MA678 Midterm Project

Airbnb Pricing Data Analysis

Boston University

Ningze Zu

12/06/2018

Airbnb Pricing Data Analysis

Ningze Zu

12/06/2018

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.

2. Introduction

2.1 Background

Airbnb is a global company that founded in 2008 San Fransisco by Brian Chesky, Joe Gebbia, and Nathan Blecharczyk as AirBed & Breakfast.[1] It is an inclusive platform, providing online hospitality service and enabling hosts to list their properties and visitors to rent short-term accommodations. Recently, more and more people start using Airbnb to plan their vacation, business and homestay. Since not many guests and hosts know well about how to fix a fair price of Airbnb rental, making a price predictor that can generate a fair rental price is a good idea.

To make this pricing tool , I built a regression model that takes some features like bedrooms, accommodates as predictors and price as response. In order to optimize the final model, I did lots of exploratory data analysis to select features as model predictors.

2.2 Datasource

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

“Inside Airbnb Project”[2] is an independent and non-commercial set of tools hosted by Airbnb that enables people to explore how Airbnb is really being used in cities around the world.

The considered data is the listings dataset in Seattle contains 8625 observations with 30 independent variables. It was scrapped in October 2018. Table 1 below shows the column names of the Airbnb Seattle dataset.

Table 1

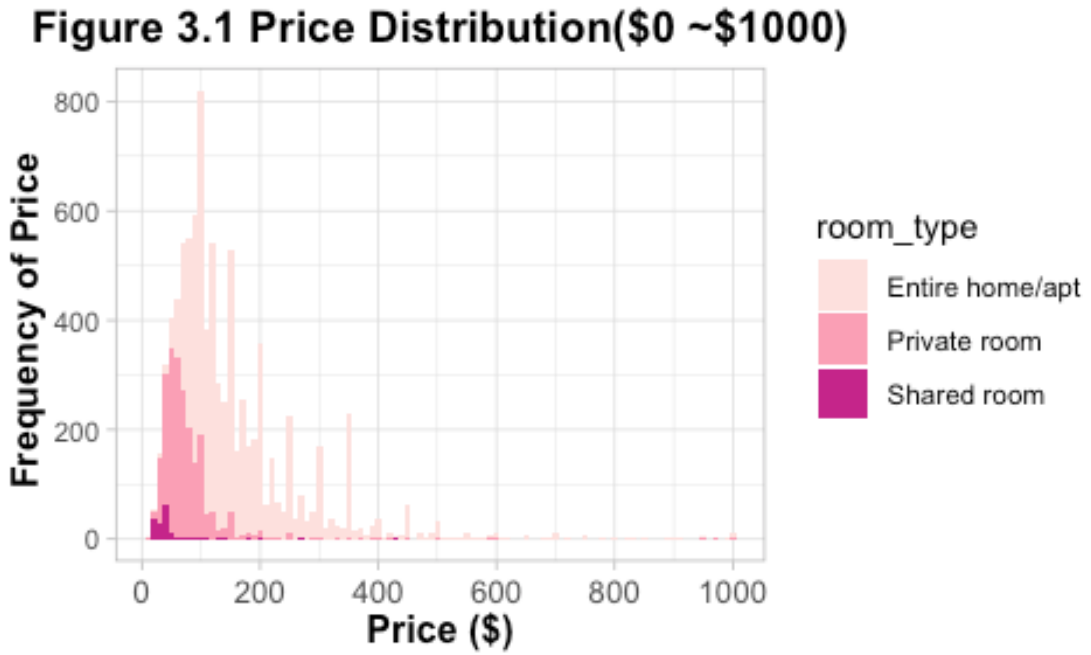
id	city	property_type	amenities	availability_365
name	state	room_type	price	number_of_reviews
host_id	zipcode	accommodates	security_deposit	review_scores_rating
host_is_superhost	country_code	bathrooms	cleaning_fee	cancellation_policy
street	latitude	bedrooms	minimum_nights	Area
neighbourhood_cleansed	longitude	beds	maximum_nights	neighbourhood

3. Exploratory Data Analysis

3.1. Price Distribution

The first thing I did is to look into the price distribution in Seattle. 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 Figure 3.1:

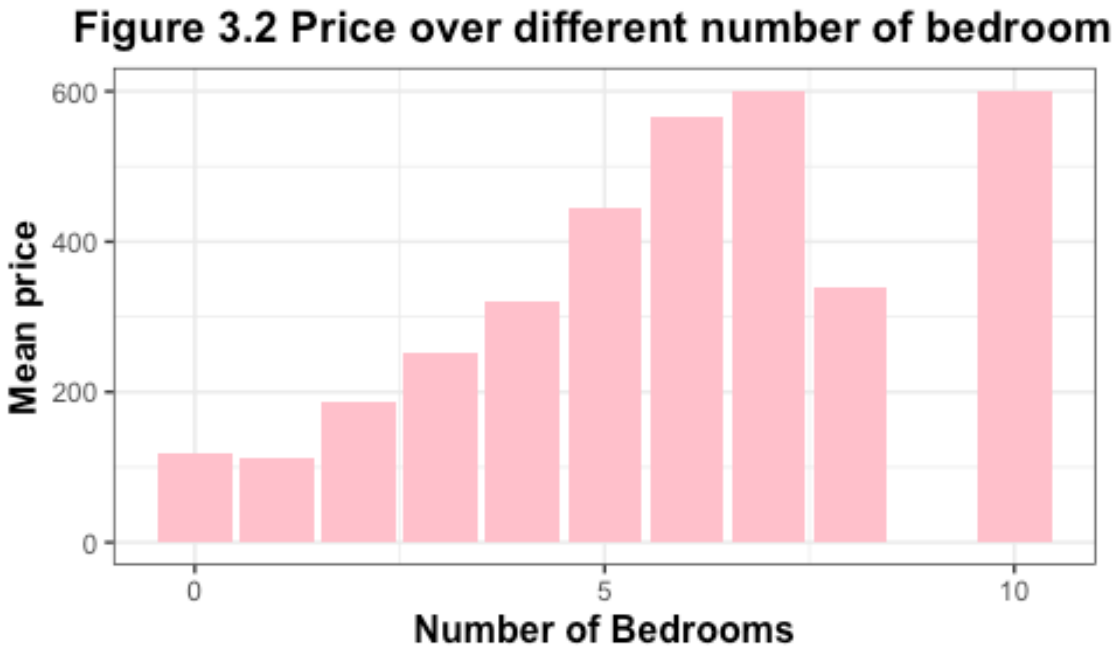
Figure 3.1 shows a high concentration of listing prices between 0 and 200 USD.



