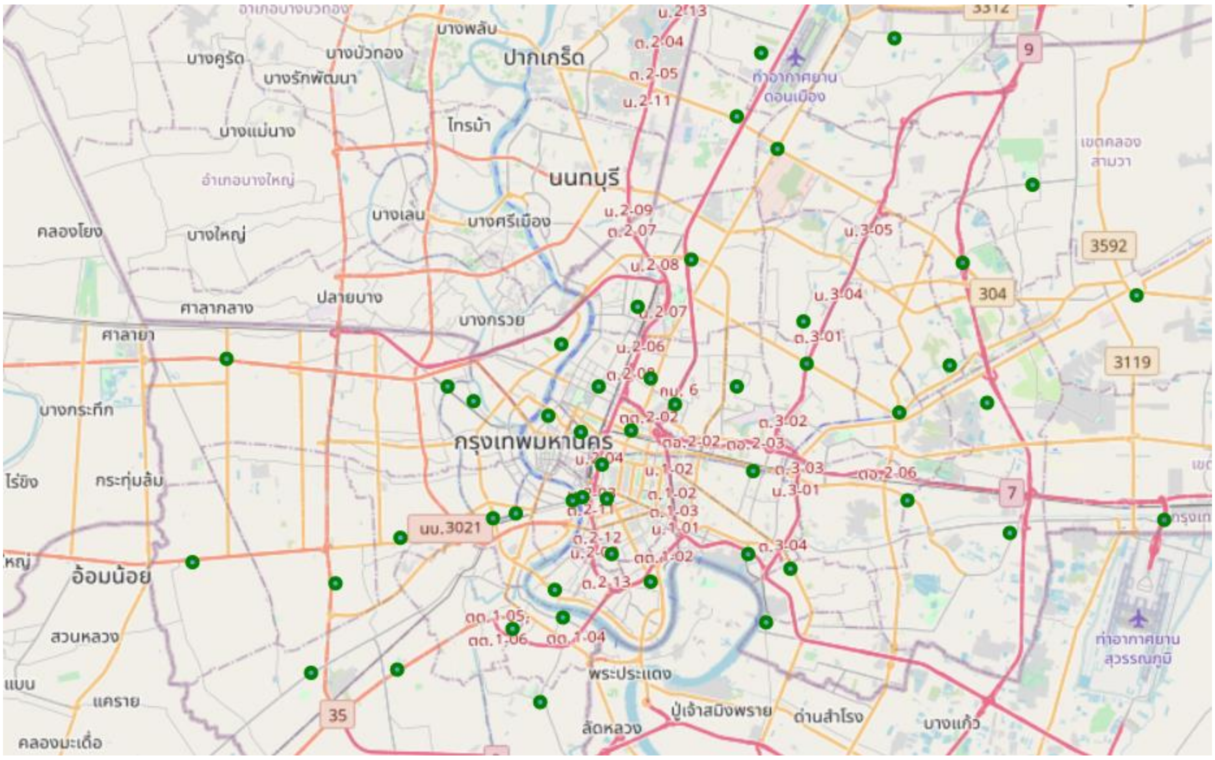


Figure 2: A map of the districts of Bangkok

3. Methodology

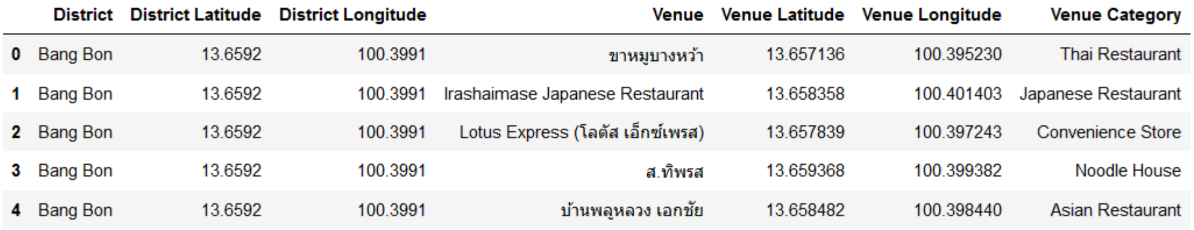
3.1 Approach

The idea here is to use a clustering algorithm to separate the districts by type. Intuition tells us that we might find typical types of district, such as commercial ones, residential ones, etc… Once we're done clustering the district, we pick the most suited one within the good cluster based on different real-life factors such as the population or the economy of that district.

