Where to Explore in Yogyakarta?

Muhammad Faris Herlansyah

August 7, 2021

1. Introduction

1.1 Background

Yogyakarta, a city in Indonesia which also serves as the capital of the Special Region of Yogyakarta, attracts many tourists each year. The city is known for its strong Javanese cultural influence due to the Special Region of Yogyakarta being ruled by a monarchy, its relatively low living cost, and its population which is dominated by young people due to the city hosting some well-known universities. Many places that attract tourists include Keraton Yogyakarta (the seat of the ruling sultan), Jalan Malioboro with its shops selling batik clothes, Kotagede which is well-known for being the centre of silver-based handicrafts, plenty of food stalls, homey hotels, et cetera. Those characteristics make visitors have plenty of choices of what to do while visiting Yogyakarta.

1.2 Problem

However, sometimes first-time visitors to Yogyakarta do not know where or in which districts to focus their vacation on, which may be due to the variety of tourist attractions and distribution of attractions and public facilities (e.g. historical attractions and places to hang out may obviously be concentrated in different parts of the town). Also sometimes when they arrive they may not have yet researched about attractions in Yogyakarta, hence a quick recommendation for first-time visitors may be needed.

1.3 Interests

This analysis is intended to facilitate first-time visitors to Yogyakarta so that they can decide where to visit, what kinds of places they would love, where to stay, where to find cheap foods, etc.

2. Data Preparation

2.1 Data Source

I scraped kodepos.nomor.net to obtain data of all districts and urban villages. The data contains not only all districts in Yogyakarta but also postal codes of each district in Yogyakarta. Besides, I also used Foursquare API to obtain the data containing all venues located within the distance of several hundred metres from each district's centre, and Nominatim from Geopy to acquire geographical coordinates of each district in Yogyakarta.

2.2 Data Cleaning

The webpage from which the districts data were scraped was, in my opinion, aesthetically untidy. Moreover, there were more than two rows representing column names, which was confusing. Also there were doubled place names and a postal code column which contained strings other than postal codes. Hence I performed data cleaning by dropping column name duplicates, resetting indexes, and changing several cells which contain duplicates and unnecessary characters.

Moreover, because we want to cluster districts, it is obvious that we need geographical coordinates data of each district in Yogyakarta. Hence, again I used Nominatim to obtain the geographical coordinates of Yogyakarta, as well as geographical coordinates of each of its districts. Then, I combined the obtained geographical coordinates data with the districts data.

1	Kode POS	Kelurahan	Kecamatan	Kota	Lat_Kecamatan	Long_Kecamatan
0	55253	Pakuncen	Wirobrajan	Yogyakarta	-7.802624	110.350447
1	55251	Patangpuluhan	Wirobrajan	Yogyakarta	-7.802624	110.350447
2	55252	Wirobrajan	Wirobrajan	Yogyakarta	-7.802624	110.350447
3	55163	Giwangan	Umbulharjo	Yogyakarta	-7.814378	110.387374
4	55165	Muja Muju	Umbulharjo	Yogyakarta	-7.814378	110.387374
5	55161	Pandeyan	Umbulharjo	Yogyakarta	-7.814378	110.387374
6	55166	Semaki	Umbulharjo	Yogyakarta	-7.814378	110.387374
7	55162	Sorosutan	Umbulharjo	Yogyakarta	-7.814378	110.387374
8	55167	Tahunan	Umbulharjo	Yogyakarta	-7.814378	110.387374
9	55164	Warungboto	Umbulharjo	Yogyakarta	-7.814378	110.387374
10	55243	Bener	Tegalrejo	Yogyakarta	-7.780455	110.355073
11	55241	Karangwaru	Tegalrejo	Yogyakarta	-7.780455	110.355073
12	55242	Kricak	Tegalrejo	Yogyakarta	-7.780455	110.355073
13	55244	Tegalrejo	Tegalrejo	Yogyakarta	-7.780455	110.355073
14	55111	Gunungketur	Pakualaman	Yogyakarta	-7.800395	110.376249
15	55112	Purwokinanti	Pakualaman	Yogyakarta	-7.800395	110.376249
16	55261	Ngampilan	Ngampilan	Yogyakarta	-7.802183	110.357603
17	55262	Notoprajan	Ngampilan	Yogyakarta	-7.802183	110.357603

