Olympic 2020 Hostel Selection

Noppadon Sakulsom

Introduction

In 2020, an Olympic game will be held in Tokyo, Japan. Beside the difficulty of purchasing any ticket, one of the problems most people would face is how to choose an accommodation. Which one will be in good quality, value for money, and close to many Olympic venues? Not only individuals who come to watch the game, but also many organizations such as news agencies which have to find accommodations for their officers. In this project, I will propose an approach to choose accommodations based on the locations of Olympic venues using data science.

Data

In this project, I used a couple set of data:

- The name and location (Latitude and Longitude) of each Olympic venue from 2020 Olympic website (https://tokyo2020.org/en/games/venue/olympic/). However, I can only scrape the name and sport of each venue. Although the locations can be found in the website, I have to directly put them into the Jupyter notebook.
- The dataset of hostel in Japan (from https://www.kaggle.com/koki25ando/hostel-world-dataset) which contains the name, location, and rating of each hostel.
- The dataset of places people check-in in Tokyo on Foursquare (from https://www.kaggle.com/chetanism/foursquare-nyc-and-tokyo-checkin-dataset).
- The data of interesting places around the hostel queried on Foursquare.

I clustered the Olympic venues into groups and clustered the hostels based on their distances from the Olympic venues. Then the hostels in each group were ranked by their score based on distance and rating. Then the groups of top-rank hostels in each cluster were re-evaluated based on their surrounding places based on Foursquare data.

Methodology

I explored the hostels dataset to select the features to represent the data and I used k-means to cluster the Olympic venues and hostels.

Exploratory Data Analysis

First, I explored the data of hostels in Japan. The main features people looking for hostels are quality and price. Therefore, I considered the summary score and price of hostels. The dataset had to be cleaned as there were some rows with unreasonable value. After deleting those rows, a boxplot displaying the distribution of price/day (yen) of each rating band is shown in Fig.1. Furthermore, a boxplot displaying the distribution of summary score of each rating band is shown in Fig.2

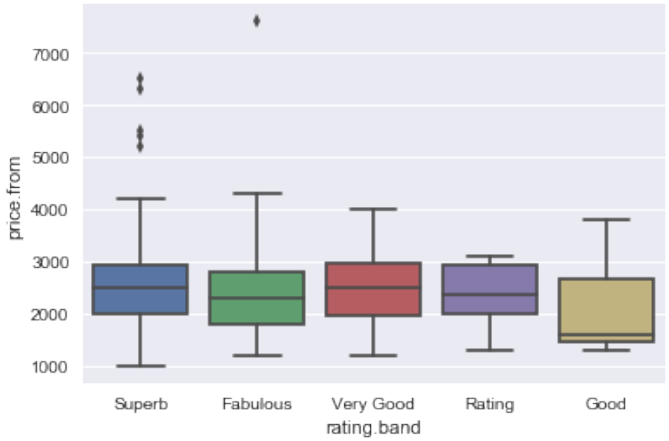

Fig.1. Boxplot displaying the distribution of price/day of each rating band

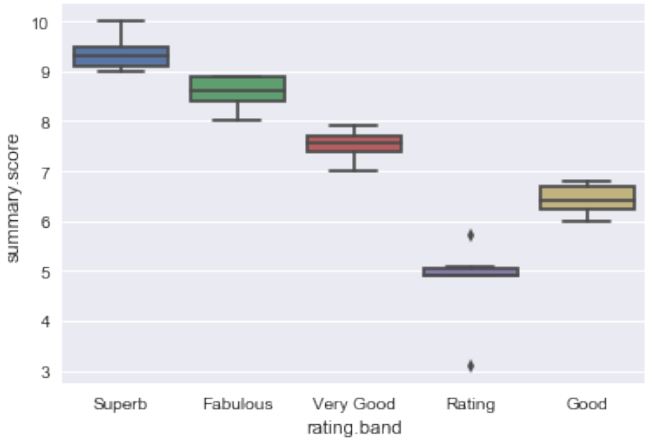

Fig.2. Boxplot displaying the distribution of summary score of each rating band

From Fig.1 and Fig.2, the distribution of price and summary score of hostels look good. However, as the rating band 'Rating' and 'Good' have the highest score less than 7/10, I decided to delete these categories.

