Airbnb in Boston

Haozhe Chen

Dec 05, 2019

Abstract

This project mainly focuses on the review data and the hotel lists and from Airbnb official website in the effort to understand which factors could have potentially affected the reviews and comments. Understanding the reviews is important for both travelers and hosts, since travelers want to know which rooms are potentially better and house hosts could enhance their competitiveness. From several different features, such as locations, room types, and hosts, I started with Exploratary Data Analysis and then utilized different statistical models to explore the potential relationships between the ratings and features. I also analyzed the text comments, which has always been a crucial part of feedback in any kind of business because it could reflect the emotions of customers.

Introduction

Background

Airbnb is a non-traditional way to connect guests looking for accommodations to hosts looking to rent their properties on both a short-term or long-term basis. Airbnb has steadily risen in terms of revenue growth and its range of service provisions. As of 2019, there 150 million users of Airbnb services in 191 countries, making it a major disruptor of the traditional hospitality industry (this is akin to how Uber and other emerging transportation services have disrupted the traditional intra-city transportation services). Airbnb generates revenue by charging its guests and hosts fees for arranging stays: hosts are charged 3% of the value of the booking, while guests are charged 6%-12% per the nature of the booking. As a rental ecosystem, Airbnb generates tons of data including but not limited to: density of rentals across regions (cities and neighborhoods), price variations across rentals, host-guest interactions in the form of reviews, and so forth.

One of the main reference for most people of choosing a hotel or apartment was the reviews and they should pay more attention to high-rating apartments I'm always interested in what factors could play a role and make diffferent rooms end up with different level of ratings. It might be locations, quality, or the consistency of the description. In this project, I am going to explore both the numerical data and text reviews, and try to look for interesting features in the business side.

Method

Data Source

I used the dataset of Boston Airbnb room lists and review lists. The raw data used on this project is taken from Inside Airbnb(Link:http://insideairbnb.com/get-the-data.html).

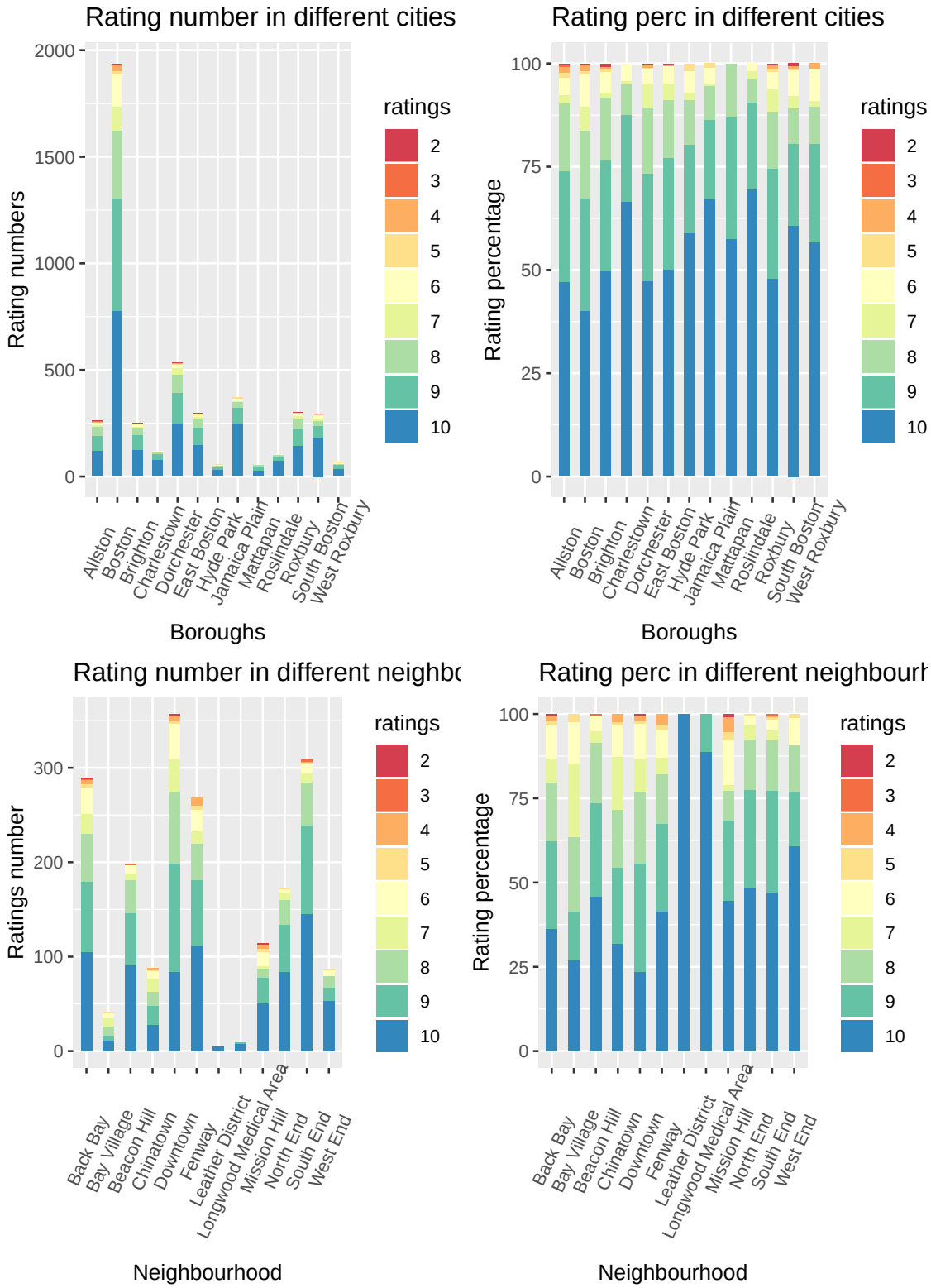
For sentiment analysis, this project used bing dictionary from (https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html). It categories words with positive and negative attitudes to binary scores.

Exploratory Data Analysis

1) Histogram of Ratings

From the visualization, we could see that people tend to give higher rating scores in general. Almost no people give ratings below 60% percent of the overall weighted score. But we should be careful with the results because it may not reflect the real thought of costumers and we do not know how they get the rating scores in the survey.

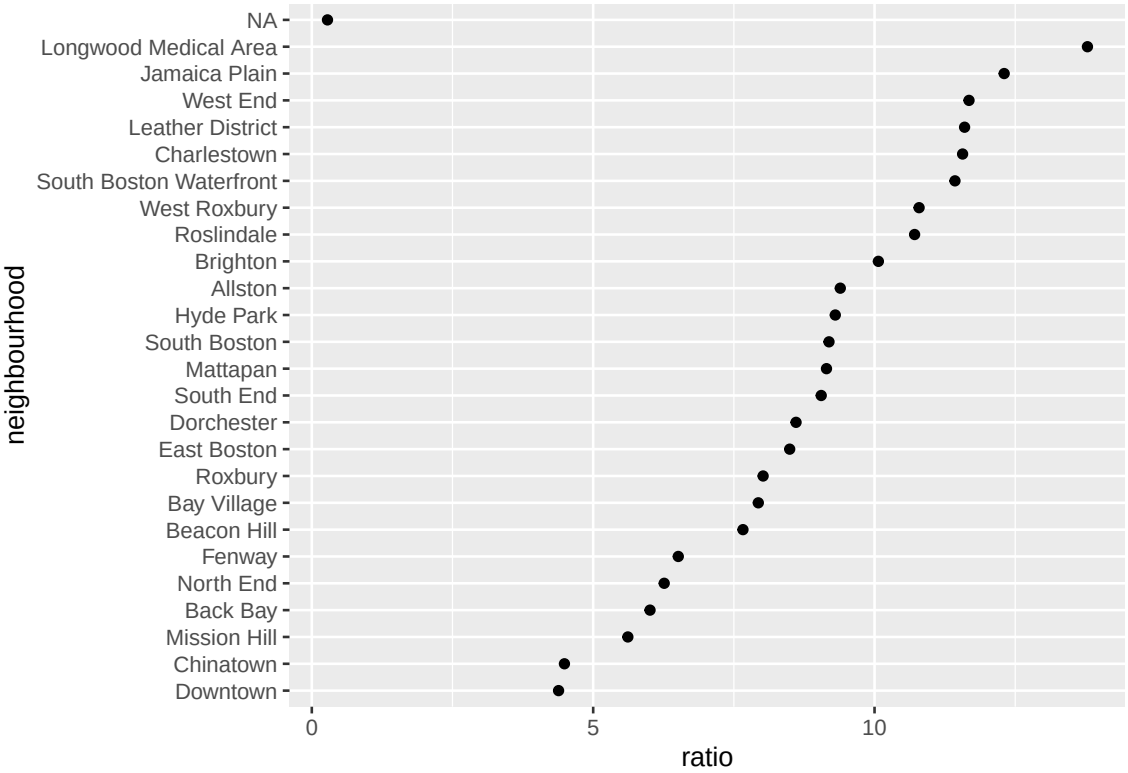
2) Demographic Distribution of Ratings



From the plot above, we could see that the Boston city has overwhelmingly more reviews than other boroughs, but the percentage of 10 scores is lower than other boroughs. And the Charlestown, Roslindale, Jamaica Plain have the highest ratings in general. In the city of Boston, Longwood Medical Area and Leather District have received better ratings. On the other hand, Chinatown and Downtown have the lowest ratings in general. (This makes sense because of the living quality are lower and downtown due to the compact space, louder

