# Determining Neighborhood Clusters in Tokyo, Japan

Kevin Goh

22 Jun 2021

## 1. Introduction

### 1.1 Background
As a businessman looking to start a business in a foreign city, one must understand where the underlying demand lies among several neighborhoods in the city. One such city which we are keen to study is Tokyo, which has over 100 neighborhoods, each with different commuters and residents.

### 1.2 Problem
Because available information about Japanese consumer businesses is not widely available in English, it may be difficult to understand which businesses are most commonly patronized by locals in the area. If a businessman wants to start a business throughout Japan and in particular Tokyo, he must have information about places which are not frequented by English-speaking people.

### 1.3 Interest
Companies intending to open businesses in Japan will be interested in understanding local consumer demand and in particular, the patterns across a large city such as Tokyo, in anticipation of increased consumer behavior during the Tokyo Olympics.

## 2. Data Acquisition and Cleaning

### 2.1 Data Sources

Data is from Wikipedia, which details all the neighborhoods in Tokyo. This is used together with the Foursquare API to retrieve the position information such as latitude and longitude. Additionally, the Foursquare API will be used to generate areas which are most commonly patronized by locals in the area.

### 2.2 Data Cleaning

It was found during data retrieval that not all neighborhoods from Wikipedia have an associated latitude and longitude in Foursquare API. These neighborhoods had to be removed from the dataset.

During retrieval of venue information, it was also found that certain neighborhoods had no venue information and produced null values. Again, these neighborhoods had to be removed from the dataset.

## 3. Exploratory Data Analysis

During the exploratory data analysis, it was found that some neighborhoods had fewer venues than others, and this could be due to the radius function, which is set to 500m from the neighborhood centre. It could also indicate that certain neighborhoods were not as dense in venues than others.

By performing one-hot encoding on the categories of the venues, the data was then processed to give the most common venues ranked from first to the tenth.

## 4. K-Means Clustering

The data was clustered using k=10. It was then found that a total of four neighborhoods had null values and thus did not have cluster labels. These neighborhoods were again dropped before visualization.

## 5. Results

The clusters were plotted and we see that there are two main clusters, Cluster 0 in red and Cluster 7 in yellow. Looking at the map, Cluster 0 is in red and is widespread across all areas of Tokyo, including the suburbs.

## 6. Conclusions and Recommendations

I would recommend to look at Cluster 0 or Cluster 7 to identify common neighborhoods to start businesses. Both clusters have a value count of 33.

Cluster 0 has most common venue is the convenience store followed by cafes and coffee shops. This could be neighborhoods with younger generations of people. Businesses can consider opening either a convenience store or a cafe to cater to this group of people.

Cluster 7 has many convenience stores but also restaurants and night spots such as sake bars. This could be neighborhoods with larger proportions of working adults or possibly adults who are more well to do and businesses can consider opening a restaurant or a bar in this cluster to cater to this group of people.

## 6. Future Work

The study can be extended to better understand the distributions of common venues within each cluster in order to make a better assessment of the neighborhood demographic. More tests can be done, including regression tests and hypothesis testing to assess if neighborhoods tend to have higher income or educational levels, etc. This will complete the assessment of the clusters and provide a more holistic recommendation to potential businessmen.