Kshitij Gurung, Rayan Sadeldin, Maggie Upjohn

Airbnb Reviews

12/16/2019

Introduction:

Airbnb is an online marketplace that provides accommodations by acting as an intermediary between hosts who want to rent out their homes and the customers who want to stay in that locale. Airbnb spans over 81,000 cities across 191 countries, and there has been over 400 million users since 2008 (Jaleesa, 2019). Among many variables, positive ratings are considered to be critical to entrepreneurial and platform success as they play a prominent role in ranking and user selection (Zervas, et. al 2015). Therefore, for online entrepreneur platforms such as Airbnb, ratings are influential factors for user selection. Moreover, previous studies have shown that a review, as a quality metric, does influence the consumer's decision when it comes to booking a property. A study published in 2018 showed that the price of an Airbnb room in Boston depends not only on the characteristics of the room itself, but also on the quality attributes of the property such as host interactions (Lawani et al., 2018). A continuously reported disadvantage among Airbnb users is that often the way in which a property is advertised does not accurately represent the property and its characteristics (Folger, 2019). As a result of the lack of assurance when booking a property, the number of reviews and their content can greatly influence the demand for the property and number of its potential customers.

In light of the aforementioned findings, in this study we are interested in exploring the factors that may compel visitors to place a review following a stay in an Airbnb in the Twin Cities, as well as the quality of the provided reviews. We aim to estimate the number of reviews over the past 12 months using the Zero-Inflated Poisson model. Additionally, we aim to predict the quality of provided reviews for listings by exploring potential determinants using Multilevel models. We consider multiple property characteristics such as price, room type, and cleaning fees, as well as various host related characteristics including the host response rate, listing counts, and if the host is a superhost. We investigate the hypothesis that visitors are more inclined to review a property when they pay higher prices, provide a higher review rating, or when they have the entire apartment or house to themselves. Additionally, we examined the hypothesis that visitors provide quality reviews when the price of the property is more affordable (lower), and they have positive interactions with the host.

Methods and Materials

Dataset Preparation

We gathered our data from Inside Airbnb, an independent, non-commercial set of tools and data, personally funded by Murray Cox. It utilizes public information scraped from the Airbnb website. The data is verified, cleansed, analyzed, and aggregated. For this project, we chose the Twin Cities Airbnb listings because it's the most relevant to traveling St. Olaf students. Overall, the dataset included 106 variables. Cleaning the dataset included parsing all prices and percentages related variables (price, security deposit fee, cleaning fee, extra people fees, and host response rate) from character to numeric variables. The total number of properties listed was higher than the total number of hosts. This suggests that there is more than one property per host. To confirm that there are multiple properties per hosts, we grouped the properties by host ID, and found that the number of properties per host ranges from 1 to 60, with the majority of hosts having only 1 property.

For ZIP exploratory data analysis, we examined the relationship between number of reviews over the past twelve months and review score rating, room type, property price, and host response rate. For multilevel analysis, we filtered out hosts with less than two properties. We then created a variable that reflects the quality of review by calculating the average review score using the six review categories that Airbnb uses (accuracy, cleanliness, check-in, communication, location, and value). We categorized the average review score into a binary variable. Review ratings for listing that received 10 points were categorized as "High", and the ones that received less than 10 points were categorized as "Low". We examined the relationships between binary rating and level 1 and level 2 variables. For level one variables, we explored different property characteristics, such as property prices, cleaning fees, location accuracy, and room types. For level two variables, we explored host characteristics, like host listing count, host response rate, and superhost status.

Model Building

This project uses two modeling approaches to explore the factors that compel customers in providing reviews, and the quality of reviews they give, following their stay in an Airbnb facility. The first approach addresses the likelihood of providing a review by modeling the total number of reviews for a listing over the past 12 months by fitting a Zero-Inflated Poisson model (ZIP). The count model coefficients predict the factors that influence the amount of reviews a listing would receive following a stay; then true zeroes are for dormant listings. The zero-inflated model coefficients would help distinguish possibly why a listing would receive reviews or not. We started by selecting the variables for count model coefficients (Poisson with log link) and the zero-inflated model coefficients (binomial with log link). In our models we used price as a predictor to see if a listing is more likely to see an increase in the number of reviews if price increases. This is stemming from the question if a person is more inclined to leave a review if they spent a lot of money on the accommodation. We also included room type as a predictor to answer if a person is more inclined to leave a review if they had to share their accommodation or have their accommodation all to themselves. Finally, we selected the host_response_rate for the zero-inflation coefficient because it would distinguish whether a listing could receive a review depended on how responsive a host is. In summary, the final model included reviews_scores_rating, finalprice, and room_type as count model coefficients and the host response rate for the zero-inflation coefficient as shown below:

$$log(\lambda) = \beta_0 + \beta_1(room\_type) + \beta_2(reviews\_score\_rating) + \beta_3(final\_price)$$

$$logit(\alpha) = \beta_0 + \beta_1(host\_response\_rate)$$

The second modeling approach focuses on the quality of reviews by fitting multi-level (glmer) models with a logistic response (bin_rating): whether a listing received a high rating (10 points) or low rating (less than 10 points). We chose multilevel modelling approach to account for the potential lack of independence between properties that have the same host. The multilevel model includes variables for hosts with two or more properties. Level one of the model includes property characteristics, and level two represent host characteristics. The models were initially unidentifiable due to the large eigenvalues for glmer function when running quantitative predictors such as price of the property, cleaning fees, and host listing counts. Therefore, we standardized all of numerical predictors into a z-score.

In the first step of building the multilevel models, we fit an unconditional multilevel model with binary rating score as response. We then incorporated z-score for the property price at level 1 and super host status at level 2. Following that we added the z-score for cleaning fees (level 1) in the model, and tested for significance by considering the coefficients p-values. Moving on, we replaced cleaning fees with is_location_exact at level 1 in the model and added an interaction terms between is_location_exact and super host status. Originally, there were four room types (hotel room, private room, shared room, and entire home/apartment) but we decided to remove shared and hotel rooms from the data set because they were only 0.9% and 1.6%, respectively, of the entire room type observation. We added host listing count to the model as a level 2 covariate. We tested for the significance of interaction between different variables, but all the interactions were statistically insignificance. We also examined models with different random effects to determine our final model. Using ANOVA Chisq test we concluded that the simpler model with (1|host_id) was better (p-value : 0.705, Chisq: 2.96) than the full model with many random effects, as the smaller model yielded lower AIC value (1898.5) and more significant fixed effects. Therefore, our final model includes z_price, is_location_exact and room type at level 1, and super_host status and listing counts at level 2 as shown below:

Level 1: $Y\ ij = a_i + b_i z\text{-}Price_{ij} + c_i is\_location\_exact_{ij} + d_i room\_type_{ij} + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim N\ (0,\sigma^2)$

