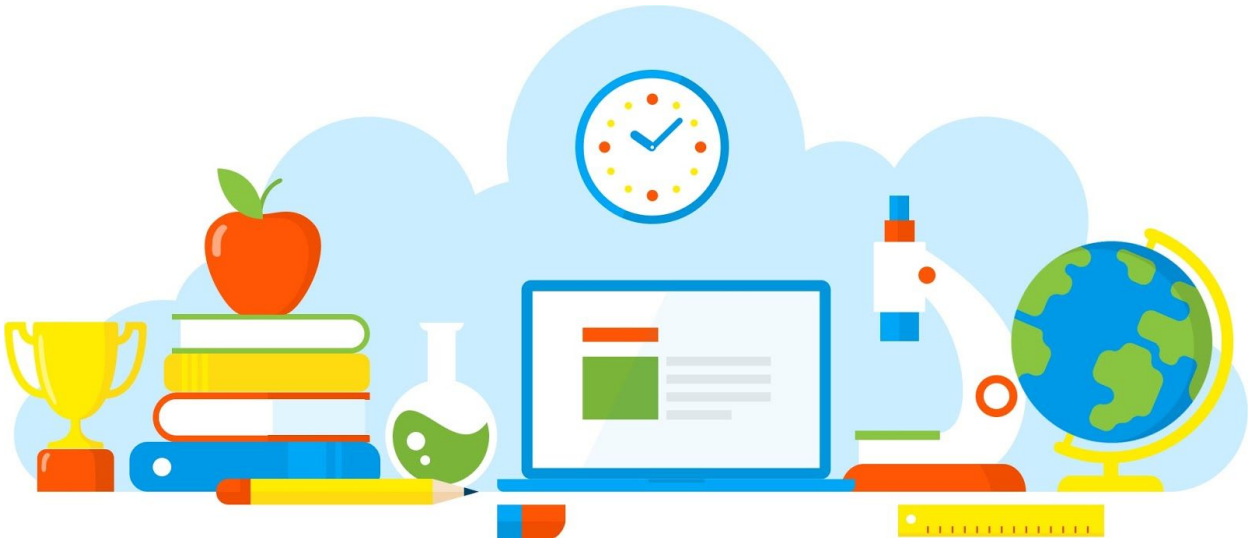


Capstone Project

Tokyo Clustering Based on Food, Entertainment, Sightseeing, Shopping

Melih Onur Temel
25.05.2020

Introduction

The capstone project is about 23 wards of City Tokyo and the municipality is gonna try to come up with a strategy to increasing demand of tourism in a best way possible. This is a imaginary scenario may be used freely anyone who wants to use it or adjust the solution to his or her case.

In the case study the municipality of Tokyo will try to classify the wards as a 3 groups and the main topics in order to determine it will be based on 4 categories.

These categories has been described based on intuition as “Shopping, Foods, Entertainment, Sightseeing and Accommodation”.

Name of the each ward listed as it is followed: 'Chiyoda', 'Chūō', 'Minato', 'Shinjuku', 'Bunkyō', ‘Taitō', 'Sumida', 'Kōtō', 'Shinagawa', 'Meguro', 'Ōta', 'Setagaya', 'Shibuya', 'Nakano', 'Suginami', 'Toshima', 'Kita', 'Arakawa', 'İtabashi', 'Nerima', 'Adachi', 'Katsushika', 'Edogawa'.

The solution will show the characteristics of the regions and the closeness between them. This may help the municipality to develop strategies such as connecting the regions having same features or regions not having the same characteristics to make the city colorful or based on specific projects as hub locations etc. The increasing number of tourists in city of Tokyo can be seen in Figure 1.

