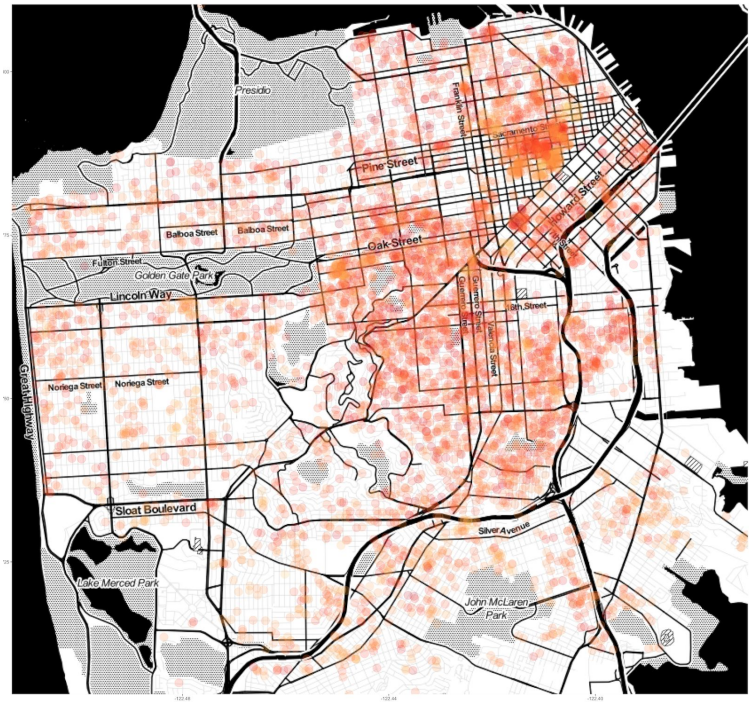


A Statistical Analysis of San Francisco Airbnb Listing Prices

Hailey Han, Jihong Lee, Josh Kang

Abstract

Given the recent increase in popularity of Airbnb as an online platform that provides lodging options for users, we wanted to understand which and how various variables that characterize an Airbnb listing affect its price per night. Data was collected for 13 different variables (price per night, host response rate, overall rating, neighborhood, superhost status, etc.) for all 8,111 Airbnb listings for the San Francisco area as of April 7, 2020. A multiple linear regression was fitted to generate the final model with room type, log of number of reviews, number of bedrooms, number of people accommodated and neighborhood (classified with the four aforementioned groups) serving as strong predictors for the log of price per night at $\alpha = 0.05$. Future extensions of this statistical analysis might consider fitting a multiple linear regression model predicting Airbnb rating, instead of price, or extending the same analysis to different major cities worldwide to observe the impact of different urban backgrounds on the regression model.

✳ *Note: Additional information and larger renditions of figures can be found in the Appendix.*

A. Background and Introduction

Airbnb is an online marketplace that provides lodging options for users. Since many users make a decision on their choice based on several characteristics, statistical analysis is useful to understand how the prices are determined. With the upsurge in popularity of the platform, we were specifically interested in using multivariate analysis to understand how location, user feedback, and accommodation characteristics can affect the pricing of Airbnbs. We conducted a statistical analysis on all the Airbnb listings in San Francisco (SF). We focused on the SF dataset as it is a major city with diverse communities and socio-economic backgrounds. Through this analysis, we hope to bring clarity to what the pricing reflects and which variables affect most, the price per night of an Airbnb experience.

B. Data and Exploratory Analysis

1. Data and Variables

To determine variables needed to predict listing price, we used the population dataset of *8,111 unique listings* in the SF area as of April 7, 2020, retrieved through publicly available information on the Airbnb website[1]. Each listing reported *7 quantitative variables*—price per night ($), host response rate (%), number of people accommodated, number of bedrooms, number of bathrooms, number of reviews, and overall rating (%)—and *4 categorical variables*—neighborhood, host superhost status, property type, and room type. *Price per night* was the response variable for this study with all other variables as potential predictors.

