Report of MA678 Midterm Project

Guangze Yu

12/1/2021

0.1 Abstract

Airbnb is an online marketplace for lodging, primarily homestays. It helps tourists to rent other homestays in other cities. Based on the willingness of the host, each house will be labeled different price. Here rise a problem: Which type of factor influences Airbnb houses' price? To address this problem, I choose the Boston area as a representative, subset some factors, and then build the multilevel model. This model shows that there is postive relationship between rating scores,listing number with price. Also, there is positive relationship between the number of bed, the number of bedroom with price. This report consists of 5 main parts: Introduction, Method, Result, and Discussion.

0.2 Introduction

Even I choose the Boston area as a representative city, the location of Airbnb houses will different from the renting price. Some key variables might influence the one-night stay price of Airbnb house. However, we don't know which factors influence the price at most or whether those factors change through different neighbourhood catogeries.

This report will start from data cleaning and then model fitting, last model discussion. The model selection code is attched in model_selection.R.

0.3 Method

0.3.1 Data cleaning and Processing

The main data set is published on Inside Airbnb:Adding data to debate. The benefit of this public Airbnb data set is that most variables are cleaned. Unique Identification include ID and house name. Although there is no relationship between ID and price,the only string interested columns are useless under our setting.

Firstly, I remove the unit of price and change the scientific notation to numeric variable. Secondly, I only keep complete case for `price` and `review scores`. There is no missing value for other varibales. Thirdly, to keep all variables into log transformation to make sure "normal" and within the same unit range.

Important variables description are shown in below chart.

column names	explanation
name	The name of specific Airbnb house
price	The price of one-night stay (dollar/night)
longitude/latitude	The geographical location
review_scores_rating	The overall review score in the scale of 5
host_listings_count	The number of listings the host has
neighbourhood_cleansed	The neighbourhood identification

column names	explanation
property_type	Host self- selected property type
bedrooms	The number of bedrooms
beds	The number of beds

After cleaning, I have a dataset with 1837 observations and 18 columns. However, not all variables are related to our problem. So only few variables are chosen based on below analysis.

0.3.2 Exploratory Data Anallysis

The distribution EDA part is attached in Appendix for reference. Here only talk about relative EDA result. Since the number of hotel room and shared room of `room_type` is limited number and doesn't include too much useful information, I only keep entire room and private room. There is total 1837 observation, while only 15 hotel room and 6 shared room. Most neighbourhood lack those two types of room type. It is meaningless to include those two room type.

Before fitting the model, I would like to investigate whether there is linear relationship between depedent variables and indepedent variables by drawing the scatter plot. Because there is multiple indepedent variables, I seperatly three scatter plots to visualize. As I mentioned before, it is important to take log transformation of origianl variables.

Figure I shows the relationship between price and overall review scores. Although not all neighbourhood have the same slope trend and intercept, most neighbourhood still follow the slope to the right-top corner. Also, entire room/apartment group seems have more similar trend than that of private room.

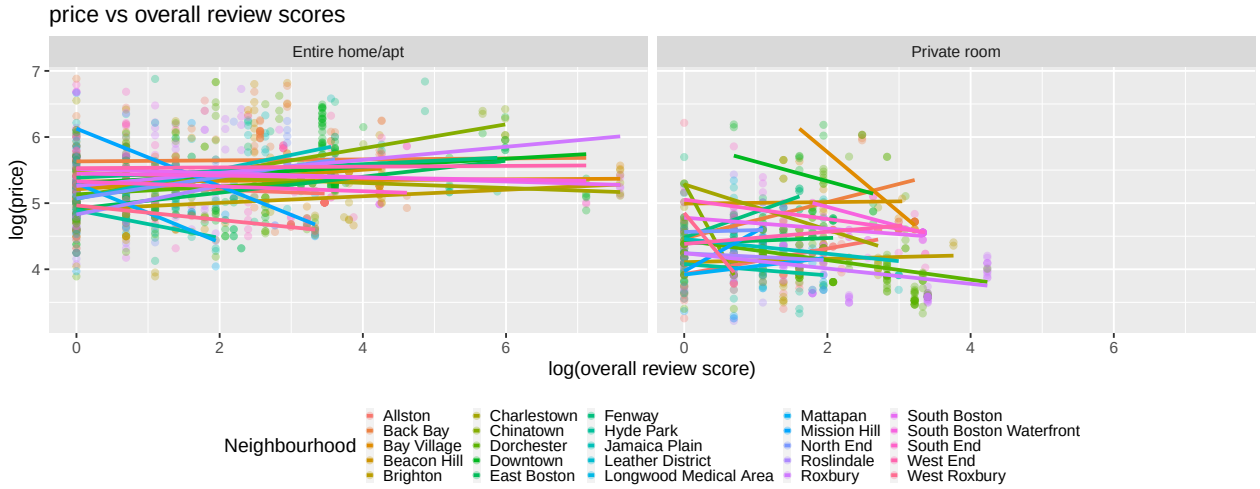


Figure 1: Data was separate into Entire home type and Private room type. Different colors represent different neighbourhood categories.

