Airbnb Midterm Project

Shangchen Han

12/3/2019

I. Abstract:

Travel becomes more popular than before, because of releasing pressure during daily life. But traveler is likely to consider the price and quality of the accommodation as their first two concern. The company Airbnb provides information about the accommodations and it is easy for travelers to choose positions by themselves. For this project, its focus on predicting the price of accommodations based on different variables, such as room types, number of reviews, neighborhoods, etc. To be specific, in order to analyze, the whole project has two parts: EDA and modeling. For EDA, it is about finding the relationships between variables, and it also shows the changes in one specific variable based on different years or different conditions. Thus, the trend of changes is visible. For the modeling part, there are some regression models, and by using these models to find the relationships between price and other variables. By compared AIC and deviance of the model, the best model will be chosen. And then travelers could predict the price by themselves.

II. Introduction:

2.1 Background:

Airbnb is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. After founded, Airbnb became famous and popular because of the traveling trend. For travelers, they tend to consider price, satisfaction, review as their concerns. Thus, I would like to do analyses such as EDA and modeling to find out the relationships between price and other factors. It will helpful for travelers to predict the price by themselves.

2.2 Data Sources:

The datasets I used for performing the analysis,'Airbnb Data Collection: Get the Data', which is particular in Vancouver, is obtained from the Tomslee website. But, actually, the dataset is not integrated, which means it only includes years from 2015 to 2017. And for the data in 2015 and 2017 have 4 months, so I cannot compare these 3 years directly.

2.3 Previous Work: Data combining and cleaning

The whole data has 20 files, so after imported I need to combine them together and choose available varibles, and then omit NAs.

```
dt <- bind_rows(dt1,dt2,dt3,dt4,dt5,dt6,dt7,dt8,dt9,dt10,dt11,dt12,dt13,dt14,dt15,dt16,dt17,dt1
8,dt19,dt20)
dt <- dt[,c(1:3,5:10,12:14)]
Van_dt <- na.omit(dt)
```

I would not like to choose that price and review equal to zero, because these data are not representative. And then select the year and months, which are overlapped, so that I could compare the factorc in these date, and see whether there are some changes or not. They are April, Augest, November, December, respectively.

```r
Van_dt <- Van_dt %>%
  filter(Van_dt$price>0) %>%
  filter(Van_dt$reviews>0)
```

```r
## Try to split last_modified data in Van_dt
Van_dt <- separate(Van_dt,last_modified,into = c("date","hour"),sep = " ")
Van_dt <- separate(Van_dt,date,into = c("year","month","day"))

Van_dt1 <- Van_dt
Van_dt1 <- Van_dt1 %>% filter(year=="2015"| year=="2016")
Van_dt1 <- Van_dt1 %>% filter(month=="04"| month=="08"| month=="11"| month=="12")
```

```
##   room_type                     neighborhood      reviews
##  Length:67092       Downtown          :14979   Min.   :  1.00
##  Class :character   West End          : 8016   1st Qu.:  4.00
##  Mode  :character   Kitsilano         : 7552   Median : 10.00
##                     Mount Pleasant    : 6140   Mean   : 21.14
##                     Grandview-Woodland: 4310   3rd Qu.: 26.00
##                     Downtown Eastside : 3629   Max.   :489.00
##                     (Other)           :22466
##  overall_satisfaction  accommodates       bedrooms         price
##  Min.   :0.000        Min.   : 1.000   Min.   :0.000   Min.   :  10.0
##  1st Qu.:4.500        1st Qu.: 2.000   1st Qu.:1.000   1st Qu.:  58.0
##  Median :5.000        Median : 2.000   Median :1.000   Median :  79.0
##  Mean   :4.407        Mean   : 2.991   Mean   :1.238   Mean   : 101.5
##  3rd Qu.:5.000        3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 116.0
##  Max.   :5.000        Max.   :16.000   Max.   :9.000   Max.   :8033.0
##
```

They are some varibles in the data, which might be included in my model. And from this summary, we could see the features of these variables.

III. EDA part:

Fig.1 Distribution of neighborhood

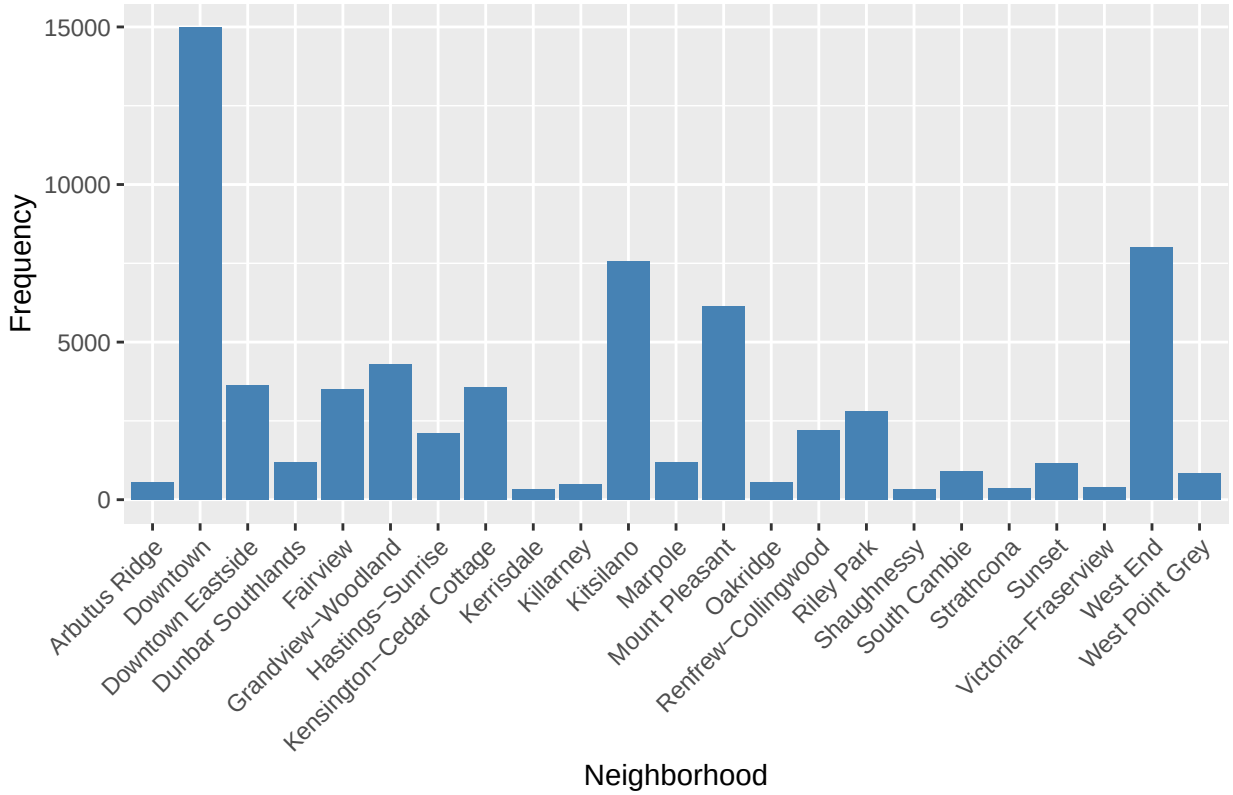
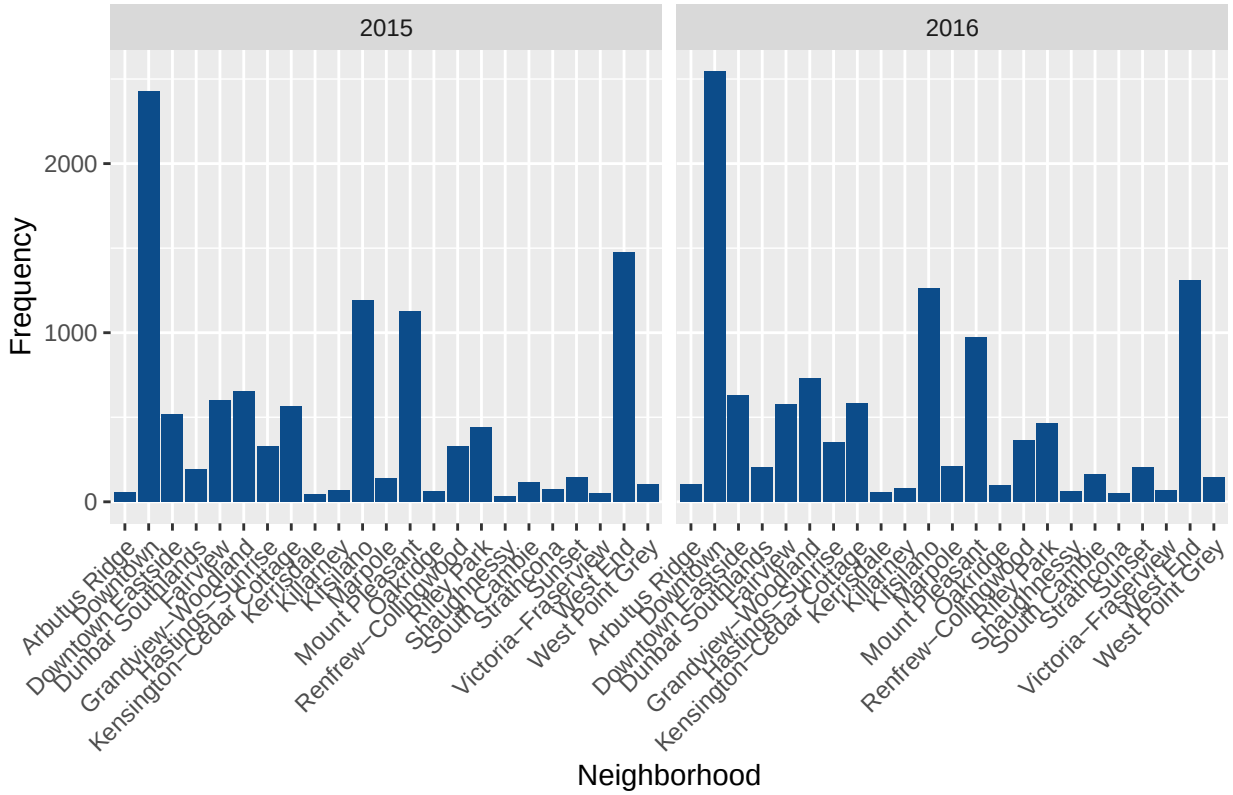


Fig.2 Distribution of neighborhood in 2015 & 2016 among 4 months



From Fig.1, as we can see the most of the neighborhoods are in the Downtown region, and then West End and Kitsilano are also popular through the whole dataset. This was the same trend in 2015 and 2016, although there were slight differences between these two years among 4 months. Thus, we can guess that in the future, there will not change too much.

