Midterm Project

Minqi Li

12/8/2020

Abstract

As sharing economy as a new business mode becomes more and more popular, Airbnb provides a rental online platform for hosts to accommodate guests with short-term lodging and tourism-related activities which differs from traditional hotel industry. Because of various lodging information offered on the Airbnb site, we build a multilevel model to explore what factors are related to price. Our model shows that room type, the number of reviews in per month, days available for renting, the number of guests have an influence on price. Through improving this model in the next steps, it is beneficial for hosts to fix a price for their rooms.

1 Introduction

Airbnb provides a rental online platform for hosts to accommodate guests with short-term lodging and tourism-related activities. Because it provides users with various lodging information such as various room type, location and hosts which is different from traditional hotel industry. Therefore, I want to explore what factors are related to price.

In this report, I downloaded data from Inside Airbnb which provides us with publicly available information about a city's Airbnb's listings around the world. For each city's dataset, there are 75 variables mostly about information of hosts, houses and reviews which are scraped in October, 2020. I chose 10 typical cities in United States to explore the relationship between price and other variables. The 10 cities is Boston, Chicago, Denver, Hawaii, Los Angeles, New Orleans, New York, Portland, Seattle and Washington, D.C.

2 Method

2.1 Data Cleaning

- Subseted the data by choosing 1000 pieces of listings from each city randomly, because the function of stan_lmer took a long time so that I needed to reduce data.
- Computed the proportion of missing values for each column and removed the columns that the proportion was greater than 0.3.
- Removed the useless variables based on the description of variables.
- Removed the rows that the number of reviews in 12 months was 0, because I thought the condition which doesn't have reviews in 12 months is not active.
- Replaced missing values with their suitable values.
- Changed the variables from percentage to decimal, from t/f to 1/0, from character to integer or factor.

2.2 Correlation Analysis

Firstly, I did a correlation analysis after removing character variables. Based on it, I selected the variables which had the 10 largest absolute value of correlation coefficients between price and other variables. The 10 variables are *accommodates, bedrooms, beds, reviews_per_month, number_of_reviews, number_of_reviews_ltm, availability_30, availability_365, number_of_reviews_l30d, minimum_nights.* The plot of correlation matrix is in the appendix.

Then, I did a correlation analysis of selected variables. According to it, I removed the redundant variables whose correlation coefficients with variables which have greater correlation coefficients in the first step is greater than 0.5. Therefore, the numeric variables I selected are *accommodates, reviews_per_month, availability_30, availability_365.* The plot of correlation matrix is in the appendix.

2.3 Variables Selection and Transformation

Finally, because I thought room type had an influence on price, I also added it into predictors. The final dataset included 6394 pieces of listings and 7 variables across 10 cities. Because I wanted to expore the relationship between price and other variables in the same city, I decided to use a linear multilevel model with random intercepts. I chose *price* as outcome, *accommodates, reviews_per_month, availability_30, availability_365,room_type* as predictors and *city* as random intercept.

Besides, because the scale of *price, availability_30* and *availability_365* was large, I took the log of price and deal with *availability_30* and *availability_365* in the z-score normalization method.

2.4 EDA

2.4.1 The Plot of Density Estimate

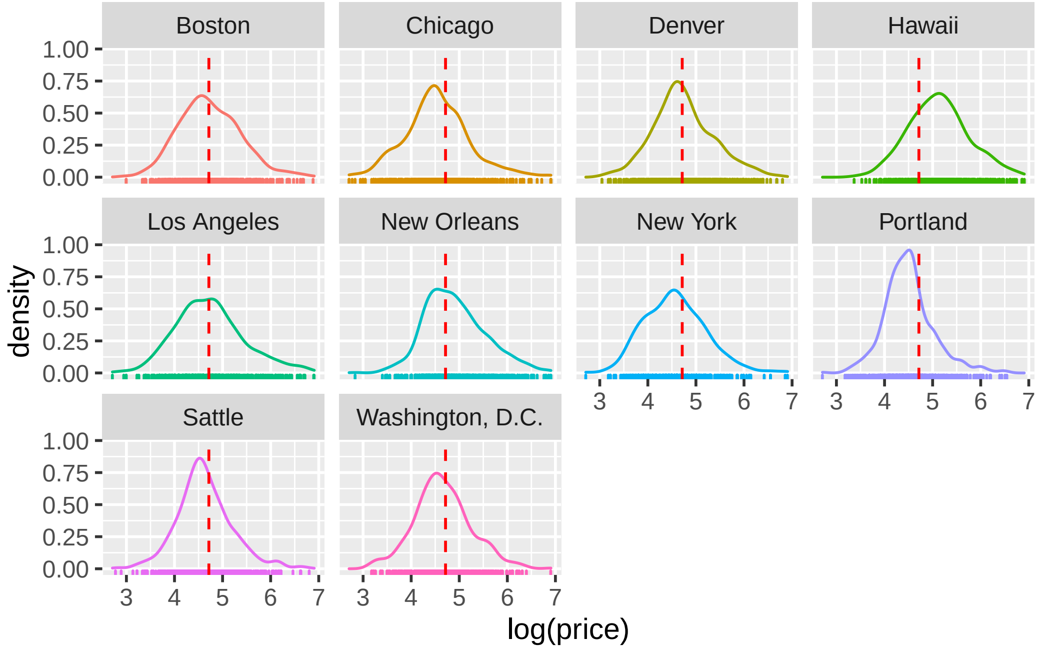


Figure 1: The plot of density estimate for price in each city.

