Factors Affecting the Price on Airbnb

Yu Du

11/30/2020

Abstract

As Airbnb becomes popular and an increasing number of hosts register on Airbnb, people who love traveling and prefer to reserve accommodations on Airbnb become to have more choices. The booking price is a major concern when people reserve accommodations. I am concerned with availability in one year, minimum nights' stays, and reviews per month for accommodations, and I developed one linear regression model and two multilevel linear mixed effect models to assess how these variables' effects on price. From the models' results, these variables seem not to have effects on price.

Introduction

Airbnb is a known online rental marketplace for people. As a faithful customer of Airbnb, I am concerned with what factors may affect price. In this report, I will analyze how availability in one year, minimum nights' stays, and reviews per month may affect price. I will start with a linear regression model and then use multilevel linear mixed effects models to investigate the relationship between these three variables and price as accounting for neighbourhoods and room types.

Methods & Results

Data

The data is a combined dataset of three datasets downloaded from the Insideairbnb website. The three datasets are summary information and metrics for listings in Hong Kong on May, June and October in 2020. No summary information on July, August and September. I cleaned the data using the tidyverse package, and subset the data from booking prices between 0 and 800 Hong Kong dollars and no missing value in columns and rows. The data contains summary information for accommodations in 17 neighbourhoods, including price, availability in one year, room type, minimum nights' stays and reviews per month.

Exploratory Data Analysis

Figure 1. The boxplot of price by neighbourhood.

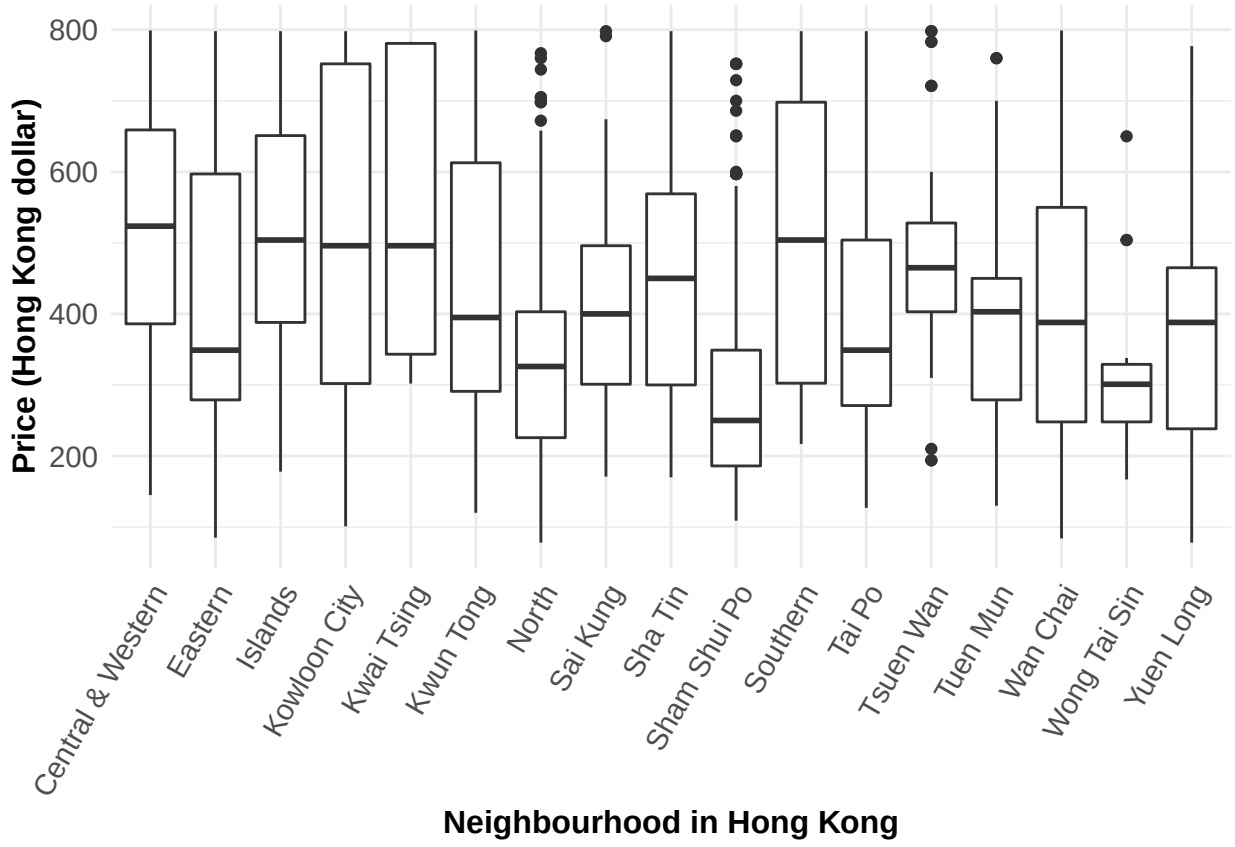


Figure 1 shows that six neighbourhoods have outliers in price.

Figure 2. The boxplot of price by room type.

