**Predictions for Sharing Economy Platforms**

**A Case Study on Airbnb**

**3/6/2019**

**Authors**

Adrian Tullock, Cathy Jia, David Wei, Monique Bi, Naman Sharma

# Abstract

There were three primary aims of this study. The first was to determine what predictor variables, among the hundreds available in our Airbnb Seattle listings data, had the highest linear association with listing prices, as well as overall review scores. Simple and stepwise linear regression would be used for price and review scores respectively for analyzing these relationships. We concluded that while higher counts of bathrooms and guest capacity associate with higher pricing, higher bedroom counts did not. Interestingly, higher bedroom counts did associate with higher ratings.

Becoming a superhost is a desirable platform achievement rewarded with elevated visibility. Therefore, we sought to answer which factors were key to attaining superhost-eligible ratings. We employed logistic regression for this analysis. Upon review of the odds-ratios, we concluded that accuracy, cleanliness, ease of check-in, and host communication were the four factors that most increased the odds of attaining desirable reviews.

A secondary aim was investigate the feasibility of a linear predictive model for price and review scores based on positively associated factors. A successful proof of concept would lend direction to future work on this study. A very small percentage of potential factors were used for this analysis and R-squared values below 0.3 were attained in both instances. Also, one factor failed to attain statistical significance in the case for pricing. We concluded that the pricing model would be more feasible than the review score model for prediction, but each presents its own challenges. Pricing would likely need a reassessment on relevant predictors. Review scores, on the other hand, are highly susceptible to confounding due to unavoidable user selection bias.

# Introduction

<Include a clear statement of the aims and questions to be addressed. The goal of the analyses should be clear from the introduction. You may include background material, such as previous research that your analyses builds on, but this should be kept brief.>

-*Consider comments from Brian based on presentation: The goals of the analysis are a little unclear, eg, whether it is focused on inferences about associations or on prediction. Try to make this clearer in the written report. In some places you focused on the associations but in others discussed results in terms of prediction, e.g., low R-squared.*

# Data Set Description

This data set is comprised of host listing details, aggregated by city and broken down by individual listings. Users are allowed to have multiple listings, so uniqueness is determined by a numerical identifier with a matching link URL. The data set is compiled of entries spanning individual cities from Airbnb. The Seattle data has 96 columns and well over 8000 rows. Should we see fit to include data from other cities, this number could easily double or triple in size. We expect the data to be relatively homogenous across locations. Sensitive data has been scrubbed to protect privacy, but other than that the data is pretty well filled out.

There is a broad mix of numerical, boolean, character values, web links and percentages across the data. There are some fields which explicitly have zero values and others which appear to use empty values to indicate ‘N/A’. Analysis of certain fields will likely need data extrapolation for aggregation.

The location data for the listings is a bit redundant, but it may be due to the data being split from a central source prior to archiving. Other data available includes the listing amenities, various pricing stages, average host response times, listing types, review ratings, availability and geospatial coordinates.

# Data Source

URL: <http://insideairbnb.com/get-the-data.html>

The data is provided by Inside Airbnb. According to the website, Inside Airbnb is *an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world*.   
  
The data is a collection of Airbnb data that is publicly available on its website, across multiple cities. The data is aimed at providing a 360-degree insight into Airbnb’s presence in a city. According to the source, *the data has been analyzed, cleansed and aggregated where appropriate to facilitate public discussion.* The data can be *copied, modified, distributed and performed work on,* even for commercial purposes, all without asking permission. A brief summary of assumptions and disclaimers in the data set is available at <http://insideairbnb.com/about.html#disclaimers>.

# Data Availability

The data set is readily available and ready for further cleaning. As presented, the listing data is fairly mature and useable for analysis. However, the amenity offerings are aggregated in a single columns as comma-separated lists. This form is not ideal for statistical analysis, so we un-nested these attributes and reformed them as binary predictors across all listings.

# Questions

**Q1.** Key factors for price and review scores

* What is the most important single factor for getting a high price on your listing?
* Similarly, what’s the factor you ought to focus on to get better reviews?
* Repeating the above with the goal of finding the secondary factors that enhance the effect of the primary factor identified.

**Q2.** Factors for becoming a Superhost[[1]](#footnote-0)

* As a host, what would be the important factors that would help a host to get higher rating to qualify for a Superhost?

# Variables (Can be moved in analysis)

Main variables to focus on and evaluate:

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Comment** |
| price | float | Daily rental price |
| host\_is\_superhost | factor | “T” or “F” |
| review\_scores\_rating | float | Total score rating  Subscores rating: accuracy, cleanliness, check-ins, communication, location, value |
| number\_of\_reviews | int |  |

Airbnb rental features to look for patterns and/or correlations:

|  |  |  |
| --- | --- | --- |
| **Feature Type** | **Variable Name** | **Main Data Types** |
| Amenities | room\_type, accommodates, bathrooms (number of), bedrooms (number of), beds, wifi, Cable TV, Washer/Dryer | int, factor (“T” or “F”) |
| Extra fee | security\_deposit, cleaning\_fee, extra\_people | float, int |
| Rental Policy | cancellation\_policy, minimum\_nights (number of) | factor, int |
| Host Verifications | email, phone, facebook, google, reviews | factor (“T” or “F”) |
| Others | neighborhood, ... (Any potential factors affecting customer rating or price change) |  |

# Potential Problems (can be removed, use Discussion section)

Given the objective of our hypothesis, we intend to work on as wide a variety of factors as possible. However, in this scenario, assessing the importance/significance of each factor may be a challenge for factors with a large number of groups. This problem will limit the tests to a sub-group of the available factors. We would need several attempts to decide which factors to use in our project.  
  
