Data Analysis for Airbnb Boston

Ruiyi Lian

2020/11/11

Abstract

This project aims to analyze a large dataset released by official website of Airbnb company. The project is divided into four parts: introduction, data cleaning, EDA, modeling, and conclusion. The Introduction part talks about the background and the goals of the project, and also includes the explanation of the dataset and variables. The data cleaning part is mainly about filtering and deleting the bad data or useless data. The EDA(exploratory data analysis) aims to analyze the dataset in order to summarize its main characteristics with visual methods. The modeling part includes building model and model selection. The conclusion part includes result and the discussion, so that I can talk about the what I found in this project and what I will do for next step.

Introduction

Background

Airbnb,Inc is an American company that provides a platform for hosts to accommodate guests with short-term lodging and tourism-related activities. The guests can choose the lodging depending on various elements such as location, equipment, environment and ect. The hosts can provide other services or equipment to attract the guests, in addition to satisfying the standard requirement of Airbnb. Certainly, for the hosts, the pricing is their prior consideration. The pricing is fluctuated by location, environment and many other elements. However, unlike some hotels, the price of the rooms are decided by the hosts who sometimes relatively lack the comprehensive information.

Goals

The goal of the project aims to provide an appropriate model which shows relationship between pricing and other elements such as location, reviews, room type and ect.

Data Explanation

The data of Airbnb Boston is downloaded from Airbnb get-data website(http://insideairbnb.com/get-the-data.html). Since I would like to analyze the data for a year, I chose the monthly listing data from October 2019 to October 2020. Then, I combined these 13 csv files into one dataset named Airbnb. The table below shows an example of the data:

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitu
3781	HARBORSIDE-Walk to subway	4804	Frank		East Boston	42.36524	-71.029
5506	**$99 Special ** Private! Minutes to center!	8229	Terry		Roxbury	42.32981	-71.095
6695	$99 Special!! Home Away! Condo	8229	Terry		Roxbury	42.32994	-71.093
6976	Mexican Folk Art Showcase in Boston Neighborhood	16701	Phil		Roslindale	42.29244	-71.135
8789	Curved Glass Studio/1bd facing Park	26988	Anne		Downtown	42.35919	-71.062
9273	Stay at "HARBORVIEW" Walk to subway	4804	Frank		East Boston	42.36461	-71.029

There are 14 variables in dataset Airbnb, and the meaning of 14 variables are below:

- id: the unique identification number for each listing
- host_id: the unique identification number for each host
- host_name: the name of each host
- neighbourhood: the geographically localised community in Boston that each listing located in
- latitude/longtitude: the detailed location of the listing
- room_type: the type of room of listing(Entirehoom/apt, Hotel room, Private room, Share room)
- Price: the price(in dollars) of each listing per night
- minimum_nights: the minimum night per booking, required by host.
- number_of_reviews: number of reviews from costumer for the listing.
- last_review: date of last review for the listing
- review_per_month: number of customer reviews per month
- calculated_host_listings_count: the number of listings belonged to a host.
- availability_365: the number of days a listing is available in a year (365 days), pre-posted by host.

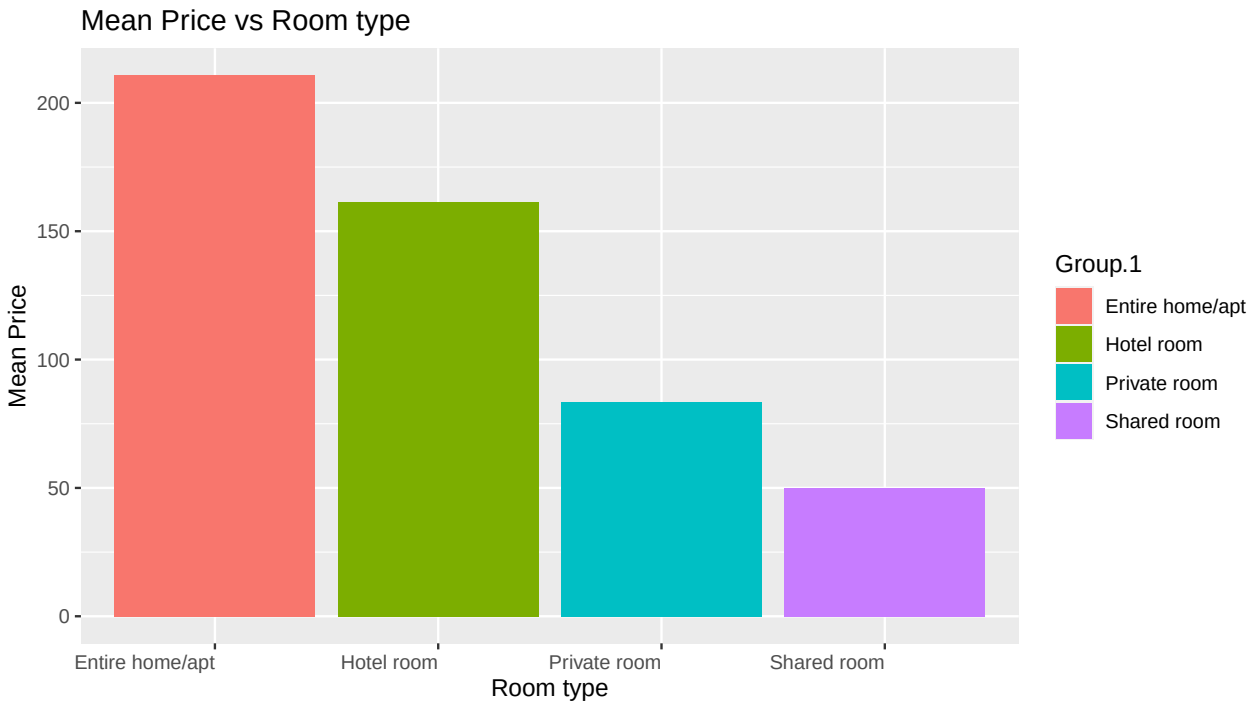
Data cleanning

For the data cleanning, I would like to deleted the useless column: name, neighbourhood_group. Also, I deletedna value. Then, I deleted some rows whose price are higher than 2500. Since these prices are much higher than other price, the result of regression may be influenced. (detailed information see the Appendix)