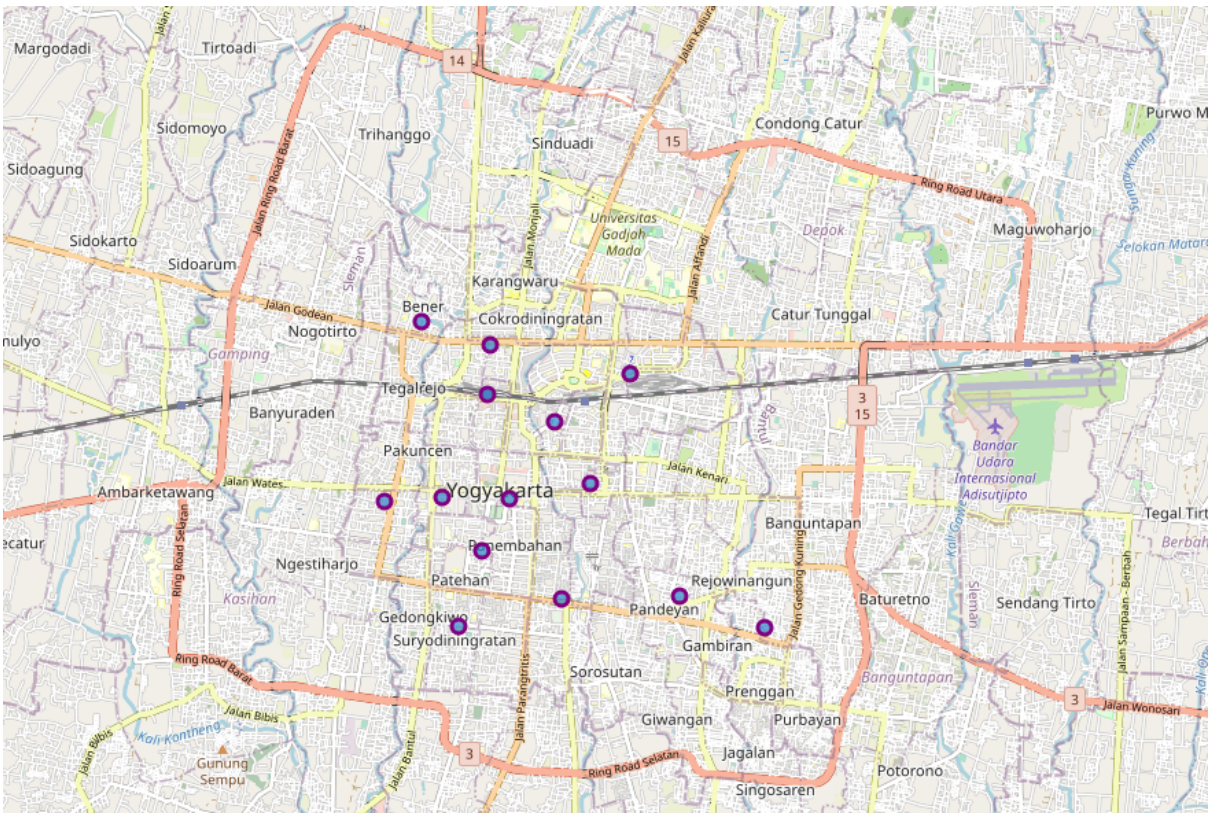
2.3 Feature selection

Given and known that urban villages in Yogyakarta have relatively small areas, which may contain only a small number of venues, I decided not to use the urban villages data; I used only postal code, districts, and city data.

The districts and urban villages data were combined with the latitudes and longitudes data. Then, the data were grouped by districts. Hence, we obtained a data frame containing 14 districts of Yogyakarta and their respective geographical coordinates.

