Airbnb

Luna Yihe Tang

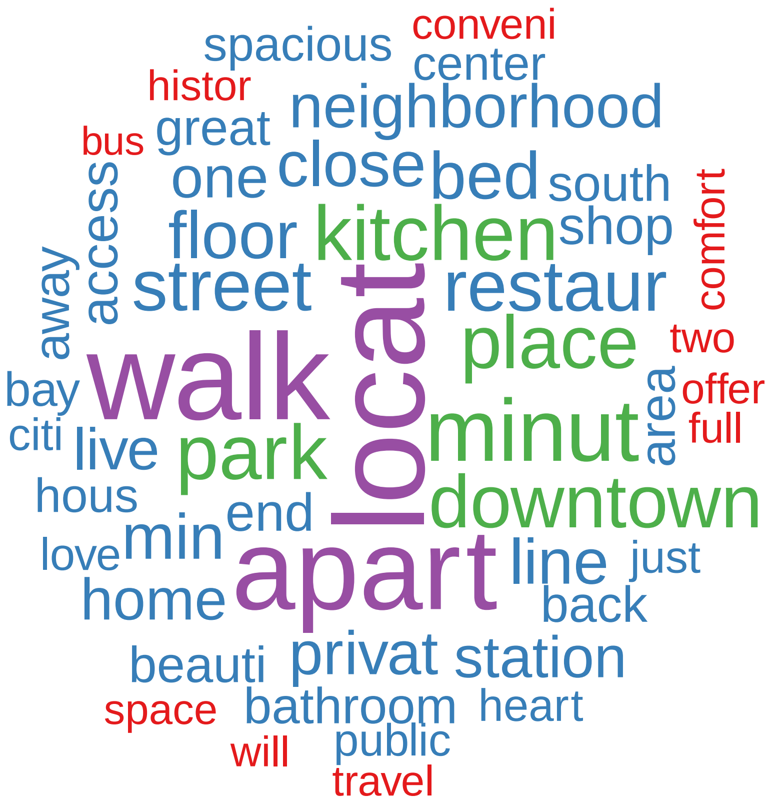
November 26, 2017

Introduction

The data I analized contains information of Boston Airbnb from Aug 2016 to July 2017. The data of Nov 2016 and Dec 2016 were removed because they do not contain how many bedrooms each property contains, while I will be comparing the price on a per-bedroom basis. My data contains room id, host id, roomtype, neighborhood, number of reviews, overall satisfaction rate, price, latitude, and longitude. I also analyzed a data sheet that contains all the detailed information by each Airbnb room in Boston in text analysis. Writer wants to figure out: 1. What are the the most mentioned words in the summary where the host describing their properties; what are the most mentioned house rules? 2. Where are the properties distributed? What types of rooms are there? 3. what are the important effects on price and satisfaction rate?

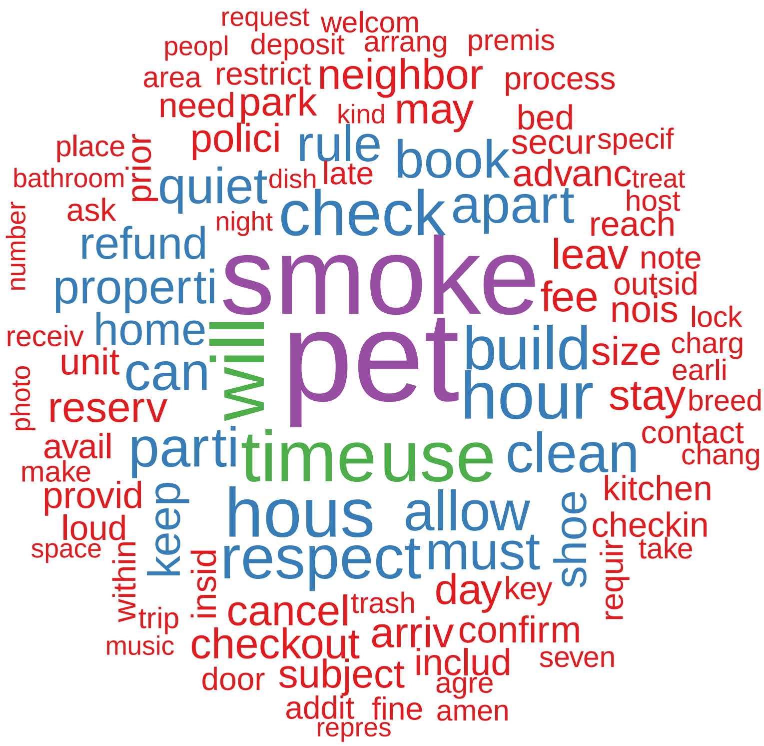
Text Analysis

```r
datatext<-read.csv("listings(1).csv",stringsAsFactors = FALSE)
#Text Analysis_Summary
jeopCorpus <- Corpus(VectorSource(datatext$summary))
jeopCorpus <- tm_map(jeopCorpus, PlainTextDocument)
jeopCorpus <- tm_map(jeopCorpus, stripWhitespace)
jeopCorpus <- tm_map(jeopCorpus, tolower)
jeopCorpus <- tm_map(jeopCorpus, removeNumbers)
jeopCorpus <- tm_map(jeopCorpus, removePunctuation)
jeopCorpus <- tm_map(jeopCorpus, removeWords, stopwords('english'))
jeopCorpus <- tm_map(jeopCorpus, stemDocument)
jeopCorpus <- tm_map(jeopCorpus, removeWords, "bedroom")
jeopCorpus <- tm_map(jeopCorpus, removeWords, "room")
jeopCorpus <- tm_map(jeopCorpus, removeWords, "boston")
jeopCorpus <- tm_map(jeopCorpus, removeWords, "pleas")
jeopCorpus <- tm_map(jeopCorpus, removeWords, "guest")
pal<-brewer.pal(4,"Set1")
wordcloud(jeopCorpus, max.words = 50, random.order = FALSE,colors=pal)
```



