Host Listing Analysis for Sharing Economy Platforms

A Case Study on Airbnb

Authors

Adrian Tullock
Cathy Jia
David Wei
Monique Bi
Naman Sharma

Abstract

This project describes the approach of analyzing Airbnb host listing data and aims to determine key predictor variables for listing price, review scores, and superhost status. We start with proposing questions, defining statistical methods, and applying regression models, and eventually analyzing results.

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. 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.

Given the nature of our dataset, we only include the preliminary test results in the current project scope. Further iterative predictor selections are required to derive a feasible prediction model, and accuracy tests are needed to validate our selection.

Introduction

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. In this project, we want to guide Airbnb hosts to maximize their profits and benefits by selecting key factors for a predictive model of price and review scores, as well as providing listing preparation advice to become Superhosts.

The first goal is to determine what predictor variables, among the hundreds available in our Airbnb Seattle listings data, have the highest linear association with listing prices, as well as overall review scores. Simple and stepwise linear regression were used for price and review scores respectively for analyzing these relationships. We aim to 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.

Our second goal is to answer which factors were key to attaining superhost-eligible ratings, given Superhost is a desirable platform achievement rewarded with elevated visibility. We extracted host data and review score data from the listings data set and employed logistic regression for this analysis. We aim to assess the strength of associations for each of the review categories and determine which factors most help hosts to qualify for Superhost status.

Data Set

Data Source and Availability

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 the webpage http://insideairbnb.com/about.html#disclaimers.

Data 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.

As presented, the listing data is fairly mature and useable for analysis. However, the amenity offerings are aggregated in a single column 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.

Part 1a - Key Factors for Price

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 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.

Key Factors Selection

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 consideration the values brought from the host itself. We believe the super_host status and host response time can boost the listing price.

RESPONSE VARIABLE		
Variable Name	**Data Type**	**Comment**
price	float	Daily rental price
PREDICTOR VARIABLES		
Variable Name	**Data Type**	**Comment**
accommodates	int	Number of guests accommodated
bathrooms	int	Number of bathrooms
bedrooms	int	Number of bedrooms
cleaning_fee	float	Amount charged for the cleaning fee

