AIRBNB PRICING STRATEGY

Linear Regression Analysis for Airbnb Price in Seattle

By Bodi Zhang (25%), Jingwei Yao (25%),

Qihan Liu (25%), Zehua Ye (25%)

Table of Contents

1 Introduction

Standing out by its affordability and uniqueness, Airbnb has already pulled more than 1 billion guests from traditional hotels to shared accommodations since its founding in 2007. Over 4 million people joined Airbnb as hosts using their vacant space to earn supplementary income and the number continues growing. For these homeowners, pricing is one of the essential decision-making when they start their Airbnb business while it is also significant for Airbnb giving pricing suggestions to be a supporter.

By analyzing 3818 detailed listing data of Airbnb in Seattle, including various attributes such as location, room type, amenities, review score, etc. we aim to get a better understanding of how these factors influence the price of stays and try to predict the optimal price for hosts based on main attributors.

2 Data Summary

2.1 Data Overview

The raw data contains 3818 records and 92 columns. Not all of them are applicable. After cleaning out the missing values and invalid data, 21 columns and 3629 records are selected and retained. One of our selected factors is amenities which including 42 features. To get better understanding of how amenities attribute to listing price, we selected 8 intuitive predictors of amenities including “TV”, “Parking”, “AC”, “Checkin_24hour”, “Pets Allowed”, “Gym”, “Pets live” and “Kid Friendly”.

2.2 Attributes Description & Summary Table

<table>
<tr><th colspan="3">Variables Summary Table</th></tr>
<tr><th>NAME</th><th>CATEGORY</th><th>EXPLANATION</th></tr>
<tr><td>price</td><td>Numerical</td><td>The nightly rate for reservation set by the host.</td></tr>
<tr><td>host_response</td><td>Numerical</td><td>The percentage of inquiries and reservation requests host response to.</td></tr>
<tr><td>bathrooms</td><td>Numerical</td><td>Number of bathrooms the property provides.</td></tr>
<tr><td>bedrooms</td><td>Numerical</td><td>Number of bedrooms the property provides.</td></tr>
<tr><td>beds</td><td>Numerical</td><td>Number of beds the property provides.</td></tr>
</table>

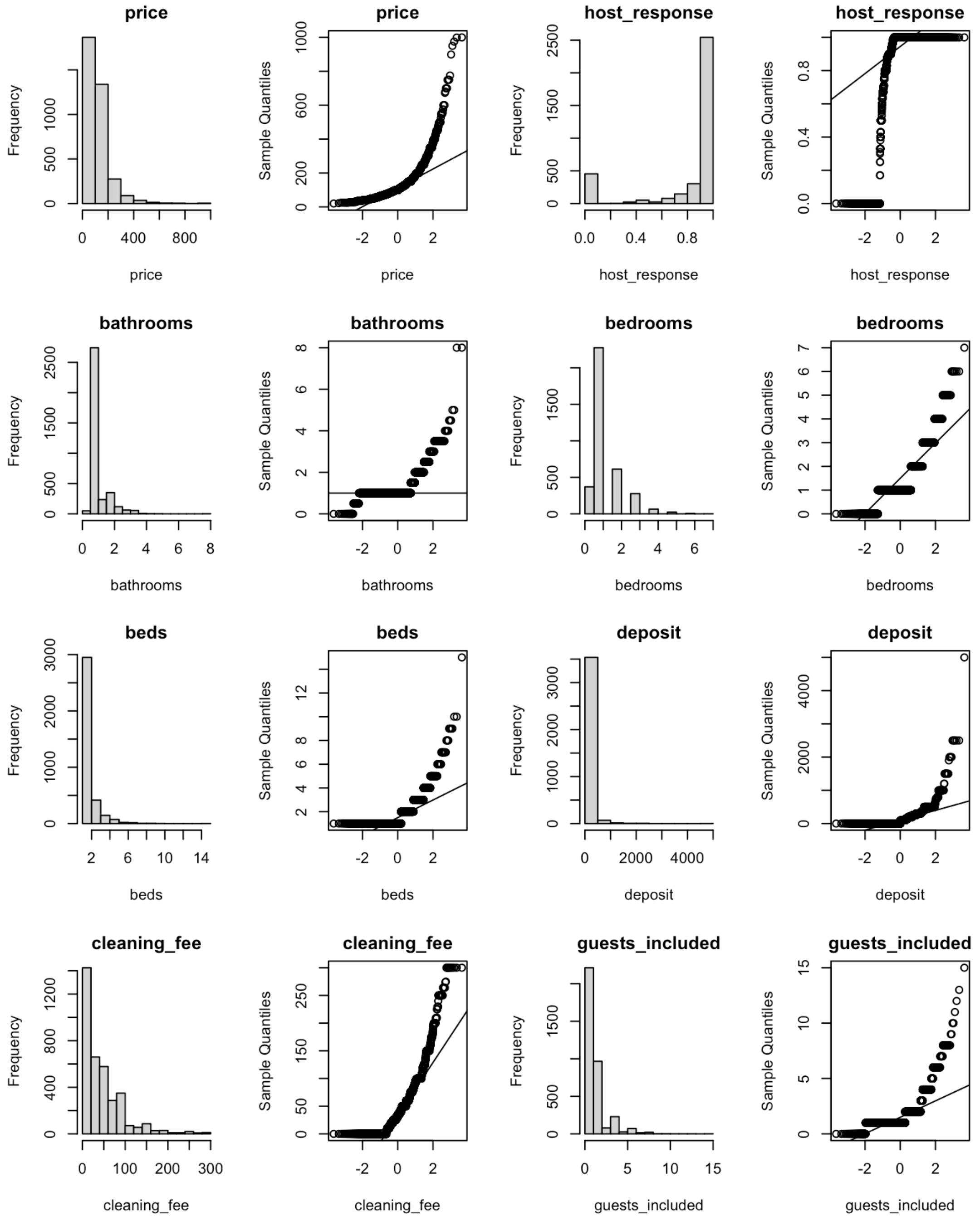
deposit	Numerical	Authorization holds that Airbnb places on a guest’s payment method. It can be required by the host.
cleaning_fee	Numerical	A one-off cleaning charge for guests and is set by the host.
guests_included	Numerical	Number of guests the host.
min_nights	Numerical	Minimum night requirements for reservation set by the host.
availability_90	Numerical	Available days for property of latest 90 days.
review_num	Numerical	Number of reviews for property received from guests.
review_score	Numerical	Rating scores from guests based on their experience. Full score of 100 points.
instant_bookable	Numerical	Instant book listings allow guests to book immediately without needing to send a request to the host for approval. It is set by the host.
superhost	Categorical	The badge showing in listing and profile of host identifies the host is welcoming and experienced. It is awarded by Airbnb.
TV	Categorical	The property provides TV.
Parking	Categorical	The property provides free parking to guests during their stays.
AC	Categorical	The property provides air conditioning.

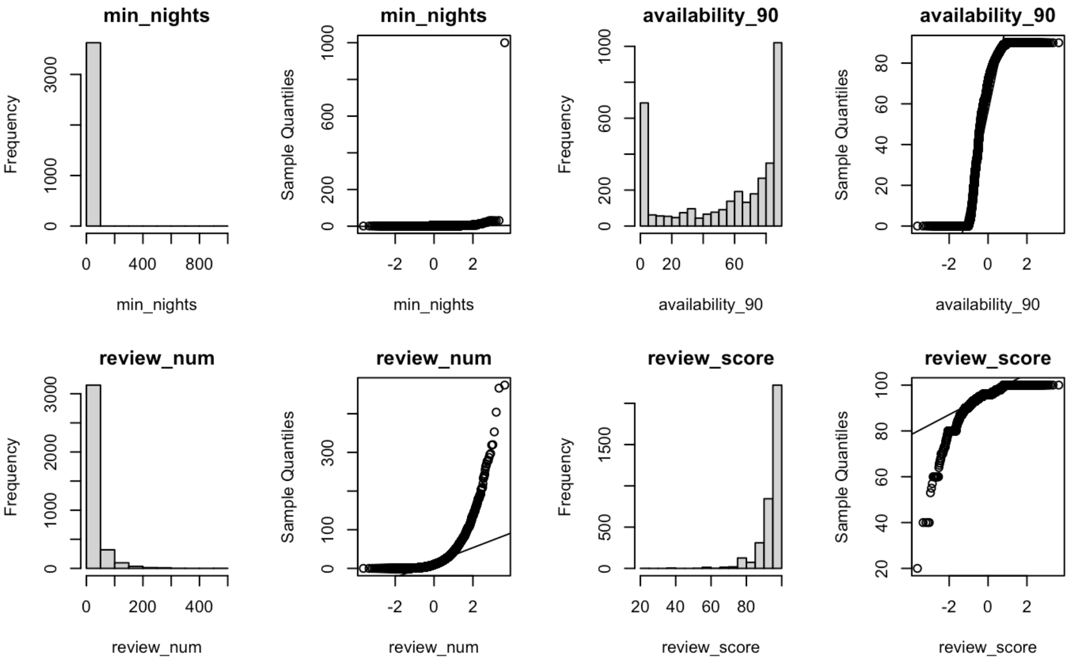
Checkin_24hour	Categorical	Guests can conveniently gain access in anytime.
Pets_allowed	Categorical	Pets are allowed to stay with guests in an Airbnb.
Gym	Categorical	The property provides gym.
Pets_live	Categorical	The property has pets in it.
Kid_friendly	Categorical	Welcome family or kids.
cancel_policy	Categorical	Cancellation Policy: Flexible, moderate, or strict
neighbourhood	Categorical	Locations of properties
property_type	Categorical	Type of property such as apartments, houses, lofts and others
room_type	Categorical	Entire room/apartment, private room, or shared room.

3 Pre-analysis

Before starting to build a model to solve this problem, we need to have a holistic view of the data distribution and make a preliminary assumption of the relationships between predictors and response variables.

