Clustering of neighborhoods in Tokyo, Japan based on popular venues by Foursquare database

Tuya Delgersaikhan

December, 2019

1. Introduction

- Background of Tokyo (23 special wards):
  - Huge city with estimated population of 9.24 million as of 2015 covering an area of about 627 square kilometers
  - Major tourism destination
  - Hundreds of neighborhoods
  - Extensive public transportation
- Problems:
  - Potentially lots of unexplored and relatively unknown neighborhoods to both tourists and the locals
  - Internet shows only popular places and neighborhoods in Tokyo, especially when searched in English!
  - The locals have major decision to make regarding where to live (and visit also)
  - Tourists and the locals need more information about neighborhoods in Tokyo in order to have a fulfilling visit and living!

1. Introduction

- Purpose of this study:
  - Cluster and explore the neighborhoods in Tokyo using K-Means clustering algorithm based on popular venues in each neighborhoods
  - Provide tourists and the locals with information to help them choose where to visit and live

2. Data preparation

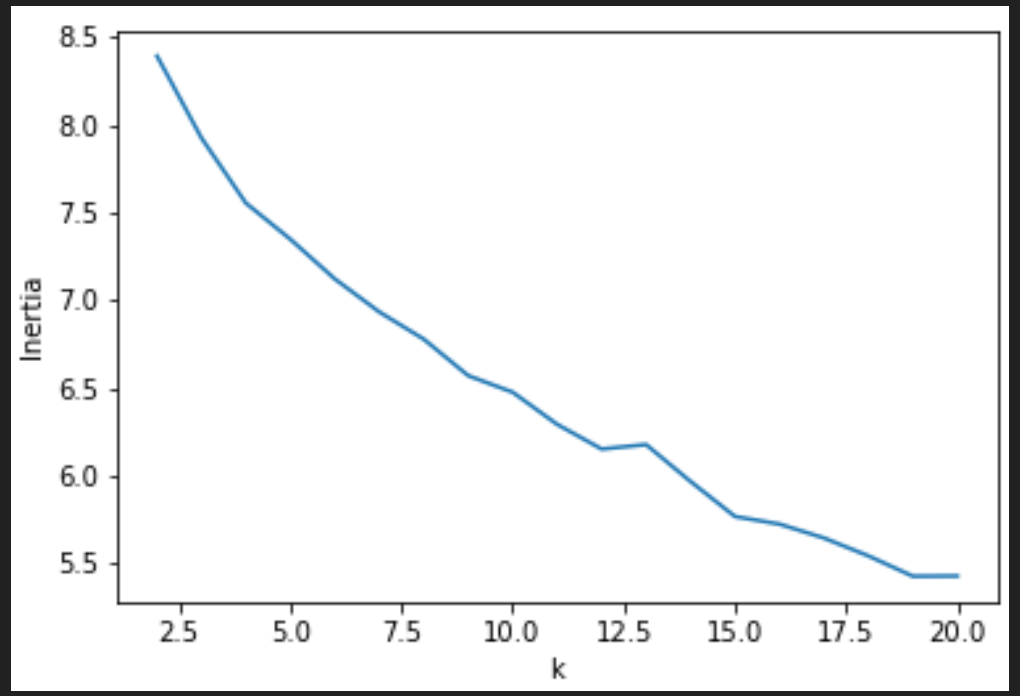
1. Initial list of names of the neighborhoods in Tokyo
   - Wikipedia page: https://commons.wikimedia.org/wiki/Category:Neighborhoods_in_Tokyo_by_ward
   - Beautiful Soup library was used to scrape the webpage and retrieve the names of each neighborhood along with the name of the ward
   - Total of 246 neighborhoods were obtained (NOTE: it is incomplete, but still good!)
2. Geographical coordinate data (i.e. latitude and longitude data) for each neighborhood
   - Geopy library was used to obtain coordinate data for each neighborhood
   - Coordinate data for some neighborhoods were manually fetched from Google Map
   - Coordinate data for all 246 neighborhoods were obtained!

2. Data preparation

3. Popular venues data by Foursquare location database:
   - Up to 100 popular venues around each neighborhood within 400 m radius
   - Total of 12952 venues obtained for 246 neighborhoods, averaging about 53 venues per neighborhood
   - Neighborhoods with less than 10 venues were then dropped from the database, leaving 12889 venues in 235 neighborhoods in 364 unique venue categories
4. Feature selection for K-Means clustering algorithm:
   - All 364 venue categories were chosen as features
   - One-hot-encoding and normalization were used to convert the 364 venue categories into features
   - Finally, dataset in the form of matrix with dimension (235, 364) were used to fit the K-Means clustering algorithm

3. Model

1. Choice of an optimal k-value for K-Means clustering algorithm:
   - So-called elbow method was utilized to choose an optimal k-value:

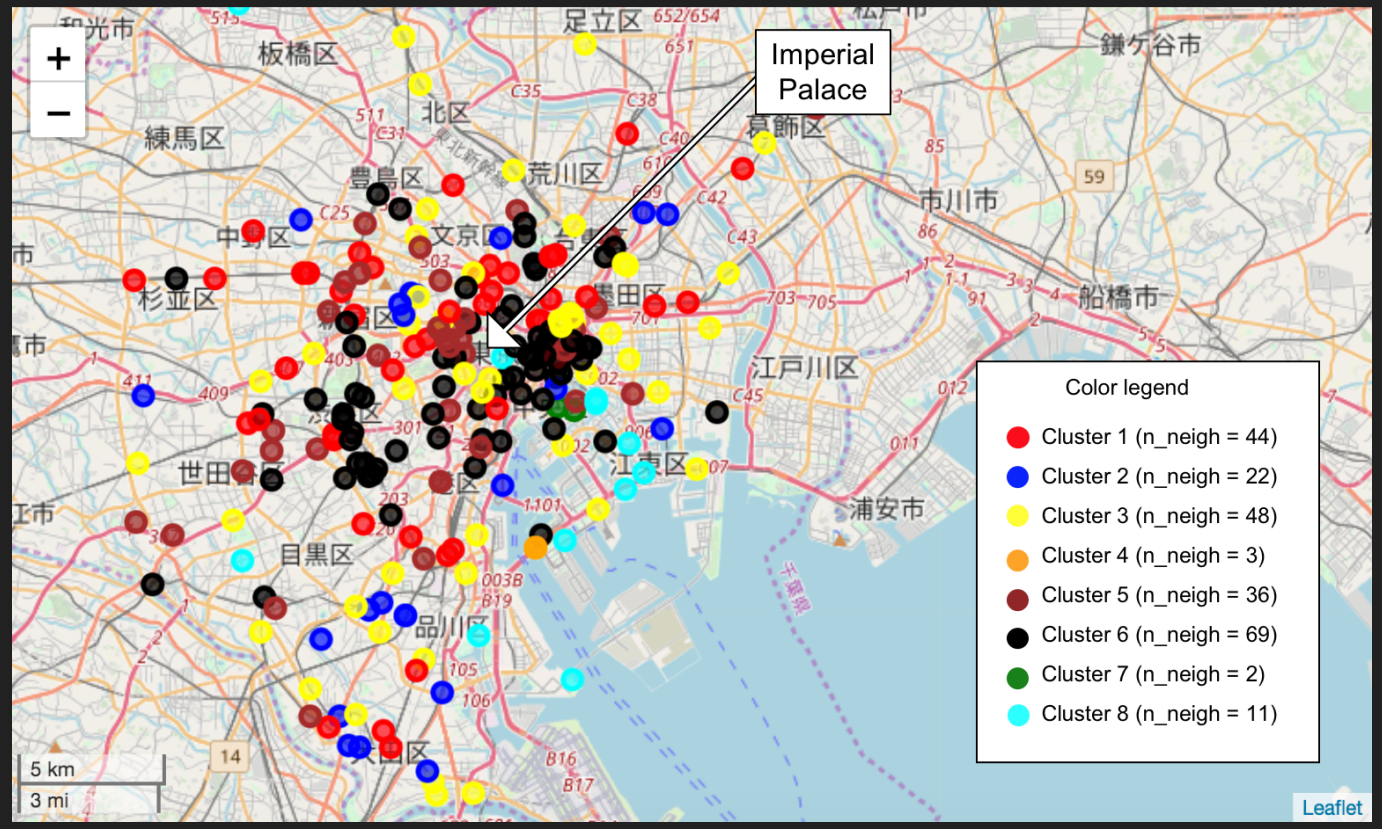


Parameters of the K-Means clustering algorithm from sklearn library:
init = 'k-means++'
n_init=12
(other parameters are default)

Although it is hard to see clearly, we can choose k=8 as an approximate optimal value!

4. Results and discussion

1. A map (folium library) showing the clustering result:



Geographically:
- Most of the neighborhoods in Cluster 6 are located around the center of Tokyo labeled as 'Imperial Palace'.
- Cluster 1and Cluster 5 appear somehow clustered near to central Tokyo also.
- Cluster 3 appears to located all over Tokyo rather uniformly.
- On the other hand, majority of Cluster 8 is located around Tokyo Bay (near the ocean).

Note: Number of neighborhoods in the cluster is shown in parentheses!

