**DATA 557**

**Final Project – Assignment 3**

**Winter 2019**

**3/2/2019**

**Project title**

Recommendation Model for Sharing Economy Platforms - Case Study on Airbnb

**Group members - Bruins**

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

**Data description**

This dataset 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 dataset 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 dataset is available [here](http://insideairbnb.com/about.html#disclaimers).

**Data availability**

The data set is readily available and ready for further cleaning.

**Questions**

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

**Q2.** Factors for the number of reviews and becoming a superhost[[1]](#footnote-0)

* As a host, what would be the important factors (among cancellation policy, days of availability, price, minimum nights, review scores, years in business, etc) that would get me more reviews to qualify for being a *superhost*.

Follow-up question: Expand our tests to more cities and compare the results.

**Variables**

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:

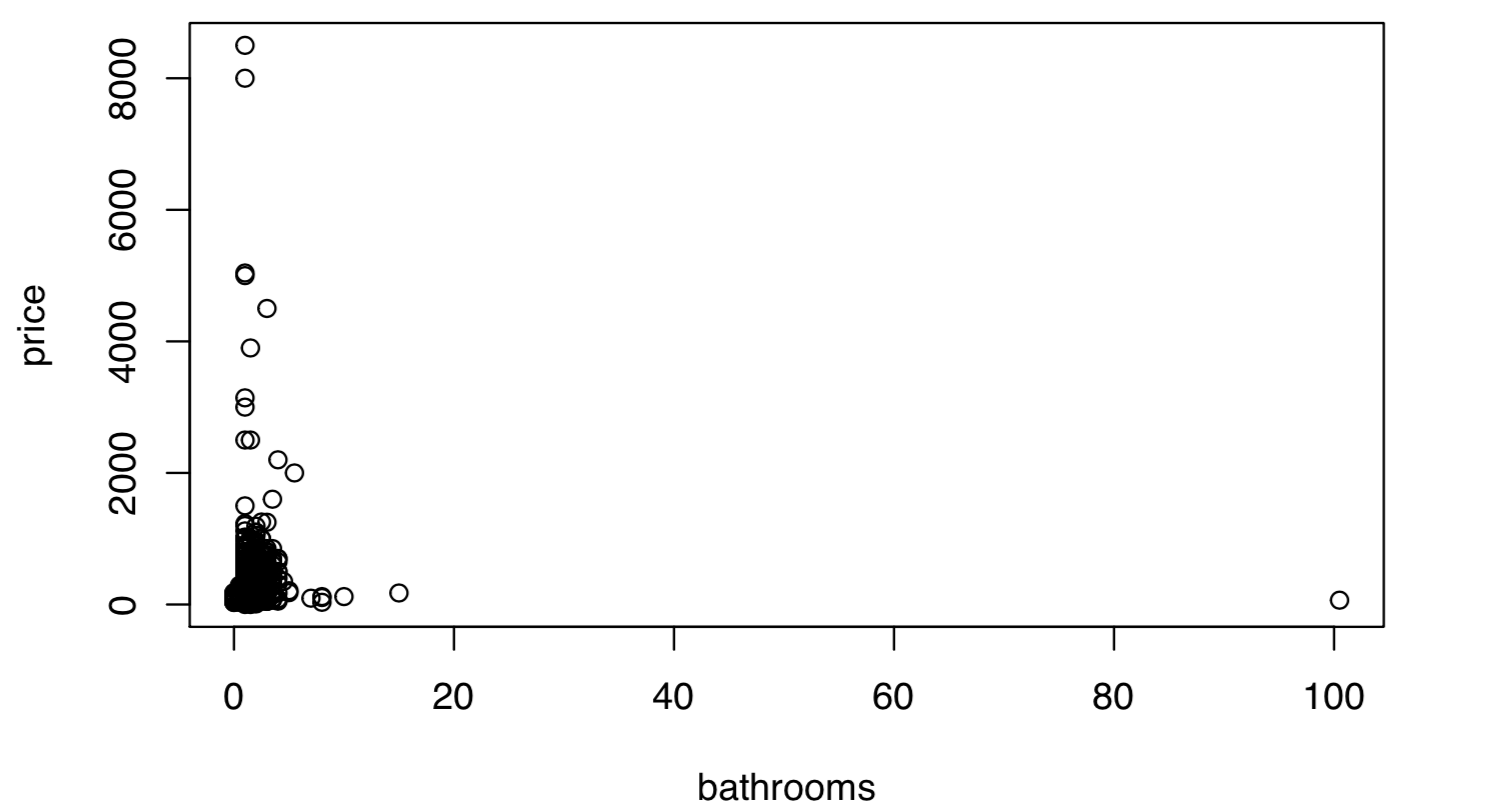
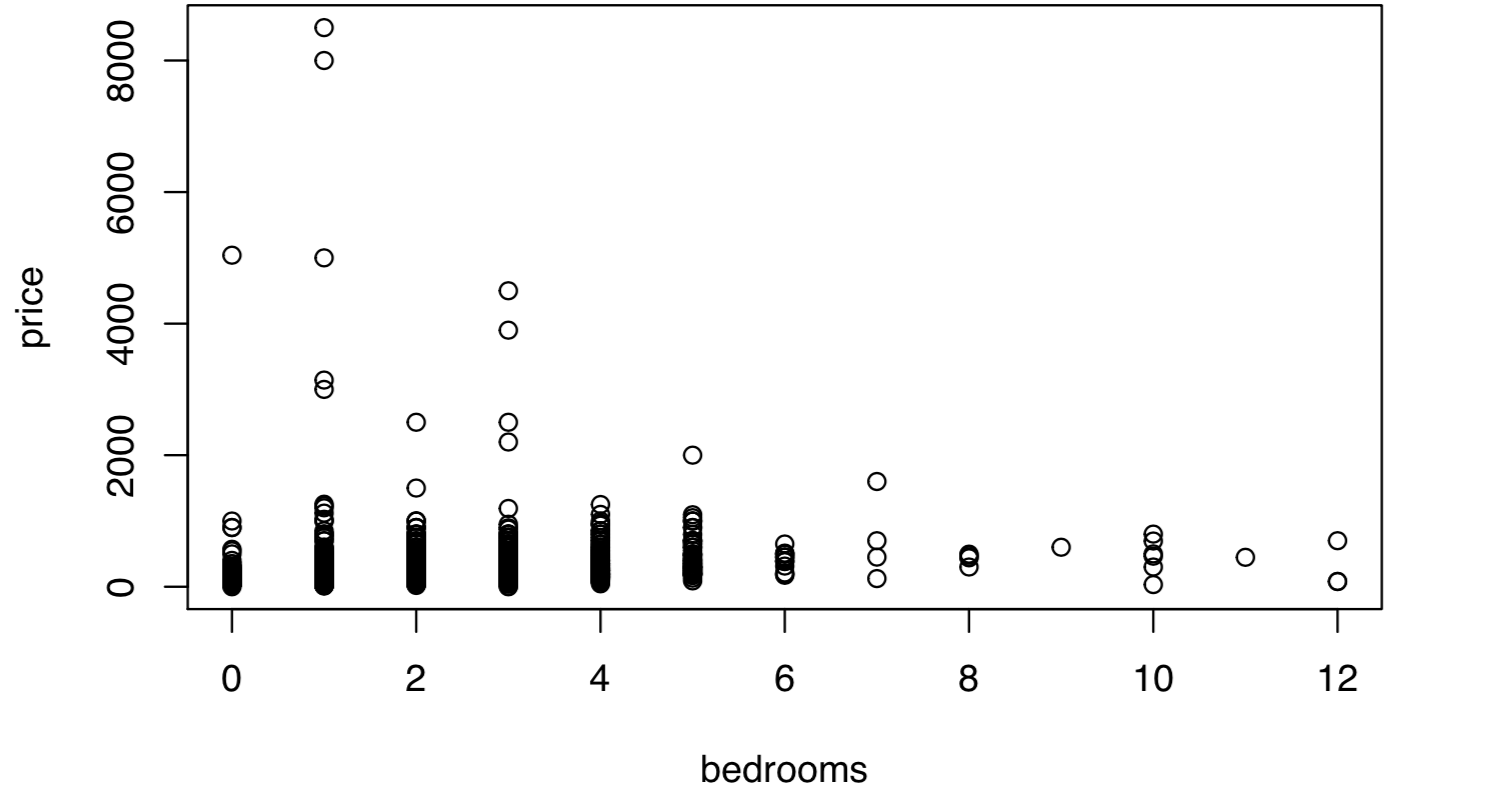
|  |  |  |
| --- | --- | --- |
| **Feature Type** | **Variable Name** | **Main Data Types** |
| Amenities | room\_type, accommodates, bathrooms (number of), bedrooms (number of), beds, wifi, Cable TV, Washer/Dryer | int, factor (“T” or “F”) |
| Extra fee | security\_deposit, cleaning\_fee, extra\_people | 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) |  |

