Tourist Recommendation Based on K-means Clustering

1.Introduction

Bangkok is a big and abundant capital city of Thailand which it has food, entertainment, night-life and etc. providing for foreign visitors. However, each district of the city has their own strength and interesting point to attract tourists.

This project points to solve problem as an assistance for tourists who want to prepare their plan for visiting Bangkok by providing district which suit with their travel styles. For example, a tourist would like to visit to Thailand for eating street food, the cluster will provide result about district which outstanding on noodle houses and Thai restaurant.

2. Data

- **Districts in Bangkok, Thailand:** this data provides information about districts name, latitudes and longitudes of each district, population and postal codes. The information can be found on this link: https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok

- **Information of venues in each district:** this data provides venues locate in each district of Bangkok, Thailand, including Venue, Venue Latitude, Venue Longitude and Venue Category. The information is retrieved by using **Foursquare** by going to this link: https://developer.foursquare.com/

3. Methodology

In this project, we would like to know outstanding point of each area from clustering analysis. Regarding to limit of my computer performance, I use the data from only 500 meters from the center of each district in Bangkok, Thailand.

3.1 Data cleaning

On the first dataset (Districts in Bangkok), it provides enough and clear information for using in analysis and clustering steps, so there is no need to be clean and adjust in this dataset. However, the second data about venues in each district need to be adjusted before going through the next process. Fortunately, both of dataset do not have any missing value or Nan, hence we do not have to cover part of cleaning missing value about of the datasets.

Due to good condition of the first dataset, this section will only focus on the second dataset. The first step of cleaning the dataset is to group the synonym of each category. In this data, it is assign manually by human, so sometimes they use different word to describe the same thing, for example, the category is coffee shop but there are two other synonym as café and cafeteria, so we have to deal with this by grouping them into one category as coffee shop.

The second step is to clean irrelevant category on travel topic out of the feature list. The dataset contains the several stuffs, so we have to drop it to reduce noise of our clustering. From this point, the second dataset is ready to use for constructing feature lists in the next step.

3.2 Constructing Feature Lists

To constructing feature lists for clustering, in this project, we leverage features as means of venue frequencies in each district or percentage of venues in each district. In converting the data, at first, we have to replace categorical data with dummy by using one hot technique then perform, then calculating mean of venues in each district, and removing the district names out of the dataset.

3.3 Clustering

In this step, we are going to perform clustering by using K-means clustering to help grouping the district to extract insight from the datasets.

K-means is the simplest clustering technique which is proofed to be useful in many situations, so it is applied to this project as well. However, an obstacle of using this technique is we have to define K number by ourselves. To find the optimum value of K, we use Elbow technique to define K for our clustering technique. We leverage library from yellowbrick as a shortcut for implementing Elbow analysis in our projects. In this dataset, the optimum K is equal to 5 based on suggestion of the library.

After getting the K, we will perform K-means clustering with our pre-processed data to grouping the district and use it in analysis step.

4.Analysis

This section provide analysis of each cluster after data is clustered into five specific group using K-means clustering technique.

4.1 Cluster 0 (1)

	District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bang Bon	0	Thai Restaurant	Japanese Restaurant	Small Shop	Convenience Store	Noodle House
2	Bang Khae	0	Convenience Store	Western Restaurant	Noodle House	BBQ Joint	Coffee
5	Bang Khun Thian	0	Small Shop	Western Restaurant	Dessert	Japanese Restaurant	Sport place
14	Chom Thong	0	Western Restaurant	Thai Restaurant	Coffee	Hotpot Restaurant	BBQ Joint
19	Khan Na Yao	0	Japanese Restaurant	Western Restaurant	Dessert	Thai Restaurant	Coffee
23	Lak Si	0	Coffee	Western Restaurant	Thai Restaurant	Market	Japanese Restaurant
28	Nong Khaem	0	Western Restaurant	Thai Restaurant	Dessert	Market	Hotpot Restaurant
29	Pathum Wan	0	Convenience Store	Noodle House	Thai Restaurant	Dessert	Asian Restaurant
30	Phasi Charoen	0	Japanese Restaurant	Western Restaurant	Small Shop	Coffee	BBQ Joint
32	Phra Khanong	0	Western Restaurant	Convenience Store	Residence	Coffee	Dessert
34	Pom Prap Sattru Phai	0	Noodle House	Chinese Restaurant	Small Shop	Western Restaurant	Coffee
47	Wang Thonglang	0	Coffee	Western Restaurant	Asian Restaurant	Sport place	Small Shop
49	Yan Nawa	0	Western Restaurant	Thai Restaurant	Coffee	Hotpot Restaurant	Chinese Restaurant