The response variable in the data set also needs to be assessed, tested and reviewed before putting in use. This is caused by the fact that the available definitions for the data set are limited in the information they provide. So, for example, in choosing an appropriate metric for *price* we would need to check the fields *listed\_price, security\_deposit,* and *cleaning\_fee*, among others, before zeroing in.

The data itself is relatively clean and well indexed however it’ll still require some cleaning, feature engineering, and reshaping.

* Missing data in key fields account for about 10% of the data and hence will have to be eliminated - which will lower the power/significance of our tests.
* Subgroups in certain fields account for too little of the population to be considered valid - they will need to be weeded out for each factor while testing.
* The amenities and owner\_verification fields are comma separated text fields that need to be normalized as they contain a plethora of possibly useful factors.

# Statistical Analysis

The following analysis has two parts. The goal is ...

## Part I. Key Factors for Price and Review Scores

For the price and review score, we go through the four-step process of:

1. Model Selection
2. Factor Selection
3. Model Assumptions Check
4. Variable Comparison Methodology and Result

### 1. Model Selection

We set out with the primary goal of identifying factors, each with a varying number of groups, that affect the value of ‘price’ and ‘review scores’ here. With this goal in mind, it is only sensible to look at ANOVA or Linear Regression (LR) in mind. We rule out Logistic regression here due to the non-binary nature of the dependent variable.

Now the eventual goal here is to come up with variables that have an effect on the value of ‘price’ and ‘review-scores’, as well as evaluate [rank] the magnitude of the effects. Additionally, we suspect a linear dose-relation effect between the subgroups, for example between ‘host\_response\_time’ and ‘review scores’. With these goals in mind, it makes more sense to proceed with a Linear Regression model to assess the relationships. The choice also supports assessment with dependent variables of non-factor nature.

### 2. Key Factors Selection

#### Factors for Price

Based on our intuition, the larger the house is, the higher the listing price is. We decided to start from this perspective and only consider the properties of the house itself. So we start with bedrooms, bathrooms, accommodates, and cleaning\_fee. The next step will be taking into consider of the values brought from host itself. We believe the super\_host status and host response time can boost the housing price.

**Response variable**

* Price

**Predictor variables**

* bedrooms
* bathrooms
* accommodates
* cleaning\_fee

#### Factors for Review Scores

The factor selection for the review scores is split in two parts: an intuitive variable selection, followed by a refinement based through Stepwise regression based on Akaike Information Criterion.

The initial factors selected were picked subjectively, and then tested for assumptions. The motivation behind the selection was to pick variables that reflect different aspects to a listing that can assumed to be independent. Post this initial selection, a forward/backward elimination process boils it down to a best-fit model with a sub-selection of the variables.

**Response variable**

* Review Scores

**Predictor variables**

* **Preliminary** 
  + Price
  + Accommodates
  + Bedrooms
  + Guests included
  + Minimum nights
  + Host listings count
  + Host response rate
  + Host is superhost
  + Total Amenities
* **Final** 
  + Host is superhost
  + Host listings count
  + Host response rate
  + Accommodates
  + Bedrooms
  + Total Amenities

### 3a. Model Assumptions Check - Price [Linear Regression]

**Independence**

Satisfied. Through our knowledge of the Airbnb pricing model (it allows renters to adjust their prices) and how tenants rate their stays, we believe the independence of our data set is valid.

We leveraged R’s plot function to visualize residuals versus fitted values. Given we have over 200 different predictors, we ran a simple linear regression on all predictors and selected Top 15 predictors with the greatest coefficient and have significant p-value.

**Linearity**

Satisfied. From Figure 1, we can see the red line being horizontal. This is good evidence on linearity.

