678 final

xiaoyanbin Cai

2022-12-10

Abstract

Since I personally love traveling a lot, I chose the database of Audemars Piguet as my project data. I wanted to know if the price of a home could be influenced by other factors. Such as the type of house or the location of the house. So in this assignment, I will use EDA and Multilevel linear model to predict the relationship between house price and other coefficients.

Introduction

I decided to choose the Airbnb data set that related to Massachusetts as my final project data resource. Since I would like to know how price and other variables will correlated. The whole data set will consist of following parts"id","host_id","neighbourhood","room_type","price","minimum_price","number_of_reviews","calculated_host_listi
Firstly I will do the visualization to find out which variable has a strong relationship with price and then I will do the multilevel regression. # Loading library

Method

Importing data

```
##        id                 name              host_id           host_name
##  Min.   :3.168e+03   Length:14803       Min.   :     3697   Length:14803
##  1st Qu.:1.936e+07   Class :character   1st Qu.: 18517776   Class :character
##  Median :4.035e+07   Mode  :character   Median : 88140341   Mode  :character
##  Mean   :6.495e+16                      Mean   :128865563
##  3rd Qu.:5.104e+07                      3rd Qu.:222978210
##  Max.   :7.203e+17                      Max.   :479130189
##
##  neighbourhood_group neighbourhood        latitude       longitude
##  Mode:logical        Length:14803       Min.   :42.23   Min.   :-71.20
##  NA's:14803          Class :character   1st Qu.:42.33   1st Qu.:-71.11
##                      Mode  :character   Median :42.35   Median :-71.08
##                                         Mean   :42.35   Mean   :-71.09
##                                         3rd Qu.:42.36   3rd Qu.:-71.06
##                                         Max.   :42.42   Max.   :-70.92
##
##   room_type             price         minimum_nights    number_of_reviews
##  Length:14803       Min.   :    0.0   Min.   :   1.00   Min.   :   0.00
##  Class :character   1st Qu.:   96.0   1st Qu.:   2.00   1st Qu.:   0.00
##  Mode  :character   Median :  165.0   Median :  28.00   Median :   8.00
##                     Mean   :  218.6   Mean   :  30.89   Mean   :  44.28
```

```
##                    3rd Qu.:  256.0   3rd Qu.:  32.00   3rd Qu.:  50.00
##                    Max.   :10000.0   Max.   :1000.00   Max.   :1489.00
##
##  last_review        reviews_per_month calculated_host_listings_count
##  Length:14803       Min.   : 0.01     Min.   :  1.00
##  Class :character   1st Qu.: 0.29     1st Qu.:  2.00
##  Mode  :character   Median : 1.00     Median :  7.00
##                     Mean   : 1.74     Mean   : 43.12
##                     3rd Qu.: 2.54     3rd Qu.: 33.00
##                     Max.   :43.37     Max.   :399.00
##                     NA's   :3840
##  availability_365 number_of_reviews_ltm   license
##  Min.   :  0.0    Min.   :  0.00        Length:14803
##  1st Qu.: 70.0    1st Qu.:  0.00        Class :character
##  Median :195.0    Median :  1.00        Mode  :character
##  Mean   :190.4    Mean   : 11.12
##  3rd Qu.:317.0    3rd Qu.: 13.00
##  Max.   :365.0    Max.   :553.00
##
```

```
##      id                                              name host_id host_name
## 1  7903         Colorful, modern 2 BR apt shared with host   14169     Stacy
## 2  8521   SunsplashedSerenity walk to Harvard & Fresh Pond  306681     Janet
## 3  8789                Curved Glass Studio/1bd facing Park   26988      Anne
## 4 10813 Back Bay Apt-blocks to subway, Newbury St, The Pru   38997  Michelle
## 5 10986 North End (Waterfront area)  CLOSE TO MGH & SUBWAY   38997  Michelle
## 6 11169    Lovely Studio Room: Available for long weekends   40965      Judy
##   neighbourhood_group   neighbourhood latitude longitude       room_type price
## 1                  NA         Agassiz 42.39031 -71.09361    Private room   118
## 2                  NA  West Cambridge 42.38258 -71.13646 Entire home/apt   300
## 3                  NA  East Cambridge 42.35867 -71.06307 Entire home/apt   110
## 4                  NA      Area 2/MIT 42.35061 -71.08787 Entire home/apt   135
## 5                  NA  East Cambridge 42.36377 -71.05206 Entire home/apt   135
## 6                  NA North Cambridge 42.39630 -71.13135    Private room   116
##   minimum_nights number_of_reviews last_review reviews_per_month
## 1              4               295  2022-09-06              1.83
## 2              2                50  2022-08-28              0.31
## 3             91                25  2020-04-15              0.25
## 4             29                 5  2020-12-02              0.06
## 5             33                 2  2016-05-23              0.03
## 6              3               160  2022-09-05              1.05
##   calculated_host_listings_count availability_365 number_of_reviews_ltm
## 1                              1               10                    16
## 2                              2              258                     5
## 3                              5              279                     0
## 4                             11              355                     0
## 5                             11              356                     0
## 6                              2              331                    11
##      license
## 1
## 2 C0121120491
## 3
## 4
## 5
```

```
## 6
```

Basic data information and data cleaning

```
## Warning in 9:11:12: numerical expression has 3 elements: only the first used
```

id	name	host_id	neighbourhood	room_type
7903	Colorful, modern 2 BR apt shared with host	14169	Agassiz	Private room
8521	SunsplashedSerenity walk to Harvard & Fresh Pond	306681	West Cambridge	Entire home/apt
8789	Curved Glass Studio/1bd facing Park	26988	East Cambridge	Entire home/apt
10813	Back Bay Apt-blocks to subway, Newbury St, The Pru	38997	Area 2/MIT	Entire home/apt
10986	North End (Waterfront area) CLOSE TO MGH & SUBWAY	38997	East Cambridge	Entire home/apt
11169	Lovely Studio Room: Available for long weekends	40965	North Cambridge	Private room

```
## Warning in 9:11:12: numerical expression has 3 elements: only the first used

##        id                 name              host_id          neighbourhood
##  Min.   :3.168e+03   Length:14803       Min.   :     3697   Length:14803
##  1st Qu.:1.936e+07   Class :character   1st Qu.: 18517776   Class :character
##  Median :4.035e+07   Mode  :character   Median : 88140341   Mode  :character
##  Mean   :6.495e+16                      Mean   :128865563
##  3rd Qu.:5.104e+07                      3rd Qu.:222978210
##  Max.   :7.203e+17                      Max.   :479130189
##
##   room_type             price         minimum_nights    number_of_reviews
##  Length:14803       Min.   :    0.0   Min.   :   1.00   Min.   :   0.00
##  Class :character   1st Qu.:   96.0   1st Qu.:   2.00   1st Qu.:   0.00
##  Mode  :character   Median :  165.0   Median :  28.00   Median :   8.00
##                     Mean   :  218.6   Mean   :  30.89   Mean   :  44.28
##                     3rd Qu.:  256.0   3rd Qu.:  32.00   3rd Qu.:  50.00
##                     Max.   :10000.0   Max.   :1000.00   Max.   :1489.00
##
##  reviews_per_month calculated_host_listings_count availability_365
##  Min.   : 0.01     Min.   :  1.00                 Min.   :  0.0
##  1st Qu.: 0.29     1st Qu.:  2.00                 1st Qu.: 70.0
##  Median : 1.00     Median :  7.00                 Median :195.0
##  Mean   : 1.74     Mean   : 43.12                 Mean   :190.4
##  3rd Qu.: 2.54     3rd Qu.: 33.00                 3rd Qu.:317.0
##  Max.   :43.37     Max.   :399.00                 Max.   :365.0
##  NA's   :3840
##  number_of_reviews_ltm
##  Min.   :  0.00
##  1st Qu.:  0.00
##  Median :  1.00
##  Mean   : 11.12
##  3rd Qu.: 13.00
##  Max.   :553.00
##

## [1] 10957
```