Overview: This group of clusters is districts that stand out on food, especially on foreign food. from most common on 2 and 3 mostly are Western Restaurant followed by coffee and Japanese Restaurant.

Recommend: These districts suit for people who want to try western food in Thailand and these areas also have coffee cafes for chilling after your meal.

4.2 Cluster 1 (2)

	District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Bang Kapi	1	Noodle House	Japanese Restaurant	Thai Restaurant	Convenience Store	Market
3	Bang Khen	1	Asian Restaurant	Park	Noodle House	Convenience Store	Thai Restaurant
4	Bang Kho Laem	1	Noodle House	Thai Restaurant	Chinese Restaurant	Coffee	Western Restaurant
6	Bang Na	1	Asian Restaurant	Coffee	Noodle House	Chinese Restaurant	Seafood Restaurant
8	Bang Rak	1	Noodle House	Residence	Thai Restaurant	Chinese Restaurant	Bar
10	Bangkok Noi	1	Noodle House	Coffee	Thai Restaurant	Western Restaurant	Park
11	Bangkok Yai	1	Noodle House	Asian Restaurant	Coffee	Dessert	Market
12	Bueng Kum	1	Asian Restaurant	Small Shop	Other Restaurant	Noodle House	Healthcare
17	Dusit	1	Noodle House	Asian Restaurant	Thai Restaurant	Coffee	Convenience Store
18	Huai Khwang	1	Noodle House	Residence	Asian Restaurant	Thai Restaurant	Other Restaurant
41	Sathon	1	Noodle House	Asian Restaurant	Dessert	Coffee	Thai Restaurant
42	Suan Luang	1	Noodle House	Asian Restaurant	Thai Restaurant	Coffee	Convenience Store
43	Taling Chan	1	Noodle House	Seafood Restaurant	Coffee	Convenience Store	Other Restaurant
45	Thon Buri	1	Noodle House	Other Restaurant	Healthcare	Seafood Restaurant	Western Restaurant

Overview: This group of clusters is districts that stand out on street food and traditional Thai food from most common on 1st, 2nd and 3rd mostly are Noodle House and Asian Restaurant followed by Thai Restaurant.

Recommend: These districts suit for people who want to taste street food and traditional Thai.

4.3 Cluster 2 (3)

	District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
7	Bang Phlat	2	Convenience Store	Bar	Western Restaurant	Coffee	Residence
9	Bang Sue	2	Thai Restaurant	Noodle House	Coffee	Sport place	Seafood Restaurant
13	Chatuchak	2	Coffee	Thai Restaurant	Sport place	Bar	Residence
20	Khlong Sam Wa	2	Japanese Restaurant	Thai Restaurant	Bar	Chinese Restaurant	Coffee
21	Khlong San	2	Small Shop	Coffee	Dessert	Thai Restaurant	Chinese Restaurant
22	Khlong Toei	2	Bar	Tourist_attraction	Western Restaurant	Thai Restaurant	Hotpot Restaurant
24	Lat Krabang	2	Thai Restaurant	Western Restaurant	Other Restaurant	Coffee	Market
25	Lat Phrao	2	Thai Restaurant	Coffee	Noodle House	Small Shop	Western Restaurant
26	Min Buri	2	Coffee	Thai Restaurant	Small Shop	Western Restaurant	Hotpot Restaurant
27	Nong Chok	2	Small Shop	Market	Thai Restaurant	Park	Convenience Store
31	Phaya Thai	2	Coffee	Thai Restaurant	Japanese Restaurant	Bar	Western Restaurant
33	Phra Nakhon	2	Coffee	Residence	Tourist_attraction	Thai Restaurant	Bar
36	Rat Burana	2	Thai Restaurant	Asian Restaurant	Sport place	Other Restaurant	Noodle House
37	Ratchathewi	2	Coffee	Residence	Western Restaurant	Thai Restaurant	Other Restaurant
38	Sai Mai	2	Bar	Western Restaurant	Thai Restaurant	Small Shop	Other Restaurant
39	Samphanthawong	2	Tourist_attraction	Bar	Coffee	Chinese Restaurant	Residence
46	Thung Khru	2	Coffee	Convenience Store	Thai Restaurant	Sport place	Small Shop
48	Watthana	2	Coffee	Japanese Restaurant	Thai Restaurant	BBQ Joint	Small Shop