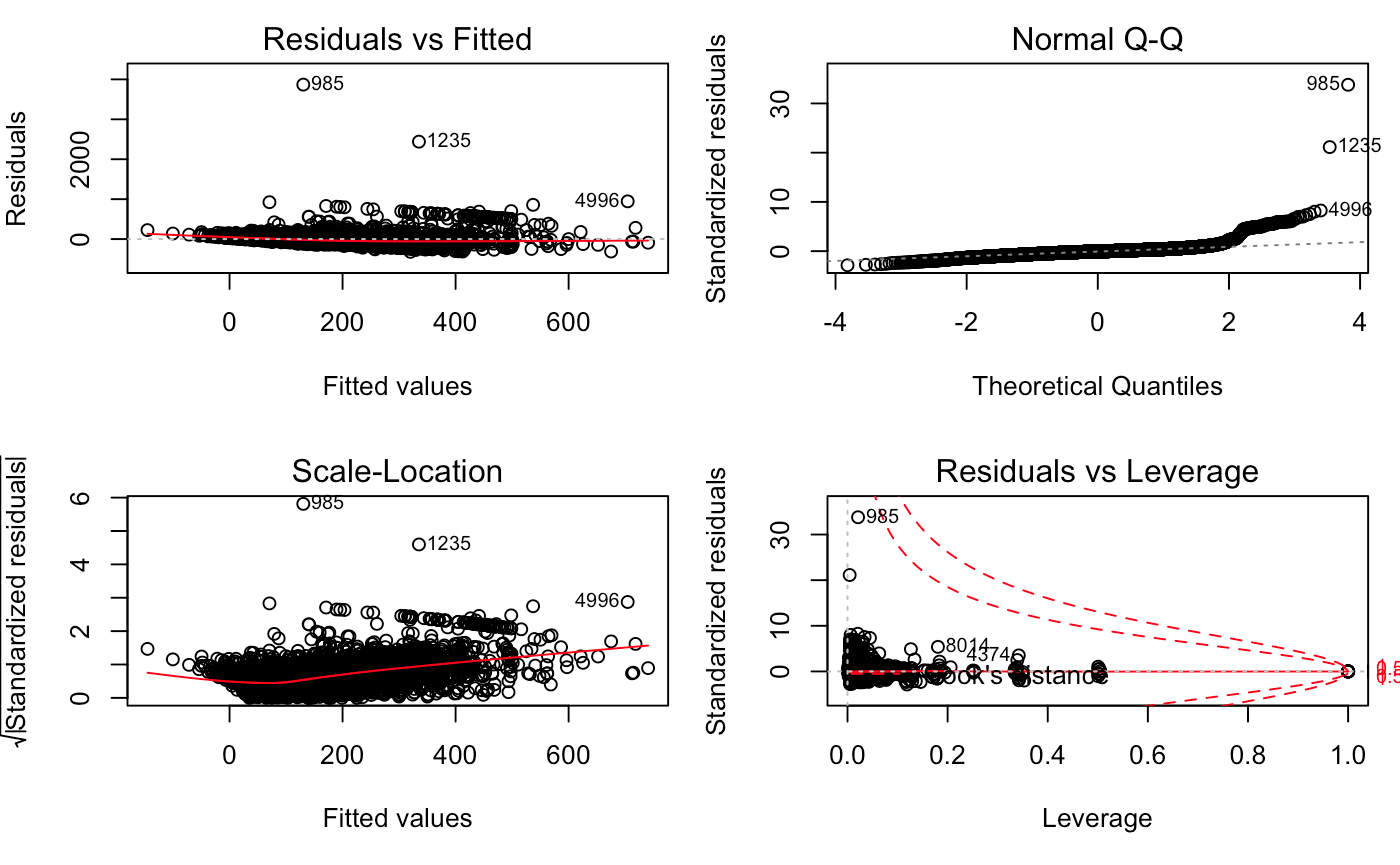
**Constant variance**

Satisfied. From Figure 1, we can see most data points are around zero-line. Although, there exists a weakly positive relationship between mean and variance from Figure3.

**Normality**

Not satisfied. Figure 2 suggested a reasonable approximation to a normal distribution, with a heavy tail. We accept this heavy tail as extreme outliers exist in our model. Given the large size of our sample (8459), we chose to not take normality strictly.

Figure 1 to Figure 4



### 3b. Model Assumptions Check - Review Scores [Linear Regression]

**Linearity**

Satisfied with adjustment. The horizontal line is mostly flat, with some variance when review ratings are low.

