MA678 Modeling Midterm Project

Yiping Jiang

12/6/2019

Contents

1. Introduction

1.1 Overview

Airbnb is the leading and rapidly growing alternative to the traditional hotel networks. It collects a large amount of data about their hosts and properties including detailed tourists reviews, and it is changing people's habits for accommodation when traveling. And there rises an problem about how customers can find the cost-effective accommodation and how owners set the the proper price for their properties.

To address this problem, I set Boston as example to analyze what factors have influences on the price of the airbnb accommodation. This report are consisted of the parts of **introduction, data preparation, Exploratory Data Analysis, modeling, discussion and appendix**, in which I mainly utlized multilevel models to conduct the regression modeling.

1.2 Outline

In this report, I will mainly investiage the following questions:

- What are the main distinctions of Airbnb superhosts from ordinary hosts?
- What are the main factors that influence Airbnb daily renting price?
- How well can one predict the Airbnb daily renting price based on the data from different neighborhoods?

To answer these questions, I have downloaded the data from Airbnb host listings and tourists reviews. From the Boston data set I chose, I will clean the data first, and then do exploratory data analysis with visualization, next analyze the data through modeling and discuss the results with interpretations.

2. Data Preparation

2.1 Data Source

The data that I used was collected from the public Airbnb web site http://tomslee.net/airbnb-data-collection-get-the-data, and there were 4705 listings in the "Boston" dataset up to July 10, 2017, which contains several variables and here are their descriptions:

- room_id: A unique number identifying an Airbnb listing.
- host_id: A unique number identifying an Airbnb host.
- room_type: One of "Entire home/apt", "Private room", or "Shared room".
- borough: A subregion of the city or search area for which the survey is carried out. The borough is taken from a shapefile of the city that is obtained independently of the Airbnb web site. For some cities, there is no borough information; for others the borough may be a number.
- neighborhood: As with borough: a subregion of the city or search area for which the survey is carried out. For cities that have both, a neighbourhood is smaller than a borough. For some cities there is no neighbourhood information.
- reviews: The number of reviews that a listing has received. Airbnb has said that 70% of visits end up with a review, so the number of reviews can be used to estimate the number of visits. Note that such an estimate will not be reliable for an individual listing (especially as reviews occasionally vanish from the site), but over a city as a whole it should be a useful metric of traffic.
- overall_satisfaction: The average rating (out of five) that the listing has received from those visitors who left a review.
- accommodates: The number of guests a listing can accommodate.
- bedrooms: The number of bedrooms a listing offers.
- price: The price (in $US) for a night stay. In early surveys, there may be some values that were recorded by month.
- minstay: The minimum stay for a visit, as posted by the host.
- latitude and longitude: The latitude and longitude of the listing as posted on the Airbnb site.
- last_modified: the date and time that the values were read from the Airbnb web site.

2.2 Data Cleaning

Loading and checking the information of the data, I plan to use visualization to show the distribution of the residences' features, from where I will then choose proper models to conduct analysis. The main steps of data preparation and clean process are:

- Extract hosts with more than 100 reviews;
- Reorganize some variables;
- Delete observations with missing values.

2.3 Praparing the data

```
# read data_final.csv
bos <- read.csv("/Users/sebas_jiang/Desktop/MA678_Modeling/Project/Air\ Bnb/Air\ Bnb/data_final.csv")

# add one column to record the price per person
bos$pricepp <- bos$price/bos$accommodates
```

```
bos %>%
  filter(!is.na(host_id)) -> bos
```

So far the dataset has 1422 rows and 16 columns.

Data Structure

```
str(bos)
```

```
## 'data.frame':    1422 obs. of  16 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ room_id             : int  6142396 6134867 3721095 3749523 4151979 3528826 7881551 497810 2881388
##  $ host_id             : int  31862615 20582119 10477638 3842205 21541399 10493424 1639654 2329611 9
##  $ room_type           : Factor w/ 3 levels "Entire home/apt",..: 3 1 1 1 1 1 1 1 1 1 ...
##  $ borough             : logi  NA NA NA NA NA NA ...
##  $ neighborhood        : Factor w/ 24 levels "Allston","Back Bay",..: 10 9 6 6 21 2 2 9 2 2 ...
##  $ reviews             : int  146 136 110 129 107 115 112 161 103 111 ...
##  $ overall_satisfaction: num  5 4.5 5 4.5 5 5 4.5 4.5 5 5 ...
##  $ accommodates        : int  2 4 6 7 8 4 4 8 7 2 ...
##  $ bedrooms            : int  1 2 3 4 4 1 0 3 3 1 ...
##  $ price               : int  56 325 375 500 500 239 210 299 295 185 ...
##  $ minstay             : logi  NA NA NA NA NA NA ...
##  $ last_modified       : Factor w/ 1422 levels "2017-01-14 08:07:54.198209",..: 178 174 177 176 175
##  $ latitude            : num  42.4 42.4 42.4 42.4 42.3 ...
##  $ longitude           : num  -71 -71.1 -71.1 -71.1 -71 ...
##  $ pricepp             : num  28 81.2 62.5 71.4 62.5 ...
```

Now from the output we have variables such as **room_type, neighborhood** that have different levels, so I then put these categorical variables into my model. Next, for numerical variables such as **reviews, overall_satisfaction, accommodates, bedrooms**, which include discrete data rather than consecutive data, so they can also be put into my model as categorical variables.