The next question was that, as the dataset had many features related to summary score, we could use only summary score to represent the quality of the hostel or not. I calculated the correlation of all features to summary score as table 1. The correlation showed that the summary score represented all other features i.e. security, staff, atmosphere, cleanliness, value for money, and facilities. Therefore, I used the summary score to represent the quality of the hostels. An example of regression plot between summary score and cleanliness is shown in Fig.3. However, the price is not correlated to any other features as shown in table 2.

Table 1 Correlation of summary score and all other features

Longitude	-0.134470
Distance	-0.049374
Latitude	-0.045817
price.from	0.102653
location.y	0.501249
security	0.710018
staff	0.732762
atmosphere	0.760614
cleanliness	0.767922
valueformoney	0.777186
facilities	0.803625
summary.score	1.000000

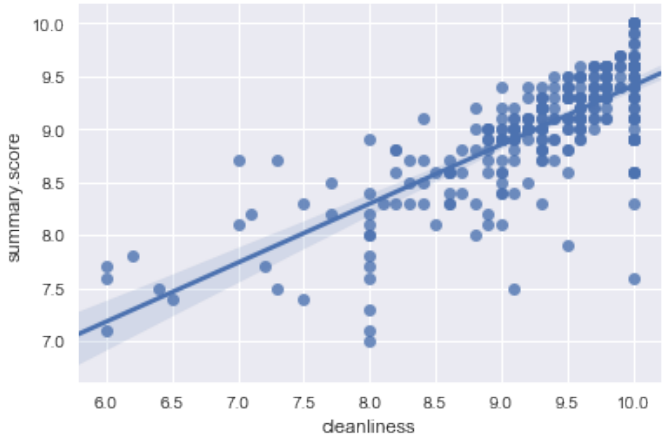

Fig.3. Regression plot between summary score and cleanliness

Table 2 Correlation of price and all other features

valueformoney	-0.065935
Distance	-0.036084
Latitude	0.012017
facilities	0.045938
security	0.064348
Longitude	0.072808
location.y	0.084771
atmosphere	0.100524
summary.score	0.102653
staff	0.105473
cleanliness	0.176612
price.from	1.000000

Machine Learning

I started by clustering the Olympic venues into 6 groups using k-means as I wanted to find the groups of venues that located close to others in the same group. The clusters are displayed in Fig. 4. There are two clusters that are too far from other clusters and two clusters that are outside Tokyo. As I wanted to focus on venues in Tokyo, I deleted these clusters. Then, I clustered the remaining venues around Tokyo into 4 groups as shown in Fig. 5.