Level 2: $a_i = \alpha_0 + \alpha_1 super\_host_i + \alpha_2 listing\_count_i + u_i$ where $u_i \sim N\ (0,\sigma_u^2)$

$$b_i = \beta_0$$

$$c_i = \gamma_0$$

$$d_i = \delta_0$$

Composite Model:

$$Yij = \alpha_0 + \alpha_1 super\_host_i + \alpha_2 listing\_count_i + \beta_0 z\text{-}Price + \gamma_0 is\_location\_exact + \delta_0 room\_type_{ij} + [\ \varepsilon_{ij} + u_i]$$

Results

From the exploratory data analysis, the total number of properties listed is 6,716 and the total number of hosts is 4,991. Property prices distribution was right skewed with many outliers (Figure 1). Comparing the distribution of the observed number of reviews within the last twelve months with the modeled number of reviews using poisson distribution, it is clear that there are many listings that have received zero reviews. Therefore, we decided to fit a Zero-Inflated Model, instead of a Poisson Model (Figure 2). Looking at the relationship between review score ratings and the number of reviews, we can see that listing with higher review scores have more reviews (Figure 3). This suggests that a visitor is more likely to leave a higher review rating than a lower one. Furthermore, hotel rooms had the highest amount of reviews, followed by entire homes and apartments, private rooms and then shared rooms.

For multilevel models, we decided to include only the prices that are below $500, which yielded average property prices of $119.4 (Table 1). Properties that received High rating had lower mean price than properties that received Low rating (Table 2, Figure 5). Private rooms received higher proportion of High reviews than entire homes/apartments (Figure 6). Moreover, the majority of properties had the exact location listed on the website (n=1634 vs. 362). Additionally, properties with exact location had higher proportion of High review relative to the ones with imprecise location (Figure 7). As expected, properties with hosts that are super-hosts receive higher proportion of High review ratings than those who are not super-hosts (Figure 8).

ZIP Model:

All of the coefficients of the final ZIP model are shown in Table 3. For those listings that had a visitor, with each additional 1 point increase in the cumulative score (review score rating), the average number of reviews given increased by 2.4%, controlling for prices and room type. Moreover, for listings that had a visitor, for each additional 10 dollars increase in price, the average number of reviews decreases 2%, controlling for the review score rating and room type. For the shared room listings that had a visitor, the average number of reviews decreases by 59.1% from the reference level of entire houses/apartments, controlling for the review score rating and price. While private room listings had a 29.67% decrease in number of reviews, hotel rooms have a 36.92% decrease from the reference level of the entire houses/apartments, controlling for review score rating and price. As the host response rate increases by 1%, the estimated odds of a listing being dormant decreases by 3.1%. All of the predictors were statistically significant and have contributed meaningfully to predicting the number of reviews in the last twelve months. We also ran a vuong test and concluded that the ZIP model is significantly better fit for our data than poisson (BIC z-statistic: -10.052, p-value= 2.22 *e^-16).

Multilevel Model:

Our final multilevel model predicts the quality of review rating (High or Low) by using price z-score, if the property location is exact, and room type at level one, and super_host status and host listing count at level two. Results from the final model are depicted in Table 4. The odds of receiving a High review decreases by 13.62% for each unit increase in the standard deviation from mean price when adjusting for the other variables. Similarly, the odds of receiving a High rating is 44.6% higher for properties with exact listed location compared to properties with inaccurate location, after controlling for price z-score, super_host status, room type, and listing count. The odds of receiving a High rating for private room is 93.6% higher than entire home/apartment, after controlling for price z-score, superhost status, location accuracy, and listing counts. In the same way, superhosts have 5.57 times higher odds for receiving a High reviews compared to hosts that are not superhost, after controlling for price z-score, listing counts, location accuracy, and room type. Finally, for a unit increase in the listing count's standard deviation from the listing count mean, the odds of receiving a High rating decreases by 20.4%, after controlling for price z-score, location accuracy, superhost status, and room type. No other predictors, including interactions between property level and host level variables, were statistically significant to the model.

Discussion

Overall, this study informed our research questions relating to the number and quality of reviews that a customer leaves and the various listings characteristics. In support of previous findings, property price, the rating of the reviews, and the room type are significant predicting the number of reviews a visitor would provide following a stay in an Airbnb facility. Consistent with our hypotheses, the results suggest that listings that rent out the entire apartment/house and listings with higher reviews received more reviews on average. However, while we initially expected that the more expensive a listing price is, the more reviews a listing would receive, our results indicate that there is actually a decrease in number of reviews with increasing price. Similarly, with higher prices, reviewers were more likely to leave a lower rating, after taking into consideration other factors such accuracy of the listed location and room type. An important determinant of Airbnb reviews is the characteristics of the property host. In our analysis, we saw that superhosts are more likely to receive a "High" review as we expected. Similarly, the higher number of listings a host has, the more likely they are to receive a High review (10 out of 10 points). Our results are consistent with previous studies examining factors that predict higher ratings. A previous concluded that superhosts that communicate effectively with visitors and provide accurate information about the property received higher ratings (Sampathi, 2019).

Therefore, our analysis could be replicated and applied to analyse the quantity and quality of reviews for Airbnb in other cities.

While our findings are conclusive and consistent with the literature, there exist several limitations that may impact the strength of the conclusions. For example, we had trouble modeling large eigenvalues. However, after standardizing our numerical variable (price and host listing count), we were able to run models without running into lack of convergence issue. Additionally, the distribution of price was highly right skewed with some very big values. Moreover, for the ZIP model, all of the p-values were statistically significant and it is unclear if it is due to overdispersion or other contributors that we cannot account for in ZIP. The Airbnb database does not provide the total number of visitors for each listing because it is not public information, which limits the accuracy of our conclusions. Another limitation to our study is that we do not have good variables that would reflect visitors interactions with hosts. Therefore, we can't fully account for hosts characteristics. Additionally, we did not consider the content of the reviews themselves; we only looked at the scores, which limits our understanding of the reviews quality. Similarly, we did not account for descriptive variables such as neighborhood description, property rules, and amenities.

We can generalize this study to property listings located in the Twin Cities, Minnesota between the years of 2018 and 2019. It is also important to note that there could be many variables that we did not include in our study. Some confounding variables that we should consider are the location of the listing and how far it is from the main tourist attractions in the city. For future studies, we suggest that the models include more variables and account for interaction between them. We would also suggest that future studies explore other cities and over longer periods of time to see if the variables that we found remain significant predictors in the models.

References

Data source: Inside Airbnb. Adding data to the debate. (2019). Retrieved 10 November 2019, from http://insideairbnb.com/get-the-data.html

Folger, J. (2019, November 18). Airbnb: Advantages and Disadvantages. Retrieved from https://www.investopedia.com/articles/personal-finance/032814/pros-and-cons-using-airbnb.asp.

