Data Visualization Mini Project – Can I find a good hostel with lower price in Japan

Group 2 – Wong Shing Fung, Dov

Project Objective

This project aims to analyze whether a hostel with higher price have better qualities and vice versa. While considering various attributes, the primary focus will be on factors related to location. Furthermore, I want to determine is it difficult to find high-quality hostels at a lower price in Japan, which can help travelers to save money in their trips.

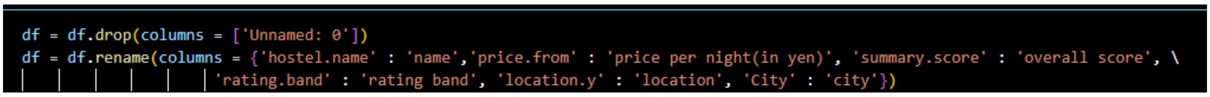
Dataset Description

The dataset is downloaded from Kaggle which extracted from Hostelworld until 2023. There are 342 entries with 16 columns. It includes hostel in five Japan cities (Tokyo, Osaka, Kyoto, Hiroshima, Fukuoka-City) with prices ranging from JPY 1000 to 7600.

In additional, the dataset included some geographic date such as city, longitude, latitude; various scores on a scale of 1 to 10 (Atmosphere, cleanliness, facilities, location etc.); distance from city centre and rating band.

Data Cleansing

I dropped some entries with “NA” columns, corrected the number in “price.from” (from 103200 to 3200), amended wrong location (amended “lan” and “lon” of one entries, and changed the “City” from Osaka to Takayama), converted “Distance” to numeric data (i.e. from “2.9km from city centre” to 2.9) and made some column to new name with more descriptive (e.g. from “price.from” to “price per night(in yen)”).



```
df = df.drop(columns = ['Unnamed: 0'])
df = df.rename(columns = {'hostel.name' : 'name','price.from' : 'price per night(in yen)', 'summary.score' : 'overall score', \
                          'rating.band' : 'rating band', 'location.y' : 'location', 'City' : 'city'})
```

Considering there is only one entry for Takayama and the analysis is city-based, this entry will be excluded due to its small sample size. Moreover, the sample sizes for Hiroshima and Fukuoka-City are relatively small compared to Tokyo, Osaka and Kyoto. Hence, the analysis will primarily focus on Tokyo, Osaka and Kyoto, while the data for Hiroshima and Fukuoka-City will serve as additional reference only.

Outlier analysis

Boxplot was utilized for outlier analysis. There are few outliers found. Given the small quantity of outliers and the limited data size, the decision was made not to remove these outliers.

Methodology

I opted for Tableau over PowerBI as it offers more appealing visualizations. In addition, Tableau is user-friendly and facilitates immediate interaction with users.

1. Selecting relevant visualization from Tableau

Type of Visualization	Description / Justifications
Map	It is more intuitive to show the distribution of location of hostels in map than textual description. It provides a clear geographical context and allows for easy comparison of hostel locations.
Bar Charts	It is effective for visualizing the proportions of different rating bands. Also, multiple bar charts per city can facilitate a comparative analysis of hostel quality across cities.
Boxplots	Boxplots are excellent for summarizing data distribution and identifying outliers, max, min, median, 1st quartile. 3rd quartile and range. They are useful for both comparing distribution and conducting outlier analysis.
Scatterplots (with regression line)	Scatterplots with regression line are ideal for correlation analysis between price and various variables. Multiple scatterplots can depict the correlation relationship between various variable in a single image, providing a comprehensive view of data.
Histograms	Histograms are used for visualizing the distribution and relationship between price (binned) and value or overall score. By binning the price, we might uncover new insights that might not be apparent in a scatterplot.

2. Applying Interactive Features from Tableau

Feature	Description/ Justification
Filters	Users can select specific or multiple cities in various graphs, providing the flexibility to tailor the view to their needs.
Tooltips	Tooltips offer additional information such as city and hostel names within the graphs, enhancing the user's understanding of the data.
Parameters	Parameters with two values (500 & 1000) are set for price bin which enable users to explore the data in more depth.
Dashboard navigator	Buttons are set in the dashboard which is user-friendly and effiecient in presentation.

Highlights of your work in THREE bullet points

1. The quality of most hostels in Japan is generally good

Fig. 1 shows that most Japan hostels have “Superb” (Score >= 9) or “Fabulous” (Score within 8 and 9) ratings.

Also, from fig. 2, which separates the data by city, shows that in all five cities, over 50% of hostels have a “Superb” rating. The ratios for the other three rating bands (Very Good, Good, and Rating) are all less than 25%.”

