MA 678 Midterm Project - Boston Airbnb

Shiyu Zhang

November, 2018

A. Abstract

The main objective of this project is to study the important factors that may have a significant impact on the ratings and price of Airbnb listing properties. The potential implication of this study is to provide suggestion for existing and potential properties owners to have a better understanding of ratings, as well as for travelers to choose a property that best fit their need.

B. Project Background

I have been using Airbnb for over three years and it has become a popular way of travelling. I have witness Airbnb develop from an unknown website to the most popular travelling website during the past several years. Many people choose Airbnb instead of hotels not only for its lower price and convenient location, but also for its humanness – travelers are able to make connections with people from all around the world. What's more, travelers are provided with more unique options compare to hotels - houses, condos, apartments, castles, houseboats, tree houses, barns, mansions, even caves! Therefore, these unique properties of Airbnb inspired me to explore more about it. For example, what the factors may have an impact on the ratings, or, what is the relationship between the occupancy rate and the neighborhood of an Airbnb apartment, etc.

C. Dataset Information

In this project, I combined two datasets - the Airbnb dataset in Boston area (http://tomslee.net/ airbnb-data-collection-get-the-data) and the crime incident dataset (https://data.boston.gov/dataset/ crime-incident-reports-august-2015-to-date-source-new-system) for the analysis. The data include the following information: (1) room_id: A unique number identifying an Airbnb listing. (2) host_id: A unique number identifying an Airbnb host. (3) room_type: One of "Entire home/apt", "Private room", or "Shared room" (4) borough: A sub-region of the city or search area for which the survey is carried out. For some cities such as Boston, there is no borough information. (5) neighborhood: a sub-region of the city or search area for which the survey is carried out. A neighborhood is smaller than a borough. (6) reviews: The number of reviews that a listing has received. The number of reviews can be used to estimate the number of visits. However, such estimation may not be reliable for an individual listing (especially as reviews occasionally vanish from the site). (7) overall_satisfaction: The average rating that the owner of the property has received. (8) accommodates: The number of guests a listing can accommodate. (9) bedrooms: The number of bedrooms a listing offers. (10) price: The price (in $US) for a night stay. In early surveys, there may be some values that were recorded by month. (11) minstay: The minimum stay for a visit, as posted by the host. (12) latitude and longitude: The latitude and longitude of the listing as posted on Airbnb web. (13) last_modified: the date and time that the values were read from the Airbnb.

D. Data Cleaning and Methodology for Models

```
library(dplyr)
library(esquisse)
```

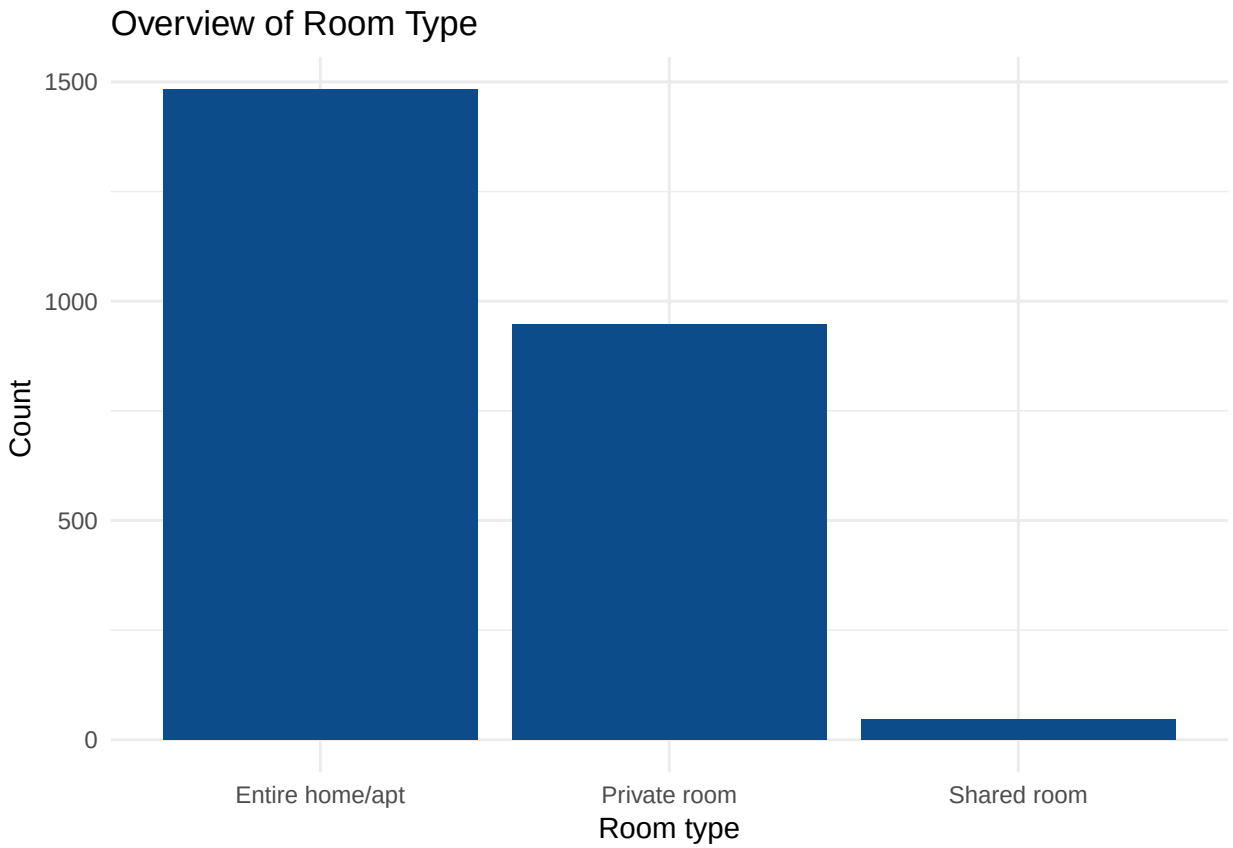
```r
library(ggplot2)
library(sqldf)
library(tidyr)
library(data.table)
library(arm)
library(knitr)
library(plyr)
#import data
Boston.airbnb<-read.csv("tomslee_airbnb_boston_0649_2016-11-21.csv")
# replace all N/A with 0
Boston.airbnb[is.na(Boston.airbnb)] <- 0
# Remove unrelevant columns
Boston.data<-Boston.airbnb[, c(-4,-9)]
#remove 0 review properties
Boston.data<-filter(Boston.data, reviews >0)
Boston.data<-filter(Boston.data, overall_satisfaction >0)
```

E. Dataset Structure & Overview

After the data cleaning process, the new dataset's structure is as follows:

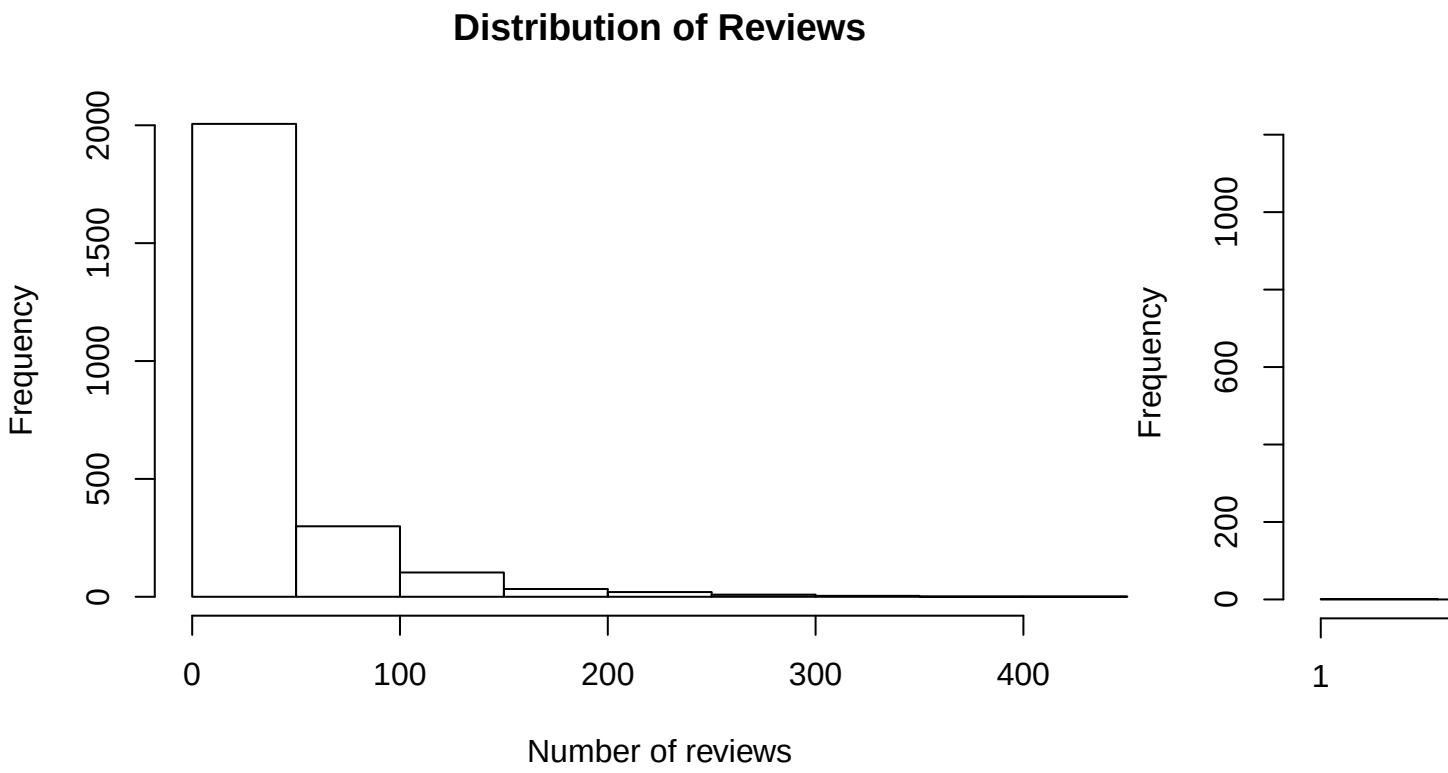
Room type overview

```
##
## Entire home/apt    Private room     Shared room
##            1483             947              46
```



From the output, we can see that in Boston area, entire home/apt is the most common type of properties for rent on the website, then is the private room. Shared room is the least common way on the website.

Check ratings and number of reviews



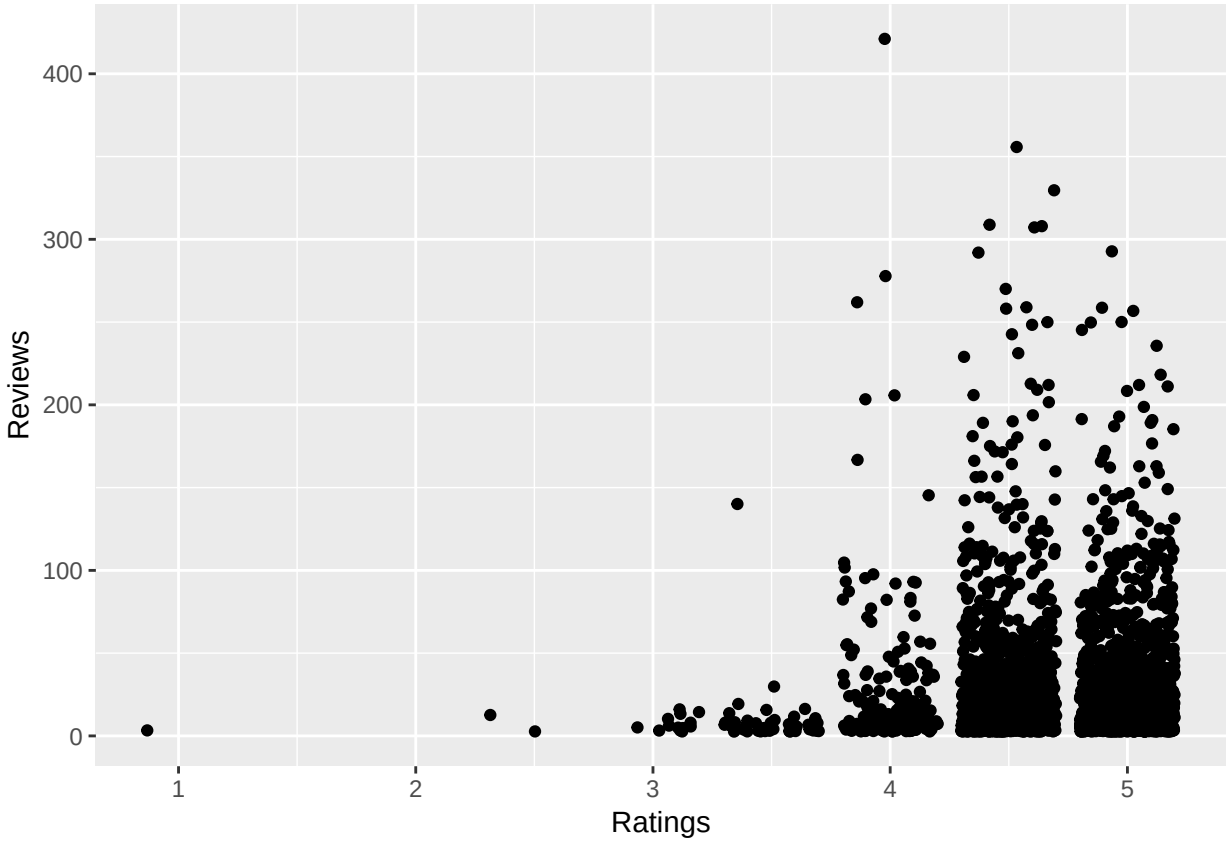

```
## [1] 0.8101777
```

```
## [1] 421
```

From the histogram for the distribution of ratings (overall_satisfaction), we can see that most of the ratings for Airbnb properties in Boston are above 4.0, the data shows right skewness. The distribution of reviews shows a left skewness. From the frequency table we can see that the majority number of reviews are less than 50 in Boston area. There are 3127 rooms in our data after cleaning, from the output, 84.9% of total Airbnb rooms have less than 50 reviews, while the maximum reviews for a room is 421.

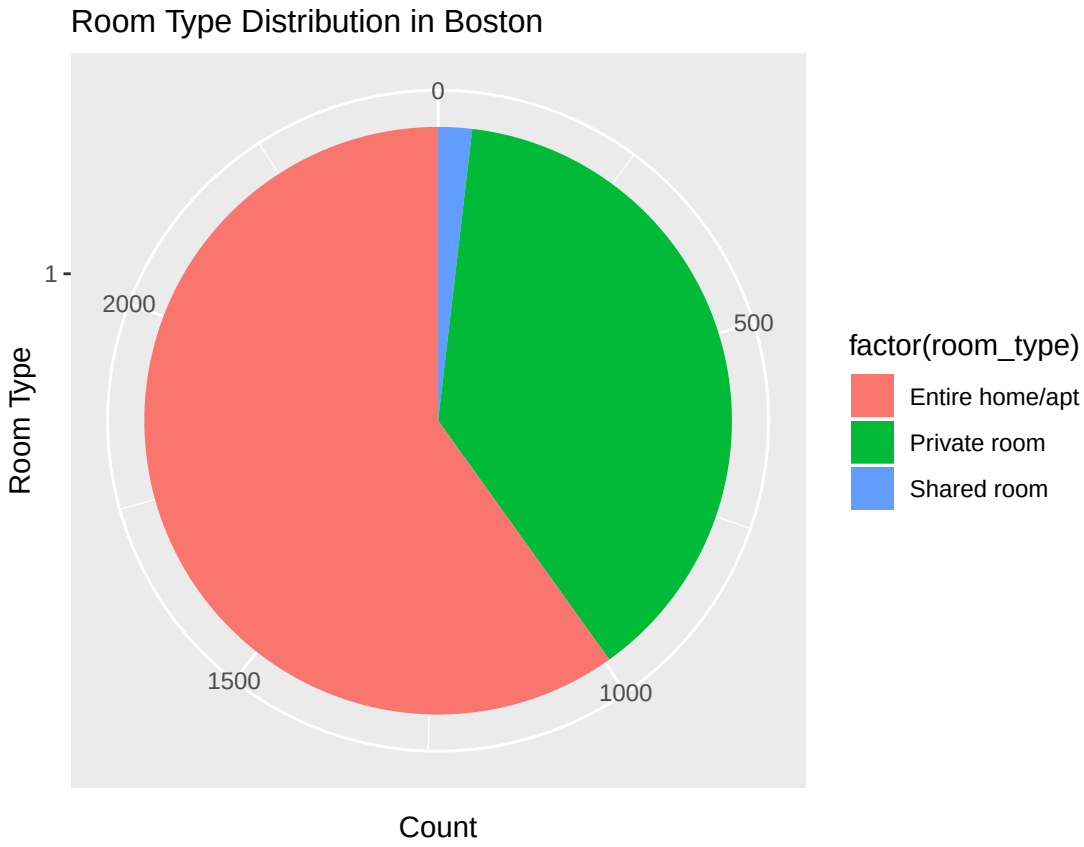
F. Data Visualization

Visualize relationship between overall satisfactions and number of reviews



Based on these output, we can tell that in general, higher ratings tend to have more reviews.