**Possible Analysis methods**

* ANOVA, with and without interactions
* Linear regression against single predictor variables, as well as variables split for interaction effects

For example, apply ANOVA and linear regression to check whether the following features have effect on Airbnb rental price:

* Price ~ Bathrooms
* Price ~ Bedrooms



**Potential problems**

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 dataset also needs to be assessed, tested and reviewed before putting in use. This is caused by the fact that the available definitions for the dataset 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**

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

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

1. Model Selection
2. Model Assumptions Check
3. 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. Model Assumptions Check [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 dataset 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Response variable**   * price | | |  |
| **Predictor variables**   * property\_type * room\_type * bathrooms * bedrooms | * beds * guests\_included * accommodates * review scores | * host\_neighbourhood * review\_scores\_checkin * host\_is\_superhost | * review\_scores\_accuracy * host\_verifications\_reviews * neighbourhood\_group\_cleansed |

**Linearity**

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

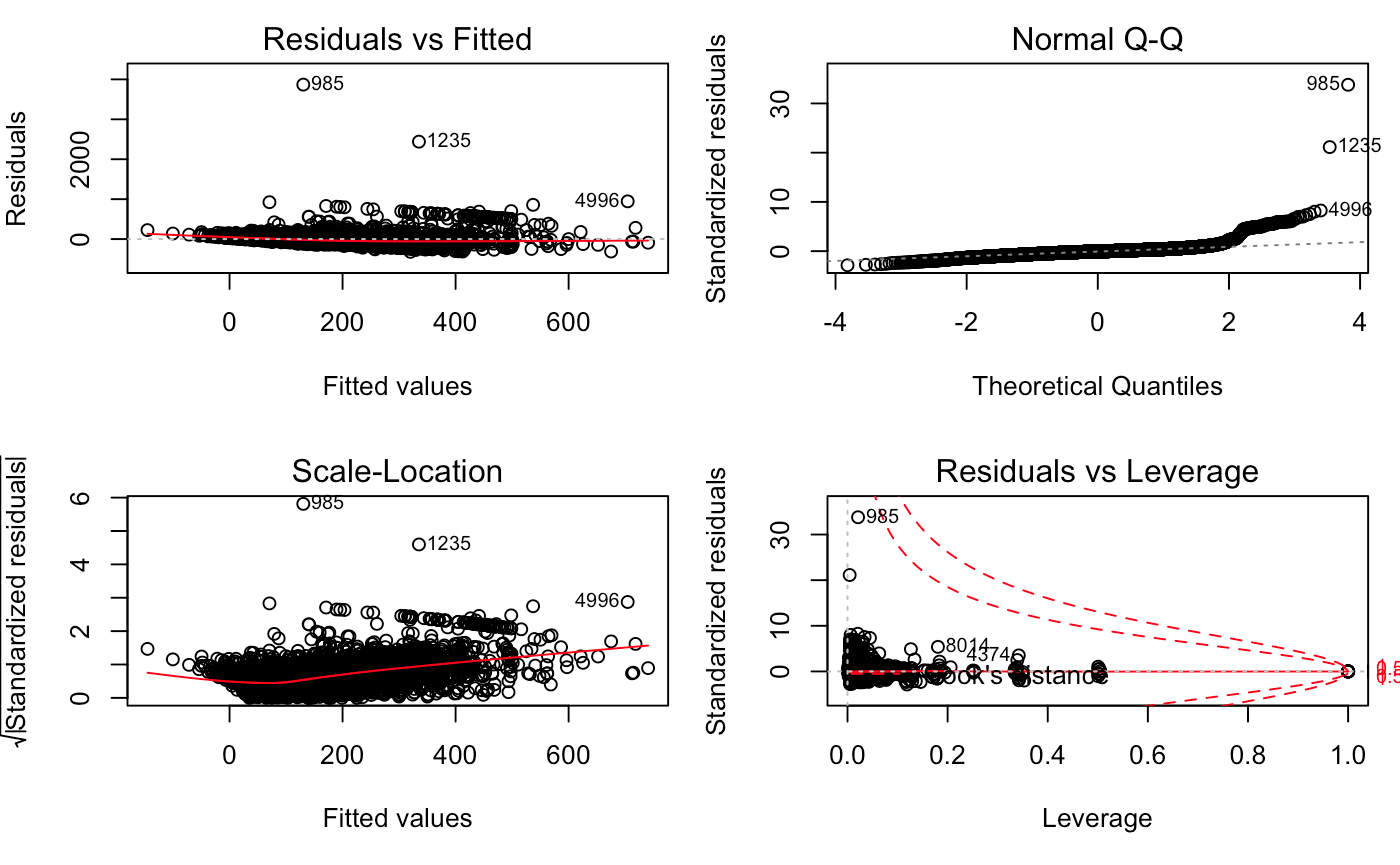
**Constant variance**

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**

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

Figure 1 to Figure 4



|  |  |  |  |
| --- | --- | --- | --- |
| **Response variable**   * review\_scores\_rating | | | |
| **Predictor variables**   * review\_scores\_communication * review\_scores\_accuracy * review\_scores\_checkin * review\_scores\_location * review\_scores\_cleanliness | * neighbourhood * host\_neighbourhood * host\_identity\_verified * host\_is\_superhost * room\_type | * require\_guest\_profile\_picture * require\_guest\_phone\_verification * host\_verifications\_identity\_manual * host\_verifications\_manual\_online * instant\_bookable | |