Fig.3 Distribution of reviews

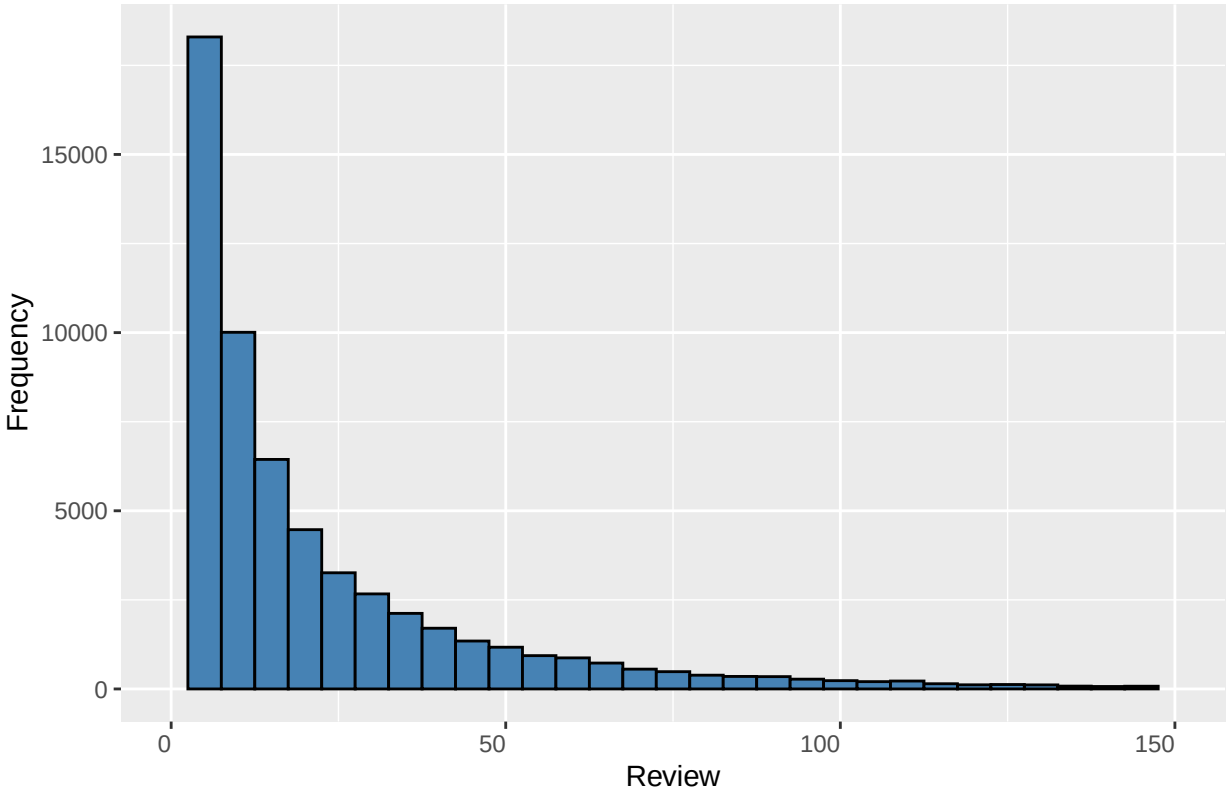
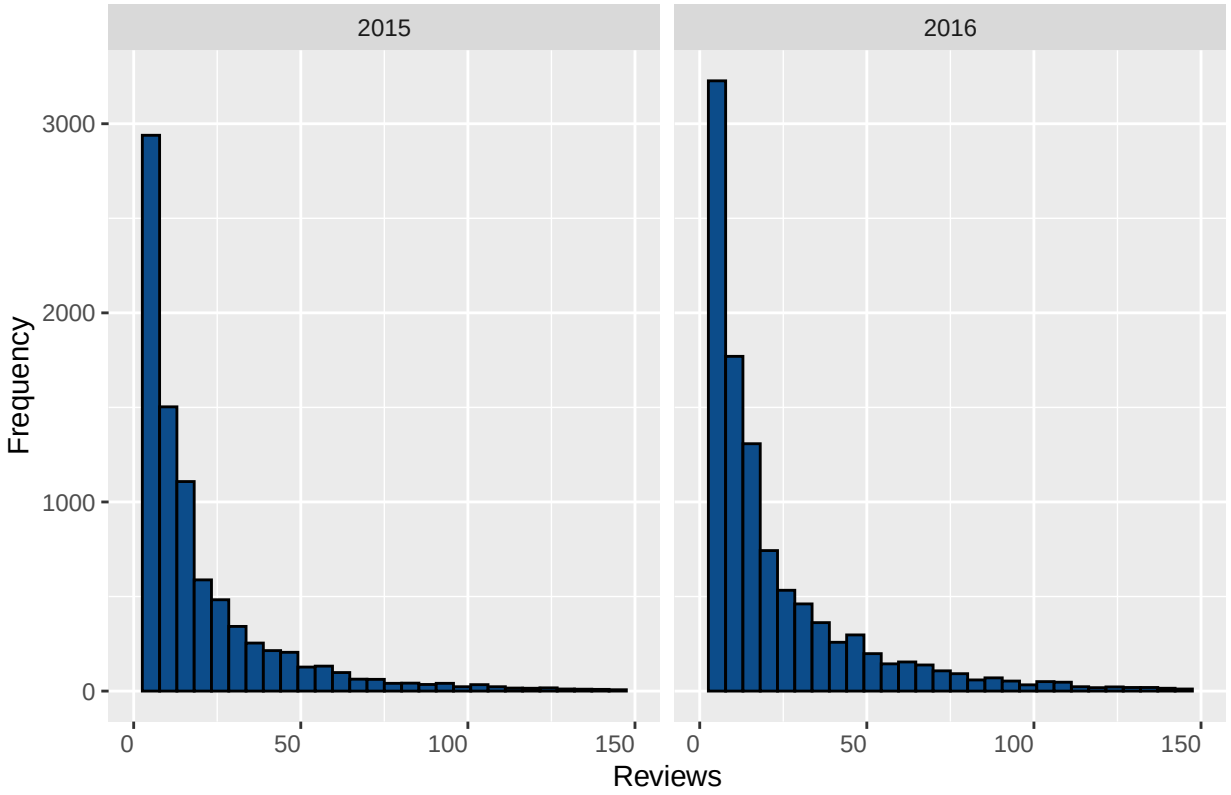


Fig.4 Distribution of reviews in 2015 & 2016 among 4 months



From the plots of the distribution of reviews, as can be seen, most of the Airbnb hosts have no reviews, and this trend does not change, because in 2015 and 2016, they are in the same shape of the distribution.

Fig.5 Distribution of overall satisfaction

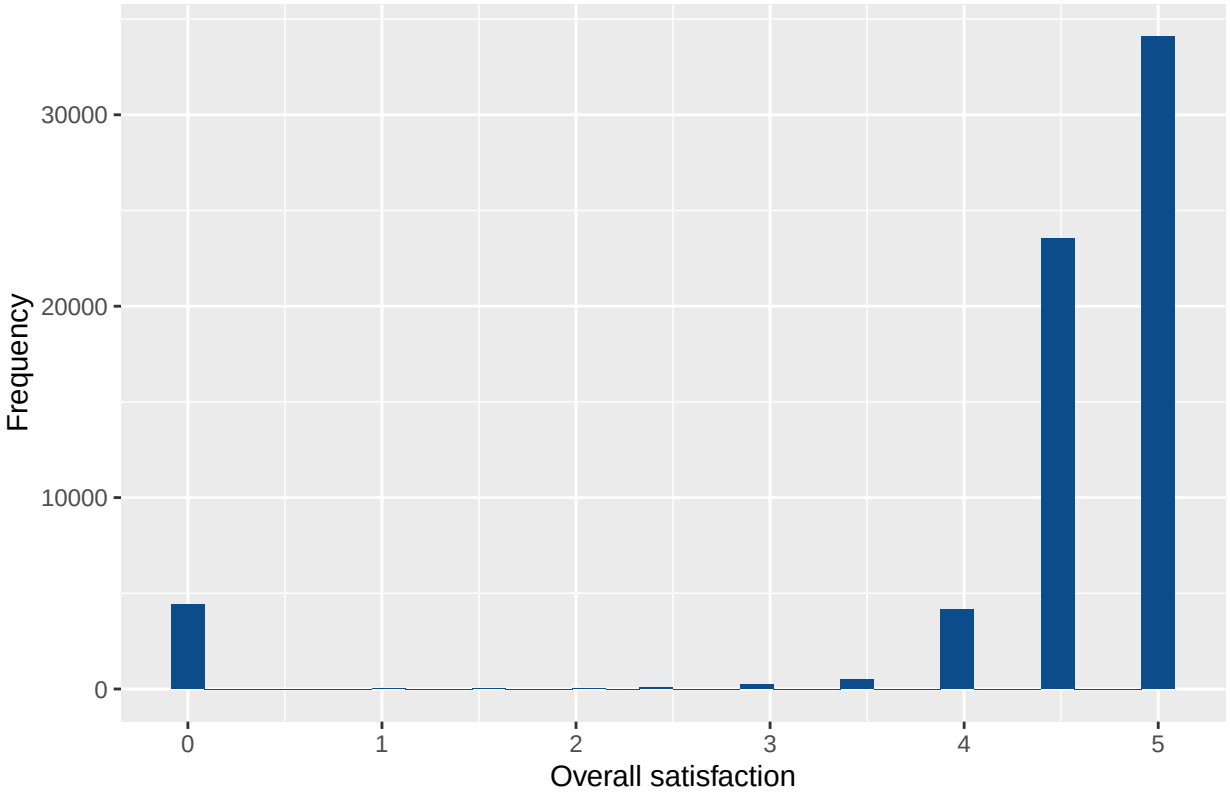
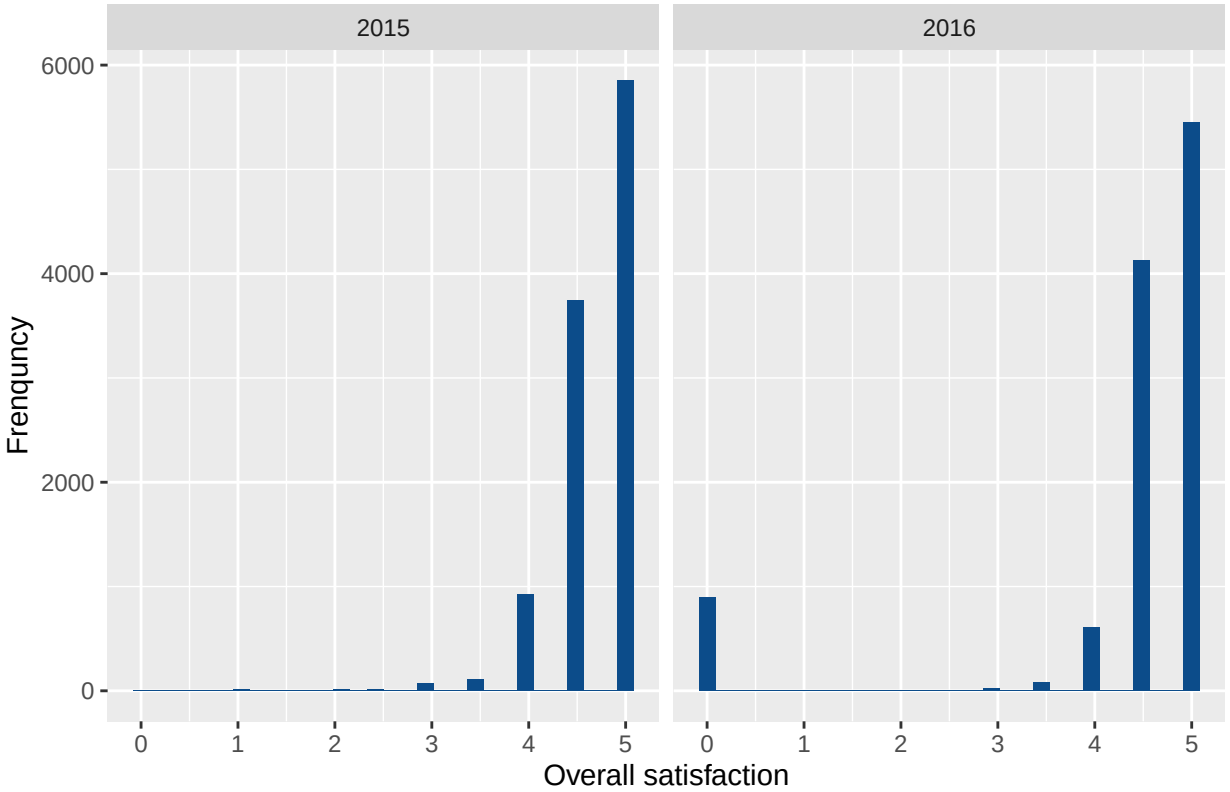


Fig.6 Distribution of overall satisfaction in 2015 & 2016 among 4 months



Most of the overall satisfaction rating is 4.5 and 5 points. To be specific, in 2015 and 2016 among 4 months,

there are not too many changes. In 2016, there are more 0 points of rating Airbnb accommodations than in 2015.