Distribution of room type

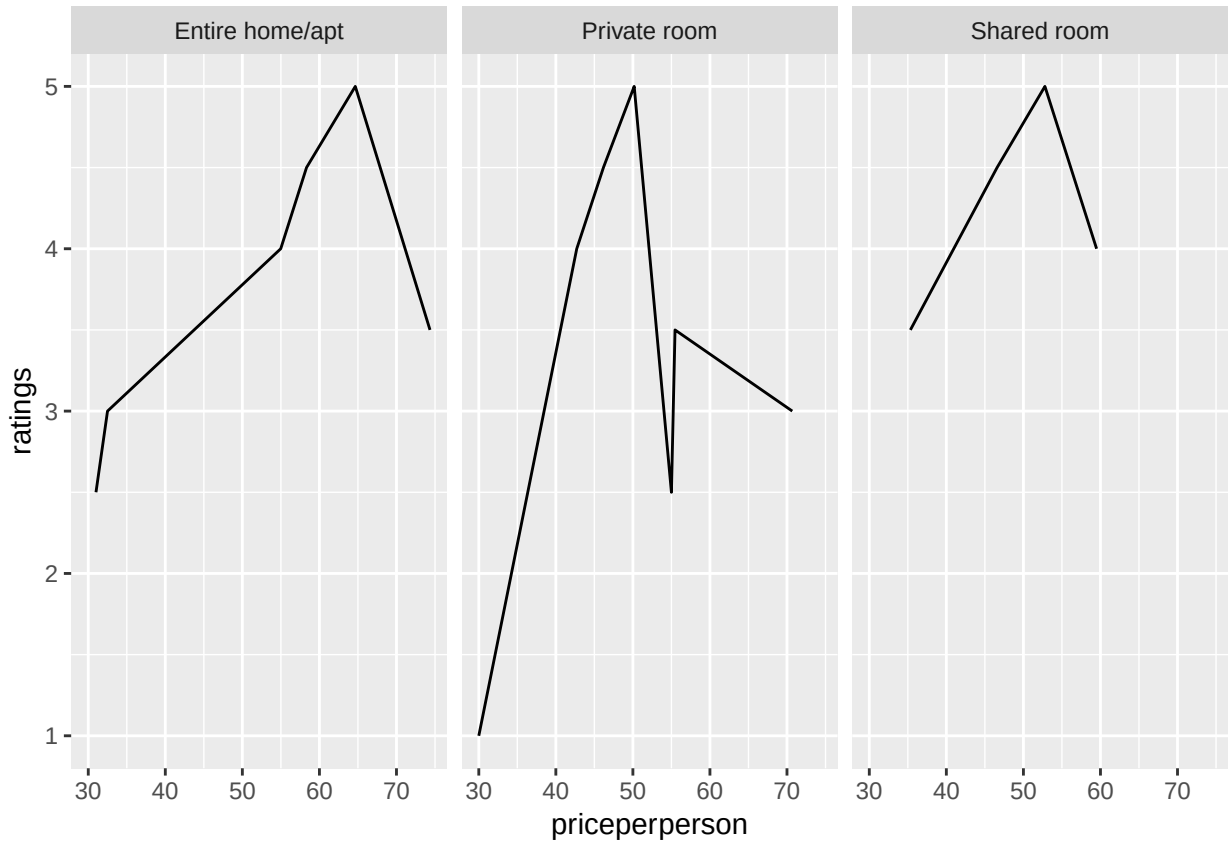


As we can tell from the pie chart, the entire home/apt has the majority proportion of the whole room types. Private room comes the next, and shared room has the least proportion among all the room types.

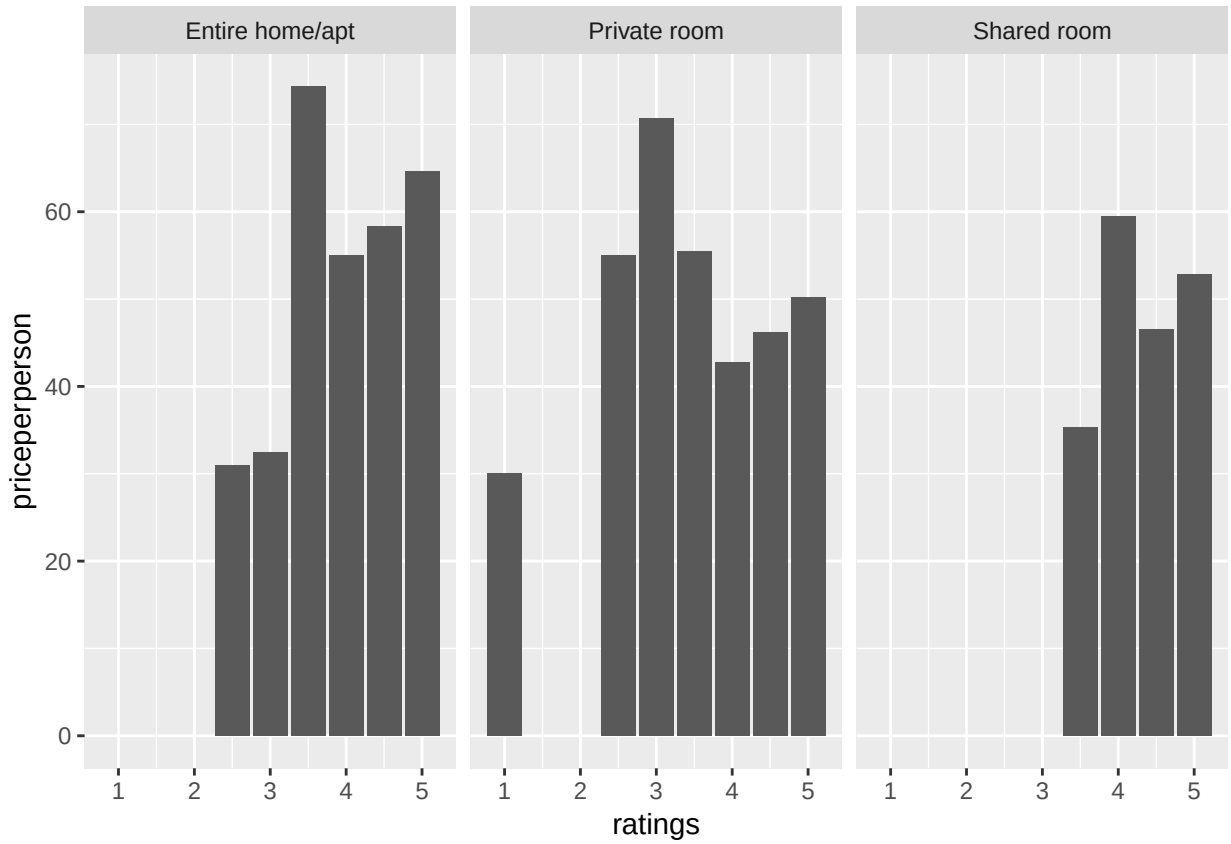
Explore Relationship Between Price per Person and Ratings & Room type

After I explored the distribution of the room type, I took a further step to explore the relationship between price per person and ratings & room type. I add one new column called " price per person" into my previous dataset, because the price is different when it comes to different room type, by introducing price per person, the variable became more comparable.

```
Boston.data$priceperperson <- (Boston.data$price)/(Boston.data$accommodates)
a1 <- aggregate( priceperperson ~ room_type+overall_satisfaction, Boston.data, mean )
a1 <- as.data.frame(a1)
names(a1) <- c("room_type","ratings","priceperperson")
ggplot(data=a1,aes(x=priceperperson,y=ratings))+geom_line()+facet_wrap(.~room_type)
```




```
ggplot(data=a1,aes(x=ratings,y=priceperperson))+geom_bar(stat="identity")+facet_wrap(.~room_type)
```




```
kable(a1, caption = "Average Price by Room Type")
```

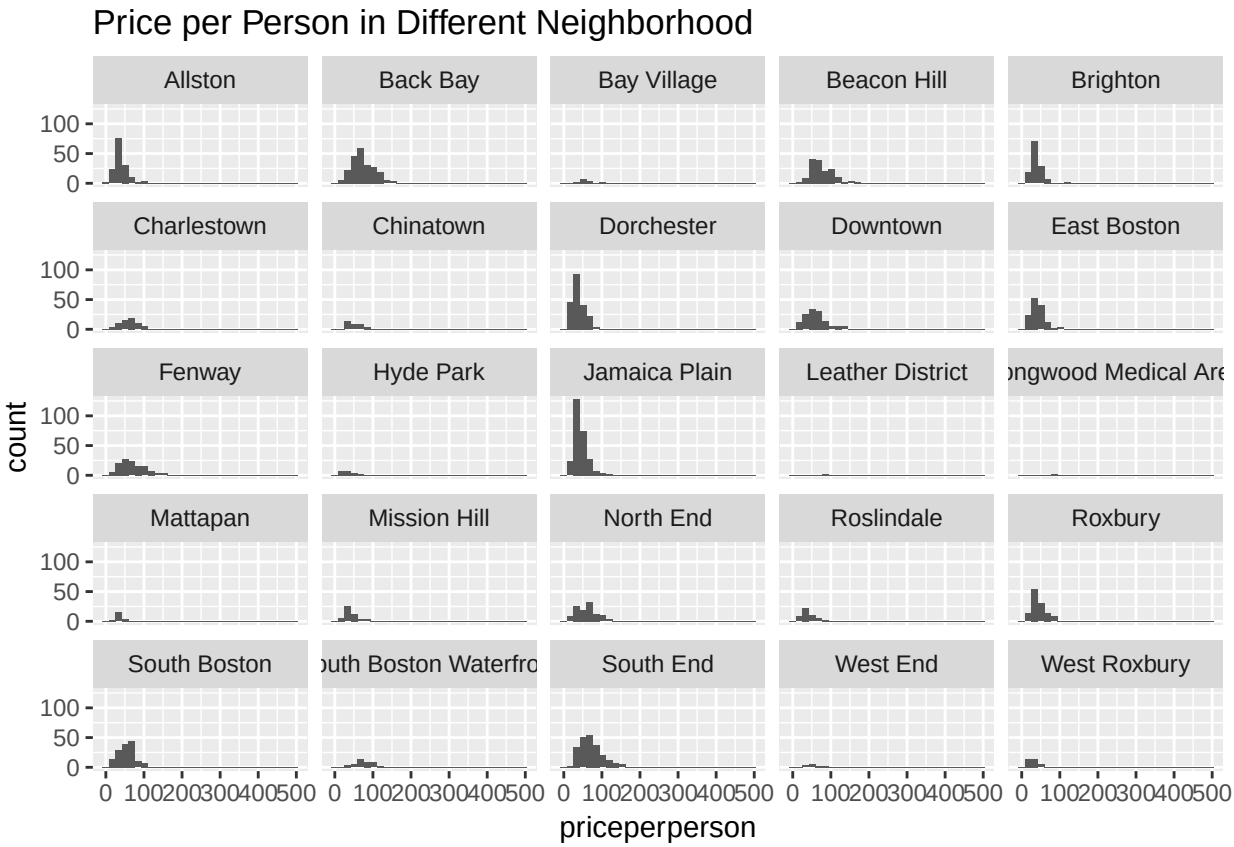
Table 1: Average Price by Room Type

room_type	ratings	priceperperson
Private room	1.0	30.00000
Entire home/apt	2.5	31.00000
Private room	2.5	55.00000
Entire home/apt	3.0	32.50000
Private room	3.0	70.68750
Entire home/apt	3.5	74.35167
Private room	3.5	55.46875
Shared room	3.5	35.33333
Entire home/apt	4.0	54.97981
Private room	4.0	42.71354
Shared room	4.0	59.50000
Entire home/apt	4.5	58.32589
Private room	4.5	46.17653
Shared room	4.5	46.58333
Entire home/apt	5.0	64.65567
Private room	5.0	50.17040
Shared room	5.0	52.78947

From the two output, we can tell that for different room type, for example, for entire home/apt, the most ratings are 3.5. For private room, ratings 3 is the most common one and ro the shared room, ratings 4 is more common. So we can tell from the graph that ratings are somehow related with the room type, i will

explore further later in the model part.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



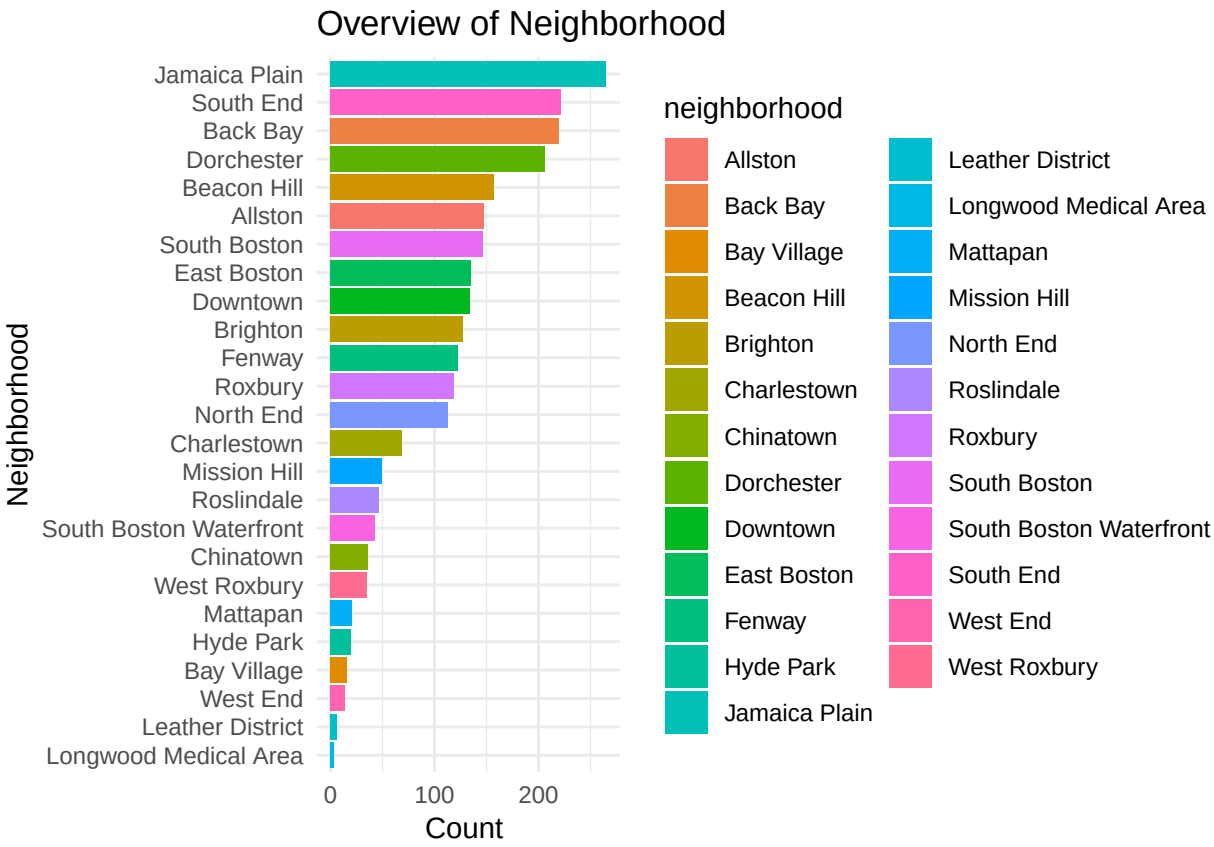
From the graph, we can see that the price in majority neighborhood ranges from 50 to 100 dollars per person, some preperties in Back Bay could be a little bit more expensive. Most of the properties in Jamaica Plain and Dorchester are 80 dollars per person.

Now we want to explore the relationship between the ratings and properties' location.