3.2 Data preprocessing

Our starting point isn't enough to make a clustering algorithm, we need to proceed to some data preprocessing.

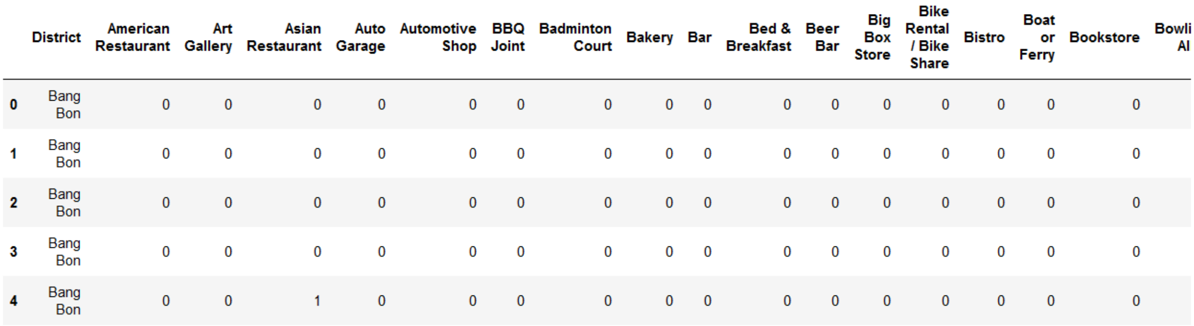
First, we will need to merge the foursquare data with our dataframe. The choice here was to select the top 100 venues of a district that are within a 500 miles radius. The result is a dataframe of 1147 rows and 7 columns (fig 3).



	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bang Bon	13.6592	100.3991	ขาหมูบางหว้า	13.657136	100.395230	Thai Restaurant
1	Bang Bon	13.6592	100.3991	Irashaimase Japanese Restaurant	13.658358	100.401403	Japanese Restaurant
2	Bang Bon	13.6592	100.3991	Lotus Express (โลตัส เอ็กซ์เพรส)	13.657839	100.397243	Convenience Store
3	Bang Bon	13.6592	100.3991	ส.ทิพรส	13.659368	100.399382	Noodle House
4	Bang Bon	13.6592	100.3991	บ้านพลูหลวง เอกชัย	13.658482	100.398440	Asian Restaurant

Fig 3: Dataframe of top 100 venues for each district

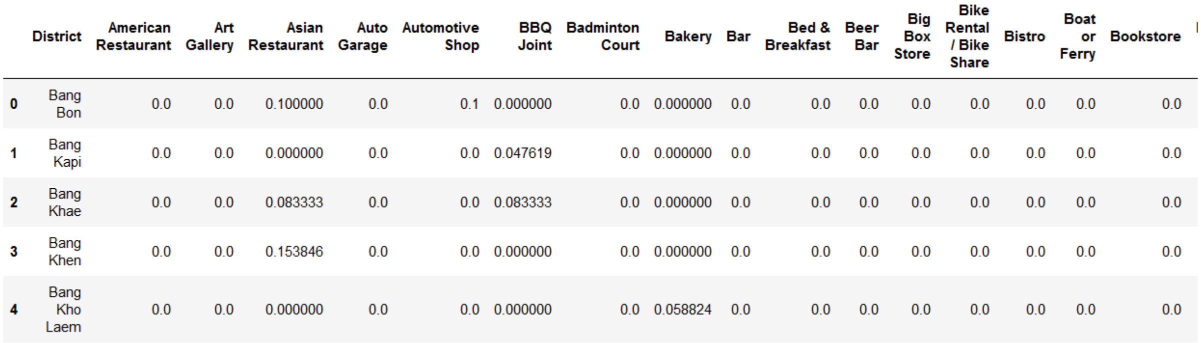
Then, as we’re facing a lot of textual data, we will perform a One-Hot Encoding in order to create a bitmap out of this dataframe. Fig 4 shows us the result, which is a dataframe with 1147 rows and 161 columns.