Fig.7 Satisfaction in different room types

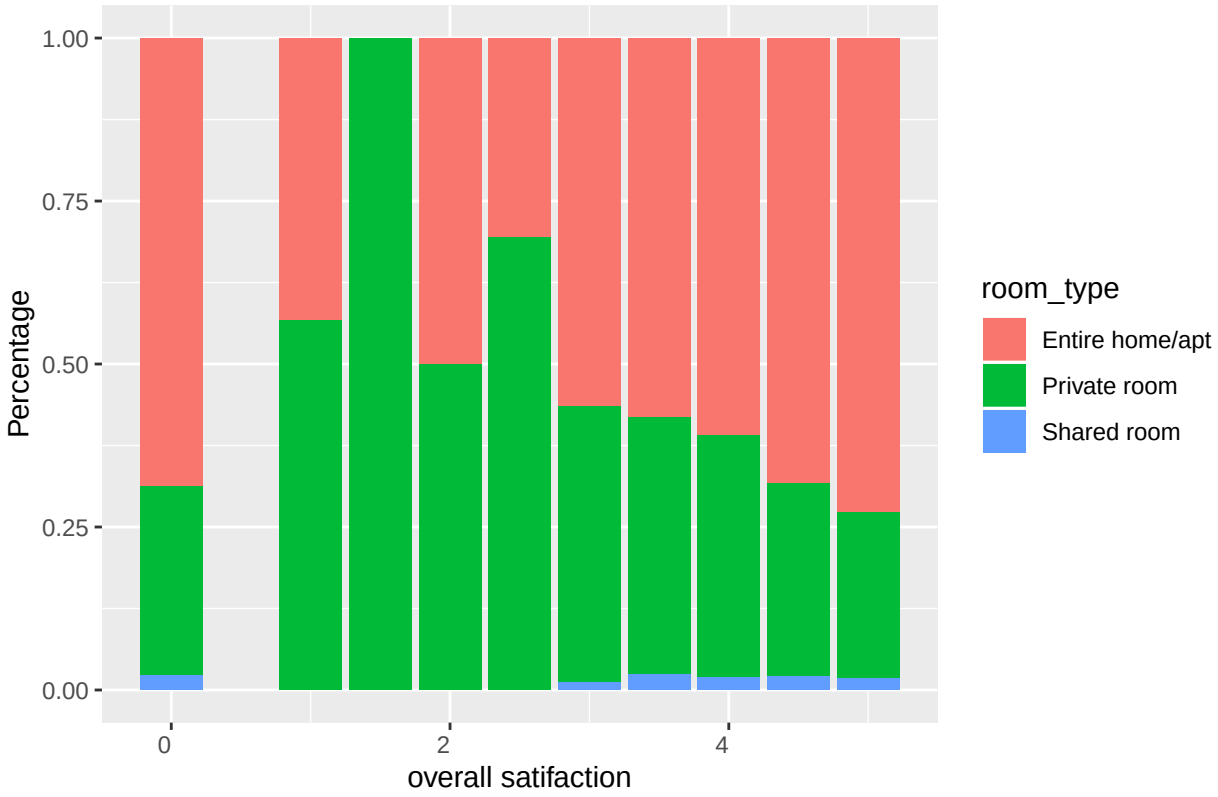
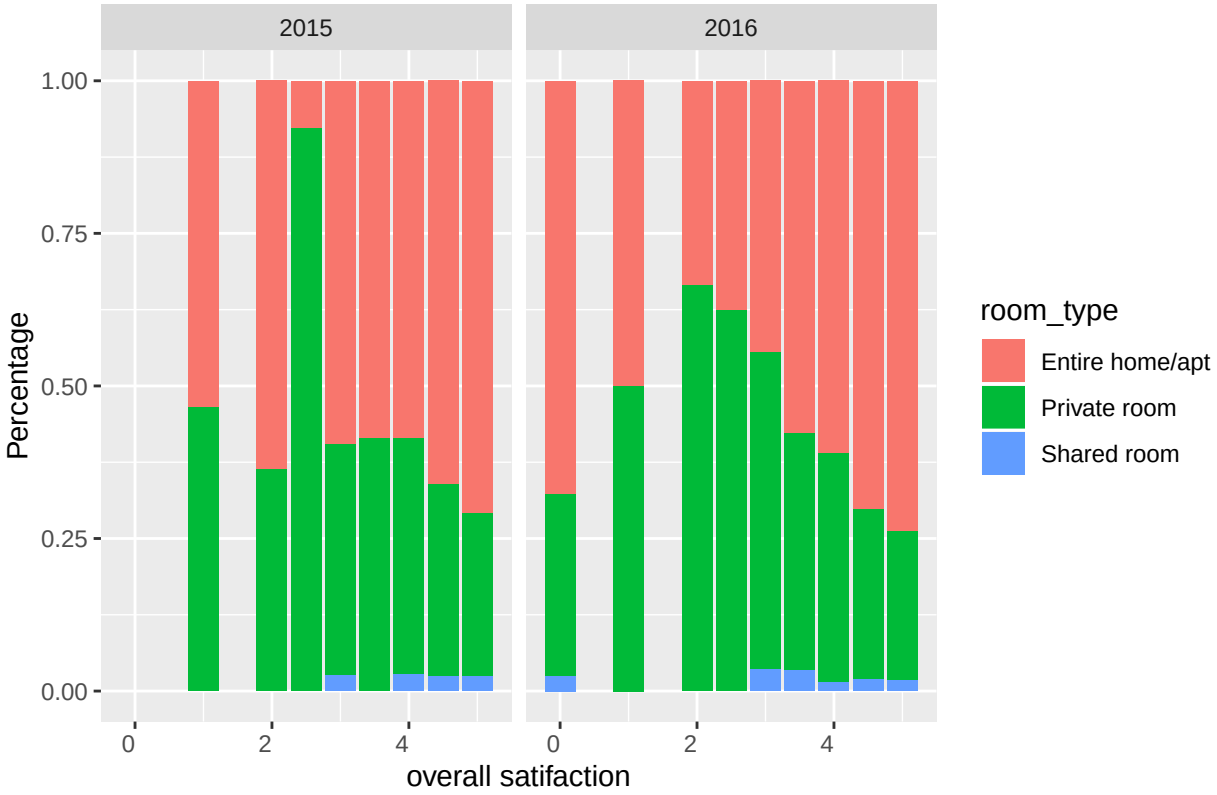


Fig.8 Satisfaction in different room types



From Fig.7 we could see the satisfaction for different room types. At first sight, the entire home/apt has the largest percentage through most of the range of ratings. But in 1.5 point field, it is only for private room. From Fig.8, in 2015, it has the shape as Fig.7. In contrast, the entire home/apt and private room are complementary, and the rate of entire home/apt increases from 2 points to 5 points.

Fig.9 Room type survey

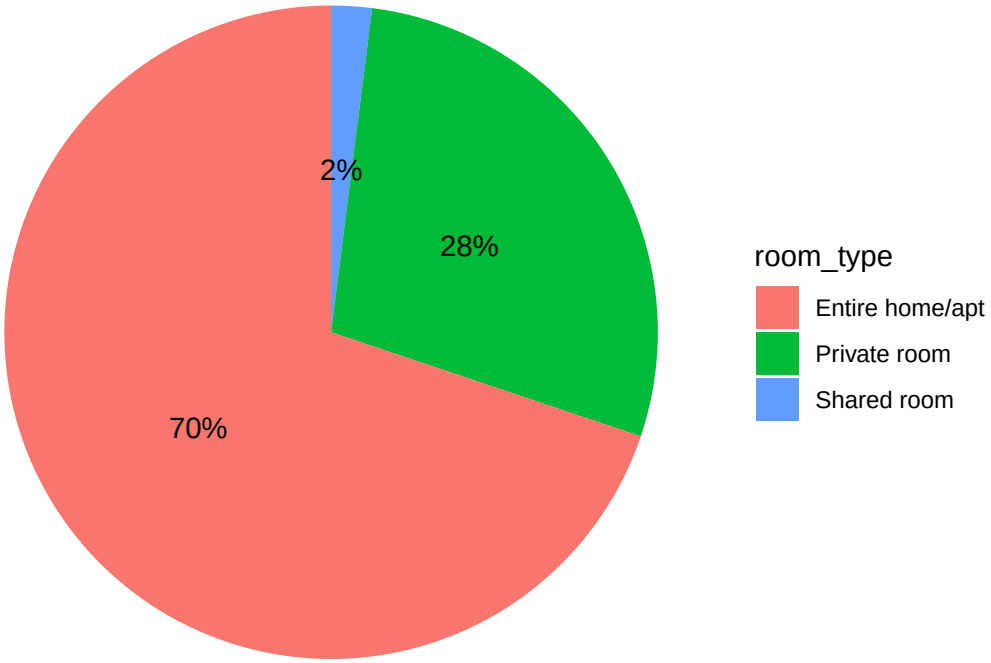


Fig.9 shows the percentage of room type, the entire home/apt occupy around 70% of whole dataset, and private room is in the second ranking around 28%. And shared room has the smallest percent(2%).