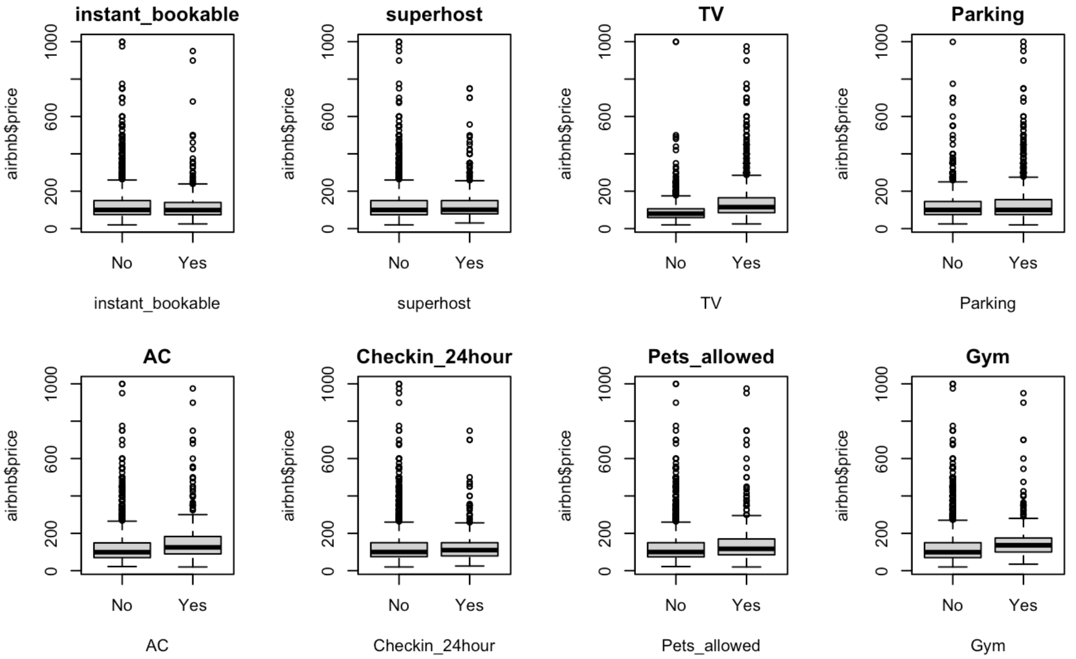
3.1 Histograms and Q-Q Plots

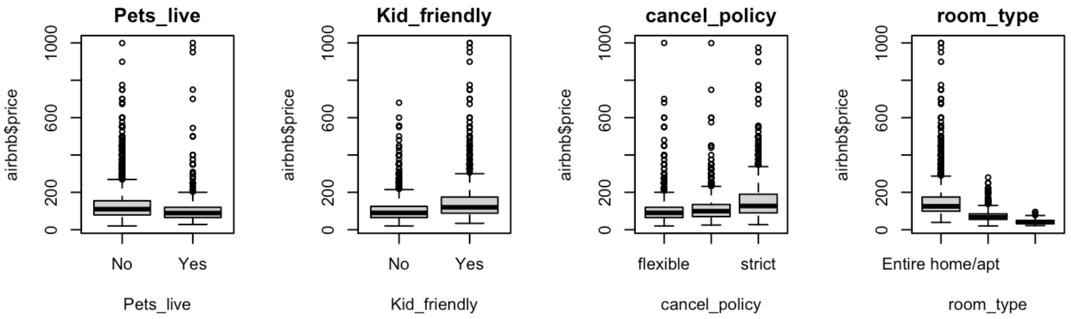


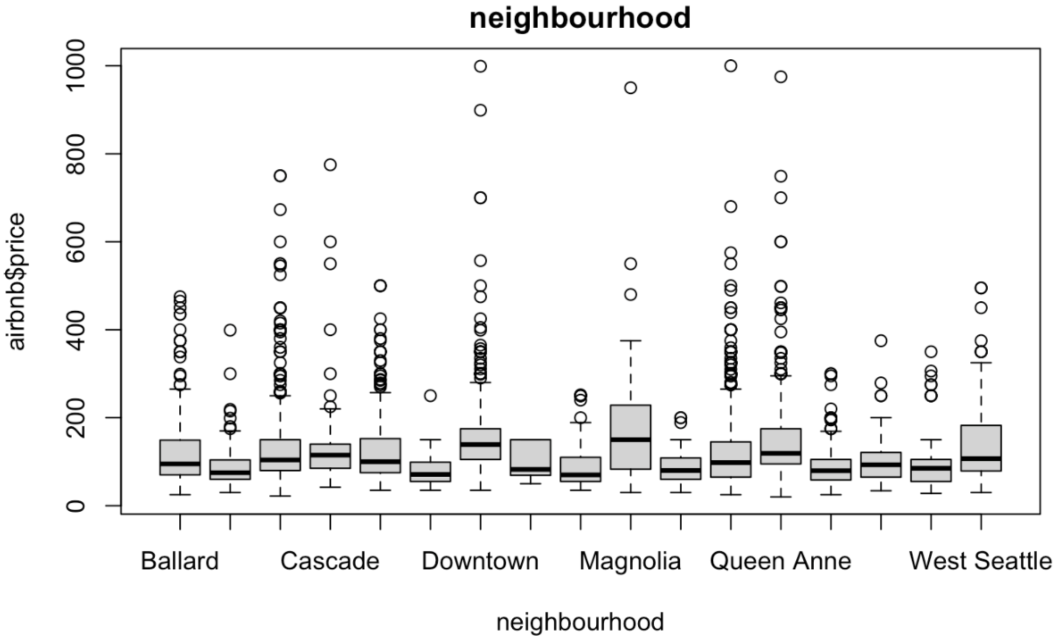


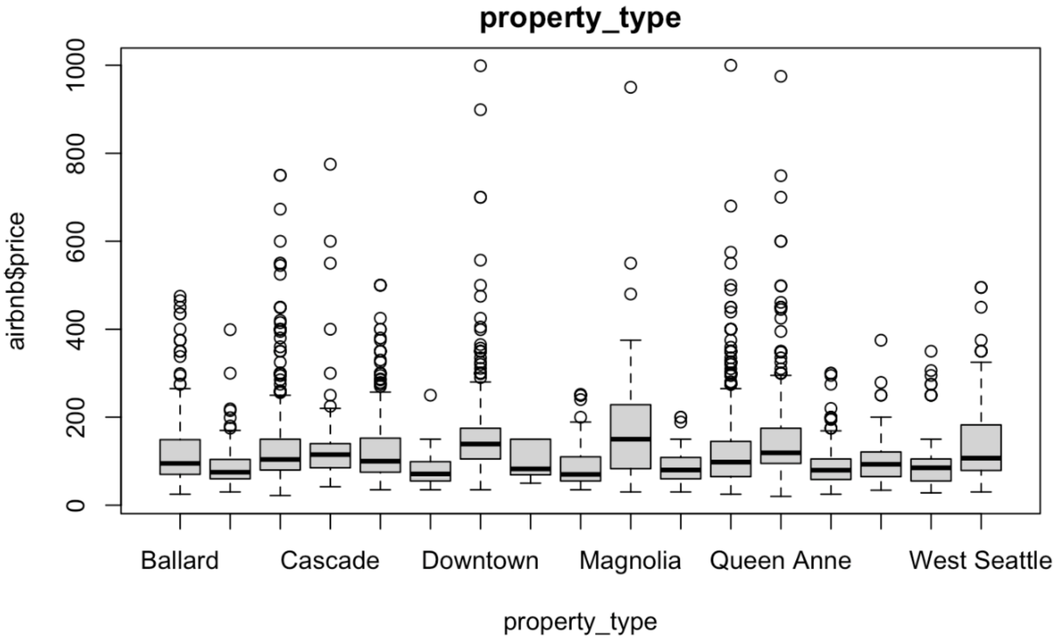
It’s apparent that all variables are in skewed distribution and for several predictors like ***min_nights*** and ***deposit***, there are obvious outliers, which need to be removed from our dataset later to reduce the bad impact on our model’s accuracy.

3.2Boxplots of Categorical Predictors



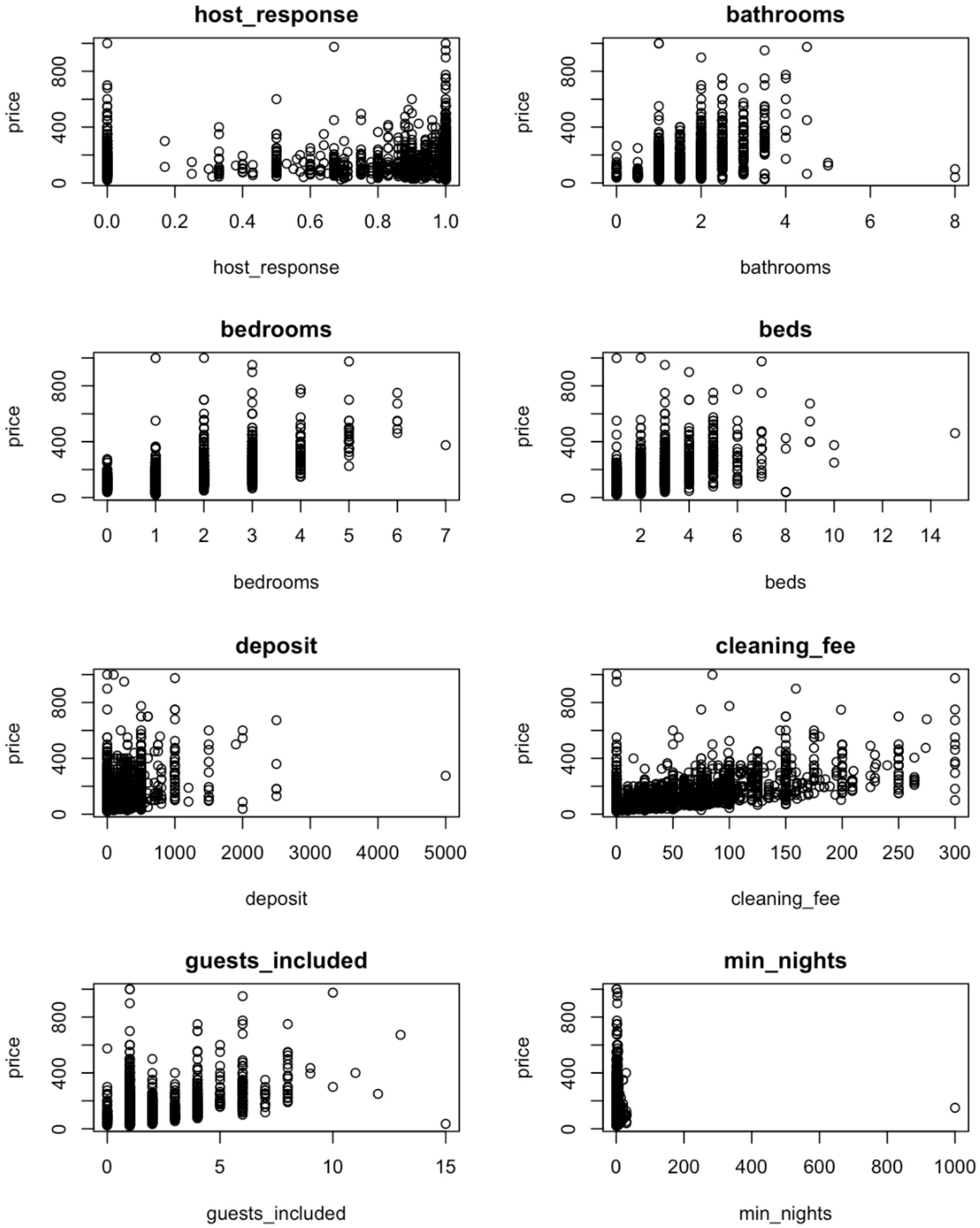

Pets_live
airbnb$price
0
200
600
1000
No
Yes
Pets_live
Kid_friendly
airbnb$price
0
200
600
1000
No
Yes
Kid_friendly
cancel_policy
airbnb$price
0
200
600
1000
flexible
strict
cancel_policy
room_type
airbnb$price
0
200
600
1000
Entire home/apt
room_type

