Final Report for Airbnb in London

Kecheng Liang

Dec 5, 2018

a.Abstract

This analysis is focus on the Airbnb rating. I choose London as my target city and the dataset is from http://tomslee.net/airbnb-data-collection-get-the-data. I mainly focus on the variables such as room type, number of reviews, price, number of bedrooms, and accommodates. There are some graphs I made to help me interpret the analysis. Finally, I made five models to find some useful information.

b.Introduction

i.Background

Nowadays, more and more people choose Airbnb when having a vocation in another city. Because of its low price and convenient location, Airbnb becomes more popular than before. London, one of the most attractive and popular cities in Europe, has a very high number of Airbnb hosts.

ii.Previous work

Airbnb is a large company and there are lots of related datasets on the website which is already organized well. I found a clearly organized dataset on the website with several variables such as ratings, price and so on.

c.Method

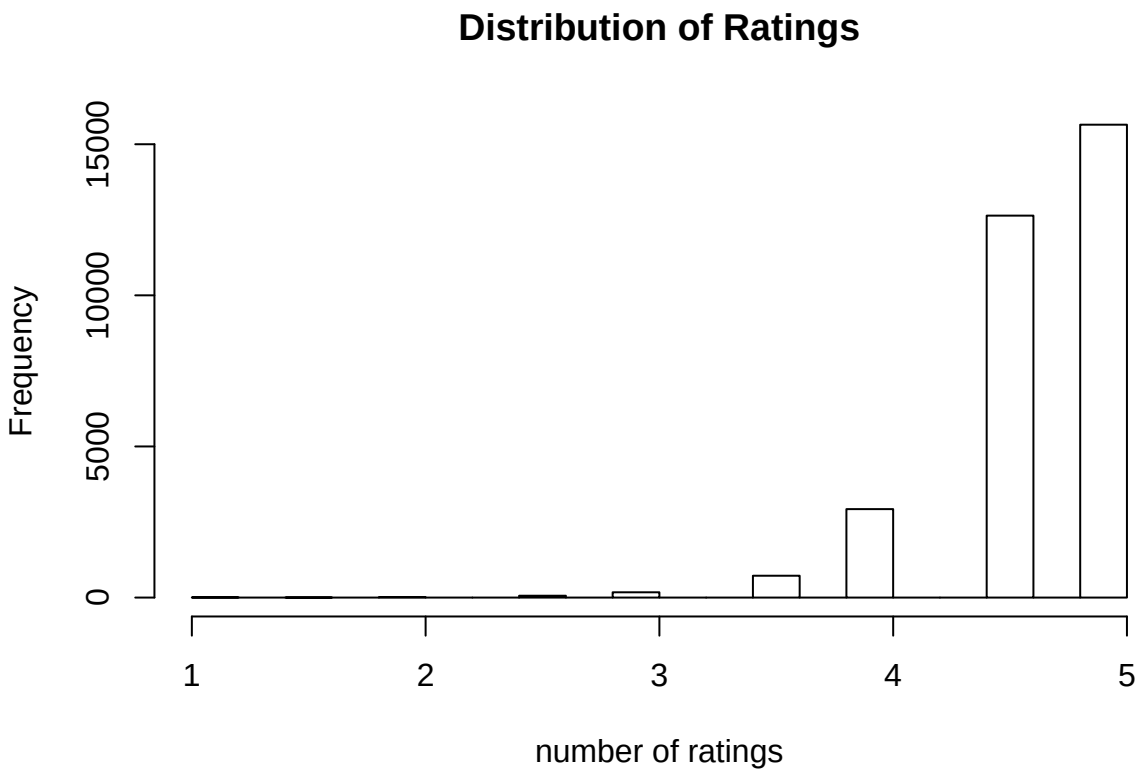
i.Data source

Data source for London with 64144 properties as of July 28, 2017 is collected from the public Airbnb website. I exclude 31947 properties with no reviews or rating score is zero.

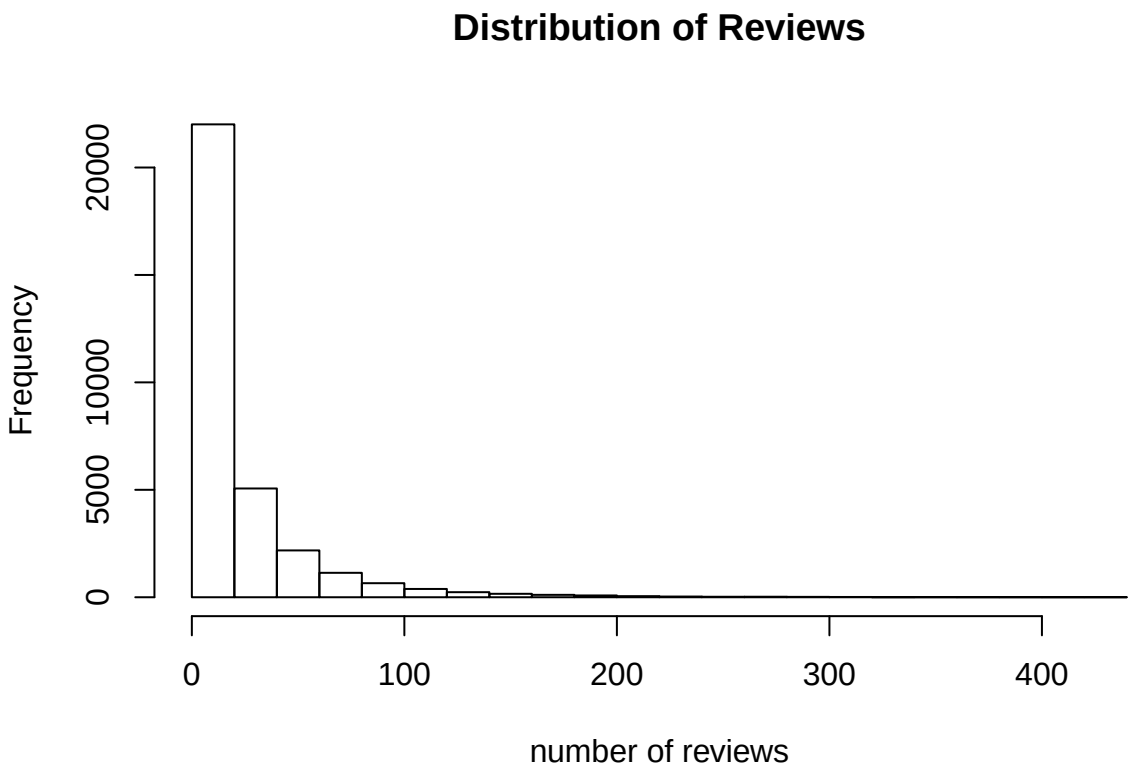
Table 1: Variables explanation

Variables	Explanation
Room_id	A unique number for an Airbnb property
Host_id	A unique number for an Airbnb host
Room_type	Room type of the Airbnb property, one of shared room, entire room/apt, or private room
Borough	A town, or part of a large city
Reviews	The number of reviews that an Airbnb property received
Overall satisfaction	The average rating that an Airbnb property received(max value is five)
Accommodates	The number of visitors can live in an Airbnb property
Bedrooms	The number of bedrooms in an Airbnb property
Price	The price for a night stay
Latitude	The latitude of the Airbnb property
Longtitude	The longtitude of the Airbnb property