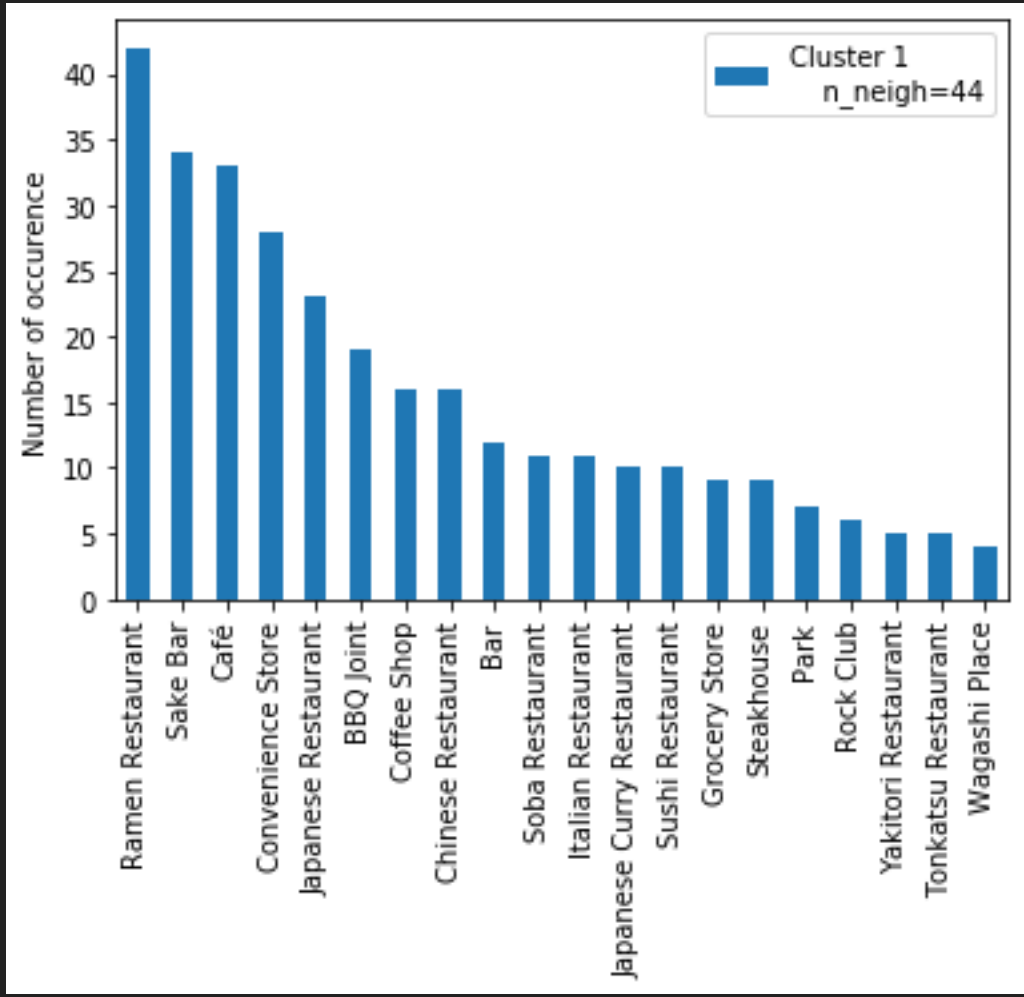
4. Results and discussion

2. Analysis of each cluster:
   - In order to analyze each cluster, first the top 10 most popular venue categories in each neighborhoods were obtained based on the venue counts.
   - Then up to top 20 venue categories for the cluster were retrieved from the union of top 10 venue categories of each neighborhood that belongs to the cluster.
   - This information will give us frequency of occurrences of the up to 20 top venue categories in each cluster.
   - In simpler words, it will let us know how many neighborhoods in the cluster have the given venue category in their top 10 venue categories.
   - Let us now start analyzing each cluster!

4. Results and discussion

Cluster 1:

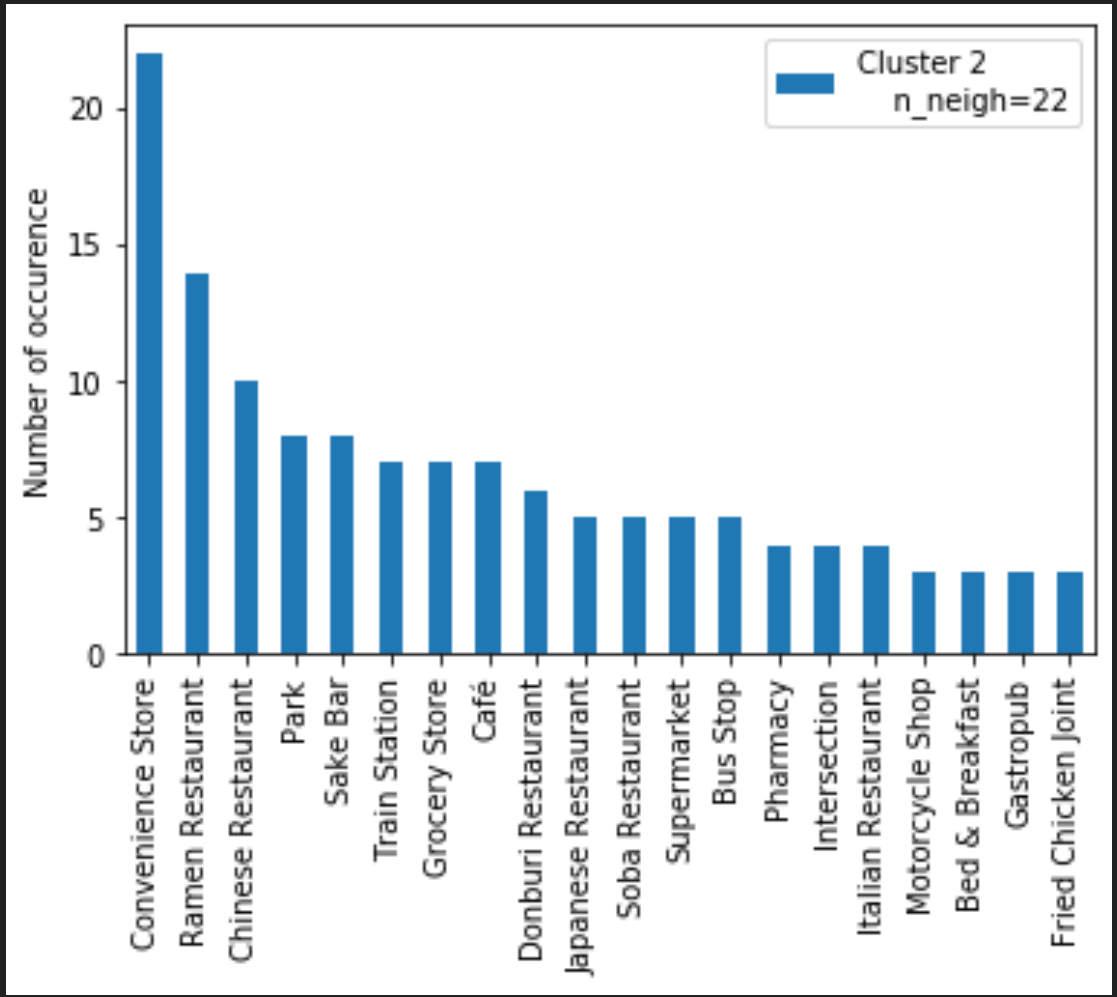
- The venue category 'Ramen Restaurant' seems very popular in this cluster as 42 out of 44 neighborhoods has it in their top 10 venue categories!
- Cluster 1 appears to be popular with other types of restaurants too.
- It also has certain popularity of coffee shops, bars, grocery stores, and parks.



