MA678 Midterm Project

Airbnb Analysis in New York

Name: Chaoqun Yin

Prof: Masanao Yajima

12/3/2018

# 1 Abstract

Airbnb is changing people’s habits for accommodation when traveling. And there rises an problem how customers can find the cost-effective accommodation and how owners set the the proper price for their properties.

To address this problem, I set New York as example to analyze what factors have a influence on the prices of the airbnb accommodation. The report are consisted of the parts of **introduction, data preparing and modeling, results, discussion and other additional information**. I mainly utlized multilevel models to conduct regression modeling.

# 2 Introduction

Several days ago when I first came to New York City for traveling, I chose Airbnb as the platform to reserve the accommodation. As for why to choose Airbnb, it is because that shared economy industry are changing the whole world, that is, it provides premium vacation rentals from certified owners - boutique hotels, homes, rooms, and more. Airbnb has 5+ million unique listings worldwide.

However, I am also puzzled that how to choose a reasonable and cost-efficient residence when facing thousands of residences in New York. Therefore, I will conduct a series of analysis to dig into the Airbnb residences to help me and other people that want to find a residence in Airbnb in New York make a decision to choose the proper living place.

# 3 Data Preparing and Modeling

## 3.1 Data Source

I got the data from the site tomslee.net/airbnb-data-collection-get-the-data that is public to everyone. The data contains much information about the residences on some specific days in New York. Here is the data description:

* oom\_id: A unique number identifying an Airbnb listing.
* host\_id: A unique number identifying an Airbnb host.
* room\_type: One of “Entire home/apt”, “Private room”, or “Shared room”
* borough: A subregion of the city or search area for which the survey is carried out. The borough is taken from a shapefile of the city that is obtained independently of the Airbnb web site.
* neighborhood: As with borough: a subregion of the city or search area for which the survey is carried out.
* reviews: The number of reviews that a listing has received. Airbnb has said that 70% of visits end up with a review, so the number of reviews can be used to estimate the number of visits.
* overall\_satisfaction: The average rating (out of five) that the listing has received from those visitors who left a review.
* accommodates: The number of guests a listing can accommodate.
* bedrooms: The number of bedrooms a listing offers.
* price: The price (in $US) for a night stay. In early surveys, there may be some values that were recorded by month.
* minstay: The minimum stay for a visit, as posted by the host.
* latitude and longitude: The latitude and longitude of the listing as posted on the Airbnb site: this may be off by a few hundred metres. I do not have a way to track individual listing locations with
* last\_modified: the date and time that the values were read from the Airbnb web site.

## 3.2 Overall of the data

In this section, I will load the data and check the basic information of the data. Then I will use some visualization to show how the residences’ features distribute. Based on the findings observing the visualization, I will choose the proper models to conduct analysis.

### Prepare the data

During this part, I load the data and it has 430783 rows and 16 columns.

0 ### Data Structure During this part, I check the data structure and data type of each variables using str().

str(ny)

## 'data.frame': 430783 obs. of 16 variables:  
## $ room\_id : int 3557831 7780413 1788989 9680791 8204392 7455635 1742471 2045660 9603803 3162817 ...  
## $ host\_id : num 17911953 4887492 7607092 3201897 11268108 ...  
## $ room\_type : Factor w/ 4 levels "","Entire home/apt",..: 3 4 2 3 3 2 2 3 3 2 ...  
## $ borough : Factor w/ 5 levels "Bronx","Brooklyn",..: 3 3 3 2 2 3 4 2 2 2 ...  
## $ neighborhood : Factor w/ 240 levels "Allerton","Arden Heights",..: 65 199 180 162 217 34 191 154 13 217 ...  
## ......

We can get from the output that these variables like **room\_type, borough, neiborhood** are factors having different level. Therefore, I can put these variables as category into my model. Then as for numerical variables such as **reviews, overall\_satisfaction, accommodates, bedrooms**, they are discrete data rather than consecutive data. So they also can be category variables in my model.

### Data overview

During this part, I use summary() to get the overview of the dataset.