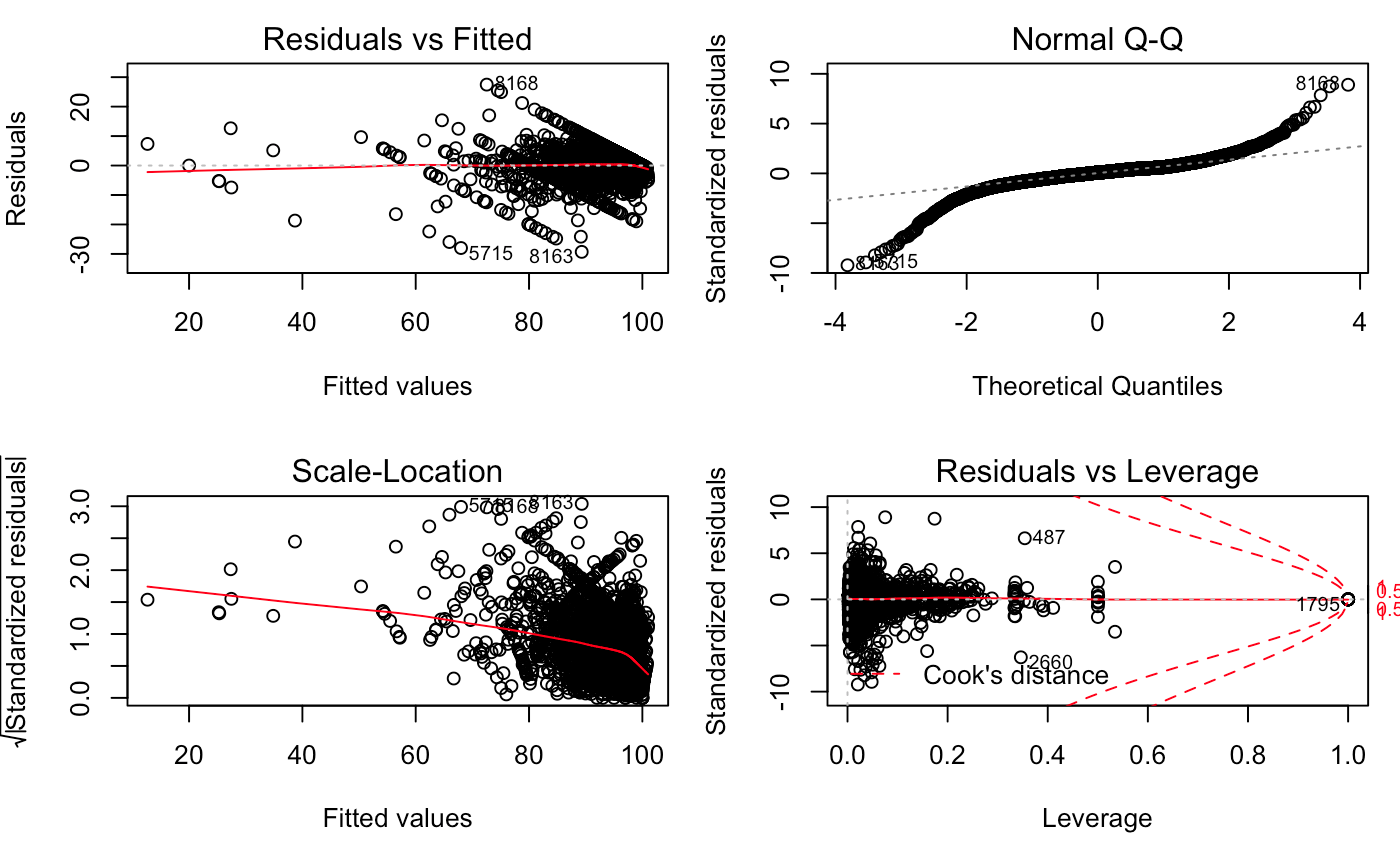
**Constant variance**

Not satisfied. The Scale-location graph (Figure 7) and Residuals vs Fitted graph (Figure 5) suggested evidence against constant variance. We believe there might be correlations between our chosen predictors, which will require adjustment while fitting our model.

**Normality**

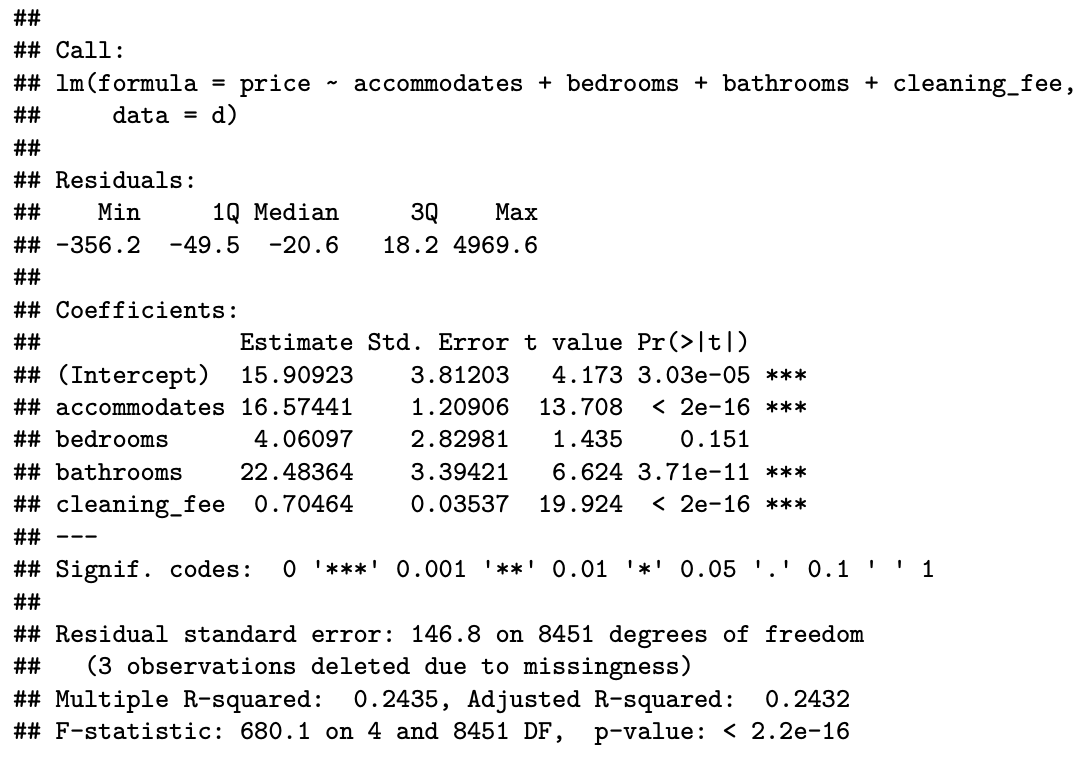
Not satisfied. The Q-Q plot showed non-normal line with heavy heads and tails. However, since our sample size is large enough (~8.5K), we can have CLT to take effect so the non-normality does not affect our method selection.