The word cloud of the summary of each property shows lacation, walk, minute, kitchen, downtown, restaurant are the most mentioned words. As wen can see, the location of the property is the most important factor by which the host used to sell their rooms. An apartment with downtown location, walking distance to major sights, close to restaurants would be considered to be the most attractive place to stay, from the host's perspective. Secondly, the structure of the apartment itself is also important. Whether it has a kitchen, a private bedroom and bathroom is also an important factor.

```
#Text Analysis_House Rules
jeopCorpus <- Corpus(VectorSource(datatext$house_rules))
jeopCorpus <- tm_map(jeopCorpus, PlainTextDocument)
jeopCorpus <- tm_map(jeopCorpus, stripWhitespace)
jeopCorpus <- tm_map(jeopCorpus, tolower)
jeopCorpus <- tm_map(jeopCorpus, removeNumbers)
jeopCorpus <- tm_map(jeopCorpus, removePunctuation)
jeopCorpus <- tm_map(jeopCorpus, removeWords, stopwords('english'))
jeopCorpus <- tm_map(jeopCorpus, stemDocument)
jeopCorpus <- tm_map(jeopCorpus, removeWords, "bedroom")
jeopCorpus <- tm_map(jeopCorpus, removeWords, "room")
jeopCorpus <- tm_map(jeopCorpus, removeWords, "boston")
jeopCorpus <- tm_map(jeopCorpus, removeWords, "pleas")
jeopCorpus <- tm_map(jeopCorpus, removeWords, "guest")
pal<-brewer.pal(4,"Set1")
wordcloud(jeopCorpus, max.words = 100, random.order = FALSE,colors=pal)
```



The word cloud of the house rules are also interested to look at. Pet and smoke are the top two most frequently mentioned words. "Use" and "respect" are also worth our attention.The host want to be clear about what are the things the guests are allowed to "use", and be respectful is the most important quality that the hosts require.

```
#Read csv files
data1608<-read.csv("2016-8.csv")
data1609<-read.csv("2016-9.csv")
data1610<-read.csv("2016-10.csv")
data1701<-read.csv("2017-1.csv")
data1702<-read.csv("2017-2.csv")
data1703<-read.csv("2017-3.csv")
data1704<-read.csv("2017-4.csv")
data1705<-read.csv("2017-5.csv")
data1706<-read.csv("2017-6.csv")
data1707<-read.csv("2017-7.csv")
```

Read data

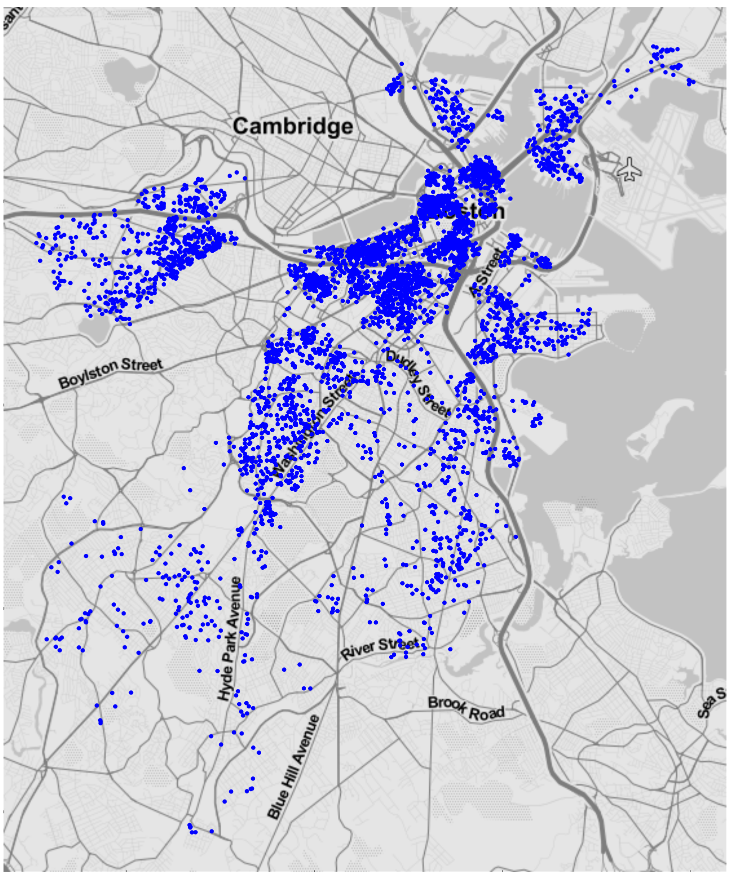
Map of a distribution of Airbnb properties in Boston using the most recent data(July 2017)

```
library(ggmap)
map1707 <- (data.frame(
  x = data1707$latitude,
  y = data1707$longitude
))
qmplot(y, x, data = map1707, colour = I('blue'), size = I(0.1), darken = .1)
```

```
## Using zoom = 12...
```

```
## Map from URL : http://tile.stamen.com/toner-lite/12/1238/1514.png
```

```
## Map from URL : http://tile.stamen.com/toner-lite/12/1239/1514.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1240/1514.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1238/1515.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1239/1515.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1240/1515.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1238/1516.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1239/1516.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1240/1516.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1238/1517.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1239/1517.png

## Map from URL : http://tile.stamen.com/toner-lite/12/1240/1517.png

## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```



From the map we can see there are two major cluster of plots. 1.Downtown Boston and Commonwealth Avenue; 2. Allston

```
#Combine monthly files to one data frame to consider the situation for a year
data<-rbind(data1608,data1609,data1610,data1701,data1702,data1703,data1704,data1705,data1706,data1707)
data$bedrooms[data$bedrooms == 0] <- NA #Turn 0 values into NAs in order to remove the properties with
data$reviews[data$reviews == 0] <- NA #I removed rows with 0 reviews becasue probably means those prope
data$overall_satisfaction[data$overall_satisfaction == 0.0] <- NA # I removed rows with 0 satisfaction
```

