MA678 Midterm Project –

Yunyi Zhang

27/11/2019

1: Introduction

All datasets are downloaded from http://tomslee.net/airbnb-data-collection-get-the-data, datasets are seperated by month and location. Each dataset contains information such as room_id, room_type, city, neighborhood, accommodates, number of bedrooms. In this project I am going to first inspect each important variable, and then do the regression on different variables against the price.

2: Load and merge the data

After merging all datasets, there are in total 7112 rows with 19 columns

3. Quick summary of my dataset

These quick summary shows that there are 13 different neighborhoods, 3 different room types, 15 different accommodates and 6 number of bedroom. I also included a detailed summary below.

```
length(unique(CAMB_2017_7$neighborhood))
```

```
## [1] 13
```

```
length(unique(CAMB_2017_7$room_type))
```

```
## [1] 3
```

```
length(unique(CAMB_2017_7$accommodates))
```

```
## [1] 15
```

```
length(unique(CAMB_2017_7$bedrooms))
```

```
## [1] 6
```

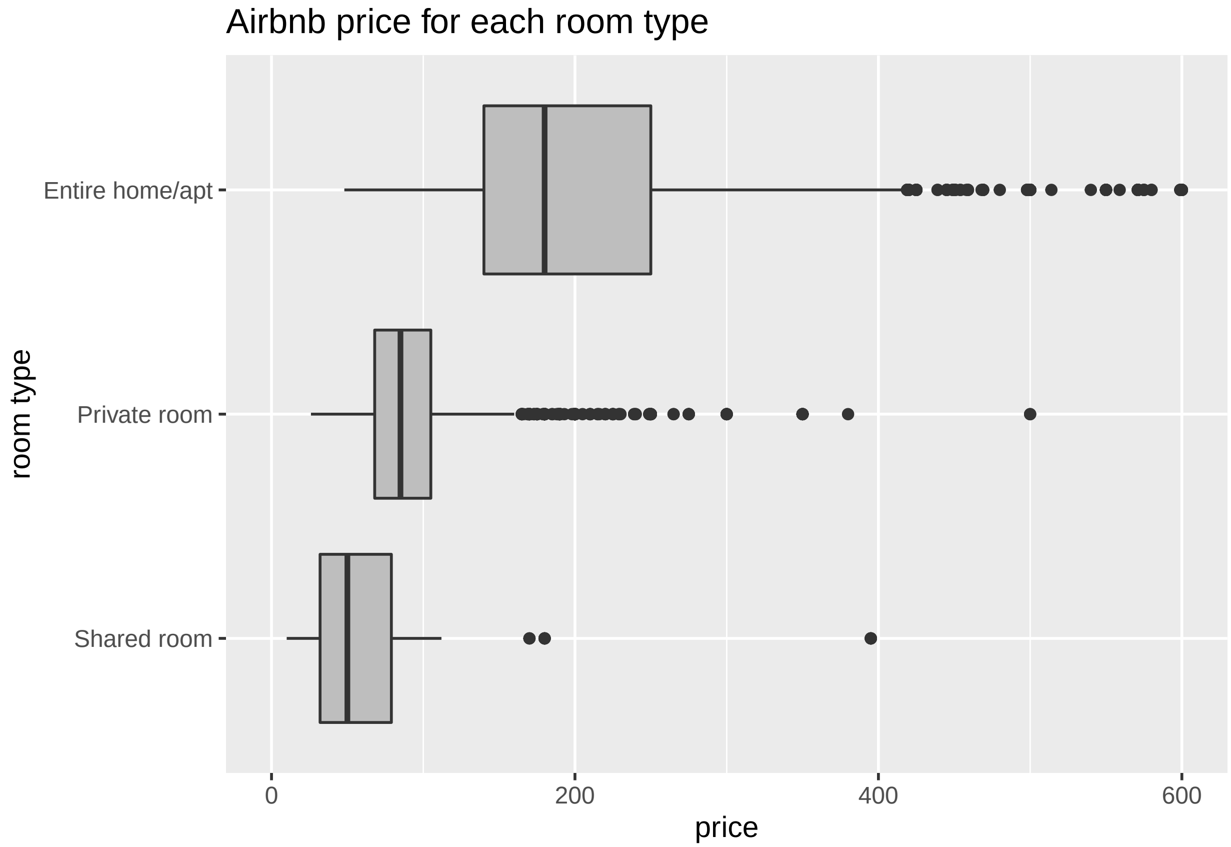
```
summary(CAMB_main)
```

```
##     room_id           survey_id       host_id
##  Min.   :    8521   Min.   :1101   Min.   :     1312
##  1st Qu.: 6466854   1st Qu.:1260   1st Qu.:  6184550
##  Median :12381809   Median :1385   Median : 21655476
##  Mean   :11091413   Mean   :1315   Mean   : 32092021
##  3rd Qu.:15816126   3rd Qu.:1502   3rd Qu.: 48159721
##  Max.   :19930909   Max.   :1502   Max.   :140659642
##
##            room_type    country                  city      borough
##  Entire home/apt:3685   Mode:logical   Cambridge MA:7112   Mode:logical
##  Private room   :3338   NA's:7112                          NA's:7112
##  Shared room    :  89
##
##
##
```

```
##
##                 neighborhood     reviews       overall_satisfaction
##  Mid-Cambridge        :1312   Min.   :  0.00   Min.   :0.000
##  Cambridgeport        :1057   1st Qu.:  1.00   1st Qu.:0.000
##  East Cambridge       : 934   Median :  6.00   Median :4.500
##  Riverside            : 760   Mean   : 23.96   Mean   :3.371
##  Area Four            : 643   3rd Qu.: 27.00   3rd Qu.:5.000
##  Wellington-Harrington: 527   Max.   :512.00   Max.   :5.000
##  (Other)              :1879                    NA's   :599
##   accommodates      bedrooms    bathrooms         price
##  Min.   : 1.00   Min.   :0.00   Mode:logical   Min.   :  10.0
##  1st Qu.: 2.00   1st Qu.:1.00   NA's:7112      1st Qu.:  85.0
##  Median : 2.00   Median :1.00                  Median : 125.0
##  Mean   : 3.02   Mean   :1.33                  Mean   : 156.7
##  3rd Qu.: 4.00   3rd Qu.:2.00                  3rd Qu.: 195.0
##  Max.   :16.00   Max.   :5.00                  Max.   :1290.0
##
##  minstay                           last_modified     latitude
##  Mode:logical   2017-04-12 12:56:41.621488:   1   Min.   :42.35
##  NA's:7112      2017-04-12 12:56:41.624354:   1   1st Qu.:42.37
##                 2017-04-12 12:56:41.626883:   1   Median :42.37
##                 2017-04-12 12:56:41.629982:   1   Mean   :42.37
##                 2017-04-12 12:56:41.635646:   1   3rd Qu.:42.38
##                 2017-04-12 12:56:41.640862:   1   Max.   :42.40
##                 (Other)                   :7106
##    longitude
##  Min.   :-71.16
##  1st Qu.:-71.12
##  Median :-71.11
##  Mean   :-71.11
##  3rd Qu.:-71.10
##  Max.   :-71.07
##
##                                                 location
##  0101000020E61000000000000000000C751C037A8FDD64E304540:   4
##  0101000020E61000000057B26323C751C0F6EFFACC592F4540:   4
##  0101000020E6100000000150C58D5BC651C0C8B60C384B2F4540:   4
##  0101000020E6100000000168942EFDC651C0B8E9CF7EA42E4540:   4
##  0101000020E61000000001C11C3D7EC651C0BA2D910BCE2E4540:   4
##  0101000020E61000000001FC53AA44C651C0068200193A2E4540:   4
##  (Other)                                           :7088
```