For *neighborhoods*, we used a map of SF and an additional dataset on SF neighborhoods' median household income[2] to divide the variable into 4 subgroups (instead of 55 distinct neighborhoods): Downtown, Northern Residential, Central, and Outskirts (used as reference). Downtown is known for its level of wealth; the Northern Residential area is known for its posh Victorian neighborhoods; Central is known for UCSF, its hipster culture, and ethnic groups; and the Outskirts include areas outside the heart of the city. For *property type*, we created 2 subcategories: "high" (villas, resorts, hotels, and boutique hotels), "low" (hostels, bed and breakfasts), and "medium" (all other property types, used as reference). For *room type*, we created 3 subcategories: private room or hotel room, shared room, and entire home/apartment (used as reference).

2. Exploratory Data Analysis

As shown in **Figure 1**, the distribution of our response variable—price per night—is heavily right-skewed and centered around $145.00. Most listings had prices < $200.00 and potential outliers may be present for listings with prices > $800.00.

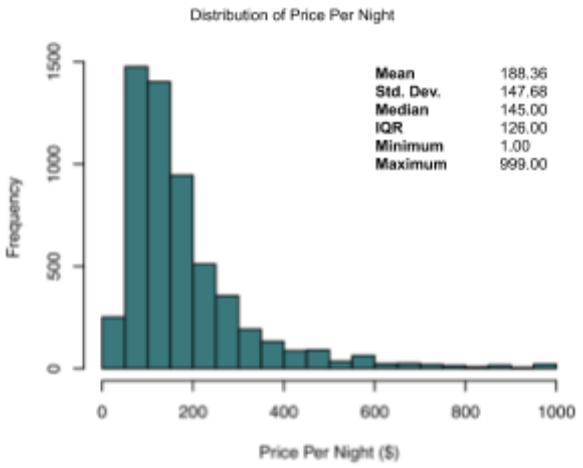


Figure 1. Histogram of the response variable, price per night ($).

Figure 2 shows the distributions of some of the predictor variables we decided to include in our final model—number of people accommodated, number of bedrooms, overall rating, neighborhood, and room type—as well as each of their scatter plots against the response variable.

The distribution for the number of people accommodated and the number of bedrooms are

heavily right-skewed while the distribution for the overall rating is heavily left-skewed with centers at 2 people accommodated, 1 bedroom, and 98% in overall rating. There seems to be a relatively strong linear relationship between the response variable, price per night, and the two variables: number of people accommodated and number of bedrooms. The association between price per night and overall rating seems to be slightly curved, suggesting that a transformation may be necessary. The neighborhoods are unequally distributed with the largest number of listings located in the Outskirts, followed by Downtown. The room types are also unequally distributed with the largest number of listings being entire homes, followed by private rooms.

C. Model and Results

1. Analytic Method

We analyzed our data with a multiple linear regression model. First, we tried using all 10 potential predictor variables and interactions of neighborhood with room type and bedroom counts, However, realized that this initial model violated all model assumptions and that many of the predictors were not useful given the others, based on the t-tests for individual coefficients yielded large p-values. The Box-Cox plot suggested that we used the log of price for each listing. Lasso regression was performed because we hoped to narrow our predictor variables and minimize error with a good prediction accuracy. We choose to use a lambda of 6.328346, which minimizes Menlo's Cp. Also, a data entry with zero price has been removed in our analysis.