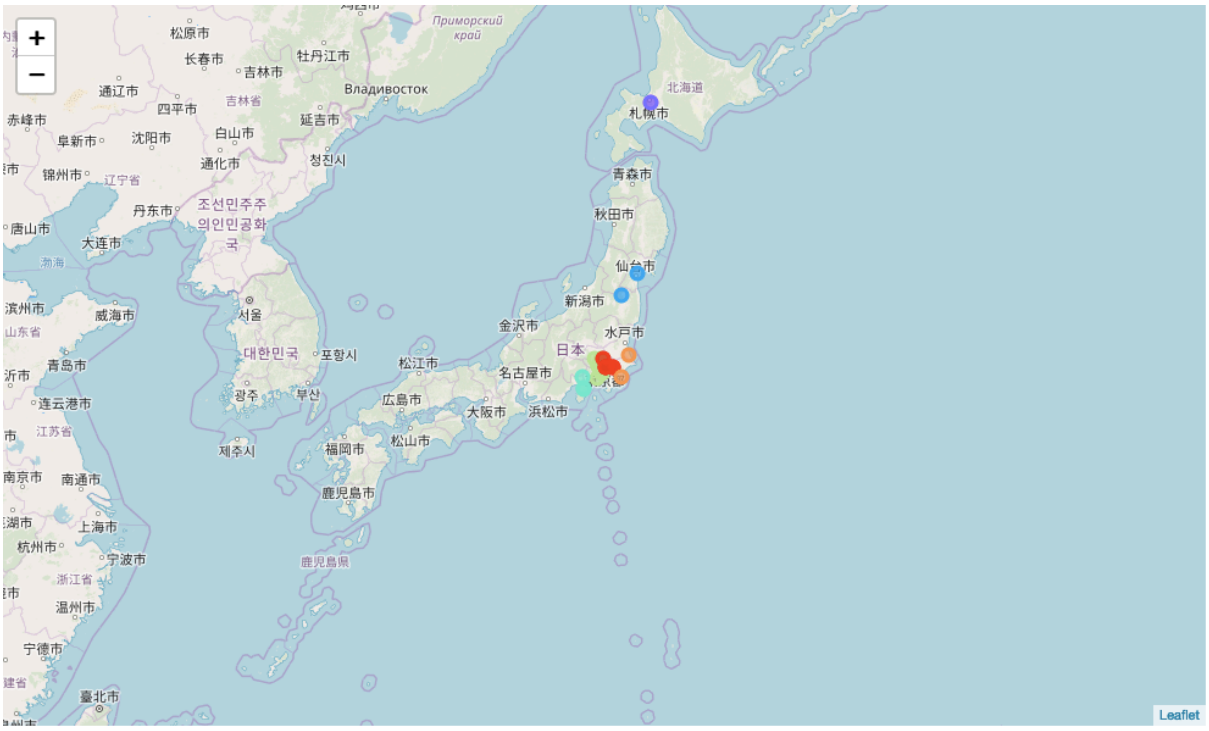

Fig.4. Clusters of Olympic venues (6 groups)