4. Results and discussion

Cluster 2:

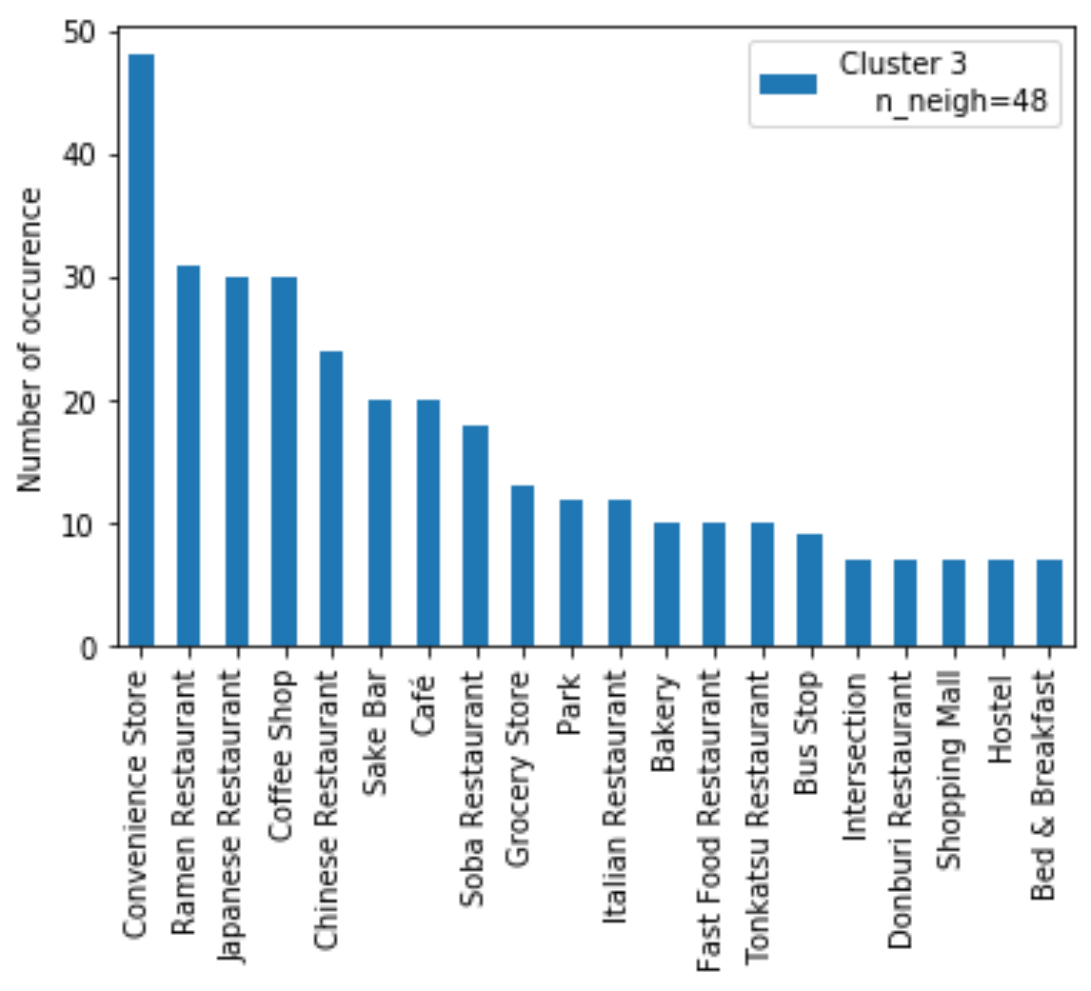
- Cluster 2 seems to be popular with convenient stores, various restaurants, train stations etc.
- In contrast to Cluster 1, it has increased popularity of train stations, supermarkets, and pharmacies.



