MA615final

Yichu Yan

12/02/2019

Airbnb property price and related factors

Background

Airbnb is a convenient website that we use to book lodging or primarily homestays. All of real estate listings are posted by property owners and these owners set unit room prices according to location, condition and many other seasonal influence factors. I plan to study elements that affect Airbnb room prices and try to find a model that could predict or provide a suggested price for Airbnb property. Then the model can be used by Airbnb hosts as a basic pricing tool.

Research question

Which factors could affect Airbnb room prices?

Data collection

I collect airbnb data of Boston in 2017. The dataset contains variables: room id, host id, room type, neighborhood, the number of reviews, the average rating, the number of guests a listing can accommodate, the number of bedrooms, the price for a night stay, latitude, longitude, the date and time that the values were read.

Read Data

Airbnb has 7 sheets of 2017 Boston dataset available online, and I just donwload and read them in RStudio first.

```
library(readr)
data1<-read_csv("~/Desktop/boston/tomslee_airbnb_boston_0779_2017-01-14.csv")
data2<-read_csv("~/Desktop/boston/tomslee_airbnb_boston_0858_2017-02-16.csv")
data3<-read_csv("~/Desktop/boston/tomslee_airbnb_boston_0931_2017-03-12.csv")
data4<-read_csv("~/Desktop/boston/tomslee_airbnb_boston_1043_2017-04-08.csv")
data5<-read_csv("~/Desktop/boston/tomslee_airbnb_boston_1187_2017-05-05.csv")
data6<-read_csv("~/Desktop/boston/tomslee_airbnb_boston_1309_2017-06-10.csv")
data7<-read_csv("~/Desktop/boston/tomslee_airbnb_boston_1429_2017-07-10.csv")
```

Data Cleaning

After viewing variables of each sheets, I find 4 sheets have 14 variables, while other 3 sheets have 20 variables. So, I decide to keep 14 variables.

```
length(colnames(data1))
```

```
## [1] 14
```

```
length(colnames(data2))
```

```
## [1] 14
```

```
length(colnames(data3))
```

```
## [1] 14
```

```
length(colnames(data4))
```

```
## [1] 14
```

```
length(colnames(data5))
```

```
## [1] 20
```

```
length(colnames(data6))
```

```
## [1] 20
```

```
length(colnames(data7))
```

```
## [1] 20
```

```
library(dplyr)
data5<-data5%>%select("room_id","host_id","room_type","borough","neighborhood","reviews","overall_satis
data6<-data6%>%select("room_id","host_id","room_type","borough","neighborhood","reviews","overall_satis
data7<-data7%>%select("room_id","host_id","room_type","borough","neighborhood","reviews","overall_satis
```

Data Organization

While combining these sheets, I find some properties have been updated and the whole sheet contains several obervations with same room_id. So, I delete duplicates and only keep the most recent observation. After data analysis, I also delete 2 blank columns and rows containing missing values. Next, I generate a new csv file containing all data.

```
mydata <- rbind(data7,data6,data5,data4,data3,data2,data1)
sample<-mydata%>%filter(room_id=="12071820")
```

```
library(tidyverse)
mydata <- distinct(mydata, room_id, .keep_all = TRUE)
sample<-mydata%>%filter(room_id=="12071820")
```

```
mydata<-mydata%>%select("room_id","host_id","room_type","neighborhood","reviews","overall_satisfaction"
library(funModeling)
data_integrity(mydata)
```

```
## $vars_num_with_NA
## [1] variable q_na     p_na
## <0 rows> (or 0-length row.names)
##
## $vars_cat_with_NA
## [1] variable q_na     p_na
## <0 rows> (or 0-length row.names)
##
## $vars_cat_high_card
## [1] variable unique
## <0 rows> (or 0-length row.names)
##
## $MAX_UNIQUE
## [1] 35
##
## $vars_one_value
## character(0)
##
## $vars_cat
## [1] "room_type"    "neighborhood"
##
## $vars_num
## [1] "room_id"              "host_id"              "reviews"
## [4] "overall_satisfaction" "accommodates"         "bedrooms"
## [7] "price"                "latitude"             "longitude"
##
## $vars_char
## [1] "room_type"    "neighborhood"
##
## $vars_factor
## character(0)
##
## $vars_other
## [1] "last_modified"
```

```
mydata <- drop_na(mydata)
write.csv(mydata,"~/Desktop/boston/Boston2017.csv", row.names = FALSE)
```

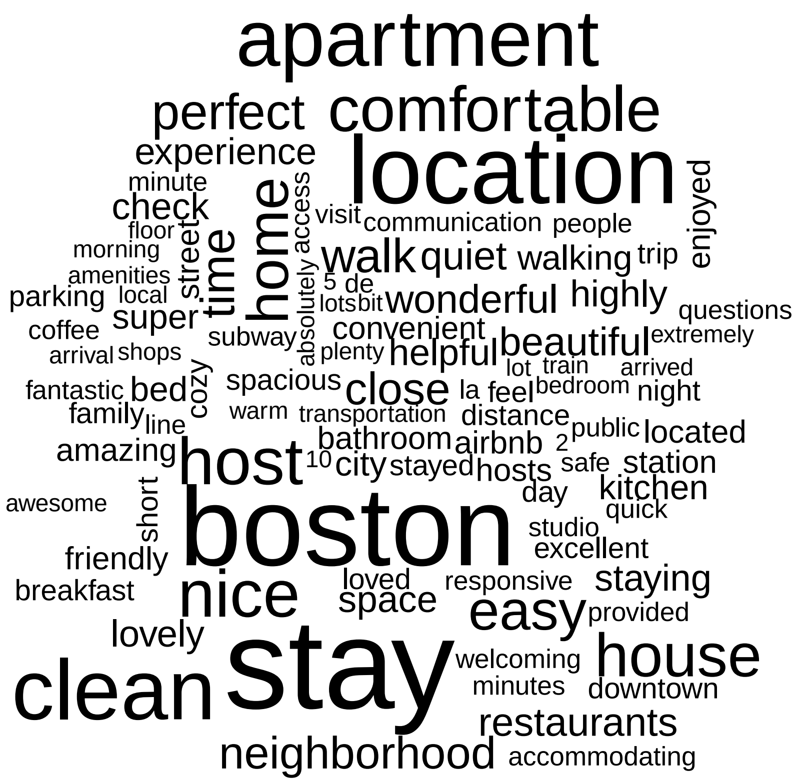
Review analysis

In order to visualize most frequent words shown in reviews, I use text mining to generate a graph.

```
library(gutenbergr)
library(tidytext)
library(knitr)
library(textdata)
library(magrittr)
library(tm)

booksource <- read.delim("~/Desktop/boston/reviews.txt", header=F, sep = "\n",stringsAsFactors = F)
booksource <- as.data.frame(booksource)
names(booksource)[1]  <- "text"
booksource <- booksource %>% mutate(gutenberg_id = 2007)

book <- "Reviews"
as.character(book)
```

```
## [1] "Reviews"
```

