MA678 Midterm Project

Airbnb Analysis

Dae Hyun Lee

12/2/2019

Contents

Abstract **2**

1. Introduction **3**
1.1 Overview . . . 3
1.2 Outline . . . 3
Project goals . . . 3

2. Exploratory Data Analysis **4**

3. Methods **7**

4. Models **8**
Model 1 . . . 8
Model 2 . . . 9
Model 3 . . . 11

5. Conclusions **12**

Appendix **13**
Additional figures for EDA . . . 13
Model Checking . . . 14

Abstract

Airbnb, which started its service in 2008, was an unconventional accommodation rental service that focused on allowing travelers to live like locals with the concept of renting my home or my room.

As one of the travel buffs who love to travel, I use Airbnb a lot even more than Hotel services these days. And, I love that I can experience the traveling area as a resident, not as a tourist, through the Airbnb platform.

The first and most mentioned, when comparing Airbnb to a hotel is the price. One of the reasons why people prefer Airbnb to a hotel is because it is cheaper. And, I've noticed that nicer hotels tend to cost more than average Airbnbs in most cities. However, there were many cases that I found booking a hotel is sometimes even less expensive than renting a house/room through Airbnb.

In many cases, the booking fee set by the host seems unreasonably expensive regarding the condition of the house. Thinking of this made me decide to investigate what are the factors that mostly affect consumers in determining Airbnb.

Also, I compare how the overall Airbnb prices change and differ by time of the year. I initially assume that the overall Airbnb price in Boston would reach the highest in May due to the regional attribute that there are always a lot of people gathering over the graduation period.

1. Introduction

1.1 Overview

Throughout the project, I look for investigaing and proving a few research questions and hypotheses, such as "What are the factors that consumers consider the most when deciding where to stay through Airbnb platform" and "In Massachussetts, the overall Airbnb prices would fluctuate according to the school schedule in Boston and reach their highest point around May, when there are many graduation ceremonies."

The data I used for the project has been collected from the site 'Inside Airbnb' http://insideairbnb.com/get-the-data.html which provides more detailed data than any other opensource data websites, such as Kaggle, Tomslee. It allows everyone else to explore how Airbnb is being used around the world.

For this project, to better evaluate the effect of factors that I intend to investigate, I restrict the region especially within the area of Boston.

Below is the data description:

Table 1: List of variables

id	host_id	accommodates	instant_bookable
scrape_id	host_name	bedrooms	cancellation_policy
last_scraped	host_response_time	bathrooms	require_guest_profile_picture
calendar_last_scraped	host_response_rate	beds	require_guest_phone_verification
name	host_is_superhost	price	number_of_reviews
summary	host_listings_count	weekly_price	reviews_per_month
space	host_total_listings_count	monthly_price	review_scores_rating
description	host_verifications	security_deposit	review_scores_accuracy
neighbourhood	host_has_profile_pic	cleaning_fee	review_scores_cleanliness
neighbourhood_cleansed	host_identity_verified	guests_included	review_scores_checkin
zipcode	calculated_host_listings_count	extra_people	review_scores_communication
latitude	property_type	minimum_nights	review_scores_location
longitude	room_type	maximum_nights	review_scores_value

1.2 Outline

The outline of this project report is as follows. First, I display the conclusions of exploratory data analyses and state a specific research question of interest. Next, I describe the methods employed to answer the research questions via statistical models. Last, I interpret the models and discuss the results.