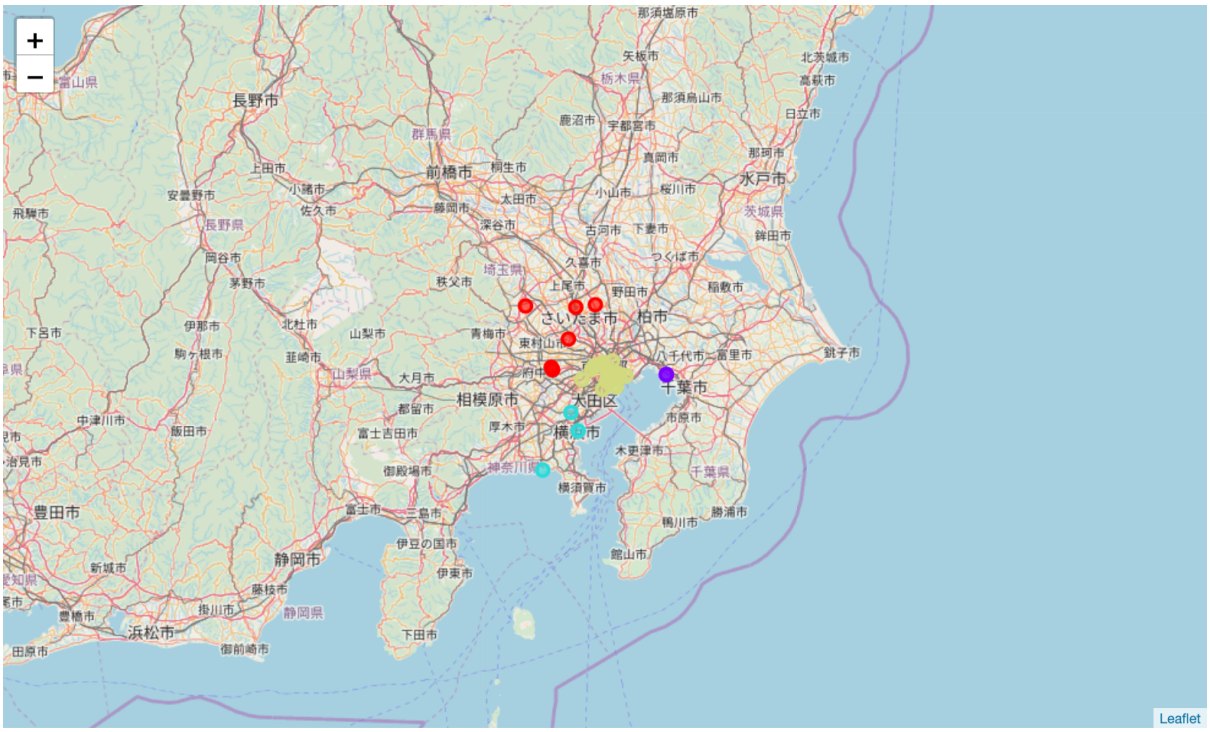

Fig.5. Clusters of remaining Olympic venues (4 groups)

Then, I worked on hostels data. I calculated the distances between each hostel and every Olympic venue. Then, I clustered Tokyo hostels into 5 groups (I tried other numbers of clusters and I found that this number worked well) based on these distances using k-means. The clusters of hostels and Olympic venues around Tokyo are shown in Fig.6. The warm-tone circles represent Olympic venues and cool-tone points represent hostels. There was a cluster located among a cluster of venues (the yellow circles) and the other clusters of hostels located between the yellow cluster of venues and other clusters of venues.