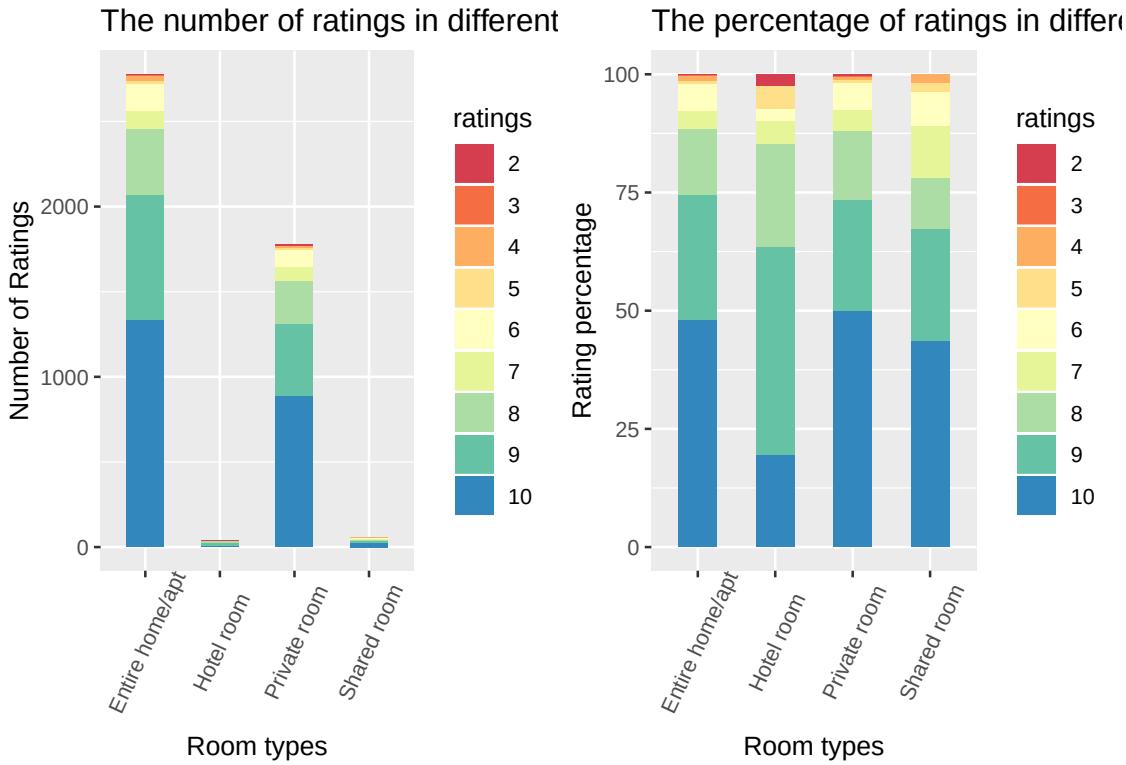
noise and higher prices)

3) The demographic differences in negative and positive reviews



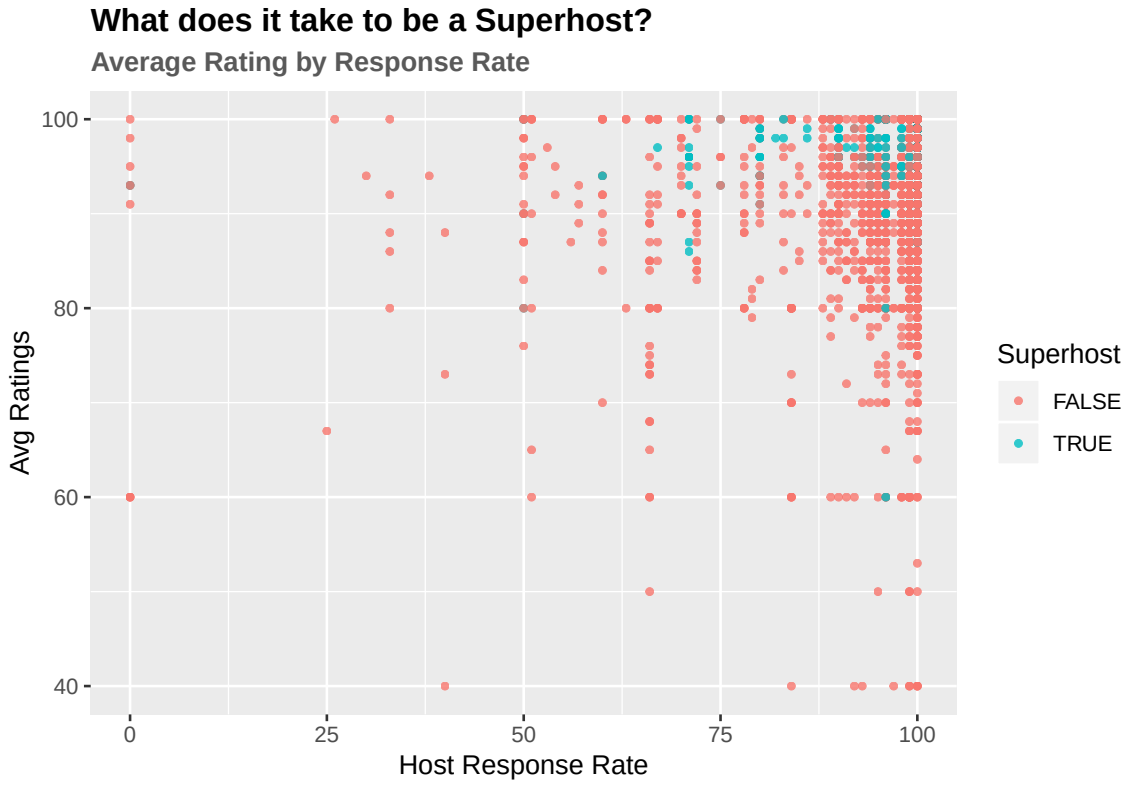
This plot indicate that there are some difference of the sentiment in different neigbourhoods. Downtown and Chinatown have the lowest ratio of positive comments to negative comments, which correspond to the previous analysis.

4) Distribution of Ratings in Room Types



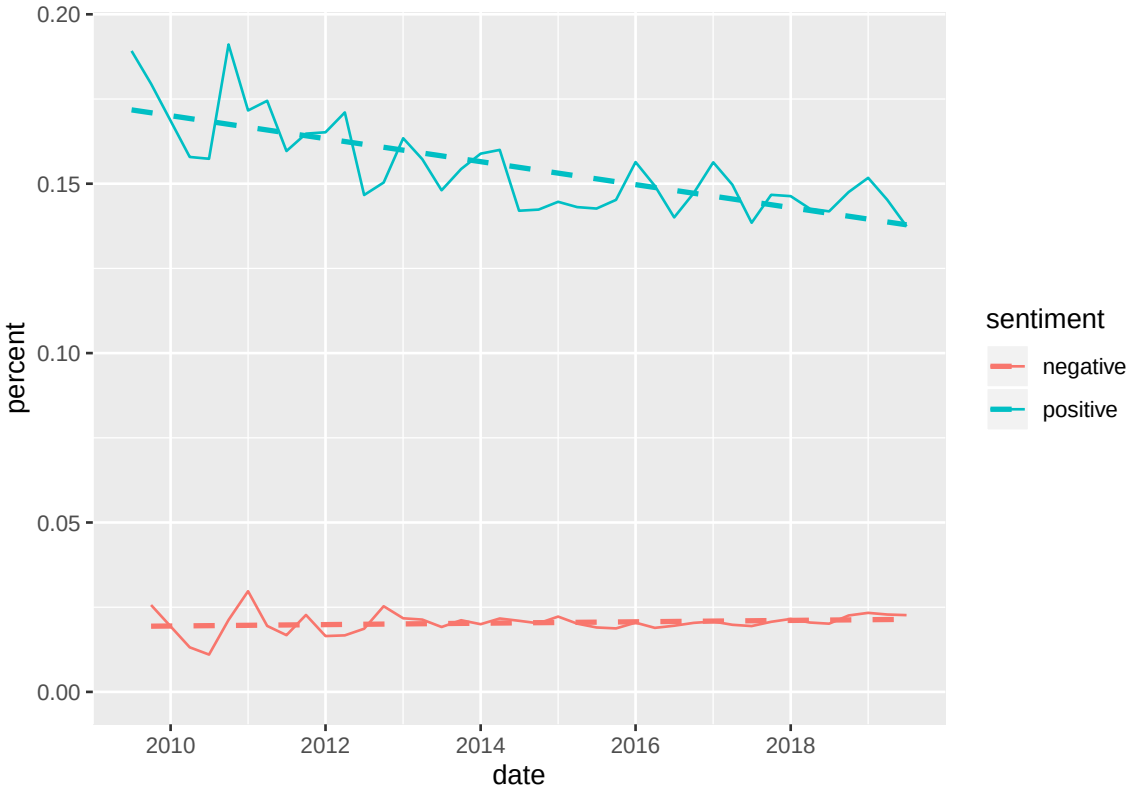
This plot indicates that there are more apartments and private rooms in the provided by Airbnb hosts. Also, hotel rooms are less popular than the other types of rooms, this more or less indicated that people who choose Airbnb are willing to look for some home-liked place to live rather than a traditional hotel.