EDA

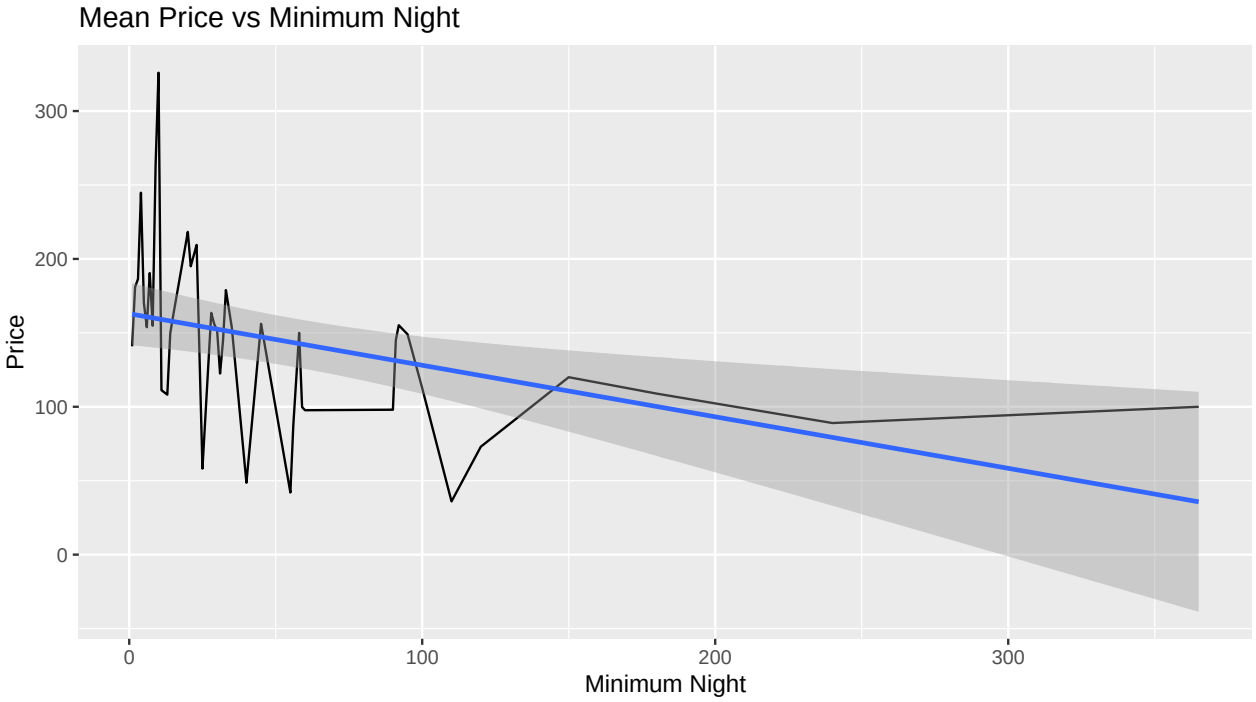
Because of the limitation of the page, more EDA shows in the More EDA part in Appendix.

Mean Price vs Room type

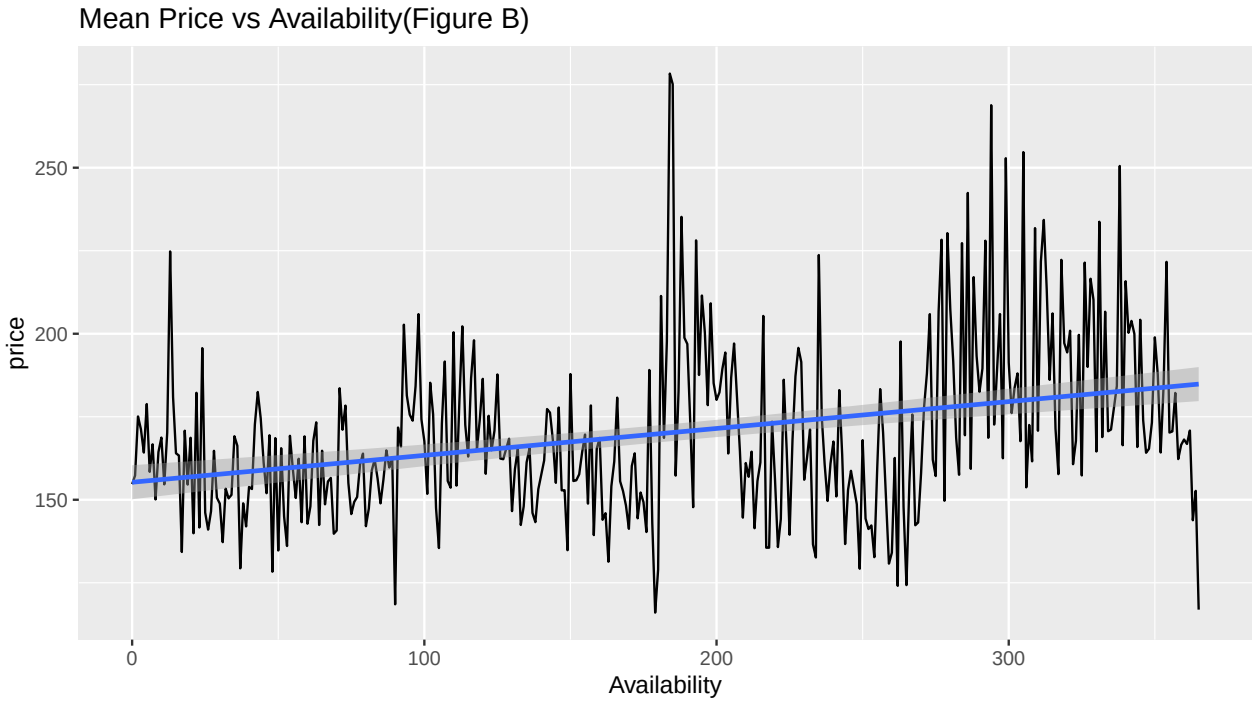


The barplot above shows the type Entired home/apt accounts for the largest proportion among the four room types.

Mean Price vs Minimum Nights



Mean Price vs Availability 365



The plot shows that there exist a positive relationship between mean price and availability.

Modeling

Since my goal is to analyze relationship between price and other variables, I chose price as response variable (dependent variable), and neighbourhood,room type,number of reviews,minimum night,availability 365 as

predictor variables (independent variable).