1	Kode POS	Kecamatan	Kota	Lat_Kecamatan	Long_Kecamatan
0	55211	Danurejan	Yogyakarta	-7.792842	110.371795
1	55271	Gedongtengen	Yogyakarta	-7.789338	110.363376
2	55221	Gondokusuman	Yogyakarta	-7.786791	110.381157
3	55121	Gondomanan	Yogyakarta	-7.802395	110.366112
4	55231	Jetis	Yogyakarta	-7.783297	110.363649
5	55171	Kotagede	Yogyakarta	-7.818311	110.397941
6	55131	Kraton	Yogyakarta	-7.808799	110.362726
7	55141	Mantrijeron	Yogyakarta	-7.818067	110.359731
8	55151	Mergangsan	Yogyakarta	-7.814734	110.372558
9	55261	Ngampilan	Yogyakarta	-7.802183	110.357603
10	55111	Pakualaman	Yogyakarta	-7.800395	110.376249
11	55241	Tegalrejo	Yogyakarta	-7.780455	110.355073
12	55161	Umbulharjo	Yogyakarta	-7.814378	110.387374
13	55251	Wirobrajan	Yogyakarta	-7.802624	110.350447

After that, I created a visualization of Yogyakarta's districts using Folium. The resulting map is as follows, with the dots representing all districts of Yogyakarta:



2.4 Using Foursquare API to Obtain Venues Data

Now we start using the Foursquare API to obtain data of venues in Yogyakarta.

First, we define our Foursquare credentials, which comprises client ID, client secret, API version, and limit value to determine the number of venues obtained. Then we try obtaining venue data at one district, to ensure that the API works well in obtaining data. If the API works well, it will return data in JSON format.

After obtaining the data in JSON format, we can reformat the data to be in the form of a data frame. Below is an example of a data frame containing venues data in the district of Gedongtengen:

	name	categories	lat	lng
0	Stasiun Yogyakarta Tugu	Train Station	-7.789425	110.363460
1	Loko Cafe	Café	-7.789171	110.363488
2	Jogja Scrummy Stasiun Tugu	Food & Drink Shop	-7.789426	110.363686
3	Roti Cane Maryam Sta. Tugu	Bakery	-7.789235	110.363395
4	Top Roof Inna Garuda Hotel	Roof Deck	-7.789274	110.363896

Because the above process has worked well, we apply the similar process for all districts in Yogyakarta. By creating a function using Foursquare APIs and method to convert JSON data into data frame, the following data frame is generated:

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Danurejan	-7.792842	110.371795	Laura's Backpackers	-7.790985	110.370988	Hostel
1	Danurejan	-7.792842	110.371795	Jagung Bakar Menthel	-7.792026	110.373640	Snack Place
2	Danurejan	-7.792842	110.371795	BAKPIAKU Gajah Mada (pusat)	-7.797168	110.372700	Bakery
3	Danurejan	-7.792842	110.371795	Bakpia Citra Premium	-7.793618	110.368021	Snack Place
4	Danurejan	-7.792842	110.371795	Rumah Makan Murni 83	-7.791199	110.372408	Asian Restaurant
5	Danurejan	-7.792842	110.371795	Sate Ayam Podomoro	-7.795822	110.369001	Asian Restaurant
6	Danurejan	-7.792842	110.371795	Indomaret Jl.Hayam Wuruk	-7.795520	110.372715	Department Store
7	Danurejan	-7.792842	110.371795	Spesial Sambal Mentah "Bu Saring"	-7.792306	110.373451	Asian Restaurant
8	Danurejan	-7.792842	110.371795	warung makan pak parno sate kambing dan tongseng	-7.792491	110.373380	Diner
9	Danurejan	-7.792842	110.371795	Kedai Kopi Mataram	-7.792888	110.367511	Coffee Shop

We can also check the shape of the above data frame:

```
[24] print(yogya_venues.shape)

     (373, 7)
```

We can check how many venues returned for each district:

District	Venue
Danurejan	16
Gedongtengen	39
Gondokusuman	33
Gondomanan	53
Jetis	66
Kotagede	13
Kraton	30
Mantrijeron	25
Mergangsan	24
Ngampilan	9
Pakualaman	21
Tegalrejo	12
Umbulharjo	15
Wirobrajan	17

3. Exploratory Data Analysis

3.1 Categorizing Venues

We applied one-hot encoding to help the model we'll create later categorize venues and hence cluster better. The resulted one-hot-encoded data frame is as follows:

	District	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bakery	Bar	Basketball Stadium	Batik Shop	Beach	Bed & Breakfast	Beer Garden	Betawinese Restaurant	Bookstore	Boutique
0	Danurejan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Danurejan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Danurejan	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3	Danurejan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Danurejan	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

5 rows × 97 columns

There are 97 columns representing types of venues e.g. bakeries, bars, hotels, Asian restaurants, etc. The value 0 represents that there is no such a venue in the corresponding district, and the value 1 represents that there is such a venue.