5) Relationship of Hosts and Rating



Airbnb awards the title of "Superhost" to a small fraction of its dependable hosts. This is designed as an incentive program that is a win-win for both the host, Airbnb, and their customers. The super hosts get more business in the form of higher bookings, the customer gets improved service and Airbnb gets happy satisfied customers. In this part, take a look at the plot of some attributes of hosts vs rating scores. There is no obvious relationship between rating scores and the host response rate. Take a clear examination of the blue plot, it indicates that the hosts are super hosts, and it is clear that they tend to get higher ratings than those who are not not super hosts.

6) Sentiment over time



Interestingly, the number of positive reviews get fewer from 2010 to 2018, and the negative review comments have not changed a lot.

Statistical analysis

The outcome can be identified as either 10 levels (ratings) and continuous numbers (original ratings), the predictors include locations, hosts, prices, room types and the percentage of positive words in text comments. For the categorical outcome, binomial or multinomial regression are good selections. I have also considered Poisson regression in this project, since the ten levels can be also identified as 10 counts, each level add 1 to the original scale; however it cannot input some variables and it is less interpretable than binomial models in this case, I excluded it. For the continuous numbers, linear regression is also a good choice, also, some different groups in the dataset, such as ids, hosts, neighborhoods, cities.. can be treatment as random effects. Through consideration and examination, I have figured that neighborhoods may have the best group effect because in this project each observation represents an individual property, the ratings are the average ratings for these properties.

Model Comparison

The BIC and AIC suggest that the binomial model is better.

```
##       df       BIC
## fit.1 42  6557.522
## fit.2 35  2958.124
## fit.3 12  2817.818
## fit.4 36 18722.839
## fit.5 13 18613.553
```

```
##       df       AIC
## fit.1 42  6304.505
## fit.2 35  2747.277
## fit.3 12  2745.528
## fit.4 36 18505.968
## fit.5 13 18535.239
```

The pooling model is better.

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: ratings_bin
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev
## NULL                                 3053     4150.9
## neighbourhood       24   267.54      3029     3883.3
## number_of_reviews    1    34.53      3028     3848.8
## price                1    14.75      3027     3834.0
## cleaning_fee         1     0.30      3026     3833.7
## room_type            3     9.50      3023     3824.2
## host_response_rate   1    39.94      3022     3784.3
## Superhost            1   758.43      3021     3025.9
## positve_perc         1   348.55      3020     2677.3
## is_location_exact    1     0.03      3019     2677.3
```

Result

Model Selection

The binomial model with random effect is seleceted to interpret