3.2. Variables

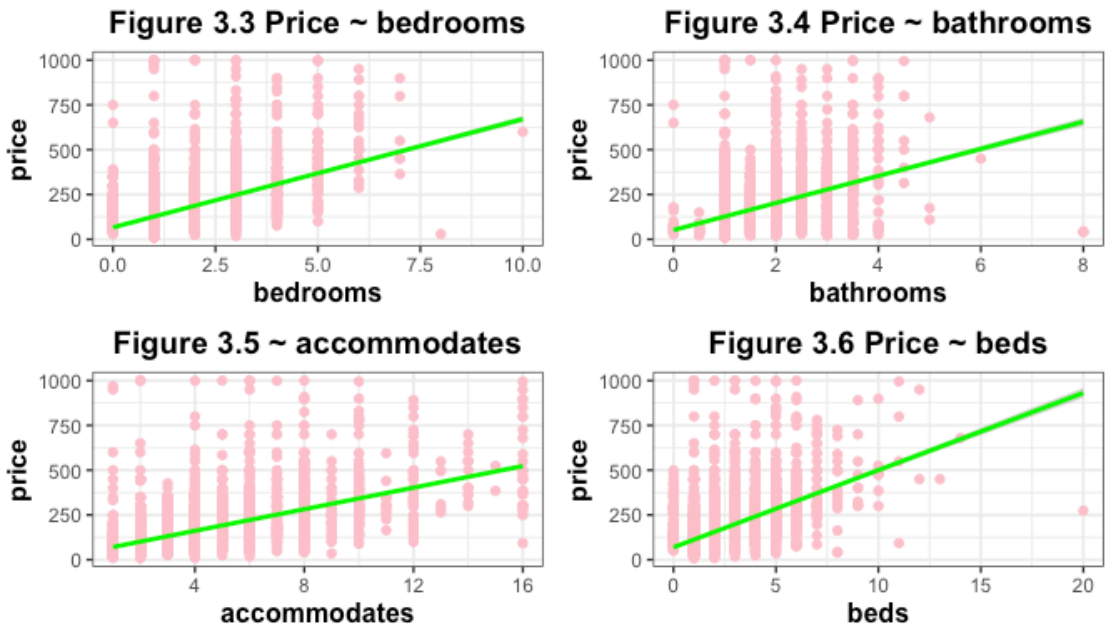
a. Bedrooms

Number of bedrooms will always be a major factor to the rental price of Airbnb. I made a barplot to show the average prices over different number of bedrooms. In this case, we can see from Figure 3.2 that there is a certain relationship between the average price of Airbnb listings and number of bedrooms.



b. Bedrooms, Bathrooms, Accommodates, Beds

Next, I did several scatterplots to show the price versus different variables such as 'bedrooms', 'bathrooms', 'accommodates'and 'beds'. In case it is hard to see if there is relationship between price and each of these vairables , I add a line to see the trend of the points. From Figure 3.3 to Figure 3.6, we can see that number of bedrooms, number of bathrooms, number of accommodates and number of beds have a certain effect on the price of Airbnb listings.



c. Property Type

Next, Figure 3.7.1 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'. To figure out more about property type, I plot a heatmap (Figure 3.7.2) to show the average room prices in different property types with different room types.

Figure 3.7.1 Count of Property Types

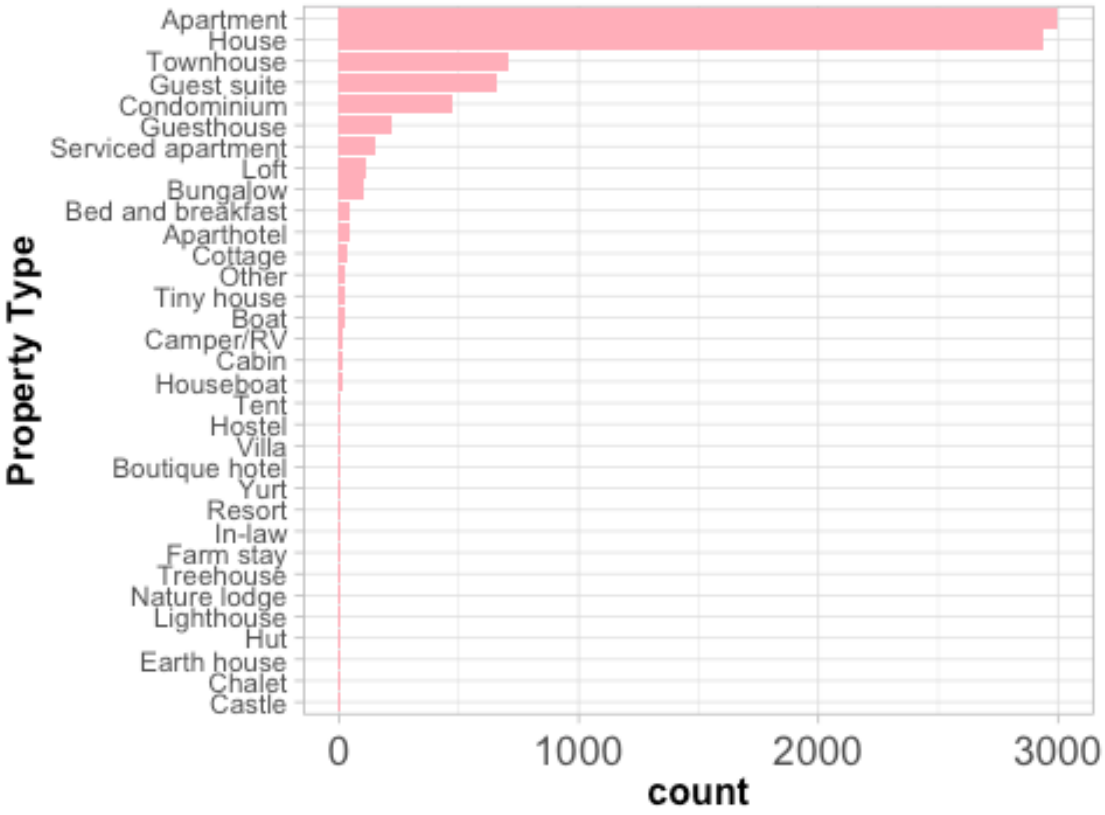
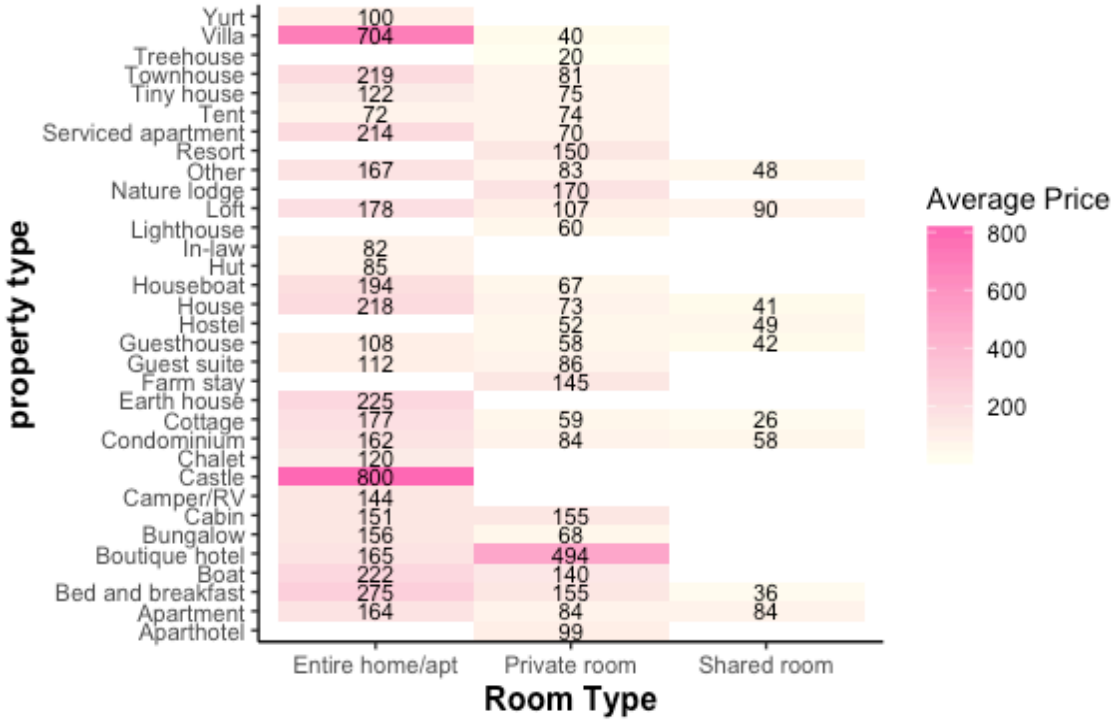