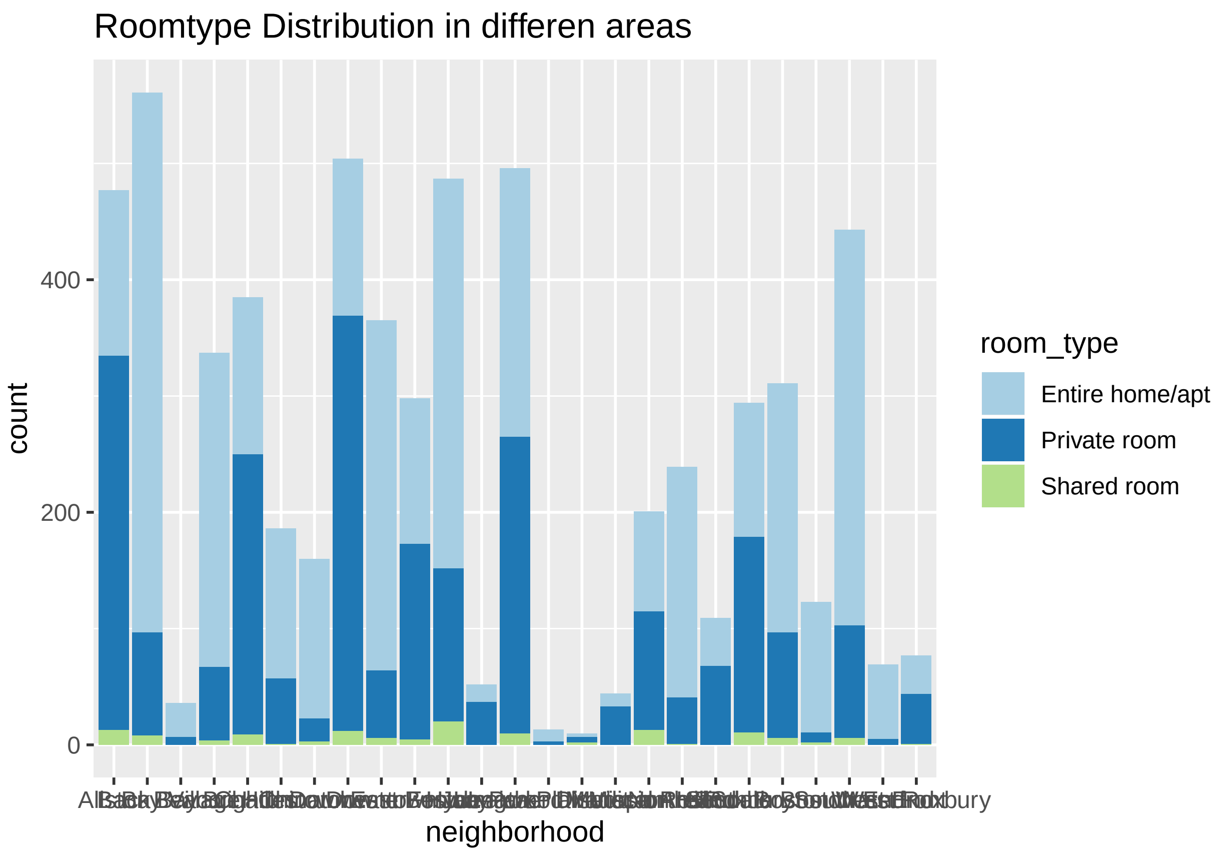
```
orignial_book <- cbind(booksource, book)

library(janeaustenr)
tidy_book <- orignial_book %>%
  group_by(book) %>%
  mutate(linenumber = row_number(),
         chapter = cumsum(str_detect(text, regex("^chapter [\\divxlc]",
                                                 ignore_case = TRUE)))) %>%
  ungroup()

tidy_book <- tidy_book %>%
  unnest_tokens(word, text)

library(wordcloud)
tidy_book %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 100))
```


apartment
perfect
comfortable
experience
location
minute
check
enjoyed
access
visit
communication
people
floor
morning
street
time
home
walk
quiet
walking
trip
amenities
absolutely
5
de
parking
local
lotsbit
wonderful
highly
questions
coffee
super
subway
convenient
beautiful
extremely
arrival
shops
plenty
helpful
lot
train
arrived
fantastic
bed
cozy
spacious
close
la
feel
bedroom
night
family
line
warm
transportation
distance
public
located
amazing
host
bathroom
airbnb
2
10
city
stayed
hosts
safe
station
kitchen
day
quick
awesome
short
boston
studio
excellent
friendly
loved
responsive
staying
breakfast
nice
space
easy
provided
lovely
stay
house
welcoming
clean
minutes
downtown
restaurants
neighborhood
accommodating

