Midterm Project

Chuning Yuan

2019/12/5

I.Introduction

Nowadays, Airbnb has became very popular for the travelers due to its unique style and creative design and competitive prices and location than the hotels. Therefore I think it will be interesting to explore the data set of Airbnb, we can look at how are the price for a night stay for each room were affected by other variables such as reviews, number of bedrooms, minimum stays etc.This project will contain analysis on Airbnb data with EDA(Exploratory Data Analysis) and modeling. From this project,we will be able to have a general idea how to predict the price of rooms on Airbnb website and what's the most influential factor in predicting process.

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 us the service to arrange or offer lodging, primarily homestays, or tourism experiences. I choose the data of San Francisco is also because it is where Airbnb has grown around the world from, and I have the experience of searching Airbnb in bay area. First, 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 multi-level regression using room type and neighborhood and few other factors to predict the price.

II. Data

Data source

The data set was extracted from the tomslee.net website– Airbnb Data Collection: Get the Data. There are zip file for many cities around the world. The zip file holds one or more csv files. Each csv file represents a single "survey" or "scrape" of the Airbnb web site for that city.

Specifically, the data used for this project was collected from the November 2013 to January 2017 in San Francisco. There are 9 variables that will be used in this project. Specifically, they are room id, host id, room type, neighborhood, number of reviews, overall satisfaction, number of accommodates, number of bedrooms and price.We are assuming the potential factors to influence the pricing are the room type, neighborhood, number of reviews, overall satisfaction, number of accommodates and number of bedrooms.

Data Cleaning

For a large dataset, we want to extract the information as more useful and informative as possible, so we filter the observation with more than 100 reviews because these could be more representative in general, and extract neighborhood with more than 30 observations for the further modeling analysis. Below is the data overview after the cleaning processes.

Overview of data

Table 1. The head of the data:

room_id	host_id	room_type	neighborhood	reviews	overall_satisfaction	price
6910758	30920210	Shared room	South of Market	125	5.0	99
259622	329072	Shared room	Financial District	117	4.5	45
229240	329072	Shared room	Financial District	194	4.5	45
70753	329072	Shared room	Financial District	206	4.5	45
4518031	22931450	Shared room	North Beach	138	4.5	56
4519780	22931450	Shared room	North Beach	101	4.5	56

Table 2. Summary of the data:

```
##     room_id           host_id                    room_type
##  Min.   :    5858   Min.   :      46   Entire home/apt:1794
##  1st Qu.:  678556   1st Qu.: 1032643   Private room   :1955
##  Median : 1827653   Median : 4393613   Shared room    :  92
##  Mean   : 2689704   Mean   : 9009894
##  3rd Qu.: 4092288   3rd Qu.:11655078
##  Max.   :13901641   Max.   :77807259
##
##  borough                        neighborhood     reviews
##  Mode:logical   Mission            : 497   Min.   :101.0
##  NA's:3841      Castro/Upper Market: 416   1st Qu.:118.0
##                 Western Addition   : 328   Median :142.0
##                 Bernal Heights     : 259   Mean   :162.5
##                 Haight Ashbury     : 232   3rd Qu.:186.0
##                 Noe Valley         : 219   Max.   :513.0
##                 (Other)            :1890
##  overall_satisfaction  accommodates       bedrooms         price
##  Min.   :4.000        Min.   : 1.000   Min.   :0.000   Min.   :  35.0
##  1st Qu.:4.500        1st Qu.: 2.000   1st Qu.:1.000   1st Qu.:  95.0
##  Median :5.000        Median : 2.000   Median :1.000   Median : 125.0
##  Mean   :4.825        Mean   : 2.813   Mean   :1.088   Mean   : 140.6
##  3rd Qu.:5.000        3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 160.0
##  Max.   :5.000        Max.   :14.000   Max.   :4.000   Max.   :1000.0
##
##  minstay
##  Mode:logical
##  NA's:3841
##
##
##
##
##
```

According to the summary we can delete the the column borough and minstay because there is no value in them. After eliminate all the NA, now we have the cleaned data we need, then we can start our EDA process.

