Clustering and Analyses of Tokyo Hostels

Massao Mitsunaga

February 19, 2020

1. Introduction

1.1 Background

With the 2020 Tokyo Olympics coming, the number of tourists planning to go to Tokyo will rise significantly. Many of those tourits don't know what to expect in Tokyo Neighborhoods or don't how to choose a hostel that could fill their needs and that could lead to a poor choice that could potentially ruin the entire experience of travelling to another country.

1.2 Problem

This Project aims to aid in the planning process for a travel to Tokyo for the Olympics dividing Tokyo Hostels in Clusters to compare them and make it available for guidance to choose the best hostel to supply its needs.

1.3 Interest

Tourist planning to go to Tokyo for the olympics would be the most interested in this kind of information, travel agencies may also be interested in it to have a better visualization in what kind of hostel or travel plan to offer to their clientes, and the hostels involved as well, to know their situation according to the FourSquare API and to help their interaction with their clientes when it comes to indication of nearby venues to visit.

2. Data acquisition and cleaning

2.1 Data Sources

Based on definition of our problem, factors that will influence our decision are:

- number and variety of venues around the hostels;
- distance to the National Olympic Stadium and to the city center;
- Hostel prices.

The Following data sources will be needed to extract/generate the required information:

- name and location of tokyo's major districts used for reference in our visualization posted by **suvoooo** on Github;
- number of venues and their type and location for every hostel coordinate in a 500m radius will be obtained using Foursquare API;
- a dataset containing Japan Hostels by **HostelWorld** posted by **thatdatastudent** on kaggle.

2.2 Data Cleaning

From the hostels dataset I filtered so that only Tokyo hostels would appear on the dataframe, removed every hostel that didn't have coordinates, created a category for the *rating.band* column so it would be represented by numbers instead of words, and removed strings from the *Distance* column rows so it would only show the distance in km. I also calculated hostels distance to the National Olympic Stadium using the *geopy* library and added to the dataframe as a new column.

For the tokyo districts dataset i removed some unnecessary columns, renamed the ones referencing Latitude and Longitude and rearranged them for better understanding.

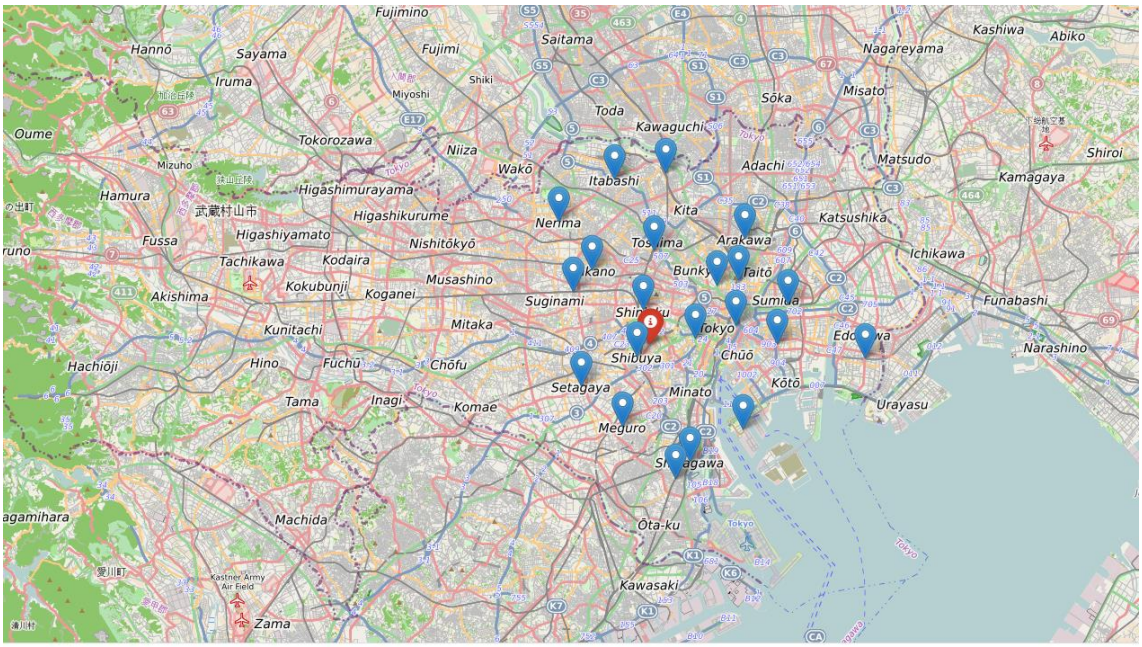
2.3 Features Selection

Afer cleaning the data there were 116 samples and 16 features in the hostels dataset and I decided to keep all the features for a better understanding of the data when comparing the clusters. As for the major districts dataset, because it will be only used as reference for the map visualization, the only attributes needed are the name of the districts and its coordinates.