Neighborhood overview

```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : num 60
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
```

```
##  - attr(*, "validate")= logi TRUE
```



From this bar plot and r output, we can see that the top 5 neighborhood for rent on Airbnb in Boston area are Jamaica Plain, Back Bay, Allston, Dorchester and Beacon Hill. On contrast, Longwood, Leather District, Bay Village, West End and Roxbury are the least popular neighborhood for Airbnb in Boston area. We can see that the ratings are highly related to the neighborhood of the properties, in order to explore further about reasons behind it, i introduced another dataset - Boston crime incident data. dataset source : https://data.boston.gov/dataset/ crime-incident-reports-august-2015-to-date-source-new-system.

```
crime<-read.csv("crime_incident_reports.csv")

# replace district code with names
distrName = c(
A1 = 'Downtown',
A15= 'Charlestown',
A7= 'East Boston',
B2= 'Roxbury',
B3= 'Mattapan',
C6= 'South Boston',
C11= 'Dorchester',
D4= 'South End',
D14= 'Brighton',
E5= 'West Roxbury',
E13= 'Jamaica Plain',
E18= 'Hyde Park',
HTU= 'Human Traffic Unit'
)
crime$ReptDistrName = as.factor(distrName[as.character(crime$DISTRICT)])
```

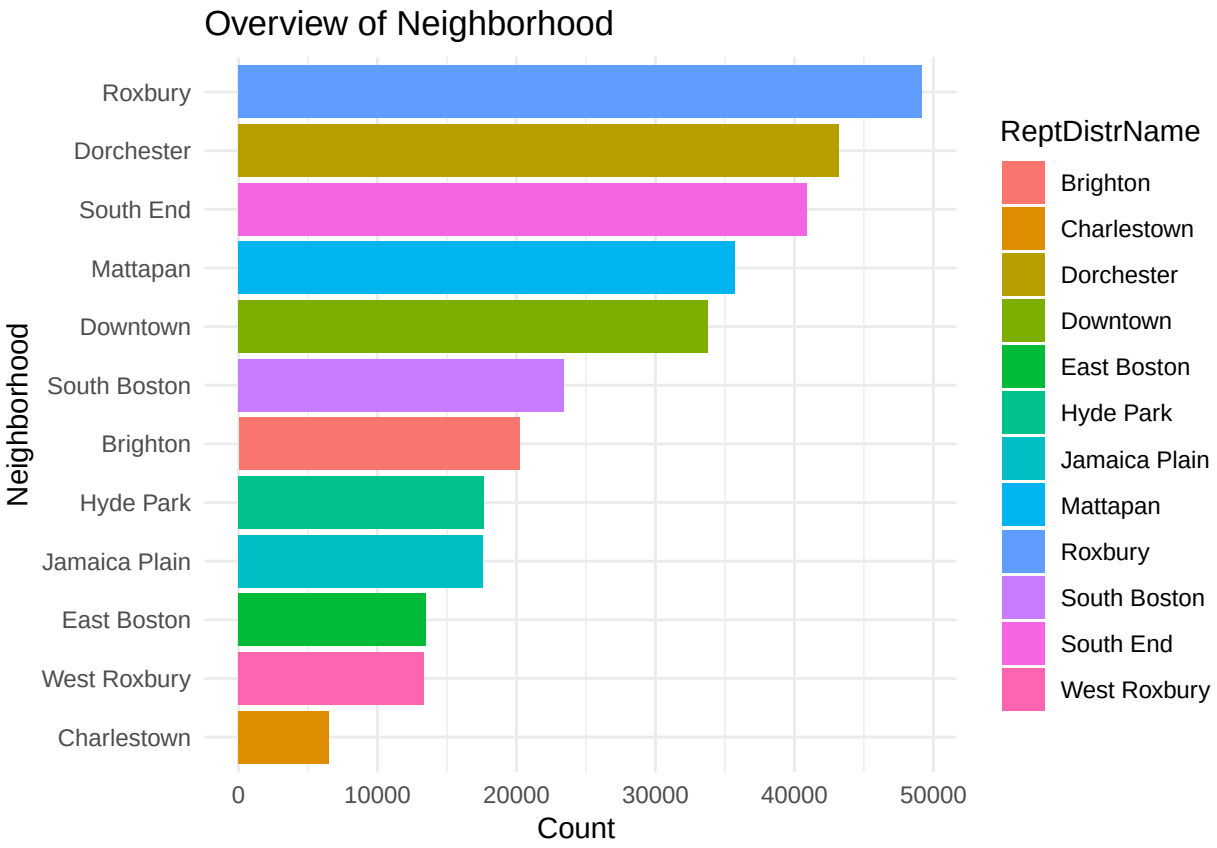
```r
crime$DISTRICT = NULL

data.crime=na.omit(crime)
dff2= count(data.crime, 'ReptDistrName')

pic2 <- ggplot(dff2, aes(x = reorder(ReptDistrName, freq), y = freq, fill = ReptDistrName)) +
         geom_bar(stat = "identity") +
  labs(title = 'Overview of Neighborhood',
    x = 'Neighborhood',
    y = 'Count ') +
  theme_minimal()+coord_flip()
 theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : num 60
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```r
print(pic2)
```



Compared to the two graphs above, we can see that ratings is highly related to the neighborhood of the property, more specific, in the neighborhood where the crime rates are lower, the ratings tends to be higher."

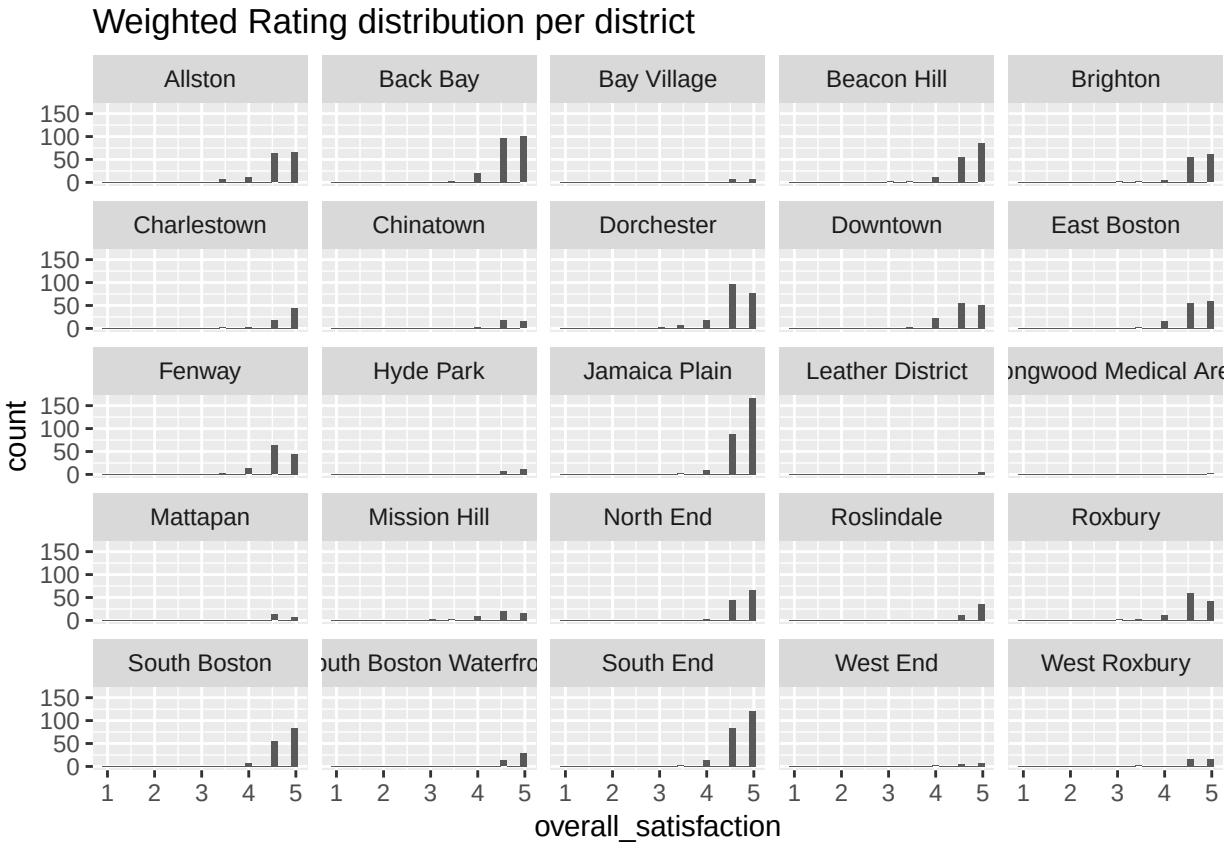
To sum up, we can say that ratings is highly related to number of reviews, price , accomodates, minimum stays and the location of the property. It has a moderate relationship with the room type."

G. EDA

Prior to the application for multilevel model, I think doing some initial EDA is helpful for a better understanding of relationship between independent variables and dependents variables.

1.Distribution of ratings in different districts

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As we can see from the output, it is more clear that Bay Village,Leather District, Longwood Medical area, west end and west roxbury have very few reviews or ratings compared to the crime plot i showed in the previous part. we can say that the area where has more crime tend to have few ratings/ reviews. On the other hand, Allston, Back Bay, Jamacia Plain and South End have most ratings/ reviews.

2. Check multicollinearity between predictors

Correlation between reviews, overall_satisfaction, accommodates, price and minstay

```
correlation<-cor(Boston.data[,c(5,6,7,8,9)])
symnum(correlation)
```

```
##                      r o a p m
## reviews              1
## overall_satisfaction   1
## accommodates             1
## price                    . 1
## minstay                      1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

From the output, we can see that the it is fine to do the model since there is no mark that shows the variables are problematic.

3.Check Model Fit and Build Model

```
# Scale price from dataset
Boston.data$price<-scale(Boston.data$price, center = TRUE, scale = TRUE)
```

(1) No random effect

Model 1 : ratings(overall_satisfaction as outcome variable)

```
no_1<-lm(overall_satisfaction ~ room_type+reviews+accommodates+minstay+price, data=Boston.data)
summary(no_1)
```

```
##
## Call:
## lm(formula = overall_satisfaction ~ room_type + reviews + accommodates +
##     minstay + price, data = Boston.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6400 -0.1847 -0.1003  0.3230  0.4550
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.7343071  0.0254746 185.844  < 2e-16 ***
## room_typePrivate room  -0.0126789  0.0201273  -0.630  0.52880
## room_typeShared room   -0.0458824  0.0593067  -0.774  0.43921
## reviews                 0.0002682  0.0001815   1.478  0.13954
## accommodates           -0.0138836  0.0052280  -2.656  0.00797 **
## minstay                -0.0070464  0.0044391  -1.587  0.11256
## price                   0.0498133  0.0104502   4.767 1.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3834 on 2469 degrees of freedom
## Multiple R-squared:  0.01336,    Adjusted R-squared:  0.01097
## F-statistic: 5.574 on 6 and 2469 DF,  p-value: 9.406e-06
```

Based on the output, we can see that only the accommodates and price are statistically significant from zero,however, i think it is still meaningful to keep the rest of the variables because it is still meaningful in reality.especially minstay factor, it should have great influence on the ratings.

with each unit increase in reviews, the ratings (overall_satisfaction) increase by 0.0002. One unit increase in accommodates, the rating decrease by 0.013. One unit increase in mininum stays, the rating decrease by 0.007, which is meaningful in reality becasue the minimum stays sometimes keep out the customers who don't meet the minimum stay requirements. For the price factor, with each unit increase in price, the rating would decrease by 0.049.

Model 2 : price as outcome variable

```
no_2<-glm(price ~ room_type+reviews+accommodates+minstay+overall_satisfaction, data=Boston.data)
summary(no_2)
```

```
##
## Call:
```

```
## glm(formula = price ~ room_type + reviews + accommodates + minstay +
##     overall_satisfaction, data = Boston.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1443  -0.3678  -0.0974   0.2535   5.4123
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.2203973  0.1874640  -6.510 9.07e-11 ***
## room_typePrivate room  -0.7085246  0.0358562 -19.760  < 2e-16 ***
## room_typeShared room   -0.8087593  0.1125343  -7.187 8.75e-13 ***
## reviews                -0.0017124  0.0003463  -4.945 8.14e-07 ***
## accommodates            0.2169535  0.0090368  24.008  < 2e-16 ***
## minstay                 0.0078057  0.0085126   0.917    0.359
## overall_satisfaction    0.1830604  0.0384038   4.767 1.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.5401428)
##
##     Null deviance: 2475.0  on 2475  degrees of freedom
## Residual deviance: 1333.6  on 2469  degrees of freedom
## AIC: 5510.6
##
## Number of Fisher Scoring iterations: 2
```

Based on the output, we can see that all the variables except minstay are statistically significant from zero, which means that this model it much better than the previous one.

Compared to entire apt, when the room type changed to private or shared room, the price will decrease by 0.7 and 0.8 respectively. It makes sense becasue the entire room should cost more than private room or shared room. With one unit increase in reviews, the price decrease by 0.0017, this number is too small so i think we can ingore the effect of reviews in this model. One unit increase in accommodates, the price increase by 0.21. One unit increase in mininum stays, the price increase by 0.007, which is meaningful in reality becasue the minimum stays sometimes keep out the customers who don't meet the minimum stay requirements. For the rating factor, with each unit increase in rating, the price would decrease by 0.183.

(2) Consider random effects - random intercept

Ratings(overall_satisfaction as outcome variable)

```
intercept_1<-lmer(overall_satisfaction ~ room_type+reviews+accommodates+minstay+price+(1|neighborhood),
summary(intercept_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## overall_satisfaction ~ room_type + reviews + accommodates + minstay +
##     price + (1 | neighborhood)
##    Data: Boston.data
##
## REML criterion at convergence: 2268.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -9.6771 -0.5158 -0.0338  0.7993  1.3895
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  neighborhood (Intercept) 0.006646 0.08153
##  Residual                 0.141296 0.37589
## Number of obs: 2476, groups:  neighborhood, 25
##
## Fixed effects:
##                          Estimate Std. Error t value
## (Intercept)             4.7438935  0.0323233 146.764
## room_typePrivate room -0.0152260  0.0211743  -0.719
## room_typeShared room  -0.0201076  0.0583896  -0.344
## reviews                0.0002200  0.0001803   1.221
## accommodates          -0.0156306  0.0054725  -2.856
## minstay               -0.0057478  0.0044029  -1.305
## price                  0.0499026  0.0110318   4.524
##
## Correlation of Fixed Effects:
##             (Intr) rm_tPr rm_tSr reviws accmmd minsty
## rm_typPrvtr -0.491
## rm_typShrdr -0.193  0.235
## reviews     -0.190 -0.022  0.047
## accommodats -0.668  0.341  0.124 -0.026
## minstay     -0.417  0.163  0.096  0.140  0.144
## price        0.199  0.214  0.117  0.081 -0.492 -0.031
```

The above model (intercept_1) was created by using the fixed effect to ratings(overall_satisfaction), controlling for by-neighborhood variability. The random effects measures of how much variability in the dependent measure that is due to the random effects "neighborhood". "Residual" which stands for the variability that's not due to the random effects. From the fixed effects output, when other variables remain constant, each unit increase of accommodates, the rating (overall_satisfaction) decreases by 0.02 on average; with every unit increase of reviews, rating remains constant ; One unit increase in mininum stays, the rating decrease by 0.01, which is meaningful in reality becasue the minimum stays sometimes keep out the customers who don't meet the minimum stay requirements. For the price factor, with each unit increase in price, the rating would increase by 0.05.

Price as outcome variable

```
intercept_2 <-lmer(price ~ room_type+reviews+accommodates+minstay+overall_satisfaction+(1|neighborhood)
summary(intercept_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## price ~ room_type + reviews + accommodates + minstay + overall_satisfaction +
##     (1 | neighborhood)
##    Data: Boston.data
##
## REML criterion at convergence: 5155.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3006 -0.5198 -0.0663  0.3297  7.6751
##
## Random effects:
```

```
##  Groups       Name        Variance Std.Dev.
##  neighborhood (Intercept) 0.1770   0.4207
##  Residual                 0.4473   0.6688
## Number of obs: 2476, groups:  neighborhood, 25
##
## Fixed effects:
##                         Estimate Std. Error t value
## (Intercept)           -1.3495736  0.1955665  -6.901
## room_typePrivate room -0.3364429  0.0376200  -8.943
## room_typeShared room  -0.5657137  0.1033584  -5.473
## reviews               -0.0013110  0.0003211  -4.083
## accommodates           0.2496074  0.0085221  29.290
## minstay                0.0141000  0.0078719   1.791
## overall_satisfaction   0.1570736  0.0358080   4.387
##
## Correlation of Fixed Effects:
##             (Intr) rm_tPr rm_tSr reviws accmmd minsty
## rm_typPrvtr -0.190
## rm_typShrdr -0.077  0.220
## reviews     -0.048 -0.040  0.037
## accommodats -0.207  0.527  0.212  0.015
## minstay     -0.141  0.173  0.101  0.141  0.150
## ovrll_stsfc -0.864  0.030  0.014 -0.017  0.014  0.022
```

Compared to entire apt, when the room type changed to private or shared room, the price will decrease by 0.34 and 0.57 respectively. It makes sense becasue the entire room should cost more than private room or shared room. One unit increase in accommodates, the price increase by 0.25. One unit increase in mininum stays, the price increase by 0.01, For the rating factor, with each unit increase in rating, the price would increase by 0.16.

(3) Consider random effects - random slope

Ratings(overall_satisfaction as outcome variable)

```
slope_1<-lmer(overall_satisfaction~room_type+reviews+accommodates+minstay+price+(0+price|neighborhood),
summary(slope_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## overall_satisfaction ~ room_type + reviews + accommodates + minstay +
##     price + (0 + price | neighborhood)
##    Data: Boston.data
##
## REML criterion at convergence: 2321.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -9.4634 -0.4984 -0.1768  0.8236  1.3483
##
## Random effects:
##  Groups       Name  Variance Std.Dev.
##  neighborhood price 0.002574 0.05074
##  Residual           0.145199 0.38105
## Number of obs: 2476, groups:  neighborhood, 25
```