Multilevel Regression Model

Before fitting the multilevel regression model, I fitted two linear regression model fit1 and fit2 (See Model part in Appendix) Since the result of linear regression fit1 shows some level of neighbourhood are not statistically significant, I would like to fit a multilevel linear regression model. Multilevel regression model were designed to analyze data generated from a nested structure because conventional linear regression models underestimate standard errors and, in turn, overestimate test statistics. I used log transformation to transform the response variable price and vary the intercept for neighbourhood.

```
# Since th 0 price exist, I define a log price without 0 price.
airbnb_log_price <- airbnb
airbnb_log_price <- filter(airbnb_log_price, airbnb_log_price$price > 0)
airbnb_log_price$log.price <- log(airbnb_log_price$price)

fit3 <- lmer(log.price ~ room_type + reviews_per_month + minimum_nights + availability_365 +  (1|neighb
display(fit3)
## lmer(formula = log.price ~ room_type + reviews_per_month + minimum_nights +
##     availability_365 + (1 | neighbourhood), data = airbnb_log_price)
##                         coef.est coef.se
## (Intercept)              5.21     0.05
## room_typeHotel room     -0.24     0.03
## room_typePrivate room   -0.77     0.01
## room_typeShared room    -1.34     0.03
## reviews_per_month       -0.01     0.00
## minimum_nights           0.00     0.00
## availability_365         0.00     0.00
##
## Error terms:
##  Groups        Name        Std.Dev.
##  neighbourhood (Intercept) 0.23
##  Residual                  0.48
## ---
## number of obs: 38671, groups: neighbourhood, 25
## AIC = 53478.2, DIC = 53317.1
## deviance = 53388.6
```

Interpretion of coefficients: Fixed effect: For the room type, the model uses Entired room/apt as baseline category against which all other groups are measured. Remaining other variables unchanged, price of hotel room is 21% (1-exp(-0.24) = 0.21) less than entired room/apt. Price of private room is 54% (1-exp(-0.77) = 0.54) less than entired room/apt. Price of shared room is 74% (1-exp(-1.33) = 0.74) less than entired room/apt. For the review per month, remaining other variables unchanged, when one unit of reviews per month increases, the price will decrease by 1.4%. For the minimum nights, remaining other variables unchanged, when one minimum night increases, the price will decrease by 18% For availability 365, remaining other variables unchanged, when one days that a listing is available in a year (365 days) increases, the price will increase by 0.14%.

Random effect: The variance among neighbourhood is 0.05.

Model Selection

Among the fit1, fit2 and fit3, The fits3 is the best model. To prove that,I used AIC (Akaike Information Criterion) for selecting the model (see Model Selection in the Appendix). For further verification, I used residual plot and binnedplot which shows below.

The points in the residual plot are randomly dispersed around the horizontal axis. Also, the points in the

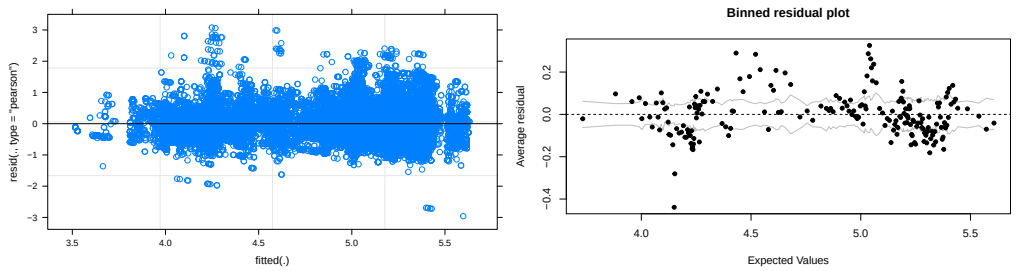


Figure 1:

binned plot are also randomly dispersed around the horizontal axis. Therefore, even though some points in the binned plot are beyond the 4 standard-error (-0.4 y-aixs in the graph), fit3 model is the most appropriate model for the data.

Conclusion

Result

The price of the Airbnb's room in Boston area is affected by location(neighbourhood), type of room, reviews, minimum night and availability. There exists a positive relationship between price and availability, which means when availability(the number of days a listing is available in a year) increases, the price will increase. Oppositely, there exists a negative relationship between price and some other variable which are review per month and minimum night. Setting A lower price by host may attract more booking and positive reviews. A room that setting a less minimum night for one booking maybe have a lower price. Also, the price varies among the different neighbourhood. Besides, the room that belongs to type of entire room or apartment has higher price than the room that belongs to the type of hotel room, private room and shared room.

Discussion

After the analysis, I found or still had some problems for the data. In the future, I want to solve these problems:

1. Some points in the binnedplot for fits 3 are still beyond the 4 standard deviation, which means that there may be some outliers in the dataset. I will try to do a more detailed process of data cleaning.
2. For the different neighbourhood or even the same neighbourhood, if there exists some other price distribution for the room. For example,in the same neighbourhood, if high-pricing room distribute in a certain neighbourhood. Drawing a room distribution map using R can display the price distribution visually better.
3. The price may be influenced by some other variables, therefore, I need to find a more detailed dataset.