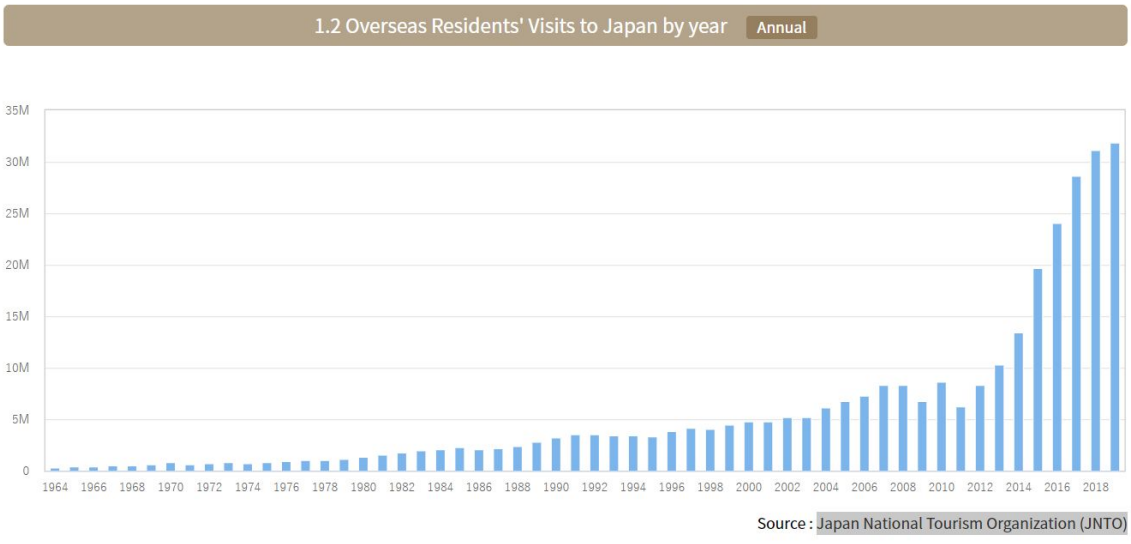


Figure 1: Tokyo Annual Tourist Numbers

Reference:

Japan-bound Statistics - Tourism Statistics. (n.d.). Retrieved from https://www.tourism.jp/en/tourism-database/stats/inbound/#annual

Business Understanding

The problem that we are trying to solve is about how to determine the wards and when making investments for further developments is will be useful to know the characteristics of the wards. This can be useful because each year the city of Tokyo attracts more tourists and also there is Olympics coming to city.

The goal is classifying the wards and clustre groups as 3 chosen randomly and it may be adjusted easily depending the decisions if it is used in real case scenarios. The solution will help the people to develop the areas based on their specific features and maybe connect them accordingly or expand the zone that does not offer the services which is offered by specific zone adjusted to other wards.

As it is said determining the areas based on their famous features will also help the tourism agencies to make better plans and routes for the tourists to go around the city and have fun and also this will increase the tourism income of the municipality.

Clustering approach has been chosen as the analytic approach. This will be used to come up with a model showing which wards has relationships with each other based on the four categories listed above.

Data Requirement

In order to solve the problem we need the geographical location data of each ward. This can be done using “geopy” library of python and 23 geographical location as latitude and longitude will be gathered after the correct implementation of the package. The format of the data in numerical values because it will be used in Foursquare API to collect the popular places around the spots given as parameters.

Data must be up to date data because the municipality areas are subject to change when time goes by.

Data collection

Data can be collected by using different sources and for the project Foursquare agent or Google can be used to obtain the geographical data and both will provide good results. For his case Foursquare API will be used.

After the location data gathered region must be inspected to find the popular places around specific geographical location.

Data Understanding

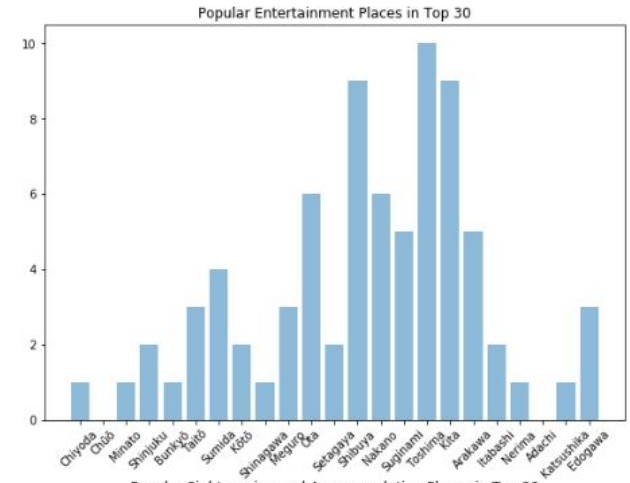
Visualization tools can be used to understand the data and geographical locations

