MA 678 Applied Statistical Modeling

Final Project:

Model to predict the Airbnb prices in New York

Professor: Luis Carvalho

Date: December 6, 2022

Author: Yueling Feng

Table of content

Abstract

The whole analysis is based on the dataset of the Airbnb market in New York in the year 2019. [1] This study generally explores the factors which might influence the price of Airbnb in New York and creates a multilevel model to predict the New York Airbnb price market. The dataset is first being cleaned to make sure that there are no outliers and missing values which might affect the future analysis. After that, Exploratory Data Analysis is implied to find the variables that might influence the New York City Airbnb price. Using those meaningful variables, multilevel models are being created and the assumptions of the multilevel model are checked through the use of ggplots.

Introduction

The dataset Airbnb market in New York in the year 2019 contains 16 columns and 48,895 rows. The 16 columns include the Airbnb id, Airbnb name, host id, host name, neighborhood group, neighborhood, latitude, longitude, room type, price, minimum rent nights, total number of reviews, last review date, reviews per month, calculated host listing count and the available days per year.

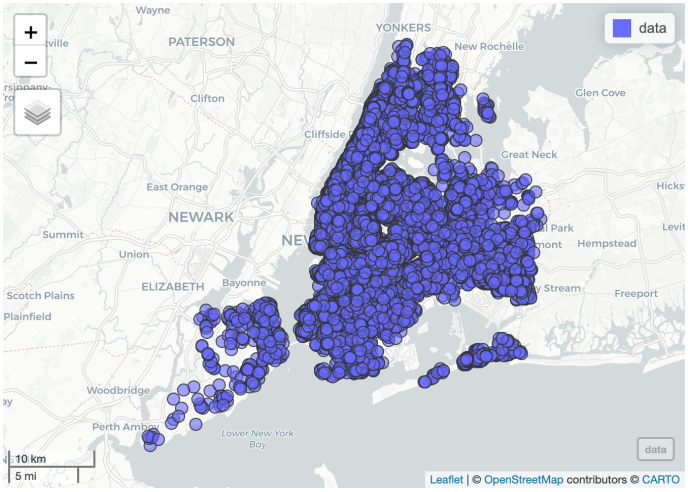


Figure1: the distribution of Airbnb in New York, 2019

The Airbnb listings are generally in five neighborhood groups: Manhattan, Brooklyn, Bronx, Staten Island and Queens. The distribution map of Airbnb in New York (Figure 1)shows that

most Airbnb listings are located in Manhattan, Brooklyn, and Queens. As a result, the model created might be more representative of those three locations.

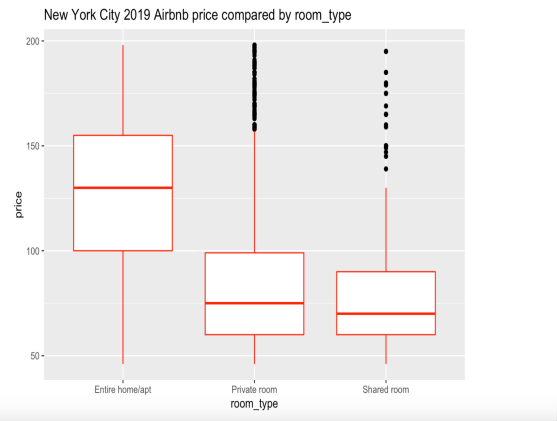


Figure 2: boxplot compare room type and price

Figure 3: ggplot compare room type and price

The boxplot and ggplot comparing room type and price (Figure 2 and Figure 3) illustrate that there is a strong correlation between the room type and Airbnb price in New York. Especially for the entire room type, the price follows a totally different distribution. For better analyzing the data, the factor room type is being transferred to numeric value (Shared room = 1, Private room = 2 , Entire home/apt = 3) as the price has an increasing trend from shared room to the Entire room/ apartment.

Looking through the dataset, there are two groups which are noticeable: availability and minimum nights. Not all apartments are available all year . As a result, for each apartment which has available days less than 365, the available time is uncertain and might vary a lot. However, the Airbnb price might be influenced by other factors such as the market prices. So it can not be concluded that each sample is independent. For the total number of minimum nights, the total minimum rent nights number which is small may have a high variation of price due to the same reason. As a result, the samples within these two groups are not independent and there might be some random effects within the groups.

Moreover, for better predicting the prices to see if there are random effects within the groups total availability and minimum rent nights, the total available days factor is being transferred into total available months and the minimum rent nights factor is being transferred to 14 levels.

Through the use of visualization and linear regression analysis, factors which might affect the prices in Airbnb Market in New York are being found. They are: minimum rent nights, number of reviews per month and room type. There is also an interactive effect being found between reviews per month and room type.

Method

1.Data cleaning

A distribution map is being created to make sure that all the Airbnb apartments in the dataset are located in New York by looking at the points created on the map.

