airbnb_model

Ningze Zu

11/25/2018

1. Abstract

In this project, I will perform an exploratory data analysis to select features and build a model to predict the Airbnb listing prices in Seattle and compare with the prices of Hotels in Seattle.

2. Introduction

2.1 Background

Airbnb is a well-known website, providing online hospitality service and enabling hosts to list their properties and visitors to rent short-term accommodations.

2.2 Previous work (Exploratory Data Analysis)

a. Price Distribution

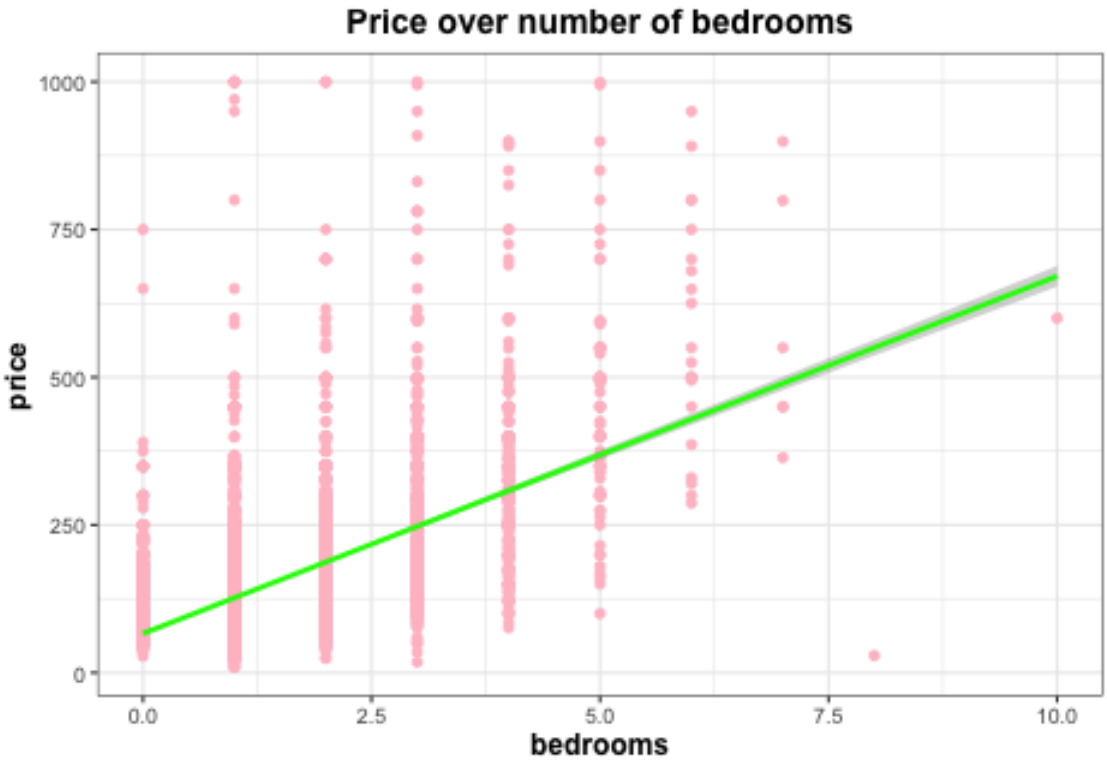
First, I looked into the price distribution among each of the cities. Then, I decided to choose price range betweem 0 and 1000 USD since the data contains outliers when the price is over 1000 USD. After excluding the price over 1000 USD, the price distributions are shown as following chart:

The plots show a high concentration of listing prices between 0 and 100 USD.