Table 1a.1: Variables Summary for Price Linear Regression Model

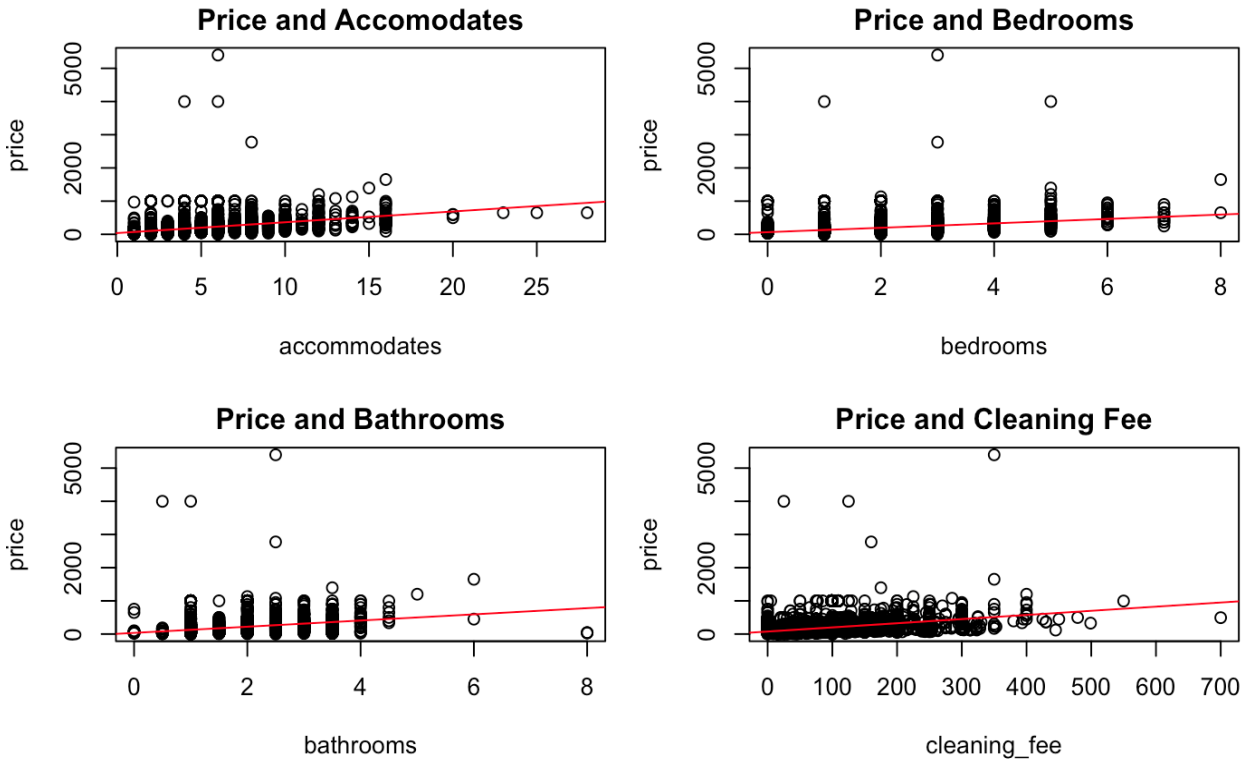


Figure 1a.1: Residual Plots for Linear Regression

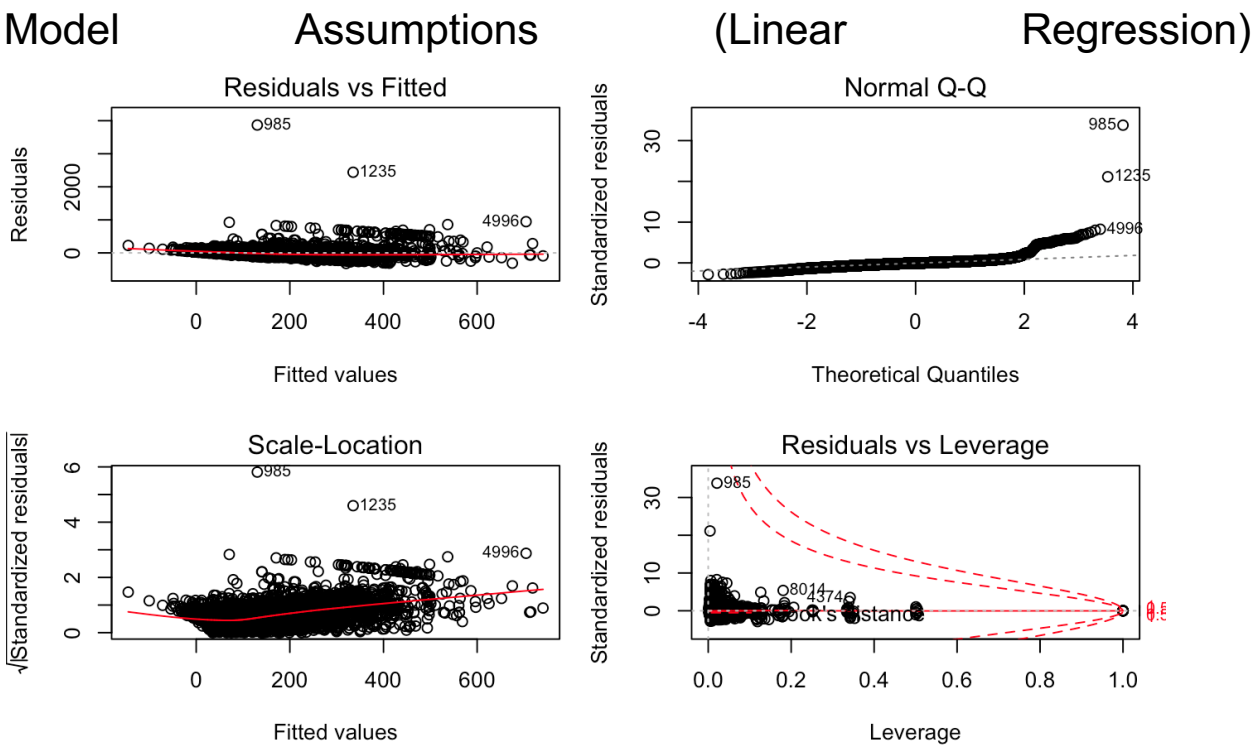


Figure 1a.2: Residual Plots for Linear Regression

From Figure 1a.2, the Residuals vs Fitted plot shows the residuals mostly hovering around the zero line. It is safe to assume a linear relationship between the predictors and their response. The deviation of residuals from the zero line remains consistent over the range of fitted values, suggesting equal variance across the error terms. The severely limited number of extreme outliers adds further confidence that the model is viable.

The Q-Q plot shows a mostly-straight line, albeit with a jump around theoretical quantile 2. The line continues in a straightforward fashion until a few outliers at the very end appear. As perfect normality is a rarity, this result does not stray far enough from the normality assumption to reject it. Also, given the large size of our sample (8459), we do not have to take normality strictly.

The Scale-Location plot is not completely straight, but there is also no distinctive pattern present. Summarily speaking, the line is horizontal enough to work under an assumption of homoscedasticity. The standardized residuals vary most where the fewest number of data points gather. Again, given that there are a very limited number of extreme outliers, this provides further confidence that the chosen model is viable.

