MA684 Midterm Project: Airbnb Data Analysis

Shuyi Jiang

12/03/2017

Project overview and goal

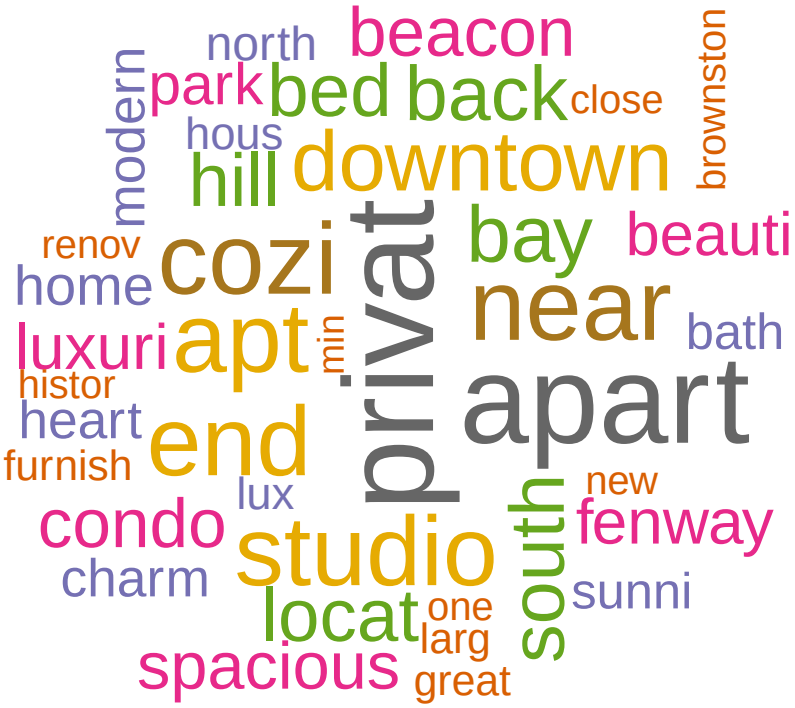
As a beginner who is new to airbnb, I really have no idea how to find an ideal place to stay through this platform. I looked up the beginner guides online and figured out that what the majority of potential houseguests care the most are price, location and host. How people choose airbnb should be an interesting angle for marketing. The project will focus on how to find a nice airbnb at a good price, in other words, to figure out the relationship bewteen price and factors that houseguests care the most.

Boston: data cleaning

I choosed Boston, the city I am the most familiar with, for this project. The original datasets are seperate monthly surveys. The data includes listing information in Boston from January to November in 2016. The sample size is 36359 (1 rows was removed because of missing data and irrelevant variables are excluded).

Word cloud

Looking at the word clouds of top 40 words about listing in name (searching key words), summary and description, I figured out that houseguests pay the most attention of price, location and room type.

