Boston Airbnb Project

Haoyu Li

12/7/2020

Project Overview

When holiday and vacation come, people would like to travel around the country more than usual. The way people choose their accommodations during vacation is no longer limited to hotel in recent years because of the rising company Airbnb. Airbnb provides accommodations for people that want to stay at a place that makes them feel like home when they travel. On Airbnb website, there are tons of hosts that provide different information regarding their places so that people can evaluate these information and choose the place that meet their requirements. For this project, we focus on predict the price of different places using information that are provided in our data. The project consists of those parts: data cleaning and exploratory data analysis, data modeling, and model checking.

1.Data Introduction & Cleaning

The data we use in this project included detailed information of hundreds of thousand Airbnb listing places in city of Boston and it also list out the neighbourhoods in which Airbnb listings are located. The detailed information includes You can view our data in the Appendix I-Data overview. For the data cleaning part, the data we found from the website "http://insideairbnb.com/get-the-data.html" has already been cleaned up, so our work would focus on the correct format of the varialbes. Since the Airbnb price is our response variable, we checked "price" variable and transform into the numerical variable using "as.numerical()" function. You can see the details in Appendix-Data cleaning. Next, we perform exploratory data analysis.

2.Exploratory data analysis

After we cleaned our data, we looked into our dataset and picked up some variables we think will have affect on the price predicting value. Those are: bedrooms, accommodates, room_type, number_of_review. We also use neighbourhood as our groups to fit the multilevel model that we need to predit Airbnb price. The distribution of price is one of the most important data analysis we need to look at. Below the Figure 2 is the distribution of Airbnb price.

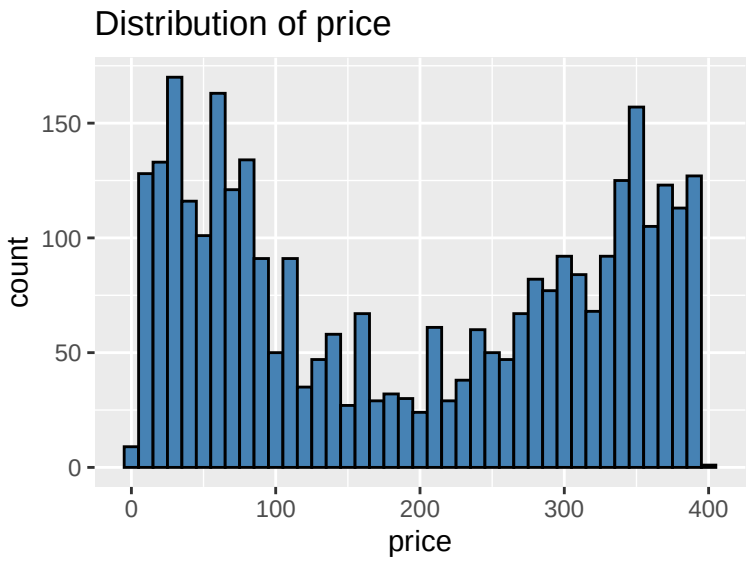


Figure 2: Distribution of price

We find that the distribution of price is skewed on both end sides, we might consider log transform the price and to use log price to improve the accuracy of our model. The distribution of log price is displayed in Appendix.

3.Modeling & Methods

Model choices

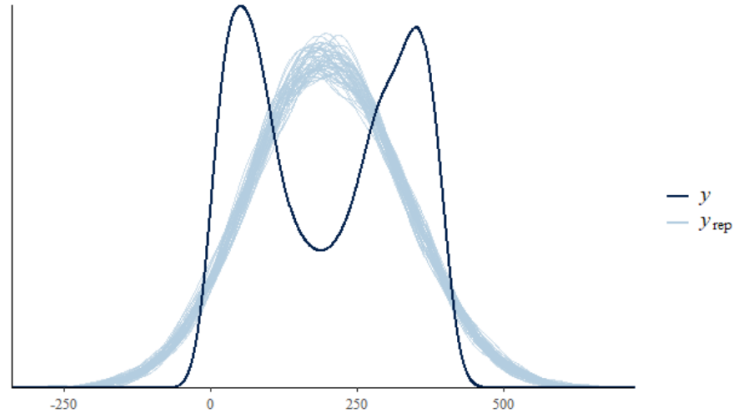
Multilevel linear model with varying intercept