```
## glmer(formula = ratings_bin ~ scale(number_of_reviews) + (1 |
##     neighbourhood) + scale(price) + scale(cleaning_fee) + room_type +
##     host_response_rate + Superhost + scale(positve_perc) + is_location_exact,
##     data = model.df.bin, family = binomial(link = "logit"))
##                           coef.est coef.se
## (Intercept)               -2.55     0.67
## scale(number_of_reviews) -0.18     0.05
## scale(price)               0.13     0.06
## scale(cleaning_fee)        0.07     0.07
## room_typeHotel room       -0.73     0.68
## room_typePrivate room     -0.11     0.13
## room_typeShared room       0.13     0.55
## host_response_rate         0.01     0.01
```

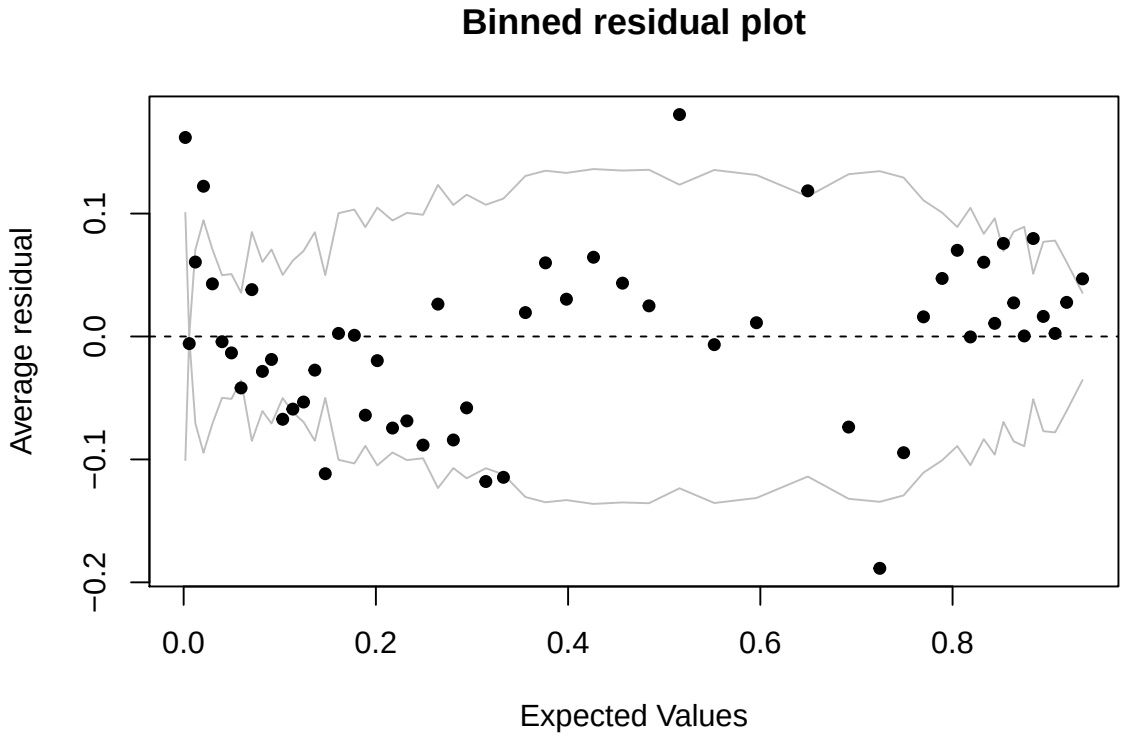
```
## SuperhostTRUE             2.08     0.11
## scale(positve_perc)       1.43     0.09
## is_location_exactTRUE     0.03     0.13
##
## Error terms:
##  Groups        Name        Std.Dev.
##  neighbourhood (Intercept) 0.28
##  Residual                  1.00
## ---
## number of obs: 3054, groups: neighbourhood, 25
## AIC = 2745.5, DIC = 2659.6
## deviance = 2690.6
```

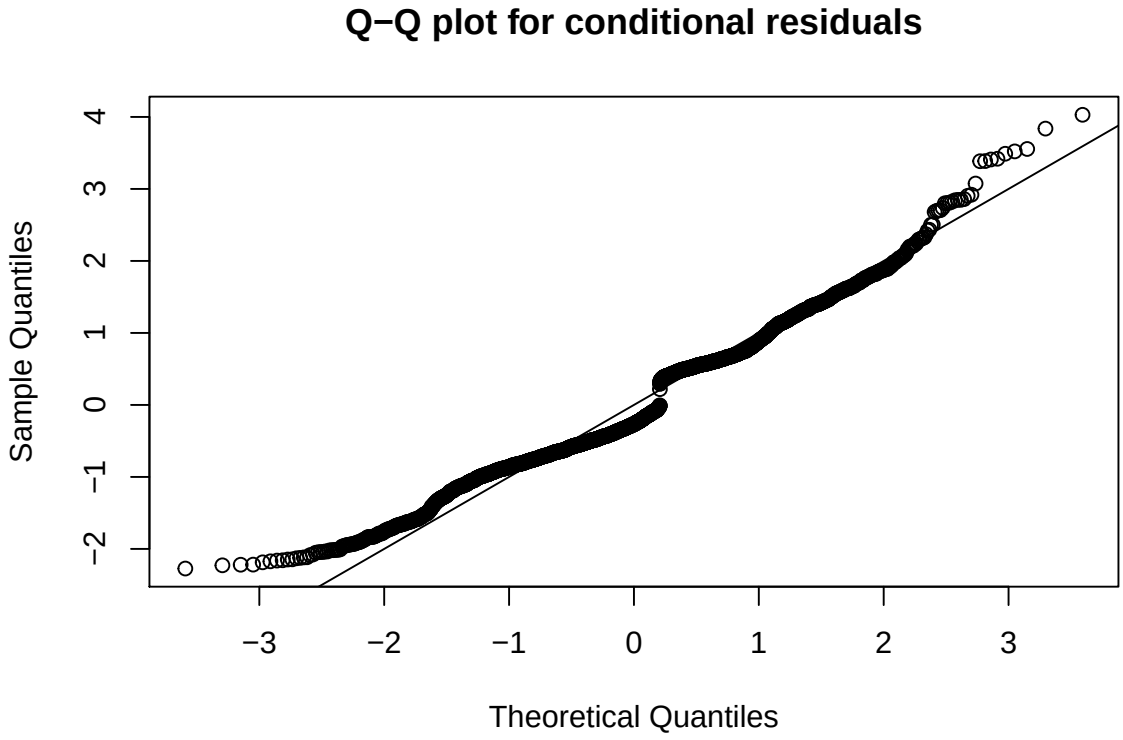
Interpretation

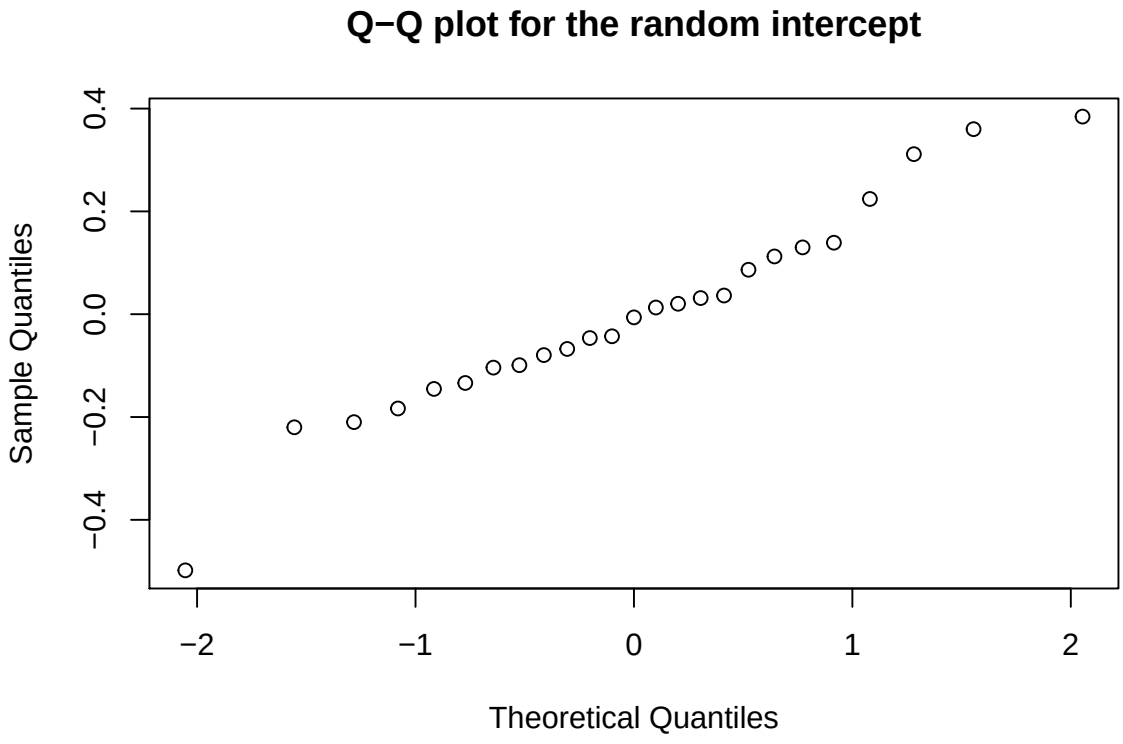
For the fixed part, the rating scores of reviews are not likely to be affected by the price of the cleaning fee. The baseline for the room type is the apartment. Compared to the baseline, on average the hotel room decreases 2 scores in the 100-score rating standard, the private room only decreases 0.38, in other words, there is no big difference, the shared room decreases 0.61 scores on average. For the host, if they respond faster, they tend to get better ratings, but not too much, since the coefficient is only 0.08; however, if the host is a super host, the chance of getting a will improve significantly, they 5 times chances than those who are not super hosts. But, we still need to be careful because the number of super hosts may not be large enough to explain the difference. The non-super are hosts are the majority.

For the random effect, $\sigma_\alpha^2 : \sigma_y^2 = 0.33^2 : 4.98^2 = 0.004391058$, the variance among the average rating scores of the different neighborhoods is lower than the within-neighborhood variance in rating scores measurements, which means the pooling effect is strong and the overall estimate is more reliable.

Model Checking







From the binned residual plot, there are still a few of points outside of the grey line area, which means this model may be not able to capture 95% of our data.