Figure 3.7.2 Property type ~ Room type



d. Room Type

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. Figure 3.8.1 shows the proportion of each of these three room type and we can see that there are nearly 74% of rooms are 'Entire Room/Apt'. Further, Figure 3.8.2 below shows that 'Entire room/apt' have a higher average price than 'Shared room' and 'Private room'.

Figure 3.8.1 Room Type Proportion

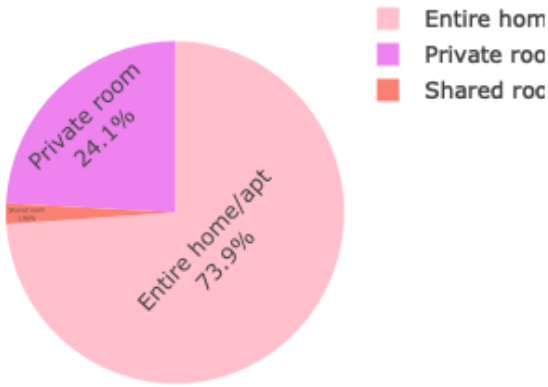
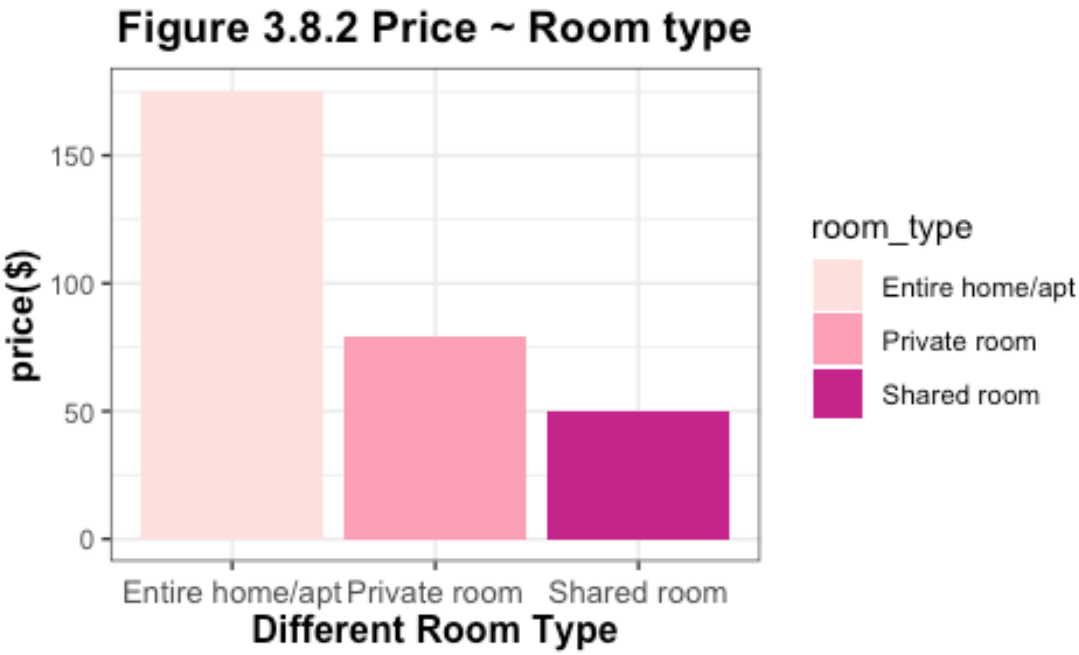


Figure 3.8.2 Price ~ Room type



e. Reviews

Next, Figure 3.9.1 and Figure 3.9.2 show the prices over review rate score and numbers of review. It can be analyzed that with the increase in number of reviews, the average rental price increase.

Figure 3.9.1 Price over Rate Scores

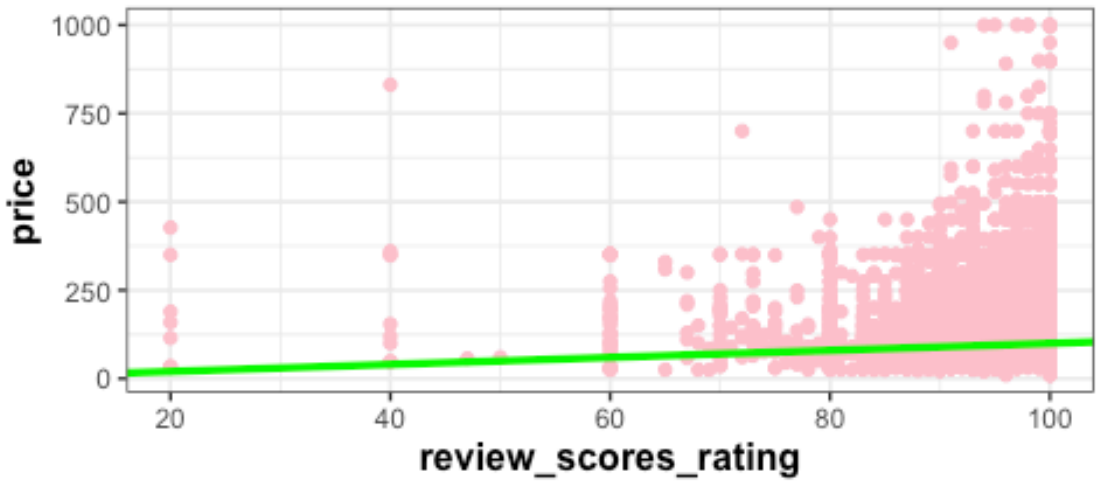
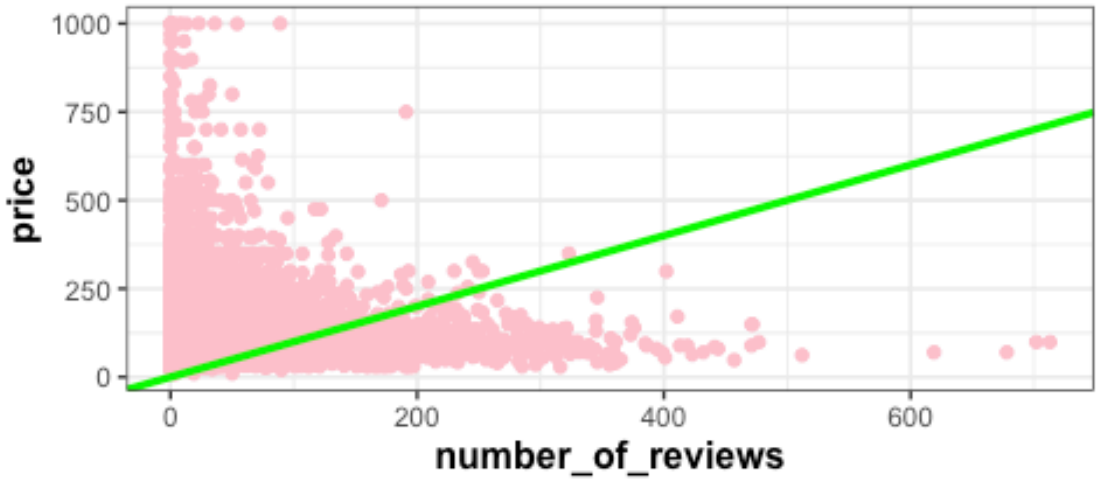