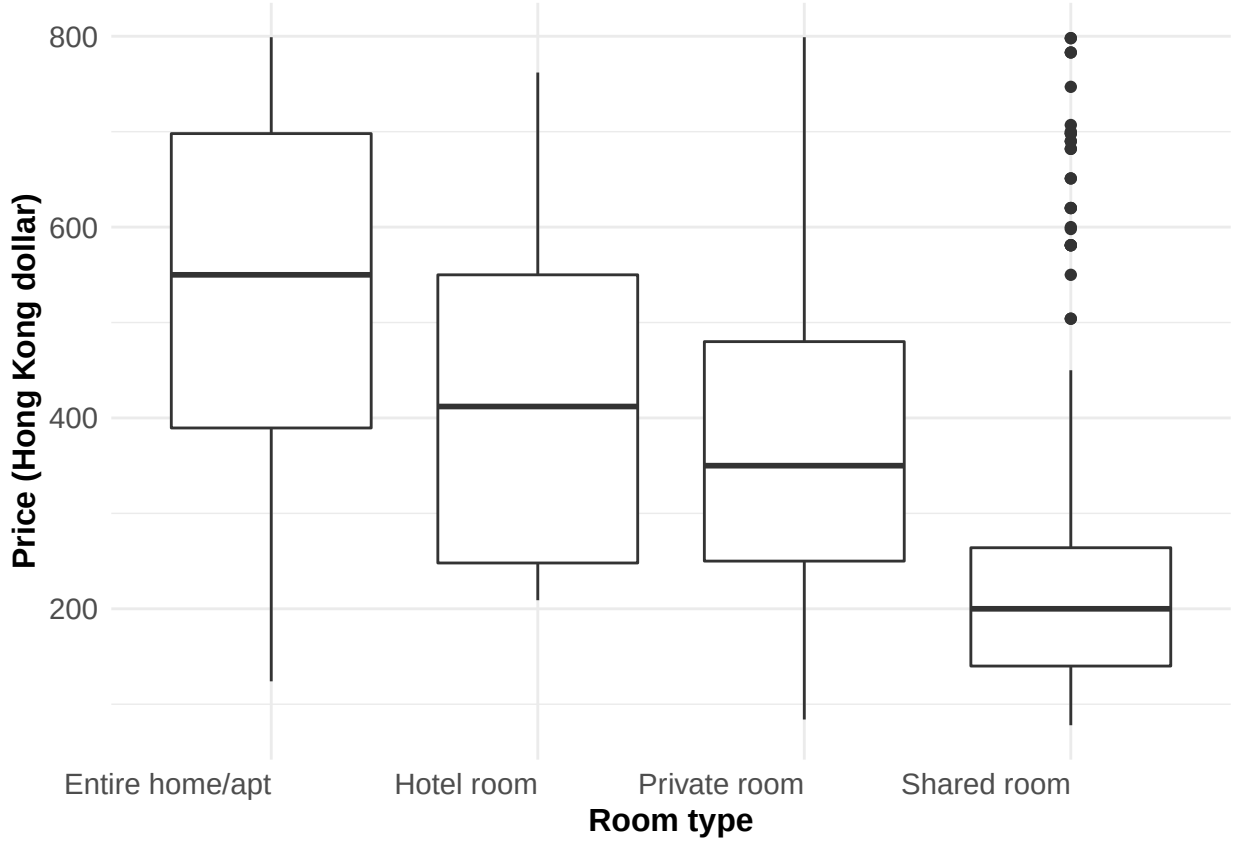


Figure 2 shows that one room type has outliers in price.

Models

Model One. Linear regression model with outcome is price and predictors are availability in one year, minimum nights' stays, and reviews per month.

The variable reviews per month has a positive effect on price, 1 unit increase in reviews per month leads to 9.5 units increase in price holding other two variables at constant. The variables minimum nights' stays and availability in one year have a little negative effect on price. The intercept is large as 473.

In order to account for the correlations of prices on the same neighbourhood and on the same room type, multilevel linear mixed effects models are developed by using the lme4 package.

Model Two. Multilevel linear mixed effects model: expand the Model One by allowing varying intercepts across neighbourhoods and room types.

The negative effect of availability in one year on price shifts to positive but this variable seems not to have effect on price, and the negative effect of minimum nights' stays on price becomes bigger. The 1 unit increase in minimum nights' stays leads to 1.2 units decreases in price. The positive effect of reviews per month on price shifts to negative, 1 unit increase in reviews per month leads to 0.8 units decrease in price.

Figure 3. Residual plot and Normal Q-Q plot for Model Two.

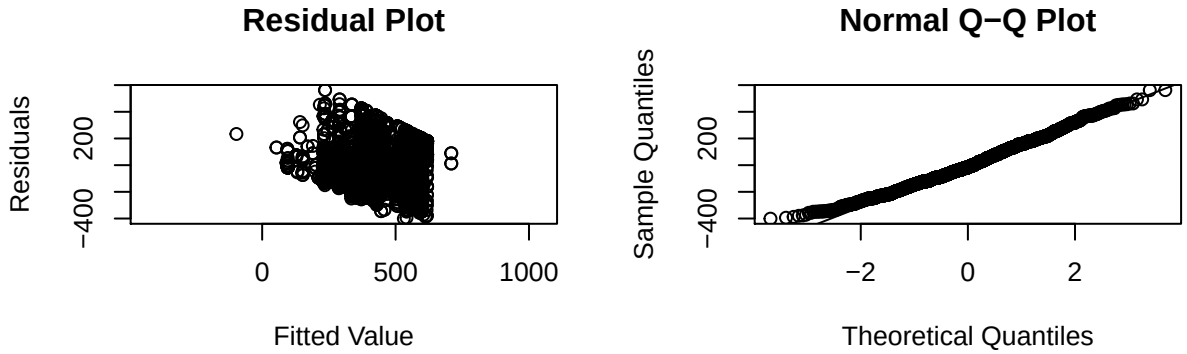


Figure 3 shows that most of the residuals centered at 0 but a few outliers exist in the residual plot. Both the ends of points in the Normal Q-Q plot deviate from the straight line and the center follows the straight line.

Model Three. Multilevel linear mixed effects model: taking log transformation of outcome in the Model Two.

The three variables do not have effects on the log transformation of price.

Figure 4. Residual plot and Normal Q-Q plot for Model Three.

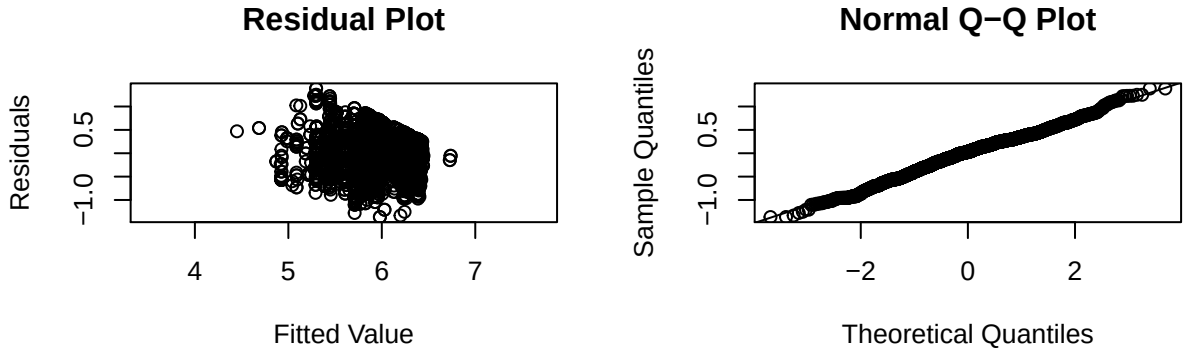


Figure 4 shows that no obvious difference between plots from previous figure. The upper end of points in the Normal Q-Q plot does not deviate from the line but the bottom end still deviates from the line.

The AIC and BIC for Model One and Model Two are over 50000. The AIC and BIC of Model Three are 4204 and 4249.

The AIC and BIC of Model Three are much lower than the previous two models. Model Three is the preferred one.