```
##
## Fixed effects:
##                          Estimate Std. Error t value
## (Intercept)             4.7360912  0.0259702 182.366
## room_typePrivate room -0.0041872  0.0205617  -0.204
## room_typeShared room  -0.0298282  0.0591536  -0.504
## reviews                0.0002951  0.0001818   1.623
## accommodates          -0.0147698  0.0053549  -2.758
## minstay               -0.0062458  0.0044411  -1.406
## price                  0.0596117  0.0160069   3.724
##
## Correlation of Fixed Effects:
##             (Intr) rm_tPr rm_tSr reviws accmmd minsty
## rm_typPrvtr -0.480
## rm_typShrdr -0.223  0.236
## reviews     -0.241 -0.014  0.049
## accommodats -0.784  0.222  0.108 -0.035
## minstay     -0.509  0.156  0.094  0.145  0.137
## price        0.134  0.291  0.105  0.066 -0.309 -0.024
```

Based on the output, when other variables remain constant, each unit increase of accommodates, the rating (overall_satisfaction) decreases by 0.01 on average; one unit increase of price deviating from mean price over standard deviation weighted rating decreases by 0.06 on average. Shared room has 0.03 lower in rating than that of entire room on average. Each unit increase in minstay, the ratings will decrease by 0.01.

Price as outcome variable

```
slope_2<-lmer(price~room_type+reviews+accommodates+minstay+overall_satisfaction+(0+overall_satisfaction
summary(slope_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## price ~ room_type + reviews + accommodates + minstay + overall_satisfaction +
##     (0 + overall_satisfaction | neighborhood)
##    Data: Boston.data
##
## REML criterion at convergence: 5157.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3489 -0.5221 -0.0684  0.3278  7.6451
##
## Random effects:
##  Groups       Name                 Variance Std.Dev.
##  neighborhood overall_satisfaction 0.007729 0.08791
##  Residual                          0.447954 0.66929
## Number of obs: 2476, groups:  neighborhood, 25
##
## Fixed effects:
##                          Estimate Std. Error t value
## (Intercept)            -1.4243357  0.1742955  -8.172
## room_typePrivate room -0.3369492  0.0376065  -8.960
## room_typeShared room  -0.5692031  0.1034365  -5.503
## reviews               -0.0012855  0.0003213  -4.001
## accommodates           0.2491535  0.0085277  29.217
```

```
## minstay                0.0144741  0.0078793   1.837
## overall_satisfaction   0.1723398  0.0398823   4.321
##
## Correlation of Fixed Effects:
##             (Intr) rm_tPr rm_tSr reviws accmmd minsty
## rm_typPrvtr -0.216
## rm_typShrdr -0.090  0.220
## reviews     -0.055 -0.036  0.039
## accommodats -0.230  0.527  0.212  0.016
## minstay     -0.158  0.175  0.102  0.141  0.151
## ovrll_stsfc -0.859  0.029  0.016 -0.015  0.011  0.019
```

Compared to entire apt, when the room type changed to private or shared room, the price will decrease by 0.34 and 0.57 respectively. It makes sense becasue the entire room should cost more than private room or shared room. One unit increase in accommodates, the price increase by 0.25. One unit increase in mininum stays, the price increase by 0.01, For the rating factor, with each unit increase in rating, the price would increase by 0.17.

(4) Combination - random slope and intercept

I added the random slope and intercept together to see if these model would be a better fit.

```
con_1<-lmer(overall_satisfaction~factor(room_type)+reviews+accommodates+minstay+price+(1+price|neighborh



# con_1_1<-lm(overall_satisfaction~factor(room_type)+reviews+accommodates+minstay+price, data=Boston.da
summary(con_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## overall_satisfaction ~ factor(room_type) + reviews + accommodates +
##     minstay + price + (1 + price | neighborhood)
##    Data: Boston.data
##
## REML criterion at convergence: 2265.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -9.6625 -0.5284 -0.0233  0.7878  1.4355
##
## Random effects:
##  Groups       Name        Variance  Std.Dev. Corr
##  neighborhood (Intercept) 0.0064612 0.08038
##               price       0.0002576 0.01605  -1.00
##  Residual                 0.1411402 0.37569
## Number of obs: 2476, groups:  neighborhood, 25
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                    4.7401901  0.0318667 148.750
## factor(room_type)Private room -0.0142881  0.0213879  -0.668
## factor(room_type)Shared room  -0.0133151  0.0584195  -0.228
## reviews                        0.0002158  0.0001805   1.195
## accommodates                  -0.0146354  0.0053735  -2.724
```

```
## minstay                      -0.0055108  0.0043975  -1.253
## price                         0.0485965  0.0112490   4.320
##
## Correlation of Fixed Effects:
##             (Intr) f(_)Pr f(_)Sr reviws accmmd minsty
## fctr(rm_)Pr -0.487
## fctr(rm_)Sr -0.195  0.235
## reviews     -0.194 -0.020  0.047
## accommodats -0.662  0.334  0.124 -0.027
## minstay     -0.417  0.161  0.097  0.141  0.143
## price       -0.015  0.233  0.118  0.077 -0.442 -0.026
```

```
#car::marginalModelPlot(con_1)
```

(1) In the first model (con_1) , (1+price|neighborhood) means that the model is expected to differ baseline-levels of price (the intercept, represented by 1) as well as differ neighborhood.

For the random effect output, it represents the estimated variability in the intercept. For the fixed effect output, the coefficient of reviews and price are positive, which indicates a positive change in unit review or unit price would lead to a positve increase in the ratings.The other variables' coefficients, room tpe, accomodates and minstay shows a negative change in one unit would lead to a decrease in ratings.

```
con_2<-lmer(price~factor(room_type)+reviews+accommodates+minstay+overall_satisfaction+(1+overall_satisfa
summary(con_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: price ~ factor(room_type) + reviews + accommodates + minstay +
##     overall_satisfaction + (1 + overall_satisfaction | neighborhood)
##    Data: Boston.data
##
## REML criterion at convergence: 5155
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3088 -0.5183 -0.0680  0.3297  7.6709
##
## Random effects:
##  Groups       Name                Variance  Std.Dev. Corr
##  neighborhood (Intercept)         0.1269583 0.35631
##               overall_satisfaction 0.0001831 0.01353  1.00
##  Residual                         0.4473259 0.66882
## Number of obs: 2476, groups:  neighborhood, 25
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                   -1.3611501  0.1899017  -7.168
## factor(room_type)Private room -0.3360808  0.0376193  -8.934
## factor(room_type)Shared room  -0.5659545  0.1033598  -5.476
## reviews                       -0.0013068  0.0003211  -4.070
## accommodates                   0.2495761  0.0085223  29.285
## minstay                        0.0141629  0.0078723   1.799
## overall_satisfaction           0.1593722  0.0358720   4.443
##
## Correlation of Fixed Effects:
##             (Intr) f(_)Pr f(_)Sr reviws accmmd minsty
## fctr(rm_)Pr -0.196
```

```
## fctr(rm_)Sr -0.080  0.220
## reviews     -0.050 -0.040  0.038
## accommodats -0.213  0.527  0.212  0.015
## minstay     -0.145  0.173  0.102  0.141  0.150
## ovrll_stsfc -0.857  0.031  0.015 -0.017  0.014  0.022
```

(2) In the second model. For the random effect output, it represents the estimated variability in the intercept. For the fixed effect output, the coefficient of accomodates, minstay and ratings (overall_satisfaction) are positive, which indicates a positive change in unit would lead to a positve increase in the price.The other variables' coefficients, room tpe and reviews shows a negative change in one unit would lead to a decrease in ratings.

To sum up, the random slope and random intercept model is better compared with only random slope or only random intercept.

(5) Consider Interaction to the Model

```
inter1<-lmer(overall_satisfaction~room_type+reviews+accommodates*minstay+price+(1+accommodates|neighborl
display(inter1)
```

```
## lmer(formula = overall_satisfaction ~ room_type + reviews + accommodates *
##     minstay + price + (1 + accommodates | neighborhood), data = Boston.data)
##                       coef.est coef.se
## (Intercept)            4.73     0.03
## room_typePrivate room -0.02     0.02
## room_typeShared room  -0.02     0.06
## reviews                0.00     0.00
## accommodates          -0.01     0.01
## minstay                0.01     0.01
## price                  0.05     0.01
## accommodates:minstay   0.00     0.00
##
## Error terms:
##  Groups       Name         Std.Dev. Corr
##  neighborhood (Intercept)  0.09
##               accommodates 0.00     -1.00
##  Residual                  0.38
## ---
## number of obs: 2476, groups: neighborhood, 25
## AIC = 2299.6, DIC = 2142
## deviance = 2208.8
```

```
inter2<-lmer(price~room_type+reviews+accommodates*minstay+overall_satisfaction+(1+accommodates|neighborl
display(inter2)
```

```
## lmer(formula = price ~ room_type + reviews + accommodates * minstay +
##     overall_satisfaction + (1 + accommodates | neighborhood),
##     data = Boston.data)
##                       coef.est coef.se
## (Intercept)           -1.20     0.18
## room_typePrivate room -0.35     0.04
## room_typeShared room  -0.59     0.10
## reviews                0.00     0.00
## accommodates           0.19     0.02
## minstay               -0.08     0.01
```

```
## overall_satisfaction   0.15     0.03
## accommodates:minstay   0.04     0.00
##
## Error terms:
##  Groups       Name         Std.Dev. Corr
##  neighborhood (Intercept)  0.23
##               accommodates 0.09     0.33
##  Residual                  0.64
## ---
## number of obs: 2476, groups: neighborhood, 25
## AIC = 5029.6, DIC = 4898.2
## deviance = 4951.9
```

H.1 ANOVA Analysis

```
anova(intercept_1,slope_1,con_1,no_1)
```

```
## Data: Boston.data
## Models:
## no_1: overall_satisfaction ~ room_type + reviews + accommodates + minstay +
## no_1:     price
## intercept_1: overall_satisfaction ~ room_type + reviews + accommodates + minstay +
## intercept_1:     price + (1 | neighborhood)
## slope_1: overall_satisfaction ~ room_type + reviews + accommodates + minstay +
## slope_1:     price + (0 + price | neighborhood)
## con_1: overall_satisfaction ~ factor(room_type) + reviews + accommodates +
## con_1:     minstay + price + (1 + price | neighborhood)
##             Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## no_1         8 2288.0 2334.5 -1136.0   2272.0
## intercept_1  9 2229.6 2282.0 -1105.8   2211.6 60.315      1  8.083e-15 ***
## slope_1      9 2281.7 2334.1 -1131.9   2263.7  0.000      0          1
## con_1       11 2230.8 2294.8 -1104.4   2208.8 54.898      2  1.200e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
"based on the output, con_1 is the best among the three models,becasue it has the least AIC and BIC."
```

```
## [1] "based on the output, con_1 is the best among the three models,becasue it has the least AIC and
```

```
anova(intercept_2,slope_2,con_2,no_2)
```

```
## Data: Boston.data
## Models:
## no_2: price ~ room_type + reviews + accommodates + minstay + overall_satisfaction
## intercept_2: price ~ room_type + reviews + accommodates + minstay + overall_satisfaction +
## intercept_2:     (1 | neighborhood)
## slope_2: price ~ room_type + reviews + accommodates + minstay + overall_satisfaction +
## slope_2:     (0 + overall_satisfaction | neighborhood)
## con_2: price ~ factor(room_type) + reviews + accommodates + minstay +
## con_2:     overall_satisfaction + (1 + overall_satisfaction | neighborhood)
##             Df    AIC    BIC  logLik deviance    Chisq Chi Df Pr(>Chisq)
## no_2         8 5510.6 5557.1 -2747.3   5494.6
## intercept_2  9 5127.5 5179.8 -2554.8   5109.5 385.0360      1     <2e-16
## slope_2      9 5130.2 5182.5 -2556.1   5112.2   0.0000      0     1.0000
```

```
## con_2       11 5131.4 5195.4 -2554.7   5109.4   2.7547      2     0.2523
##
## no_2
## intercept_2 ***
## slope_2
## con_2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
"based on the output, con_2 is the best among the three models,becasue it has the least AIC and BIC."
```

```
## [1] "based on the output, con_2 is the best among the three models,becasue it has the least AIC and
```

```
anova(inter1,inter2)
```

```
## Data: Boston.data
## Models:
## inter1: overall_satisfaction ~ room_type + reviews + accommodates * minstay +
## inter1:     price + (1 + accommodates | neighborhood)
## inter2: price ~ room_type + reviews + accommodates * minstay + overall_satisfaction +
## inter2:     (1 + accommodates | neighborhood)
##        Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## inter1 12 2232.8 2302.6 -1104.4   2208.8
## inter2 12 4975.9 5045.7 -2476.0   4951.9     0      0          1
```

```
"from the output, when consider interaction, inter1 is better at the others, so inter1 is the best fit
```

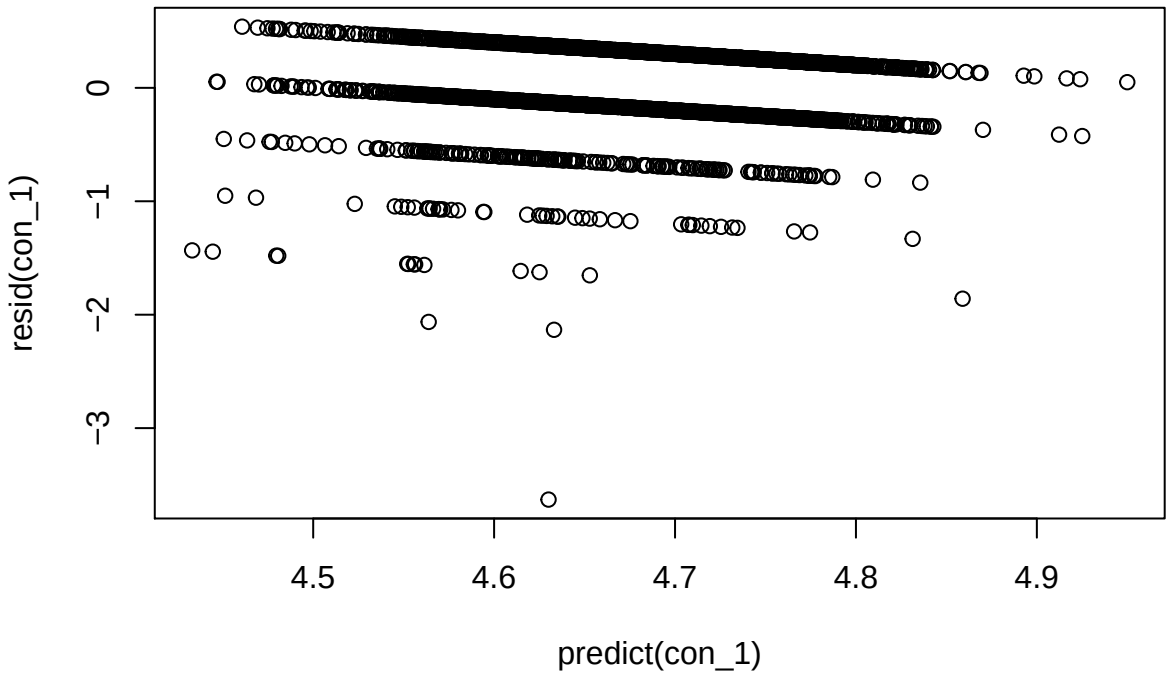
```
## [1] "from the output, when consider interaction, inter1 is better at the others, so inter1 is the be
```

H.2 Check assumptions for multilevel model

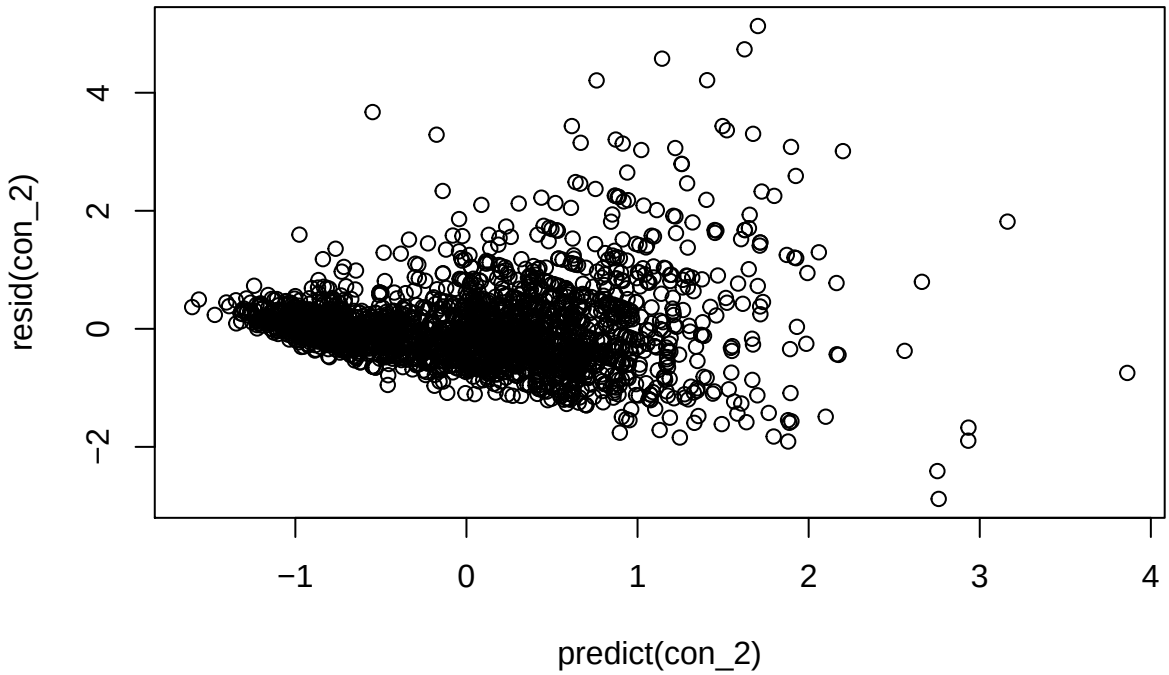
```
"1. Check Linearity - residual analysis "
```