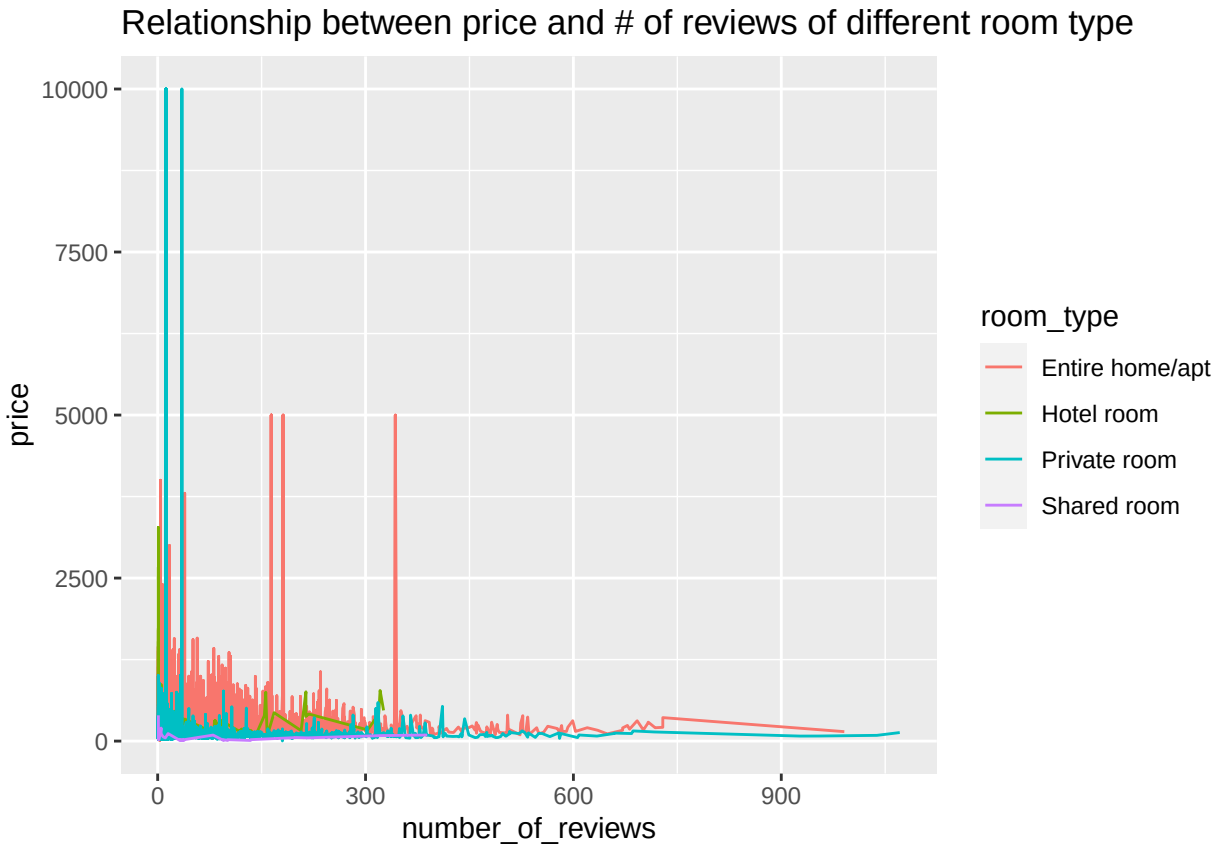
- As we can see from new data frame summary, there're some NA values in neighbourhood_group and reviews_per_month. So that I need to elimate all the NA values that include in price and number_of_reviews. After that we get 10957 results in total

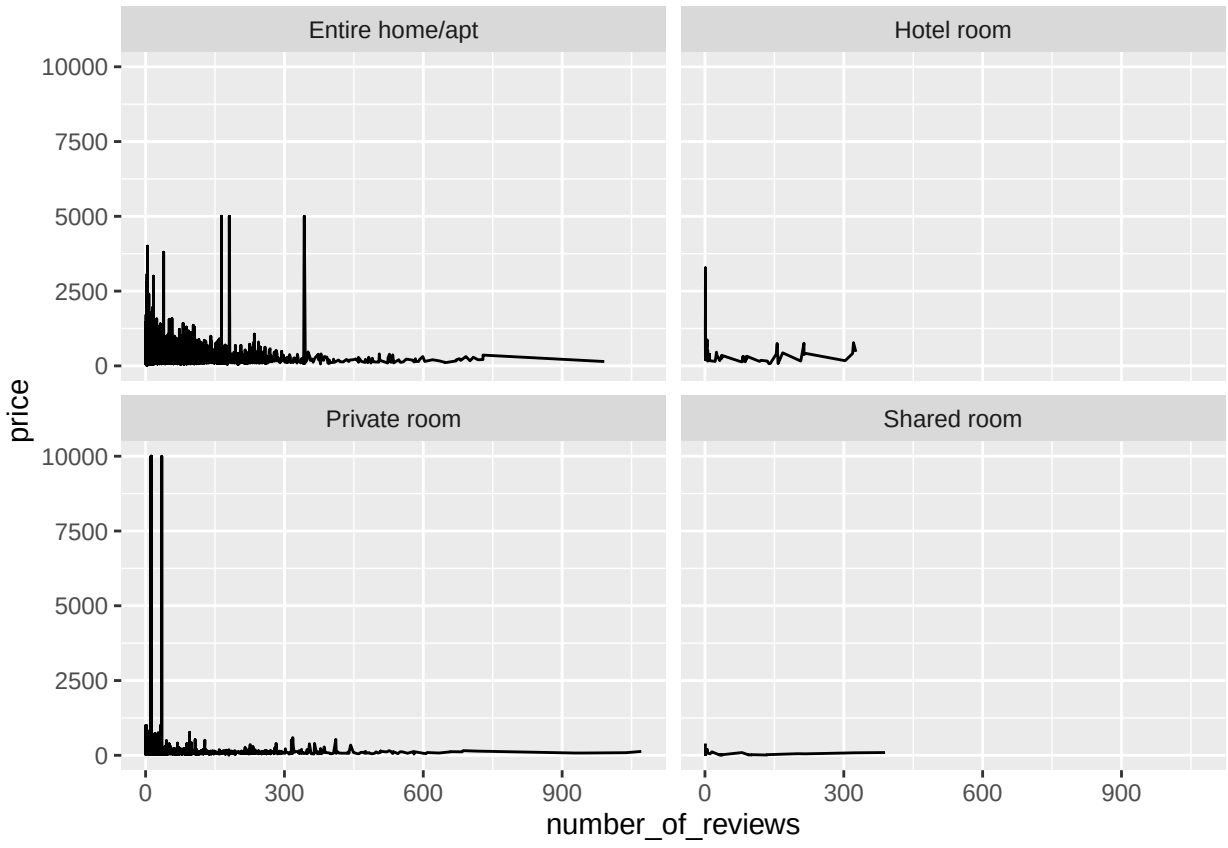
Table about room type

Var1	Freq
Entire home/apt	6937
Hotel room	64
Private room	3919
Shared room	37

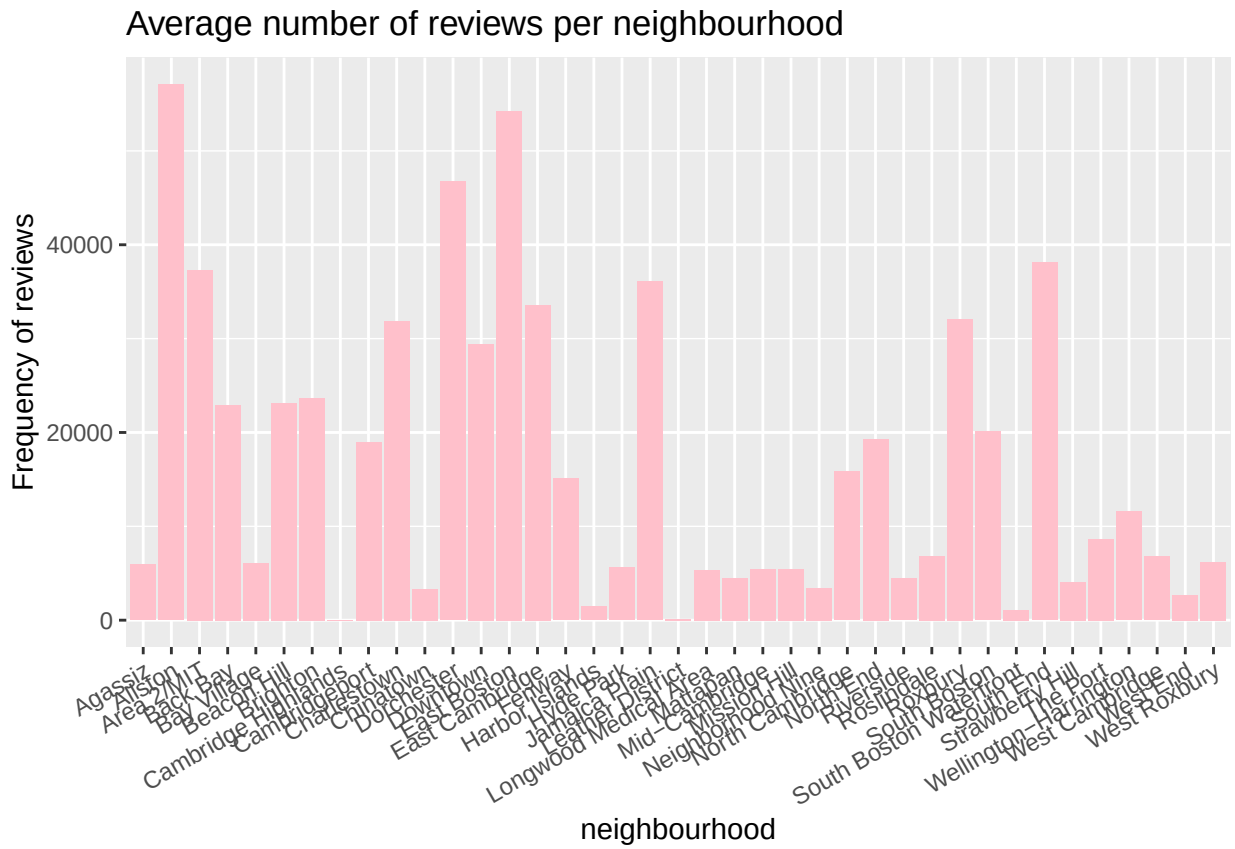
- As we can see from above,home/apt has the largest number in the table which is 6937. So that we can see how the relationship between the most-used room type and the price.

