RICE UNIVERSITY
Department of Statistics
605 Final Project: Recommendation of Airbnb Housing

December 6, 2019

Genyi Huang
Yawen Guo
Yuetong Yang
Yutong Su

Contents

Chapter 1

1.1 Introduction of Background and Method

Airbnb is a widely used online marketplace for housestays and tourism experiences. In this report, our main goal is to give housing recommendations for tourists. Besides, We also interested in how to become a superhost. To achieve our goals, first, we use Neural Network to fit a model help us classify superhost. Then, we choose five most attractive cities to visualize renting records using packages like ggplot2, ggmap, leaflet, grid to find if there are some relationships between variables. Finally, we choose whether host is superhost, longitude & latitude, amenities, price,cancellation policy and accommodates (6 variables) from the raw as helpful criterion. And we give weights to each variable according to its importance of renting a house. Next, we use these six criterion to give top ten houses for those five cities.

1.2 Description of Data

Our dataset includes three kind of information: Listings, Reviews and Calendar. Listings include the information of host and description variables of the houses. Reviews include feedback of guests. Calender includes booking information of each house. Table 1.1 shows some of the variables we used in our analysis. In the following parts, we will introduction those variables in detail. In addition, our dataset covers 101 cities, including some European cities, some North American cities, a few Australian and Asian cities. In each listings table, there are a total of 106 columns containing detailed information about various orders for a user. Each city has an average of 20,000 or 30,000 order information per day.

1.3 Motivation

1.3.1 Preliminary price analysis

1. We aggregate the total price of each city, since the price of data is obtained from already finished users' order and all data collected time is in 2019. Compute the total price of each hotels and group by cities. We can clearly see each city's airbnb revenue distribution.

2. Because of the limited space, we choose 5 most popular city to further analyze in details, they are: Amsterdam, New York, Paris, London and Barcelona.

1.3.2 Recommendation

1. Choose 6 helpful criteria and appropriately transform each criteria to numeric form. We can analyse each criteria and calculate the recommendation index to present a list of top 10 recommend airbnb hotels. Then for future tourists, they can easily choose their preferred hotels or home-stays. This part will make full considerations and save a lot of time for each tourist.

1.3.3 Killer Plot

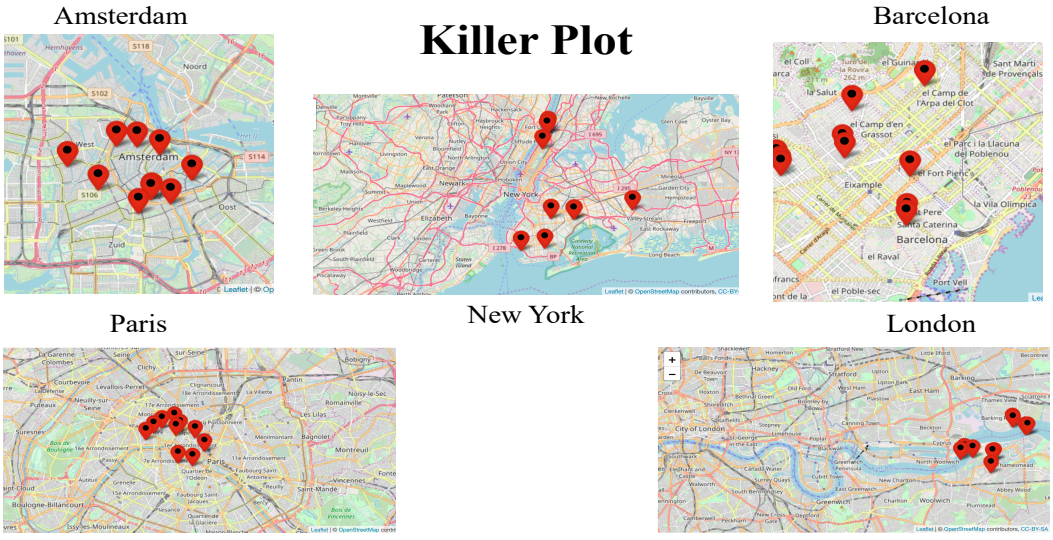

Figure 1.1: Killer Plot

Chapter 2

2.1 Methodology

2.1.1 Neural Network

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. Neural networks help us cluster and classify. You can think of them as a clustering and classification layer on top of the data you store and manage. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on. The following figure shows a neural network with two hidden layers.