Figure 5 to Figure 8

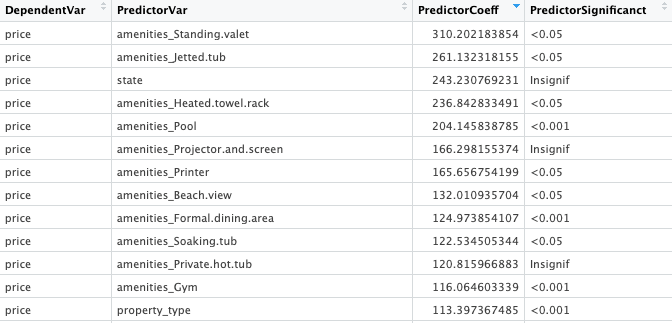


### 4. Result and Variable Comparison Methodology

**Results for Price Linear Regression Test:**

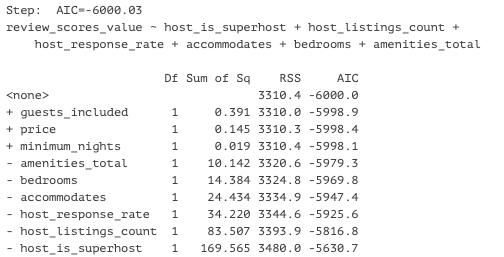
****

Each Dependent Variable and Predictor Variable combination is tested for the NULL hypothesis [same expected response observed in the dependent variable for any value of the predictor variable]. The factors are first filtered for relevance through p-value significance [PredictorSignificance], and then ranked with the expected movement in the dependent variable [PredictorCoeff] against a unit movement of the factor itself.



This evaluation will be broadened in the final phase with the consideration for Std Error/Residual Std Errors that the model delivers for each of the Dependent Variable - Predictor Variable combination. Further, we utilize the results from here to move on to the next phase of our assessment, i.e., factor selection with adjustments.

**Results for Review Scores Linear Regression Test:**

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** |
| (Intercept) | 9.53E+00 | 2.51E-02 | 379.3 | < 2.00E-16 \*\*\* |
| host\_is\_superhost | 3.45E-01 | 1.77E-02 | 19.504 | < 2.00E-16 \*\*\* |
| host\_listings\_count | -6.85E-04 | 5.01E-05 | -13.688 | < 2.00E-16 \*\*\* |
| host\_response\_rate | -1.85E-03 | 2.11E-04 | -8.762 | < 2.00E-16 \*\*\* |
| accommodates | -4.31E-02 | 5.82E-03 | -7.404 | 1.47E-13 \*\*\* |
| bedrooms | 7.40E-02 | 1.30E-02 | 5.681 | 1.39e-08 \*\*\* |
| amenities\_total | 3.87E-03 | 8.10E-04 | 4.77 | 1.88e-06 \*\*\* |

Residual standard error: 0.6676 on 7427 degrees of freedom

Multiple R-squared: 0.1142, Adjusted R-squared: 0.1134

F-statistic: 159.5 on 6 and 7427 DF, p-value: < 2.2e-16

### 5. Discussion - Adjustment Factors and Interactions

After identifying which factors to use as main effects, those effects will be compared against other potential adjustment factors, individually and in multiple factor scenarios. The decision to keep or eliminate potential adjustment factors will be based on how correlated factors are. Predictors that may be considered secondary candidates as the main effect will also receive priority consideration to be adjustment factors. The goal of the model is to explain the variation in the response variable without overcomplicating the interpretation of the model. The multiple and adjusted r-squared values will help assess viable adjustment candidates.

We must also be cautious to account for the interaction between chosen predictors. Tested interactions that are statistically significant are the most likely to be included for testing. Adjustment and interaction models that maintain or minimize residuals will be judged as useful in model selection. For our current question set, most of our predictors are binary in nature. This should somewhat simplify evaluations on what variables are relevant to answering our questions.

It is not expected that price, one of our response variables, will be best explained by a sole predictor. Once the main effect is determined, additional predictors that better explain the price variability will be tested for adjustments and interactions. The process of selecting the main effect will provide intel on ideal candidates as secondary factors. Baring the idea of dividing price ranges into categories, we will look for secondary factors that enhance the effect of the chosen primary factor.

Early testing hints that categorical review scores may highly correlate between categories, as well as with the overall listing review score. A more comprehensive test on this may cause us to remove categorical review scores as predictors of overall review scores due to their interrelated dependence. On the other hand, this may provide new question opportunities; such as determining if whether the correlation between scores is attributed to similar factors.