EDA

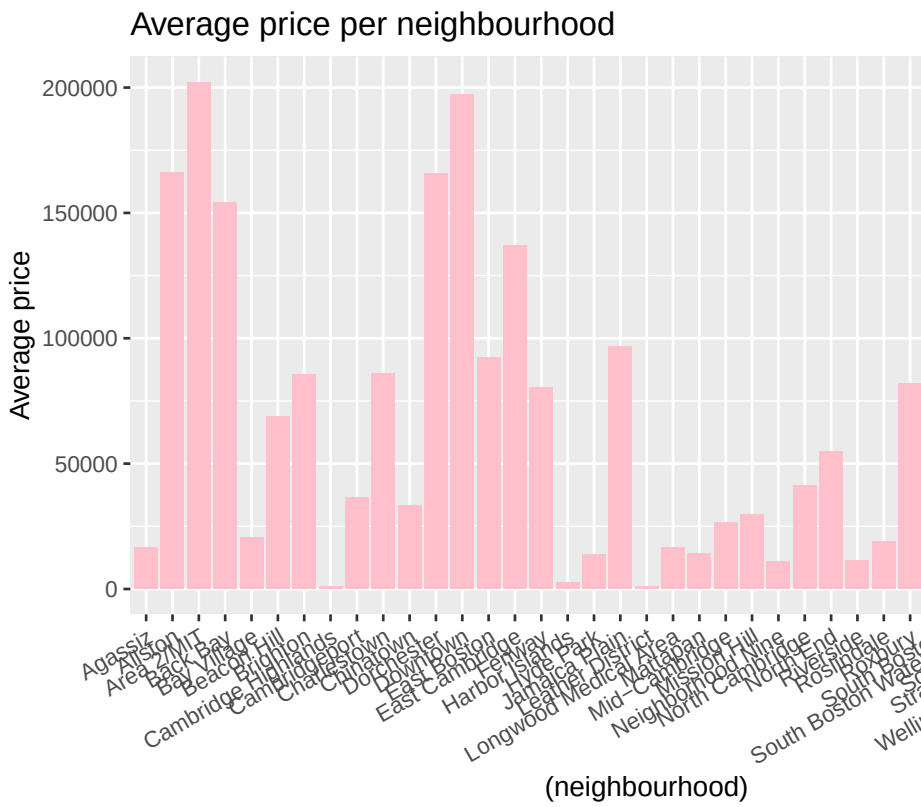

Relationship between price and # of reviews of different room type
price
number_of_reviews
room_type
Entire home/apt
Hotel room
Private room
Shared room
0
2500
5000
7500
10000
0
300
600
900


Entire home/apt
Hotel room
Private room
Shared room
price
number_of_reviews
0
2500
5000
7500
10000
0
300
600
900

* I made two linear graphs about relation between price and number of reviews of different room types. As we can see from 'plot2'entire home/apt has the most number of reviews and change of price is relative stable than that of other room type. And for shared room which has the most small number of reviews and the price of shared room is almost the same.



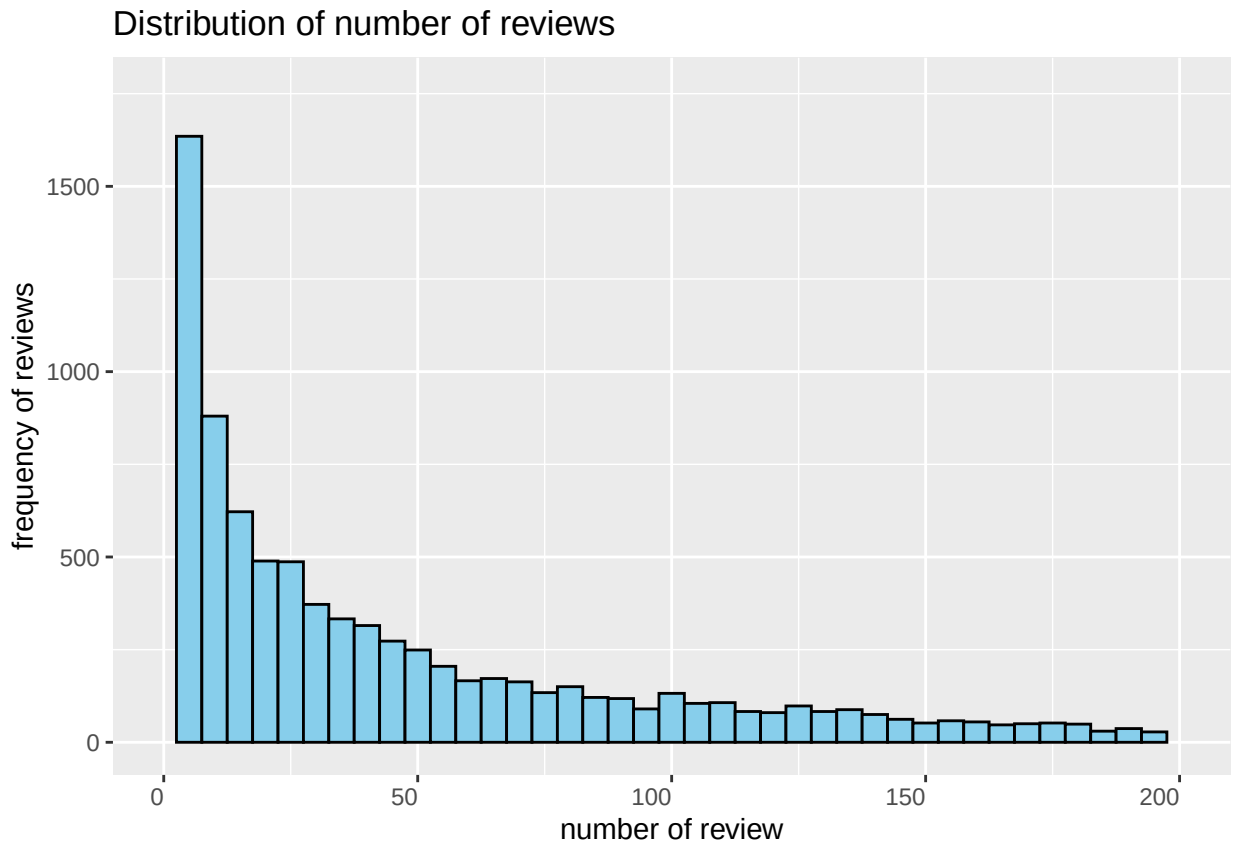
* As we can see from the graph above, Allston has the highest number of reviews and East Boston is second highest and Dorchester is the third one.



- Fig of average price per neighbourhood
- Fig of distribution of number of reviews

```
## Warning: Removed 827 rows containing non-finite values (‘stat_bin()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```



* According to the graph above, we can see that the majority of airbnb have less than 100 reviews

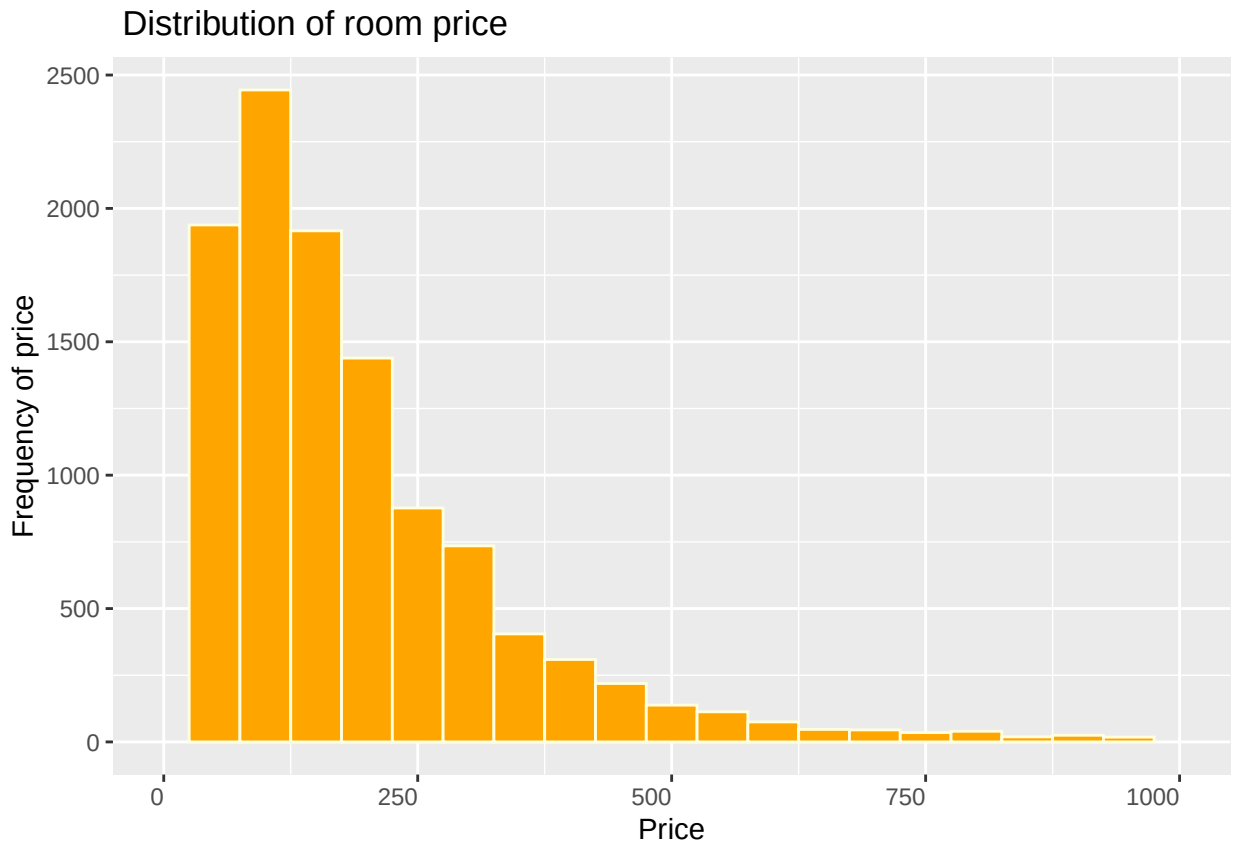
- Using leaflet to make a map of Airbnb in MA
- To see how the correlation between price and minimum nights. Because higher price may affect the number of reviews.

