Tokyo Hostels

Jonathan Ong
Final Capstone
August 18, 2020

Overview

Hostel accomodation is a booming industry in tourism. Considering this once the state of the world is back to normal, those who are looking to travel and still save money might opt for hostels. It is important to consider what type of hostel the traveller would like to stay at, and with this model I intend to make that decision easier for the traveller. A few questions that might come up would be:

- How does the price vary with location?
- Where are the ‘highest value’ hostels located?
- How does proximity to public transit affect the hostels rating?
- Which hostels are in the safest area?
- What hostels are similar to the one I am interested in?

Target and Approach

This particular project will serve two target audiences:

1. Travellers: To help them make an informed decision when choosing a hostel and providing an in-depth analysis of hostels and their neighborhood.
2. Business Person: Provide useful information and insight using models which can help them open up their first or next hostel.

I intend to do this using both exploratory data analysis and prescriptive analytics. Through EDA, I will uncover hidden properties from the data I access and provide useful insights to the reader. Through prescriptive analytics, I will cluster similar hostels together, which can then be arranged by different parameters deemed useful by our audience

Data Requirements

Japan Hostel Dataset

This dataset was webscraped from Hostelworld by Koki Ando and is available on Kaggle. This is the base dataset we are workin with.

Foursquare API

This API helped me get the venues around the hostel which I used for both EDA and clustering analyses

Tokyo Land Prices

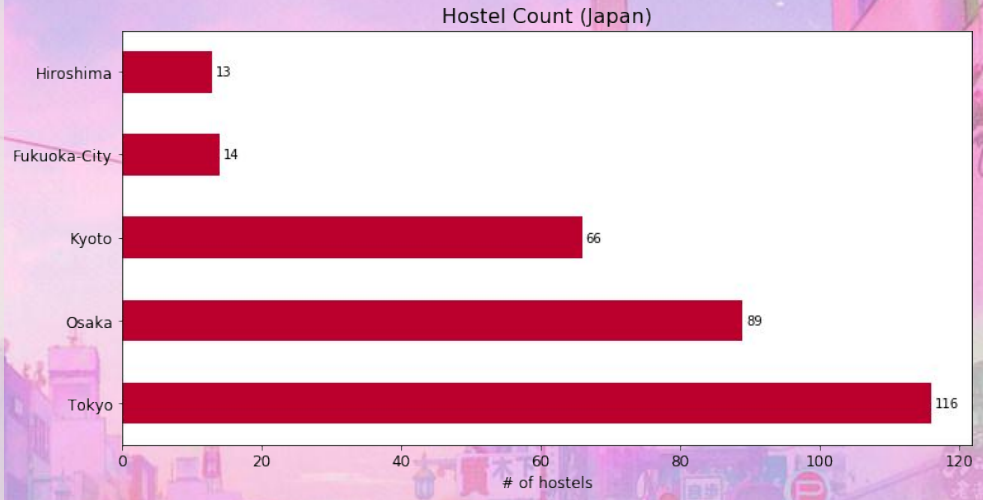
I will webscrape this website to get land prices per square meter for various neighborhoods in Tokyo.

Data Preparation and Pre-Processing

- Hostel Data
  - 342 hostels total, 44 missing latitude and longitude
  - Pull only Tokyo hostels, leaving a remaining 116 hostels
- Foursquare for Neighborhood Data
  - Use Foursquare for venues in the surrounding area
- Land Price Data
  - For each neighborhood in Tokyo, we gather it’s price per square meter
  - Tokyo has many neighborhoods considering how dense it is
- Use OpenCageGeocode to reverse geocode our hostels
  - Obtain the neighborhood from the reverse geocode and assign the neighborhood to the hostel
  - Remove any hostels without neighborhood listed