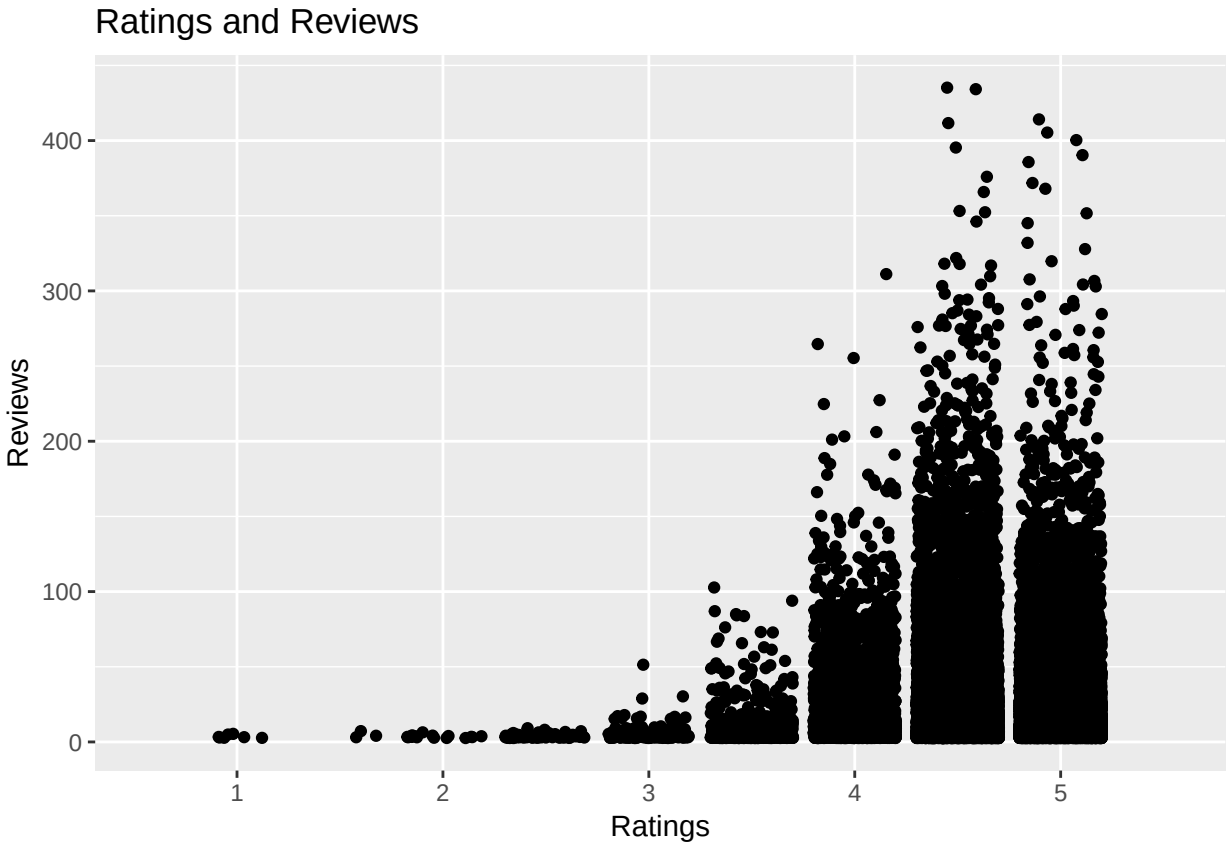
ii.Model used



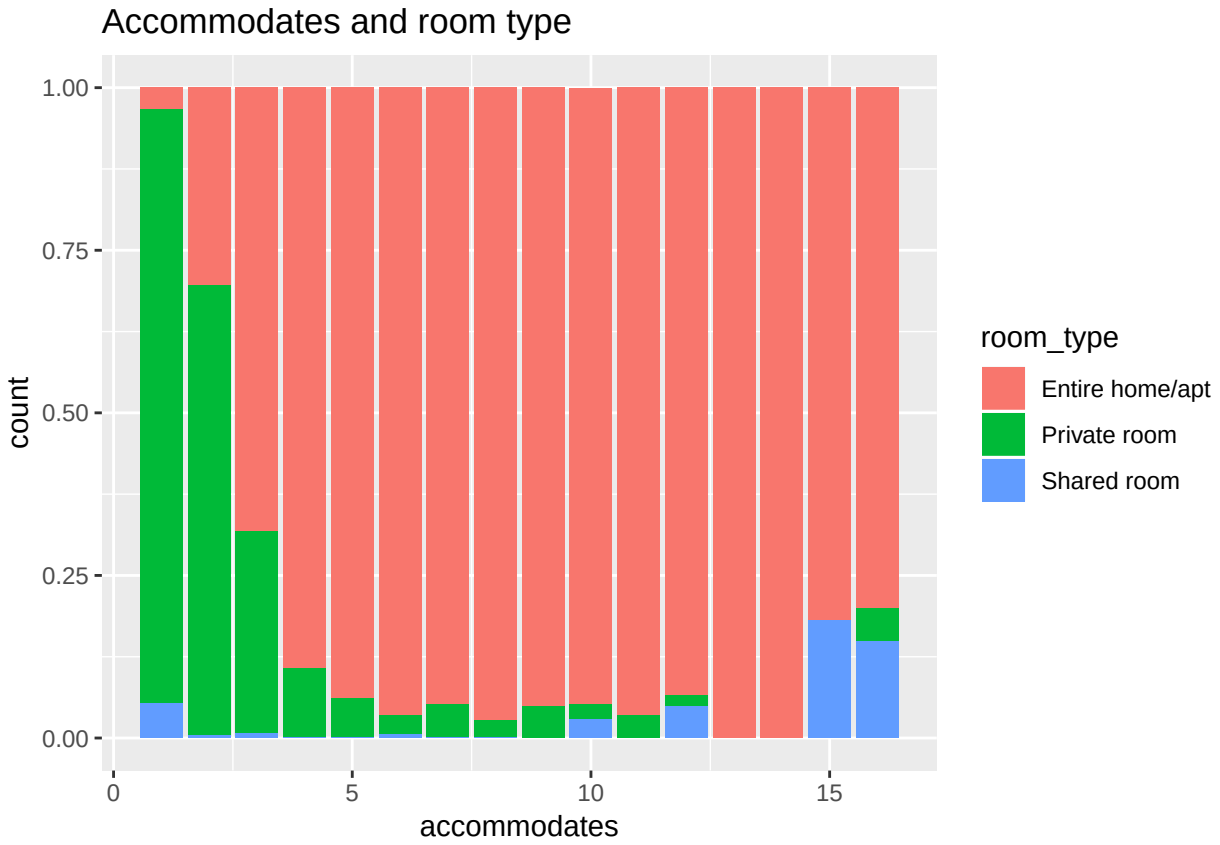
From the above graph, we can see that most of the Airbnb properties ratings have more than 4 points.



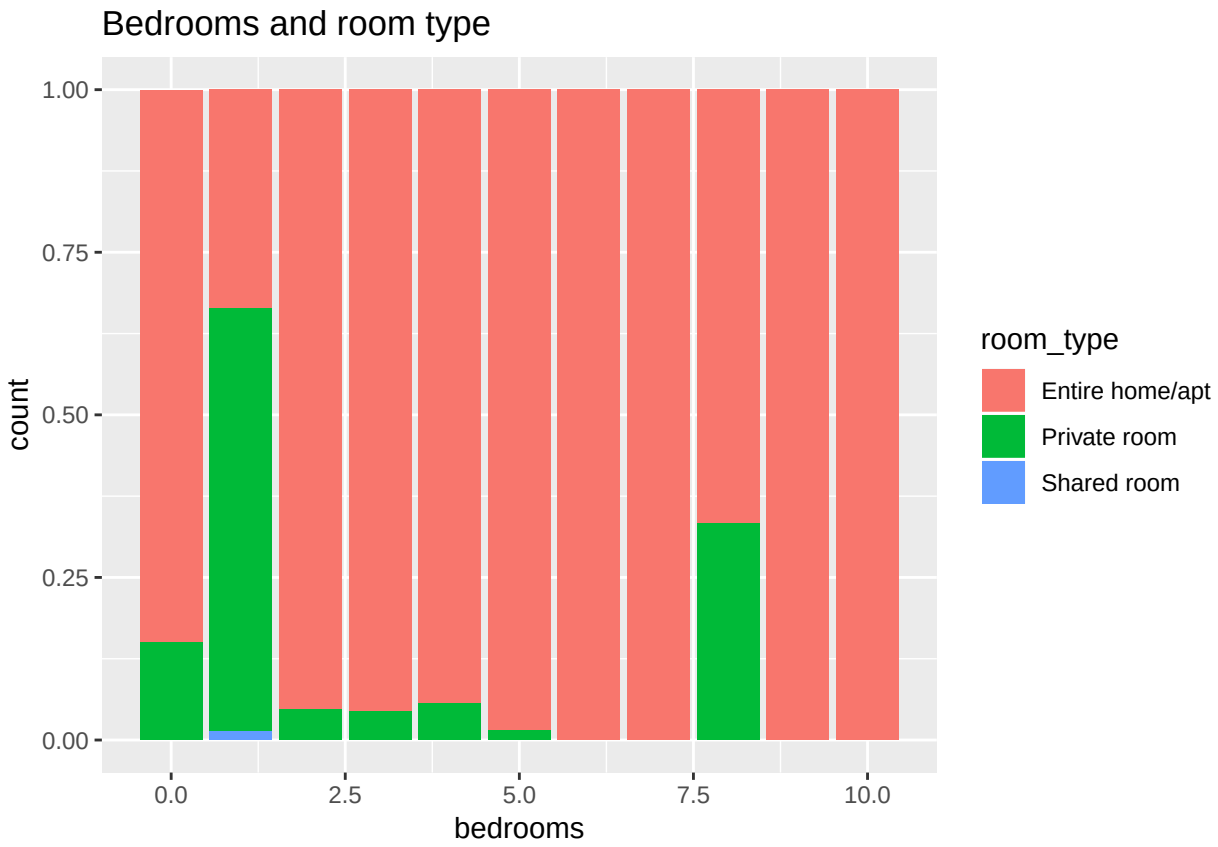
From the above graph, we can see that most of the Airbnb properties have less than 100 reviews.