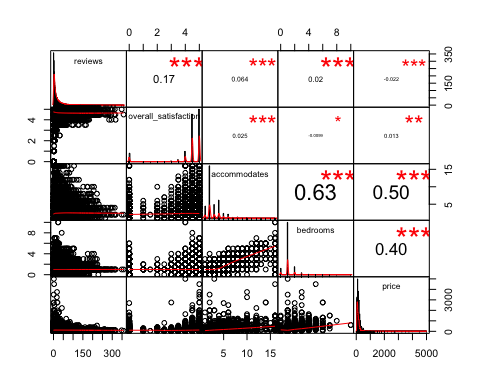
summary(ny)

## room\_id host\_id room\_type   
## Min. : 105 Min. : 43 : 0   
## 1st Qu.: 3907848 1st Qu.: 4071690 Entire home/apt:220697   
## Median : 7708956 Median : 13972066 Private room :197044   
## Mean : 7488534 Mean : 22168318 Shared room : 13042   
## 3rd Qu.:10686406 3rd Qu.: 35322092   
## Max. :16544971 Max. :108817370   
##   
## borough neighborhood reviews   
## Bronx : 6342 Williamsburg : 38146 Min. : 0.00   
## Brooklyn :174455 Bedford-Stuyvesant: 27952 1st Qu.: 1.00   
## Manhattan :209213 Harlem : 25489 Median : 3.00   
## Queens : 38367 Upper West Side : 20398 Mean : 13.75   
## Staten Island: 2406 East Village : 19989 3rd Qu.: 15.00   
## Bushwick : 19643 Max. :387.00   
## (Other) :279166   
## ......

We can see from the output that these variables may have correlation with the price of each listing: **room\_type, borough, neighborhood, reviews, overall\_satisfaction, accommodates, bedrooms**. It is a suppose and I will explore the correlation of variables mainly with the prices in the exploratory data analysis and based the outcome to help choose proper model to analyze the price.

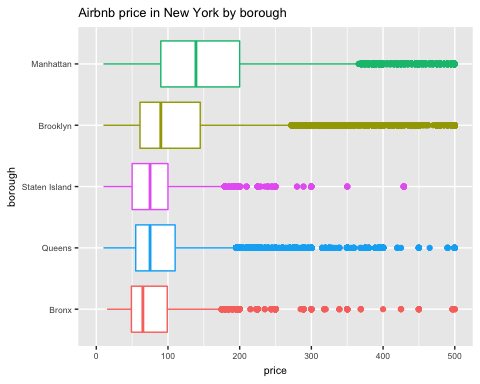
## 3.3 Exploratory Data Analysis

### The correlation plot



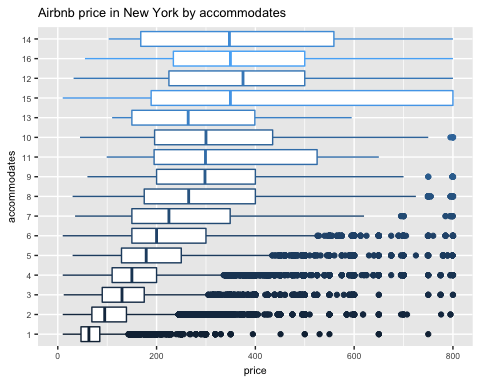
From the correlation plot, firstly we can get the distribution of the five variables on the diagonal. Then we can see the bivariate scatter plots with a fitted line on the bottom of the diagonal that there are correlation between these variables. Finally, we can see the numbers on the top of the diagonal that the correlation and the significant level for price are high. Therefore, I consider to put these variables into my model to predict the accommodation price.

### The boxplot of borough and the prices

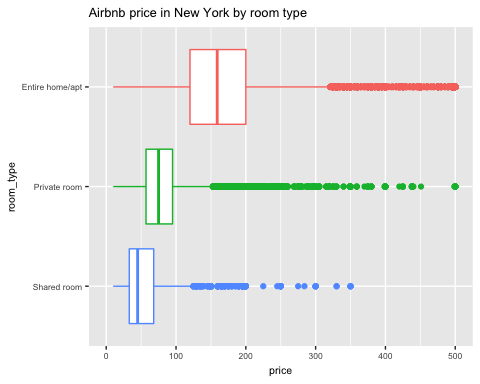


We can see from the boxplot that the median price of each borough has *obvious difference*. Manhattan borough has the highest median price rather than other borough while Bronx has the lowest median price. So we can infer from the plot that *borough* is a feature that influences the price of one property.

### Boxplot of accommodates and the prices

 Similarly, we can see from the boxplot that the median price of each number of accommodates has obvious difference. The accommodates of 14 has the highest median price rather than other number while accommodation of 1 has the lowest median price. So we can infer from the plot that *accommodation* is a feature that influences the price of one property.