Based on the figure 1, we could find that the distribution of price in each city has a few differences. For example, the peak in each city has the biggest difference.

2.4.2 The Scatterplot and Linear Regression for each city

```
## `geom_smooth()` using formula 'y ~ x'
```

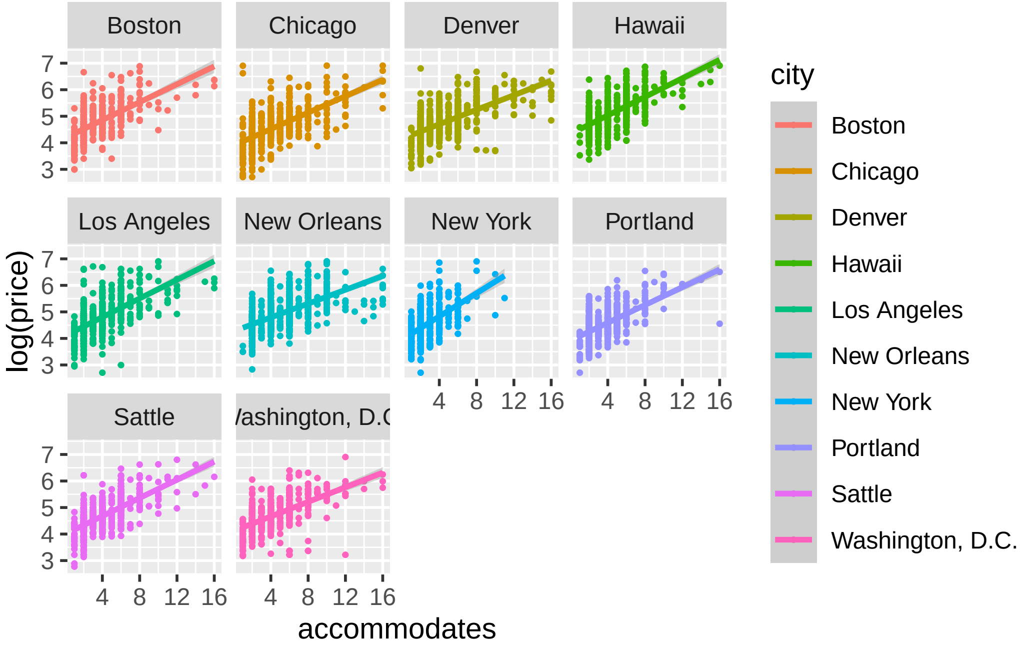


Figure 2: The scatterplot and linear Regression between the number of guests and log(price) for each city.

Because *accommodates* has maximum correlation coefficient with price, I made a scatterplot and linear regression between *accommodates* and *price_log* for each city. Based on the plot, the points are distributed in a linear trend. The scatterplots and linear regression between other variables and *price_log* are in the appendix.

In conclusion, based on explortary data analysis, there is no problem with model and feature selection.

3 Results

3.1 Model Coefficients & Estimates

According to the model's fixed effect and their 95% confidence interval. *Accommodates*, *availability_30_scale* has a positive effect on price. *reviews_per_month*, *availability_365_scale* have a negative effect on price. *Room_type* has the biggest impact on price. The order of price of room type from the highest to the lowest is entire home/apartment, hotel room, private room, shared room. All the coefficients of variables are statistically significant.

3.2 Residual plot

Based on the residual plot, most data is fitted well. A few outliers are mostly in Washington, D.C. and Chicago.

3.3 Posterior predictive checks

From the plot of posterior predictive checks, the model doesn't capture the data very well around the peak and the part of high price. This is because that the peak has a big difference in different city which could be seen in the plot of density estimate of EDA part. Besides, the part of high price is more different than the part of low price among cities.

