Final report

Zhaobin Liu

2018 12 12

Abstract

The following analysis is about the Airbnb dataset. I choose the Las_vegas dataset from http://tomslee.net/airbnb-data-collection-get-the-data. I will mainly focus on analyzing some significant variables, such as price, rating and the number of bedrooms. It will be important to know which factor will affect the most on each other. There are six models building in this analysis, and I will interpret each of them to pick the better model to give some useful information of the dataset.

Introduction

Background

Nowadays, choosing airbnb over hotel has been more frequencies because of its flexibility, price and many other factors. We all want to pick the best choice for our trip. Then we need to learn how to look at the data and how the factors relate with each others. Sometimes, data are deceptive and misleading in some ways. Building models can avoid many obvious problems. Thus, we can get a sense of which factors are the most persuasive in the dataset.

Previous work

There are a lot of dataset on the website. Many people have already collected some specific variables, such as room_id, reviews, overall_satisfaction of airbnb apartments. These data can give people a basic information of airbnb apartments.

Method

Data source

The airbnb information of Las vegas dataset will be from the Airbnb website: http://tomslee.net/airbnb-data-collection-get-the-data. I am using R to combine all of five separate csv files into one csv file in order to prepare for analyzing the model doing the EDA.

The variables I am using in the dataset are: room_id, host_id, room_type, neighborhood, reviews, overall_satisfaction, accommodates, bedrooms, price, latitude and longitude. I am going to the study two interesting variables: overall_satisfaction and the price. The former is the average rating (1-5) that the listing has received from those visitors who left a review. The latter is the price for a night stay.

Model used

```
library(arm)

# GLM for Overall_satisfaction
rating_reg <- glm(overall_satisfaction ~ factor(room_type) + log(price)  +
                    (accommodates) + bedrooms,data = overall_satisfaction_data)

# random slope
rating_reg_2 <- lmer(overall_satisfaction ~ factor(room_type) + (accommodates) + bedrooms
                     + (0 + log(price) | neighborhood),data = overall_satisfaction_data)

# random slope and intercept with interaction
rating_reg_3 <- lmer(overall_satisfaction ~ factor(room_type)*accommodates + bedrooms + log(price) +
                      (1 + log(price) | neighborhood),data = overall_satisfaction_data)

# GLM for Price
price_reg <- glm(log(price) ~ factor(room_type) + overall_satisfaction + accommodates +
                   bedrooms, data = overall_satisfaction_data)

# random intercept
price_reg_2 <- lmer(log(price) ~ factor(room_type) + overall_satisfaction + accommodates
                     + bedrooms + (1|neighborhood), data = overall_satisfaction_data)

# random intercept and slope
price_reg_3 <- lmer(log(price) ~ factor(room_type) + accommodates + bedrooms +
                      (1 + overall_satisfaction|neighborhood), data = overall_satisfaction_data)
```

Result

Model choice and interpretation

```
display(price_reg)
```

```
## glm(formula = log(price) ~ factor(room_type) + overall_satisfaction +
##     accommodates + bedrooms, data = overall_satisfaction_data)
##                               coef.est coef.se
## (Intercept)                    3.21     0.05
## factor(room_type)Private room -0.78     0.01
## factor(room_type)Shared room  -1.40     0.02
## overall_satisfaction           0.22     0.01
## accommodates                   0.06     0.00
## bedrooms                       0.12     0.00
## ---
##   n = 16041, k = 6
##   residual deviance = 3414.7, null deviance = 10366.2 (difference = 6951.5)
##   overdispersion parameter = 0.2
##   residual sd is sqrt(overdispersion) = 0.46
```

According to the model, we can see all the coefficients are statistically significant. With each unit increase of overall_satisfaction, log(price) will increase by 0.22. With each unit increase of accommodates and bedrooms, log(price) will increase by 0.06 and 0.12. Private room has 0.78 lower weighted price than the Entire home/apt. Shared room has 1.4 lower weighted price than the Entire home/apt.

```
display(price_reg_2)
```

```
## lmer(formula = log(price) ~ factor(room_type) + overall_satisfaction +
##     accommodates + bedrooms + (1 | neighborhood), data = overall_satisfaction_data)
##                               coef.est coef.se
## (Intercept)                    3.01     0.06
## factor(room_type)Private room -0.71     0.01
## factor(room_type)Shared room  -1.35     0.02
## overall_satisfaction           0.23     0.01
## accommodates                   0.05     0.00
## bedrooms                       0.16     0.00
##
## Error terms:
##  Groups       Name        Std.Dev.
##  neighborhood (Intercept) 0.12
##  Residual                 0.45
## ---
## number of obs: 16041, groups: neighborhood, 31
## AIC = 19862.7, DIC = 19752.7
## deviance = 19799.7
```

According to the model, all the coefficients are statistically significant. With each increase unit of overall_satisfaction, log(price) will increase by 0.23. With each increase unit of accommodates and bedrooms, log(price) will increase by 0.05 and 0.16. Private room has 0.71 lower weighted price than the Entire home/apt. Shared room has 1.35 lower weighted price than the price of Entire home/apt. The neighborhood variation has the standard deviation of 0.45 and the intercept of 0.12.

```
display(price_reg_3)
```

```
## lmer(formula = log(price) ~ factor(room_type) + accommodates +
##     bedrooms + (1 + overall_satisfaction | neighborhood), data = overall_satisfaction_data)
##                               coef.est coef.se
## (Intercept)                    4.16     0.02
## factor(room_type)Private room -0.70     0.01
## factor(room_type)Shared room  -1.34     0.02
## accommodates                   0.05     0.00
## bedrooms                       0.16     0.00
##
## Error terms:
##  Groups       Name                 Std.Dev. Corr
##  neighborhood (Intercept)          1.30
##               overall_satisfaction 0.26     -1.00
##  Residual                          0.45
## ---
## number of obs: 16041, groups: neighborhood, 31
## AIC = 19764.2, DIC = 19666.6
## deviance = 19706.4
```

All the coefficients are tatistically significant. All signs do not change comparing to the previous model. The residual is still 0.45. the slope of the overall_satisfaction is 0.26. There is correlation with intercept of -1.

```
display(rating_reg)
```

```
## glm(formula = overall_satisfaction ~ factor(room_type) + log(price) +
##     (accommodates) + bedrooms, data = overall_satisfaction_data)
##                               coef.est coef.se
```

```
## (Intercept)                    4.34     0.02
## factor(room_type)Private room  0.11     0.01
## factor(room_type)Shared room  -0.01     0.02
## log(price)                     0.10     0.01
## accommodates                   0.00     0.00
## bedrooms                      -0.02     0.00
## ---
##   n = 16041, k = 6
##   residual deviance = 1622.3, null deviance = 1675.0 (difference = 52.8)
##   overdispersion parameter = 0.1
##   residual sd is sqrt(overdispersion) = 0.32
```