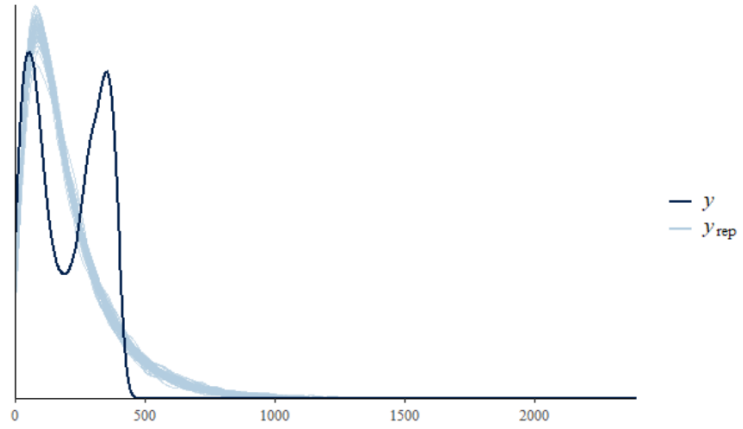
$$price = \alpha_i + \beta_1 x_{bedrooms} + \beta_2 x_{accommodates} + \beta_3 x_{numberofreviews} + \beta_4 x_{roomtype}$$

Multilevel negative binomial model with varying intercept and varying slopes

$$log(price) = \alpha_i + \beta_{1[i]} x_{bedrooms} + \beta_{2[i]} x_{accommodates} + \beta_3 x_{roomtype}$$

Below is the posterior predictve model check from first model to second model.





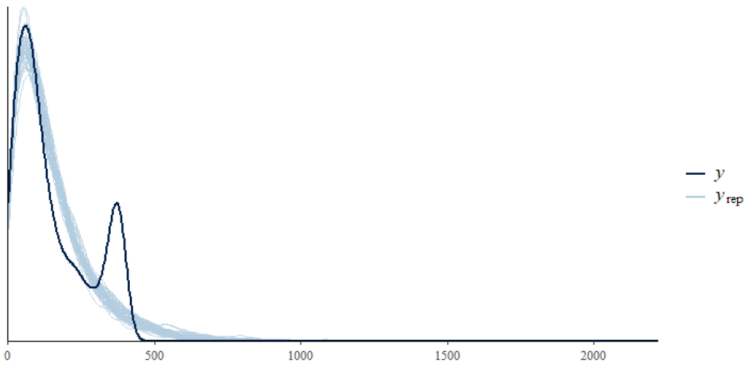
From the first model to second model, we improve our model by moving from multilevel linear model to multilevel possion/negative binomial model, which means we use log transformation to improve model accuracy, from the model check we can see a obvious model accuracy has improved.

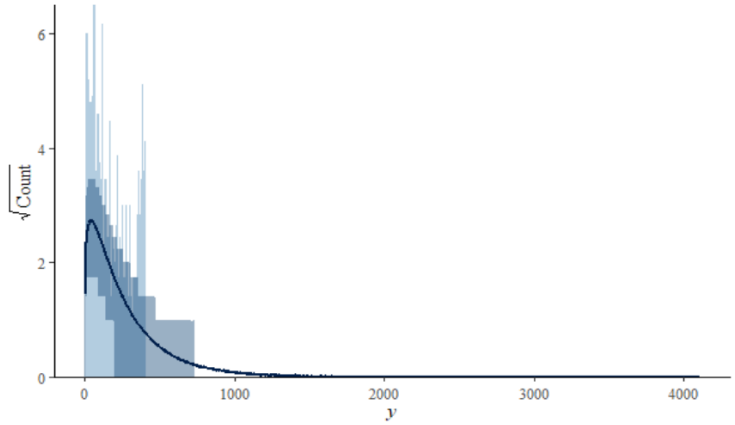
Multilevel negative binomial model focus on one room type