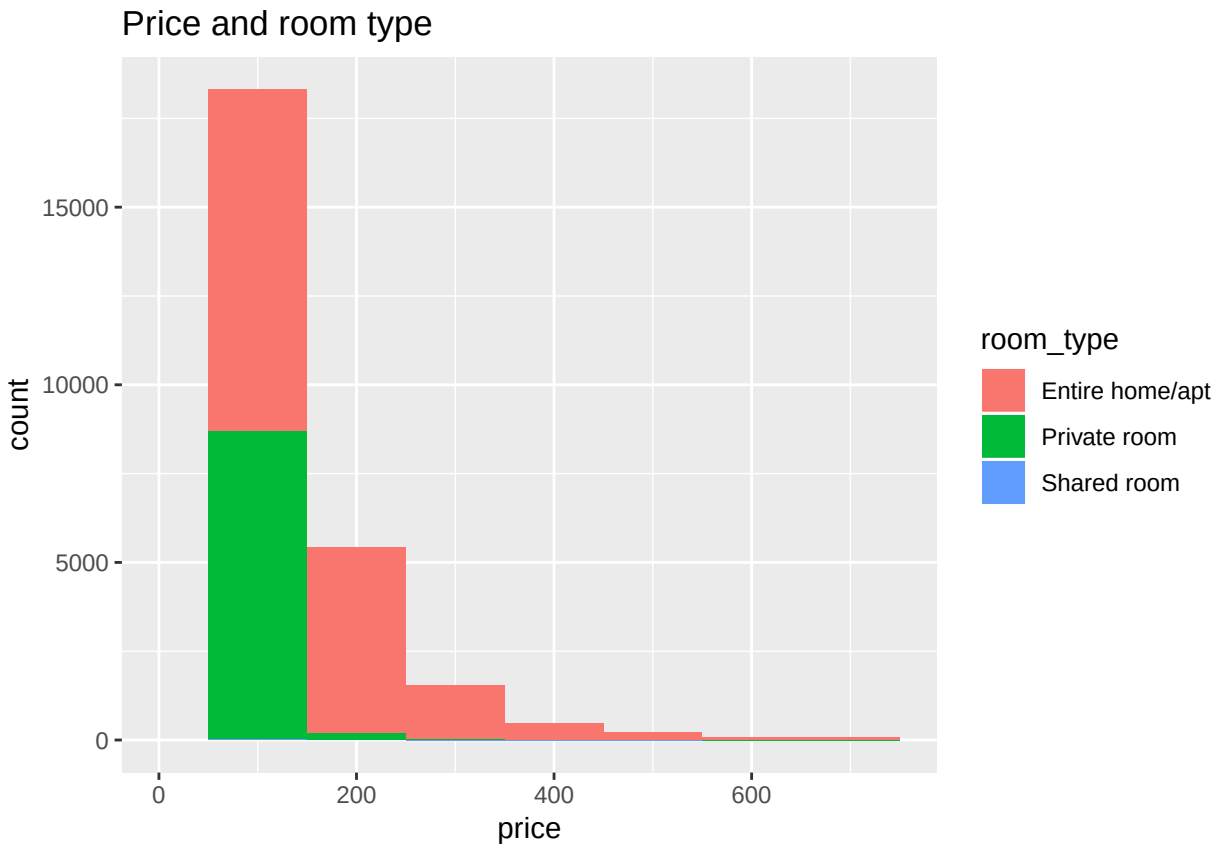
From the above graph, we can see that generally when the ratings are higher, there are more reviews for this Airbnb property.



From the above graph we can see that entire room tends to allow more accommodates compare to the private room.



From the graph above we can see that shared rooms have only one bedroom, and entire room usually have more rooms.



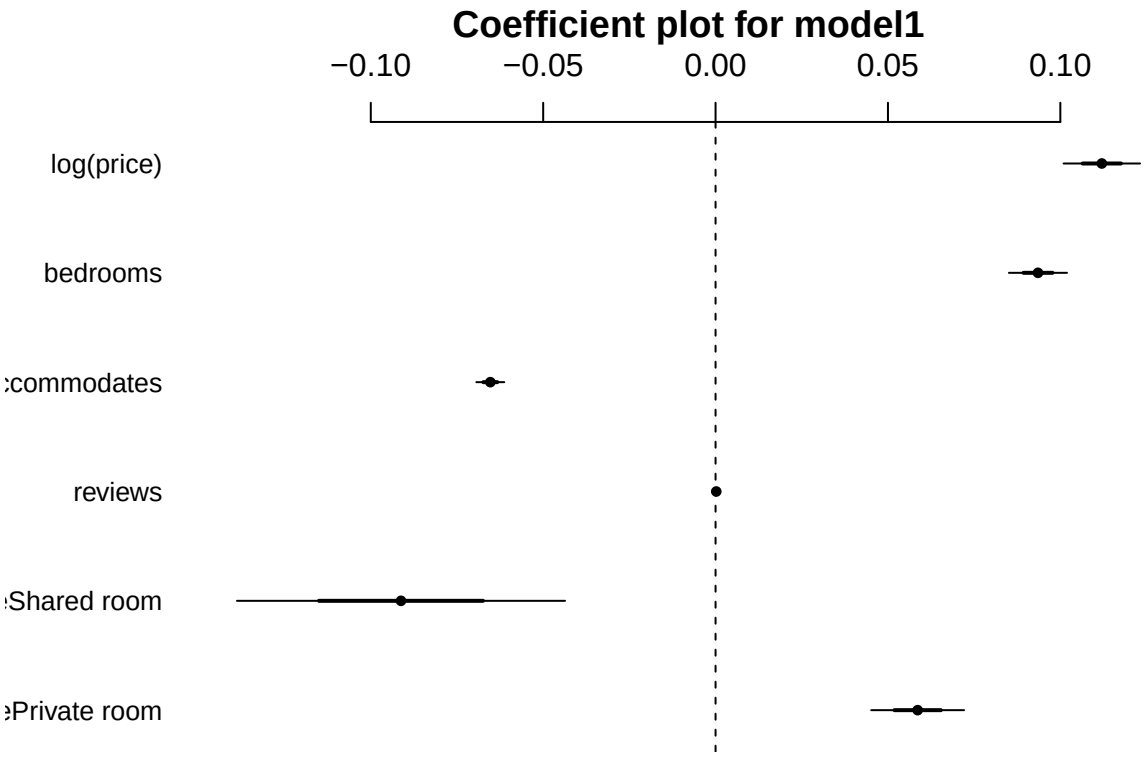
From the graph we can see that entire room are more likely have higher price.

```
#There are five models I use.
#model1 has no random effect and I use glm.
#model2 has no random effect, but I remove reviews which is not significant in the first model.
#model3 has random intercept.
#model4 has random slope.
#model5 has random intercept and slope.

#model1 no random effect
model1 <- glm(overall_satisfaction~room_type+reviews+accommodates+bedrooms+log(price),
              data = LondonAirbnb)
#remove reviews, no relationship

#model2 remove reviews
model2 <- glm(overall_satisfaction~factor(room_type)+accommodates+bedrooms+log(price),
              data = LondonAirbnb)

#random intercept borough
model3 <- lmer(overall_satisfaction~factor(room_type)+accommodates+bedrooms+log(price)+
                 (1|borough),data = LondonAirbnb)

#random slope
model4 <- lmer(overall_satisfaction~factor(room_type)+accommodates+bedrooms+log(price)+
                 (0+log(price)|borough),data = LondonAirbnb)

#random slope and intercept
```

```
model5 <- lmer(overall_satisfaction~factor(room_type)+accommodates+bedrooms+log(price)+(1+log(price)|bo
```

d.Result

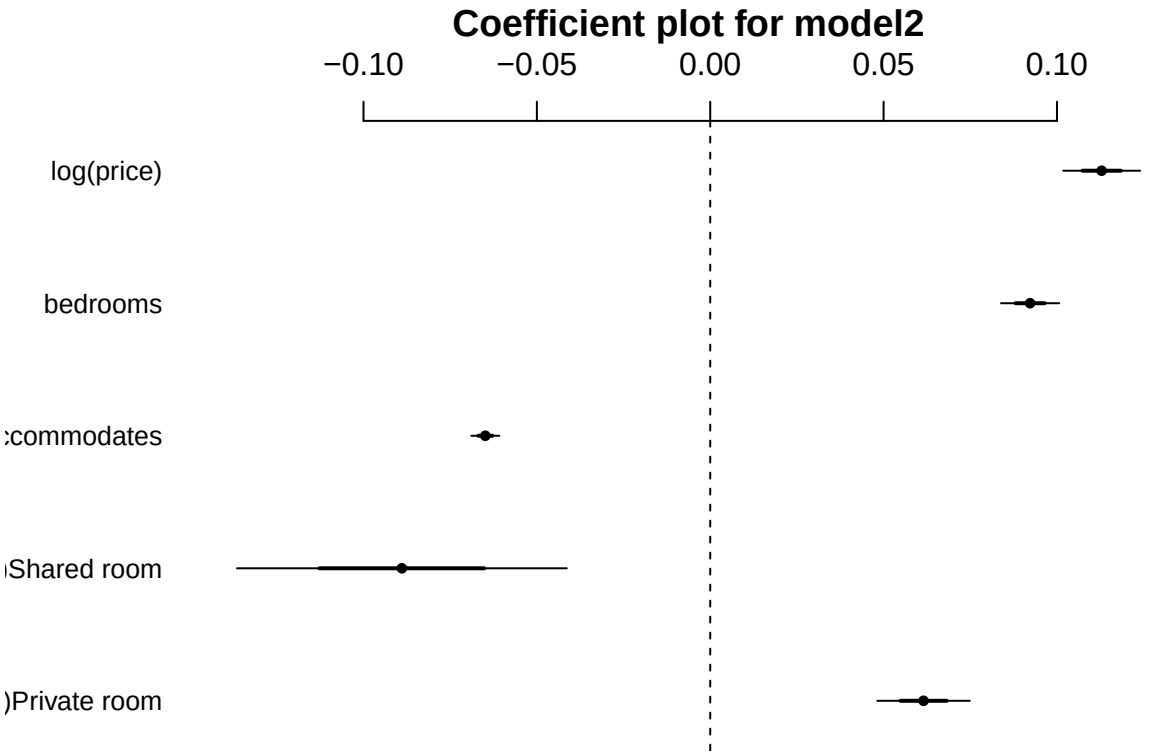
i.Model choice and interpretation




```
## glm(formula = overall_satisfaction ~ room_type + reviews + accommodates +
##     bedrooms + log(price), data = LondonAirbnb)
##                       coef.est coef.se
## (Intercept)            4.20     0.03
## room_typePrivate room  0.06     0.01
## room_typeShared room  -0.09     0.02
## reviews                0.00     0.00
## accommodates          -0.07     0.00
## bedrooms               0.09     0.00
## log(price)             0.11     0.01
## ---
##   n = 32197, k = 7
##   residual deviance = 5153.8, null deviance = 5383.9 (difference = 230.2)
##   overdispersion parameter = 0.2
##   residual sd is sqrt(overdispersion) = 0.40
```