Jaleesa, & Bustamante. (2019, November 7). Airbnb Statistics: User & Market Growth Data [November 2019]. Retrieved from https://ipropertymanagement.com/airbnb-statistics.

Lawani, A., Reed, M. R., Mark, T., & Zheng, Y. (2018, November 21). Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. Retrieved from https://www.sciencedirect.com/science/article/pii/S016604621730340X

Sampathi, A. (2019). Analyzing Airbnb reviews using SAS® Text Miner and Predicting the factors contributing for higher ratings. *SESUG Paper*. Retrieved from https://www.lexjansen.com/sesug/2019/SESUG2019_Paper-229_Final_PDF.pdf

Zervas, Georgios, Davide, Byers, & John. (2015, January 25). A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2554500.

Appendix:

Table 1: Variables chart

Description	Variable	Role	Type	Units
Rental price of the property	Price	Predictor (ZIP and multilevel)	Quantitative	US dollars
the holistic average rating score for a listing.	Review_scores_rating	Explanatory (ZIP)	Quantitative	1-100
the number of reviews a listing has over the past twelve months.	Number_of_reviews_ltm	Response (ZIP)	Quantitative	Reviews
what space is available to the customer (all of the property, a room, etc)	room_type	Explanatory (ZIP and Level 1)	Categorical	Private Room, Entire House/Apt etc.
the rate a host responds to a query from a customer.	host_response_rate	Explanatory (ZIP)	Quantitative	1-100
Average review score	bin_rating	Response (multilevel model)	Categorical	Low High
Precision of property location	is_location_exact	Predictor (level 1)	Binary	True False
Cleaning fees	cleaning_fee	Predictor (level 1)	Quantitative	US dollars
How many listings a host has	host_listing_count	Predictor (level 2)	Quantitative	Number of listings (at least 2)

Table 2: Summary statistics for all quantitative variables:

Variable	**Minimum**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**	**SD**	**N**
Total price	0	78	151	528.61	600	15000	908.23	6716
Price (High ratings)	10	50	85	106.625	129	499	81.983	904
Price (Low ratings)	13	50	99	120.51	150	495	93.24	846
Review score rating	20	96	99	96.46	100	100	6.53	4152
Host response rate	1	10	10	9.65	10	10	1.03	3667
Host listing count	0	2	4	17.709	10	375	47.721	1994
Host listing count (High)	0	2	3	11.228	6	375	34.823	882
Host listing count (Low)	0	3	5	23.0085	17	375	54.571	821
Cleaning fee	0	20	50	75.99	100	1000	83.60	4583
Cleaning fee (High rating)	0	20	35	51.087	75	385	48.113	803
Cleaning fee (Low rating)	0	25	50	65.306	90	600	58.534	780

Table 3: ZIP Final Model

	Estimate	Exp(Estimate)	Std. Error	Z value	Pr(>|z|)
Count: (Intercept)	1.128	3.090	0.090	12.488	<2e-16
Count: review_scores_rating	0.0237	1.024	0.000929	25.472	<2e-16
Count:	-0.00192	0.998	0.0000325	-59.213	<2e-16

finalprice					
Count: room_type Hotel room	-0.461	0.631	0.046	-9.915	<2e-16
Count: room_typePrivate room	-0.352	0.703	0.00904	-38.911	<2e-16
Count: room_typeShared room	-0.8936	0.409	0.0738	-12.112	<2e-16
Logistic: (Intercept)	-0.124	0.883	0.341	-0.365	0.715
Logistic: host_response_rate	-0.031	0.969	0.00364	-8.615	<2e-16

Table 4: Final multilevel model results predicting the quality of review (High or Low)

	Estimate	exp(Estimate)	Std. Error	Z-value	P-value
Intercept	-1.39	0.248	0.21	-6.55	5.47e-11
Price (z-score)	- 0.146	0.863	0.072	-2.012	0.044
is_location_exact	0.369	1.446	0.177	2.084	0.037
Room_tpe (Private room)	0.66	1.936	0.165	4.00	6.31e-5
Super_host	1.717	5.573	0.177	9.65	2e-16
Listing count (z-score)	-0.228	0.796	0.114	-1.986	0.047

Figure 1: Property price distribution before and after removing outliers (prices larger than $500.0).

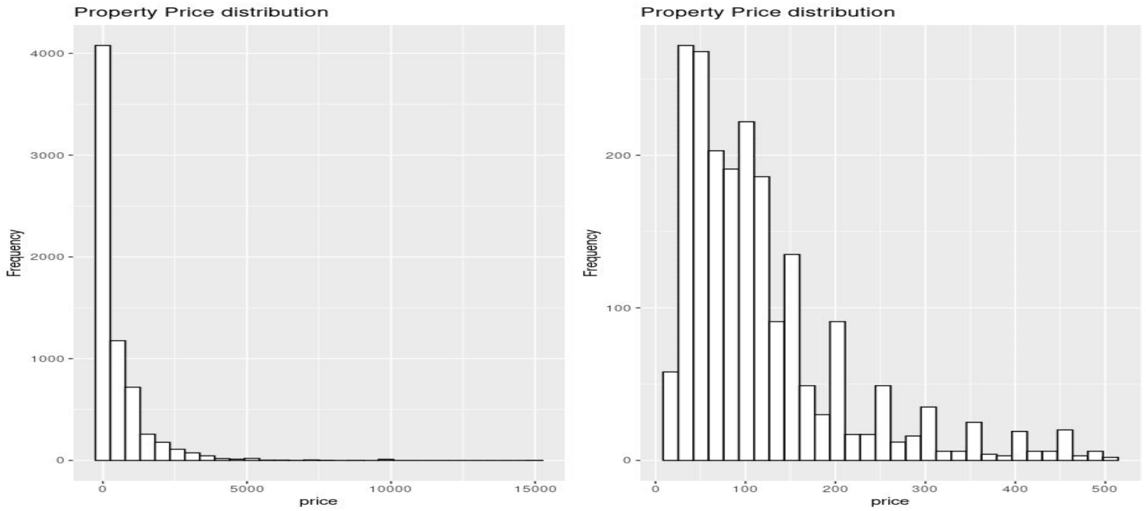


Figure 2: Histogram showing the observed number of reviews vs the modeled number of reviews