```
newdata<-na.omit(data) # Removed all the unwanted data.
newdata$price_per_bedroom<-round(newdata$price/newdata$bedrooms,0)# Add price per bedroom to the origin
```

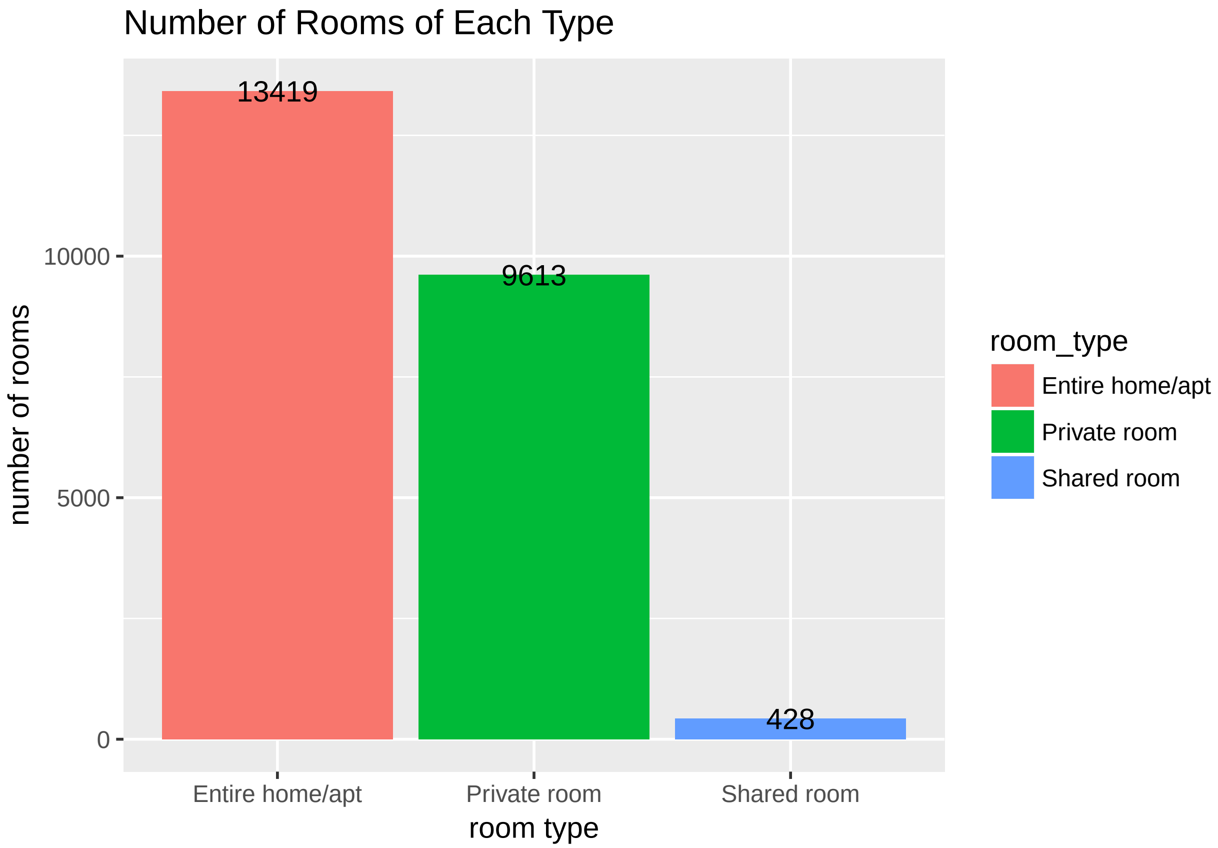
EDA

What many rooms are there by room types?