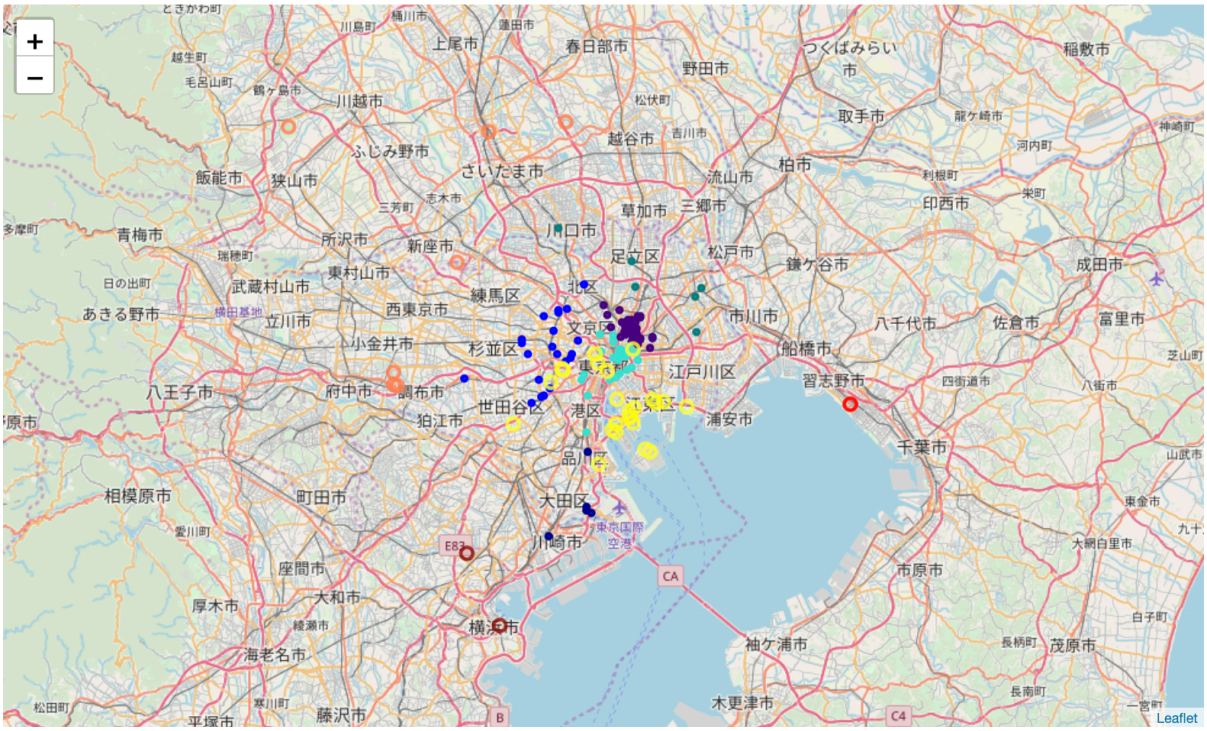

Fig.6. Clusters of hostels and Olympic venues around Tokyo

However, I thought that grouping venues based on distances would not be reasonable. I tried clustering the hostels based on new criterion - the hostels with shorter distance will have higher points. I re-calculated the features as follows:

$$NewCriterion_{ij} = (MaxDistance_j - Distance_{ij})/ (MaxDistance_j - MinDistance_j)$$

Where,
$Distance_{ij}$ = distance from hostel i to venue j
$MaxDistance_j$ = maximum distance from every hostel to venue j
$MinDistance_j$ = minimum distance from every hostel to venue j
$NewCriterion_{ij}$ = new criterion between hostel i and venue j

The clusters of hostels and Olympic venues around Tokyo based on new criterion are shown in Fig.7. The new clusters of hostels were a bit different from the old ones but I believed that on a bigger dataset this would lead to a big difference.