III. EDA

Figure 1. Distribution of room price

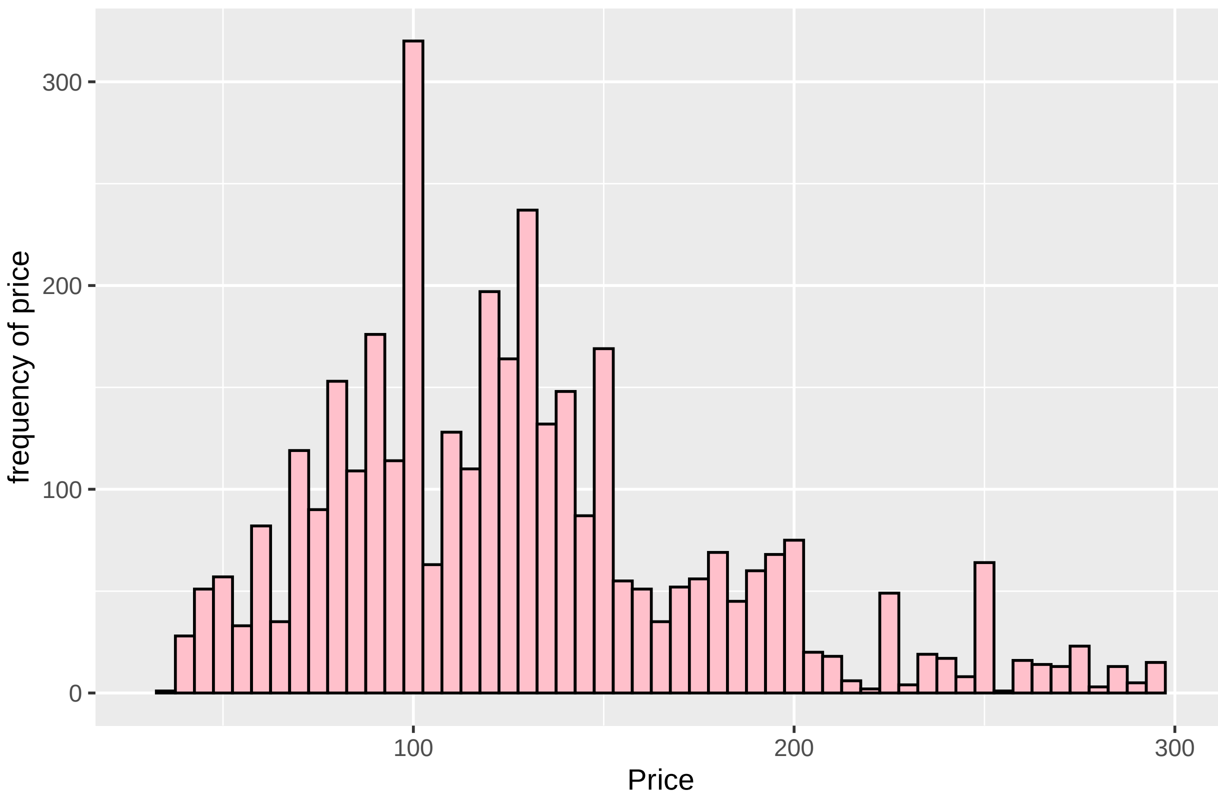
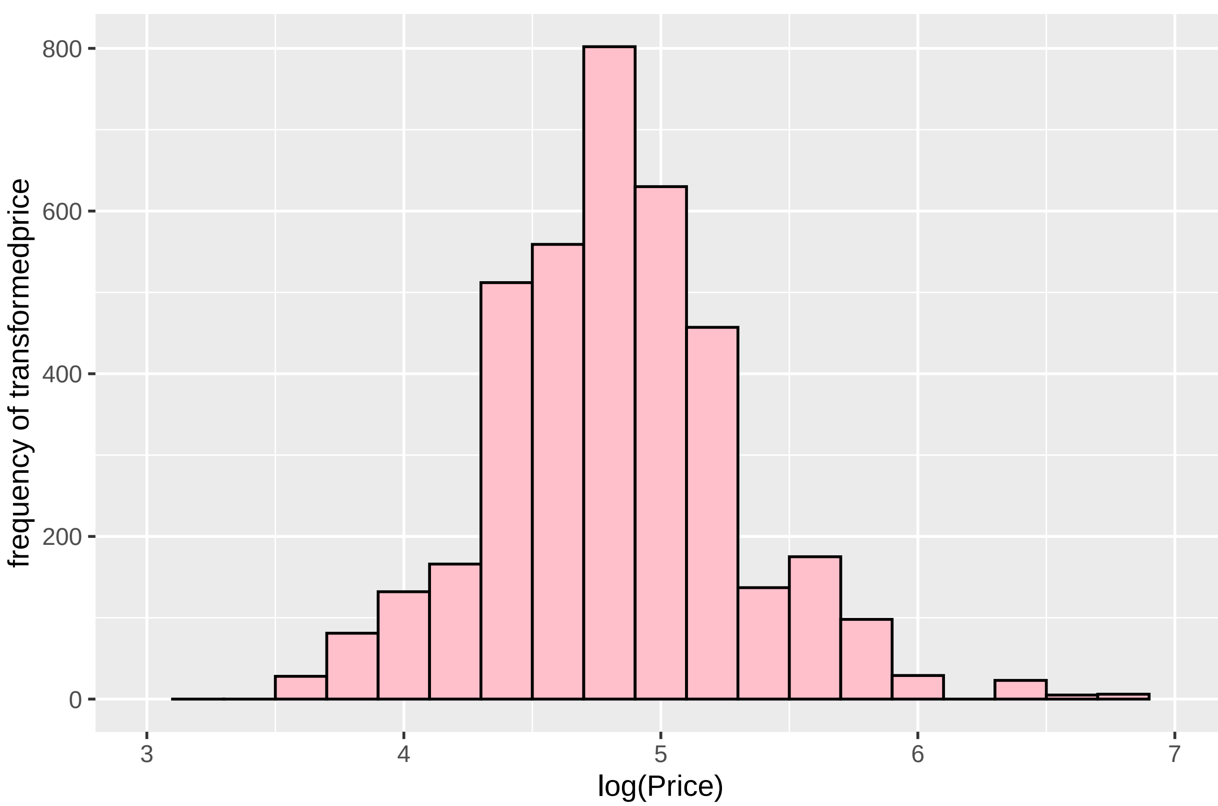
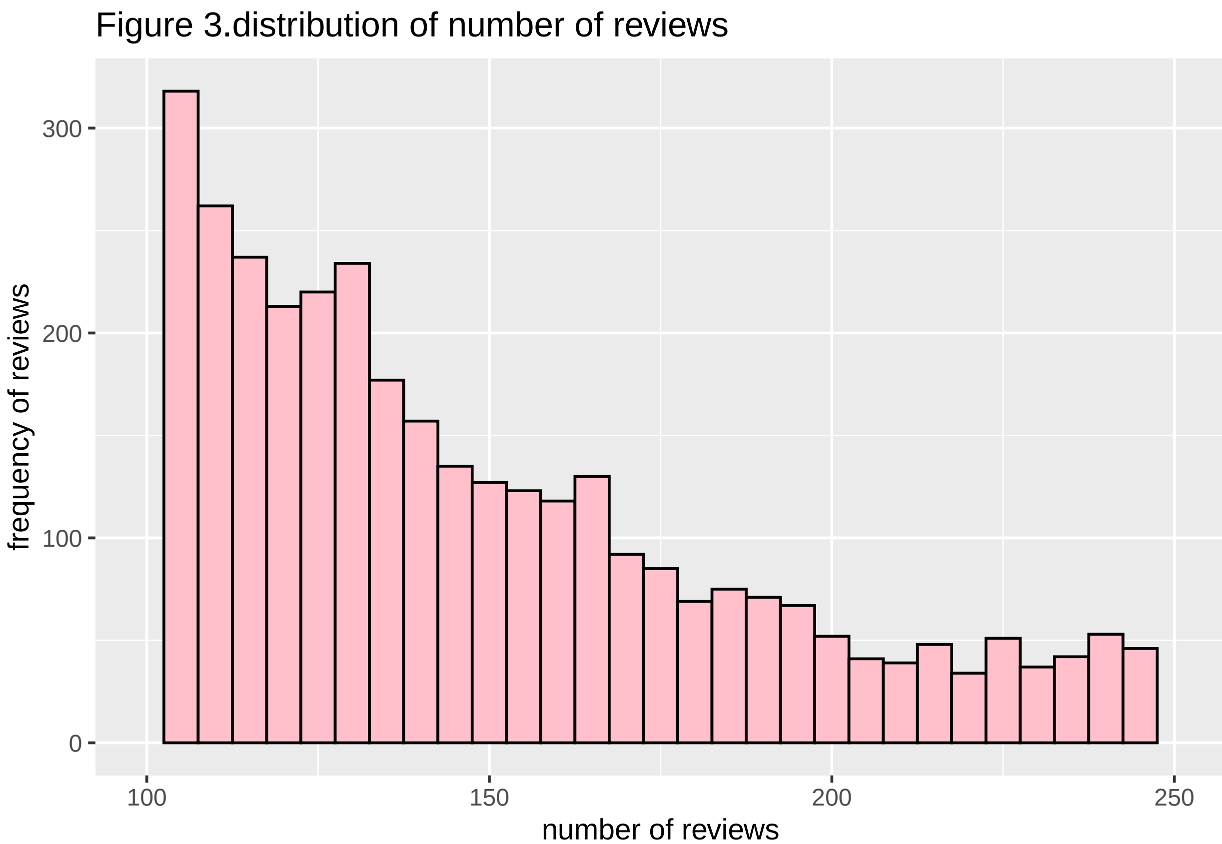