	District	American Restaurant	Art Gallery	Asian Restaurant	Auto Garage	Automotive Shop	BBQ Joint	Badminton Court	Bakery	Bar	Bed & Breakfast	Beer Bar	Big Box Store	Bike Rental / Bike Share	Bistro	Boat or Ferry	Bookstore	Bowli Al
0	Bang Bon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Bang Bon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Bang Bon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Bang Bon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Bang Bon	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 4: One-Hot encoded dataframe

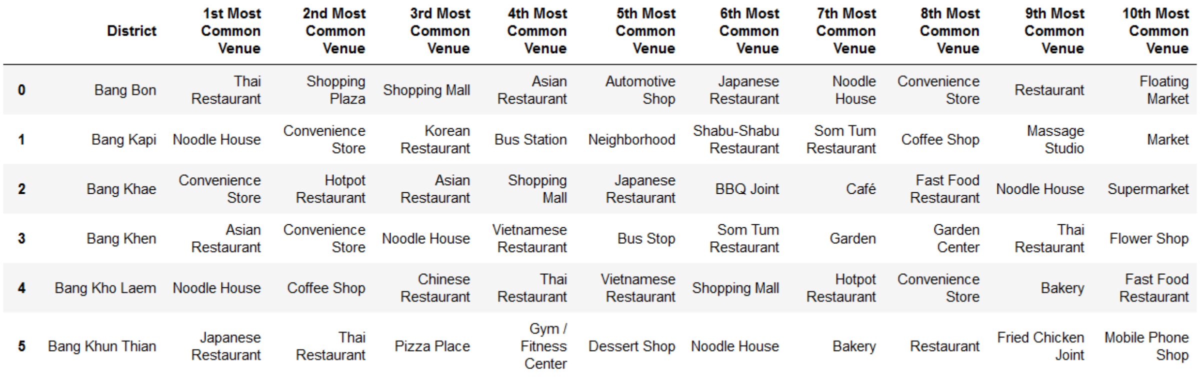
Then, we group the rows by district and we aggregate the attributes by mean frequency of occurrence, to show which type of venue appears the most for each district. Fig 5 shows us the result.



	District	American Restaurant	Art Gallery	Asian Restaurant	Auto Garage	Automotive Shop	BBQ Joint	Badminton Court	Bakery	Bar	Bed & Breakfast	Beer Bar	Big Box Store	Bike Rental / Bike Share	Bistro	Boat or Ferry	Bookstore	I
0	Bang Bon	0.0	0.0	0.100000	0.0	0.1	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	Bang Kapi	0.0	0.0	0.000000	0.0	0.0	0.047619	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	Bang Khae	0.0	0.0	0.083333	0.0	0.0	0.083333	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	Bang Khen	0.0	0.0	0.153846	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	Bang Kho Laem	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.058824	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Figure 5: Grouped dataframe

Finally, we sorted the venues in a descending order in order to create a suitable dataframe for the clustering, as it shows the 10 first most occurring venue for each district (fig 6).