Reference

1.https://machinelearningmastery.com/probabilistic-model-selection-measures/

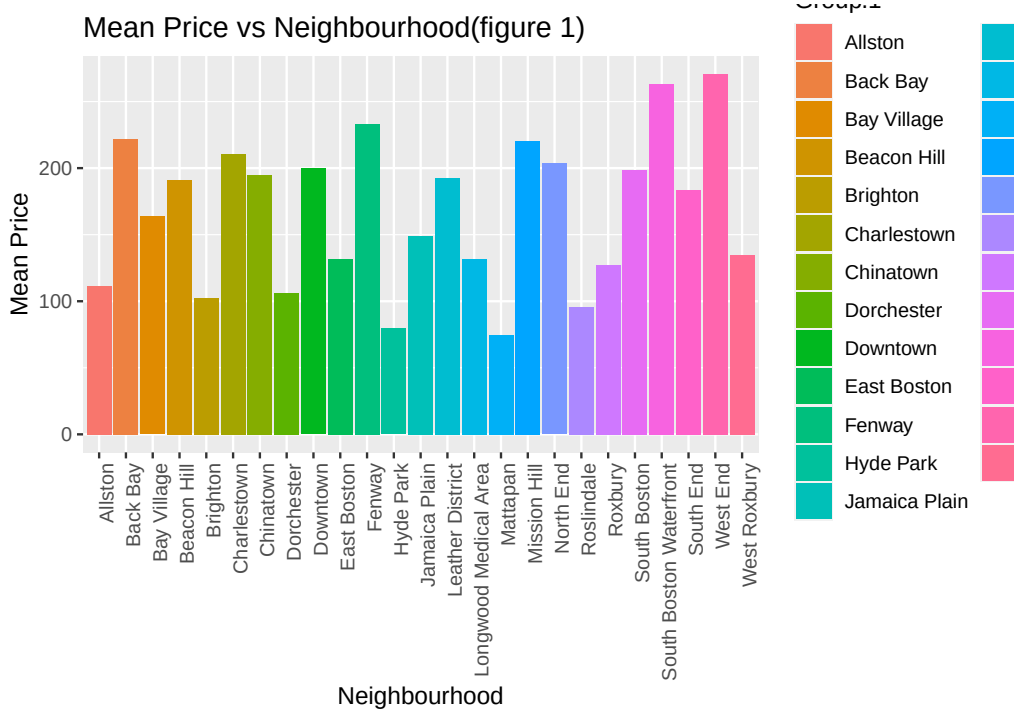
2.https://en.wikipedia.org/wiki/Exploratory_data_analysis#:~:text=In%20statistics%2C%20exploratory%20data%20analy

3.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1540459/

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

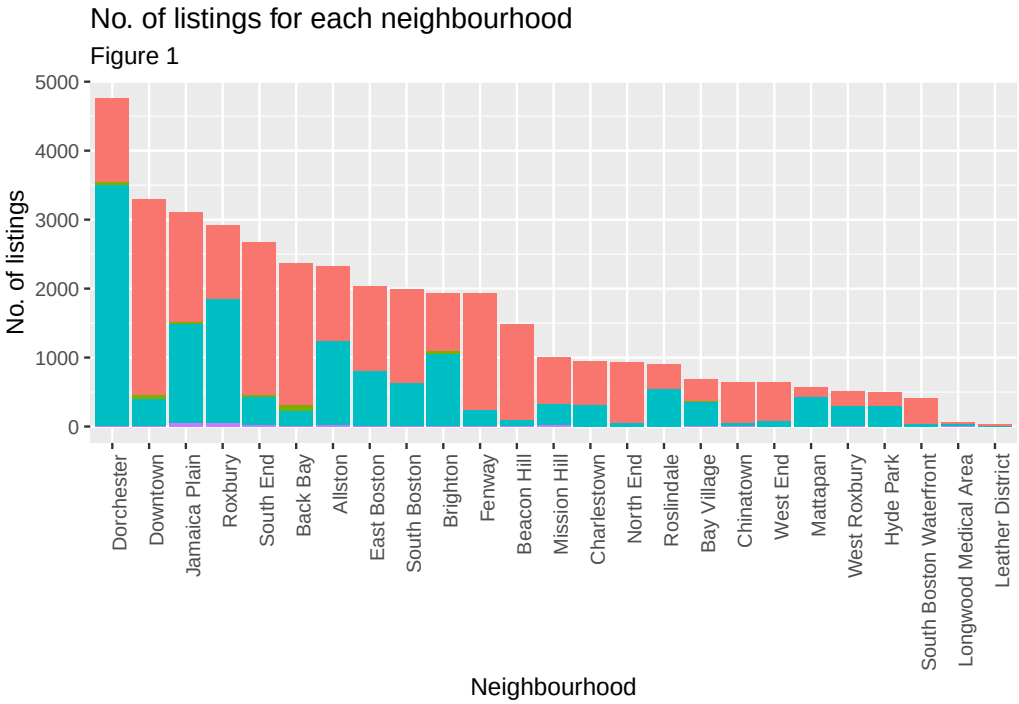
Appendix

More EDA



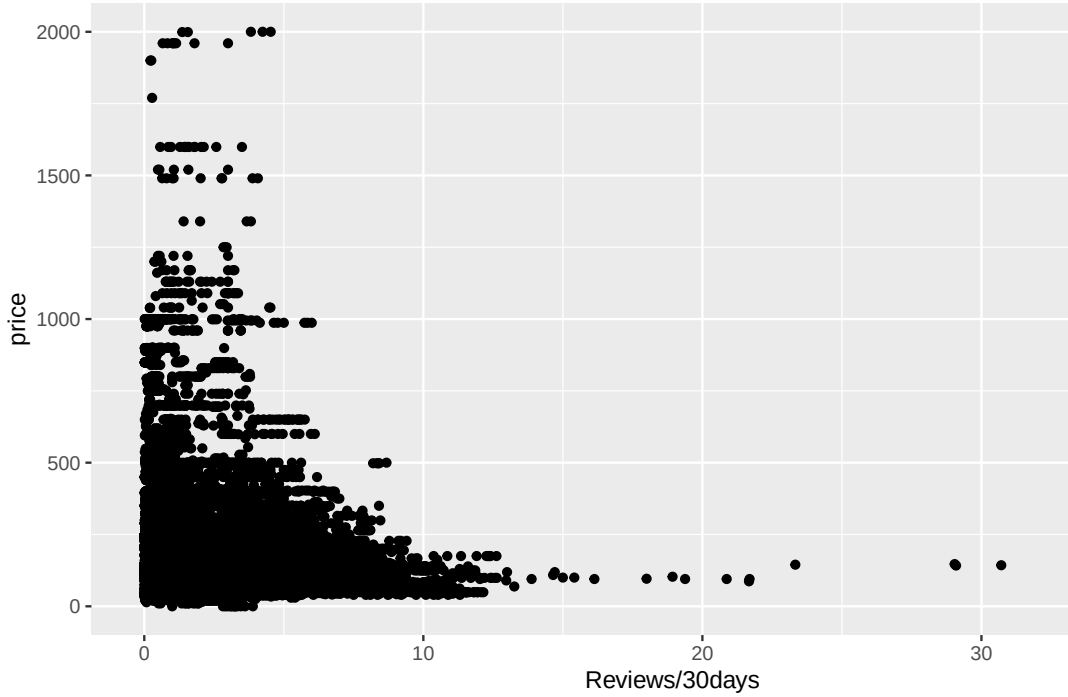
Mean Price vs Neighbourhood

The barplot above shows mean price of each neighbourhood. The West End neighbourhood has the highest mean price, and Mattapan has the lowest mean price.