**Linearity**

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

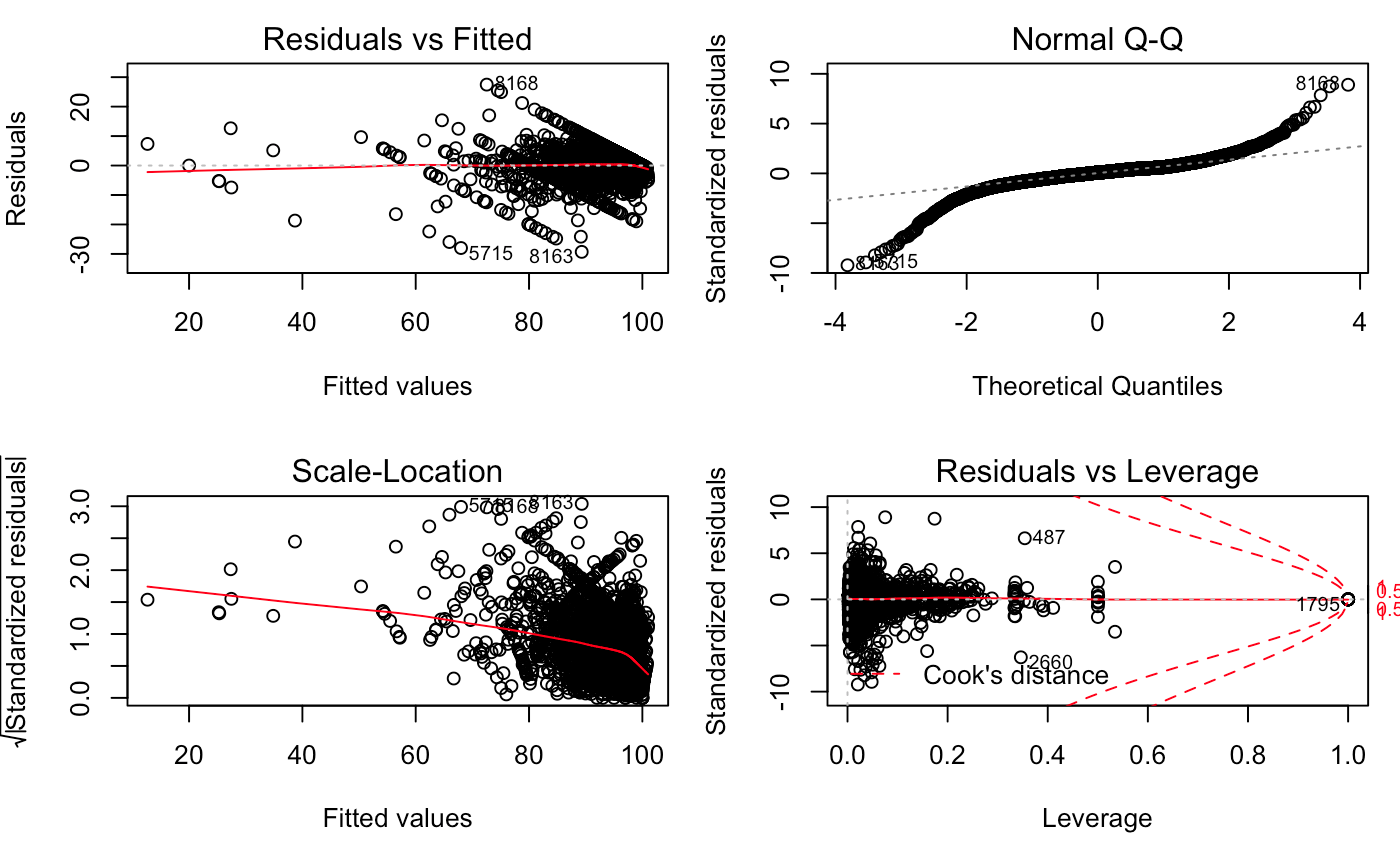
**Constant variance**

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**

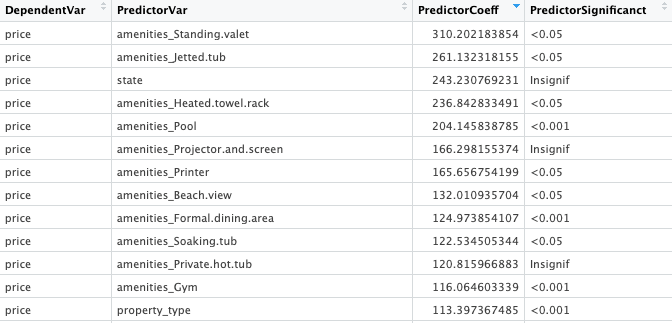
Not satisfied. The Q-Q plot showed non-normal line with heavy heads and tails. However, since our sample size is large enough (8459), we can have CLT to take effect so the non-normality does not affect our method selection.

Figure 5 to Figure 8



**3. Variable Comparison Methodology and Result**

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.

**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.

**Q2. Key Factor for Superhost Qualification**

As listed on the official website, the requirements for being a Superhost are listed as follows :

1. Hosted at least 10 trips
2. Maintained a 90% response rate or higher
3. Received a 5-star review at least 80% of the time being reviewed, as long as at least half of the guests who stayed at the property left a review