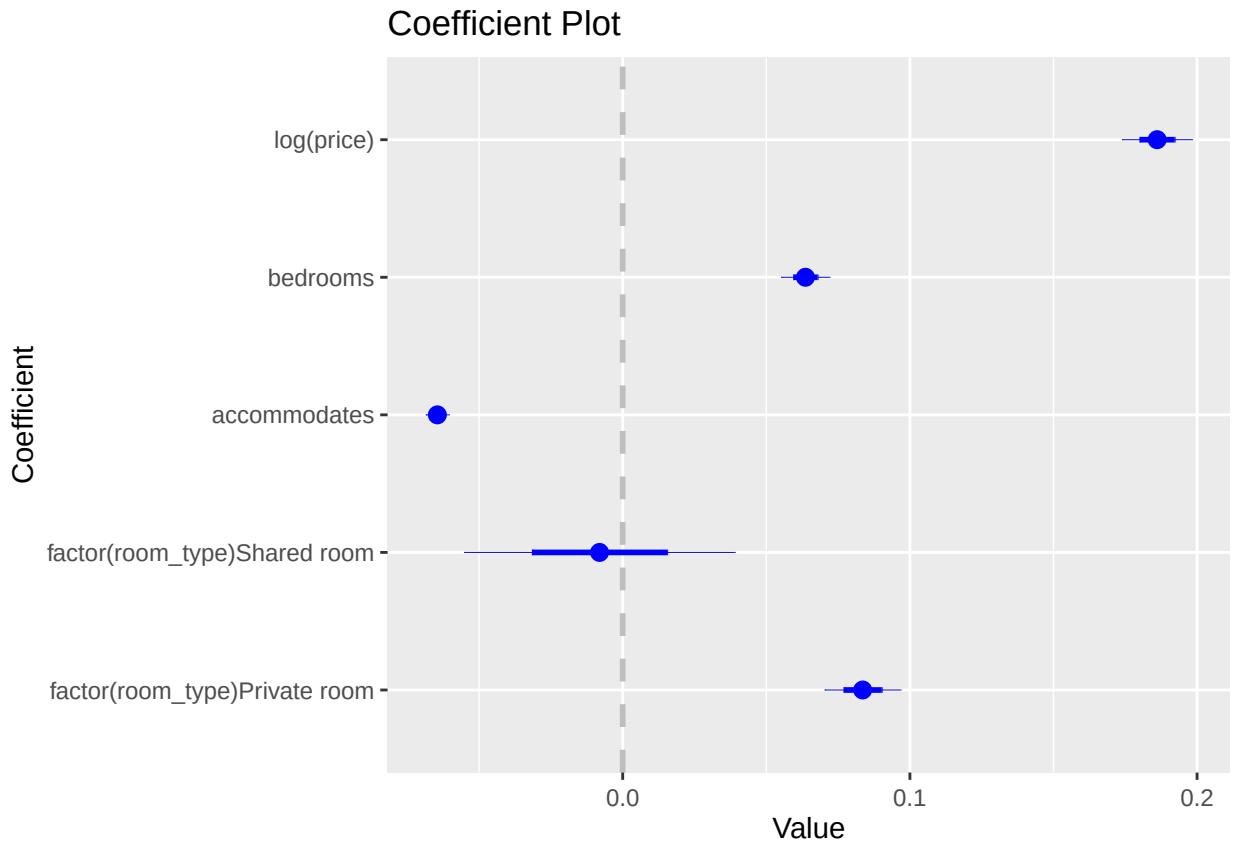
Based on the output, we can see that all the coefficients are significant except "reviews". With each unit increase of accommodates, the rating will decrease 0.07. With each unit increase of bedroom, the rating will

increase 0.09. When the log(price) increase by one, the rating will increase 0.11. Pivate room has 0.06 higher than Entire room and Shared room has 0.09 lower than the Entire room.



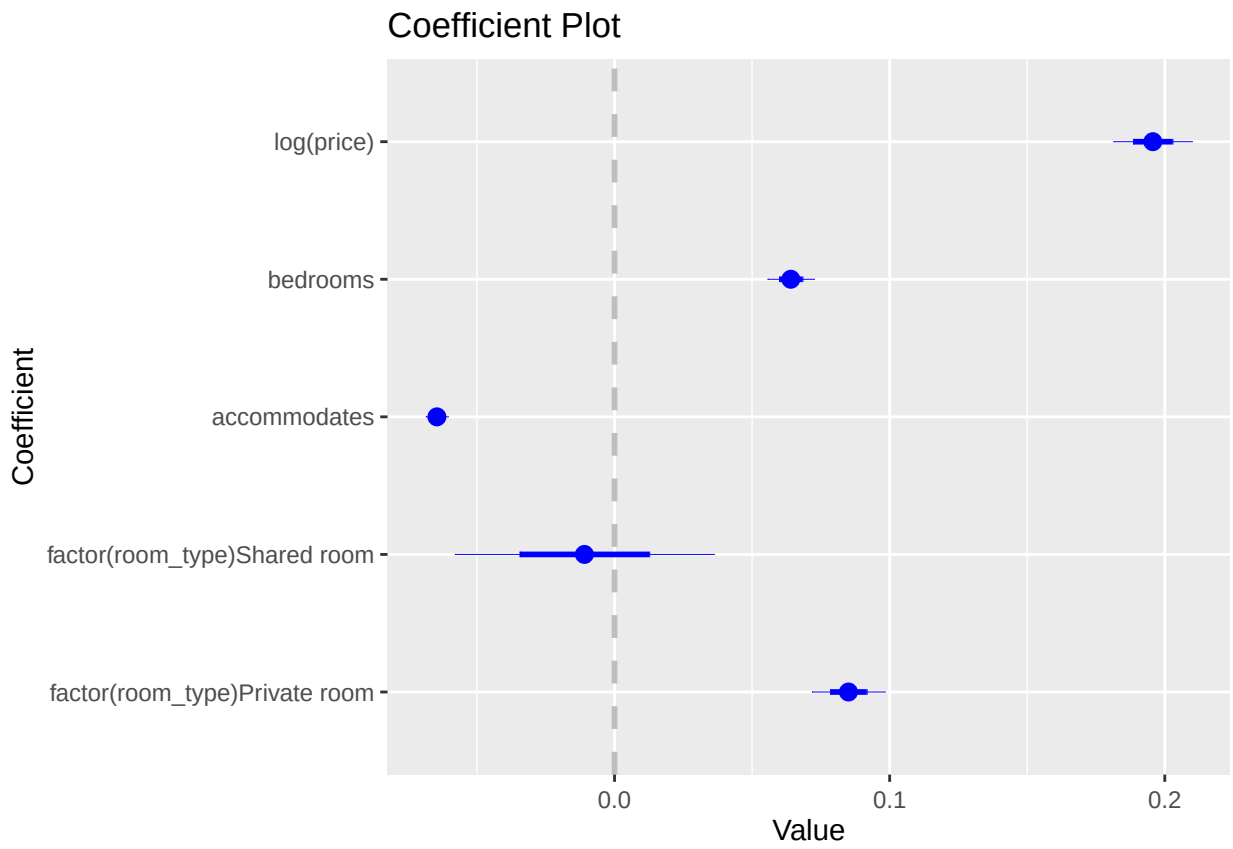

```
## glm(formula = overall_satisfaction ~ factor(room_type) + accommodates +
##     bedrooms + log(price), data = LondonAirbnb)
##                               coef.est coef.se
## (Intercept)                    4.20     0.03
## factor(room_type)Private room  0.06     0.01
## factor(room_type)Shared room  -0.09     0.02
## accommodates                  -0.06     0.00
## bedrooms                       0.09     0.00
## log(price)                     0.11     0.01
## ---
##   n = 32197, k = 6
##   residual deviance = 5155.2, null deviance = 5383.9 (difference = 228.7)
##   overdispersion parameter = 0.2
##   residual sd is sqrt(overdispersion) = 0.40
```

Because of the previous model, I remove reviews out. Based on the output, we can see that all the coefficients are significant except "reviews". With each unit increase of accommodates, the rating will decrease 0.06. With each unit increase of bedroom, the rating will increase 0.09. When the log(price) increase by one, the rating will increase 0.11. Pivate room has 0.06 higher than Entire room and Shared room has 0.09 lower than the Entire room.