Figure 3.9.2 Price over numbers of review



f. Neighbourhood

Speaking of housing, most of the rental prices depend on the geographic feature. Houses in downtown are more expensive than those in suburb area. Further, the distance to shopping center or traffic station is also innegligible if we try to list a fair rental price.

The bar plot (Figure 3.10) below shows the average prices in different neighbourhood.

To give a better understanding of the price in different neighbourhoods, I plotted a heatmap (Figure 3.11) to show the average room prices in different neighbourhoods with different room types.

Next, from the heatmap (Figure 3.12), with the increase in the number of bedrooms, the average price of Airbnb listings in Seattle increase. Although it depends upon the neighbourhoods as well.

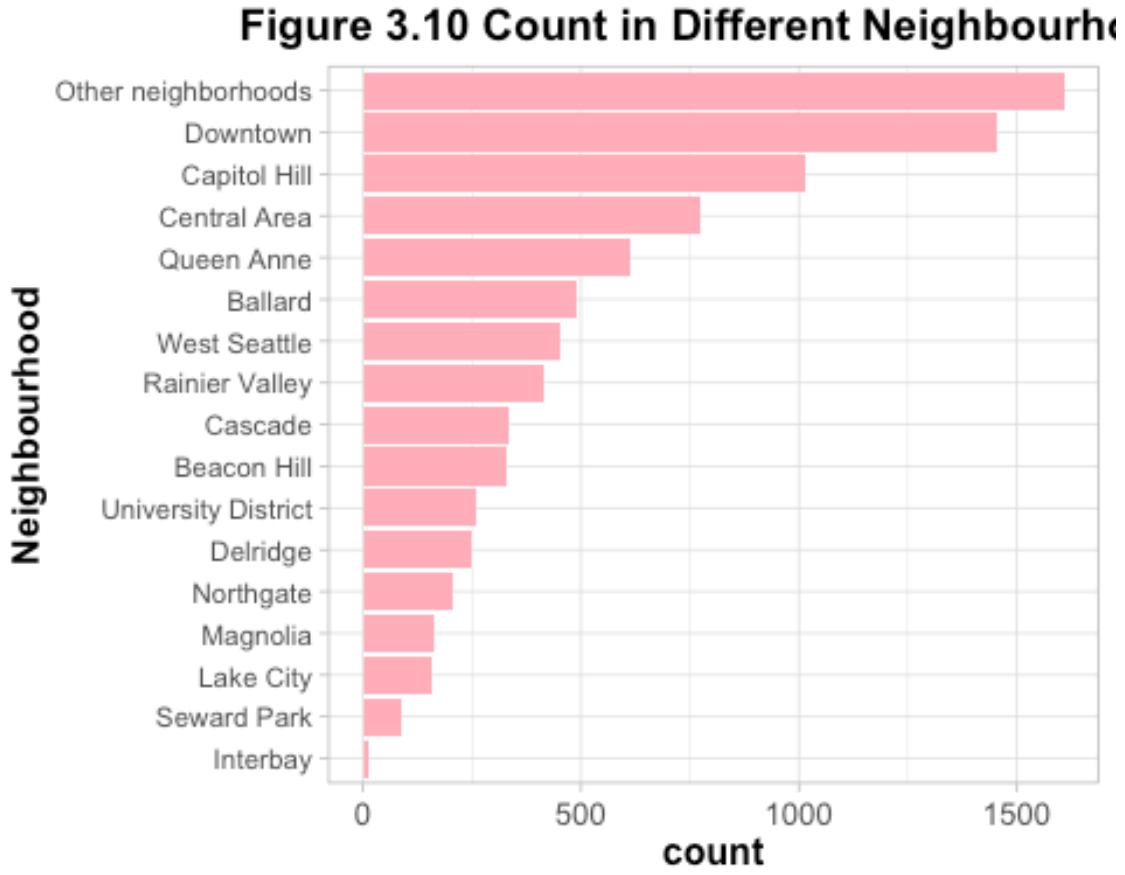


Figure 3.11 Neighbourhood ~ Room type

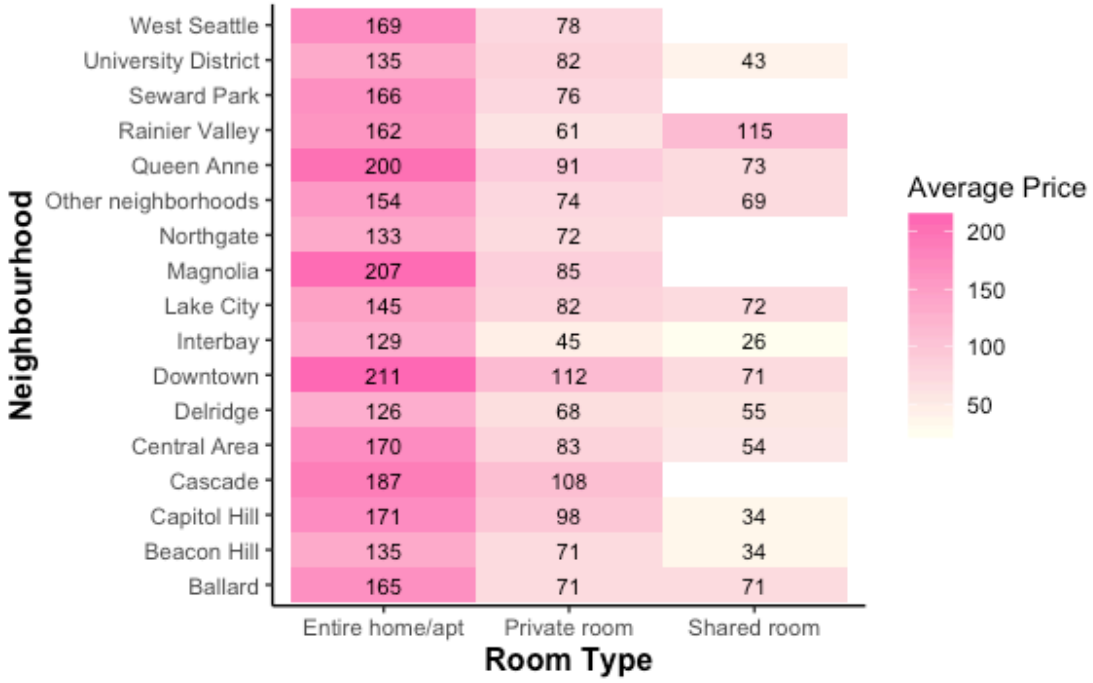
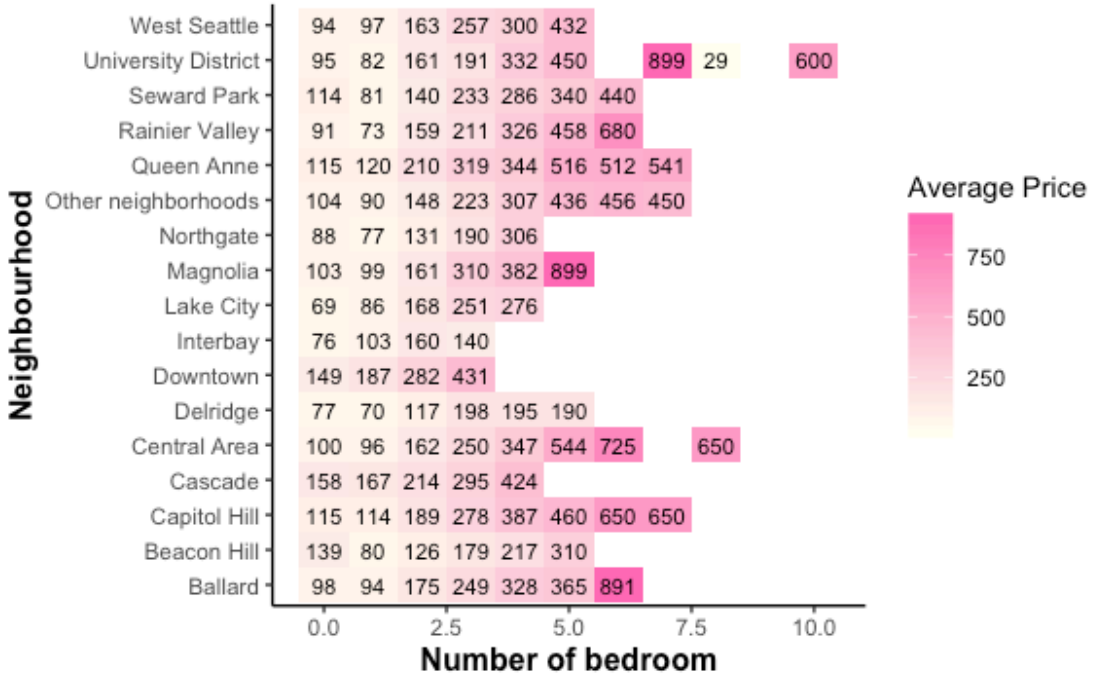
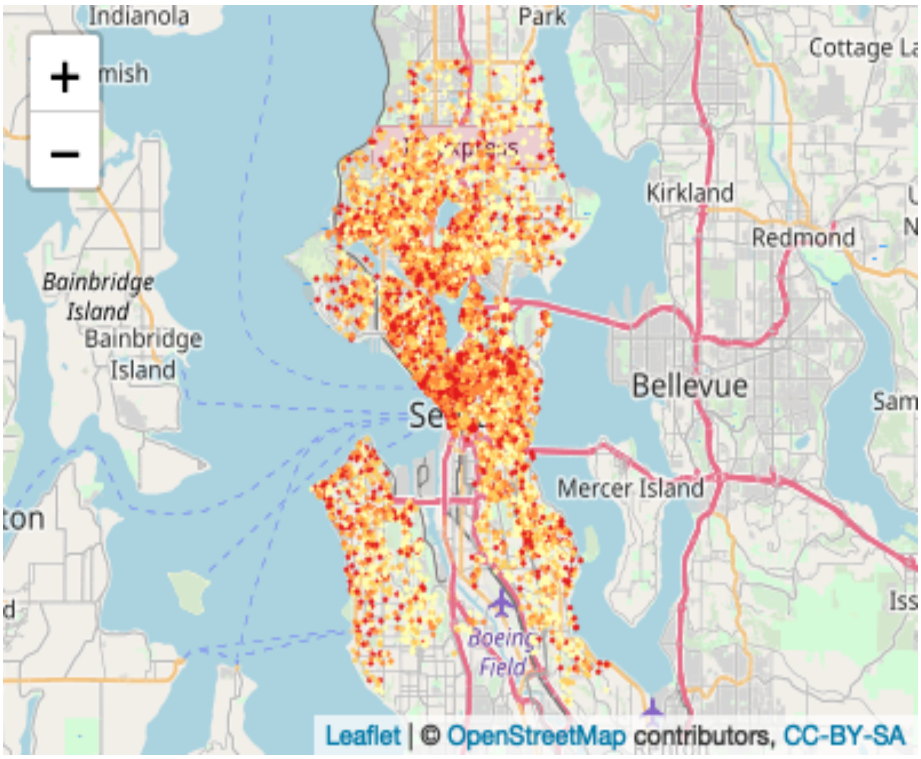


Figure 3.12 Neighbourhood ~ Bedroom



We can also see the price changes depend on neighbourhood by the mapping below: the rental prices in 'Downtown' are higher than other neighbourhoods and neighbourhoods that are far from the urban area have less listings and lower prices.



g. Others: Amenities

Word cloud below shows the most frequent amenities among the Airbnb price lower than 500 USD.

Based on the explortary data analysis, the prices of listings on Airbnb depends upon the number of bedrooms, number of bathrooms, number of accommodates, number of reviews, room type, property type, and neighbourhood.

4. Method

4.1 Model used

a. Initial Model (model 1)

First, I fit a classic linear regression model without group variabls and do a coefficient plot. Table 2.1 and 2.2 below shows the coefficient of this model and AIC & R-square value. Then, I use residual plots and marginal model plots to check the fit of my initial model. Most of the variables seemed significant. However, the R-square is lower than 0.5. From the marginal model plots, we can see the marginal relationships between the response (log(price)) and each predictors. From the residual plot, we can see that there is no non-linear relationship, which indicates that this initial model is not bad.