According to the model, we can see all the coefficients are statistically significant except the shared room coefficient. With each unit increase of log(price), overall_satisfaction will increase by 0.1. Accommodates are not affecting the model. With each unit increase of bedrooms, overall_satisfaction will decrease by 0.02. Private room has 0.11 higher weighted rating than the Entire home/apt. Shared room has 0.01 lower weighted rating than the Entire home/apt.

```
display(rating_reg_2)
```

```
## lmer(formula = overall_satisfaction ~ factor(room_type) + (accommodates) +
##     bedrooms + (0 + log(price) | neighborhood), data = overall_satisfaction_data)
##                              coef.est coef.se
## (Intercept)                   4.35     0.02
## factor(room_type)Private room  0.09     0.01
## factor(room_type)Shared room  -0.02     0.02
## accommodates                   0.00     0.00
## bedrooms                      -0.04     0.00
##
## Error terms:
##  Groups       Name       Std.Dev.
##  neighborhood log(price) 0.11
##  Residual                0.31
## ---
## number of obs: 16041, groups: neighborhood, 31
## AIC = 8675.8, DIC = 8576.9
## deviance = 8619.3
```

According to the model, we can see all the coefficients are statistically significant except the shared room coefficient. Accommodates are not affecting the model. With each unit increase of bedrooms, overall_satisfaction will decrease by 0.04. Private room has 0.09 higher weighted rating than the Entire home/apt. Shared room has 0.02 lower weighted rating than the Entire home/apt. The neighborhood variation has the standard deviation of 0.31 and the intercept of 0.11.

```
display(rating_reg_3)
```

```
## lmer(formula = overall_satisfaction ~ factor(room_type) * accommodates +
##     bedrooms + log(price) + (1 + log(price) | neighborhood),
##     data = overall_satisfaction_data)
##                                          coef.est coef.se
## (Intercept)                               4.43     0.06
## factor(room_type)Private room             0.17     0.01
## factor(room_type)Shared room             -0.07     0.02
## accommodates                              0.00     0.00
## bedrooms                                 -0.04     0.00
## log(price)                                0.10     0.01
```

```
## factor(room_type)Private room:accommodates -0.03     0.00
## factor(room_type)Shared room:accommodates   0.01     0.00
##
## Error terms:
##  Groups       Name        Std.Dev. Corr
##  neighborhood (Intercept) 0.30
##               log(price)  0.06     -0.99
##  Residual                 0.31
## ---
## number of obs: 16041, groups: neighborhood, 31
## AIC = 8178.8, DIC = 8015.9
## deviance = 8085.3
```

All the coefficients are statistically significant. Some coefficients are changing slightly due to the effect of the interaction. The AIC and DIC are lower than the previous model that this model should be a better fit.

Model checking

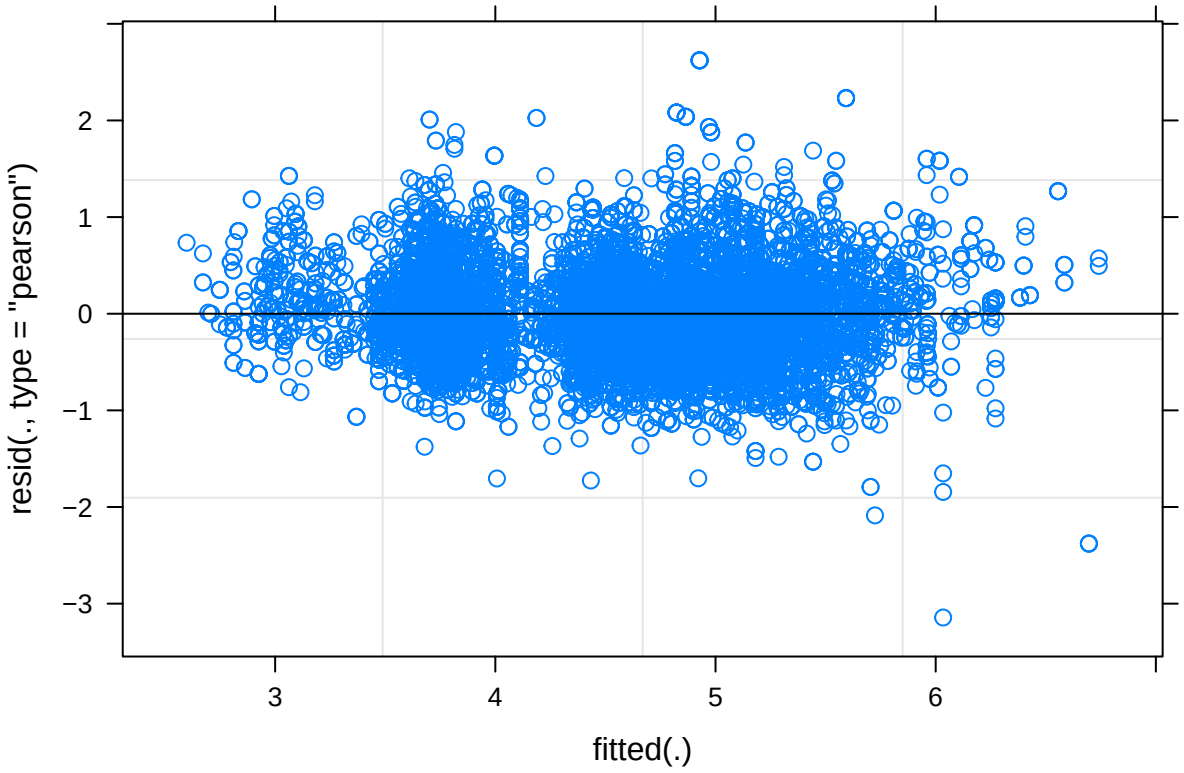
```
anova(price_reg_3,price_reg_2,price_reg)
```

```
## refitting model(s) with ML (instead of REML)
```