Project goals

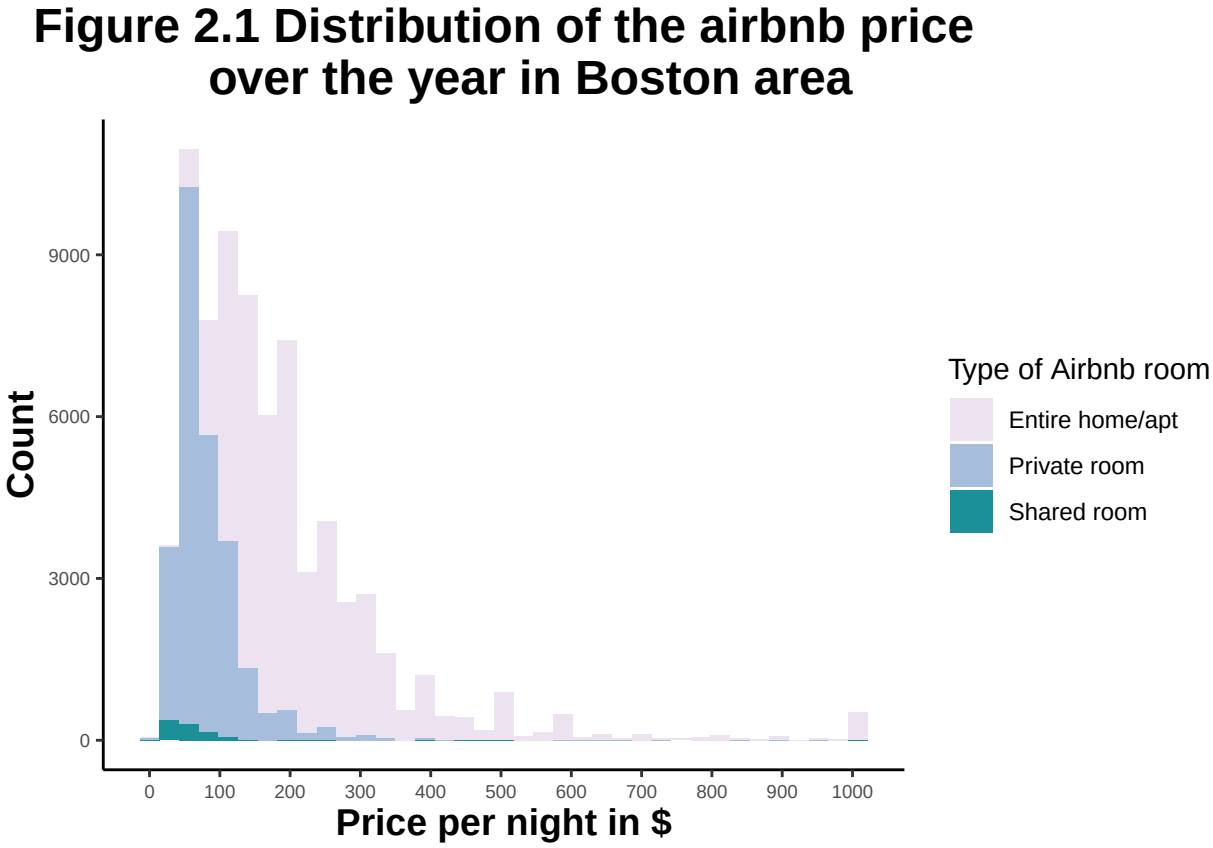
1. Test and check the factors that mostly effect consumers in determining Airbnb.
2. Investigate how the overall Airbnb prices change and differ by month within the year between Oct.2018 and Sept.2019
3. Figure out the effect of cleaning fee on other variables including the price/day.

Hypotheses

1. Airbnb prices in Boston would reach the highest in the month of May.
2. Airbnb prices in Boston would reach the lowest in the month of December and January.