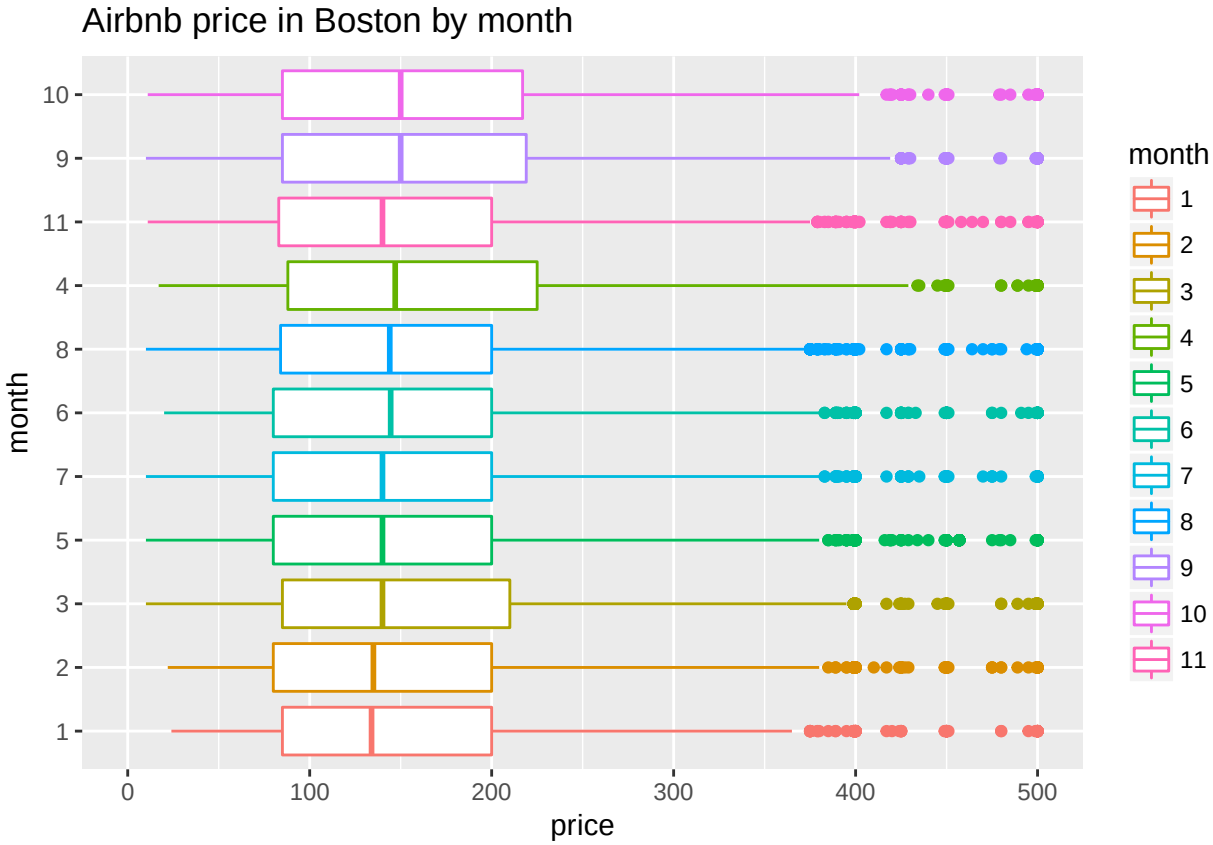


Airbnb daily price in Boston

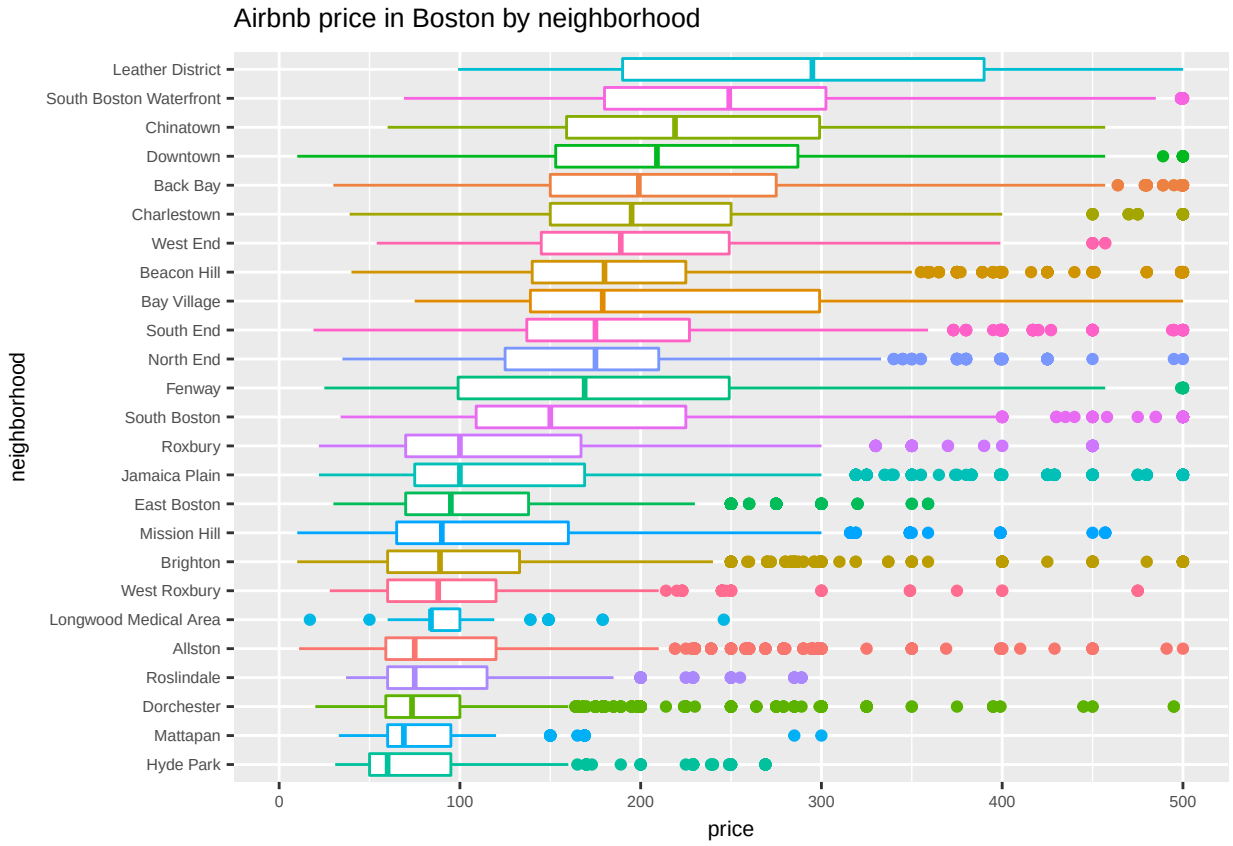
Based on what discussed about, I am interested in how location (neighborhood), room type and season affect price.