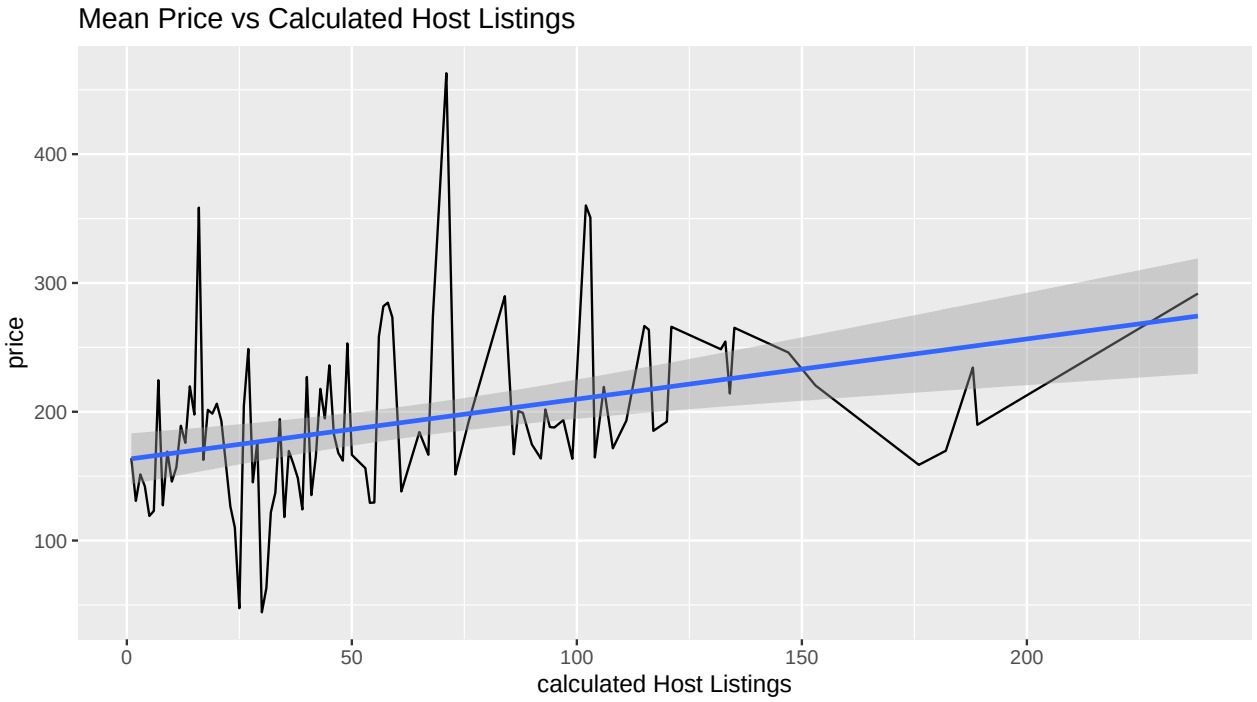
Neighbourhood

Figure 1 shows that Neighbourhood Dorchester has the largest amount of lodges among all the neighbourhoods, and the Neighbourhood Leather District has the smallest amount of lodge. Also, type Entired home/apt and type Private room accounts for the largest proportion among the four room types.



Price vs Numer of Reviews

The plot above shows that there are several points whose reviews per month are higher than 30. However, after check the table, there are only two rows that reviews per month are higher than 30. Therefore, I would not like to delete them, since they have no large influence on my analysis. Also. the plot shows that the the low-price-lodgings are more likely to gain higher reviews per month.



availability

Model

Simple Linear Regression Model Before fitting the multilevel regression model, I tried to build a simple linear regression model.

```
fit1 <- lm(price ~ neighbourhood + room_type + reviews_per_month + minimum_nights + availability_365,da
summary(fit1)
##
## Call:
## lm(formula = price ~ neighbourhood + room_type + reviews_per_month +
##     minimum_nights + availability_365, data = airbnb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -274.34  -57.57  -19.13   20.71 1802.44
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.819e+02  3.015e+00  60.339  < 2e-16 ***
## neighbourhoodBack Bay             5.900e+01  3.745e+00  15.757  < 2e-16 ***
## neighbourhoodBay Village          4.094e+01  5.494e+00   7.451 9.45e-14 ***
## neighbourhoodBeacon Hill          1.918e+01  4.254e+00   4.509 6.54e-06 ***
## neighbourhoodBrighton            -1.246e+01  3.885e+00  -3.208  0.00134 **
## neighbourhoodCharlestoum          6.778e+01  4.870e+00  13.917  < 2e-16 ***
## neighbourhoodChinatoum            2.285e+01  5.657e+00   4.039 5.37e-05 ***
## neighbourhoodDorchester           5.266e+00  3.246e+00   1.622  0.10479
## neighbourhoodDowntoum             3.394e+01  3.490e+00   9.726  < 2e-16 ***
## neighbourhoodEast Boston          6.523e+00  3.842e+00   1.698  0.08957 .
## neighbourhoodFenway               6.467e+01  3.974e+00  16.276  < 2e-16 ***
## neighbourhoodHyde Park           -3.317e+01  6.224e+00  -5.330 9.87e-08 ***
## neighbourhoodJamaica Plain        2.374e+01  3.476e+00   6.828 8.71e-12 ***
## neighbourhoodLeather District     2.763e+01  2.350e+01   1.176  0.23976
## neighbourhoodLongwood Medical Area  2.336e+01  1.593e+01   1.466  0.14265
## neighbourhoodMattapan            -2.573e+01  5.925e+00  -4.343 1.41e-05 ***
## neighbourhoodMission Hill         7.340e+01  4.813e+00  15.249  < 2e-16 ***
## neighbourhoodNorth End            3.343e+01  4.933e+00   6.776 1.25e-11 ***
## neighbourhoodRoslindale          -2.310e+01  4.974e+00  -4.644 3.43e-06 ***
## neighbourhoodRoxbury              2.098e+01  3.516e+00   5.968 2.42e-09 ***
## neighbourhoodSouth Boston         5.530e+01  3.868e+00  14.299  < 2e-16 ***
## neighbourhoodSouth Boston Waterfront  1.016e+02  6.750e+00  15.056  < 2e-16 ***
## neighbourhoodSouth End            2.363e+01  3.626e+00   6.518 7.22e-11 ***
## neighbourhoodWest End             1.003e+02  5.704e+00  17.583  < 2e-16 ***
## neighbourhoodWest Roxbury         1.687e+01  6.138e+00   2.748  0.00600 **
## room_typeHotel room              -5.476e+01  7.912e+00  -6.922 4.53e-12 ***
## room_typePrivate room            -1.102e+02  1.521e+00 -72.456  < 2e-16 ***
## room_typeShared room             -1.583e+02  8.499e+00 -18.630  < 2e-16 ***
## reviews_per_month                -4.288e+00  3.554e-01 -12.067  < 2e-16 ***
## minimum_nights                   -2.923e-01  1.910e-02 -15.304  < 2e-16 ***
## availability_365                  7.075e-02  4.906e-03  14.421  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 125.7 on 38653 degrees of freedom
## Multiple R-squared:  0.2338, Adjusted R-squared:  0.2332
## F-statistic: 393.1 on 30 and 38653 DF,  p-value: < 2.2e-16
```