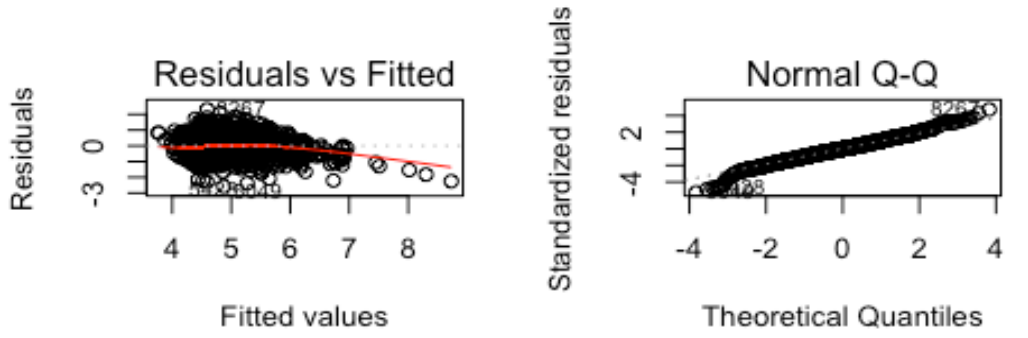
```
lm1 <- lm(log(price) ~ bedrooms + bathrooms + accommodates +
       review_scores_rating + number_of_reviews, data = airbnb_se_m)
```

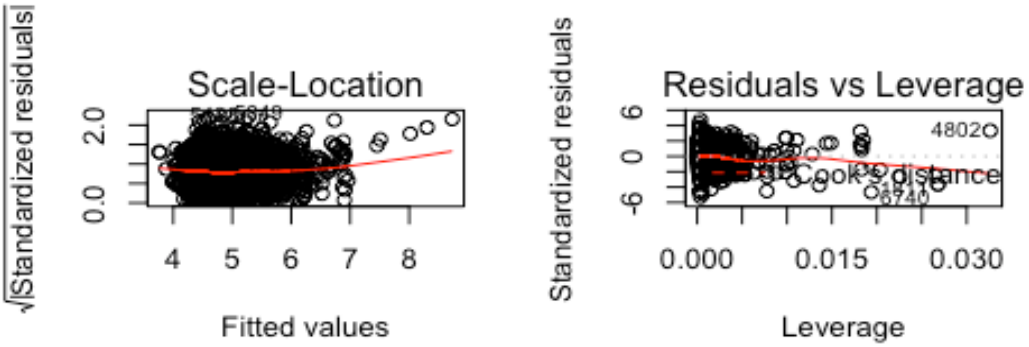
Table 2.1

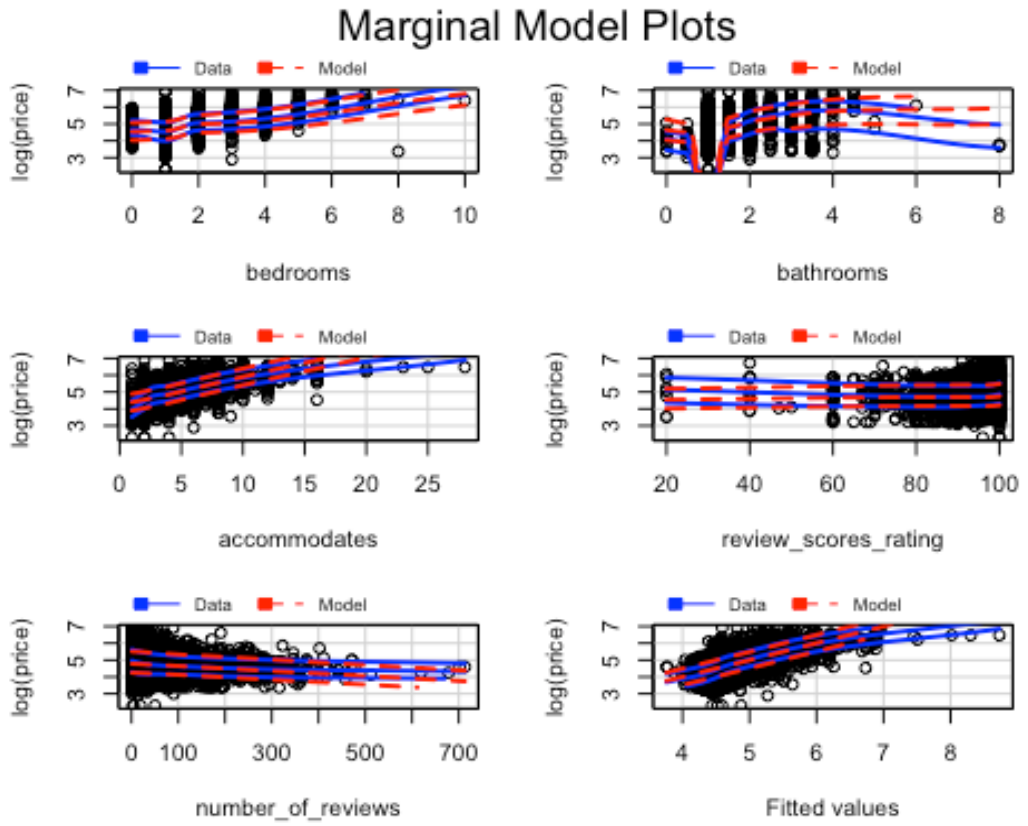
AIC	Rsquare
10717.01	0.3828073

Table 2.2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.9216564	0.0839895	46.692246	0.0000000
bedrooms	0.0360751	0.0099232	3.635436	0.0002793
bathrooms	0.0173823	0.0115909	1.499655	0.1337454
accommodates	0.1503913	0.0041087	36.602714	0.0000000
review_scores_rating	0.0031813	0.0008759	3.631833	0.0002833
number_of_reviews	-0.0011533	0.0000884	-13.046323	0.0000000


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Marginal Model Plots
Data
Model
log(price)
bedrooms
bathrooms
accommodates
review_scores_rating
number_of_reviews
Fitted values

b. Second Model (model 2)

Next, I adjusted model by adding group variable 'neighbourhood' and 'room type' and take log of price. Most variables seemed significant: bedrooms, bathrooms, accommodates, review_scores_rating, room_type, numbers_of_reviews. Table 3.1 and 3.2 below show part of coefficients of the second model and AIC & R-square value. R-square is over 0.5 and p-value is pretty small. Also, from the residual plot we can see that there is no non-linear relationships and it is much better than the residual plot of model 1. Therefore, this model is slightly better than the initial model in most cases.