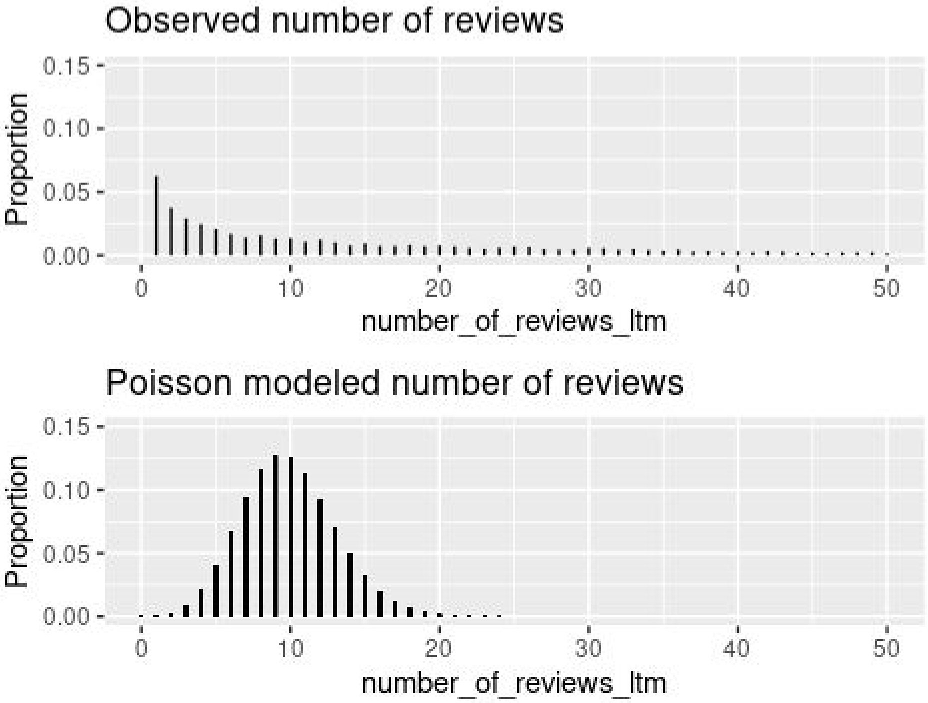


Figure 3: The relationship between review score rating and the number of reviews provided with the last 12 months for all properties.

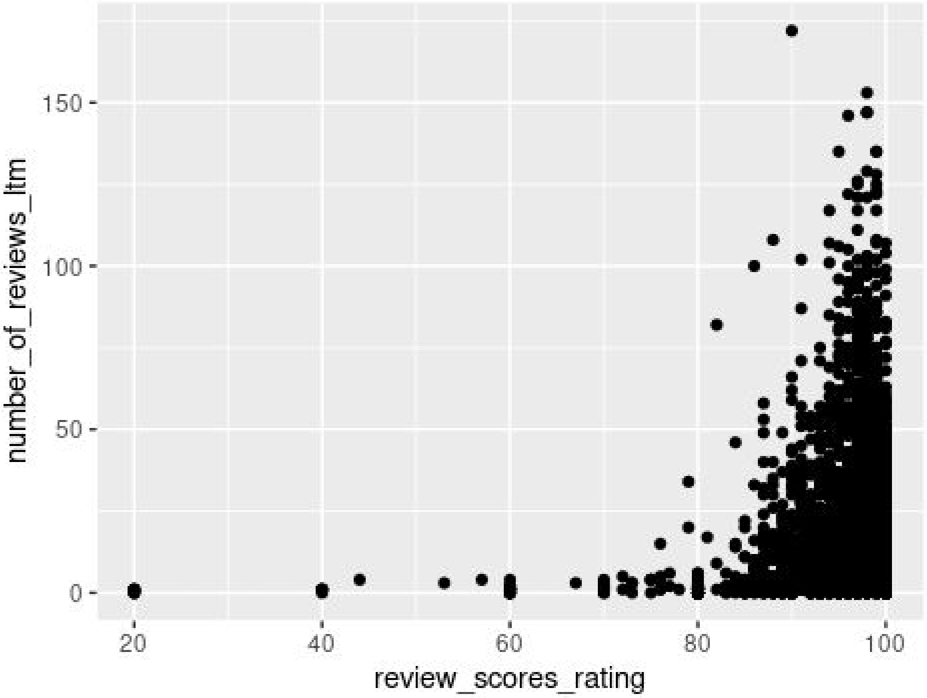


Figure 4: Number of reviews for each room type category

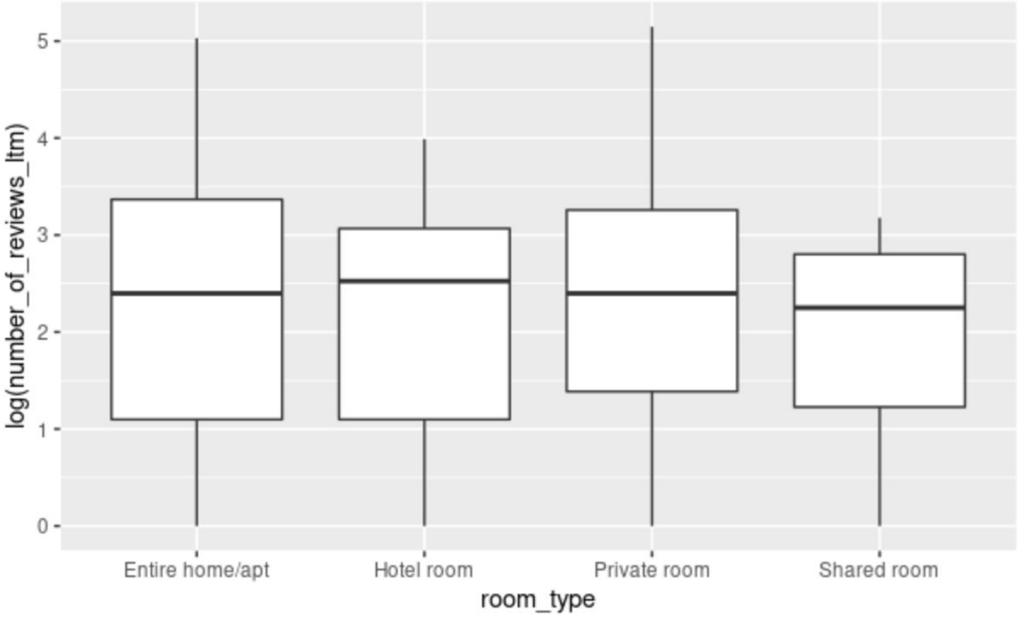


Figure 5: Boxplot of property price by rating category.