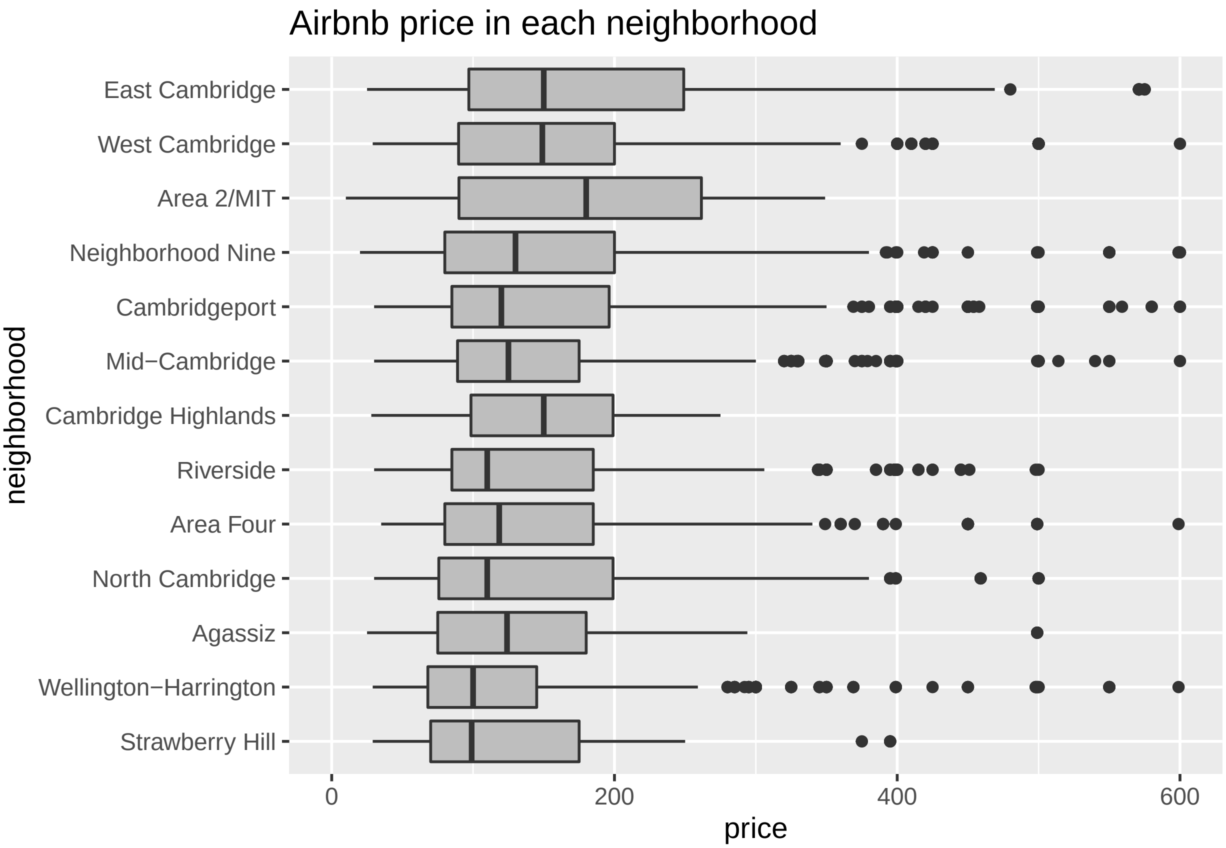
4. Boxplots for different variables vs rating