Figure II shows the relationship between price and the number of bed. Most neighbourhood have the the trend is almost the same,pointing to top-right, although there exists some expectation. Also, entire room/apartment group seems have more similar trend than that of private room. When I fit the model, `beds` should take in account.

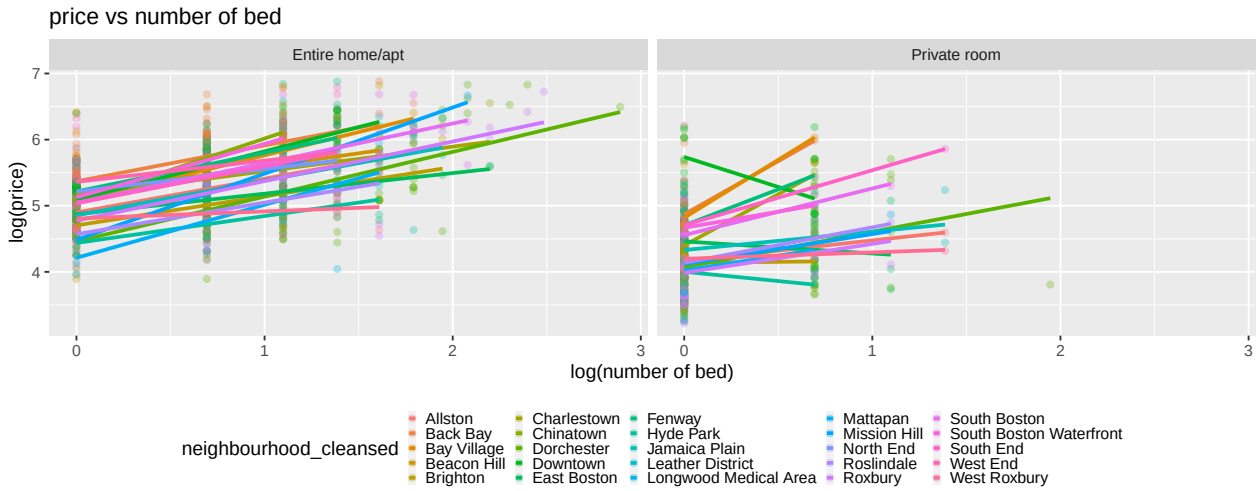


Figure 2: Data was separate into entire home/apt and private room. Different colors represent different neighbourhood categories.

Figure III shows the relationship between price and the number of bedroom. Most neighbourhood have the the trend is almost the same,pointing to top-right Also, entire room/apartment group seems have more similar trend than that of private room.Some neighbourhood don't have enough private room to investigate the relationship with price. When I fit the model, `bedrooms` should take in account.