### Boxplot of the room types and the price

 Similarly, we can see from the boxplot that the median price of each type of room has *obvious difference*. The entire apartment room type has the highest median price rather than other types while the shared room has the lowest median price. So we can infer from the plot that *room\_type* is a feature that influences the price of one property.

## 3.4 Model Used

**The multilevel model is chosen to analyze the prices.**

### Model 1

* Regress price on accommodates and bedrooms and treat all other intercepts as random.

fit.1 <- lmer(data = ny.m, price ~ accommodates + bedrooms + (1 | Date) +

(1 | room\_type) + (1 | borough), REML = FALSE)  
display(fit.1)

## lmer(formula = price ~ accommodates + bedrooms + (1 | Date) +   
## (1 | room\_type) + (1 | borough), data = ny.m, REML = FALSE)  
## coef.est coef.se  
## (Intercept) -0.89 26.31   
## accommodates 16.02 0.40   
## bedrooms 44.37 0.93   
##   
## Error terms:  
## Groups Name Std.Dev.  
## Date (Intercept) 1.75   
## borough (Intercept) 29.14   
## room\_type (Intercept) 39.45   
## Residual 94.34   
## ---  
## number of obs: 43078, groups: Date, 11; borough, 5; room\_type, 3  
## AIC = 514056, DIC = 514042.4  
## deviance = 514042.4

### Model 2

* Based on model 1, include a between-group correlation between accommodates and borough.

fit.2 <- lmer(data = ny.m, price ~ accommodates + bedrooms + (1 | room\_type)  
 + (1 | Date) + (1 + bedrooms | borough), REML = FALSE)  
display(fit.2)

## lmer(formula = price ~ accommodates + bedrooms + (1 | room\_type) +   
## (1 | Date) + (1 + bedrooms | borough), data = ny.m, REML = FALSE)  
## coef.est coef.se  
## (Intercept) 19.76 24.12   
## accommodates 16.33 0.39   
## bedrooms 25.53 11.17   
##   
## Error terms:  
## Groups Name Std.Dev. Corr   
## Date (Intercept) 1.81   
## borough (Intercept) 16.12   
## bedrooms 24.44 -0.25   
## room\_type (Intercept) 39.54   
## Residual 93.44   
## ---  
## number of obs: 43078, groups: Date, 11; borough, 5; room\_type, 3  
## AIC = 513254, DIC = 513236.5  
## deviance = 513236.5

### Model 3

* Then, I regress price on accomodates, room type, reviews and bedrooms while including one between-group and other variables as random effect.

fit.3 <- lmer(data = ny.m, price ~ accommodates + reviews + room\_type +

bedrooms + (1 | Date) + (1 | neighborhood) + (1 + accommodates | borough), REML = FALSE)  
display(fit.3)

## lmer(formula = price ~ accommodates + reviews + room\_type + bedrooms +   
## (1 | Date) + (1 | neighborhood) + (1 + accommodates | borough),   
## data = ny.m, REML = FALSE)  
## coef.est coef.se  
## (Intercept) 62.27 3.80   
## accommodates 10.78 4.19   
## reviews -0.14 0.01   
## room\_typePrivate room -52.56 1.05   
## room\_typeShared room -75.44 2.70   
## bedrooms 44.23 0.88   
##   
## Error terms:  
## Groups Name Std.Dev. Corr   
## neighborhood (Intercept) 27.18   
## Date (Intercept) 2.16   
## borough (Intercept) 5.39   
## accommodates 9.27 0.78   
## Residual 88.63   
## ---  
## number of obs: 43078, groups: neighborhood, 220; Date, 11; borough, 5  
## AIC = 509027, DIC = 509003.4  
## deviance = 509003.4

ranef(fit.3)

## $neighborhood  
## (Intercept)  
## Allerton 4.42680713  
## Arden Heights -18.87761479  
## Arrochar -10.34137213  
## Arverne 19.33899013  
## Astoria 10.66402955  
## ……

##   
## $Date  
## (Intercept)  
## 2016-01-01 0.23755858  
## 2016-02-01 -3.15536992  
## 2016-03-01 -1.73711019  
## 2016-04-01 -0.03904789  
## 2016-05-01 0.20893168  
## 2016-06-01 -0.12305216  
## ……