```
##
##  Pearson's product-moment correlation
##
## data:  Airbnb2$minimum_nights and Airbnb2$price
## t = -7.6544, df = 10955, p-value = 2.105e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09153639 -0.05428648
## sample estimates:
##         cor
## -0.07293687
```

- The p-value is 2.105e-14. So it reject the null hypothesis. So the correlation between price and minimum nights is significant. I might want to add the correlation term into the model to test whether this influence is significant.
- Distribution of room price

```
## Warning: Removed 78 rows containing non-finite values (`stat_bin()`).
```

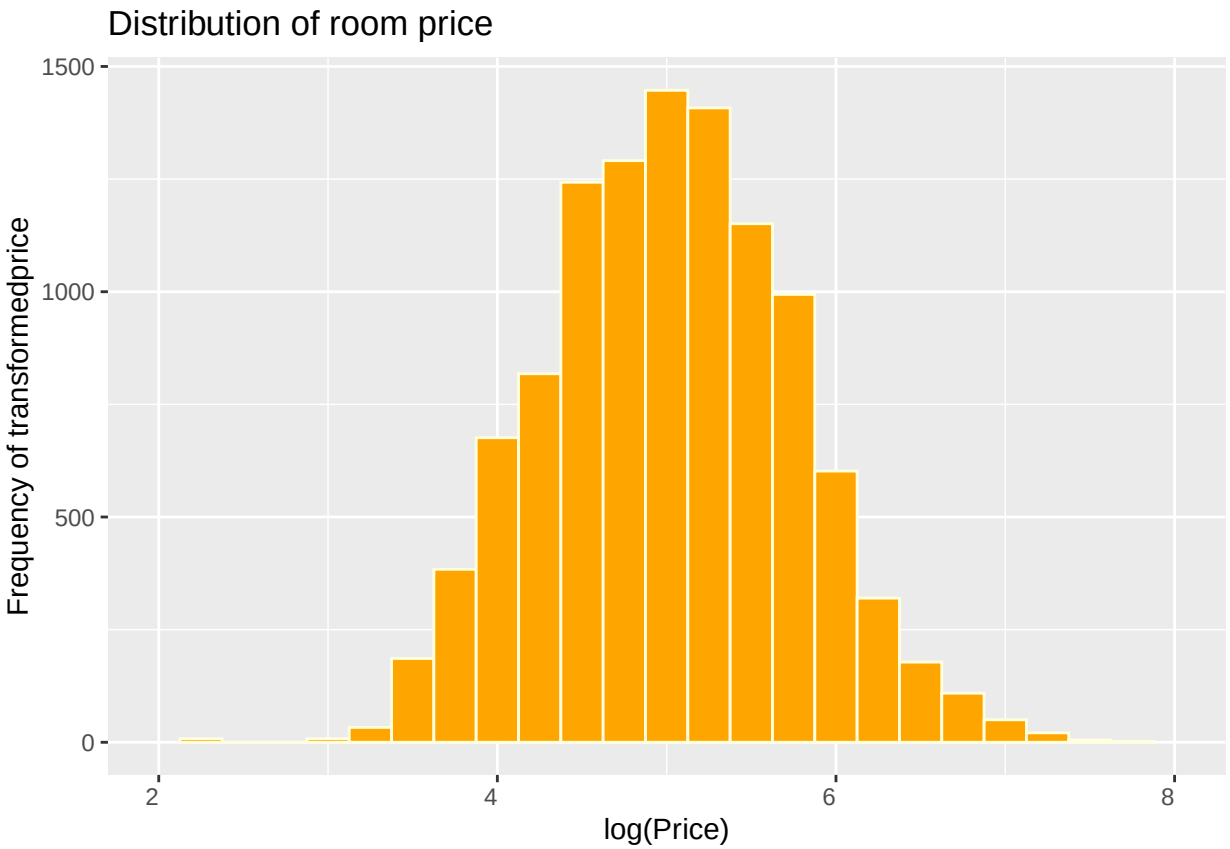
```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```



- Since from the graph above, price does not present as normal distribution so that I need to do the log transformation.