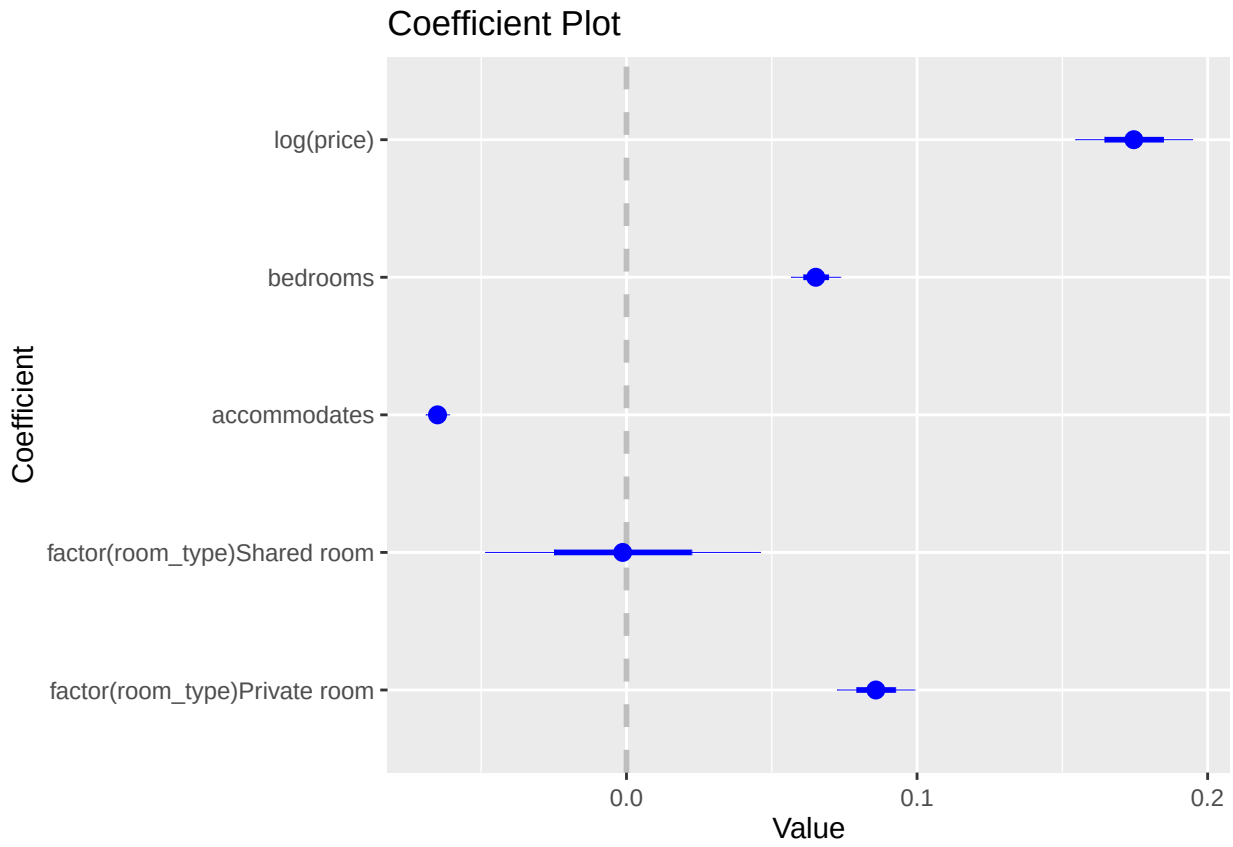

```
## lmer(formula = overall_satisfaction ~ factor(room_type) + accommodates +
##     bedrooms + log(price) + (1 | borough), data = LondonAirbnb)
##                               coef.est coef.se
## (Intercept)                    3.94     0.03
## factor(room_type)Private room  0.08     0.01
## factor(room_type)Shared room  -0.01     0.02
## accommodates                  -0.06     0.00
## bedrooms                       0.06     0.00
## log(price)                     0.19     0.01
##
## Error terms:
##  Groups   Name        Std.Dev.
##  borough  (Intercept) 0.09
##  Residual             0.39
## ---
## number of obs: 32197, groups: borough, 33
## AIC = 31379.7, DIC = 31263.5
## deviance = 31313.6
```

Based on the output, we can see that all the coefficients are significant except "shared room". With each unit increase of accommodates, the rating will decrease 0.06. With each unit increase of bedroom, the rating will increase 0.06. When the log(price) increase by one, the rating will increase 0.19. Pivate room has 0.08 higher than Entire room and Shared room has 0.01 lower than the Entire room. Thr borough variation has the standard deviation of 0.39 an d the intercept of 0.09.




```
## lmer(formula = overall_satisfaction ~ factor(room_type) + accommodates +
##     bedrooms + log(price) + (0 + log(price) | borough), data = LondonAirbnb)
##                               coef.est coef.se
## (Intercept)                    3.90     0.03
## factor(room_type)Private room  0.09     0.01
## factor(room_type)Shared room  -0.01     0.02
## accommodates                  -0.06     0.00
## bedrooms                       0.06     0.00
## log(price)                     0.20     0.01
##
## Error terms:
##  Groups   Name       Std.Dev.
##  borough  log(price) 0.02
##  Residual            0.39
## ---
## number of obs: 32197, groups: borough, 33
## AIC = 31417.6, DIC = 31301.4
## deviance = 31351.5
```

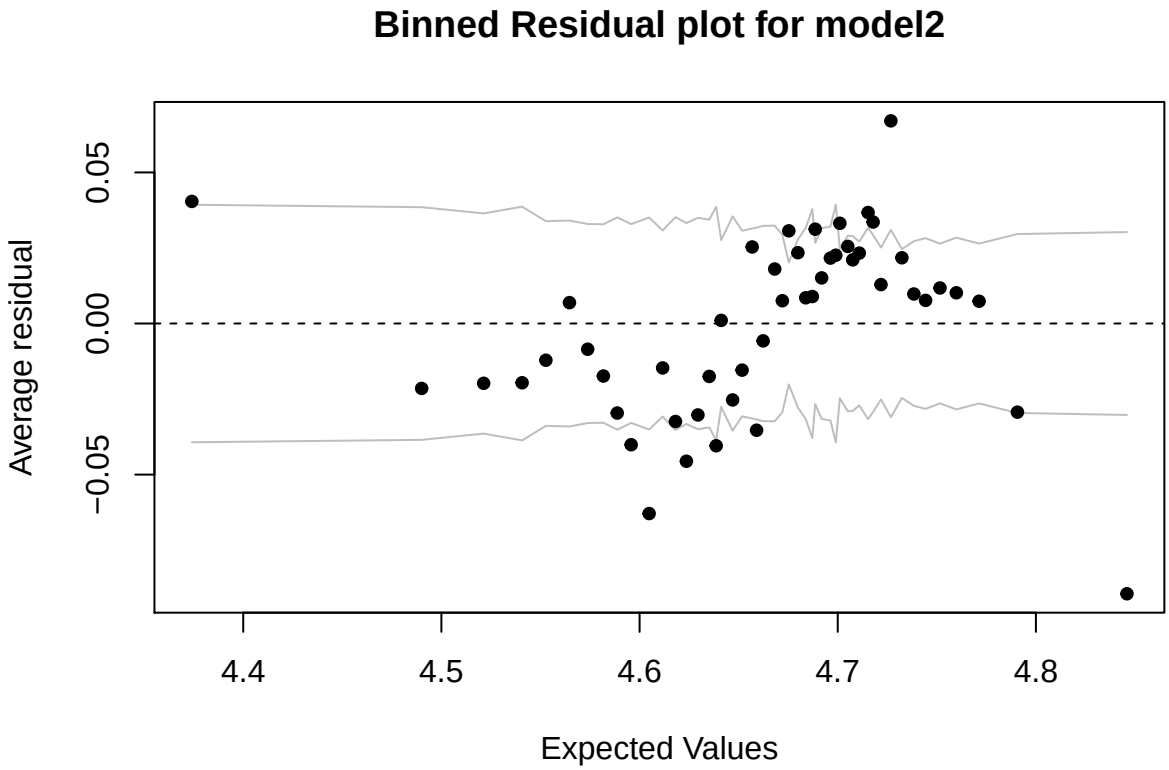
Based on the output, we can see that all the coefficients are significant except "shared room". With each unit increase of accommodates, the rating will decrease 0.06. With each unit increase of bedroom, the rating will increase 0.06. When the log(price) increase by one, the rating will increase 0.20. Pivate room has 0.09 higher than Entire room and Shared room has 0.01 lower than the Entire room. Thr borough has the residual 0.39 and the intercept of 0.02.