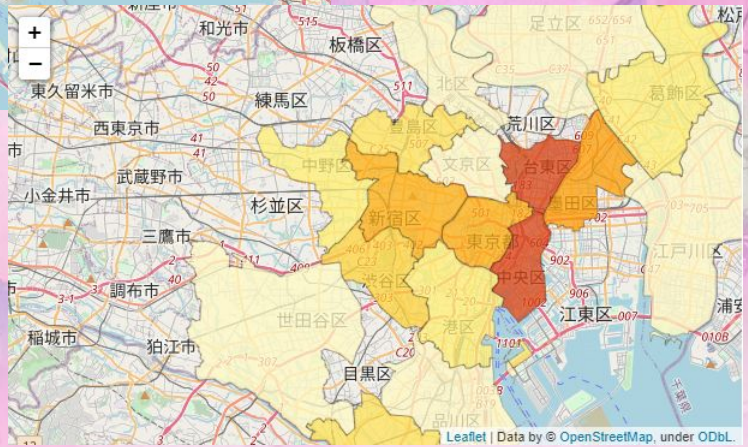
Data Exploration

- Most hostels are have listed are from Tokyo
- Quite a few more in Osaka and Kyoto, which can be used for more analysis later on, but solely will focus on Tokyo for this project



Data Exploration (cont.)

- We can then plot each hostel on a map of Tokyo to see the density of each area or neighborhood
- We can view this both by area and by density, indicated by the darker colors on the second map



Note for Investors

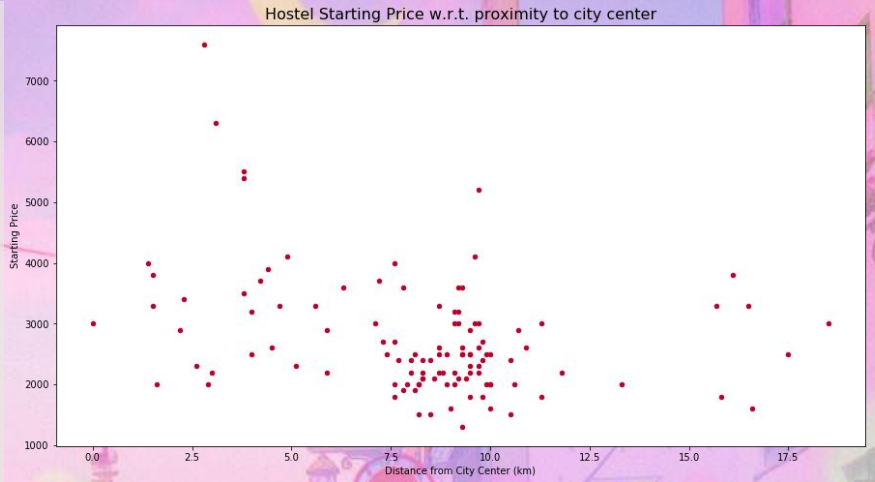
- One main observation for investors we see here is the popularity of a particular area known as Sumida-Ku
  - Sumida-Ku is 3rd in terms of hostel density and 5th least expensive
  - The price is almost 43% less than that of Taito-ku, which has by far the highest density of hostels in all of Tokyo
- Sumida-Ku could still be extremely profitable to start a new business in

```
cnt_price_df.sort_values(['Count', 'PricePerSqMeter'])
```

	Neighborhood	Count	PricePerSqMeter
12	Edogawa	1	332511
13	Adachi	2	295750
8	Setagaya	2	664106
6	Shinagawa	2	767398
5	Bunkyo	2	945155
10	Nakano	3	596154
2	Minato	3	2121252
11	Ota	4	560106
7	Toshima	4	731392
3	Shibuya	5	1360332
0	Chiyoda	9	2705898
9	Sumida	11	617190
1	Chuo	13	2699719
4	Taito	39	1064759

Price versus the distance from city center?