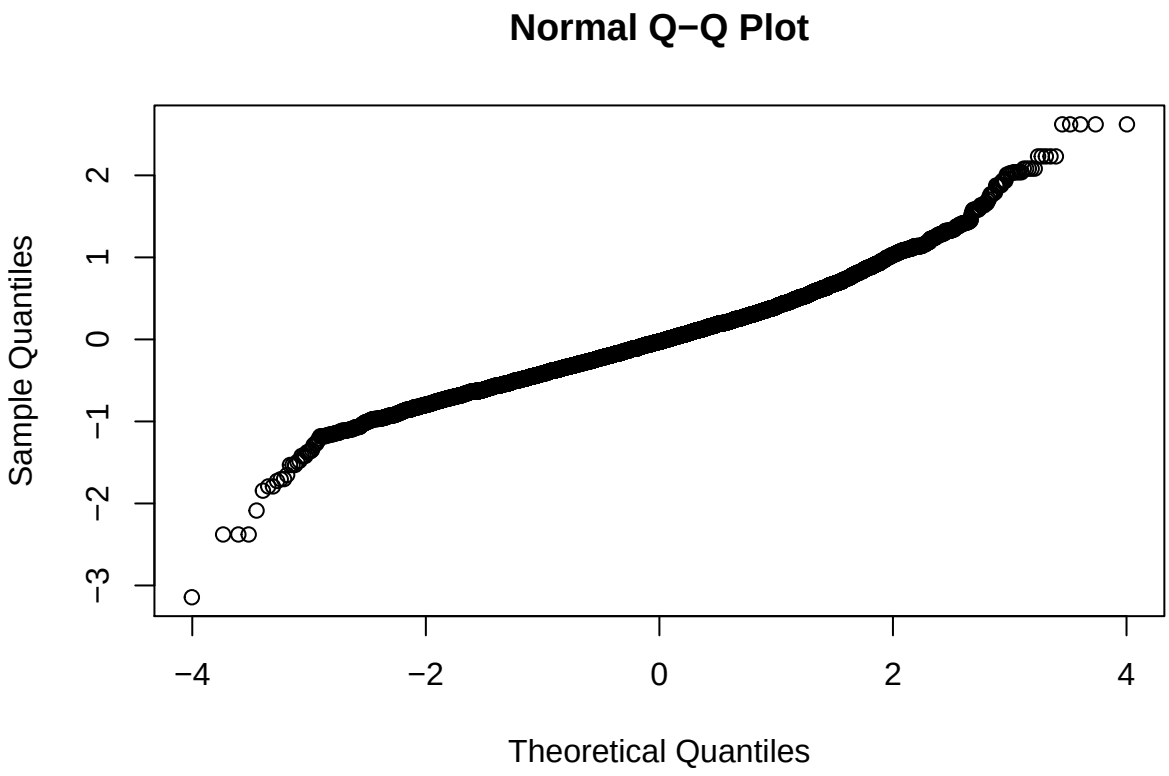
```
## Data: overall_satisfaction_data
## Models:
## price_reg: log(price) ~ factor(room_type) + overall_satisfaction + accommodates +
## price_reg:     bedrooms
## price_reg_2: log(price) ~ factor(room_type) + overall_satisfaction + accommodates +
## price_reg_2:     bedrooms + (1 | neighborhood)
## price_reg_3: log(price) ~ factor(room_type) + accommodates + bedrooms + (1 +
## price_reg_3:     overall_satisfaction | neighborhood)
##             Df   AIC   BIC   logLik deviance   Chisq Chi Df Pr(>Chisq)
## price_reg    7 20720 20774 -10353.1    20706
## price_reg_2  8 19816 19877  -9899.8    19800 906.415      1  < 2.2e-16 ***
## price_reg_3  9 19724 19794  -9853.2    19706  93.317      1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that price_reg_3 is a better fit comparing to the other two model with lower AIC and BIC values.

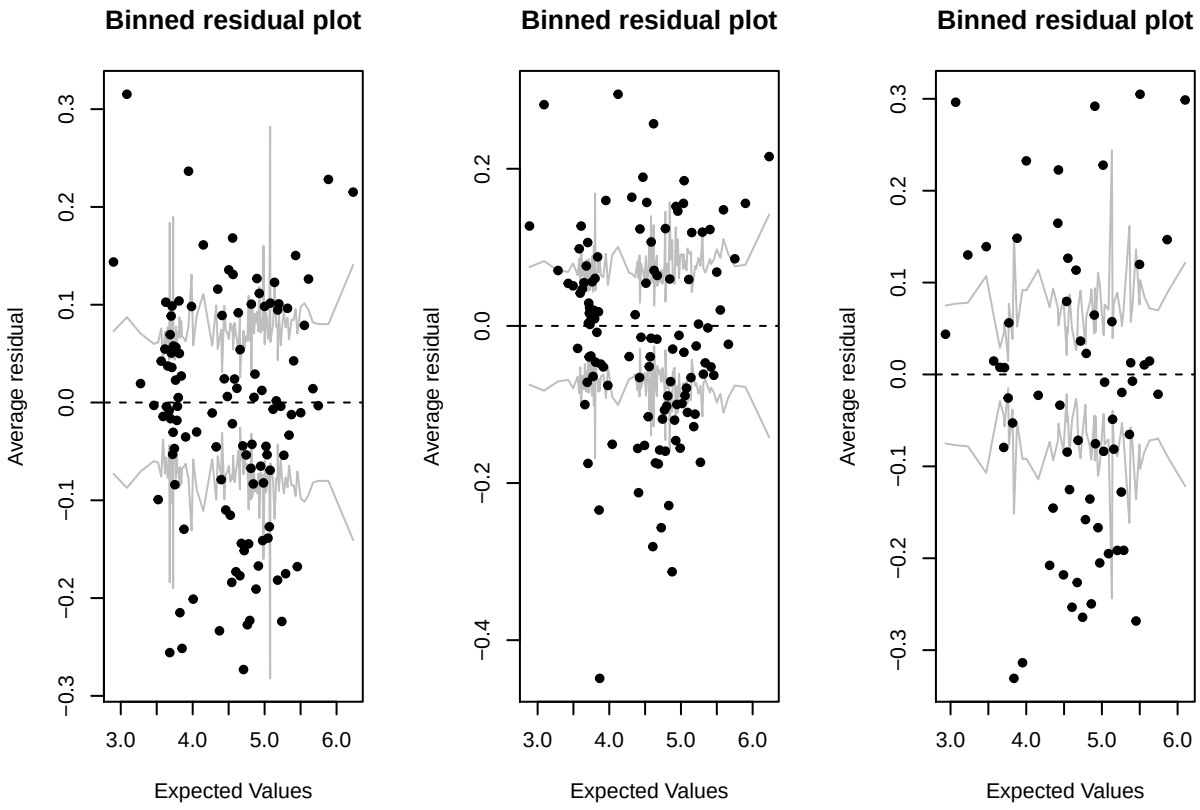
```
library(arm)
plot(price_reg_3)
```




```
qqnorm(resid(price_reg_3))
```




```
par(mfrow=c(1,3))
binnedplot(fitted(price_reg_3),residuals(price_reg_3, type="response"))
binnedplot(fitted(price_reg_2),residuals(price_reg_2, type="response"))
binnedplot(fitted(price_reg),residuals(price_reg, type="response"))
```



The resid plot and qqplot which looks like normal of the third model of Price seems fine. However, when we compare three binnedplots of all models of price, the first two graphs do not show the significant difference. There are still decent amounts of points outside of the range.

```
anova(rating_reg_3,rating_reg_2,rating_reg)
```

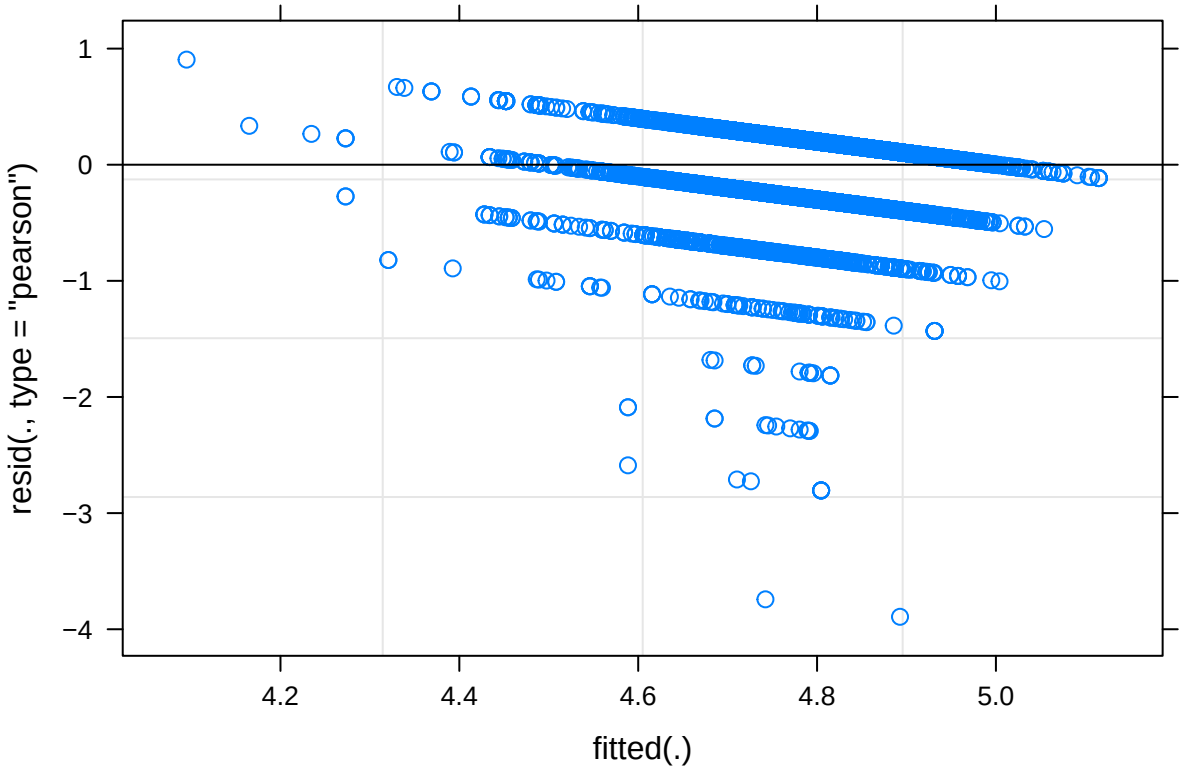
```
## refitting model(s) with ML (instead of REML)
```