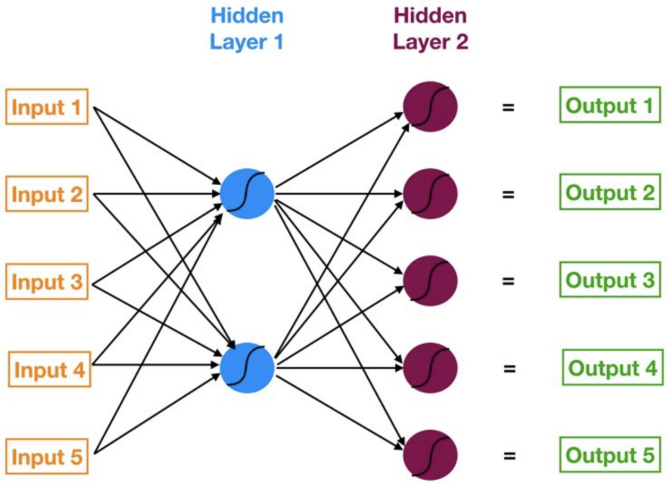

Figure 2.1: Neural network with two hidden layers

In our Neural Network, inputs are the variable which we are interested in and outputs

are is or isn't superhost.

2.1.2 Preliminary price analysis

- Making use of database, we write queries to aggregate by each city's name and compute all the home-stays' price as the total price. Since the price sum is very large, in the plot, we set the y-axis(price) unit as hundreds dollars.

- Since there are large amounts of data, we separate the whole data with location distribution into 6 subsets and save in our database. So we will obtain 6 plots each represent the price trend of airbnb home-stay in one area's cities. And use grid to gather 6 plots together.

2.1.3 Methods of Recommendation for each city

1. First we choose whether host is superhost, longitude & latitude, amenities, price, cancellation policy and accommodates (6 variables) from the raw data.

    - Superhost is a tag for best host in airbnb, with this criteria whether it is a superhost user can clearly know the quality of the hotel or the home-stay generally. We define the data that it is a superhost as 1 while it is not a superhost as 0.
    - Distance is the average straight line distance from each hotels to 5 each most popular attractions. Use distHaversine function to calculate two points' longitude and latitude and then we can obtain the straight line distance. Next, compute the average value of 5 attractions and transform the meters to miles. If the average distance is smaller than 1.5 miles, we set the value as 1 otherwise set the value as 0.
    - Amenities is the description of amenities each hotel will offer. We use library(stringr) and count each description's string length. If the length is longer, it represents the hotel offer more amenities and more considerate. Then transform the string length value into standard format for computing in future and better output.
    - Price is the daily price of each hotels. User of course will consider the price when they are choosing hotels. Transform the price value into standard format for computing in future and better output.

- Cancellation policy is different from hotel to hotel. Some hotel or home-stay is very flexible toward user's cancellation while others are very strict. Cities may have 5 or 6 different levels of cancellation policy. We cast the different levels into numbers and calculate the standard format of it for further handling.
- Accommodates is the maximum number that each hotel can contain. Averagely, the capacity is around 3. So we set accommodates lesser than 3 as 1, more than 3 as 0.

2. Second we need to give weights to each variable. Our given weight policy is : 0.10 for whether host is superhost 0.30 for distance 0.25 for amenities 0.20 for price 0.10 for cancellation policy 0.05 for accommodates With the weights value above, we can calculate the recommendation index for each hotel in each city. Also, we set equal weights to each variable and compute the recommendation index with equal weights.

3. Third we sort the recommendation index and present the top 10 hotels. Besides, we choose 5 popular tourism city to carry out above methods.

2.2 Results

2.2.1 How to Become a Superhost?

Superhosts are experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests. Once a host reaches Superhost status, a badge will automatically appear on their listing and profile to help guests identify them. Superhosts may attract more guests and make more profits. In Airbnb official website, they give four requirements to become a superhost:

- Completed at least 10 trips OR completed 3 reservations that total at least 100 nights.
- Maintained a 90% response rate or higher.
- Maintained a 1% percent cancellation rate (1 cancellation per 100 reservations) or lower.
- Maintained a 4.8 overall rating (this rating looks at the past 365 days of reviews, based on the date the guest left a review, not the date the guest checked out).