```
## [1] "1. Check Linearity - residual analysis "
```

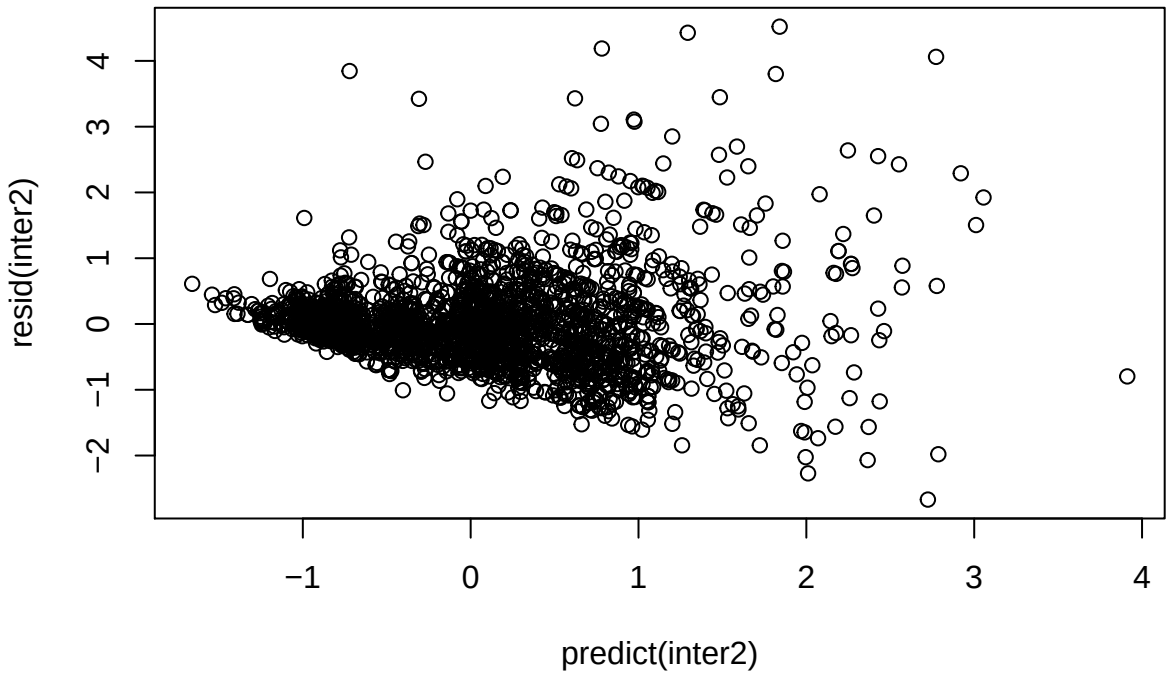
```
plot(predict(con_1),resid(con_1))
```




```
plot(predict(con_2),resid(con_2))
```

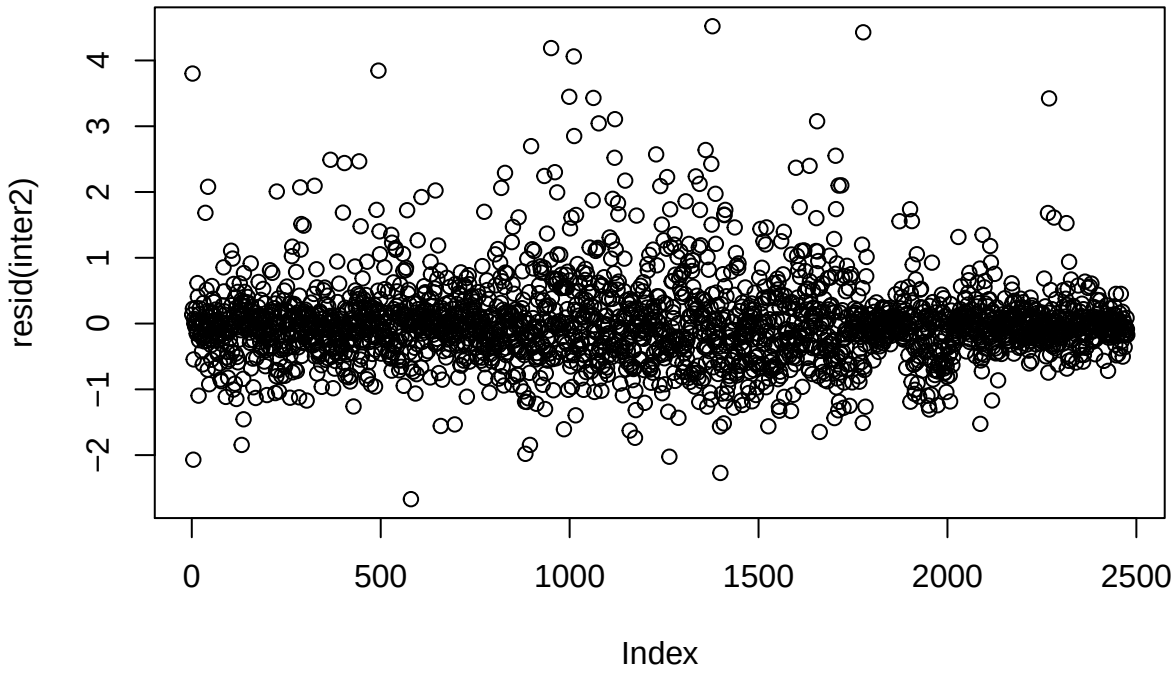



```
plot(predict(inter2),resid(inter2))
```




```
# con_1,con_2 and inter1 residual plot is not normal pattern, so i reject the inter1 model, and use int
```

```
plot(resid(inter2))
```




```
"2.Check Homogeneity of Variance - ANOVA"
```

```
## [1] "2.Check Homogeneity of Variance - ANOVA"
```

```
anova(con_1)
```

```
## Analysis of Variance Table
##                   Df  Sum Sq Mean Sq F value
```

```
## factor(room_type)  2 0.25331 0.12665  0.8974
## reviews            1 0.14917 0.14917  1.0569
## accommodates       1 0.07912 0.07912  0.5606
## minstay            1 0.18410 0.18410  1.3044
## price              1 2.63409 2.63409 18.6630
```

```
anova(con_2)
```

```
## Analysis of Variance Table
##                      Df Sum Sq Mean Sq F value
## factor(room_type)     2 407.41  203.70 455.379
## reviews               1   7.53    7.53  16.841
## accommodates          1 384.06  384.06 858.558
## minstay               1   1.30    1.30   2.901
## overall_satisfaction  1   8.83    8.83  19.738
```

```
anova(inter2)
```

```
## Analysis of Variance Table
##                      Df Sum Sq Mean Sq  F value
## room_type             2 85.293  42.647 102.5460
## reviews               1  6.770   6.770  16.2796
## accommodates          1 64.006  64.006 153.9049
## minstay               1  0.954   0.954   2.2946
## overall_satisfaction  1  7.626   7.626  18.3371
## accommodates:minstay  1 28.255  28.255  67.9401
```

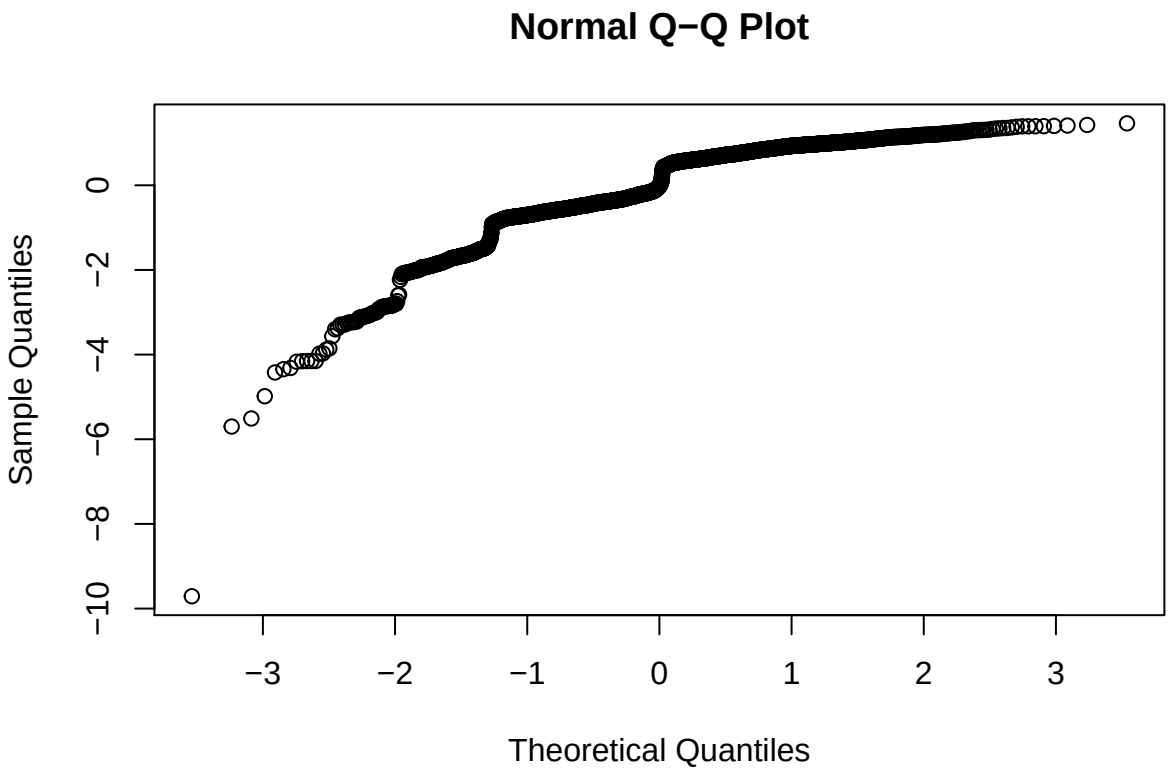
```
"Since the all the p values are greater than 0.05, we can say that the variance of the residuals
is equal and therefore the assumption of homoscedasticity is met."
```

```
## [1] "Since the all the p values are greater than 0.05, we can say that the variance of the residuals
```

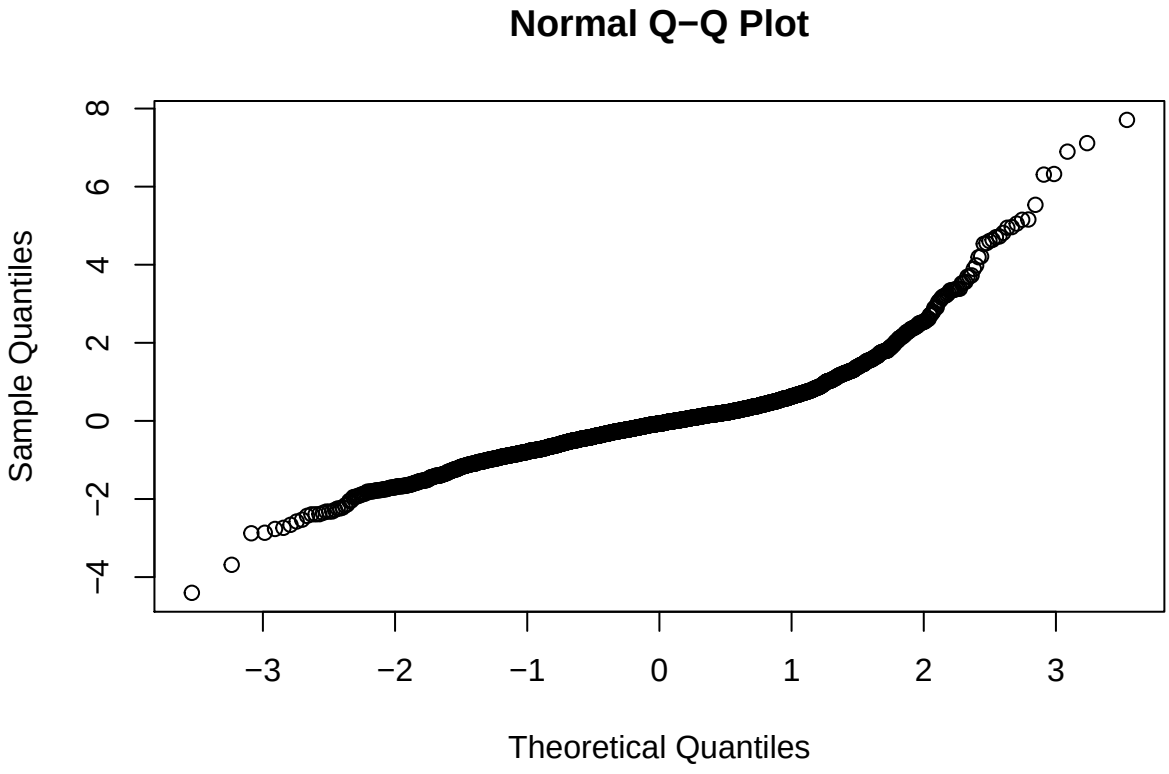
```
"3. Check if the residuals of the model are normally distributed."
```

```
## [1] "3. Check if the residuals of the model are normally distributed."
```

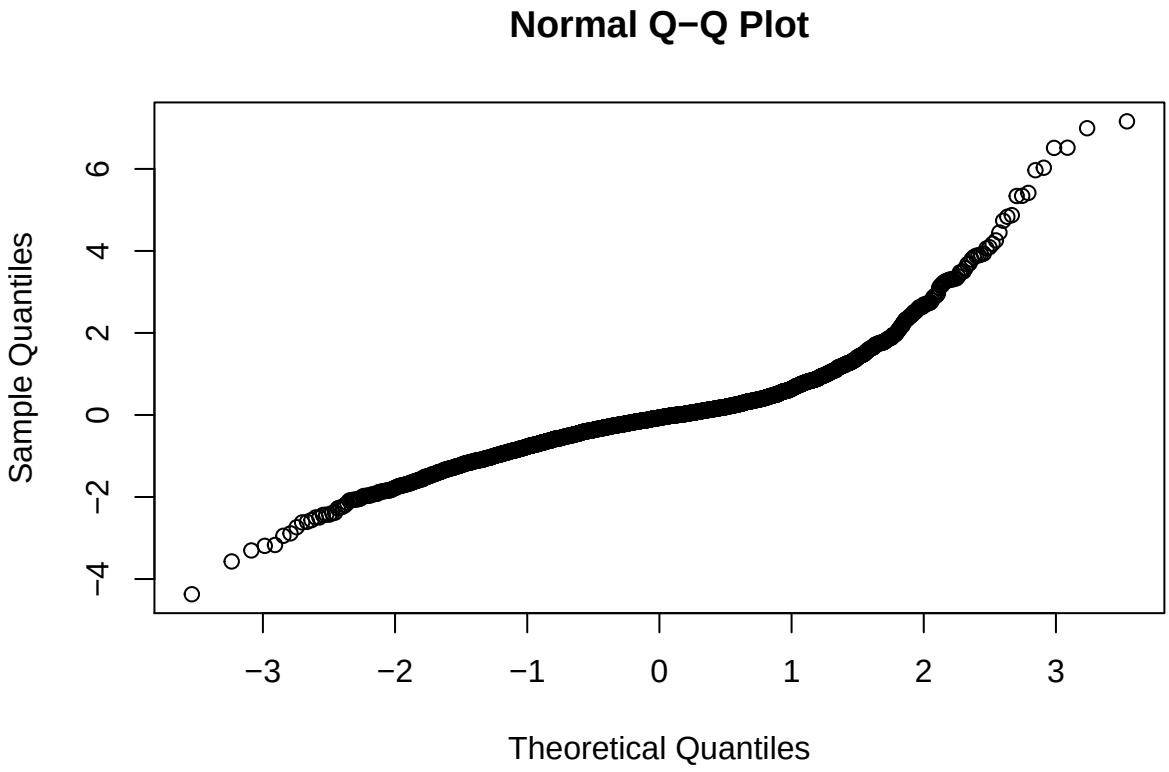
```
s1=rstudent(con_1)
qqnorm(s1)
```




