Proposal of Statistical Analysis Project

Shiyu Zhang

Nov, 2018

1. Project Background

I have been using Airbnb for over three years and it has become a popular way of travelling. I have witness Airbnb develop from an unknown website to the most popular travelling website during the past several years. Many people choose Airbnb instead of hotels not only for its lower price and convenient location, but also for its humanness – travelers are able to make connections with people from all around the world. What's more, travelers are provided with more unique options compare to hotels - houses, condos, apartments, castles, houseboats, tree houses, barns, mansions, even caves! Therefore, these unique properties of Airbnb inspired me to explore more about it. For example, what the factors may have an impact on the ratings, or, what is the relationship between the occupancy rate and the neighborhood of an Airbnb apartment, etc.

2. Goal of the Project

The main objective of this project is to study the important factors that may have a significant impact on the ratings of Airbnb listing properties. The potential implication of this study is to provide suggestion for existing and potential properties owners to have a better understanding of ratings, as well as for travelers to choose a property that best fit their need.

3. Description of the dataset

Personally I prefer to choose Boston data from Airbnb dataset website (http://tomslee.net/airbnb-data-collection-get-the-data). The data used in this project is a single survey for Boston with 3,864 observations with 14 variables (from 11/21/2016 dataset).

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 such as Boston, there is no borough information.

(5) neighborhood: 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. The number of reviews can be used to estimate the number of visits. However, such estimation may not be reliable for an individual listing (especially as reviews occasionally vanish from the site).

(7) overall_satisfaction: The average rating that the owner of the property has received.

(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 Airbnb web.

(13) last_modified: the date and time that the values were read from the Airbnb.

4. Limitations of the dataset

I chose data only contained the listing properties of Nov 2016 , which couldn't represent the full picture of the ratings in terms of trend over time. What's more, the reason I chose the data from 2016 instead of the most updated one (from July 2017) is that the newest data doesn't have the "minimum stay" data, I think minimum stay is important for evaluating the rating for properties.

5. Statistical Methodology for the Project

I will apply EDA, multilevel regression model to study the distribution of ratings within neighborhood and cross neighborhood, and explore other factors that may have significant impact on the ratings. Multilevel regression model allows examination of between neighborhood variation, within neighborhood variation and their interactions simultaneously.

After fitting the model, I will assess the model fit by conducting residual and deviance analysis, check outliers, and measure parameter significance.