Airbnb Midterm Project

Xiangliang Liu

12/8/2018

I. Abstract

Nowadays, as the development of tourism, more and more people are willing to spend money traveling around the world. Airbnb is one of the good choices for travelers to find appropriate accommodation to meet different requirements. This report will mainly focus on predicting the price of different rooms on the Airbnb website. Specifically, the report will do analysis on Airbnb data with two method: EDA(Exploratory Data Analysis) and modeling. After reading this report, you will have a general idea how to predict the price of rooms on Airbnb website and what's the most influential factor in predicting process.

II. Introduction

This project will be focusing on the analysis on the relationship between price and other factors on Airbnb website. Airbnb is a privately held global company headquartered in San Francisco that operates an online marketplace and hospitality service which is accessible via its websites and mobile apps. Members can use the service to arrange or offer lodging, primarily homestays, or tourism experiences. The whole project will consist of following parts "abstract", "introduction", "Method", "Result","Discussion", "Acknowledgement", "reference" and "Appendix. Firstly, I will read in the data and do some visualization to see which predictor will contributes more to the prediction of price. And then the modeling will be conducting multi-level regression using room type and neighborhood as factors to predict the price.

III. Method

Data source:

The data was extracted from http://tomslee.net, which can also be extracted from official Airbnb website. Specifically, the data was collected from the September 2014 to July 2017 in Boston area. There are 10 variales that will be used in this project. Specifically, they are room id, host, id room type, neighborhood, number of reviews, overall satisfaction(rating), number of accommodates, number of bedrooms and minimum stay for a visit those will be the potential factors to influence the pricing.

overview of Airbnb data:

Table 1. The head of the data:

room_id	host_id	room_type	neighborhood	reviews	overall_satisfaction
5453	8021	Private room	Jamaica Plain	53	5.0
5506	8229	Private room	Roxbury	30	4.5
6695	8229	Entire home/apt	Roxbury	39	5.0
6976	16701	Private room	Roslindale	26	5.0
8789	26988	Entire home/apt	Downtown	1	5.0
8792	26988	Entire home/apt	Downtown	11	4.5

features of the dataset:

Table 2. features of the dataset:

```
##     room_id           host_id           room_type
##  Min.   :    3353   Min.   :     4240   Length:76941
##  1st Qu.: 4409653   1st Qu.:  5695034   Class :character
##  Median : 8227206   Median : 18517776   Mode  :character
```

```
##  Mean   : 8543440   Mean   : 26142364
##  3rd Qu.:12949309   3rd Qu.: 36128699
##  Max.   :19777573   Max.   :139551362
##                     NA's   :6
##  neighborhood          reviews       overall_satisfaction  accommodates
##  Length:76941       Min.   :  0.00   Min.   :0.000        Min.   : 1.000
##  Class :character   1st Qu.:  1.00   1st Qu.:4.000        1st Qu.: 2.000
##  Mode  :character   Median :  5.00   Median :4.500        Median : 2.000
##                     Mean   : 18.95   Mean   :3.836        Mean   : 3.034
##                     3rd Qu.: 20.00   3rd Qu.:5.000        3rd Qu.: 4.000
##                     Max.   :470.00   Max.   :5.000        Max.   :16.000
##                                      NA's   :15510        NA's   :2339
##     bedrooms          price            minstay          latitude
##  Min.   : 0.00   Min.   :    0.0   Min.   :  1.00   Min.   :42.24
##  1st Qu.: 1.00   1st Qu.:   82.0   1st Qu.:  1.00   1st Qu.:42.33
##  Median : 1.00   Median :  140.0   Median :  2.00   Median :42.35
##  Mean   : 1.27   Mean   :  173.6   Mean   :  2.68   Mean   :42.34
##  3rd Qu.: 2.00   3rd Qu.:  218.0   3rd Qu.:  3.00   3rd Qu.:42.35
##  Max.   :10.00   Max.   :10000.0   Max.   :365.00   Max.   :42.39
##  NA's   :5018                      NA's   :32314
##    longitude
##  Min.   :-71.17
##  1st Qu.:-71.10
##  Median :-71.08
##  Mean   :-71.08
##  3rd Qu.:-71.06
##  Max.   :-70.99
##
```

According to the summary we can see there are some missing values in the column of host_id, overvall_statisfaction, accommodates, bedrooms and minstay.

After eliminating all the miss values, we filtered the data with 0 reviews, or price = 0. the observation is meaningless if it has 0 value in number of reviews and price

Table 3. Stat about the room type

Var1	Freq
Entire home/apt	14246
Private room	10706
Shared room	561

*We can see from table 2 that entire home/apt has the largest number in boston area, and shared room has smallest proportion.