4.1 Room Type and Price



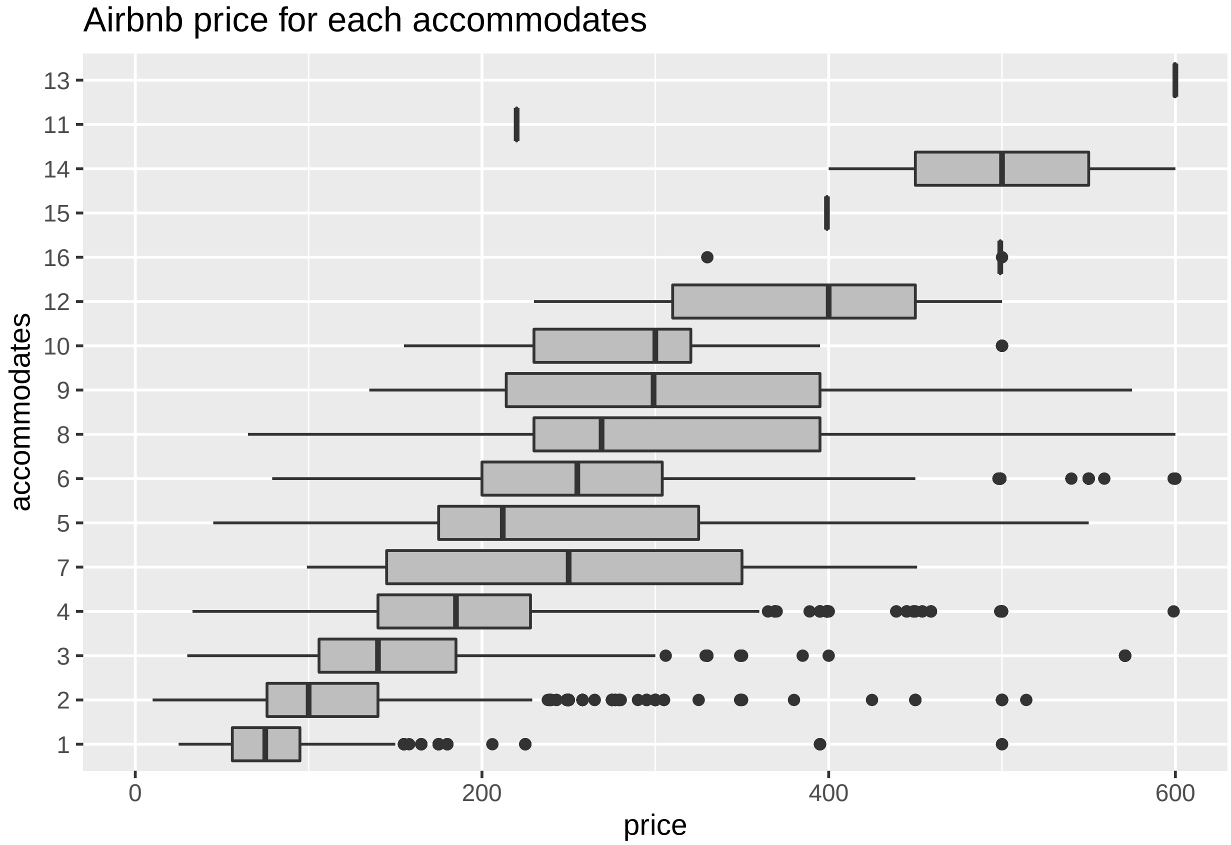
room type	average price
Entire home/apt	216.98
Private room	92.30
Shared room	75.98