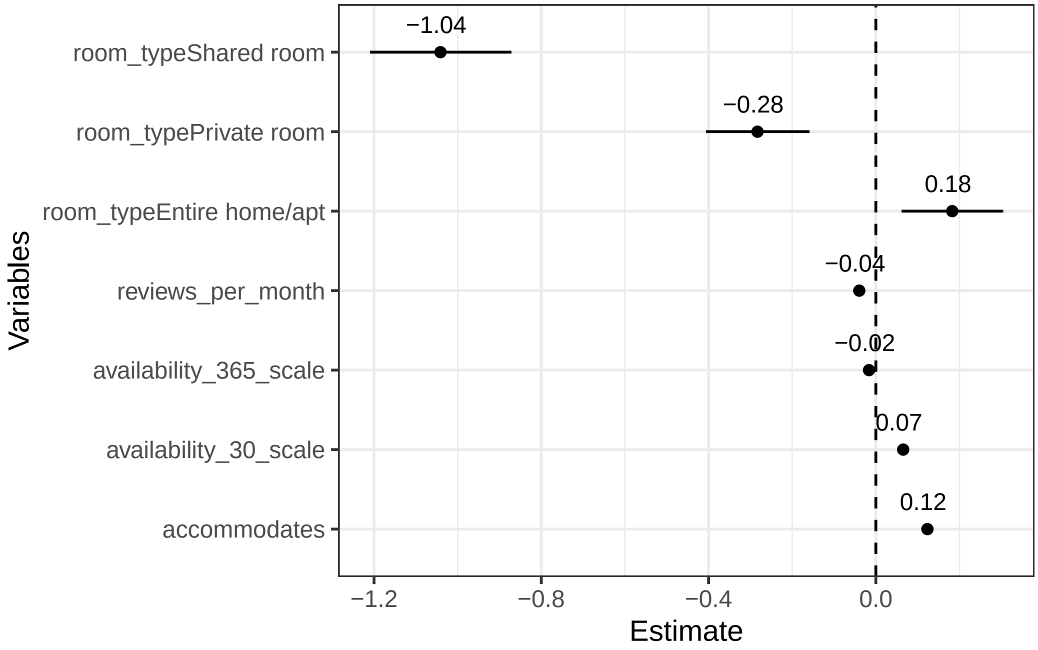


Figure 3: The model's fixed effect and their 95% confidence interval.

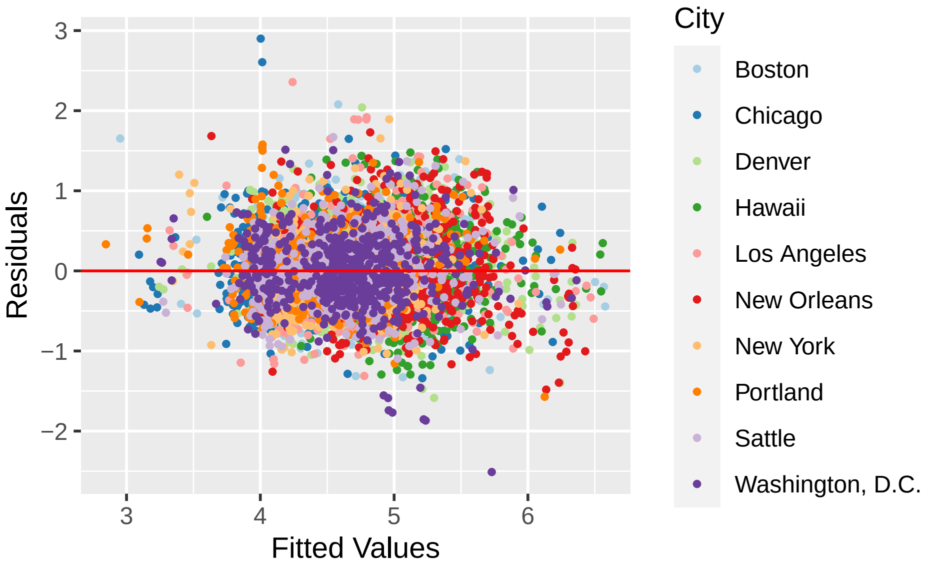


Figure 4: Residual plot

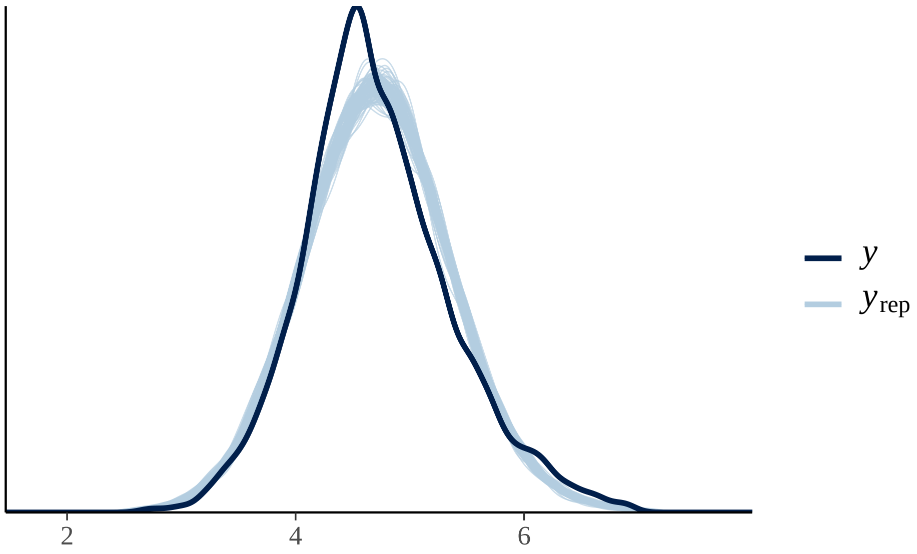


Figure 5: Density estimates of posterior predicted values and the fitted values.

4 Discussion

The model results mostly lines up people's perception. Foe example, *availability_30_scale* has a positive effect on price. On the contrary, *availability_365_scale* has a negative effect on price. It conforms the rule that the earlier we reserve rooms, the cheaper the price is. Besides, *reviews_per_month* has a negative effect on price which means that people prefer cost-effective rooms.

However, in this model, I didn't add neighborhood into predictors which may have a big influence on price. This is because that the cleanliness score of location is hardly correlated with price in the correlation analysis. And the character variable of neighborhood has too many factors so that it is difficult to add it into predictors.