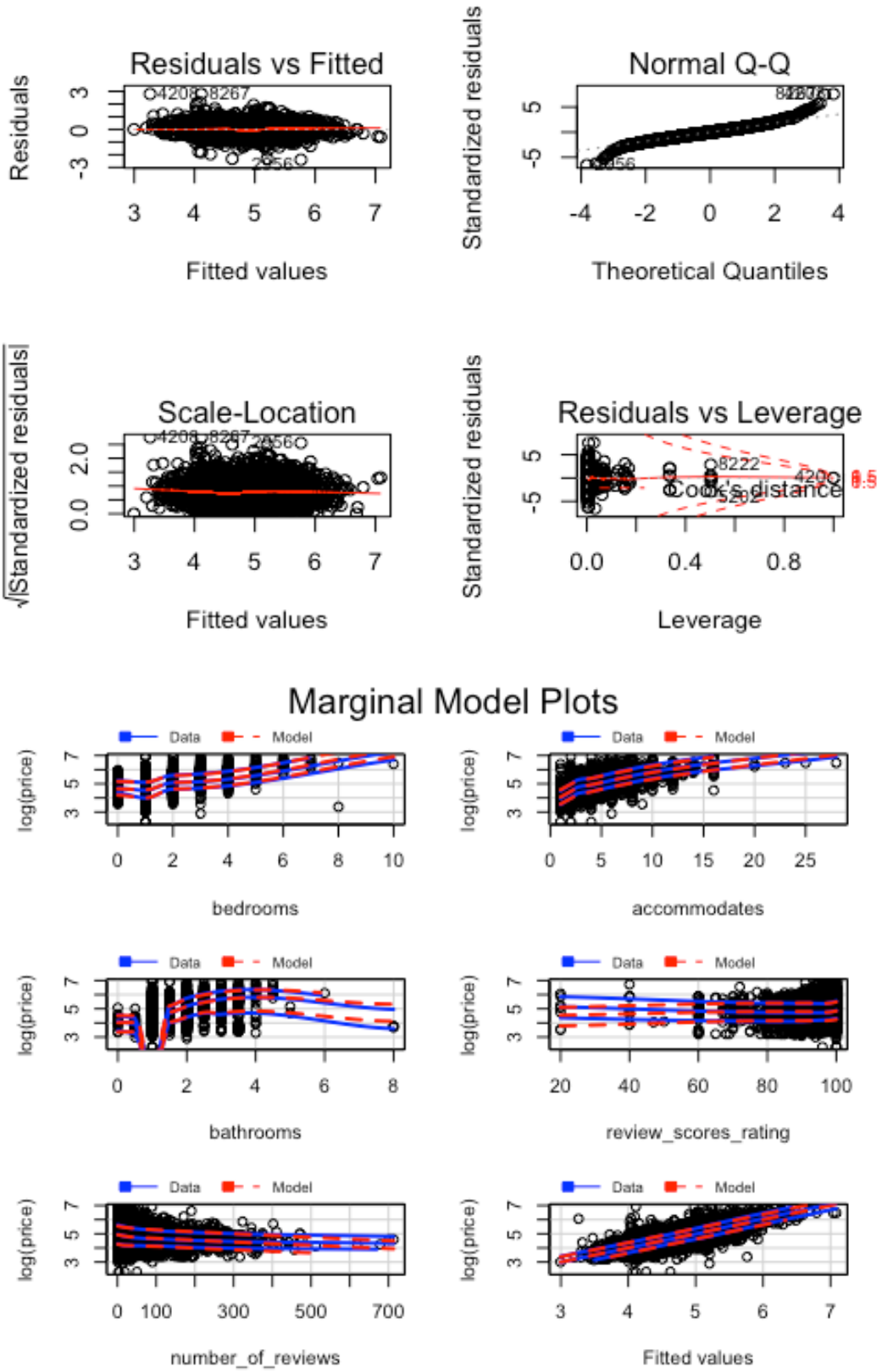
```
lm2 <- lm(log(price) ~ bedrooms + accommodates + bathrooms + review_scores_ra
ting +
          neighbourhood +property_type + room_type + number_of_reviews, data
= airbnb_se_m)
```

Table 3.1

AIC	Rsquare
6663.859	0.6398352

Table 3.2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.2534481	0.0940070	45.2460677	0.0000000
bedrooms	0.1715835	0.0094702	18.1182585	0.0000000
accommodates	0.0682009	0.0049704	13.7213338	0.0000000
bathrooms	0.1089394	0.0095770	11.3751552	0.0000000
review_scores_rating	0.0045005	0.0006909	6.5135206	0.0000000
neighbourhoodBeacon Hill	-0.1192375	0.0283200	-4.2103585	0.0000258
neighbourhoodCapitol Hill	0.1821312	0.0223729	8.1407207	0.0000000
neighbourhoodMagnolia	0.0845913	0.0358358	2.3605228	0.0182745
… … …	-0.1637295	0.1158302	-1.4135305	0.1575410
property_typeHouseboat				
property_typeHut	-0.6365067	0.3790279	-1.6793135	0.0931323
property_typeIn-law	-0.5253951	0.2723222	-1.9293148	0.0537292
… … …	-0.5068674	0.3786089	-1.3387625	0.1806883
property_typeLighthouse				
room_typePrivate room	-0.4638434	0.0138173	-33.5697652	0.0000000
room_typeShared room	-1.0892515	0.0370162	-29.4263421	0.0000000
number_of_reviews	-0.0009313	0.0000693	-13.4342631	0.0000000
bedrooms:accommodates	-0.0042487	0.0011859	-3.5828414	0.0003420


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Marginal Model Plots
Data
Model
log(price)
bedrooms
accommodates
bathrooms
review_scores_rating
number_of_reviews
Fitted values

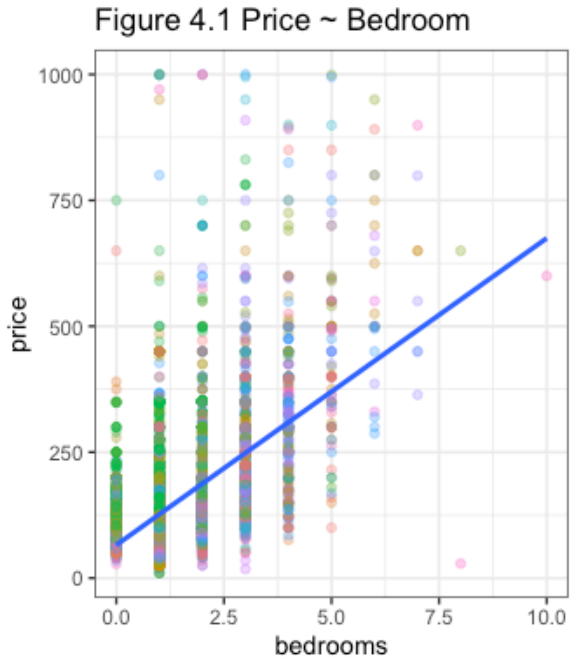
Model Comparison

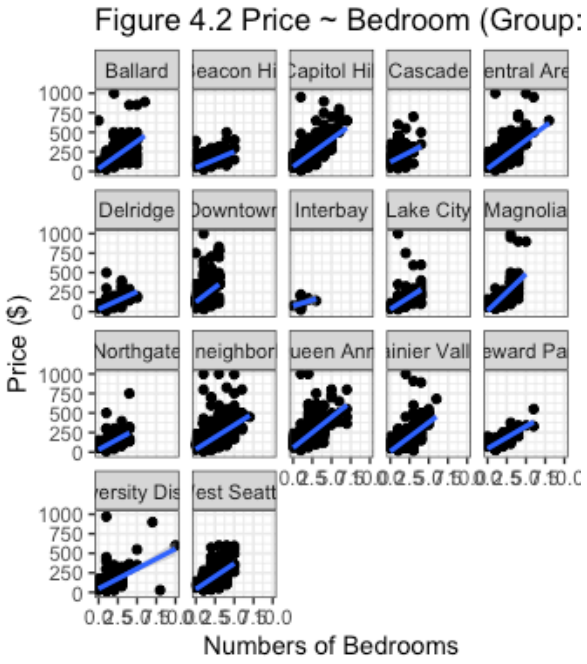
According to the Anova test, model 2 is better than model 1 because of the lower residual diviance and small p-valiue. Table 4 below show the result of anova test.