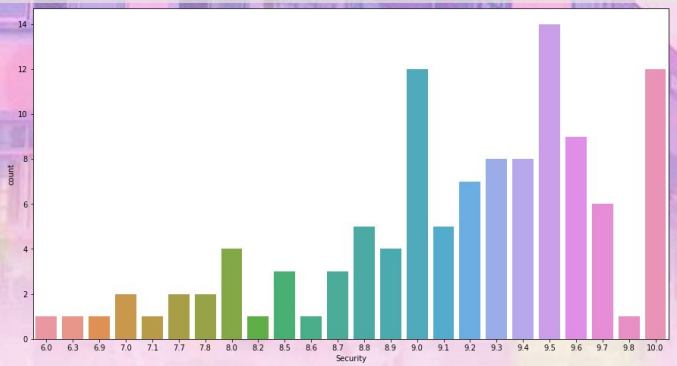
- As far as can be observed here, there is not much correlation between the price of a hostel versus the distance from the center of Tokyo City.
- Some negative correlation here is shown, but is likely negligible due to the impact being very little

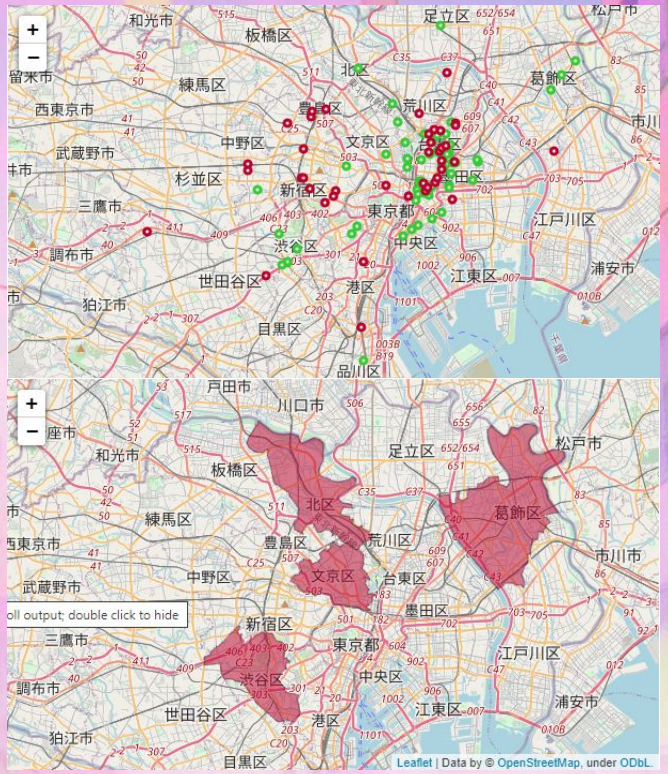


	StartPrice	DistanceFromCityCentre
StartPrice	1.00000	-0.32931
DistanceFromCityCentre	-0.32931	1.00000

Most secure hostels, where are they located?

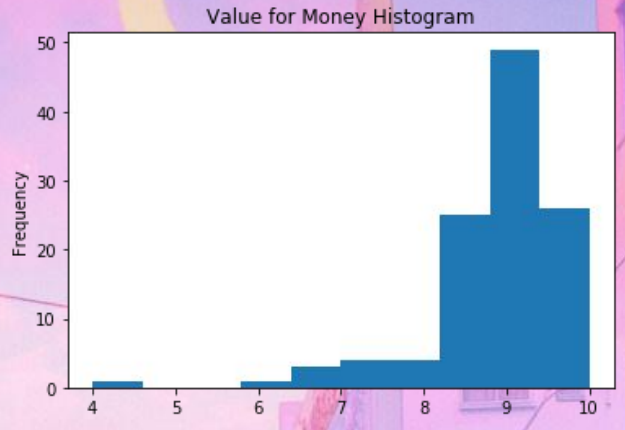
- Most hostels scored 9 and above on security, so anything below 9 we consider low security
- Hostels with the very high security score are in Shibuya, Katsushika, Bunkyo, and Kita





Where are the highest value hostels located?

- Using the Value for Money rating, we can see intuitively that the outskirts of Tokyo hold the best value for your money.
  - Even though the ones closest to the city center are not the most expensive, there seems to be other factors that affect the outer hostels' higher value for money
- The exception seems to be Chiyoda-Ku which is located very close to the center of the city.