While plotting the distribution of the prices for Airbnb Market in New York, there are a lot of outliers being found. The outliers are being eliminated by calculating the quantile of price and removing the sample which has a price not within the quantile.

There are also many samples which have a price and available days of 0. This implies that actually those Airbnbs are not available. As a result, all the samples which have a price and available days of 0 or NA values are being removed from the dataset to ensure that all the remaining data contains available Airbnb apartments.

Also, the factor room type is being transformed to numeric values of 1 to 3, the factor available days is being transformed into available months and the factor minimum nights is being transferred to 14 levels for better study of the data.

2.Exploratory Data Analysis

The EDA is generally used to see whether there are correlations between those factors and price. For each comparison which includes the categorical data, a boxplot and a ggplot are being created to see whether the factors are correlated. The anova and Kruskal-Wallis test[2] are also involved to see if there is difference between the means within the groups of the categorical data. For numeric factors, a linear regression model is being fitted to see if there is correlation between the factor and the price. Also, a fit vs residual plot is used to check the effectiveness of the regression model.

All the factors which are found to be correlated with price are being chosen to include in the model. After selecting the variables, whether there is correlation between these variables are also analyzed to see if there are interactive effects between those variables.

3.**Multilevel Model Fitting**

Plots are being created using ggplot2 to see if there are random effects within the groups available month and minimum rent nights. After choosing the more effective group, ggplot2 plots graphs to see the correlation between the factors and the price of Airbnb in New York for each group level.

Two models are being fitted: one contains the interactive effect and the other does not. For both models, the assumptions of the multilevel model are checked to see whether the model fits the requirement.[3] For each multilevel model, the linearity, homogeneity of variance, and the normality assumption are being checked using the plots. The residual plot is used to confirm the linearity of the model, the model plot is used to check the homogeneity of variance and the Q-Q plot is used to ensure the normality of the model.

Result

Neighborhood group

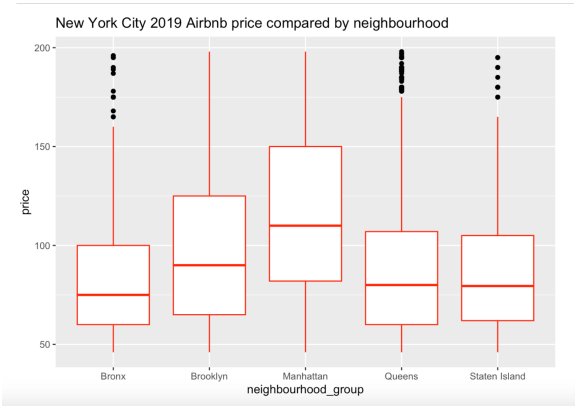


Figure 4

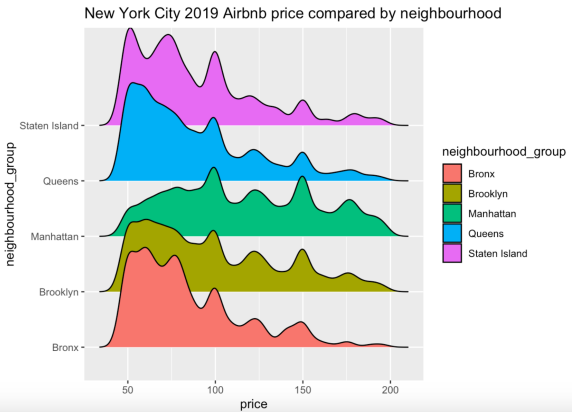


Figure 5

Figure 4 and figure 5 are used to see the distribution of Airbnb prices in New York for different neighborhood groups. Although the ANOVA and Kruskal-Wallis test implies that there is a difference in mean price for various neighborhood groups.The graph states that other than Manhattan, the other four neighborhood groups generally follow a similar pattern and mean. As a result, the neighborhood group cannot be counted into a meaningful variable.

Room type

Figure 2 and 3 illustrate that there is a difference in mean price for different room types and the ANOVA and Kruskal-Wallis test confirm the assumption. By turning the room type into numeric value, the regression model shows that the price will increase if the room type increases from shared room to entire room.

Available month

Fitting the regression model, one unit increase in available month will only result in 0.1156 increase value of price. The p-value is greater than 0.05 which implies that there is almost no correlation between available month and price. So the factor available month is being removed from the multilevel model and it is also not considered as a group which might have random effects.