4. Results and discussion

Cluster 3:

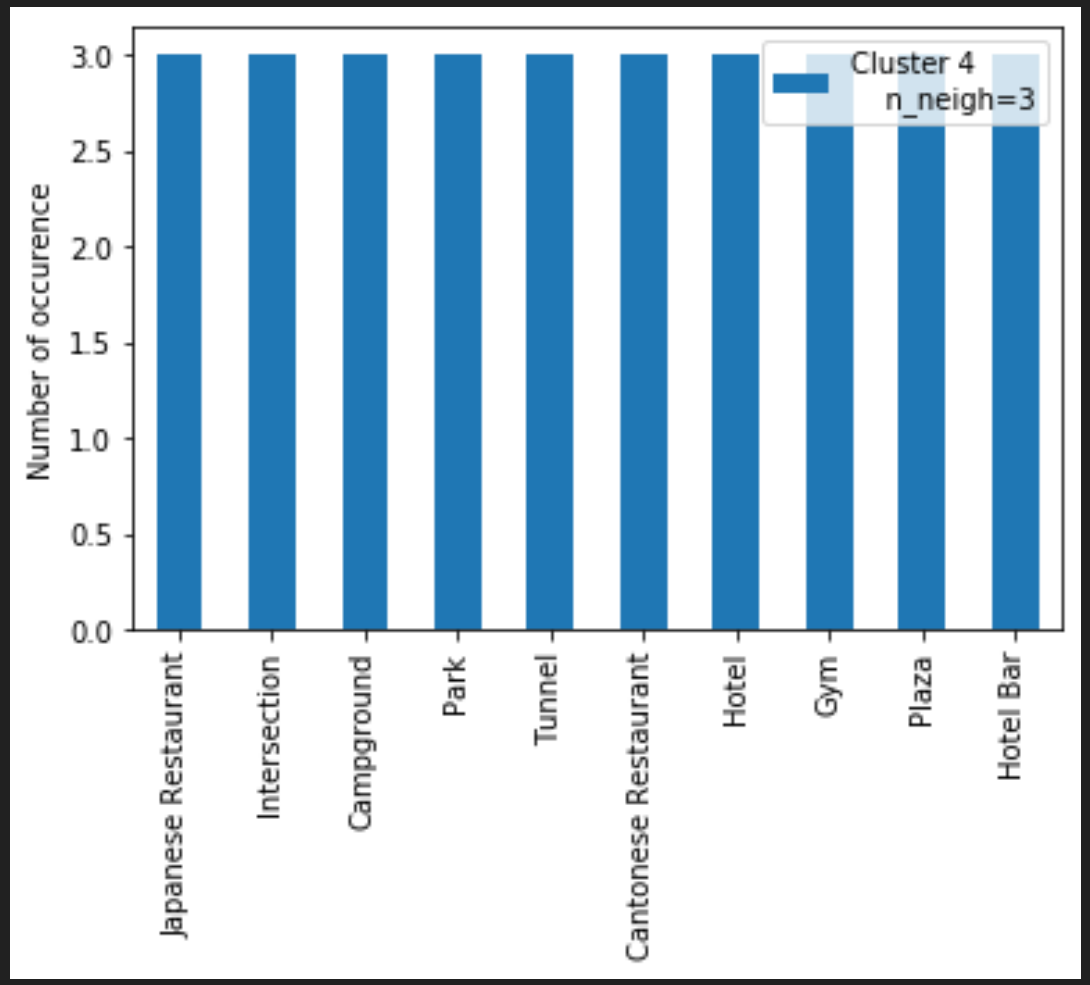
- Similar with previous two clusters, it is popular with various restaurants and convenient stores.
- However, unlike them, it has increased popularity of venue categories 'Bakery', 'Shopping Mall', and 'Hostel' etc.



4. Results and discussion

Cluster 4:

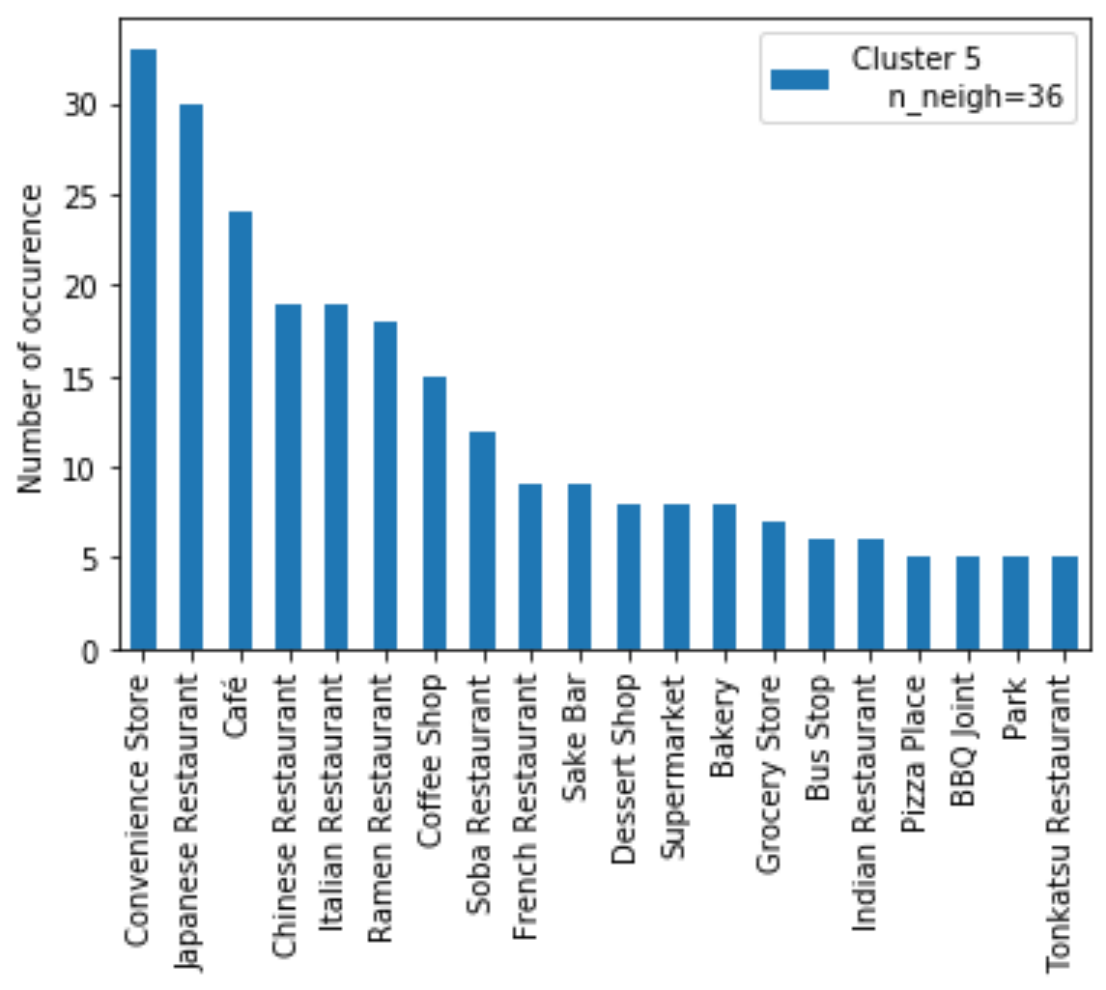
- All three neighborhoods in this cluster happened to have the same coordinate data by Geopy library, thus they have the same venues also, which means that Cluster 4 can be thought to have only one neighborhood.
- In that neighborhood, Japanese restaurants, intersections, campgrounds, parks etc seem to be popular.



4. Results and discussion

Cluster 5:

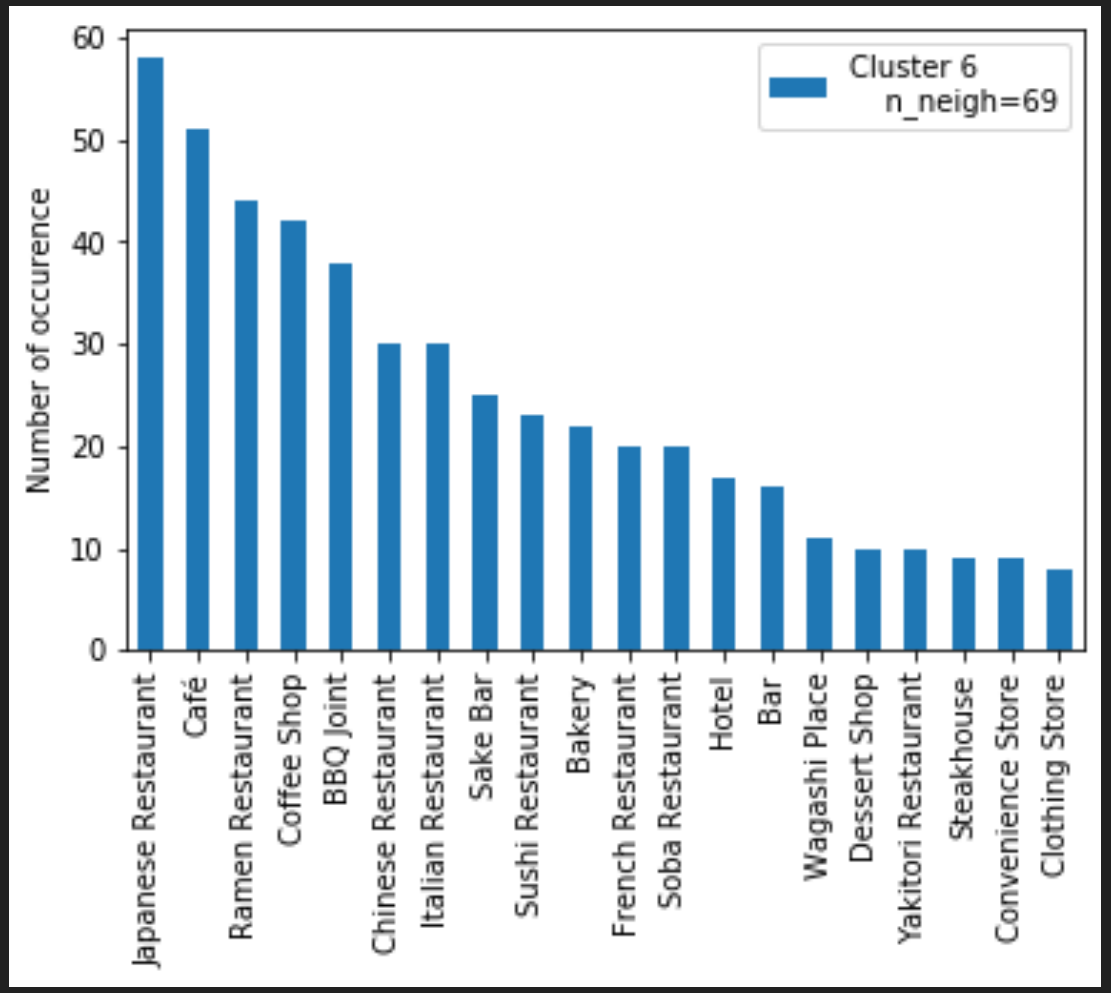
- Similar to the previous clusters, Cluster 5 appears to be popular with various types of restaurants.
- Unlike the previous clusters, it has increased popularity of dessert shops & pizza places.