	District	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	Bang Bon	Thai Restaurant	Shopping Plaza	Shopping Mall	Asian Restaurant	Automotive Shop	Japanese Restaurant	Noodle House	Convenience Store	Restaurant	Floating Market
1	Bang Kapi	Noodle House	Convenience Store	Korean Restaurant	Bus Station	Neighborhood	Shabu-Shabu Restaurant	Som Tum Restaurant	Coffee Shop	Massage Studio	Market
2	Bang Khae	Convenience Store	Hotpot Restaurant	Asian Restaurant	Shopping Mall	Japanese Restaurant	BBQ Joint	Café	Fast Food Restaurant	Noodle House	Supermarket
3	Bang Khen	Asian Restaurant	Convenience Store	Noodle House	Vietnamese Restaurant	Bus Stop	Som Tum Restaurant	Garden	Garden Center	Thai Restaurant	Flower Shop
4	Bang Kho Laem	Noodle House	Coffee Shop	Chinese Restaurant	Thai Restaurant	Vietnamese Restaurant	Shopping Mall	Hotpot Restaurant	Convenience Store	Bakery	Fast Food Restaurant
5	Bang Khun Thian	Japanese Restaurant	Thai Restaurant	Pizza Place	Gym / Fitness Center	Dessert Shop	Noodle House	Bakery	Restaurant	Fried Chicken Joint	Mobile Phone Shop

Figure 6: Final dataframe

As of now, we were ready to perform a clustering of the data in order to differentiate the different districts and pick the cluster that is the most appropriate for the problem.

3.3 Clustering

As we want to go from a given number of clusters to perform the separation, we are going to use a k-means algorithm.

To know the optimal value for k, the most common technique is to use the elbow method. However, fig 7 shows us that the given dataset doesn't allow us to perform this method. In fact, the distortion seems to be decreasing linearly as the value of k raises.

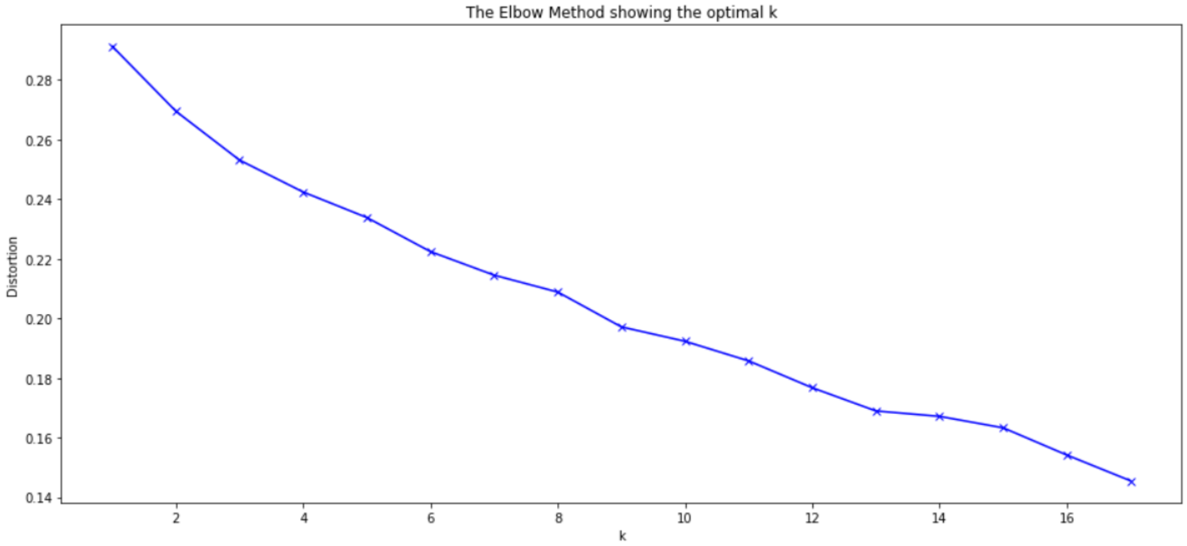


Figure 7: Elbow Method for k

This means there is no real optimal value for k. We can either see it pessimistically and conclude that there is no good clustering, or see it optimistically and conclude that any value of k could generate a valid clustering. Let's be optimistic.

As there are 50 districts and we will have to analyse each cluster individually for our business problem, a lower value of k is preferable. On the other hand, it is better to keep it high enough to have a useful clustering of the data. Some tests with the map rendering show us that a good value of k for this purpose is 4.

4. Results

The clustering gave us 4 different clusters, in which the districts were distributed (fig 8). We can already notice 3 big clusters, and 1 cluster of only one district. This specific cluster seems to appear even with lower values such as k=3, so it will be treated as an outsider.