Figure 2.Distribution of room price



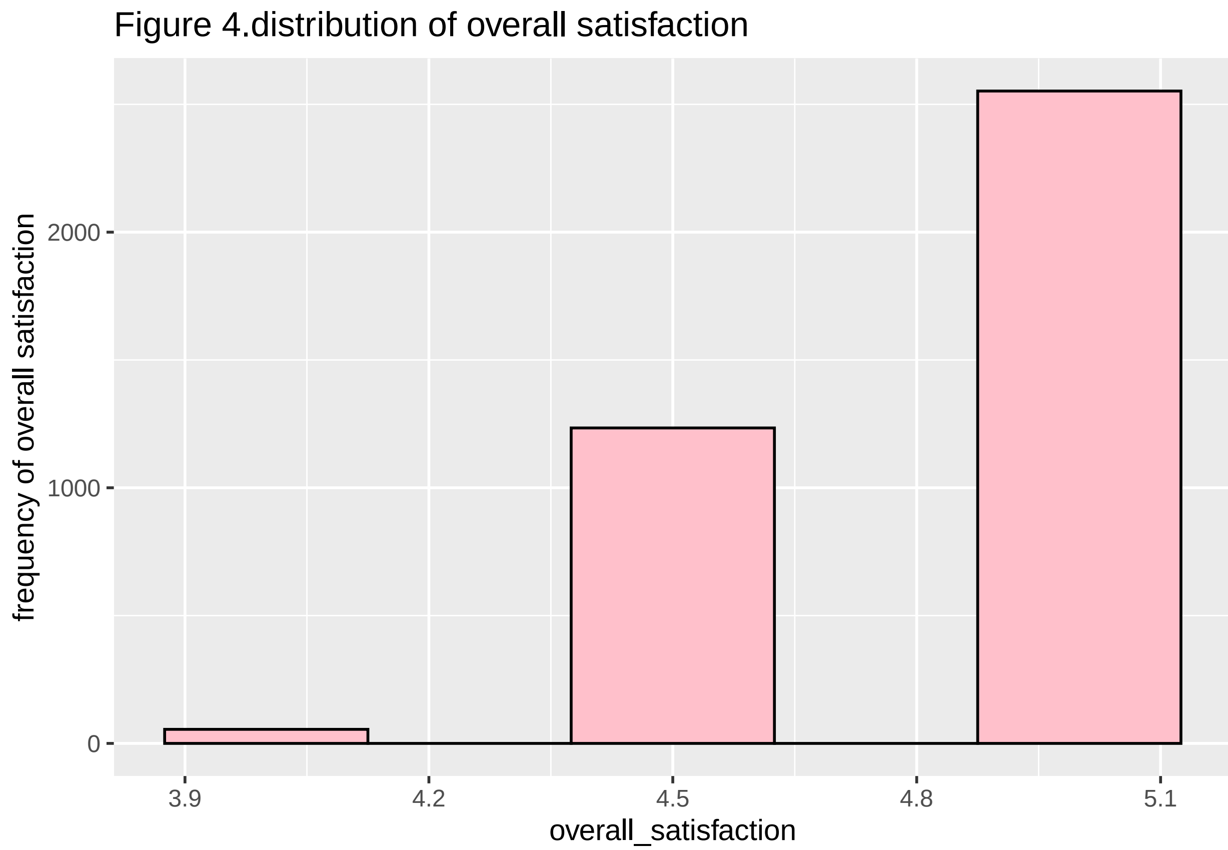
What is the price range for SF Airbnb, what would be a common price we should expect when we are searching the Airbnb in SF?

We take a look at the room price distribution, and it is obvious that the most popular price are around 100, and between 100-150.



How many reivew would be a common amount we should expect when we are searching the Airbnb in SF?

Since we have already filter the reviews that is less than 100, here the distribution plot can tell us that most of the reviews are around 100 to 175,most of the reivews are ubder 200. This provide us some general ideas of number of reivews for the San Francisco Airbnb, so we can know how much reviews to expect when we are choosing the Airbnb from the website based on the reivews.



How are the Airbnb rated in SF Airbnb, is there any extreme good or bad rating we should be aware of?

In this histgram plot, most of overall rating is aroud 4.5 and 5. There are also a small portion of people rate the room 3.9 to 4 star. Overall, customers are satisfied with most of rooms in San Francisco area.