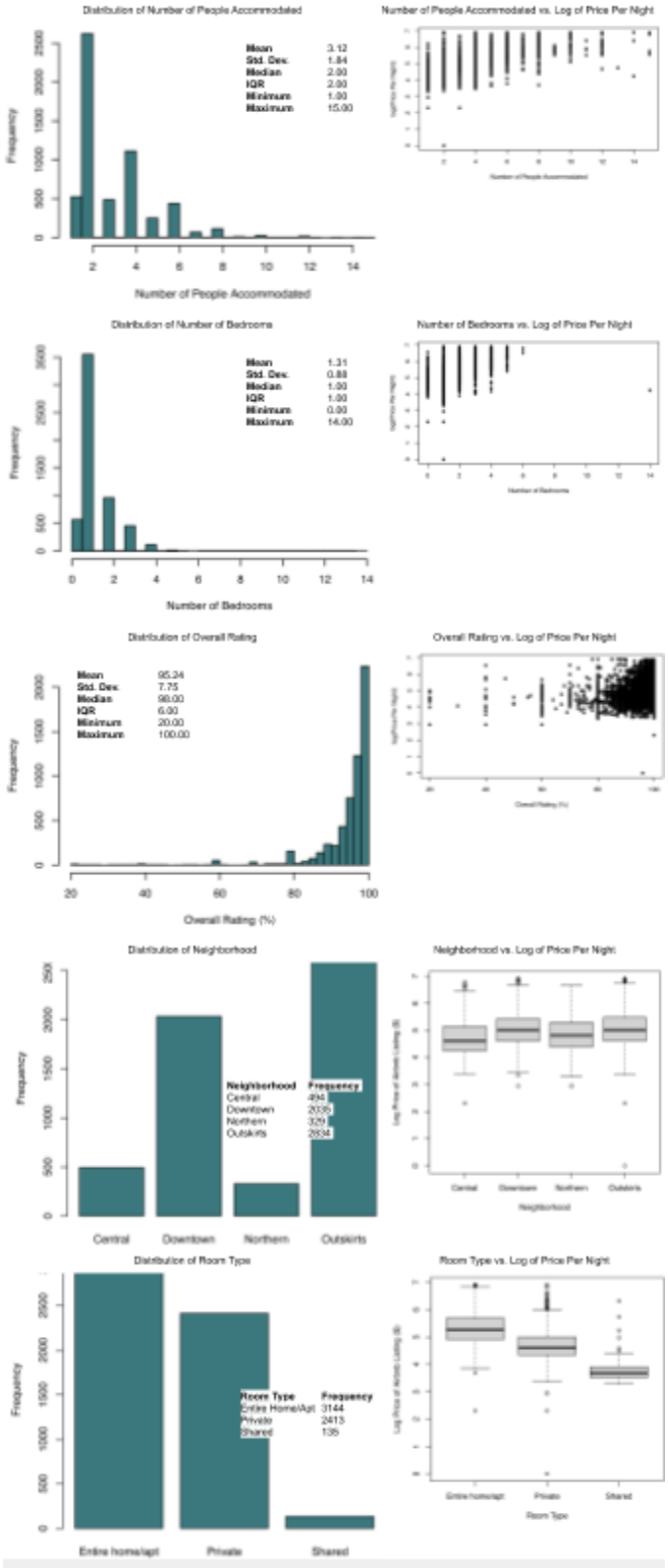


Figure 2. Histogram of the predictor variables used in the final model: number of people accommodated, number of bedrooms, overall rating (%), neighborhood, and room type. Relationship between Log of Price Per Night and each of the final predictor variables.

2. Final Model and Results

Our final model is summarized in the right.

Given our diagnostic plots in **Figure 4**, this multivariate model seems appropriate as all model assumptions are satisfied. There are no clear patterns in the residual plots that suggest constant variance and a normal distribution. The points are well-fitted to the line in the Normal QQ plot.