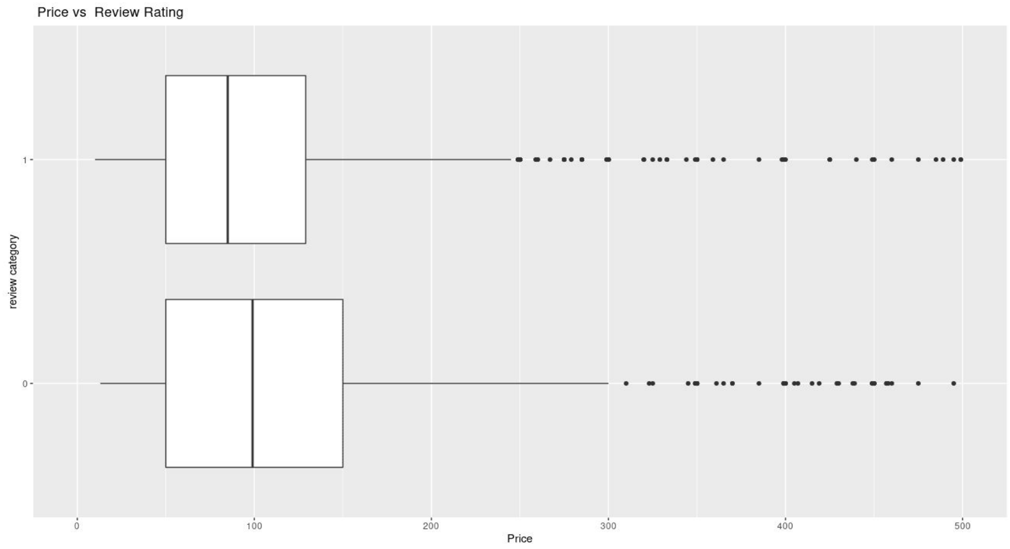


Figure 6: Barplot showing the proportion of High and Low review ratings in Private room and Home/Entire apartment.

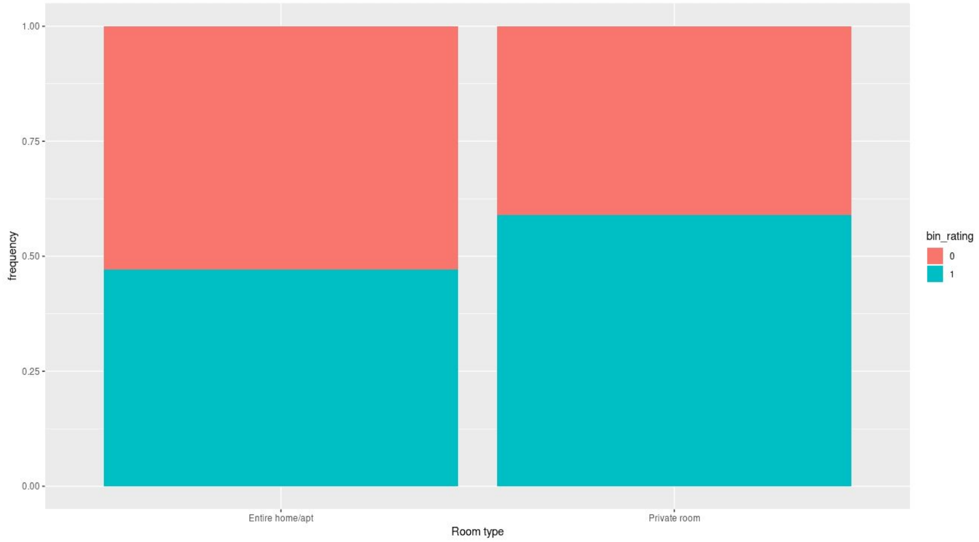


Figure 7: Barplot of the proportion of Binary review rating by location accuracy.

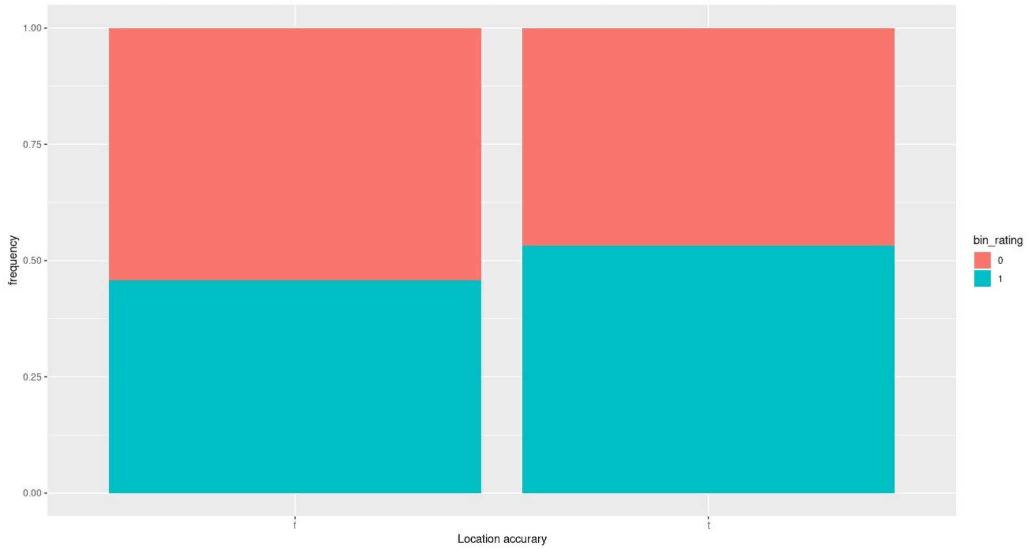


Figure 8: Barplot of the Binary review rating by super host status.