From the result of fit1, most of P-value is smaller than 0,05, which means most of coefficients are statistically significant. However, three of neighbourhoods'p-value are larger than 0.05. The model may not fit the data. Before I solve the problem of variable neighbourhood, I want to make sure that other variable have

no problem. Therefore, I try one more simple linear regression model without variable neighbourhood.

```
fit2 <- lm(price ~ room_type + reviews_per_month + minimum_nights + availability_365,data = airbnb)
summary(fit2)
##
## Call:
## lm(formula = price ~ room_type + reviews_per_month + minimum_nights +
##     availability_365, data = airbnb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -224.15  -60.19  -22.03   22.47 1860.22
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.203e+02  1.447e+00  152.25  < 2e-16 ***
## room_typeHotel room     -5.743e+01  8.032e+00   -7.15 8.86e-13 ***
## room_typePrivate room -1.282e+02  1.357e+00  -94.47  < 2e-16 ***
## room_typeShared room  -1.635e+02  8.623e+00  -18.96  < 2e-16 ***
## reviews_per_month      -6.177e+00  3.475e-01  -17.78  < 2e-16 ***
## minimum_nights         -3.252e-01  1.921e-02  -16.93  < 2e-16 ***
## availability_365        6.298e-02  4.882e-03   12.90  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128.3 on 38677 degrees of freedom
## Multiple R-squared:  0.2001, Adjusted R-squared:    0.2
## F-statistic:  1612 on 6 and 38677 DF,  p-value: < 2.2e-16
```

According to the result of fit2, every p-value is smaller than 0.05. Every coefficient is statistically significant. Interpretion of coefficients For the room type, the model uses Entired room/apt as baseline category against which all other groups are measured. Remaining other variable unchanged, the price of hotel room is 0.574 lower than the price of entired room/apt. The price of private room is 0.013 lower than the price of entired room/apt. The price of shared room is 0.016 lower than the price of entired room/apt. For the reviews per month, remaining other variables unchanged, when one unit of reviews per month increases, the price will decrease by 6.177. For the minimum_nights, remaining other variables unchanged,when one minimum night increases, the price will decrease by 0.325. For the availabilty 365, remaining other variables unchanged, when one days that a listing is available in a year (365 days) increases, the price will increase by 0.062.

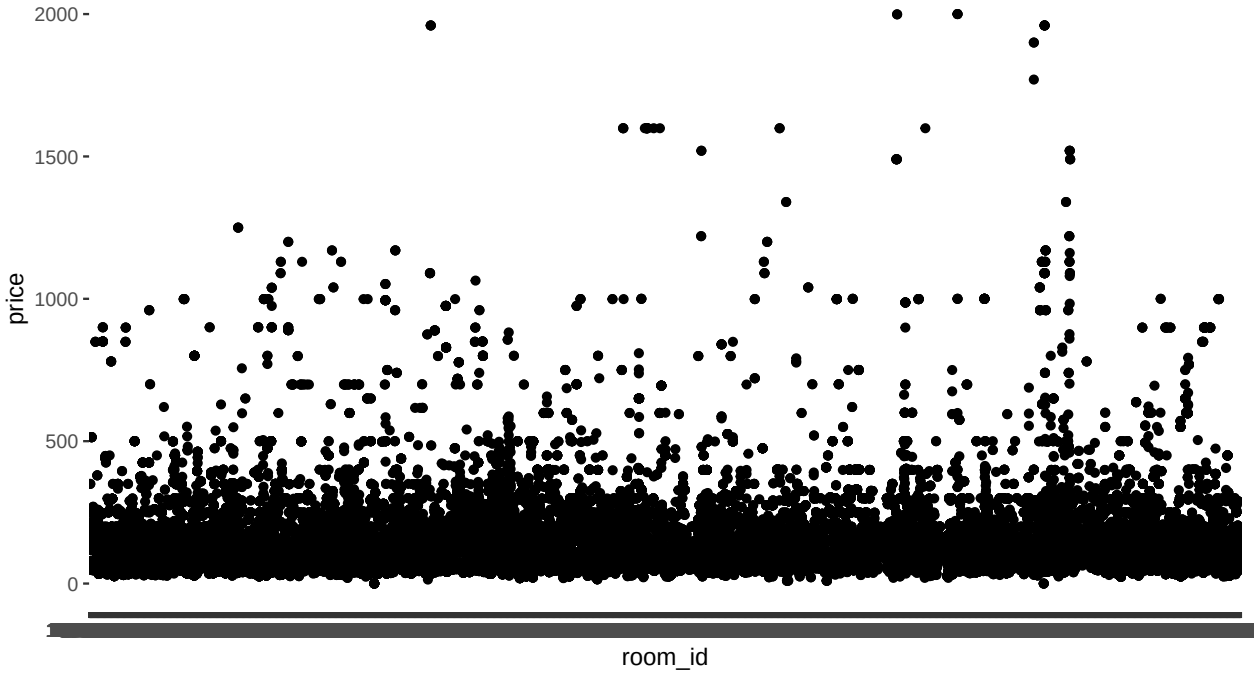
Model Selection

I use Akaike Information Criterion to select the best model among fit1, fit2 and fit3.To use AIC for model selection, we simply choose the model giving smallest AIC over the set of models considered.