Data overview

```
##        X             room_id            host_id
##  Min.   :   1.0   Min.   :   20000   Min.   :    8229
##  1st Qu.: 356.2   1st Qu.: 1147871   1st Qu.: 1600541
##  Median : 711.5   Median : 2898226   Median : 6608084
##  Mean   : 711.5   Mean   : 3806340   Mean   :13376608
##  3rd Qu.:1066.8   3rd Qu.: 6089865   3rd Qu.:19357187
##  Max.   :1422.0   Max.   :15535385   Max.   :92630065
##
##           room_type   borough              neighborhood    reviews
##  Entire home/apt:698   Mode:logical   Jamaica Plain:231   Min.   :101.0
##  Private room   :717   NA's:1422      Dorchester   :192   1st Qu.:115.0
##  Shared room    :  7                  East Boston  :141   Median :137.5
##                                       North End    :125   Mean   :162.1
##                                       South End    :103   3rd Qu.:191.0
##                                       South Boston : 99   Max.   :470.0
##                                       (Other)      :531
##  overall_satisfaction  accommodates       bedrooms         price
##  Min.   :3.500        Min.   : 1.000   Min.   :0.000   Min.   : 20.0
##  1st Qu.:4.500        1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 70.0
##  Median :5.000        Median : 2.000   Median :1.000   Median : 99.0
##  Mean   :4.726        Mean   : 2.902   Mean   :1.141   Mean   :121.2
```

```
##  3rd Qu.:5.000        3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.:150.0
##  Max.   :5.000        Max.   :16.000   Max.   :4.000   Max.   :500.0
##
##  minstay                           last_modified     latitude
##  Mode:logical   2017-01-14 08:07:54.198209:   1   Min.   :42.26
##  NA's:1422      2017-01-14 08:07:54.231862:   1   1st Qu.:42.32
##                 2017-01-14 08:07:54.750934:   1   Median :42.34
##                 2017-01-14 08:07:56.822212:   1   Mean   :42.34
##                 2017-01-14 08:07:56.832572:   1   3rd Qu.:42.36
##                 2017-01-14 08:07:56.844632:   1   Max.   :42.38
##                 (Other)                   :1416
##    longitude         pricepp
##  Min.   :-71.17   Min.   :  6.667
##  1st Qu.:-71.10   1st Qu.: 30.000
##  Median :-71.07   Median : 41.000
##  Mean   :-71.08   Mean   : 46.129
##  3rd Qu.:-71.06   3rd Qu.: 60.833
##  Max.   :-71.00   Max.   :144.500
##
```

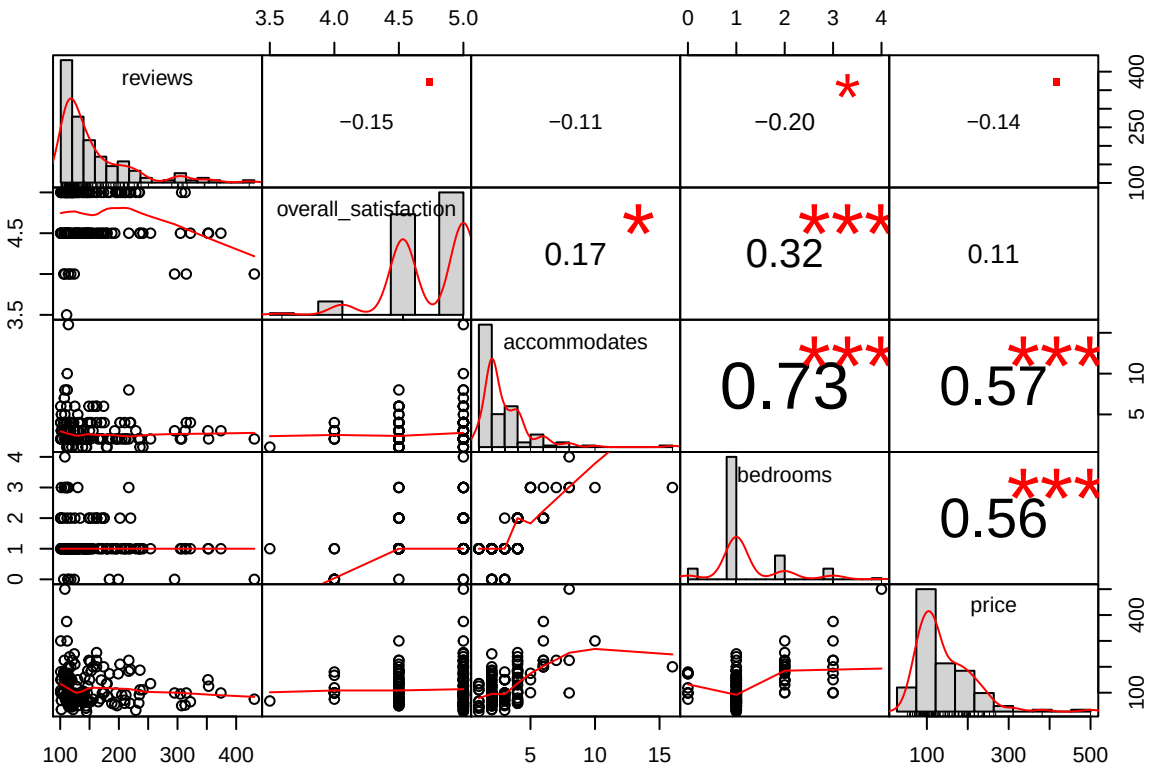
From the summary we can see these variables may have correlation with the price of each listing, such as **room_type, neighborhood, reviews, overall_satisfaction, accommodates, bedrooms**, from which I can explore the correlation of variables mainly with the price in the exploratory data analysis by choosing proper model to do analyses.

3. Exploratory Data Analysis

3.1 Correlation plot

```
# create the subset
set.seed(2019)
bos %>%
  dplyr::select(price, room_type, reviews, overall_satisfaction,
                neighborhood, accommodates, bedrooms) %>%
  na.omit(accommodates, bedrooms, room_type, neighborhood, borough, reviews) %>%
  sample_n(0.1 * nrow(bos))-> bos.m
bos.m %>%
  dplyr::select(reviews, overall_satisfaction, accommodates, bedrooms, price) -> cor.p

library("PerformanceAnalytics")
chart.Correlation(cor.p, histogram = TRUE, pch = 20)
```