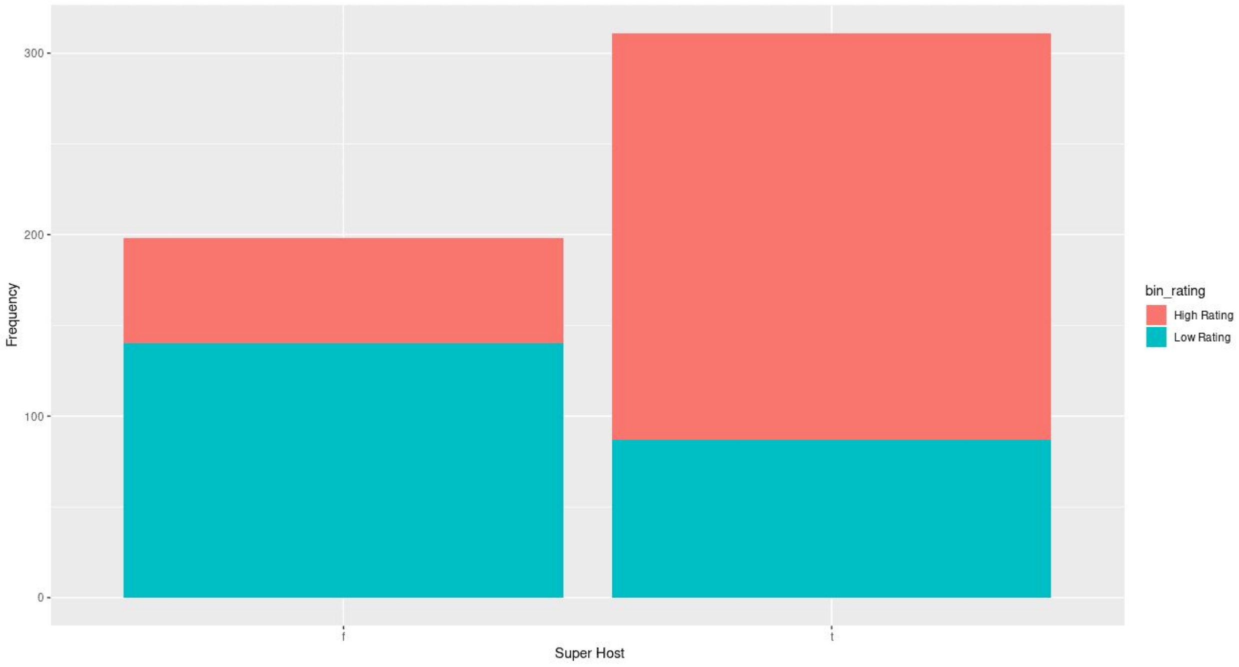
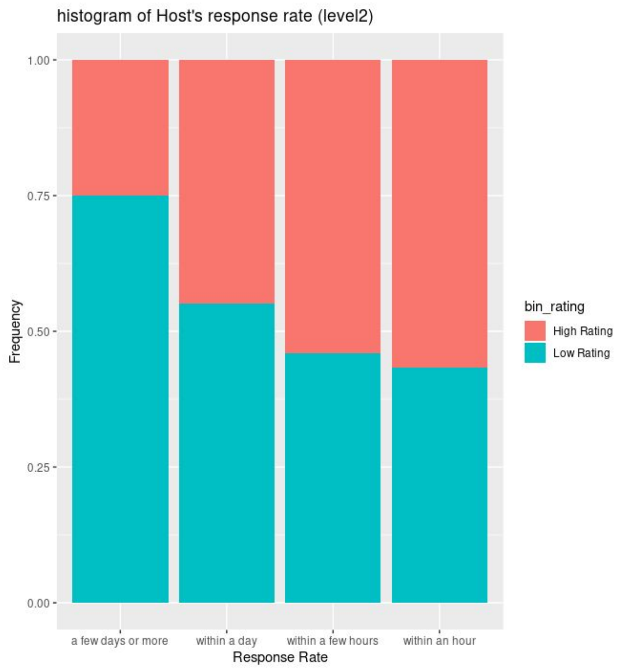
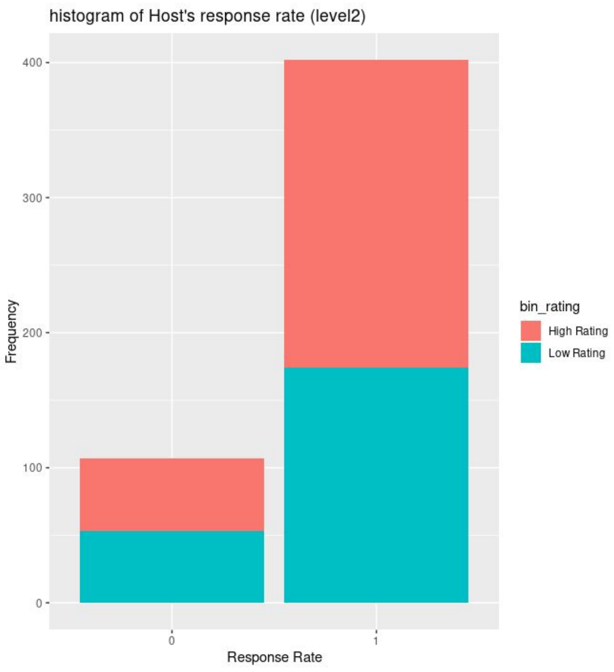


Figure 9: Barplot of the Binary review rating by host response times.





R Script:

ZIP Model:

Our strategy for modeling is to use our hypothesis of interest and what we have learned from our exploratory data analysis. In our model, all of the predictors of the review score rating, price, and room type were significant. The model summary is provided below:

```
Call:
zeroinfl(formula = number_of_reviews_ltm ~ review_scores_rating + finalprice +
    room_type | host_response_rate, data = airbnbzipmodel)

Pearson residuals:
    Min      1Q  Median      3Q     Max
 -3.698  -2.447  -1.120   1.240 225.267

Count model coefficients (poisson with log link):
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.128e+00  9.035e-02  12.488   <2e-16 ***
review_scores_rating   2.366e-02  9.290e-04  25.472   <2e-16 ***
finalprice            -1.922e-03  3.246e-05 -59.213   <2e-16 ***
room_typeHotel room   -4.607e-01  4.646e-02  -9.915   <2e-16 ***
room_typePrivate room -3.519e-01  9.044e-03 -38.911   <2e-16 ***
room_typeShared room  -8.936e-01  7.378e-02 -12.112   <2e-16 ***
```

```
Zero-inflation model coefficients (binomial with logit link):
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -0.124474   0.340999  -0.365    0.715    
host_response_rate -0.031348   0.003639  -8.615   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Number of iterations in BFGS optimization: 15 
Log-likelihood: -3.725e+04 on 8 Df
```

We exponentiated the coefficients of the ZIP model to see if they contribute a large difference in the number of reviews even though they're significant.

```
    count_(Intercept)  count_review_scores_rating 
            3.0902571                   1.0239457 
           count_finalprice   count_room_typeHotel room 
            0.9980800                   0.6308589 
count_room_typePrivate room  count_room_typeShared room 
            0.7033395                   0.4091707 
           zero_(Intercept)     zero_host_response_rate 
            0.8829609                   0.9691378 
```

The Poisson model contains the same predictors as the ZIP model and we run a vuong test and conclude that the ZIP model is a significantly better fit for the data than the Poisson model.

```
Call:
glm(formula = number_of_reviews_ltm ~ review_scores_rating + 
    finalprice + room_type, family = poisson, data = airbnbzipmodel)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
-7.698  -4.093  -1.630   1.556  26.182  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            9.539e-01  8.863e-02  10.762   <2e-16 ***
review_scores_rating   2.540e-02  9.122e-04  27.846   <2e-16 ***
finalprice            -2.105e-03  3.526e-05 -59.704   <2e-16 ***
room_typeHotel room   -4.360e-01  4.647e-02  -9.384   <2e-16 ***
room_typePrivate room -3.845e-01  9.095e-03 -42.282   <2e-16 ***
room_typeShared room  -8.775e-01  7.379e-02 -11.892   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72482  on 3290  degrees of freedom
```

```
Residual deviance: 65024  on 3285  degrees of freedom
  (484 observations deleted due to missingness)
AIC: 78311

Number of Fisher Scoring iterations: 6
```