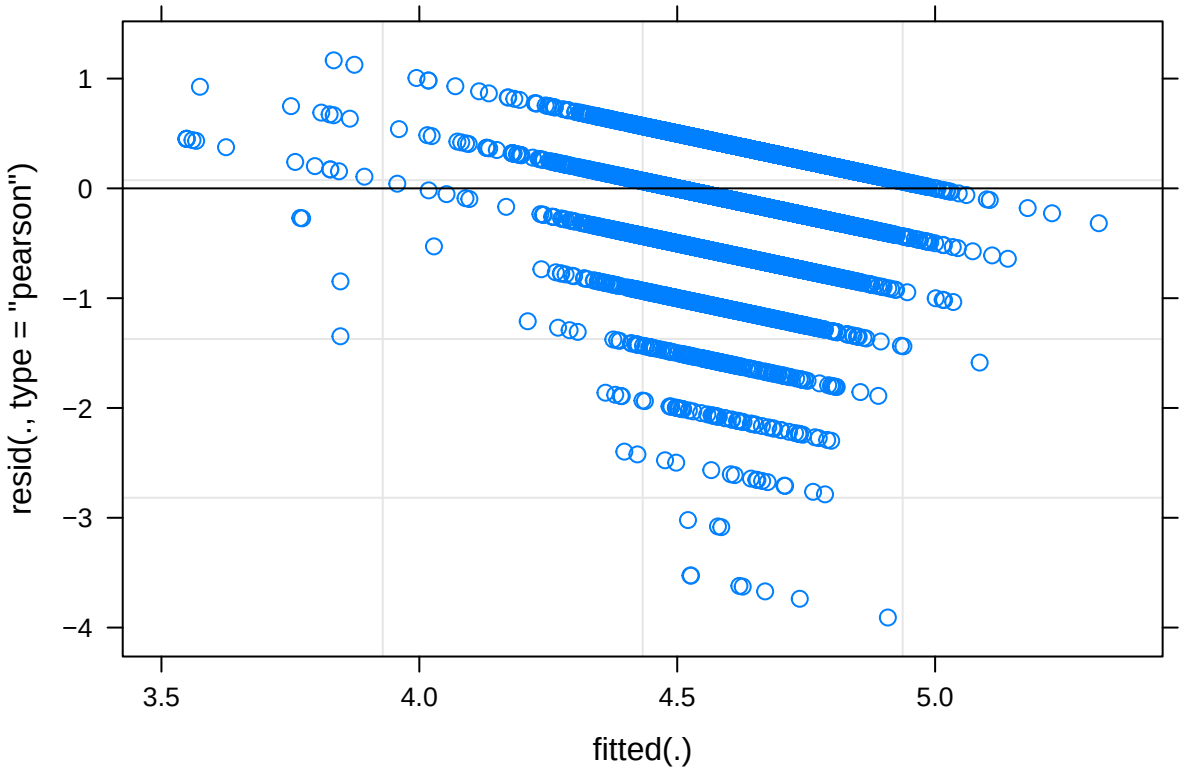

```
## lmer(formula = overall_satisfaction ~ factor(room_type) + accommodates +
##     bedrooms + log(price) + (1 + log(price) | borough), data = LondonAirbnb)
##                               coef.est coef.se
## (Intercept)                    3.99     0.05
## factor(room_type)Private room  0.09     0.01
## factor(room_type)Shared room   0.00     0.02
## accommodates                  -0.07     0.00
## bedrooms                       0.07     0.00
## log(price)                     0.17     0.01
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  borough  (Intercept) 0.20
##           log(price)  0.04     -0.92
##  Residual             0.39
## ---
## number of obs: 32197, groups: borough, 33
## AIC = 31339.1, DIC = 31222.3
## deviance = 31270.7
```

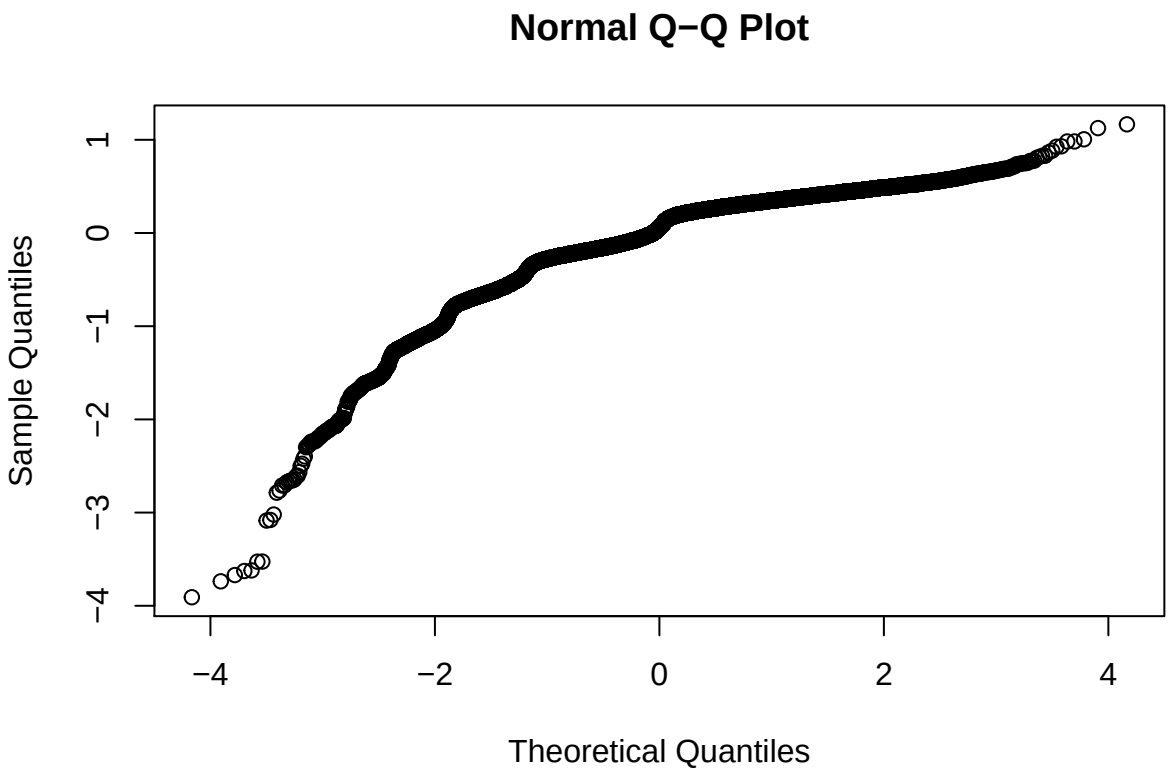
Comparing to the previous model, the signs do not change. The residual is still 0.39 and the slope is 0.04. The correlation with intercept is -0.92.