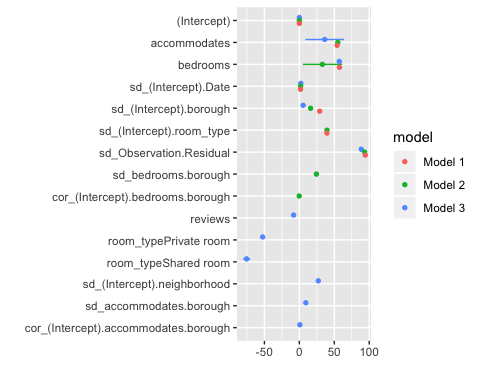
##   
## $borough  
## (Intercept) accommodates  
## Bronx -3.166484 -9.9028010  
## Brooklyn -1.465529 1.1980235  
## Manhattan 9.043088 16.3226865  
## Queens -2.518231 -7.5113859  
## Staten Island -1.892845 -0.1065231

# 4 Result

## 4.1 Model Choice

### Coeficient plot

dwplot(list(fit.1, fit.2, fit.3), show\_intercept = TRUE)



From the coefficient plot, I can easily find whether the coefficients of the three models are significant or not. And most of coeficients of model 3 are significant.

### ANOVA

anova(fit.1, fit.2)

## Data: ny.m  
## Models:  
## fit.1: price ~ accommodates + bedrooms + (1 | Date) + (1 | room\_type) +   
## fit.1: (1 | borough)  
## fit.2: price ~ accommodates + bedrooms + (1 | room\_type) + (1 | Date) +   
## fit.2: (1 + bedrooms | borough)  
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)   
## fit.1 7 514056 514117 -257021 514042   
## fit.2 9 513254 513332 -256618 513236 805.91 2 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(fit.1, fit.3)

## Data: ny.m  
## Models:  
## fit.1: price ~ accommodates + bedrooms + (1 | Date) + (1 | room\_type) +   
## fit.1: (1 | borough)  
## fit.3: price ~ accommodates + reviews + room\_type + bedrooms + (1 |   
## fit.3: Date) + (1 | neighborhood) + (1 + accommodates | borough)  
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)   
## fit.1 7 514056 514117 -257021 514042   
## fit.3 12 509027 509131 -254502 509003 5039 5 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(fit.2, fit.3)

## Data: ny.m  
## Models:  
## fit.2: price ~ accommodates + bedrooms + (1 | room\_type) + (1 | Date) +   
## fit.2: (1 + bedrooms | borough)  
## fit.3: price ~ accommodates + reviews + room\_type + bedrooms + (1 |   
## fit.3: Date) + (1 | neighborhood) + (1 + accommodates | borough)  
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)   
## fit.2 9 513254 513332 -256618 513236   
## fit.3 12 509027 509131 -254502 509003 4233 3 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(fit.1, fit.2, fit.3)

## Data: ny.m  
## Models:  
## fit.1: price ~ accommodates + bedrooms + (1 | Date) + (1 | room\_type) +   
## fit.1: (1 | borough)  
## fit.2: price ~ accommodates + bedrooms + (1 | room\_type) + (1 | Date) +   
## fit.2: (1 + bedrooms | borough)  
## fit.3: price ~ accommodates + reviews + room\_type + bedrooms + (1 |   
## fit.3: Date) + (1 | neighborhood) + (1 + accommodates | borough)  
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)   
## fit.1 7 514056 514117 -257021 514042   
## fit.2 9 513254 513332 -256618 513236 805.91 2 < 2.2e-16 \*\*\*  
## fit.3 12 509027 509131 -254502 509003 4233.05 3 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

From ANOVA, the result shows a Df of 3 comparing model 3 to model 1, and a small p-value. This means adding the 3 variables to the models does lead to a significantly improved fit. Also, the AIC of model 3 is a little smaller than the model 1 and model 2.

Therefore, the final model choosen is model 3, a multilevel model varying both intercepts and slopes.