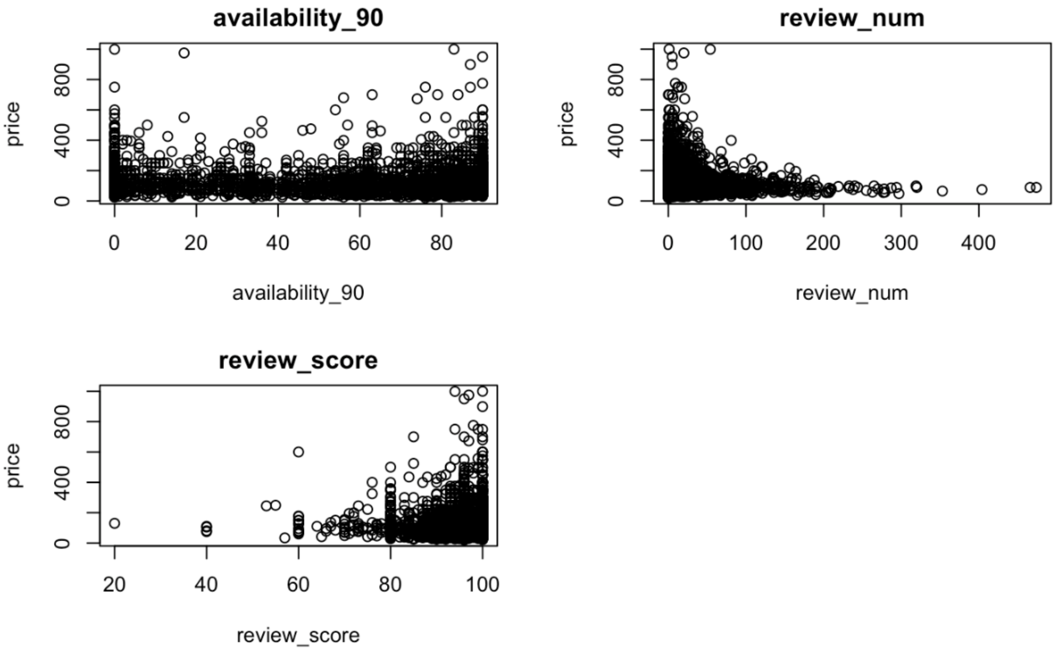

neighbourhood
airbnb$price
0
200
400
600
800
1000
Ballard
Cascade
Downtown
Magnolia
Queen Anne
West Seattle
neighbourhood


property_type
airbnb$price
0
200
400
600
800
1000
Ballard
Cascade
Downtown
Magnolia
Queen Anne
West Seattle
property_type

From these boxplots, we could find that for ***TV***, ***AC***, ***Kid_friendily***, ***cancel_policy***, ***room_type***, ***neighbourhood*** and ***property_type***, there are significant price difference within these predictors. Therefore, we believe that these predictors would influence a lot on our subsequent model.

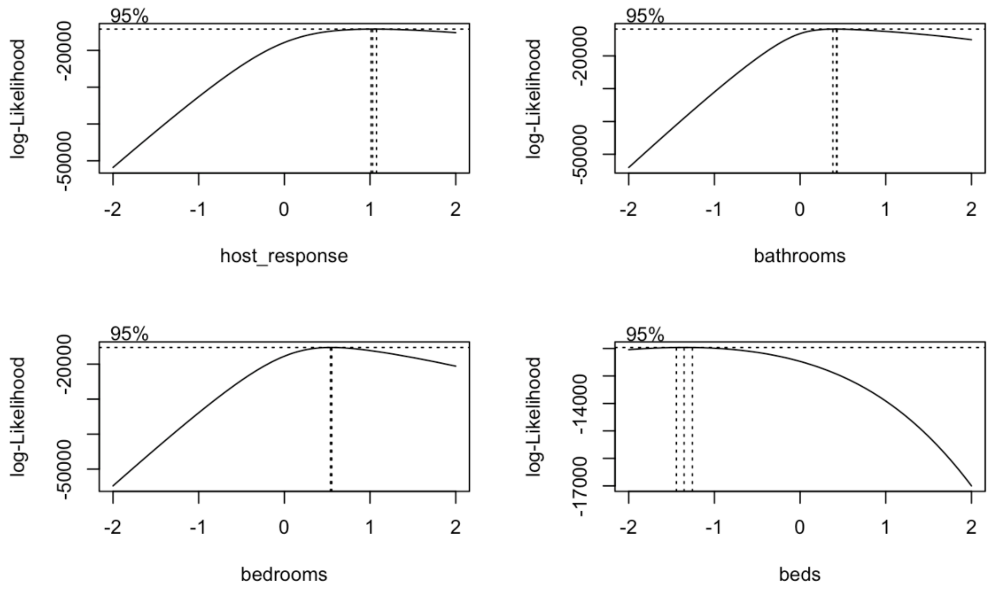
3.3 Scatterplots

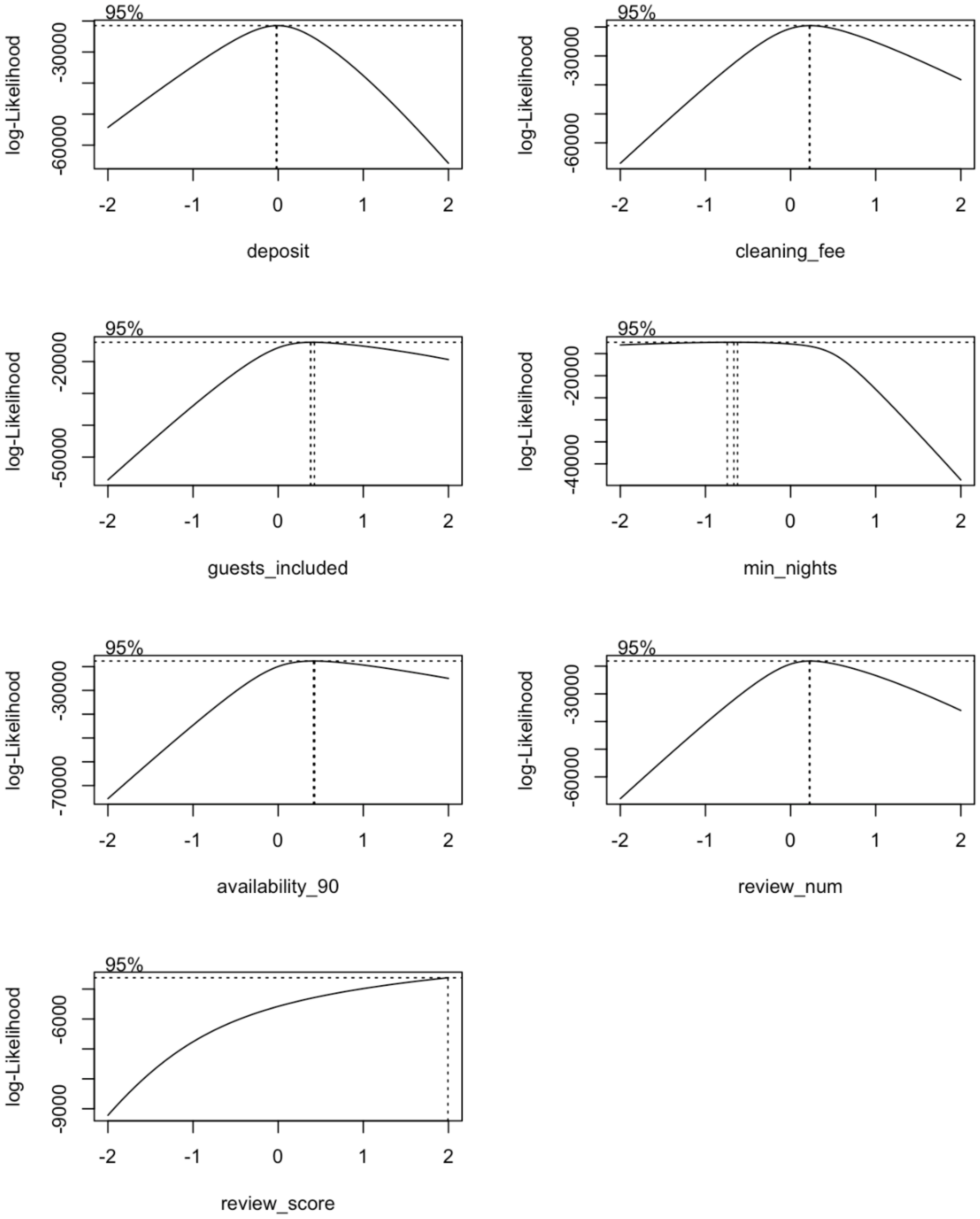




It’s a pity that only ***cleaning_fee*** appears to have a close linear relationship with price, which means we are supposed to do some box-cox transformation of the response variable and these continuous predictors to improve the model.