Minimum nights

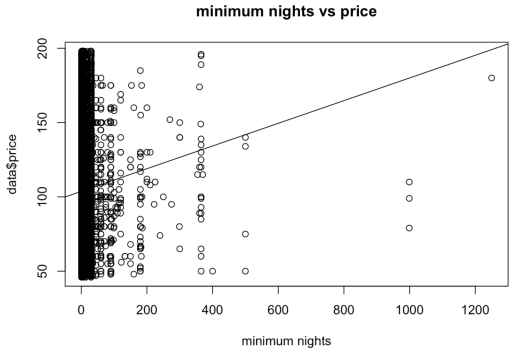


Figure 6 : minimum night vs price

The regression model implies that one day increase in minimum rent nights will increase the price by 0.08. Although the value is small, the maximum minimum rent nights is 1200 which will cause a very influential impact on the price. After turning minimum rent nights into 14 levels, one level increase in minimum rent nights will cause a 2.23 increase in Airbnb price in

New York and the p-value is smaller than 0.05 which means beta is not equal to 0. As a result, there is a correlation between the minimum rent nights and price of Airbnb in New York.

Total number of reviews and Reviews per month

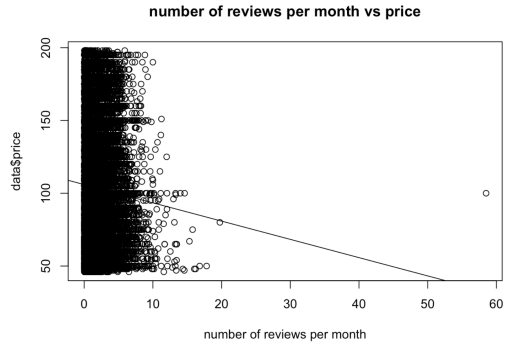


Figure 7: review per month vs price

No correlation is being found between total number of reviews and price. However, there is correlation being found between reviews per month and price. As the regression model has a p-value smaller than 0.05. This might be because all Airbnb apartments are open for different amounts of days, so the total number of reviews is meaningless. However, the reviews per month represent the popularity of the Airbnb apartments as it shows how many people book this Airbnb apartment each month.

Interactive Effects & Potential groups

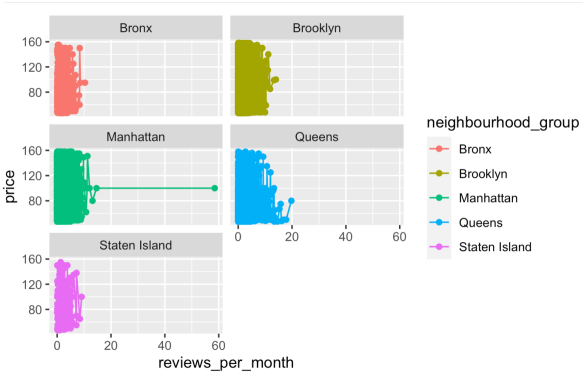


Figure 8: review per month vs price grouped by neighborhood

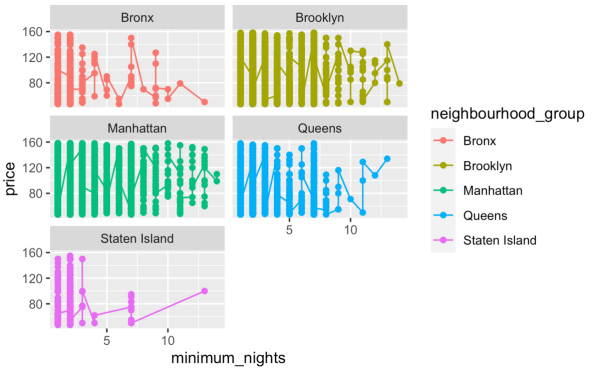


Figure 9: minimum nights vs price group by neighborhood

Figure 8 and 9 shows that there is no difference in pattern for reviews per month and minimum nights compared with price in Airbnb in New York when dividing the data into different

neighborhoods. As a result, the factor neighborhood group can be excluded from the multilevel model.

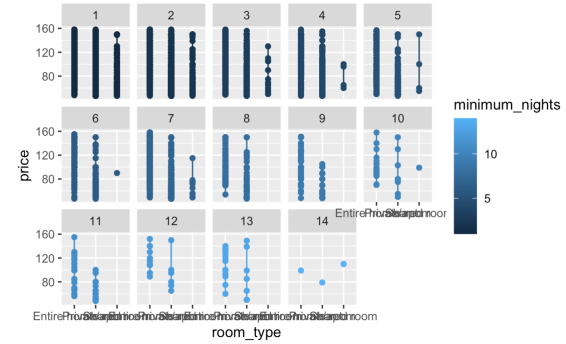


Figure 10: room type vs price grouped by minimum night

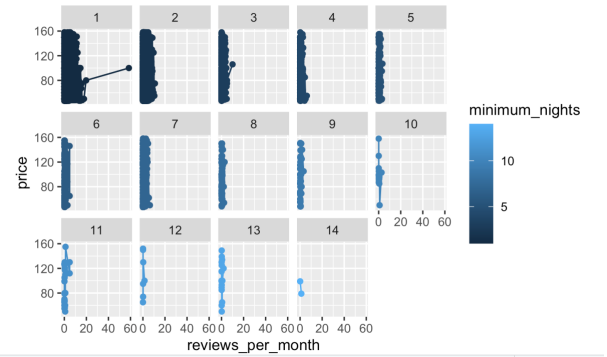


Figure 11: review per month vs price group by minimum nights

Figure 10 and 11 shows there is small difference in correlation within the groups when the data is grouped by minimum nights. As a result, the random effect within the minimum night group might not be big in the multilevel model.