Table 4

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
7606	1818.333	NA	NA	NA	NA
7558	1052.749	48	765.5837	114.5074	0

Figure 4.1 and 4.2 give a visualized comparison of price changes over number of bedrooms and how it changes within different neighbourhoods.



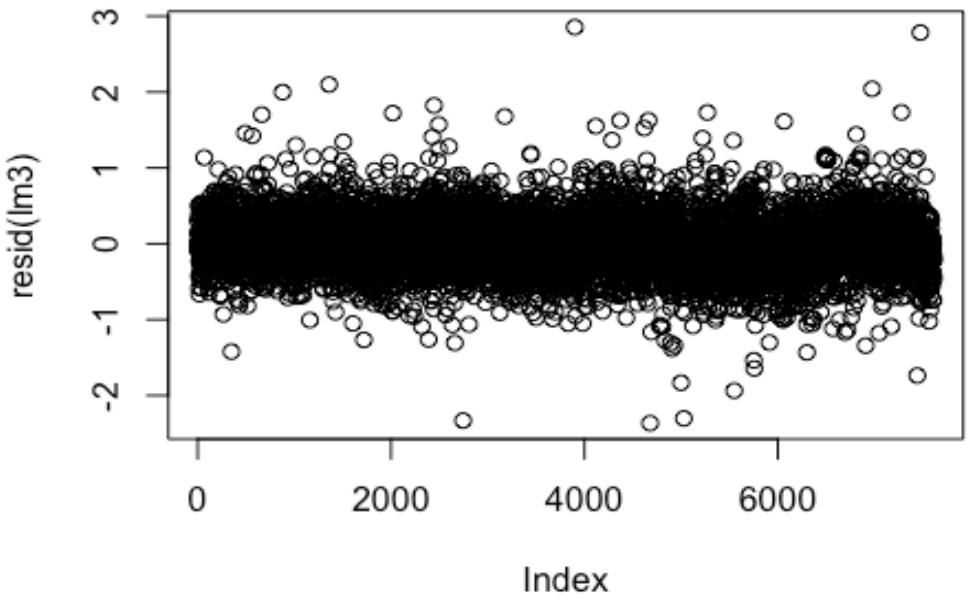


b. Third Model (model 3)

For the third model, I fitted a varying interpect model by grouping level vairables "neighbourhood" with 'lmer'. The residual plot looks pretty random. However, both the deviance and AIC value are bigger than the second model.

```
lm3 <- lmer(log(price) ~ bedrooms + bathrooms + accommodates + review_scores_
    rating + (1 | neighbourhood) + room_type + number_of_reviews, data = airb
nb_se_m)
display(lm3)
```

```
## lmer(formula = log(price) ~ bedrooms + bathrooms + accommodates +
##     review_scores_rating + (1 | neighbourhood) + room_type +
##     number_of_reviews, data = airbnb_se_m)
##                       coef.est coef.se
## (Intercept)            3.91     0.08
## bedrooms               0.16     0.01
## bathrooms              0.11     0.01
## accommodates           0.05     0.00
## review_scores_rating   0.00     0.00
## room_typePrivate room -0.46     0.01
## room_typeShared room  -1.14     0.04
## number_of_reviews      0.00     0.00
##
## Error terms:
##  Groups        Name        Std.Dev.
##  neighbourhood (Intercept) 0.18
##  Residual                  0.38
## ---
## number of obs: 7612, groups: neighbourhood, 17
## AIC = 6949.3, DIC = 6784.1
## deviance = 6856.7
```



5. Result

5.1 Model choice

I began to predict the rental price by fitting a linear regression and checking the residuals plots. The initial model was fitted by using bedrooms, bathrooms, accommodates, review_scores_rating and number_of_reviews as predictors and log(price) as reponse. Except 'bathrooms', all other variables seemed significant. The residual plots did not show any non-linear relationships. From the marginal plots, we can see that 'bedrooms' and 'accommodates' show higher importance to log(price) rather than other predictors. However, the R-square is lower than 0.5 and AIC is 10717.01.

Then, I fitted second model by adding group vairables 'neighbourhood', 'property_type' and 'room type'. Most of the variables seemed significant and the residual plots is better than the initial models. Same as the inital model, the marginal model plots show higher importance of variable 'bedrooms' and 'accommocdates' rather than other variables. R-square is 0.64 which is higher than the R-square of initial model. In addition, the AIC decrease by nearly 4000, which is 6652.942. Further, based on the result of Anova test, the second model has lower residual deviance than the initial model. Therefore, second model is much better than the initial model.

Lastly, I fitted a varying interpect model by grouping level vairable "neighbourhood" with 'lmer'. However, both deviance and AIC value are bigger than the second model.

Based on the results I got from these three model, I chose the second model as the final model to predict the rental price of Airbnb listings in Seattle.

5.2 Interpretations

According to the summary of second model, percentage changes in price were driven by the following variables: room_type, property_type, neighborhood, number_of_reviews, review_scores_rating, accommodates, bedrooms and bathrooms.

5.3 Model checking

Predictd vs. Actual Plot

By using the second model to predict the rental price in Seattle, I splited the dataset into 2 part. 80% of the dataset was used as train data, and 20% of the dataset was used as test data to for predicting. After I predicted the price of test data and compare with the origin data. The lowest RMSE of the prediction came out to be USD 63 and the MAE is around 39. On a closer look at the prediction error, I made two plots, one is the price distribution of

test data with origin price （Figure5.1), the other one is the price distribution of test data with predict price (Figure 5.2). The difference between these two plots is very apparent.

RMSE	MAE
63.7325	39.3258

6. Discussion

6.1 Implications

By using the Airbnb listings dataset in Seattle, I builded a linear regression model to predict the the rental price. This pricing tool could be applicable since hosts can make use of available data and get a reference about how to fix a fair rental price. Guests could use it to check if the room they would like to rent has a reasonable price.

6.2 Limitation

Airbnb listings contain different kinds of rooms. There could be two rooms with exactly same features but totally different prices. Price not only depends on the features of room, but also other factors such as furnitures, new or old houses, amenities or level of the rooms.

Since there is no classification system for Airbnb like star rating for Hotels, predicting the rental price only depends upon the features in historical data is inaccurate sometimes.

6.3 Future direction

Because of the limit of time, some features were not used in this project. In future studies, I may include more features such as 'minimum_nights', 'maximum_nights', 'cleansing_fee' and time vairables to optimize the model's accuracy. Second, I would do some text mining by using some text based features such as 'amenities', 'name' and 'reviews' and convert some texts into features. Further, I might compare the rental price of Airbnb listings in Seattle with the price of Hotel in different neighbourhood.

7. Reference

https://www.airbnb.com/diversity

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

https://github.com/ruchigupta19/Boston-Airbnb-data-analysis