3.4 Box-Cox Transformation of Numerical Predictors



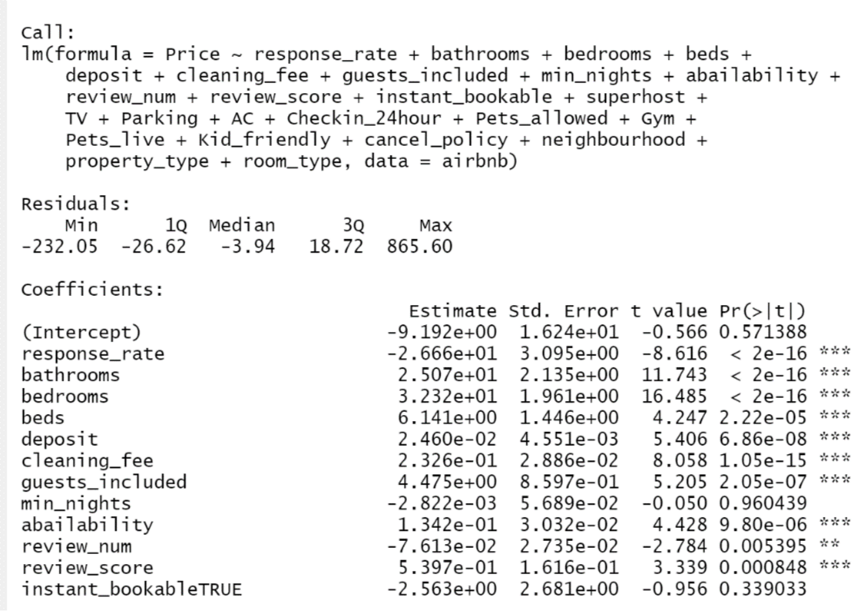

95%
log-Likelihood
-60000
-30000
-2
-1
0
1
2
deposit
95%
log-Likelihood
-60000
-30000
-2
-1
0
1
2
cleaning_fee
95%
log-Likelihood
-50000
-20000
-2
-1
0
1
2
guests_included
95%
log-Likelihood
-40000
-20000
-2
-1
0
1
2
min_nights
95%
log-Likelihood
-70000
-30000
-2
-1
0
1
2
availability_90
95%
log-Likelihood
-60000
-30000
-2
-1
0
1
2
review_num
95%
log-Likelihood
-9000
-6000
-2
-1
0
1
2
review_score

4 Model Building

4.1 Full Model

$$\begin{aligned} Price = \beta_0 &+ \beta_1 response\_rate + \beta_2 bathrooms + \beta_3 bedrooms + \beta_4 beds + \beta_5 deposit \\ &+ \beta_6 cleaning\_fee + \beta_7 guests\_included + \beta_8 min\_nights + \beta_9 availability \\ &+ \beta_{10} review\_num + \beta_{11} review\_score + \beta_{12} instant\_bookable \\ &+ \beta_{13} superhost + \beta_{14} TV + \beta_{15} Parking + \beta_{16} AC + \beta_{17} 24\_Check\_in \\ &+ \beta_{18} Pet\_allowed + \beta_{19} Gym + \beta_{20} Pet\_live + \beta_{21} Kid\_Friendly \\ &+ \beta_{22} cancel\_policy + \beta_{23} neighbourhood + \beta_{24} property\_type \\ &+ \beta_{25} room\_type + \varepsilon_t \end{aligned}$$

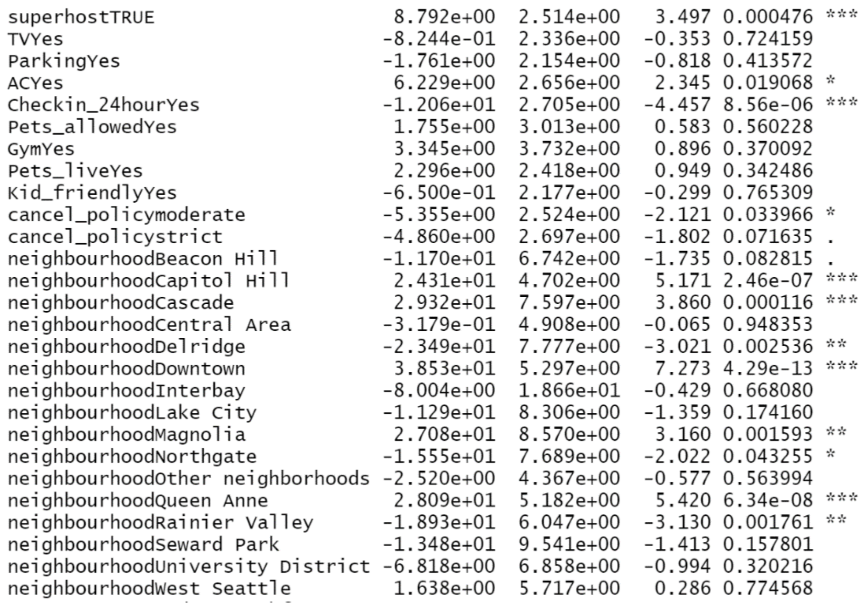
4.1.1 Summary of Full Model



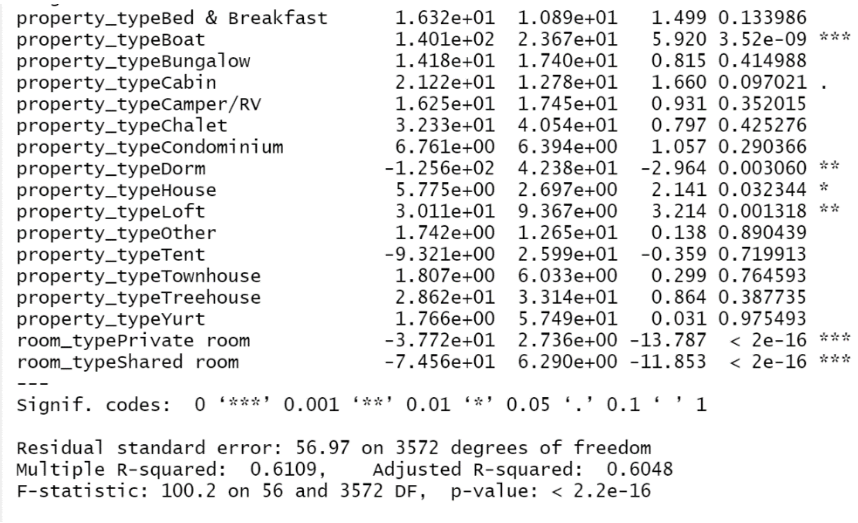
```
Call:
lm(formula = Price ~ response_rate + bathrooms + bedrooms + beds +
    deposit + cleaning_fee + guests_included + min_nights + abailability +
    review_num + review_score + instant_bookable + superhost +
    TV + Parking + AC + Checkin_24hour + Pets_allowed + Gym +
    Pets_live + Kid_friendly + cancel_policy + neighbourhood +
    property_type + room_type, data = airbnb)

Residuals:
    Min      1Q  Median      3Q     Max
-232.05  -26.62   -3.94   18.72  865.60

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -9.192e+00  1.624e+01  -0.566 0.571388
response_rate                   -2.666e+01  3.095e+00  -8.616  < 2e-16 ***
bathrooms                        2.507e+01  2.135e+00  11.743  < 2e-16 ***
bedrooms                         3.232e+01  1.961e+00  16.485  < 2e-16 ***
beds                             6.141e+00  1.446e+00   4.247 2.22e-05 ***
deposit                          2.460e-02  4.551e-03   5.406 6.86e-08 ***
cleaning_fee                     2.326e-01  2.886e-02   8.058 1.05e-15 ***
guests_included                  4.475e+00  8.597e-01   5.205 2.05e-07 ***
min_nights                      -2.822e-03  5.689e-02  -0.050 0.960439
abailability                     1.342e-01  3.032e-02   4.428 9.80e-06 ***
review_num                      -7.613e-02  2.735e-02  -2.784 0.005395 **
review_score                     5.397e-01  1.616e-01   3.339 0.000848 ***
instant_bookableTRUE            -2.563e+00  2.681e+00  -0.956 0.339033
```



```
superhostTRUE                     8.792e+00  2.514e+00   3.497 0.000476 ***
TVYes                            -8.244e-01  2.336e+00  -0.353 0.724159
ParkingYes                       -1.761e+00  2.154e+00  -0.818 0.413572
ACYes                             6.229e+00  2.656e+00   2.345 0.019068 *
Checkin_24hourYes                -1.206e+01  2.705e+00  -4.457 8.56e-06 ***
Pets_allowedYes                   1.755e+00  3.013e+00   0.583 0.560228
GymYes                            3.345e+00  3.732e+00   0.896 0.370092
Pets_liveYes                      2.296e+00  2.418e+00   0.949 0.342486
Kid_friendlyYes                  -6.500e-01  2.177e+00  -0.299 0.765309
cancel_policymoderate            -5.355e+00  2.524e+00  -2.121 0.033966 *
cancel_policystrict              -4.860e+00  2.697e+00  -1.802 0.071635 .
neighbourhoodBeacon Hill         -1.170e+01  6.742e+00  -1.735 0.082815 .
neighbourhoodCapitol Hill         2.431e+01  4.702e+00   5.171 2.46e-07 ***
neighbourhoodCascade              2.932e+01  7.597e+00   3.860 0.000116 ***
neighbourhoodCentral Area        -3.179e-01  4.908e+00  -0.065 0.948353
neighbourhoodDelridge            -2.349e+01  7.777e+00  -3.021 0.002536 **
neighbourhoodDowntown             3.853e+01  5.297e+00   7.273 4.29e-13 ***
neighbourhoodInterbay            -8.004e+00  1.866e+01  -0.429 0.668080
neighbourhoodLake City           -1.129e+01  8.306e+00  -1.359 0.174160
neighbourhoodMagnolia             2.708e+01  8.570e+00   3.160 0.001593 **
neighbourhoodNorthgate           -1.555e+01  7.689e+00  -2.022 0.043255 *
neighbourhoodOther neighborhoods -2.520e+00  4.367e+00  -0.577 0.563994
neighbourhoodQueen Anne           2.809e+01  5.182e+00   5.420 6.34e-08 ***
neighbourhoodRainier Valley      -1.893e+01  6.047e+00  -3.130 0.001761 **
neighbourhoodSeward Park         -1.348e+01  9.541e+00  -1.413 0.157801
neighbourhoodUniversity District -6.818e+00  6.858e+00  -0.994 0.320216
neighbourhoodWest Seattle         1.638e+00  5.717e+00   0.286 0.774568
```