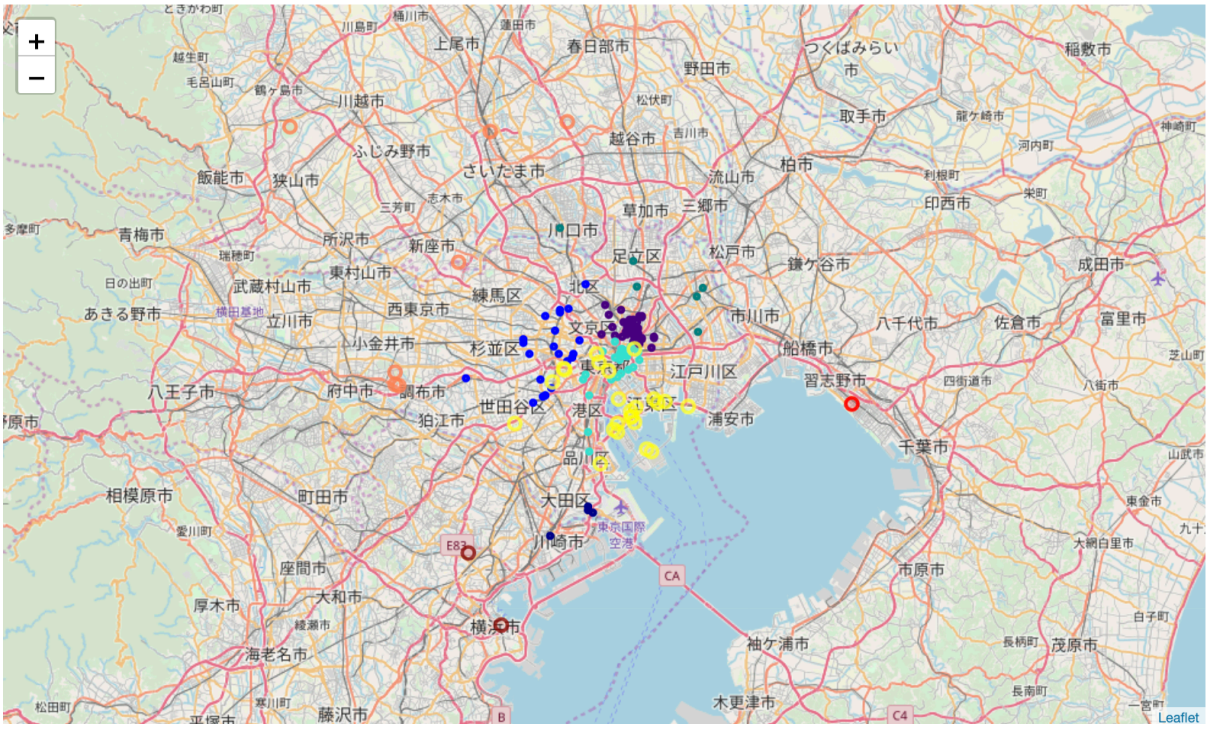

Fig.7 Clusters of hostels and Olympic venues around Tokyo based on new criterion

After I had clusters of hostels, I chose hostels from each cluster. I focused on the hostel's summary score and distances from clusters of Olympic venues. We would like to choose a hostel that has high quality (which means high summary score) and locates close to Olympic venues. Therefore, I calculated the distances from hostels to the centroid of each cluster of venues. Then, I calculated the ratio of hostel's summary score divided by these distances (the summary rating was the summation of all clusters). Therefore, I had this rating on each cluster of venues for every hostel. However, this rating cannot be directly used since it has different ranges of values on each cluster. A boxplot displaying the distribution of rating of each cluster is shown in Fig.8.

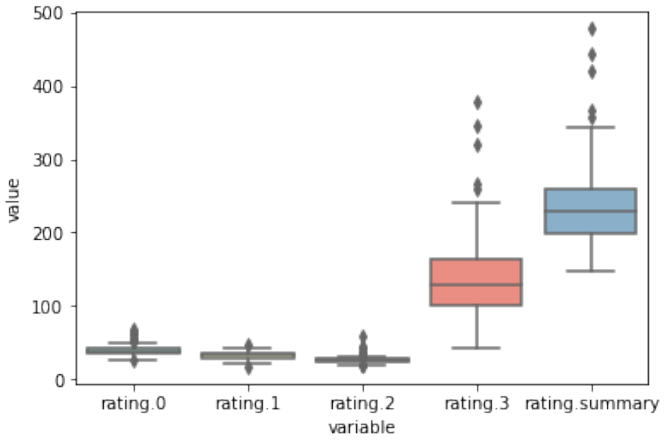

Fig.8. A boxplot of the distribution of rating of each cluster

The rating on each cluster was normalized by dividing by the maximum value of each group and the range of a new rating was between 0 to 1 (the summary rating was the average value of all clusters). A boxplot displaying the distribution of new rating of each cluster is shown in Fig.9.

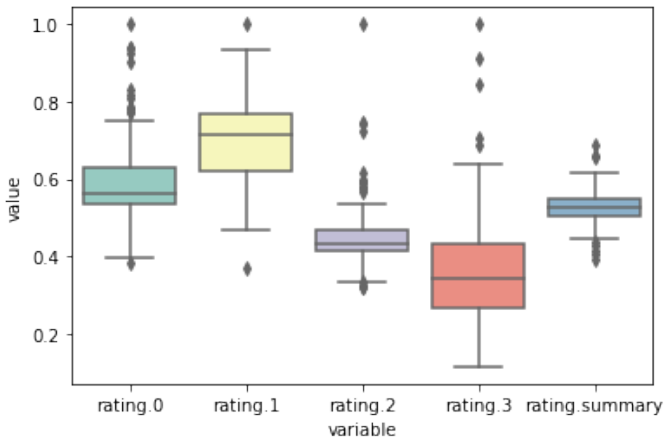

Fig.9. A boxplot of the distribution of new rating of each cluster

I chose 5 candidates that were the top-rating hostels from each cluster of hostels (except cluster 2 which has only 4 members). Then, the final decision was based on the dataset of places people check-in in Tokyo on Foursquare. I calculated the number of check-ins on each category of places. If the hostel was surrounded by places in popular category, it would have high score. Then, I queried the places around each candidate hostel on Foursquare API and gave points based on the popularity of places around the hostel. From the first 10,000 rows, the number of check-ins in top 20 categories is shown in Fig.10.