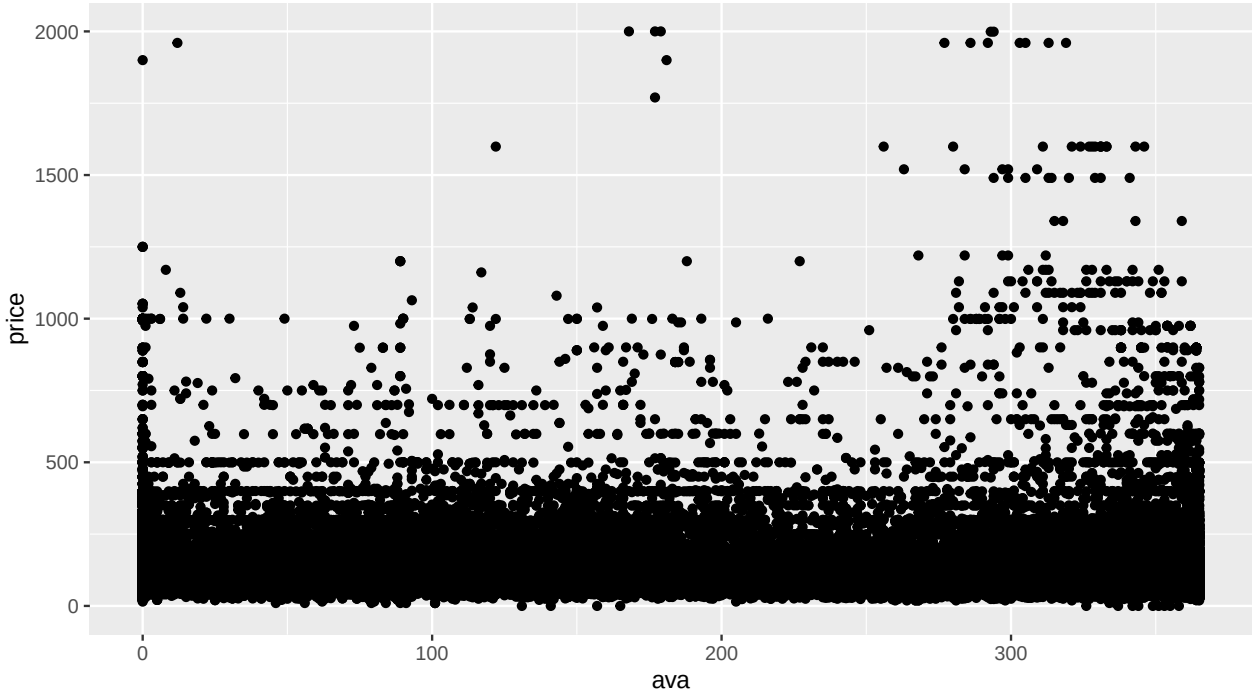
```
aic_value <- c(AIC(fit1), AIC(fit2), AIC(fit3))
knitr::kable(aic_value,col.names = c('AIC value'))
```

AIC value
483774.63
485389.96
53478.21

Some other plots

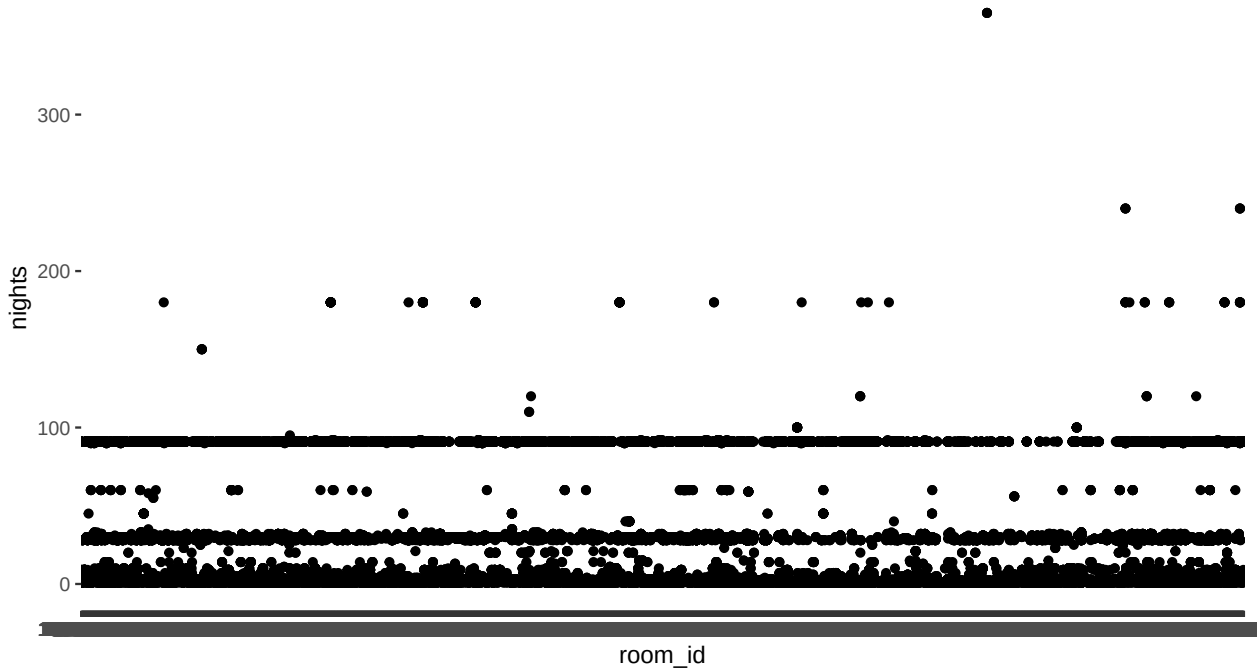


I would like to delete some rows whose price are higher than 2500. Since these prices are much higher than other price, the result of regression may be influenced.