To reiterate the methodology, we will pair the main effect predictor with adjustments and iteratively determine which factors improve the model’s accuracy; ideally increasing the r-squared values. As interactions are tested for, statistically significant interactive predictors will be eligible for the model.

## 

## Part II. Key Factors for Superhost Qualification

### Background

Airbnb has a Superhost program in order to reward Airbnb’s top-rated and most experienced hosts. Superhost are not only recognized as “the best in hospitality”, but also receive financial benefits such as earning 22% more on average, attracting more guests and accessing exclusive rewards from Airbnb.

As a highly valued status, Superhost status is checked every three months with the following criteria. <https://www.airbnb.com/superhost>

1. Have a 4.8 or higher average overall rating based on reviews from at least 50% of their Airbnb guests in the past year.
2. Have hosted at least 10 trips in the past year.
3. Have no cancellations in the past year.
4. Respond to 90% of new messages within 24 hours.

The requirements 2, 3 and 4 can be easily quantified, while the first one is related to various factors. Therefore we decided to concentrate on review ratings to help the hosts to better prepared for the listings in Superhost analysis.

As listed on the website, Airbnb has six rating categories:

1. Accuracy: How accurately did the listing page represent the space?
2. Communication: How well did hosts communicate with guests before and during their stay?
3. Cleanliness: Did guests feel that the space was clean and tidy?
4. Location: How did guests feel about the listing neighborhood?
5. Check-in: How smoothly did guests’ check-in go?
6. Value: Did guests feel the listing provided good value for the price?

Based on the Airbnb rating categories and listing data set, we chose these six rating categories as predictors, for each rating score ranging from 0 to 10.

### 

### Hypothesis

Considering the Superhost requirements and review schema, we built our hypothesis about the association between being a Superhost and six review categories.

H0: There is no association between being a Superhost and each of the six review categories.

H1: There is an association between being a Superhost and each of the six review categories.

### Model Selection: Logistic Regression

In order to assess the factors that determine whether the host can qualify to be a Superhost, we chose to use *host\_is\_superhost* as response and six rating categories as predictors. Variables used are summarized in Table X. Since the response *host\_is\_superhost* is a binary response, we selected to apply the Logistic Regression model to our analysis. The assumptions for this model will be assessed in the next section.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Type** | **Feature Name** | **Data Type** | **Comments** |
| Response | *host\_is\_superhost* | Numeric | 0 or 1 |
| Predictor | *revew\_rating\_accuracy* | Numeric | 0 - 10 |
| Predictor | *revew\_rating\_communication* | Numeric | 0 - 10 |
| Predictor | *revew\_rating\_cleanliness* | Numeric | 0 - 10 |
| Predictor | *revew\_rating\_location* | Numeric | 0 - 10 |
| Predictor | *revew\_rating\_checkin* | Numeric | 0 - 10 |
| Predictor | *revew\_rating\_value* | Numeric | 0 - 10 |

*Table X: Variables Summary for Superhost Logistic Regression Model*

### Model Assumptions Assessment

**Binary dependent variable**

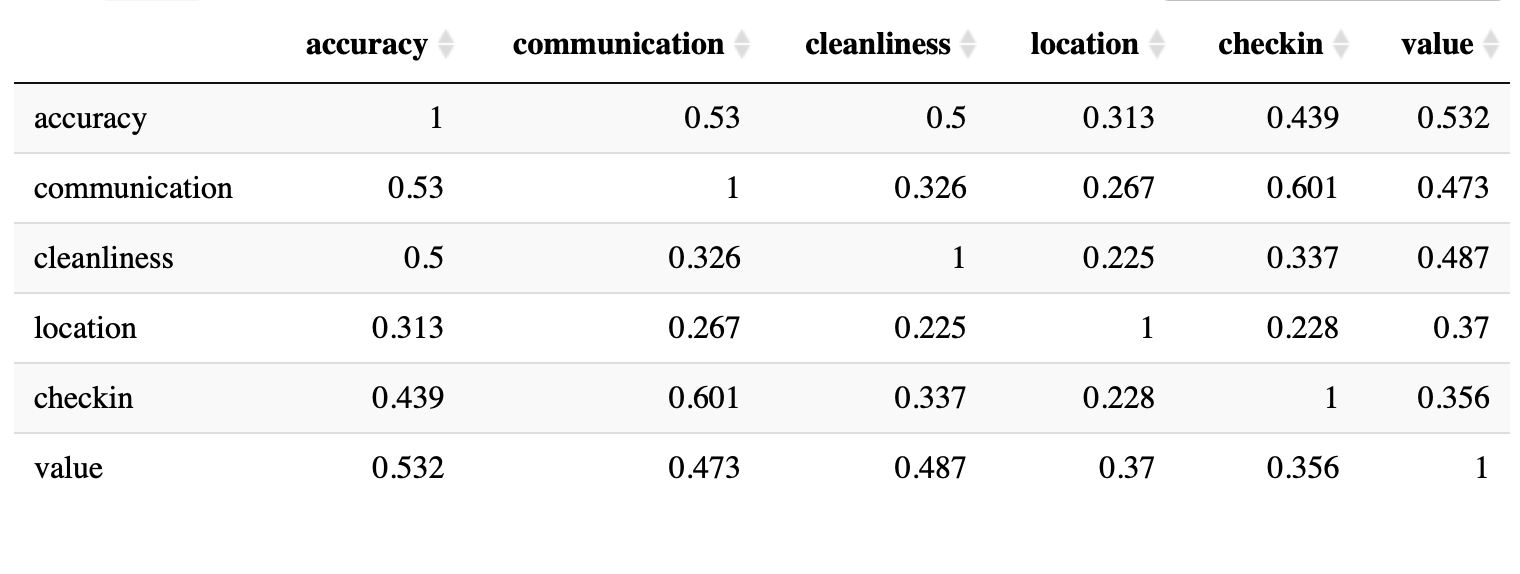
Satisfied.The *host\_is\_superhost* are binary response variables 0/1, with 0 representing “not Superhost”, and 1 representing “Superhost”.