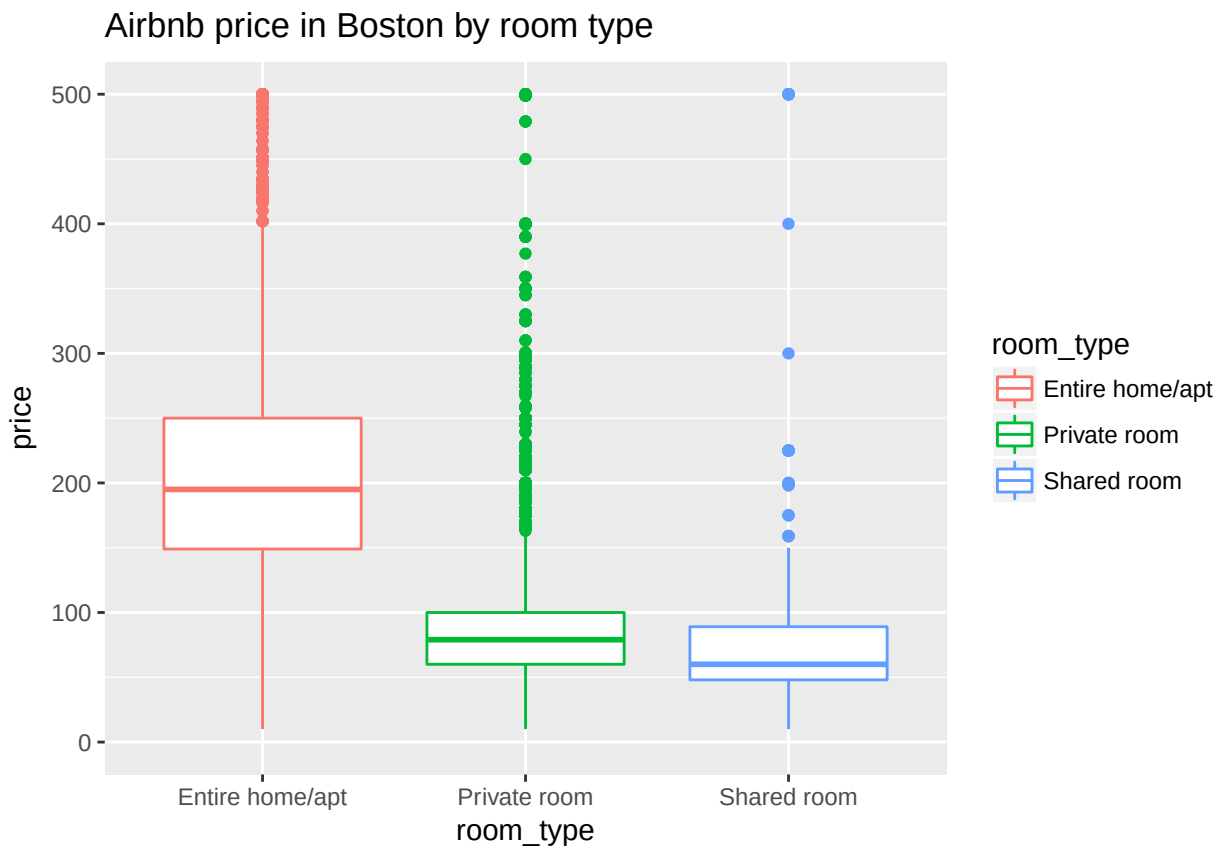
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0    85.0   147.0   173.5   215.0 10000.0
```

Only 769 out of 36359 samples have price over 500. Thus the following boxplots will only show samples with price under 500.



The boxplots by month indicates that season (month) has very limited influence on price. It seems to have some seasonal change, but fluctuation within 10 dollars is probably not what really matters at this point of study. I also looked into the seasonal price change of every single room listed. The result shows that the price of the same room is quite stable. Therefore, I decide to disregard seasonal affect of the airbnb price in the first stage of my study.


Airbnb price in Boston by neighborhood
neighborhood
Leather District
South Boston Waterfront
Chinatown
Downtown
Back Bay
Charlestown
West End
Beacon Hill
Bay Village
South End
North End
Fenway
South Boston
Roxbury
Jamaica Plain
East Boston
Mission Hill
Brighton
West Roxbury
Longwood Medical Area
Allston
Roslindale
Dorchester
Mattapan
Hyde Park
0
100
200
300
400
500
price


Airbnb price in Boston by room type
price
500
400
300
200
100
0
Entire home/apt
Private room
Shared room
room_type
room_type
Entire home/apt
Private room
Shared room

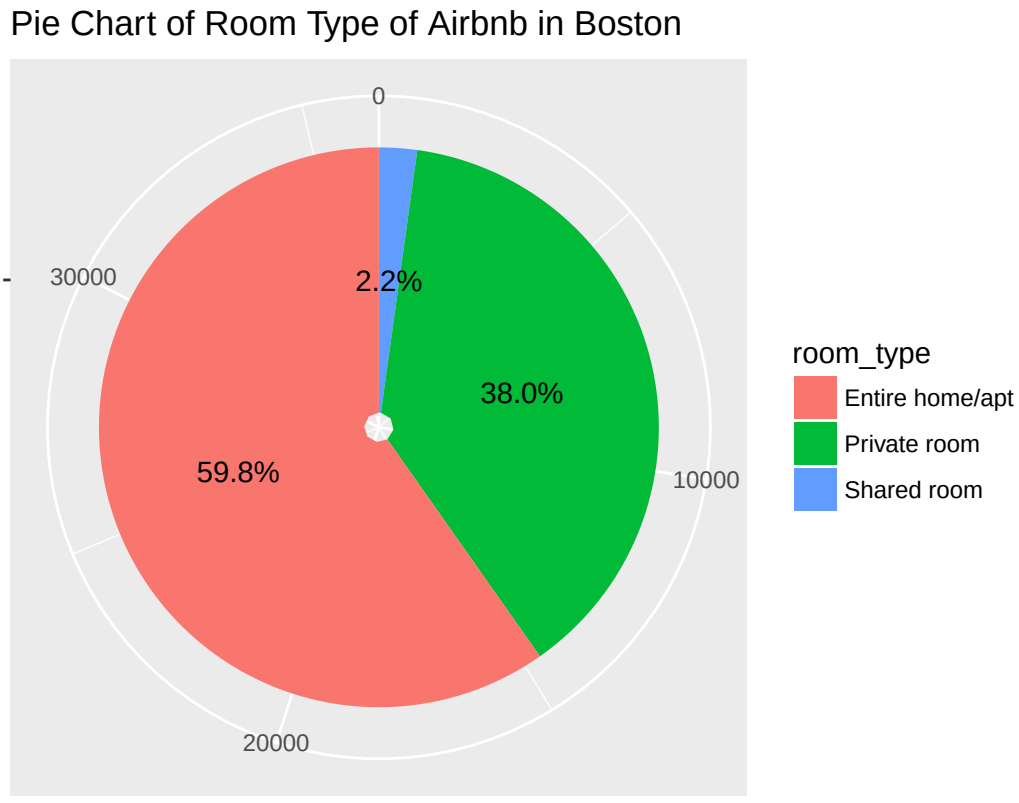
Compared to month, neighborhood and room type make relatively great influence on price. The boxplots indicates that there is very noticeable difference between airbnb prices in various neighborhood and those of distinct room types. I would like to dig deeper in these 2 indicators and get more detail about how neighborhood and room types affect airbnb price.