4. Completed each of the confirmed reservations without canceling

Based on the current dataset we have, we can not access the number of trips, the response rate (review rate) and the cancellation status. In this study, we focus on analyzing the review scores rating, through several perspectives including accuracy, cleanliness, check-ins, communication, location, value, among which is more important.

For the price and review score, we will go through the three-step process of same as above,

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

**1. Model Selection**

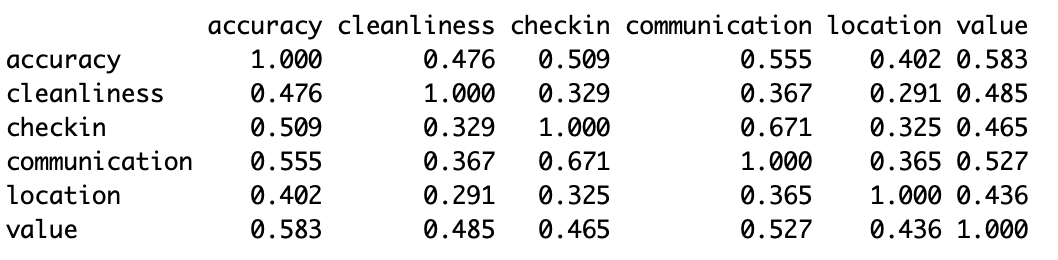
In order to assess the factors that determine whether the host qualifies for a superhost, we choose to use the Logistic Regression model because the host\_is\_superhost is a binary response. We also need to later assess the assumptions of this model.