Figure 5.Average number of reviews per neighbor

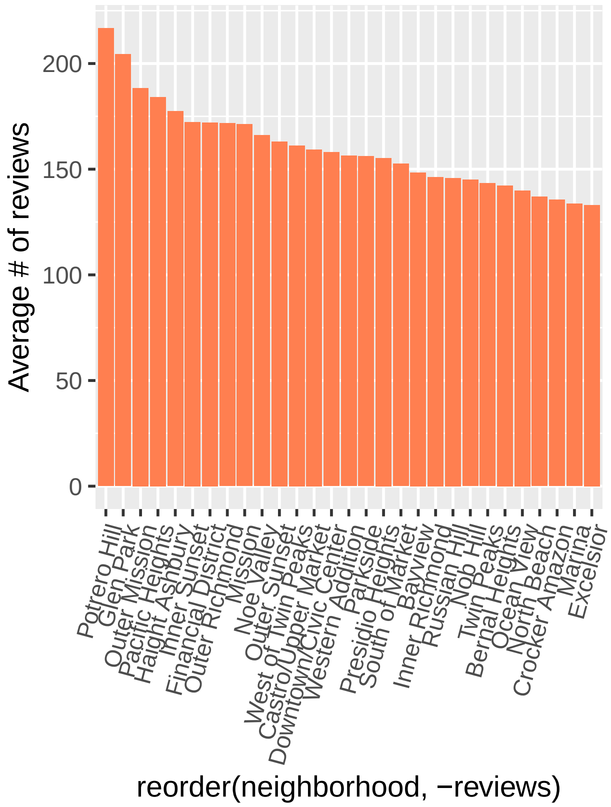


Figure 6.Average rating of airbnb rooms per neighbo

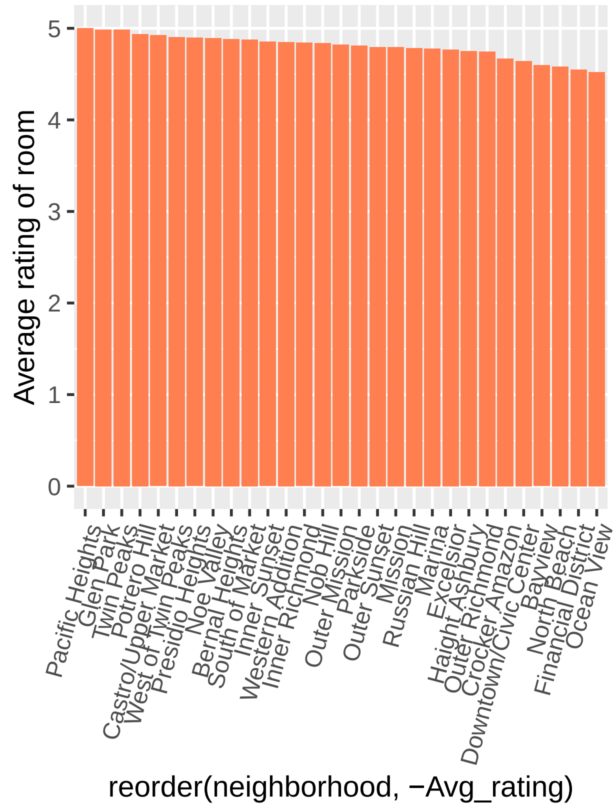
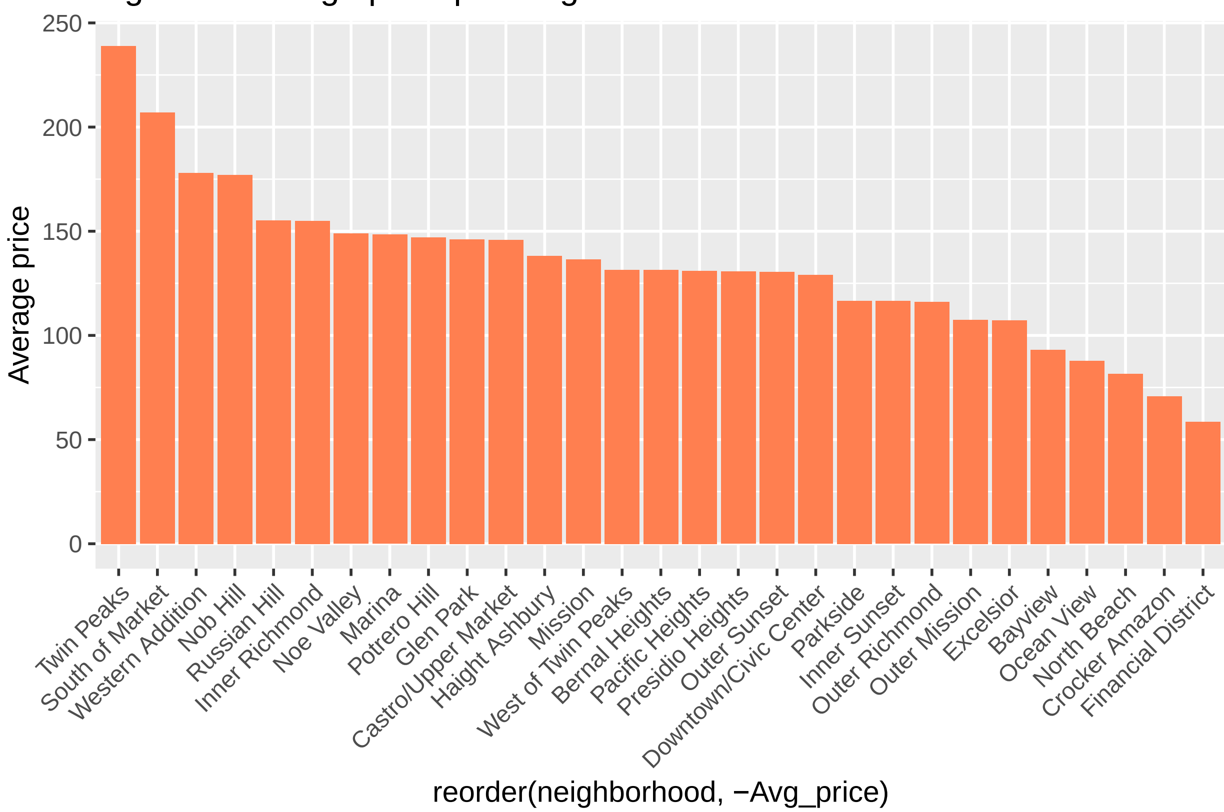
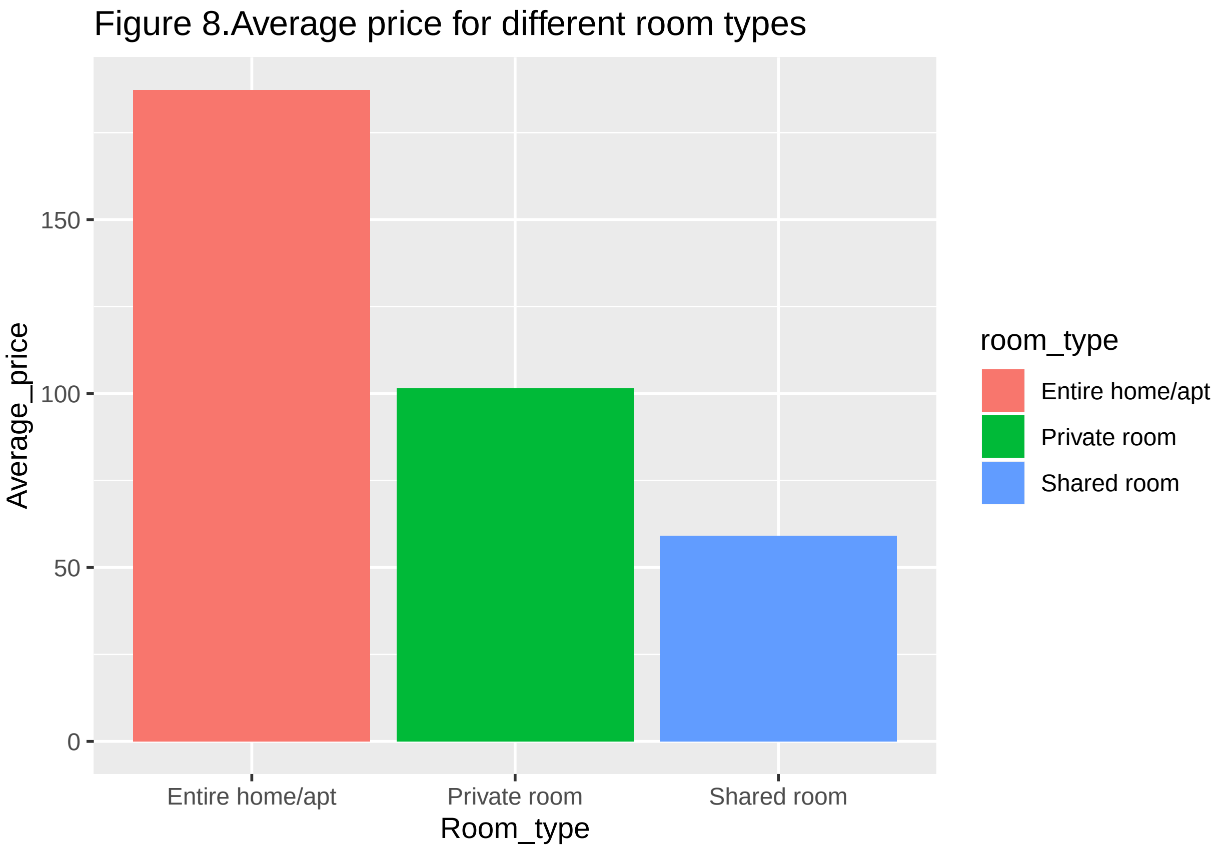


Figure 7.Average price per neighborhood



How are the number of reviews, price and rating different in each neighborhood?

From this plot, the Potrero Hill area has the highest number of review (over 200), while Excelsior has the lowest average number of reviews beside those less than 100 reviews. We can observe that the average number of review do vary a lot by neighborhood. Although there is not much diffence of rating among different neighborhood, still the neighborhood of the Airbnb room could be an influence predictor based on figure 6. We need to include this predictor in the model to see whether rating is a significant for predicting price of rooms.