Room type of airbnb in Boston

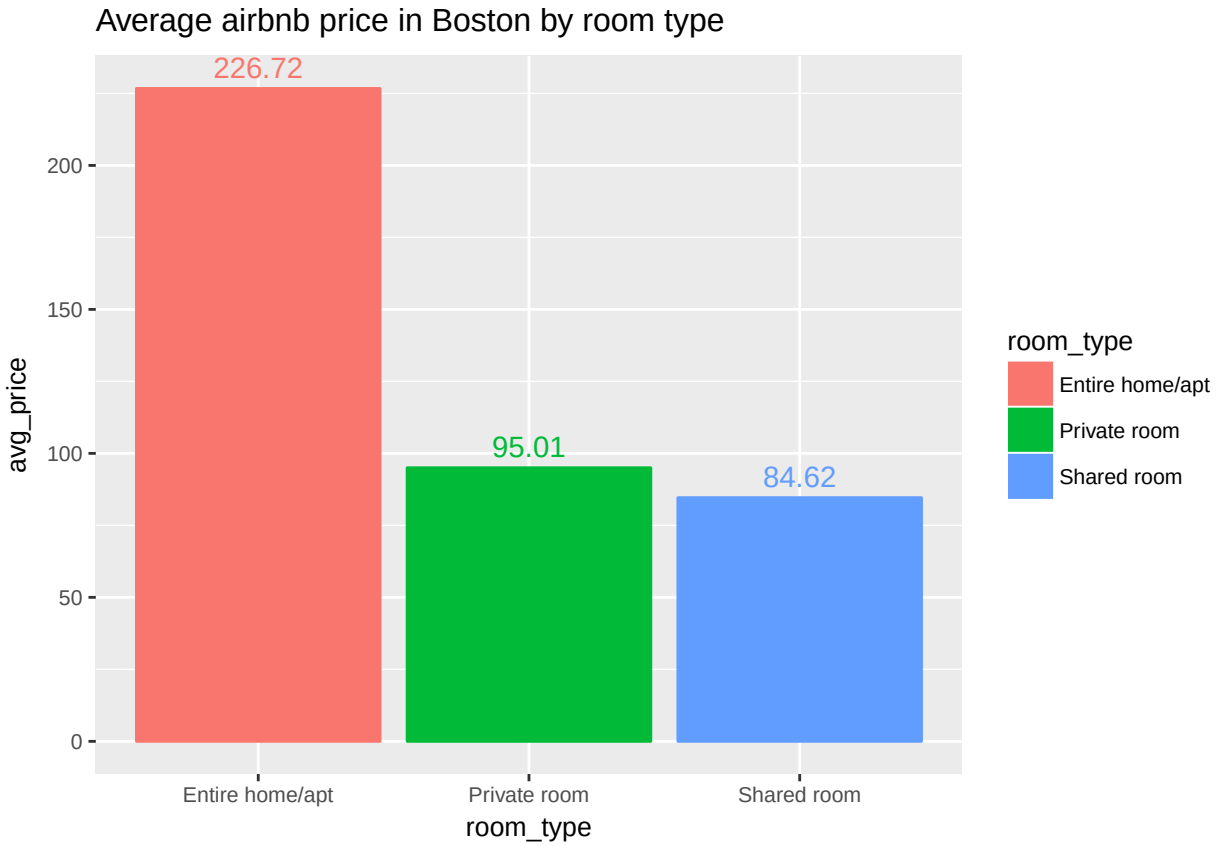
There are 3 room types: entire home/apt, private room and shared room. Entire home/apt is the most popular room type (more than half of 36359 listings) and the average price of this type of room is more than twice of the other 2. Share room is only about 2% of all the listing airbnb and it has the lowest average price.

room_type	freq
Entire home/apt	21727
Private room	13834
Shared room	798

room_type	avg_price
Entire home/apt	226.72154
Private room	95.00860
Shared room	84.61529



Entire home/apt is most popular room type. Almost 60% of houseguests choose entire home/apt. The pie chart also shows that people do not like shared room since only about 2% of them choose shared room.



Although the price of entire home/apt is more than twice of the other 2, people are much more likely to choose entire home/apt, which implies that maybe price is not the most important factor that people consider for airbnb.

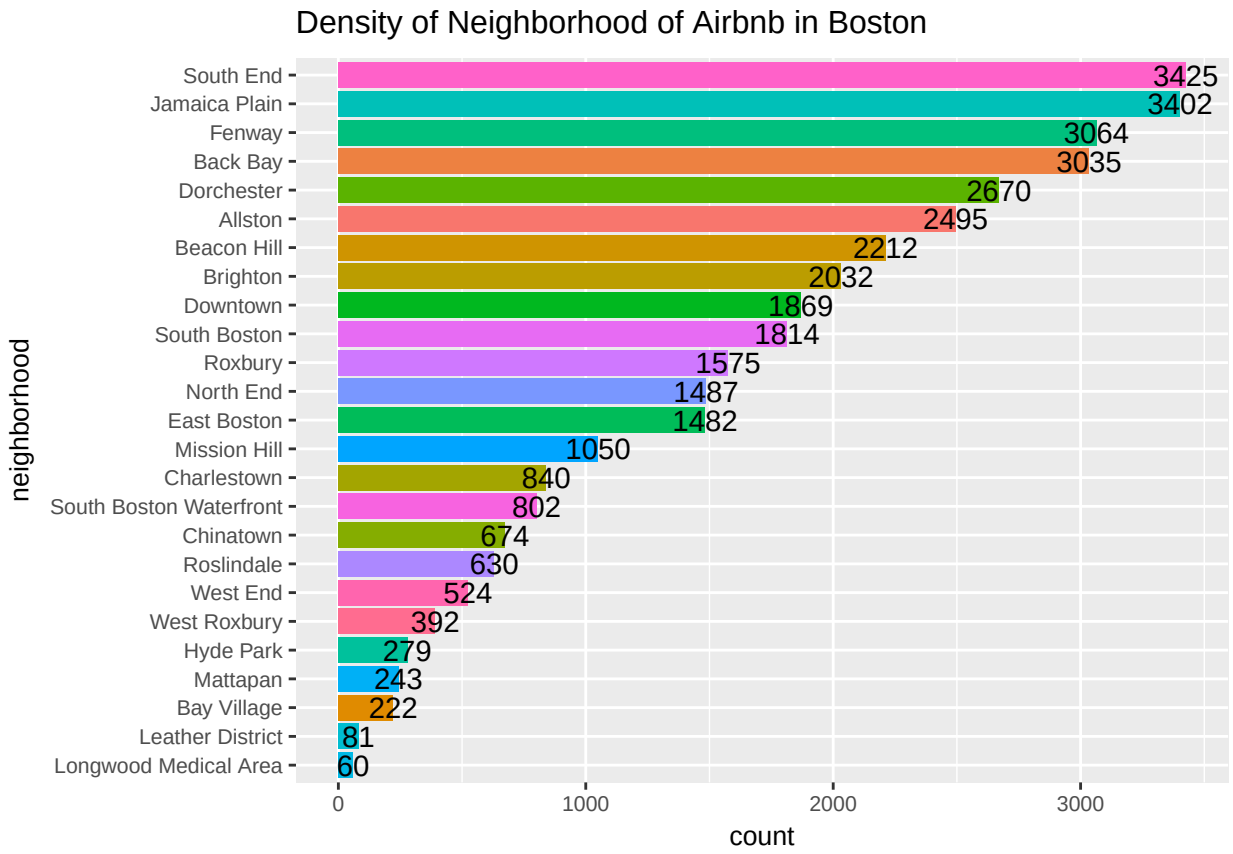
Neighborhood of airbnb in Boston

There are 25 different neighborhoods with airbnb listings in boston. The triangle area on the south riverbank of Charles River, including Allston, Fenway, Back Bay and South Boston is the most popular place for airbnb. This area does not have the lowest price but has relatively more convenient transportation.

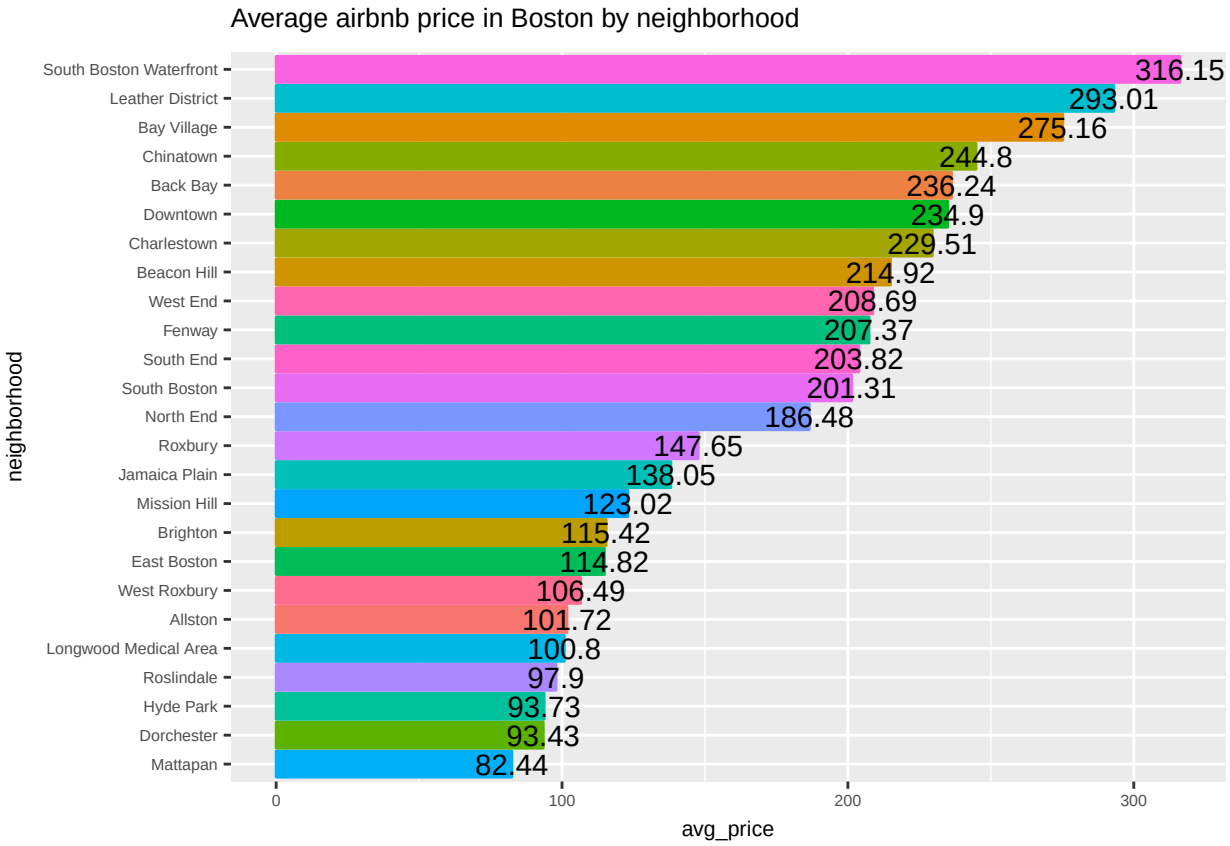
neighborhood	freq
Allston	2495
Back Bay	3035
Bay Village	222
Beacon Hill	2212
Brighton	2032
Charlestown	840
Chinatown	674
Dorchester	2670
Downtown	1869