The Residuals vs Leverage plot has a few outliers, but appears to only have one point outside of Cook's distance lines. Also, given that this point is shown to have approximately 0 residual, its removal would likely result in a minimal change of the slope coefficient. Although this point appears to have high leverage, it is not interpreted to be very influential. Further analysis could reveal that this point is affecting the statistics in a way that alters our interpretation of results. For the purposes of prediction, this is a possible candidate for removal to procure more desirable responses.

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 a significant p-value

Results

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	15.9092254	3.81202822	4.173428	3.030574e-05
accommodates	16.5744129	1.20906101	13.708500	2.556050e-42
bedrooms	4.0609667	2.82980637	1.435069	1.5130450e-01
bathrooms	22.4836387	3.39421351	6.624109	3.707020e-11
cleaning_fee	0.7046361	0.03536567	19.924295	2.356354e-86
Model Statistics				
Residual standard error	**Multiple R-squared**	**Adjusted R-squared**	**F-statistic**	**p-value**
146.8 (8451 DF)	0.2435	0.2432	680.1	2.2e-16

Table 1a.2: Results table for Price

Analysis of the model predicting price via predictors of the number of bedrooms, bathrooms, guests accommodated, and the cleaning fee yielded a residual standard error of 146.8 and an adjusted R-squared value of 0.2432. This leads to an interpretation that there is much variability unaccounted for by this model. A few likely reasons are the existence of a few extreme outliers, the lack of relevant predictors in the model, and possibly some interactions being unaccounted for. It is also possible that the selected predictors are not the ideal combination, as some collinearity may exist.

Regarding the coefficients, the bedrooms coefficient alone did not attain significance for its estimate. This implies it potentially has little to no value in predicting listing price. That said, a possible interpretation is that the base price is estimated at $15.90 for a listing without bedrooms or bathrooms, no cleaning fee, and that accommodates no guests. On average, respectively while holding all other predictors constant, each unit increase in the number of guests accommodated increases the price by $16.57, unitary increase in bedrooms increases price by $4.06, unitary increase in bathrooms adjusts price by $22.48, and changes in the cleaning fee adjust the price by a 0.70 multiplier on that factor.

One possible explanation of why a unitary increase in bathrooms has the highest price increase is that the number of bathrooms could be a real indicator of number of guests to accommodate. People traveling with a large group might prefer a bit more privacy when they have their own private bathroom. The price thus will be higher, as each additional bathroom can add more value to the entire group when the number of bedrooms stays the same.

As for the low coefficient of cleaning fee, it makes sense as not every host charges cleaning fees. Many hosts tend to use low cleaning fees to attract potential tenants, whereas some hosts use high cleaning fees to encourage longer stays of their property. It's reasonable to treat cleaning fee as a trivial factor of price and exclude that from our future iterations of predictor selection.

Discussion and Future Steps Required for Prediction

If the assumptions of this model are to be taken, then higher listing prices are most associated with higher bathroom counts and the capacity to accommodate more guests. That said, we are hesitant to say we have determined the ideal primary effect. Nonetheless, we feel we have a good foundation to build on and that a viable predictive model is certainly feasible.

In selecting secondary effects, we must be careful to choose predictors that are predictive of the outcome yet are not potentially dependent on it. Educated exclusion in adherence of statistical evidence will help reduce bias in our conclusion respectful to dubious relationships. Empirical selection will be key in determining our secondary effects, and analysis will be used to help check selection bias.

The predictive model we used is a multiple regression model that accounts for the number of bedrooms, the number of bathrooms, the cleaning fee, and the number of guests accommodated. At the moment, it does not account for interactions and/or extreme outliers. Based on data seen previously, the model applies the computed regression equation to new data to determine an estimated listing price. At this stage, it is little more than a proof concept. In order to generate a more viable predictive tool, more testing needs to be done around the available predictors; of which there are hundreds.

The results for this question will serve as the basis for choosing a primary effect. Selecting secondary effects will involve iteratively analyzing plots comparing predictors versus residuals. Additional predictors that extend the tool's ability to explain the variability in the response will be added. Other steps that can help generate more precise estimates will be to test and account for interactions, confounding relationships, and predictors that either associate or highly correlate with the response. A backward selection procedure may be a more comprehensive approach to the latter step.

The final consideration for a predictive model is that other factors less associated with a specific listing could be more relevant to prices, such as seasonal effects, competitive pricing and proximity to events and attractions. For instance, listings close to amusement parks or tourist attractions may see pricing adjustments during peak seasons of operation or travel. These are noted as considerations for future work.

