To review or not to review? That's the question

Team members: Doria Wengting Wang, Inge Oostveen, Ralph Delsing, Paulus Hovens, Wouter Floors

---

Introduction

We all know the feeling of opening an e-mail that asks you to leave a review on a product that you just bought or a listing that you stayed the night at. But do more reviews lead to a higher review score? Furthermore, this effect is examined at cities in Spain that differ in sizes, such as Malaga compared to the capital Madrid. This code is written in order to find answers to these questions.

Motivation

The reason for diving into this problem is because review scores are an important factor for customers when they decide which listing to stay at. Hosts want a good review score and some push their guests to leave a review. The issue that arises is that more reviews does not necessarily result into a higher review score. This study is conducted in order to show hosts if they should or should not push their guests for reviews by analyzing the relationship between review score and number of reviews, while taking moderators and control variables into account.

The study focuses on seven different cities in Spain. To be able to examine the differences between small and big cities, there are three categories; Big, Medium or Small. The number of residents determine the category of the city:

A Big city has more than 1,000,000 residents –> Barcelona and Madrid

A Medium city has more than 500,000 residents –> Sevilla, Valencia and Malaga

A Small city has less than 500,000 residents –> Menorca and Girona

Data

Data Collection

The data is collected from AirBNB in order to analyze if the number of reviews has an effect on the review score. The data contains all listings in a specific city with data about the listing, the host, review scores and much more. The data is downloaded from the web and stored in a Google Drive in order to be accessed at any point in time (see src/download_data.R).

Data Processing

In order to analyze the data, a new dataset was created by combining the separate cities' listings. Missing values were deleted after being inspected. On average, 26% of each cities listings contained missing values on our variables. These were deleted because the dataset still has enough observations, and because the missing values could not be used for an analysis. A city ID column is added to keep track of which city a listing belongs to. Irrelevant columns have been removed, while dummy variables for the moderators have been added.

Methodology

The relationship between the number of reviews and actual review score may be influenced by other variables, which are integrated in the model as moderators. In this study, three moderators are considered which are city size, room type and superhost status.

1. City size: Each of the seven cities in Spain falls in either one of three categories; Big, Medium or Small, depending on the number of residents as was mentioned before. The base level of this variable is the "big" size.
2. Room type: There are four categories in the dataset; Entire home/apartment, Hotel room, Private room or Shared room. In the dummy encoding, shared room is the base level.
3. Superhost status. The dataset includes information on whether the host of the listing is a superhost or not. Experienced hosts can become a superhost upon fulfilling certain criteria, such as having a low cancellation rate, a high response rate as well as a high overall rating based on the last 365 days.

Results

First, an overall descriptives table was made for both the total dataset and the dataset grouped by size, showing variables such as the mean rating, mean number of reviews, ratio of superhosts etc. This table looks as follows:

city_size	mean_rating	median_number_reviews	number_private_rooms	number_hotel_rooms	number_entire_rooms	number_shared_rooms	number_of_superhosts	superhost_ratio
Big	4.524	45	10237	283	15106	278	5746	25904	0.222
Medium	4.620	50	2546	97	10464	33	4131	13140	0.314
Small	4.539	16	879	70	13396	18	2458	14363	0.171
Total	4.552	38	13662	450	38966	329	12335	53407	0.231

Moreover, three individual tables were made that show the average rating and number of reviews for the cleaned dataset, grouped by the three moderators.

host_is_superhost	mean_number_reviews	mean_rating
0	30	4.47
1	67	4.82

room_type	mean_number_reviews	mean_rating
Entire home/apt	39	4.57
Private room	37	4.51
Hotel room	24	4.57
Shared room	22	4.35

city_size	mean_number_reviews	mean_rating
Big	45	4.52
Medium	50	4.62
Small	16	4.54

The first table shows that the number of reviews for listings that belong to a superhost receive more than double the number of reviews (67 vs 30) and receive quite higher ratings (4.82 vs 4.47). the second table

shows that room types “Entire home” and “private room” get almost double the number of reviews compared to “shared room” and “hotel room”. No significant difference in terms of ratings can be seen between these room types, although shared rooms stands out a little bit in last place. The first table that was created including the entire dataset did show that there way more entire homes and private rooms than hotel rooms and shared rooms. The reason for these latter 2 being less common could also explain why these receive less reviews. Finally, the last table shows that listings in large and medium sized cities receive more than double the amount of reviews than listings that are established in smaller cities. In terms of the overall rating, the three sizes score about the same.

Next, a plot of the cleaned dataset was made to visually inspect the data. As the number of reviews goes up, it seems that the rating goes up. The plot would suggest some sort of exponential function.