```
s2=rstudent(con_2)
qqnorm(s2)
```

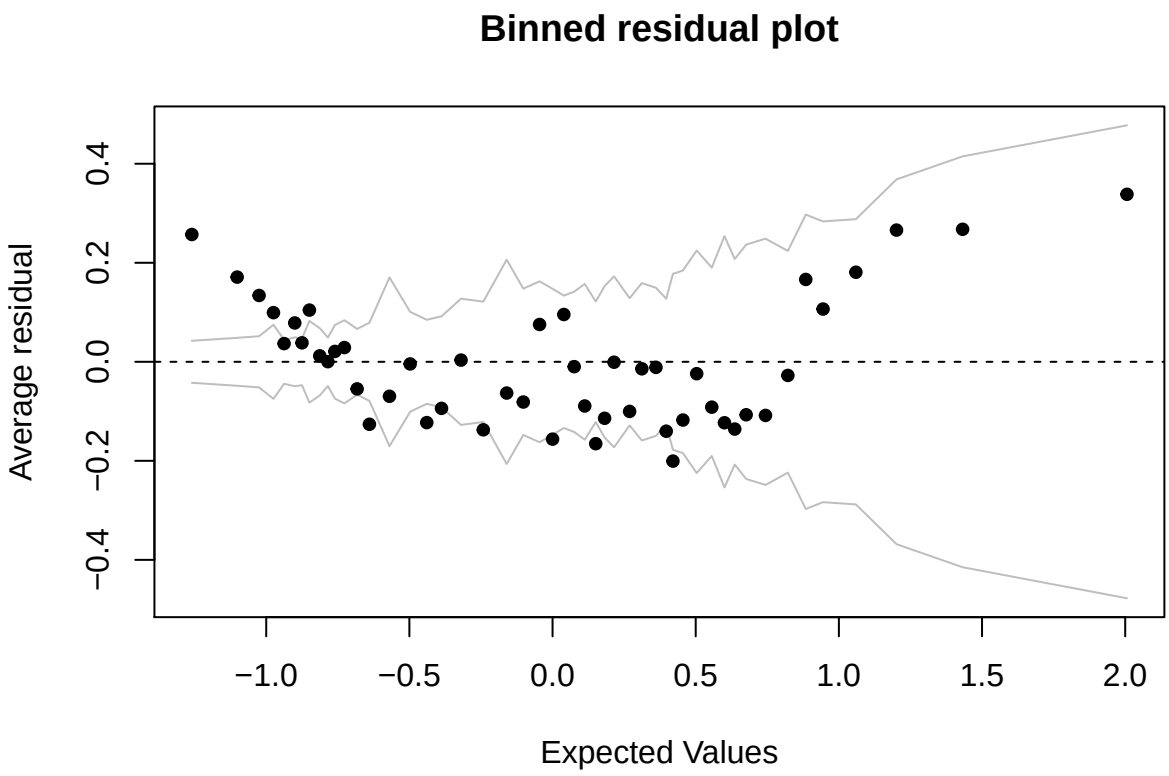



```
s3=rstudent(inter2)
qqnorm(s3)
```

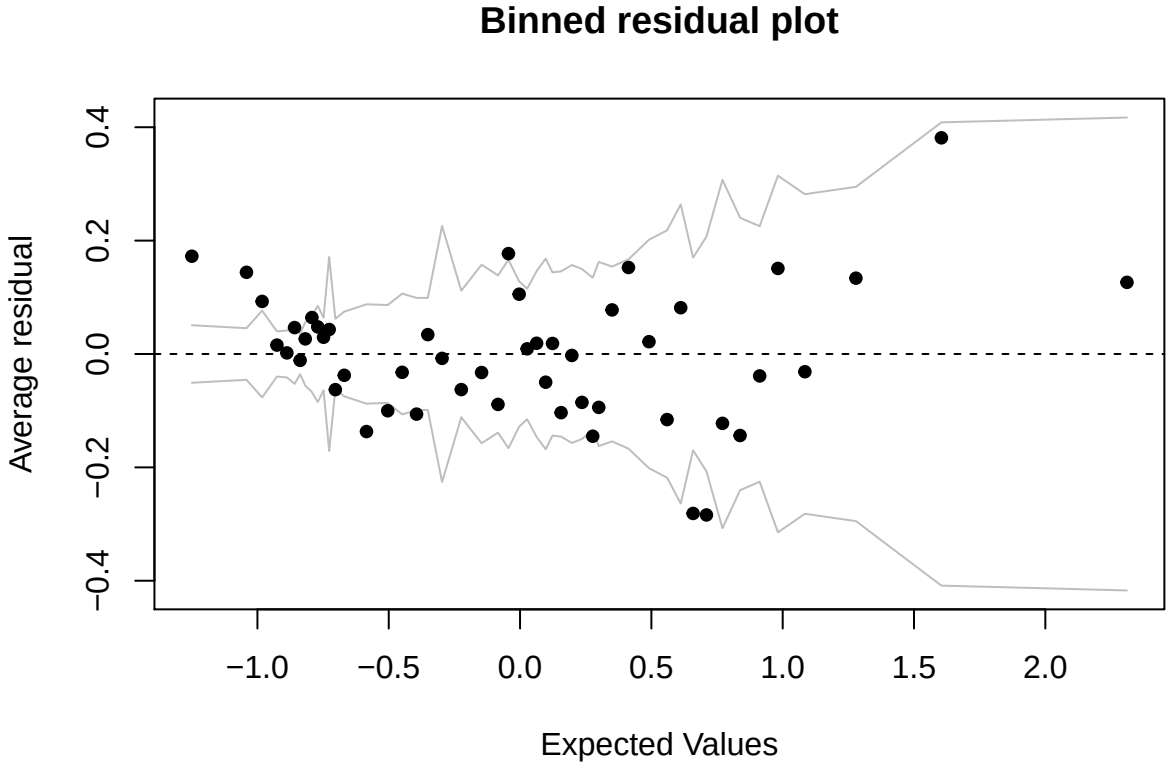


From the plot we can see that first model (con_1) model, data distribution is not linear, so it doesn't meet my model selection critia, so ratings as outcome variable is not suitable for building the model in general. The second one and the last one is good, so i will continue my multilevel model analysis based on the last two models.

```
#Binned plot for random intercept and random slope - price as outcome variable
binnedplot(fitted(con_2),residuals(con_2, type="response"))
```




```
# Binned plot for random intercept and random slope with interactions - price as outcome variable
binnedplot(fitted(inter2),residuals(inter2, type="response"))
```



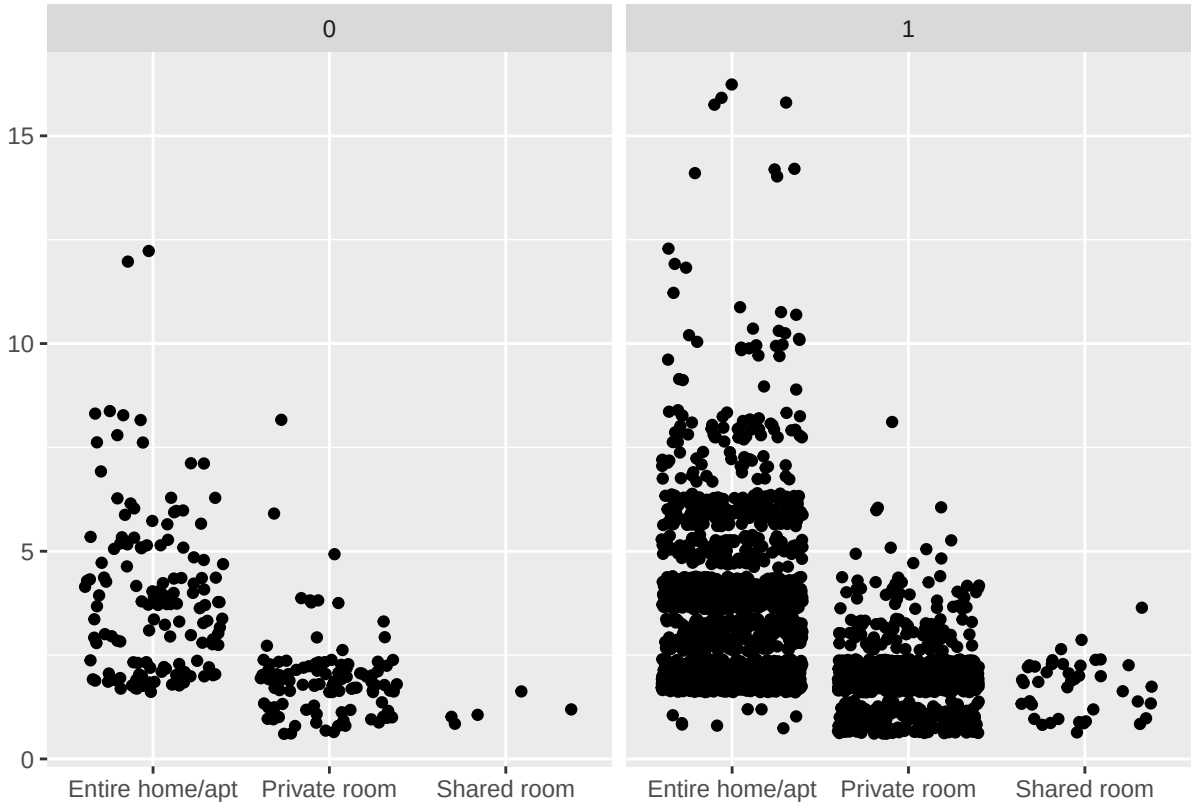
I. Model Improvement

I improved the model (ratings as outcome variable) by adding a new column according to the ratings data - form a multilevel logistic model:

```
binarydata<-function(data){ result=data
for (row in 1:nrow(result)){
for (n in names(result)[6]){ if (result[row,n]>=4.5){
result$binarydata[row] = 1 }else{
result$binarydata[row] = 0 }
} }
return(result)
}

improv_data<-binarydata(Boston.data)

ggplot(improv_data)+aes(x=room_type,y=accommodates)+ geom_jitter()+facet_grid(.~binarydata)+ scale_fill_
```



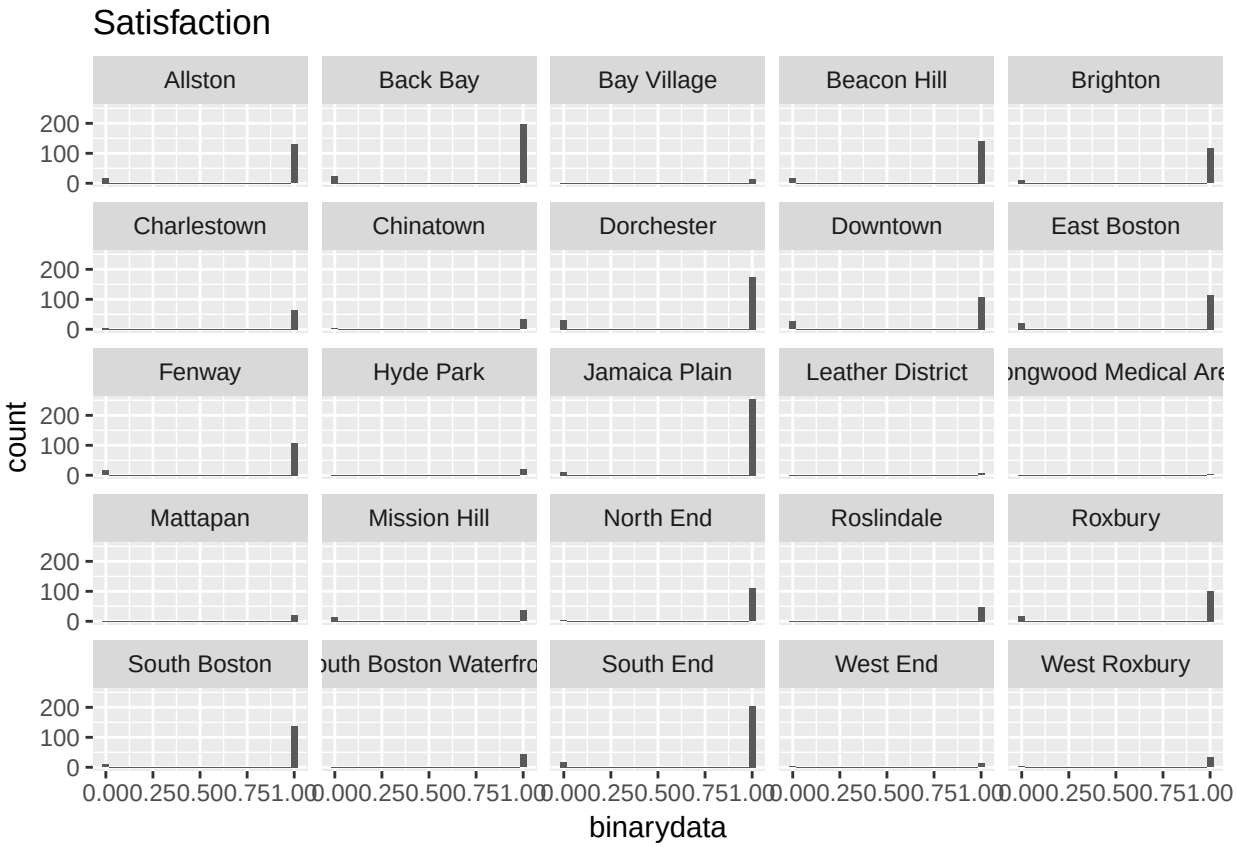

```
# creating logistic model
log1<-glm(binarydata~room_type+reviews+accommodates+minstay+price, family = binomial, data=improv_data)
display(log1)
```

```
## glm(formula = binarydata ~ room_type + reviews + accommodates +
##     minstay + price, family = binomial, data = improv_data)
##                       coef.est coef.se
## (Intercept)            2.25     0.23
## room_typePrivate room -0.08     0.18
## room_typeShared room   0.09     0.50
## reviews                0.01     0.00
```

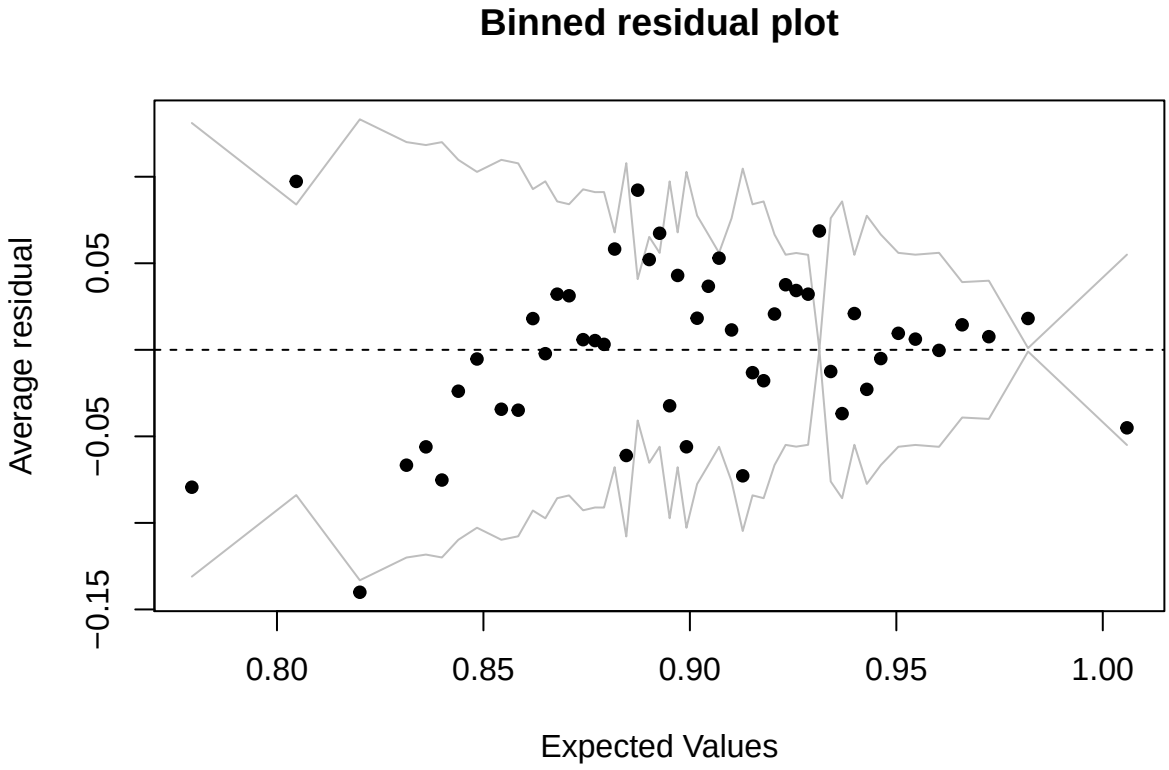
```
## accommodates          -0.04     0.05
## minstay               -0.02     0.03
## price                  0.29     0.11
## ---
##   n = 2476, k = 7
##   residual deviance = 1603.8, null deviance = 1624.7 (difference = 20.9)
```