```
## Data: overall_satisfaction_data
## Models:
## rating_reg_2: overall_satisfaction ~ factor(room_type) + (accommodates) + bedrooms +
## rating_reg_2:     (0 + log(price) | neighborhood)
## rating_reg: overall_satisfaction ~ factor(room_type) + log(price) + (accommodates) +
## rating_reg:     bedrooms
## rating_reg_3: overall_satisfaction ~ factor(room_type) * accommodates + bedrooms +
## rating_reg_3:     log(price) + (1 + log(price) | neighborhood)
##              Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## rating_reg_2  7 8633.3 8687.1 -4309.7   8619.3
## rating_reg    7 8781.1 8834.9 -4383.6   8767.1   0.00      0          1
## rating_reg_3 12 8109.3 8201.5 -4042.7   8085.3 681.77      5     <2e-16
##
## rating_reg_2
## rating_reg
## rating_reg_3 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

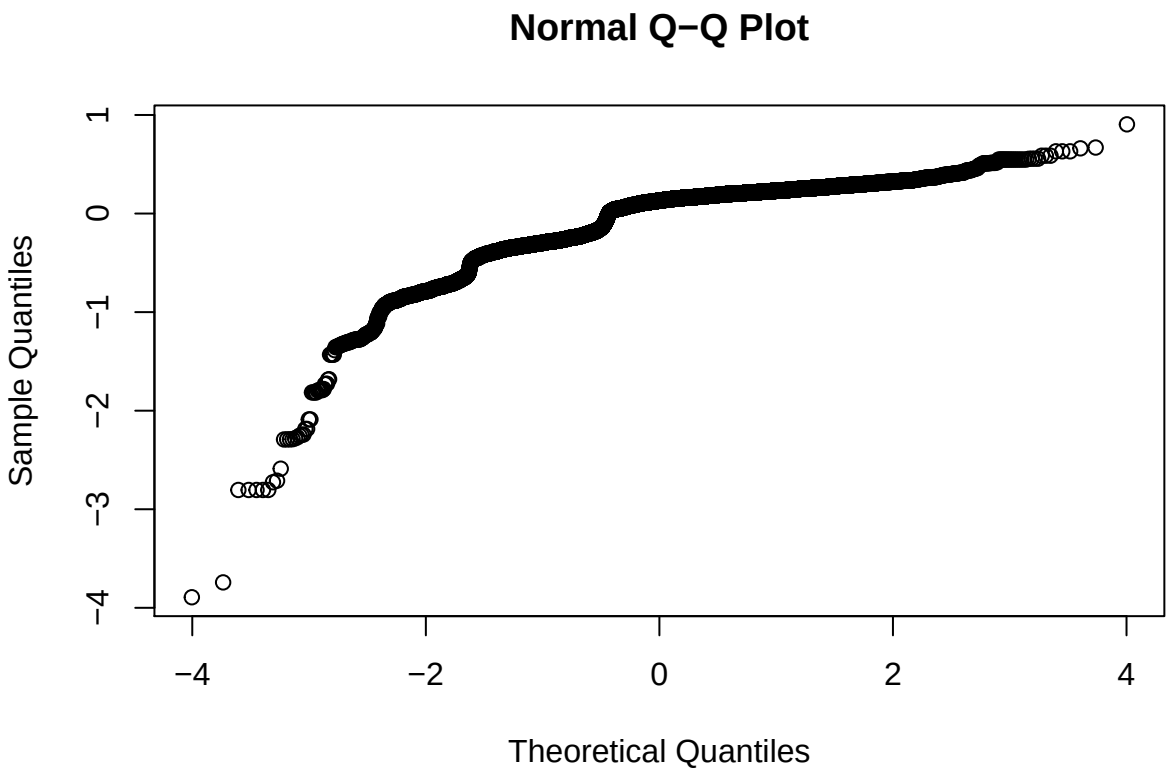
We can see that rating_reg_3 is a better fit comparing to the other two model with lower AIC and BIC

values.

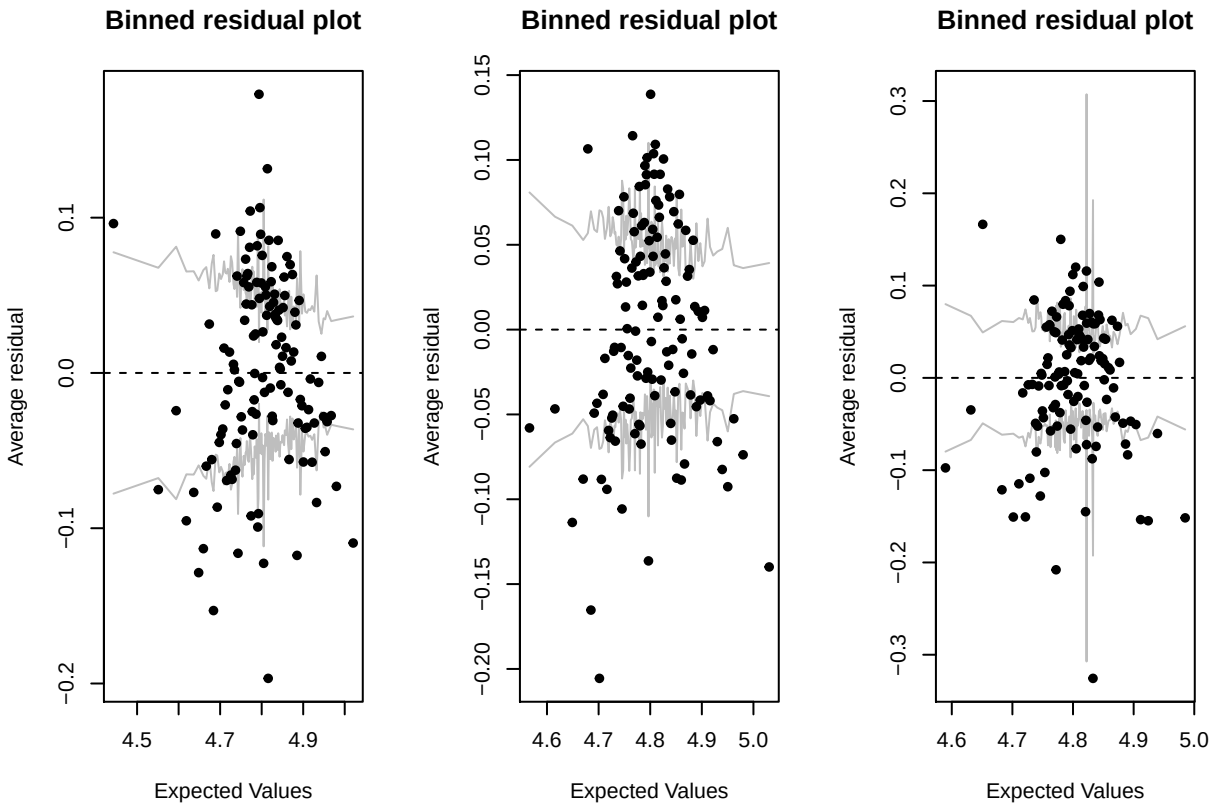
```
library(arm)
plot(rating_reg_3)
```




```
qqnorm(resid(rating_reg_3))
```




```
par(mfrow=c(1,3))
binnedplot(fitted(rating_reg_3),residuals(rating_reg_3, type="response"))
binnedplot(fitted(rating_reg_2),residuals(rating_reg_2, type="response"))
binnedplot(fitted(rating_reg),residuals(rating_reg, type="response"))
```