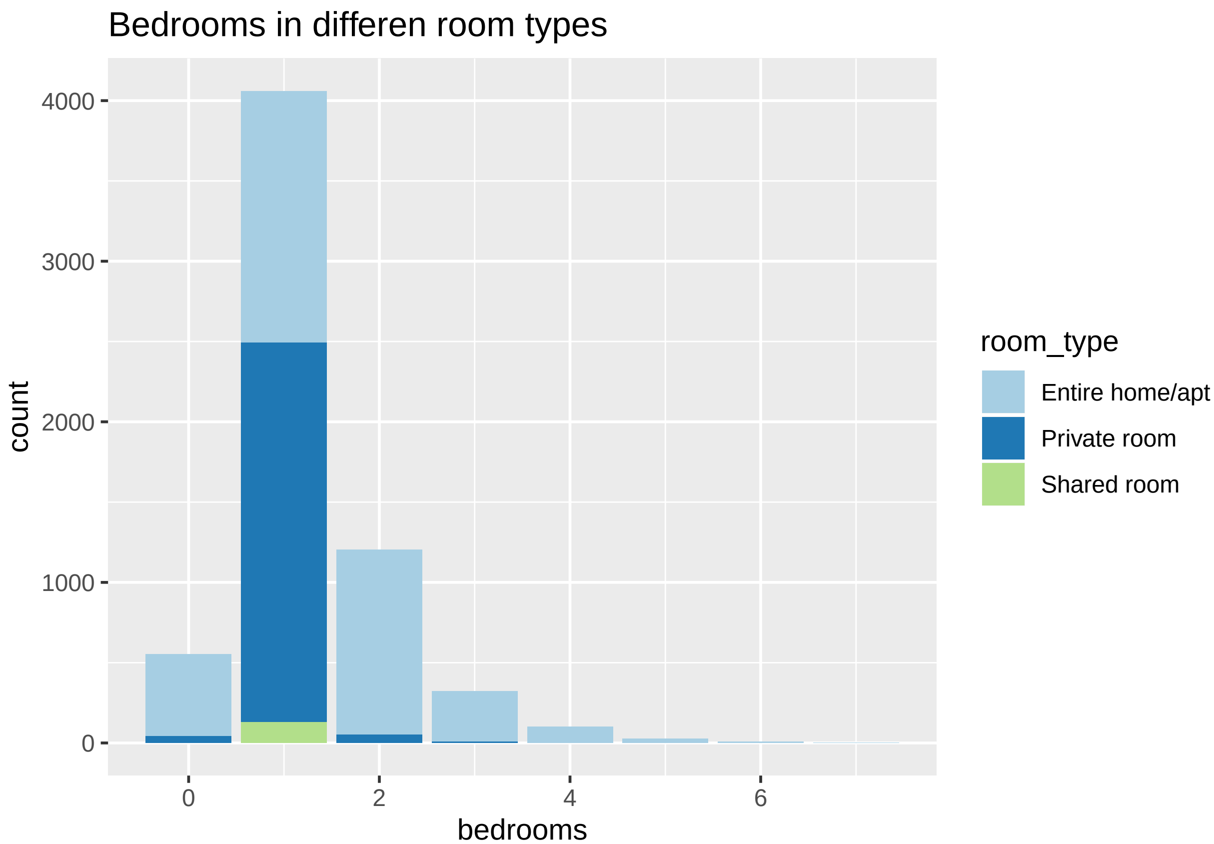
EDA

Data plots and table

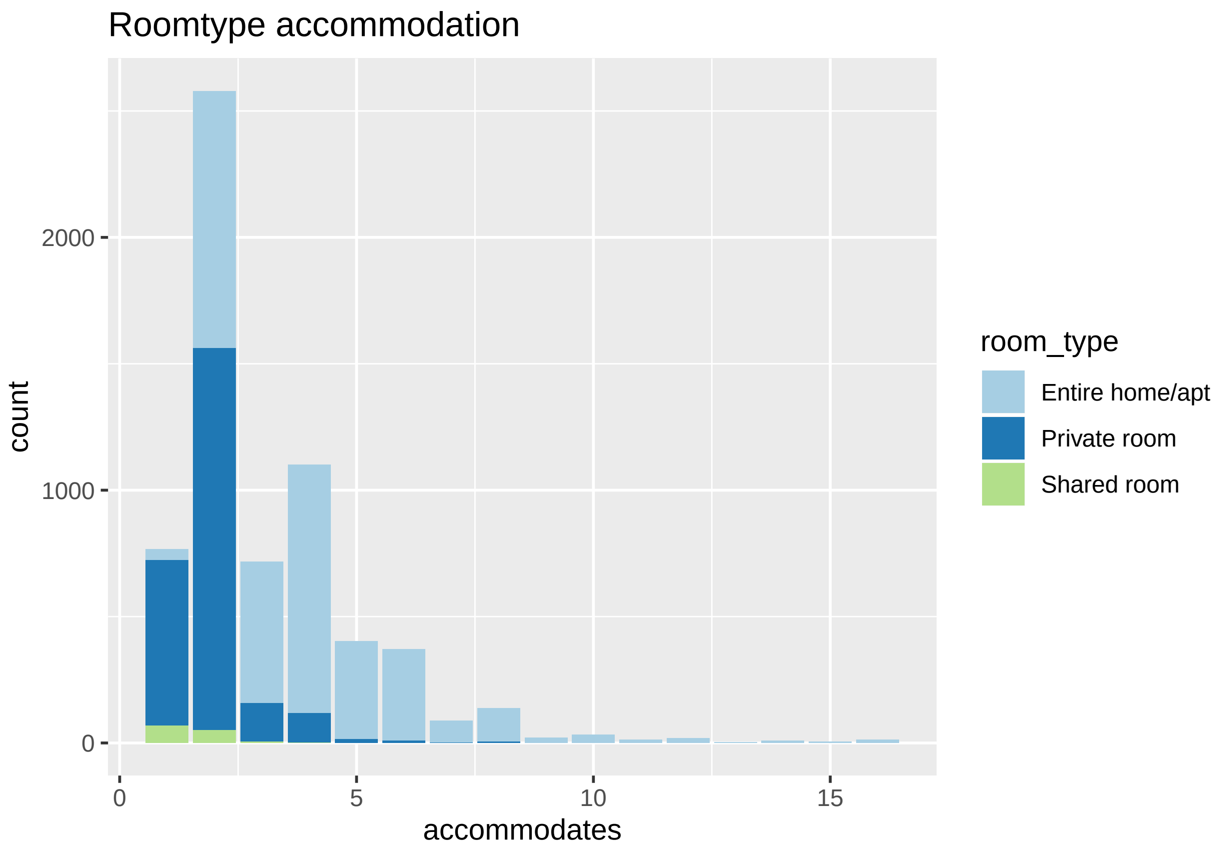
```
library(ggplot2)
library(stringi)
ggplot(mydata, aes(x = neighborhood, fill = room_type)) +
  geom_bar() +
  ggtitle("Roomtype Distribution in differen areas") +
  scale_fill_brewer(palette = "Paired")
```



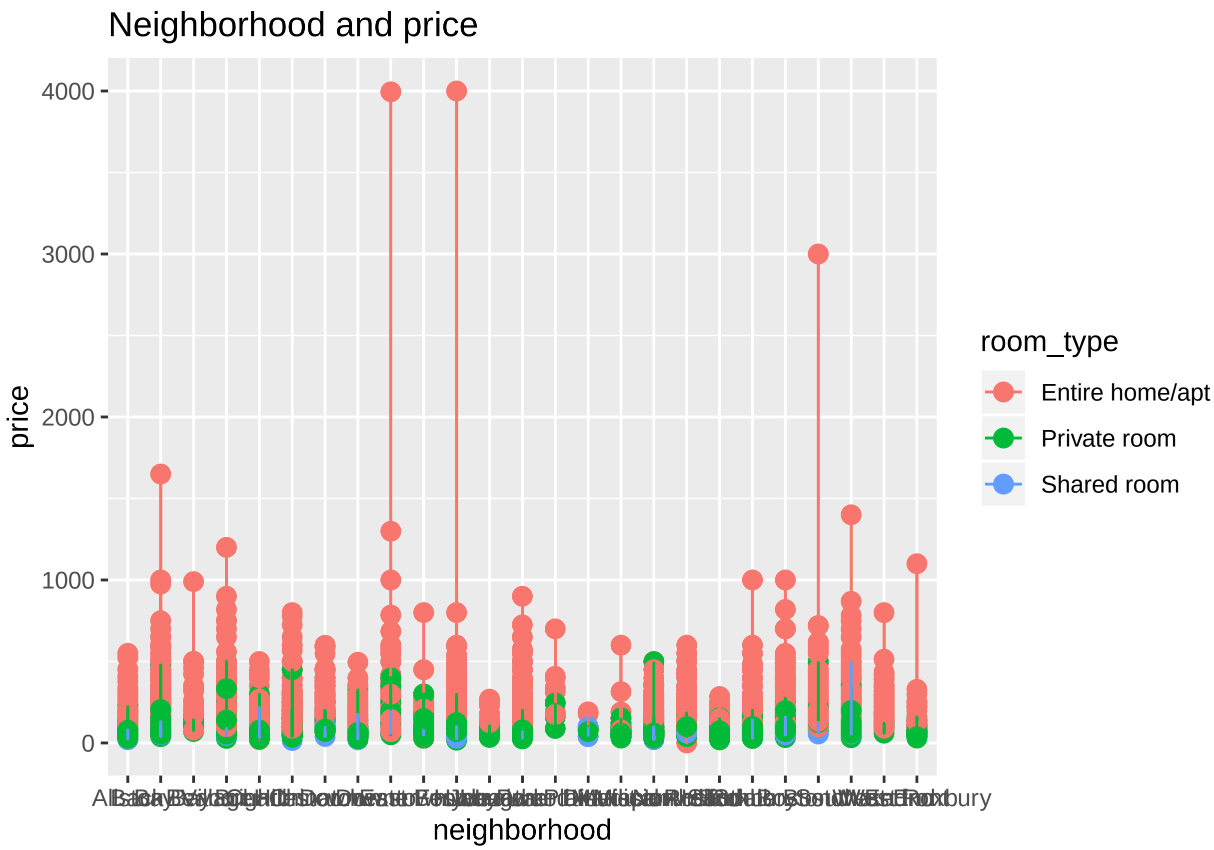

```
ggplot(mydata, aes(x = bedrooms, fill = room_type)) +
  geom_bar() +
  ggtitle("Bedrooms in differen room types") +
  scale_fill_brewer(palette = "Paired")
```