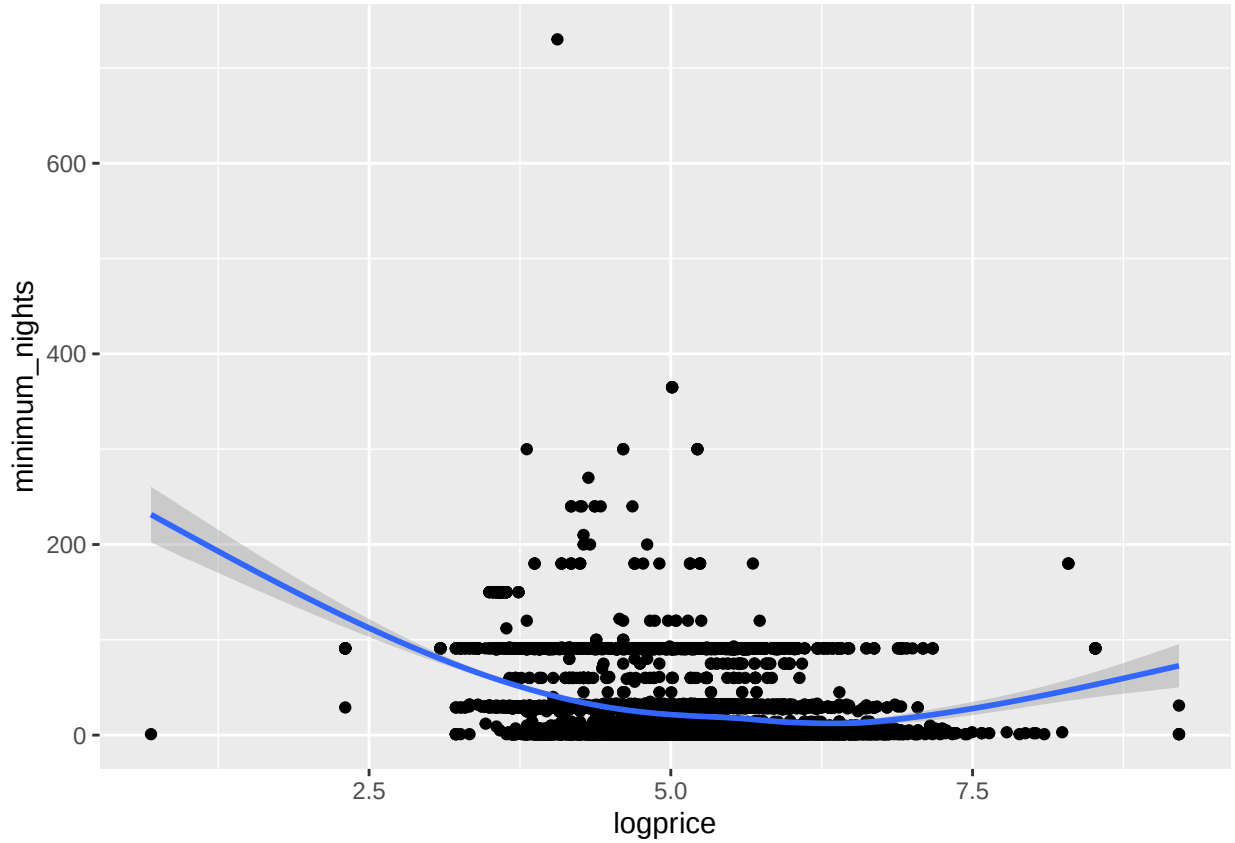
```
## Warning: Removed 21 rows containing non-finite values (‘stat_bin()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```



* Visualize the relationship between logprice and minimum nights

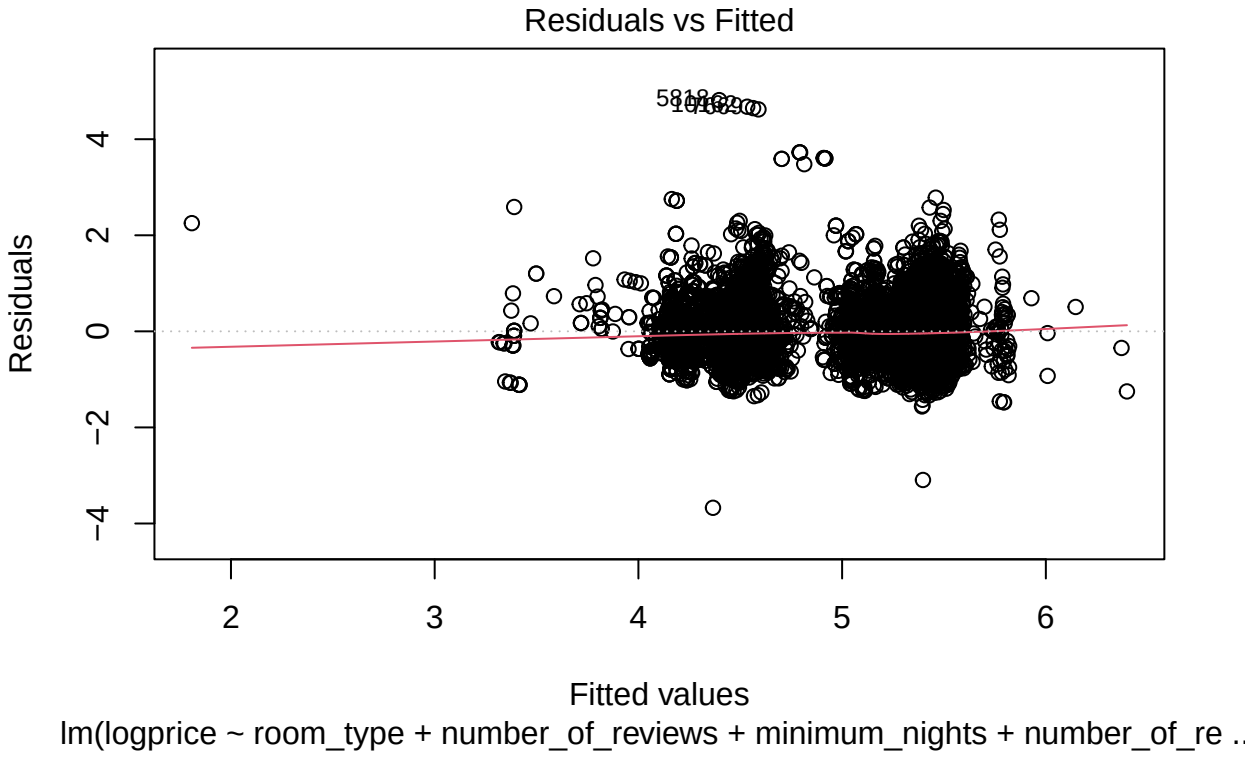
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

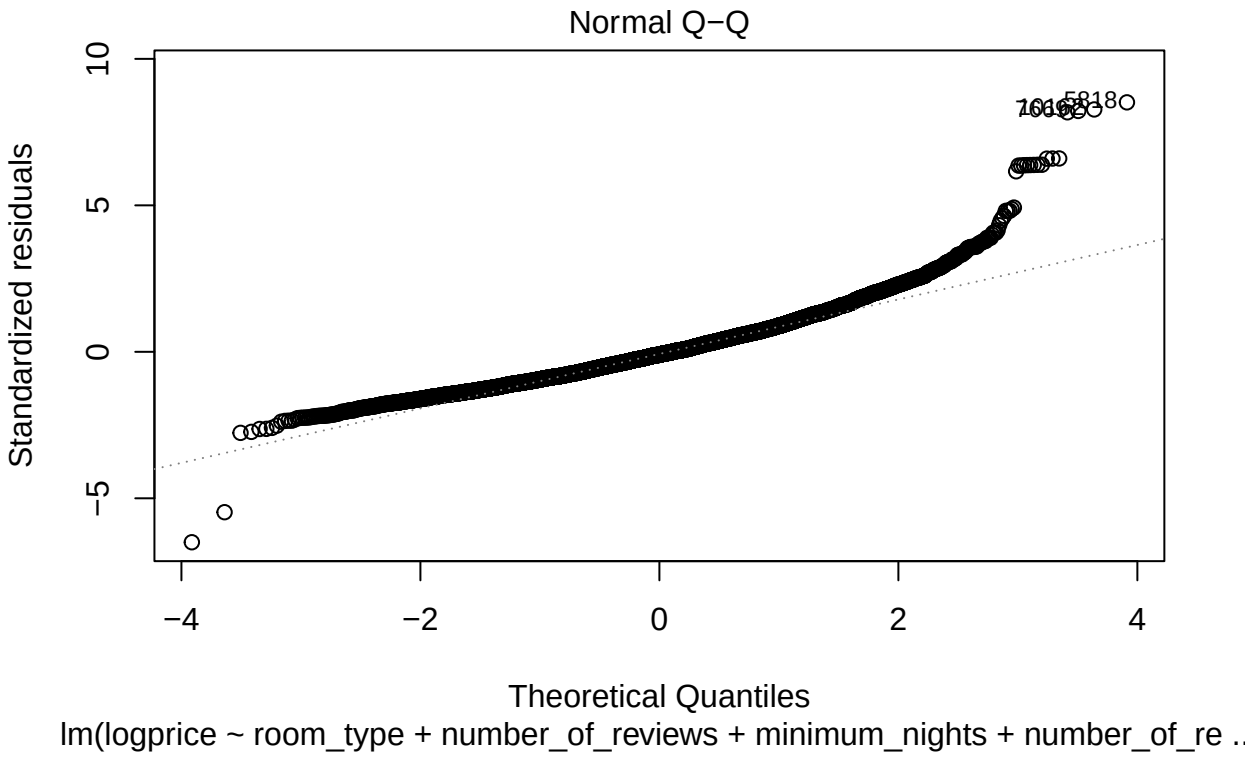


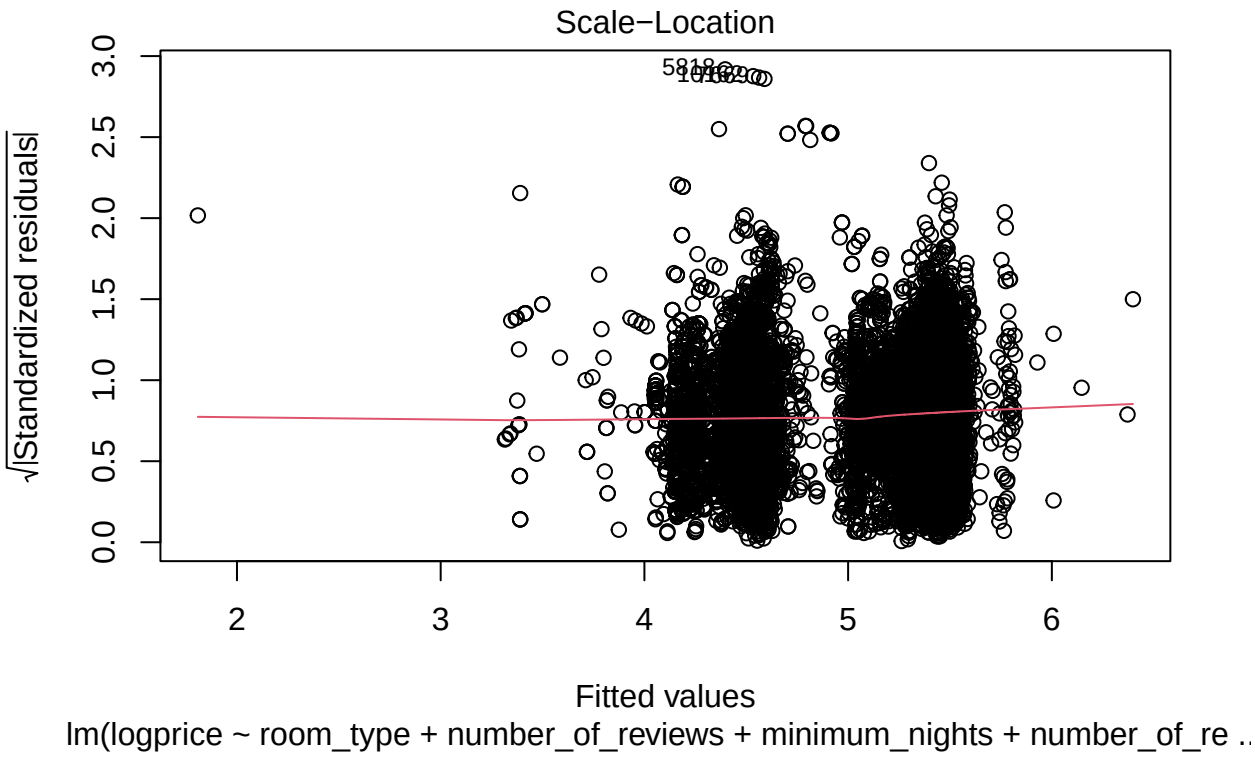
* As we can see from the graph above, the graph does not suggest a linearly increasing relationship between the logprice and minimum nights. So the assumption that lower price may lead to more days to stay may not be a right assumption ## Modeling * Doing model 1 of Simple Linear Regression

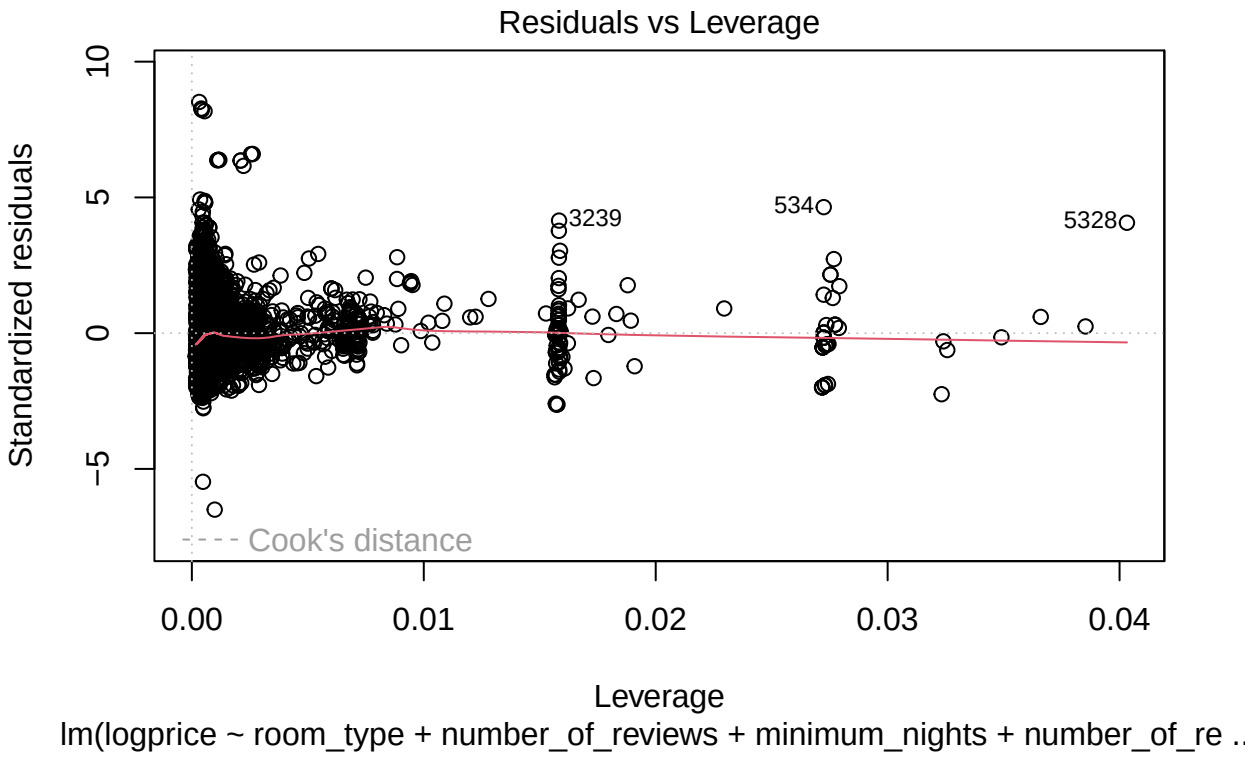
```
##
## Call:
## lm(formula = logprice ~ room_type + number_of_reviews + minimum_nights +
##     number_of_reviews_ltm + reviews_per_month + calculated_host_listings_count +
##     availability_365, data = price_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6730 -0.3951 -0.0537  0.3137  4.8137
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.395e+00  1.316e-02 409.993  < 2e-16 ***
## room_typeHotel room             3.072e-01  7.129e-02   4.310 1.65e-05 ***
## room_typePrivate room          -8.861e-01  1.146e-02 -77.332  < 2e-16 ***
## room_typeShared room           -1.656e+00  9.359e-02 -17.695  < 2e-16 ***
## number_of_reviews              -6.136e-04  7.448e-05  -8.240  < 2e-16 ***
## minimum_nights                 -3.825e-03  1.623e-04 -23.568  < 2e-16 ***
## number_of_reviews_ltm           3.240e-03  4.094e-04   7.913 2.75e-15 ***
## reviews_per_month              -1.201e-02  4.877e-03  -2.463 0.013780 *
## calculated_host_listings_count  4.316e-04  1.273e-04   3.390 0.000702 ***
## availability_365                3.047e-04  4.397e-05   6.929 4.48e-12 ***
## ---
```

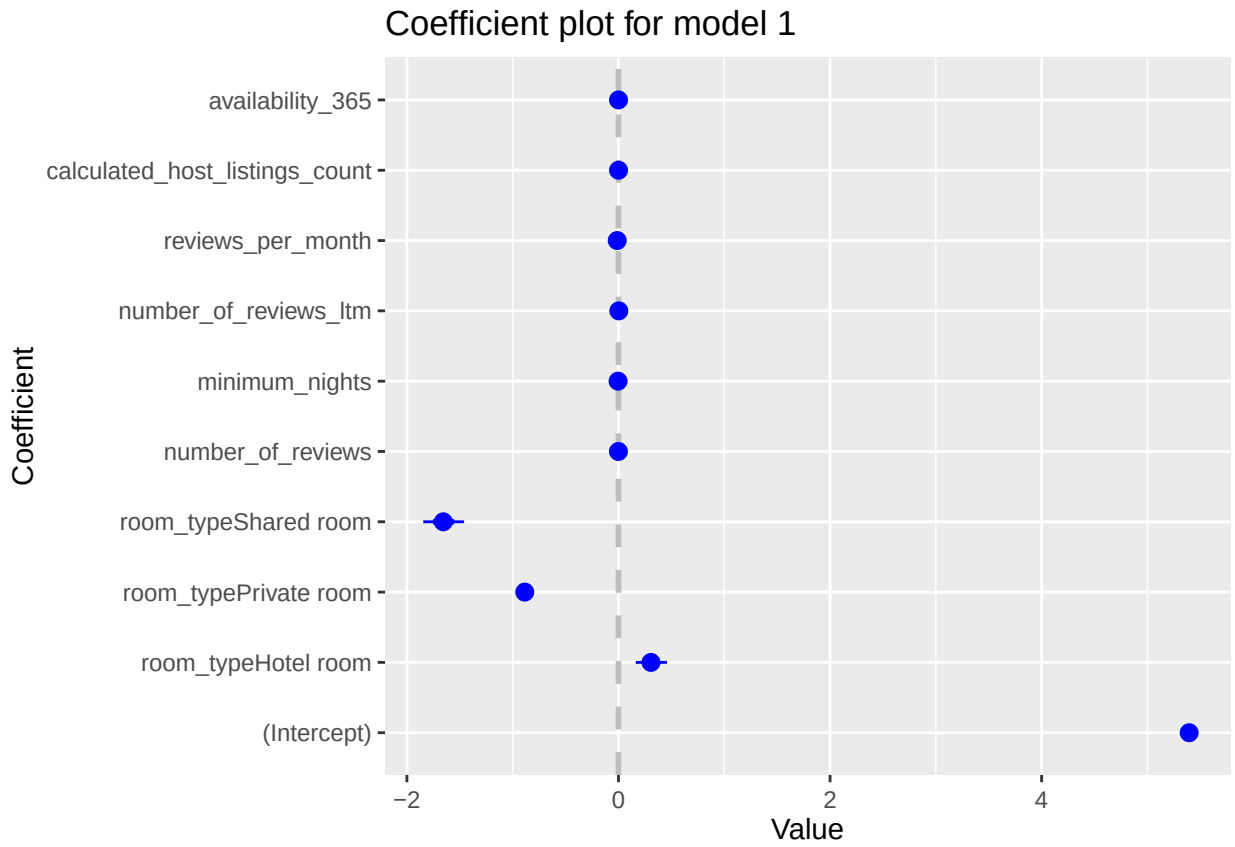
```
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5654 on 10947 degrees of freedom
## Multiple R-squared:  0.4212, Adjusted R-squared:  0.4207
## F-statistic: 885.1 on 9 and 10947 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
10
5
0
−5
5818
−4
−2
0
2
4
Theoretical Quantiles
lm(logprice ~ room_type + number_of_reviews + minimum_nights + number_of_re ..