## 4.2 Interpretation

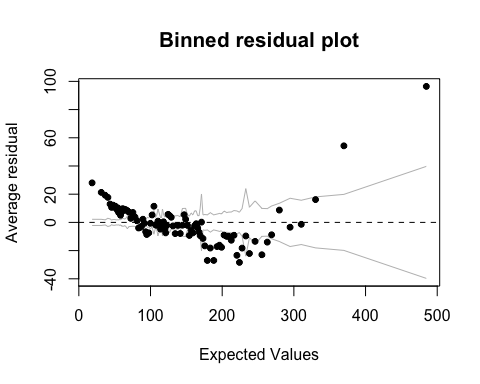
Because it is a multilevel model, I choose one circumstance to interprete. For example, when it is in January and located in Manhattan district near Manhattan Beach, and the price can be expressed as this formula:

**Interpretation:** When the accommodate is 0, the number of bedrooms is 0, the reviews are 0 and the room type is apartment, the price of the commodation is 52.73 (the accommodate cannot be 0). When accommodates increase by 1 person while controling other variables, the price will increase by 27.10. When the number of reviews increase by 1 while controling other variables, the price will decrease by 0.14. When the number of bedrooms increases by 1 while controling other variables, the price will increase by 44.23. If the room type is private room while controling other variables, the price will decrease by 52.56. If the room type is shared room while controling other variables, the price will decrease by 75.44.

As for the standard error and standard deviation of the variable **accommodates**, the *standard error* is 4.19, which tells us the precision of the “10.78” estimate: we are essentially certain that more accommodates lead to higher price. The *estimated standard* deviation of the slope is 9.27, which implies that the individual slopes vary greatly.

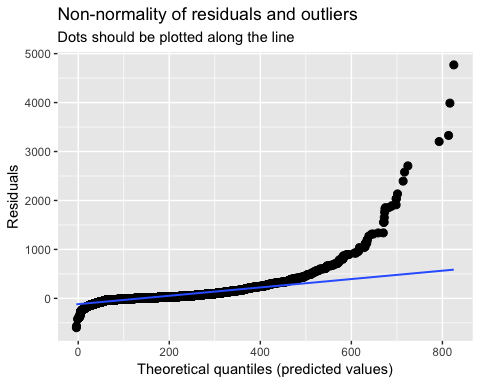
## 4.3 Model Check

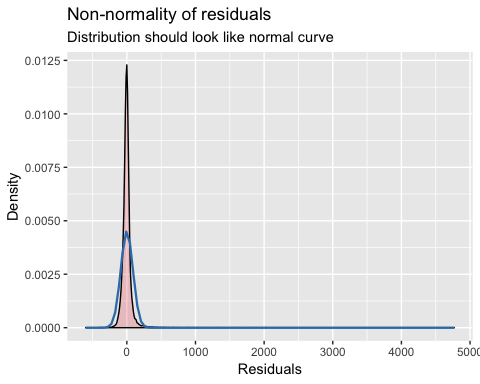
### Binned plot

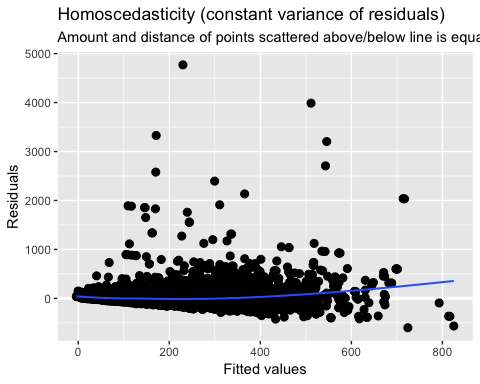


From the binned plot, we can see the average residual distributed versus expected values.

### Diagnostic plots



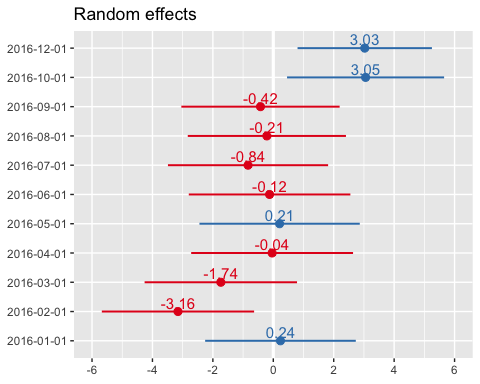


 Observing the dianostic plots, we can get a lot of information from it. For the distribution of the residuals, it is non-normality distributed. And from the residuals versus predicted values plot, they seem align with the line. But when the predicted values increase, the points are more far from the line. Similar is the residuals versus fitted values plot. There are also some plots describing the random effect quantiles versus standard normal quantiles.(put in the appendix)