4.2 Neighorhood and Price



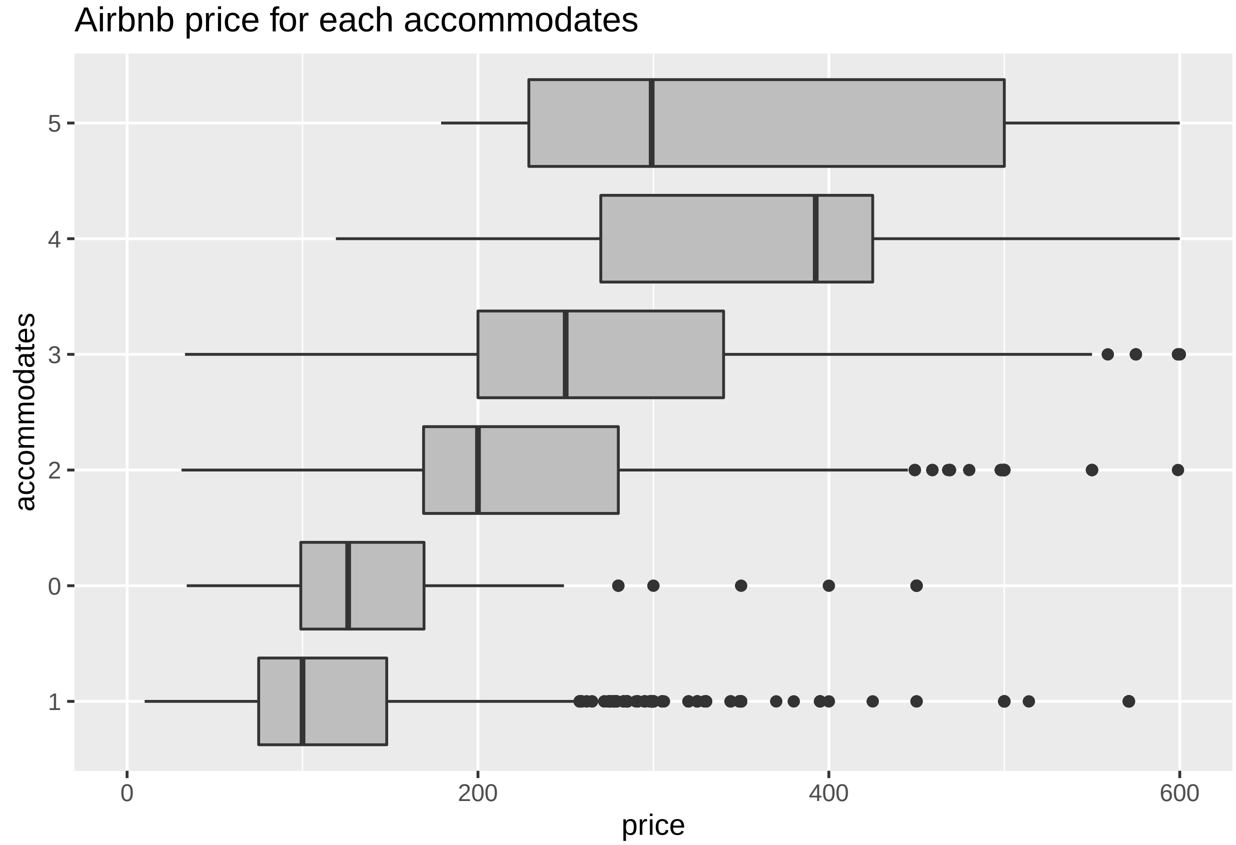
Neighborhood	average price
Agassiz	143.15
Area 2/MIT	172.16
Area Four	146.28
Cambridge Highlands	150.49
Cambridgeport	159.56
East Cambridge	191.00
Mid-Cambridge	151.24
Neighborhood Nine	169.90
North Cambridge	144.85
Riverside	146.50
Strawberry Hill	127.29
Wellington-Harrington	129.05
West Cambridge	176.91