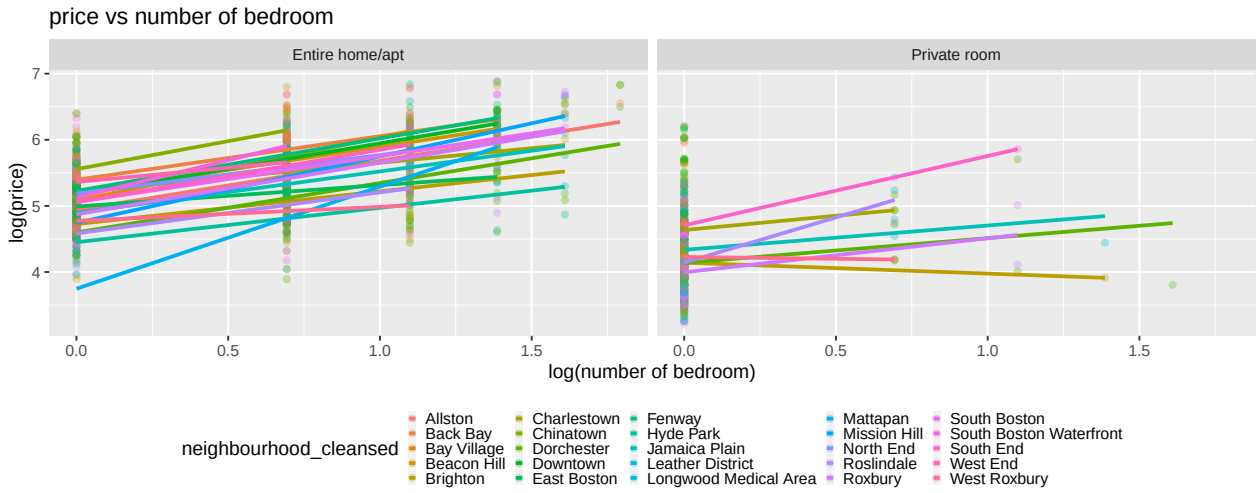


Figure 3: Data was separate into entire home/apt and private room. Different colors represent different neighbourhood categories.

Figure IV shows the relationship between price and the number of house that this host list on Airbnb. Most neighbourhood have the the trend is almost the same,pointing to top-right Also, entire room/apartment group seems have more similar trend than that of private room.Some neighbourhood don't have enough private room to investigate the relationship with price.

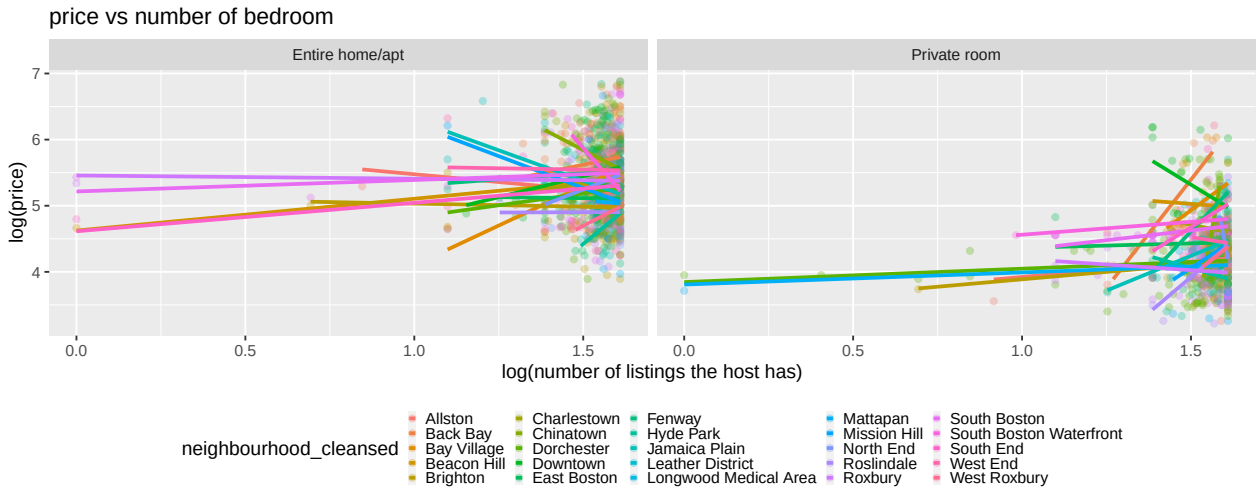


Figure 4: Data was separate into entire home/apt and private room. Different colors represent different neighbourhood categories.

0.3.3 Model Fitting

Considering different neighbourhood categories, I will use multilevel model to fit the data. For `log_bed` and `log_bedroom`, the log of bed number and the log of the bedroom number, from EDA part, it is clear enough to include them as predictors, so I use varying slope and varing intercept in multilevel models. Besides, since I also have `room_type` as another level of cluster, when I fit the model, the AIC and BIC of using `room_type` as second level or first level will extremely big. So, I decide to only keep one category variable and keep `room_type` as dummy variable. Among all possible models, I choose the final model below.

```
model11 <- lmer(log_price ~ log_house + log_score + log_bed + log_bedroom + factor(room_type) +
     (1 + log_bed + log_bedroom | neighbourhood_cleansed), subset_listings)
```

And to see the fixed effects below, all variables are significant at alpha = 0.05 level and the value is round to two digits.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.58	0.14	408.07	33.72	0.00 ***
log_house	0.02	0.01	1783.86	2.61	0.00245 **
log_score	0.20	0.08	1777.45	2.62	0.00870 **
log_bed	0.27	0.05	17.27	5.89	6.91e-09 ***
log_bedroom	0.42	0.05	16.94	8.22	2e-16 ***
room_type/private	-0.46	0.03	1729.1	-16.75	0.00 ***

0.4 Result

0.4.1 Model Coefficient

The full list of coefficient, fix effect and random effect is attached in the appendix. Here I just take Allston as an example, the formula should look that:

$$log(price) = 4.031 + 0.023 \cdot log(host\_listings\_count) + 0.213 \cdot log(review\_scores\_rating) + 0.532 \cdot log(beds) + 0.541 \cdot log(bedrooms) - 0.45 \cdot room\_type$$

where `room_type` = 1 means private room,otherwise entire apartment.