**Independence**

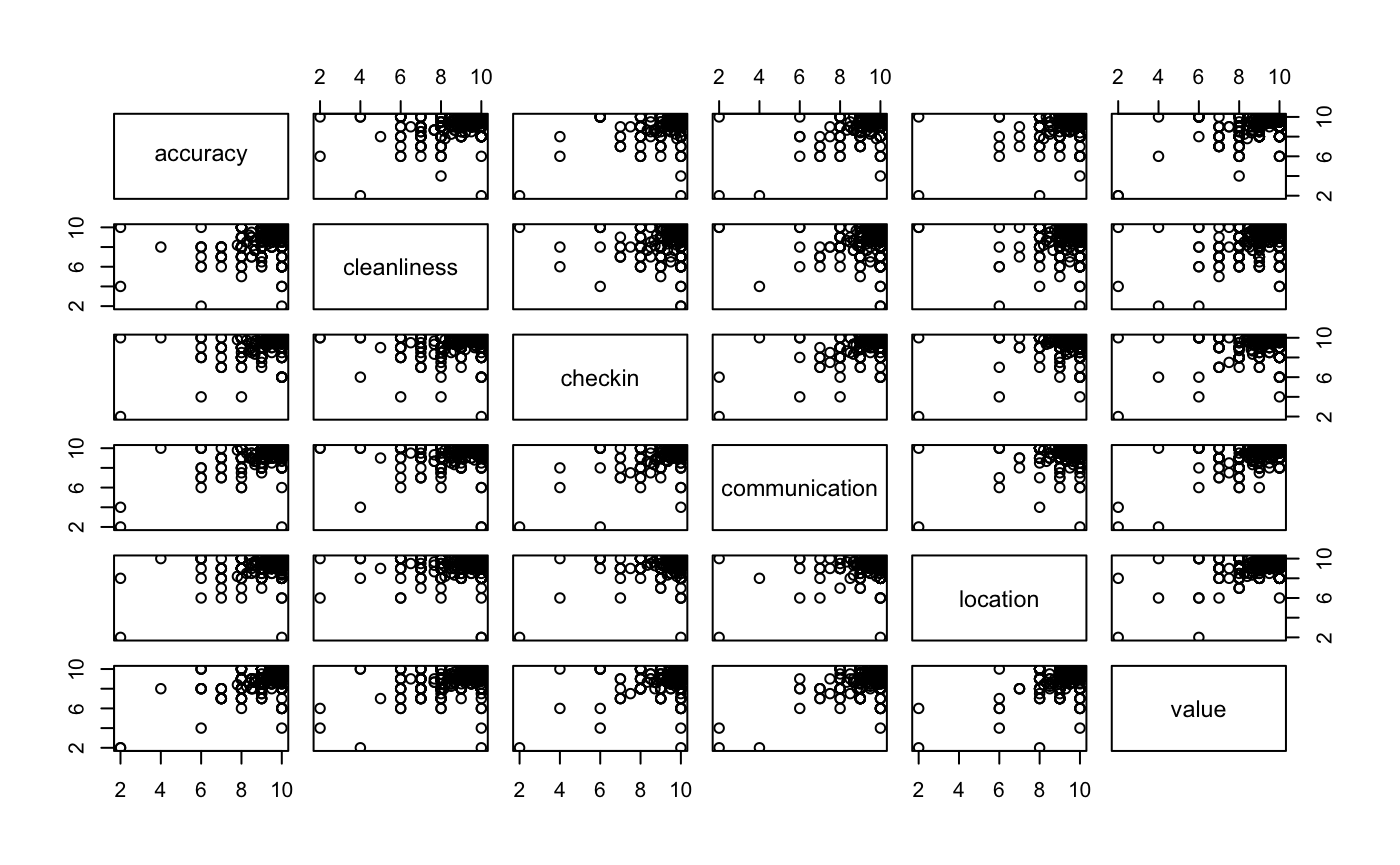
Satisfied. The data set comes from an observational study, we could assume that each booking is independent of each other. So the observations are independent of each other.

**No strong intercorrelations among independent variables**

Satisfied. In order to check the intercorrelations among, we used Pearson correlation table and Correlation matrix plot. From table X, we observed that the largest correlation coefficient is 0.601 between check-in and communication, which indicates a moderate correlation.