The resid and qqplot do not look very well. There are clearly a trend away with the fitted line. By looking at the three binnedplot plots, the rating_reg_3 does not look very good comparing to the others despite being the best model analyzing by the anova table.

Discussion

Implication and Limitation

For the six models, I did three for predicting price and three for predicting overall_satisfaction.

For the price model, the "price_reg_3" looks more reasonable comparing to the other two. Room_type, accomodates, bedrooms, and some random intercept and slope between overall_satisfaction and neighborhood will be the main effect of the price.

For the overall_satisfaction model, the "rating_reg_3" seems like the best choice according to the anova table; However, the binnedplot, residual and qqplot plots suggest that this model still has a lot of problems because a good model fit should have random patterns of the residual. It would be best to refit a better model for the overall_satisfaction.

Future direction

For the price model, in the binnedplot, there are still a lot of points outside of the fitted line. One of the reason might be that the dataset has many outliers. Maybe I can clean the data in a beter way next time.

For the overall_satisfaction model, since all models look not very well, it is best to try some other models, such as mult-level logistic model. I can do the same step of the mult-level linear model for the logistic model

by testing random slope and intercept.

Reference

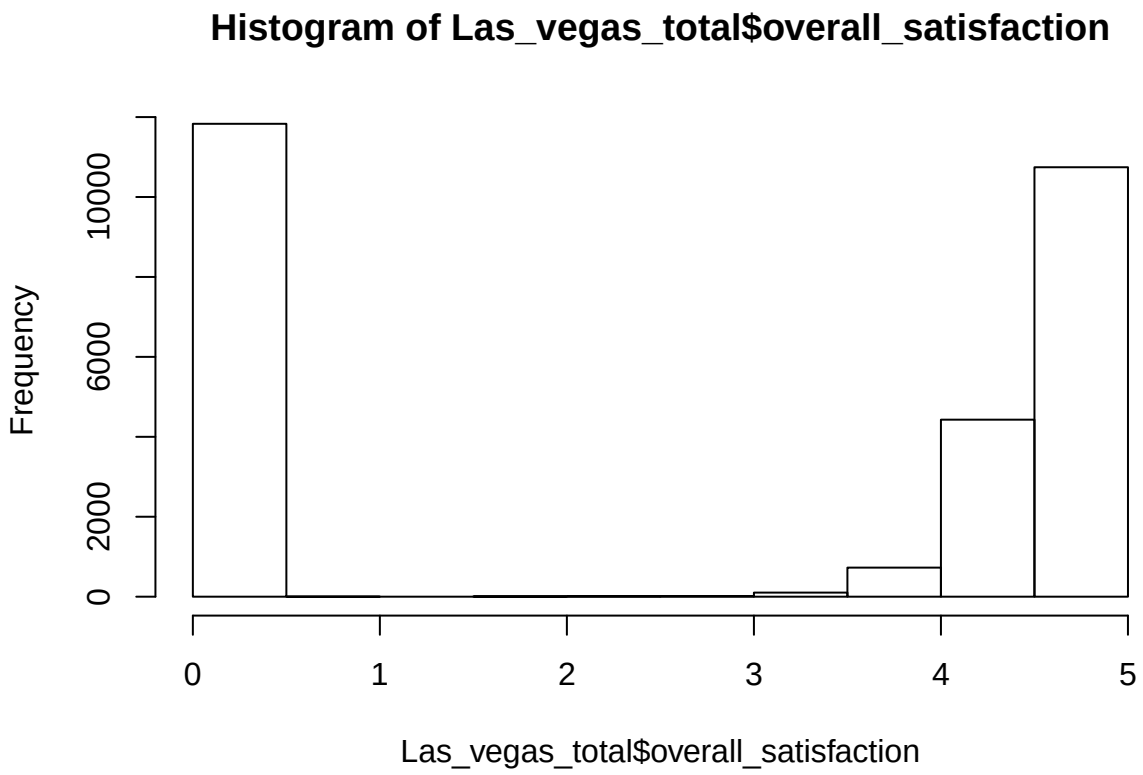
http://tomslee.net/airbnb-data-collection-get-the-data

Appendix

Some interesting EDA

```
library(ggplot2)
library(dplyr)
library(arm)
```

```
# Distribution of rating
hist(Las_vegas_total$overall_satisfaction)
```