4.3 Accommodate and Price



Accommodates	average price
1	79.90
2	114.08
3	162.51
4	195.88
5	255.95
6	282.82
7	251.33
8	341.92
9	356.82
10	386.53
11	584.00
12	436.15
13	599.67
14	531.25
15	469.25
16	465.40

4.4 Bedrooms and Price



Bedrooms	average price
0	138.20
1	116.17
2	234.23
3	299.10
4	397.35
5	477.68

5 Multilevel Model analysis

5.1 Regress price on Neighborhood and Bedrooms, no between group

```
## lmer(formula = price ~ neighborhood + bedrooms + (1 | room_type) +
##     (1 | accommodates), data = CAMB_main, REML = FALSE)
##                                 coef.est coef.se
## (Intercept)                     175.18    33.61
## neighborhoodArea 2/MIT           29.46    13.12
## neighborhoodArea Four             4.17     5.63
## neighborhoodCambridge Highlands -25.86    13.22
## neighborhoodCambridgeport         9.13     5.24
## neighborhoodEast Cambridge       35.02     5.33
## neighborhoodMid-Cambridge        10.53     5.10
## neighborhoodNeighborhood Nine    10.63     5.81
## neighborhoodNorth Cambridge        -4.87     5.87
## neighborhoodRiverside               5.85     5.45
## neighborhoodStrawberry Hill       -26.67     9.17
## neighborhoodWellington-Harrington  -9.86     5.83
## neighborhoodWest Cambridge         -7.11     6.37
## bedrooms                           40.07     2.01
##
## Error terms:
##  Groups       Name        Std.Dev.
##  accommodates (Intercept) 93.16
##  room_type    (Intercept) 38.59
##  Residual                 81.15
## ---
## number of obs: 7112, groups: accommodates, 16; room_type, 3
## AIC = 82834.4, DIC = 82800.4
## deviance = 82800.4
```

5.2 Regress price on Neighborhood, Bedrooms, Room Types, with one between-group of bedrooms and accommodates.

```
## lmer(formula = price ~ neighborhood + bedrooms + room_type +
##     (1 + bedrooms | accommodates), data = CAMB_main, REML = FALSE)
##                                   coef.est coef.se
## (Intercept)                       136.22     5.48
## neighborhoodArea 2/MIT             27.99    13.03
## neighborhoodArea Four               5.79     5.59
## neighborhoodCambridge Highlands   -26.37    13.14
## neighborhoodCambridgeport           9.34     5.21
## neighborhoodEast Cambridge         35.91     5.29
## neighborhoodMid-Cambridge          11.66     5.07
## neighborhoodNeighborhood Nine      11.02     5.78
## neighborhoodNorth Cambridge        -5.28     5.83
## neighborhoodRiverside               6.44     5.42
## neighborhoodStrawberry Hill       -26.94     9.11
## neighborhoodWellington-Harrington  -8.95     5.80
## neighborhoodWest Cambridge         -8.99     6.34
## bedrooms                           59.35     6.93
## room_typePrivate room             -61.54     2.73
## room_typeShared room              -77.15     9.00
##
## Error terms:
##  Groups       Name        Std.Dev. Corr
##  accommodates (Intercept)  2.44
##               bedrooms    26.76    1.00
##  Residual                 80.64
## ---
## number of obs: 7112, groups: accommodates, 16
## AIC = 82733.6, DIC = 82693.6
## deviance = 82693.6
```

6. Interpretation

For Model 2, if I would like to rent an airbnb MIT, I would expect the price to be 136.22 + 27.99 + 59.35*Bedrooms - 61.54*Private room - 77.15*shared room

By increasing one bedroom, I would expect to price to rise by 59.35, if it is a private room, the price would drop by 61.54, and if it is a shared room, the price would drop by 77.15. It makes sense because the shared room is normally cheaper than a private one.

6. Model check

```
binnedplot(fitted(mulre1), resid(mulre1), cex.main=1.2, model_name = "Model 1", main="residual plot", n
```

```
## Warning in plot.window(...): "model_name" is not a graphical parameter
```