2. Exploratory Data Analysis

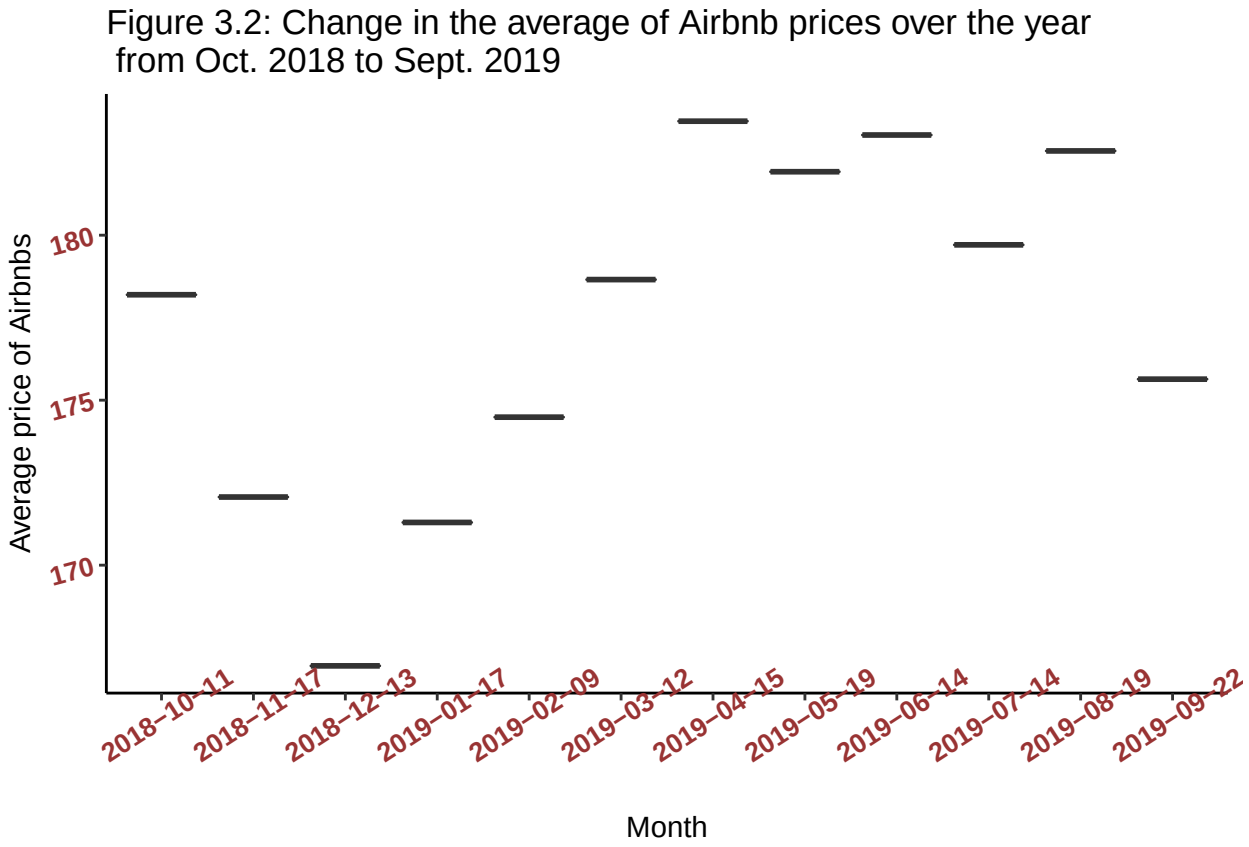
Figure 2.1




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    80.0   149.0   195.2   225.0 10000.0
```

This plot gives the general impression of how the nightly price for Airbnb listings are distributed in Boston. Most of the plots and models that I use are restricted within the price range of between zero to a hundred dollars per night. In this way, I believe would be more reliable for others as well as myself to read the trends or the patterns from the dataset I use.

Figure 2.2



This is the plot in which the average of the total Airbnb nightly rate within the Boston Area over the year between the month of October in 2018 and the September in 2019. As I expected prior to the project, the average nightly price for the Airbnb listings in Boston reached the highest roughly around the month of May.

Many visitors or tourists visit Boston every year because of its regional attributes of having numerous schools and colleges around it. Since most schools are holding their graduation ceremonies roughly from April to May, Boston is mostly crowded with many tourists, visitors and locals trying to attend graduation ceremonies for their friends or families. As a result, we can see that the price of accommodation such as hotels and Airbnb increases as well.

Also, in the winter, most students, professors, and people who study or work at universities return to their home or travel from vacation time, and most people are expected to leave Boston by December. Therefore, as in the graph above shows, the price of Airbnb reaching the lowest in the month of December within the year seems reasonable.

The variable that I used to indicate the timeline of the year is actually stored in the raw datasets as 'last_scraped' which seemed highly reliable to indicate the time-varying prices within the combined datasets that contain redundant airbnb listings within Boston district.

Figure 2.3

The simple linear regression line fitted to the dataset between the nightly rate and the cleaning fee, which both set by the host, represents that they are positively correlated. According to the regression plot, one unit increase in the cleaning fee would lead to the two-unit increase in the nightly price rate.

The brighter the color of the scatterplots indicates the higher review scores given by the guest. Also, the size of the dots represents the number of reviews that the host received from their guests. One of the impressive notions that I had while I was looking at this plot was that most of the plots at the very end of the right lower corner received high review scores for the rating. My initial assumption was that there would be many dishonest Airbnb hosts who deliberately set the price lower and additional fees, such as cleaning fees, relatively high to attract consumers. Based on the review scores, I assume those Airbnb listings with a cleaning fee set above 400 dollars would probably be the luxurious Airbnb houses. I presume that the hosts of the previously mentioned listings may have lowered prices and increased costs as part of their marketing plans. Otherwise, they would not have received that much review scores, which seems relatively high.

At the same time, there are lots of listings that are colored in dark greys below the regression line that I fitted in red color. Those were the ones that I wanted to see even before I planned for this project. Although the price for staying one night is set below 250 dollars, the hosts set the cleaning fee roughly around or above 150 dollars, which seems to be highly unreasonable as well as unacceptable.

Since the total price of a reservation on Airbnb is based on the nightly rate set by the host plus other fees or costs also determined by the host, I assume many of the hosts would willingly set the price/day lower and cleaning fee higher. Because the higher the cleaning fee is, the more the percentage of tax would be imposed on the guests than to the hosts.

To see the general patterns of what Airbnb users mostly concerns about when deciding where to stay,

3. Methods

I had initially decided to use Boston Airbnb open data from Kaggle, multiple regional Airbnb data from Tom Slee. However, one of the problems that I found from those datasets is that the price for individual Airbnb house/room is not explicitly indicated. It was not clear if the Airbnb price in the datasets is whether the total booking price set by the host, including the fee for the cleaning and service all together. I believed that it might not be reliable to make careful analyses or predictions based on the price listed in the previous datasets.