However, these information are not included in our dataset expect response rate. So we choose some of the variables related to host to find how to become a superhost. From the requirements, we find that three requirements except response rate are all related to the customers' feedback. Therefore, we choose "host response time", "description" and "host response rate" as our predictors and use neural network to fit a model to see how do these predictors affect our response-"superhost".

We find it is difficult for us to find a model that both predict is or isn't superhost. To solve this problem, we fitted two models, one has high accuracy rate which is 0.935001 to predict is superhost and we denote this model as Model One, and other has high accuracy rate which is 0.9610487 to predict isn't superhost and we denote this model as Model Two. Model One has one hidden layer and has four units in the layer. Model Two also has one hidden layer but five units. Figure 2.2 is flow chart to decide if a host is superhost.

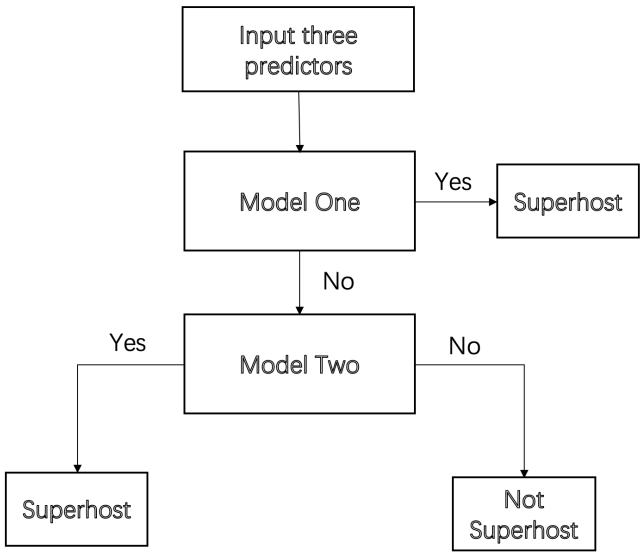

Figure 2.2: Flow Chat of Superhost

2.2.2 Map Plots

We categorized variable "price" into four levels :

- Not expensive: Price <= 300
- Moderate: 300 < Price <= 500
- Expensive: 500 < Price <= 1000
- Deluxy: Price >1000

Categorized variable "Accommodates" into three levels:

- Small: Accommodates <= 2
- Median: 2 < Accommodates <= 4
- Large: Accommodates > 4

1. Amsterdam
   Randomly subset Paris dataset with size 50 and plot it on the map, we can see there is no "Not Expensive" house show up on the map, houses with "Moderate", "Expensive", and "Deluxy" spread out on the map. Apparently, it has most red points on the map, which are houses with price level labeled as "Deluxy".

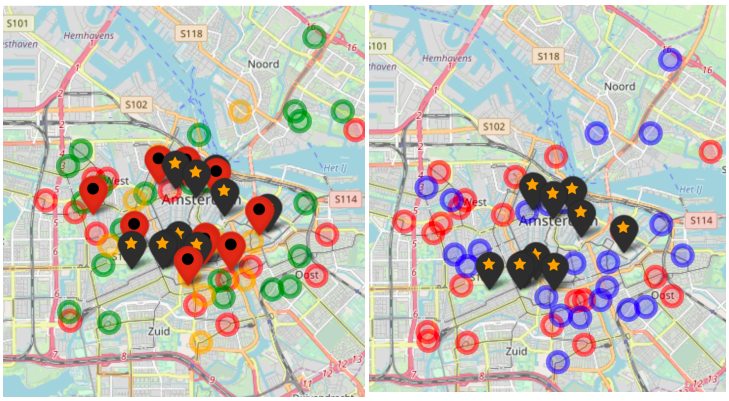

Figure 2.3: Price Map of Amsterdam

2. New York City
   Although New York City is comprised of 5 boroughs, most of houses located at

central Manhattan area. From the map plot. We can see all top ten iconic sites locate at Manhattan area, however, according to our recommendation criterion, houses in other boroughs such as Brooklyn are also recommended. Those houses stand out because their cheaper rents and dynamic neighbourhoods.