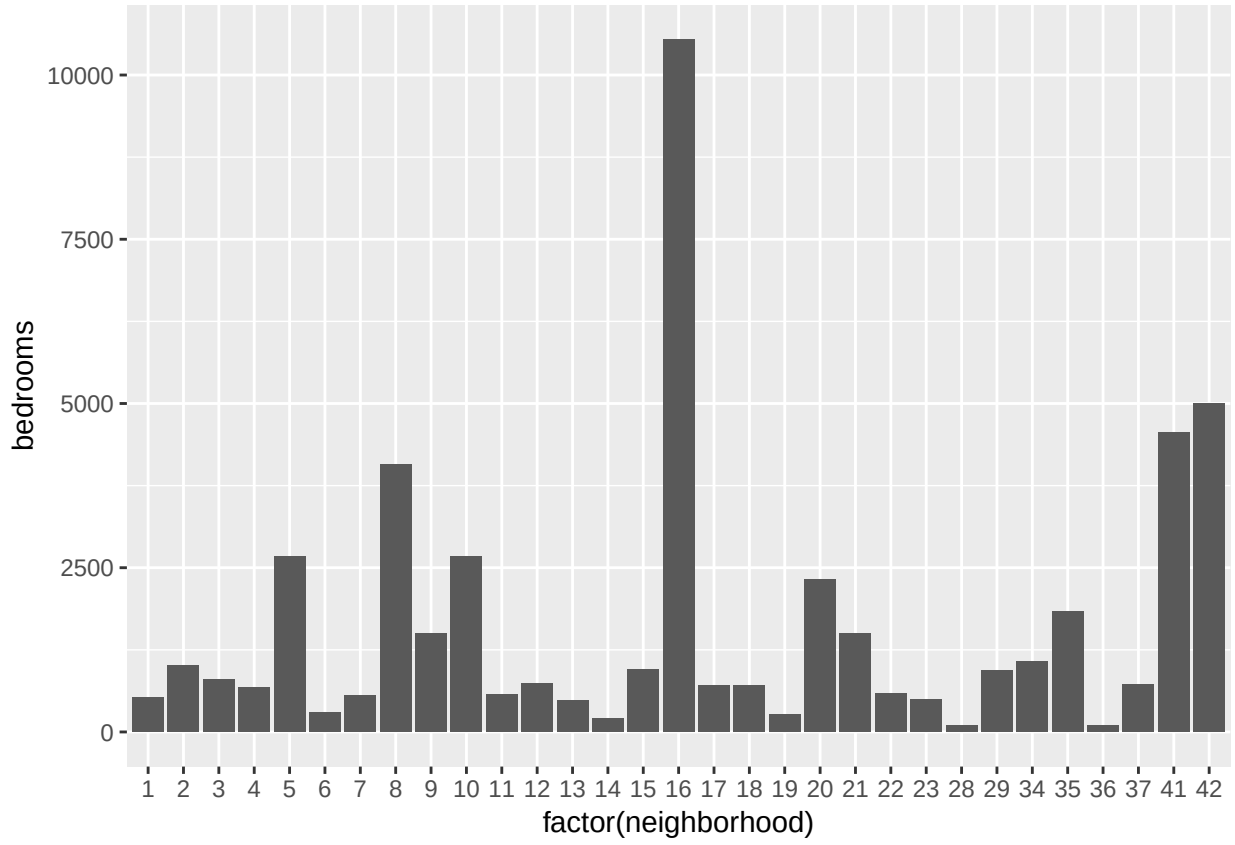



```
summary(Las_vegas_total$overall_satisfaction)
```

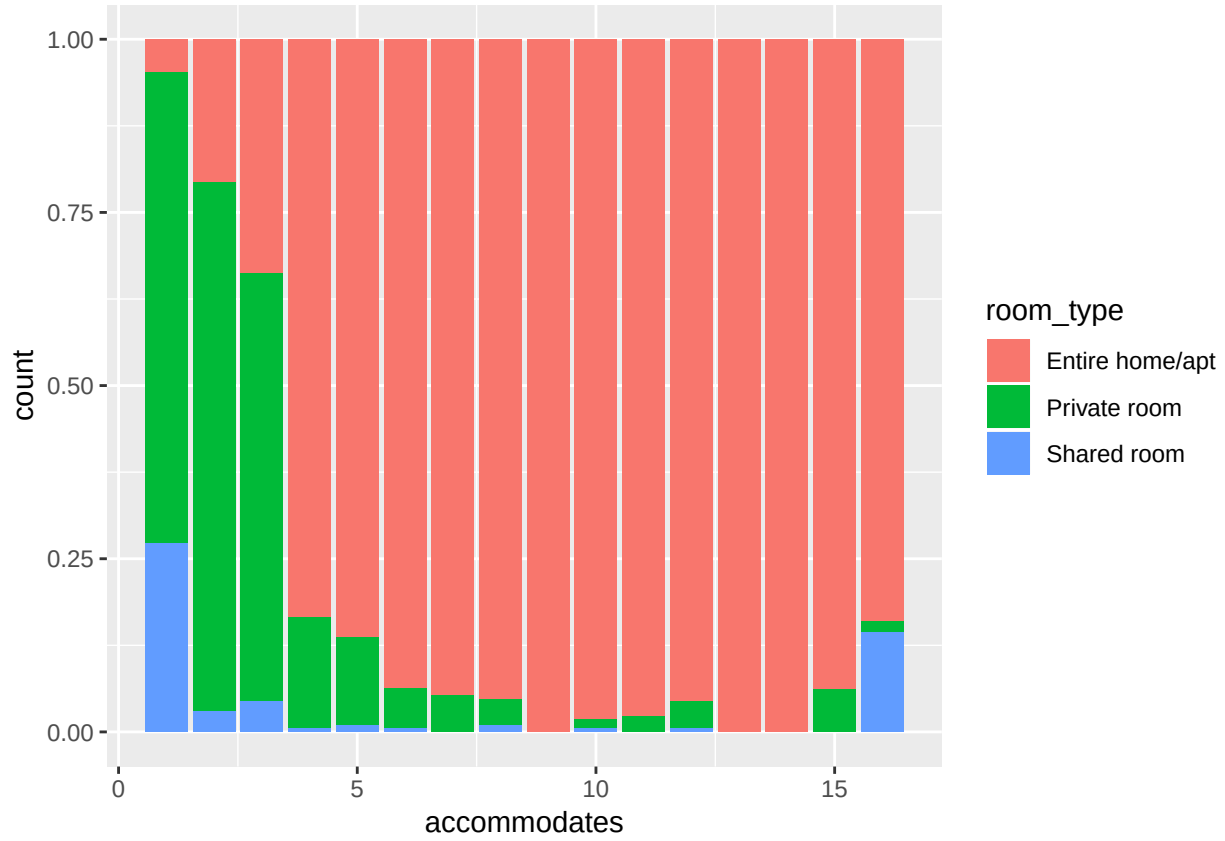
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   4.500   2.763   5.000   5.000
```

According to the graph, the rating of airbnb is between 1 and 5. Zero values mean people did not rate on it. The median is 4.5. Thus, we can see overall the airbnb properties in Las vegas have high rating.

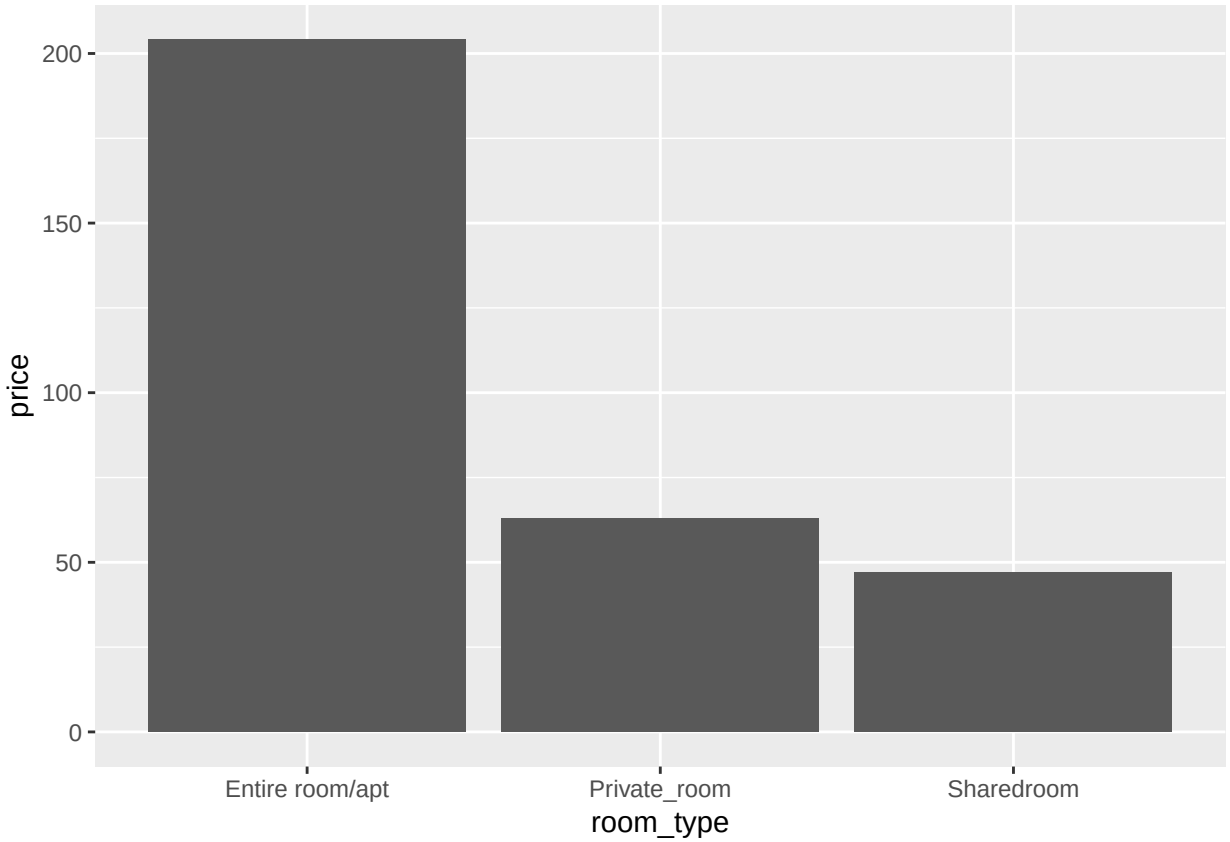
```r
library(ggplot2)
# Distribution of bedrooms
# Each number represents an neighborhood