Overview: This group of clusters is districts which stand out on beverage shop. From the table, Coffee shop and Bar, beverage shop, are most members of this cluster, and the number of Thai restaurants just come after the beverage shop.

Recommend: These districts suit for people who want to taste coffee and other kinds of beverages in Thailand, and this area also has Thai restaurants to welcome visitors.

4.4 Cluster 3 (4)

	District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
16	Don Mueang	3	Other Restaurant	Convenience Store	Healthcare	Thai Restaurant	Residence
35	Prawet	3	Convenience Store	Other Restaurant	Noodle House	Western Restaurant	Hotpot Restaurant
44	Thawi Watthana	3	Convenience Store	Asian Restaurant	Other Restaurant	Healthcare	Thai Restaurant

Overview: This cluster has common venues on Convenience Store and Healthcare-service, these areas may be areas for working which are not suited for traveling

Recommend: these areas are not recommended on visiting, however, this cluster has Healthcare-service (i.e, massage and spa), so this would be matched for travelers who want to try famous Thai massage in its origin

4.5 Cluster 4 (5)

	District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
15	Din Daeng	4	Sport place	Convenience Store	Residence	Park	Western Restaurant
40	Saphan Sung	4	Japanese Restaurant	Thai Restaurant	Sport place	Convenience Store	Hotpot Restaurant

Overview: This cluster does not have any outstanding attributes, so I guess this is resident zones of local people from having high numbers of sport places and convenience stores which they are the most commonplace Thai people usually visiting.

Recommend: these areas are not recommended for visiting.

5. Results and Discussion

From the analysis, I found that there are three outstanding clusters on the Western restaurant, street food and traditional Thai, and beverage on cluster zero, one and two, respectively. On number three cluster is not very outstanding in any venue but if comparing this cluster to the other clusters, the cluster has a higher number of healthcare venues more than the others. So, the cluster can be assumed to be outstanding from the other in the healthcare field. In the last cluster, there is no outstanding attribute exist so districts in this cluster are not recommended for traveling.

From the result, it can be seen that the cluster technique can only provide and clustering information based on restaurant categories. This because of imbalanced data.

	Venue Category
Coffee	135
Noodle House	126
Thai Restaurant	103
Western Restaurant	72
Convenience Store	70
Small Shop	61
Japanese Restaurant	59
Dessert	55
Asian Restaurant	49
Other Restaurant	45

According to the table, Restaurant and beverage data overwhelm most of the top ten frequency of the dataset. So, when we perform analyst tables the data frequently provide restaurant and coffee on top 5 most common venues on the districts.

To figure out the problem, we can use the bigger dataset by retrieving the higher number of the radius when using FourSquare, because I only collected only venues located within 500 meters (radius = 500) from the center of each district.

However, the nature of Bangkok city is a standing out on food which people mostly come to Bangkok for the reason of tasting street food. In addition, Bangkok does have only a few tourist attractions for tourism. Therefore, this could be the reason why gathering data mostly consists of restaurant data.

6.Conclusion

This project point to solve the problem of tourists who want to plan their visit to Bangkok, Thailand, by using results of cluster analysis to help as assistance in making decisions on planning their vacation. The analysis results are from analyzing the clustering result from K-means (k=5 which is picked from using the elbow technique). Then, the result will provide outstanding attributes of each cluster, containing a district list, so tourist can pick the cluster that suits for their traveling style.

In future work, this data is not to provide the dominant conclusion on clustering analysis. because I just use only from a dataset from FourSquare which setting the radius equal to 500 meters which is a small number. So, it would be able to find more useful insight from the same methodology but a bigger dataset.