# Clustering and Analyses of Tokyo Hostels

Massao Mitsunaga

February 19, 2020

## 1. Data acquisition and cleaning

### 1.1 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.

### 1.2 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.

### 1.3 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.