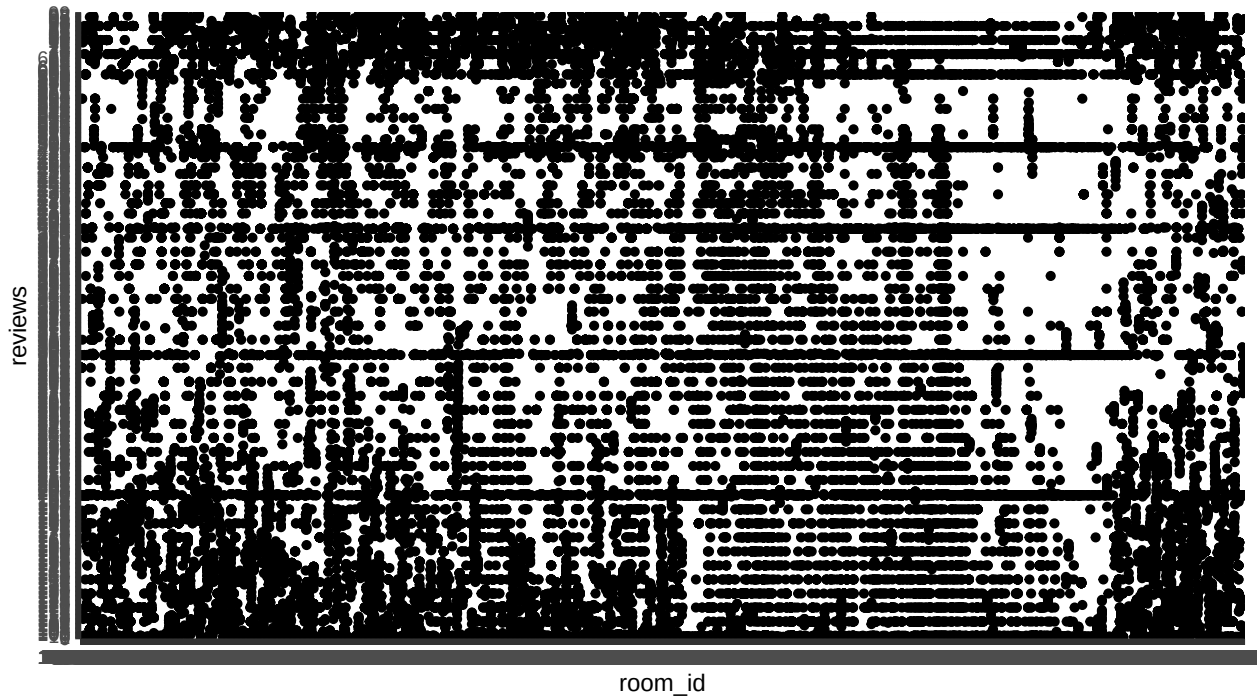


Availability_365 = 0 means that the listing is not available at all or in other words 0 days out of the year (365). Which can mean : 1) At the moment of collecting data for this dataset, those hosts had their listing availability set to 0 . available'(https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/discussion/111835).

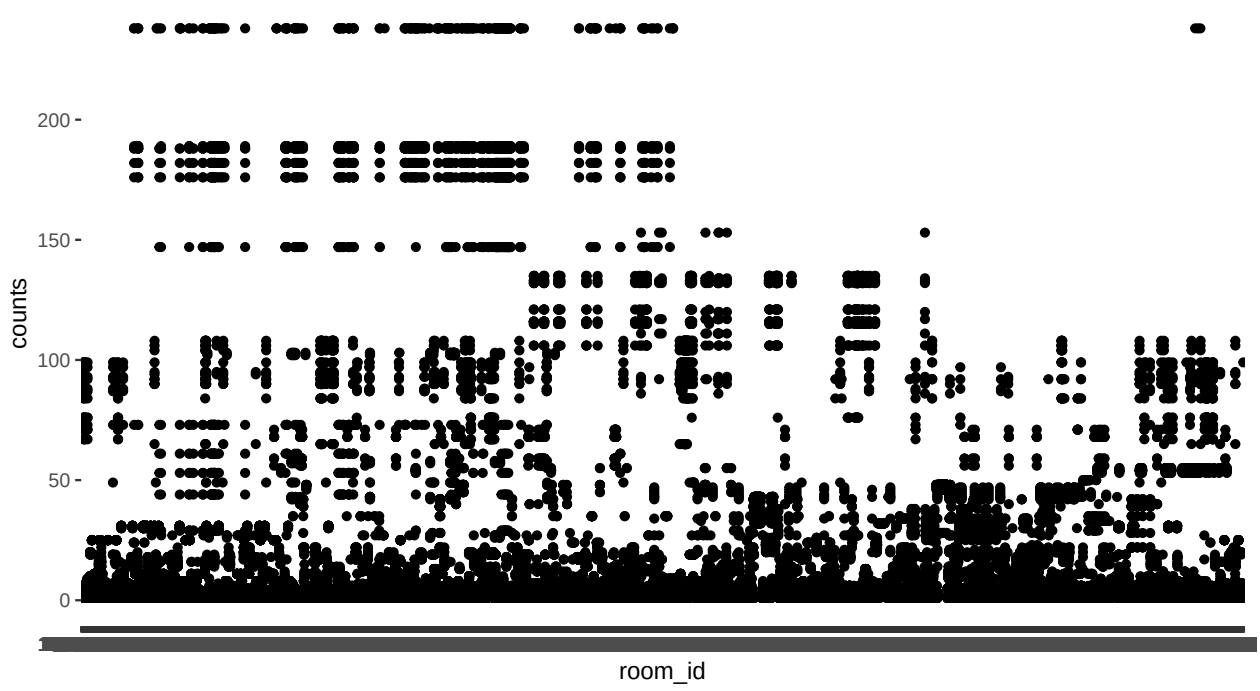
```
# minimum nights
ggplot(data=airbnb, aes(x=id, y=minimum_nights)) + geom_point() +
  labs(x = "room_id", y = "nights")
```




```
#number of reviews
ggplot(data=airbnb, aes(x=id, y=number_of_reviews)) + geom_point() +
  labs(x = "room_id", y = "reviews")
```



```
#Calculated host listings count
ggplot(data=airbnb, aes(x=id, y=calculated_host_listings_count)) + geom_point() +
  labs(x = "room_id", y = "counts")
```




```
ggplot(data=airbnb, aes(x=price, y=calculated_host_listings_count)) + geom_point() +
  labs(x = "price", y = "counts")
```