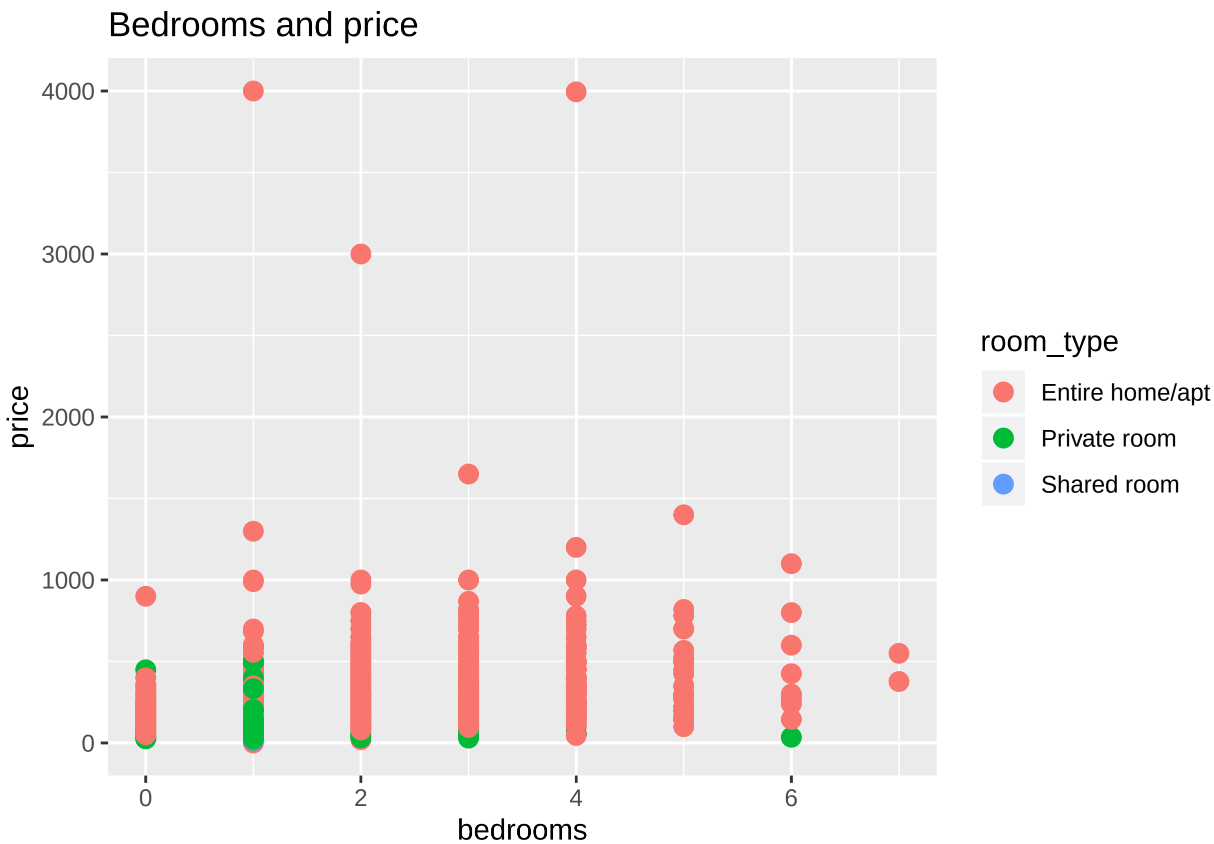

```r
ggplot(mydata, aes(x = accommodates, fill = room_type)) +
  geom_bar() +
  ggtitle("Roomtype accommodation") +
  scale_fill_brewer(palette = "Paired")
```



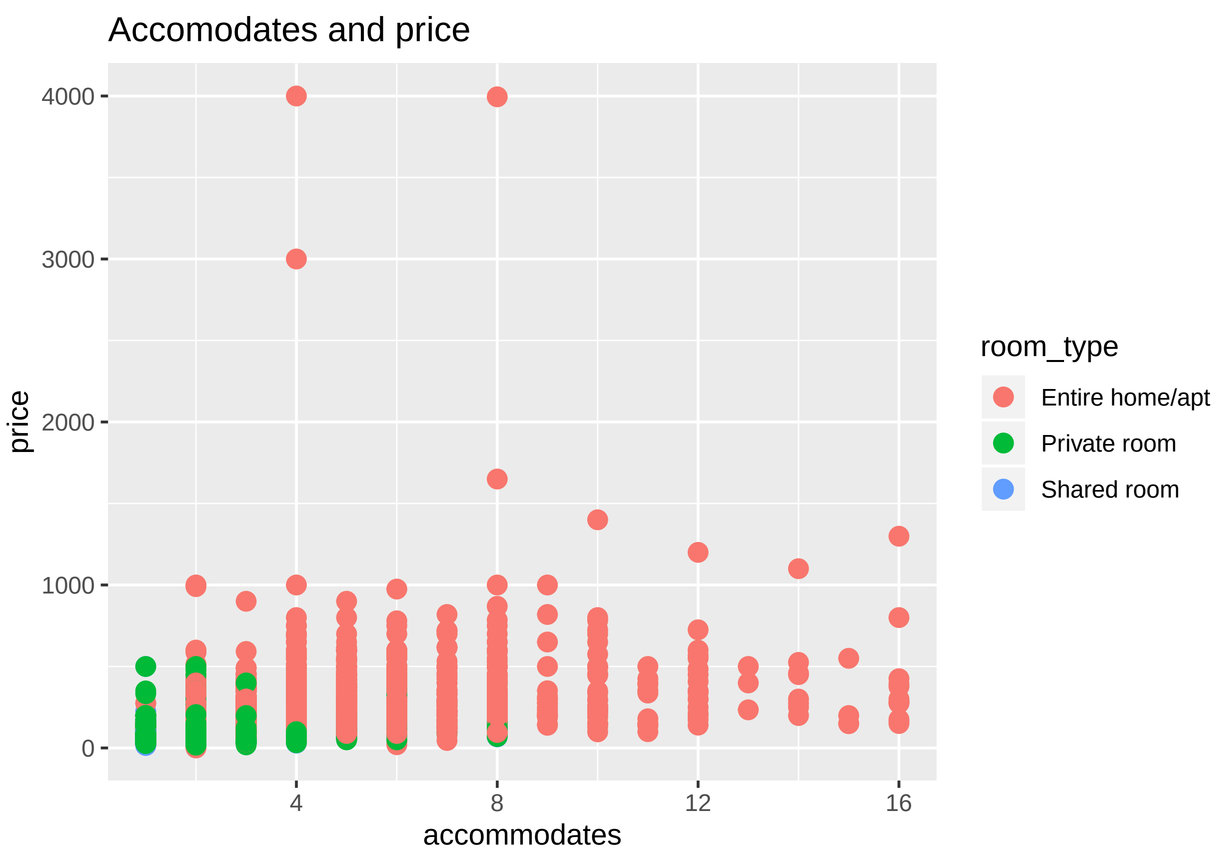

```r
ggplot(data = mydata, aes(x = neighborhood, y = price, color = room_type)) +
    geom_point(size = 3) +
    geom_line() +
  ggtitle("Neighborhood and price")
```




```r
ggplot(data = mydata, aes(x = bedrooms, y = price, color = room_type)) +
    geom_point(size = 3) +
  ggtitle("Bedrooms and price")
```




```r
ggplot(data = mydata, aes(x = accommodates, y = price, color = room_type)) +
    geom_point(size = 3) +
  ggtitle("Accomodates and price")
```