East Boston	1482
Fenway	3064
Hyde Park	279
Jamaica Plain	3402
Leather District	81

neighborhood	freq
Longwood Medical Area	60
Mattapan	243
Mission Hill	1050
North End	1487
Roslindale	630
Roxbury	1575
South Boston	1814
South Boston Waterfront	802
South End	3425
West End	524
West Roxbury	392

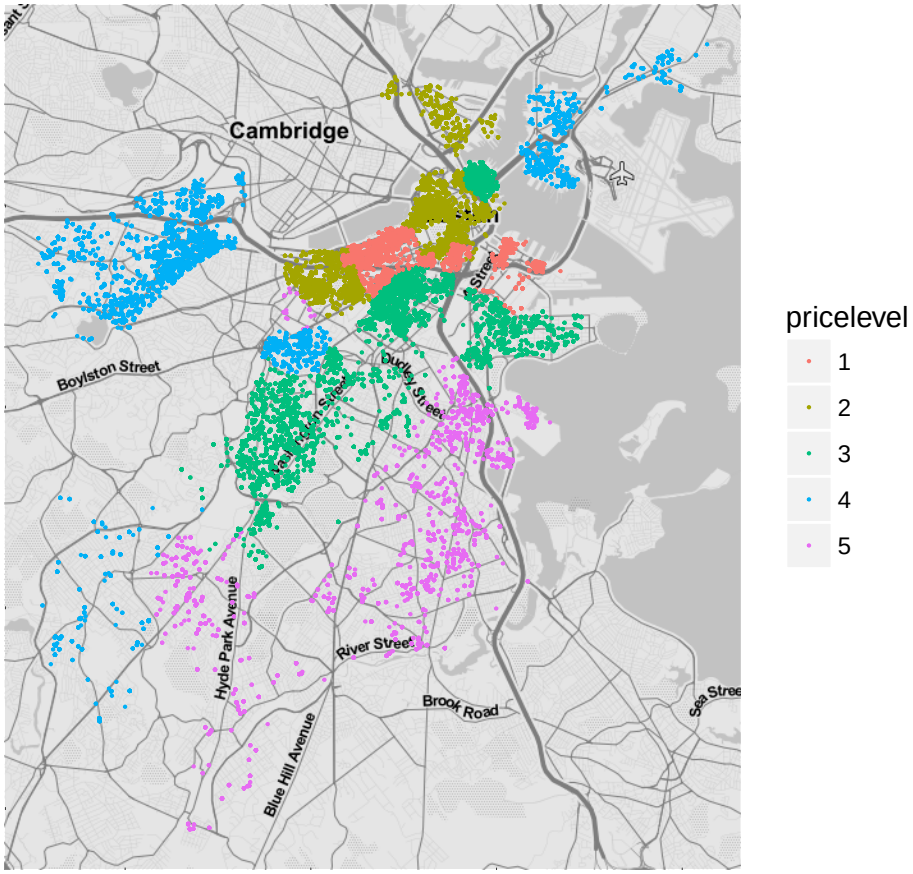
neighborhood	avg_price
Allston	101.71543
Back Bay	236.24217
Bay Village	275.16216
Beacon Hill	214.91953
Brighton	115.42470
Charlestown	229.51071
Chinatown	244.79970
Dorchester	93.43408
Downtown	234.89781
East Boston	114.81984
Fenway	207.37304
Hyde Park	93.73118
Jamaica Plain	138.05409
Leather District	293.01235
Longwood Medical Area	100.80000
Mattapan	82.44444
Mission Hill	123.02095
North End	186.47613
Roslindale	97.89683
Roxbury	147.64508
South Boston	201.30706
South Boston Waterfront	316.14713
South End	203.82161
West End	208.68893
West Roxbury	106.49490



The barplot shows that houseguests do have preference of location. They are more likely to stay in the center of Boston.



The 4 most expensive neighborhood are all not popular among houseguests. However, relative high price (the 5th most expensive) does not hider Back Bay to be one of the most favorite (rank 4 in popularity).

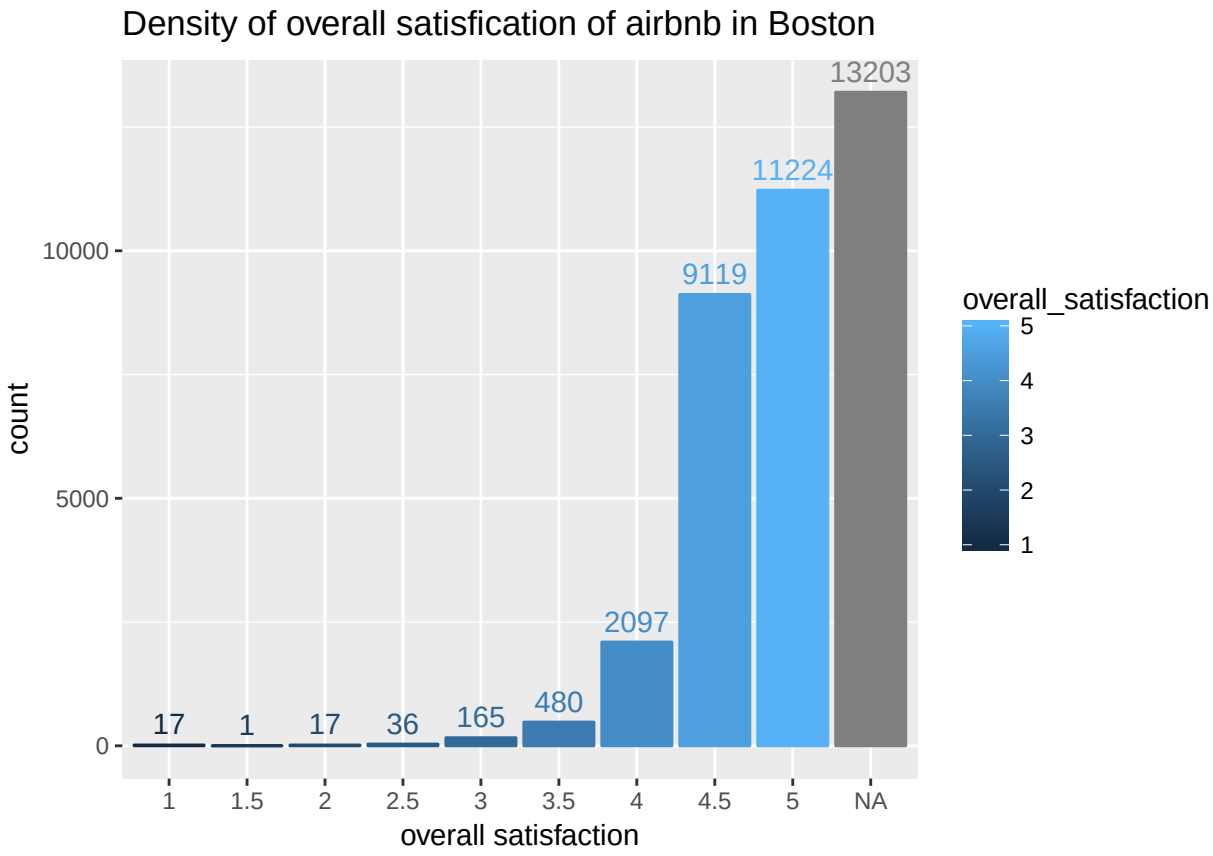


The prices of different neighborhood are divided into 5 levels by 20 percentile (level 1 is the most expensive and level 5 is the cheapest). The center of Boston (south bank of the Charles River) is the most expensive, but it is the favorite location for houseguests (the color block is more condensed).

Overall satisfication of airbnb in Boston