Comparing the factors which will be included in the model. Interactive effects are being found between minimum rent nights and number of reviews. Because minimum rent nights might influence the number of reviews. If the minimum number of rent days is high, the number of reviews will be small since the Airbnb apartment will have only a few guests.

Multilevel Model & Model checking

1.Excluding Interactive Effects

The model includes the factors minimum nights, reviews per month, room type , and the random effect grouped by minimum nights.

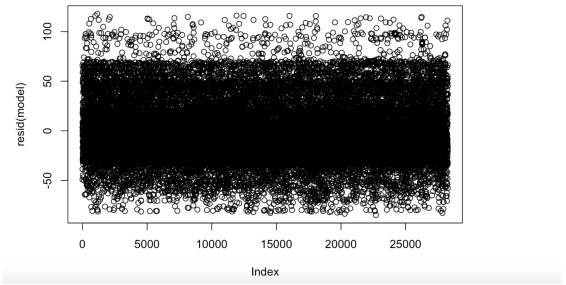


Figure 12: residual plot of model 1

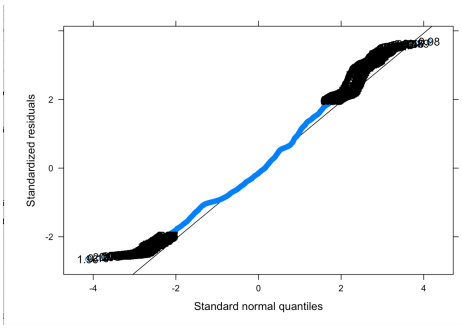


Figure 13: QQ plot for model 1

The QQ plot (Figure 13) and residual plot (Figure 12) generally confirm the linearity and normality of the model. However, the homogeneity of variance is not confirmed. Although the residuals are randomly distributed around 0, there is loss of points between price 90 to 110.

2.Including Interactive Effects

The model includes the factors minimum nights, reviews per month, room type , the interactive effect between minimum nights and reviews per month, and the random effect grouped by minimum nights.

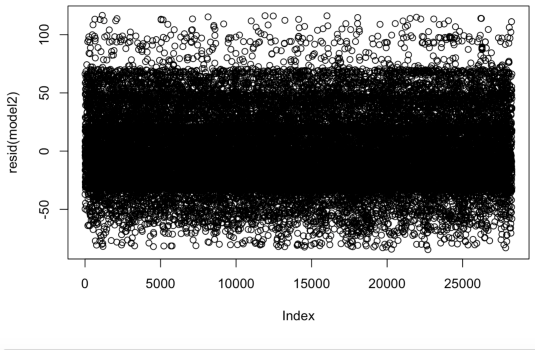


Figure 14: residual plot of model 2

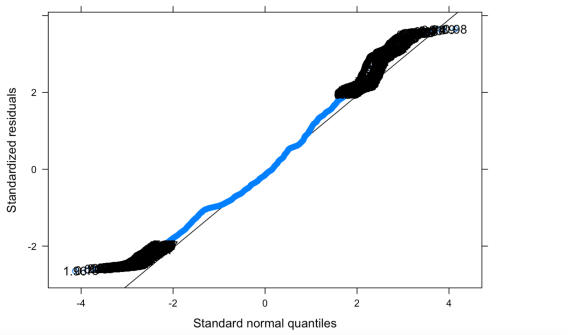


Figure 15: QQ plot for model 2

The QQ plot (Figure 15) and residual plot (Figure 14) generally confirm the linearity and normality of the model. But the new model generally does not improve the linearity. However, the model helps to improve the homogeneity of variability as the points are more normality distributed in the fit and residual plot for model 2.

Discussion

There are two models being created, one contains the interactive effect and the other does not. The result shows that other than confirmation of assumptions of linearity and normality of the model, the homogeneity of variance is being improved while adding interactive effects between minimum nights and reviews per month.

For the model, y equals to the price of Airbnb in New York, x variables include minimum nights, room type, reviews per month, minimum nights*reviews per month and the random effects

within the group minimum nights. The factor room type is divided into room type 1(shared room), 2(private room), and 3(Entire room) and are binary (no= 0, yes= 1).

Price= 79.92 – 1.08*minimum nights – 2.53*reviews per month + 1.2* room type 2 +49.64*room type 3 + 2.3*minimum nights*room type 2 + 1.3*minimum nights*room type 3

There is an intercept of variance of 18.97 for the random effect within the group minimum rent nights. And the mixed effect model shows that the room type has a huge influence on price of Airbnb in New York especially for renting entire room. Renting the entire room in Airbnb generally increases the price by 49.64 dollars.