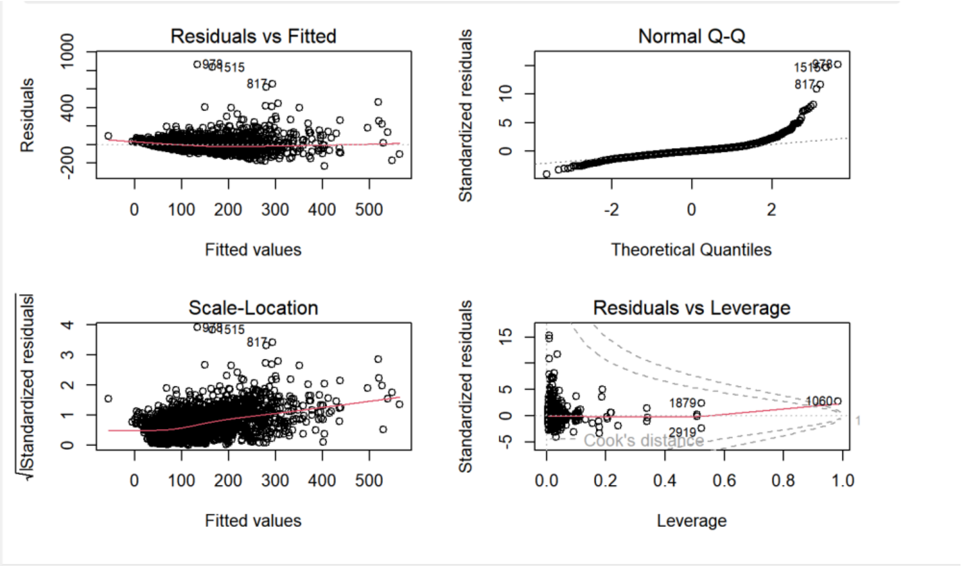


```
property_typeBed & Breakfast      1.632e+01  1.089e+01   1.499 0.133986    
property_typeBoat                 1.401e+02  2.367e+01   5.920 3.52e-09 ***
property_typeBungalow             1.418e+01  1.740e+01   0.815 0.414988    
property_typeCabin                2.122e+01  1.278e+01   1.660 0.097021 .  
property_typeCamper/RV            1.625e+01  1.745e+01   0.931 0.352015    
property_typeChalet               3.233e+01  4.054e+01   0.797 0.425276    
property_typeCondominium          6.761e+00  6.394e+00   1.057 0.290366    
property_typeDorm                -1.256e+02  4.238e+01  -2.964 0.003060 ** 
property_typeHouse                5.775e+00  2.697e+00   2.141 0.032344 *  
property_typeLoft                 3.011e+01  9.367e+00   3.214 0.001318 ** 
property_typeOther                1.742e+00  1.265e+01   0.138 0.890439    
property_typeTent                -9.321e+00  2.599e+01  -0.359 0.719913    
property_typeTownhouse            1.807e+00  6.033e+00   0.299 0.764593    
property_typeTreehouse            2.862e+01  3.314e+01   0.864 0.387735    
property_typeYurt                 1.766e+00  5.749e+01   0.031 0.975493    
room_typePrivate room            -3.772e+01  2.736e+00 -13.787  < 2e-16 ***
room_typeShared room             -7.456e+01  6.290e+00 -11.853  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56.97 on 3572 degrees of freedom
Multiple R-squared:  0.6109,    Adjusted R-squared:  0.6048 
F-statistic: 100.2 on 56 and 3572 DF,  p-value: < 2.2e-16
```

From the summary, we can see our full model has 25 predictors. Value of R Square and adjusted R Square is 0.6109 and 0.6048 respectively, which is not bad for its debut. But depending on so many predictors the model concludes, it is not the result we expect. And there are many predictors that are insignificant, most related to “neighborhood”, “property type” and “amenities”.

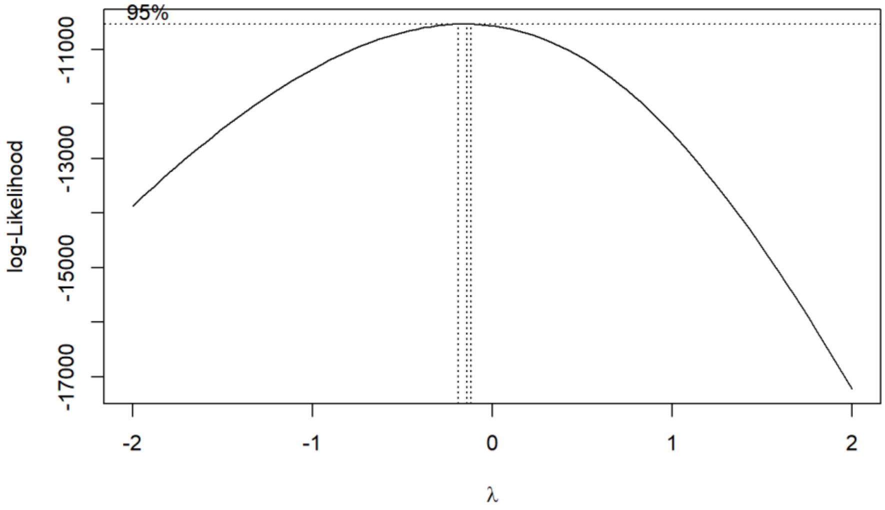
4.1.2 Plots of Full Model



Above are plots of the full model, we can see the residual plot is not random, the QQ-plot is not close to a straight line, suggesting a transformation of the response variable in demand. The leverage plot also shows some potential influential points (1060), which need to pay attention to further.

4.2 Transformation of Model

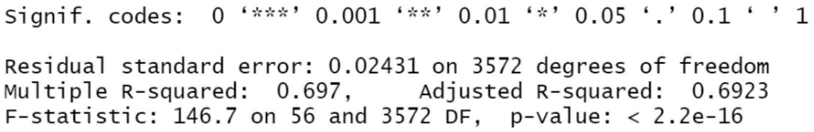
4.2.1 Box-cox transformation for response variable (Price)



In the end of Pre-analysis, we showed the box-cox plots of our numerical predictors. After calculating the likelihood value of each variable, we have a preliminary plan of the box-cox transformation afterwards, as shown in the following table.

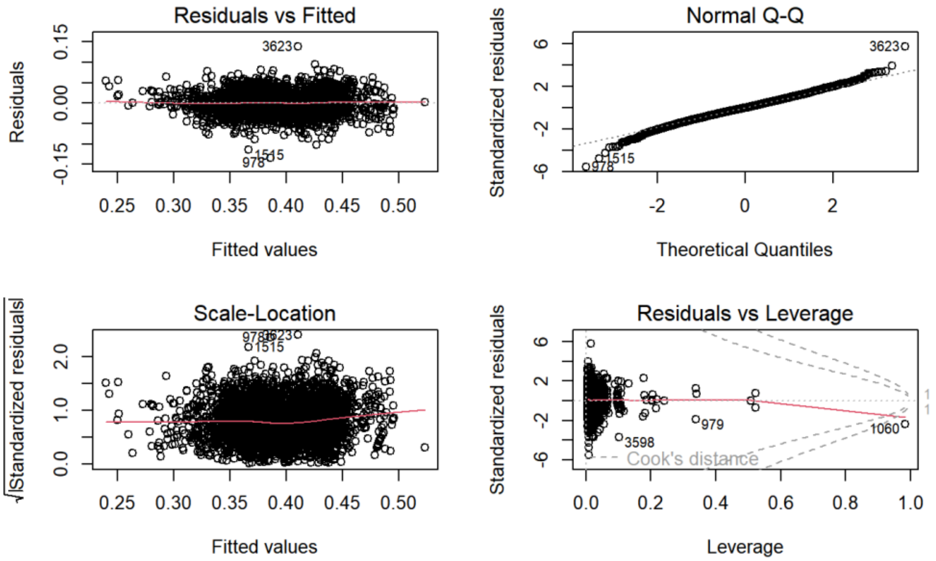
Variable name	Transformation type	Lambda value
Price (response variable)	t	-0.2
deposit	log	0
cleaning_fee	t	0.2
review_num	t	0.2
review_score	t	2
availabity_90	t	0.4

4.2.2 Summary and Plots after Transformation



```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02431 on 3572 degrees of freedom
Multiple R-squared:  0.697,	Adjusted R-squared:  0.6923 
F-statistic: 146.7 on 56 and 3572 DF,  p-value: < 2.2e-16
```



As a result of the transformation, the new model has a better R Square, increasing from 0.6048 to 0.6923, and more predictors become significant. The performance of the new model is also improved on residual plots and QQ plots. Even though the residual plot has not achieved the randomness we expect, it is better and QQ_plot has moved closer to the straight line.

5 Model Selection

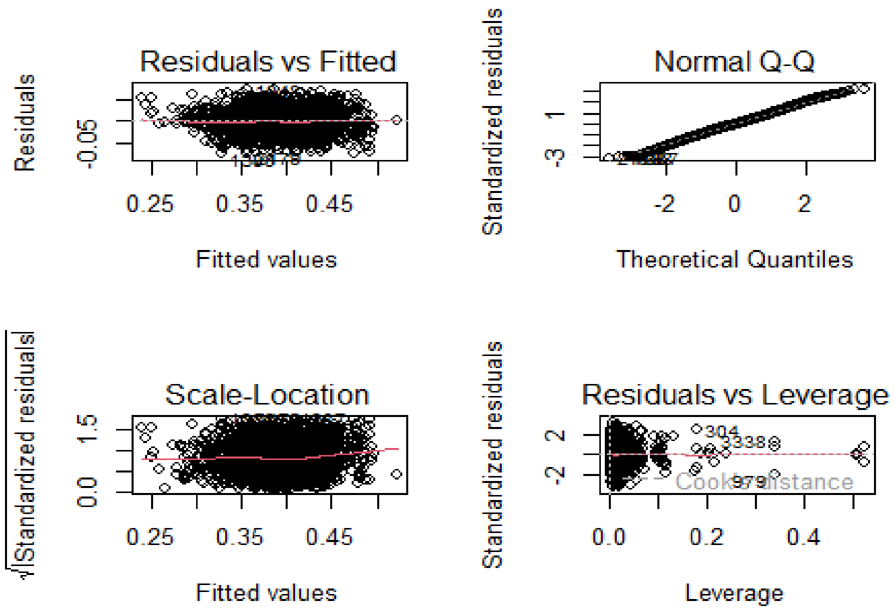
After transferring response variable and predictors, we use studentized residuals to test if there are potential outlier points influencing the accuracy in our model. And we find that there are 27 variables whose studentized residuals are greater than 3.