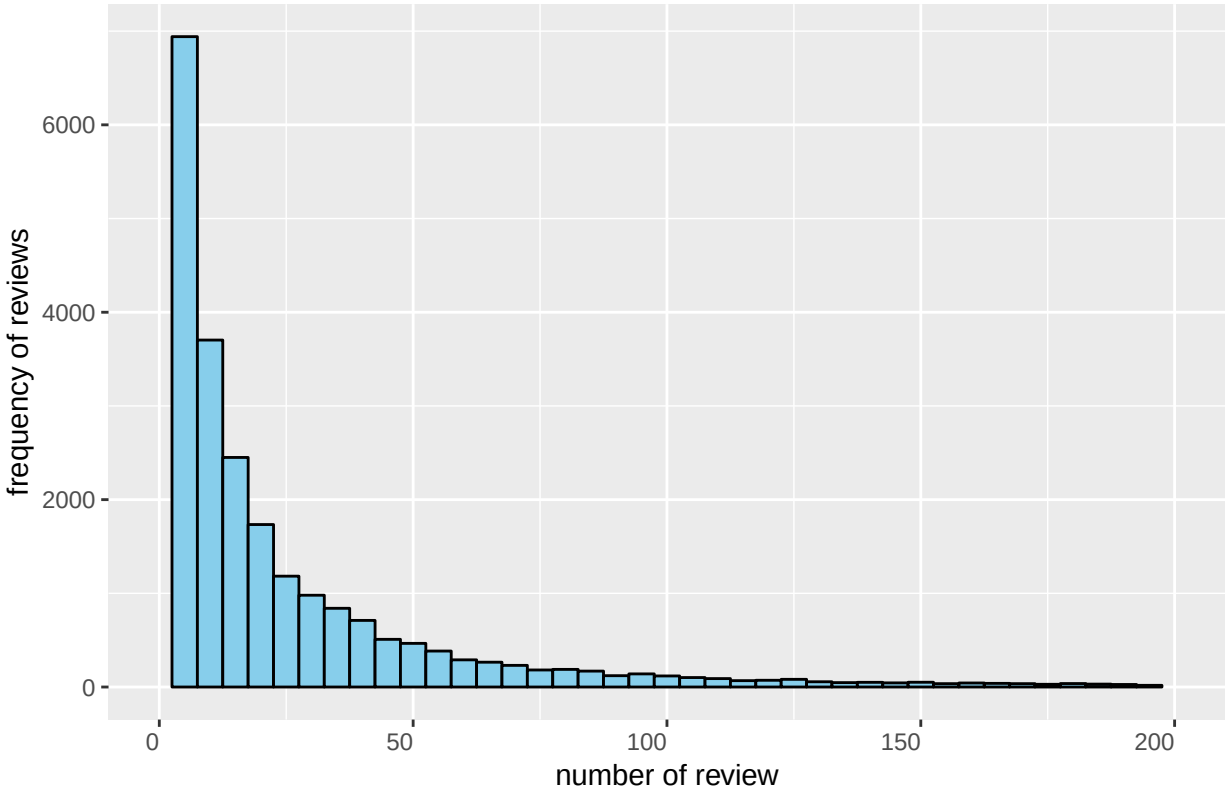
EDA

Distribution of number of reviews

```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

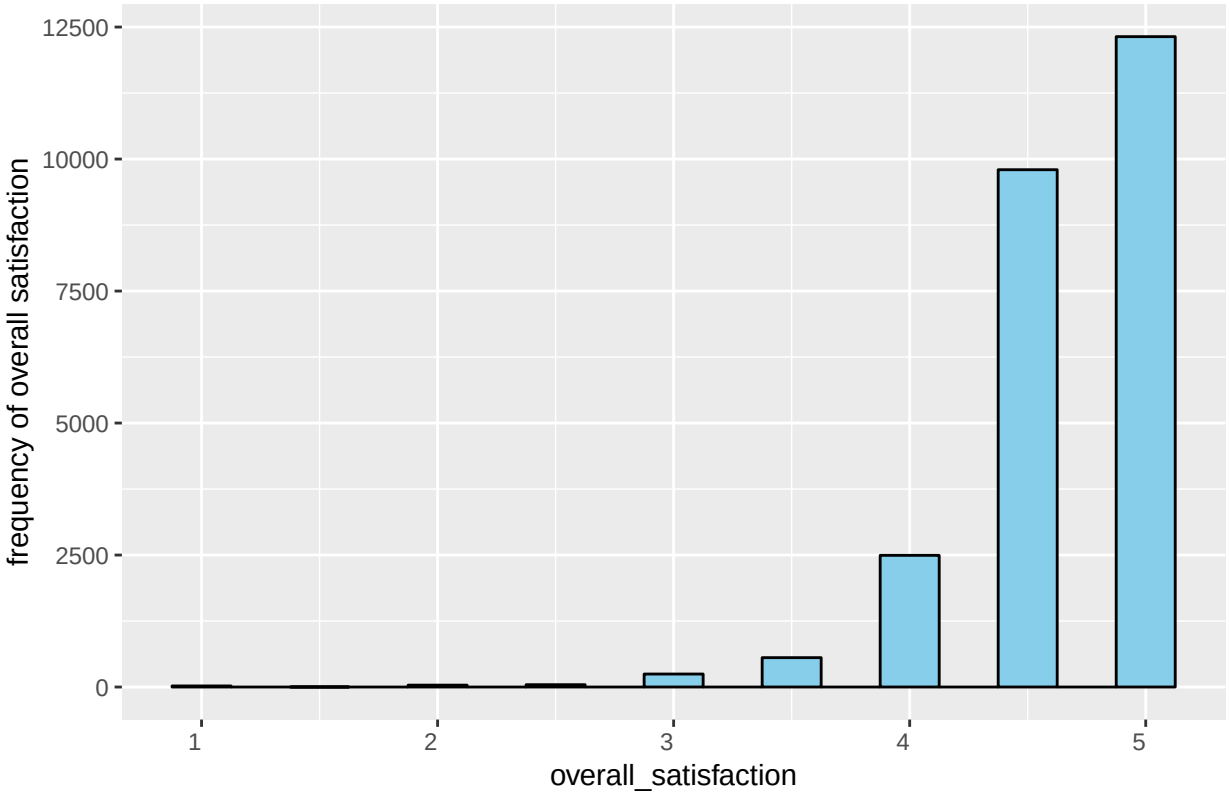
Fig 1.distribution of number of reviews



as we can see from the distribution plot of number of reviews, most of airbnb hosts have less than 100 reviews.

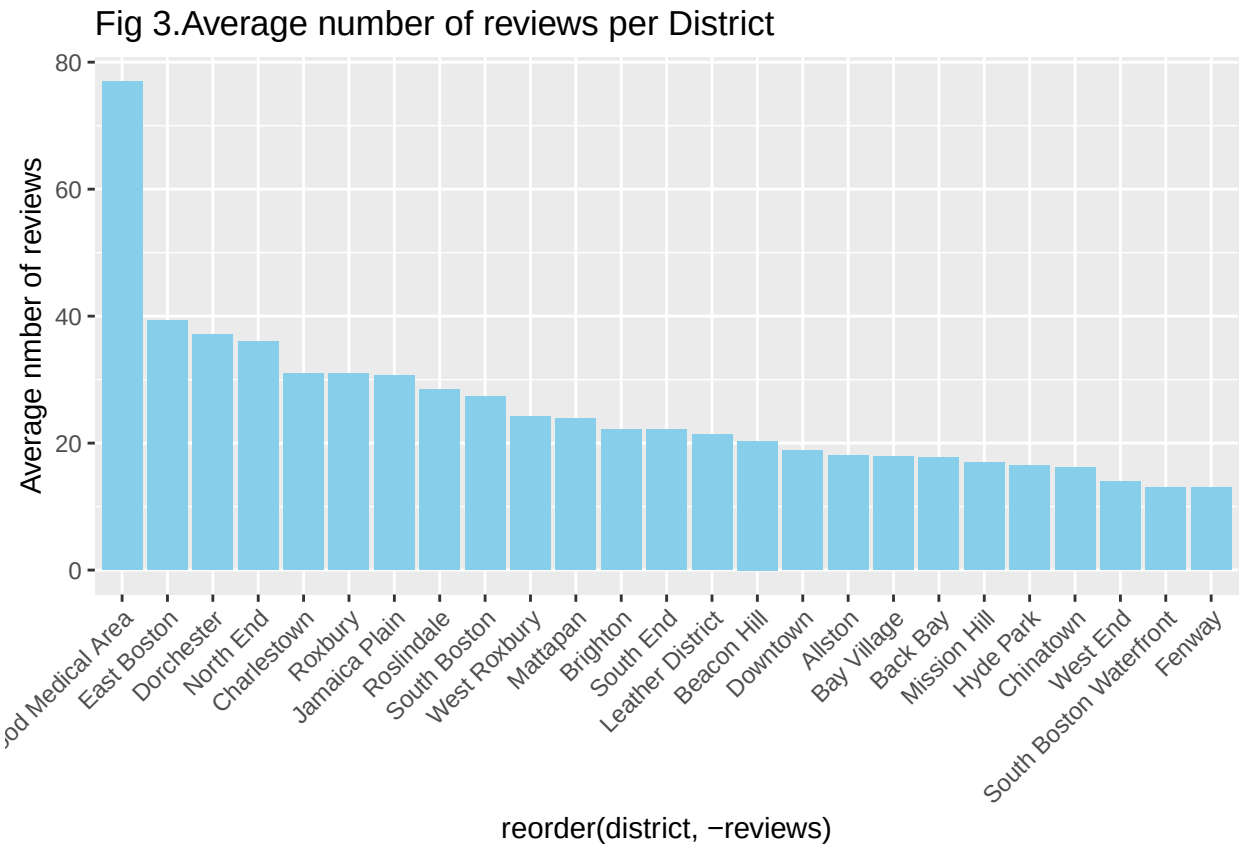
Distribution of overall satisfaction:

Fig 2.distribution of overall satisfaction



in th histgram plot, most of overall rating is aroud 4.5 and 5. There are also a small portion of people rate the room 4 star. Overall, customers are satisfied with most of rooms in boston area

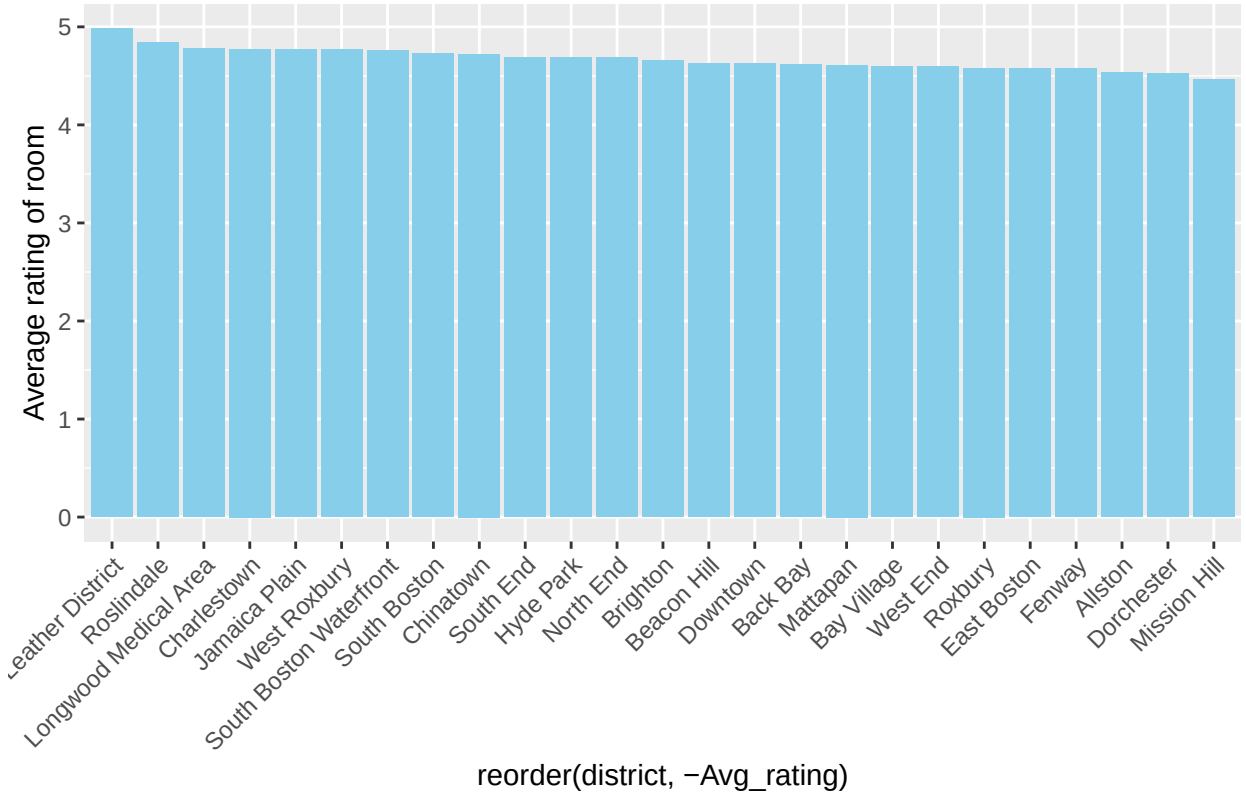
Number of room, pricing and average rating with different districts:



From this plot, the district Longwood Medical area has the highest number of review (around 78), while Fenway has the lowest average number of reviews in Boston areas(around 15). So the average number of review do vary a lot by district.

Create a map with leaflet:

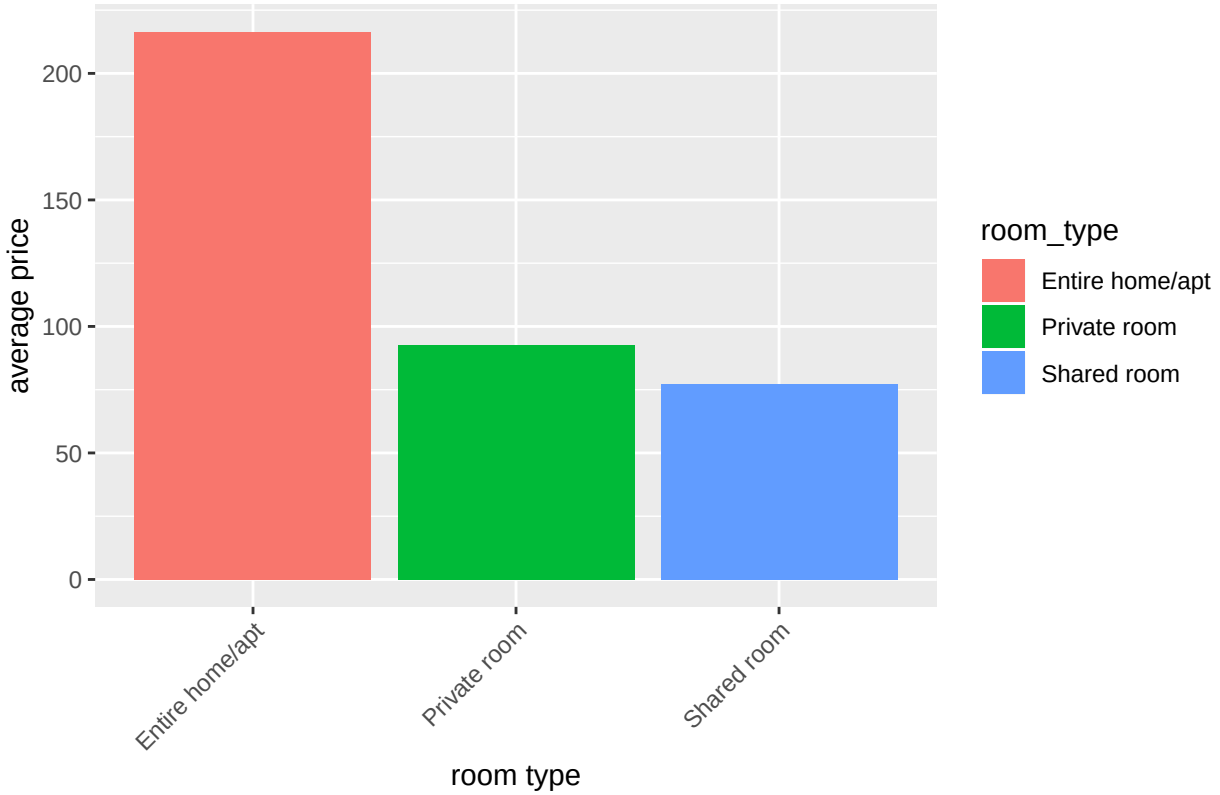
Fig 4.Average rating of airbnb rooms per neighborhood