	Vuong z-statistic <dbl>	**H_A** <fctr>	**p-value** <fctr>
Raw	-10.09541	model2 > model1	< 2.22e-16
AIC-corrected	-10.08476	model2 > model1	< 2.22e-16
BIC-corrected	-10.05228	model2 > model1	< 2.22e-16

Multilevel Model:

Model 1: Unconditional Means Model

The first model we fit was the unconditional means model. We include no predictors at either level in order to assess the amount of variation at each level. Model 1 can be fit to the review rating data, producing the following output:

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [

glmerMod]

 Family: binomial  ( logit )

Formula: bin_rating ~ 1 + (1 | host_id)

   Data: try


     AIC      BIC   logLik deviance df.resid

  2259.6   2270.5  -1127.8   2255.6     1748


Scaled residuals:

    Min      1Q  Median      3Q     Max

-2.0026 -0.6991  0.4279  0.6466  2.4666


Random effects:
 Groups   Name        Variance Std.Dev.

 host_id (Intercept) 1.724    1.313

Number of obs: 1750, groups:  host_id, 558



Fixed effects:

            Estimate Std. Error z value Pr(>|z|)

(Intercept)  0.36378    0.08784   4.141 3.45e-05 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Intercept)

   1.438755
```

We incorporated both the z-score for the property price at level 1 and super host status at level 2, both were statistically significant, but the interaction between prices and super host status was not significant.

```
> model.b2 <- glmer(bin_rating ~ z_price + super_host+
z_price:super_host+(z_price|host_id),data=try, family =binomial(link="logit"))
boundary (singular) fit: see ?isSingular
> summary(model.b2)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
Family: binomial  ( logit )
Formula: bin_rating ~ z_price + super_host + z_price:super_host + (z_price |
host_id)
   Data: try

     AIC      BIC   logLik deviance df.resid
  1959.1   1996.7   -972.5   1945.1     1576

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4774 -0.6758  0.3779  0.6541  3.0603

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 host_id (Intercept) 1.004681 1.00234
         z_price     0.005178 0.07196  -1.00
Number of obs: 1583, groups:  host_id, 473

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.7239     0.1270  -5.698 1.21e-08 ***
```

```
z_price              -0.2311     0.1169  -1.977   0.0481 *  
super_hostt           1.6884     0.1750   9.649  < 2e-16 ***
z_price:super_hostt  -0.0233     0.1657  -0.141   0.8881    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) z_pric spr_hs
z_price     -0.019              
super_hostt -0.745  0.006       
z_prc:spr_h -0.001 -0.743 -0.070
convergence code: 0
boundary (singular) fit: see ?isSingular

> exp(fixef(model.b2))
        (Intercept)             z_price         super_hostt z_price:super_hostt 
          0.4848584           0.7936459           5.4106445           0.9769663 
```

We included if the property location is accurate (level 1) and tested for its interaction with z-price. We tested for coefficients significance by considering the coefficients p-values. All variables except super host status and the interaction between z-price and location exact.

```
> model.b3 <- glmer(bin_rating ~ z_price + super_host + 
is_location_exact+z_price:is_location_exact+
(z_price+is_location_exact|host_id),data=try, family =binomial(link="logit"))
> summary(model.b3)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) 
['glmerMod']
 Family: binomial  ( logit )
Formula: bin_rating ~ z_price + super_host + is_location_exact + 
z_price:is_location_exact +  
    (z_price + is_location_exact | host_id)
   Data: try

     AIC      BIC   logLik deviance df.resid 
  1960.0   2019.1   -969.0   1938.0     1572 

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-2.0573 -0.6728  0.3670  0.6576  2.9846 

Random effects:
 Groups  Name               Variance Std.Dev. Corr       
 host_id (Intercept)        0.36349  0.60290             
         z_price            0.00392  0.06261  -1.00      
         is_location_exactt 0.22090  0.46999   1.00 -1.00
Number of obs: 1583, groups:  host_id, 473

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)    
```

```
(Intercept)                -1.048460   0.178086  -5.887 3.92e-09 ***
z_price                    -0.240693   0.171487  -1.404    0.160
super_hostt                 1.629826   0.174315   9.350  < 2e-16 ***
is_location_exactt          0.430180   0.171216   2.512    0.012 *
z_price:is_location_exactt -0.001858   0.187694  -0.010    0.992
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) z_pric spr_hs is_lc_
z_price      0.011
super_hostt -0.499 -0.023
is_lctn_xct -0.705 -0.018 -0.045
z_prc:s_lc_ -0.008 -0.889 -0.008  0.003
convergence code:
```

We removed the interaction between price z-score and if the location is exact. All variables were significant except if location is exact.

```
> model.b4 <- glmer(bin_rating ~z_price+ z_listing_count+super_host+
+                     is_location_exact+ (1|host_id), data=try, family =
binomial(link="logit"))
> summary(model.b4)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: bin_rating ~ z_price + z_listing_count + super_host + is_location_exact +
(1 | host_id)
   Data: try

     AIC      BIC   logLik deviance df.resid
  1947.5   1979.7   -967.8   1935.5     1577

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0255 -0.6750  0.3983  0.6468  3.6073

Random effects:
 Groups  Name        Variance Std.Dev.
 host_id (Intercept) 0.9174   0.9578
Number of obs: 1583, groups:  host_id, 473

Fixed effects:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.01661    0.18944  -5.366 8.03e-08 ***
z_price            -0.22413    0.07045  -3.182  0.00146 **
z_listing_count    -0.28574    0.11236  -2.543  0.01099 *
super_hostt         1.64406    0.17109   9.609  < 2e-16 ***
is_location_exactt  0.30395    0.17333   1.754  0.07951 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Correlation of Fixed Effects:
            (Intr) z_pric z_lst_ spr_hs
z_price      0.054
z_lstng_cnt  0.074  0.001
super_hostt -0.480 -0.052 -0.011
is_lctn_xct -0.748 -0.051  0.037 -0.020
```

We included host response rate variable (level 2) but it turned out to be clearly insignificant. One reason could be the distribution of response rate is very unequal with 73% of the response being within an hour.

```
 a few days or more              N/A       within a day within a few hours
          2              35             164              76              254
   within an hour
          1465
```

```
 Formula: bin_rating ~ z_price + is_location_exact + room_type + super_host +
    host_response_time + (1 | host_id)
   Data: try1

     AIC      BIC   logLik deviance df.resid
  1902.0   1955.5   -941.0   1882.0     1545

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5134 -0.6858  0.3450  0.6481  3.0986

Random effects:
 Groups  Name        Variance Std.Dev.
 host_id (Intercept) 0.994    0.997
Number of obs: 1555, groups:  host_id, 471

Fixed effects:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -2.02843    0.92956  -2.182   0.0291 *
z_price                              -0.14052    0.07271  -1.933   0.0533 .
is_location_exactt                    0.37632    0.17776   2.117   0.0343 *
room_typePrivate room                 0.69623    0.16541   4.209 2.56e-05 ***
super_hostt                           1.79591    0.18670   9.619  < 2e-16 ***
host_response_timeN/A                 1.28641    0.97858   1.315   0.1887
host_response_timewithin a day        0.27510    0.99184   0.277   0.7815
host_response_timewithin a few hours  0.81041    0.94114   0.861   0.3892
host_response_timewithin an hour      0.55100    0.91999   0.599   0.5492
```

We filtered out Shared and Hotel room, and added host listing count as level 2, all predictors remained significant. This is our final model.