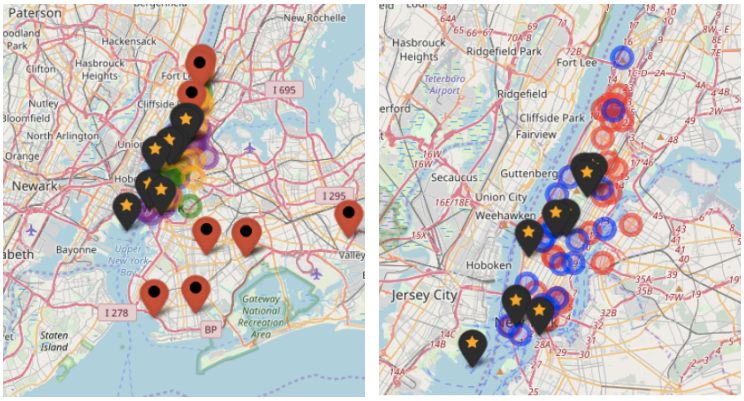

Figure 2.4: Price Map of NYC

3. Paris

   Houses at all price level in Paris spread out on the map, as well as the tourist attractions in Paris. Recommended houses clustered around tourist attractions. Most of houses have price level at "Moderate" and "Expensive"

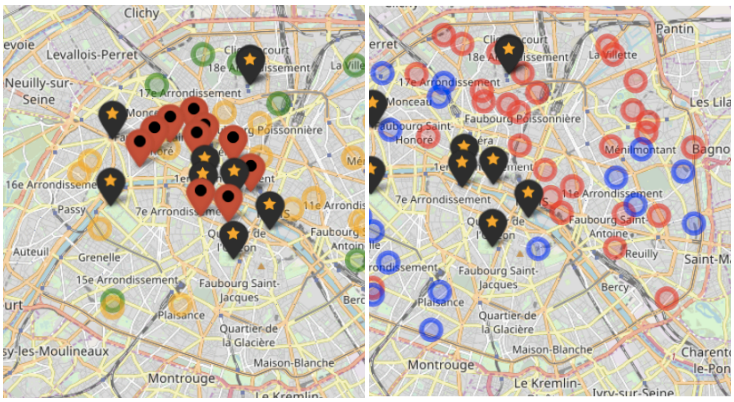

Figure 2.5: Price Map of Paris

4. Barcelona

   According to the plot, recommended houses and tourist attractions gathered together in the center of Barcelona, while the houses spread out in the city, which

can meet different needs for location. Since there are no "not expensive" house in Barcelona, we suppose that the expense of Barcelona is relatively higher than other four cities.

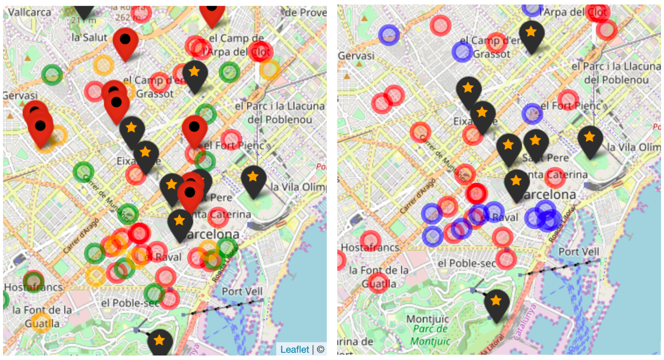

Figure 2.6: Price Map of Barcelona

5. London

   Surprisingly, all of our recommendations and houses jump out of the cluster of tourist attractions, considering the high living expense, it is not hard to understand this results. However, there are no" deluxy" houses in London, most of the houses are "expensive". According to the Size Map, most of people who travel to London prefer to go with group.