However, this study has two limitations. Firstly, although the factor neighborhood is not included in the dataset since the four neighborhoods except Manhattan follow a similar pattern and mean, Manhattan does have an obvious difference in mean and pattern compared with those four neighborhoods. Future analysis might need to see how to put neighborhood as a factor into the model.

Also, after 2019, the covid-19 pandemic exists. This implies that the Airbnb price market might be shocked a lot. As a result, based on the influence of market prices due to consequences such as covid-19. The model might not be useful while predicting the recent Airbnb price in New York City.

What's more, the reviews per month also implies the popularity of the Airbnb apartments, so the future study can be done to see which factors might affect the popularity of the Airbnb market in New York and a model might be fitted to predict the popularity.

Reference

1.*https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data/code?resource=download*

2.*https://bookdown.org/daniel_dauber_io/r4np_book/comparing-groups.html*

3.*https://ademos.people.uic.edu/Chapter18.html*

Appendix

Plots

Figure 16: available month vs price

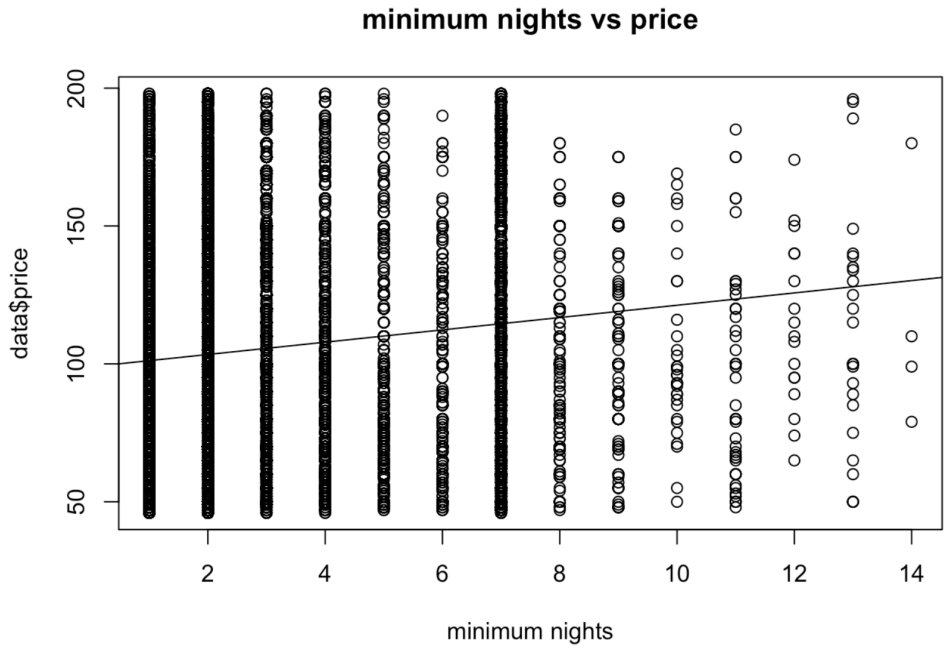


Figure 17: minimum night vs price

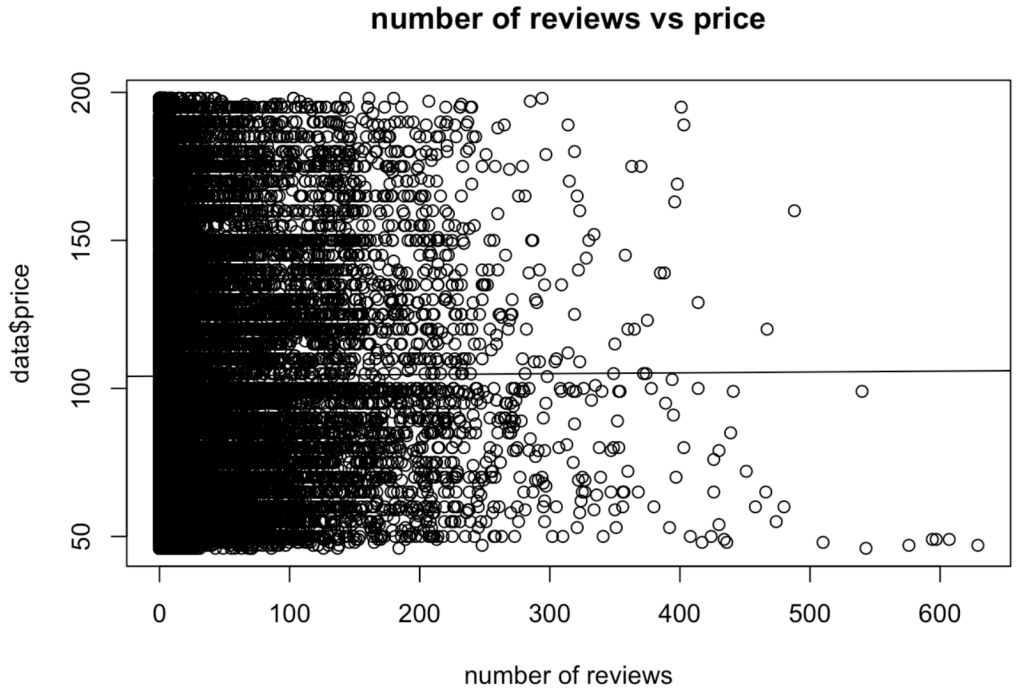
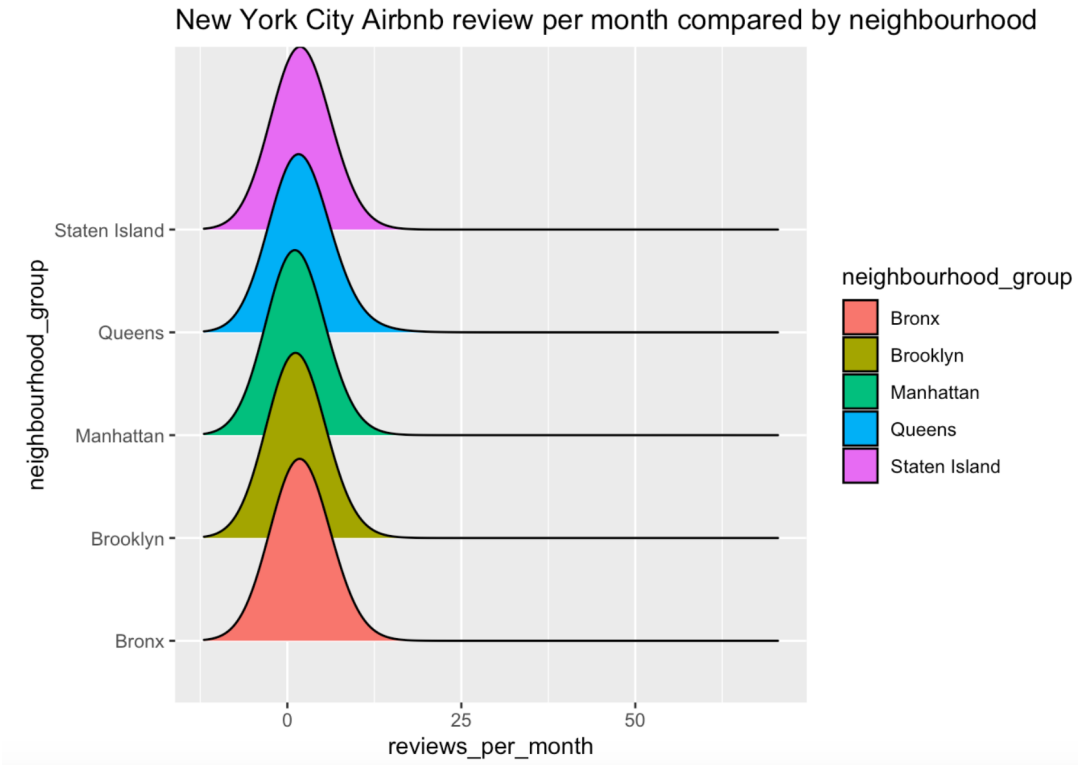
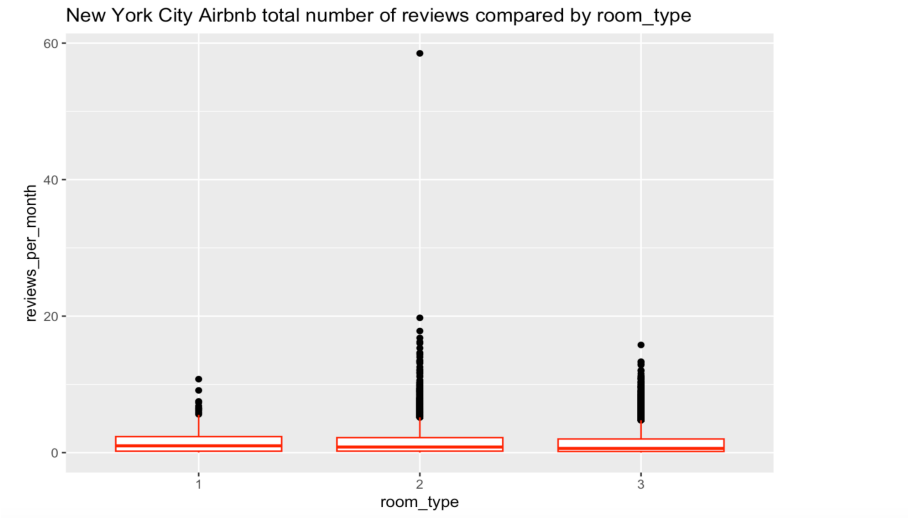