What the Average price for different room types:

This result is consistent with our common sense and meaning the pricing of Airbnb in San Francisco are reasonable as the bigger the room type has the higher average price.

Testing other predictors

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:  SFAirbnb$accommodates and SFAirbnb$bedrooms
## t = 43.983, df = 3839, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5574291 0.5995021
## sample estimates:
##       cor
```

```
## 0.5788508
```

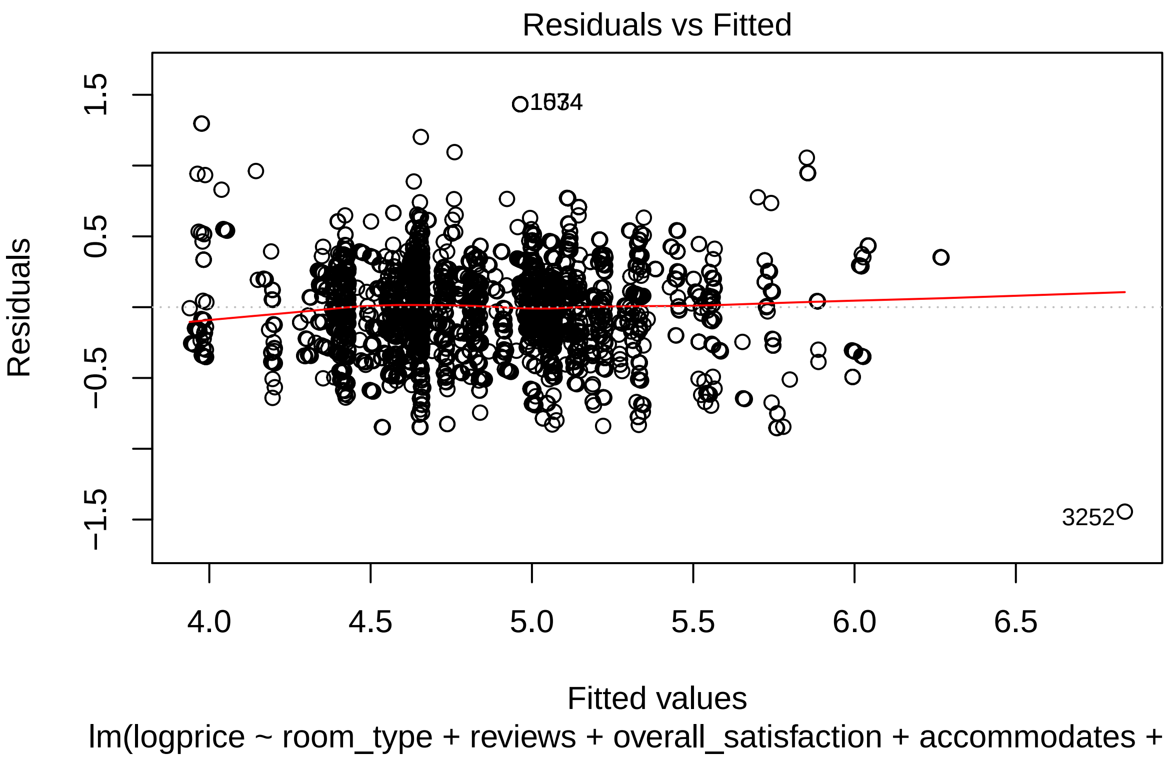
The accommodates and bedrooms could be two correlated terms in the model, because the number of bedroom is related to the number of customers served for the room. So we can do a correlation test for this two predictors. The p-value of this test is 2.2e-16. Reject the null hypothesis. So correlation between the those two variables is significant.Therefore we can add the correlation term into the model to test whether this influence term is significant.

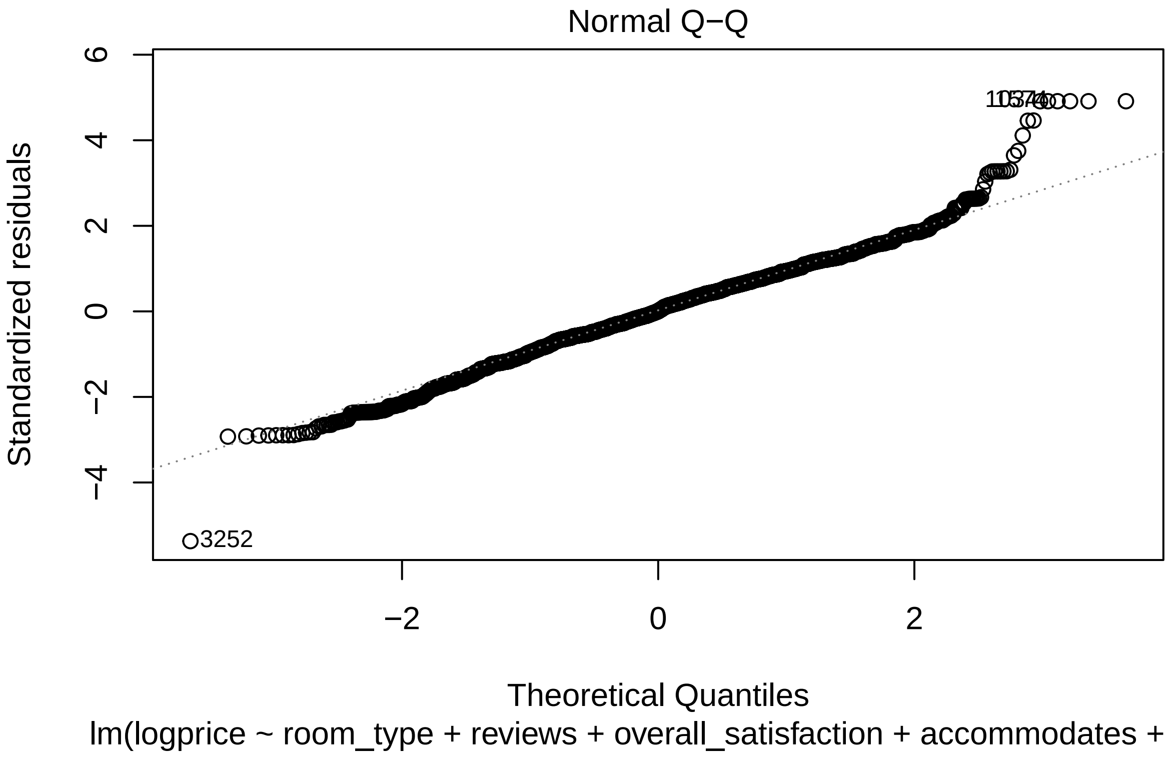
IV. Modelling:

Model1: Simple linear regression:

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