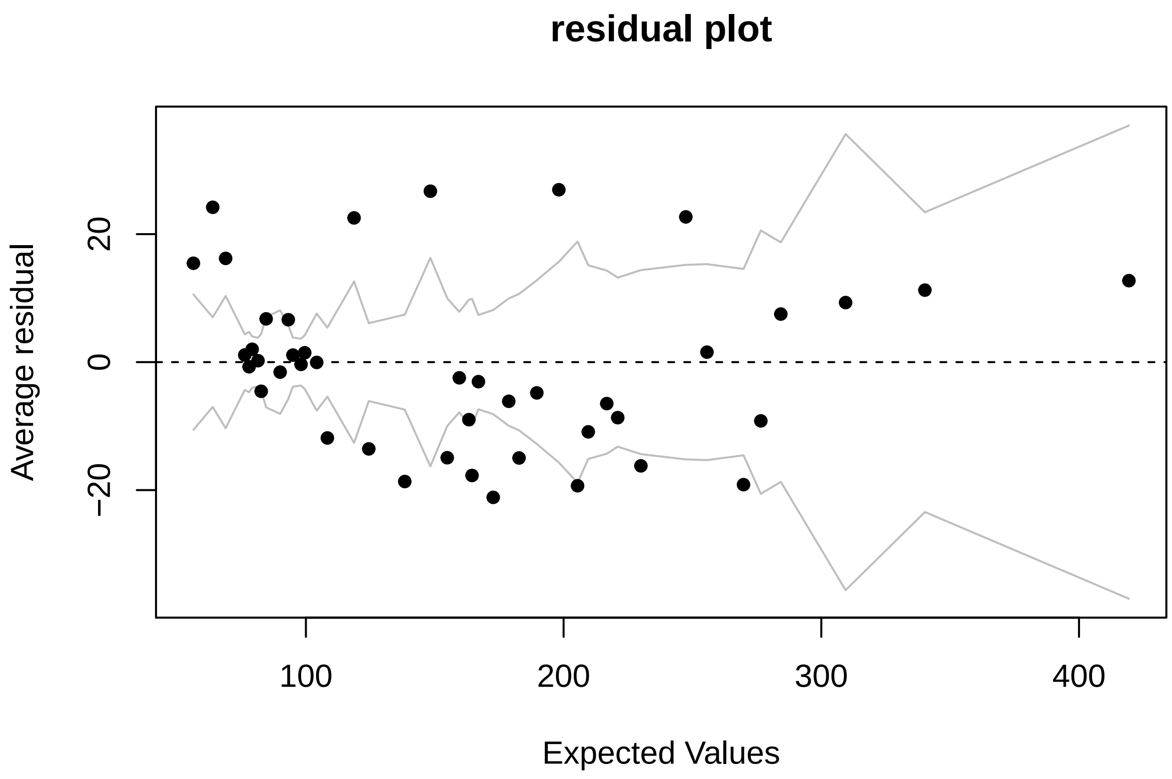
```
## Warning in plot.xy(xy, type, ...): "model_name" is not a graphical
## parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "model_name"
## is not a graphical parameter
```

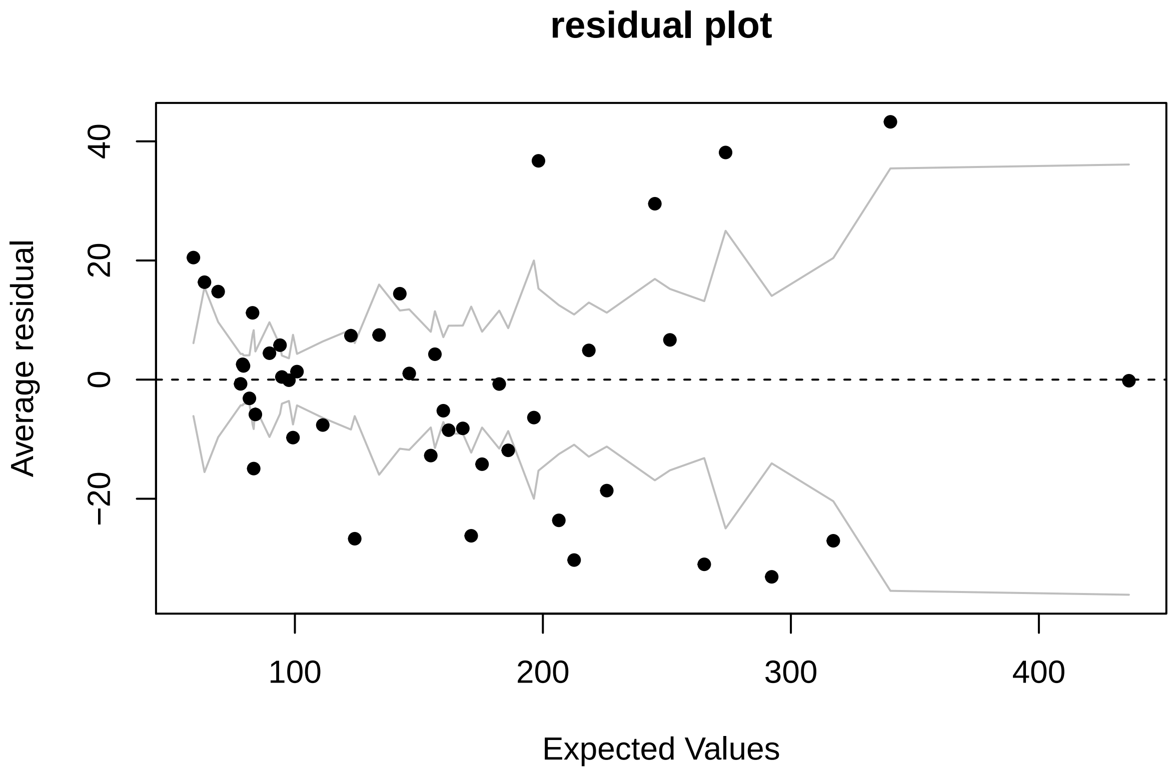
```
## Warning in axis(side = side, at = at, labels = labels, ...): "model_name"
## is not a graphical parameter
```

```
## Warning in box(...): "model_name" is not a graphical parameter
```

```
## Warning in title(...): "model_name" is not a graphical parameter
```



```
binnedplot(fitted(mulre2), resid(mulre2), cex.main=1.2, main="residual plot", nclass = 50)
```