By reviewing the outlier variables and data summary, we find that except for outliers found by studentized residual, there is also an influencer in predictor, minimum_nights, with the maximum value 1000. Therefore, after removing all outliers, we can get the summary of lmod3.new as below.

```
lmod3.new = lm(I((price)^(-1/5))~host_response+log.deposit+t.cleaning_fee+t.availabilit
y_90+t.review_num+t.review_score+guests_included+min_nights+bathrooms+bedrooms+beds+ins
tant_bookable+TV+Parking+AC+Checkin_24hour+Pets_allowed+Gym+Pets_live+Kid_friendly+supe
rhost+property_type+cancel_policy+room_type+neighbourhood, data=airbnb)
summary(lmod3.new)

##
## Call:
## lm(formula = I((price)^(-1/5)) ~ host_response + log.deposit +
##     t.cleaning_fee + t.availability_90 + t.review_num + t.review_score +
##     guests_included + min_nights + bathrooms + bedrooms + beds +
##     instant_bookable + TV + Parking + AC + Checkin_24hour + Pets_allowed +
##     Gym + Pets_live + Kid_friendly + superhost + property_type +
##     cancel_policy + room_type + neighbourhood, data = airbnb)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.072523 -0.014659 -0.000277  0.015005  0.072043
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.318e-01  4.274e-03 101.027  < 2e-16 ***
## host_response                 1.661e-02  1.290e-03  12.869  < 2e-16 ***
## log.deposit                   1.117e-05  7.339e-05   0.152 0.879013
## t.cleaning_fee               -1.725e-03  4.860e-04  -3.549 0.000391 ***
## t.availability_90            -1.267e-03  2.000e-04  -6.334 2.69e-10 ***
## t.review_num                  1.037e-03  6.593e-04   1.573 0.115895
## t.review_score               -1.317e-06  3.884e-07  -3.390 0.000706 ***
## guests_included              -1.775e-03  3.504e-04  -5.067 4.25e-07 ***
## min_nights                    3.580e-04  1.773e-04   2.019 0.043564 *
## bathrooms                    -5.457e-03  8.615e-04  -6.334 2.69e-10 ***
## bedrooms                     -1.496e-02  7.900e-04 -18.931  < 2e-16 ***
## beds                         -2.819e-03  5.811e-04  -4.851 1.28e-06 ***
## instant_bookableYes           4.626e-03  1.097e-03   4.217 2.53e-05 ***
## TVYes                        -4.774e-03  9.517e-04  -5.017 5.51e-07 ***
## ParkingYes                    1.092e-03  8.771e-04   1.245 0.213300
## ACYes                        -2.141e-03  1.079e-03  -1.983 0.047422 *
## Checkin_24hourYes             8.370e-03  1.115e-03   7.509 7.50e-14 ***
## Pets_allowedYes              -7.714e-04  1.229e-03  -0.627 0.530418
## GymYes                       -4.991e-04  1.526e-03  -0.327 0.743625
## Pets_liveYes                     6.753e-04  9.850e-04   0.686 0.493030    
## Kid_friendlyYes                 -1.880e-03  8.869e-04  -2.120 0.034070 *  
## superhostYes                    -7.784e-03  1.045e-03  -7.446 1.20e-13 ***
## property_typeBed & Breakfast    -2.365e-02  4.418e-03  -5.353 9.19e-08 ***
## property_typeBoat               -3.019e-02  9.600e-03  -3.145 0.001677 ** 
## property_typeBungalow           -5.210e-03  7.059e-03  -0.738 0.460561    
## property_typeCabin              -6.105e-03  5.161e-03  -1.183 0.236919    
## property_typeCamper/RV           1.062e-02  7.418e-03   1.432 0.152175    
## property_typeChalet             -1.226e-02  1.645e-02  -0.745 0.456147    
## property_typeCondominium        -3.232e-03  2.601e-03  -1.243 0.214010    
## property_typeDorm                4.526e-02  1.717e-02   2.636 0.008437 ** 
## property_typeHouse              -3.772e-03  1.098e-03  -3.437 0.000595 ***
## property_typeLoft               -1.258e-02  3.792e-03  -3.318 0.000915 ***
## property_typeOther              -2.860e-03  5.133e-03  -0.557 0.577419    
## property_typeTent                2.629e-02  1.056e-02   2.491 0.012798 *  
## property_typeTownhouse          -7.062e-03  2.466e-03  -2.864 0.004209 ** 
## property_typeTreehouse          -5.325e-03  1.345e-02  -0.396 0.692261    
## property_typeYurt               -2.350e-03  2.332e-02  -0.101 0.919764    
## cancel_policymoderate           -1.661e-04  1.060e-03  -0.157 0.875482    
## cancel_policystrict             -1.582e-03  1.134e-03  -1.395 0.163027    
## room_typePrivate room            3.855e-02  1.120e-03  34.432  < 2e-16 ***
## room_typeShared room             8.285e-02  2.546e-03  32.541  < 2e-16 ***
## neighbourhoodBeacon Hill         7.939e-03  2.740e-03   2.898 0.003781 ** 
## neighbourhoodCapitol Hill       -1.306e-02  1.918e-03  -6.811 1.13e-11 ***
## neighbourhoodCascade            -1.407e-02  3.094e-03  -4.547 5.61e-06 ***
## neighbourhoodCentral Area       -4.395e-03  1.997e-03  -2.200 0.027838 *  
## neighbourhoodDelridge            1.751e-02  3.157e-03   5.546 3.14e-08 ***
## neighbourhoodDowntown           -2.226e-02  2.158e-03 -10.315  < 2e-16 ***
## neighbourhoodInterbay           -4.289e-03  7.568e-03  -0.567 0.570940    
## neighbourhoodLake City           1.009e-02  3.391e-03   2.977 0.002929 ** 
## neighbourhoodMagnolia           -5.534e-03  3.530e-03  -1.568 0.117046    
## neighbourhoodNorthgate           9.740e-03  3.138e-03   3.104 0.001927 ** 
## neighbourhoodOther neighborhoods  1.512e-03  1.781e-03   0.849 0.395927    
## neighbourhoodQueen Anne         -1.539e-02  2.111e-03  -7.292 3.75e-13 ***
## neighbourhoodRainier Valley      1.099e-02  2.464e-03   4.462 8.39e-06 ***
## neighbourhoodSeward Park         9.444e-03  3.873e-03   2.439 0.014786 *  
## neighbourhoodUniversity District  6.654e-03  2.788e-03   2.387 0.017045 *  
## neighbourhoodWest Seattle       -3.898e-03  2.323e-03  -1.678 0.093451 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.02311 on 3543 degrees of freedom
## Multiple R-squared:  0.717,  Adjusted R-squared:  0.7125 
## F-statistic: 160.3 on 56 and 3543 DF,  p-value: < 2.2e-16
```

After removing the outliers, the adjusted R-squared of our model increases to 0.7125. And the diagnostic plot of the model also pretty good.



However, since there are still many insignificant variables after removing outliers, we use stepwise selection to help us pick up important variables.

The model AIC selected is as follows:

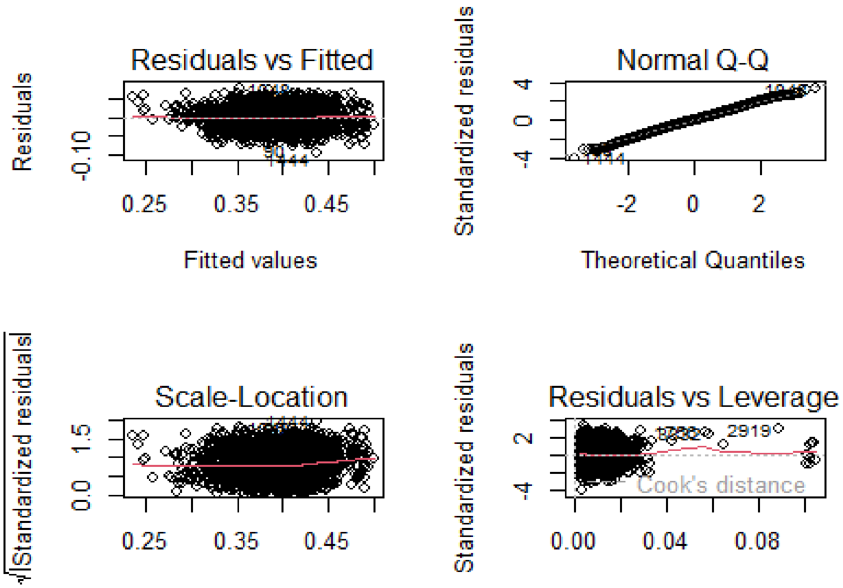
$$\begin{aligned} Price^{-1/5} = \beta_0 &+ \beta_1 host\_response + \beta_2 t.cleaning\_fee + \beta_3 t.availability\_90 \\ &+ \beta_4 t.review\_num + \beta_5 t.review\_score + \beta_6 guests\_included \\ &+ \beta_7 min\_nights + \beta_8 bathrooms + \beta_9 bedrooms + \beta_{10} beds \\ &+ \beta_{11} instant\_bookable + \beta_{12} TV + \beta_{13} AC + \beta_{14} Checkin\_24hour \\ &+ \beta_{15} Kid\_friendly + \beta_{16} superhost + \beta_{17} property_{type} \quad + \beta_{18} room\_type \\ &+ \beta_{19} neighborhood + \varepsilon_t \end{aligned}$$

The model BIC selected is as follows:

$$\begin{aligned} Price^{-1/5} = \beta_0 &+ \beta_1 host\_response + \beta_2 t.cleaning\_fee + \beta_3 t.Availability\_90 \\ &+ \beta_4 t.review\_score + \beta_5 guests\_included + \beta_6 bathrooms + \beta_7 bedrooms \\ &+ \beta_8 beds + \beta_9 instant\_bookable + \beta_{10} TV + \beta_{11} Checkin\_24hour \\ &+ \beta_{12} superhost + \beta_{13} room\_type + \beta_{14} neighborhood + \varepsilon_t \end{aligned}$$

The model AIC chooses has more variables than the one BIC choose. And the adjusted R-square of both model is 0.7126 and 0.7068. Since we aim to have a good degree of model fit, and there's not a big difference in R-square value between AIC and BIC, we select BIC model.