Discussion

Implication

Through EDA, text mining and modeling of the data from Airbnb insights, most people tend to give better rating scores to their living experience and service from Airbnb hosts. Also, the number of positive comments is much larger than negative comments.

What factors affect people when they determine give how many stars to a hotel or B&B? Through the analysis, I 've found that the demographic location, room types are good indicators and hosts are good indications. Specifically, in the Boston city, Downtown and Chinatown are not good choices to live, apartments and private rooms are more popular than hotel rooms. The most intuitionistic plots have shown this trend; moreover, from the word cloud of comments, the most negative comments are about noise, the most positive comments have reflected that Airbnb users love a clean and a quiet place. All of these suggest that people are not that much care about the downtown location, possibly because Airbnb users are looking for someplace to live for serval days or months, not just stop by or for business purposes. In the traditional hotel industry, people pay more attention to the convenience of transportation or the location near to the downtown area, where companies clustered. Do not forget. a title of a super host always gets higher ratings.

Tips for Airbnb users:1. Choose an apartment or private house, it may give you better service and living quality. 2. The super host is not a bad choice, because their interaction with the official Airbnb platform may force them to provide better service. 3. If you come to Boston, not for business, Charlestown, Brighton, Roslindale, Jamaica Plain might become better choices for you. 4. Pay attention to the number of reviews, do not trust a 5.0 star's house only have 1 or 2 reviews.

Suggestions for the first clients of Airbnb (property hosts): 1. Try to become a super host 2. Response more and faster. 3. Do not do hotels, you will not find your target travelers and you are not able to compete with brand hotels. 4.Provide the service as best as you can, sufficient numbers of good reviews will bring you more and more customers in the future.

Limitation & Future Direction