Discussion

From the results, the effect of availability in one year, the effect of minimum nights' stays, and the effect of reviews per month on price are slightly small. As I improve the model from the linear regression model to the second multilevel model, its predictors' effects become much smaller. In conclusion, the availability in one year, minimum nights' stays, and reviews per month seem not to have effects on price in Hong Kong.

Limitations

One limitation is all variables in the data which I am concerned with and can be analyzed as factors are not seem to be important for price. The second limitation is the difference between the counts of neighbourhoods is large. The third limitation is the price in the data varies a lot from May to October. This situation happens might be because of COVID-19, the highest price is almost 80000 Hong Kong dollars. The data I use in this report is a subset from prices between 0 to 800 Hong Kong dollars.

Future Direction

From a customer's aspect, I can choose a larger dataset that contains more variables I can include as factors for analysis. Another future work I can do is to analyze the data from the past years to find if availability in one year, minimum nights' stays, and reviews per month still seem not to have effects on price in the past years without the COVID-19 situation.

Reference

Summary information and metrics for listings in Hong Kong, http://insideairbnb.com/get-the-data.html

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5

Appendix

Figure 5. The distribution of neighbourhood.

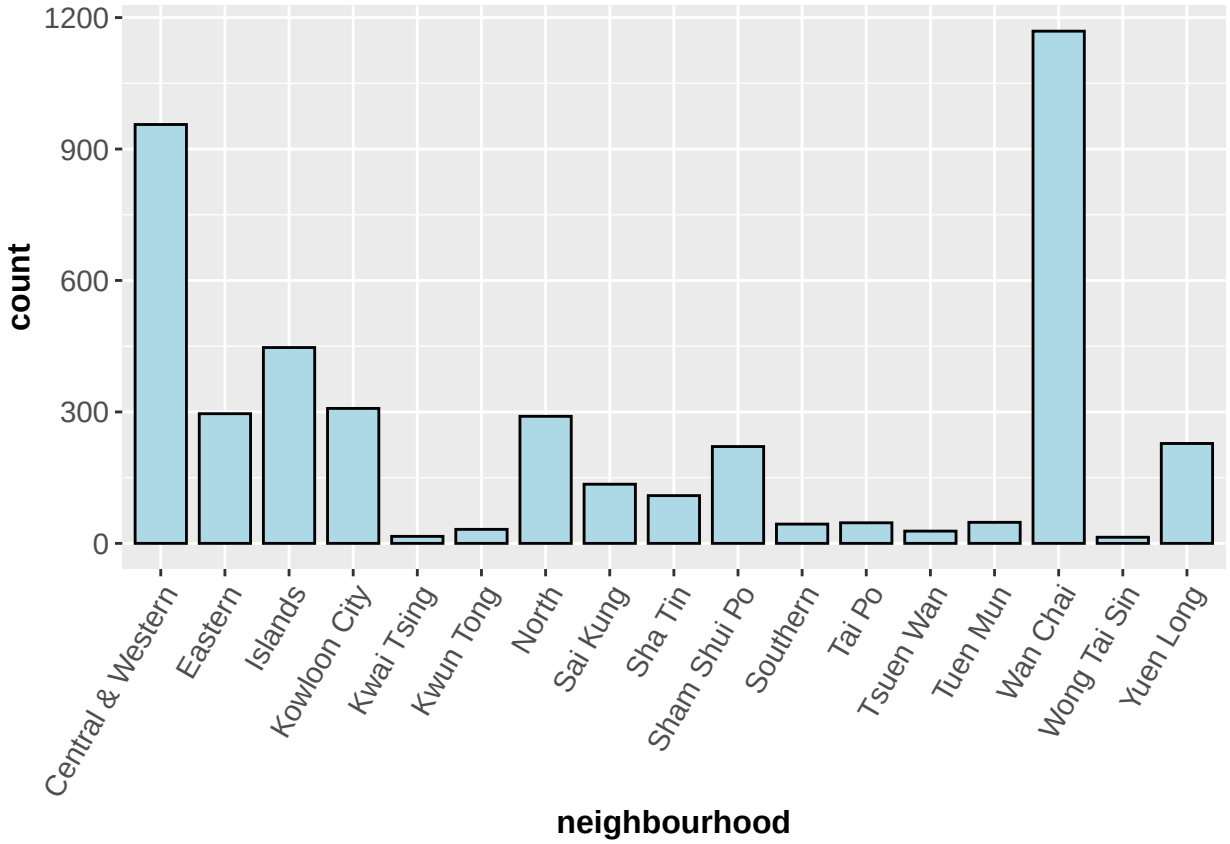


Figure 6. The distribution of room type.