However, the problem with both models is that there are two modes there, it might be caused by an usual large number of one variable. When we look into the distriution of room types, we find that the numbers of each room type are pretty uneven, so we decide to focus on one room type, which is entire home/apt, then we fit our model.

$$log(price) = \alpha_i + \beta_{1[i]}x_{bedrooms} + \beta_{2[i]}x_{accommodates}$$

Below is the posterior predictive model check and rootogram





4.Discussion

Assess result of improved model

From the posterior check, we can see on the lower side, the accuracy of our model improves significantly, but one the higher side, we still have discrepancy between observed value and predict value. We can also see the same problem from our rootogram, there are still some data on the high end of range that our model does not capture.

Limitation

One thing we might consider why our model has the problem that we mention above is to look at the distribution of the log price. On the high log price, the frequency is extremely high compared to lower log price. It makes our model hard to capture those values due to the unusual pattern in our response variable.

Further direction

For our further research, we could look into how to deal with the unusual pattern in small proportion of Airbnb price, and we can also consider include more predictors to improve the accuracy of our model.

Reference

Data source: http://insideairbnb.com/get-the-data.html

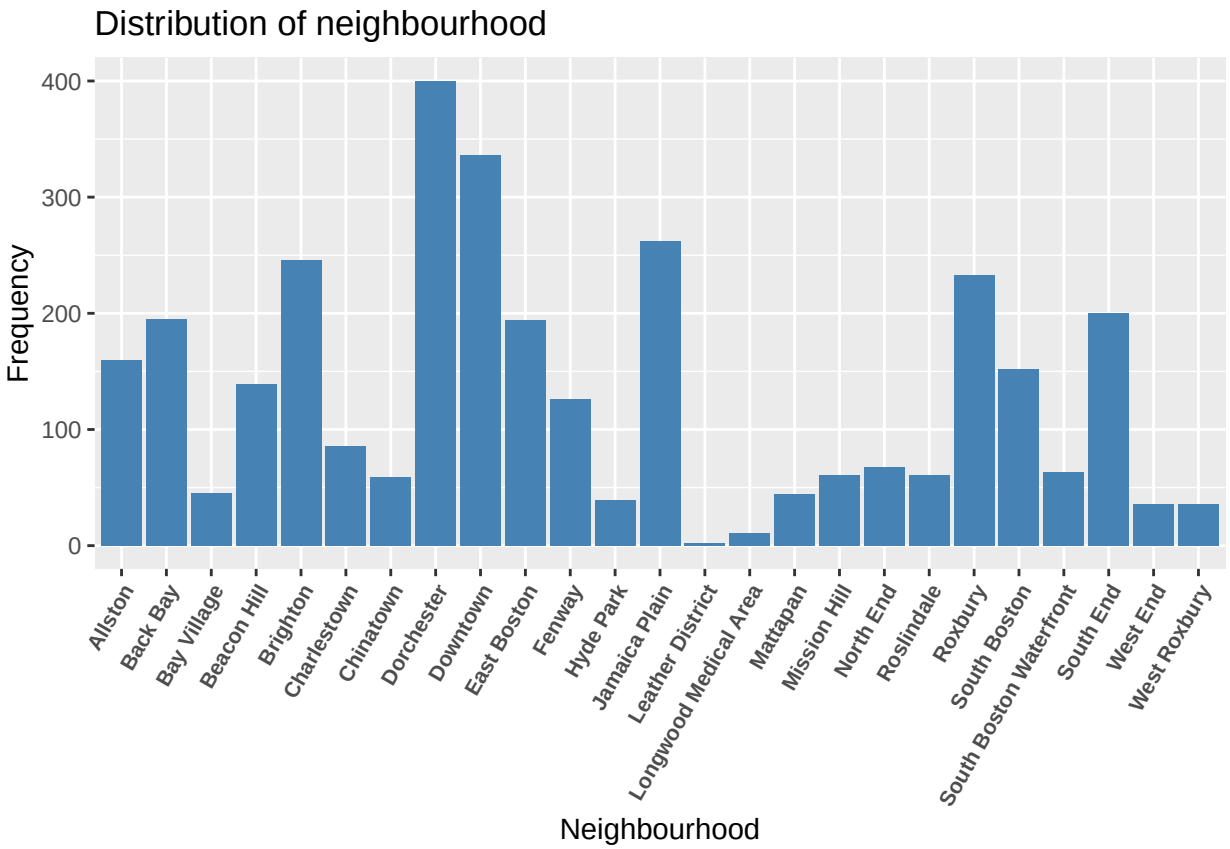
Appendix

Part I:Data Cleaning

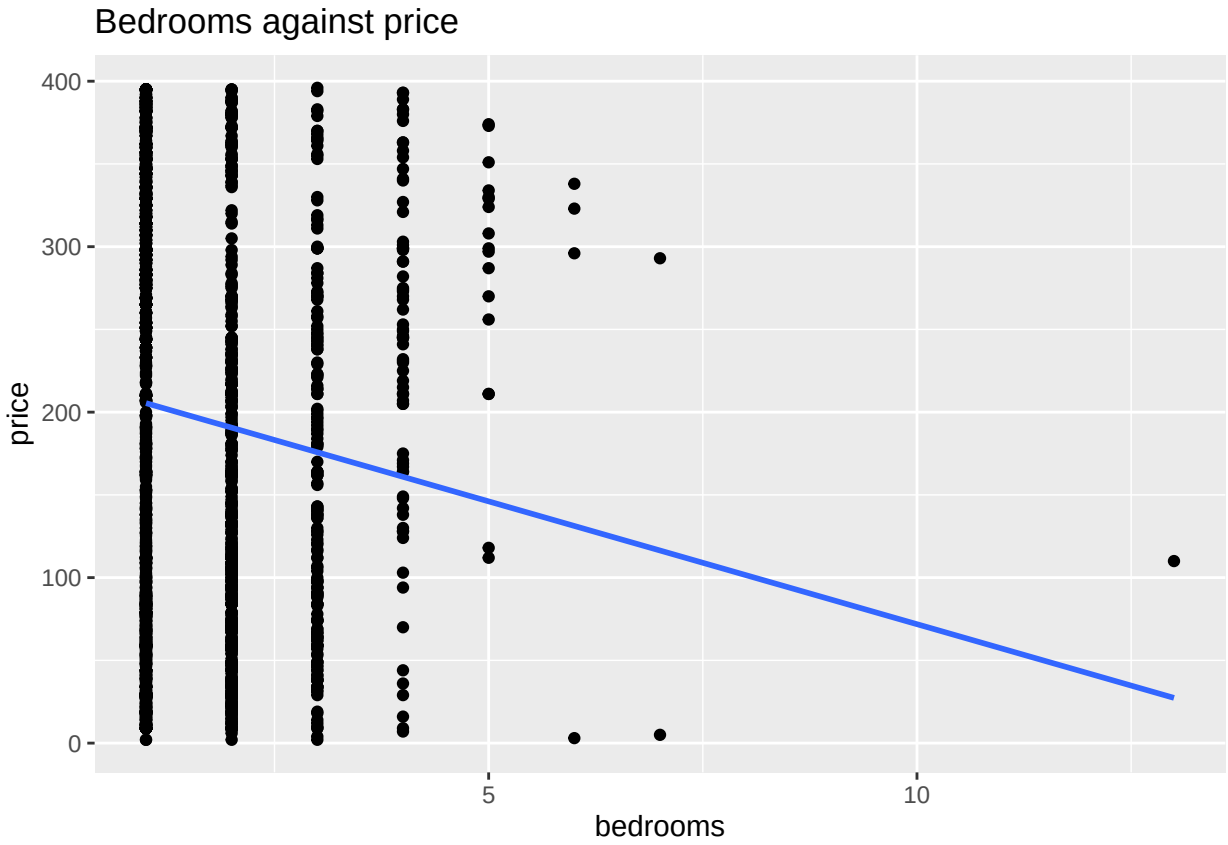
Part II:EDA

Distribution of Neibourhood

```
## # A tibble: 25 x 2
##    neighbourhood_cleansed number_of_neighbourhood
##    <fct>                                   <int>
##  1 Dorchester                                400
##  2 Downtown                                  336
##  3 Jamaica Plain                             262
##  4 Brighton                                  246
##  5 Roxbury                                   233
##  6 South End                                 200
##  7 Back Bay                                  195
##  8 East Boston                               194
##  9 Allston                                   160
## 10 South Boston                              152
## # ... with 15 more rows
```