We can check the frequency of occurrence of those types of venues in each district by grouping the above data frame by districts, and calculate the mean of the occurrences.

	District	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bakery	Bar	Basketball Stadium	Batik Shop	Beach	Bed & Breakfast	Beer Garden	Betawinese Restaurant	Bookstore
0	Danurejan	0.000000	0.000000	0.000000	0.000000	0.062500	0.250000	0.000000	0.062500	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0625	0.000000
1	Gedongtengen	0.000000	0.000000	0.000000	0.000000	0.000000	0.025641	0.000000	0.000000	0.051282	0.000000	0.000000	0.000000	0.025641	0.025641	0.0000	0.000000
2	Gondokusuman	0.030303	0.030303	0.000000	0.000000	0.000000	0.060606	0.090909	0.030303	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
3	Gondomanan	0.000000	0.018868	0.056604	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.018868	0.018868	0.018868	0.000000	0.0000	0.018868
4	Jetis	0.000000	0.000000	0.000000	0.000000	0.000000	0.106061	0.000000	0.045455	0.000000	0.015152	0.000000	0.000000	0.000000	0.030303	0.0000	0.000000
5	Kotagede	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
6	Kraton	0.000000	0.000000	0.000000	0.000000	0.033333	0.066667	0.000000	0.066667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
7	Mantrijeron	0.000000	0.000000	0.160000	0.000000	0.000000	0.040000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
8	Mergangsan	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333	0.041667	0.083333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
9	Ngampilan	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.111111	0.000000	0.000000	0.000000	0.000000	0.111111	0.000000	0.0000	0.000000
10	Pakualaman	0.000000	0.000000	0.000000	0.000000	0.000000	0.047619	0.047619	0.047619	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
11	Tegalrejo	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
12	Umbulharjo	0.000000	0.000000	0.000000	0.133333	0.066667	0.133333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
13	Wirobrajan	0.000000	0.000000	0.000000	0.058824	0.058824	0.058824	0.000000	0.058824	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000

14 rows × 97 columns

To make it easier to see top venues in each district, we can sort the frequency of occurrence values and group by districts. Below is an example of the sorting result.

```
----Danurejan----
                  venue  freq
0       Asian Restaurant  0.25
1            Snack Place  0.12
2                 Hostel  0.06
3  Betawinese Restaurant  0.06
4                 Diner  0.06


----Gedongtengen----
                  venue  freq
0                 Hotel  0.26
1  Indonesian Restaurant  0.13
2                  Café  0.08
3          Noodle House  0.08
4           Coffee Shop  0.08
```

We can conclude that if we take the Gedongtengen district as an example, the most frequent venue that occurs in said district is hotels, followed by Indonesian restaurants.

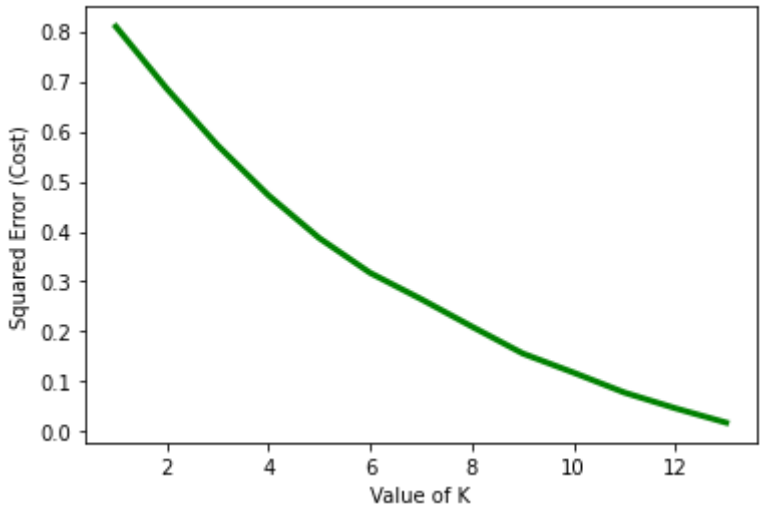
It would be easier if we turn the above data into a data frame like this:

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Danurejan	Asian Restaurant	Snack Place	Bakery	Department Store	Diner
1	Gedongtengen	Hotel	Indonesian Restaurant	Noodle House	Coffee Shop	Café
2	Gondokusuman	Food Truck	Indonesian Restaurant	BBQ Joint	Diner	Breakfast Spot
3	Gondomanan	Indonesian Restaurant	Historic Site	Clothing Store	Art Gallery	Chinese Restaurant
4	Jetis	Hotel	Asian Restaurant	Coffee Shop	Indonesian Restaurant	Food Truck

4. Clustering

Now, let's start clustering the above data into clusters. Remember, we want to cluster districts to help the users, who are visitors, to determine which areas to explore in accordance to their focus and/or desire. Here we'll use K-Means clustering. First, we prepare the clustering data, which is the above venues data without the districts column, and don't forget to choose the best k value.

For choosing the best k value, we will apply the elbow method, which is a common way to determine the best k. Choosing the range between 1 and 14 (the maximum number of districts), the elbow plot looks like this:

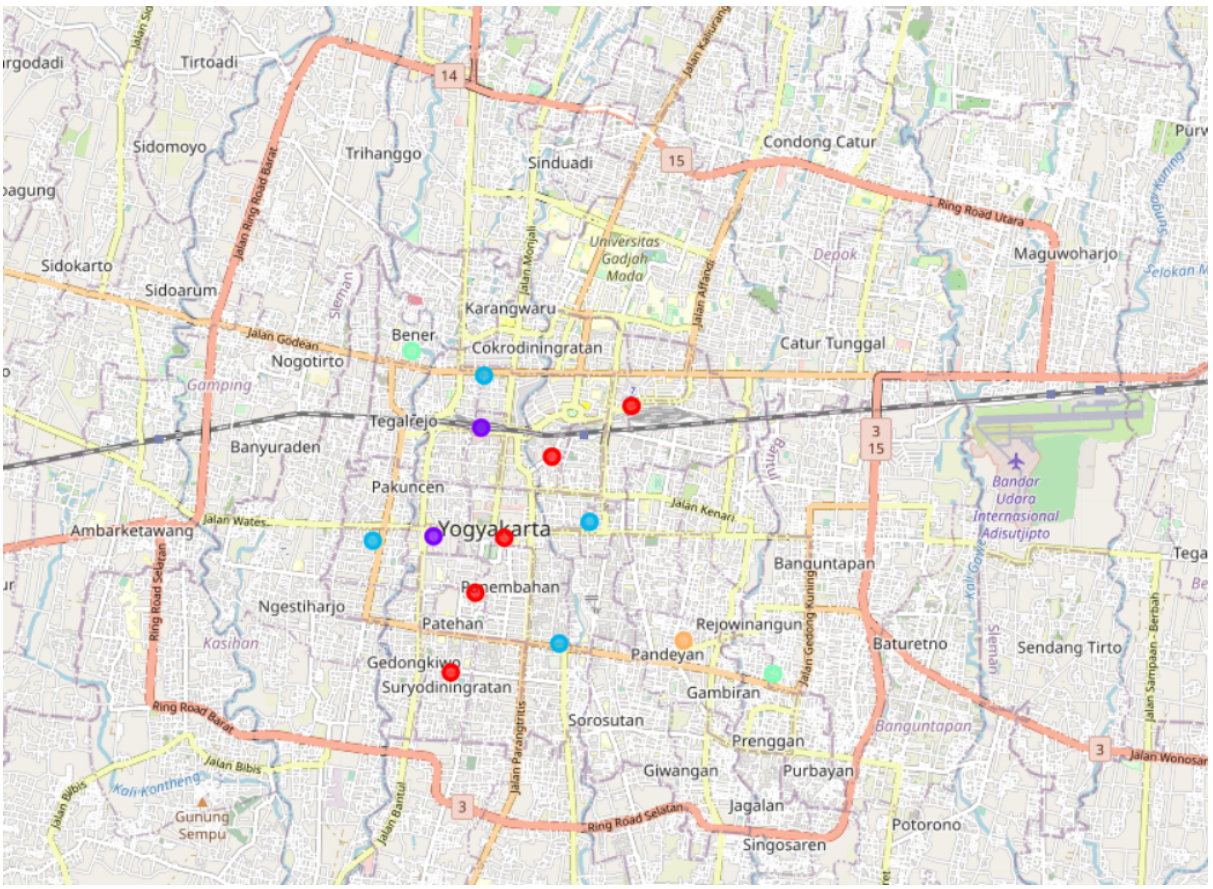


Although the “elbow” looks not too obvious, we can choose 5 as our best k value since it is located between the steeper part and the less steep part of the curve.

After choosing k = 5, we can build the clustering model, train with the data, and apply cluster labels to each district. The resulted data frame with the cluster labels is as follows:

	Kode POS	Kecamatan	Kota	Lat_Kecamatan	Long_Kecamatan	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	55211	Danurejan	Yogyakarta	-7.792842	110.371795	0	Asian Restaurant	Snack Place	Bakery	Department Store	Diner
1	55271	Gedongtengen	Yogyakarta	-7.789338	110.363376	1	Hotel	Indonesian Restaurant	Noodle House	Coffee Shop	Café
2	55221	Gondokusuman	Yogyakarta	-7.786791	110.381157	0	Food Truck	Indonesian Restaurant	BBQ Joint	Diner	Breakfast Spot