```
##
## Call:
## lm(formula = logprice ~ room_type + reviews + overall_satisfaction +
##     accommodates + accommodates * bedrooms, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44414 -0.17748  0.00088  0.19270  1.43330
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.662e+00  9.788e-02  27.196  < 2e-16 ***
## room_typePrivate room -4.109e-01  1.083e-02 -37.936  < 2e-16 ***
## room_typeShared room  -1.009e+00  3.160e-02 -31.926  < 2e-16 ***
## reviews              -2.824e-04  7.562e-05  -3.734 0.000191 ***
## overall_satisfaction  4.565e-01  1.894e-02  24.103  < 2e-16 ***
## accommodates          4.791e-02  7.338e-03   6.529 7.48e-11 ***
## bedrooms             -2.776e-03  2.270e-02  -0.122 0.902668
## accommodates:bedrooms  3.073e-02  4.411e-03   6.967 3.79e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2925 on 3833 degrees of freedom
## Multiple R-squared:  0.6137, Adjusted R-squared:  0.613
## F-statistic: 869.9 on 7 and 3833 DF,  p-value: < 2.2e-16
```





From this simple linear regression,from the results we can see most of the predictoirs are significant. The preditcor bedrooms is not significant to price,We may want to eliminate those predictors in the multilevel models. The R-square in the model is 0.6137, So the model is not well fitted. However, in the residual plot, there are some points having big residual: 1674,3252. Those might be the prices are very high that lead to big residuals.The rest of points are symmetrically distributed around the line $h = 0$. In the QQ plot, we can see in the middle most dots falls on the line. However, the data have more extreme values on the tail of the *distribution.Now let's expand simple linear model to multilevel linear model.*

Model2: Multilevel linear model with random intercept:

$$log(price) = \alpha_i + \beta_1 x_{roomtype} + \beta_2 x_{reviews} + \beta_3 x_{overallsatisfaction} + \beta_4 x_{accommodates*bedrooms} + \beta_5 x_{accommodates}$$

Model3 : Multilevel linear model with random slope:

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

Model4 : Multilevel linear model with random slope and random intercept:

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

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00411589
## (tol = 0.002, component 1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: logprice ~ room_type + overall_satisfaction + accommodates +
##     accommodates * bedrooms + (1 + overall_satisfaction | neighborhood) -
##     1
##    Data: df
##
## REML criterion at convergence: 583.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7273 -0.6419  0.0319  0.5713  6.1523
##
## Random effects:
##  Groups       Name                 Variance Std.Dev. Corr
##  neighborhood (Intercept)          1.86755  1.3666
##               overall_satisfaction 0.08130  0.2851   -0.99
##  Residual                          0.06449  0.2539
## Number of obs: 3841, groups:  neighborhood, 29
##
## Fixed effects:
##                          Estimate Std. Error t value
## room_typeEntire home/apt 2.950101   0.287960  10.245
## room_typePrivate room    2.544891   0.287803   8.842
## room_typeShared room     1.989895   0.289515   6.873
## overall_satisfaction     0.361664   0.059558   6.072
## accommodates             0.074336   0.006685  11.120
## bedrooms                 0.040599   0.020606   1.970
## accommodates:bedrooms    0.019349   0.003985   4.855
##
## Correlation of Fixed Effects:
##             rm_Eh/ rm_tPr rm_tSr ovrll_ accmmd bedrms
## rm_typPrvtr  0.999
## rm_typShrdr  0.989  0.989
## ovrll_stsfc -0.989 -0.989 -0.979
## accommodats -0.104 -0.098 -0.099  0.031
## bedrooms    -0.085 -0.091 -0.093  0.017  0.543
## accmmdts:bd  0.094  0.097  0.099 -0.022 -0.802 -0.880
```

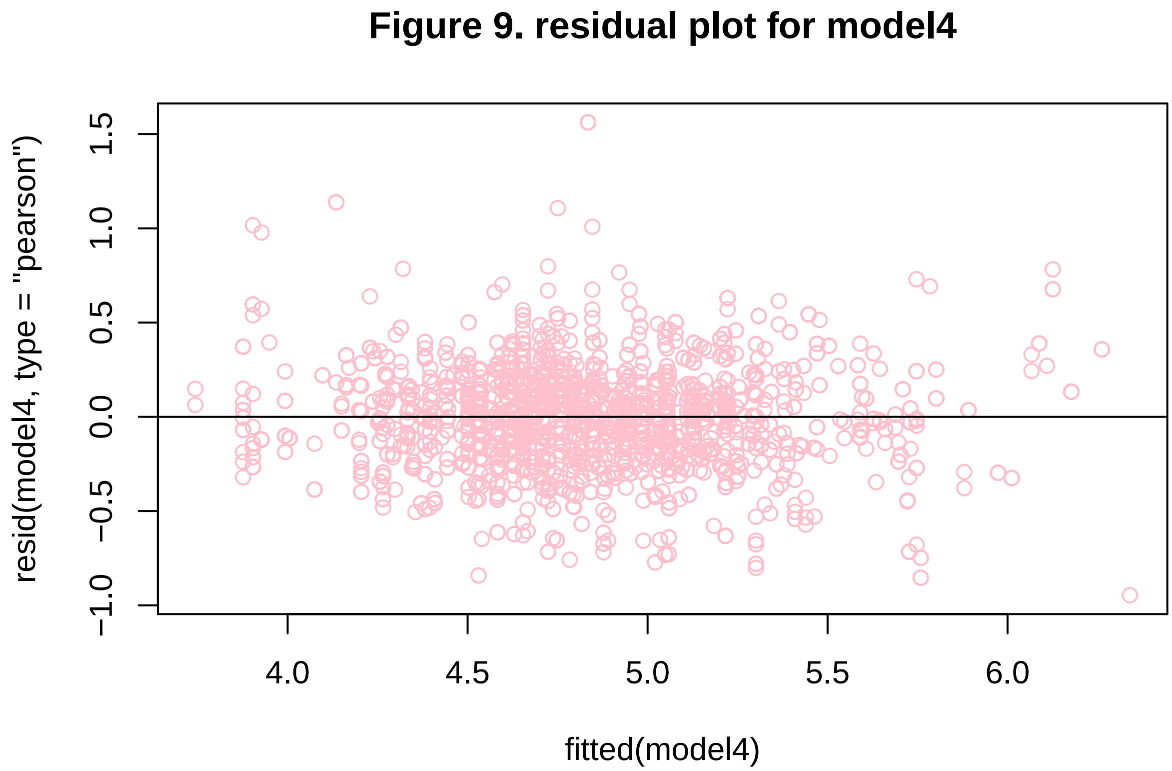
```
## convergence code: 0
## Model failed to converge with max|grad| = 0.00411589 (tol = 0.002, component 1)

## Computing profile confidence intervals ...

## Warning in nextpar(mat, cc, i, delta, lowcut, upcut): unexpected decrease
## in profile: using minstep

## Warning in FUN(X[[i]], ...): non-monotonic profile for .sig03

## Warning in confint.thpr(pp, level = level, zeta = zeta): bad spline fit
## for .sig03: falling back to linear interpolation

##                                 2.5 %      97.5 %
## .sig01                   0.9424766290  1.88414758
## .sig02                  -0.9926812541 -0.97714462
## .sig03                   0.1965420753  0.39785803
## .sigma                   0.2481679760  0.25961447
## room_typeEntire home/apt  2.3685574472  3.52199751
## room_typePrivate room     1.9637948284  3.11674585
## room_typeShared room      1.4070261997  2.56680488
## overall_satisfaction      0.2437658766  0.48248678
## accommodates              0.0612101788  0.08740651
## bedrooms                  0.0001592506  0.08100124
## accommodates:bedrooms     0.0115477115  0.02717413
```