As the graph showing below, we can't tell significant diffence of rating among different neighborhoods. But we can see the average rating varys by district. Leather District has highest average rating which is close to 5. So neighborhood of the Airbnb room could be an influence predictor. We might want to include this predictor in the model to see whether rating is a significant for predicting price of rooms.

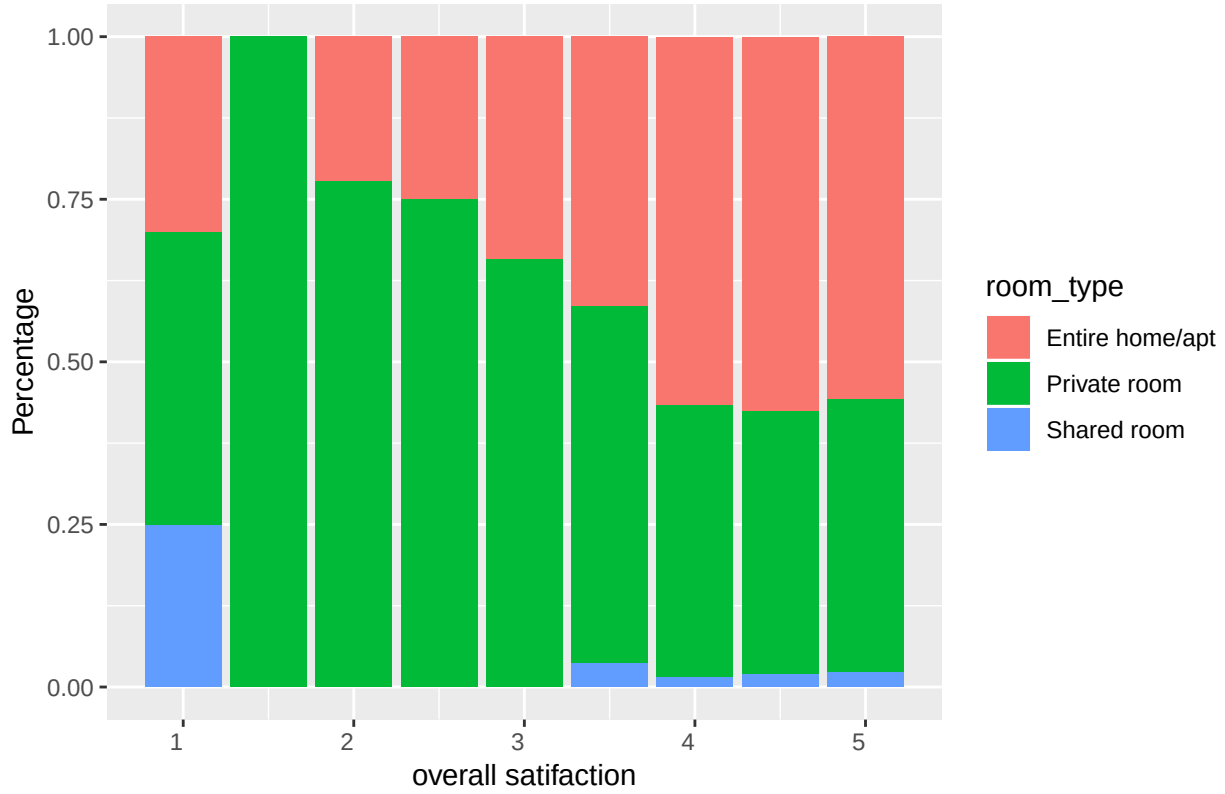
Average rating and price with different room types:

Fig 5.Average price with different room types



From the plot below, we can see entire home/Apt have highest average price among all types of room. This result is reasonable because entire home/apt have more space that can serve more people, so the price is higher.

Fig 6. Average rating with different room types



The plot showing above indicates the proportion of room types in different score of rating. We can tell from the graph that when overall satisfaction = 1 share room has relatively larger proportion compared situations when rating equal other values. However, we can't tell that "Entire home/apt" have higher proportion than other room type, because we found in previous EDA that the number of entire home/apt is higher than either private room or shared room.

The accommodates and bedrooms could be two correlated terms in the model, because the number of bedroom will limit the number of customers served. So the correlation test will be conducted in next step.