4. Results and discussion

Cluster 6:

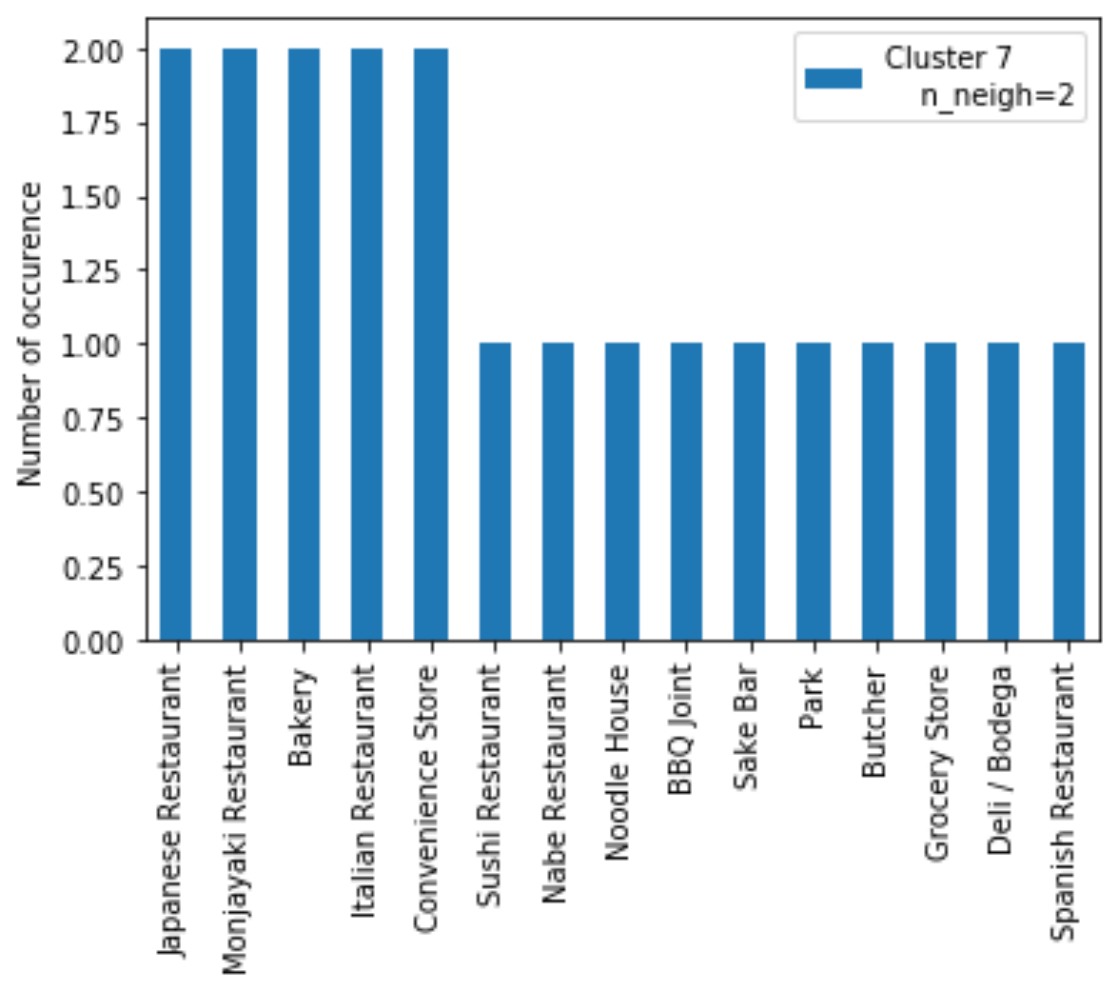
- Cluster 6 has the largest number of neighborhoods among all other clusters.
- And yet again, it is popular with various restaurants, bars, and coffee shops.
- However, the venue categories 'French Restaurant', 'Hotel', and 'Clothing Store' etc are popular also in this cluster.



