

IBM Applied Data Science Capstone Project

The Battle of Neighborhoods – Which Tokyo Ward(s) Should a Tourist Visit?

Introduction

Covid-19 has dealt a serious blow to tourism in countries around the world, and Japan is no exception. Tokyo, the capital and the most visited city in Japan, has suffered an 87% drop in number of tourists in 2020 alone. Even with a vaccine, a return to pre-Covid days is unlikely, or slow at best. In the latter scenario, future travelers are probably limited in the areas they could explore, and it would be best if they could have an efficient travel plan where they are able to enjoy the most number of experiences in the shortest period of time. One strategy to achieve this would be to visit areas with a high concentration of attractions, food places, and entertainment where it is easy to navigate by foot, and the venues are close to each other.

The Business Problem

As the most popular city among tourists in Japan, Tokyo must undoubtedly have the largest concentration of attractions in Japan. Each Ward, as Tokyo is split into, correspondingly has its own cluster of districts filled with attractions, and food and entertainment places. The question for the **future traveler** is then:

Which Ward(s) should they explore so that they can have the most experiences in the shortest amount of time?

The Dataset

To address this problem, we could start with a dataset that examines the number of venues each Ward has in its major districts. It stands to reason that Wards with more attractions and food and entertainment places allow for more experiences and more efficient exploration. We would also be examining the latitude and longitude of the Wards and districts.

To get this dataset, information about the Wards and major districts in Tokyo was scrapped from the Wikipedia page https://en.wikipedia.org/wiki/Special_wards_of_Tokyo using BeautifulSoup. Python's geopy package was then used to obtain the latitude and longitude of all the major districts and merged with the former data.

Using Pandas, I have created a dataframe from this dataset:

	Ward Name	Major District	Latitude	Longitude
0	Chiyoda	Nagatacho	35.675618	139.743469
1	Chuo	Nihonbashi	35.684068	139.774503
2	Minato	Odaiba	35.619050	139.779364
3	Shinjuku	Shinjuku	35.693763	139.703632
4	Bunkyo	Hongo	35.175376	137.013476
5	Taito	Ueno	35.711795	139.776075
6	Sumida	Kinshicho	35.696312	139.815043
7	Koto	Kiba	23.013134	-80.832875
8	Shinagawa	Shinagawa	35.599252	139.738910
9	Meguro	Meguro	35.621250	139.688014
10	Ota	Omori	35.588473	139.727933
11	Setagaya	Shimokitazawa	35.661678	139.666335
12	Shibuya	Shibuya	35.664596	139.698711
13	Nakano	Nakano	35.718123	139.664468
14	Suginami	Koenji	35.704942	139.649909
15	Toshima	Ikebukuro	35.731084	139.708916
16	Kita	Akabane	35.778139	139.720800
17	Arakawa	Arakawa	35.737529	139.781310
18	Itabashi	Itabashi	35.774143	139.681209
19	Nerima	Nerima	35.748360	139.638735
20	Adachi	Ayase	35.446047	139.430823
21	Katsushika	Tateishi	33.481791	131.478154

Methodology

There are 23 Wards in Tokyo, but we will only be focusing on the central 5 wards, or affectionately referred to as the C5W, of Minato, Shinjuku, Shibuya, Chuo, and Chiyoda:

	Ward Name	Major District	Latitude	Longitude
0	Chiyoda	Nagatacho	35.675618	139.743469
1	Chuo	Nihonbashi	35.684068	139.774503
2	Minato	Odaiba	35.619050	139.779364
3	Shinjuku	Shinjuku	35.693763	139.703632
12	Shibuya	Shibuya	35.664596	139.698711

These Wards are so popular that rent and property prices in these Wards are the highest among all Wards in Tokyo. The plan then is to use the Foursquare Place API to generate the top 100 most popular venues and their categories in each of the 5 Wards across all their major districts.

An issue I faced earlier on was that each Ward comprises of many major districts, but due to the limitations of BeautifulSoup, I could only scrap the name of one major district from each Ward. To get around this problem and explore other major districts in the same Ward, I set the radius for Foursquare API to be 5000 meters. By doing so, I was able to capture the venues in other major districts in the Ward to create a comprehensive dataframe as the major districts are close to each other in their Wards.

Another issue was that the naming of the Wards and their major districts is not consistent. Some Ward and major district names were in Romaji while others were in English. Some also contained two words in their names while others consisted of only one. As such, I did some basic data cleaning to standardise the naming convention and convert the names to English. For example, I changed 'Chiyoda, Tokyo' to 'Chiyoda' and 'Chūō' to 'Chuo' for the Wards, and 'Nagotachō' to 'Nagatacho' for the major districts.