```
table(newdata$room_type)#Count by room type
```

```
##
## Entire home/apt    Private room     Shared room
##           13419            9613             428
```

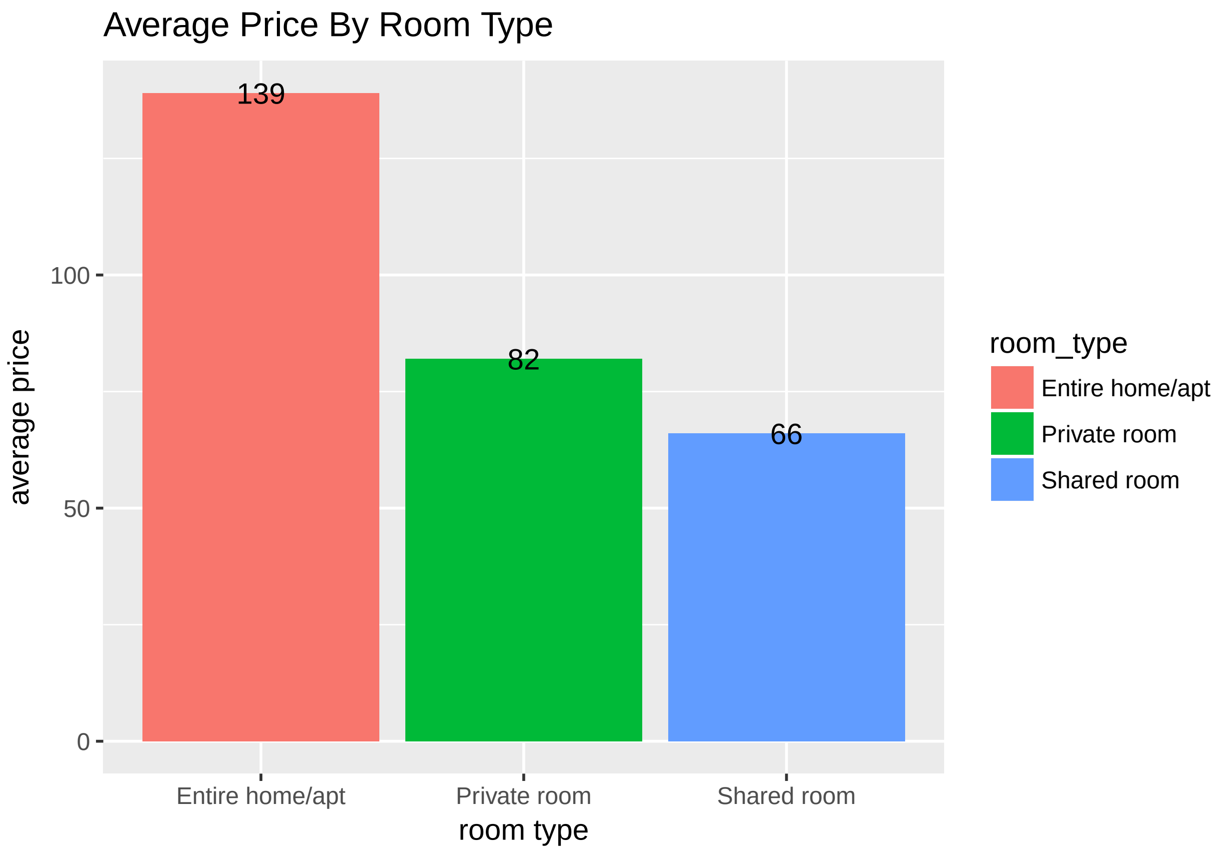
```
count_roomtype<-sqldf("SELECT COUNT(room_id) as number_of_rooms, room_type FROM newdata GROUP BY room_ty
#distribution of property type
ggplot(count_roomtype,aes(x=room_type,y=number_of_rooms,fill=room_type))+geom_bar(stat="identity")+geom_
```



Most of the hosts rent their entire apartment. Only a few of hosts are willing to share their room with guests.

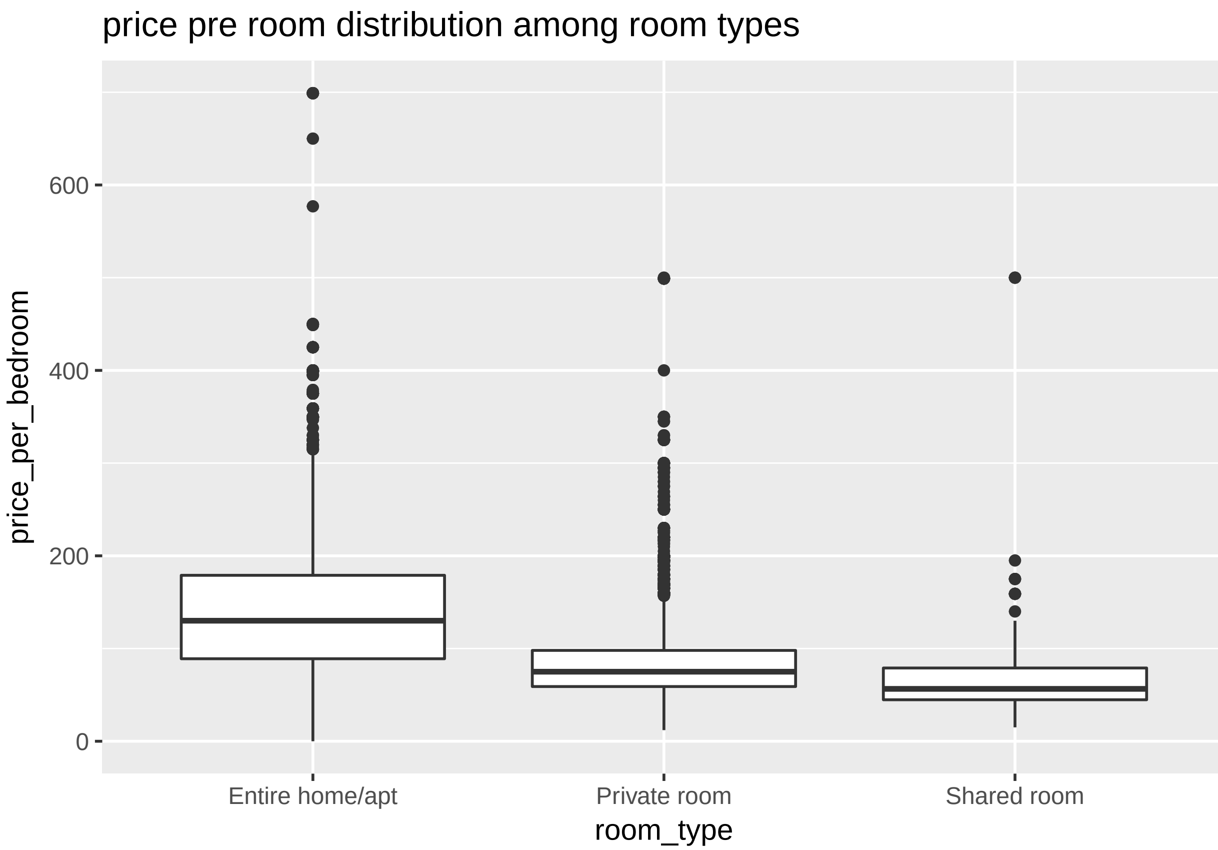
What is the cheapest room type?

```
bdprice_by_roomtype<-sqldf("SELECT room_type,avg(price_per_bedroom) as avg_bdprice FROM newdata GROUP BY
#Average Price Per Bedroom of Each Room Type
bdprice_by_roomtype$avg_bdprice<-round(bdprice_by_roomtype$avg_bdprice)
ggplot(bdprice_by_roomtype,aes(x=room_type,y=avg_bdprice,fill=room_type))+geom_bar(stat="identity")+labs
```



After deviding the price of entire home/apartment by how many bedrooms it has, the unit price of a bedroom of an entire home/apartment is still the highest. That's probabaly becasue usually entire apartments have a living room which can also accomodate some guests, and it is perfect for a group of travellers to have their own space as a group. Not surprisingly, share room has the lowest prices becasue who doesn't want their own room?