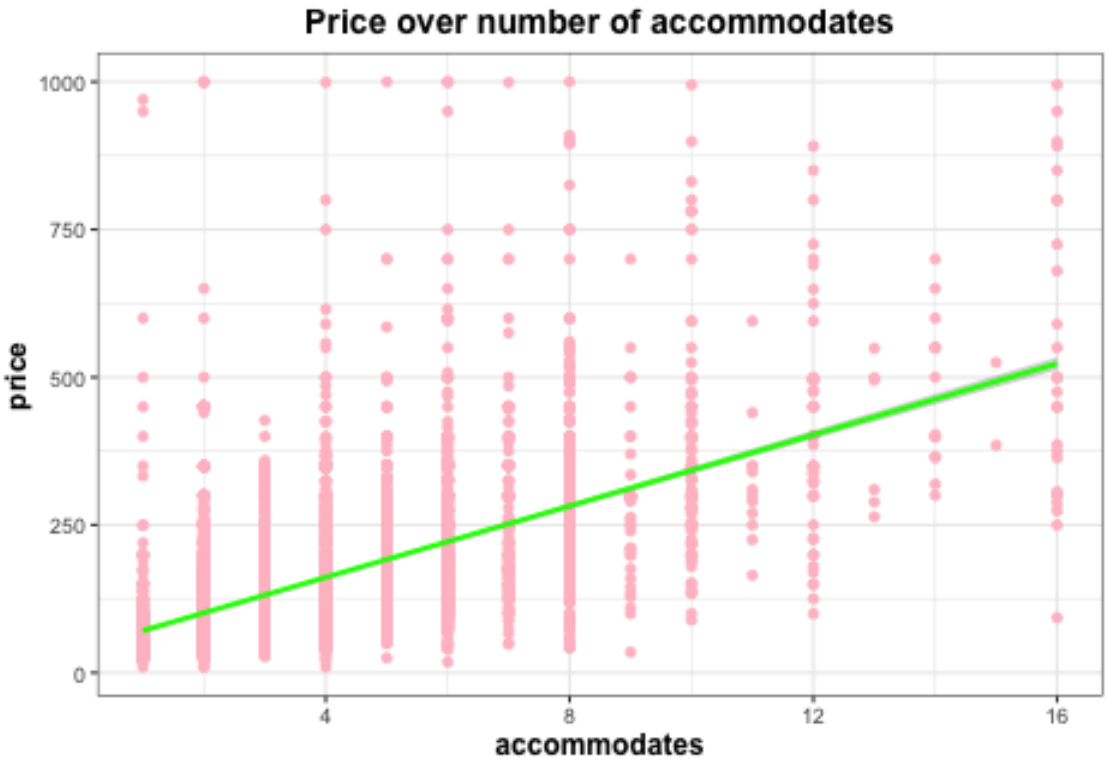
b. Variables

Number of bedrooms will always be a major factor to price of Airbnb. I made a barplot to show the average prices over different number of bedroom. In this case, we can see that there is a certain relationship between the average price of Airbnb listings and number of bedrooms.

Next, I did several scatterplot to show the price versus different variables such as 'bedrooms', 'bathrooms', 'accommodates'. In case it is hard to see if there is relationship between price and each of these vairables , so I add a line to see the trend of the points. From these plots, we can also see that number of bedrooms, number of bathrooms and accommodates have a certain effect on the price of Airbnb listings.

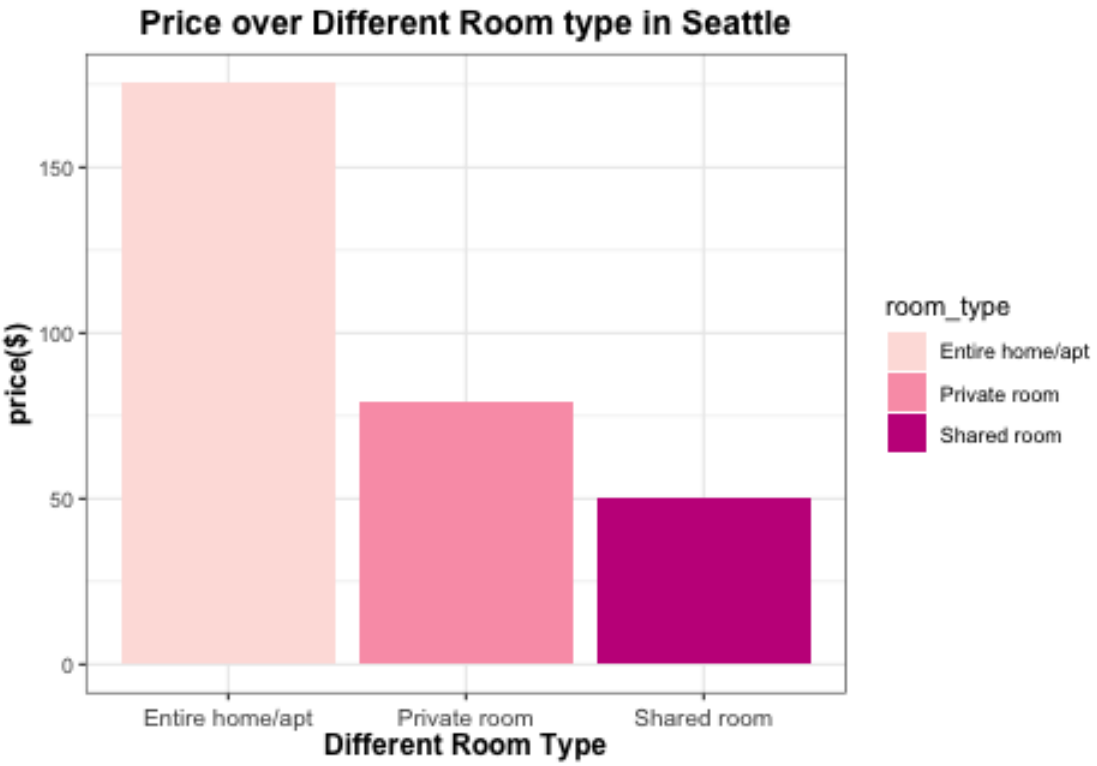

Price over number of bedrooms
price
1000
750
500
250
0
0.0
2.5
5.0
7.5
10.0
bedrooms