Distribution of neighbourhood
Frequency
400
300
200
100
0
Allston
Back Bay
Bay Village
Beacon Hill
Brighton
Charlestown
Chinatown
Dorchester
Downtown
East Boston
Fenway
Hyde Park
Jamaica Plain
Leather District
Longwood Medical Area
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End
West End
West Roxbury
Neighbourhood

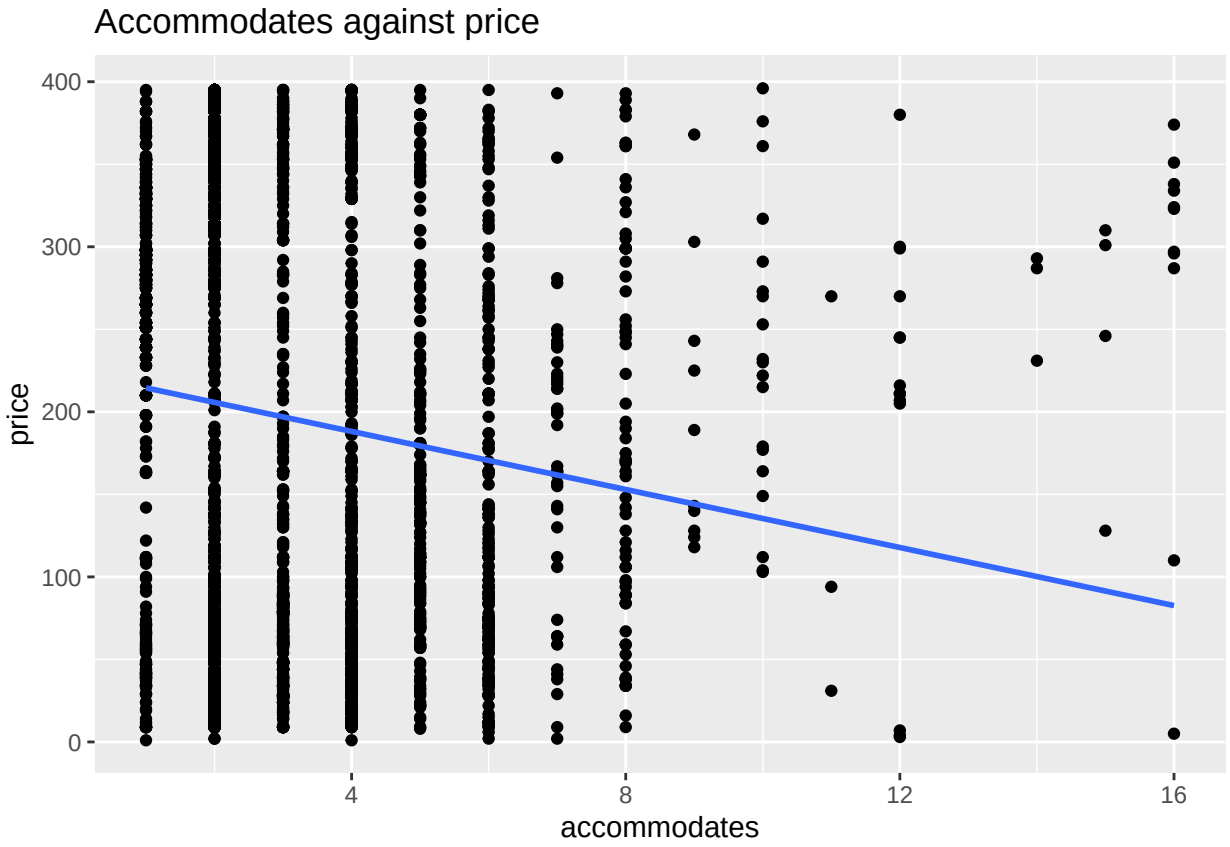
Relationship between price and bedrooms



We can see a clear linear relationship between price and bedrooms

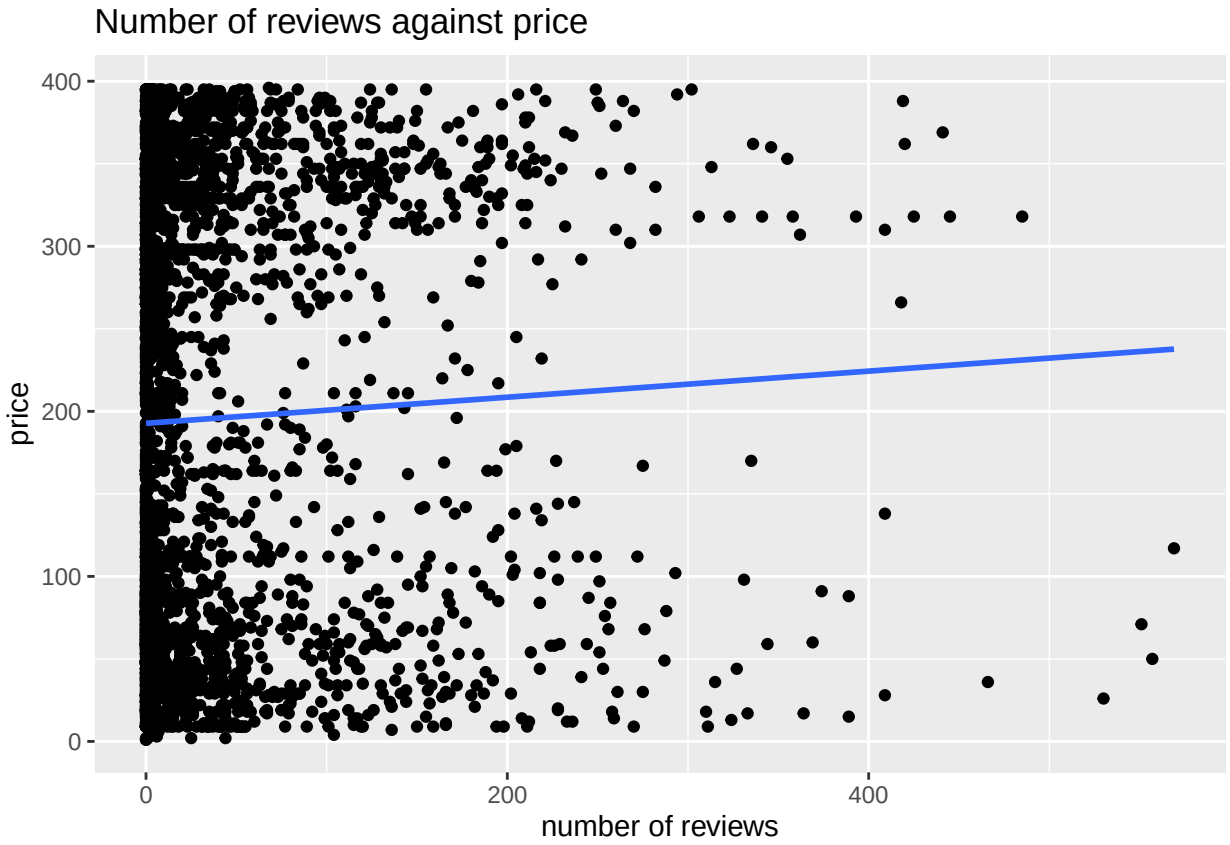
Relationship between price and accommodates

The variable accommodates means the number of people an Airbnb place can hold.