```
#Room type-price per room distribution
p4 <- ggplot(newdata, aes(x=room_type, y=price_per_bedroom, fill=price_per_bedroom)) + geom_boxplot() +
p4
```



There are a lot of high prices of each type, while not much low prices according to the box plot. Next time if you want to save money, get a shared room in Airbnb.

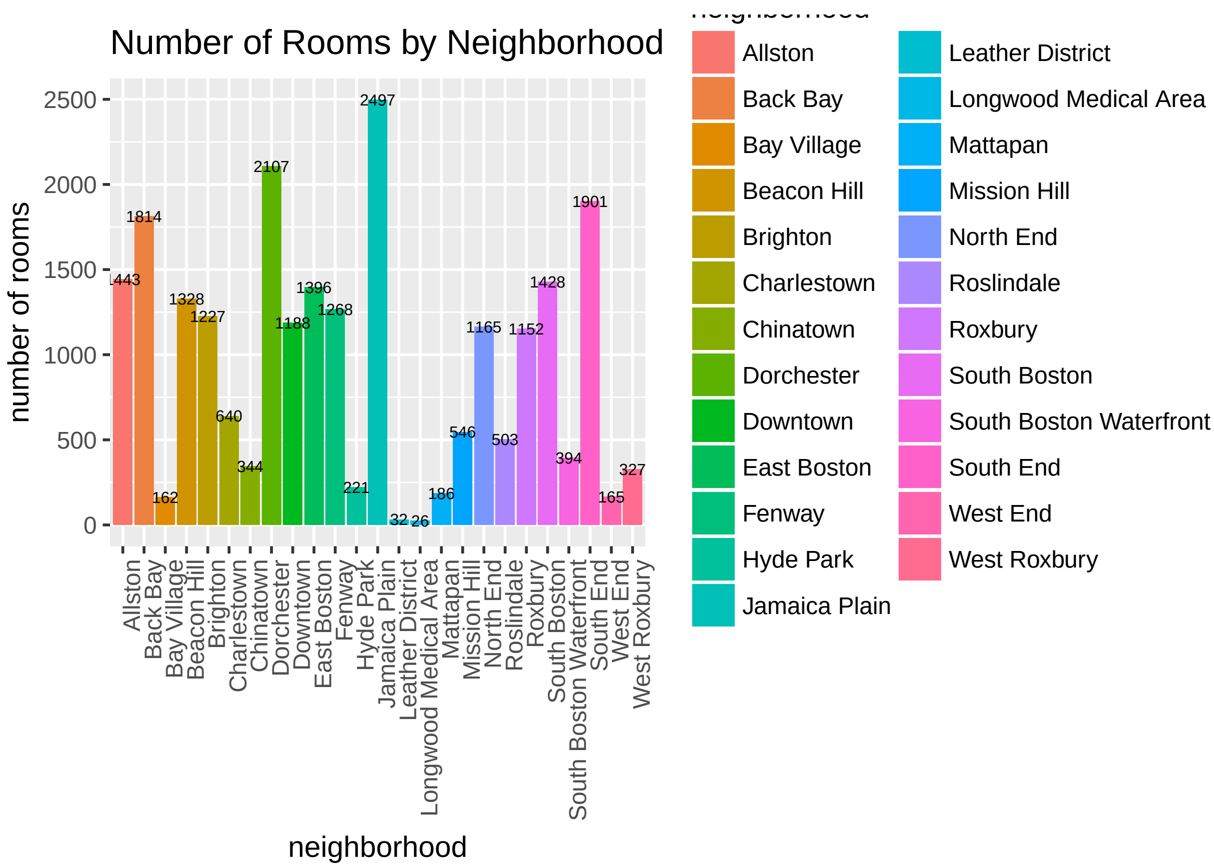
What neighborhood has the most Airbnb rooms?

```
table(newdata$neighborhood)
```

```
##
##                  Allston                 Back Bay              Bay Village
##                     1443                     1814                      162
##              Beacon Hill                 Brighton              Charlestown
##                     1328                     1227                      640
##                Chinatown               Dorchester                 Downtown
##                      344                     2107                     1188
##              East Boston                   Fenway                Hyde Park
##                     1396                     1268                      221
##            Jamaica Plain         Leather District    Longwood Medical Area
##                     2497                       32                       26
##                 Mattapan             Mission Hill                North End
##                      186                      546                     1165
##               Roslindale                  Roxbury             South Boston
##                      503                     1152                     1428
## South Boston Waterfront                South End                 West End
##                      394                     1901                      165
##             West Roxbury
##                      327
```