Price over number of bathrooms
price
1000
750
500
250
0
0
2
4
6
8
bathrooms



The next histogram chart shows the count of different property types in Seattle. We we can observe the disparity in count of different property types. Some property types have more Airbnb listings than others, most of the Airbnb listings' property types are 'Apartment', 'House' and 'Condo'.

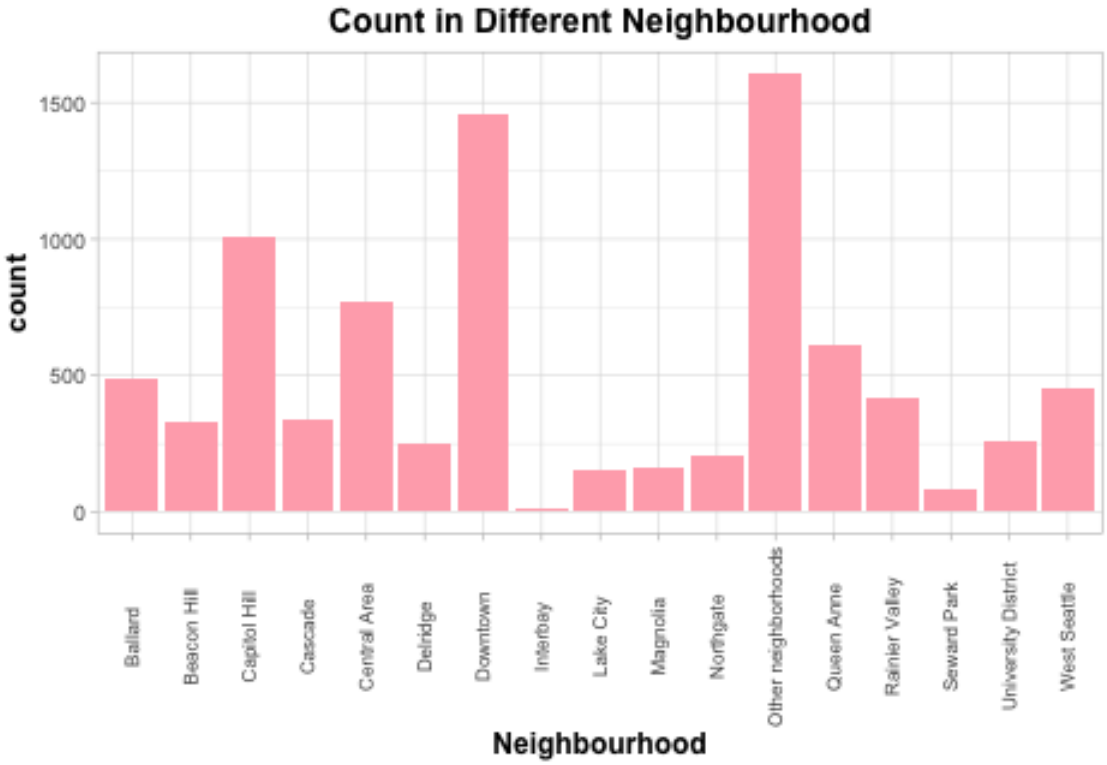
In Airbnb listings, there are three types of room: Entire home/apt, Shared room and Private room. Room type is also a major factor of the price.