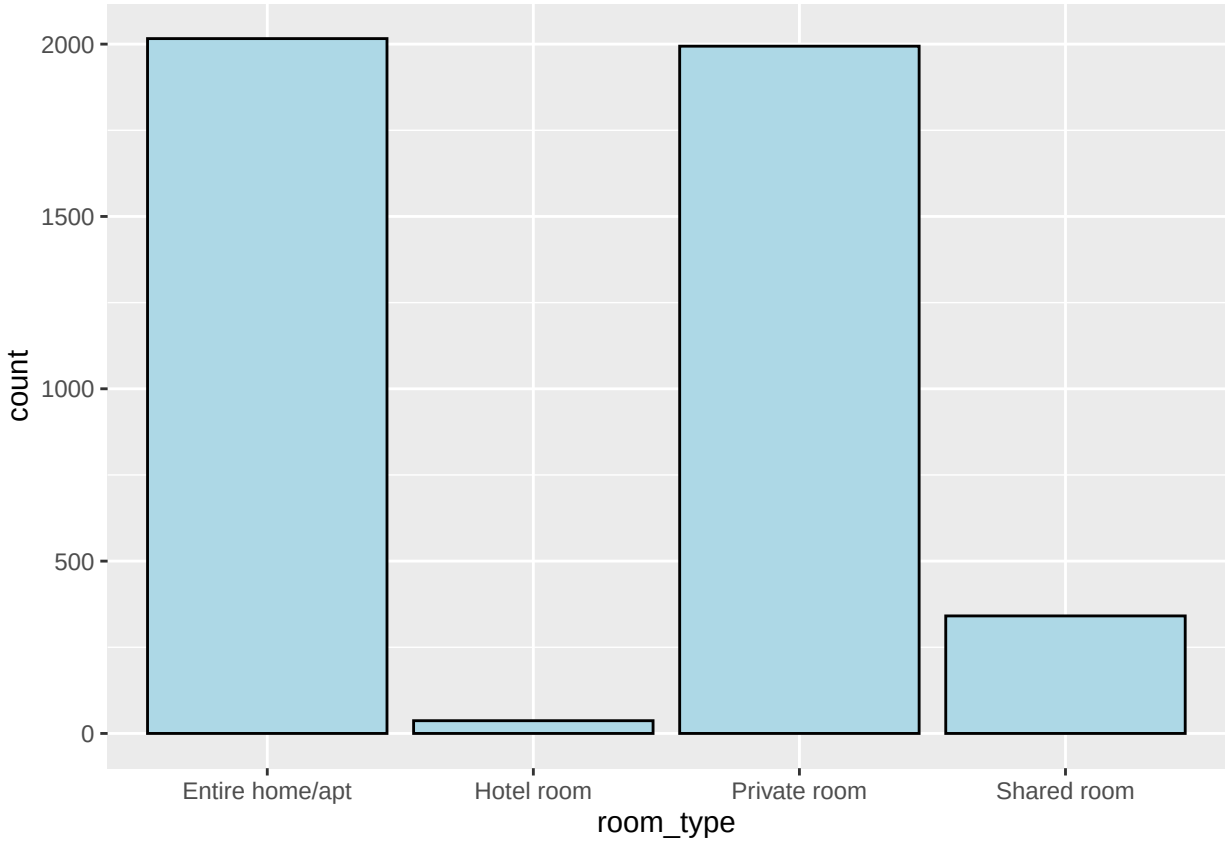


Figure 7. The relationships between price, reviews per month, and neighbourhood.

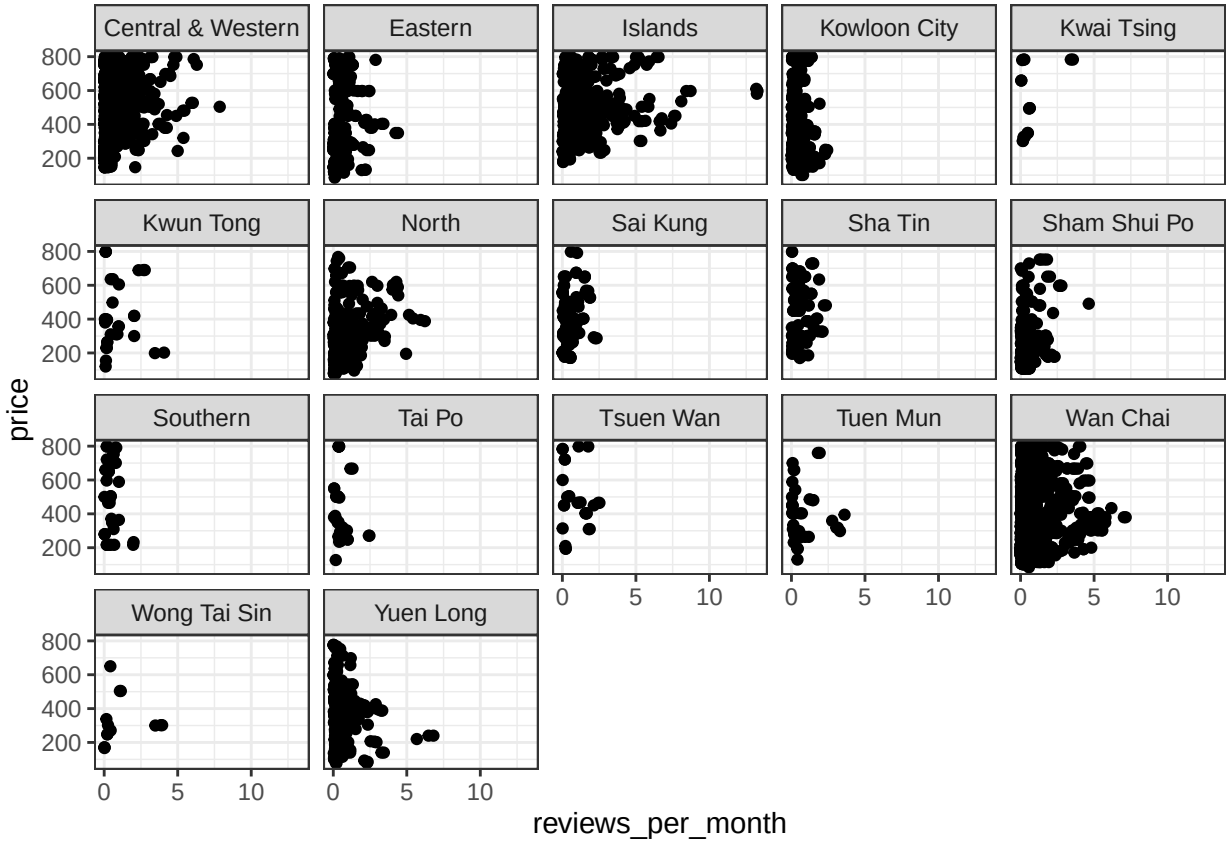


Figure 8. The relationships between price, reviews per month, and room type.