For example, when booking through Airbnb, it is often irritating to see how different the final amount is from the price shown on the website when proceeding with the payment. Because in many cases, hosts set the rental amount low, and the cleaning or service fee unreasonably high. Therefore, there could be a considerable gap between the rental price and the final amount customers ended up paying.

Therefore, I decided to find other data that holds many more observations with variables in the set. However, the problem that I had encountered while I was cleaning and wrangling the datasets is that the data from the 'Inside Airbnb' was very much untidy than I expected. In the dataset, most of the variables were treated as factor levels. For example, the price per day for the listing is even treated as a factor level with a '$' sign in addition to ',' sign. Therefore, I had to change most of the variables one by one as a numeric value or character value. Unlike the Airbnb datasets from Kaggle or Tomslee that I had initially worked with, the datasets downloaded from 'Inside Airbnb' was not organized in a way that is helpful to make analyses. However, I had to use these datasets, especially because I wanted to see the effect of the cleaning fee on other variables.

The cleaning and wrangling process that I had conducted for these datasets is stored in other Rmd file. Due to the amount of the work that I had made for just cleaning and wrangling, I decided to separate the part of the cleaning process from this analysis file.

4. Models

Model 1

Test and check the factors that mostly effect consumers in determining Airbnb.

```
##
## Call:
## lm(formula = review_scores_rating ~ number_of_reviews + reviews_per_month +
##     require_guest_profile_picture + cleaning_fee, data = list_tot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.072  -2.789   1.644   5.002  11.803
##
## Coefficients:
##                                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                   92.8052137  0.0747023 1242.335  < 2e-16 ***
## number_of_reviews              0.0060655  0.0007243    8.374  < 2e-16 ***
## reviews_per_month              0.2611386  0.0208647   12.516  < 2e-16 ***
## require_guest_profile_picture -4.1669946  0.2369896  -17.583  < 2e-16 ***
## cleaning_fee                  -0.0033540  0.0007217   -4.647 3.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.974 on 51890 degrees of freedom
##   (22070 observations deleted due to missingness)
## Multiple R-squared:  0.01842,    Adjusted R-squared:  0.01834
## F-statistic: 243.4 on 4 and 51890 DF,  p-value: < 2.2e-16
```

- (Intercept): Expected review score for the Airbnb listing which has a non-cleaning fee, non-required guest profile pic, zero review per month and zero number of reviews is 92.81
- number of reviews: On average review score for the Airbnb listing increase by 0.006 with every one more review on the listing.
- reviews per month: one more review per month increases 0.26 in the rating score.
- require guest profile picture: Difference in requiring the guest for the profile picture with the listing which does not require for the guest picture is -4.16.
- cleaning fee: On average increase in one dollar for the cleaning fee decreases guest's rating score for the host.

Model 2

The likelihood of booking Airbnb would reach the highest in the month of May.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.340   1.180   2.014   3.090  53.080    5372
```

```
##
## Call:
## glm(formula = popularity ~ factor(last_scraped), data = l_Min_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4085  -0.3650  -0.3519   0.6219   0.6510
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.391568   0.008845  44.270  < 2e-16 ***
## factor(last_scraped)2018-11-17  0.016924   0.012433   1.361 0.173462
## factor(last_scraped)2018-12-13  0.001359   0.012417   0.109 0.912850
## factor(last_scraped)2019-01-17 -0.013428   0.012406  -1.082 0.279101
## factor(last_scraped)2019-02-09 -0.026598   0.012405  -2.144 0.032034 *
## factor(last_scraped)2019-03-12 -0.037055   0.012402  -2.988 0.002812 **
## factor(last_scraped)2019-04-15 -0.039698   0.012396  -3.202 0.001364 **
## factor(last_scraped)2019-05-19 -0.042573   0.012385  -3.437 0.000588 ***
## factor(last_scraped)2019-06-14 -0.037543   0.012379  -3.033 0.002424 **
## factor(last_scraped)2019-07-14 -0.037260   0.012373  -3.011 0.002601 **
## factor(last_scraped)2019-08-19 -0.038684   0.012369  -3.128 0.001764 **
## factor(last_scraped)2019-09-22 -0.040130   0.012379  -3.242 0.001189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2319633)
##
##     Null deviance: 8548.9  on 36805  degrees of freedom
## Residual deviance: 8534.9  on 36794  degrees of freedom
## AIC: 50685
##
## Number of Fisher Scoring iterations: 2
```