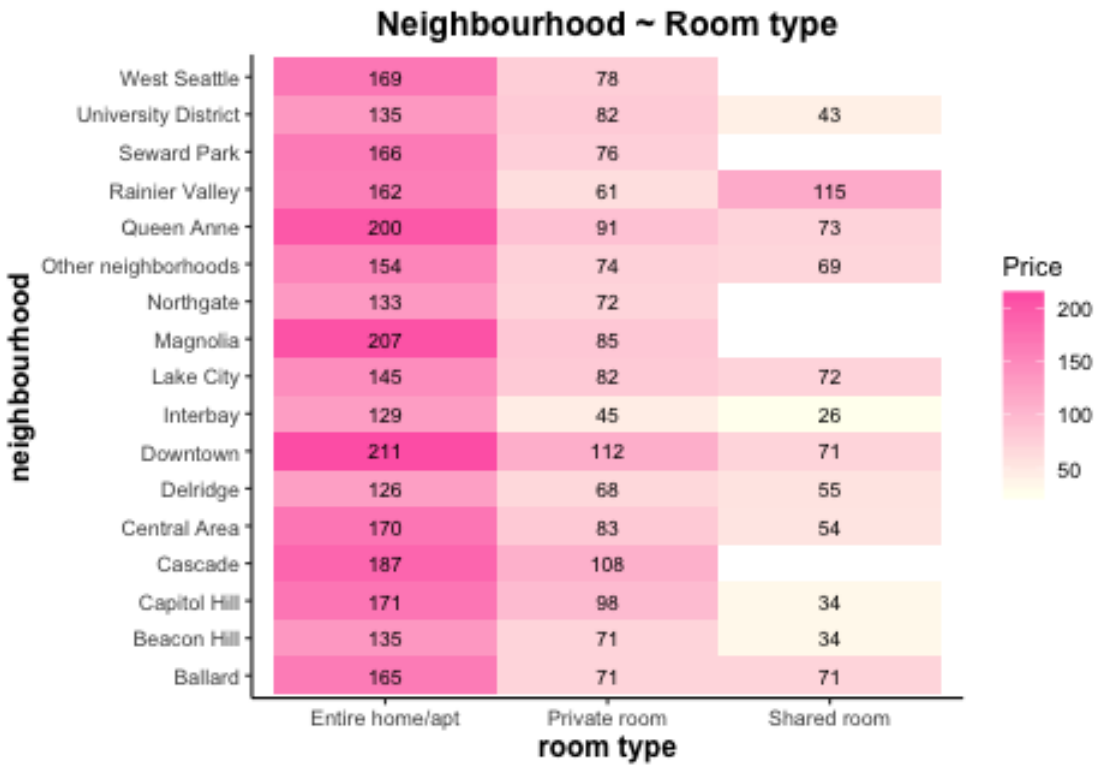


Next, the scatterplot shows the prices over review rate score.

```
## Warning: Removed 993 rows containing missing values (geom_point).
```

Lastly, The bar plot below shows the average prices in different neighbourhood. To give a better understanding of the price in different neighbourhoods, I plotted a heatmap to show the room price in different neighbourhoods with different room types.






```
###leaflet
```

3. Method

3.1 Datasource

Data source: http://insideairbnb.com/get-the-data.html

Inside Airbnb Project is an independent and non-commercial set of tools that enables people to explore how Airbnb is really being used in cities around the world.

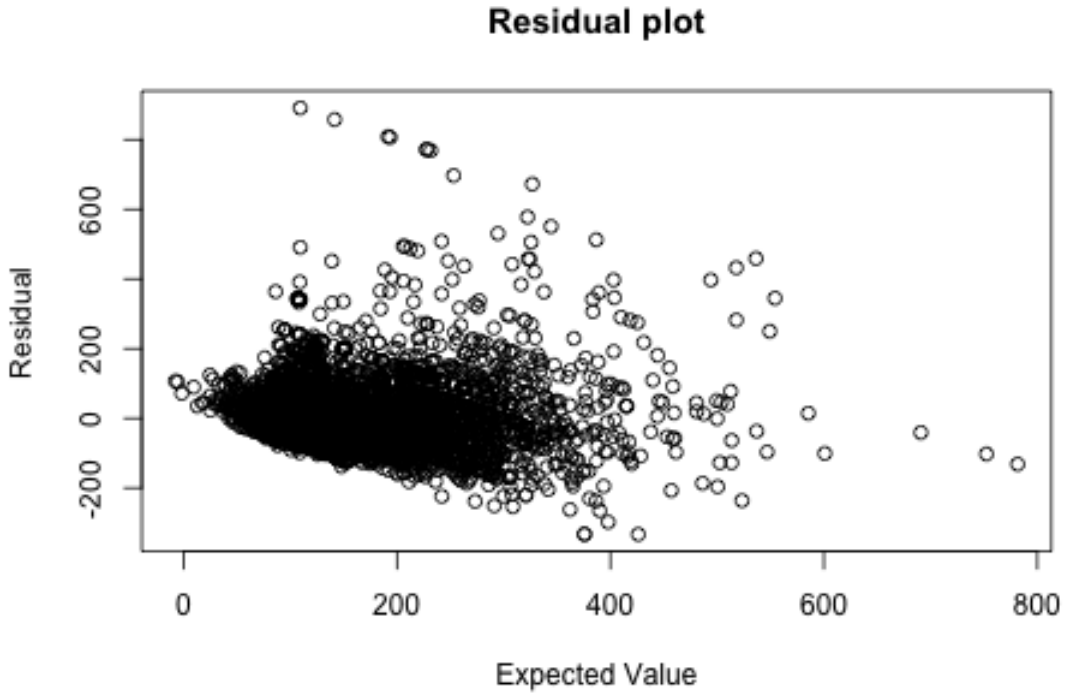
Features of dataset: room_id, property_type, room_type, city, neighbourhood, accommodates, review_score_rating, bedrooms, bathrooms, numbers_of_reviews, latitude, longitude.