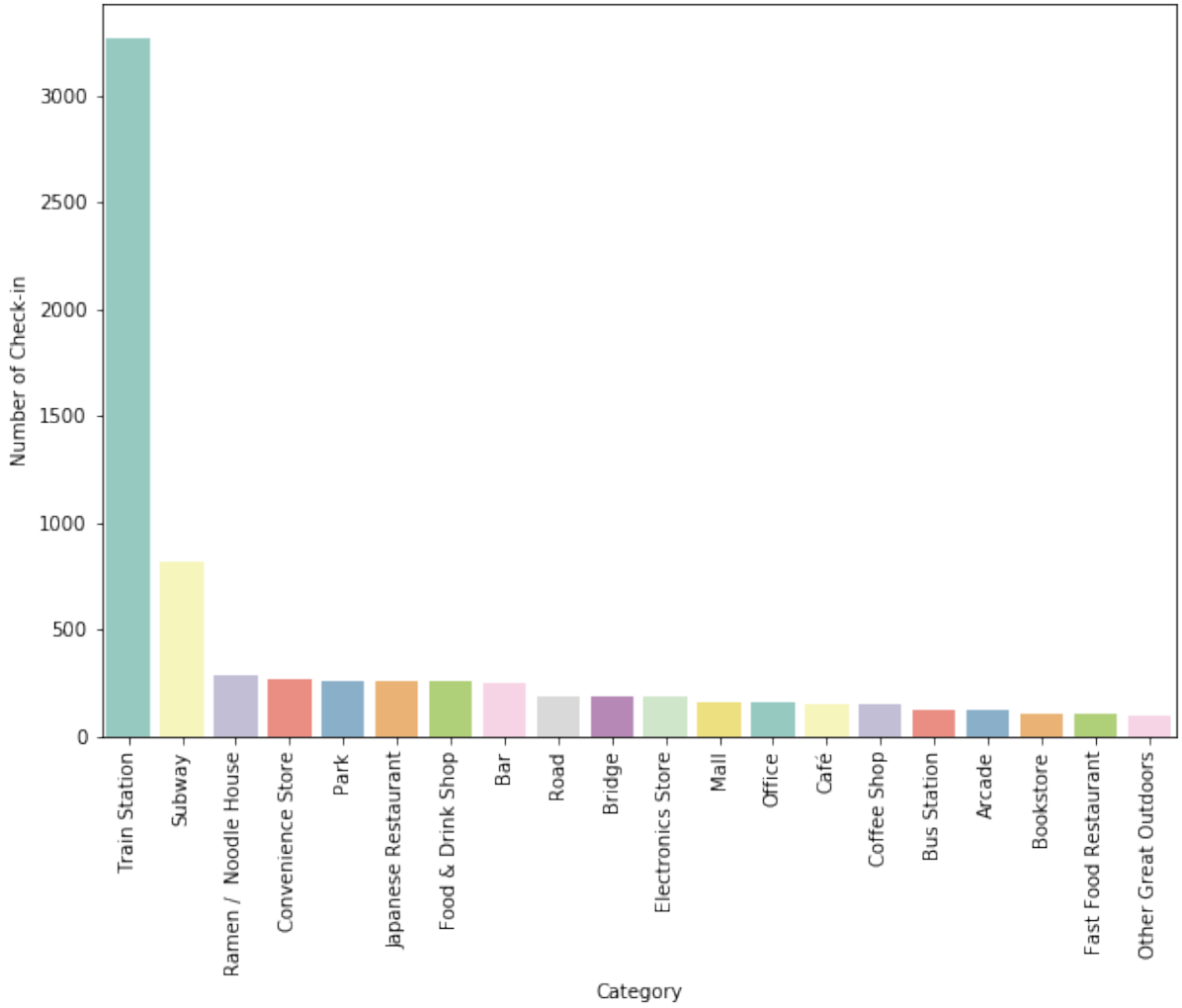

Fig.10. The number of check-ins in top 20 categories

For each place in these categories around the hostel, one point was given to the hostel. However, as 'Train Station' and 'Subway' were actually related to the same transportation system and they had far more check-in number than other categories, two points were given for each spot in these categories. Then, the result is as follows.

cluster: 0
There are 30 around the hostel The Prime Pod Ginza Tokyo with score of 4
There are 30 around the hostel Imano Tokyo Ginza Hostel with score of 7
There are 30 around the hostel Wise Owl Hostels Tokyo with score of 5
There are 30 around the hostel Oakhostel Cabin with score of 6
There are 30 around the hostel Zabutton Hostel with score of 6
cluster: 1
There are 30 around the hostel YADOYA Guesthouse Orange with score of 5
There are 22 around the hostel Tokyo Hostel Fuji with score of 9
There are 29 around the hostel Sheena and Ippei with score of 17
There are 30 around the hostel Hiromas Inn Chitose-Karasuyama with score of 4
There are 30 around the hostel YADOYA Guesthouse Green with score of 6
cluster: 2
There are 30 around the hostel Hotel&Hostel On the Marks Tokyo Kawasaki with score of 4
There are 30 around the hostel Guesthouse Sensu with score of 12
There are 30 around the hostel 328 Hostel & Lounge with score of 10
There are 30 around the hostel Beagle Tokyo Hostel & Apartments with score of 12
cluster: 3
There are 30 around the hostel Hostel Chapter Two Tokyo with score of 8
There are 30 around the hostel Oakhostel Sakura with score of 6
There are 29 around the hostel Guesthouse Fete with score of 18
There are 30 around the hostel Khaosan Tokyo Origami with score of 5
There are 30 around the hostel Iza Tokyo Asakusa Guesthouse with score of 10
cluster: 4
There are 23 around the hostel YAWP! backpackers with score of 12
There are 30 around the hostel Edo Tokyo Hostel with score of 8
There are 7 around the hostel J's Backpackers with score of 2