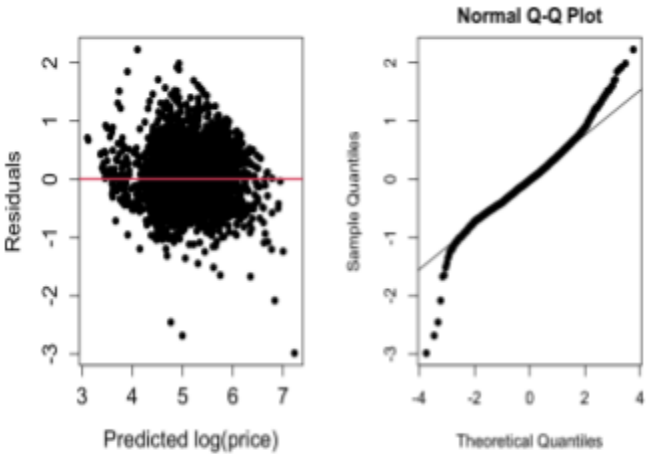


Figure 4. Diagnostic plots for the final model with the predicted log of price per night vs. residuals on the left and the normal QQ plot on the right.

```
Call:
lm(formula = log(price) ~ accommodates + bedrooms + bathrooms +
    log(number_of_reviews) + review_scores_rating + room_type +
    property_high + property_low + neigh_NO + neigh_DT + neigh_CT +
    room_type * neigh_NO + room_type * neigh_DT + room_type *
    neigh_CT + bedrooms * neigh_NO + bedrooms * neigh_DT + bedrooms *
    neigh_CT, data = listings)

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.2816218  0.0758657  43.256  < 2e-16 ***
accommodates                   0.1011842  0.0049908  20.274  < 2e-16 ***
bedrooms                       0.2011059  0.0157096  12.801  < 2e-16 ***
log(number_of_reviews)         0.0071359  0.0034565   2.064  0.03902 *
review_scores_rating           0.0117171  0.0007434  15.762  < 2e-16 ***
room_typeHotel room           -0.4837286  0.0756960  -6.390 1.79e-10 ***
room_typePrivate room         -0.4157664  0.0231962 -17.924  < 2e-16 ***
room_typeShared room          -1.1344775  0.0947315 -11.976  < 2e-16 ***
property_high                  0.4475387  0.0427742  10.463  < 2e-16 ***
property_low                  -0.2122659  0.0530428  -4.002 6.37e-05 ***
neigh_NO                       0.4622830  0.0429504  10.763  < 2e-16 ***
neigh_DT                       0.2710494  0.0371381   7.298 3.31e-13 ***
neigh_CT                       0.2147728  0.0347189   6.186 6.60e-10 ***
room_typeHotel room:neigh_NO   0.4865961  0.1335204   3.644  0.00027 ***
room_typePrivate room:neigh_NO 0.0234263  0.0418556   0.560  0.57571
room_typeShared room:neigh_NO  0.1386327  0.1964328   0.706  0.48037
room_typeHotel room:neigh_DT   0.1345153  0.0893402   1.506  0.13221
room_typePrivate room:neigh_DT 0.1736157  0.0343670   5.052 4.51e-07 ***
room_typeShared room:neigh_DT -0.0914587  0.1065687  -0.858  0.39081
room_typeHotel room:neigh_CT           NA         NA      NA       NA
room_typePrivate room:neigh_CT 0.0710825  0.0305071   2.330  0.01984 *
room_typeShared room:neigh_CT  0.0468388  0.1364810   0.343  0.73147
bedrooms:neigh_NO             -0.0881893  0.0206774  -4.265 2.03e-05 ***
bedrooms:neigh_DT              0.0132490  0.0209361   0.633  0.52687
bedrooms:neigh_CT              0.0025412  0.0171522   0.148  0.88223
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4143 on 5667 degrees of freedom
Multiple R-squared:  0.5992,    Adjusted R-squared:  0.5976
F-statistic: 368.4 on 23 and 5667 DF,  p-value: < 2.2e-16
```

Figure 3. Summary output of the final model that displays the estimated coefficients, t-test statistics, F-statistic, residual standard error, and adjusted R2.

As seen in **Figure 5,** we have a residual standard error of 0.4143 with a relatively large adjusted R-squared value of 0.5976. All the final predictors— accommodates, bedrooms, log of number of reviews, review score rating, room type, property type, and neighborhoods— are useful at $\alpha = 0.05,$ since the t-tests for each of the variables return small p-values <0.05. We also note that many of our interaction terms, representing the spatial correlation between room type and number of bedrooms, yield substantially low p-values when evaluated using t-tests.

D. Discussion and Conclusions

Our main goal for this study was to understand what characteristics of an Airbnb listing in SF help predict its price per night. Based on our results and final model, it appears that room type, log of number of reviews, number of bedrooms, number of people accommodated and neighborhood (classified with the four aforementioned groups) are strong predictors for the log of price per night at $\alpha = 0.05$.