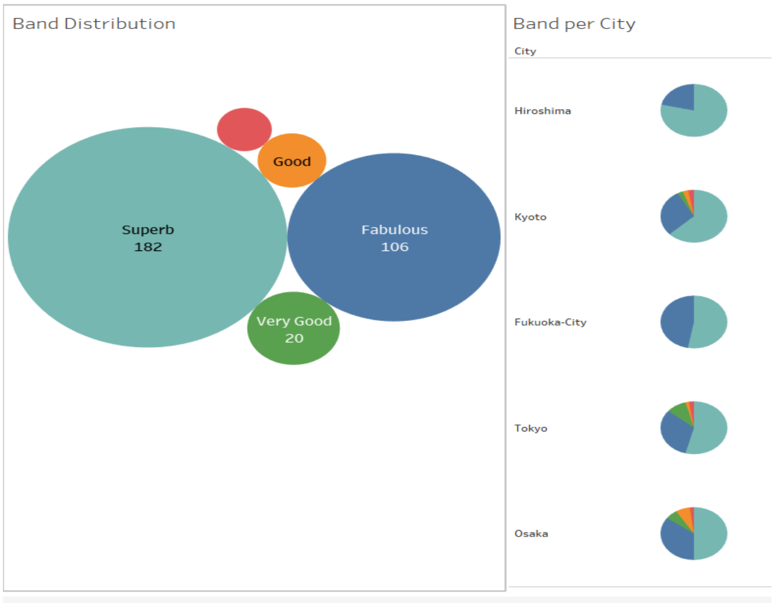


Fig. 1 (left), fig. 2 (middle)

2. The correlations between different attributes and prices are low

Fig. 3 shows that the slopes of regression lines between all attributes (including distance from city centre, all scores and price of hostel) are relatively low. Although the p-value of some correlations are lower than 0.05, indicating statistical significance, the effect sizes are very small.

For the most important factor, location or distance from city centre, there is no significant correlation relationship with price. This suggest that it is possible to find a relatively cheap hostel in a good location (e.g. near a rail station).

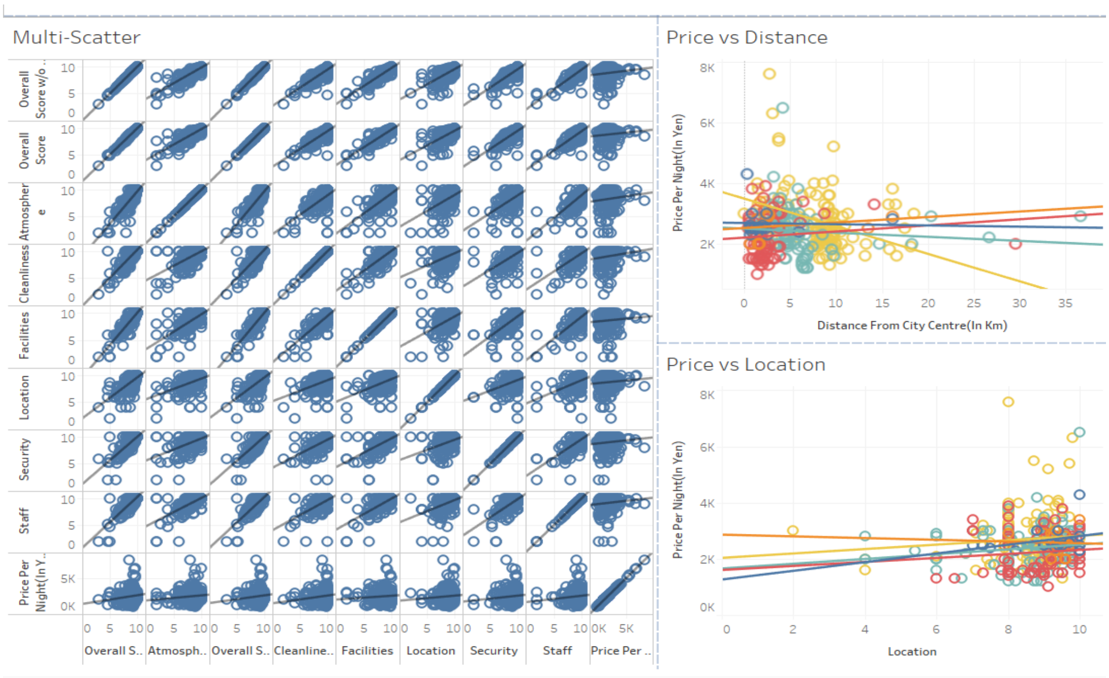


Fig. 3

3. Hostel in Kyoto showed lower price with better quality

Fig. 2 shows that, compared to Tokyo and Osaka, Kyoto has a higher ratio of hostels rated as "Superb" and "Fabulous" of Kyoto.