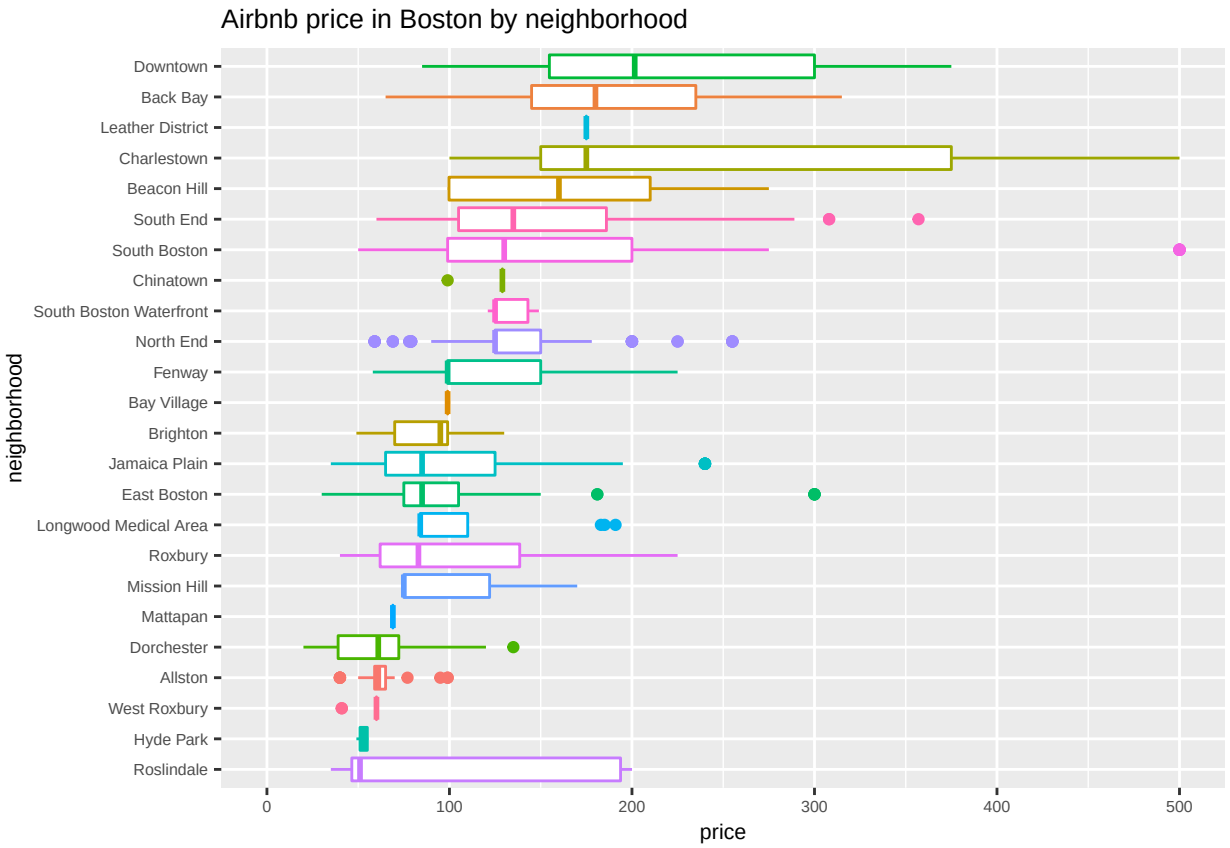


From the correlation plot, we have distributions of these five variables on the diagonal, and we can see the bivariate scatter plots with fitted lines in the lower left triangular area that there are correlations between each of the first four variables and the price (accomodates and bedrooms in particular). Next from the numbers on the upper right triangular area we can see correlations and significant level among these four variables and price are high, so I consider to put them into my model to predict the accommodation price.

3.2 Boxplot of neighborhood and price

```
ggplot(bos, aes(x = reorder(neighborhood, price, median),
                y = price, group = neighborhood, colour = neighborhood)) +
  geom_boxplot() +
  ylim(0,500) +
```

```
theme(text=element_text(size = 8),legend.position = "none") +
labs(title = "Airbnb price in Boston by neighborhood", x = "neighborhood") +
coord_flip()
```



From the boxplot we can see that the difference of the price median among all neighborhoods are significant, so I consider to use multilevel model. Downtown has the highest median price than other neighborhoods and Roslindale has the lowest median price. As a result, we can infer from the plot that *neighborhood* is one of the factors that influences the price.

4. Modeling

4.1 Multilevel Model

Model 1

Regress **price** on **accommodates and bedrooms** and treat all other intercepts as random.

```
fit.1 <- lmer(data = bos.m, log(price) ~ accommodates +
                (1 | room_type) + (1 | neighborhood), REML = FALSE)
display(fit.1)
```

```
## lmer(formula = log(price) ~ accommodates + (1 | room_type) +
##     (1 | neighborhood), data = bos.m, REML = FALSE)
##              coef.est coef.se
## (Intercept)  4.54     0.19
## accommodates 0.06     0.02
##
## Error terms:
##  Groups       Name        Std.Dev.
##  neighborhood (Intercept) 0.23
##  room_type    (Intercept) 0.24
##  Residual                 0.31
## ---
## number of obs: 142, groups: neighborhood, 18; room_type, 2
## AIC = 115, DIC = 105
## deviance = 105.0
```

Model 2

Based on Model 1, we add a between-group correlation between accommodates and neighborhood.

```
fit.2 <- lmer(data = bos.m, log(price) ~ accommodates + bedrooms +
                (1 | room_type) + (1 + bedrooms | neighborhood), REML = FALSE)
display(fit.2)
```

```
## lmer(formula = log(price) ~ accommodates + bedrooms + (1 | room_type) +
##     (1 + bedrooms | neighborhood), data = bos.m, REML = FALSE)
##              coef.est coef.se
## (Intercept)  4.49     0.19
## accommodates 0.02     0.02
## bedrooms     0.15     0.06
##
## Error terms:
##  Groups       Name        Std.Dev. Corr
##  neighborhood (Intercept) 0.21
##               bedrooms    0.01     1.00
##  room_type    (Intercept) 0.24
##  Residual                 0.31
## ---
## number of obs: 142, groups: neighborhood, 18; room_type, 2
## AIC = 114.9, DIC = 98.9
## deviance = 98.9
```

Model 3

Then try regressing price on accomodates, room_type, reviews and bedrooms with remaining one between-group and other variables as random.

```
fit.3 <- lmer(data = bos.m, log(price) ~ accommodates + reviews + room_type +
                bedrooms + (1 + accommodates | neighborhood), REML = FALSE)
display(fit.3)
```

```
## lmer(formula = log(price) ~ accommodates + reviews + room_type +
##     bedrooms + (1 + accommodates | neighborhood), data = bos.m,
##     REML = FALSE)
##                        coef.est coef.se
## (Intercept)             4.97     0.16
## accommodates            0.01     0.04
## reviews                 0.00     0.00
## room_typePrivate room -0.55     0.08
## bedrooms                0.12     0.06
##
## Error terms:
##  Groups       Name         Std.Dev. Corr
##  neighborhood (Intercept)  0.36
##               accommodates 0.09     -0.93
##  Residual                  0.29
## ---
## number of obs: 142, groups: neighborhood, 18
## AIC = 99.4, DIC = 81.4
## deviance = 81.4
```