Accomodates and price
price
4000
3000
2000
1000
0
4
8
12
16
accommodates
room_type
Entire home/apt
Private room
Shared room

```r
library(kableExtra)
df1 <- mydata[,c("neighborhood","accommodates","price")]
df2 <- aggregate(df1[,2:3],by=list(df1$neighborhood),mean)
kable(df2, digits = 2,        ## call kable to make the table
      col.names = c("Location", "Average Rating", "Price"),
      caption = "Location and price by average rating" ,align = 'c') %>%
  kable_styling(latex_options = 'hold_position',font_size = 12,full_width = F,position = "center")%>%
  column_spec(1,bold = T)
```

Table 1: Location and price by average rating

Location	Average Rating	Price
Allston	2.53	89.40
Back Bay	3.13	218.80
Bay Village	3.67	270.31
Beacon Hill	3.02	194.54
Brighton	2.74	93.73
Charlestown	3.53	201.94
Chinatown	4.10	232.96
Dorchester	2.87	91.88
Downtown	3.67	230.74
East Boston	3.02	111.36
Fenway	2.93	184.14
Hyde Park	2.75	81.75
Jamaica Plain	3.21	132.41
Leather District	3.38	278.77
Longwood Medical Area	2.20	98.70
Mattapan	3.00	98.36
Mission Hill	2.73	124.71
North End	3.58	172.13
Roslindale	3.13	87.96
Roxbury	3.22	119.08
South Boston	3.99	188.60
South Boston Waterfront	3.81	306.25
South End	2.93	192.68
West End	3.68	241.94
West Roxbury	3.23	114.34

The data has 25 subregions in the neighborhood variable and plots compare unit room prices in different locations. The table also show summary of price and accomodates regardless of room types.

Concerns

Zero values in “the number of reviews” and “the average rating” may lead to potential problems. Usually, living spots with unattractive appearance or location probably have few or no reviews. But new posted houses also have zero review since no one has stayed before. If I keep these zero values in the fitted model, the model will predict relatively low prices for those new lodgings. In addition, the plot shows 3 outliers with pretty high price above $3000, which might make regression less reliable. In this way, I remove these observations.

```r
mydata$reviews[mydata$reviews=="0"] <- NA
mydata <- drop_na(mydata)
mydata <- mydata[!(mydata$price==max(mydata$price)),]
mydata <- mydata[!(mydata$price==max(mydata$price)),]
mydata <- mydata[!(mydata$price==max(mydata$price)),]
```

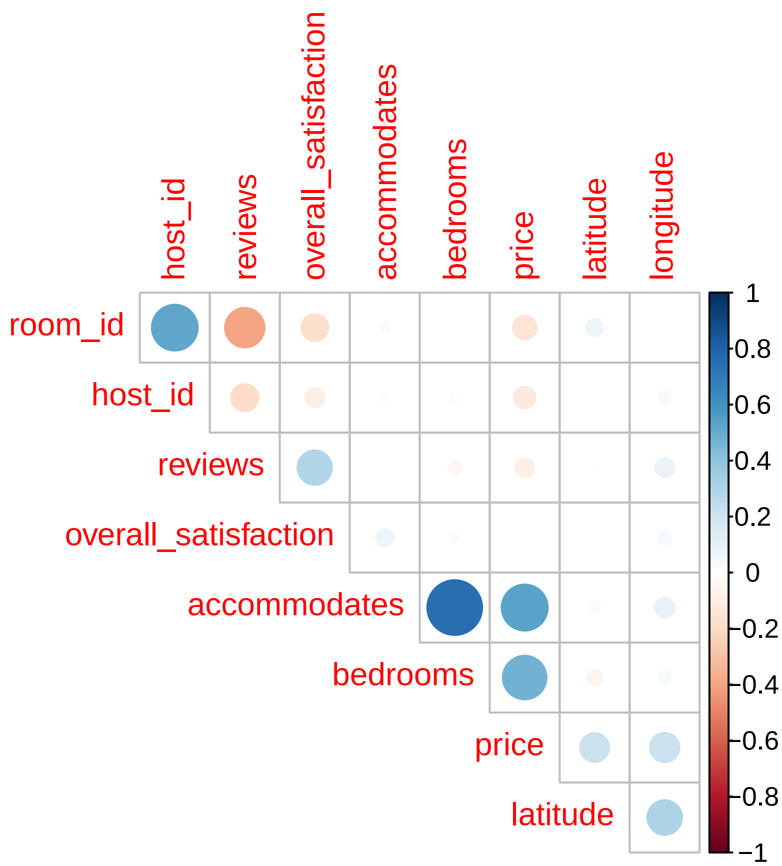
Methods

Correlation

```
mydata$reviews<-as.numeric(mydata$reviews)
mydata$overall_satisfaction<-as.numeric(mydata$overall_satisfaction)
mydata$latitude<-as.numeric(mydata$latitude)
mydata$longitude<-as.numeric(mydata$longitude)
sapply(mydata, is.numeric)
```

```
##               room_id              host_id            room_type
##                  TRUE                 TRUE                FALSE
##          neighborhood              reviews overall_satisfaction
##                 FALSE                 TRUE                 TRUE
##          accommodates             bedrooms                price
##                  TRUE                 TRUE                 TRUE
##              latitude            longitude        last_modified
##                  TRUE                 TRUE                FALSE
```

```
cordata <- mydata[, sapply(mydata, is.numeric)]
cor.ma <- cor(cordata, method = "pearson")
corrplot::corrplot(cor.ma, method = "circle", type = "upper", diag = F)
```