3	55121	Gondomanan	Yogyakarta	-7.802395	110.366112	0	Indonesian Restaurant	Historic Site	Clothing Store	Art Gallery	Chinese Restaurant
4	55231	Jetis	Yogyakarta	-7.783297	110.363649	2	Hotel	Asian Restaurant	Coffee Shop	Indonesian Restaurant	Food Truck

And again it’s easier for us to visualize the labeled clusters on a map.



We can see that there are five clusters, namely Cluster 0, 1, 2, 3, and 4. Cluster 0, colored red, is the largest cluster, and Cluster 4, colored orange, is the smallest cluster.

To make it clearer, let's observe each cluster.

- Cluster 0

	Kode POS	Kecamatan	Kota	Lat_Kecamatan	Long_Kecamatan	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	55211	Danurejan	Yogyakarta	-7.792842	110.371795	0	Asian Restaurant	Snack Place	Bakery	Department Store	Diner
2	55221	Gondokusuman	Yogyakarta	-7.786791	110.381157	0	Food Truck	Indonesian Restaurant	BBQ Joint	Diner	Breakfast Spot
3	55121	Gondomanan	Yogyakarta	-7.802395	110.366112	0	Indonesian Restaurant	Historic Site	Clothing Store	Art Gallery	Chinese Restaurant
6	55131	Kraton	Yogyakarta	-7.808799	110.362726	0	Historic Site	Indonesian Restaurant	Javanese Restaurant	Café	Asian Restaurant
7	55141	Mantrijeron	Yogyakarta	-7.818067	110.359731	0	Café	Art Gallery	Coffee Shop	Indonesian Restaurant	Pizza Place

Cluster 0, colored red, contains five districts: Danurejan, Gondokusuman, Gondomanan, Kraton, and Mantrijeron. If we inspect the most common venues in Cluster 0, we see historic sites and Indonesian restaurants dominating the top venues; for example, the Keraton Yogyakarta, Fort Vredeburg, and the North and South Squares of Yogyakarta are all located in Cluster 0. Hence, we can recommend visitors to visit areas included in Cluster 0 if they would like to have more cultural and/or historic experiences during their visits.