Part 1b - Key Factors for Review Scores

Model & Key Factors Selection

The LR model selection process for the factors was split into 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, so that can be assumed to be independent. Post this initial selection, a forward/backward elimination process boiled it down to a best-fit model with a sub-selection of the variables, through targeted Akaike Information Criterion minimization.

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

Table 1b.1: Variables Summary for Review Score Linear Regression Model

Model Assumptions (Linear Regression)

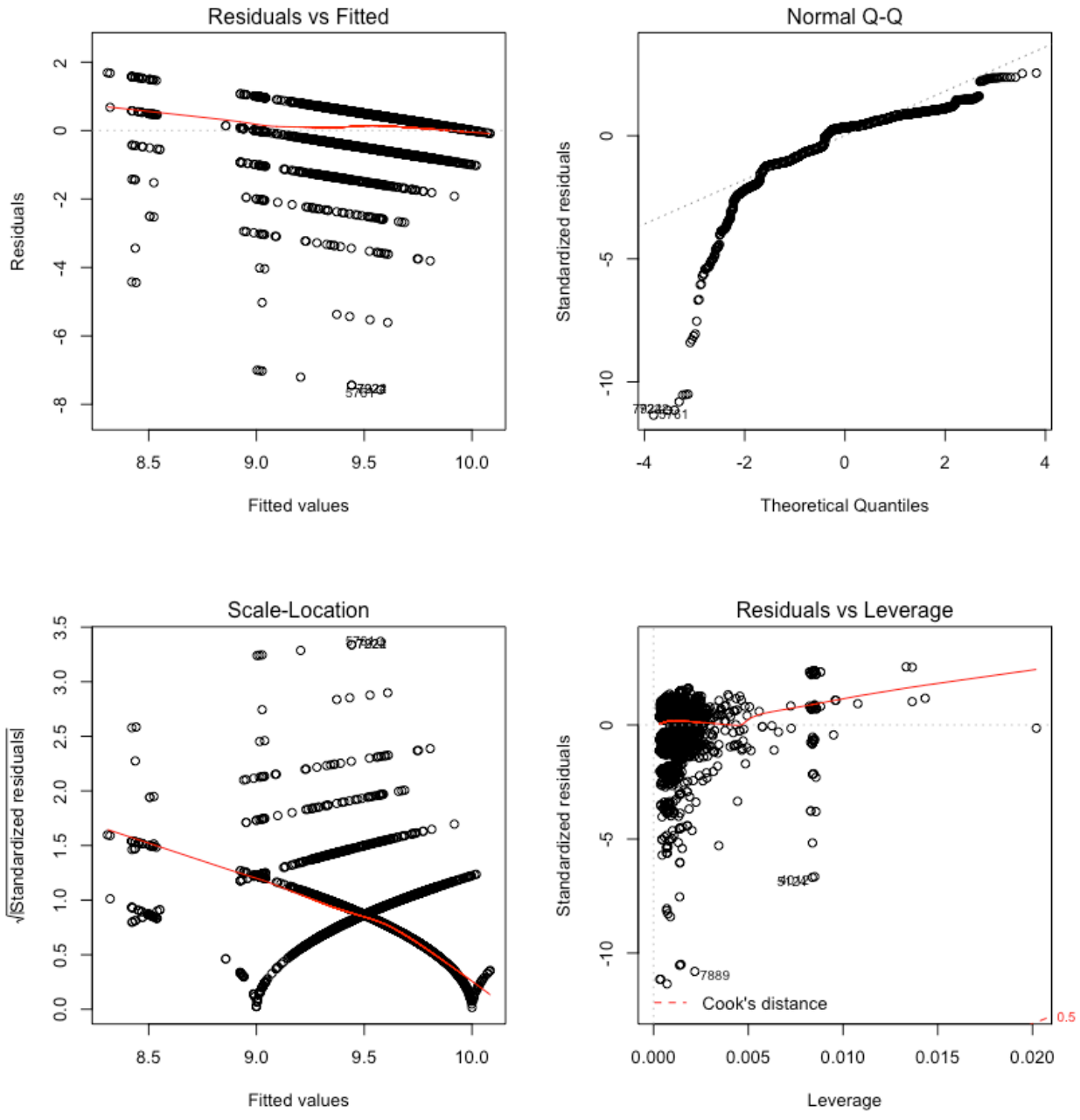


Figure 1b.1 Residual Plots for Linear Regression - Review Scores

The Residuals vs Fitted plot appears to have multiple tracks of linearity on a decreasing slope. The pattern can partially be explained by the impossibility of some fitted values; as the observed values only take integers. The fitted line is partially straight, which is a good sign. Also, the presence of some outliers is noticeable, but there doesn't appear to be an extreme deviation from homoscedasticity.