Fig.10 Distribution of accommodates

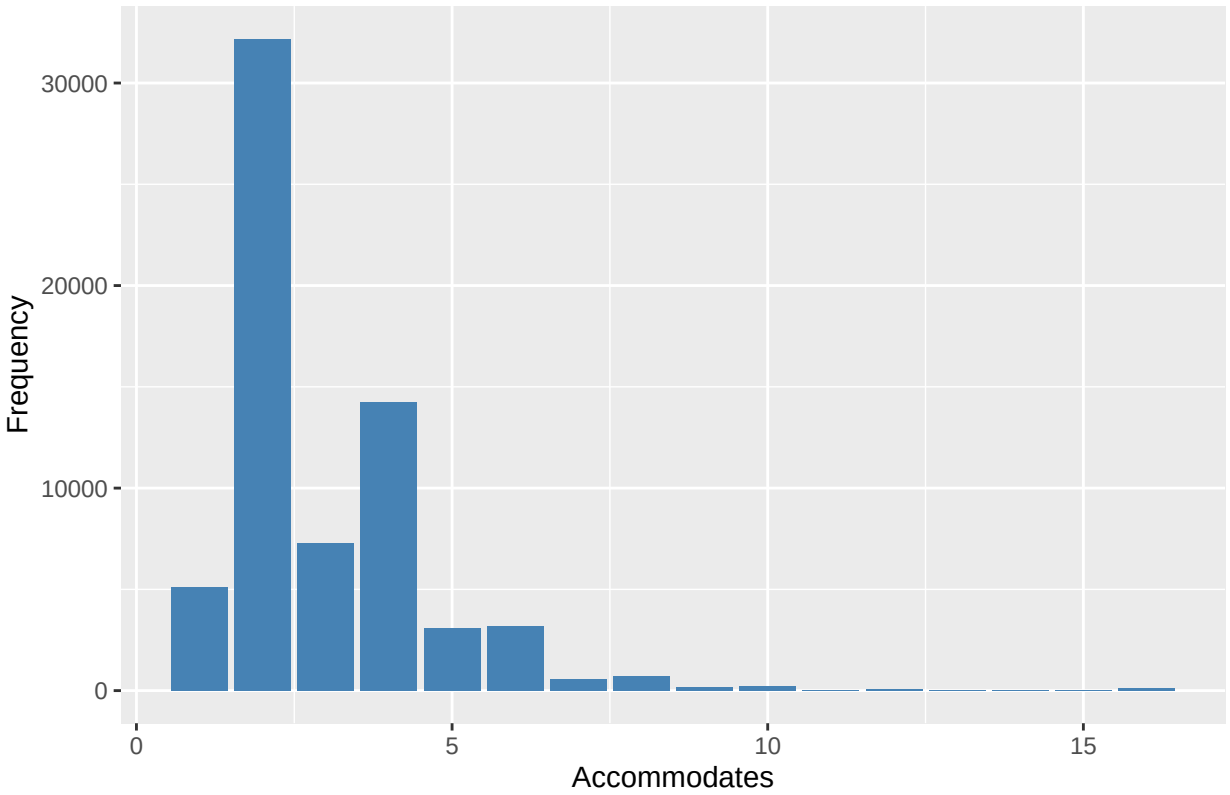
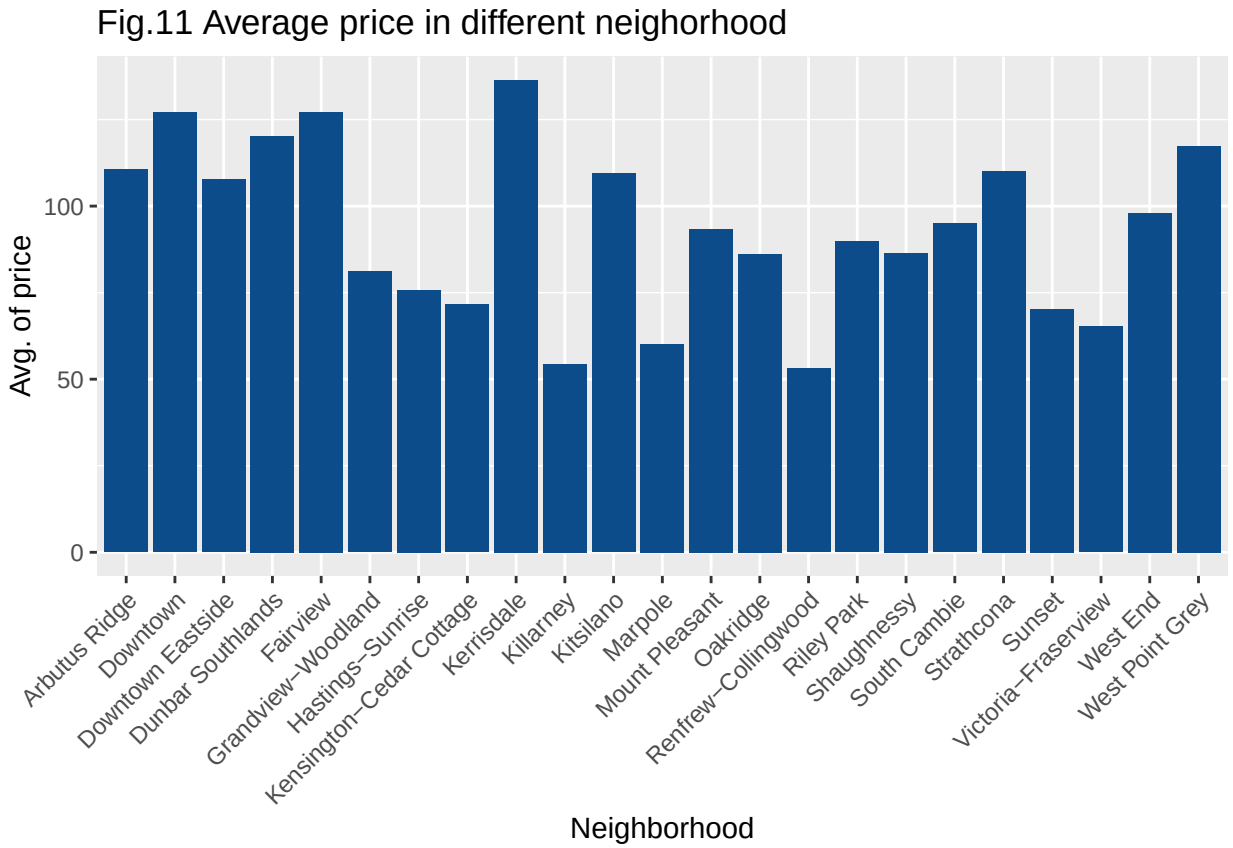
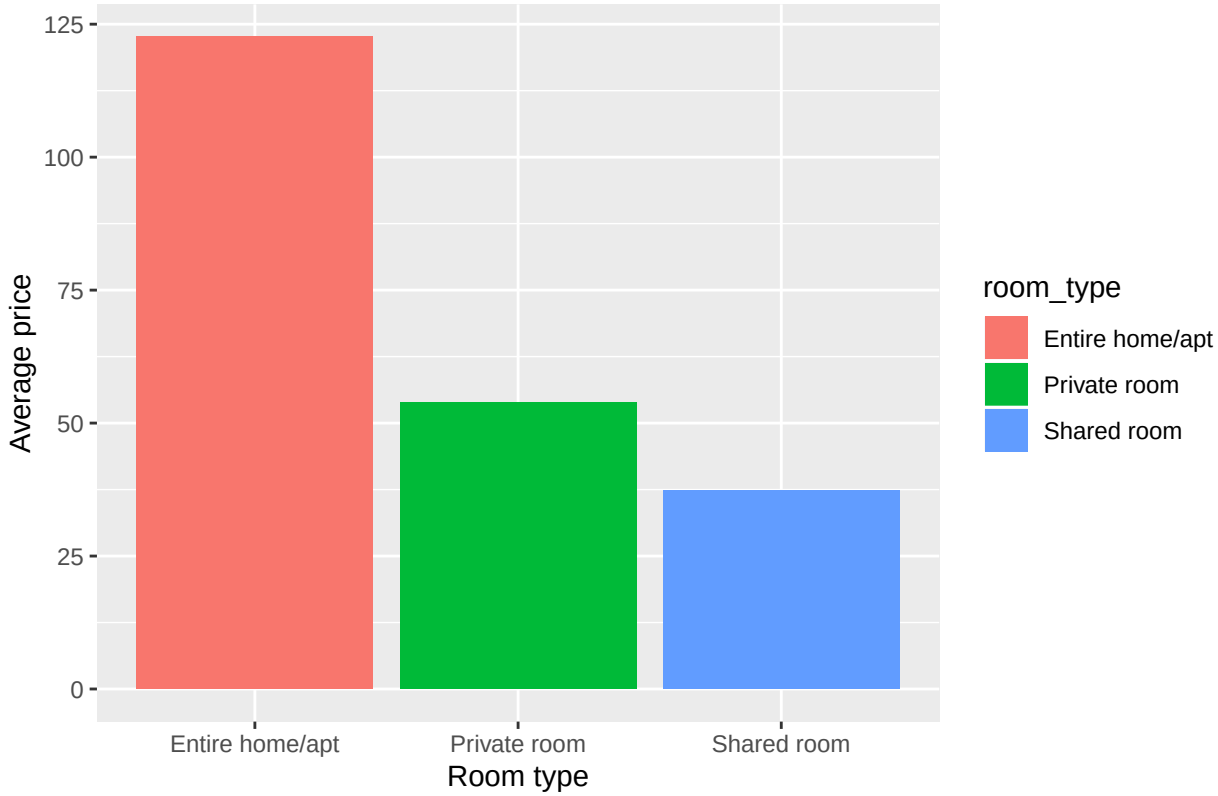


Fig.10 shows the distribution of accommodates. The most common range of accommodates is from 1 to 6, and the top three accommodates are 2,4,3.

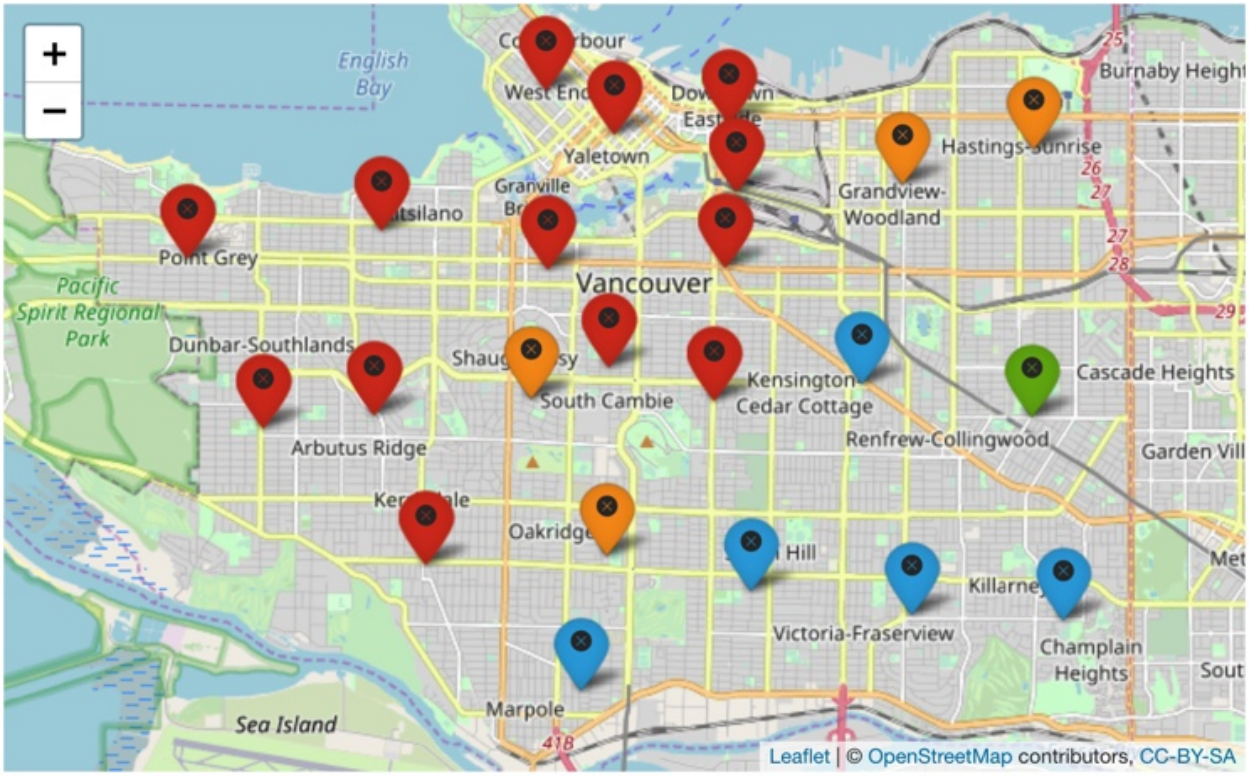


From Fig.11 we could see that the most expensive Airbnb position is in Kerrisdale around 175 per night, and Downtown, Fairview are also expensive. In contrast, Killarney and Renfrew-Collingwood have the cheapest position for living around 50 dollars.

Fig.12 Average price among different room types



From Fig.12, the average price for different room types could be seen. The entire home/apt has the highest price of around 125 per night, and the price of the private room is half of the entire home/apt. The price of the shared room is lowest at about 37.5 per night.



This map shows the different prices in different districts. I used median, mean, 3rd quartile to classify the regions. From green, blue, orange to red represents the average price is about 54, 72, 89 and above 89, respectively. Relatively costly positions are in the top left.

Fig.13 Distribution of room price

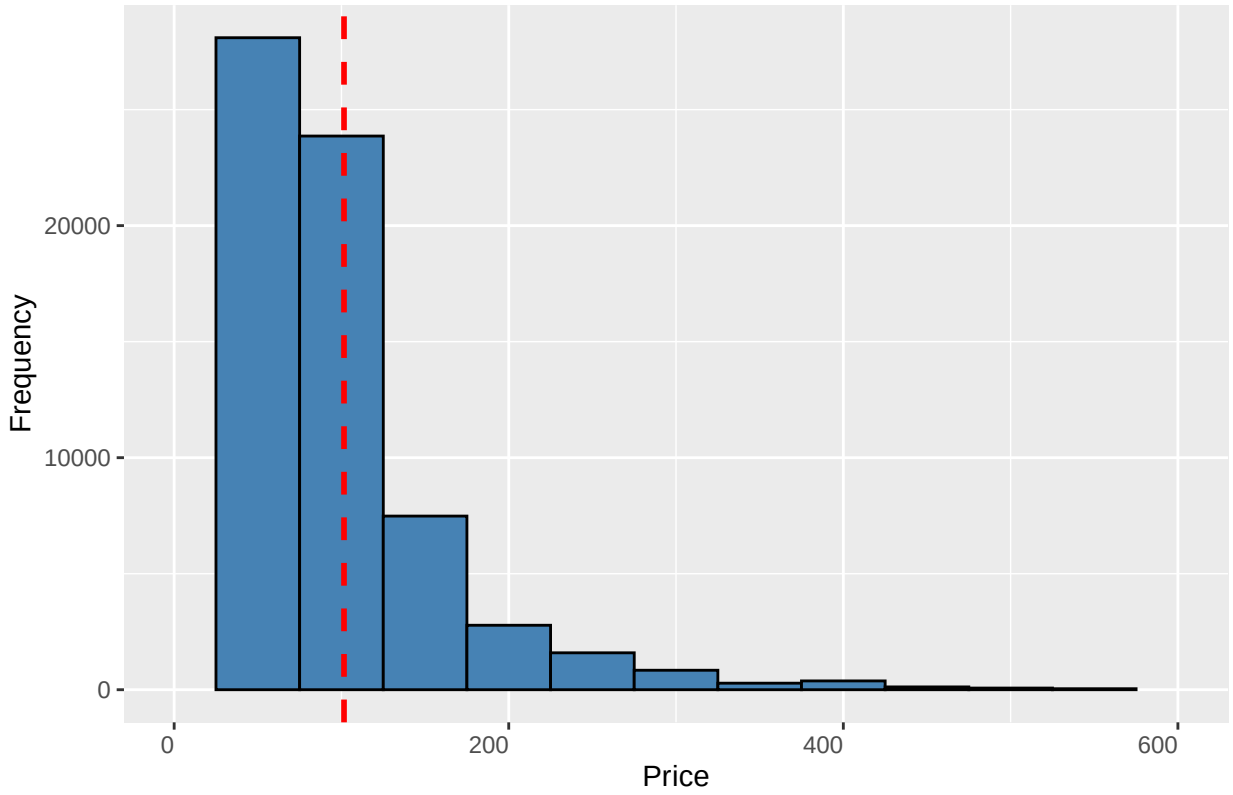
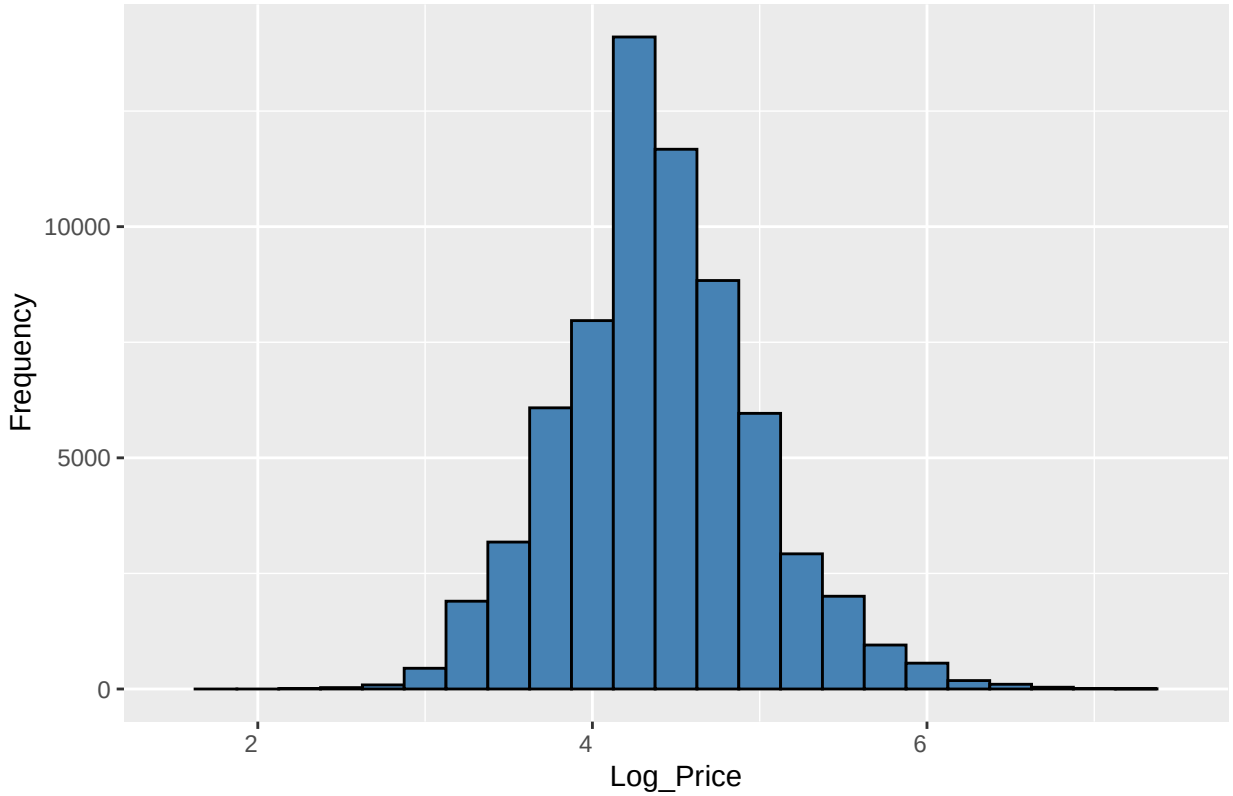