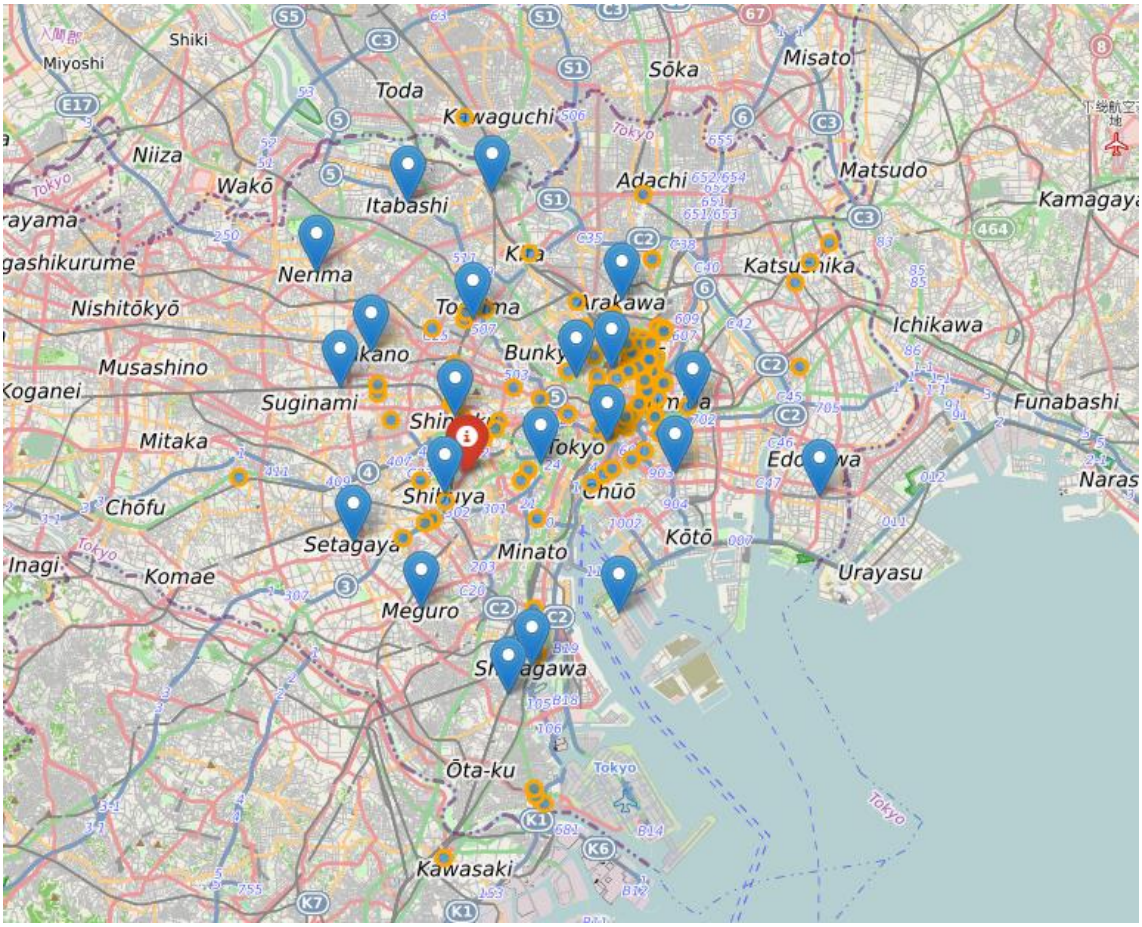
3. Exploratory Data Analysis

3.1 Map visualizations

First I made a map using the folium library using the data from the major districts dataset. The blue markers are the major districts and the red marker is the National Olympic Stadium.



Then, using the hostels dataset I added the hostels as circle markers in the map.



3.2 Gathering venues data with the FourSquare API

Using the FourSquare API I was able to get all the Hotels nearby venues in a 500m radius and found 276 unique categories of vênus surrounding Tokyo Hostels. After merging the datasets and grouping by the hostels, the resulting dataframe had 116 samples and 289 features and in the process I was also abled to find the 10 most common venues for each Hostel in the dataset.