The listed airbnb in this data set is rated in a scale of 1 to 5 with 0.5 interval. The majority (9119 + 11224 = 20343) of 36359 samples is rated as 4.5 or 5 and 13203 samples, which is about 36%, are missing rating data. Thus, the overall satisfaction will be DISREGARDED in the analysis because it is obviously NOT a representitive indicator.

overall_satisfaction	n
1.0	17
1.5	1
2.0	17
2.5	36
3.0	165
3.5	480
4.0	2097
4.5	9119
5.0	11224
NA	13203



Model Testing

Multilevel linear model was choosen after several trials of different models. While fitting the multilevel, samples have NA in reviews, accommodates and bedrooms are removed.

Model fit1, fit2, fit3 and fit4 are varing-intercept models with predictors. They are using different combination of 3 variables (reviews, accommodates, bedrooms) and room type to see whether it is better to include any of the 4 variables in the model. The models also include neighborhood and host_id as 2 multilevel variables (one host could have more than 1 room listed and I am using host id as an indicator of service quality and the appearence of host_id decreases AIC by over 20000). It turns out that fit1 with all 4 variables is the best among these 4 models.

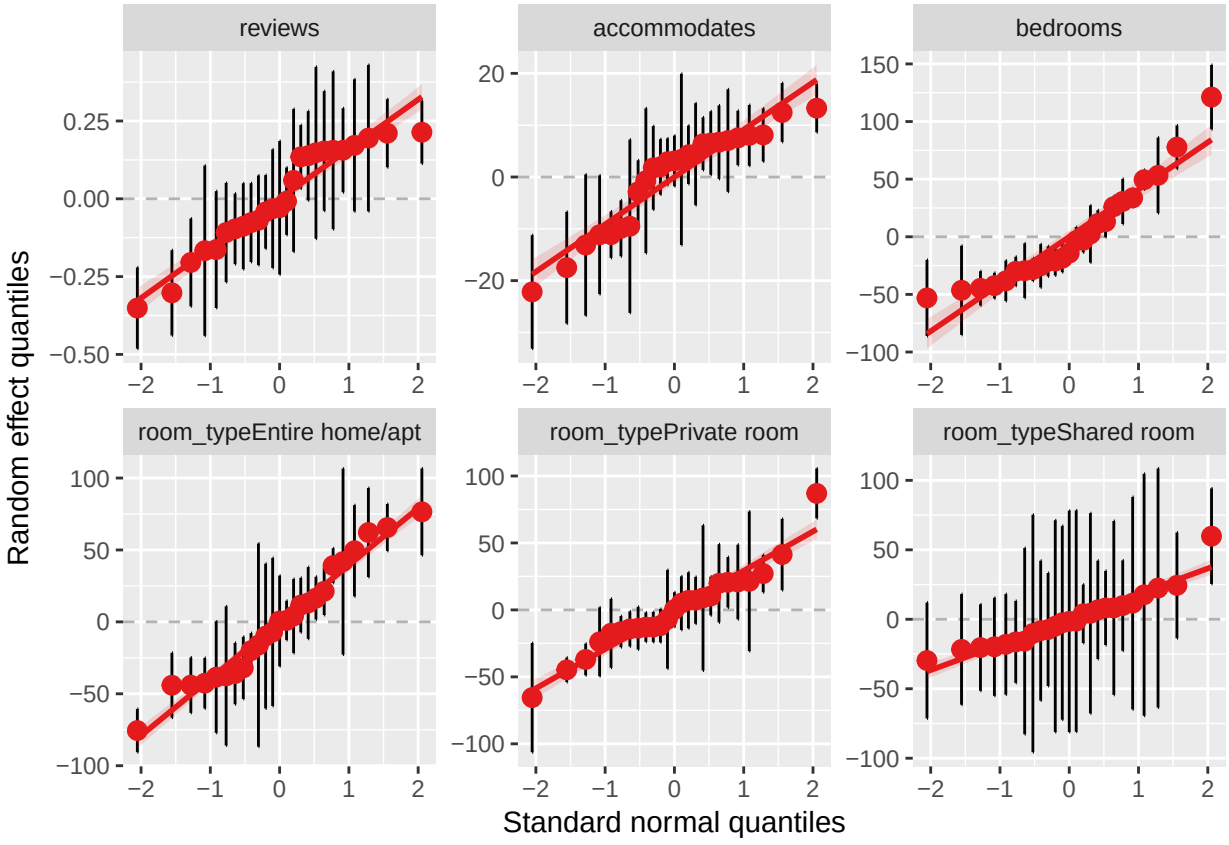
	fixef(fit1)
reviews	-0.2024038
accommodates	11.1748611
bedrooms	42.8091934
room_typeEntire home/apt	103.2439071
room_typePrivate room	50.9499199
room_typeShared room	42.6401771

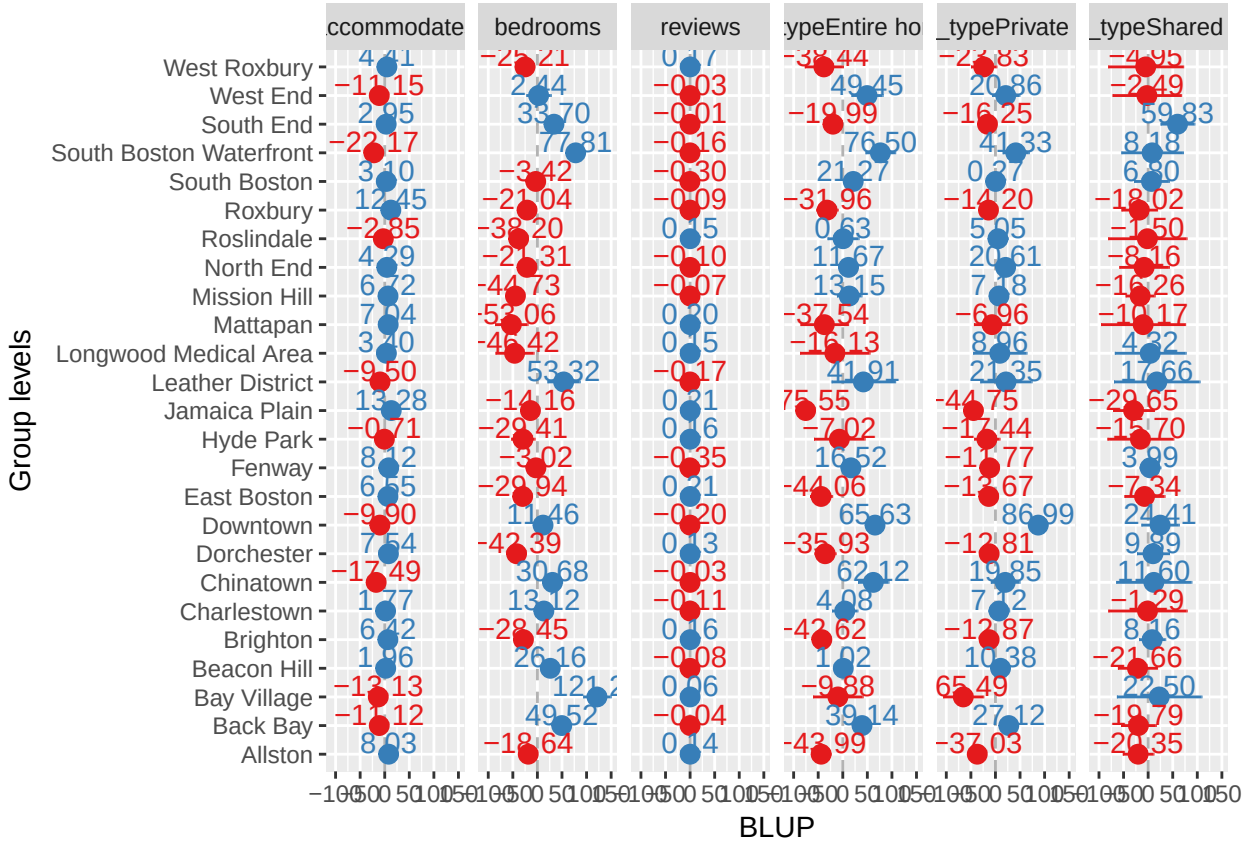
```
## Data: boston.model
## Models:
## fit2: price ~ reviews + accommodates + room_type + (1 | neighborhood) +
## fit2:     (1 | host_id)
```