- Cluster 1

	Kode POS	Kecamatan	Kota	Lat_Kecamatan	Long_Kecamatan	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	55271	Gedongtengen	Yogyakarta	-7.789338	110.363376	1	Hotel	Indonesian Restaurant	Noodle House	Coffee Shop	Café
9	55261	Ngampilan	Yogyakarta	-7.802183	110.357603	1	Hotel	Indonesian Restaurant	Bed & Breakfast	Hostel	Dessert Shop

There are two districts located in Cluster 1: Gedongtengen and Ngampilan. We see that the most common venues in Cluster 1 are hotels, Indonesian restaurants, as well as bed and breakfasts and hostels. Knowing that the Yogyakarta railway station is located in Cluster 1, it makes sense that there are many kinds of accommodations located in Cluster 1, because it will be easier for visitors to get around if they stay at hotels with easy access to public transportation. Hence, we can recommend visitors to stay around areas belonging to Cluster 1.

- Cluster 2

	Kode POS	Kecamatan	Kota	Lat_Kecamatan	Long_Kecamatan	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	55231	Jetis	Yogyakarta	-7.783297	110.363649	2	Hotel	Asian Restaurant	Coffee Shop	Indonesian Restaurant	Food Truck
8	55151	Mergangsan	Yogyakarta	-7.814734	110.372558	2	Indonesian Restaurant	Convenience Store	Noodle House	Hotel	Asian Restaurant
10	55111	Pakualaman	Yogyakarta	-7.800395	110.376249	2	Hotel	Convenience Store	Ice Cream Shop	Pizza Place	Department Store
13	55251	Wirobrajan	Yogyakarta	-7.802624	110.350447	2	Hotel	Food Truck	Clothing Store	Bakery	Convenience Store

Cluster 2 comprises four districts: Jetis, Mergangsan, Pakualaman, and Wirobrajan. If we see the above data frame, we can see that the most common venues are similar to Cluster 1: hotels and Indonesian restaurants. However, if we see the third, fourth, and fifth most common venues, we can see types of venues e.g. ice cream shops, coffee shops, bakery, pizza places, and food trucks. Hence, if visitors would like to look for light bites and places for a mini hang-out in the afternoon, for example, we can recommend them to stay in hotels located in and around areas belonging to Cluster 2.

- Cluster 3

	Kode POS	Kecamatan	Kota	Lat_Kecamatan	Long_Kecamatan	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	55171	Kotagede	Yogyakarta	-7.818311	110.397941	3	Indonesian Restaurant	Convenience Store	Grocery Store	Department Store	Supermarket
11	55241	Tegalrejo	Yogyakarta	-7.780455	110.355073	3	Indonesian Restaurant	University	Electronics Store	Convenience Store	Pizza Place

Cluster 3 comprises two districts: Kotagede and Tegalrejo. From the above data frame, we can see that the most common venues in Cluster 3 are Indonesian restaurants, convenience stores, grocery stores, and department stores. This means that we can recommend visitors to go to these two districts if they need to shop for daily needs.

Moreover, many Indonesians know that Kotagede is known for being the centre of silver-based handicrafts in Yogyakarta. That means in Cluster 3 there would also be many shops selling handicrafts made of silver. Hence visitors can go there to buy souvenirs aside from daily needs.

- Cluster 4

	Kode POS	Kecamatan	Kota	Lat_Kecamatan	Long_Kecamatan	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
12	55161	Umbulharjo	Yogyakarta	-7.814378	110.387374	4	Café	Art Museum	Soup Place	Asian Restaurant	Food Truck

Cluster 4 is the smallest cluster, containing only the district of Umbulharjo. The most common venues in Cluster 4, which we can retrieve from the above data frame, are cafés, art museums, soup places, Asian restaurants, and food trucks. Hence, we can recommend visitors to explore Cluster 4 if they want to eat or buy foods, or in other words, to have a culinary experience. Some snack shops are located in Cluster 4. Moreover, the famous Gembira Loka zoo is located in this cluster, and Cluster 4 is located beside Kotagede of Cluster 3 which means this cluster is not far from shopping places.

5. Conclusion

Although some of the clustering results still have similar characteristics of common venues, we can see that clustering may help us in identifying patterns of the data set. The model we built has been able to categorize areas with similar characteristics, with the help of Foursquare API in obtaining geographical coordinates data and encoding to help the model categorize data. The results of this clustering method can be used to help us recommend visitors to Yogyakarta about where to do particular activities.