Fig.14 Distribution of log_price

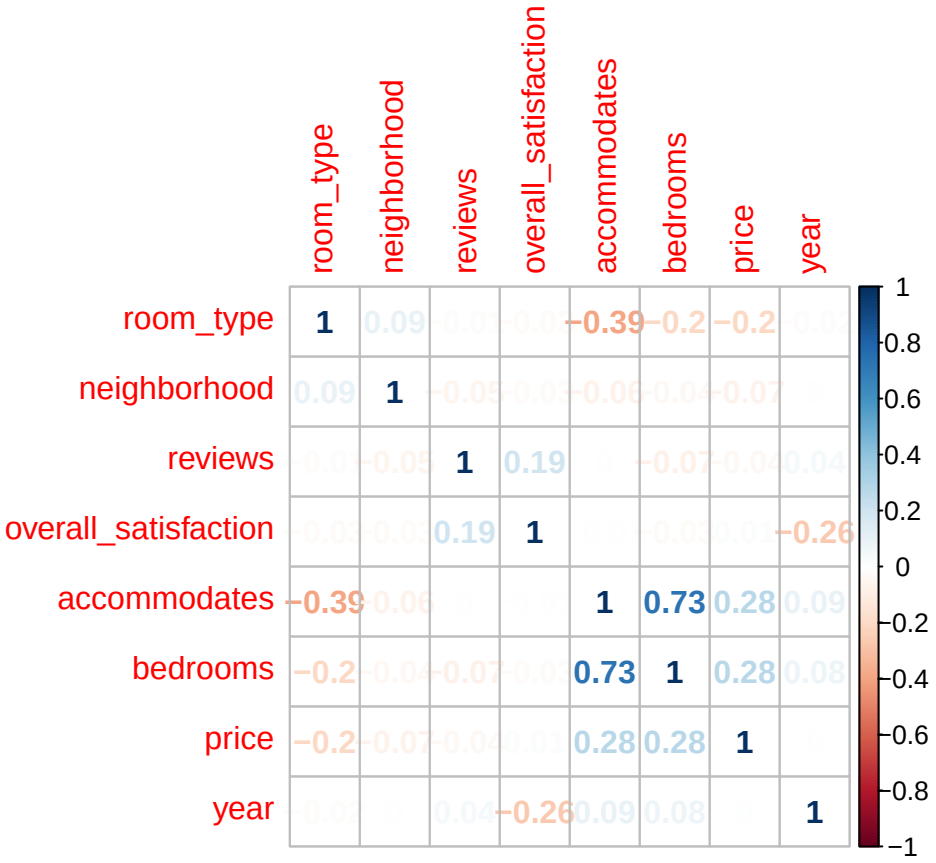


From Fig.13, the distribution of price does not follow the normal distribution. Thus, it is better to transform the data. After log-transformation, the shape of distribution looks better than before, it roughly follows the normal distribution.

IV.Modelling:

Model choice for linear regression

```
##                      room_type neighborhood reviews overall_satisfaction
## room_type                 1.00         0.09   -0.01                -0.03
## neighborhood              0.09         1.00   -0.05                -0.03
## reviews                  -0.01        -0.05    1.00                 0.19
## overall_satisfaction     -0.03        -0.03    0.19                 1.00
## accommodates             -0.39        -0.06    0.00                -0.01
## bedrooms                 -0.20        -0.04   -0.07                -0.03
##                      accommodates bedrooms price  year
## room_type                   -0.39    -0.20 -0.20 -0.02
## neighborhood                -0.06    -0.04 -0.07  0.00
## reviews                      0.00    -0.07 -0.04  0.04
## overall_satisfaction        -0.01    -0.03  0.01 -0.26
## accommodates                 1.00     0.73  0.28  0.09
## bedrooms                     0.73     1.00  0.28  0.08
```



Since accommodate could decide how many people could live in, therefore I want to check whether accommodates and bedrooms have a high correlation. From the output of the correlation plot, we could see the correlation coefficient between these two is 0.73. But actually, it can not determine whether to use accommodates to replace bedrooms or not, vice versa.

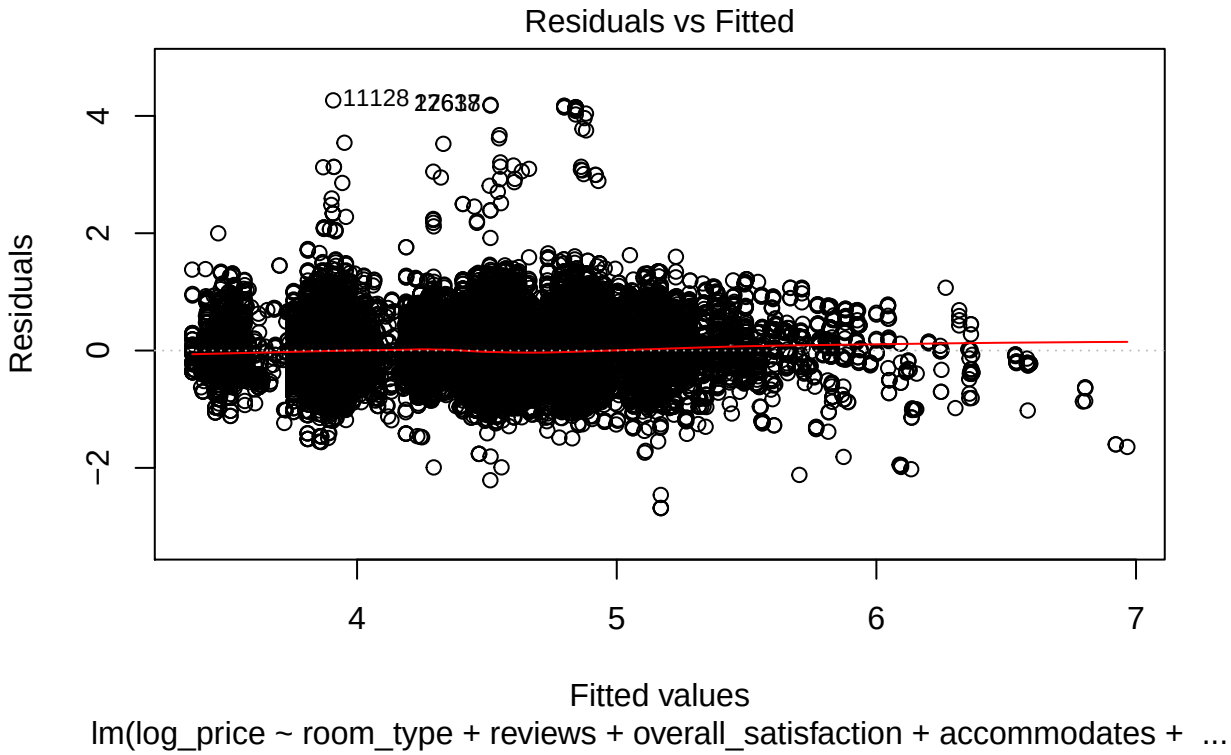
```
## Analysis of Variance Table
##
## Model 1: log_price ~ room_type + reviews + overall_satisfaction + accommodates +
##     year
## Model 2: log_price ~ room_type + reviews + overall_satisfaction + accommodates +
##     bedrooms + year
## Model 3: log_price ~ room_type + reviews + overall_satisfaction + bedrooms +
##     year
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  67084 11866
## 2  67083 11070  1    796.13 4824.6 < 2.2e-16 ***
## 3  67084 11295 -1   -225.50 1366.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

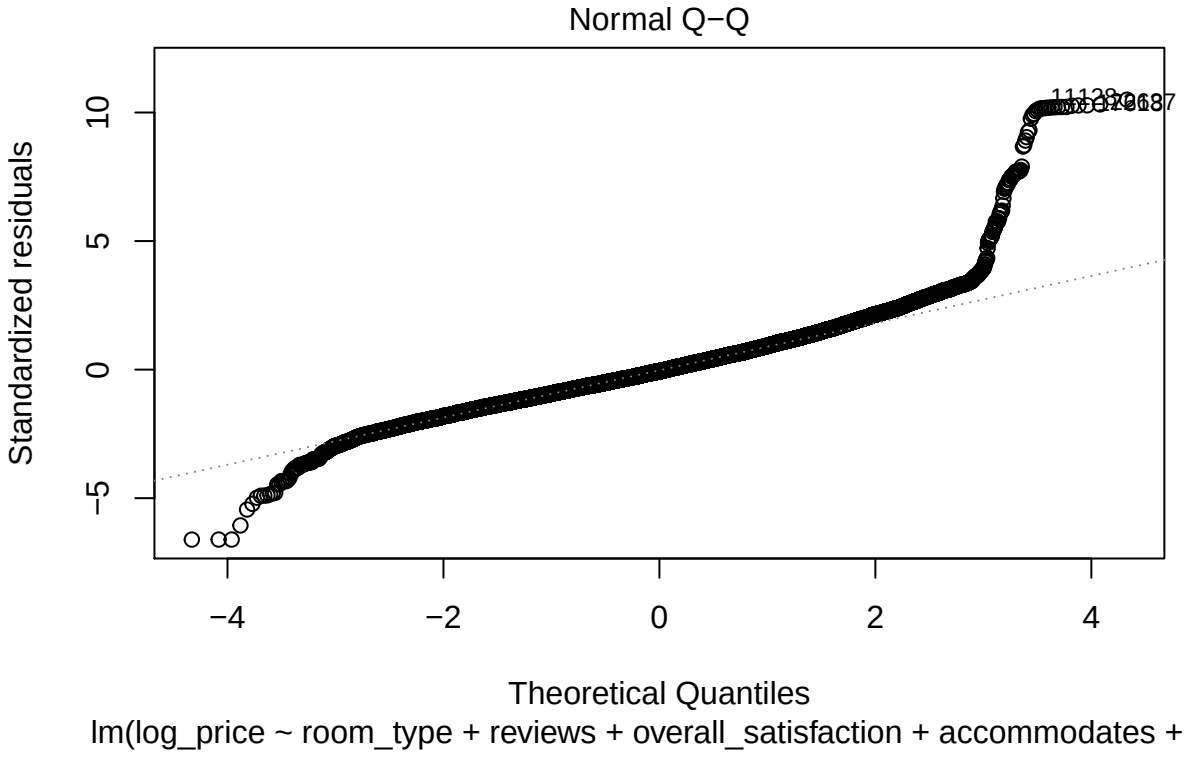
They are three linear models. The first one only has accommodates, and the second one has both accommodates and bedrooms, and the third one only has bedrooms as their predictors. From the result, the second model has the least RSS 11070, which means it has the least residuals. Thus, the linear model needs to include both predictors.

Linear regression model

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