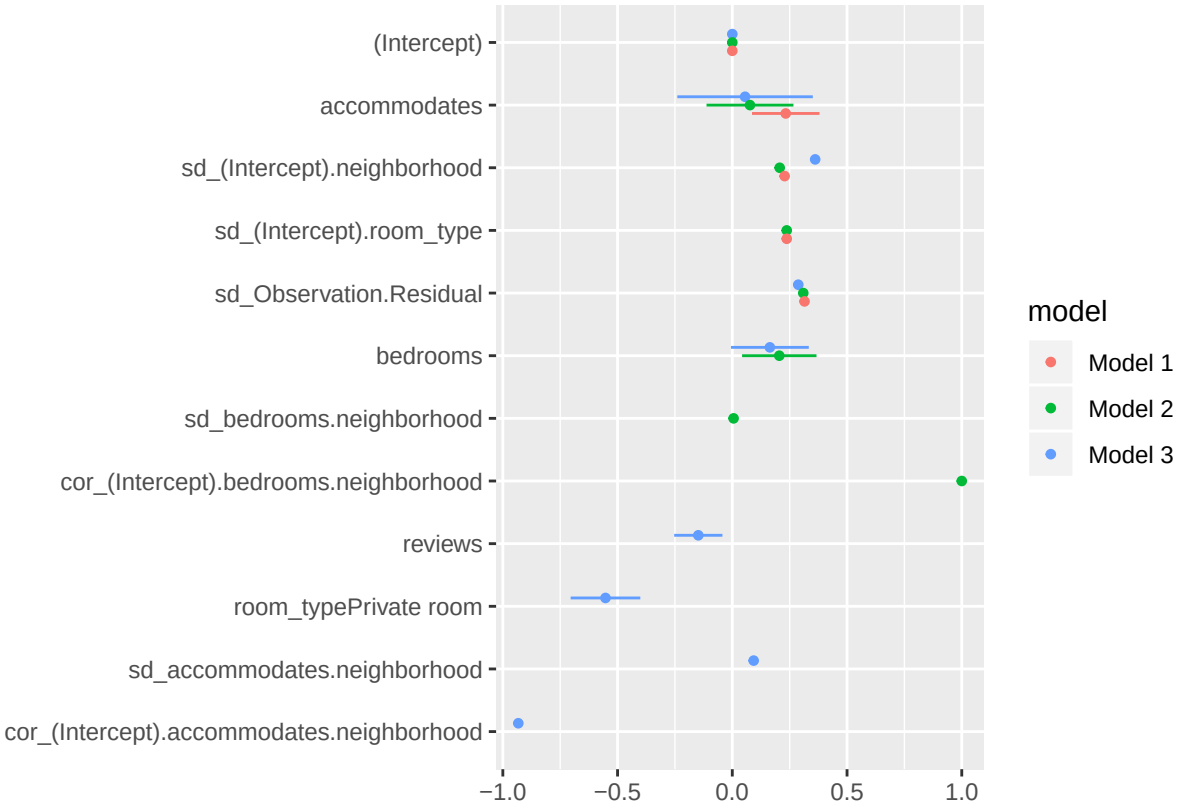
```
ranef(fit.3)
```

```
## $neighborhood
##                          (Intercept) accommodates
## Allston                 -0.22817280  0.047245990
## Back Bay                 0.08344884  0.004074514
## Bay Village              0.07992116 -0.016548660
## Beacon Hill              0.08148332  0.005603148
## Brighton                 0.34542416 -0.115527769
## Charlestown             -0.24817551  0.088797190
## Dorchester              -0.66975769  0.150427077
## Downtown                -0.14399961  0.047596989
## East Boston              0.05121202 -0.022047376
## Fenway                  -0.11381676  0.009999657
## Jamaica Plain           -0.12612447  0.006390602
## Mission Hill            -0.09964572  0.020632874
## North End                0.10677852 -0.073592400
## Roslindale               0.03332934 -0.009634443
## Roxbury                 -0.19223322  0.063112645
## South Boston            -0.04559075  0.021633459
## South Boston Waterfront  0.47670863 -0.098708396
## South End                0.60921053 -0.129455101
##
## with conditional variances for "neighborhood"
```

4.2 Model Choice

Coefficient plot

```
dwplot(list(fit.1, fit.2, fit.3), show_intercept = TRUE)
```



From the above plot, we cannot tell whether model 2 or 3 is better, because although Model 3 has more points than Model 2, it might because Model 3 has more observations. For Model 2, although it has less points but it shows more precision than Model 3.