Therefore, in the next step, I try to add the factor of neighborhood into predictors to explore the relationship between location and price. What's more, I need to find and remove outliers in the residual plot to fit the model better.

5 References

1. Inside Airbnb [online]. Available from: http://insideairbnb.com/get-the-data.html [accessed 3 December 2020]

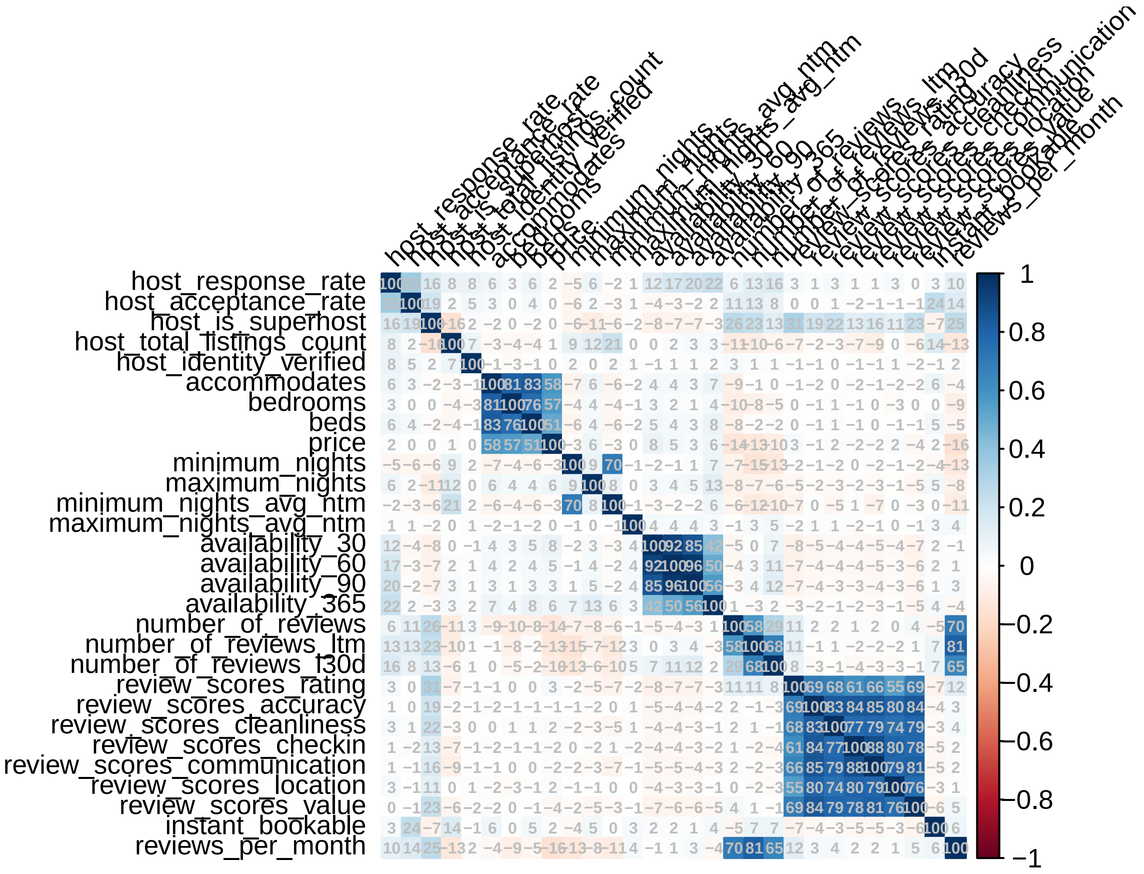
2. Wikipedia [online]. Available from: https://en.wikipedia.org/wiki/Airbnb [accessed 9 December 2020]

3. Hadley Wickham (2019). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.3.0. https://cloud.r-project.org/package=tidyverse