```
##
## Call:
## lm(formula = log_price ~ room_type + reviews + overall_satisfaction +
##     accommodates + bedrooms + year, data = Van_dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6852 -0.2622 -0.0250  0.2408  4.2647
##
## Coefficients:
##                            Estimate Std. Error  t value Pr(>|t|)
## (Intercept)               4.163e+00  8.216e-03  506.778  < 2e-16 ***
## room_typePrivate room    -5.969e-01  3.859e-03 -154.686  < 2e-16 ***
## room_typeShared room     -9.864e-01  1.134e-02  -86.952  < 2e-16 ***
## reviews                  -3.920e-04  5.481e-05   -7.152 8.63e-13 ***
## overall_satisfaction      1.260e-02  1.371e-03    9.192  < 2e-16 ***
## accommodates              5.484e-02  1.483e-03   36.967  < 2e-16 ***
## bedrooms                  2.184e-01  3.144e-03   69.460  < 2e-16 ***
## year2016                 -3.936e-02  4.460e-03   -8.824  < 2e-16 ***
## year2017                 -8.379e-02  5.154e-03  -16.258  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4062 on 67083 degrees of freedom
## Multiple R-squared:  0.5246, Adjusted R-squared:  0.5245
## F-statistic:  9252 on 8 and 67083 DF,  p-value: < 2.2e-16
```





From linear regression output, the adjusted R-squared is about 0.5245, which means this model does not fit very well. And all of the predictors are significant because the p-values are small enough. As can be seen, the first plot is about residual plots, although it is symmetric, it still has some outliers. For the Q-Q plot, most of the points are in the line, but for the tail part, there are some outliers and even out of the line. Thus, from the plots, the model is not compatible with the data.

Multilevel linear regression with random intercept

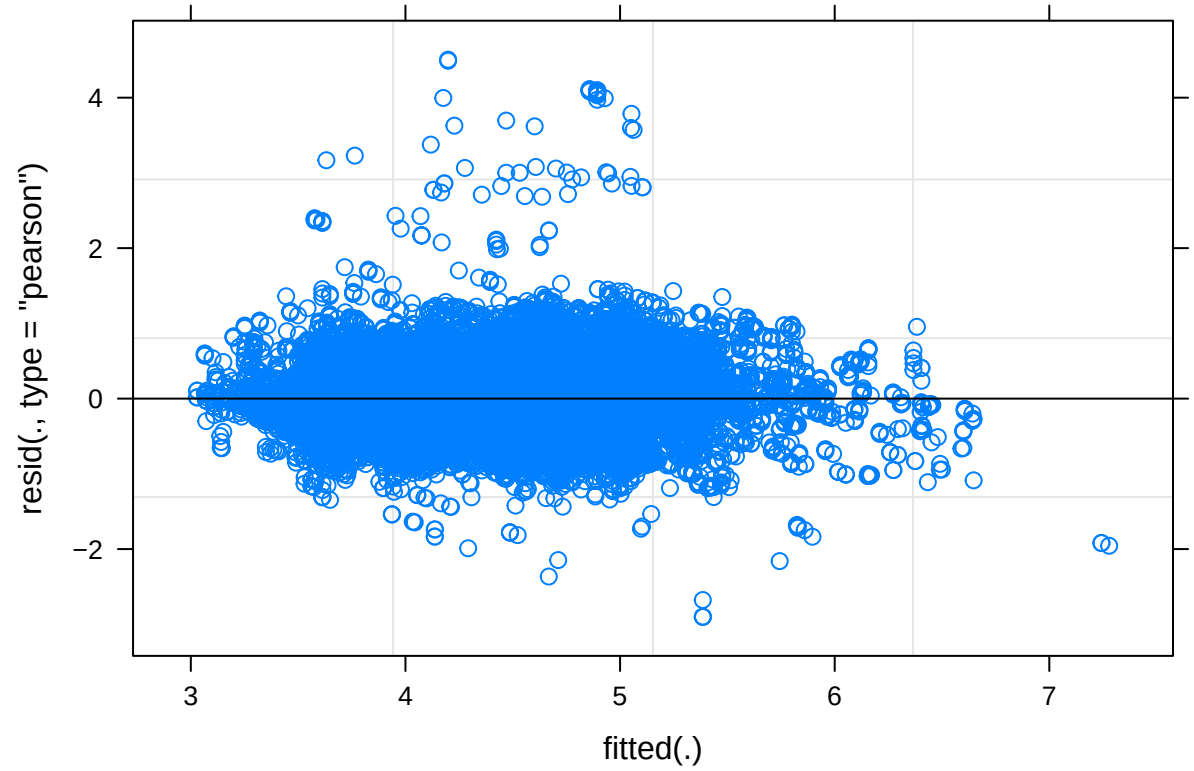
$log(price) = \alpha_i + \beta_1 x_{roomtype} + \beta_2 x_{reviews} + \beta_3 x_{overallsatisfaction} + \beta_4 x_{accommodates} + \beta_5 x_{bedrooms} + \beta_6 x_{year}$

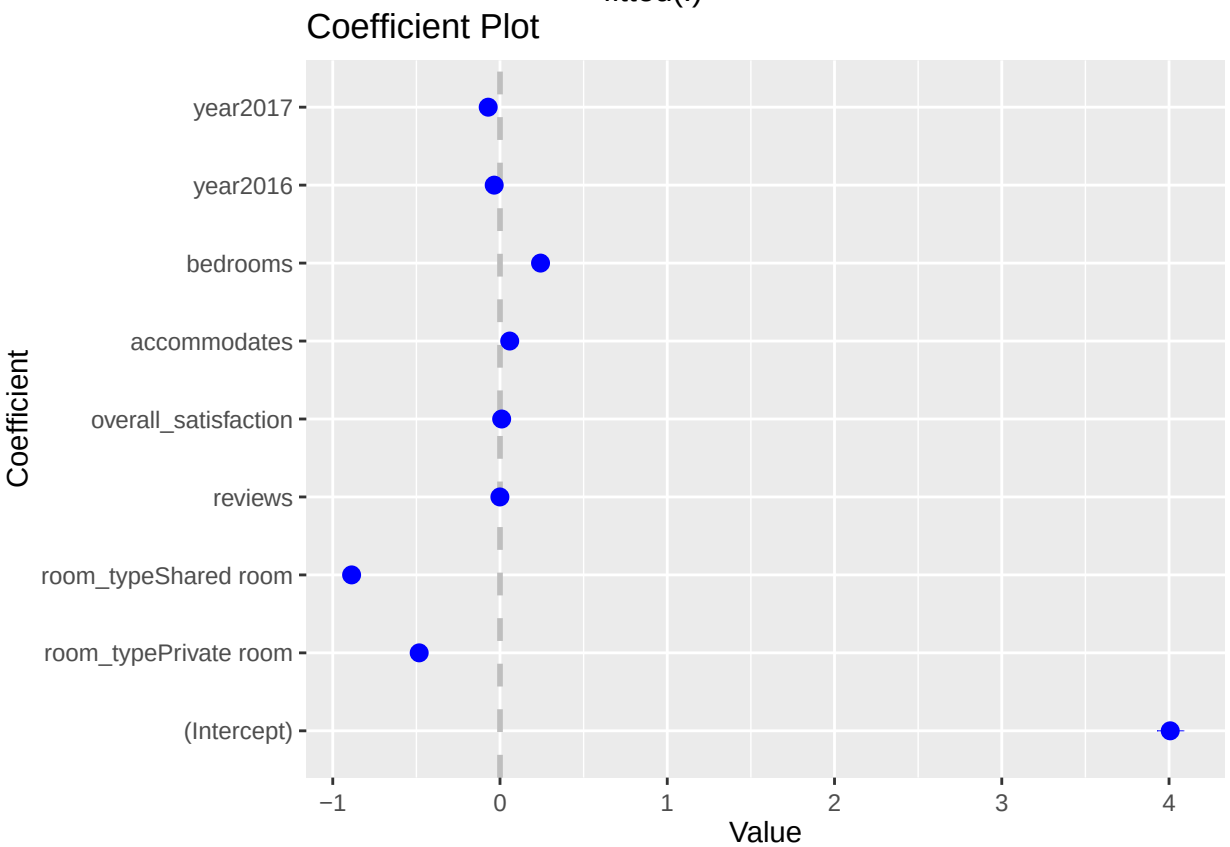
```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log_price ~ room_type + reviews + overall_satisfaction + accommodates +
##     bedrooms + year + (1 | neighborhood)
##    Data: Van_dt
##
## REML criterion at convergence: 57613.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.8175 -0.6250 -0.0395  0.5716 12.1392
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  neighborhood (Intercept) 0.03392  0.1842
##  Residual                 0.13773  0.3711
## Number of obs: 67092, groups:  neighborhood, 23
##
## Fixed effects:
##                          Estimate Std. Error  t value
## (Intercept)            4.007e+00  3.923e-02  102.151
```

```
## room_typePrivate room -4.832e-01  3.746e-03 -129.004
## room_typeShared room  -8.873e-01  1.046e-02  -84.866
## reviews               -6.700e-04  5.057e-05  -13.251
## overall_satisfaction   9.917e-03  1.254e-03    7.906
## accommodates           5.803e-02  1.364e-03   42.530
## bedrooms               2.424e-01  2.903e-03   83.497
## year2016              -3.470e-02  4.077e-03   -8.512
## year2017              -7.062e-02  4.715e-03  -14.979
##
## Correlation of Fixed Effects:
##             (Intr) rm_tPr rm_tSr reviws ovrll_ accmmd bedrms yr2016
## rm_typPrvtr -0.070
## rm_typShrdr -0.023  0.161
## reviews      0.009 -0.030  0.006
## ovrll_stsfc -0.144  0.016  0.009 -0.203
## accommodats -0.045  0.370  0.128 -0.059 -0.011
## bedrooms    -0.021 -0.104 -0.040  0.090  0.006 -0.710
## year2016    -0.081 -0.008  0.002 -0.067  0.042 -0.051 -0.002
## year2017    -0.100 -0.004  0.006 -0.099  0.247 -0.058 -0.008  0.690
```

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