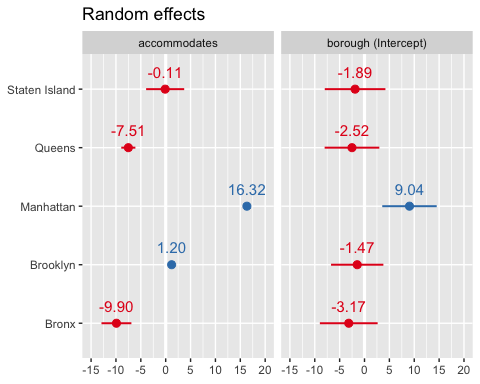
### Random effect estimate plot

plot\_model(fit.3, type = "re", show.values = TRUE, value.offset = .3)

## [[1]]



##   
## [[2]]



We can check the random effect coeficients and the significance of them in this plot. It seems that many estimates are not significant.

# 5 Discussion

## 5.1 Implication

* From the model I give a try to predict the prices of the accommodation in airbnb, and the prices can be a reference for the people who want to reserve the accommodation in airbnb.
* The variables involved in the model are general features of the accommodation, but other features like the decoration, the furnitures, the area and so on can be important to predict the prices. Therefore, this can be a reason why the model I used is not that good.

## 5.2 Limitation

* Continuing the implication, the dataset lacks necessary features of the accommodation to get better fitting model.
* The errors are mroe distributed around the higher prices of the accommodation. Therefore I can clarify all the accommodation based their prices to get a better fitting model.

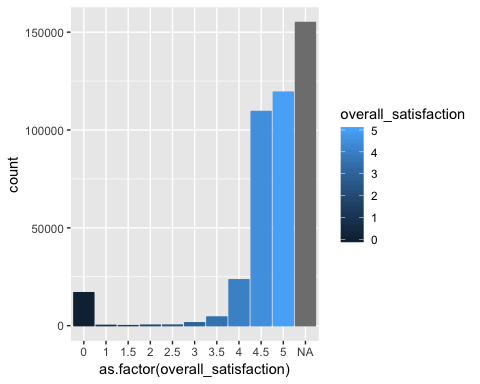
## 5.3 Future Direction

* I will try to use Zillow API to get more information of the accommodation to bring in more features. By doing this the model can get better fitting.
* I will try to use deep learning to fit models to make prediction.

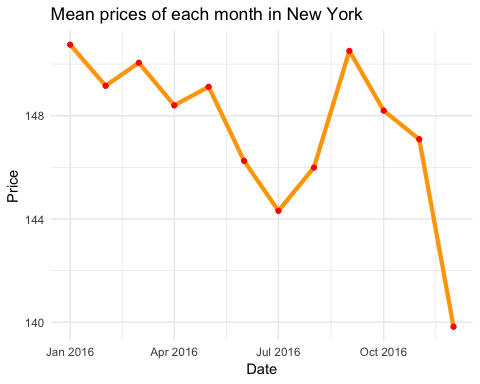
# 6 Appendix

## More EDA

### Diagram of reviews

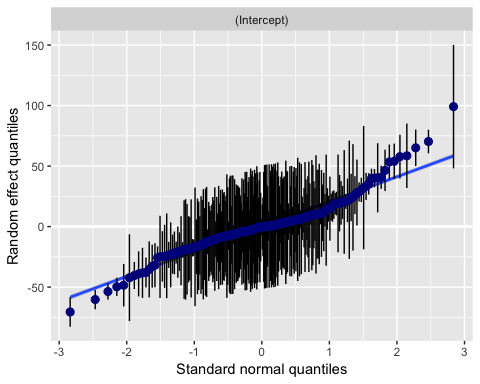


### Mean prices of each month in New York

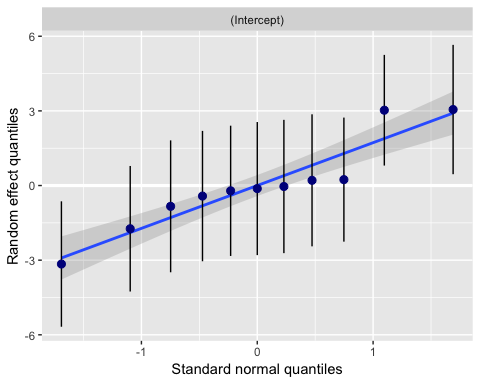


## Diagnostic Plots of Model 3

### Random Effect neighborhood



**Date**



**borough**