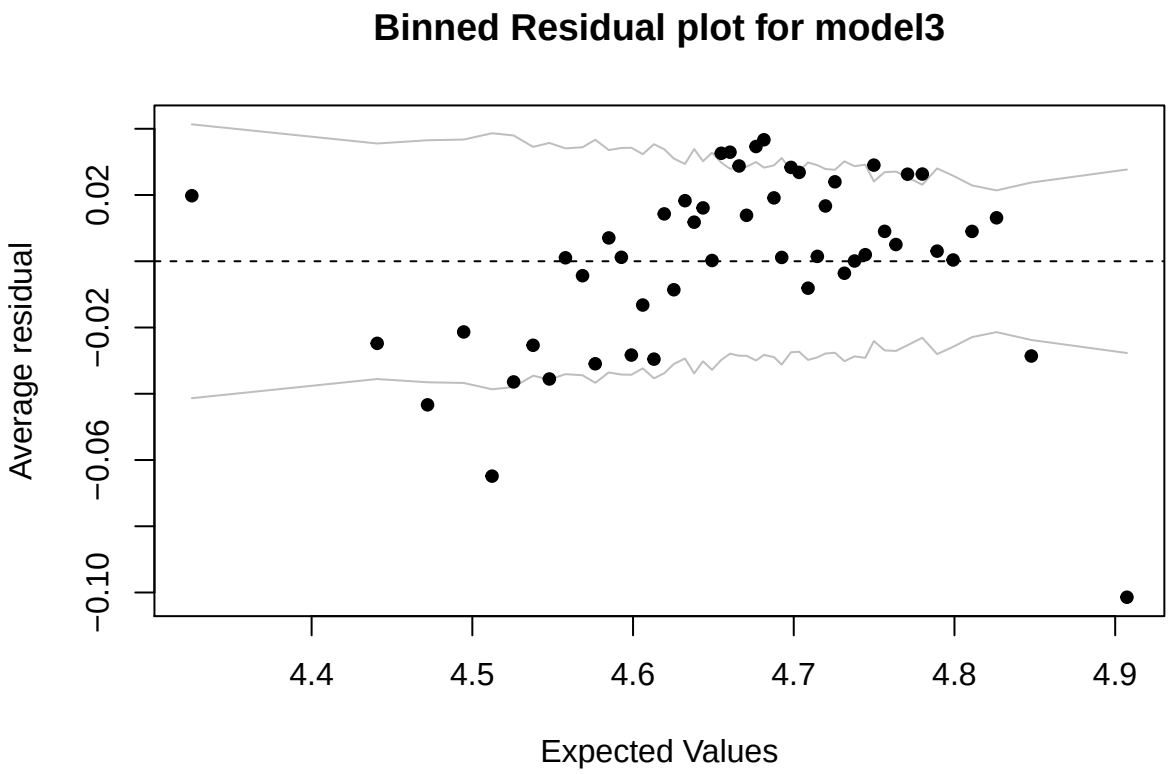
ii.Model checking

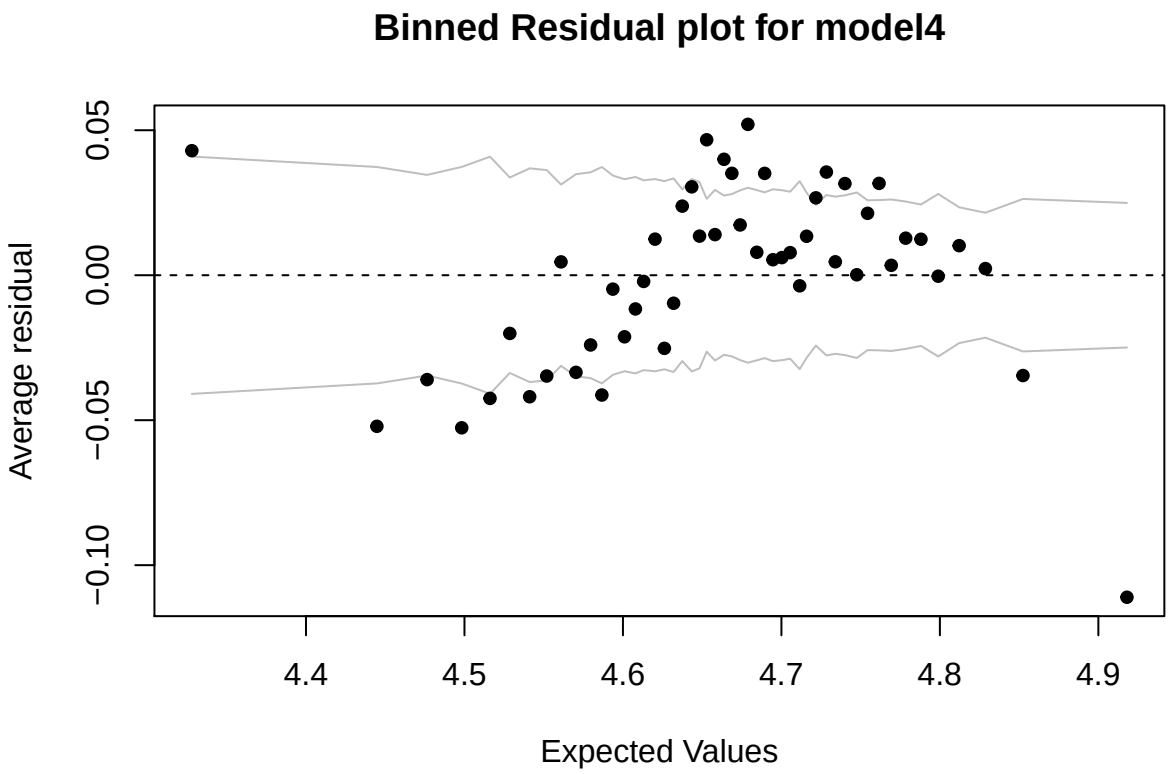


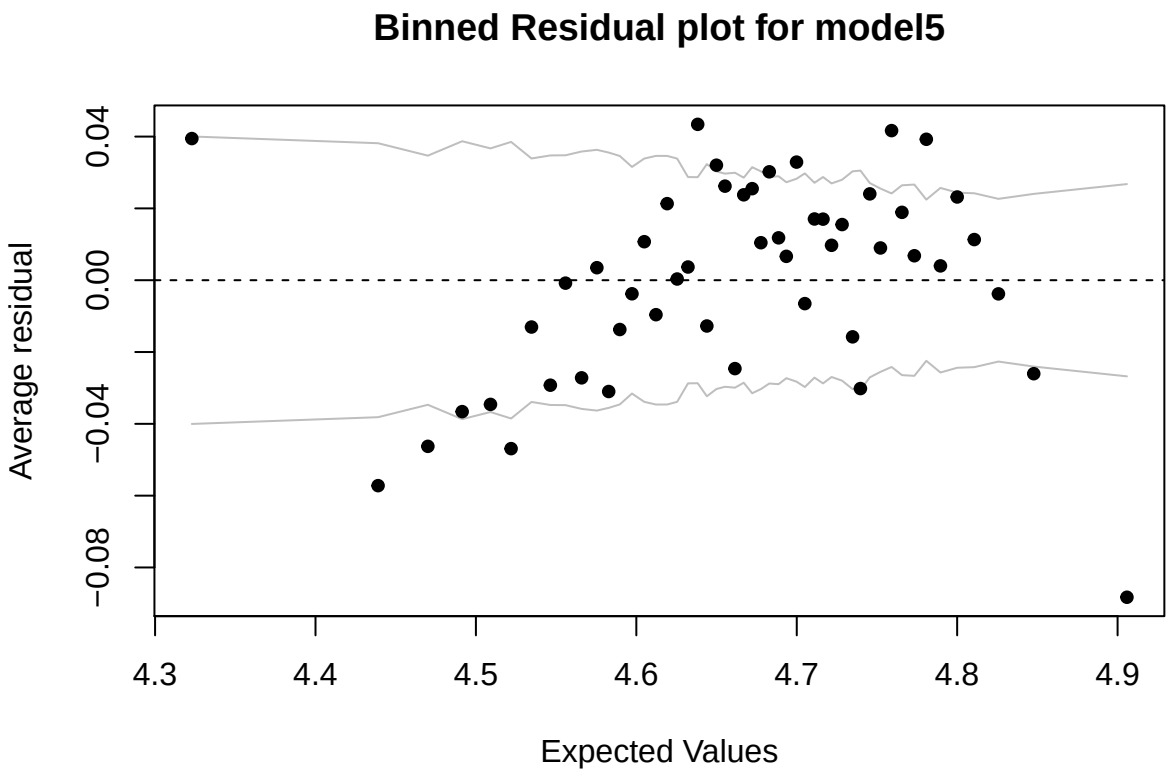

```
## refitting model(s) with ML (instead of REML)
## Data: LondonAirbnb
## Models:
## model3: overall_satisfaction ~ factor(room_type) + accommodates + bedrooms +
## model3:     log(price) + (1 | borough)
## model4: overall_satisfaction ~ factor(room_type) + accommodates + bedrooms +
## model4:     log(price) + (0 + log(price) | borough)
## model5: overall_satisfaction ~ factor(room_type) + accommodates + bedrooms +
## model5:     log(price) + (1 + log(price) | borough)
##        Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model3  8 31330 31397 -15657    31314
## model4  8 31368 31435 -15676    31352  0.00      0          1
## model5 10 31291 31375 -15635    31271 80.83      2     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


resid(., type = "pearson")
1
0
−1
−2
−3
−4
3.5
4.0
4.5
5.0
fitted(.)


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−4
−3
−2
−1
0
1
−4
−2
0
2
4


Binned Residual plot for model3
Average residual
0.02
−0.02
−0.06
−0.10
4.4
4.5
4.6
4.7
4.8
4.9
Expected Values


Binned Residual plot for model4
Average residual
0.05
0.00
−0.05
−0.10
4.4
4.5
4.6
4.7
4.8
4.9