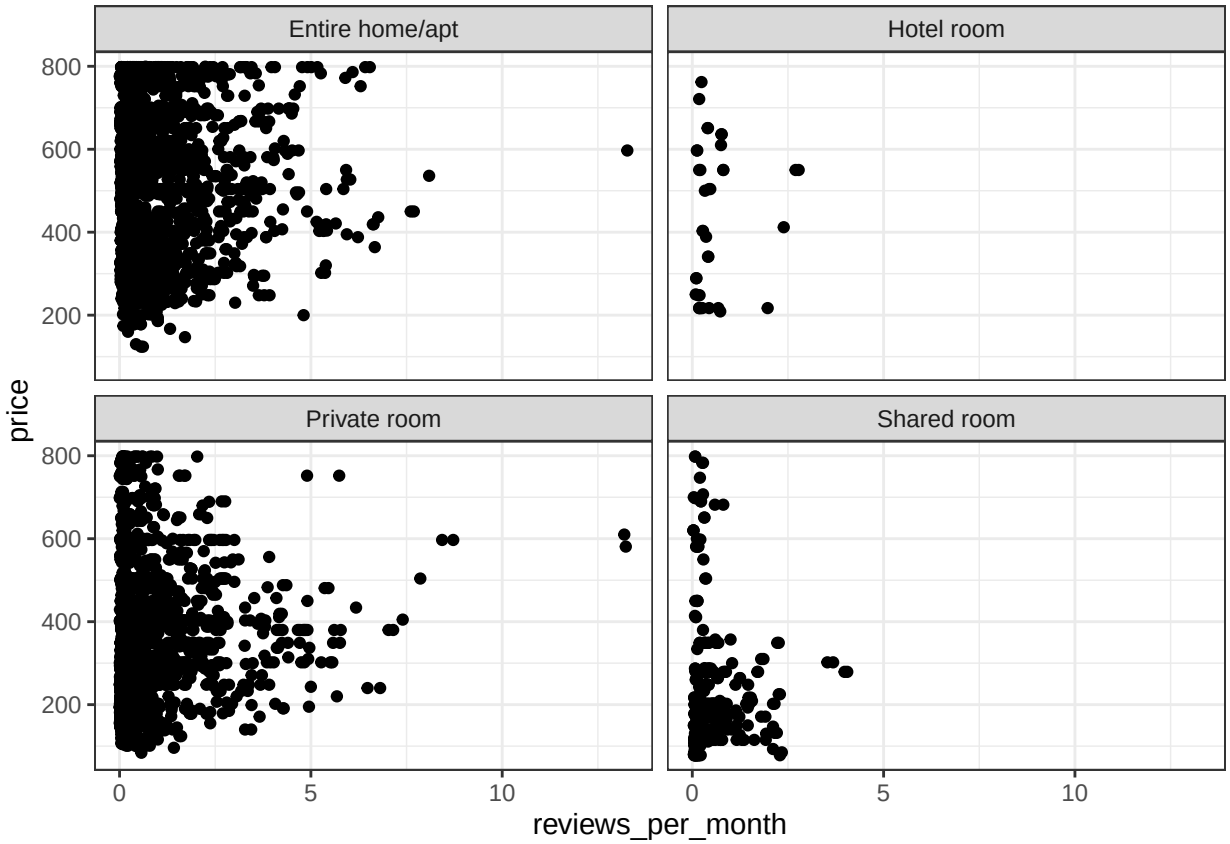


Figure 9. The coefficient plot for Model One.

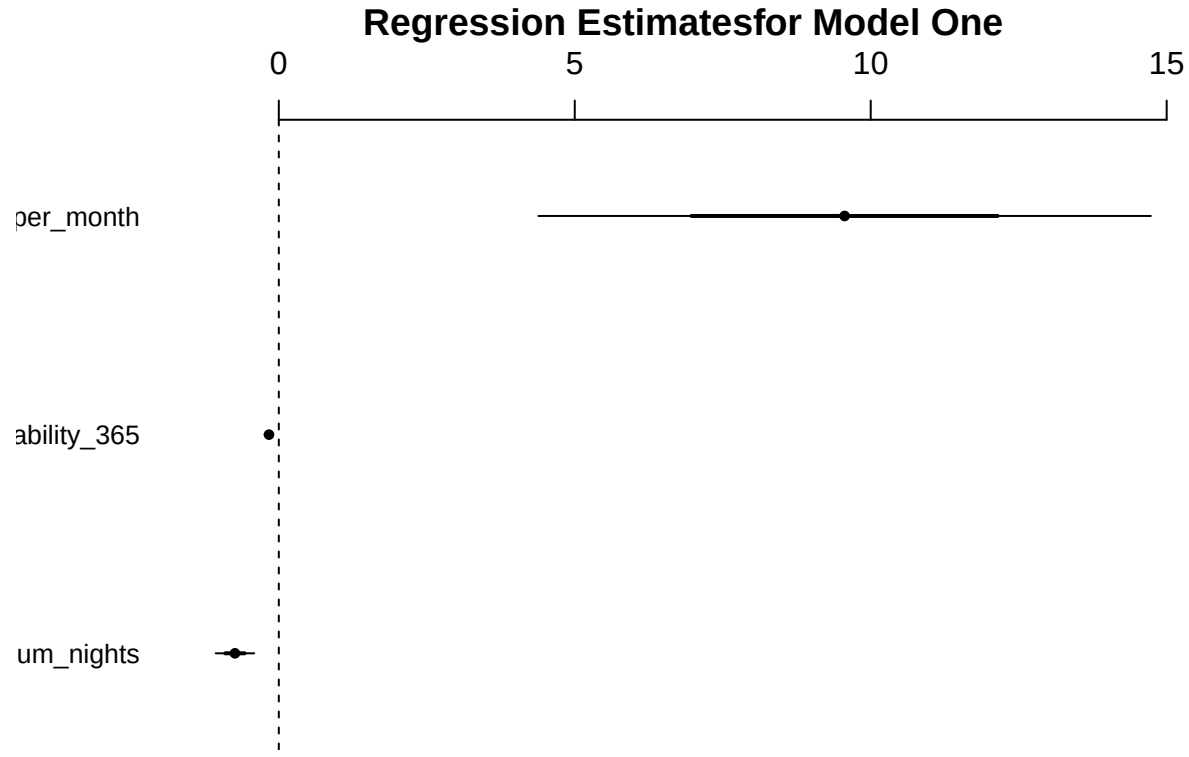


Figure 10. The residual plot and normal Q-Q plot for Model One.