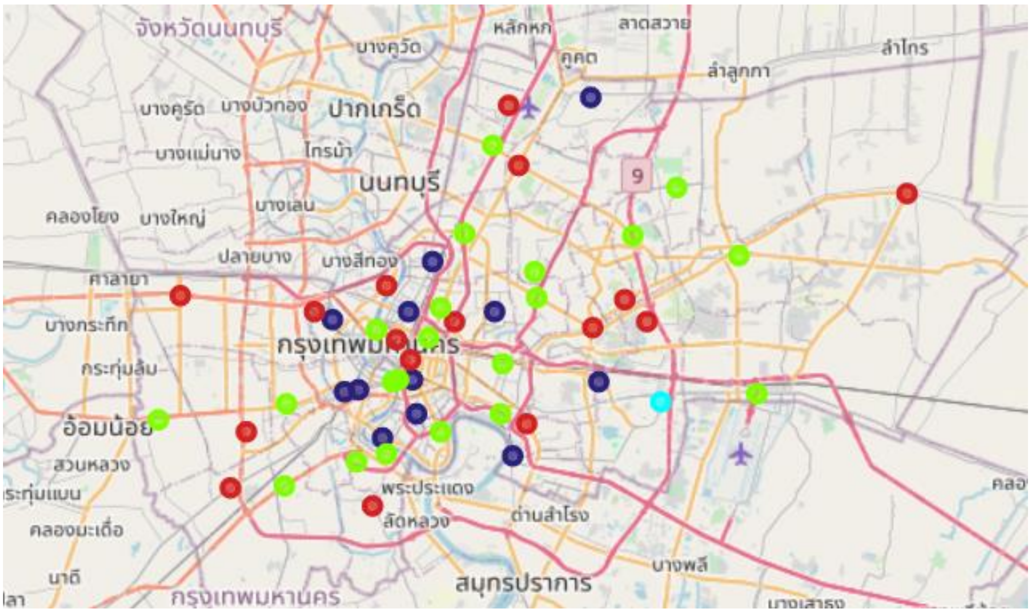


Figure 8: Districts with their respective clusters

4.1 Cluster 0: The outsider

This cluster is the outsider that we will ignore.

4.2 Cluster 1: The dynamic districts

This seems to be the cluster with the most diversified venues, either coffee shops, art galleries, noodle houses, markets... This means these districts are probably the ones with the highest economic activity.

As one of Bangkok's main sources of income is tourism and expat businesses, the appropriate demographics to open a non-local restaurant might be found there. Moreover, many of the district in this cluster are where you'll find popular touristic markets or huge malls (Chatuchak, Phayathai, Ratchathewi...).

4.3 Cluster 2: The residential areas

Cluster of convenience stores, hotels and basic entertainement. Seems like it could be from a residential local area.

4.4 Cluster 3: The food areas

This cluster is mainly focused on noodle houses. In Bangkok, noodle houses are small local restaurants, often held by a couple or a family, where the capacity is around 20 people. Therefore, these are district where people go to eat. Local food-focused districts. This isn't what we're looking for either.

5. Discussions

We are looking for a district in which we can open our restaurant. Since we aren't locals, we need a district popular to the expats, and preferably dynamic. This brings us to cluster 3 (label 2). Any district of this cluster close to the center of Bangkok might be a fit for what we're looking for.

Let's aim for the districts with the highest number of hostels/hotels. **Phayathai (10400)** seems to be a good pick: it has hostels and hotels at the 3rd and 4th places, which means it hosts many tourists, and steakhouses (something that isn't local) are among the most common venues.

6. Conclusion

In this study, I analysed the different districts of Bangkok by separating them by nature, and then incorporated real-life questions in order to determine which one is the most suited to open a foreign restaurant. I collected the data from public sources such as Wikipedia and performed a clustering algorithm to separate the district. This project could be of use for anyone willing to open a foreign restaurant in Bangkok and wondering which district could be classified as residential, dynamic or local food area.