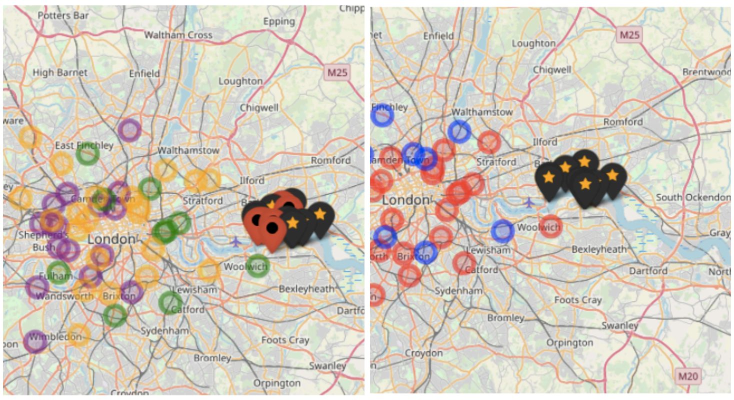

Figure 2.7: Size Map of London

2.2.3 Preliminary price analysis

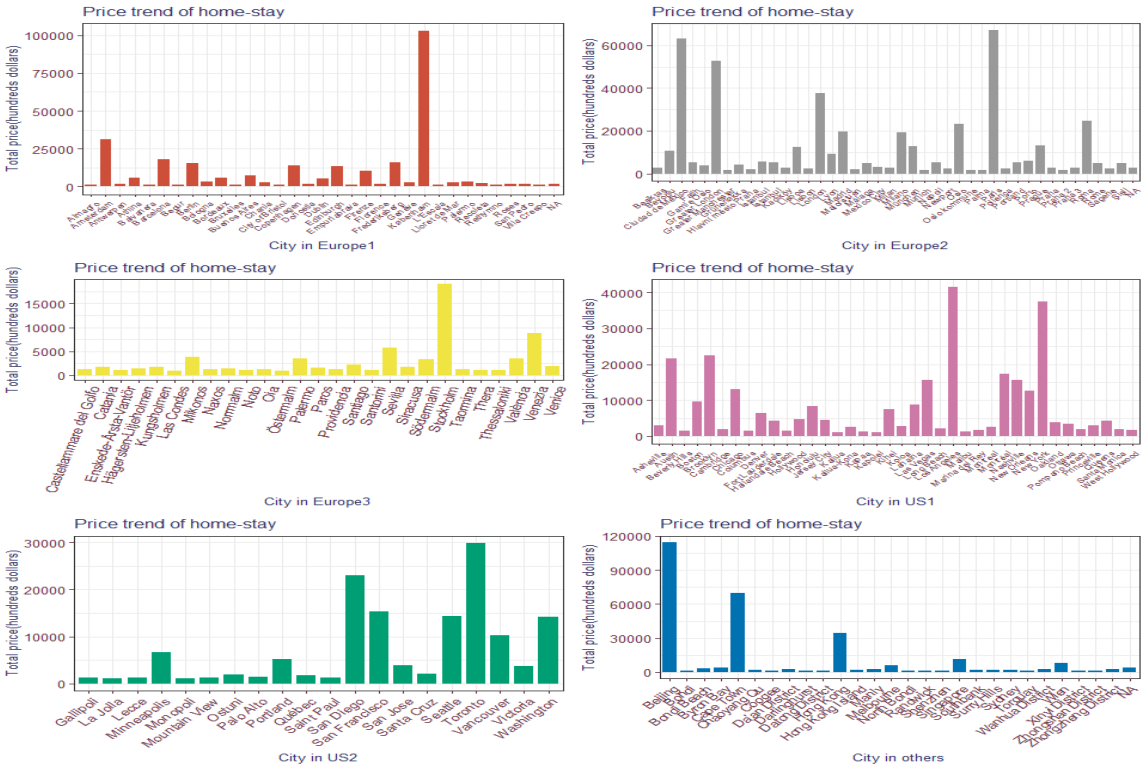

Figure 2.8: Price trend of home-stays

2.2.4 Recommendation of Top 10 house for Five Cities

1. For Amsterdam City

   - Figure 2.9 shows the results of given weight top 10 in Amsterdam, the lowest given weight is 0.88(Penthouse Studio with Roof terrace and city view) and the highest given weight is 0.91(Trendy apartment in Oud-West).

2. For Paris City

   - Figure 2.9 shows the results of given weight top 10 in Paris, the lowest given weight is 0.84(50 m^2 in the heart of the capital SPA) and the highest given weight is 0.89(Avenue Montaigne Apartment in the Luxury Heart of Paris).