The Q-Q plot shows a significant deviation from normality. There are significant outliers at the left tail and numerous jumps occur as the data progresses to the right. Also, roughly only half of the standardized residuals hover around 0. This is not a safe assumption and will certainly affect the interpretation of results. The large sample size may be able to offset the non-normality.

A few distinctive patterns are observed in the Scale-Location plot, making it a challenge to interpret. Part of this is explained by the fact that observed values only take integers from 0 to 10; making some fitted values impossible to attain. The fitted line is not at all ideal, but we decided to offset this evidence against some confidence provided by the Residuals vs Fitted plot in Figure 1b.1.

For this model, the Residuals vs Leverage plot maintains all points within Cook's distance lines. There is little concern here regarding points having impactful leverage and influence. This is expected in part because of the limited range of responses, and the clustering of a large percentage of data points.

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.

Results

```
Step:  AIC=-6000.03
review_scores_value ~ host_is_superhost + host_listings_count +
    host_response_rate + accommodates + bedrooms + amenities_total

                     Df Sum of Sq    RSS     AIC
<none>                             3310.4 -6000.0
+ guests_included      1     0.391 3310.0 -5998.9
+ price                1     0.145 3310.3 -5998.4
+ minimum_nights       1     0.019 3310.4 -5998.1
- amenities_total      1    10.142 3320.6 -5979.3
- bedrooms             1    14.384 3324.8 -5969.8
- accommodates         1    24.434 3334.9 -5947.4
- host_response_rate   1    34.220 3344.6 -5925.6
- host_listings_count  1    83.507 3393.9 -5816.8
- host_is_superhost    1   169.565 3480.0 -5630.7
```

	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 ***
Model Statistics				
Residual standard error	**Multiple R-squared**	**Adjusted R-squared**	**F-statistic**	**p-value**
0.6676 (7427 DF)	0.1142	0.1134	159.5 (6 and 7427 DF)	< 2.2e-16

Table 1b.2 Results for Review Scores

The review scores model managed to attain significance for all of its predictors comfortably. Interpretation of the results though, requires some prerequisite understanding of the data. The residual standard error indicates an average deviation of just over half a point of a score for fitted values. This interpretation, however, and the rest that follow should be mindful of the underlying assumptions.

The low adjusted R-squared value of 0.1134 informs us that much of the variability in the data is unaccounted for. This would be problematic for purposes of predictability. The choice of predictors, lack of other relevant ones, and weak adherence to assumptions are some of the more probable candidates for this result. There is also some confounding that requires mitigating and some relationships in need of more investigating.

While the coefficient for superhost status is the most correlated with changes in the response per unit difference, it is also the most suspected of confounding. The number of bedrooms was the next most associated predictor per unit difference. The number of listings and the response rate of the host were least associated with overall review scores and were seen to have a negative relationship in this model.

On its face, one interpretation for this model is that superhosts have an average 0.345 higher review score rating over non-superhosts when all other predictors assume 0. For the remaining predictors, we interpret them as the average effect while assuming all other predictors constant. Each additional listing negatively impacts a host's review score by ~0.0007, their response rate has an effect of -0.00185 per unit increase, and each additional guest accommodated drops the score on average by 0.0431. Bedroom counts and the total number of amenities offered average increases to the review score by 0.074 and 0.00387, respectively.

Discussion and Future Steps Required for Prediction

There is an average of 45 reviews per listing for the data set, which gives us a layer of confidence in the findings. One interpretable finding being that Seattle Airbnb hosts are very competent. The intercept is just over 9.5 on a response scale of 0 to 10. This somewhat positions the more relevant question to be what can hosts do to keep from having their scores docked instead of how to raise them.

Regarding potential as a predictive model, this one has fundamental issues to iron out. This is compiled on top of a low adjusted R-squared value. Of the selected predictors, as well as many not selected, there is an inherent confounding element due to unavoidable user selection bias. However, we proceed with this premise understanding the nature of an observational study.

The challenges in building a predictive model for the question of review scores are both, plentiful and complex. For instance, it is plausible that a relatively low price for a listing correlates with high-value ratings. On the other hand, it may not. It is not expected that users will book listings they deem to have low value in the first place. Therefore, their perception of value in comparison to other listings is filtered for prior to the user selecting the unit they will eventually rate. Also, the selected predictor of superhost status is considered confounding since becoming a superhost is partially dependent on high review scores. Only because it is not the only factor determinant of superhost status did we justify its inclusion in our test. It would likely be disqualified as a predictor in a viable prediction model.

Similar to the pricing model, this one currently does not account for interactions and/or extreme outliers. This does not qualify as a proof of concept, but we are not certain one is unattainable based on these results and conditions. In future work on this topic, a better approach may be to investigate individual review score categories prior to testing on the overall user rating.