The intercept is larger than 0, which make sense because no matter which condition of this Airbnb house is, there still exists a rental price. Also, even if there doesn't exist rating score or beds. The price still valid. However, because the number of bed and bedroom can be 0, it is hard to explained. I forced to change those negative infinite value to 0. For each 1% difference in review rating, the predicted difference in price is 0.213% for every neighbourhood, which other variables constant. It make sense that the coefficient of `host_listing_cout` is the same and positive for all neighbourhood because more house that the host own more money that he/she want to make. Beside, the coefficient of `review_scores_rating` is the same and positive because higher rating means higher quality of this house. I decide to keep `log_score` and `log_house` as fixed effect variables because it doesn't vary through the change of categories and also high AIC/BIC value. The coefficient of `log_bed` is positive might because with the increasment of bed, more people can live in this house, which will lead the price rise. The coefficient of `log_bedroom` is positive might because with the increasment of bedrooms, more people can live in this house, which will lead the price rise. The coefficient is the same and negative for all neighbourhood because entire apartment have better living experience.

0.4.2 Model Validation

From the Residual plots in Figure V we can see that the mean of residuals is almost 0 and dots spread almost the same from the right to the leftAnd for the Q-Q plot in Figure IV, majority dots are on the lines so the normality is good. Figure V shows that there are not obvious leverage point. Figure VI and V is attached in Appendix.

0.5 Discussion

Based on my model and under my knowledge, the price of Airbnb house is positively related with the review rating score, the number of listings the host has, the number of bed and the number of bedrooms. In average, the price for entire apartment is higher than that of private room. Also, beside the fix effect, the random effect still play an important role when predicting.

Beside the convicing predition result, I would like to talk about the limitation of this model. Firstly, because all catogies have postive,this model cannot have solid conclusion about the trend. Secondly, the first rule that I choose model is based on AIC/BIC value. Involving more variables will explain more about variance. However, it might overstate the generalizability of the average treatment effect. Thirdly, the sample is drawn from small dataset. I cannot conclude that this sample is randomly picked and whether it represent the whole population. Lastly, the ICC value among the cluster is lower than the normal value, which might cause some problems.

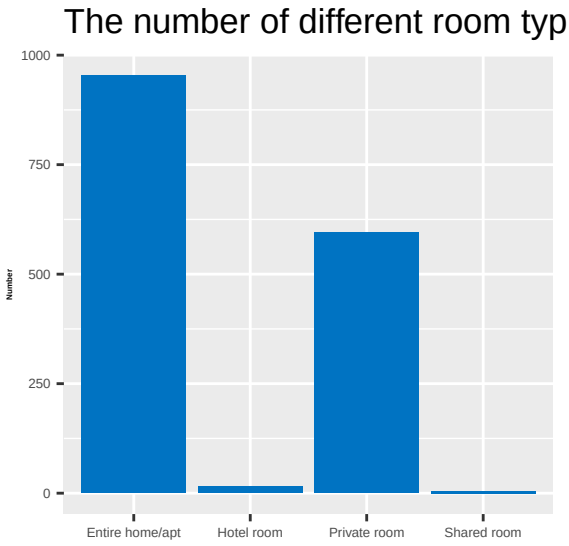
For the future, I would like to add more variables into the model. Also, combined with other dataset which contain other time spots price, it can help to study the population problem. Moreover, try to read more case-study to look for a better solution for this dataset not only limited to multilevel model.