4. Results and discussion

Cluster 7:

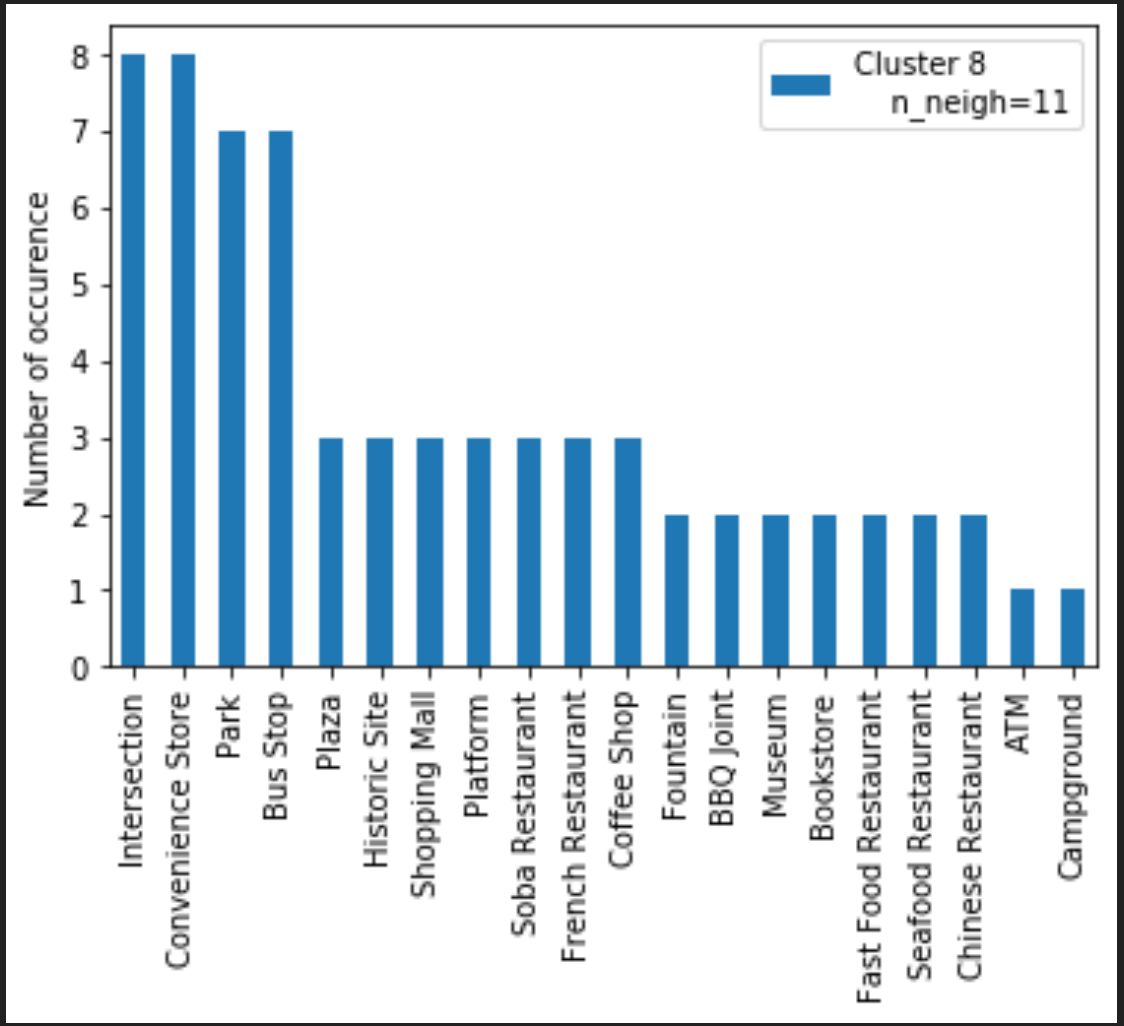
- Cluster 7 contains only two neighborhoods that are geographically close.
- Yet again, it is popular with Japanese restaurants and convenience stores.
- The distinct feature of these neighborhoods, however, is Monjayaki restaurant.
- As this cluster has only two neighborhoods, it is easy to confirm if Monjayaki restaurants are indeed popular around them.
- A 'Monjayaki' search around Tokyo using Google Map easily proved that Monjayaki is indeed popular around these neighborhoods.



4. Results and discussion

Cluster 8:

- In contrast to other clusters, Cluster 8 shows quite a distinct profile.
- It appears to be popular with parks, historic sites, shopping malls, museums, bookstores, and campgrounds etc.



5. Conclusions

- Total of 235 neighborhoods in Tokyo were clustered into 8 groups using K-Means clustering based on 12889 venues in 364 unique venue categories obtained using Foursquare location database.
- Each cluster showed some similar and some distinct characteristics to one another.
- In general, it can be said that Tokyo is a great place for food lovers as various types of restaurants dominated the popular venues around almost all neighborhoods.

6. Suggestions for future improvements of the clustering

- It was easily observable from the clustering results that various restaurants dominated database.
- In fact, 87 out of 364 venue categories included the word 'Restaurant'.
- One **idea** to possibly improve the clustering is to collapse these 87 venue categories into single category 'Restaurant'. Although the total number of restaurants are still the same, collapsing and clustering can bring other venue categories into top 10 venue categories, thus will make them more visible.
- Another **idea** to improve the clustering is to reduce the dimension of features by applying principal component analysis or by dropping highly correlated features.
- Yet another **idea** to improve the overall clustering is to use more neighborhoods. As discussed earlier, the initial source of neighborhoods in Tokyo misses many neighborhoods. For example, there are two wards missing (2 out of 23) and some wards contain only few neighborhoods. Including more neighborhoods will certainly improve the clustering and produce more realistic clusters.