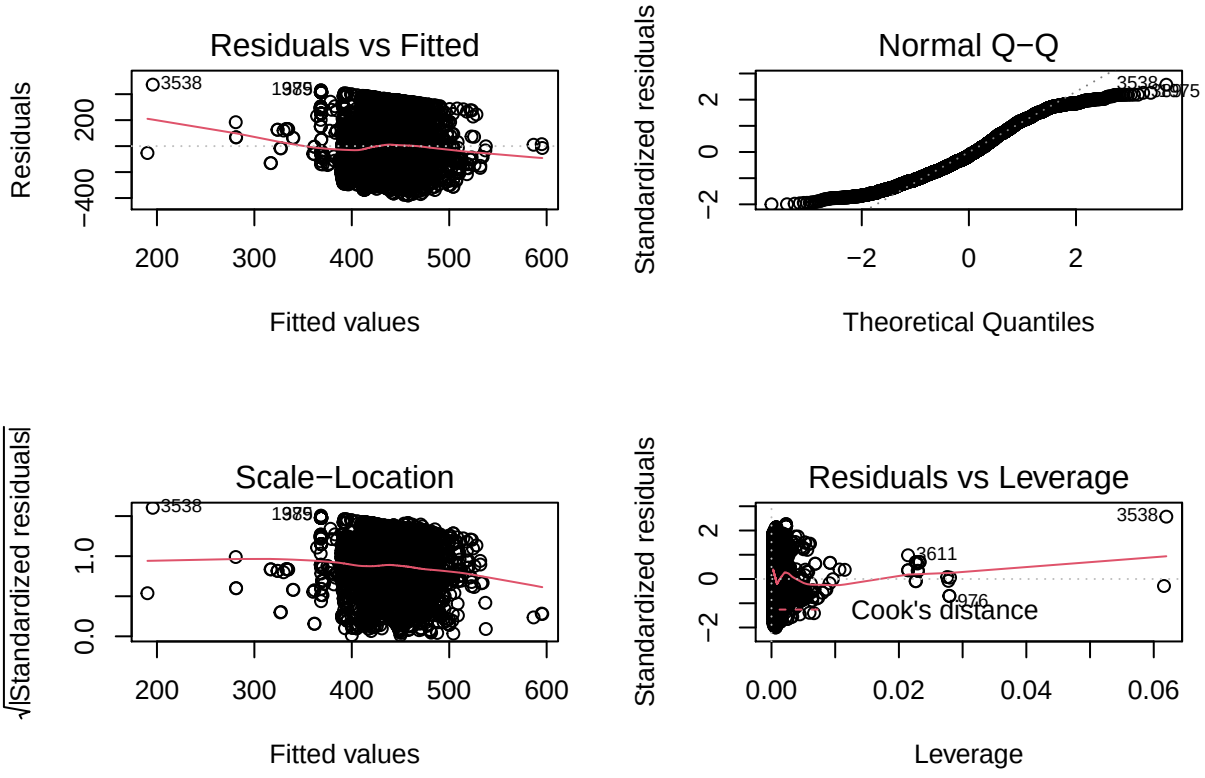
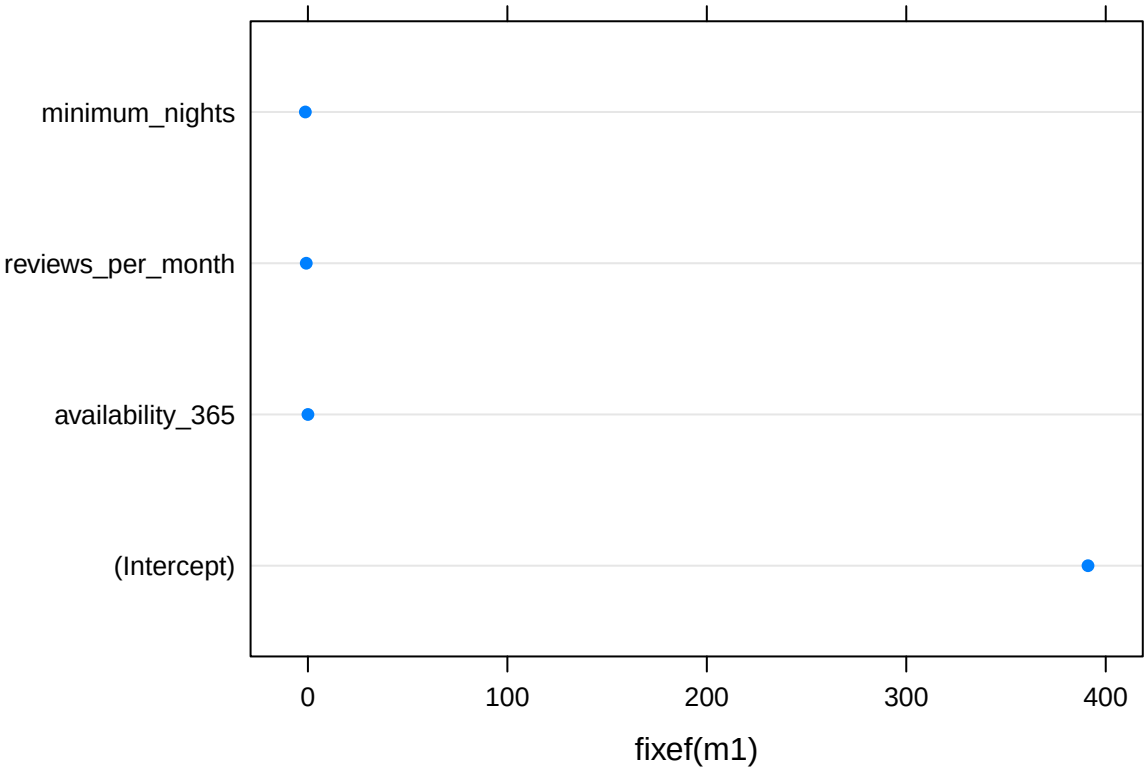
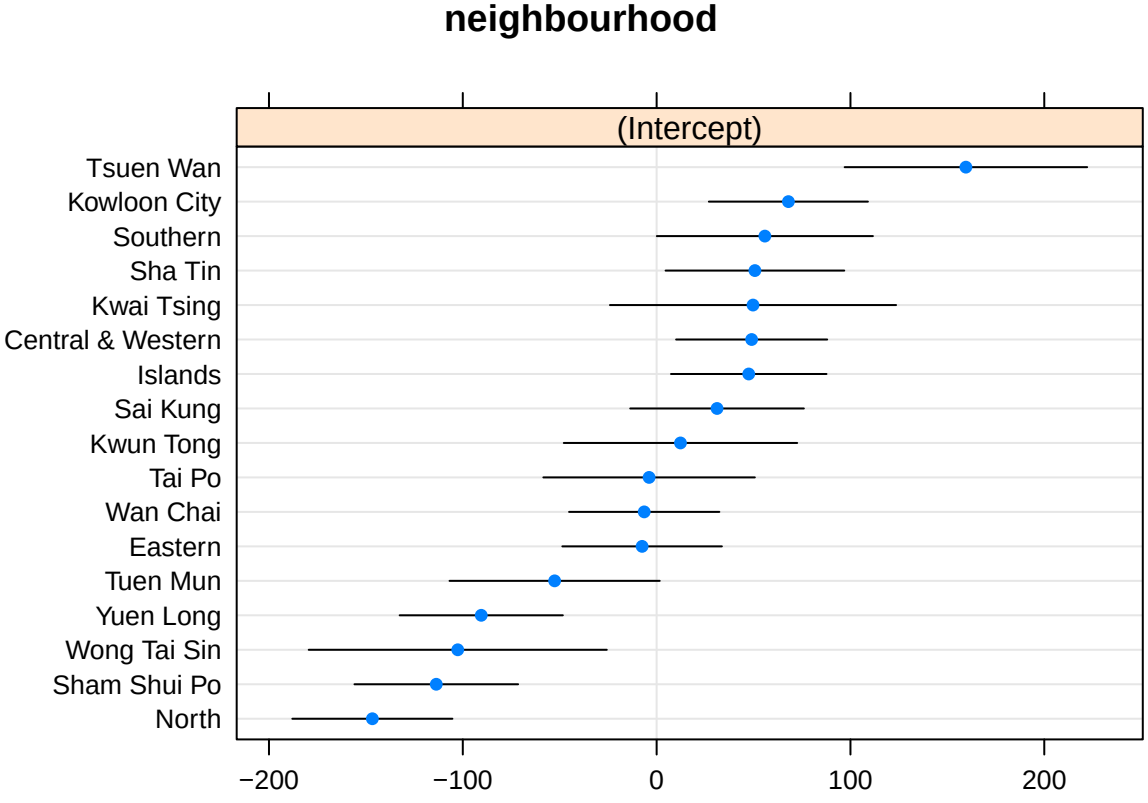