ggplot(Las_vegas_total, aes(x = factor(neighborhood), y = bedrooms)) + geom_bar(stat = "identity")
```




```r
# Distribution of accommodates in room types
ggplot(Las_vegas_total, aes(x = accommodates, fill = room_type)) + geom_bar(position = "fill")
```

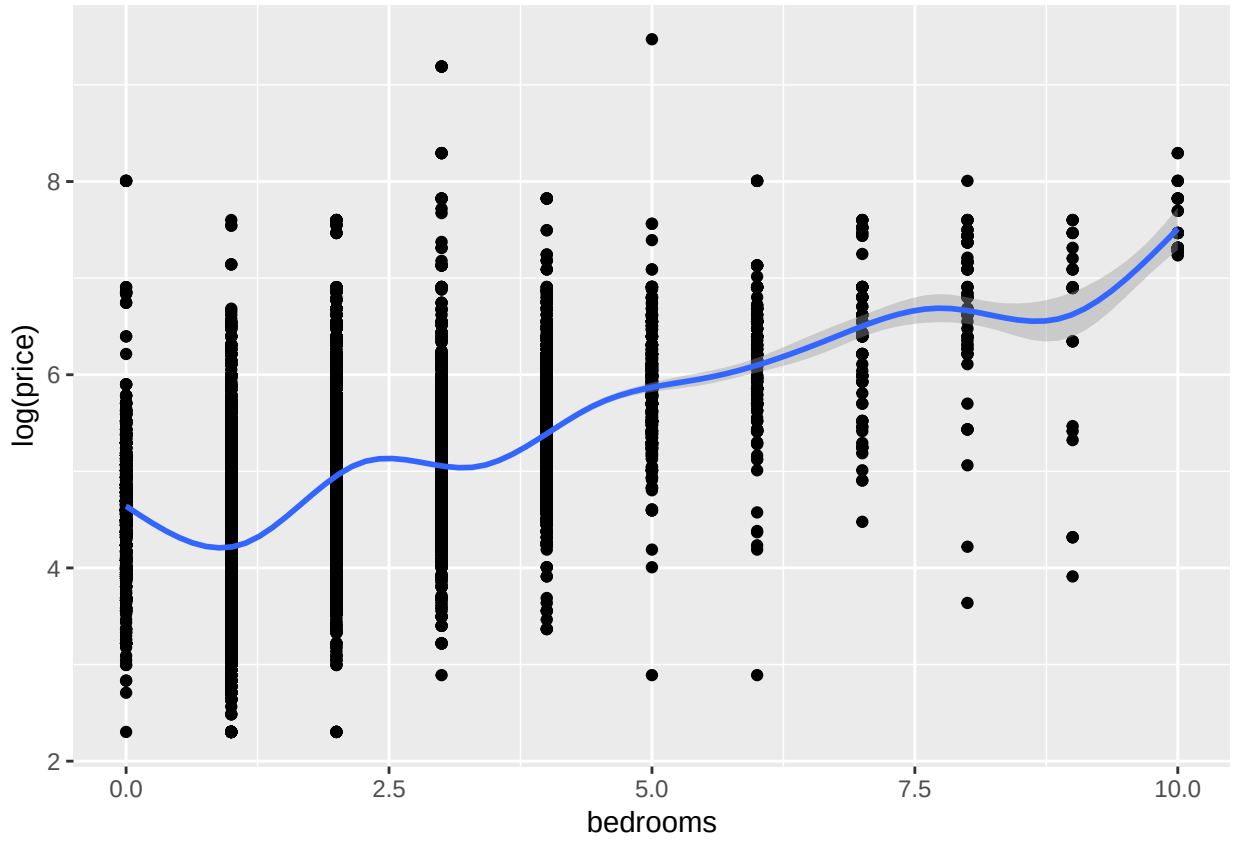