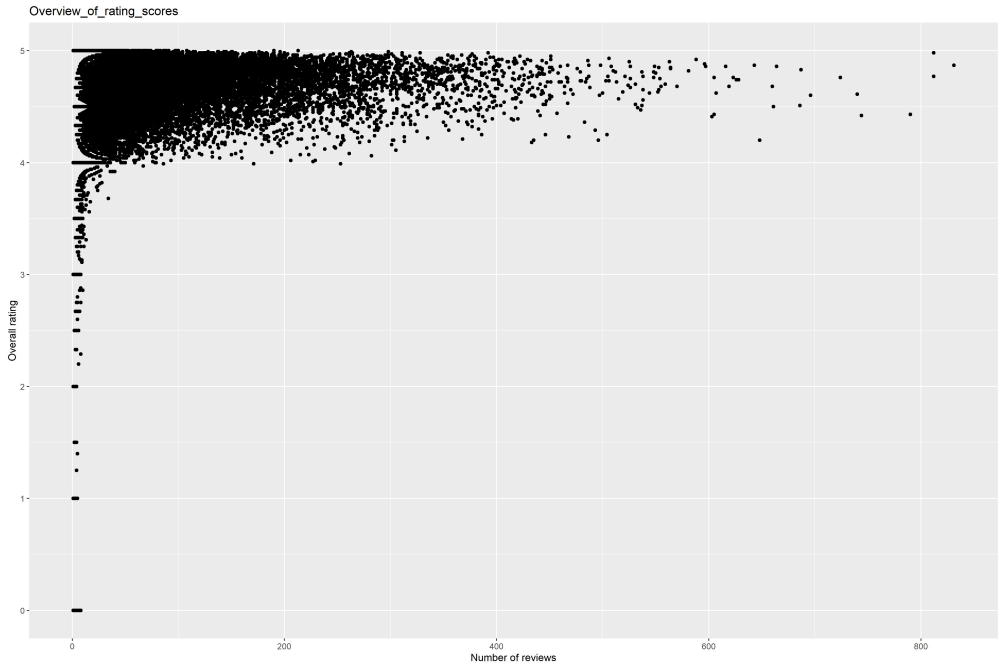

Figure 1: Rating scores

Finally, a regression model is created, shown below:

Regression model

	Model 1
(Intercept)	4.2724*** | (0.0424)
host_is_superhost	0.3871*** | (0.0091)
number_of_reviews	0.0015+ | (0.0008)
room_type_private	0.1024* | (0.0429)
room_type_hotel	0.2154*** | (0.0564)
room_type_entire	0.1423*** | (0.0427)
city_size_small	0.0372*** | (0.0090)
city_size_medium	0.0717*** | (0.0092)
host_is_superhost × number_of_reviews	-0.0013*** | (0.0001)
number_of_reviews × room_type_private	0.0001 | (0.0008)
number_of_reviews × room_type_hotel	-0.0010 | (0.0011)
number_of_reviews × room_type_entire	-0.0002 | (0.0008)
number_of_reviews × city_size_small	0.0007*** | (0.0002)
number_of_reviews × city_size_medium	-0.0003** | (0.0001)
Num.Obs.	53407
R2	0.054
R2 Adj.	0.054

From the regression model, the individual outputs for the moderators are interpreted:

Moderator 1: The predicted rating including only the first moderator, host_is_superhost becomes: rating = 4.272 + .387 x host_is_superhost + .0015 x nr_of_reviews - .0013 x host_is_superhost x nr_of_reviews. If the host of a listing is a superhost, this variable is coded as a "1". Thus, the rating for a superhost is 4.272 + .387 + nr_of_reviews x (.0015 - .0013) = 4.659 + .0002 * nr_of_reviews, whereas the rating of a listing from someone who is not a superhost is 4.272 + .0015 x nr_of_reviews. Despite the interaction effect being negative (-.0013), the predicted ratings for superhosts are clearly higher. Moreover, this effect is highly significant ($p<.01$)

Moderator 2: For the second moderator, room_type, the results are statistically insignificant ($p=.95$, $p=.34$, $p=.83$). Thus, no further work is needed on this moderator, as this moderator does not influence the

relationship between the rating and the number of reviews.

Moderator 3: Finally, the last moderator includes the city size. As for moderator 1, these interaction effects are highly significant ($p<.01$). The predicted rating including this moderator is as follows: 4.272 + .0015 x nr_of_reviews + .037 x size_small + .072 x size_medium + .0007 x nr_of_reviews x size_small - .0003 x nr_of_reviews * size_medium. From the output based on the moderators implemented as independent variables, it can be seen that they are significantly larger than for the base level (which is size_large). Thus, these small and medium sized cities are rated significantly higher than larger cities. However, looking at the interaction effects for these two columns, despite being statistically significant, are so small (.0007 and -.0003) that they only influence the relationship between the rating and number of reviews minimally.

Conclusion

In conclusion, the moderators all have different impacts on the research question. Having the status superhost gives a higher predicted rating and it is shown that listings that have this status have more reviews. Although the room type does matter for the number of reviews a listing gets (a listing with a "shared room" of "hotel room" has almost half of the reviews an "entire home" or "private room" has), the effect is insignificant and does not influence the relationship between the rating and the number of reviews. Finally, the number of reviews is different per city size. Mostly, the smaller cities get less reviews. However, in the model it is shown that small and medium sized cities are rated significantly higher than larger cities. Despite the significant difference, the effect is small and influences the relationship between the rating and the number of reviews minimally. To conclude, the moderators can impact the relationship between the rating and the number of reviews but is for the most part significantly small.