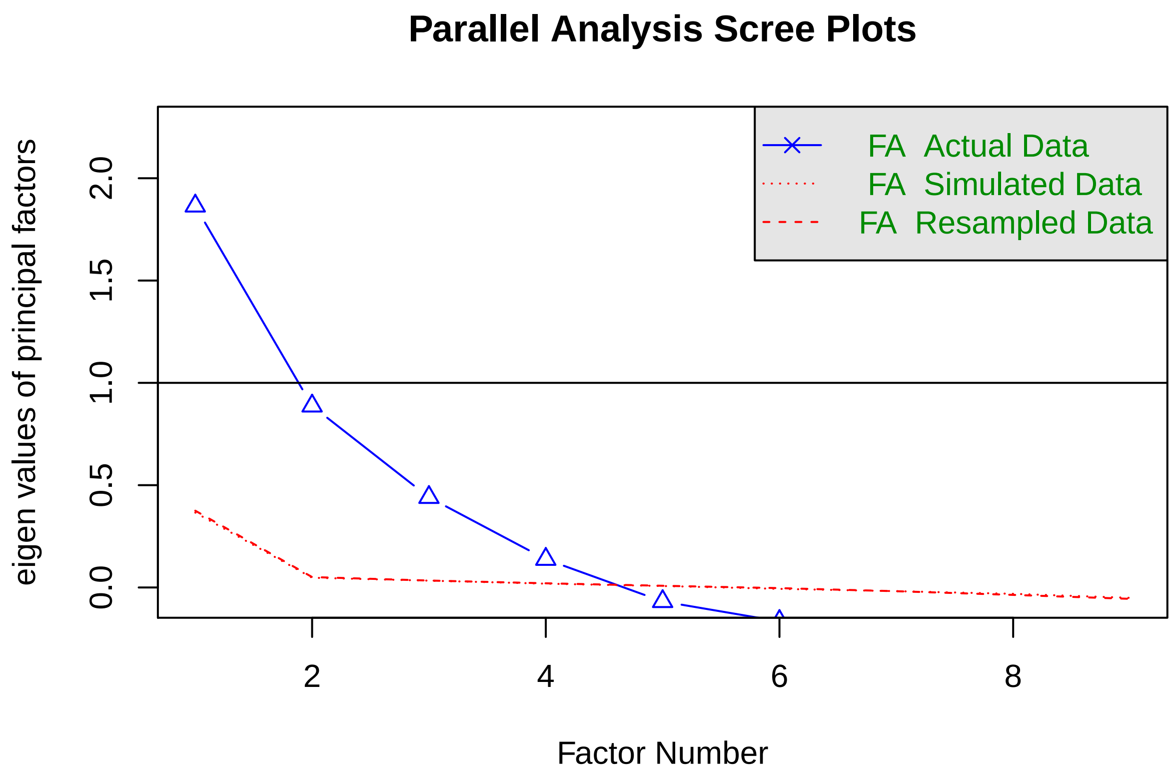


Some variables like accomodates and bedrooms have high correlations, so I need to consider only use part of them in models.

EFA

Dataset has 12 variables and I want to find out the number of factors that will be selected for later analysis.

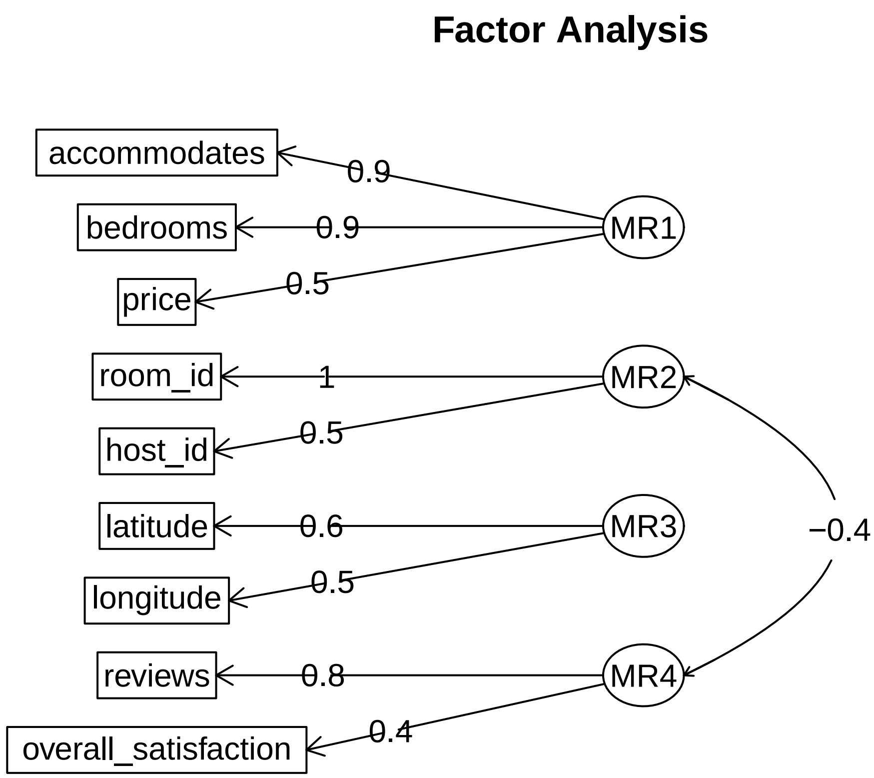
```
library(psych)
library(GPArotation)
parallel <- fa.parallel(cordata, fm = 'minres', fa = 'fa') # parallel analysis
```




```
## Parallel analysis suggests that the number of factors =  4  and the number of components =  NA
```

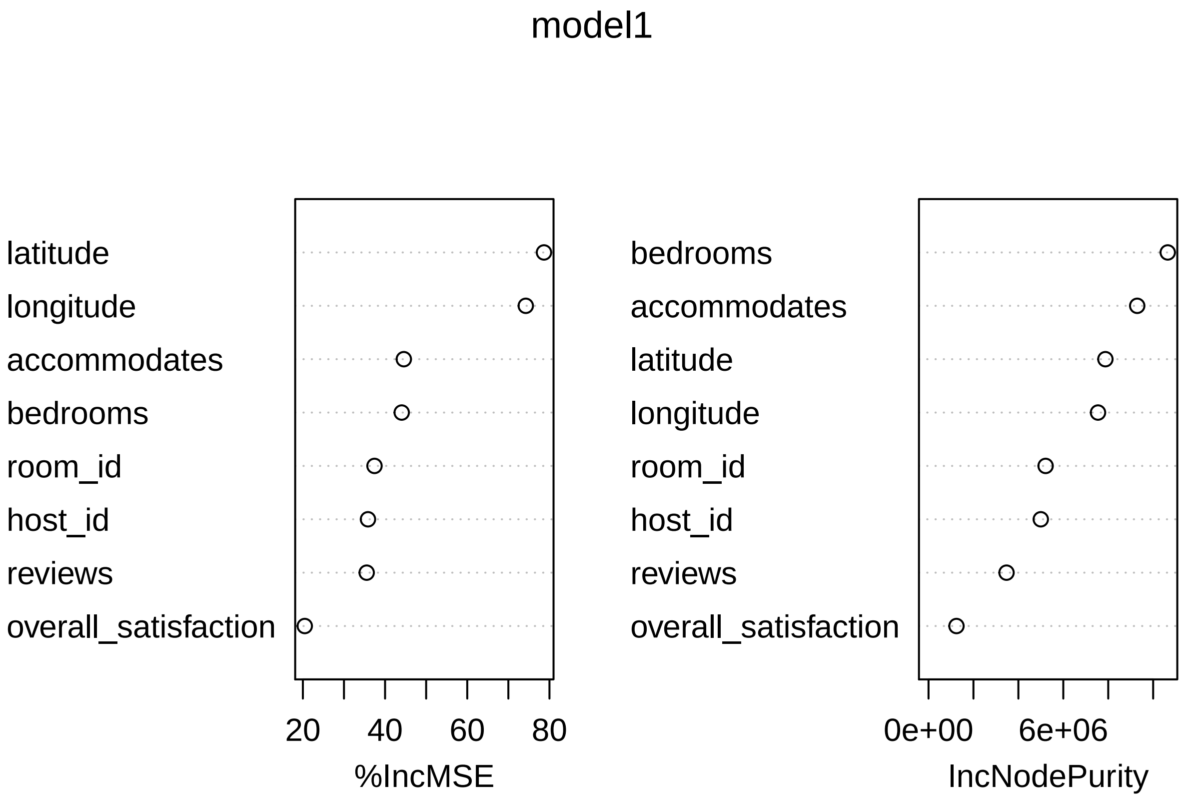
The blue line shows eigenvalues of actual data and the two red lines (placed on top of each other) show simulated and resampled data. Here we look at the large drops in the actual data and spot the point where it levels off to the right. Also we locate the point of inflection – the point where the gap between simulated data and actual data tends to be minimum.

```
fourfactor <- fa(cordata,nfactors = 4,rotate = "oblimin",fm="minres") # 4 factor analysis
fa.diagram(fourfactor)
```