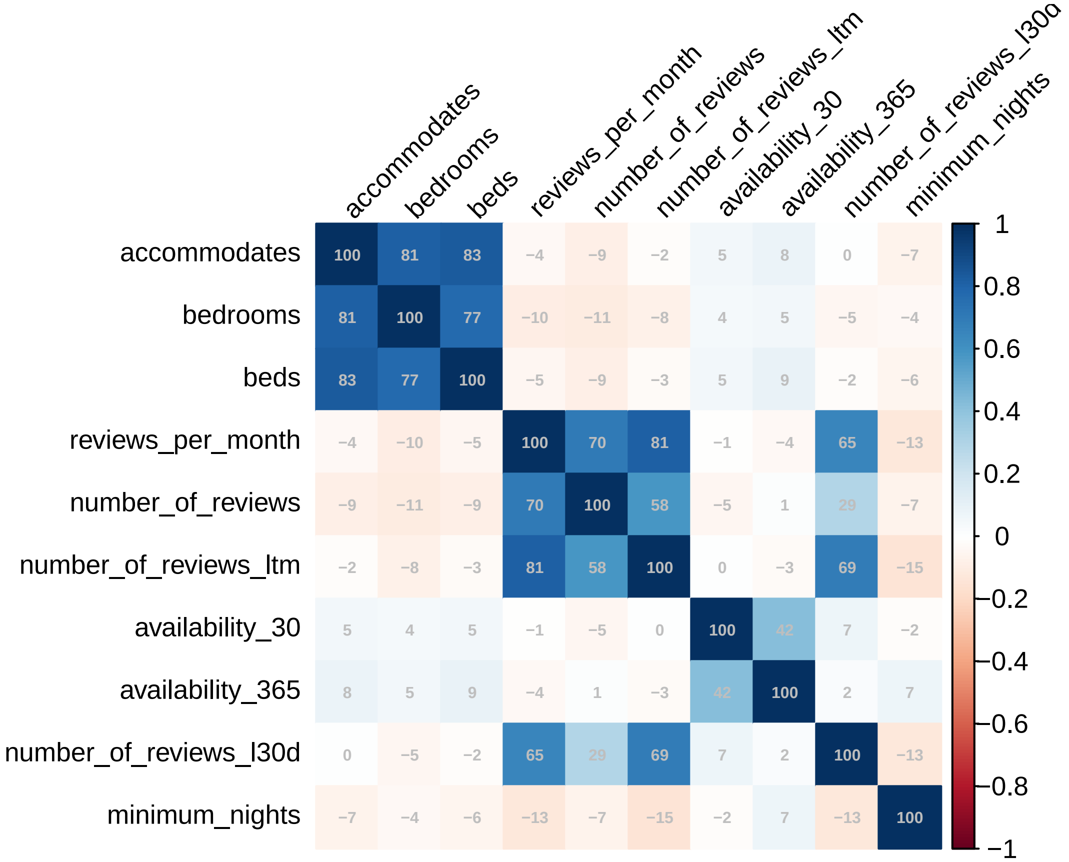
4. Jonah Gabry, Imad Ali, Sam Brilleman, etc (2020). rstanarm: Bayesian Applied Regression Modeling via Stan. R package version 2.21.1. https://cloud.r-project.org/web/packages/rstanarm/index.html

Appendix

1. The correlation matrix among variables



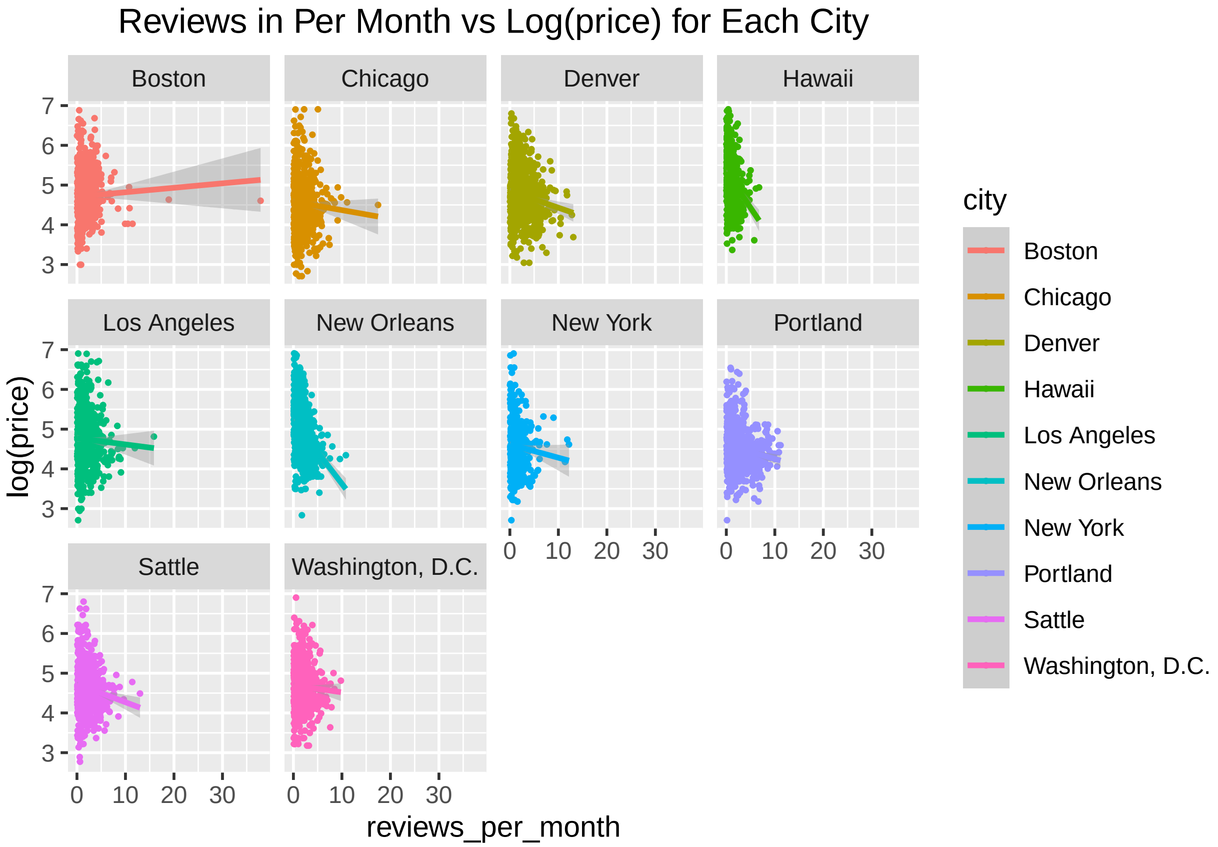
2. The correlation matrix among selected variables