4. Clusterization

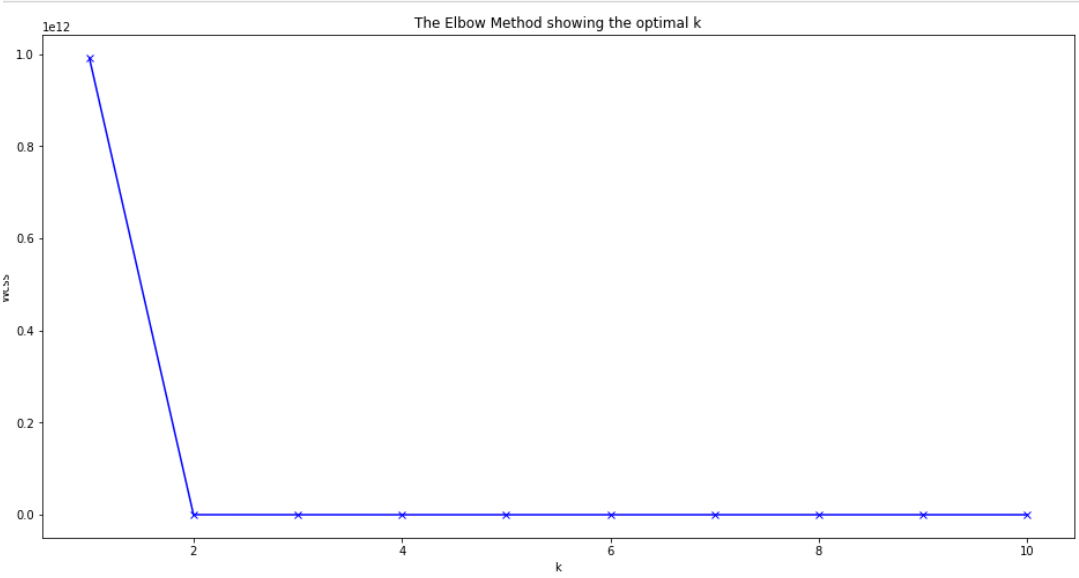
4.1 Creating Clusters

Calculating the distance from a point to a line defined by two points I was able to get the optimal number of clusters needed for the data and later validated the results through the visualization of the elbow method.

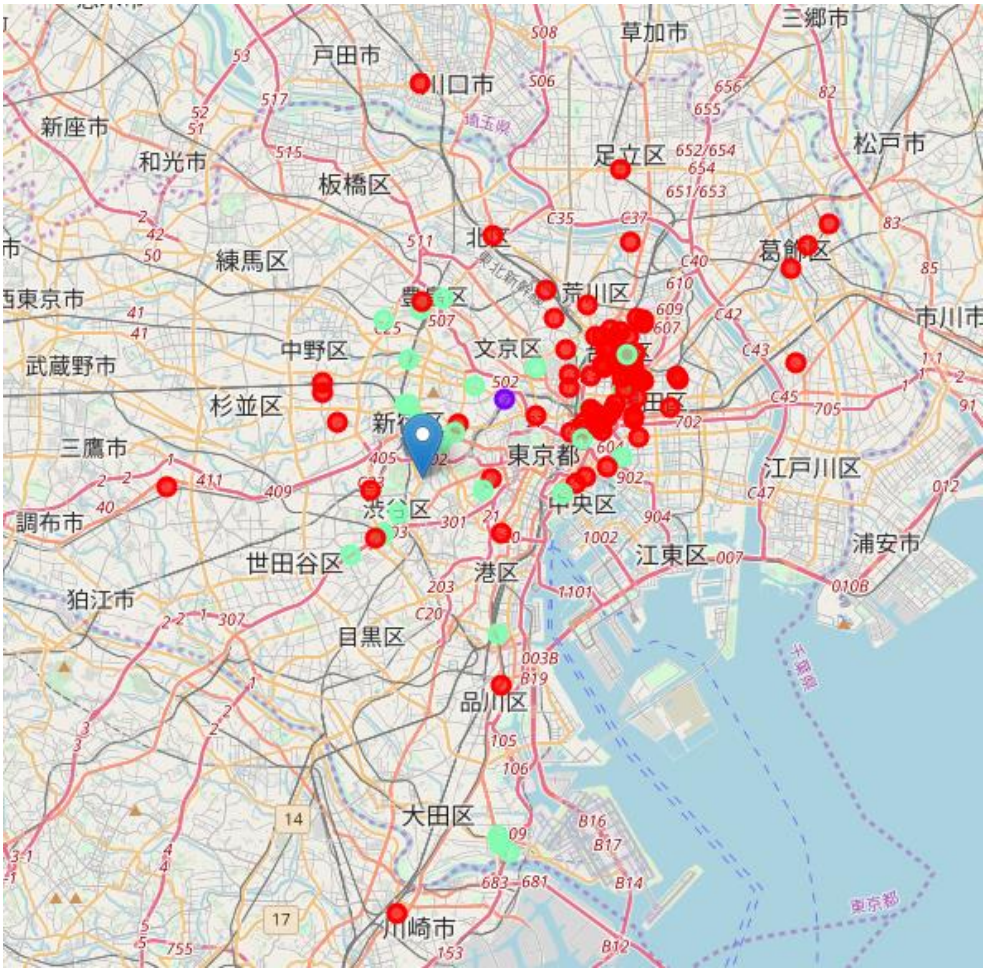
The formula used was the following and resulted in a optimal number of clusters equal to three.