The probability of making reviews during the specific month for Airbnb listings in Boston is,

$Pr(popularity = 1) = logit^{-1}(0.391568 + 0.016924 * Nov. + 0.001359 * Dec. - 0.013428 * Jan. - 0.026598 * Feb. - 0.037055 * Mar. - 0.039698 * Apr. - 0.042573 * May. - 0.037543 * Jun - 0.037260 * Jul - 0.038684 * Aug. - 0.040130 * Sep.)$.

The regression coefficient for making reviews on May is -0.0426, which indicates that the month of May has a multiplicative effect of $e^{-0.042573} = 0.9583$ on the odds of booking 'Airbnb' compared to other months over the year in Boston.

The value of $invlogit(0.391568 - 0.042573) = 58.63738\%$ is the estimated probability of making reviews on that given month. To quickly interpret the coefficient for May on probability scale, I divide the coefficient estimate for male by four: $\frac{0.042573}{4} = 0.01064325$.

Thus, on May people are 1.06% more likely to book Airbnbs and make reviews after their stays. Regressing the 'review per month' onto the 'month' variable is highly reliable to check whether to check the number of people using Airbnb over month because to make any reviews for Airbnb, the customers must stay in the listed houses/rooms before making any reviews. Based on the regression results, I have strong evidence

that the number of booking Airbnb, as well as leaving reviews, would reach the highest in May, especially in Boston. Therefore, I would like to conclude that the model supports the hypothesis that Airbnb prices in Boston would reach the highest in May.

Model 3

Investigate the effect of cleaning fee on the popularity variable including the review scores.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## popularity ~ cleaning_fee + scale(review_scores_value) + scale(review_scores_accuracy) +
##     scale(review_scores_cleanliness) + scale(review_scores_checkin) +
##     scale(review_scores_communication) + scale(review_scores_location) +
##     (1 | id)
##    Data: l_Min_2
##
##      AIC      BIC   logLik deviance df.resid
##   9274.5   9349.8  -4628.3   9256.5    31666
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4382 -0.0115 -0.0081  0.0816  4.6224
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 170.7    13.07
## Number of obs: 31675, groups:  id, 2696
##
## Fixed effects:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -8.680542   0.239979 -36.172  < 2e-16
## cleaning_fee                      -0.003544   0.002035  -1.741   0.0816
## scale(review_scores_value)         0.025803   0.112594   0.229   0.8187
## scale(review_scores_accuracy)      0.073390   0.113206   0.648   0.5168
## scale(review_scores_cleanliness)   0.192383   0.118064   1.629   0.1032
## scale(review_scores_checkin)       0.907498   0.140119   6.477 9.38e-11
## scale(review_scores_communication) -0.638912   0.127244  -5.021 5.14e-07
## scale(review_scores_location)      0.158232   0.099166   1.596   0.1106
##
## (Intercept)                       ***
## cleaning_fee                      .
## scale(review_scores_value)
## scale(review_scores_accuracy)
## scale(review_scores_cleanliness)
## scale(review_scores_checkin)      ***
## scale(review_scores_communication) ***
## scale(review_scores_location)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##                 (Intr) clnng_ scl(rvw_scrs_v) scl(rvw_scrs_cc)
## cleaning_fe     -0.654
## scl(rvw_scrs_v) -0.052  0.075
## scl(rvw_scrs_cc)  0.018 -0.009 -0.214
## scl(rvw_scrs_cl)  0.041 -0.062 -0.179          -0.169
## scl(rvw_scrs_ch) -0.051  0.019 -0.045          -0.084
```