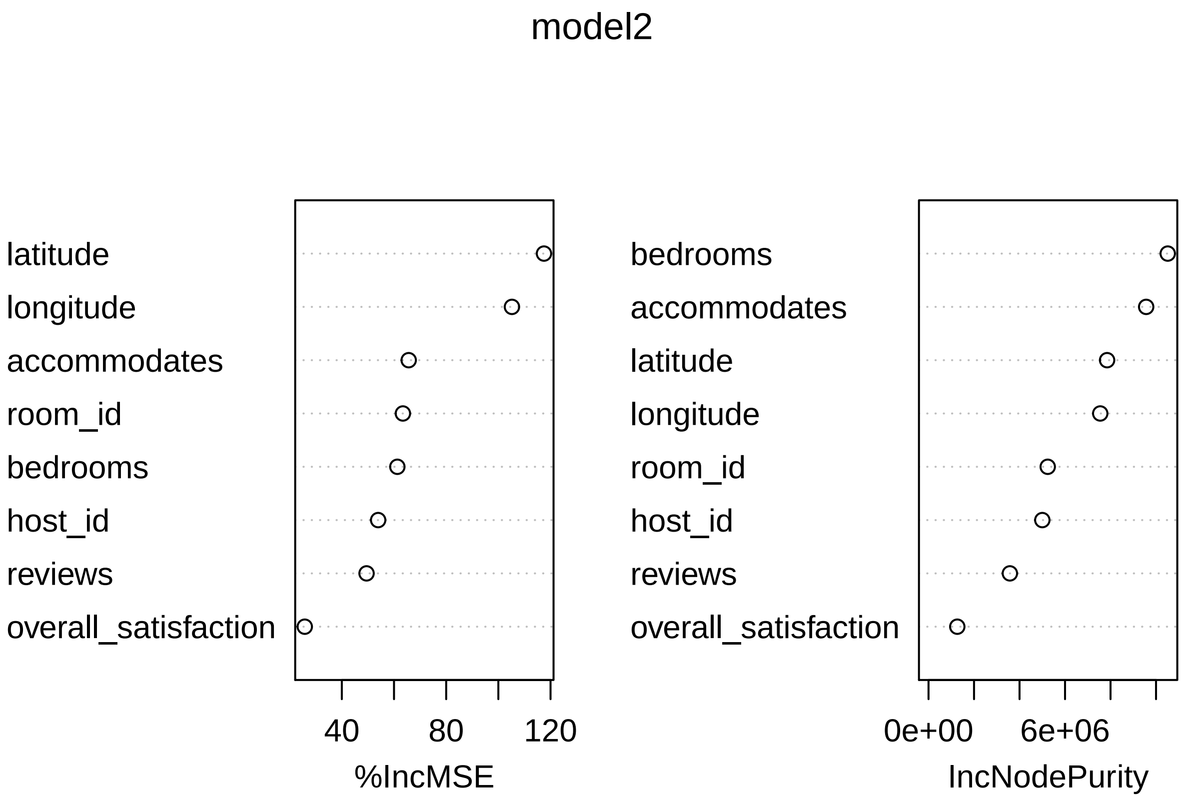

Factor Analysis
accommodates
bedrooms
price
room_id
host_id
latitude
longitude
reviews
overall_satisfaction
0.9
0.9
0.5
1
0.5
0.6
0.5
0.8
0.4
MR1
MR2
MR3
MR4
−0.4

Random Forest

```
library(randomForest)
model1 <- randomForest(price~., data=cordata, importance=T, ntree=500)
model2 <- randomForest(price~., data=cordata, importance=T, ntree=1000)
varImpPlot(model1)
```




```
varImpPlot(model2)
```



Try to fit models using top factors

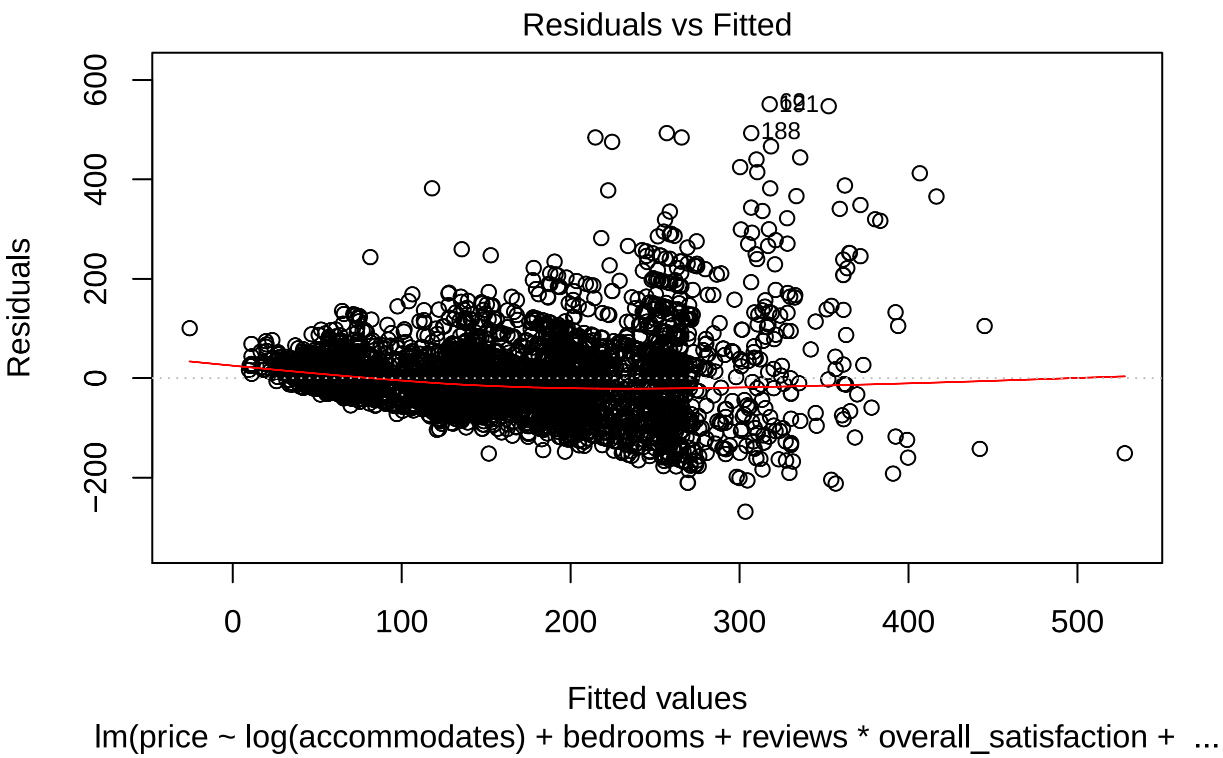
Models

```
fit1 <- lm(price~log(accommodates)+bedrooms+reviews*overall_satisfaction+as.factor(neighborhood)+as.fac
summary(fit1)
```