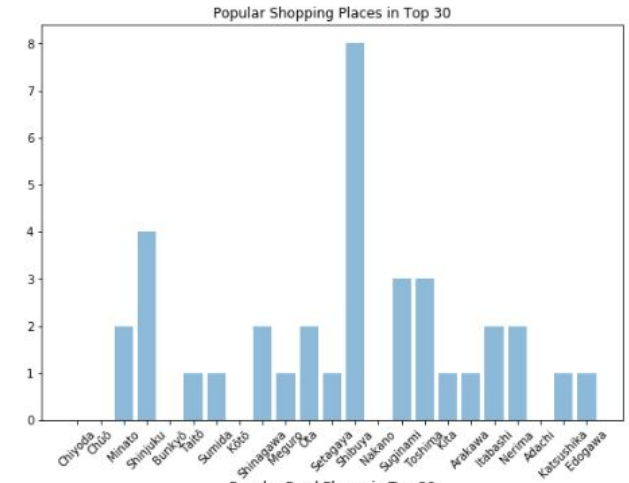
To find the popular venues for each category bar charts or other graphs may help us and they can depict the popular venue types

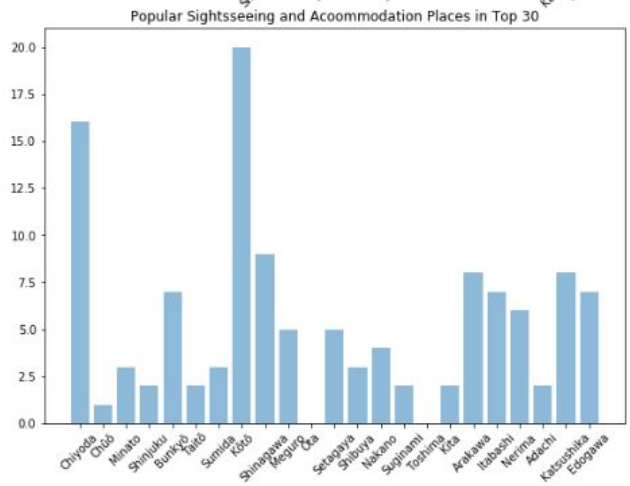
Methodology and Exploratory Data Analysis

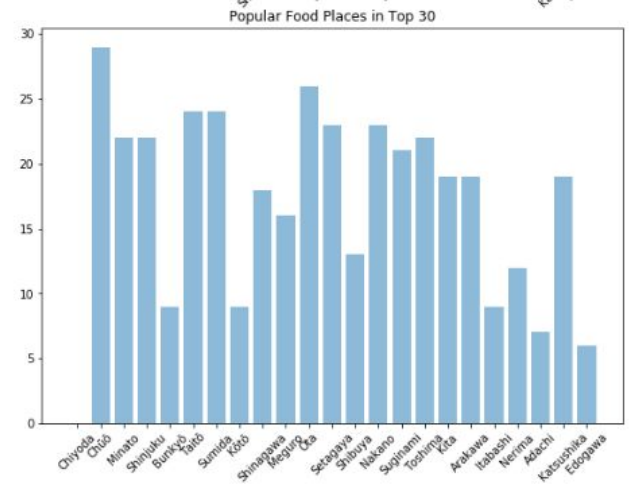
The geographical locations for each wards has been found. After that the popular spots and the types of the spot can be found. The assumption used here is that if the type of the location is described as 'Botanical Garden' or ‘Canal’ this is aggregated as Sightseeing spot or if the location is described as restaurant such as Japanese restaurant or Indian restaurant is is

aggregated as Foods. The bar charts for each category described below and the details can be found for each ward. The radius has been taken as 1000 meters and the location number for each ward taken as 30 as inputs. This is subjected to change and it can be done easily for further analysis to create more robust models in case it is needed.