```
## scl(rvw_scrs_cm)  0.018  0.034 -0.163          -0.180
## scl(rvw_scrs_l)   0.044 -0.062 -0.178          -0.112
##                  scl(rvw_scrs_cl) scl(rvw_scrs_ch) scl(rvw_scrs_cm)
## cleaning_fe
## scl(rvw_scrs_v)
## scl(rvw_scrs_cc)
## scl(rvw_scrs_cl)
## scl(rvw_scrs_ch) -0.015
## scl(rvw_scrs_cm) -0.090           -0.349
## scl(rvw_scrs_l)  -0.041            0.009           -0.001
## convergence code: 0
## Model failed to converge with max|grad| = 19.0722 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

Fiting a classical logistic regression predicting $Pr(y_{ij} = 1)$ given the Airbnb listing i's 6 review scores j.

5. Conclusions

This project entailed testing a set of hypotheses pertaining to the effect that a particular set of factors had on the Airbnb prices.

I took the approach of first cleaning and manipulating the data so as to facilitate its use with visualization techniques that would, then, go on to serve as a prelude to model fitting. Performing analysis on each of the hypotheses resulted in each hypothesis being strongly supported by the data.

In summary, I can infer that the regional attributes of Boston within the scope of the provided hypotheses do affect the price of Airbnb listings.

Appendix

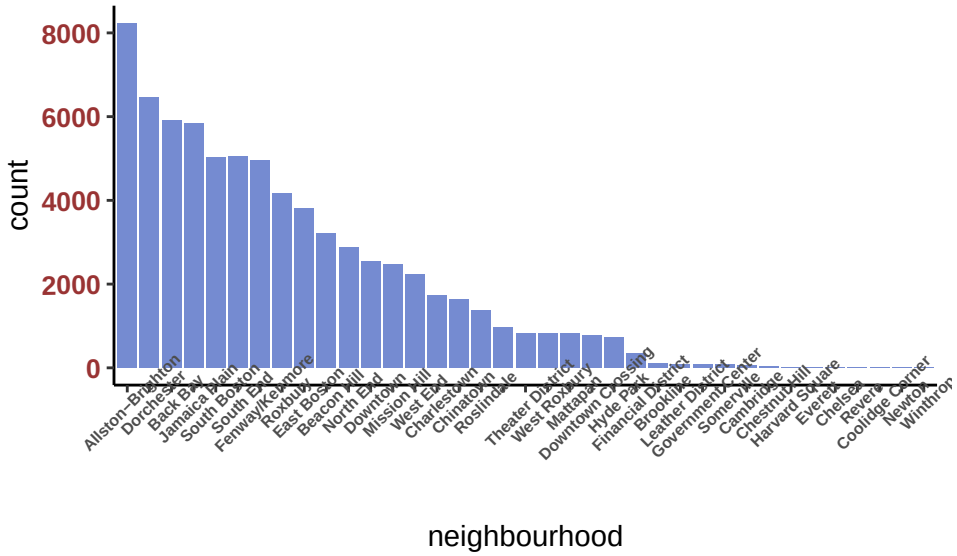
Additional figures for EDA

Summary.Reg_2.3

```
##
## Call:
## lm(formula = price ~ cleaning_fee, data = list_tot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1213.4   -75.3   -30.2    25.1  9942.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.54806    1.92401   29.91   <2e-16 ***
## cleaning_fee  1.99328    0.02197   90.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 271.5 on 62339 degrees of freedom
##   (11624 observations deleted due to missingness)
## Multiple R-squared:  0.1166, Adjusted R-squared:  0.1166
## F-statistic:  8229 on 1 and 62339 DF,  p-value: < 2.2e-16
```