Figure 18: number of reviews vs price


New York City Airbnb review per month compared by neighbourhood
neighbourhood_group
Staten Island
Queens
Manhattan
Brooklyn
Bronx
0
25
50
reviews_per_month
neighbourhood_group
Bronx
Brooklyn
Manhattan
Queens
Staten Island


New York City Airbnb total number of reviews compared by room_type
60
40
20
0
reviews_per_month
1
2
3
room_type

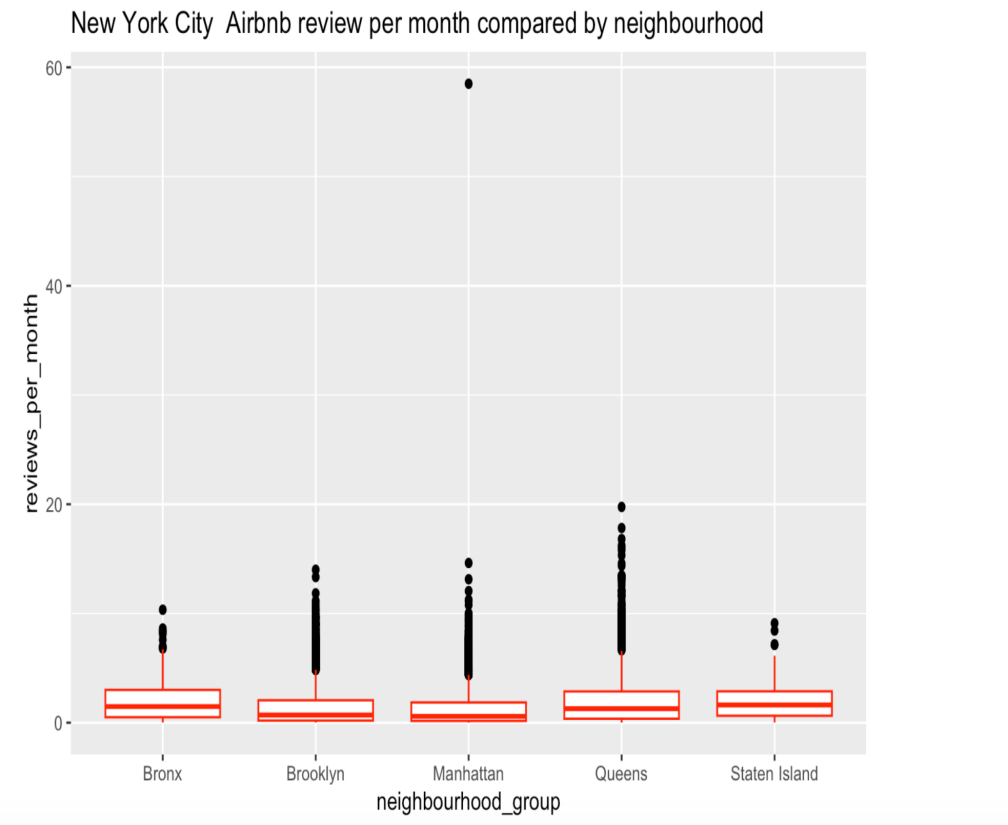


Figure 19: review per month compared by neighborhood group

Figure 20: review per month distribution compared by neighborhood group

Figure 21: total number of reviews compared by room type

Figure 22: total number of reviews distribution compared by room type

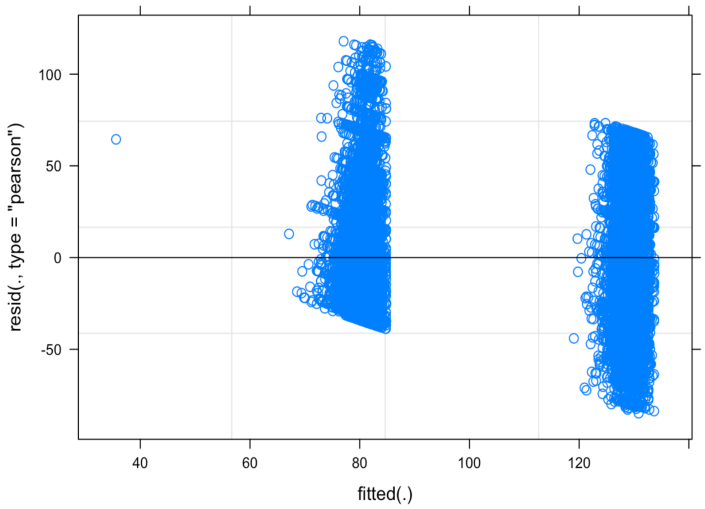


Figure 23: fit vs residual plot for model 1

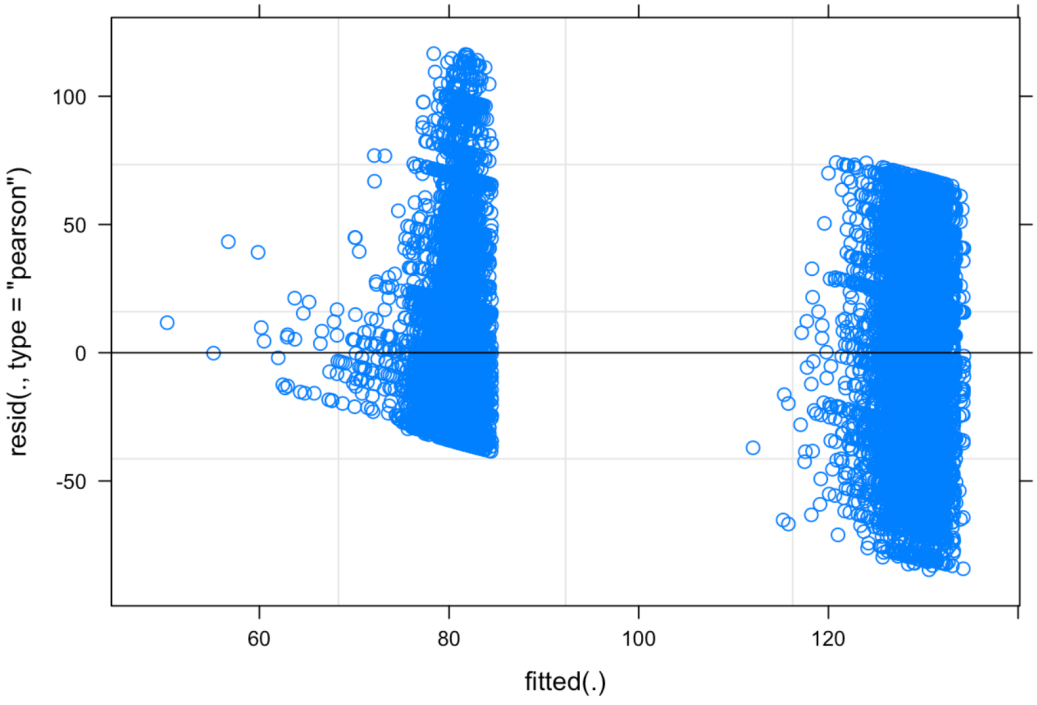


Figure 24:Fit vs Residual plot for model 2

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: price ~ minimum_nights + reviews_per_month + room_type + (1 |
##     minimum_nights)
##    Data: data
##
## REML criterion at convergence: 271254.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6291 -0.7332 -0.1390  0.6489  3.6571
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  minimum_nights (Intercept)  19.25    4.387
##  Residual                    970.94   31.160
## Number of obs: 27916, groups:  minimum_nights, 14
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        76.8858     3.2956  23.330