```
##
##  Pearson's product-moment correlation
##
## data:  BostonAirbnb$accommodates and BostonAirbnb$bedrooms
## t = 150.06, df = 25511, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6781395 0.6911761
## sample estimates:
##       cor
## 0.6847126
```

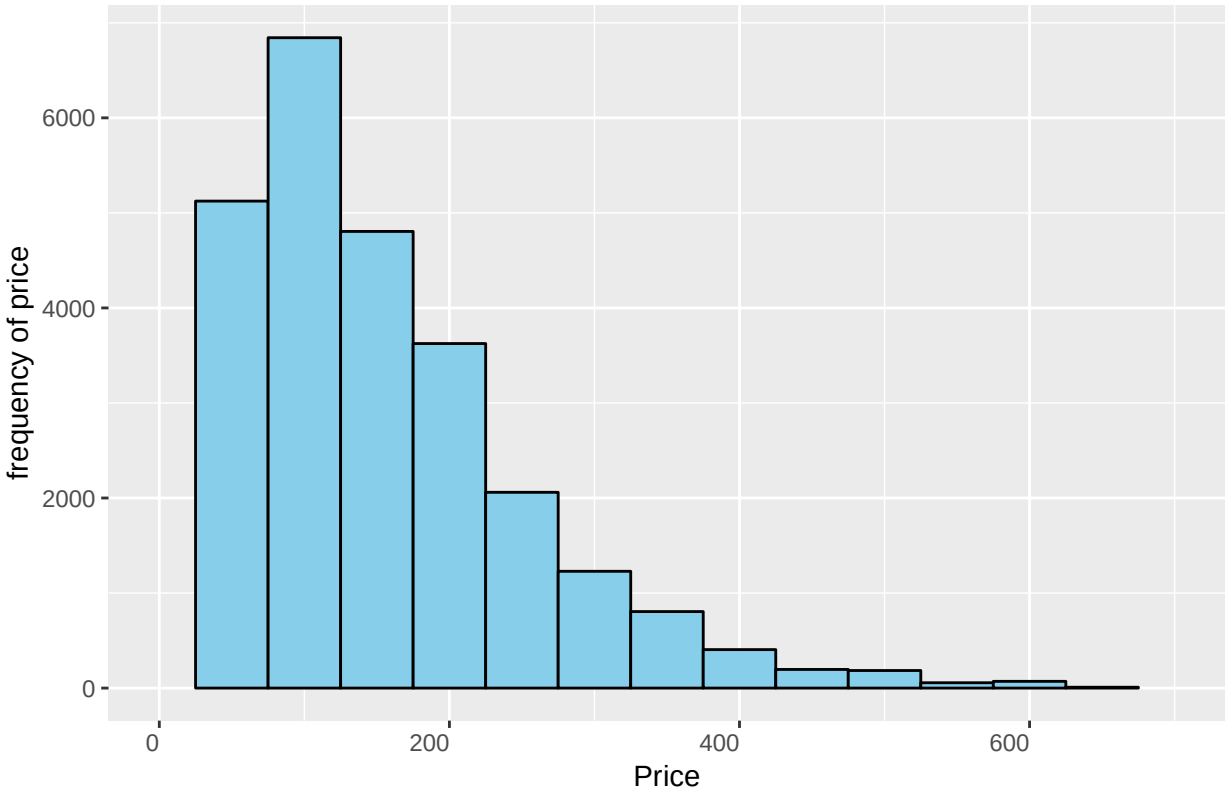
the p-value of this test is 2.2e-16. Reject the null hypothesis. So correlation between the those two variables is significant. we might want to add the correlation term into the model to test whether this influence term is significant

Before going further, we want to verify the distribution of the response variable–price, to test the assumption that it is normal distributed.

```
## Warning: Removed 40 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

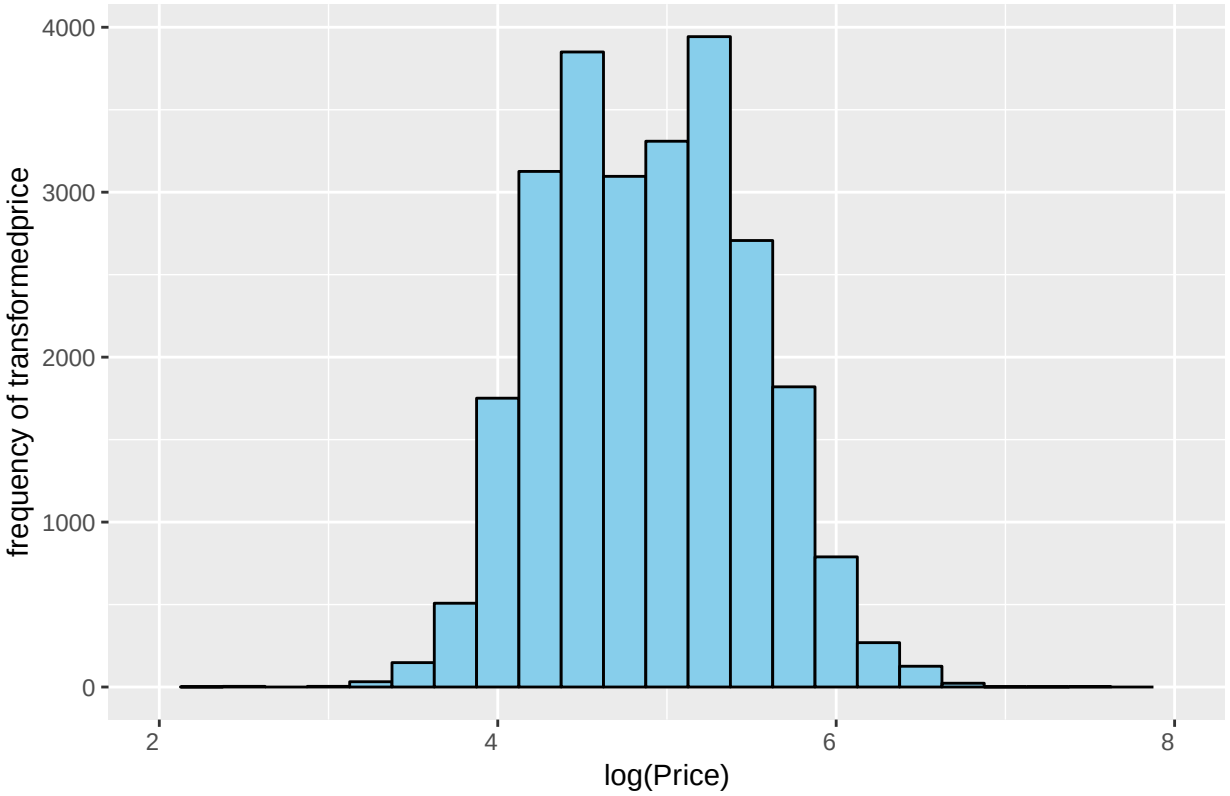
Fig 7. Distribution of room price




```
## Warning: Removed 3 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).
```