```
#Count by neighborhood
count_neighborhood<-sqldf("SELECT COUNT(room_id) as number_of_rooms, neighborhood FROM newdata GROUP BY
```

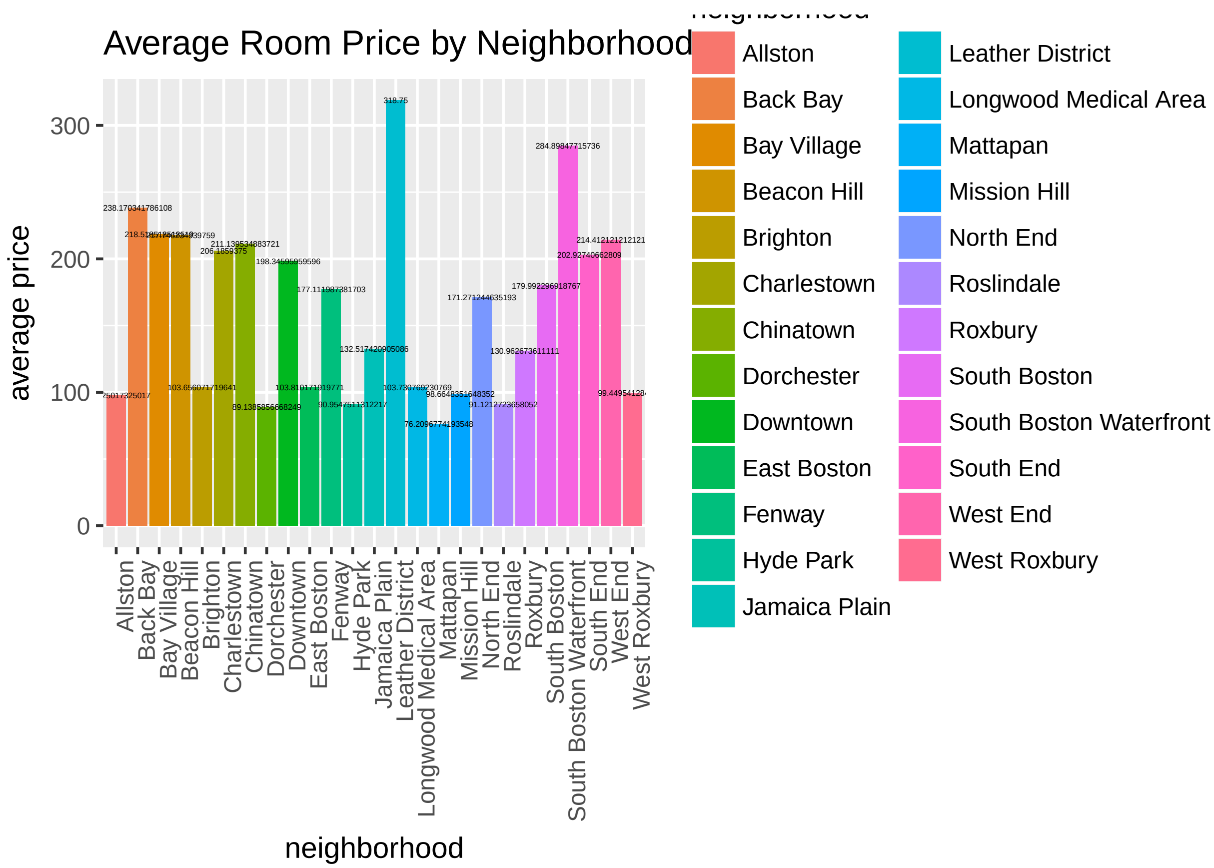
```
ggplot(count_neighborhood,aes(x=neighborhood,y=number_of_rooms,fill=neighborhood))+geom_bar(stat="identi
```



Jamaica Plain, Dorchester, and South End have the most properties.

Where to stay to get the cheapest price?

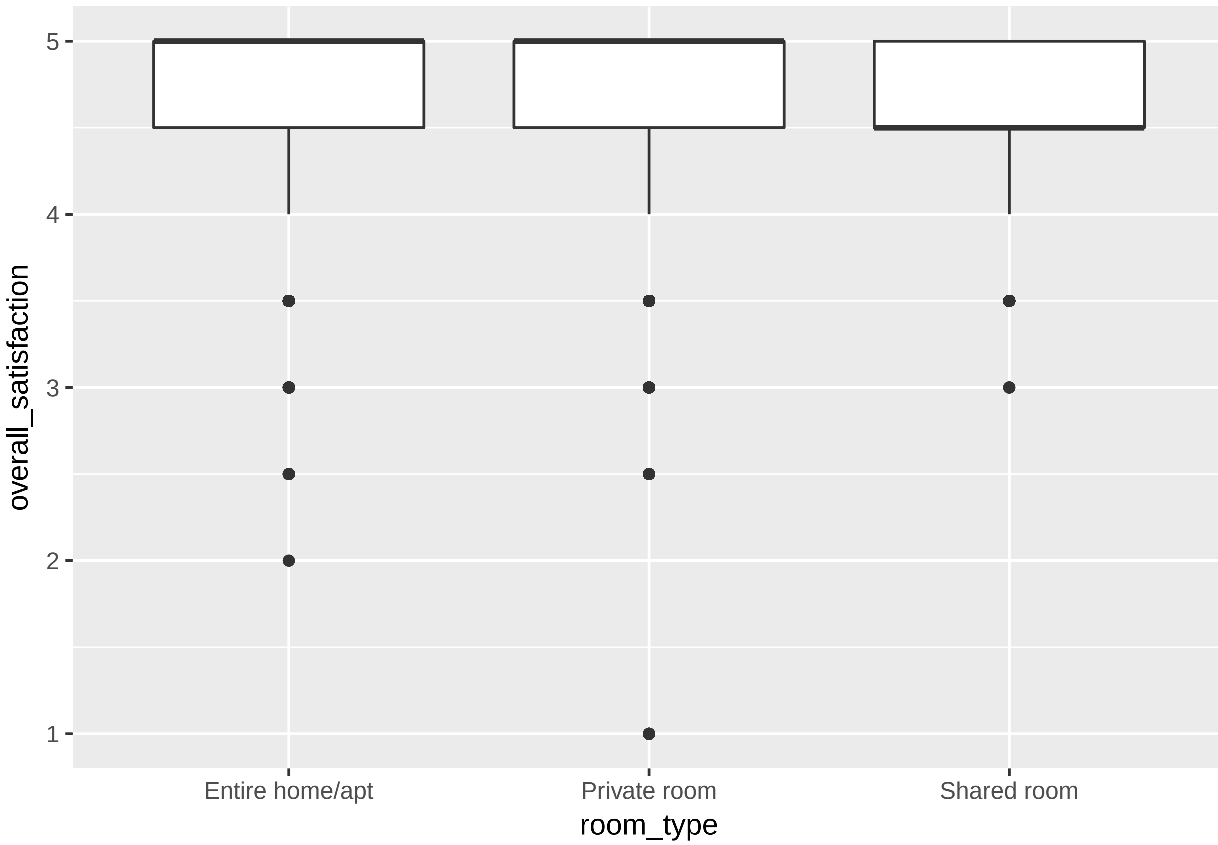
```
price_by_neighborhood<-sqldf("SELECT neighborhood,avg(price) as average_price FROM newdata GROUP BY neig
ggplot(price_by_neighborhood,aes(x=neighborhood,y=average_price,fill=neighborhood))+geom_bar(stat="ident
```



We can see from the graph that Allston, Dorchester, Mattapan, and Roslindale have the lowerst prices. However, consider the number of good restaurants nearby, I highly recommend Allston!

What roomtype gets the highest satisfaction rate?