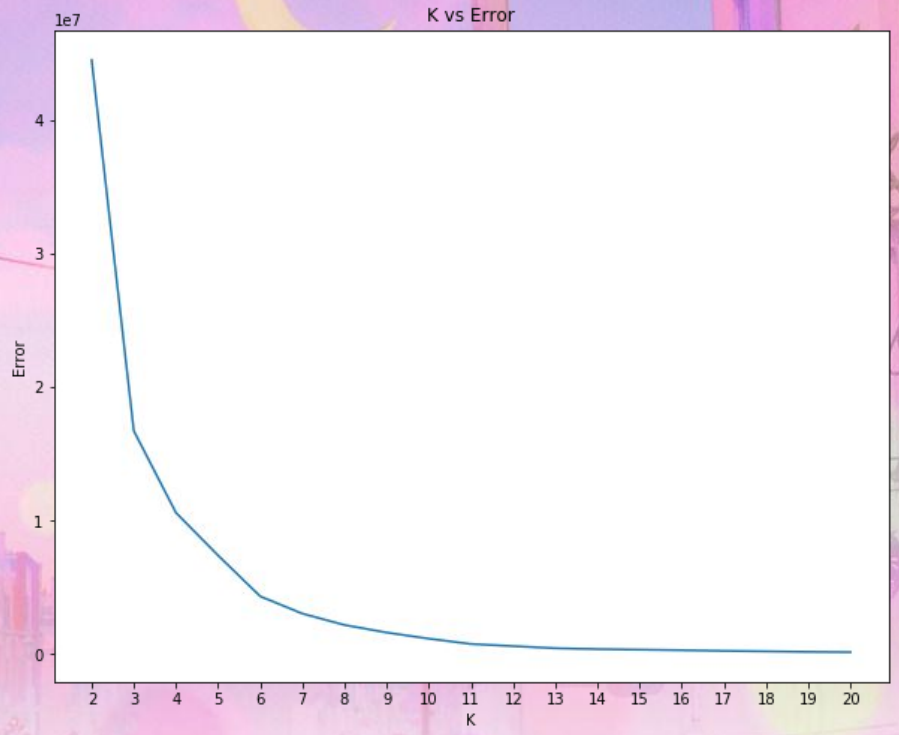


Other Questions Answered by EDA

- Worse hostels near metro stations?
  - The proximity of the hostel to any metro or train station has not much bearing on the rating of the hostel.
  - Slight negative correlation to the overall rating of the hostel,
- Which neighborhood venues affect a user's rating for location of hostel?
  - Hostels with proximity to a **park** are rated the same as in general; 30% rated “Fabulous” near a mall versus 33% overall
  - Hostels nearby **convenience stores** also achieved the same rating; 30% versus 33%.
  - The proximity to **historic sites** and venues has a positive effect on the rating of the hostel; 45% versus 33%.
  - The proximity of the hostels nearby **museums** has a slightly higher than average rating, 37.5% to 33%.

Clustering by Ratings

- The clustering method I ultimately chose after going through a few different methods was K-Means++
  - Provided me with more centric points and definitive clusters than other methods
  - Makes more sense due to the proximity of the points graphed, most variation can be explained by distance alone
  - Allows more for the elbow method to be applied, inertia clearly defines the amount of clusters present, 6



Cluster by Ratings Visualization

- Each separate cluster is represented by a different color
- Each cluster has somewhat specific traits associated with it, for example:
  - Cluster 0, the red markers, are all generally high cost, high rating, and mostly closer to city center. Outliers seem to have high ratings and high cost as well
  - Cluster 4 have moderately high cost, averages slightly further from city center, rated moderately high

Cluster Uses

- Now what this clustering allows us to do now is to select a specific hostel and find similar hostels to it
- For example, if I wanted to stay at ***Retrometro Backpackers*** but it was fully booked, I can use this model to find similar hostels to it
  - Sorting by 'Overall Score' we can see that Hotel Bedgasm is the highest rated hostel in that same cluster
  - Can use this method for any other sorting parameter as well

```
tokyo_hostels_df[tokyo_hostels_df.RatingCluster == 3].sort_values(['OverallScore
```