Our study encompassed the population of listings in SF as of April 2020, so our findings can definitely change in the future. In coding the variable of neighborhoods into our model, we divided it into 4 subgroups to make our final model more stable. There was difficulty doing so based merely on geographic location. For further research, we would apply additional datasets to our model that detail more characteristics (i.e. crime rate and median income) of each neighborhood to group them more accurately. Also, we could study these predictors' relationship with rating, instead of price, exclusively. Finally, to make this statistical analysis more useful and applicable, these questions could be easily applied to data on other cities worldwide.

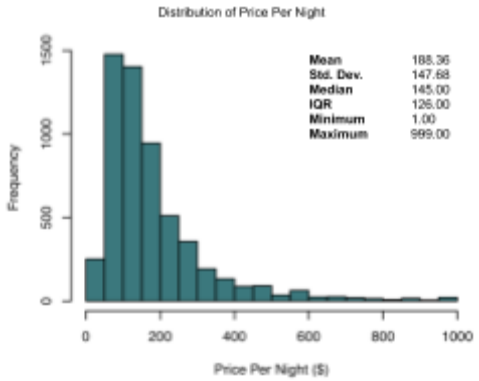
References

1. Loretizo, John Elmer. “San Francisco Airbnb.” *Kaggle*, 12 Nov. 2019, www.kaggle.com/jeploretizo/san-francisco-Airbnb-listings.
2. San Francisco Planning Department. *San Francisco Neighborhoods: Socio-economic Profiles*. San Francisco: San Francisco Planning Department, 2017. Web.

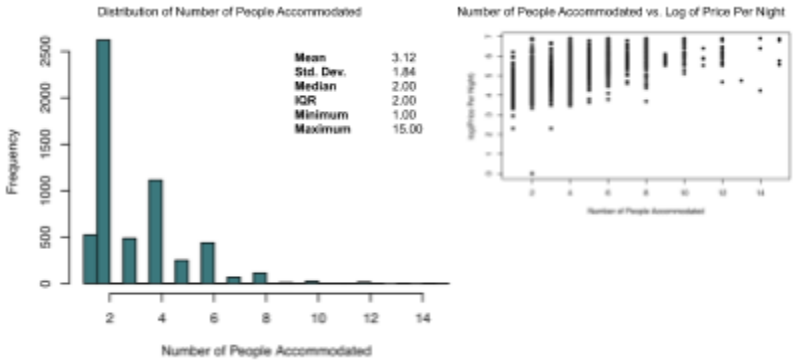
Appendix

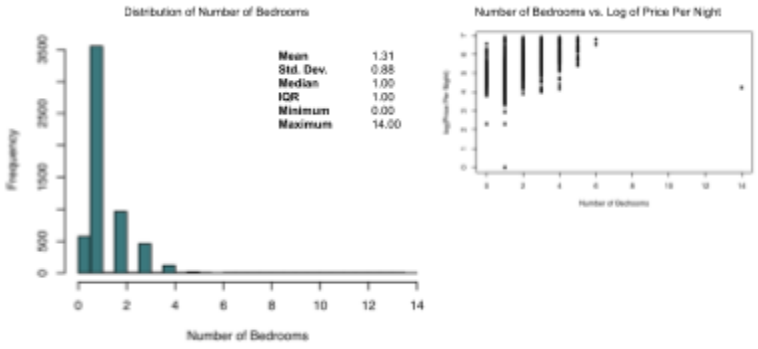
A. Larger Versions of Figures from Exploratory Data Analysis

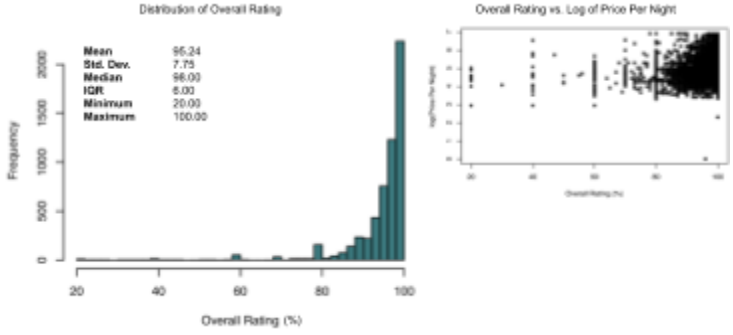
1. Histogram and Numerical Summary of Response Variable: Price Per Night ($)