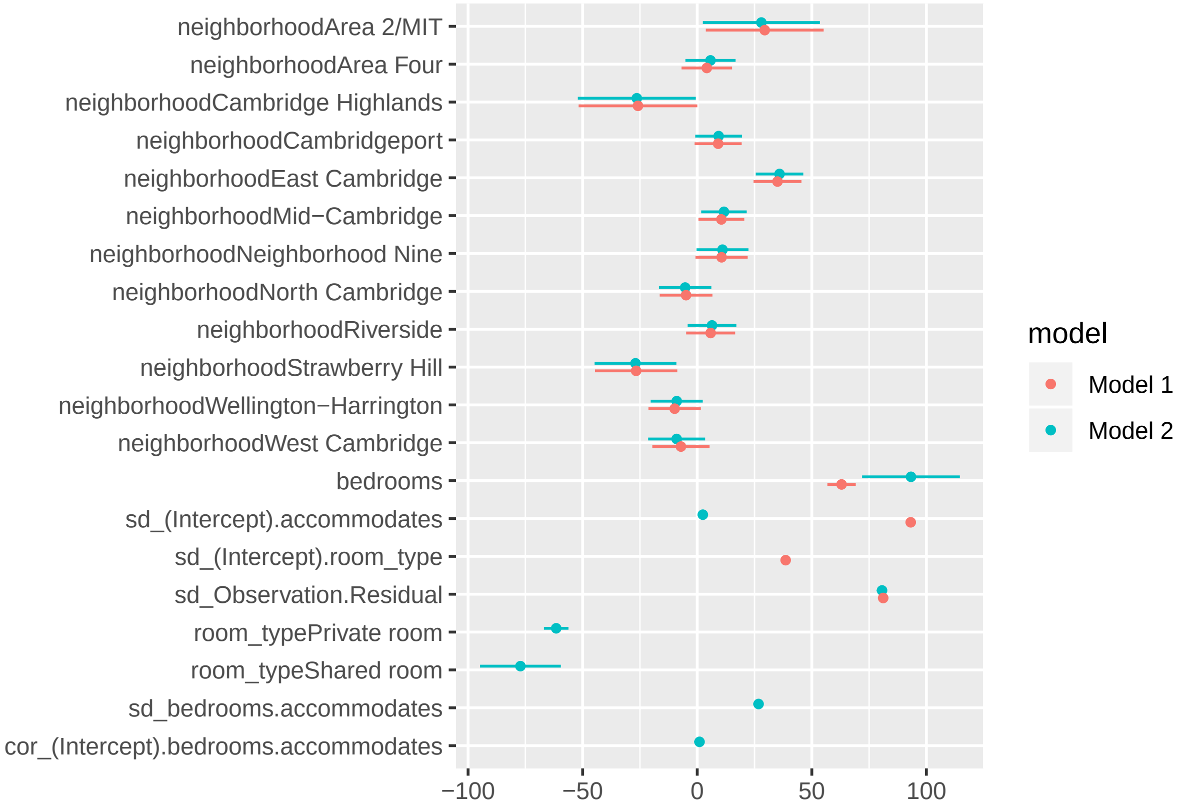
```
dwplot(list(mulre1, mulre2), dodge_size = 0.4, show_intercept = FALSE)
```

```
## Warning in bind_rows_(x, .id): binding factor and character vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding factor and character vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```



From the plots we can see that the second regression model have a better and more reliable result.

7. ANOVA

```
anova(mulre1, mulre2)
```

```
## Data: CAMB_main
## Models:
## mulre1: price ~ neighborhood + bedrooms + (1 | room_type) + (1 | accommodates)
## mulre2: price ~ neighborhood + bedrooms + room_type + (1 + bedrooms |
## mulre2:     accommodates)
##        Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## mulre1 17 82834 82951 -41400    82800
## mulre2 20 82734 82871 -41347    82694 106.85      3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model two has a DF of 3 and a smaller p-value which lead to the conclusion that Model 2 has a better fit.