	accommodates	bedrooms	beds	reviews_per_month	number_of_reviews	number_of_reviews_ltm	availability_30	availability_365	number_of_reviews_l30d	minimum_nights
accommodates	100	81	83	−4	−9	−2	5	8	0	−7
bedrooms	81	100	77	−10	−11	−8	4	5	−5	−4
beds	83	77	100	−5	−9	−3	5	9	−2	−6
reviews_per_month	−4	−10	−5	100	70	81	−1	−4	65	−13
number_of_reviews	−9	−11	−9	70	100	58	−5	1	29	−7
number_of_reviews_ltm	−2	−8	−3	81	58	100	0	−3	69	−15
availability_30	5	4	5	−1	−5	0	100	42	7	−2
availability_365	8	5	9	−4	1	−3	42	100	2	7
number_of_reviews_l30d	0	−5	−2	65	29	69	7	2	100	−13
minimum_nights	−7	−4	−6	−13	−7	−15	−2	7	−13	100

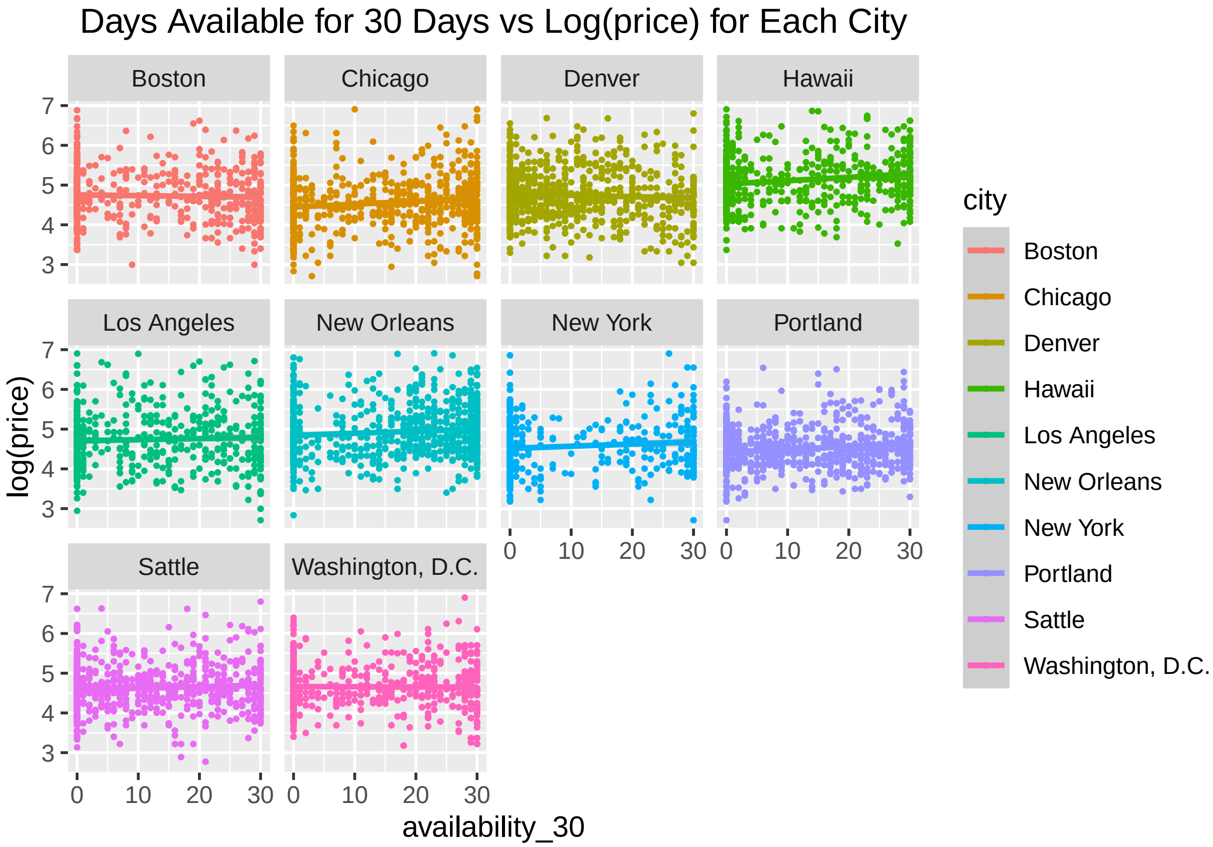
3. The plot of reviews in per month vs Log(price) for each city

```
## `geom_smooth()` using formula 'y ~ x'
```