This is the model is based on model2 and model3 with random slope as well random intercept. According to the output from this model, we can interpret coeiffients as the average log.price of entire home/apt is 2.95, the confidence interval did not cross zero which mean the predictors is significant to our result. And for the coefficent of the overall satisfaction can be understand as for each Airbnb the rating of increase by 1, the log price of this Airbnb will increase 0.36, although the range of the rating is not very big according the previous plot(figure.6), we still consider this predictors significant snice it also contain the confidence interval that not cross zero

V. Result:

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:
## model3: logprice ~ room_type + overall_satisfaction + accommodates +
## model3:     accommodates * bedrooms + (0 + overall_satisfaction | neighborhood) -
## model3:     1
## model2: log(price) ~ room_type + reviews + overall_satisfaction + accommodates +
## model2:     accommodates * bedrooms + (1 | neighborhood) - 1
## model4: logprice ~ room_type + overall_satisfaction + accommodates +
## model4:     accommodates * bedrooms + (1 + overall_satisfaction | neighborhood) -
## model4:     1
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model3  9 686.99 743.28 -334.50   668.99
## model2 10 672.14 734.68 -326.07   652.14 16.851      1  4.042e-05 ***
## model4 11 605.33 674.12 -291.67   583.33 68.809      1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
## [1] 2.816359
```

According to the output of the test, we can see that model 4 with random intercept and random slope is the best fit among three multilevel models. It has lowest deviance with 583.33. The second plot is a residual plot form model 4. As we can see the plot the points are symmetrically distributed around the line h = 0. The neighborhood with the maximum intercept is Parkside with intercept 2.816. From the analysis above we found the best model among is the Model4. From the model,we can found that the most significant predictor is the factor of room type. The second significant predictors is the district of neighborhood. Also, higher overall satisfaction will lead to a higher price. The ideas of building these model is to predicting models by different levels of neighborhood. The model in each level will have a unique intercept and a unique slope for the predictor overall satisfaction, in this way it will lead the last model minimize the deviance comparing to the previous two multilevel linear model.

VI. Discussion:

The result terms out to be not very surprirsing, although I though the reviews would so how affect the price of Airbnb in San Francisco. The reality here is that the factor of room type will be determine the price the most, which is reasonable because from what my knowledge, San Francisco is one of the most expensive living environment in the U.S. especially the bay area has the highest housing price. Therefore it is a place the every inch is like the price as gold. The room type of Airbnb is more like the different room size for the hotel. The entire home/apt tends to serve more people and have bigger space 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 based on the living environment in San Francisco there are many tech company and university in the city, as well as the neighborhood around union square, twin peak is close to downtown area and there are many sight-seeing spots for tourists. We can conclude that our findings are reasonable.

However, this analysis is definitely not perfect. There are only 4 predictors in the multilevel model. Those predictors are the most significant variables we found in the dataset. This is just a very basic analysis of the Airbnb data, there are many other factors that can take in to account, such as the academic institution in the neighborhood or the transportation condition, and also the geographic difference in countries will be interesting to explore in the future. That maybe the future direction of this project.

VII.Reference

http://tomslee.net

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

https://www.airbnb.com/

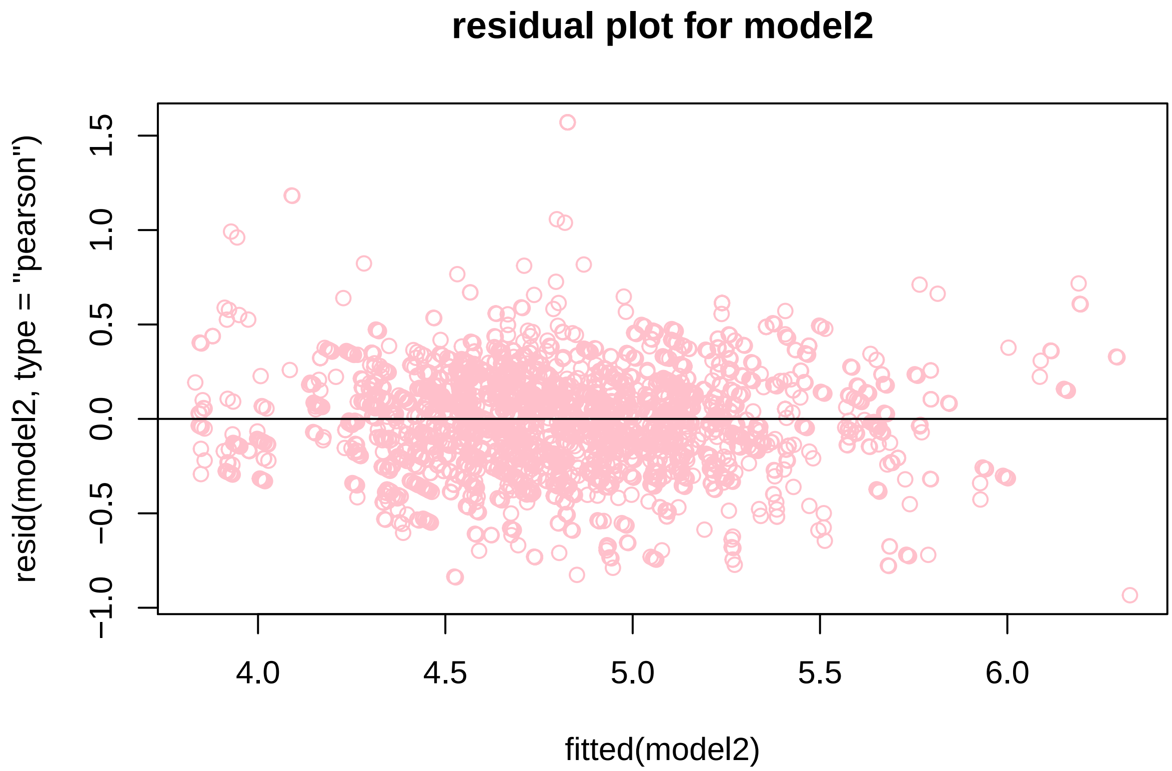
VIII. Appendix

Output and residual plot for model 2

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log(price) ~ room_type + reviews + overall_satisfaction + accommodates +
##     accommodates * bedrooms + (1 | neighborhood) - 1
##    Data: df
##
## REML criterion at convergence: 652.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6255 -0.6322  0.0431  0.5941  6.0988
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  neighborhood (Intercept) 0.03487  0.1867
##  Residual                 0.06632  0.2575
## Number of obs: 3841, groups:  neighborhood, 29
##
## Fixed effects:
##                            Estimate Std. Error t value
## room_typeEntire home/apt  3.017e+00  9.728e-02  31.008
## room_typePrivate room     2.613e+00  9.707e-02  26.924
## room_typeShared room      2.021e+00  1.049e-01  19.268
## reviews                  -3.974e-04  6.911e-05  -5.750
## overall_satisfaction      3.576e-01  1.777e-02  20.124
## accommodates              7.496e-02  6.671e-03  11.237
## bedrooms                  4.696e-02  2.051e-02   2.289
## accommodates:bedrooms     1.906e-02  3.970e-03   4.802
##
## Correlation of Fixed Effects:
##             rm_Eh/ rm_tPr rm_tSr reviws ovrll_ accmmd bedrms
## rm_typPrvtr  0.995
## rm_typShrdr  0.915  0.915
## reviews     -0.064 -0.070 -0.053
## ovrll_stsfc -0.891 -0.892 -0.826 -0.054
## accommodats -0.263 -0.244 -0.237  0.013  0.051
## bedrooms    -0.265 -0.281 -0.271 -0.028  0.075  0.541
## accmmdts:bd  0.257  0.267  0.258  0.024 -0.056 -0.800 -0.879
```

```
## Computing profile confidence intervals ...
```

```
##                                  2.5 %        97.5 %
## .sig01                     0.1429034414  0.2444516626
## .sigma                     0.2516234569  0.2631786577
## room_typeEntire home/apt   2.8260397676  3.2066884226
## room_typePrivate room      2.4232281236  2.8030337606
## room_typeShared room       1.8157086836  2.2263119360
## reviews                   -0.0005326705 -0.0002619241
## overall_satisfaction       0.3228740310  0.3925677058
## accommodates               0.0618700559  0.0880053494
## bedrooms                   0.0067551356  0.0871202090
## accommodates:bedrooms      0.0112923800  0.0268447733
```