Part 2 - Factors for becoming a Superhost

Background

As a highly valued status, Superhost status is checked every three months with the following criteria[1].

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.

Requirements 2, 3 and 4 can be easily quantified, while the first requirement is related to various factors. Therefore, we decided to concentrate on review ratings to help hosts better prepare their listings in our Superhost analysis. As listed on the website, Airbnb has six rating categories [2].

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?

These six review rating scores are accessible in our listing data. So, we processed the listing data to our Superhost analysis, which will be described in later section.

Hypothesis

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

H_0: There is no association between being a Superhost and each of the six review categories.

H_1: There is an association between being a Superhost and each of the six review categories.

[1] Airbnb Superhost Criteria: https://www.airbnb.com/superhost
[2] Airbnb Rating Categories: https://www.airbnb.com/help/article/1257/how-do-star-ratings-work

Factors Selection and Data Preparation

In order to assess the factors that determine whether the host qualifies to be a Superhost, we chose to use *host_is_superhost* as the response and the six rating categories as predictors.

RESPONSE VARIABLE		
Variable Name	**Data Type**	**Comment**
host_is_superhost	factor	0 or 1
PREDICTOR VARIABLES		
Variable Name	**Data Type**	**Comment**
review_rating_accuracy	int	0 - 10
revrew_rating_communication	int	0 - 10
revrew_rating_cleanliness	int	0 - 10
revrew_rating_location	int	0 - 10
revrew_rating_checkin	int	0 - 10
revrew_rating_value	int	0 - 10

Table 2.1: Variables Summary for Superhost Logistic Regression Model

Note that there are 8459 individual listings in the data set, while there are 5302 distinct host IDs. We should consider the case that one host may have more than one listing. Since the unit of Superhost analysis focuses on hosts, we flattened our data by individual hosts. Each unique host's categorical review score is the mean review score for all listings under that host. For example, say a host has three listings with corresponding cleanliness review scores 10, 8 and 9. Then, the mean cleanliness review score for this host is 9. So, 9 is the cleanliness review score we would use for this host.

Model Selection

Since the response *host_is_superhost* is a binary response, we applied the Logistic Regression model to our analysis. The assumptions for this model will be assessed in the next section. For our Superhost logistic regression model, the log-odds equation is:

$$l = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Here, the log-odds l of being a Superhost is a linear combination of predictors x_i with coefficients β_i, where i =1, 2, 3, 4, 5, 6 denotes the six Airbnb review categories respectively for accuracy, communication, cleanliness, location, check-in, and value.

Model Assumptions (Logistic Regression)

The dependent variable *host_is_superhost* is a binary response variable, with 0 representing "not Superhost" and 1 representing "Superhost". Therefore, the binary dependent variable assumption for logistic regression model is met.

Our data set contains hosts that have more than one listing. Therefore, we manipulated the dataset so that an individual host is not accounted for multiple times (as mentioned in the data preparation section). Within 8459 individual listings, we have 5302 distinct hosts. This satisfies the large sample size assumption. Also, since the data set comes from an observational study, we can assume that bookings are independent of each other. This validates our independence assumption for this model.

To check the intercorrelations among the six predictors, we used Pearson correlation table and correlation matrix plot. From Table 2.2, we observed that the largest correlation coefficient is 0.601 between check-in and communication, which indicates a moderate correlation.

	accuracy	communication	cleanliness	location	check-in	value
accuracy	1	0.530	0.500	0.313	0.439	0.532
communication	0.530	1	0.326	0.267	0.601	0.473
cleanliness	0.500	0.326	1	0.225	0.337	0.487
location	0.313	0.267	0.225	1	0.228	0.370
check-in	0.439	0.601	0.337	0.228	1	0.356
value	0.532	0.473	0.487	0.370	0.356	1

Table 2.2: Pearson Correlation Table

From Figure 2.1, we also observed that there is no strong intercorrelation. The check-in procedure could be confusing, although the space may be pretty tidy. Listings with convenient location but extremely negative host communication are also possible. Guests make different ratings for different aspects. So, there are no strong intercorrelations among predictors.