```
#Box Plot Overall Satisfaction by room types.
ggplot(newdata, aes(x=room_type, y=overall_satisfaction, fill=overall_satisfaction)) + geom_boxplot()
```



From the box plot, Nearly all the properties got 4.5 or 5 rating, no matter what roomtype it is. That's becasue people usually want to give a positive feedback, and people rate their Airbnb either they receive a very nice stay, or something extremenly horrible happened. That's why there are a few low satisfaction rates of each room type. Overall, most of the people are ok with their stay at Airbnb. But some people will just give a 5 star to a ok stay just for their won convenience. I personally do that too. I gave every Ok Uber driver a 5 star just becasue I won't bother rating them. A suggestion to Airbnb is, in order to get more detailed feedback, use discount or other benefits to encourage guests filling out more detailed feedbacks.

Multilevel Models

Since we found out ratings are almost all 4.5 to 5 stars, there is no need to build a model use rating as outcome variables. Let's focus on what affects price per bedroom at this moment.

First I built a model using price per bedroom as outcome, number of reviews and overall satisfaction as numeric random variables, use room type, neighborhood, and room-id as groups with various intercepts.

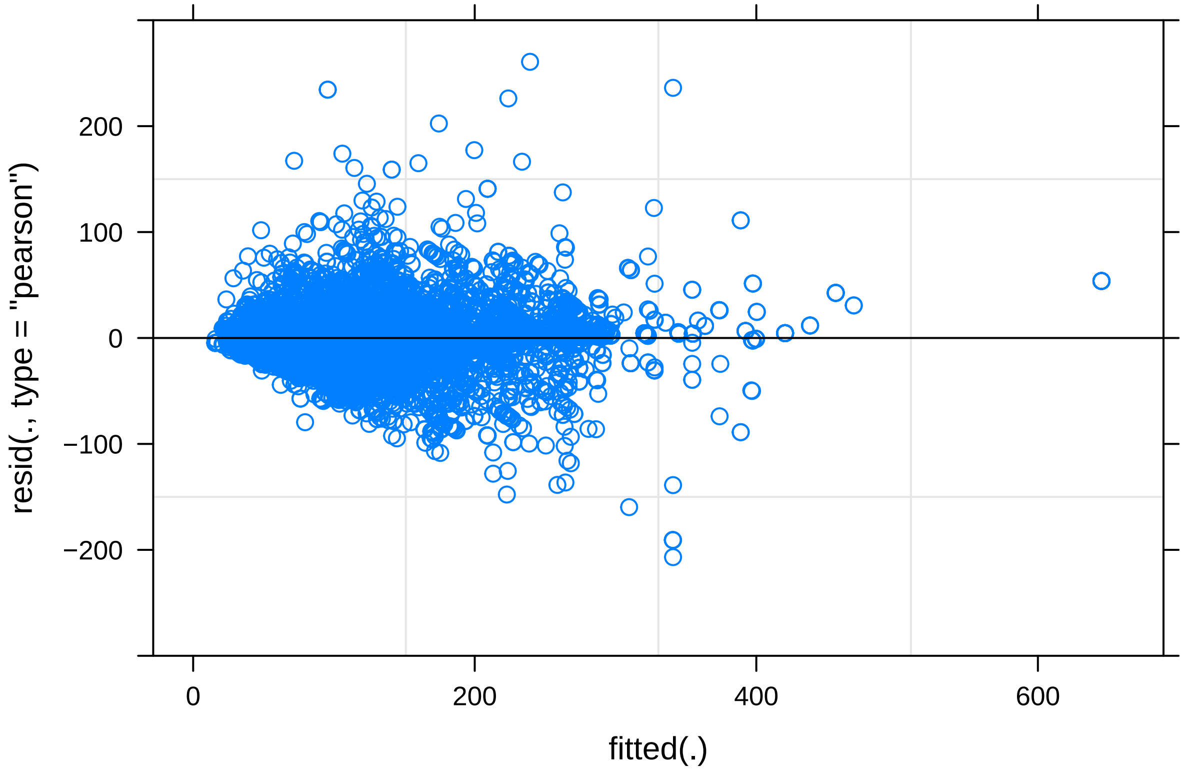
```
model1<-lmer(price_per_bedroom~reviews+overall_satisfaction+(1|room_type)+(1|neighborhood)+(1|room_id),
summary(model1)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula:
## price_per_bedroom ~ reviews + overall_satisfaction + (1 | room_type) +
##     (1 | neighborhood) + (1 | room_id)
##    Data: newdata
##
##       AIC       BIC    logLik  deviance  df.resid
##  220925.8  220982.2 -110455.9  220911.8     23453
##
```

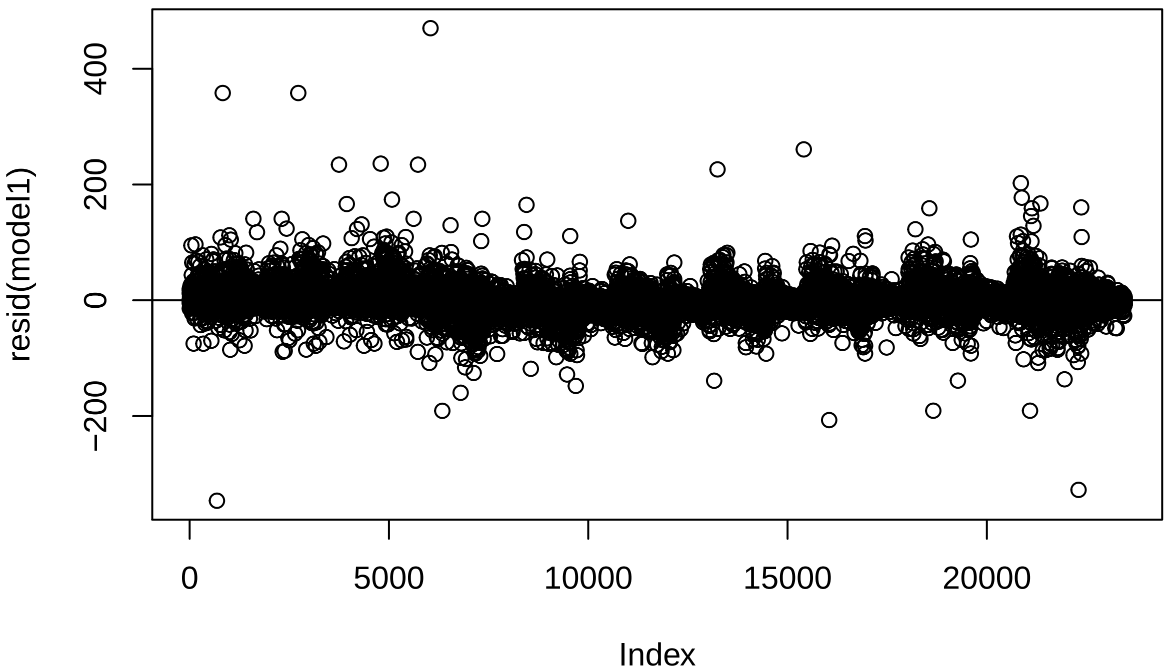
```
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -17.0653  -0.1840  -0.0030   0.1621  23.1748
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  room_id      (Intercept) 2330.9   48.28
##  neighborhood (Intercept) 1146.4   33.86
##  room_type    (Intercept)  364.8   19.10
##  Residual                  411.5   20.28
## Number of obs: 23460, groups:
## room_id, 3917; neighborhood, 25; room_type, 3
##
## Fixed effects:
##                      Estimate Std. Error t value
## (Intercept)          99.88002   14.05343   7.107
## reviews              -0.16760    0.01236 -13.558
## overall_satisfaction  1.39866    1.09438   1.278
##
## Correlation of Fixed Effects:
##             (Intr) reviws
## reviews     -0.012
## ovrll_stsfc -0.365 -0.026
```

From the model summary, the overall satisfaction has a positive effect on the price-per-bedroom, which means a 5-star-rating room are more likely to have a 1.4 higher price than a 4-start-rating property with all the other factors remaining the same, because its quality and popularity. However, number of reviews has a negative effect on price per bedroom, which might because cheaper rooms got the most guests, and then got the most reviews. Therefore, the model doesn't mean if a host want to raise the price of his property, he needs to somehow get less number of reviews.