```
# Room_type vs price

Shareroom <- Las_vegas_total %>%
  filter(Las_vegas_total$room_type == "Shared room")
Entire_room <- Las_vegas_total %>%
  filter(Las_vegas_total$room_type == "Entire home/apt")
Private_room <- Las_vegas_total %>%
  filter(Las_vegas_total$room_type == "Private room")
avg_price <- data.frame(room_type=c("Sharedroom","Entire room/apt","Private_room"),
                price=c(mean(Shareroom$price), mean(Entire_room$price),mean(Private_room$price)))
ggplot(data=avg_price, aes(x=room_type,y = price)) + geom_bar(stat="identity")
```



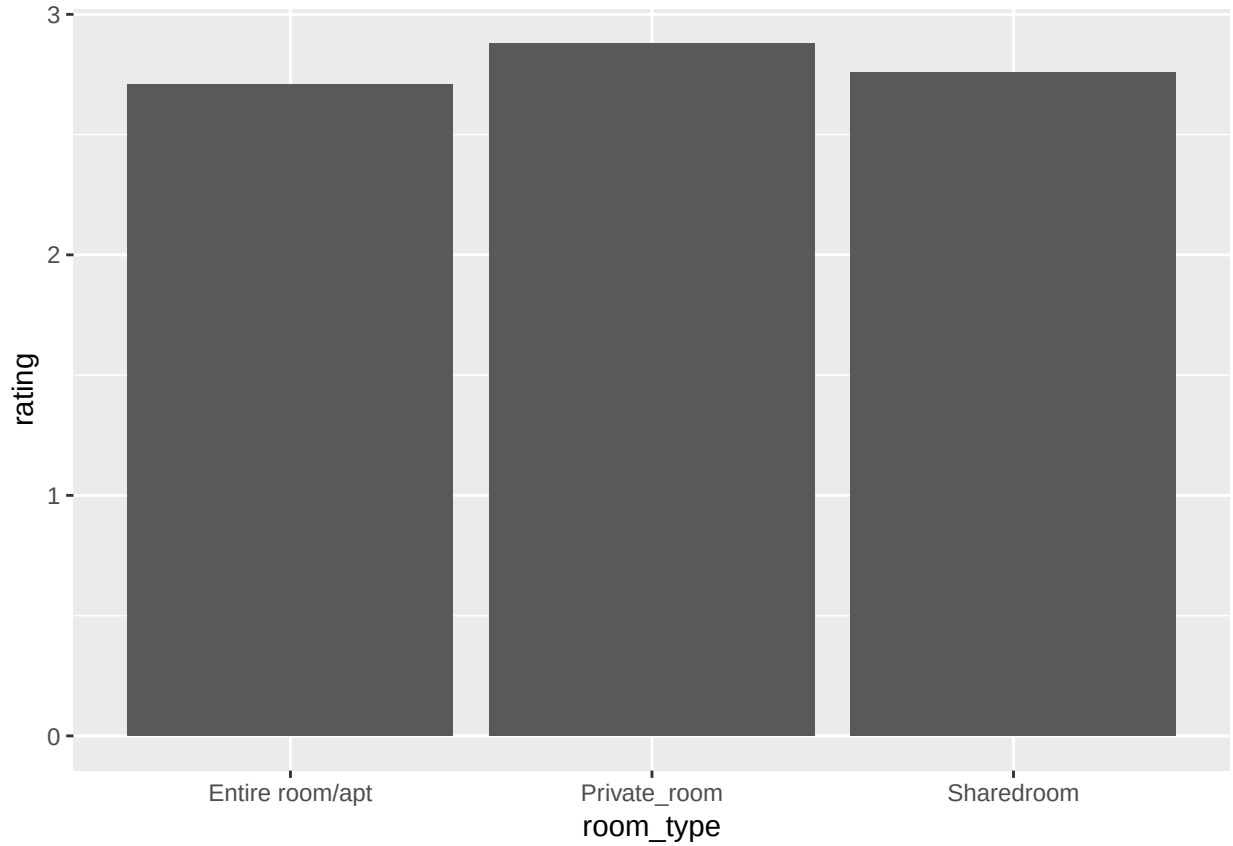

```
# Bedrooms  vs price
ggplot(Las_vegas_total,aes(x = bedrooms,y = log(price))) + geom_point() +geom_smooth()
```

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

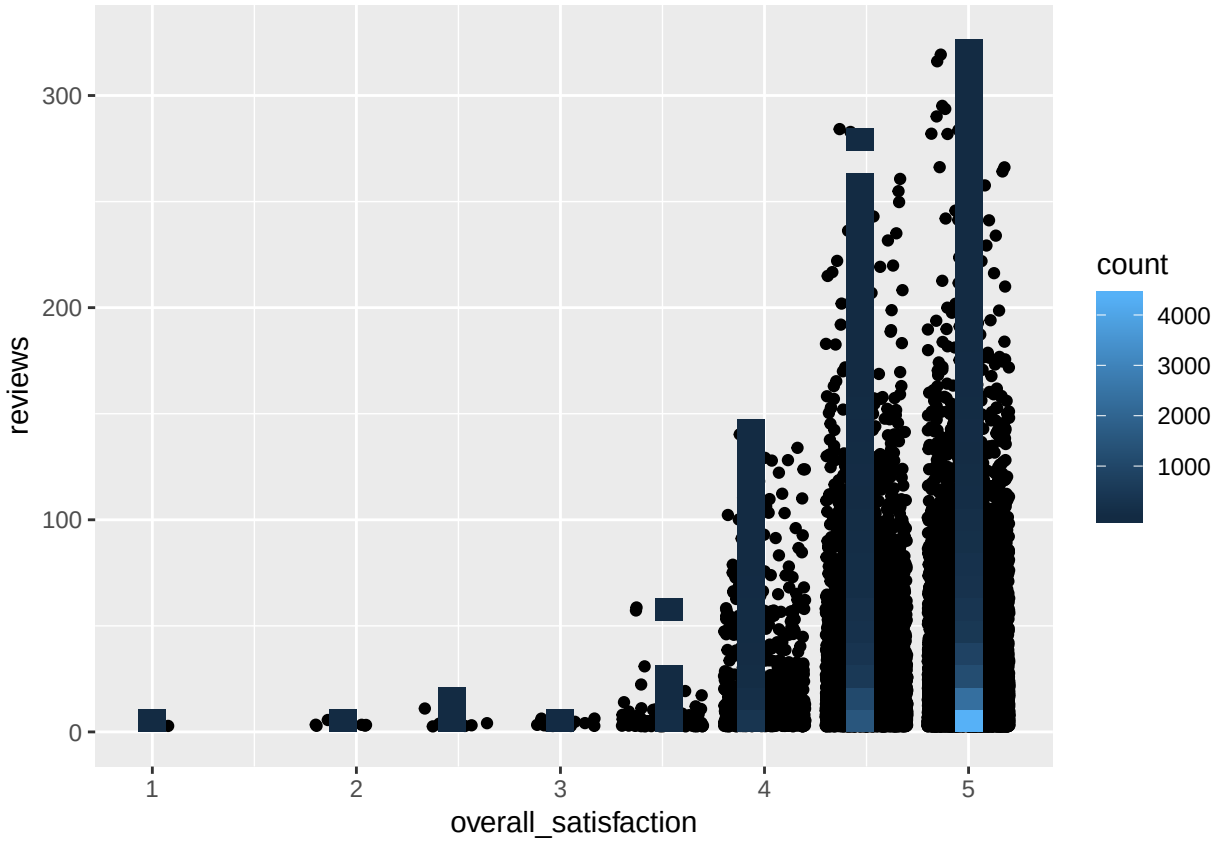


Apartments with more bedrooms tend to have higher price.

```
library(dplyr)
library(ggplot2)
# Rating vs room_type
Shareroom_rating <- overall_satisfaction_data %>%
  filter(room_type == "Shared room")

Entire_rating <- overall_satisfaction_data %>%
  filter(room_type == "Entire home/apt")

Privateroom_rating <- overall_satisfaction_data %>%
  filter(room_type == "Private room")

avg_rating <- data.frame(room_type=c("Sharedroom","Entire room/apt","Private_room"),
                rating=c(mean(Shareroom$overall_satisfaction), mean(Entire_room$overall_satisfaction),
                         mean(Private_room$overall_satisfaction)))

ggplot(data=avg_rating, aes(x=room_type,y = rating)) + geom_bar(stat="identity")
```




```
# Rating vs reviews
# Higher over_satisfaction have more reviews
ggplot(overall_satisfaction_data,aes(x = overall_satisfaction,y =reviews )) +geom_jitter() +
  geom_bin2d()
```


300
200
100
0
reviews
1
2
3
4
5
overall_satisfaction
count
4000
3000
2000
1000