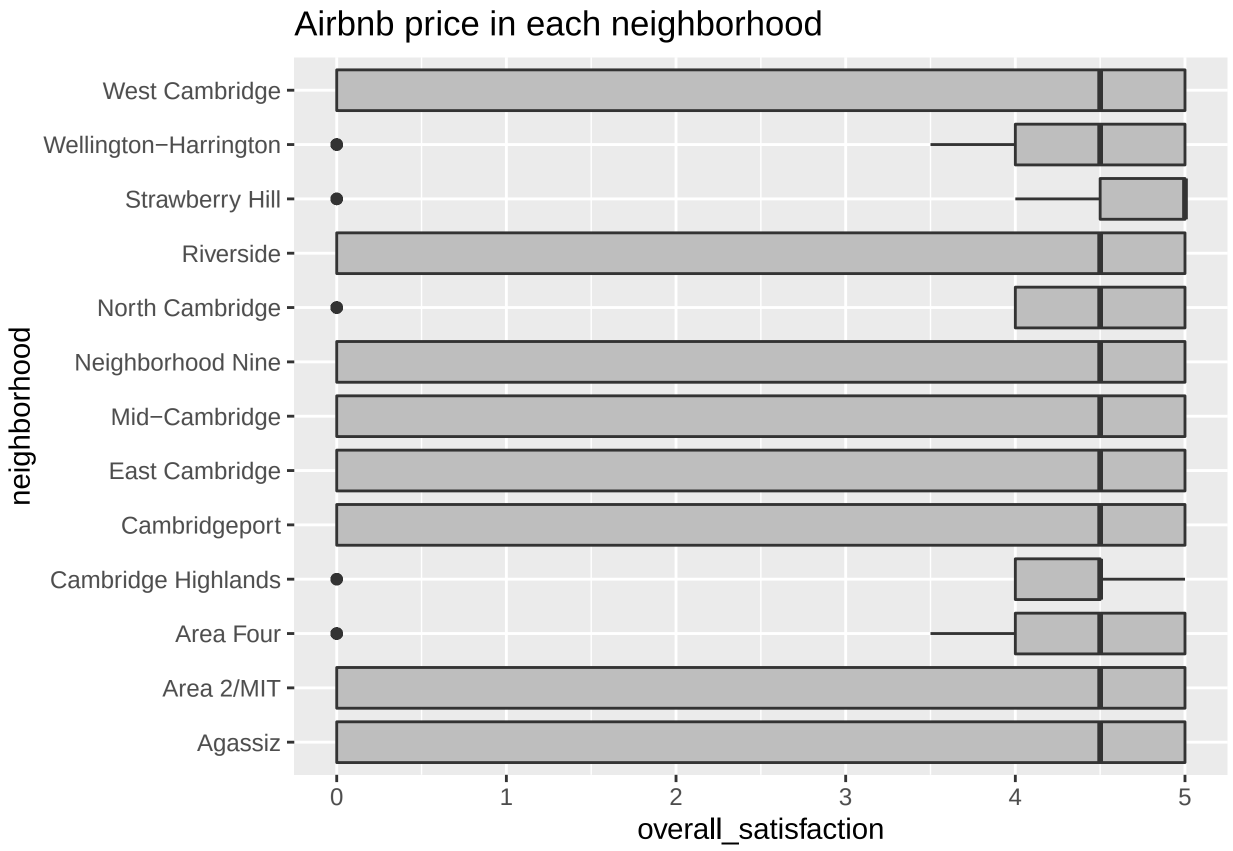
8. Discussion

#8.1 From futher research on google map, airbnb near Harvard and MIT tend to have a higher price comparing to other areas in Cambridge, which makes sense considering that those areas are more popular for visiters

#8.2 The result of the fit shows some residuals outside the model which indicate some limitation of the model

#8.3 My initial thought is to analyze the relationship between neighborhood and the rating, however, as I looked into the dataset and tried to do some plot, I found out that the rating is not sufficient enough for me to keep my work (see the graph below), also the result does not make any sense.

```
## Warning: Removed 599 rows containing non-finite values (stat_boxplot).
```



9. Future Direction

I am still very interested in the relationship between rating and other factors. I would like to find some more reliable dataset if possible, and keep up the work.