Fig 8.Distribution of room price

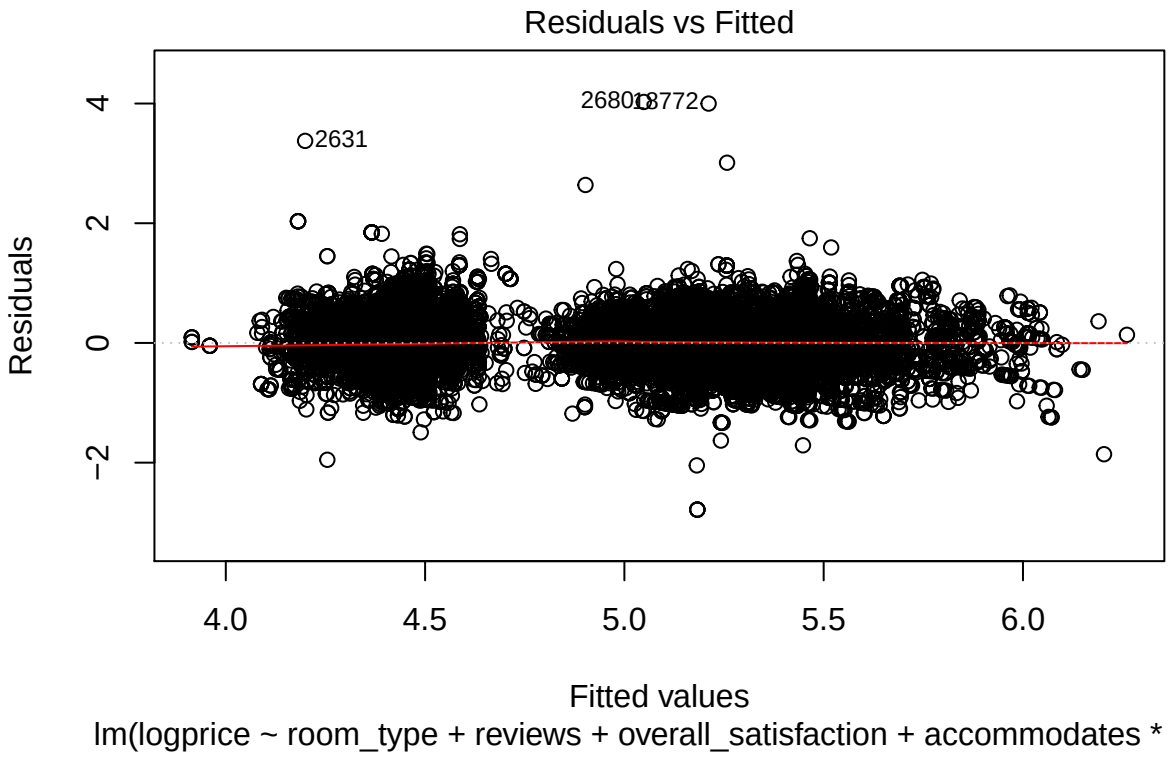


The first histogram distribution plot obviously does not follow normal distribution. After doing the log transformation, we got the second histogram plot. Now it looks like it follows normal distribution.

Modelling:

Model1: Simple linear regression:

$$log(price) = \alpha + \beta_1 x_{roomtype} + \beta_2 x_{reviews} + \beta_3 x_{reviews} + \beta_4 x_{accommodates*bedrooms} + \beta_5 x_{accommodates} + \beta_6 x_{bedrooms} + \beta_7 x_{minstay}$$


Residuals vs Fitted
Residuals
Fitted values
lm(logprice ~ room_type + reviews + overall_satisfaction + accommodates * b ...
2631
4.0
4.5
5.0
5.5
6.0
−2
0
2
4

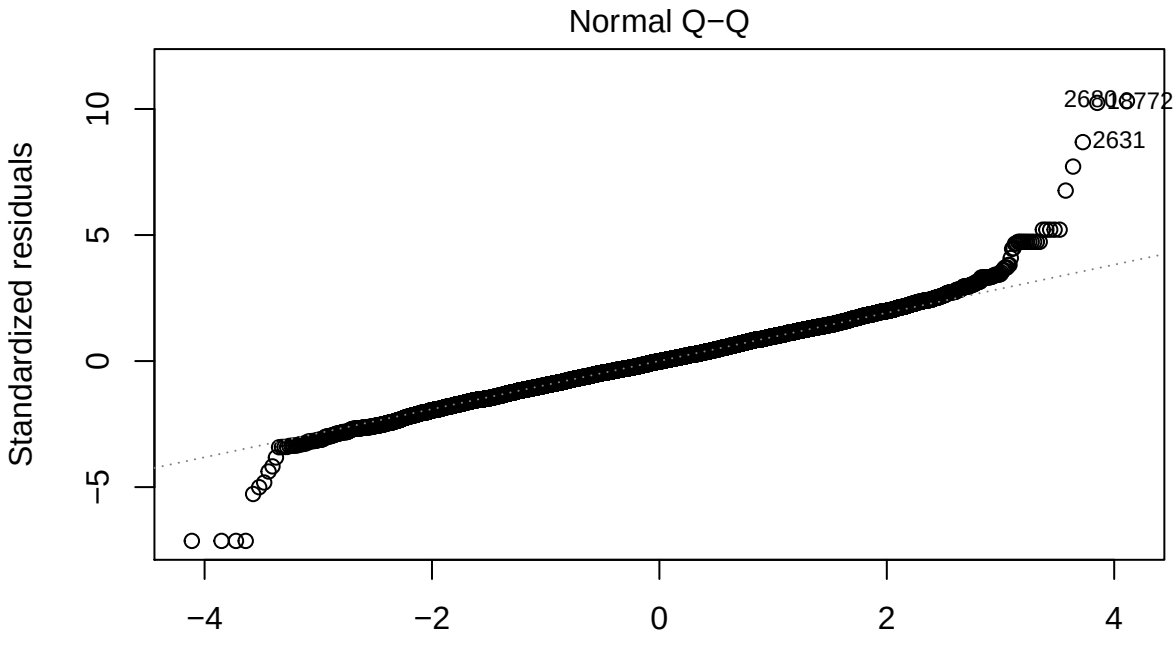
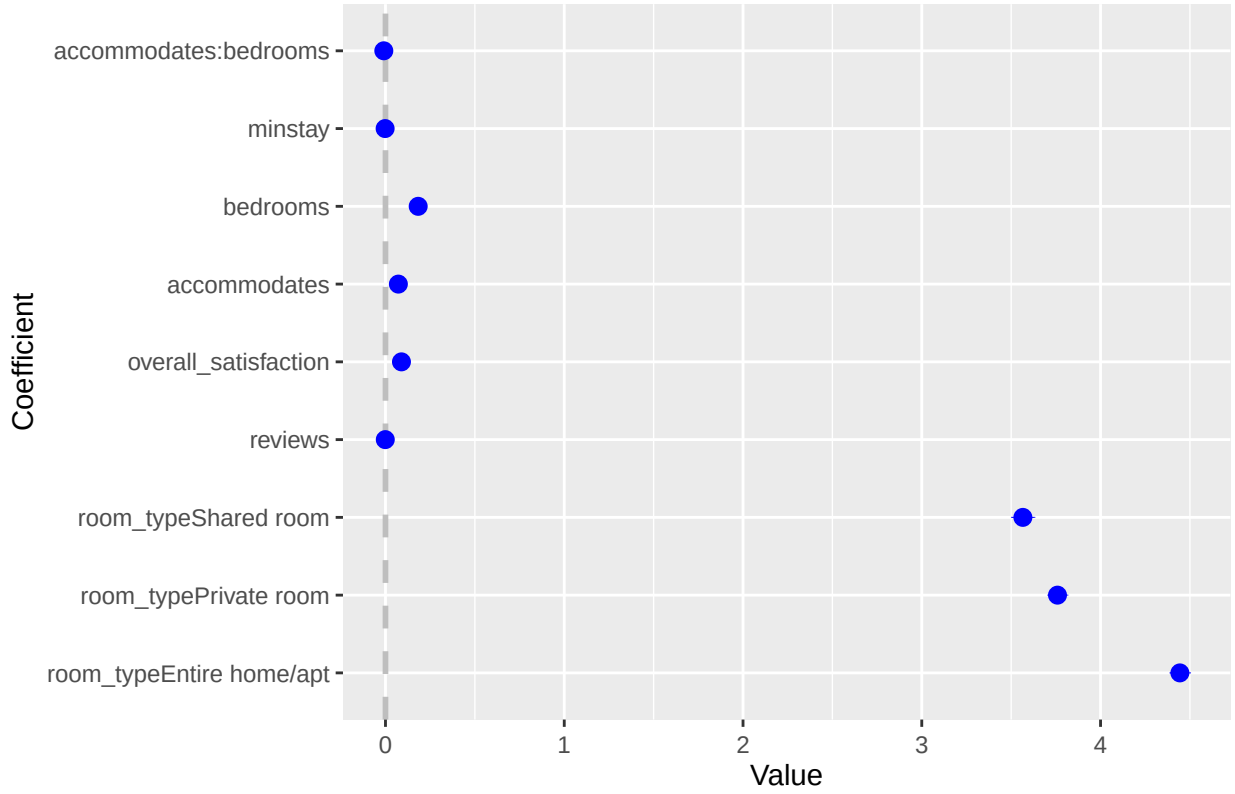