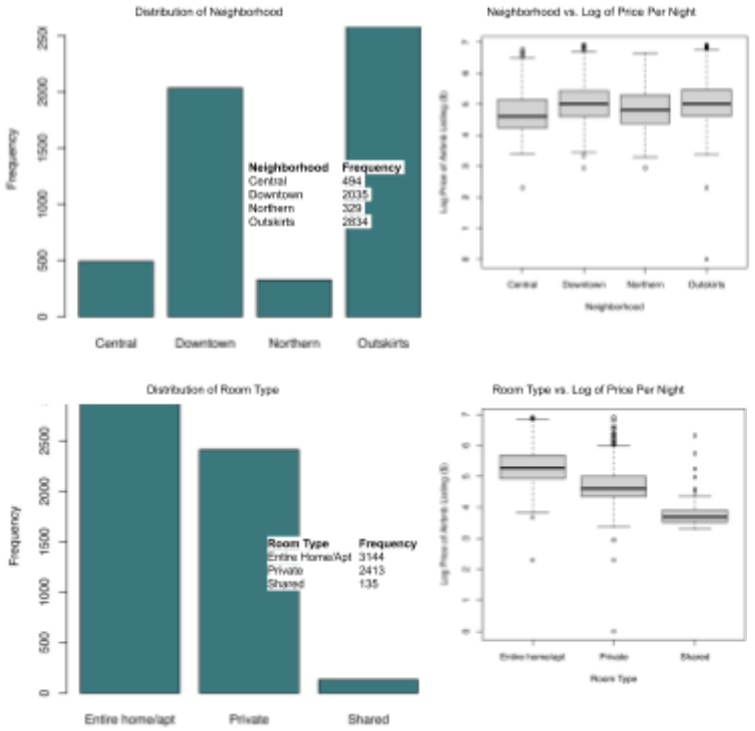


2. Histogram + Summary of Predictor Variables and Scatterplot Against Price Per Night ($)

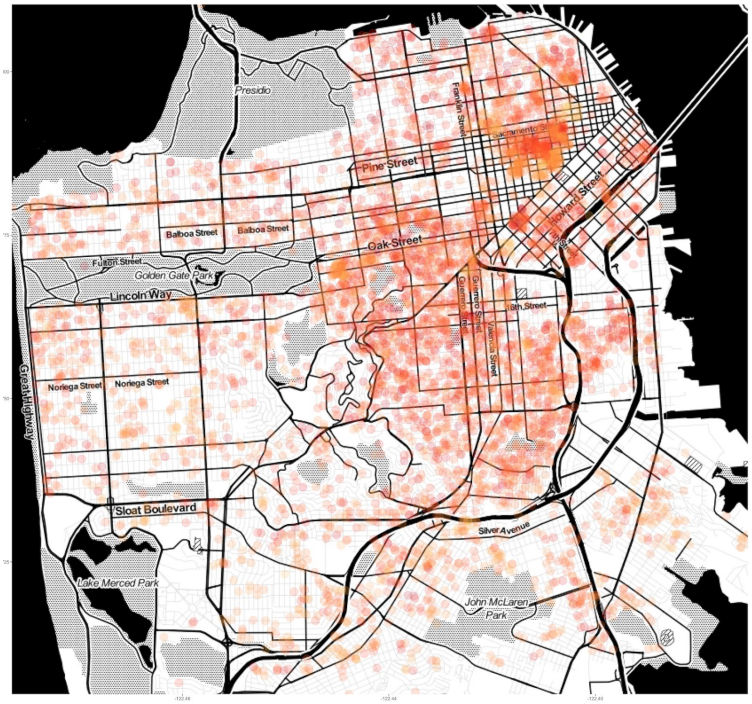








B. Map plots

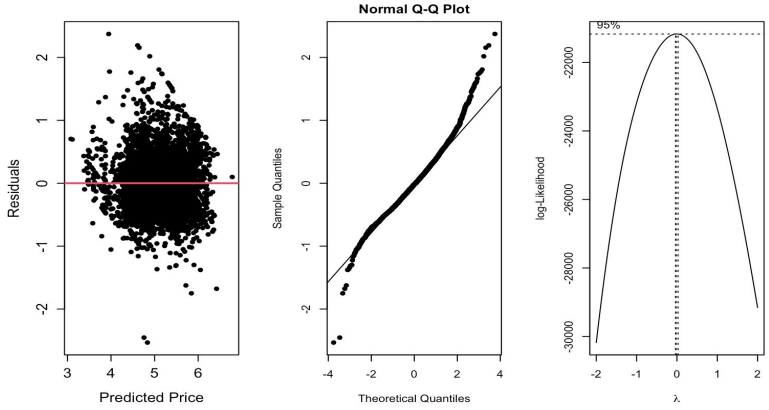


Map plot of log(prices); lighter to darker red indicates price range from low to high.

Map plot of log(prices) for each of the four neighborhood groups;
starting from top left, clockwise: Central, Downtown, Northern Residential, Other neighorhoods