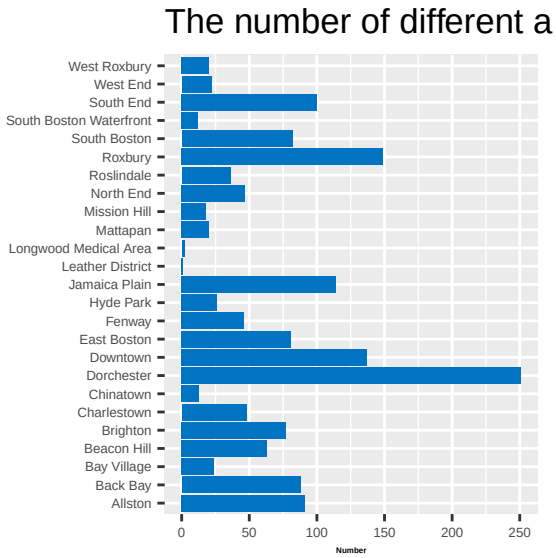
0.6 Reference

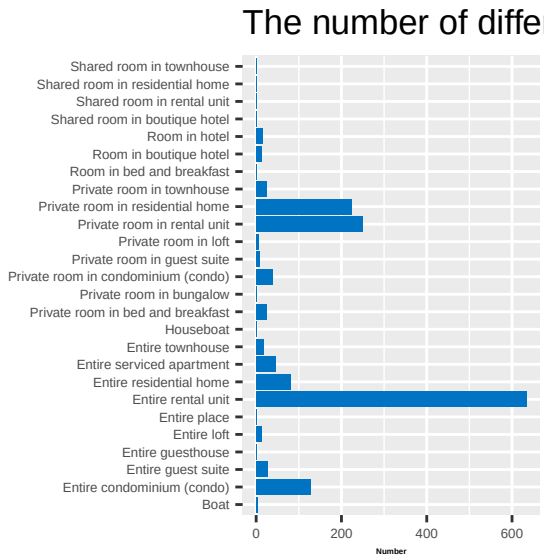
1. Bingenheimer, J. B., & Raudenbush, S. W. (2004). Statistical and substantive inferences in public health: Issues in the application of Multilevel Models. Annual Review of Public Health, 25(1), 53–77. https://doi.org/10.1146/annurev.publhealth.25.050503.153925
2. Finch, H. W., Bolin, J. E., & Kelley, K. (2019). Multilevel modeling using R. CRC Press.

0.7 Appendix

0.7.1 EDA part







0.7.2 Model validation

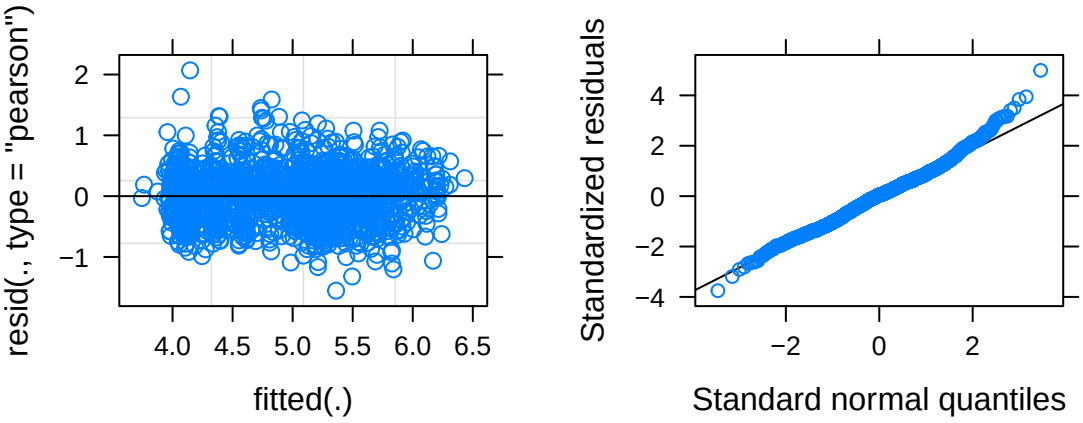


Figure 5: Residual plot and Q-Q plot.

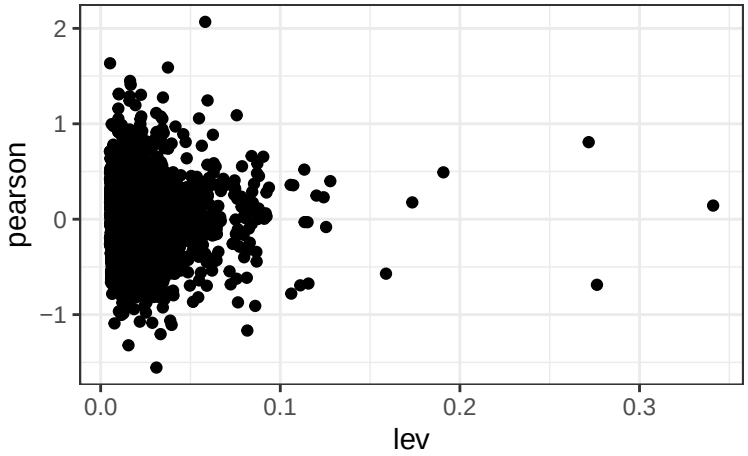


Figure 6: Residuals vs Leverage.