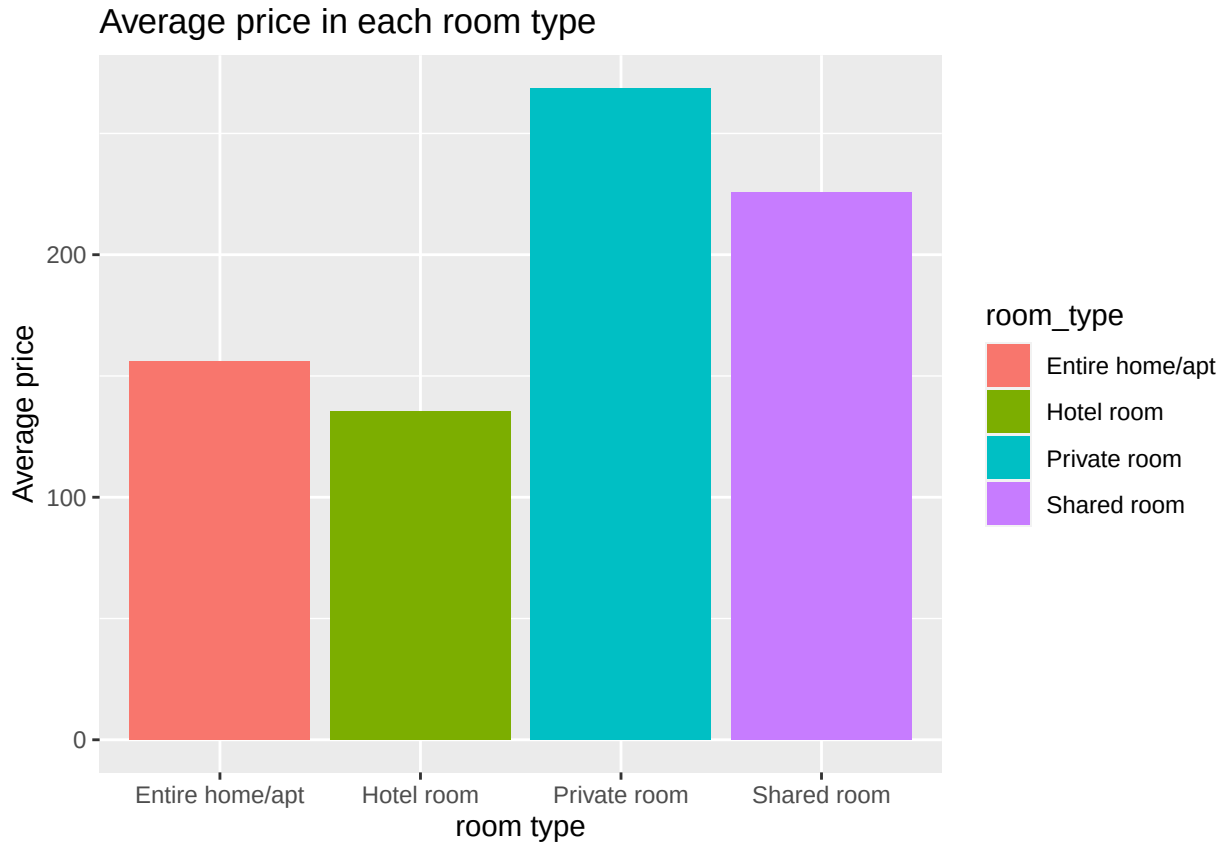
We can see a clear linear relationship between price and accommodates.

Relationship between price and number of reviews



We might be careful include number of reviews in our model since the relationship between number of reviews and price is not obvious compared to bedrooms and accommodates.