```
##                              2.5 %        97.5 %
## .sig01                0.1381274442  0.2482161995
## .sigma                0.3691265971  0.3730985452
## (Intercept)           3.9287193755  4.0852723745
## room_typePrivate room -0.4905523046 -0.4758684270
## room_typeShared room  -0.9077711322 -0.8667897392
## reviews               -0.0007691292 -0.0005709205
## overall_satisfaction   0.0074589606  0.0123755802
## accommodates           0.0553555026  0.0607037850
## bedrooms               0.2367159877  0.2480957828
## year2016              -0.0426940842 -0.0267132300
## year2017              -0.0798608360 -0.0613805766
```





From the result of the confidence interval, it could be concluded that the whole predictors are significant because the intervals do not include zero point. Besides constant-coefficient, room_type plays the most important role in this model. As can be seen, the residual plot is not very well, because there are some outliers, although most of the points are among the baseline.

Multilevel linear regression with random slope

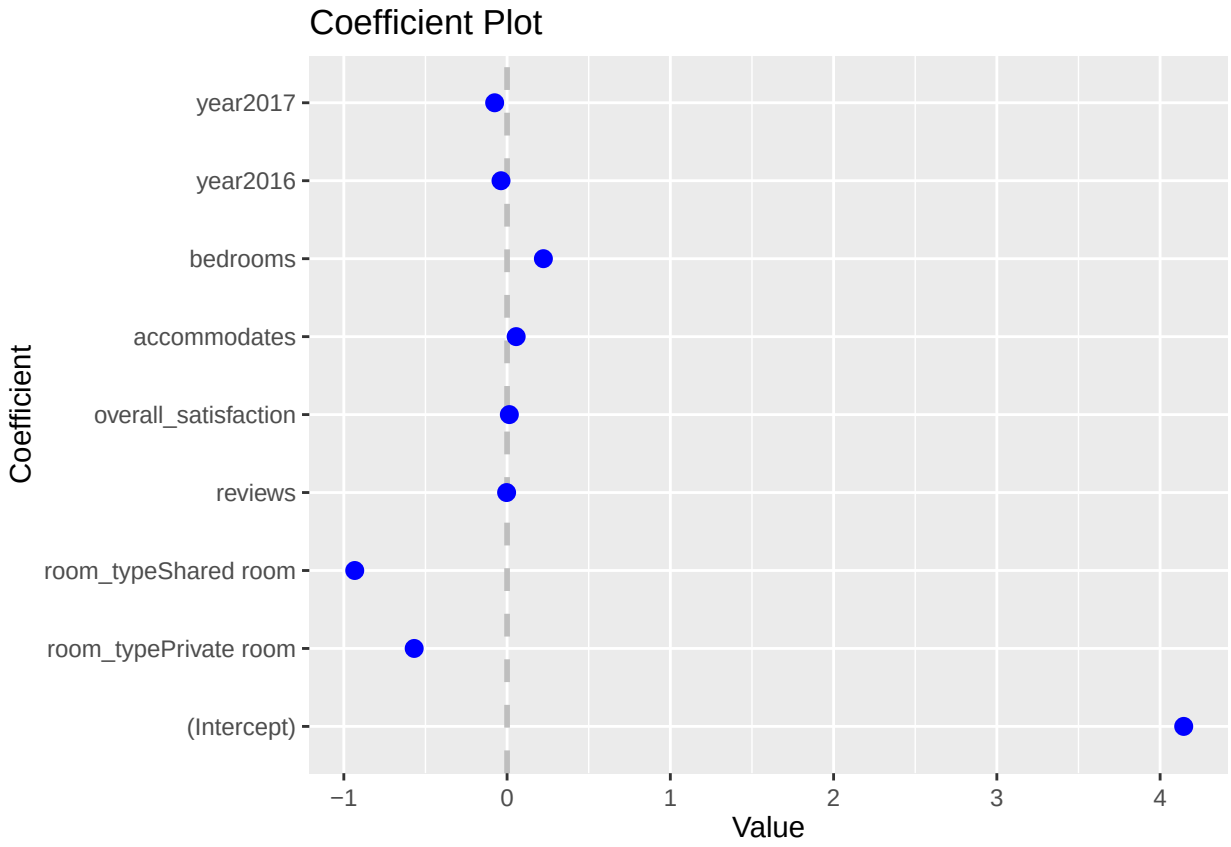
$log(price) = \alpha + \beta_1 x_{room_type} + \beta_{2[i]} x_{reviews} + \beta_3 x_{overallsatisfaction} + \beta_4 x_{accommodates} + \beta_5 x_{bedrooms} + \beta_6 x_{year}$

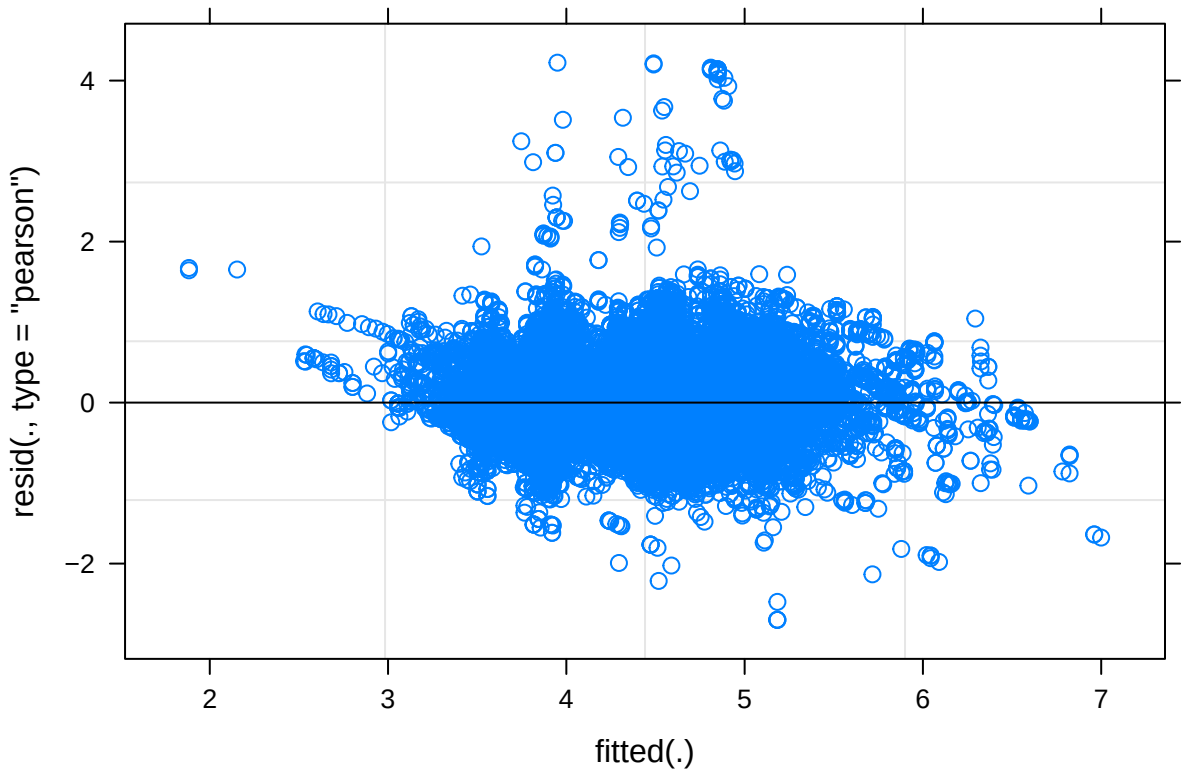
```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log_price ~ room_type + reviews + overall_satisfaction + accommodates +
##     bedrooms + year + (0 + reviews | neighborhood)
##    Data: Van_dt
##
## REML criterion at convergence: 65958.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.8331 -0.6391 -0.0610  0.5737 10.6905
##
## Random effects:
##  Groups       Name    Variance  Std.Dev.
##  neighborhood reviews 2.003e-05 0.004475
##  Residual             1.560e-01 0.394987
## Number of obs: 67092, groups:  neighborhood, 23
##
## Fixed effects:
##                          Estimate Std. Error  t value
## (Intercept)             4.144562   0.008008  517.548
## room_typePrivate room -0.568996   0.003811 -149.316
## room_typeShared room  -0.933127   0.011119  -83.923
## reviews               -0.002954   0.000940   -3.142
## overall_satisfaction   0.013801   0.001336   10.334
## accommodates           0.055648   0.001447   38.456
## bedrooms               0.222423   0.003069   72.475
## year2016              -0.037222   0.004339   -8.578
## year2017              -0.076120   0.005018  -15.170
##
## Correlation of Fixed Effects:
##             (Intr) rm_tPr rm_tSr reviws ovrll_ accmmd bedrms yr2016
## rm_typPrvtr -0.287
## rm_typShrdr -0.106  0.151
## reviews      0.006 -0.012 -0.001
## ovrll_stsfc -0.763  0.030  0.013 -0.015
## accommodats -0.210  0.371  0.125 -0.005 -0.010
## bedrooms    -0.084 -0.133 -0.043  0.003  0.011 -0.720
## year2016    -0.416 -0.013  0.002 -0.005  0.044 -0.053 -0.003
## year2017    -0.511 -0.012  0.006 -0.009  0.250 -0.060 -0.009  0.690
## convergence code: 0
## Model failed to converge with max|grad| = 0.008829 (tol = 0.002, component 1)
```

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

```
##                               2.5 %       97.5 %
## .sig01                 0.003328779  0.006066558
## .sigma                 0.392858845  0.397086184
## (Intercept)            4.128874584  4.160264821
## room_typePrivate room -0.576473439 -0.561534778
## room_typeShared room  -0.954922706 -0.911339369
## reviews               -0.004841506 -0.001076770
```

```
## overall_satisfaction   0.011183284  0.016418018
## accommodates           0.052810873  0.058482939
## bedrooms               0.216407234  0.228436851
## year2016              -0.045727546 -0.028718619
## year2017              -0.085958119 -0.066289256
```





From the result of the confidence interval, all predictors are significant. Room type has the most significant influence on this model, and then it is bedroom. As can be seen, this model does not fit very well, because, from the residual plot, some regions have outliers.

Multilevel linear regression with random intercept and random slope

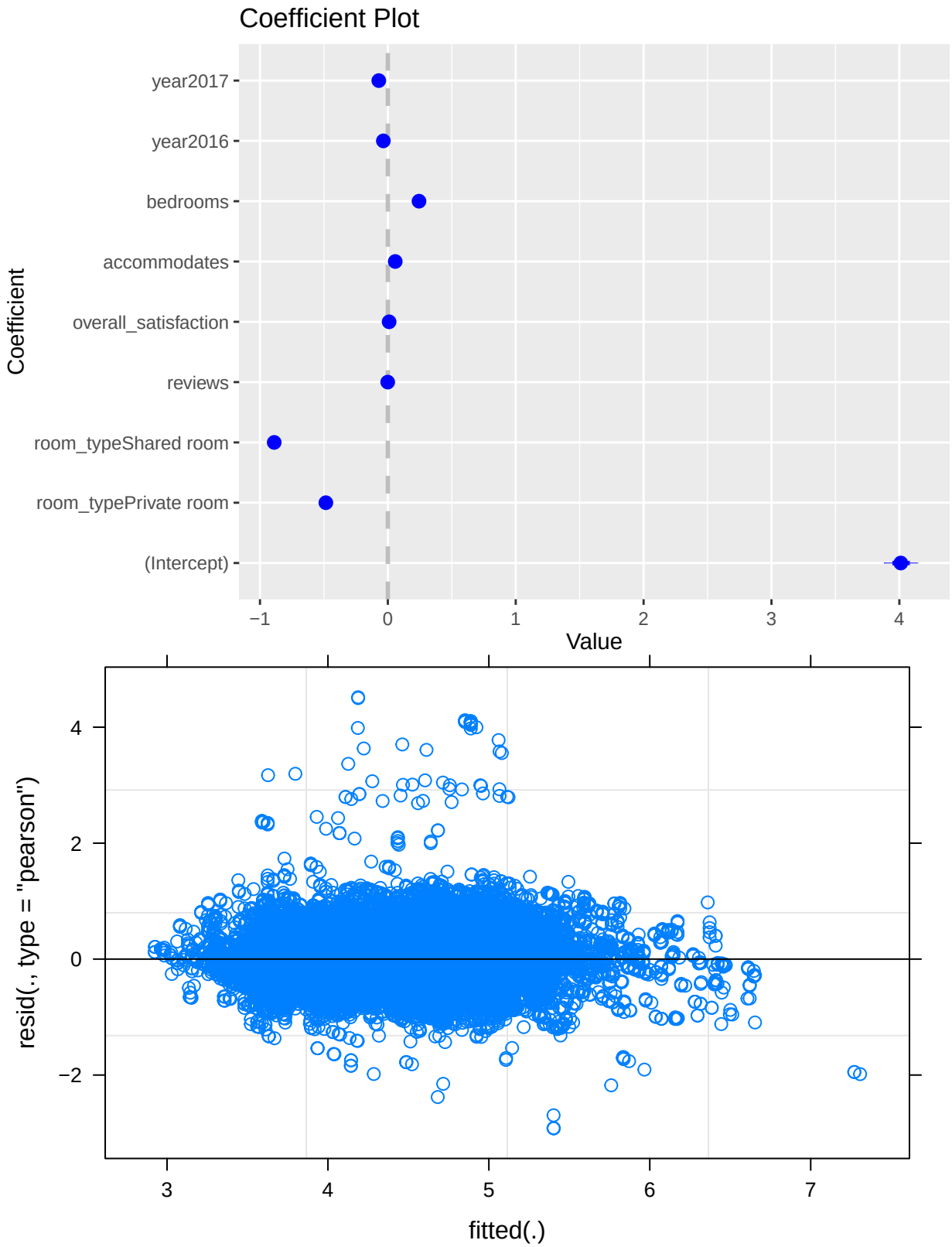
$log(price) = \alpha_i + \beta_1 x_{room_type} + \beta_{2[i]} x_{reviews} + \beta_3 x_{overallsatisfaction} + \beta_4 x_{accommodates} + \beta_5 x_{bedrooms} + \beta_6 x_{year}$

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log_price ~ room_type + reviews + overall_satisfaction + accommodates +
##     bedrooms + year + (1 + reviews | neighborhood)
##    Data: Van_dt
##
## REML criterion at convergence: 57404.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.8816 -0.6246 -0.0384  0.5695 12.1962
##
## Random effects:
##  Groups       Name        Variance  Std.Dev. Corr
##  neighborhood (Intercept) 9.602e-02 0.309866
##               reviews     2.962e-06 0.001721 -0.04
##  Residual                 1.371e-01 0.370295
## Number of obs: 67092, groups:  neighborhood, 23
##
## Fixed effects:
##                            Estimate Std. Error  t value
## (Intercept)              4.0113500  0.0651232   61.596
## room_typePrivate room -0.4847507  0.0037584 -128.978
```

```
## room_typeShared room  -0.8886793  0.0104575  -84.980
## reviews               -0.0010429  0.0003778   -2.761
## overall_satisfaction   0.0099112  0.0012552    7.896
## accommodates           0.0578274  0.0013634   42.413
## bedrooms               0.2436337  0.0029053   83.859
## year2016              -0.0350807  0.0040699   -8.620
## year2017              -0.0708487  0.0047082  -15.048
##
## Correlation of Fixed Effects:
##             (Intr) rm_tPr rm_tSr reviws ovrll_ accmmd bedrms yr2016
## rm_typPrvtr -0.043
## rm_typShrdr -0.014  0.159
## reviews     -0.042  0.004  0.001
## ovrll_stsfc -0.086  0.015  0.010 -0.041
## accommodats -0.027  0.368  0.127 -0.004 -0.012
## bedrooms    -0.013 -0.101 -0.041  0.013  0.006 -0.709
## year2016    -0.049 -0.007  0.002 -0.011  0.044 -0.051 -0.002
## year2017    -0.060 -0.004  0.006 -0.017  0.249 -0.059 -0.008  0.690
## convergence code: 0
## Model failed to converge with max|grad| = 15.5174 (tol = 0.002, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?

## Computing profile confidence intervals ...

##                              2.5 %       97.5 %
## .sig01                 0.137810414  0.2479984283
## .sig02                -0.423127787  0.4033968845
## .sig03                 0.001039789  0.0025804432
## .sigma                 0.368343511  0.3723083824
## (Intercept)            3.933062709  4.0895433770
## room_typePrivate room -0.492247560 -0.4775123910
## room_typeShared room  -0.909211806 -0.8682165720
## reviews               -0.001855313 -0.0003012244
## overall_satisfaction   0.007450342  0.0123720593
## accommodates           0.055150049  0.0604948300
## bedrooms               0.237913411  0.2493025304
## year2016              -0.043072410 -0.0271175144
## year2017              -0.080091139 -0.0616338986
```