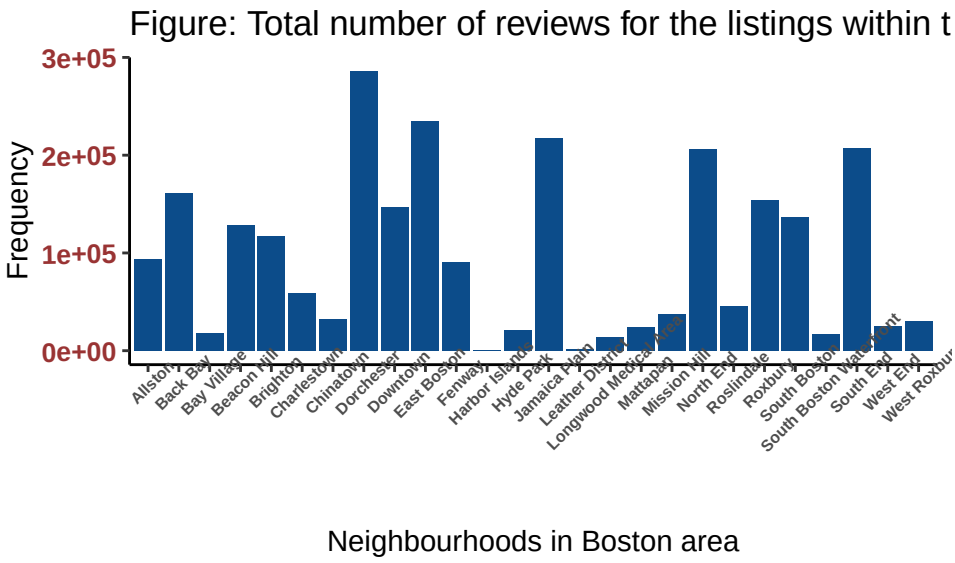
Summary of the Linear Regression fitted to test the relationship between the price and the cleaning fee.

Figure.A



It is clear to see the overall number of listings around Boston Area.