I then used the Folium library to create a map of Tokyo with the 5 Wards superimposed on it with their corresponding major district as shown below:



As mentioned above, I used the Foursquare Place API to generate the top 100 most popular venues and their categories within a 5000m radius of each of the 5 Wards across all their major districts. I then created a Pandas dataframe with one hot encoding for the venue categories, grouped them on the 'Major District' column and calculated their means. The analysis results for the top 5 venues categories of each Ward are as below:

----Nagatacho----

	venue	freq
0	Convenience Store	0.16
1	Coffee Shop	0.12
2	Japanese Restaurant	0.09
3	Hotel	0.06
4	Intersection	0.06

----Nihonbashi----

	venue	freq
0	Café	0.06
1	Japanese Restaurant	0.05
2	BBQ Joint	0.05
3	Hotel	0.04
4	Department Store	0.04

----Odaiba----

	venue	freq
0	Japanese Restaurant	0.08
1	Convenience Store	0.08
2	Bus Stop	0.05
3	Coffee Shop	0.05
4	Bar	0.05

----Shibuya----

	venue	freq
0	Café	0.11
1	Record Shop	0.07
2	Ramen Restaurant	0.04
3	Rock Club	0.04
4	Nightclub	0.03

----Shinjuku----

	venue	freq
0	Bar	0.08
1	Sake Bar	0.08
2	Ramen Restaurant	0.07
3	BBQ Joint	0.05
4	Japanese Restaurant	0.04

Next, I put the top 10 venue categories of each Ward into the dataframe below:

	Major District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Nagatacho	Convenience Store	Coffee Shop	Japanese Restaurant	Intersection	Hotel	Ramen Restaurant	Soba Restaurant	Café	Steakhouse	Supermarket
1	Nihonbashi	Café	BBQ Joint	Japanese Restaurant	Yoshoku Restaurant	Department Store	Coffee Shop	Chinese Restaurant	Bakery	Hotel	Sake Bar
2	Odaiba	Japanese Restaurant	Convenience Store	Coffee Shop	Hot Spring	Bus Stop	Bar	Creperie	Smoke Shop	Science Museum	Scenic Lookout
3	Shibuya	Café	Record Shop	Ramen Restaurant	Rock Club	Italian Restaurant	Nightclub	Japanese Restaurant	Clothing Store	BBQ Joint	Indie Movie Theater
4	Shinjuku	Bar	Sake Bar	Ramen Restaurant	BBQ Joint	Japanese Restaurant	Dessert Shop	Pub	Chinese Restaurant	Rock Club	Department Store

Finally, I clustered the 5 Wards using the dataframe above with the K-Means clustering algorithm from the scikit-learn library. It had been expected that the clustering would be based on the similarities of venue categories.

Results

With K = 3, 3 clusters were obtained as shown in the leaflet map below:



Result: Shibuya, Shinjuku, and Chuo were grouped in one cluster, while Minato and Chiyoda had their own clusters.

	Ward Name	Major District	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Chiyoda	Nagatacho	35.675618	139.743469	1	Convenience Store	Coffee Shop	Japanese Restaurant	Intersection	Hotel	Ramen Restaurant	Soba Restaurant	Café	Steakhouse	Supermarket
1	Chuo	Nihonbashi	35.684068	139.774503	0	Café	BBQ Joint	Japanese Restaurant	Yoshoku Restaurant	Department Store	Coffee Shop	Chinese Restaurant	Bakery	Hotel	Sake Bar
2	Minato	Odaiba	35.619050	139.779364	2	Japanese Restaurant	Convenience Store	Coffee Shop	Hot Spring	Bus Stop	Bar	Creperie	Smoke Shop	Science Museum	Scenic Lookout
3	Shinjuku	Shinjuku	35.693763	139.703632	0	Bar	Sake Bar	Ramen Restaurant	BBQ Joint	Japanese Restaurant	Dessert Shop	Pub	Chinese Restaurant	Rock Club	Department Store
12	Shibuya	Shibuya	35.664596	139.698711	0	Café	Record Shop	Ramen Restaurant	Rock Club	Italian Restaurant	Nightclub	Japanese Restaurant	Clothing Store	BBQ Joint	Indie Movie Theater

Discussion

The results suggest that the Wards of Shibuya, Shinjuku, and Chuo have similar categories of food and entertainment places, and the time-pressed traveller need only visit one of them to have the same experience. As hotels are one of the most common venues in Chuo, it might be the best Ward for a traveler to stay and explore.

Another Ward of interest to visit might be Minato. Unlike other Wards, hot springs and scenic lookouts are among its most common venues and are not in the top 10 venue categories of other Wards. Minato therefore offers a unique experience that other Wards cannot provide and can be considered in the traveler's itinerary.

The results indicate then that Chuo and Minato Wards should be visited by a traveler to maximise his experiences in the shortest period of time.

However, there are some caveats to consider. First, the prices of these venues were not considered. While Chuo and Minato are suggested as the Wards to explore by venue categories, prices would undoubtedly play a huge role in the traveler's decision to explore and could perhaps be included in future studies. Another area to consider is the population density of the Wards. As mentioned in the introduction, it is unlikely that Covid-19 would have been fully eradicated in the near future, and the traveler would do well to visit only areas with lower population densities to keep himself or herself safe. As such, future studies should consider the population densities as well.

Conclusion

This project attempts to suggest which Tokyo Wards a traveler should visit to get the most experience within the shortest time as part of an efficient travel itinerary in a world ravaged by Covid-19. To do this, we have scrapped information of the Wards in Tokyo from the web, before

merging it with the longitude and latitude data of the Wards. We then used this dataset as inputs to the Foursquare Place API to generate the most common venue categories of each Ward and clustered them with the K-means algorithm. The results suggest that the **Chuo and Minato** Wards are the ideal Wards to visit. It is important to note, however, that this project is still preliminary and other factors such as the average prices of travel activity and population densities were not included to generate the results. This is something future projects could consider to offer a more comprehensive recommendation to the Tokyo traveler.