ANOVA

```
anova(fit.1, fit.2, fit.3)
```

```
## Data: bos.m
## Models:
## fit.1: log(price) ~ accommodates + (1 | room_type) + (1 | neighborhood)
## fit.2: log(price) ~ accommodates + bedrooms + (1 | room_type) + (1 +
## fit.2:     bedrooms | neighborhood)
## fit.3: log(price) ~ accommodates + reviews + room_type + bedrooms +
## fit.3:     (1 + accommodates | neighborhood)
##       Df     AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
## fit.1  5 114.985 129.76 -52.492  104.985
## fit.2  8 114.934 138.58 -49.467   98.934  6.0507      3     0.1092
## fit.3  9  99.397 126.00 -40.698   81.397 17.5373      1  2.817e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From ANOVA, comparing among fit 1, 2 and 3, we can see both the AIC and BIC of fit 3 are smaller than Fit 1 and 2, and through fit 1 to 3, the deviance is decreasing, and the p-value is 2.817e-05, which is

significant enough. As a result, we decide to use fit 3 and add the those variables to the model, which does lead to a significantly improved fit.

Therefore, the final model choosen is model 3, a multilevel model varying both intercepts and slopes.

$$log(y_i) \sim \ N(\alpha_{j[i]} + \beta_{j[i]} x_i,\ \sigma_y^2),\ \ for\ i = 1, \ldots, n$$

4.3 Interpretation

This multilevel model can be explained as below. Taking reserving airbnb accommodation in Allston as example:

$$\begin{aligned} log(Y_{price}) &= (4.97 - 0.23) + (0.01 + 0.05) X_{accommodates} + 0 * X_{reviews} + 0.12 X_{bedrooms} - 0.55 X_{privateroom} \\ &= 4.74 + 0.06 X_{accommodates} + 0 * X_{reviews} + 0.12 X_{bedrooms} - 0.55 X_{privateroom} \end{aligned}$$

Taking a accommodation with the accommodate is 0, the number of bedrooms is 0, the reviews are 0 and the room type is apartment, the price gives $e^{4.74}$ which is 114.43 (though the accommodate cannot be 0).

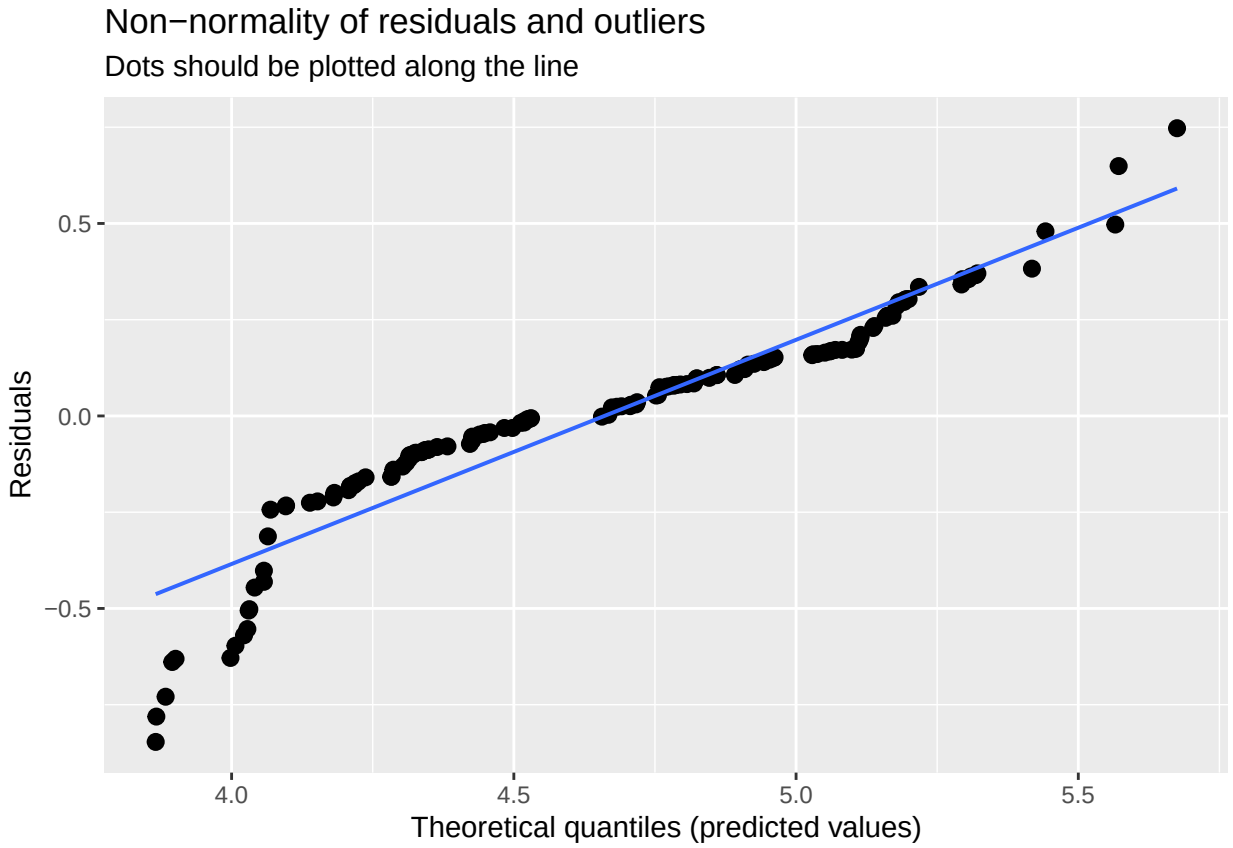
With every one person increased in accommodate and other variables ramain the same, the price will increase by $e^{0.06}$ which is 1.06; with every one unit increased in the number of reviews and controlling others, the price will not change; with every one unit increased in the number of bedrooms and others do not change, the price will increase by $e^{0.12}$ which is 1.13. And if the room type is private room while controling other variables, the price will decrease by $e^{-0.55}$ which is 0.57.

4.4 Model Check

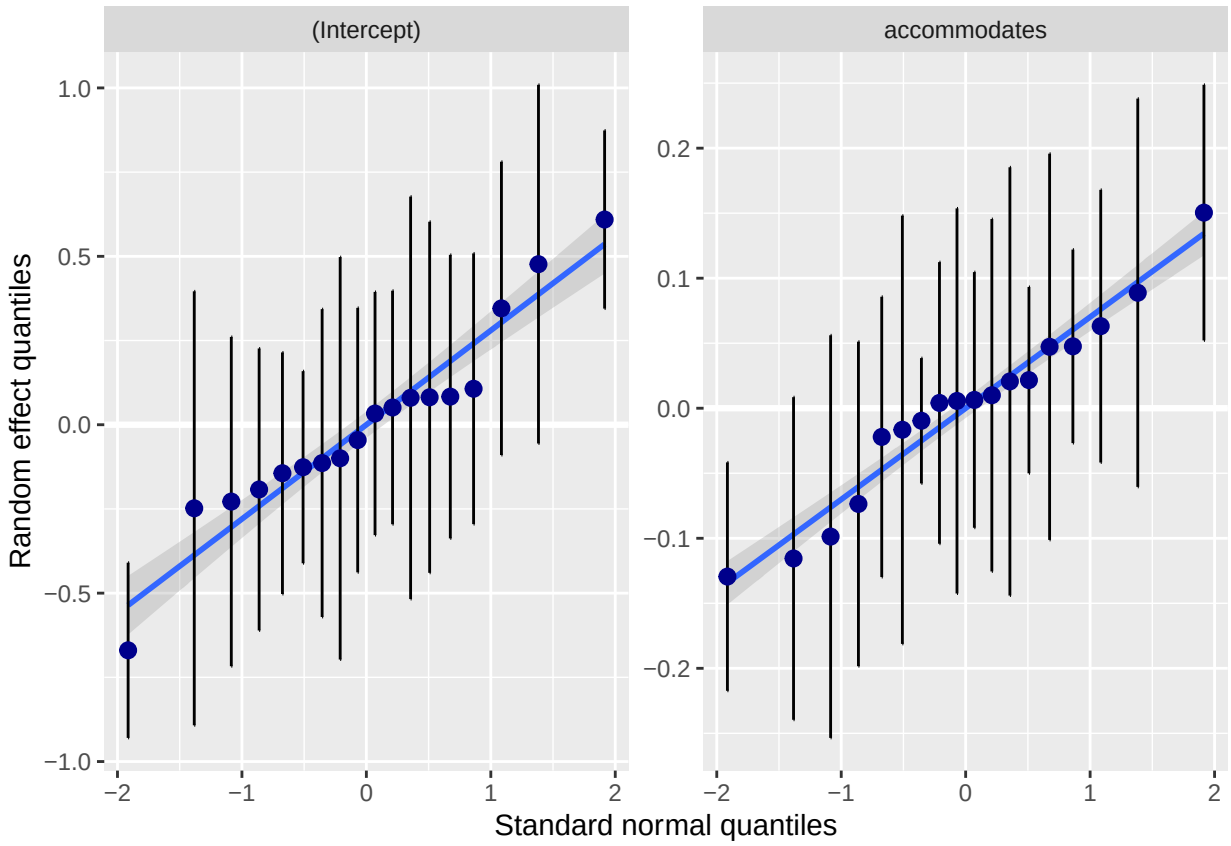
Diagnostic plots