*Table X: Pearson Correlation Table*

And from Figure X, we also observed that there is no strong intercorrelation. The assumption is checked.

*Figure X: Intercorrelation Matrix Plot*

**Linearity of independent variables and log odds**

Satisfied. For linearity between the logit of the outcome and each predictor variables, we leveraged the plot function to visualize residuals vs fitted values plot and observed a linear relationship between the logit (logit(p) = log(p/(1-p)), where p is the probabilities of the outcome) of the outcome and each predictor variables.

**Large sample size**

Satisfied. Since our dataset contains hosts that have more than one listing, we manipulated the dataset further: for an individual host, we took the average reviewing score for each of the review categories, thus ensure that the unit of analysis for this question is one observation per unique host. Within 8459 individual listings, we have 5302 distinct hosts. Therefore, the large sample size assumption is checked.

### Results

By performing logistic regression, we got the following coefficient estimate table for six review categories. We computed exponentiated coefficient estimates in order to interpret the association between being a Superhost and different review categories.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Category Name** | **Coefficient Estimate** | **Exp. Coefficient Estimate** | **Std. Error** | **Z value** | **Pr(>|z|)** |
| accuracy | 1.379 | 3.971 | 0.173 | 7.965 | 0.000 |
| communication | 0.893 | 2.442 | 0.275 | 3.249 | 0.001 |
| cleanliness | 0.961 | 2.614 | 0.092 | 10.481 | 0.000 |
| location | 0.253 | 1.287 | 0.087 | 2.895 | 0.004 |
| checkin | 0.946 | 2.577 | 0.237 | 3.987 | 0.000 |
| value | 0.31 | 1.364 | 0.084 | 3.707 | 0.000 |

*Table X: Coefficient Estimate Table*

Since the p-values for all six categories are very small, then we can reject the null hypothesis. There is evidence of the associations between becoming a Superhost and each review category. We can interpret the associations as follows:

Accuracy exp(𝜷1) = 3.971: Holding other predictors fixed, there is a 297.1% increase in odds of becoming a Superhost for per unit increase in accuracy rating score.

Communication exp(𝜷2) = 2.442: Holding other predictors fixed, there is a 144.2% increase in odds of becoming a Superhost for per unit increase in communication rating score.

Cleanliness exp(𝜷3) = 2.614: Holding other predictors fixed, there is a 161.4% increase in odds of becoming a Superhost for per unit increase in cleanliness rating score.

Location exp(𝜷4) = 1.287: Holding other predictors fixed, there is a 28.7% increase in odds of becoming a Superhost for per unit increase in location rating score.

Check-in exp(𝜷5) = 2.577: Holding other predictors fixed, there is a 157.7% increase in odds of becoming a Superhost for per unit increase in check-in rating score.

Value exp(𝜷6) = 1.364: Holding other predictors fixed, there is a 36.4% increase in odds of becoming a Superhost for per unit increase in value rating score.

We ranked these categories by the exponentiated coefficients, which shows a rank of strength in association between being a Superhost and review categories.

|  |  |  |
| --- | --- | --- |
| **Category Name** | **Exp. Coefficient Estimate** | **% Increase in Odds\*** |
| Accuracy | 3.971 | 297.1 |
| Cleanliness | 2.614 | 161.4 |
| Check-in | 2.577 | 157.7 |
| Communication | 2.442 | 144.2 |
| Value | 1.364 | 36.4 |
| Location | 1.287 | 28.7 |

*Table X: Ranking of Significance of Six Rating Categories*

*\* Percentage increase in odds of becoming a Superhost for per unit increase in review score, holding other factors fixed.*

This ranking is reasonable when we consider real life. The first 4 categories, accuracy, cleanliness, check-in, and communication, leads more than 100% increase in the possibility to become a Superhost for each unit increase in rating when other factors are fixed. These 4 factors are more related to how the guests will feel after they live in the space and how smoothly they can interact with the hosts. So they have strong association with becoming a Superhost. For the left 2 factors, value and location, have a lower association with becoming a superhost. Usually, guests should have already evaluated the location or the value of this listing before they make a reservation.

Therefore, it is safer to summarize that Airbnb hosts, who aim to become Superhosts, should well prepare their listings majorly for information accuracy, space cleanliness, check-in procedure, and positive communication. A reasonable price and a great location can attract more guests to make reservations and help to raise the overall satisfaction.

# Conclusion

<Describe your results clearly and concisely. Use graphical displays and tables to convey descriptive information about the data and the results of the analysis.>

# Discussion

<This section should briefly summarize the results and conclusions. Also describe limitations of the analyses, including limitations of the data set as well as of the statistical analyses>

# Reference

<List books or articles you consulted. (It is not necessary to do a lot of background research, so the reference list should be short.) References for statistical methods used in class (e.g., t-tests, and linear regression) are not required, but references should be given for advanced methods not covered in class.>

1. Superhost: <https://blog.beyondpricing.com/how-do-i-become-an-airbnb-superhost/> [↑](#footnote-ref-0)