```
## fit3: price ~ reviews + (bedrooms - 1) + room_type + (1 | neighborhood) +
## fit3:     (1 | host_id)
## fit4: price ~ accommodates + (bedrooms - 1) + room_type + (1 | neighborhood) +
## fit4:     (1 | host_id)
## fit1: price ~ reviews + accommodates + (bedrooms - 1) + room_type +
## fit1:     (1 | neighborhood) + (1 | host_id)
##      Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
## fit2  8 372940 373007 -186462   372924
## fit3  8 372434 372501 -186209   372418 506.370      0  < 2.2e-16 ***
## fit4  8 372291 372358 -186138   372275 142.613      0  < 2.2e-16 ***
## fit1  9 372239 372314 -186110   372221  54.807      1   1.33e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

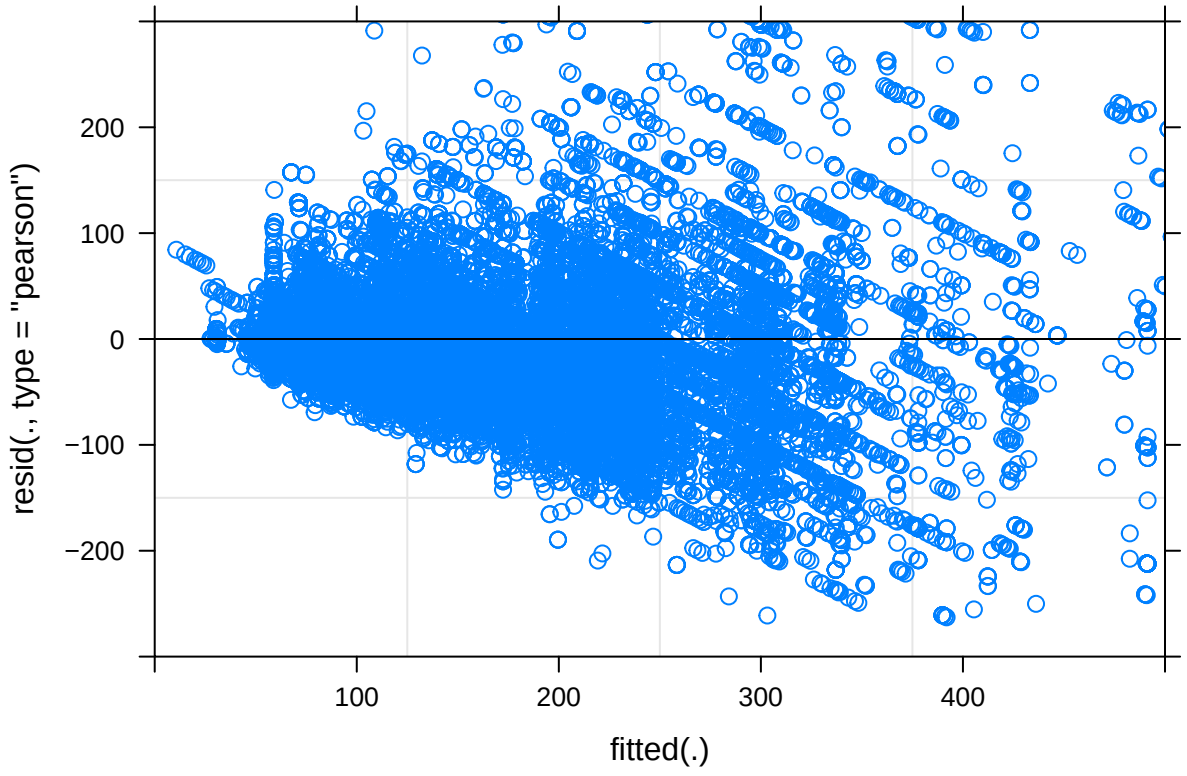
Different from fit1, fit5 has mixed random effect and varying both slope and intercept. Since host_id has more than 1000 levels and it is messy to show the mixed effect of both neighborhood and host_id, fit5 only includes mixed random effect of neighborhood to illustrate how mixed effect works. The final model will also includes host_id as a multilevel random variable.