```
library(sjPlot)
plot_model(fit.3, type = "diag", show.values = TRUE, value.offset = 0.3)
```

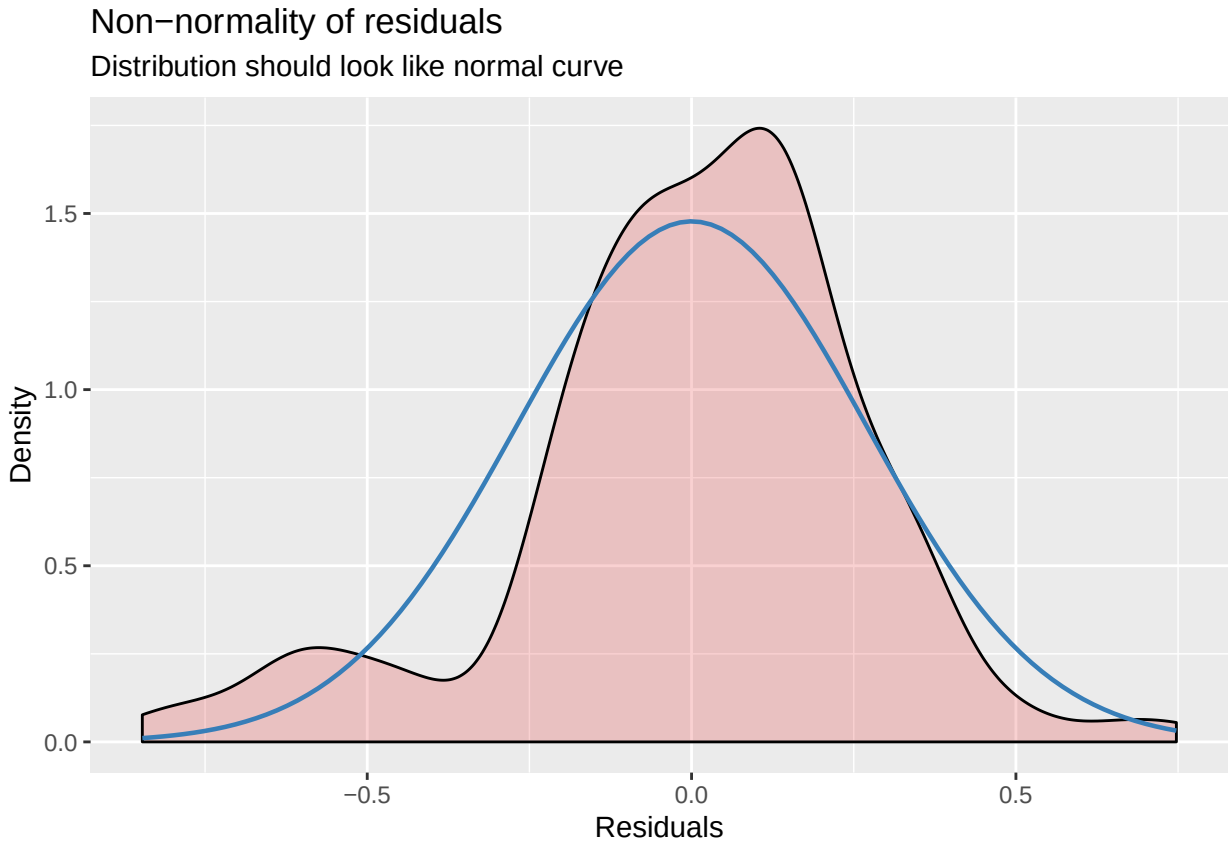
```
## [[1]]
```



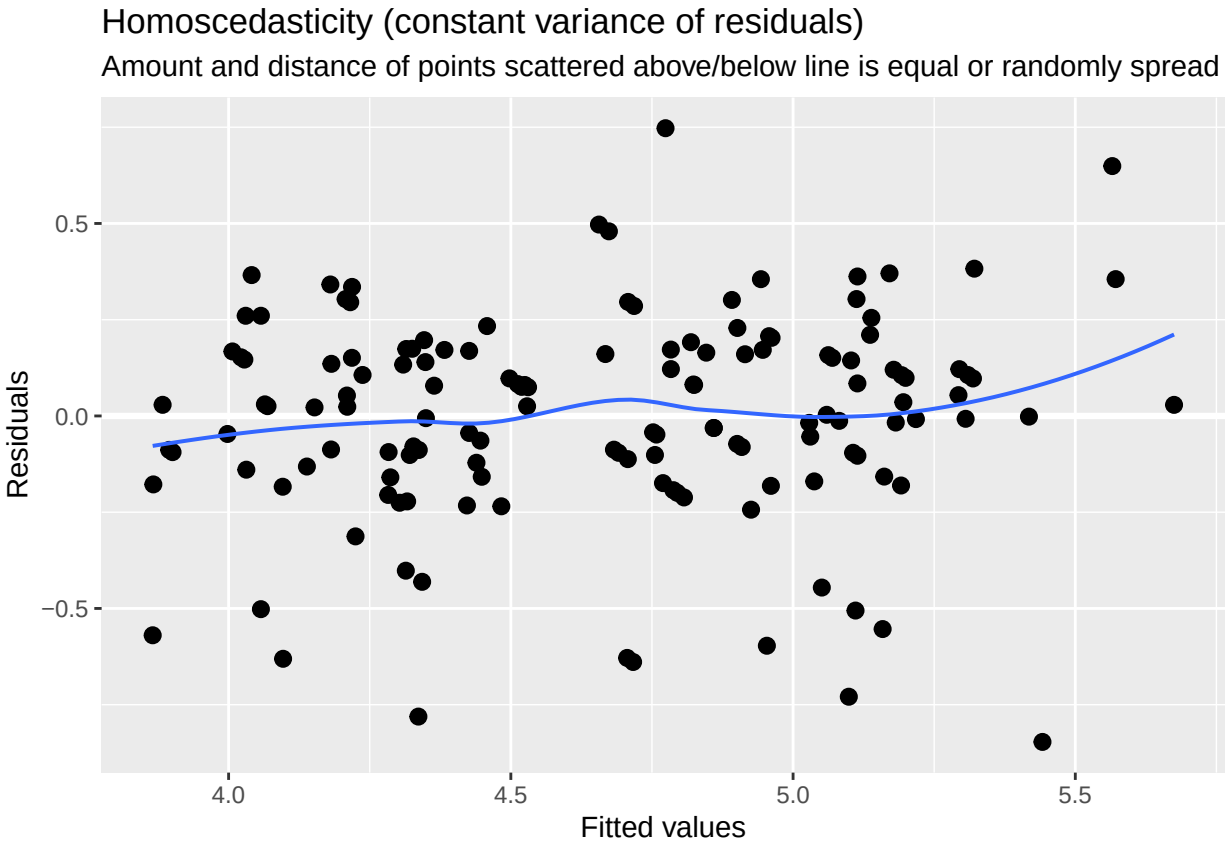