Fig. 4 shows that, although the median and 1st quartile of hostel prices in Kyoto are higher than those in Osaka, the range, max, 3rd quartile and min hostel prices in Kyoto is lower than both Osaka and Tokyo.

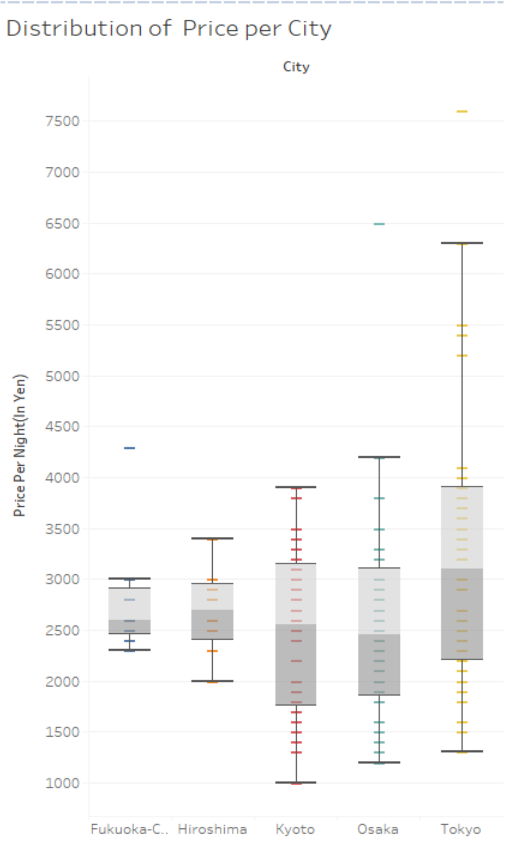


Fig. 4

Effectiveness of the visualizations

Attractive and functional graphics

Tableau provided various types of graphics, each with different functions. Map can provide an intuitive view for data distribution. Scatterplots with regression line can show the correlation between variables. By simply adding filter for cities, the scatterplots can show the correlation for each city individually or multiple cities, rather than for the entire dataset.

Rich information

Tooltips are a powerful tool for providing detailed information. For example, in the scatterplot between price and distance from city centre, I can just point to the dots to check out the name of hostel which providing good distance and cheaper prices.

Moreover, by hovering over the regression line, the R-squared value and p-value are displayed, which is very convenient.

Challenges and Limitations

1. **New to Tableau**

   - Tableau is a powerful data visualization tool. As I am not familiar with Tableau, the analysis could be presented in more decent and appropriate way.

2. **Small sample size**

   - This dataset only included hostel in five or six cities, which is not enough. More data with more cities could be analyzed in deeper way.

3. **Potential bias in score analysis**

   - The scores (e.g., facilities, staff, etc.) in the dataset are mean scores from all customers. However, the differences between each increment of one in the score may be not equal. For instance, improving the 'staff' score from 9 to 10 could be much harder than improving it from 5 to 6. Therefore, using the median of scores might provide a more accurate representation than mean scores.

Future Work

1. **Include more types of accommodation**

   - In the past, hostels might have been a cheaper choice for accommodation for travelers. However, there are many different choices such as Airbnb, which also provide an economical accommodation. It would be more useful if the analysis includes those data.

2. **Include more cities and variable such as "Month"**

   - The price of accommodations varies due to many factors. One of the most important factors is the time period. By integrating the price with different months and more cities, we can show the price of accommodation per month and city, which can help travelers choose the city with lower-priced accommodation in different months.

Due to the time constraint and the unavailability of a proper dataset, the proposed works proposed were not implemented in this report.

Conclusion

Firstly, the average quality of hostels in Japan is relatively high, with most of them having an overall score of 8 or above. Next, there is no or weak correlation between various attributes or quality with the prices. Travelers can choose a good hostel with reasonable price instead of focusing on expensive hostels. In addition, compared to Tokyo and Osaka, travelers can easily find a cheaper hostel with higher quality in Kyoto.

To conclude, Japan is good country for travelling. Travelers can save money on accommodation but still enjoy a good experience in hostel.

Reference

http://www.jyh.gr.jp/tcyh/eng/index.php

https://maps.app.goo.gl/dbGgD7PUmJJQihA96

https://maps.app.goo.gl/XwjLKTS2iKJyDaKWA

https://www.kaggle.com/datasets/koki25ando/hostel-world-dataset