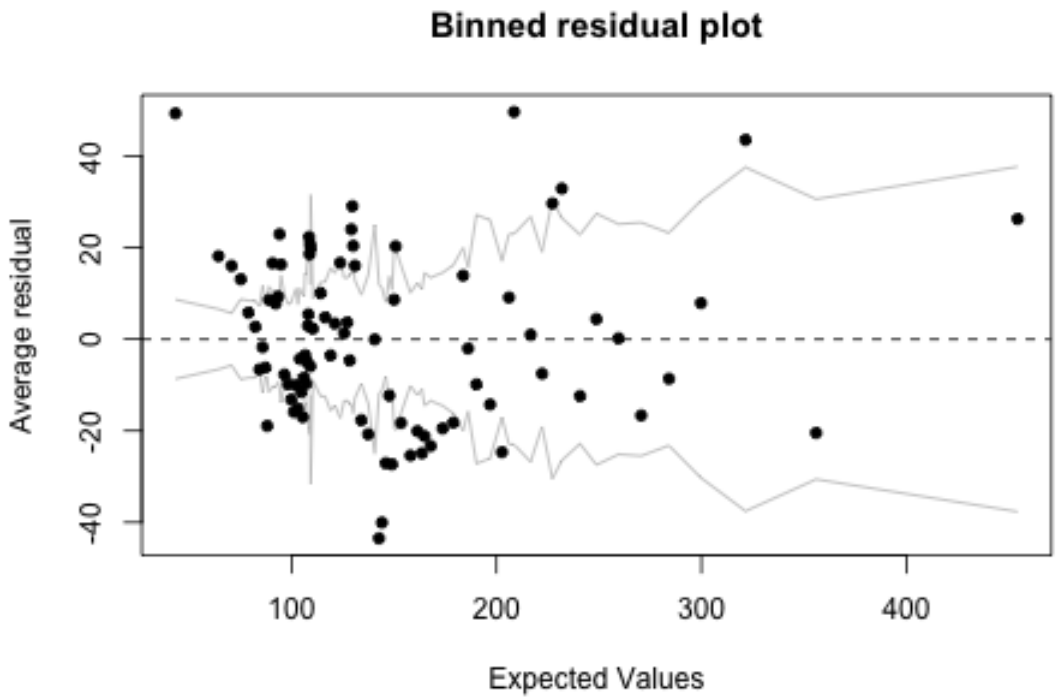
3.2 Model used

First, I fit a classic linear regression model without group variabls and draw a binned residual plot.

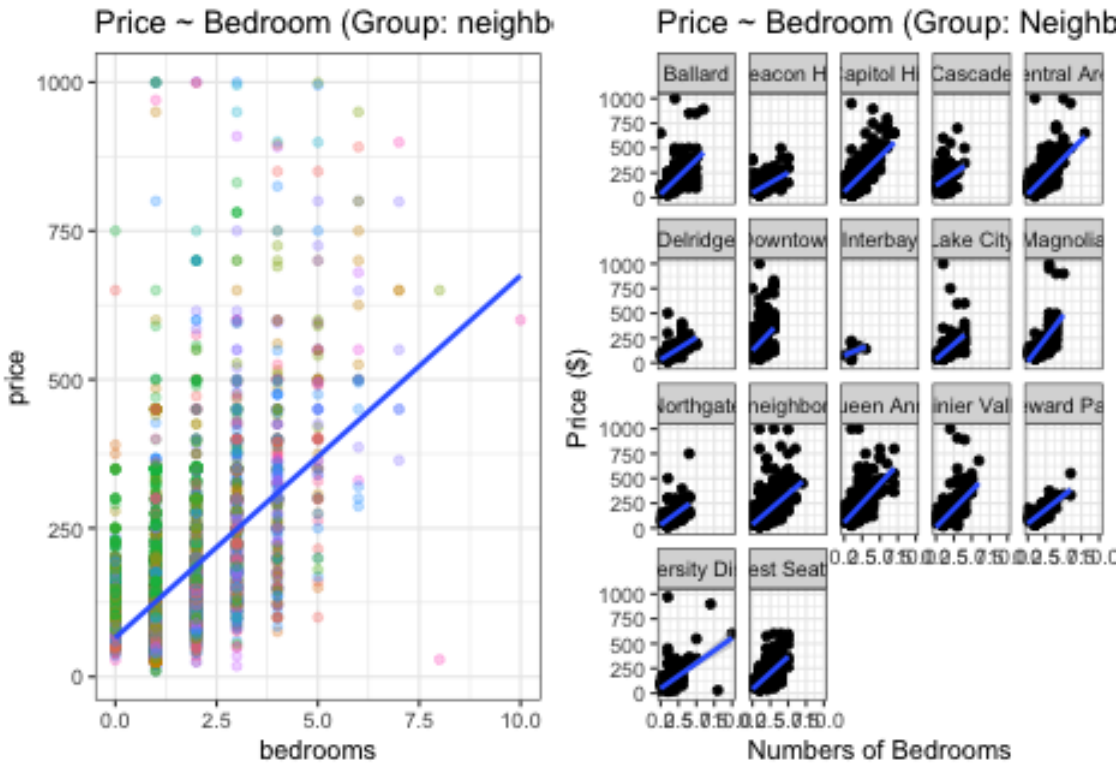
```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + accommodates + review_scores_rating +
##     number_of_reviews, data = airbnb_se_m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -333.14  -45.93  -13.87   29.21  890.65
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          28.89163   14.15206   2.042   0.0412 *
```

```
## bedrooms             14.44083    1.67204   8.637   <2e-16 ***
## bathrooms            22.15939    1.95304  11.346   <2e-16 ***
## accommodates         20.86557    0.69232  30.139   <2e-16 ***
## review_scores_rating  0.02491    0.14759   0.169   0.8660
## number_of_reviews    -0.17979    0.01490 -12.071   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.39 on 7606 degrees of freedom
##   (998 observations deleted due to missingness)
## Multiple R-squared:  0.4173, Adjusted R-squared:  0.4169
## F-statistic:  1089 on 5 and 7606 DF,  p-value: < 2.2e-16
```