Figure 11. The plots for fixed effects and random effects of Model Two.




```
## $neighbourhood
```




```
##
```

```
## $room_type
```

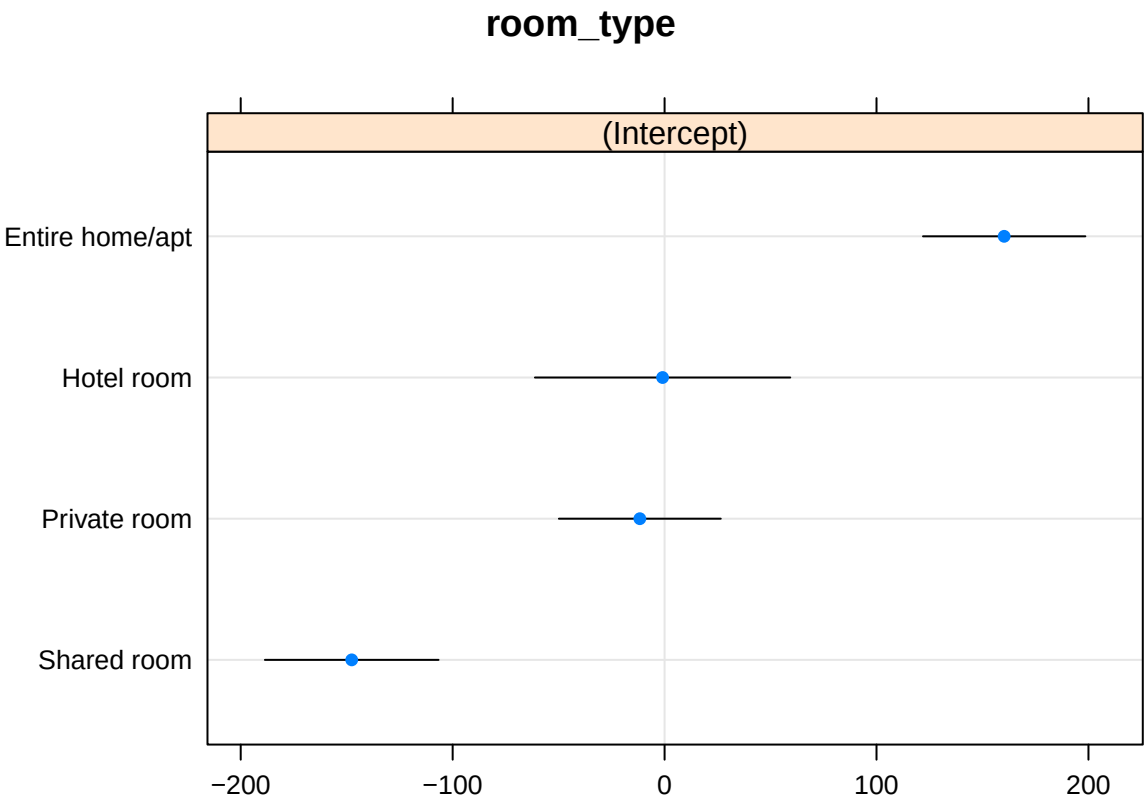
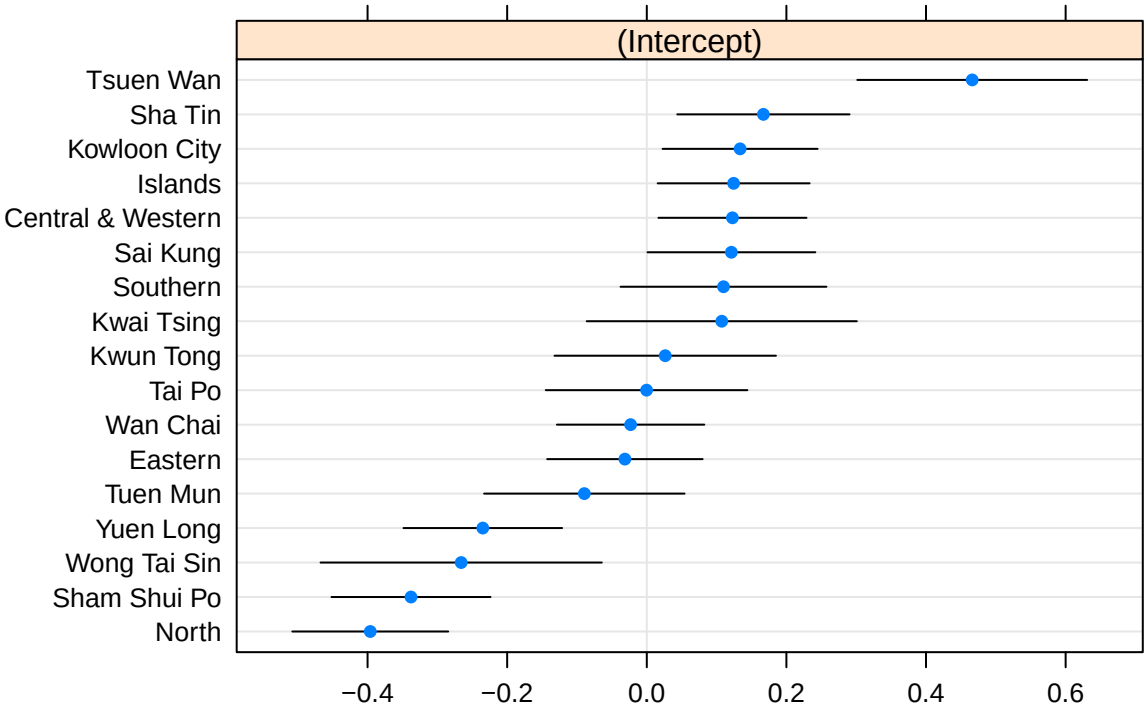


Figure 12. The plots for random effects of Model Three.

```
## $neighbourhood
```


neighbourhood
(Intercept)
Tsuen Wan
Sha Tin
Kowloon City
Islands
Central & Western
Sai Kung
Southern
Kwai Tsing
Kwun Tong
Tai Po
Wan Chai
Eastern
Tuen Mun
Yuen Long
Wong Tai Sin
Sham Shui Po
North
−0.4
−0.2
0.0
0.2
0.4
0.6

```
##
## $room_type
```

