Project Proposal COGS 109

On the relativity of Airbnb star ratings

Christopher Jensen (cajensen@ucsd.edu)
Lana Andreasyan (landreas@ucsd.edu)
Amine M'Charrak (amine.mcharrak@tum.de)

Abstract

In this paper we aim to uncover counterintuitive trends and associations in the star ratings of Airbnb from New York, Copenhagen, and Sri Lanka to understand how people in different parts of the world make decisions. We want to understand if the requirements for a five star rating are the same or alternatively, how do they vary based on geographic locations? What are the main requirements affecting the rating? We are interested in finding a general and simplified model that will possibly predict similar datasets independent from locations.

First we will get familiar with the data set by playing around with various fits. Then we will decide which variables we want to consider in our final analysis. We will also want to consider what assumptions are we will need to make and what our hypothesis will be. Next we will do the analysis. Finally, we will conclude by summarizing our findings and reflecting on our hypotheses.

The two central hypotheses in the paper are that geographic locations affect the requirements people rely on for ratings. The duration of the stay positively affects the rating.

1 The Team

We have two Math/CS majors and one computer and information technology major in our team. We hope to profit from each others knowledge. The work breakdown will be as follows. Together we will discuss different approaches and decide what statistical techniques we are going to settle with and then each team member will execute the analysis on a specific city. In the end we will have analyzed three different geographical/social regions. Eventually, we will compare and combine our results in order to make reasonable conclusions for our hypotheses.

2 Background

We all agreed, that we are heavily influenced by rating systems and often let ratings take over huge parts of our final decision. It goes so far that we even exclude possible candidates (accommodations) by simply relying on single scores (average rating value).

3 Data Source

We are going to choose the datasets for each city from the following link below:
`http://tomslee.net/airbnb-data-collection-get-the-data`
There we can choose between different cities and download the corresponding csv file. Each file contains parsed features of the corresponding Airbnb site for that city. An overview of predictors the dataset contains is listed in the section below.

4 Predictors to consider

- **reviews:** The number of reviews that a listing has received.
- **room_type:** One of "Entire home/apt", "Private room", or "Shared room"
- **borough:** A sub region of the city or search area for which the survey is carried out.

- **neighborhood:** As with borough: a sub region of the city or search area for which the survey is carried out. For cities that have both, a neighborhood is smaller than a borough.
- **overall_satisfaction:** The average rating (out of five) that the listing has received from those visitors who left a review.
- **accommodates:** The number of guests a listing can accommodate.
- **bedrooms:** The number of bedrooms a listing offers.
- **price:** The price (in US) for a night stay.
- **minstay:** The minimum stay for a visit, as posted by the host.

5 Data Analysis and Modeling Methods

We will most likely be fitting a general linear model to our data, but this might change after our exploratory phase. For model assessment we are considering either doing cross validation or test our model onto a totally different dataset which might be a similar city. For visualization purposes we are also considering to apply principle component analysis (PCA) onto our model in order to get a better overview of our high-dimensional dataset.

6 Expected Results

We would be really happy if we are able to uncover counterintuitive trends and associations in our dataset and perhaps make out interesting relationships which disprove our common sense of how we make decisions based on rating systems. All in all, we want to gain a better understanding of the real value of this five star based rating hierarchy. We want to repeat the analysis on a time shifted dataset and see if our conclusion still holds. If this is the case, then our analysis was successful and the conclusions are reliable.