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(logprice ~ room_type + reviews + overall_satisfaction + accommodates * b ...
2631
−4
−2
0
2
4
−5
0
5
10

Fig 9.Coefficient plot for model 1



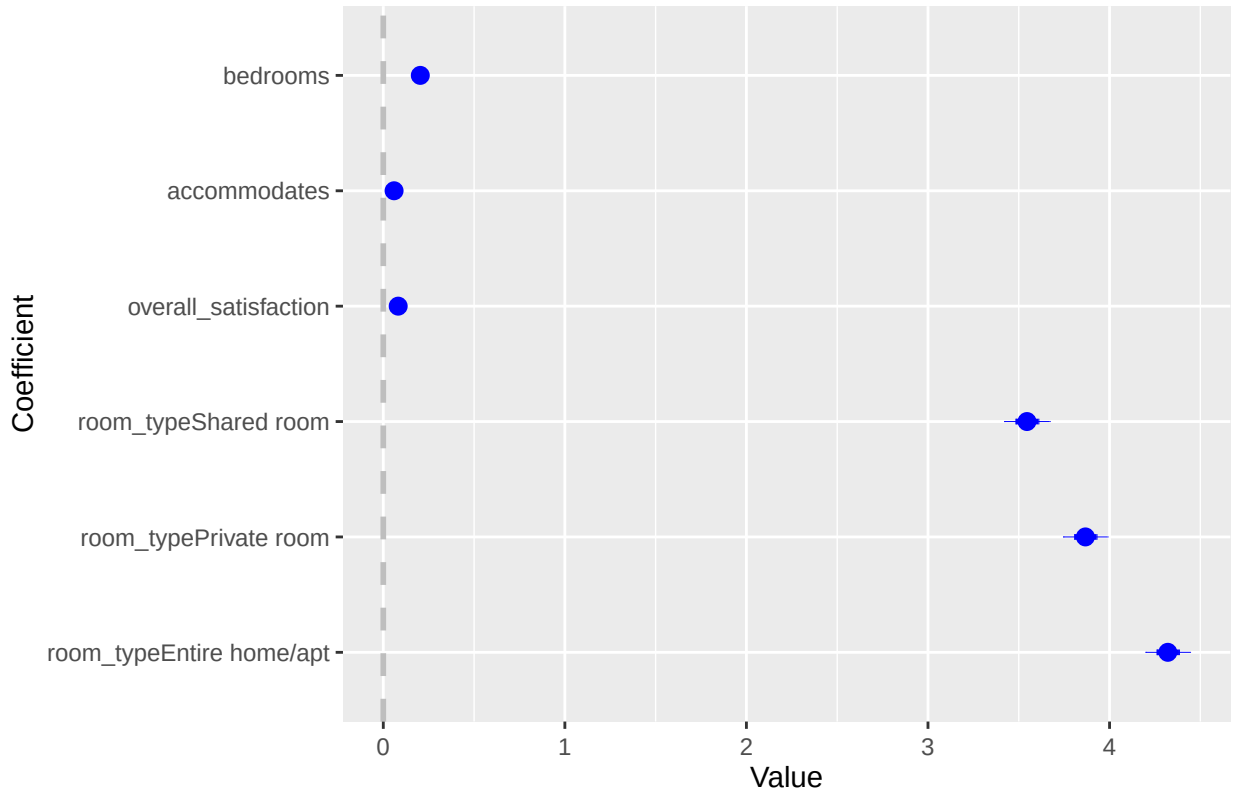
From this simple linear regression, I found all the predictors are significant. Also the R-square in the model is 0.57, So the model is not well fitted. However, in the residual plot, there are some points having big residual: 18772, 2680 and 2631. Those prices are very high that lead to huge residuals.The rest of points are symmetrically distributed around the line $h = 0$. In the QQ plot, we can see most dots in the middle falls on the line. However, the data have more extreme values on the tail of the distribution. So the model overestimates the low values and underestimate the high values.

Though all the coefficients are significant in previous summary checking. The coefficient plot tell us the predictor: "reviews", "minstays" and correlation accommodates:bedrooms coefficient exactly fall on the zero point. We may want to eliminate those predictors in the multilevel models. Now let's expand simple linear model to multilevel linear model.

Model2: Multilevel linear model with random intercept:

$$log(price) = \alpha_i + \beta_1 x_{roomtype} + \beta_2 x_{overallsatisfaction} + \beta_3 x_{accommodates*bedrooms} + \beta_4 x_{accommodates} + \beta_6 x_{bedrooms}$$