C. Figures from the Initial Model

1. *Model Assumptions:*
   Residual Plot + Normal QQ-Plot + Box-Cox Plot

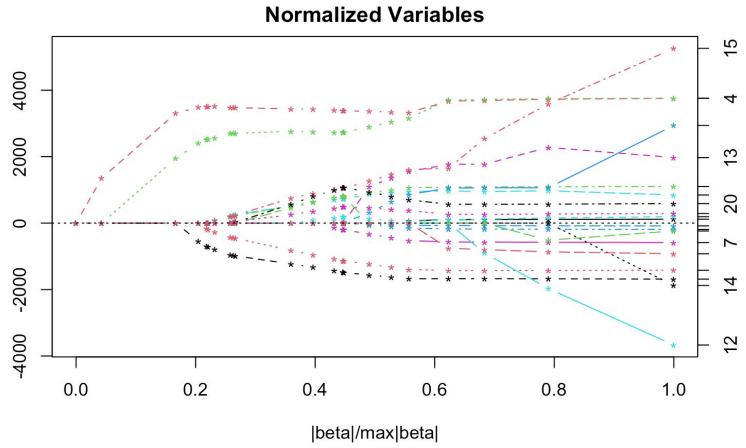


2. *Inferential Results*

A. Many predictors are insignificant as seen by their large p-values. The interaction terms are only useful for certain neighborhoods.
B. Large Adjusted R-squared value, but this could be because we have a lot of predictors.
C. LASSO regression performed as below.