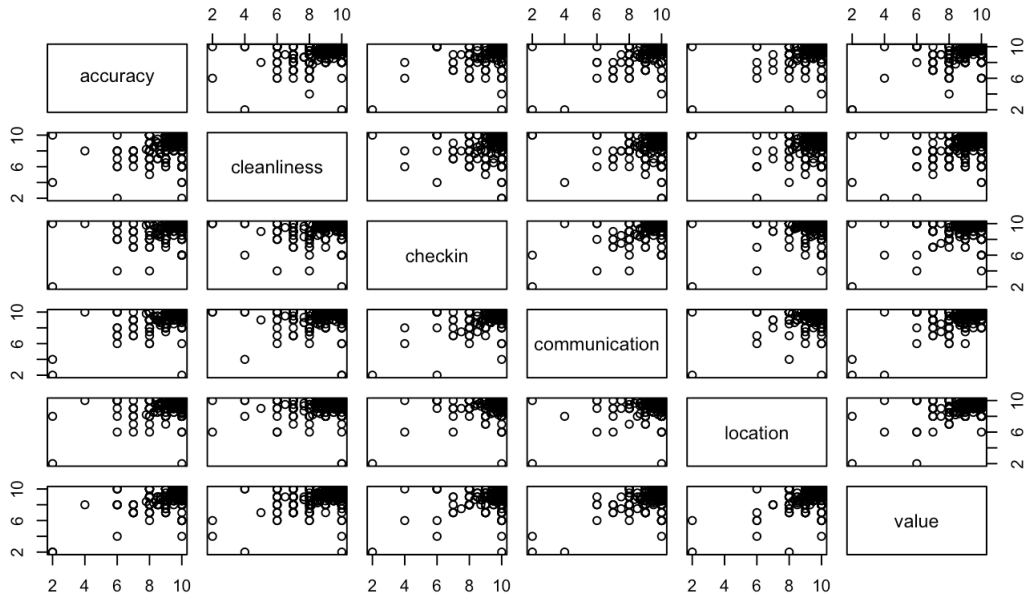
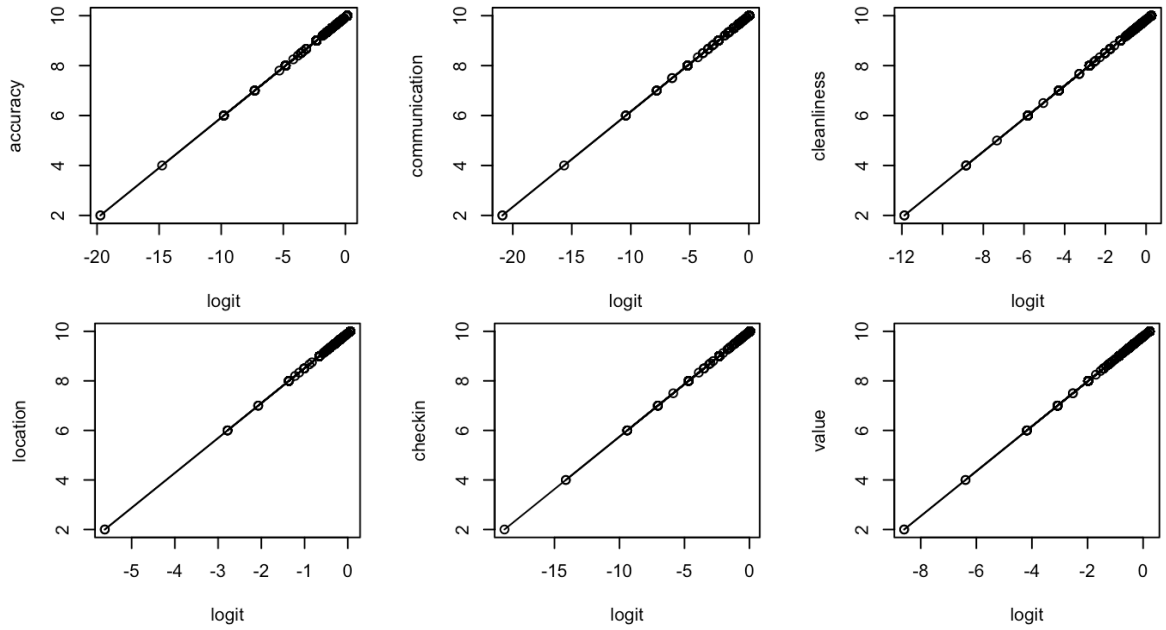


Figure 2.1: Intercorrelation Matrix Plot

The Predictors vs Logit plots in Figure 2.2 show the linear relationship between each predictor variable and the logit of the outcome $logit(p) = log(\frac{p}{1-p})$, where *p* is the probability of the



outcome. So, the linearity of independent variables and log odds assumption is satisfied.
Figure 2.2: Relationships between Predictor Variables and Logit

Results

Our logistic regression model delivered the following coefficient estimate table for the six review categories. We computed the 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
check-in	0.946	2.577	0.237	3.987	0.000
value	0.310	1.364	0.084	3.707	0.000

Table 2.3: Coefficient Estimate Table for Superhost Model

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

Accuracy $\exp(\boldsymbol{\beta}_1) = 3.971$: Holding other predictors fixed, there is a 297.1% increase in odds of becoming a Superhost per unit increase in accuracy rating score.