In this project, I have not yet answered the question that I have proposed: "Do previous comments affect new costumers?" since such analysis follows a completely data-wrangling method, which will mess up the current dataset, but it can be done in the future and I' ll completed that because it's an interesting topic which could provide insights for both Airbnb users and Airbnb hosts. Also, I have not figured out how to do seasonality trends for the rating scores because the list only gives average ratings for each individual property. If there is a rating score in the comments dataset with a specific date, this could have been finished. Instead, I did the sentiment analysis and found that the comments tend to become less positive over time.

In the part of the modeling and statistical analysis, several models are used and they can tell stories. However, I hope I can have more tricks to analysis the model and the goodness of fit in the future.

Referece

[1] Gelman, Andrew, and Jennifer Hill. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, 2017.

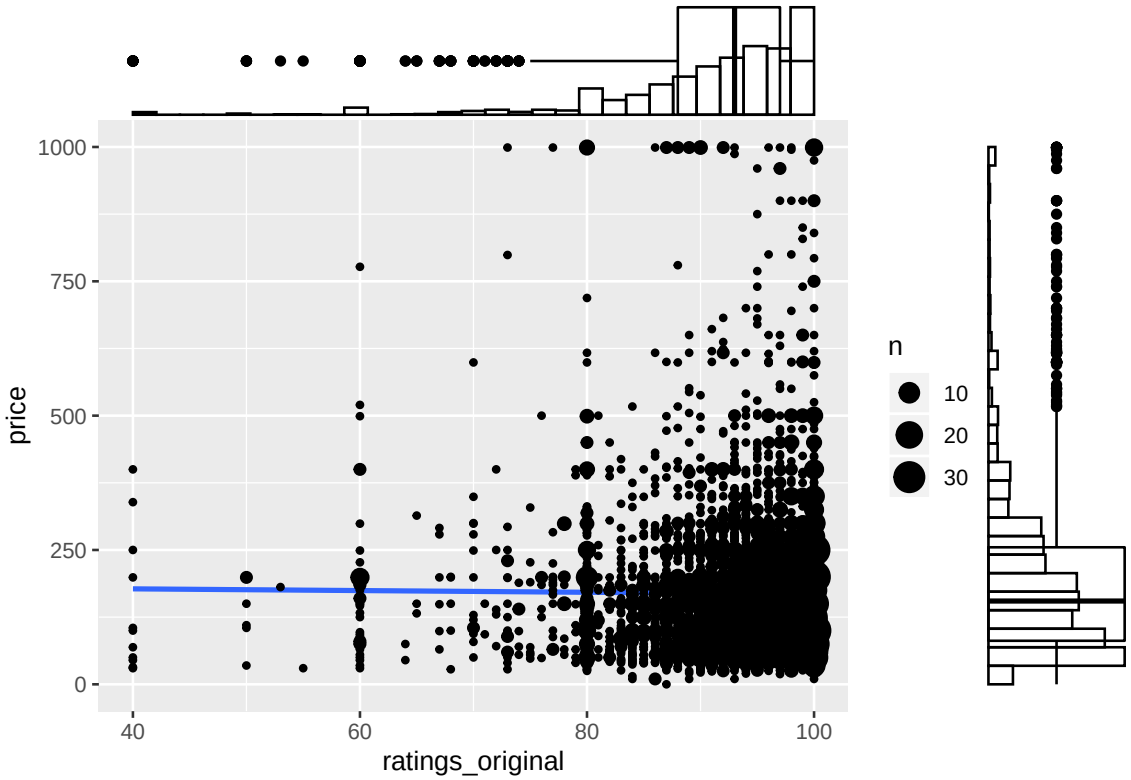
[2] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, Brian Marx. Regression Models, Methods and Applications.

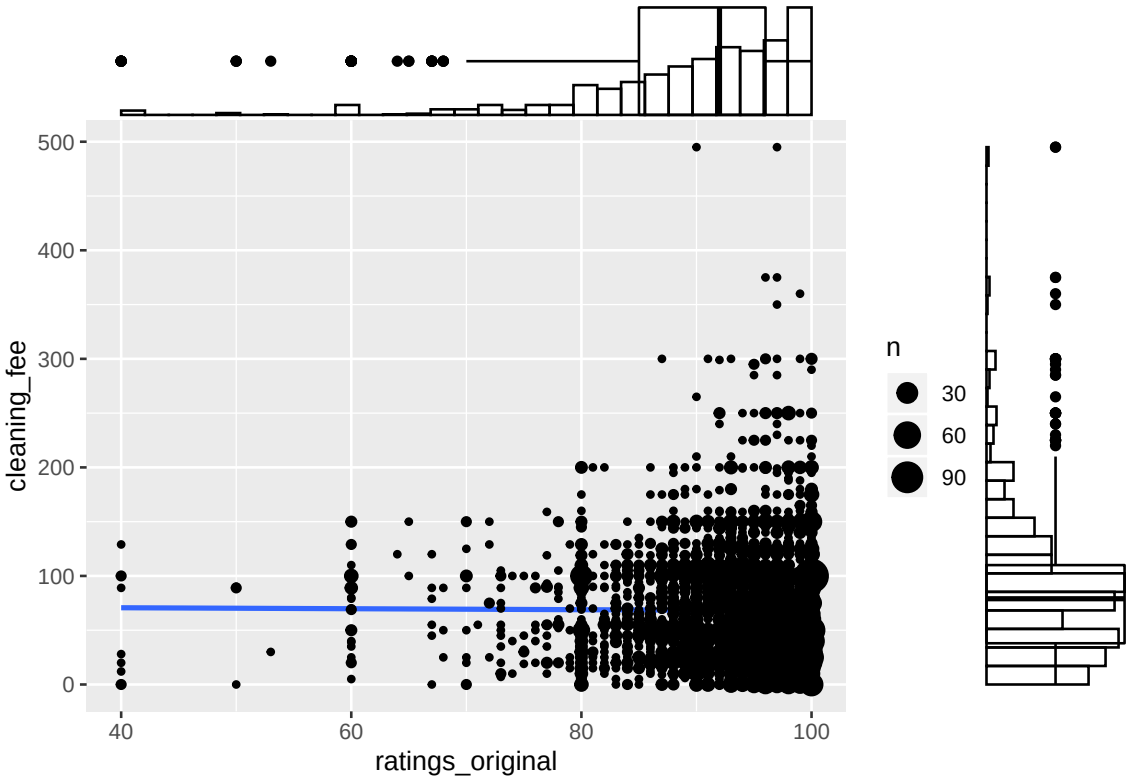
[3] Sarang Gupta. Airbnb Rental Listings Dataset Mining. https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec

Appendix

Other EDAs

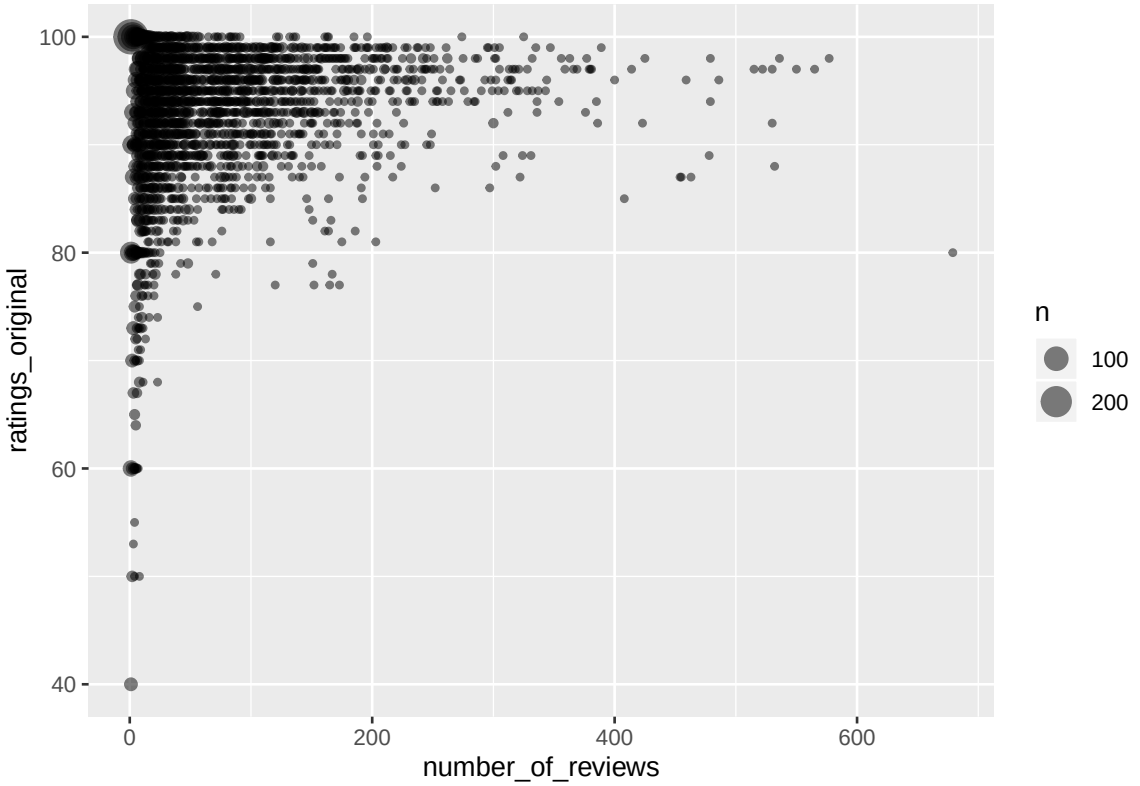
1) Relationship of Price and Rating





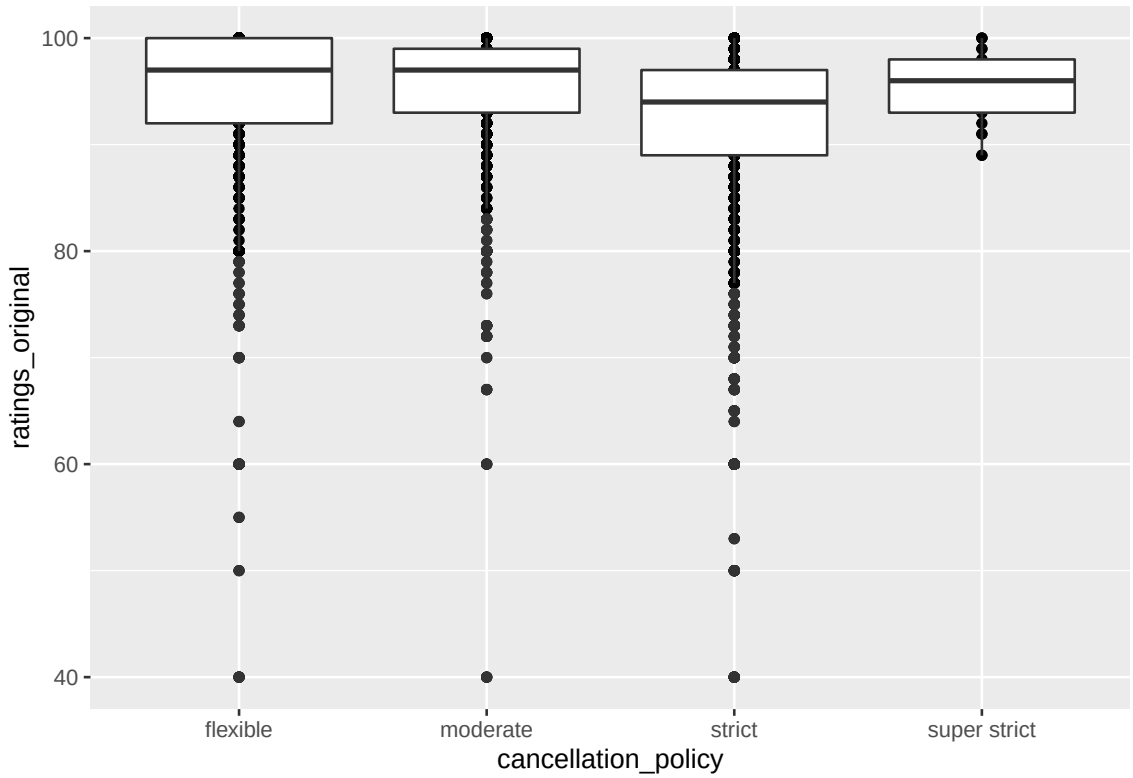
This analysis focuses on the original rating scores (0-100 scores). There's no obvious pattern indicating that price will affect the ratings or living experience. However, from the plot of service fee vs rating, we could examine that between 90 and 100 scores, there are more points representing a higher cleaning fee, possibly because of higher cleaning fees means better service, therefore better feedback from customers.

2) Number of reviews may also be a good indication



Intuitively, higher ratings with more reviews are usually the best selections and the number of these top selections is usually very small. In this part, take a look at the plot, high ratings are clustered at the top left of the plot, which is also the points of rooms with fewer reviews. Only a few points have 90 or higher scores with 400+ reviews.

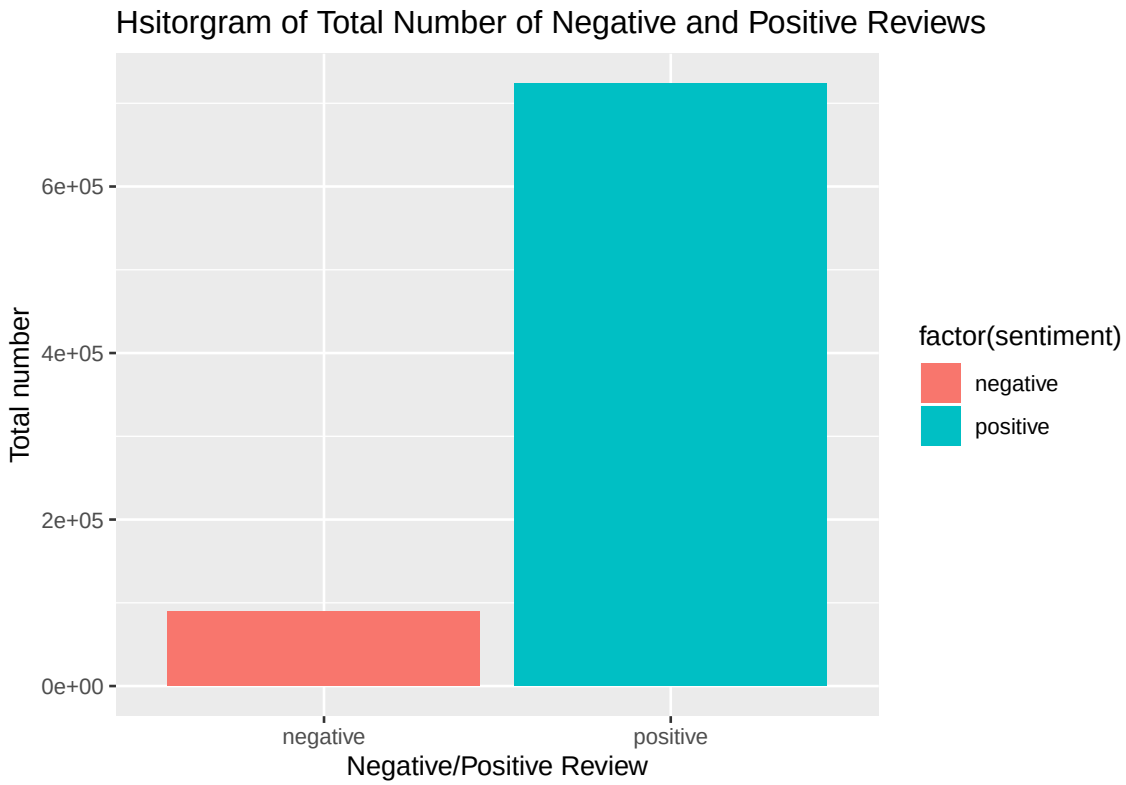
3) Whether strict plices affect the living experience and ratings?



There's no indications in this part, so that exclue this part in model buildings.

Other text Analysis of Reviews

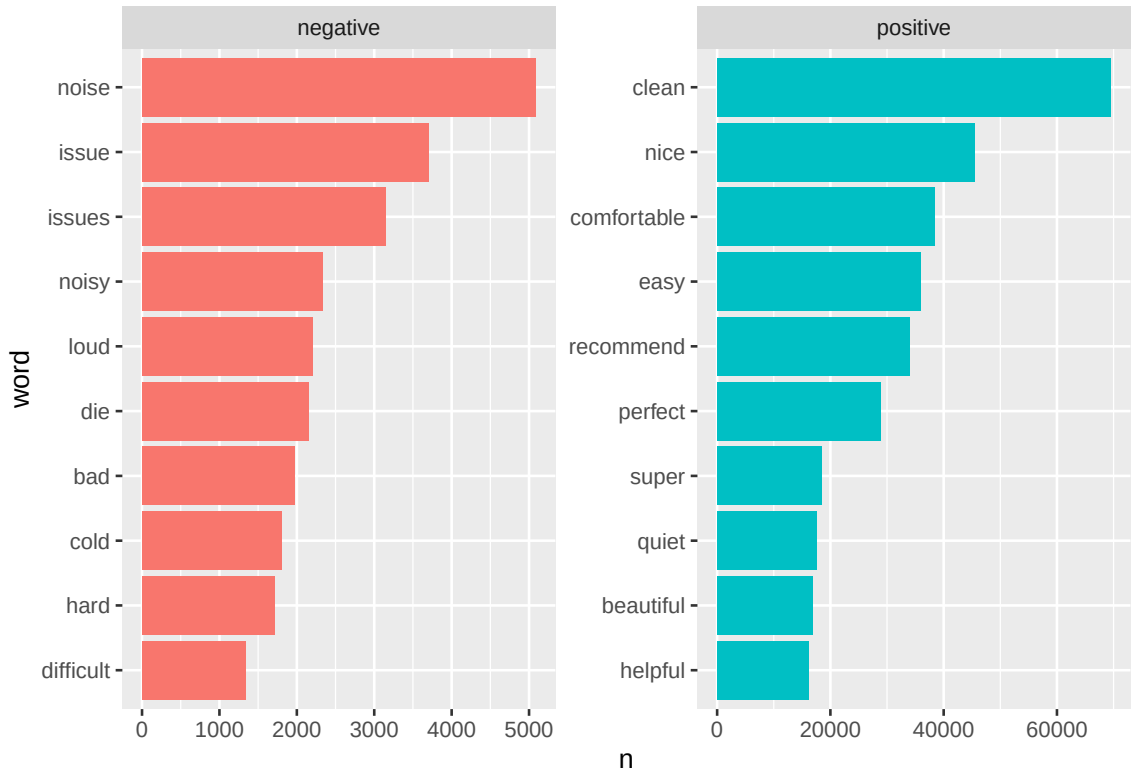
1) The total number of negative and positive reviews



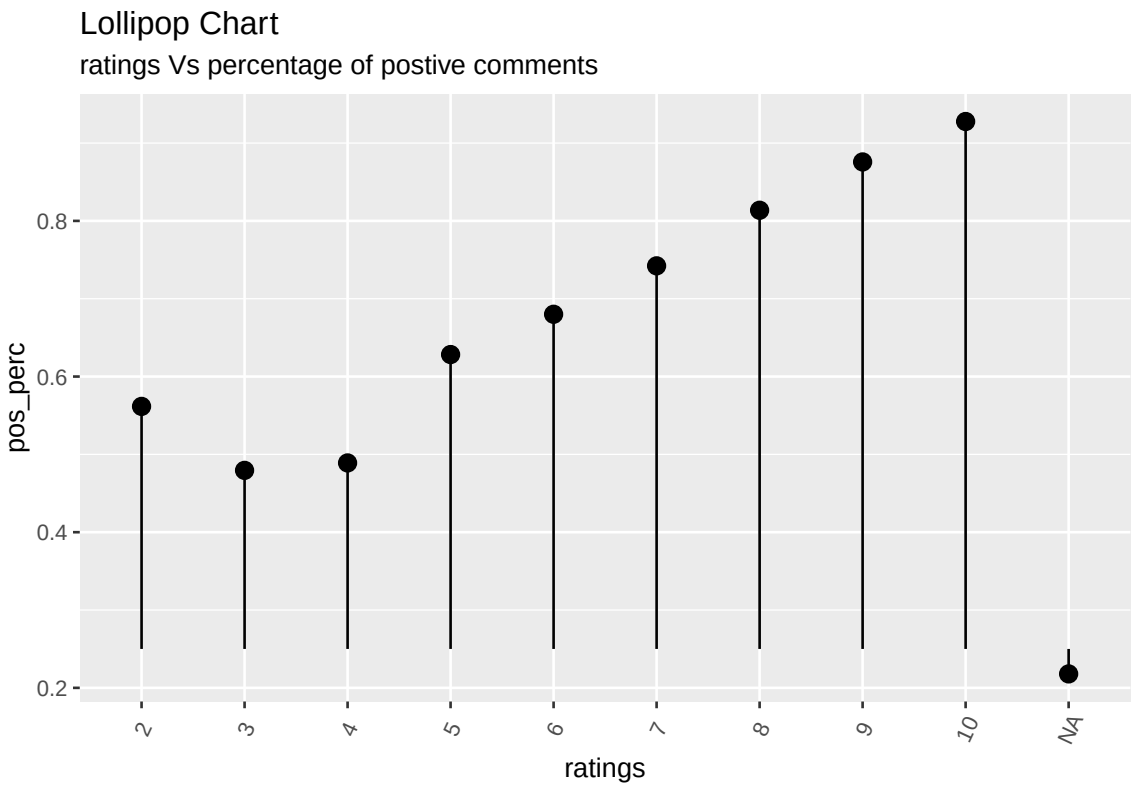
There are more positive feedbacks than negative feedbacks for the Airbnb in the Great Boston Area.

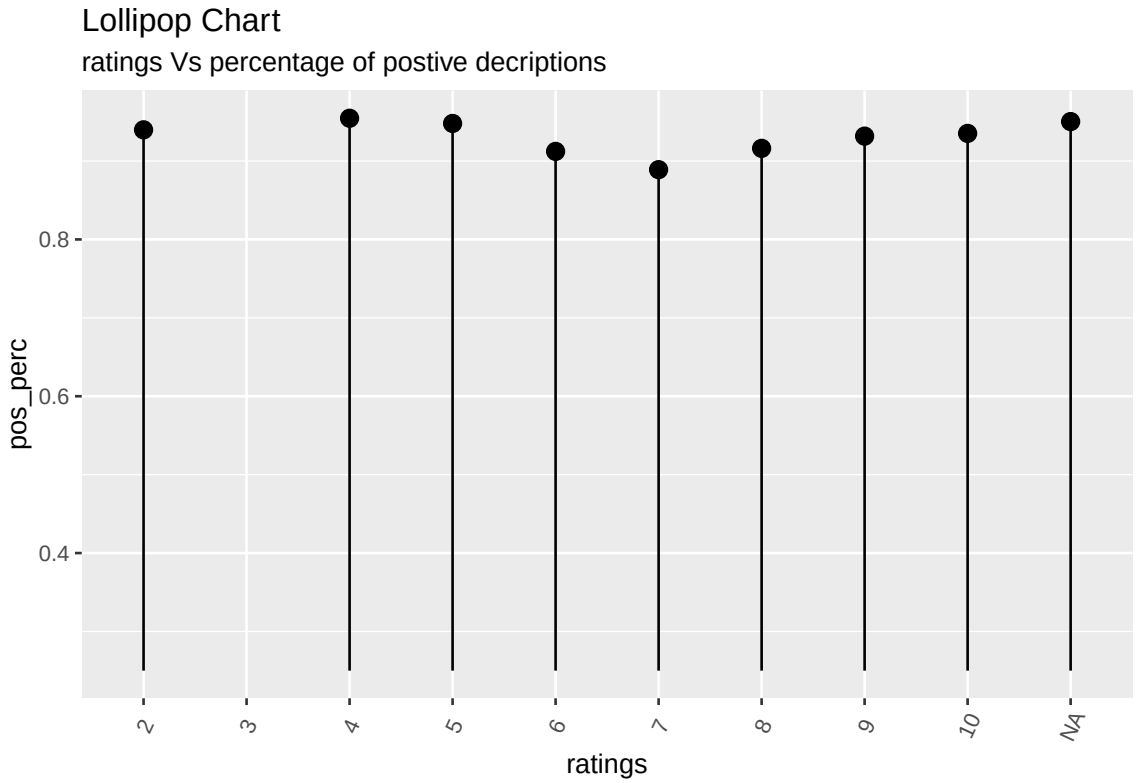
2) Wordclouds reviews


