Statistical Analysis Plan of Airbnb Data

Ran Tao

Nov 2017

1. Project Background

Airbnb is getting popular as a substitute of hotel. It's an online platform where people can lease or rent short-term lodging. Nowadays, more and more travelers are choosing Airbnb over hotel because of its low cost, convenient location and household amenities especially in Europe. Therefore, I am motivated to study the ratings of Airbnb listing properties, to find out the factors that may influence the ratings. I choose Paris, one of the most popular touristic city in Europe as my study city. It's also interesting to study the ratings of Airbnb listing properties within each neighborhood and cross neighborhood in Paris.

2. Research Objective

The main research objective of this project is to study the ratings of Airbnb listing properties within each neighborhood and cross neighborhood in Paris and the factors that may influence the ratings. The potential implication of this study could provide suggestion for Airbnb leasers on how to improve their ratings as well as for travelers on how to choose the good Airbnb properties when they are visiting Paris.

3. Approach

3.1 Description of data

The data used in this project is a single survey for Paris with 70,158 listing properties as of 25th July 2017, which is collected from the public Airbnb website. The data include the following information:

1) room_id: A unique number identifying an Airbnb listing.
2) host_id: A unique number identifying an Airbnb host.
3) room_type: One of "Entire home/apt", "Private room", or "Shared room"
4) borough: A sub-region of the city or search area for which the survey is carried out. For some cities, there is no borough information, which is the case of Paris.
5) neighborhood: As with borough: a sub-region of the city or search area for which the survey is carried out. A neighborhood is smaller than a borough.
6) reviews: The number of reviews that a listing has received. Airbnb has said that 70% of visits end up with a review, so the number of reviews can be used to estimate the number of visits. Note that such an estimate will not be reliable for an individual listing (especially as reviews occasionally vanish from the site), but over a city as a whole it should be a useful metric of traffic.
7) overall_satisfaction: The average rating (out of five) that the listing has received from those visitors who left a review.
8) accommodates: The number of guests a listing can accommodate.
9) bedrooms: The number of bedrooms a listing offers.

10) price: The price (in $US) for a night stay. In early surveys, there may be some values that were recorded by month.
11) minstay: The minimum stay for a visit, as posted by the host.
12) latitude and longitude: The latitude and longitude of the listing as posted on the Airbnb
13) last_modified: the date and time that the values were read from the Airbnb

3.2 Data limitations

The data only contained the listing properties as of 25th July, which couldn't display a full picture of the ratings in terms of trend over time.

3.3 Study population

The 51,055 listing properties with at least one review as of 25th July 2017 in Paris. As the properties with no review don't have any rating, we need to remove them from our study population.

3.4 List of attributions

The dependent variable will be the weighted overall satisfaction, which is equal to the multiple of the significance level of number of reviews and average rating (out of five) that the listing has received from those visitors who left a review. As a rating of 5 with only 1 review is difference from a rating of 5 with 100 reviews. We need a significance level corresponding to difference range of review numbers. And then multiple with the average rating to get a weighted overall satisfaction.
The independent variables are listed as below:

Room ID	Unique number identifying room
Host ID	Unique number identifying host
Room Type	Entire home/apt, Private room, Shared room
Neighborhood	80 different neighborhoods
Accommodates	Integer from 1 to 16
Bedrooms	Integer from 0 to 10
Price	From 10 to 11323 USD per night
Property type	24 different types

3.5 Description of statistical methodology

The multilevel regression model will be applied to study the distribution of ratings within neighborhood and cross neighborhood, as the dataset has hierarchical structure with two or more different dimensions of sampling units. So we should have different neighborhoods with different characteristics that impact the ratings differently. Multilevel regression model allows examination of between

neighborhood variation, within neighborhood variation and their interactions simultaneously.

3.6 Output interpretation

After fitting the model, we will assess the model fit by conducting residual and deviance analysis, check outliers, and measure parameter significance. Based on this, we could modify and improve our model.