```
##
## [[2]]
## [[2]]$neighborhood
```


(Intercept)
accommodates
Random effect quantiles
Standard normal quantiles
1.0
0.5
0.0
−0.5
−1.0
0.2
0.1
0.0
−0.1
−0.2
−2
−1
0
1
2

```
##
##
## [[3]]
```


Non−normality of residuals
Distribution should look like normal curve
Density
1.5
1.0
0.5
0.0
−0.5
0.0
0.5
Residuals

```
##
## [[4]]
```



From above diagnostic plots, in plot [1], observing the non-normality of residuals and outliers, we can see the dots are mainly plotted along the line; in plot [2], the random effect shows normal distribution; in plot [3], the blue line closely comply normal distribution, and the pink reresents the normal residual. As we can see that these two mainly accord with each other; in plot [4], the homoscedasticity shows the dispersion degree is even, and there is no heteroscedasticity.

5. Discussion

5.1 Summary

By predicting the price of accommodation in airbnb, people who are reserving accommodations can take it as a reference. In this model, the number of accommodates and bedrooms are the significant factors that influence Airbnb daily renting price. While except the room type, reviews, overall satisfication, accommodates, and bedrooms, there are other variables can also affect the price while predicting, which explains why the model I chose is not that good.

5.2 Limitation

The dataset lacks some necessary features of the accommodation to get the model with better fitting, and the errors are more distributed around the higher prices of the accommodation. Therefore I can clarify all the accommodation based their prices to get a better fitting model.

5.3 Future Direction

In the future I will try using Zillow API to get more information of the accommodation to bring in more variables to better fit the model with more ideal results.

Acknowledgments

I would like to thank my classmate, Harry Cao, for answering questions and providing suggestions to help me better understand this project.

6. Appendix

Random effect plot