Whole predictors are signigicant, and room type is the most influential term. From the residual plot, it seems better than model 3.

ANOVA Test:

In order to check which model is the best, I prefer to use ANOVA to test these three models.

```
## Data: Van_dt
## Models:
## m2: log_price ~ room_type + reviews + overall_satisfaction + accommodates +
## m2:     bedrooms + year + (1 | neighborhood)
## m3: log_price ~ room_type + reviews + overall_satisfaction + accommodates +
## m3:     bedrooms + year + (0 + reviews | neighborhood)
## m4: log_price ~ room_type + reviews + overall_satisfaction + accommodates +
## m4:     bedrooms + year + (1 + reviews | neighborhood)
##    Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## m2 11 57636 57736 -28807    57614
## m3 11 65981 66081 -32979    65959    0.0      0          1
## m4 13 57431 57549 -28702    57405 8554.4      2     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the output of the ANOVA test, because the model 4 that has random intercept and the random slope has the least AIC 57431, and it also has the least number of deviance 57405. Thus the fourth model is the best among these 4 models.

V. Discussion:

Implication

From this analysis of Airbnb in Vancouver, it can be concluded that the room type is the most influential factor in terms of price. And entire home/apt is the most expensive type in room type choices. And also, when travelers are choosing Airbnb, they should concern the number of bedrooms and the neighborhoods. From my survey, the number of bedrooms has a significant influence on the price of Airbnb. In terms of neighborhoods, downtown has the most number of accommodations, but the price is in the top three. Thus, when people plan to travel and want to choose Airbnb, they need to consider the room type, number of bedrooms and the location of accommodations.

Limitation

My dataset is only from 2015 to 2017, and the data in 2015 and 2017 are not complete. So, I cannot compare these three years directly. Thus, the result may have deviations, and I am not sure whether it is useful for 2019 or not. Besides, my report is only about Vancouver. Thus, it may not be compatible with another region. For the prediction part, the result could not be precise, since predictors are limited.

Future direction

To improve the precision, I would like to search for another bigger dataset, which includes predictors like the number of facilities in accommodations, and the conditions of transportation near the locations and so on.

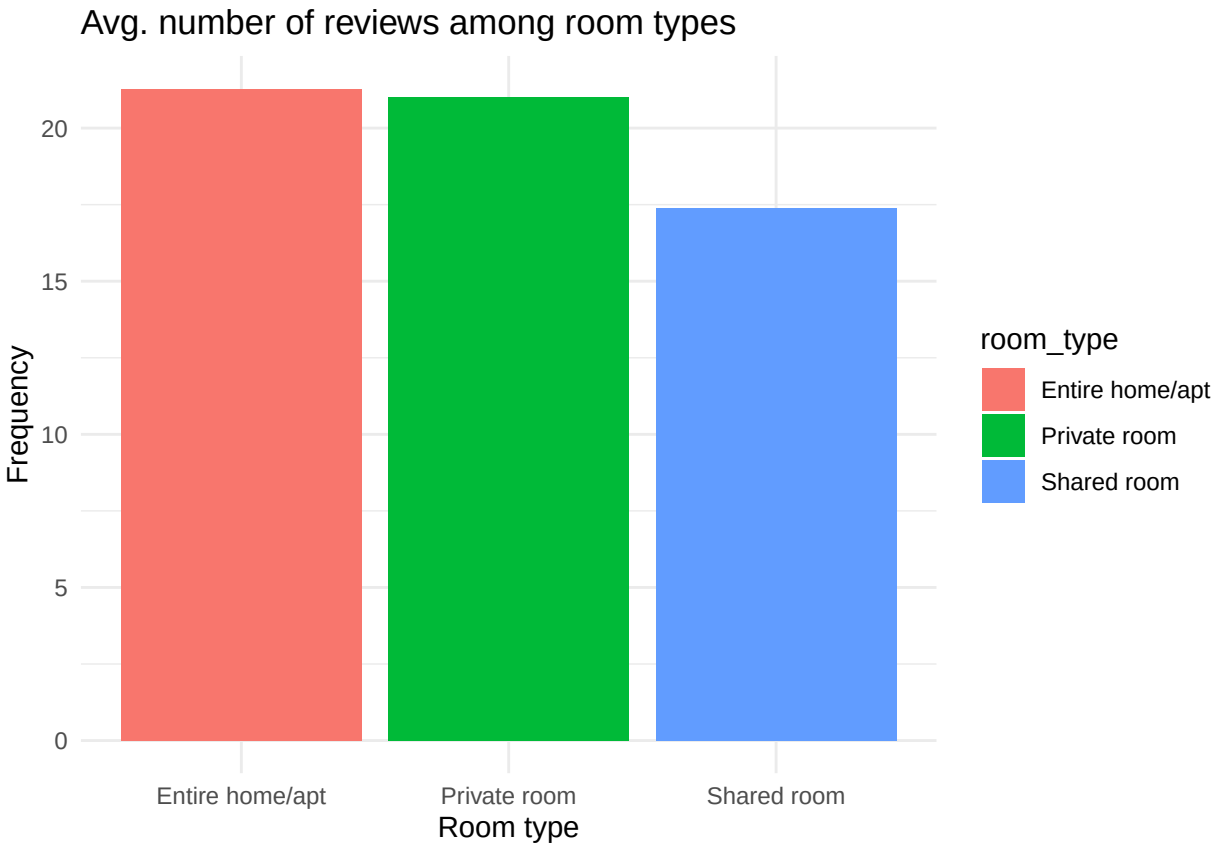
VI. Reference:

http://tomslee.net

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

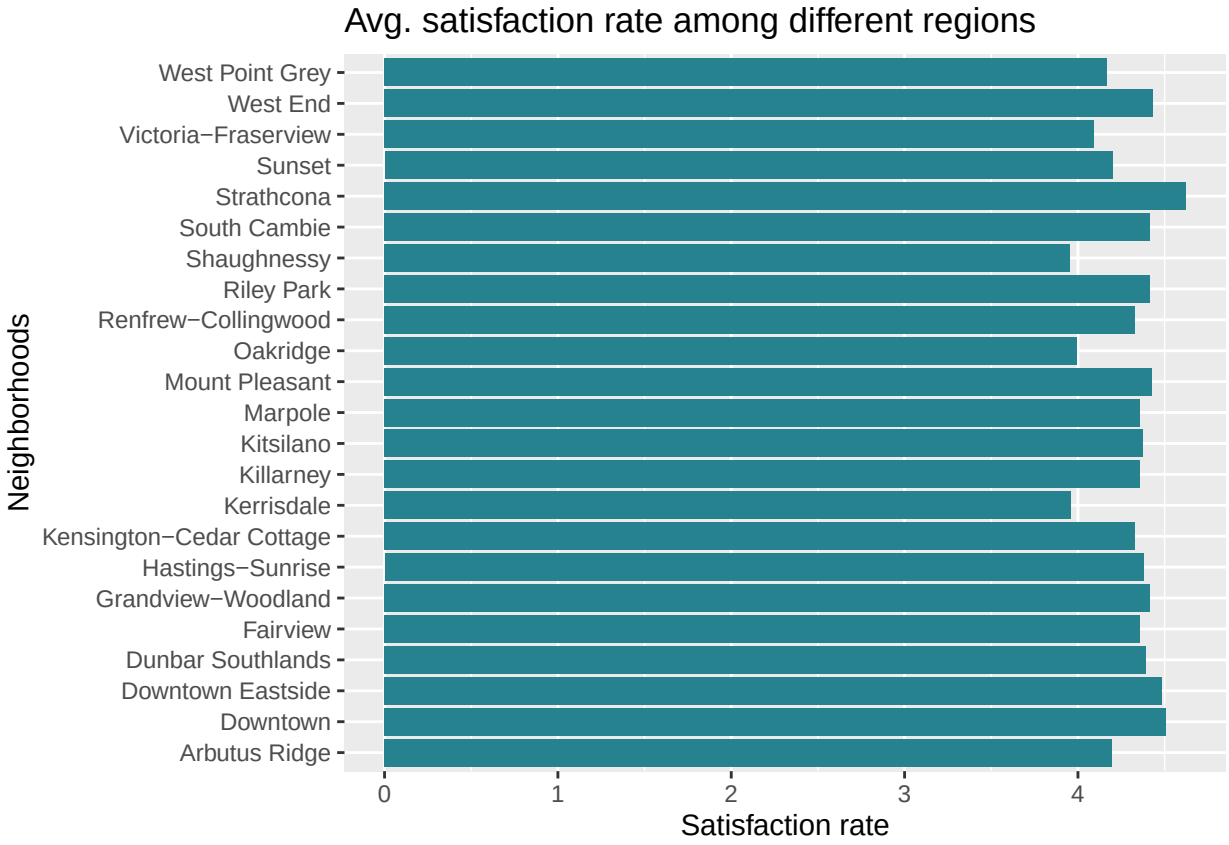
VII. Appendix:

Appendix I



The entire home/apt has the highest number of average reviews. Conversely, the type of shared room has the least information about reviews.

Appendix II



It could be found, Strathcona, Downtown, and Downtown Eastside have the top three high satisfaction rates.