Expected Values



Model5 is better because of the lower AIC and BIC. The residual plot looks great for the models.

e.Discussion

i.Implication andLimitation

The price has a lot positive influence for the rating. There are some time limitation, because the data only contained the properties as of July 28, 2017. The data cannot show the big picture of the whole Airbnb properties.

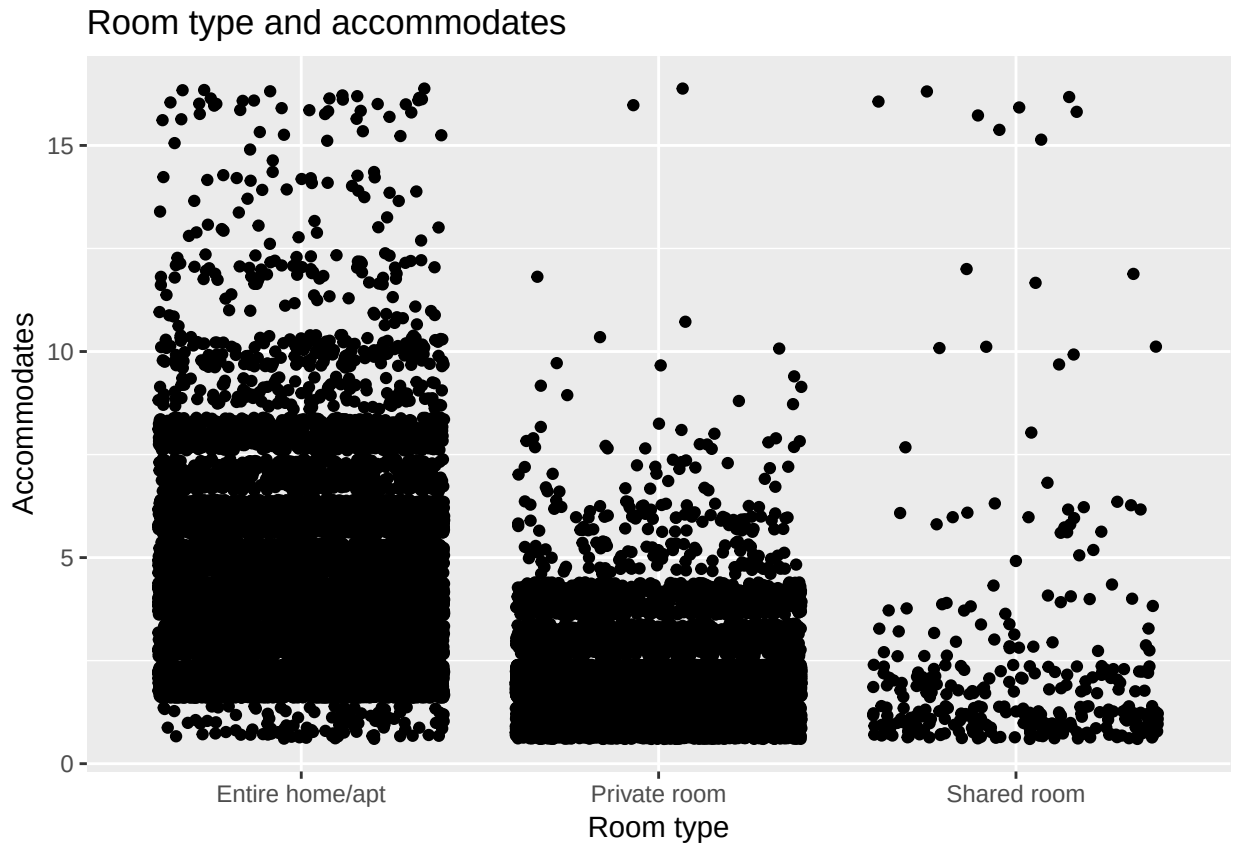
ii.Future direction

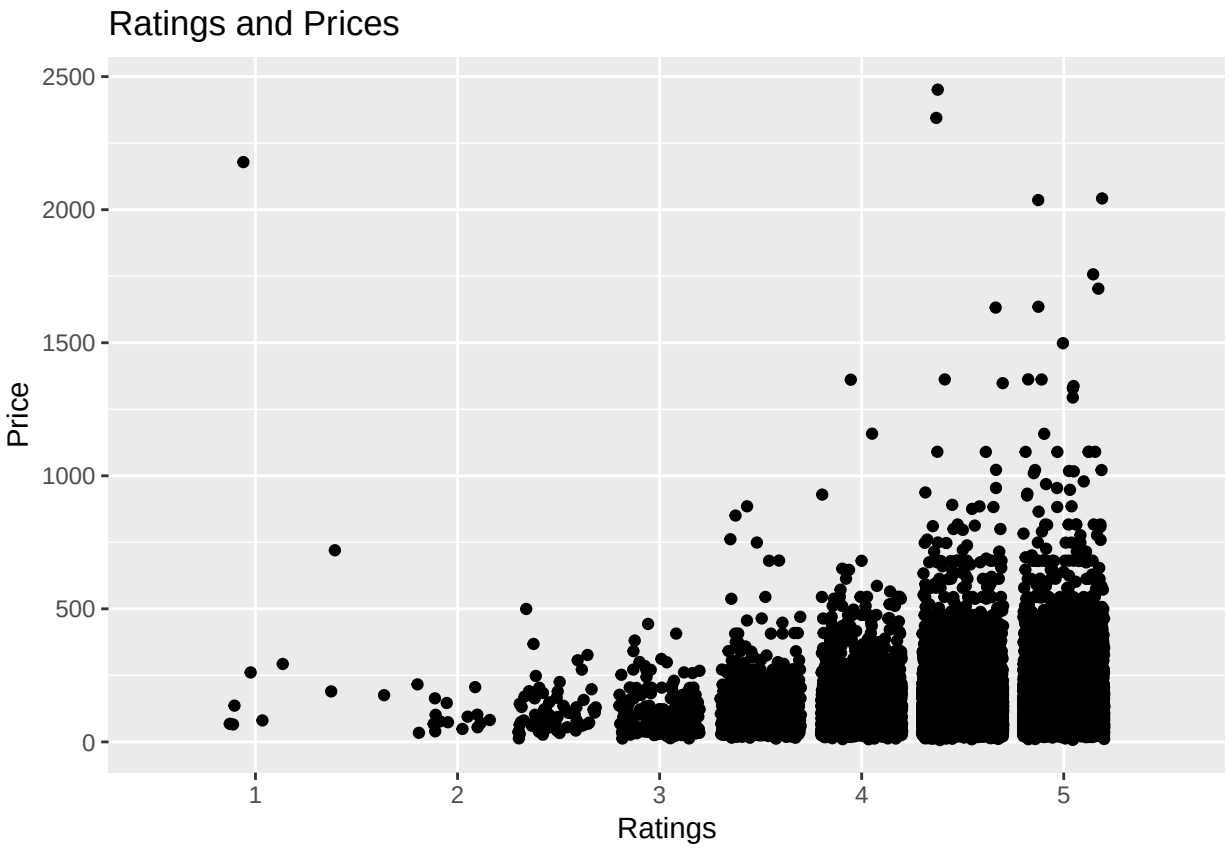
For the rating model, I can do the multilevel logistic model in the future.

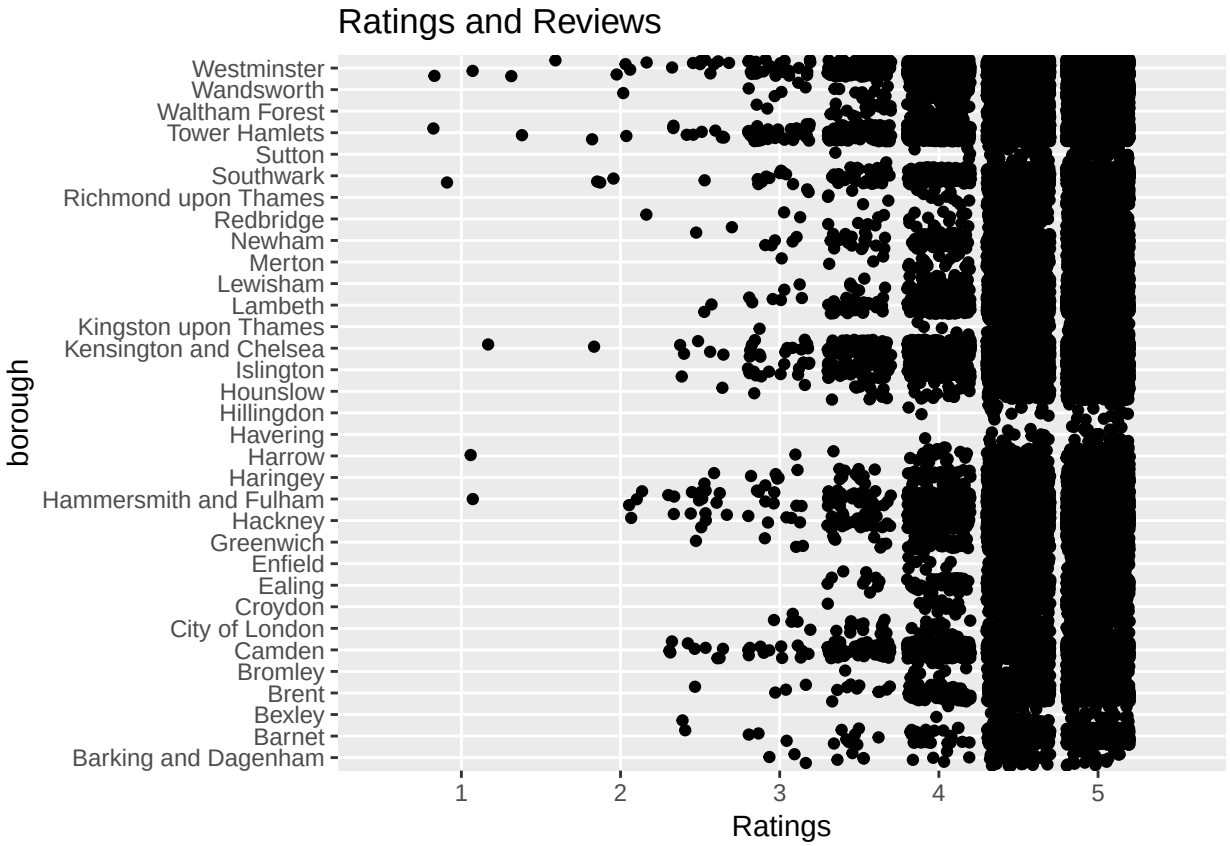
f.Reference

http://tomslee.net/airbnb-data-collection-get-the-data

g.Appendix


Room type and accommodates
Accommodates
15
10
5
0
Entire home/apt
Private room
Shared room
Room type


Ratings and Prices
2500
2000
1500
1000
500
0
Price
1
2
3
4
5
Ratings


Ratings and Reviews
borough
Westminster
Wandsworth
Waltham Forest
Tower Hamlets
Sutton
Southwark
Richmond upon Thames
Redbridge
Newham
Merton
Lewisham
Lambeth
Kingston upon Thames
Kensington and Chelsea
Islington
Hounslow
Hillingdon
Havering
Harrow
Haringey
Hammersmith and Fulham
Hackney
Greenwich
Enfield
Ealing
Croydon
City of London
Camden
Bromley
Brent
Bexley
Barnet
Barking and Dagenham
1
2
3
4
5
Ratings