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
3
4
5
6
Fitted values
lm(logprice ~ room_type + number_of_reviews + minimum_nights + number_of_re ..

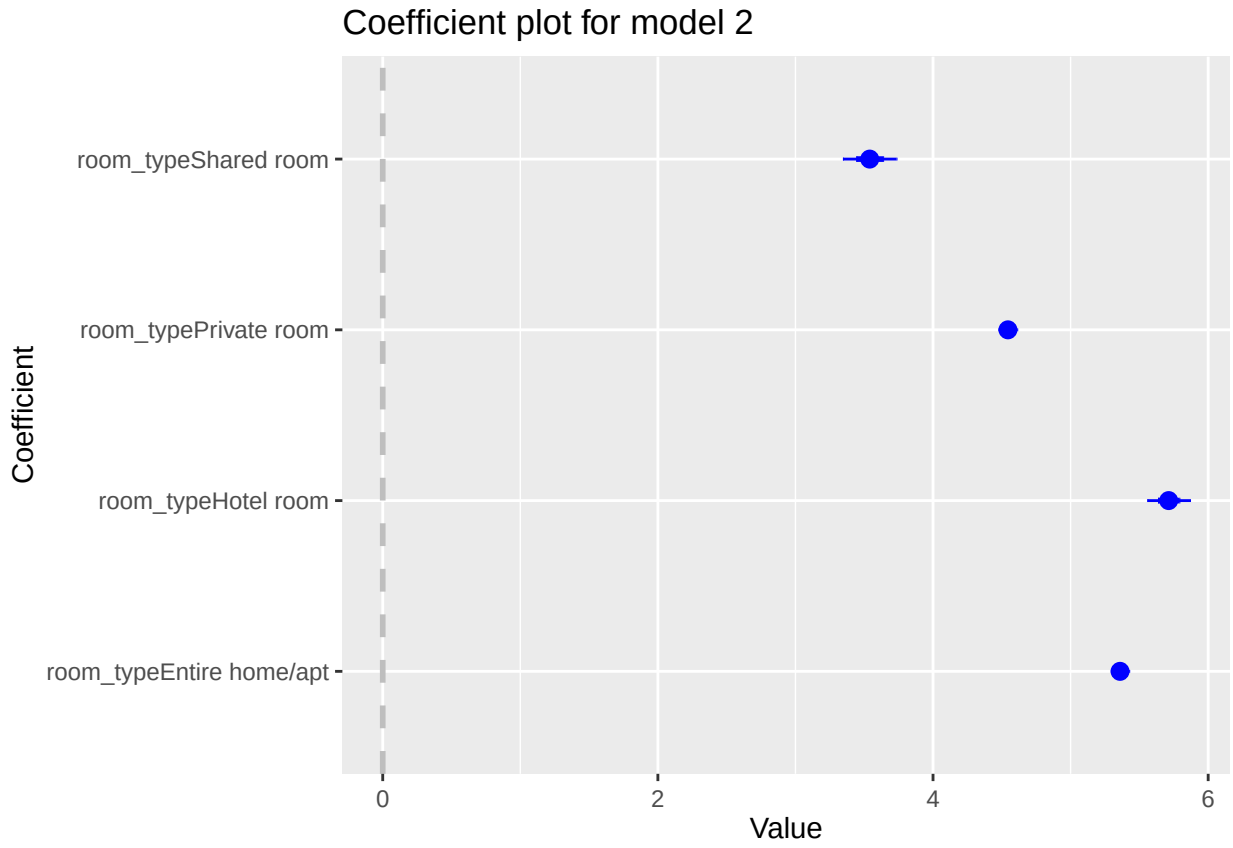

Residuals vs Leverage
Standardized residuals
10
5
0
−5
3239
534
5328
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage
lm(logprice ~ room_type + number_of_reviews + minimum_nights + number_of_re ..