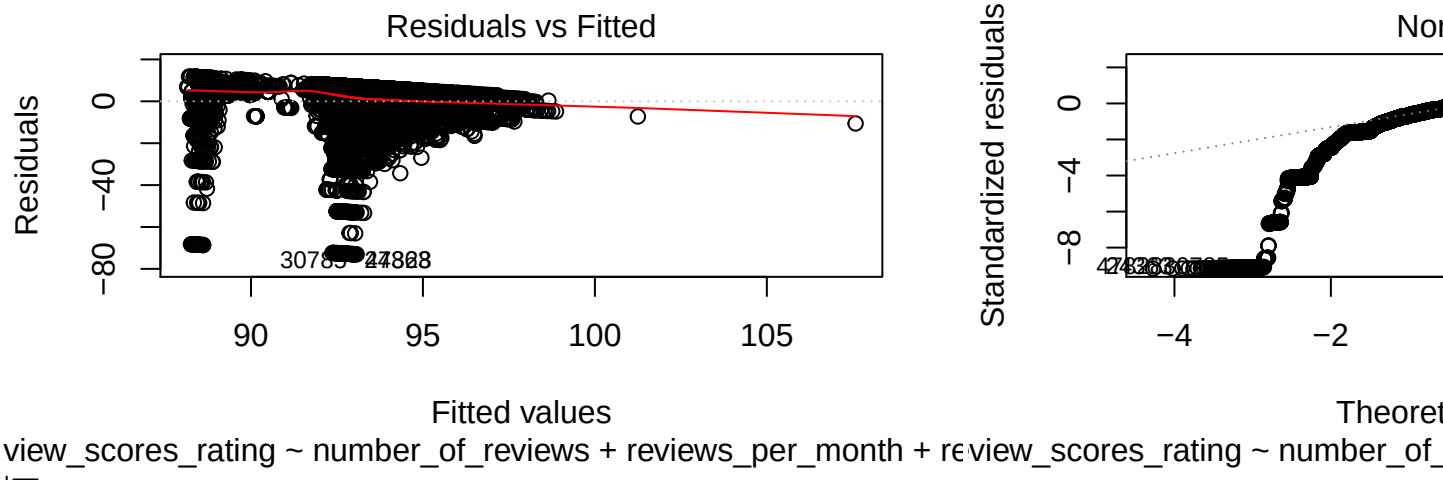
Figure.B

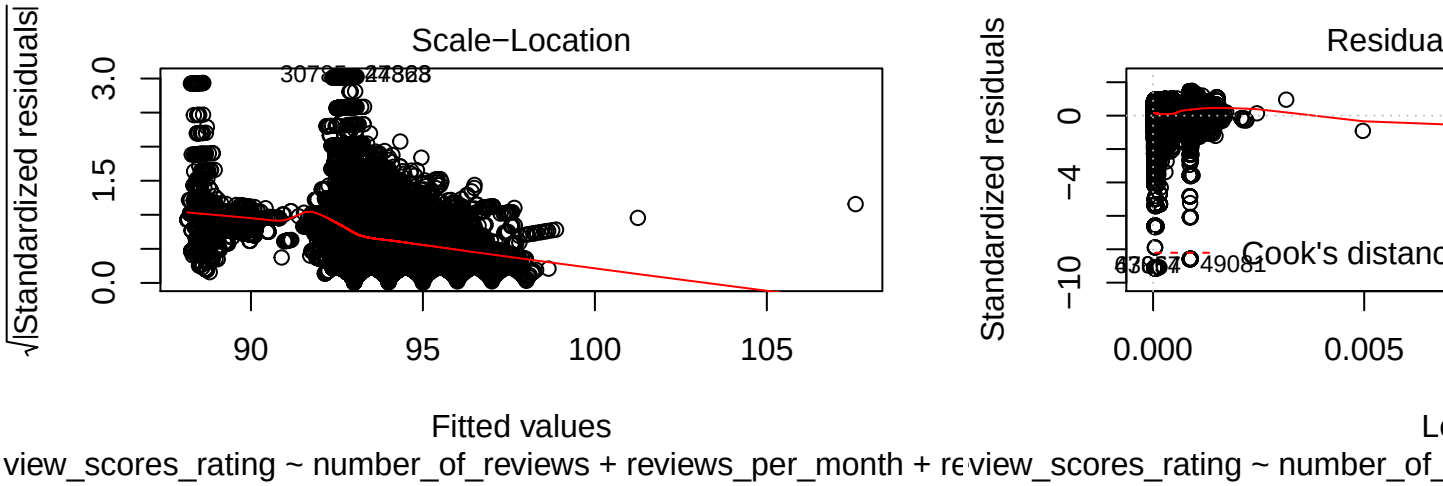


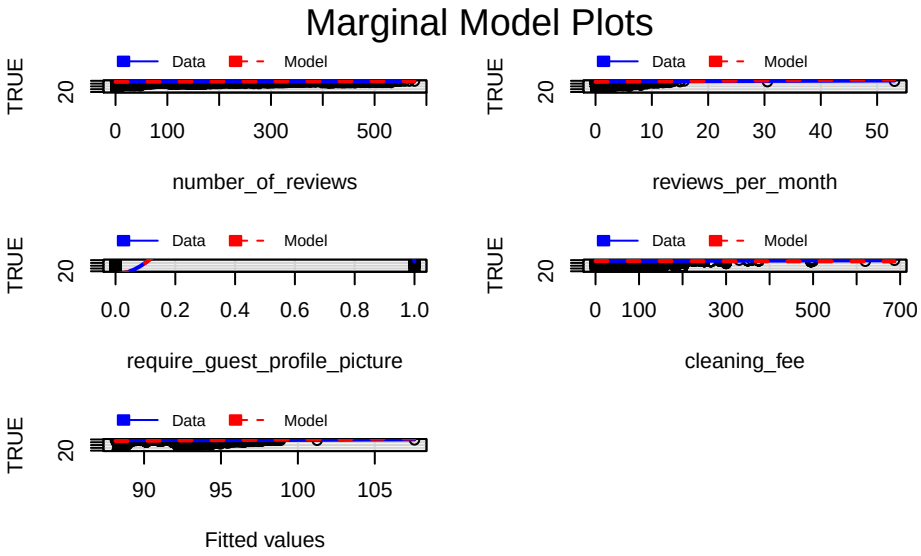
It shows the total number of reviews over the neighbourhoods in Boston. It was useful to see which areas are typically famous to stay among the users.

Model Checking

Model.1






Marginal Model Plots
Data
Model
TRUE
20
0
100
300
500
number_of_reviews
0
10
20
30
40
50
reviews_per_month
0.0
0.2
0.4
0.6
0.8
1.0
require_guest_profile_picture
0
100
300
500
700
cleaning_fee
90
95
100
105
Fitted values