```
# fit multilevel logistic model
ggplot(data=improv_data, aes(x=binarydata))+geom_histogram()+ ggtitle("Satisfaction")+facet_wrap(~neighb
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Adding interaction
log2<-lmer(binarydata~room_type+reviews+accommodates*price+minstay+(1+accommodates|neighborhood), data=
binnedplot(fitted(log2),residuals(log2, type="response"))
```



J. Conclusion

The goal of this project is to explore factors that may have impact on ratings or price. After EDA and multilevel model building process, i think ratings as outcome is not suitable for this project for fitting the multilevel model. So i focus more on the price as outcome factors, i found out that number of reviews, ratings , accomodates, neighborhood and room type are factors that may affact the price. Accomodates and ratings have positive effect on the price, other variables have negative effect. Among all the variables, room type has the most significant effect on the price on Airbnb because it would cost a lot more when it comes to entire house/ apt than shared room or private room.

Although ratings as outcome doesn't fit my multilevel model well, but i did some improvement on the model (refer the model improvement part) and i did some useful points during the EDA process. Price, accomodates and neighborhood are the major factors that may influence the ratings on Airbnb. These all make sense since the neighborhood is related to crime rate in the area, properties in a nice place tend to have more high ratings. I think these are all useful points for Airbnb users or owners to know about.

K. Future Implication

I will do further research and study for the categorical multilevel model building to improve the ratings model from this project. I learnt about from doing this project and it also help me get to know better about Airbnb and the whole industry. I think for users and property owners, these analysis are useful and important to know as a reference. For my future exploration of this dataset, I would divide the dataset into training and testing, to better test whether the regressions and models.

Appendix

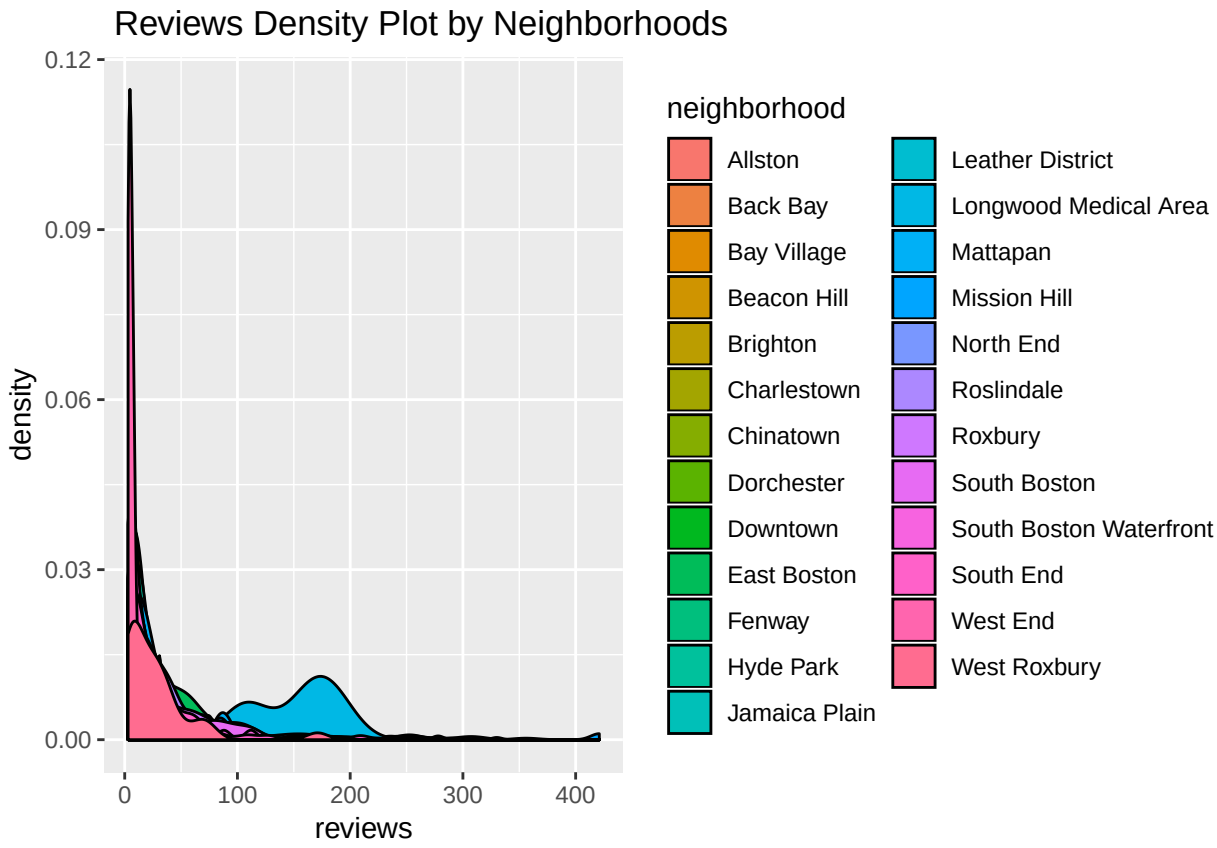
Summary of Boston data

```
summary(Boston.data)
```

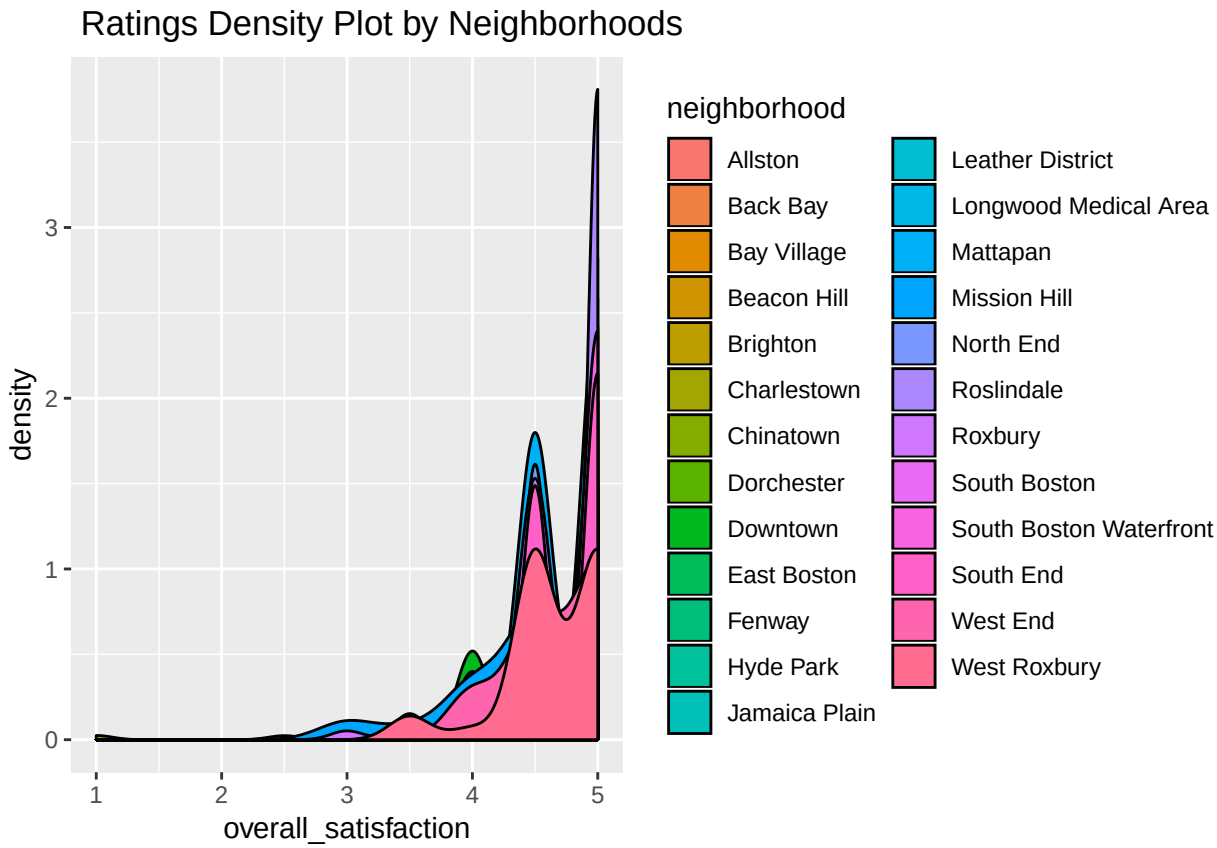
```
##     room_id            host_id                    room_type
##  Min.   :    3781   Min.   :     4240   Entire home/apt:1483
##  1st Qu.: 4301413   1st Qu.:  5618949   Private room   : 947
##  Median : 8404690   Median : 17904726   Shared room    :  46
##  Mean   : 8485211   Mean   : 26030825
##  3rd Qu.:13219410   3rd Qu.: 39803926
##  Max.   :15916518   Max.   :103145926
##
##         neighborhood     reviews       overall_satisfaction
##  Jamaica Plain: 265   Min.   :  3.00   Min.   :1.000
##  South End    : 222   1st Qu.:  6.00   1st Qu.:4.500
##  Back Bay     : 220   Median : 15.00   Median :4.500
##  Dorchester   : 206   Mean   : 31.69   Mean   :4.679
##  Beacon Hill  : 157   3rd Qu.: 38.00   3rd Qu.:5.000
##  Allston      : 148   Max.   :421.00   Max.   :5.000
##  (Other)      :1258
##   accommodates         price.V1          minstay          latitude
##  Min.   : 1.000   Min.   :-1.410904   Min.   : 0.000   Min.   :42.24
##  1st Qu.: 2.000   1st Qu.:-0.733880   1st Qu.: 1.000   1st Qu.:42.33
##  Median : 2.000   Median :-0.214519   Median : 2.000   Median :42.35
##  Mean   : 3.174   Mean   : 0.000000   Mean   : 1.975   Mean   :42.34
##  3rd Qu.: 4.000   3rd Qu.: 0.341939   3rd Qu.: 2.000   3rd Qu.:42.36
##  Max.   :16.000   Max.   : 6.833951   Max.   :27.000   Max.   :42.39
##
##    longitude                         last_modified  priceperperson
##  Min.   :-71.17   2016-11-21 04:37:48.873977:   1   Min.   :  5.50
##  1st Qu.:-71.10   2016-11-21 04:38:05.185518:   1   1st Qu.: 35.00
##  Median :-71.08   2016-11-21 04:38:23.807348:   1   Median : 49.50
##  Mean   :-71.08   2016-11-21 04:38:50.511911:   1   Mean   : 56.15
##  3rd Qu.:-71.06   2016-11-21 04:38:53.731677:   1   3rd Qu.: 71.54
##  Max.   :-71.00   2016-11-21 04:39:01.808247:   1   Max.   :500.00
##                   (Other)                   :2470
```

(1) Density Plot

```
ggplot(data=Boston.data, aes(x=reviews, fill=neighborhood))+geom_density()+ ggtitle(" Reviews Density Pl
```

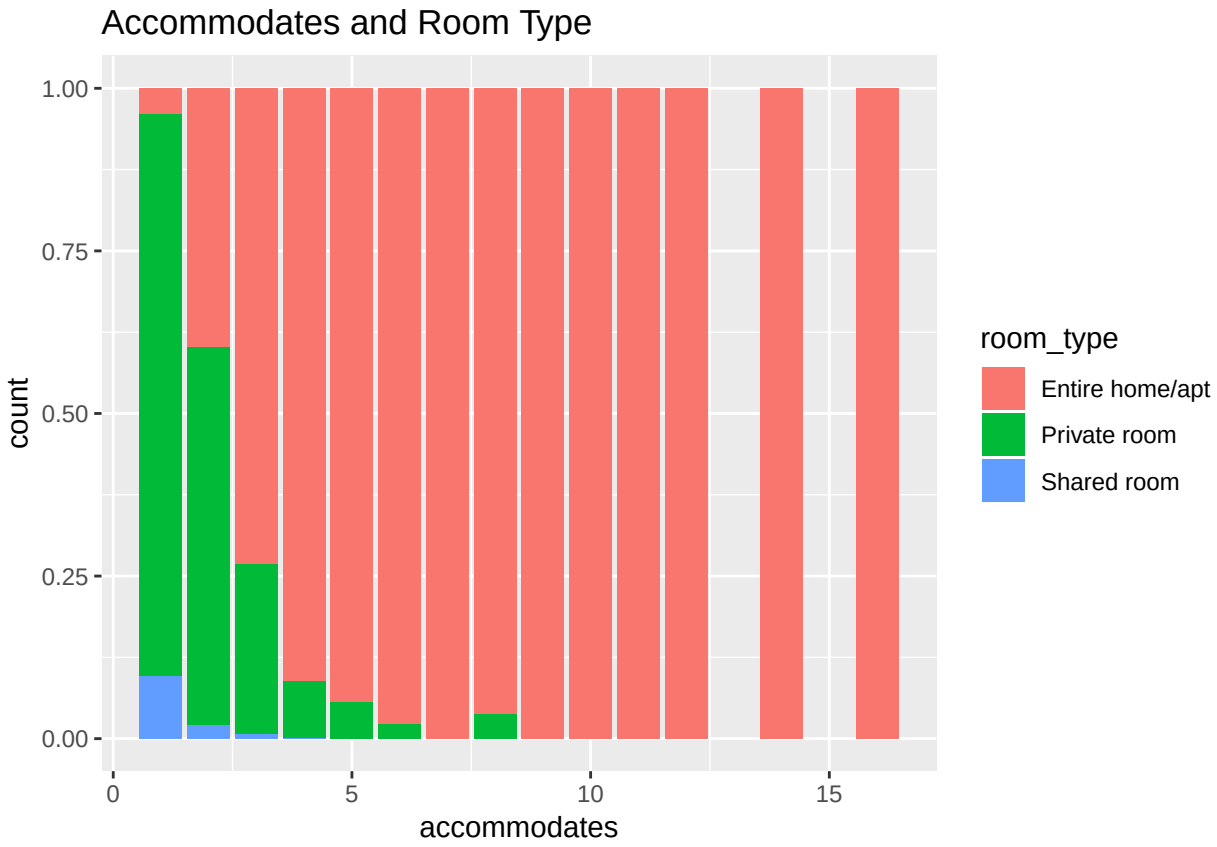