There are 30 around the hostel Emblem Hostel Nishiarai with score of 14
There are 25 around the hostel Guesthouse Kagaribi with score of 9

The hostel with the highest score is the hostel "Guesthouse Fete" with score of 18 from cluster 3.

Results

Four datasets are used to choose the hostel. Clusters of venues and clusters of hostels are obtained using k-means method. The k-means method can give a rough sense of how venues and hostels are related and visualize that relationship. However, as the distance from venues is not the only criterion, rating ratio (hostel's summary score divided by distance from cluster of venues) is calculated. The clusters from k-means and this rating ratio give candidates of hostels with more detail. Then, the final decision is made based on the surroundings obtained from Foursquare. I put this step as the final step of the proposed approach because this step can be change based on user's preference. For example, a user who likes Indian food can give more points to category Indian restaurant.

Discussion

In this study, I used k-means to cluster the hostels based on distances from Olympic venues. When I directly used the distances, the result is different from when I used the normalized distances. Accordingly, I calculated the rating ratio of hostel's summary score divided by distance from each cluster of venues. this rating ratio also has to be normalized as it has different range among clusters. Therefore, I observe that I should always normalize the data before using any machine learning technique.

Conclusion

An approach to choose hostels based on price, quality, location and surroundings using data science is proposed. The proposed approach gives a group of candidates where users can then choose based on their preferences of surroundings such as types of restaurants or tourist attractions.

The approach can be extended to cover other accommodation such as hotels, service apartment or even AirBNB.