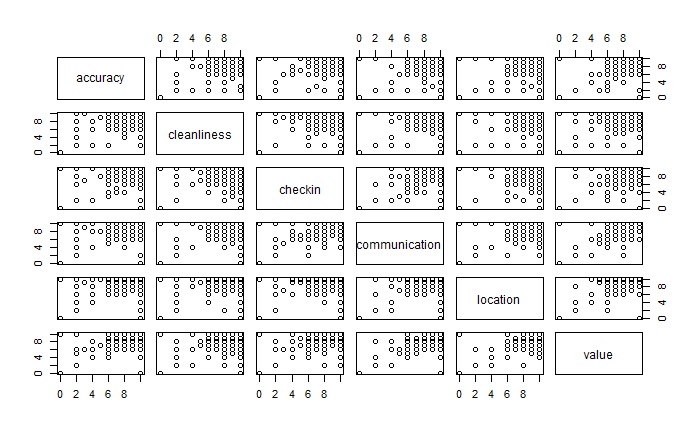
**Sample data**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **superhost** | **accuracy** | **cleanliness** | **checkin** | **communication** | **location** | **value** |
| 1 | 1 | 10 | 10 | 10 | 10 | 10 | 10 |
| 2 | 0 | 9 | 9 | 9 | 10 | 9 | 9 |
| 3 | 1 | 10 | 10 | 10 | 10 | 9 | 10 |
| 4 | 1 | 9 | 9 | 10 | 10 | 10 | 9 |
| 5 | 1 | 10 | 10 | 10 | 10 | 9 | 9 |

**2. Model Assumptions Check [Logistic Regression]**

1. Dependent variable to be binary - Satisfied
2. The observations to be independent of each other - Suppose each reservation were booked independently because customers don’t know each other. So the rating scores are independent.
3. Little or no intercorrelations among independent variables (Use matrix and scatter plot and compute Pearson correlation)

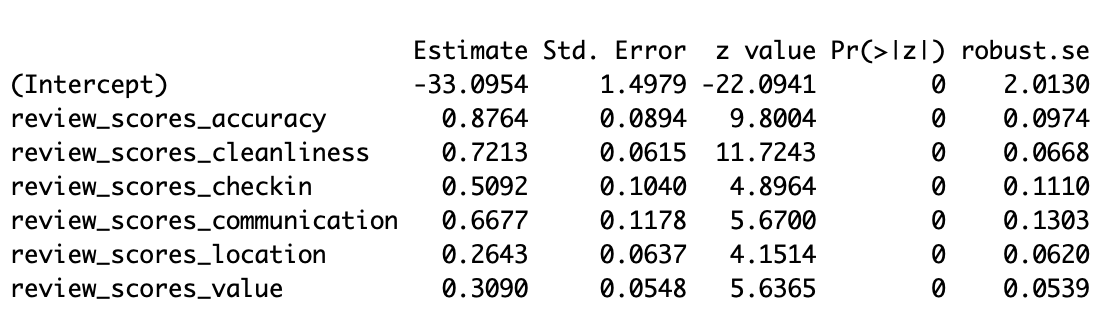




1. Assuming linearity of independent variables and log odds - There is a linear relationship between the logit of the outcome and each predictor variables. Recall that the logit function is logit(p) = log(p/(1-p)), where p is the probabilities of the outcome.
2. Large sample size - 8459 objects

**3. Variable Comparison Methodology and Result**

By performing logistic regression, we expect to get the following coefficient table.

Rank the estimated coefficients to find out the most important factors. In the example table above, we can conclude that hosting information accuracy, cleanliness and communication are the top 3 important factors.

Problem need to be considered: One host can have more than one host listing. For improvement, we need to generate average review scores for each host.

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