	fixef(fit5)
reviews	-0.2482487
accommodates	8.1912471
bedrooms	49.6148049
room_typeEntire home/apt	109.0935892
room_typePrivate room	46.8227468
room_typeShared room	21.9454694





The model has q multilevel predictor: neighborhood. As shown in the graphs above, the influence of variables varies by different levels of the predictor. For example, bedrooms increases the price more with the same change in neighborhood with higher price, such as Back bay.



Model

The final model choosen is a multilevel model varying both intercepts and slopes.

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

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: price ~ reviews + accommodates + (bedrooms - 1) + room_type +
##     (1 + reviews + accommodates + (bedrooms - 1) + room_type |
##         neighborhood) + (1 | host_id)
##    Data: boston.model
##
##       AIC       BIC    logLik  deviance  df.resid
##  371705.7  371948.2 -185823.8  371647.7     31618
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -17.947  -0.136  -0.014   0.102 114.771
##
## Random effects:
##  Groups       Name                       Variance  Std.Dev. Corr
##  host_id      (Intercept)                1.443e+04 120.1193
##  neighborhood reviews                    8.432e-02   0.2904
##               accommodates               5.347e+01   7.3126  0.06
##               bedrooms                   1.060e+03  32.5580 -0.20 -0.67
##               room_typeEntire home/apt   1.517e+03  38.9453 -0.33 -0.72
##               room_typePrivate room      6.095e+02  24.6882 -0.23 -0.78
##               room_typeShared room       2.454e+03  49.5414 -0.70 -0.14
##  Residual                                5.342e+03  73.0868
##
##
##
##
##
##   0.07
##   0.22  0.91
##  -0.01  0.48  0.48
##
## Number of obs: 31647, groups:  host_id, 3772; neighborhood, 25
##
## Fixed effects:
##                             Estimate Std. Error t value
## reviews                     -0.18642    0.06967  -2.676
## accommodates                 8.73073    1.73712   5.026
## bedrooms                    39.13384    6.94667   5.633
## room_typeEntire home/apt   118.74018    8.84568  13.424
## room_typePrivate room       56.35124    5.91069   9.534
## room_typeShared room        37.22649   12.68932   2.934
##
## Correlation of Fixed Effects:
##             reviws accmmd bedrms rm_Eh/ rm_tPr
## accommodats -0.010
## bedrooms    -0.129 -0.642
## rm_typEnhm/ -0.242 -0.647  0.025
```

```
## rm_typPrvtr -0.139 -0.631  0.148  0.846
## rm_typShrdr -0.513 -0.067 -0.052  0.406  0.421
## convergence code: 1
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 2 negative eigenvalues
```

	fixef(fit.final)
reviews	-0.1864155
accommodates	8.7307329
bedrooms	39.1338436
room_typeEntire home/apt	118.7401841
room_typePrivate room	56.3512351
room_typeShared room	37.2264882

The "reviews" is scaled by subtracting the mean and "bedrooms" is shifted to the left by 1. As showen below, the coefficients of fitted model are quite significant, especially the number of rooms and room type. Reviews (the number of total reviews of a specific room, which is used to predict the popularity of the this room) has much less influence than the other 3 variables. In general, with everything else remains the same, 1 increase in accommodates will lead to 8.73 higher daily price. One more room in airbnb will increase the daily price by 39.13. On average, an entire home/apt will be 62.39 dollars more expensive than a private room and 81.51 than a shared room.

Discussion and concerns

The original data has lots of limitation, such as missing guest_id, which is necessary to do analysis of houseguests' review of rooms they have stayed in. Although the final model is better than the other testing models, it has huge AIC. Digging deeper with help of data other than the original data is definitely what I need to do after this first step of research. More variables that will make the model better has to be determined. More importantly, I am insterested in the factors, such as transportance, convinience, sightseeing, food and shopping center, that make the neighborhood popular. That is to say, I will need more data, not limited to just airbnb data, to build up a scale or maybe scoring system to measure the popularity of a specific location.