The machine learning model has been used to come up with a solution and for the case clustering model which is K-Means model has been chosen among the other models. We had to use clustering model because we did not have the target label and this is one of the unsupervised machine learning models.

Results

Image provided by using Folium library.

Wards_number	Wards	Latitude	Longitude	Entertainment	Shopping	Sightseeing_Accommodation	Foods	Cluster Labels
0	Chiyoda	35.685402	139.752731	-0.814266	-0.899438	2.260594	-2.318431	2
1	Chūō	35.666256	139.775558	-1.161083	-0.899438	-0.909751	1.626450	1
2	Minato	35.643227	139.740051	-0.814266	0.249844	-0.487039	0.674238	1
3	Shinjuku	35.688362	139.699081	-0.467449	1.399126	-0.698395	0.674238	1
4	Bunkyō	35.718811	139.744736	-0.814266	-0.899438	0.358387	-1.094158	2
5	Taitō	35.717449	139.790863	-0.120632	-0.324797	-0.698395	0.946298	1
6	Sumida	35.700428	139.805023	0.226185	-0.324797	-0.487039	0.946298	1
7	Kōtō	35.649155	139.812790	-0.467449	-0.899438	3.106019	-1.094158	2
8	Shinagawa	35.599251	139.738907	-0.814266	0.249844	0.781100	0.130116	2
9	Meguro	35.621250	139.688019	-0.120632	-0.324797	-0.064326	-0.141945	1
10	Ōta	35.561207	139.715836	0.919819	0.249844	-1.121108	1.218359	1
11	Setagaya	35.646095	139.656265	-0.467449	-0.324797	-0.064326	0.810268	1
12	Shibuya	35.664597	139.698715	1.960269	3.697690	-0.487039	-0.550036	0
13	Nakano	35.718124	139.664474	0.919819	-0.899438	-0.275682	0.810268	1
14	Suginami	35.699493	139.636292	0.573002	0.824485	-0.698395	0.538207	1
15	Toshima	35.736156	139.714218	2.307086	0.824485	-1.121108	0.674238	1
16	Kita	35.755836	139.736694	1.960269	-0.324797	-0.698395	0.266146	1
17	Arakawa	35.737530	139.781311	0.573002	-0.324797	0.569743	0.266146	1
18	Itabashi	35.774143	139.681213	-0.467449	0.249844	0.358387	-1.094158	2
19	Nerima	35.748360	139.638733	-0.814266	0.249844	0.147031	-0.686066	2
20	Adachi	35.783703	139.795319	-1.161083	-0.899438	-0.698395	-1.366218	2
21	Katsushika	35.751732	139.863815	-0.814266	-0.324797	0.569743	0.266146	2
22	Edogawa	35.737705	139.896118	-0.120632	-0.324797	0.358387	-1.502249	2

The cluster labels can be seen on the column positioned on the left. The name of the locations stated in Japanese therefore the image of the data frame added as well. The values for entertainment, shopping etc listed as normalized values around N(0,1) which is mean of 0 and standard deviation of 1 because for K- Means algorithm using the non transformed values will give wrong clusters due to weight difference and it is done by “Z transformation”

Discussions and Conclusion

Results show us the relationships between the wards and according to that 9 wards depicted in green dots, 1 ward which is Shibuya ward depicted in red dot and all the remainings are depicted in blue dots based on the 4 different groups made by aggregation of type of the popular placed on the ward selected. This is interesting because when we look at the northeast region the wards shows nearly same pattern and the most of the center show us the other pattern except Shibuya ward of the Tokyo. The unique region can be promoted differently than the other regions if the municipality decides to promote them in future. The further analysis such as more

than 30 places to explore or wider radius choice must be implemented if it is really done on actual project but this small frame gives hint about how it may be done and good illustration of the basic points. connecting the regions sharing same characteristics with different means of transportation could be another choice if the desire is based on the centralization of unique characteristics such as Chiyoda and Koto for sightseeing and accommodation spots. Increasing the tourism agencies or implementing the western or pacific region country type hotels may be another strategy but to decide the efficacy and effectiveness of the project other socio-economic perspectives must be taken into account. This is just a small frame but it gives precious hints for the decision makers and illustrates the situation.