```
> try1<- try%>%
+  filter(room_type == 'Entire home/apt' | room_type=='Private room')
> model.e1<- glmer(bin_rating ~z_price+ is_location_exact+ room_type+ super_host+
z_listing_count
+                     + (1|host_id), data=try1, family = binomial(link="logit"))
> summary(model.e1)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: bin_rating ~ z_price + is_location_exact + room_type + super_host +
    z_listing_count + (1 | host_id)
   Data: try1

     AIC      BIC   logLik deviance df.resid
  1898.5   1936.0   -942.3   1884.5     1548

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4090 -0.6874  0.3560  0.6428  3.0646

Random effects:
 Groups  Name        Variance Std.Dev.
 host_id (Intercept) 0.9707   0.9852
Number of obs: 1555, groups:  host_id, 471

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.39276    0.21239  -6.558 5.47e-11 ***
z_price               -0.14634    0.07272  -2.012   0.0442 *
is_location_exactt     0.36920    0.17713   2.084   0.0371 *
room_typePrivate room  0.66076    0.16515   4.001 6.31e-05 ***
super_hostt            1.71796    0.17787   9.658  < 2e-16 ***
z_listing_count       -0.22811    0.11483  -1.986   0.0470 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) z_pric is_lc_ rm_tPr spr_hs
z_price     -0.032
is_lctn_xct -0.702 -0.036
rm_typPrvtr -0.401  0.187  0.048
super_hostt -0.499 -0.022 -0.013  0.151
z_lstng_cnt  0.024  0.018  0.042  0.104  0.006
```

We added random effects for our model, and tested the difference in the model using ANOVA test. Including more random effect did not improve the final model, and made fixed effects seem less significant. We also had convergence issue and some of the correlation coefficients in the fuller model are 1 and -1, so we decided that the previous model is better.

```
> model.final.1 <- glmer(bin_rating ~ z_price + room_type + is_location_exact + 
super_host + z_listing_count + (z_price+is_location_exact|host_id),data=try1, family
=binomial(link="logit"))
> summary(model.final.1)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: bin_rating ~ z_price + room_type + is_location_exact + super_host +  
    z_listing_count + (z_price + is_location_exact | host_id)
   Data: try1

     AIC      BIC   logLik deviance df.resid 
  1905.6   1969.8   -940.8   1881.6     1543 

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-2.4089 -0.6918  0.3501  0.6396  2.7181 

Random effects:
 Groups  Name               Variance  Std.Dev. Corr       
 host_id (Intercept)        0.3990996 0.63174             
         z_price            0.0003113 0.01764  -1.00      
         is_location_exactt 0.1795360 0.42372   1.00 -1.00
Number of obs: 1555, groups:  host_id, 471

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -1.41572    0.19982  -7.085 1.39e-12 ***
z_price               -0.14833    0.07676  -1.932   0.0533 .  
room_typePrivate room  0.64536    0.16497   3.912 9.15e-05 ***
is_location_exactt     0.43797    0.17147   2.554   0.0106 *  
super_hostt            1.69238    0.17743   9.538  < 2e-16 ***
z_listing_count       -0.19194    0.11202  -1.713   0.0866 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) z_pric rm_tPr is_lc_ spr_hs
z_price     -0.062                            
rm_typPrvtr -0.420  0.202                     
is_lctn_xct -0.648 -0.025  0.027              
super_hostt -0.540 -0.019  0.161 -0.009       
z_lstng_cnt -0.049  0.063  0.108  0.146 -0.013
convergence code: 0
Model failed to converge with max|grad| = 0.00949315 (tol = 0.001, component 1)
```

```
> # FINAL MODEL
> model.final<-glmer(bin_rating ~ z_price + is_location_exact + room_type+ super_host+
+                        z_listing_count+ (1|host_id),data=try1, family
+                  =binomial(link="logit"))
> model.final.1 <- glmer(bin_rating ~ z_price + room_type + is_location_exact +
super_host +
+                          z_listing_count + (z_price+is_location_exact|host_id),
+                        data=try1, family
+                  =binomial(link="logit"))
> summary(model.final.1)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: bin_rating ~ z_price + room_type + is_location_exact + super_host +
    z_listing_count + (z_price + is_location_exact | host_id)
   Data: try1

     AIC      BIC   logLik deviance df.resid
  1905.6   1969.8   -940.8   1881.6     1543

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4089 -0.6918  0.3501  0.6396  2.7181

Random effects:
 Groups  Name               Variance  Std.Dev. Corr
 host_id (Intercept)        0.3990996 0.63174
         z_price            0.0003113 0.01764  -1.00
         is_location_exactt 0.1795360 0.42372   1.00 -1.00
Number of obs: 1555, groups:  host_id, 471

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.41572    0.19982  -7.085 1.39e-12 ***
z_price               -0.14833    0.07676  -1.932   0.0533 .
room_typePrivate room  0.64536    0.16497   3.912 9.15e-05 ***
is_location_exactt     0.43797    0.17147   2.554   0.0106 *
super_hostt            1.69238    0.17743   9.538  < 2e-16 ***
z_listing_count       -0.19194    0.11202  -1.713   0.0866 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) z_pric rm_tPr is_lc_ spr_hs
z_price     -0.062
rm_typPrvtr -0.420  0.202
is_lctn_xct -0.648 -0.025  0.027
super_hostt -0.540 -0.019  0.161 -0.009
z_lstng_cnt -0.049  0.063  0.108  0.146 -0.013
```

```
convergence code: 0
Model failed to converge with max|grad| = 0.00949315 (tol = 0.001, component 1)

> # Anova Chisq test
> anova(model.final.1, model.final)
Data: try1
Models:
model.final: bin_rating ~ z_price + is_location_exact + room_type + super_host + 
model.final:     z_listing_count + (1 | host_id)
model.final.1: bin_rating ~ z_price + room_type + is_location_exact + super_host + 
model.final.1:     z_listing_count + (z_price + is_location_exact | host_id)
              Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
model.final    7 1898.5 1936.0 -942.27   1884.5                         
model.final.1 12 1905.6 1969.8 -940.78   1881.6 2.9661      5     0.7052
```