Average price in different room type

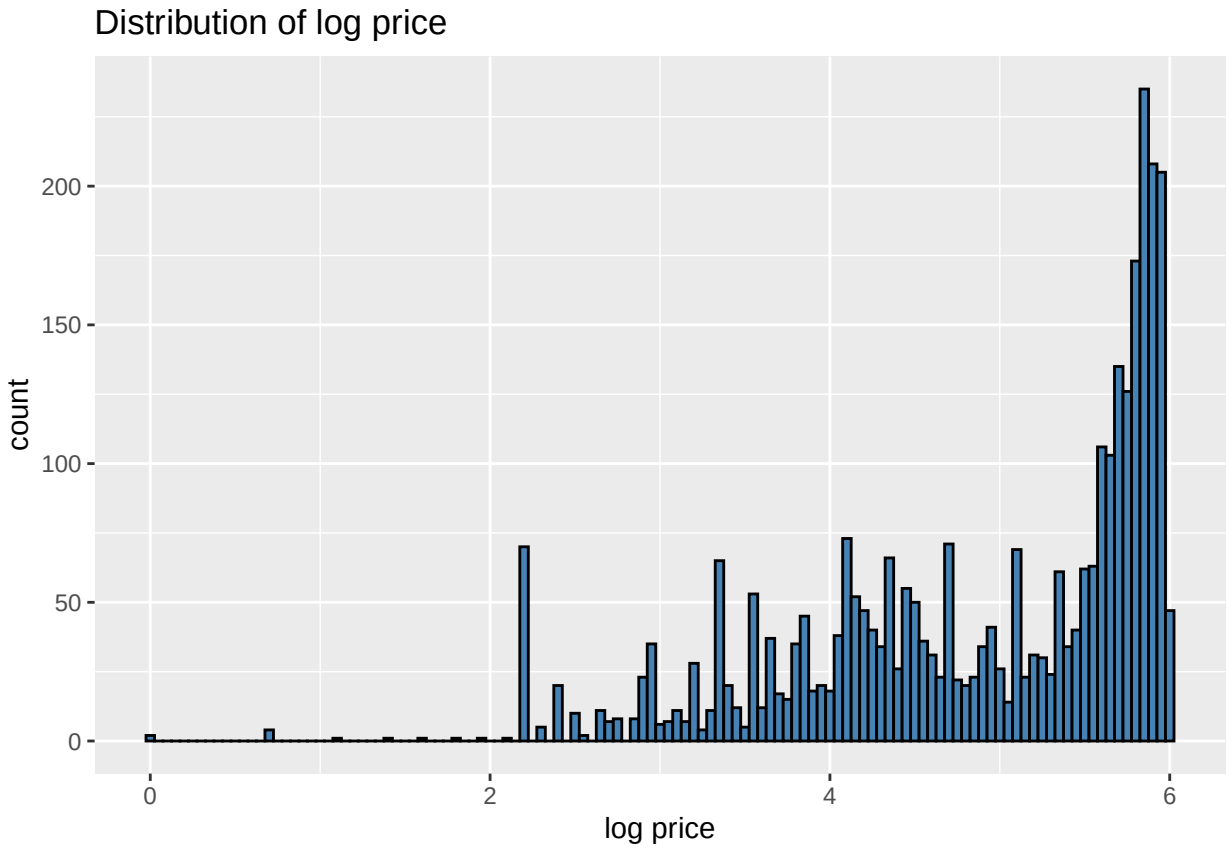


We can consider include room type as predictor in our model since we see different room type has relatively differenr average price.

Plot lienar model between price and review scores

There is no such significant relationship between review scores rating and price, we choose to ignore this variable.

Distribution of log price



Part III: Models choices

Before narrow down to one room type

```
### Multilevel linear model with varying intercept
boston_fit_1<-stan_lmer(price~bedrooms+accommodates+room_type+number_of_reviews+
(1|neighbourhood_cleansed),data=model_boston)

### Multilevel negative binomial model with varying intercept and slopes
boston_fit_4<-stan_glmer(price~bedrooms+accommodates+room_type+
(1+bedrooms+accommodates|neighbourhood_cleansed),
data=model_boston,family = neg_binomial_2(link = "log"))
```

After narrow down to one room type

```
### Multilevel negative binomial model with one varying slope and one varying intercpet
boston_fit_entire_3<-stan_glmer(price~bedrooms+accommodates+
(1+bedrooms|neighbourhood_cleansed),data=model_boston_entire,family = neg_binomial_2(link="log"))

### Multilevel negative binomial model with two varying slope and one varying intercpet
```

```
boston_fit_entire_4<-stan_glmer(price~bedrooms+accommodates+
(1+bedrooms+accommodates|neighbourhood_cleansed),
data=model_boston_entire,family = neg_binomial_2(link="log"))
```