```
## Selecting by n
```



3) To see if the comments and descriptions are related to the ratings





The lollipop chart shows that higher scores tend to have higer positive percentage of comments. However, it seems that rating scores have no obvious relationship with the positive percentage of comments of house descriptions.

Take a look the models

1) Multinonial model.

```
##
## Re-fitting to get Hessian

## polr(formula = ratings ~ neighbourhood + price + cleaning_fee +
##     room_type + host_response_rate + Superhost + positve_perc +
##     is_location_exact, data = model.df)
##                                    coef.est coef.se
## neighbourhoodBack Bay               -0.32     0.23
## neighbourhoodBay Village            -0.34     0.42
## neighbourhoodBeacon Hill            -0.06     0.26
## neighbourhoodBrighton                0.02     0.24
## neighbourhoodCharlestown             0.33     0.33
## neighbourhoodChinatown              -0.05     0.30
## neighbourhoodDorchester              0.37     0.20
## neighbourhoodDowntown               -0.10     0.21
## neighbourhoodEast Boston             0.04     0.22
## neighbourhoodFenway                  0.25     0.23
## neighbourhoodHyde Park               0.13     0.40
## neighbourhoodJamaica Plain           0.62     0.24
## neighbourhoodLeather District       14.92     0.00
## neighbourhoodLongwood Medical Area   0.75     0.08
## neighbourhoodMattapan                0.62     0.37
```

```
## neighbourhoodMission Hill             0.18     0.32
## neighbourhoodNorth End                0.24     0.25
## neighbourhoodRoslindale               1.11     0.35
## neighbourhoodRoxbury                  0.04     0.22
## neighbourhoodSouth Boston             0.29     0.25
## neighbourhoodSouth Boston Waterfront  0.97     0.51
## neighbourhoodSouth End                0.14     0.23
## neighbourhoodWest End                -0.26     0.39
## neighbourhoodWest Roxbury             0.10     0.36
## price                                 0.00     0.00
## cleaning_fee                          0.00     0.00
## room_typeHotel room                   0.32     0.39
## room_typePrivate room                -0.23     0.11
## room_typeShared room                  0.19     0.40
## host_response_rate                    0.02     0.00
## SuperhostTRUE                         2.01     0.10
## positve_perc                         14.45     0.48
## is_location_exactTRUE                 0.15     0.10
## 1|2                                   3.25     0.05
## 2|3                                   6.58     0.67
## 3|4                                   7.12     0.63
## 4|5                                   8.49     0.60
## 5|6                                   9.25     0.59
## 6|7                                  11.36     0.61
## 7|8                                  12.46     0.62
## 8|9                                  14.27     0.63
## 9|10                                 16.53     0.65
## ---
## n = 3054, k = 42 (including 9 intercepts)
## residual deviance = 6220.5, null deviance is not computed by polr
```

2) Binomial Model.