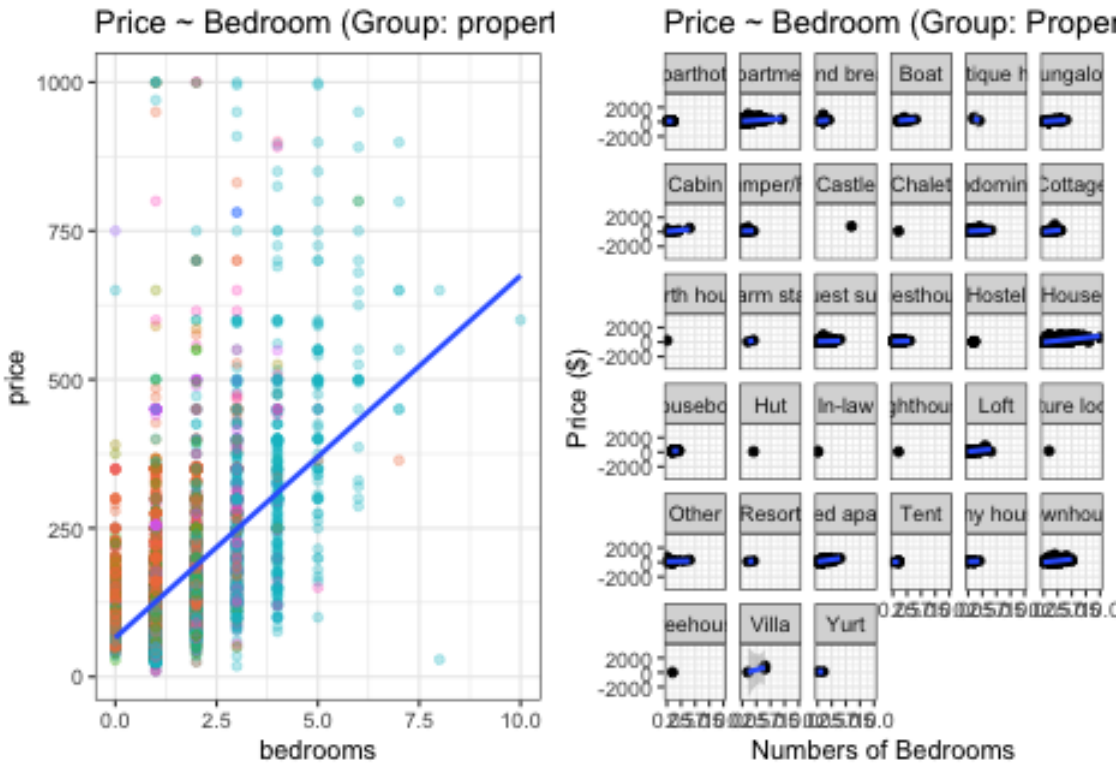





```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
## Warning: Removed 1 rows containing missing values (geom_point).
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
## Warning: Removed 1 rows containing missing values (geom_point).
```