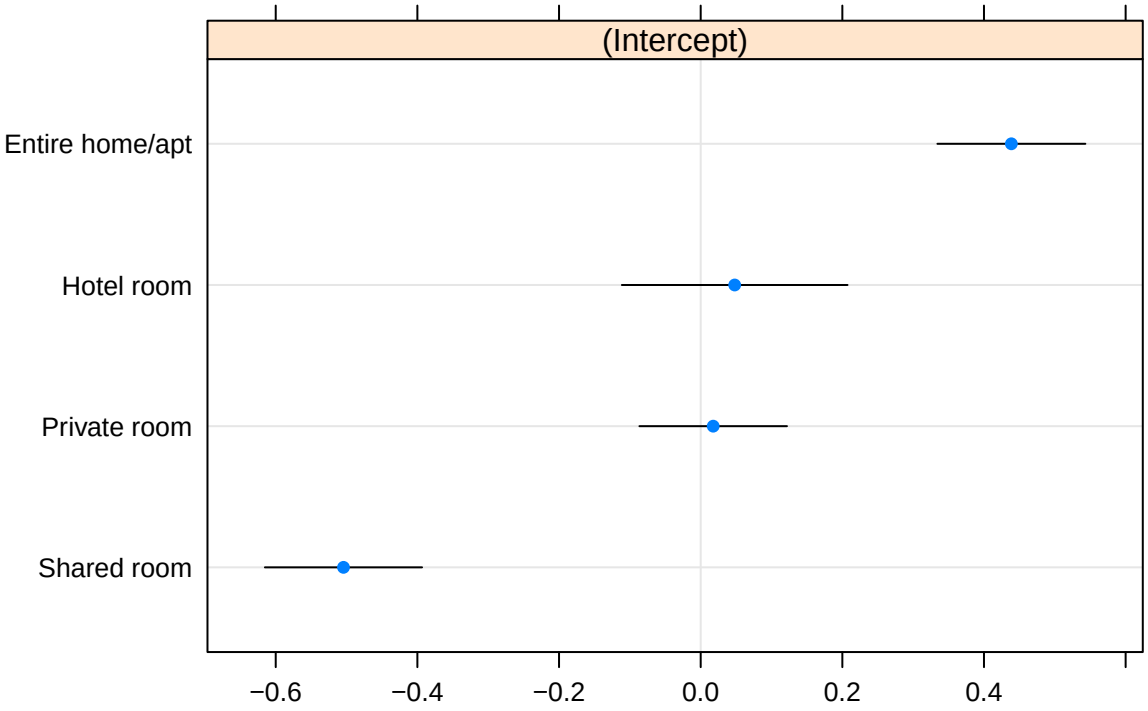
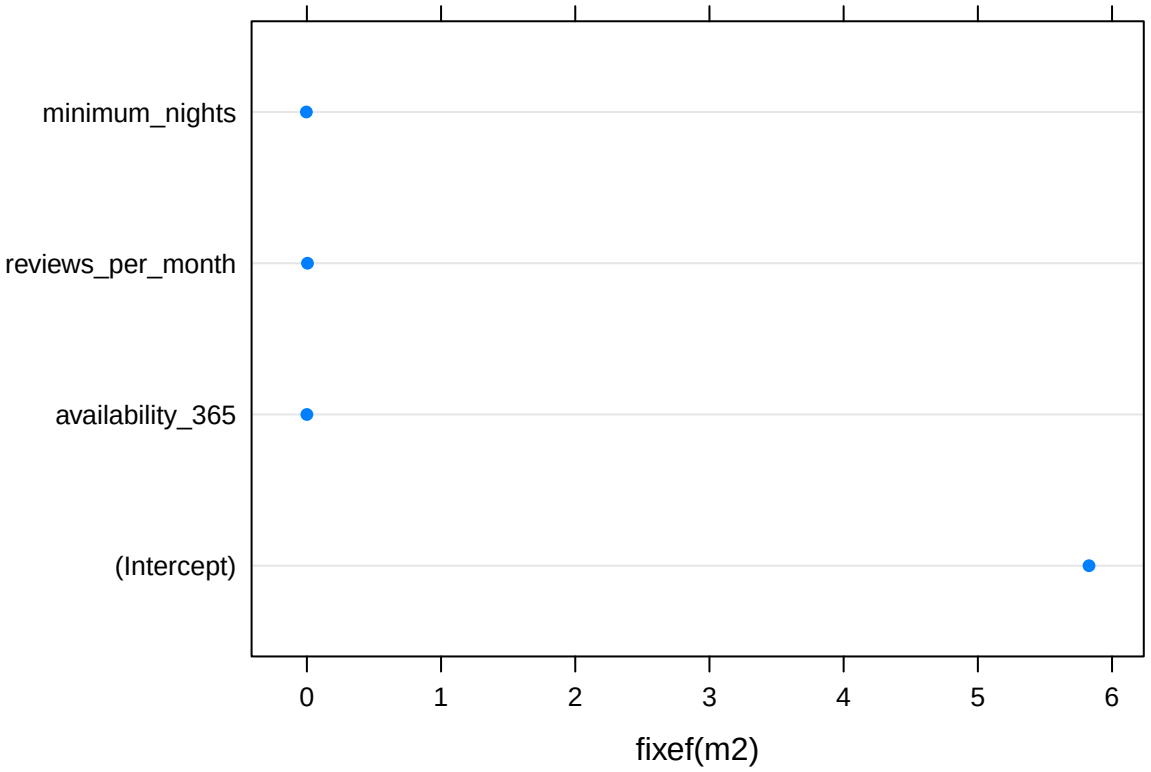

room_type
(Intercept)
Entire home/apt
Hotel room
Private room
Shared room
−0.6
−0.4
−0.2
0.0
0.2
0.4

Figure 13. The plot of fixed effects of Model Three.



AIC & BIC for models:

```
##    df       AIC
## m0  5 58513.824
## m1  7 56526.386
## m2  7  4203.834
```

```
##    df      BIC
## m0  5 58545.76
## m1  7 56571.09
## m2  7  4248.54
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: airbnb
## Models:
## m1: price ~ availability_365 + reviews_per_month + minimum_nights +
## m1:     (1 | neighbourhood) + (1 | room_type)
## m2: log_price ~ availability_365 + reviews_per_month + minimum_nights +
## m2:     (1 | neighbourhood) + (1 | room_type)
##    npar   AIC   BIC   logLik deviance Chisq Df Pr(>Chisq)
## m1    7 56531 56576 -28258.7    56517
## m2    7  4162  4206  -2073.8     4148 52370  0  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

m0: Model One m1: Model Two m2: Model Three

Residual plots and Normal Q-Q plots for Model Two and Model Three as shown in the report. (clearer version)