$$\text{distance}(P_1, P_2, (x_0, y_0)) = \frac{|(y_2 - y_1)x_0 - (x_2 - x_1)y_0 + x_2 y_1 - y_2 x_1|}{\sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}}.$$

Validation through visualization:



After creating the cluster a map is generated with each cluster with its own color and the blue marker being the National Olympic Stadium:



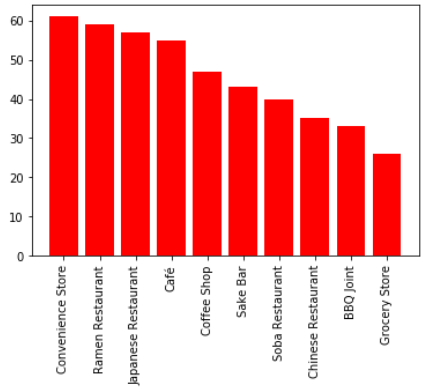
4.2 Analysing Clusters

Below we have the average of the features grouped by cluster. The main fator used to divide the clusters was the hostels prices:

Cluster Labels	price.from	distance.from.city.centre	summary.score	rating.band	atmosphere	cleanliness	facilities	location.y	security	staff	valueformoney	distance.from.stadium
0	2275.58	8.96	8.78	3.40	8.22	8.98	8.54	8.57	9.14	9.16	8.87	16.43
1	1003200.00	4.80	8.00	3.00	6.00	10.00	10.00	8.00	8.00	8.00	6.00	14.50
2	3993.10	6.68	8.89	3.48	8.39	9.25	8.77	8.74	8.93	9.31	8.85	12.82

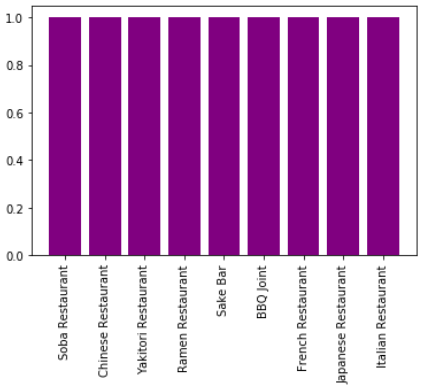
Analysing the clusters we found the most common venues by cluster:

- Cluster 0 - Red



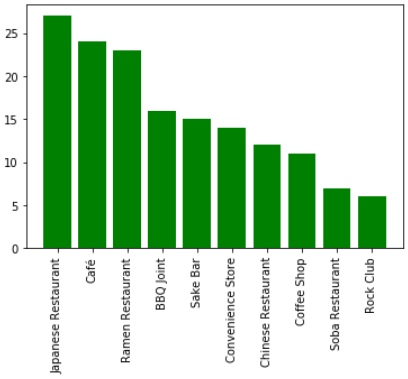
With the red cluster embracing the biggest part of the hostels it's clear that it would have more volume in it's surrounding venues with convenience stores as the most common venue with ramen restaurant coming as the second most common.

- Cluster 1 – Purple



The purle cluster can be considered na outliers because of the huge price difference comparing to the others, making ir a cluster of 1 hostel only.

- Cluster 2 – Green



The best cost-benefit hostels are in the Third (Green) cluster since they have intermediate price, higher hatings and score and tend to be closer to the center of the city as well as to the Olympic Stadium, also there are good amounts of venues considering the lower number of hostels in this cluster.

5. Conclusions

The purpose of this study was to aid our StakeHolders to find hostels that could best fill their needs based on a different number of features when visiting Tokyo for the Olympics. To do that, we gathered data from different datasets and the FourSquare API and then created clusters based on their attributes and compared them, narrowing the StakeHolder choices in the trip planning process.