```
ggplot(data=Boston.data, aes(x=overall_satisfaction, fill=neighborhood))+geom_density()+ ggtitle(" Rati
```

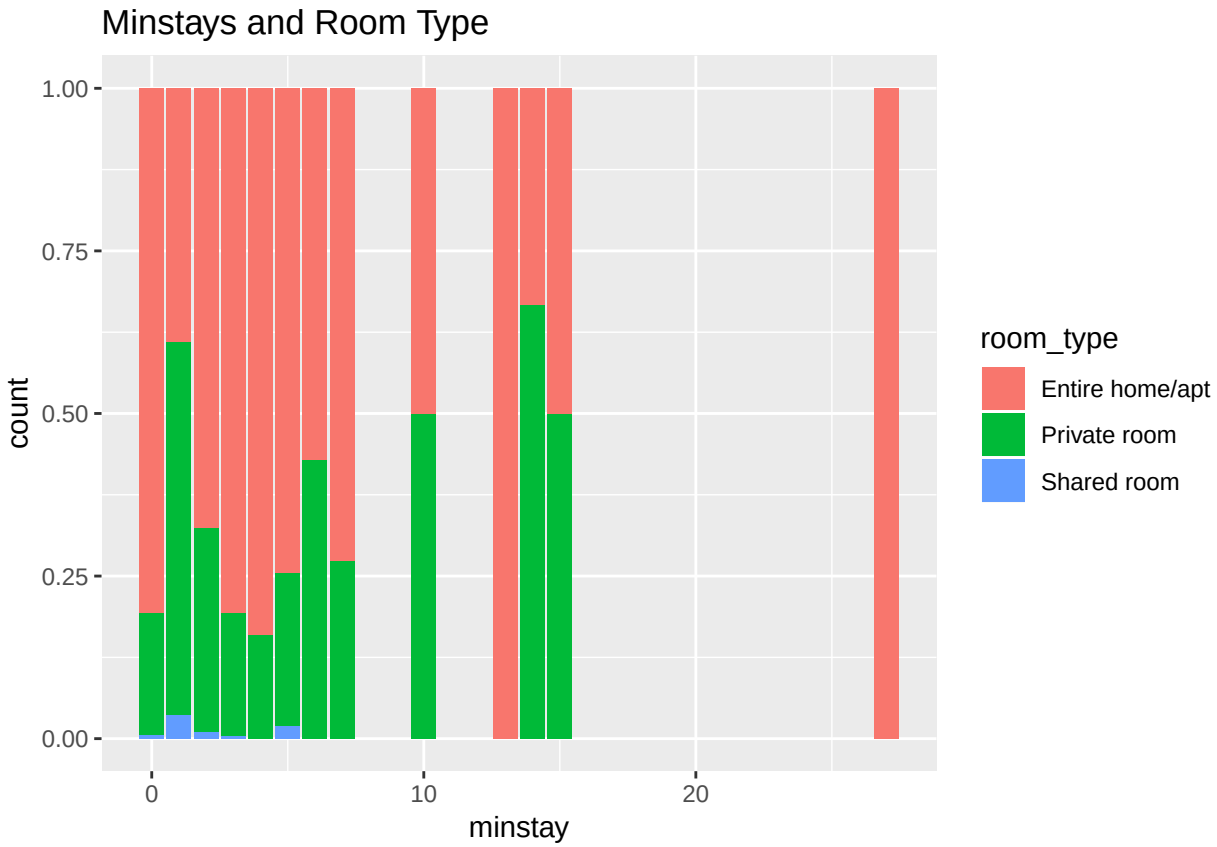


we can see that the distribution of ratings and number of reviews are different between neighborhoods.

(2) relationship between accomodates and room type

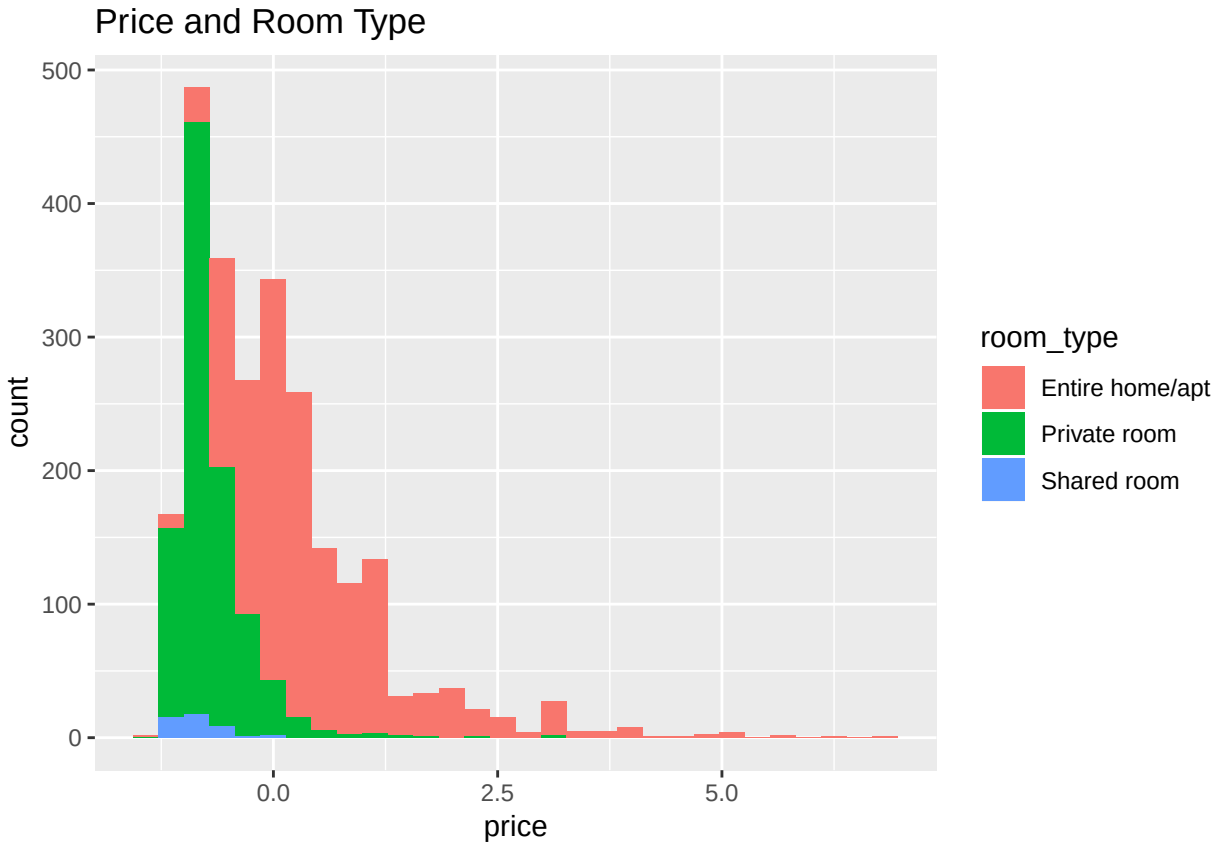


(3) relationship between mininum stay dates and room type



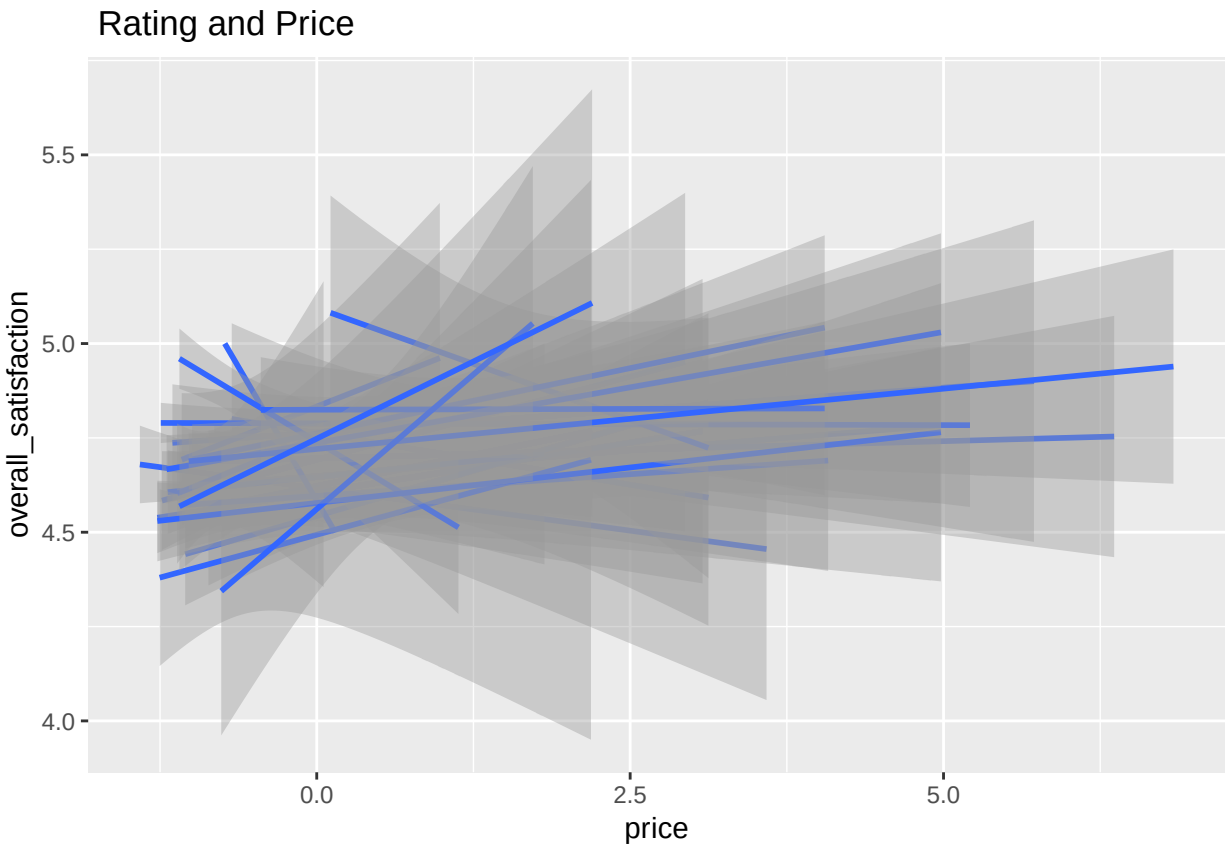
(4) relationship between prices per night versus room type

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



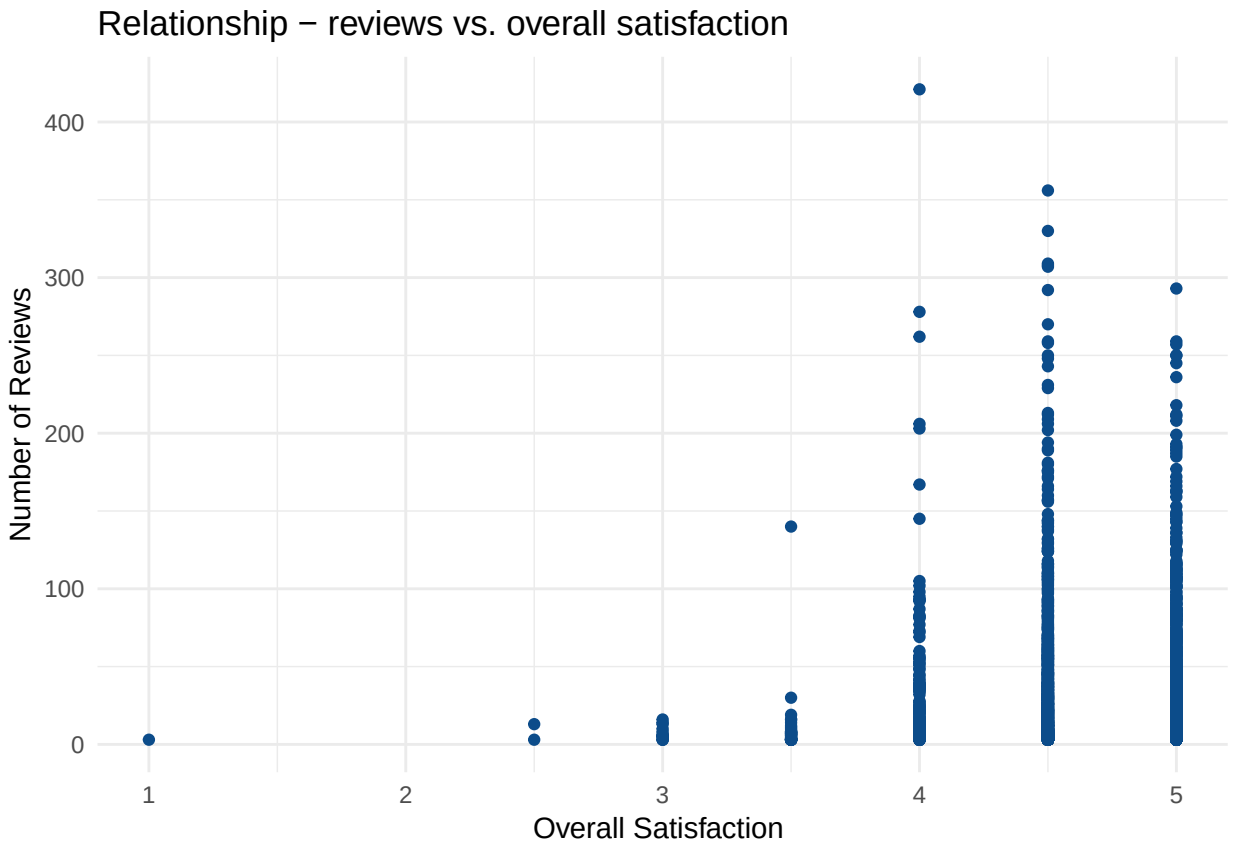
(5) Plot for Rating and Price

```
ggplot(data=Boston.data, aes(x=price, y=overall_satisfaction, group=neighborhood))+ geom_smooth(method
```

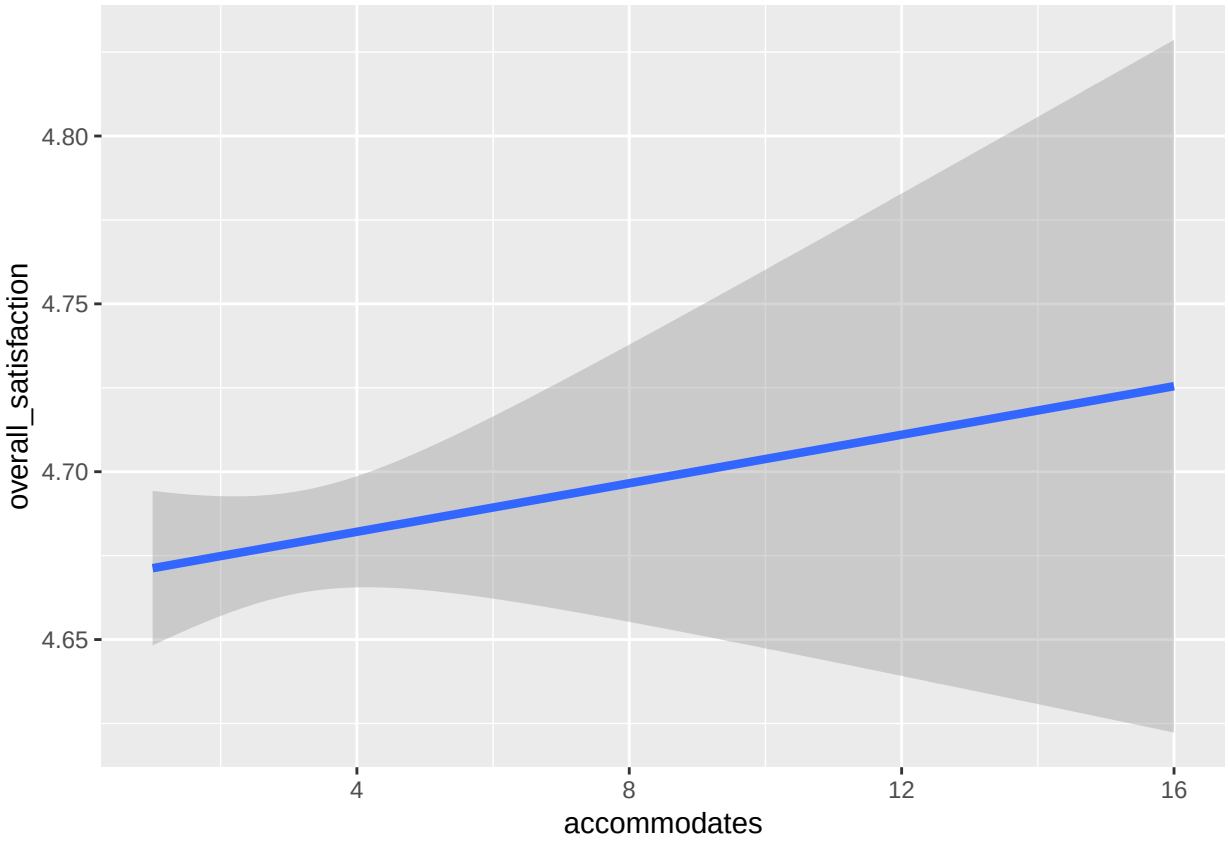


(6)

```
ggplot(data = Boston.data) +
  aes(x = overall_satisfaction, y = reviews) +
  geom_point(color = '#0c4c8a') +
  labs(title = 'Relationship - reviews vs. overall satisfaction',
    x = 'Overall Satisfaction ',
    y = 'Number of Reviews') +
  theme_minimal()
```


Relationship − reviews vs. overall satisfaction
Number of Reviews
400
300
200
100
0
1
2
3
4
5
Overall Satisfaction

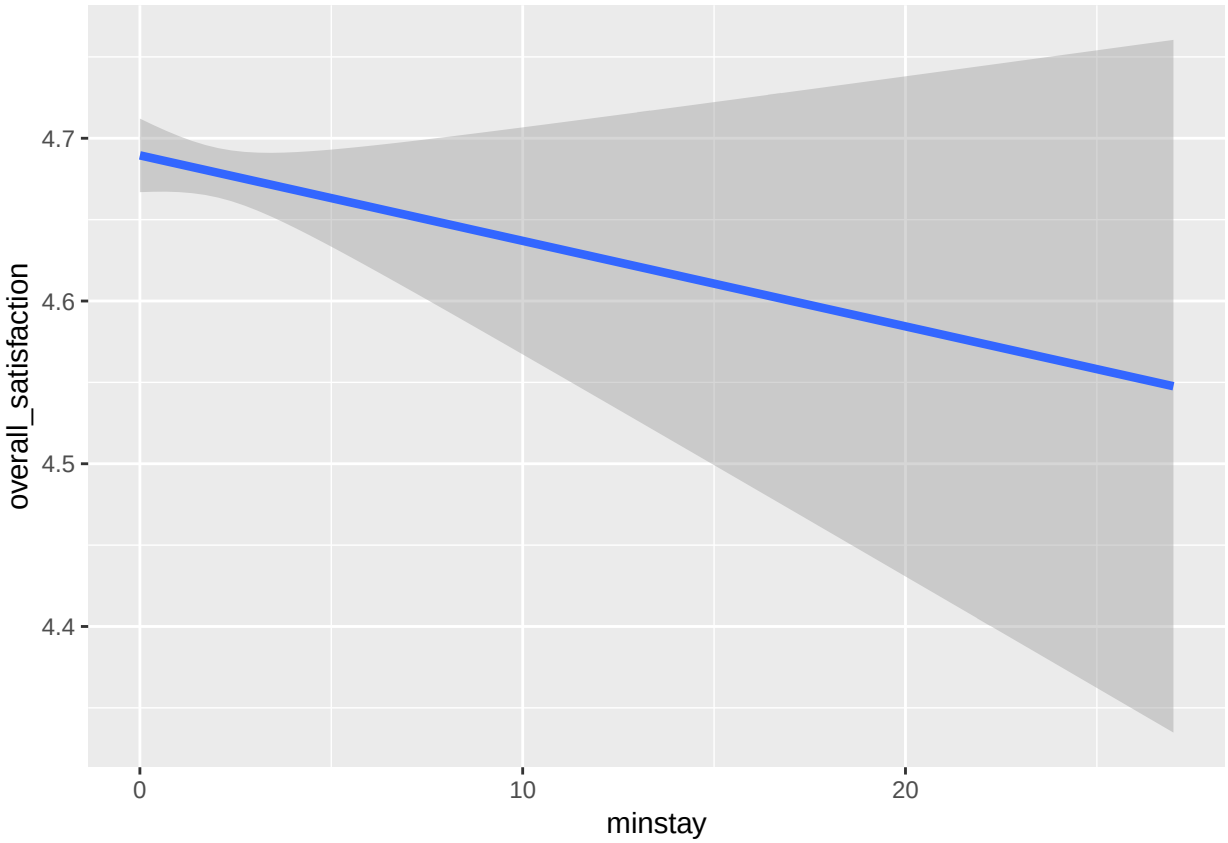
(7)Explore relationship between ratings and accommodates



I tried the plot with geom_point and geom_hex, they are look pretty weird and doesn't reflect what I was looking for. Since my focuses are what kind of relationship (positive or negative or no relationship) between the two variables, so i will just use geom_sooth line this this case.

As we can see from the graph, there is a moderate positive relationship betwwen the ratings and the accomodates

(8) Explore relationship between ratings and minstay



As we can see from the graph, there is a negative relationship betwwen the ratings and the minimum stays, it makes sense becasue the minimum stays will block other cumstomers (who want to stay less than the minimum nights) from choosing the property.