The diagnostic plot of the model is down below, and all diagnostic plots are pretty good.



Then we conduct VIF test to check the variables in the BIC model, and all values of variables are below 10.

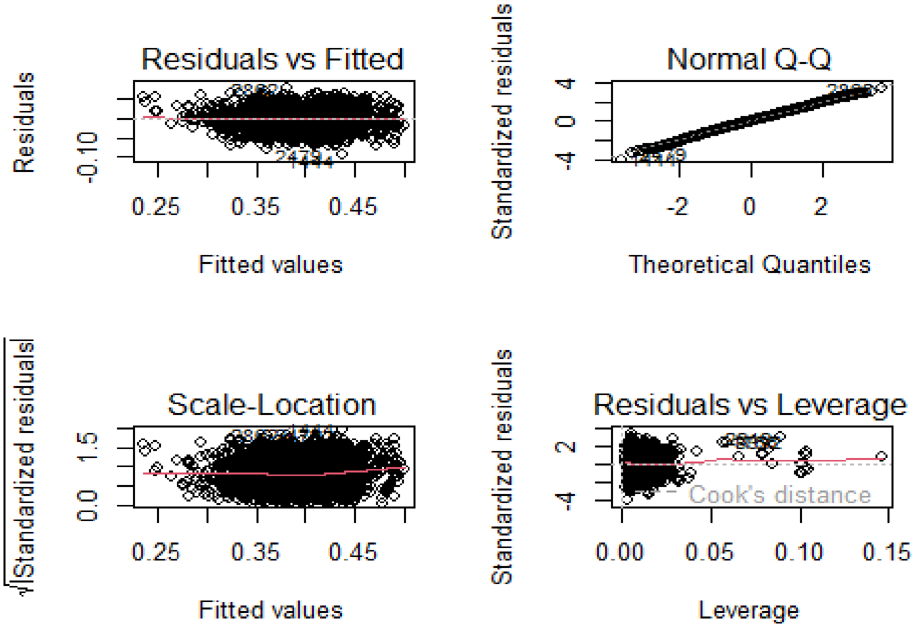
```
##                    host_response                   t.cleaning_fee
##                         1.182930                         1.309119
##                t.availability_90                   t.review_score
##                         1.199181                         1.107982
##                  guests_included                        bathrooms
##                         1.373128                         1.682379
##                         bedrooms                             beds
##                         3.031448                         2.795581
##              instant_bookableYes                            TVYes
##                         1.060363                         1.173752
##                Checkin_24hourYes                     superhostYes
##                         1.072228                         1.115765
##            room_typePrivate room             room_typeShared room
##                         1.422903                         1.111628
##         neighbourhoodBeacon Hill        neighbourhoodCapitol Hill
##                         1.460756                         2.980348
##             neighbourhoodCascade        neighbourhoodCentral Area
##                         1.354868                         2.343635
##            neighbourhoodDelridge            neighbourhoodDowntown
##                         1.305456                         3.016450
##            neighbourhoodInterbay           neighbourhoodLake City
##                         1.046579                         1.254390
##            neighbourhoodMagnolia           neighbourhoodNorthgate
##                         1.232883                         1.313275
## neighbourhoodOther neighborhoods          neighbourhoodQueen Anne
##                         3.484824                         2.119760
##      neighbourhoodRainier Valley         neighbourhoodSeward Park
##                         1.618377                         1.182529
## neighbourhoodUniversity District        neighbourhoodWest Seattle
##                         1.498136                         1.770844
```

Therefore, according to the analysis above, we select ***lmod.bic*** as our final model with adjusted R-square value 0.7068.

Since some of transformed variables are still non-linear, we tried to add polynomial variables to fix this problem. We squared predictors $t.Avi\_90\ and\ t.RSR$, then add them into model. And the summary of new model ***lmod_bic_new*** is as follows.

```
lmod_bic_new = lm(formula = I((price)^(-1/5)) ~ host_response + t.cleaning_fee + t.avai
lability_90 + t.review_score + guests_included + bathrooms + bedrooms + beds + instant_
bookable + TV + Checkin_24hour + superhost + room_type + neighbourhood + I((t.availabil
ity_90)^2) + I((t.review_score)^2), data = airbnb)
summary(lmod_bic_new)

## 
## Call:
## lm(formula = I((price)^(-1/5)) ~ host_response + t.cleaning_fee + 
##     t.availability_90 + t.review_score + guests_included + bathrooms + 
##     bedrooms + beds + instant_bookable + TV + Checkin_24hour + 
##     superhost + room_type + neighbourhood + I((t.availability_90)^2) + 
##     I((t.review_score)^2), data = airbnb)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.091960 -0.014337  0.000229  0.015314  0.077942 
## 
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                       4.039e-01  9.978e-03  40.484  < 2e-16 ***
## host_response                     1.500e-02  1.278e-03  11.738  < 2e-16 ***
## t.cleaning_fee                   -1.990e-03  4.419e-04  -4.504 6.88e-06 ***
## t.availability_90                 5.003e-03  8.365e-04   5.981 2.44e-09 ***
## t.review_score                    6.214e-06  2.472e-06   2.513 0.011998 *  
## guests_included                  -1.960e-03  3.425e-04  -5.723 1.13e-08 ***
## bathrooms                        -6.188e-03  8.375e-04  -7.388 1.85e-13 ***
## bedrooms                         -1.535e-02  7.520e-04 -20.415  < 2e-16 ***
## beds                             -2.831e-03  5.606e-04  -5.051 4.62e-07 ***
## instant_bookableYes               3.987e-03  1.087e-03   3.668 0.000248 ***
## TVYes                            -4.983e-03  9.334e-04  -5.339 9.95e-08 ***
## Checkin_24hourYes                 7.099e-03  1.089e-03   6.521 7.94e-11 ***
## superhostYes                     -6.714e-03  9.977e-04  -6.729 1.98e-11 ***
## room_typePrivate room             3.731e-02  1.011e-03  36.885  < 2e-16 ***
## room_typeShared room              8.394e-02  2.475e-03  33.912  < 2e-16 ***
## neighbourhoodBeacon Hill          8.891e-03  2.707e-03   3.285 0.001030 ** 
## neighbourhoodCapitol Hill        -1.172e-02  1.870e-03  -6.269 4.07e-10 ***
## neighbourhoodCascade             -1.282e-02  3.028e-03  -4.233 2.36e-05 ***
## neighbourhoodCentral Area        -3.656e-03  1.991e-03  -1.836 0.066465 .  
## neighbourhoodDelridge             1.725e-02  3.146e-03   5.483 4.48e-08 ***
## neighbourhoodDowntown            -2.156e-02  1.922e-03 -11.218  < 2e-16 ***
## neighbourhoodInterbay            -6.789e-03  7.492e-03  -0.906 0.364897    
## neighbourhoodLake City            1.056e-02  3.373e-03   3.132 0.001752 ** 
## neighbourhoodMagnolia            -4.919e-03  3.524e-03  -1.396 0.162903    
## neighbourhoodNorthgate            1.021e-02  3.113e-03   3.281 0.001044 ** 
## neighbourhoodOther neighborhoods  2.031e-03  1.775e-03   1.144 0.252705    
## neighbourhoodQueen Anne          -1.431e-02  2.089e-03  -6.853 8.46e-12 ***
## neighbourhoodRainier Valley       1.082e-02  2.446e-03   4.422 1.01e-05 ***
## neighbourhoodSeward Park          1.037e-02  3.859e-03   2.688 0.007219 ** 
## neighbourhoodUniversity District  8.770e-03  2.747e-03   3.192 0.001424 **
## neighbourhoodWest Seattle        -3.855e-03  2.306e-03  -1.672 0.094642 .
## I((t.availability_90)^2)         -1.016e-03  1.313e-04  -7.740 1.29e-14 ***
## I((t.review_score)^2)            -5.007e-10  1.572e-10  -3.185 0.001460 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02312 on 3567 degrees of freedom
## Multiple R-squared:  0.7147, Adjusted R-squared:  0.7122
## F-statistic: 279.3 on 32 and 3567 DF,  p-value: < 2.2e-16
```



As we can see from the summary, the R-squared improves to 0.7122. And diagnostic plots of new model also seem good. Then we can get the VIF test result of the model.

```
vif(lmod_bic_new)

##                      host_response                    t.cleaning_fee
##                           1.210279                          1.309357
##                   t.availability_90                    t.review_score
##                          22.276595                         46.227903
##                    guests_included                         bathrooms
##                           1.373500                          1.684977
##                           bedrooms                              beds
##                           3.034670                          2.805015
##                 instant_bookableYes                             TVYes
##                           1.063368                          1.174674
##                   Checkin_24hourYes                       superhostYes
##                           1.073915                          1.130188
##                 room_typePrivate room              room_typeShared room
##                           1.431160                          1.114034
##              neighbourhoodBeacon Hill          neighbourhoodCapitol Hill
##                           1.461336                          2.983896
##                 neighbourhoodCascade          neighbourhoodCentral Area
##                           1.357651                          2.349331
##                 neighbourhoodDelridge            neighbourhoodDowntown
##                         1.305768                         3.018894
##            neighbourhoodInterbay           neighbourhoodLake City
##                         1.046998                         1.255454
##            neighbourhoodMagnolia           neighbourhoodNorthgate
##                         1.235683                         1.313354
## neighbourhoodOther neighborhoods          neighbourhoodQueen Anne
##                         3.490231                         2.120680
##      neighbourhoodRainier Valley         neighbourhoodSeward Park
##                         1.618966                         1.183183
## neighbourhoodUniversity District        neighbourhoodWest Seattle
##                         1.505273                         1.771697
##         I((t.availability_90)^2)            I((t.review_score)^2)
##                        21.738082                        46.563097
```

Only the VIF value of polynomial variables and their original variables get improved, which is within our expectation and is reasonable.