```
## glm(formula = ratings_bin ~ neighbourhood + number_of_reviews +
##     price + cleaning_fee + room_type + host_response_rate + Superhost +
##     positve_perc + is_location_exact, family = binomial(link = "logit"),
##     data = model.df.bin)
##                                   coef.est coef.se
## (Intercept)                       -13.86     1.02
## neighbourhoodBack Bay              -0.13     0.31
## neighbourhoodBay Village           -0.07     0.69
## neighbourhoodBeacon Hill           -0.05     0.33
## neighbourhoodBrighton              -0.11     0.30
## neighbourhoodCharlestown            0.51     0.37
## neighbourhoodChinatown             -0.13     0.47
## neighbourhoodDorchester             0.37     0.26
## neighbourhoodDowntown              -0.55     0.31
## neighbourhoodEast Boston            0.11     0.28
## neighbourhoodFenway                 0.34     0.31
## neighbourhoodHyde Park              0.00     0.47
## neighbourhoodJamaica Plain          0.77     0.29
## neighbourhoodLeather District      12.42   324.74
## neighbourhoodLongwood Medical Area -0.12     1.98
## neighbourhoodMattapan               0.26     0.44
## neighbourhoodMission Hill           0.32     0.42
```

```
## neighbourhoodNorth End                 0.39     0.33
## neighbourhoodRoslindale                1.07     0.40
## neighbourhoodRoxbury                   0.08     0.29
## neighbourhoodSouth Boston              0.56     0.31
## neighbourhoodSouth Boston Waterfront   1.69     0.61
## neighbourhoodSouth End                -0.05     0.29
## neighbourhoodWest End                  0.07     0.45
## neighbourhoodWest Roxbury              0.28     0.43
## number_of_reviews                      0.00     0.00
## price                                  0.00     0.00
## cleaning_fee                           0.00     0.00
## room_typeHotel room                   -0.89     0.68
## room_typePrivate room                 -0.13     0.14
## room_typeShared room                   0.15     0.56
## host_response_rate                     0.01     0.01
## SuperhostTRUE                          2.07     0.11
## positve_perc                          12.83     0.84
## is_location_exactTRUE                  0.02     0.13
## ---
##   n = 3054, k = 35
##   residual deviance = 2677.3, null deviance = 4150.9 (difference = 1473.6)
```

The price and the clearning fee do not have much effect on the ratings.The positive comment percentage is highly correlated to positive feedbacks. The coeficient of other variables have changed when it becomes binary outcomes. In this model, the room type hotel indicate that it will decreases the chances that a costumer gives a good feedback, which corresponds to the EDA results. Also, if the host is a super host and response more, the chance of getting a good feedback increases 2 times. It seems that this model is more accurate than the multinomial model

3) Binomial regression with random effect.

```
## glmer(formula = ratings_bin ~ scale(number_of_reviews) + (1 |
##     neighbourhood) + scale(price) + scale(cleaning_fee) + room_type +
##     host_response_rate + Superhost + scale(positve_perc) + is_location_exact,
##     data = model.df.bin, family = binomial(link = "logit"))
##                           coef.est coef.se
## (Intercept)               -2.55     0.67
## scale(number_of_reviews) -0.18     0.05
## scale(price)               0.13     0.06
## scale(cleaning_fee)        0.07     0.07
## room_typeHotel room       -0.73     0.68
## room_typePrivate room     -0.11     0.13
## room_typeShared room       0.13     0.55
## host_response_rate         0.01     0.01
## SuperhostTRUE              2.08     0.11
## scale(positve_perc)        1.43     0.09
## is_location_exactTRUE      0.03     0.13
##
## Error terms:
##  Groups        Name        Std.Dev.
##  neighbourhood (Intercept) 0.28
##  Residual                  1.00
## ---
## number of obs: 3054, groups: neighbourhood, 25
## AIC = 2745.5, DIC = 2659.6
```

```
## deviance = 2690.6
```

When introduce the random effect, the model improves.First of all, $\sigma_\alpha^2 : \sigma_y^2 = 0.28^2 : 1^2 = 0.0784$, that means the pooling effect is strong, i.e.there are group difference between neighbourhoods. Besides, the deviance is slightly lower than the previous model.

4) Linear regression model.

```
## lm(formula = ratings_original ~ number_of_reviews + neighbourhood +
##     price + cleaning_fee + room_type + host_response_rate + Superhost +
##     positve_perc + is_location_exact, data = model.df)
##                                    coef.est coef.se
## (Intercept)                        50.35     1.41
## number_of_reviews                   0.00     0.00
## neighbourhoodBack Bay              -1.37     0.59
## neighbourhoodBay Village           -1.03     1.03
## neighbourhoodBeacon Hill           -0.11     0.65
## neighbourhoodBrighton              -0.15     0.59
## neighbourhoodCharlestown            0.42     0.72
## neighbourhoodChinatown             -0.63     0.79
## neighbourhoodDorchester             0.70     0.51
## neighbourhoodDowntown              -0.21     0.55
## neighbourhoodEast Boston            0.26     0.56
## neighbourhoodFenway                 0.26     0.60
## neighbourhoodHyde Park              0.27     0.97
## neighbourhoodJamaica Plain          0.25     0.56
## neighbourhoodLeather District       1.93     5.00
## neighbourhoodLongwood Medical Area  0.81     3.58
## neighbourhoodMattapan               1.14     0.91
## neighbourhoodMission Hill          -1.21     0.79
## neighbourhoodNorth End              0.36     0.63
## neighbourhoodRoslindale             1.55     0.77
## neighbourhoodRoxbury                0.32     0.56
## neighbourhoodSouth Boston           0.17     0.60
## neighbourhoodSouth Boston Waterfront -0.38    1.05
## neighbourhoodSouth End              0.04     0.57
## neighbourhoodWest End              -0.19     0.90
## neighbourhoodWest Roxbury           0.12     0.85
## price                               0.00     0.00
## cleaning_fee                        0.01     0.00
## room_typeHotel room                 1.58     1.06
## room_typePrivate room              -0.43     0.26
## room_typeShared room               -0.07     1.03
## host_response_rate                  0.07     0.01
## SuperhostTRUE                       2.64     0.22
## positve_perc                       39.04     0.95
## is_location_exactTRUE               0.51     0.24
## ---
## n = 3054, k = 35
## residual sd = 4.98, R-Squared = 0.51
```

The simple linear regression model also makes sense on its' coefficients.

5) Linear model with random effect.

```
## lmer(formula = ratings_original ~ (1 | neighbourhood) + number_of_reviews +
##     price + cleaning_fee + room_type + host_response_rate + Superhost +
```

```
##     positve_perc + is_location_exact, data = model.df)
##                         coef.est coef.se
## (Intercept)             49.86     1.35
## number_of_reviews        0.00     0.00
## price                    0.00     0.00
## cleaning_fee             0.01     0.00
## room_typeHotel room      1.43     1.05
## room_typePrivate room   -0.31     0.25
## room_typeShared room     0.08     1.01
## host_response_rate       0.08     0.01
## SuperhostTRUE            2.69     0.22
## positve_perc            39.51     0.93
## is_location_exactTRUE    0.52     0.24
##
## Error terms:
##  Groups        Name        Std.Dev.
##  neighbourhood (Intercept) 0.33
##  Residual                  4.98
## ---
## number of obs: 3054, groups: neighbourhood, 25
## AIC = 18535.2, DIC = 18424.3
## deviance = 18466.8
```

Still, the model including random effect has strong group effect, which means neighbourhood might be an important factor that affect the ratings and living experience. $\sigma_\alpha^2 : \sigma_y^2 = 0.33^2 : 4.98^2 = 0.004391058$;