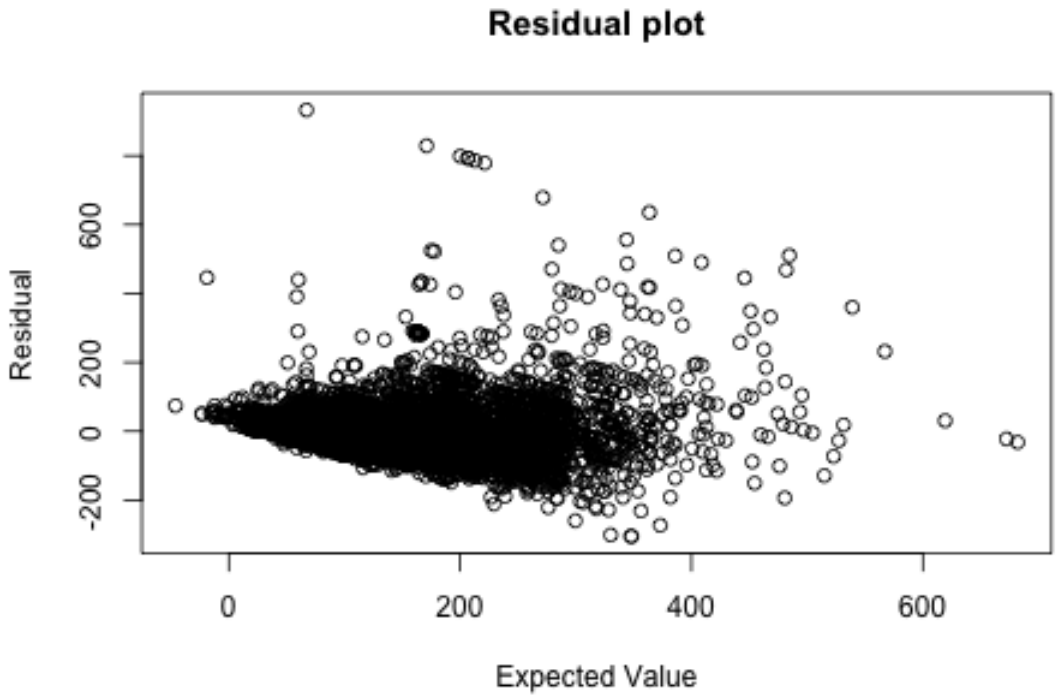

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning in qt((1 - level)/2, df): NaNs produced

## Warning in qt((1 - level)/2, df): NaNs produced

## Warning in qt((1 - level)/2, df): NaNs produced

## Warning: Removed 1 rows containing missing values (geom_point).
```

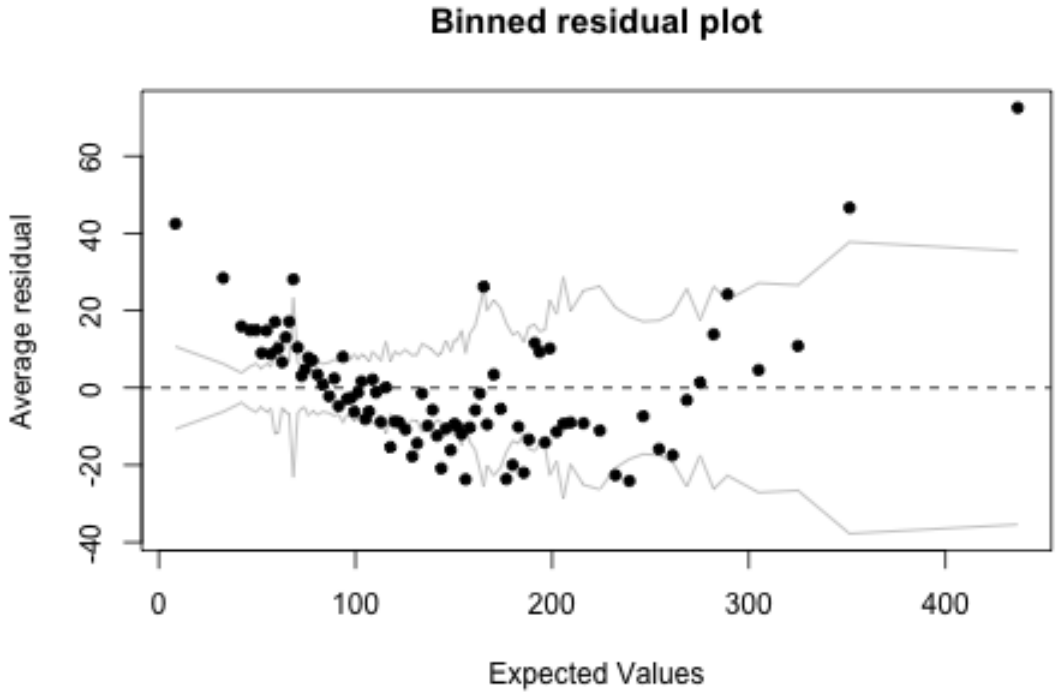


Next, I adjusted model by adding group variable 'neighbourhood' and 'property type' and draw the binned residual plot. Most variables seemed significant: bedrooms, bathrooms, accommodates, review_scores_rating, room_type, numbers_of_reviews. R-square is over 0.5 and p-value is pretty small.