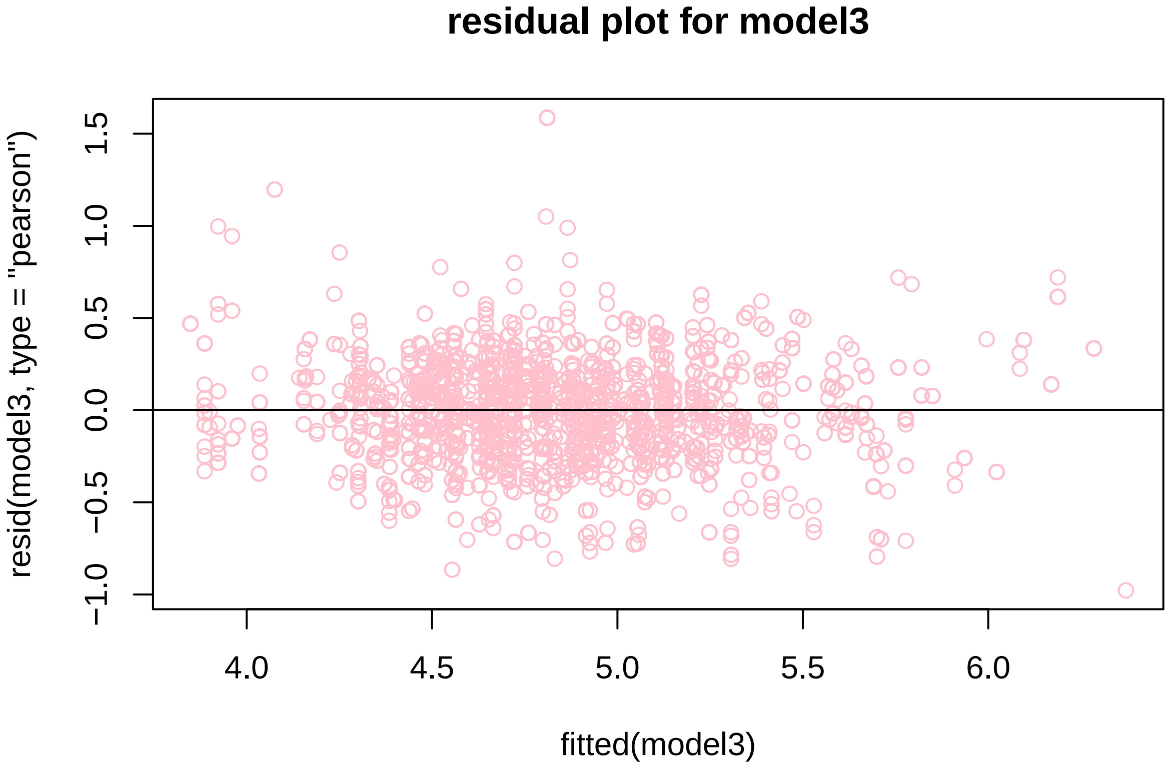


In our second model, we gets rid of non-significant terms bedrooms.The factor of room type plays the most important part in the model. The coefficent of reviews is zero so we don't need to keep it for the next model. Besides, the overall satisfaction is the second influential term in this model other than the room type.

Output and residual plot for model 3

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: logprice ~ room_type + overall_satisfaction + accommodates +
##     accommodates * bedrooms + (0 + overall_satisfaction | neighborhood) -
##     1
##    Data: df
##
## REML criterion at convergence: 669
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7803 -0.6196  0.0436  0.5948  6.1335
##
```

```
## Random effects:
##  Groups       Name                 Variance Std.Dev.
##  neighborhood overall_satisfaction 0.001515 0.03892
##  Residual                          0.066898 0.25865
## Number of obs: 3841, groups:  neighborhood, 29
##
## Fixed effects:
##                          Estimate Std. Error t value
## room_typeEntire home/apt 2.988264   0.090711  32.943
## room_typePrivate room    2.579167   0.090405  28.529
## room_typeShared room     1.984792   0.099459  19.956
## overall_satisfaction     0.349509   0.019229  18.176
## accommodates             0.075942   0.006693  11.346
## bedrooms                 0.049136   0.020612   2.384
## accommodates:bedrooms    0.018831   0.003986   4.724
##
## Correlation of Fixed Effects:
##             rm_Eh/ rm_tPr rm_tSr ovrll_ accmmd bedrms
## rm_typPrvtr  0.994
## rm_typShrdr  0.906  0.907
## ovrll_stsfc -0.890 -0.892 -0.819
## accommodats -0.276 -0.257 -0.248  0.043
## bedrooms    -0.283 -0.300 -0.285  0.064  0.541
## accmmdts:bd  0.273  0.284  0.271 -0.046 -0.800 -0.879

## Computing profile confidence intervals ...

##                                2.5 %     97.5 %
## .sig01                   0.029772925 0.05095609
## .sigma                   0.252750959 0.26435799
## room_typeEntire home/apt 2.809961398 3.16559622
## room_typePrivate room    2.401449234 2.75589727
## room_typeShared room     1.789302925 2.17917757
## overall_satisfaction     0.311936110 0.38720061
## accommodates             0.062808603 0.08903638
## bedrooms                 0.008721213 0.08948098
## accommodates:bedrooms    0.011030174 0.02664824
```


residual plot for model3
resid(model3, type = "pearson")
−1.0
−0.5
0.0
0.5
1.0
1.5
4.0
4.5
5.0
5.5
6.0
fitted(model3)