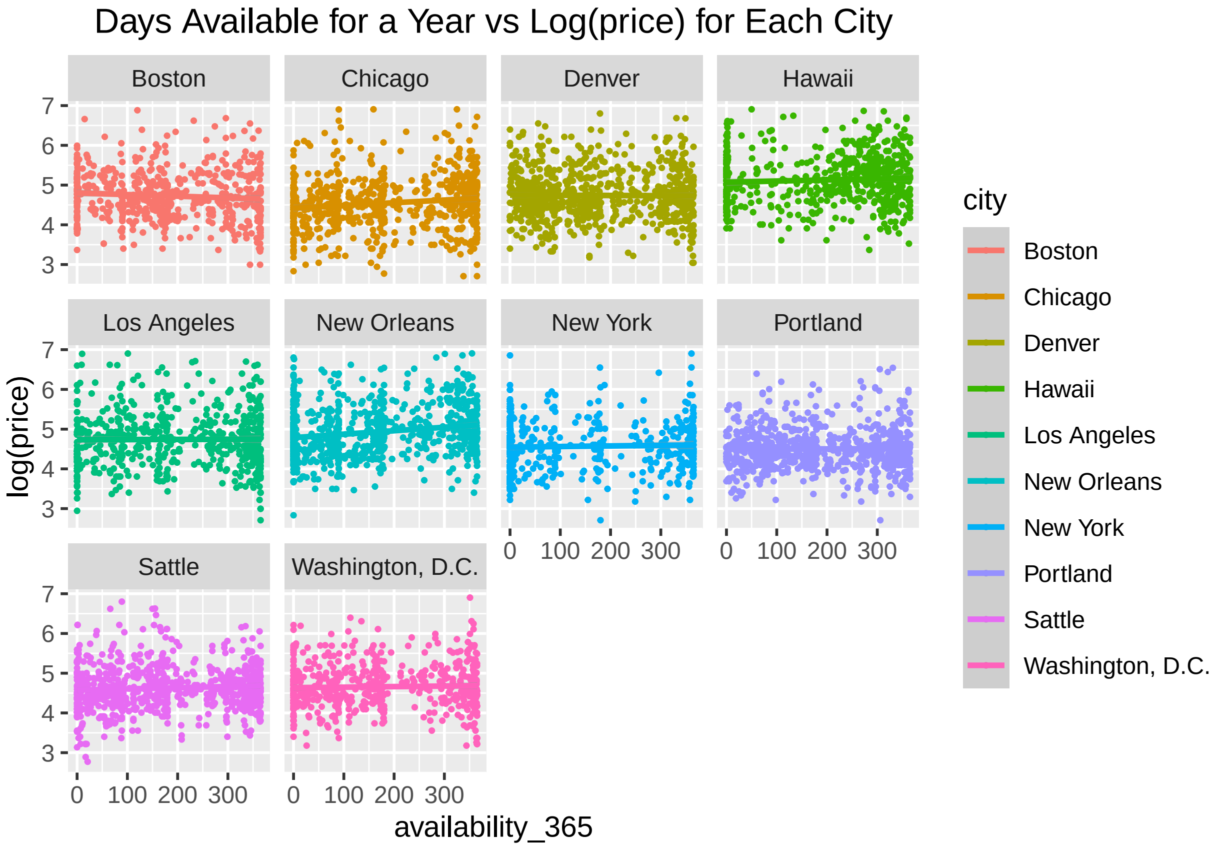
4. The plot of days available for 30 days vs Log(price) for each city

```
## `geom_smooth()` using formula 'y ~ x'
```