```
plot(model1,ylim=c(-300,300))
```




```
plot(resid(model1))+abline(0, 0)
```




```
## integer(0)
```

From the two plots above, the model fit is not ideal.

What if we delete room-id from our model?

```
model2<-lmer(price_per_bedroom~reviews+overall_satisfaction+(1|room_type)+(1|neighborhood), data=newdat
summary(model2)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula:
## price_per_bedroom ~ reviews + overall_satisfaction + (1 | room_type) +
##     (1 | neighborhood)
##    Data: newdata
##
##       AIC       BIC    logLik  deviance  df.resid
##  252035.1  252083.5 -126011.6  252023.1     23454
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9115 -0.5594 -0.0686  0.4137 10.4529
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  neighborhood (Intercept) 1083.8   32.92
##  room_type    (Intercept)  649.4   25.48
##  Residual                 2692.9   51.89
## Number of obs: 23460, groups:  neighborhood, 25; room_type, 3
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)          50.186611  16.750636   2.996
## reviews              -0.039917   0.007848  -5.086
## overall_satisfaction 10.560547   0.938841  11.248
##
## Correlation of Fixed Effects:
##             (Intr) reviws
## reviews     -0.003
## ovrll_stsfc -0.264 -0.046
```

The AIC is still pretty big, so it didn't help.

What if we delete number of reviews since it's effect cannot be correctly shown by the previous model?

```
model3<-lmer(price_per_bedroom~+overall_satisfaction+(1|room_type)+(1|neighborhood), data=newdata,REML=
summary(model3)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: price_per_bedroom ~ +overall_satisfaction + (1 | room_type) +
##     (1 | neighborhood)
##    Data: newdata
##
##       AIC       BIC    logLik  deviance  df.resid
##  252059.0  252099.3 -126024.5  252049.0     23455
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9014 -0.5573 -0.0735  0.4120 10.4482
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  neighborhood (Intercept) 1088     32.99
##  room_type    (Intercept)  646     25.42
```

```
##  Residual                 2696     51.92
## Number of obs: 23460, groups:  neighborhood, 25; room_type, 3
##
## Fixed effects:
##                      Estimate Std. Error t value
## (Intercept)           49.9555    16.7227   2.987
## overall_satisfaction  10.3424     0.9384  11.022
##
## Correlation of Fixed Effects:
##             (Intr)
## ovrll_stsfc -0.265
```

We still got a similar model summary except for the high coefficient of overall satisfaction since we deleted room_id as a variable. However, we found out room_id cannot be deleted because we can't ignore the fact that the same properties were recorded repeatedly in each month. Overall, model 1 is still the best model among these three, although it has a large AIC, we can still learn meaningful things about Airbnb from that model.

Conclusion

From this analysis, I got a general sense of Airbnb Boston, including what's in the property summary and house rules, what type of room and neighborhood to choose to get the best deal, the problem of current rating system and how to improve that, and from the model, how much effect does rating have on the avreage per bedroom.