```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + accommodates + review_scores_rating +
##     se.neighbourhood_group_cleansed + room_type + number_of_reviews,
##     data = airbnb_se_m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -305.13  -36.57   -6.09   23.71  932.84
##
## Coefficients:
##                                                 Estimate Std. Error
## (Intercept)                                    -10.21687   13.56409
## bedrooms                                        29.70951    1.53793
## bathrooms                                       29.55371    1.81085
## accommodates                                    11.47512    0.69068
## review_scores_rating                             0.38126    0.13438
## se.neighbourhood_group_cleansedBeacon Hill     -20.14131    5.50317
## se.neighbourhood_group_cleansedCapitol Hill     29.44399    4.25493
## se.neighbourhood_group_cleansedCascade          56.93294    5.55645
## se.neighbourhood_group_cleansedCentral Area      3.66982    4.40667
## se.neighbourhood_group_cleansedDelridge        -26.86445    6.12634
## se.neighbourhood_group_cleansedDowntown         76.79852    4.14284
## se.neighbourhood_group_cleansedInterbay         -0.29216   19.23089
## se.neighbourhood_group_cleansedLake City        -7.65133    7.03810
```

```
## se.neighbourhood_group_cleansedMagnolia              16.71207    7.00564
## se.neighbourhood_group_cleansedNorthgate            -14.36268    6.46575
## se.neighbourhood_group_cleansedOther neighborhoods   -7.88345    3.97266
## se.neighbourhood_group_cleansedQueen Anne            34.52708    4.66752
## se.neighbourhood_group_cleansedRainier Valley       -14.09466    5.13412
## se.neighbourhood_group_cleansedSeward Park           -6.02884    9.13799
## se.neighbourhood_group_cleansedUniversity District    1.84831    6.18566
## se.neighbourhood_group_cleansedWest Seattle           1.64214    5.03995
## room_typePrivate room                               -35.03587    2.31342
## room_typeShared room                                -101.65352    6.96405
## number_of_reviews                                    -0.13871    0.01337
##                                                    t value Pr(>|t|)
## (Intercept)                                         -0.753 0.451335
## bedrooms                                            19.318  < 2e-16 ***
## bathrooms                                           16.320  < 2e-16 ***
## accommodates                                        16.614  < 2e-16 ***
## review_scores_rating                                 2.837 0.004565 **
## se.neighbourhood_group_cleansedBeacon Hill          -3.660 0.000254 ***
## se.neighbourhood_group_cleansedCapitol Hill          6.920 4.89e-12 ***
## se.neighbourhood_group_cleansedCascade              10.246  < 2e-16 ***
## se.neighbourhood_group_cleansedCentral Area          0.833 0.404991
## se.neighbourhood_group_cleansedDelridge             -4.385 1.18e-05 ***
## se.neighbourhood_group_cleansedDowntown             18.538  < 2e-16 ***
## se.neighbourhood_group_cleansedInterbay             -0.015 0.987879
## se.neighbourhood_group_cleansedLake City            -1.087 0.277014
## se.neighbourhood_group_cleansedMagnolia              2.386 0.017079 *
## se.neighbourhood_group_cleansedNorthgate            -2.221 0.026357 *
## se.neighbourhood_group_cleansedOther neighborhoods  -1.984 0.047244 *
## se.neighbourhood_group_cleansedQueen Anne            7.397 1.54e-13 ***
## se.neighbourhood_group_cleansedRainier Valley       -2.745 0.006060 **
## se.neighbourhood_group_cleansedSeward Park          -0.660 0.509431
## se.neighbourhood_group_cleansedUniversity District   0.299 0.765097
## se.neighbourhood_group_cleansedWest Seattle          0.326 0.744566
## room_typePrivate room                              -15.145  < 2e-16 ***
## room_typeShared room                               -14.597  < 2e-16 ***
## number_of_reviews                                  -10.371  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.23 on 7588 degrees of freedom
##   (998 observations deleted due to missingness)
## Multiple R-squared:  0.5407, Adjusted R-squared:  0.5393
## F-statistic: 388.4 on 23 and 7588 DF,  p-value: < 2.2e-16
```





4. Result

4.1 Model choice