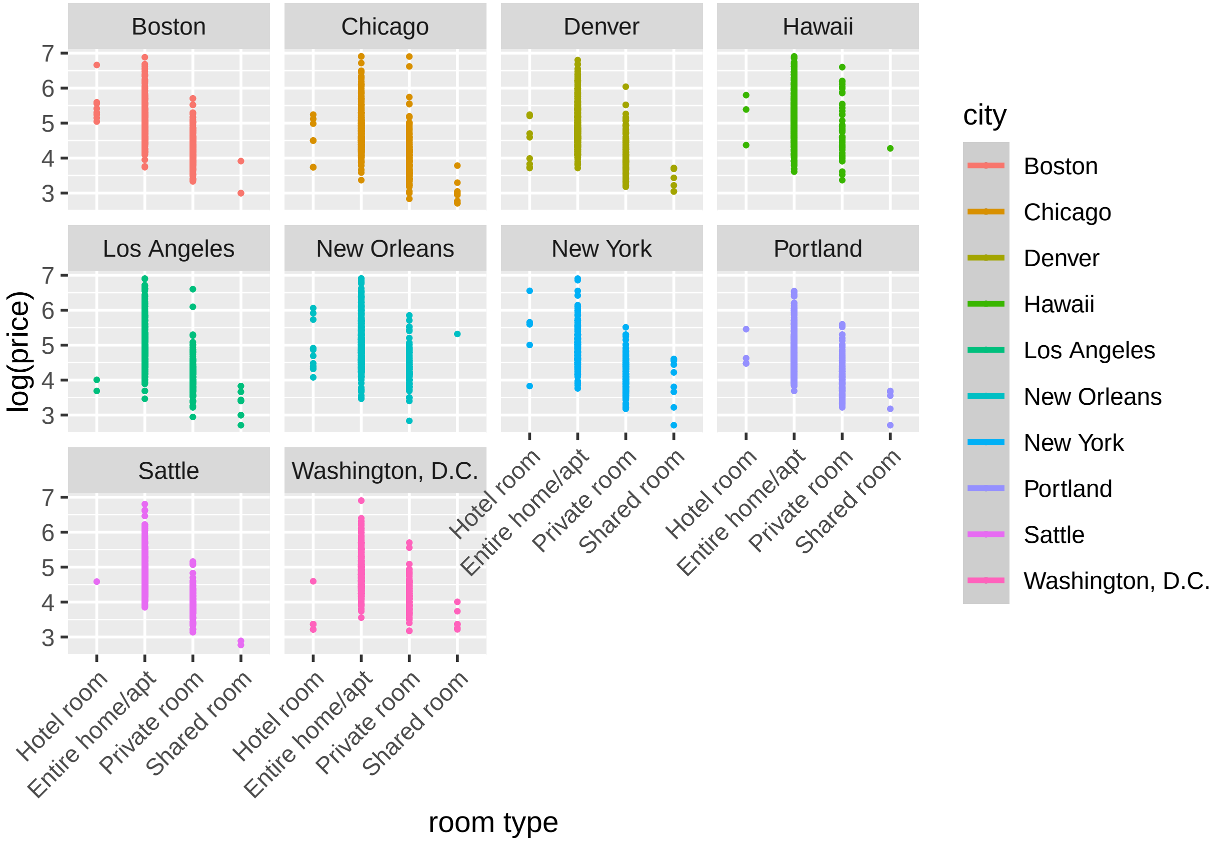


5. The plot of days available for a year vs Log(price) for each city

```
## `geom_smooth()` using formula 'y ~ x'
```

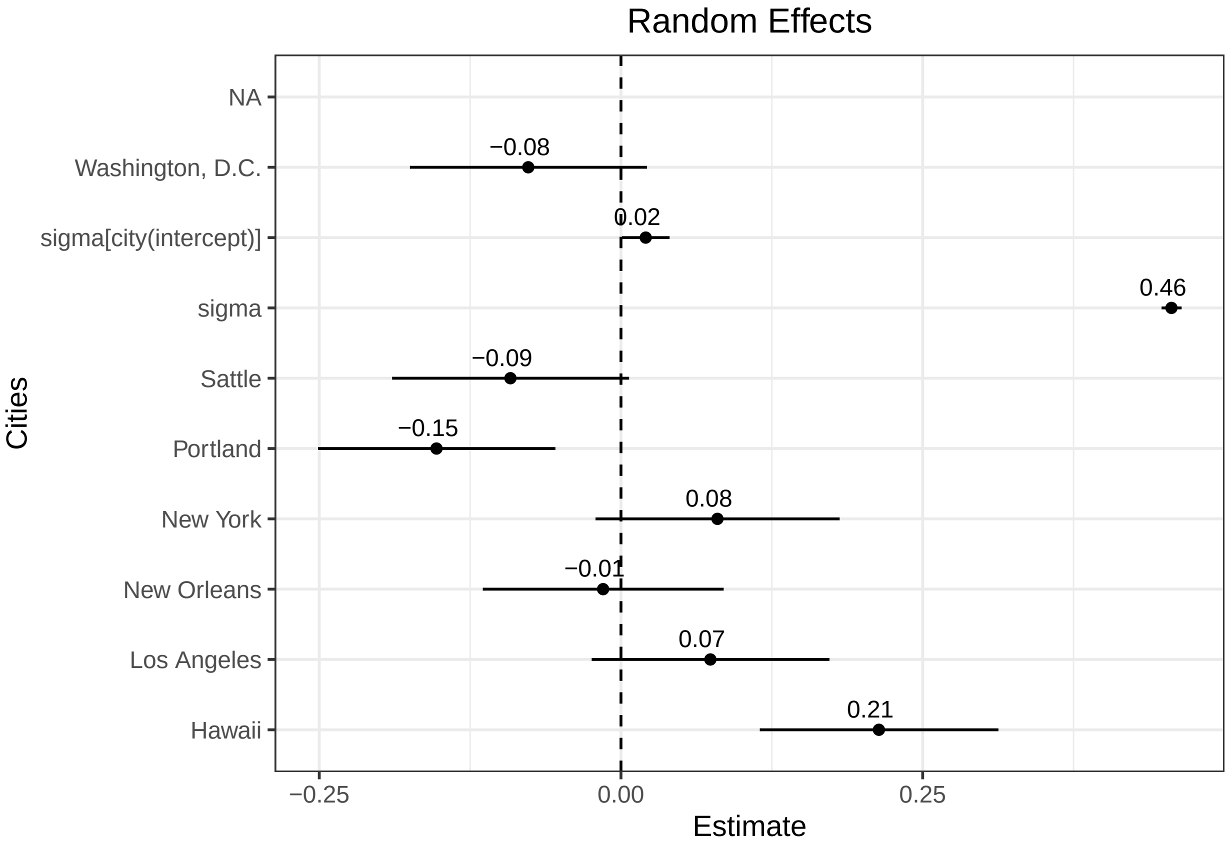


6. The plot of room type vs Log(price) for each city



```
## NULL
```

7. Random effects


Random Effects
Cities
Estimate
NA
Washington, D.C.
sigma[city(intercept)]
sigma
Sattle
Portland
New York
New Orleans
Los Angeles
Hawaii