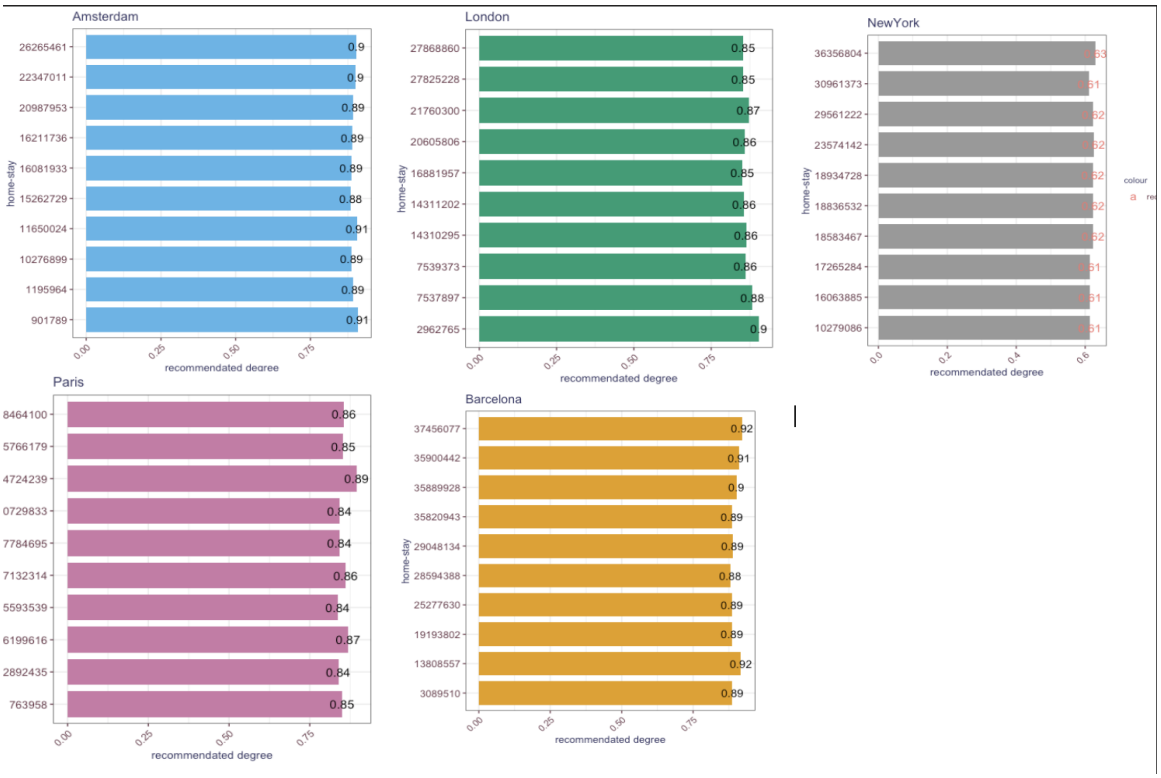

Figure 2.9: Recommendation Results

3. For London City

   - Figure 2.9 shows the results of given weight top 10 in London, the lowest given weight is 0.85(Great apartment, central location, stunning views) and the highest given weight is 0.90(Explore Borough Markets from a Room in an Award Winning Loft).

4. For New York City

   - Figure 2.9 shows the results of given weight top 10 in New York, the lowest given weight is 0.61(Master Bedroom in Spacious Washington Heights Apt.) and the highest given weight is 0.63(Bright quiet and Comfortable).

5. For Barcelona City

   - Figure 2.9 shows the results of given weight top 10 in Barcelona, the lowest given weight is 0.88(Estudio en Gracia /Studio flat) and the highest given weight is 0.92(Business Apartment Diagonal Francesc Macia).

2.3 Conclusions

- Based our "good houses" criteria, we highly recommend Airbnb users to travel in Paris, Barcelona and Amsterdam among five top travel cities, where our recommended houses surround around tourist attractions. Reasonable price with great lively and dynamic neighborhoods would provide a better traveling experience
- In terms of traveling expense, New York and Barcelona would be the top two recommendation.
- From the recommendation index we calculated, all index of home-stay in New York city is relatively low. Since all home-stays are kind of far from the most popular attraction and the price in New York is very high. Compare with other cities, New York is not quite much recommended. In Paris, London, Amsterdam and Barcelona, our recommended home-stays are all close to the attractions.
- Almost all the home-stays we recommend are super-host except for one hotel in Paris.