```
# Check random effect coeficients and the significance.
plot_model(fit.3, type = "re", show.values = TRUE, value.offset = 0.3)
```

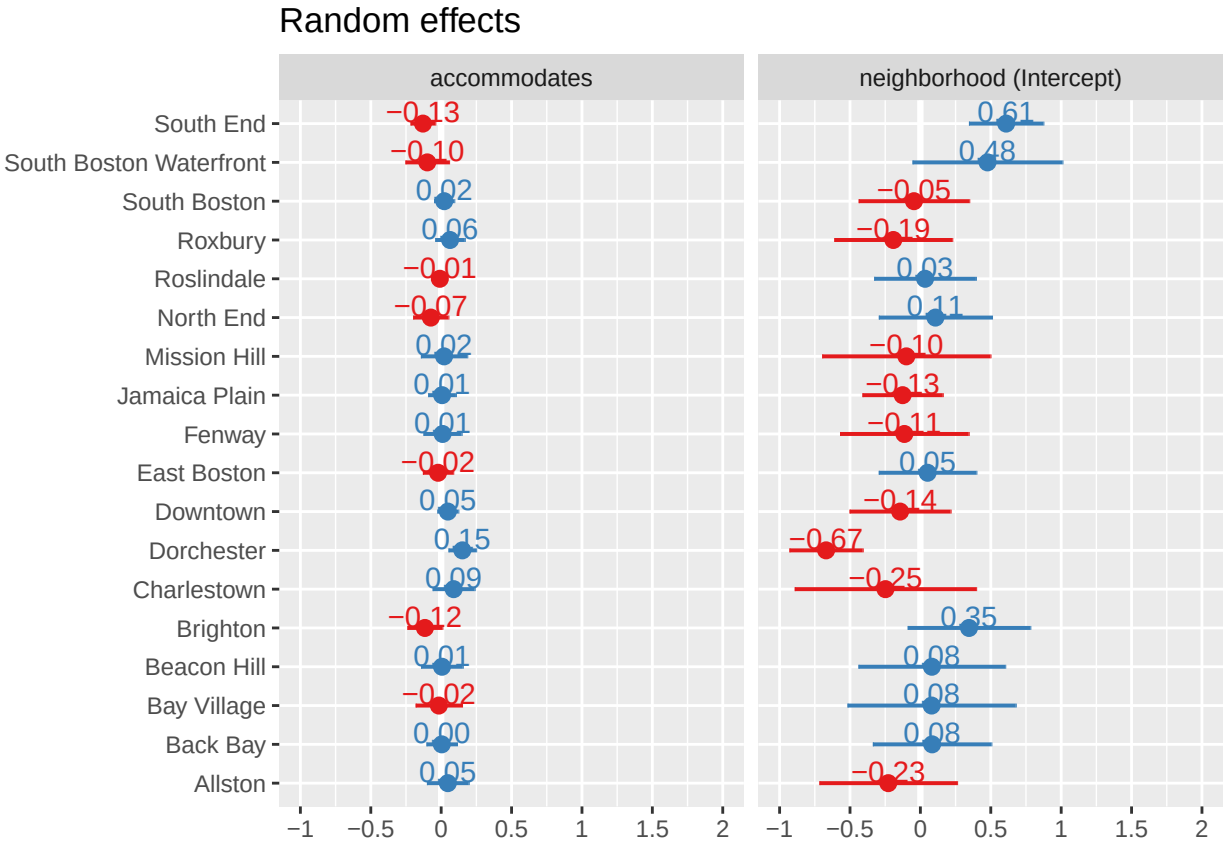
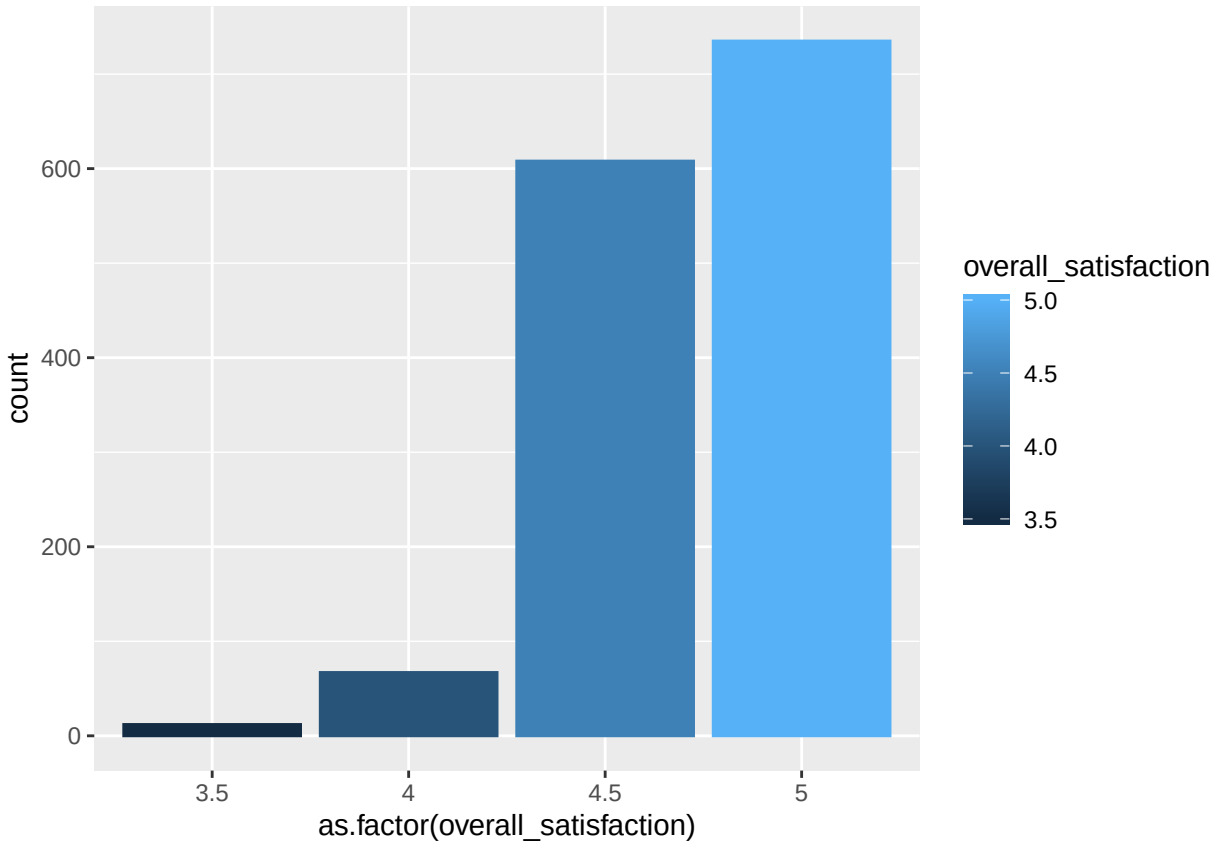


Diagram of reviews

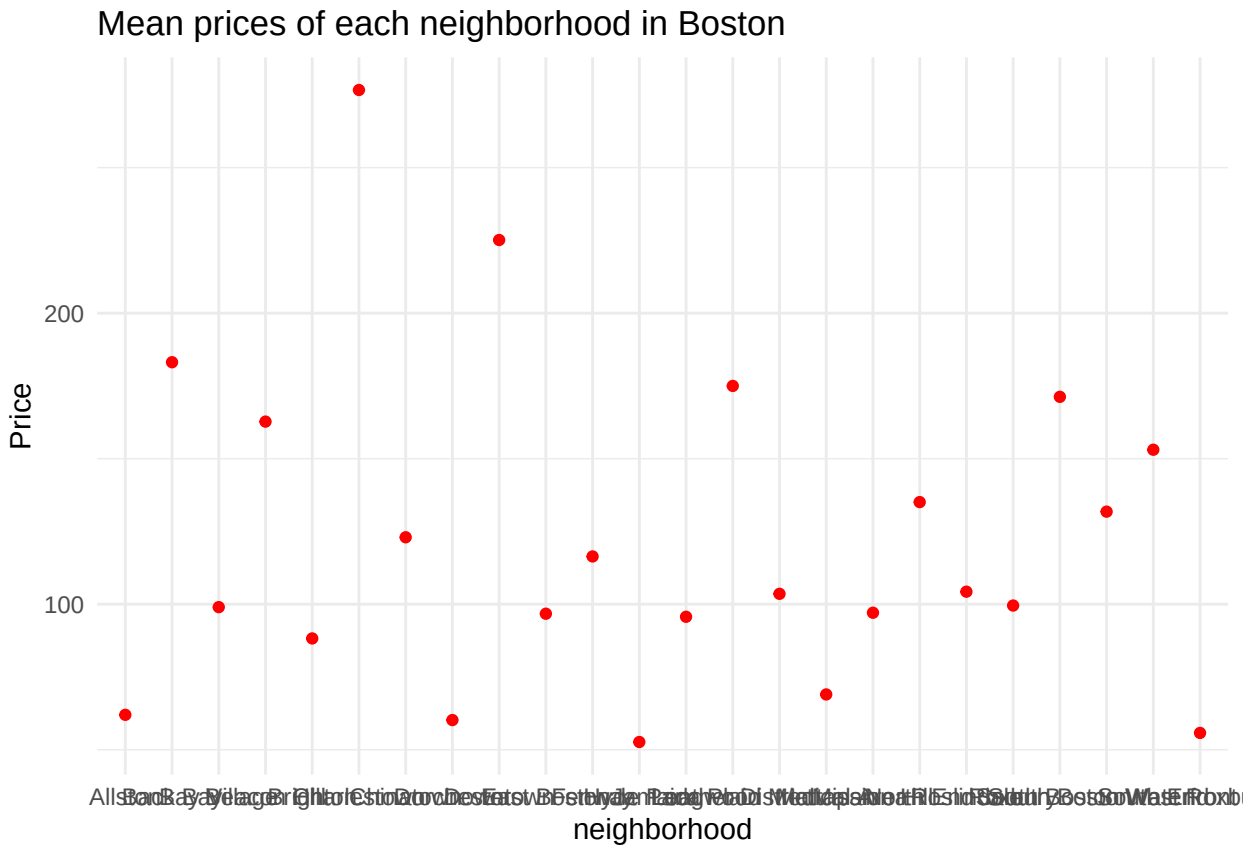
```
ggplot(data = bos, aes(x = as.factor(overall_satisfaction),
                       col = overall_satisfaction,
                       fill = overall_satisfaction)) +
  geom_histogram(stat = "count")
```



Mean prices of each neighborhood in Boston

```
#Calculate the mean of home values
MHV <- aggregate(price ~ neighborhood, bos, mean)
# Calculate the mean of home values
MHV <- aggregate(price ~ neighborhood, bos, mean)
ggplot(data = MHV, aes(x = neighborhood, y = price)) +
  geom_line(col = "orange", size = 1.5) +
  geom_point(col = "red") +
  labs(y = "Price", title = "Mean prices of each neighborhood in Boston") +
  theme_minimal()
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```


Mean prices of each neighborhood in Boston
Price
200
100
neighborhood