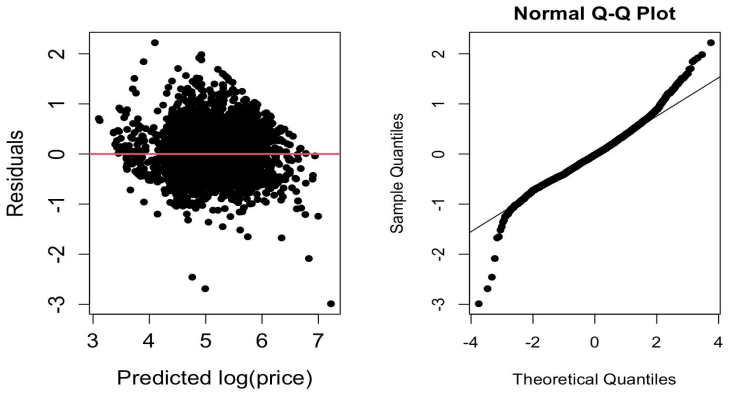
```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -136.36144   40.23355  -3.389 0.000706 ***
host_response_rate             0.15771    0.13660   1.155 0.248337
accommodates                  27.56056    1.69928  16.219  < 2e-16 ***
bedrooms                      55.26558    4.93737  11.193  < 2e-16 ***
bathrooms                      6.21031    1.79917   3.452 0.000561 ***
number_of_reviews             -0.08329    0.01676  -4.970 6.88e-07 ***
review_scores_rating           1.48629    0.39361   3.776 0.000161 ***
room_typeHotel room          -51.92617   10.35147  -5.016 5.43e-07 ***
room_typePrivate room        -45.97467    3.37611 -13.618  < 2e-16 ***
room_typeShared room         -130.36366   10.30116 -12.655  < 2e-16 ***
property_high                 104.33398   10.67336   9.775  < 2e-16 ***
property_low                  -20.87668   12.27858  -1.700 0.089138 .
neigh_NO                     -151.99547   70.09151  -2.169 0.030160 *
neigh_DT                      57.97062   45.38730   1.277 0.201569
neigh_CT                     -55.33604   53.67955  -1.031 0.302651
review_scores_rating:neigh_NO  2.26359    0.72770   3.111 0.001876 **
review_scores_rating:neigh_DT -0.08244    0.47672  -0.173 0.862711
review_scores_rating:neigh_CT  0.88031    0.55848   1.576 0.115022
bedrooms:neigh_NO              4.07233    5.04617   0.807 0.419693
bedrooms:neigh_DT              7.73796    5.20095   1.488 0.136860
bedrooms:neigh_CT              8.20186    4.08936   2.006 0.044940 *
accommodates:bedrooms         -1.92228    0.60014  -3.203 0.001367 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 105.1 on 5669 degrees of freedom
Multiple R-squared:  0.4955,    Adjusted R-squared:  0.4936
F-statistic: 265.1 on 21 and 5669 DF,  p-value: < 2.2e-16
```



C. Figures from the Final Model

1. *Model Assumptions:*
   Residual Plot + Normal QQ-Plot
2. *Inferential Results*




```
Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.2816218  0.0758657  43.256  < 2e-16 ***
accommodates                   0.1011842  0.0049908  20.274  < 2e-16 ***
bedrooms                       0.2011059  0.0157096  12.801  < 2e-16 ***
log(number_of_reviews)         0.0071359  0.0034565   2.064  0.03902 *
review_scores_rating           0.0117171  0.0007434  15.762  < 2e-16 ***
room_typeHotel room           -0.4837286  0.0756960  -6.390 1.79e-10 ***
room_typePrivate room         -0.4157664  0.0231962 -17.924  < 2e-16 ***
room_typeShared room          -1.1344775  0.0947315 -11.976  < 2e-16 ***
property_high                  0.4475387  0.0427742  10.463  < 2e-16 ***
property_low                  -0.2122659  0.0530428  -4.002 6.37e-05 ***
neigh_NO                       0.4622830  0.0429504  10.763  < 2e-16 ***
neigh_DT                       0.2710494  0.0371381   7.298 3.31e-13 ***
neigh_CT                       0.2147728  0.0347189   6.186 6.60e-10 ***
room_typeHotel room:neigh_NO   0.4865961  0.1335204   3.644  0.00027 ***
room_typePrivate room:neigh_NO 0.0234263  0.0418556   0.560  0.57571
room_typeShared room:neigh_NO  0.1386327  0.1964328   0.706  0.48037
room_typeHotel room:neigh_DT   0.1345153  0.0893402   1.506  0.13221
room_typePrivate room:neigh_DT 0.1736157  0.0343670   5.052 4.51e-07 ***
room_typeShared room:neigh_DT -0.0914587  0.1065687  -0.858  0.39081
room_typeHotel room:neigh_CT          NA         NA      NA       NA
room_typePrivate room:neigh_CT 0.0710825  0.0305071   2.330  0.01984 *
room_typeShared room:neigh_CT  0.0468388  0.1364810   0.343  0.73147
bedrooms:neigh_NO             -0.0881893  0.0206774  -4.265 2.03e-05 ***
bedrooms:neigh_DT              0.0132490  0.0209361   0.633  0.52687
bedrooms:neigh_CT              0.0025412  0.0171522   0.148  0.88223
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4143 on 5667 degrees of freedom
Multiple R-squared:  0.5992,    Adjusted R-squared:  0.5976
F-statistic: 368.4 on 23 and 5667 DF,  p-value: < 2.2e-16
```