6 Model Interpretation

The summary of our final model is down below.

```
summary(lmod_bic_new)

## 
## Call:
## lm(formula = I((price)^(-1/5)) ~ host_response + t.cleaning_fee + 
##     t.availability_90 + t.review_score + guests_included + bathrooms + 
##     bedrooms + beds + instant_bookable + TV + Checkin_24hour + 
##     superhost + room_type + neighbourhood + I((t.availability_90)^2) + 
##     I((t.review_score)^2), data = airbnb)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.091960 -0.014337  0.000229  0.015314  0.077942 
## 
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                       4.039e-01  9.978e-03  40.484  < 2e-16 ***
## host_response                     1.500e-02  1.278e-03  11.738  < 2e-16 ***
## t.cleaning_fee                   -1.990e-03  4.419e-04  -4.504 6.88e-06 ***
## t.availability_90                 5.003e-03  8.365e-04   5.981 2.44e-09 ***
## t.review_score                    6.214e-06  2.472e-06   2.513 0.011998 *  
## guests_included                  -1.960e-03  3.425e-04  -5.723 1.13e-08 ***
## bathrooms                        -6.188e-03  8.375e-04  -7.388 1.85e-13 ***
## bedrooms                         -1.535e-02  7.520e-04 -20.415  < 2e-16 ***
## beds                             -2.831e-03  5.606e-04  -5.051 4.62e-07 ***
## instant_bookableYes               3.987e-03  1.087e-03   3.668 0.000248 ***
## TVYes                            -4.983e-03  9.334e-04  -5.339 9.95e-08 ***
## Checkin_24hourYes                 7.099e-03  1.089e-03   6.521 7.94e-11 ***
## superhostYes                     -6.714e-03  9.977e-04  -6.729 1.98e-11 ***
## room_typePrivate room             3.731e-02  1.011e-03  36.885  < 2e-16 ***
## room_typeShared room              8.394e-02  2.475e-03  33.912  < 2e-16 ***
## neighbourhoodBeacon Hill          8.891e-03  2.707e-03   3.285 0.001030 ** 
## neighbourhoodCapitol Hill        -1.172e-02  1.870e-03  -6.269 4.07e-10 ***
## neighbourhoodCascade             -1.282e-02  3.028e-03  -4.233 2.36e-05 ***
## neighbourhoodCentral Area        -3.656e-03  1.991e-03  -1.836 0.066465 .  
## neighbourhoodDelridge             1.725e-02  3.146e-03   5.483 4.48e-08 ***
## neighbourhoodDowntown            -2.156e-02  1.922e-03 -11.218  < 2e-16 ***
## neighbourhoodInterbay            -6.789e-03  7.492e-03  -0.906 0.364897    
## neighbourhoodLake City            1.056e-02  3.373e-03   3.132 0.001752 ** 
## neighbourhoodMagnolia            -4.919e-03  3.524e-03  -1.396 0.162903    
## neighbourhoodNorthgate            1.021e-02  3.113e-03   3.281 0.001044 ** 
## neighbourhoodOther neighborhoods  2.031e-03  1.775e-03   1.144 0.252705    
## neighbourhoodQueen Anne          -1.431e-02  2.089e-03  -6.853 8.46e-12 ***
## neighbourhoodRainier Valley       1.082e-02  2.446e-03   4.422 1.01e-05 ***
## neighbourhoodSeward Park          1.037e-02  3.859e-03   2.688 0.007219 ** 
## neighbourhoodUniversity District  8.770e-03  2.747e-03   3.192 0.001424 ** 
## neighbourhoodWest Seattle        -3.855e-03  2.306e-03  -1.672 0.094642 .  
## I((t.availability_90)^2)         -1.016e-03  1.313e-04  -7.740 1.29e-14 ***
## I((t.review_score)^2)            -5.007e-10  1.572e-10  -3.185 0.001460 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.02312 on 3567 degrees of freedom
## Multiple R-squared:  0.7147, Adjusted R-squared:  0.7122 
## F-statistic: 279.3 on 32 and 3567 DF,  p-value: < 2.2e-16
```

Firstly, from the table, we can see that almost all variables in the model are significant to the price, and the R-squared value is 0.7122. Since the response variable has been transformed, after re-analyzing the coefficient, we find that those predictors with *negative* coefficients have a *negative* influence on the transformed response variable, which means that they have a **positive** influence on the original response variable. For example, the more bedrooms the property has, the higher price it has. Based on our result, we suggest the host can provide more amenities to guests such as TV, which can improve the price of the property. Also, the certification of the super host from Airbnb is beneficial for the price increase.

Appendix

```r
airbnb = read.csv('/Users/cosmoser/Desktop/airbnb.csv')
head(airbnb)
airbnb = airbnb %>%
  mutate(cancel_policy = as.factor(cancel_policy),
         neighbourhood = as.factor(neighbourhood),
         property_type = as.factor(property_type),
         room_type = as.factor(room_type),
         instant_bookable = as.factor(instant_bookable),
         superhost = as.factor(superhost),
         TV = as.factor(TV),
         AC= as.factor(AC),
         Parking = as.factor(Parking),
         Checkin_24hour = as.factor(Checkin_24hour),
         Pets_allowed = as.factor(Pets_allowed),
         Gym = as.factor(Gym),
         Pets_live = as.factor(Pets_live),
         Kid_friendly = as.factor(Kid_friendly)
)
summary(airbnb)

par(mfrow=c(2,4))
Name=names(airbnb)
##airbnb$minimum_nights[minimum_nights = 1000] = 0
for(i in c(1:12)){
  hist(airbnb[,i], main=Name[i], xlab=Name[i])
  qqnorm(airbnb[,i], main=Name[i], xlab=Name[i])
  qqline(airbnb[,i])}

par(mfrow=c(2,4))
for (i in c(13:23, 26)){
  boxplot(airbnb$price~airbnb[,i], main=Name[i], xlab=Name[i])
}

boxplot(airbnb$price~airbnb$neighbourhood, main=Name[24], xlab=Name[2
4])
boxplot(airbnb$price~airbnb$neighbourhood, main=Name[25], xlab=Name[2
5])

par(mfrow=c(2,2))
for (i in c(2:12)) {
  lmod0=lm(I(abs(airbnb[,i])+0.001)~1)
  b=boxcox(lmod0,xlab=Name[i])
}

par(mfrow=c(2,2))
for (i in c(2:12)) {
  plot(airbnb[ , i], airbnb$price, main=Name[i], xlab=Name[i], ylab="pr
ice")
}

lmod1 = lm(price~. ,data=airbnb)
summary(lmod1)
par(mfrow=c(2,2))
plot(lmod1)
boxcox(lmod1)

lmod2=lm(I(price^(-1/5))~host_response+deposit+cleaning_fee+availabilit
y_90+review_num+review_score+guests_included+min_nights+bathrooms+bedro
oms+beds+instant_bookable+TV+Parking+AC+Checkin_24hour+Pets_allowed+Gym
+Pets_live+Kid_friendly+superhost+room_type+property_type+cancel_policy
+neighbourhood,data=airbnb)
summary(lmod2)
par(mfrow=c(2,2))
plot(lmod2)

Name=names(airbnb)
par(mfrow=c(2,2))
for (i in c(2:12)) {
  plot(airbnb[ , i], (airbnb$price)^(-1/5), main=Name[i], xlab=Name[i],
 ylab="price^(-1/5)")
}

airbnb$log.deposit = log(airbnb$deposit+0.001)
airbnb$t.cleaning_fee = airbnb$cleaning_fee^(0.2)
airbnb$t.availability_90 = airbnb$availability_90^(0.4)
airbnb$t.review_num = airbnb$review_num^(0.2)
airbnb$t.review_score = airbnb$review_score^2
Name = names(airbnb)
par(mfrow=c(2,2))
for (i in c(27:31)) {
  plot(airbnb[,i],(airbnb$price)^(-1/5),main=Name[i],xlab=Name[i],ylab=
"price^(-1/5)")
}

lmod3 = lm(I((price)^(-1/5))~host_response+log.deposit+t.cleaning_fee+
t.availability_90+t.review_num+t.review_score+guests_included+min_night
s+bathrooms+bedrooms+beds+instant_bookable+TV+Parking+AC+Checkin_24hour
+Pets_allowed+Gym+Pets_live+Kid_friendly+superhost+property_type+cancel
_policy+room_type+neighbourhood, data=airbnb)
summary(lmod3)
par(mfrow=c(2,2))
plot(lmod3)

studres=rstudent(lmod3)
range(studres)
out.idx = which(abs(studres)>3)
out.idx
```

```
length(out.idx)airbnb[c(out.idx),]
summary(airbnb)

out.ind = c(out.idx,31)
airbnb = airbnb[-out.ind,]
airbnb[1055,]
airbnb = airbnb[-1055,]
lmod3.new = lm(I((price)^(-1/5))~host_response+log.deposit+t.cleaning_f
ee+t.availability_90+t.review_num+t.review_score+guests_included+min_ni
ghts+bathrooms+bedrooms+beds+instant_bookable+TV+Parking+AC+Checkin_24h
our+Pets_allowed+Gym+Pets_live+Kid_friendly+superhost+property_type+can
cel_policy+room_type+neighbourhood, data=airbnb)
summary(lmod3.new)
par(mfrow=c(2,2))
plot(lmod3.new)

library(MASS)
lmod.aic = stepAIC(lmod3.new,direction="both",k=2,trace = F)
summary(lmod.aic)
n=dim(airbnb)[1]
lmod.bic = stepAIC(lmod3.new,direction="both",k=log(n),trace = F)
summary(lmod.bic)
AIC(lmod.aic)
AIC(lmod.bic)
BIC(lmod.aic)
BIC(lmod.bic)

library(faraway)
vif(lmod.bic)
vif(lmod.aic)
par(mfrow=c(2,2))
plot(lmod.bic)

summary(lmod.bic)
par(mfrow=c(2,2))
lmod_bic_new = lm(formula = I((price)^(-1/5)) ~ host_response + t.clean
ing_fee +
    t.availability_90 + t.review_score + guests_included +
    bathrooms + bedrooms + beds + instant_bookable + TV +
    Checkin_24hour + superhost + room_type +
    neighbourhood + I((t.availability_90)^2) + I((t.review_score)^2), d
ata = airbnb)
summary(lmod_bic_new)
plot(lmod_bic_new)
```

Reference

1. Seattle Airbnb Open Data | Kaggle