* We can see that R-square is 0.42 so it does not will fitted. We can see from residual graph, there're some dots are having big residuals. The rest of points are symmetric distributed around the line h=0. We can see from the QQ plot, the model overestimate the low values and underestimate the high value. The coefficient plot tell us that coefficients "availability_365","calculated_host_listings_count","reviews_per_month","number_of_reviews_ltm","
are fall on zero point. So we need to eliminate those values in the mutilevel regression model.

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: logprice ~ room_type + (1 | neighbourhood) - 1
##    Data: price_t
##
## REML criterion at convergence: 18350.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.6840 -0.6762 -0.0843  0.5374  8.8967
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  neighbourhood (Intercept) 0.03853  0.1963
##  Residual                  0.30868  0.5556
## Number of obs: 10957, groups:  neighbourhood, 39
##
## Fixed effects:
##                            Estimate Std. Error        df t value Pr(>|t|)
## room_typeEntire home/apt 5.360e+00  3.324e-02 4.006e+01  161.26   <2e-16 ***
## room_typeHotel room      5.713e+00  7.751e-02 1.096e+03   73.71   <2e-16 ***
```

```
## room_typePrivate room    4.545e+00  3.378e-02 4.265e+01  134.55   <2e-16 ***
## room_typeShared room     3.540e+00  9.734e-02 2.412e+03   36.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             rm_Eh/ rm_tHr rm_tPr
## rm_typHtlrm 0.418
## rm_typPrvtr 0.935  0.402
## rm_typShrdr 0.325  0.138  0.322
```



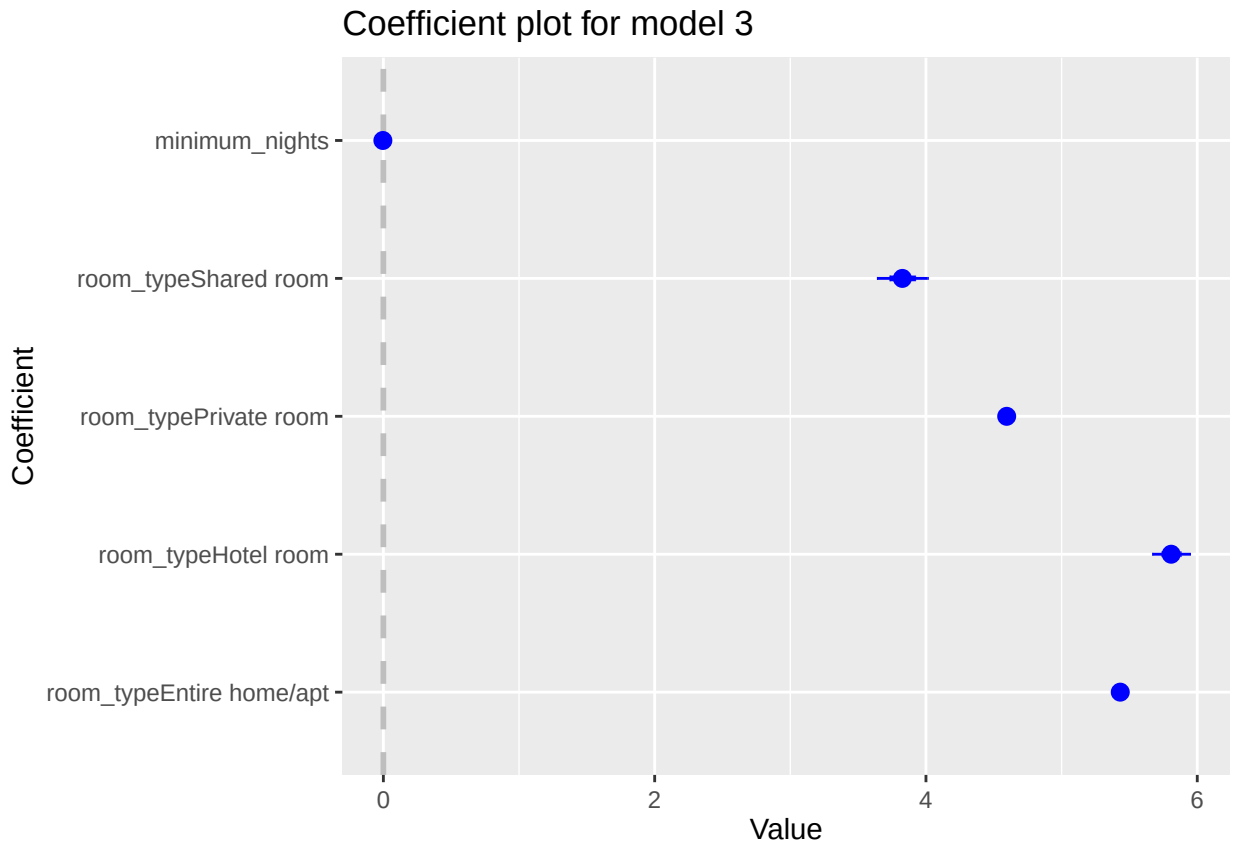
* Afer we get rid of non-significant terms, we find out room type have the most obvious correlation with price.

- Multilevel linear model with random slope

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: logprice ~ room_type + minimum_nights + (0 + minimum_nights |
##     neighbourhood) - 1
##    Data: price_t
##
## REML criterion at convergence: 18356.7
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.0126 -0.7064 -0.0872  0.5226  8.5828
##
## Random effects:
##  Groups        Name           Variance  Std.Dev.
##  neighbourhood minimum_nights 9.596e-06 0.003098
##  Residual                     3.092e-01 0.556063
## Number of obs: 10957, groups:  neighbourhood, 39
##
## Fixed effects:
##                             Estimate Std. Error         df t value Pr(>|t|)
## room_typeEntire home/apt  5.432e+00  7.607e-03  1.095e+04 714.105  < 2e-16 ***
## room_typeHotel room       5.808e+00  6.951e-02  1.092e+04  83.552  < 2e-16 ***
## room_typePrivate room     4.595e+00  1.004e-02  1.095e+04 457.675  < 2e-16 ***
## room_typeShared room      3.826e+00  9.380e-02  1.095e+04  40.788  < 2e-16 ***
## minimum_nights           -3.518e-03  5.495e-04  3.839e+01  -6.402 1.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             rm_Eh/ rm_tHr rm_tPr rm_tSr
## rm_typHtlrm  0.001
## rm_typPrvtr  0.155  0.001
## rm_typShrdr  0.048  0.000  0.045
## minmm_nghts -0.133 -0.001 -0.112 -0.031
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```



* We can see from modle 3, room type still be the most obvious coefficients with price. Minimum nights still lays on the point zero.

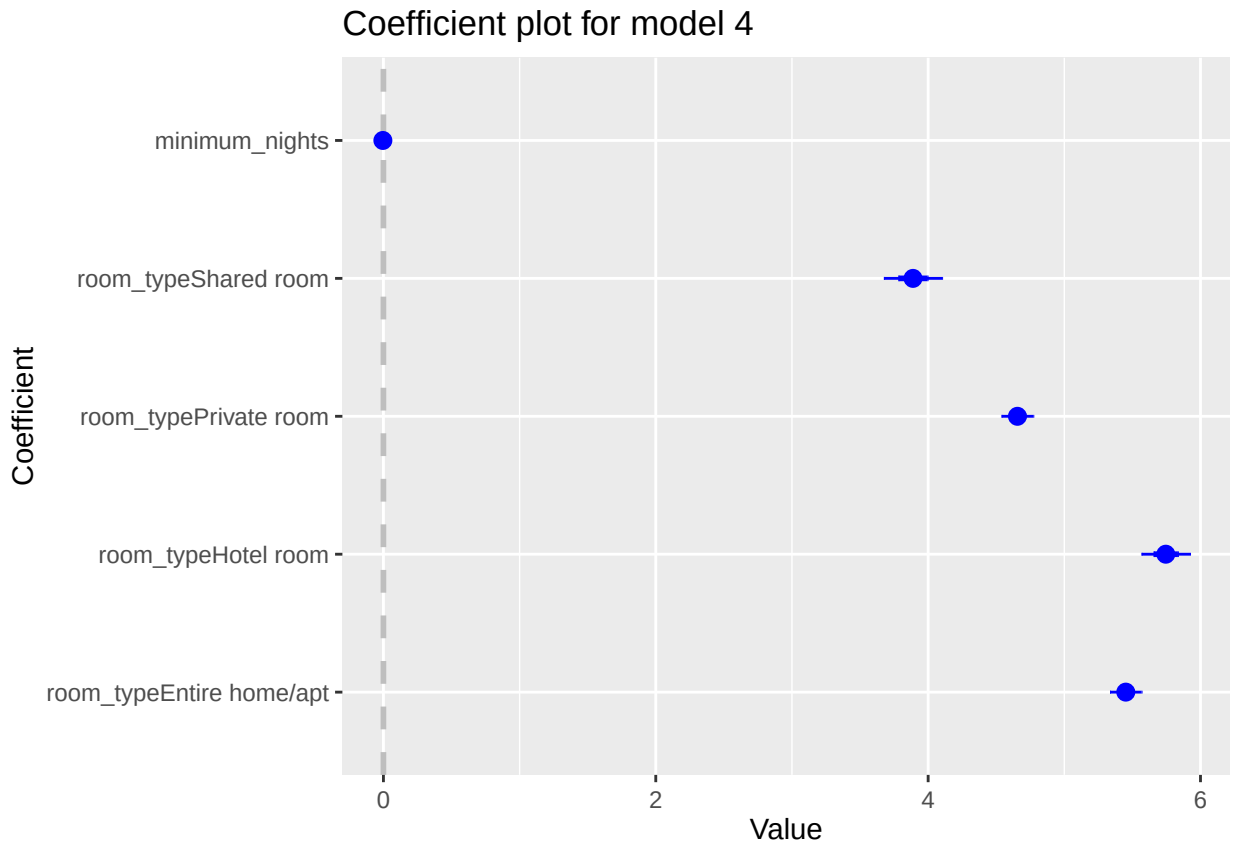
- Multilevel linear model with random slope and random intercept

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 23.7985 (tol = 0.002, component 1)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: logprice ~ room_type + minimum_nights + (1 + minimum_nights |
##     neighbourhood) - 1
##    Data: price_t
##
## REML criterion at convergence: 17565.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.0141 -0.6652 -0.1006  0.5201  9.0183
##
## Random effects:
##  Groups        Name           Variance  Std.Dev. Corr
##  neighbourhood (Intercept)    1.236e-01 0.351588
```

```
##                minimum_nights 8.147e-06 0.002854 -0.58
##  Residual                     2.842e-01 0.533082
## Number of obs: 10957, groups:  neighbourhood, 39
##
## Fixed effects:
##                             Estimate Std. Error         df t value Pr(>|t|)
## room_typeEntire home/apt  5.4513433  0.0578904  2.5898100  94.167 1.19e-05 ***
## room_typeHotel room       5.7452271  0.0889297 14.4311636  64.604  < 2e-16 ***
## room_typePrivate room     4.6558334  0.0582137  2.6459639  79.978 1.49e-05 ***
## room_typeShared room      3.8888916  0.1068219 30.0732998  36.405  < 2e-16 ***
## minimum_nights           -0.0040037  0.0005147 11.0024604  -7.779 8.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             rm_Eh/ rm_tHr rm_tPr rm_tSr
## rm_typHtlrm  0.645
## rm_typPrvtr  0.980  0.637
## rm_typShrdr  0.537  0.348  0.536
## minmm_nghts -0.553 -0.353 -0.551 -0.316
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 23.7985 (tol = 0.002, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

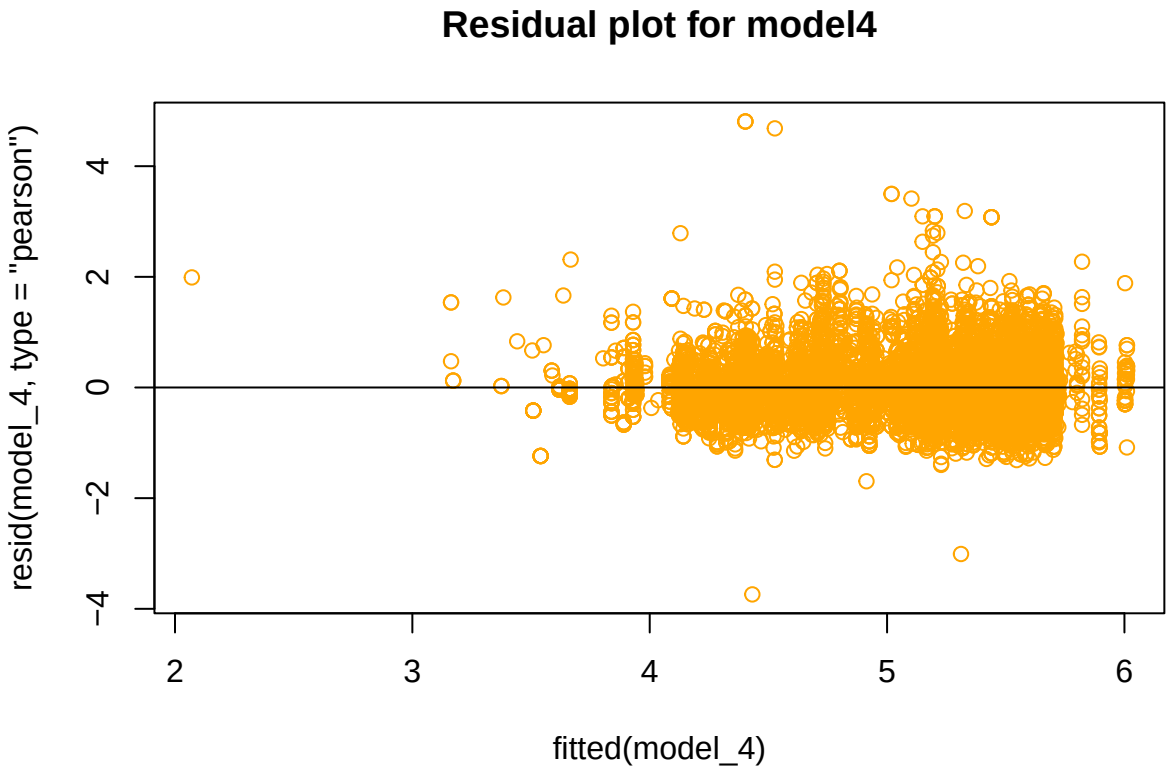


* For modle 4, room type still be the most significance predictors.

Result

- Model interpretation