## minimum_nights     -1.0472     0.4544  -2.304
## reviews_per_month  -0.6845     0.1137  -6.021
## room_type2          4.8044     1.6267   2.954
## room_type3         51.9093     1.6375  31.700
##
## Correlation of Fixed Effects:
##             (Intr) mnmm_n rvws__ rm_ty2
## minmm_nghts -0.747
## rvws_pr_mnt -0.042  0.029
## room_type2  -0.481 -0.002 -0.011
## room_type3  -0.479 -0.009 -0.015  0.972
```

Figure 25: multilevel model1 results

```
## price ~ minimum_nights + reviews_per_month + room_type + reviews_per_month *
##     room_type + (1 | minimum_nights)
##    Data: data
##
## REML criterion at convergence: 271231.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6532 -0.7310 -0.1456  0.6324  3.6789
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  minimum_nights (Intercept)  18.97    4.355
##  Residual                    970.24   31.149
## Number of obs: 27916, groups:  minimum_nights, 14
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                     79.9156     3.5863  22.284
## minimum_nights                  -1.0769     0.4524  -2.381
## reviews_per_month               -2.5303     0.9024  -2.804
## room_type2                       1.1956     2.1949   0.545
## room_type3                      49.6366     2.2046  22.515
## reviews_per_month:room_type2     2.2596     0.9132   2.474
## reviews_per_month:room_type3     1.3111     0.9183   1.428
##
## Correlation of Fixed Effects:
##             (Intr) mnmm_n rvws__ rm_ty2 rm_ty3 r__:_2
## minmm_nghts -0.682
## rvws_pr_mnt -0.406  0.003
## room_type2  -0.598 -0.001  0.662
## room_type3  -0.595 -0.008  0.658  0.974
## rvws_pr_:_2  0.403 -0.001 -0.987 -0.672 -0.654
## rvws_pr_:_3  0.400  0.003 -0.982 -0.653 -0.670  0.970
```

Figure 26: multilevel model2 results