	Name	City	StartPrice	DistanceFromCityCentre	OverallScore	RatingCategory	Atmosp
132	Hostel bedgasm	Tokyo	2900	9.5	9.6	Superb	
329	Unplan Kagurazaka	Tokyo	3200	4.0	9.5	Superb	
275	Retrometro Backpackers	Tokyo	3000	9.2	9.4	Superb	
227	Lyuro Tokyo Kiyosumi - The Share Hotels-	Tokyo	3200	9.1	9.4	Superb	
6	328 Hostel & Lounge	Tokyo	3300	16.5	9.3	Superb	

Cluster by Venues in Neighborhood

- Many people also travel not for the stay itself, but for the surrounding area. Sights to see, places to visit are very important to a tourist.
- Here we will cluster by the most frequently appearing venues nearby the hostel and cluster by the most similar hostels
  - We can categorize the top 5 most frequently appearing venues around each hostel. From what I found, convenience stores, coffee shops, and any type of restaurant are most frequent, so 3 clusters makes sense

```
---- &And Hostel Akihabara ----
                venue  freq
0            Sake Bar  0.14
1    Ramen Restaurant  0.09
2           BBQ Joint  0.04
3          Hobby Shop  0.03
4  Chinese Restaurant  0.03


---- &And Hostel Ueno ----
                venue  freq
0   Convenience Store  0.17
1    Ramen Restaurant  0.10
2               Hotel  0.05
3  Chinese Restaurant  0.04
4         Coffee Shop  0.04


---- &And Hostel-Asakusa North- ----
                venue  freq
0  Convenience Store  0.19
1              Hotel  0.16
2           Sake Bar  0.11
3             Hostel  0.05
4               Park  0.05
```

Cluster by Venues Visualization

- The clusters are very much based on area, as visualized in the map here. Which makes sense, but there are few outliers or overlapping clusters which is a good thing
  - If someone is looking for a similar hostel to one in the city but wants to be less centralized in the heart of Tokyo, the option is there for them with a similarly surrounded hostel

Cluster by Venues Properties

- Each cluster has its own properties, but there's much more than just it's 1st Most Common Venue. There are 2nd, 3rd, 4th, and 5th most common as well, which can be taken into account, but for sake of simplicity, we consider the 1st Most Common Venue first
  - Cluster 1: Cafes and ramen shops by far. So if the user is looking to staying nearby more food cafes, authentic Japanese ramen shops, or the like.
  - Cluster 2: Convenience Stores. For a more on-the-go type stay, users may search for convenience stores to purchase simple items before heading off to busy parts of the city. This might prove ideal for them.
  - Cluster 3: Strip Mall. After a bit of research, these areas are surrounded more by a combination of stores, restaurants, and convenience stores. Users looking for more of a variety of shopping types, this cluster would be useful for these types of travellers

Recap

Overall, we got a glimpse of Tokyo's hostel scene with some interesting insights useful to those with either travelling or business interests.

Additionally, we clustered hostels in two ways:

1. By rating - for travellers to find the highest quality stays possible within their preferences and budget; for business person to find best area to start a hostel or what to improve with their current hostel
2. By venue - for travellers to find hostels with particular surrounding venues to visit during their stay; for business person to find most strategic starting area to build their first/next hostel

Conclusion

To further this research, I would want to pursue other methods of clustering, especially for our second cluster. I tried multiple clustering methods and multiple numbers of clusters, but K-means and 3 clusters worked the best in what I had tried.

I would like to have refined it further by grouping Japanese restaurants into one category, foreign restaurants into one category, and sightseeing attractions into one category. It might then further diversify our clusters and make them stand out more.

I would also like to extend this research to all of Japan. We have more data for Osaka and Kyoto than can easily be implemented in this research.

Thank you!

Notebook: https://github.com/mynameisjc/thinkful/tree/master/Final%20Capstone