```
## Data: price_t
## Models:
## model_2: logprice ~ room_type + (1 | neighbourhood) - 1
## model_3: logprice ~ room_type + minimum_nights + (0 + minimum_nights | neighbourhood) - 1
## model_4: logprice ~ room_type + minimum_nights + (1 + minimum_nights | neighbourhood) - 1
##         npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model_2    6 18363 18407 -9175.5    18351
## model_3    7 18371 18422 -9178.4    18357   0.00  1          1
## model_4    9 17583 17649 -8782.6    17565 791.53  2     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



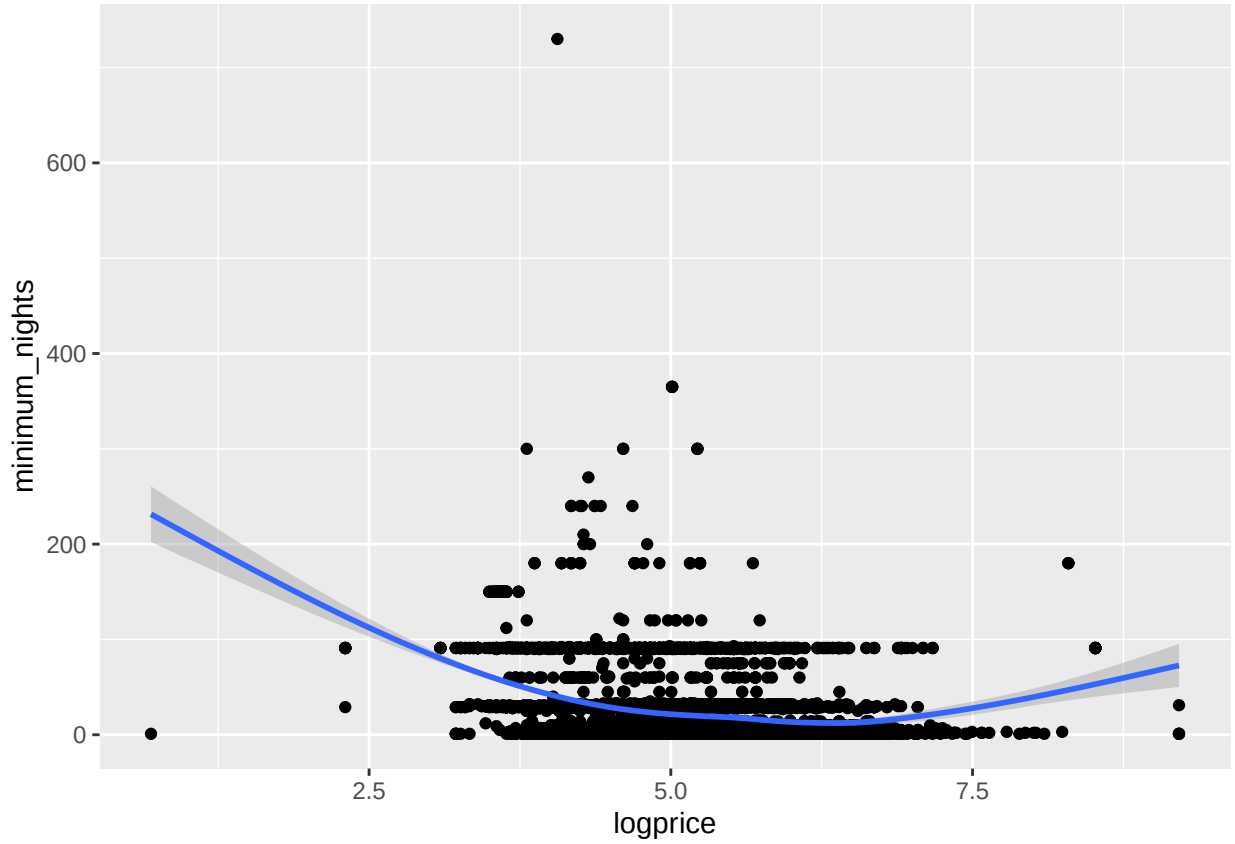
Disscussion

- We can see from the ANOVA table, three tables are highly different since p-value is less than 0.05. Since model 4 has the lowest deviance so model 4 is the best model among other three models. Residual plot is also spread symmetrically around line h=0

Appendix

- Visualize the relationship between logprice and minimum nights

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



* As we can see from the graph above, the graph does not suggest a linearly increasing relationship between the logprice and minimum nights. So the assumption that lower price may lead to more days to stay may not be a right assumption.

Supplement

- http://insideairbnb.com/get-the-data/
- https://quantdev.ssri.psu.edu/tutorials/r-bootcamp-introduction-multilevel-model-and-interactions