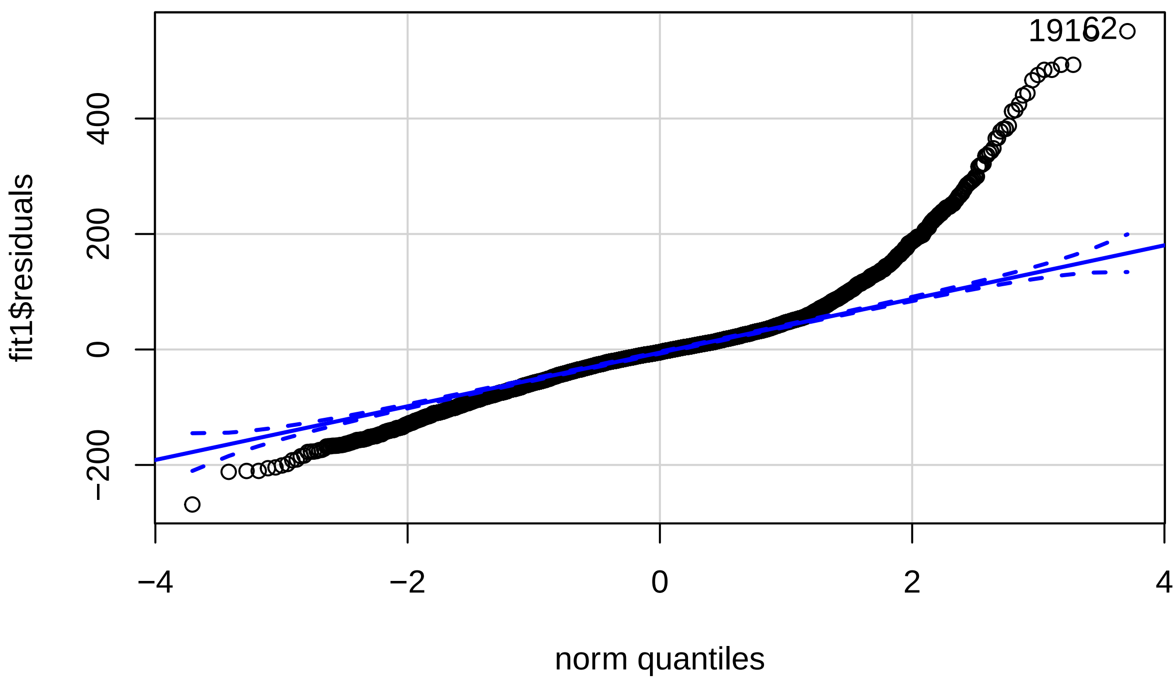
```
##
## Call:
## lm(formula = price ~ log(accommodates) + bedrooms + reviews *
##     overall_satisfaction + as.factor(neighborhood) + as.factor(room_type),
##     data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -268.45  -36.80   -4.81   25.88  551.18
##
## Coefficients:
##                                                 Estimate Std. Error
## (Intercept)                                     66.72498    5.53226
## log(accommodates)                               21.84811    3.07796
## bedrooms                                        46.82777    1.79909
## reviews                                         -1.07118    0.31741
## overall_satisfaction                            -1.50798    0.59976
## as.factor(neighborhood)Back Bay                 82.71440    5.43277
## as.factor(neighborhood)Bay Village              86.36800   13.41012
## as.factor(neighborhood)Beacon Hill              71.28623    6.01118
## as.factor(neighborhood)Brighton                 -7.41221    5.85838
## as.factor(neighborhood)Charlestown              50.47330    7.37905
## as.factor(neighborhood)Chinatown                65.34073    8.43886
## as.factor(neighborhood)Dorchester               -8.04243    5.30369
## as.factor(neighborhood)Downtown                 67.03606    6.14315
## as.factor(neighborhood)East Boston               0.81827    6.05857
## as.factor(neighborhood)Fenway                   57.15089    5.63847
## as.factor(neighborhood)Hyde Park               -20.16416   12.52559
## as.factor(neighborhood)Jamaica Plain            10.29973    5.35887
## as.factor(neighborhood)Leather District        138.44788   22.18337
## as.factor(neighborhood)Longwood Medical Area    45.86323   29.94530
## as.factor(neighborhood)Mattapan                -10.12923   12.51801
## as.factor(neighborhood)Mission Hill             13.91584    7.72606
## as.factor(neighborhood)North End                30.54092    6.57196
## as.factor(neighborhood)Roslindale              -21.91381    8.53485
## as.factor(neighborhood)Roxbury                   8.99643    6.14965
## as.factor(neighborhood)South Boston             41.61851    6.08265
## as.factor(neighborhood)South Boston Waterfront 122.90569    8.83656
## as.factor(neighborhood)South End                73.92047    5.55521
## as.factor(neighborhood)West End                 62.26874   12.74232
## as.factor(neighborhood)West Roxbury            -18.35354   10.35129
## as.factor(room_type)Private room               -56.01553    2.98953
## as.factor(room_type)Shared room                -86.59251    8.33934
## reviews:overall_satisfaction                     0.20649    0.06761
##                                                t value Pr(>|t|)
## (Intercept)                                     12.061  < 2e-16 ***
## log(accommodates)                                7.098 1.45e-12 ***
## bedrooms                                        26.029  < 2e-16 ***
## reviews                                         -3.375 0.000745 ***
```

```
## overall_satisfaction                             -2.514 0.011959 *
## as.factor(neighborhood)Back Bay                  15.225  < 2e-16 ***
## as.factor(neighborhood)Bay Village                6.441 1.31e-10 ***
## as.factor(neighborhood)Beacon Hill               11.859  < 2e-16 ***
## as.factor(neighborhood)Brighton                  -1.265 0.205850
## as.factor(neighborhood)Charlestown                6.840 8.92e-12 ***
## as.factor(neighborhood)Chinatown                  7.743 1.18e-14 ***
## as.factor(neighborhood)Dorchester                -1.516 0.129489
## as.factor(neighborhood)Downtown                  10.912  < 2e-16 ***
## as.factor(neighborhood)East Boston                0.135 0.892570
## as.factor(neighborhood)Fenway                    10.136  < 2e-16 ***
## as.factor(neighborhood)Hyde Park                 -1.610 0.107500
## as.factor(neighborhood)Jamaica Plain              1.922 0.054666 .
## as.factor(neighborhood)Leather District           6.241 4.73e-10 ***
## as.factor(neighborhood)Longwood Medical Area      1.532 0.125696
## as.factor(neighborhood)Mattapan                  -0.809 0.418457
## as.factor(neighborhood)Mission Hill               1.801 0.071742 .
## as.factor(neighborhood)North End                  4.647 3.46e-06 ***
## as.factor(neighborhood)Roslindale                -2.568 0.010272 *
## as.factor(neighborhood)Roxbury                    1.463 0.143557
## as.factor(neighborhood)South Boston               6.842 8.79e-12 ***
## as.factor(neighborhood)South Boston Waterfront  13.909  < 2e-16 ***
## as.factor(neighborhood)South End                 13.307  < 2e-16 ***
## as.factor(neighborhood)West End                   4.887 1.06e-06 ***
## as.factor(neighborhood)West Roxbury              -1.773 0.076282 .
## as.factor(room_type)Private room                -18.737  < 2e-16 ***
## as.factor(room_type)Shared room                 -10.384  < 2e-16 ***
## reviews:overall_satisfaction                      3.054 0.002270 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.31 on 4733 degrees of freedom
## Multiple R-squared:  0.5433, Adjusted R-squared:  0.5403
## F-statistic: 181.6 on 31 and 4733 DF,  p-value: < 2.2e-16
```

```r
plot(fit1,which=1)
```




```r
car::qqPlot(fit1$residuals)
```




```
## [1]  62 191
```

Citation

http://tomslee.net/airbnb-data-collection-get-the-data

http://insideairbnb.com/get-the-data.html