Fig 10.Coefficient plot for model 2

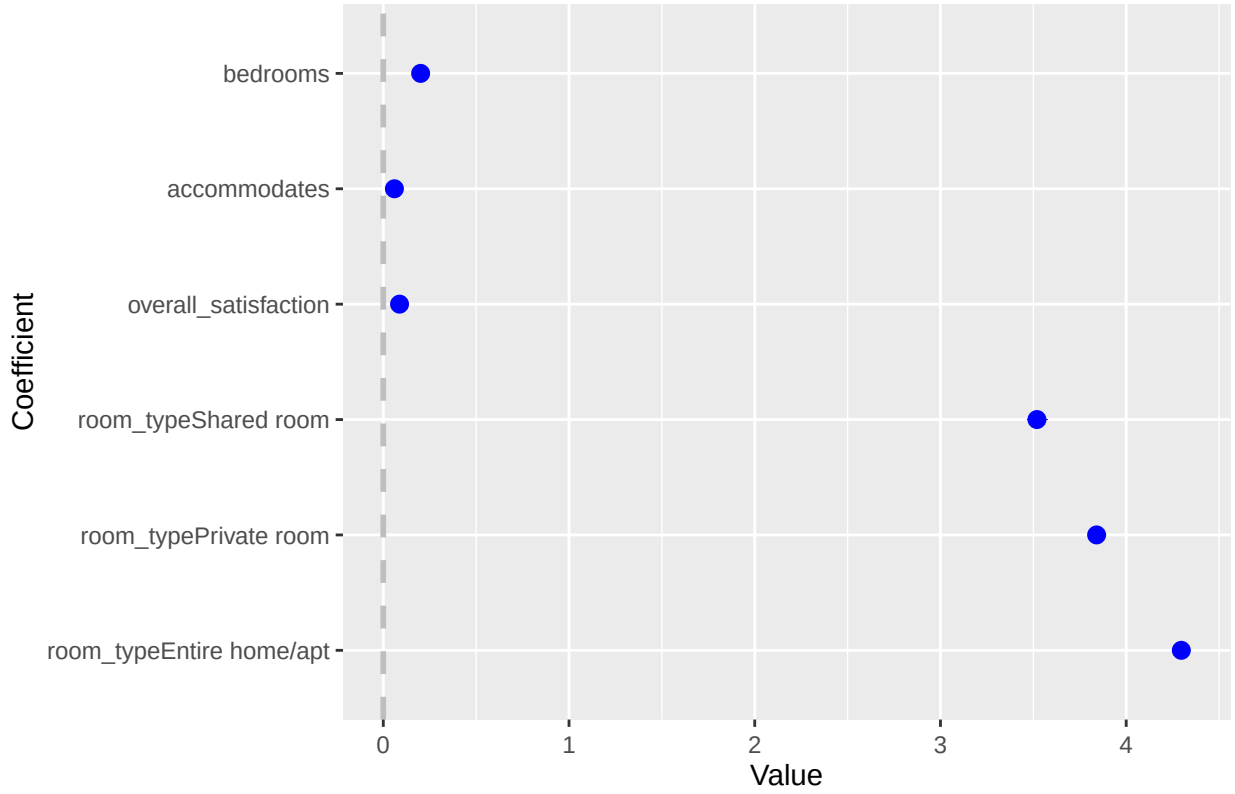


This model gets rid of a few non-significant terms we discussed previously such as interaction between accommodate and bedrooms. In the coefficient plot, all the predictors are significant. And the facotor of room type, room type plays the most important part in the model. Besides, the room types the number of bedrooms is the second influential term in this model

Model3 : Multilevel linear model with random slope:

$$log(price) = \alpha + \beta_1 x_{room_type} + \beta_{2[i]} x_{overallsatisfaction} + \beta_3 x_{accommodates} + \beta_4 x_{bedrooms}$$

Fig 11.Coefficient plot for model 3

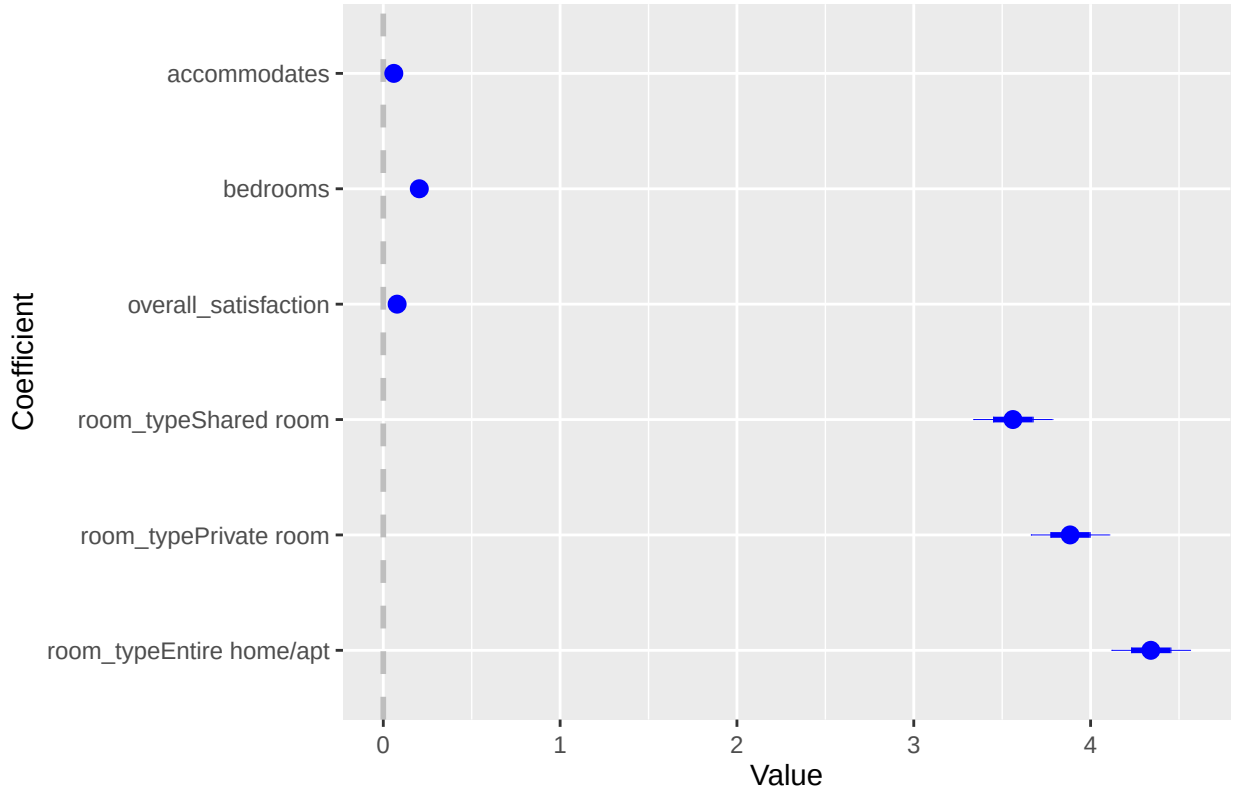


This is a multilevel linear model with random slope. As we can see from the coefficient plot, all the predictors are significant.

model4 : Multilevel linear model with random slope and random intercept.

$$log(price) = \alpha_i + \beta_1 x_{room_type} + \beta_{2[i]} x_{overallsatisfaction} + \beta_3 x_{accommodates} + \beta_4 x_{bedrooms}$$

Fig 12.Coefficient plot for model 4



This is the model is based on model2 and model3 with random slope as well random intercept. Again from the coefficient plot, all the predictors are significant.

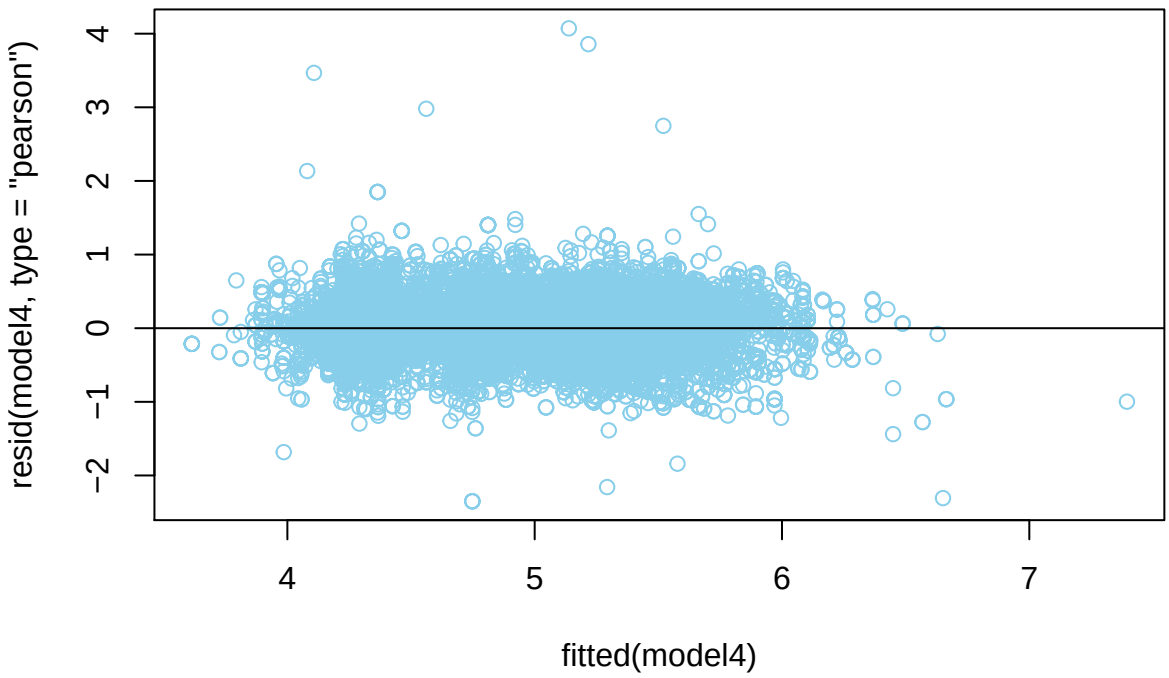
IV. Result:

Model Choice and Interpretation:

Since we have 3 multilevel models with similar structures, we want to run ANOVA test to test whether there's any difference among the models and which model has best goodness of fit.

```
## Data: df
## Models:
## model2: logprice ~ room_type + overall_satisfaction + accommodates +
## model2:     bedrooms + (1 | neighborhood) - 1
## model3: logprice ~ room_type + overall_satisfaction + accommodates +
## model3:     bedrooms + (0 + overall_satisfaction | neighborhood) - 1
## model4: logprice ~ room_type + overall_satisfaction + bedrooms + accommodates +
## model4:     (1 + overall_satisfaction | neighborhood) - 1
##        Df   AIC   BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## model2  8 16536 16601 -8259.8    16520
## model3  8 16661 16726 -8322.4    16645   0.0      0          1
## model4 10 16313 16395 -8146.5    16293 351.8      2     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig 13. residual plot for model4



According to the output of the test, three models are highly different, because the p-value is less than 0.05. Also, we found that model 4 with random intercept and random slope is the best fit among three multilevel models. It has lowest deviance with 16293. The second plot is a residual plot form model 4. As we can see the plot is pois are symmetrically distributed around the line h = 0(you can check the rest of residual plots in the appendix). The neighborhood with the maximum intercept is bay village with intercept 0.7954577

From the analysis above we found the best model among is the follow one:

$$log(price) = \alpha_i + \beta_1 x_{room_type} + \beta_{2[i]} x_{overallsatisfaction} + \beta_3 x_{accommodates} + \beta_4 x_{bedrooms}$$

The strategy of this model is predicting models by different levels of neighborhood. The model in each level will contain a unique intercept and a unique slope for the predictor overall satisfaction. This structure of will lead this model to minimize the deviance comparing to the other two multilevel linear model. From the coefficient plot, the most influential term is the factor of room type. The second influential term is the district of neighborhood. Also, more bedrooms and higher overall satisfaction will lead to a higher price.

V. Discussion:

Implication

The result of modeling is what we expected at initial. The factor of room type will determine the number of customers served. For example, entire home/apt tends to serve more people at one time. So the price of entire home/apt should be higher than the other two room type. Also, the second term neighborhood is obvious to be significant. This is because different districts have different environment condition. For example, Bay village is close to downtown area and there are many sight-seeing spots for tourists. We can conclude that our findings are reasonable.

Limitation

First, the dataset is limited in the Boston area. So, it is not enough to predict the Airbnb room price in other area. Second, there are only 4 predictors in the multilevel model. Those predictors are the most significant variables we found in the dataset. However, there are many other factors that may influence the price of rooms. For example, the crime rate in the neighborhood or the transportation condition. Those could be potential influence terms.

Future direction

In the future, I will try to include more predictors with a bigger dataset. This will help predict the price more precisely and find more factors that influence the room price.

VI. Acknowledgement

Special thanks to Professor Yajima's suggestion on model fitting and solutions to the problem "do not converge" when I was fitting a multilevel model. And thank the data provider on tomslee's website.

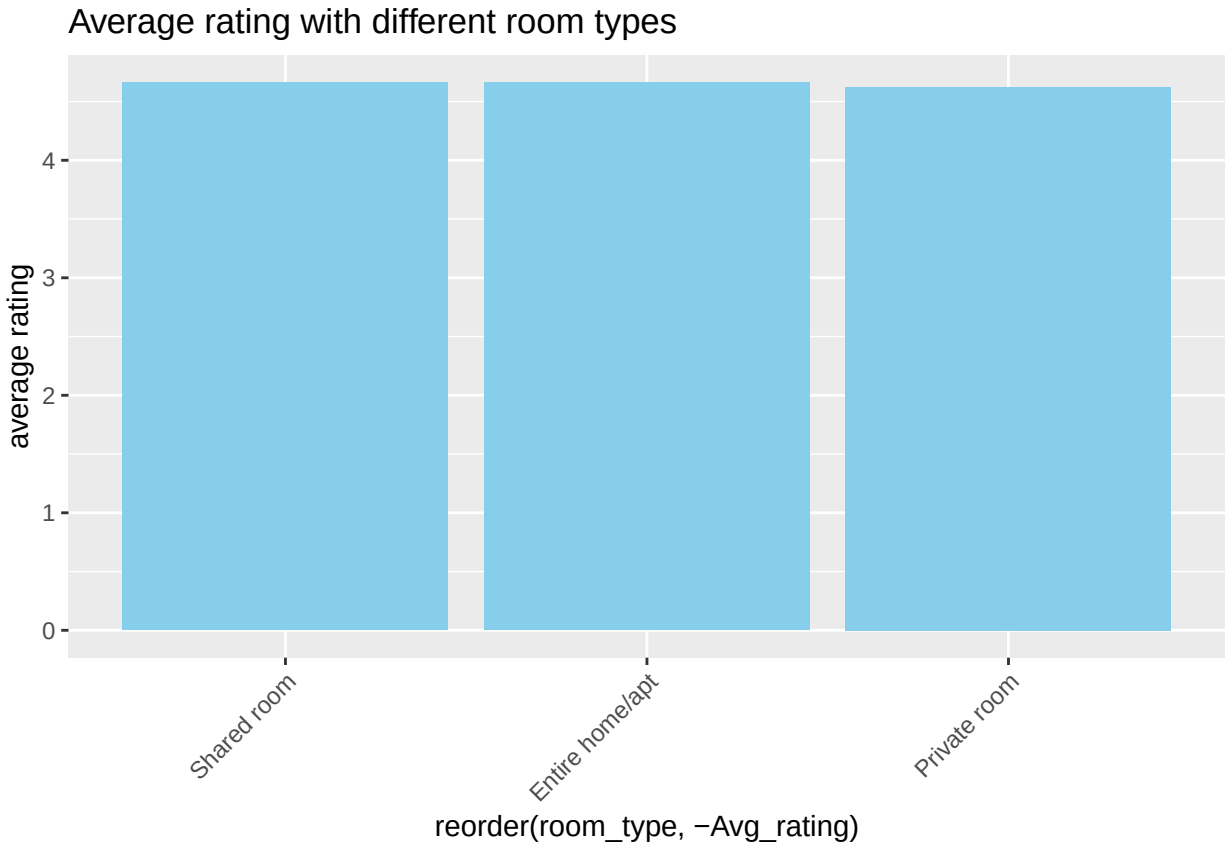
VII.Reference

http://tomslee.net

https://en.wikipedia.org/wiki/Airbnb

https://www.airbnb.com/

VIII. Appendix



The graph above shows the average rating with different room types. However we can't tell any difference

among those three types. This is because the overall rating in Boston area are all very hight(above 4.5).