Communication $\exp(\boldsymbol{\beta}_2) = 2.442$: Holding other predictors fixed, there is a 144.2% increase in odds of becoming a Superhost per unit increase in communication rating score.

Cleanliness $\exp(\boldsymbol{\beta}_3) = 2.614$: Holding other predictors fixed, there is a 161.4% increase in odds of becoming a Superhost per unit increase in cleanliness rating score.

Location $\exp(\boldsymbol{\beta}_4) = 1.287$: Holding other predictors fixed, there is a 28.7% increase in odds of becoming a Superhost per unit increase in location rating score.

Check-in $\exp(\boldsymbol{\beta}_5) = 2.577$: Holding other predictors fixed, there is a 157.7% increase in odds of becoming a Superhost per unit increase in check-in rating score.

Value $\exp(\boldsymbol{\beta}_6) = 1.364$: Holding other predictors fixed, there is a 36.4% increase in odds of becoming a Superhost per unit increase in value rating score.

Category Name	Exp. Coefficient Estimate	% Increase in Odds*
Accuracy	3.971	297.1
Communication	2.442	144.2
Cleanliness	2.614	161.4

Location	1.287	28.7
Check-in	2.577	157.7
Value	1.364	36.4

Table 2.4: Interpretation of Associations with Being a Superhost and Six Review Categories
** Percentage increase in odds of becoming a Superhost for per unit increase in review score, holding other factors fixed.*

Discussion and Future Improvements

After fitting the logistic regression model to 5302 distinct Airbnb hosts, we evaluated the six categories for their associations with becoming a Superhost. The categories with stronger association with being a Superhost are accuracy, communication, cleanliness, and check-in, which lead to more than 100% increase in the probability of becoming a Superhost for each unit increase in rating, while holding other factors fixed.

It is a reasonable outcome when we think about real-life scenarios: after guests make a reservation, the determining factors of the overall satisfaction rate depends more on whether the properties match the guests' expectation (mainly depends on the listing accuracy and cleanliness of the space), and how smoothly they interact with the host(s) (evaluated by the communication with the host(s) and check-in experience). Value and location are the factors guests would have considered before making the reservation in the first place.

Therefore, we conclude that for Airbnb hosts who aim to maximize profits by becoming Superhosts, the model recommends prioritizing information accuracy, space cleanliness, check-in procedure, and communication experience when preparing the listings and properties. On top of that, value-based pricing and a great location (consider that guests have different preferences on location based on their needs) can attract more guests to make reservations and help to raise the overall satisfaction level.

There are some limitations due to data availability and processing. First of all, we created a host data set with 5302 individual hosts, each having one set of review scores. These hosts may have multiple listings in the selected city (Seattle, WA) according to the listing data set. In the current study, we calculated and assigned the average score under each review category from all the listings an individual host owns. However, there may exist more appropriate methods of dealing with the review scores for multiple listings, such as taking weighted averages instead.

Secondly, the analysis for Airbnb Superhost factors can be extended a little further if we have more available data. As mentioned in the Background section, the criteria we are working on is the first one: "Having a 4.8 or higher average overall rating based on reviews from at least 50% of their Airbnb guests in the past year". In this analysis, we focused on the association among different review scores and the qualification of being a Superhost. Based on our current data set, we were not able to evaluate the factors that would ensure hosts getting more than 50% review rate. Our next step is to potentially get access to the Airbnb booking data, which enables us to analyze key factors that would drive a 50% plus review rate.

Lastly, the overall review rating score is not applied in our current analysis. The data set gives a column named 'review_scores_rating', which is on a scale of 0 to 100 that is separate from the six review categories that we included in our analysis. We would assume this factor to be the

overall review score and is one of the determining factors towards superhost qualification. In our next steps, we would hope to evaluate the association among this overall rating and the six individual review categories, as wells as the association between being a Superhost and the overall score.

Conclusion

The overall goal of this study was to find evidence of relationships that can, in turn, help Seattle hosts succeed on the Airbnb platform. The assumed host priorities were maximizing income via price listing, and attaining Superhost status to attain specialized treatment from the company. Since maintaining high review ratings is relevant to becoming a Superhost, review scores were considered germane to host success.

This analysis has provided the foundation for a viable predictive model, particularly in the case of pricing. As mentioned, review score prediction presents a bigger challenge, and potentially a lower value-add; especially given that review scores appear to trend high already. However, through a careful and comprehensive approach, this is also possible. The results on the analysis for becoming a superhost breakdown our recommended strategy in further detail. The short of it, though, comes down to keeping the unit clean and the listing accurate and thorough.

For hosts looking to maximize profit on the platform, becoming a Superhost should be a priority. It rewards the host with preferred visibility, increased confidence from renters, and priority assistance from Airbnb services. Hosts who can benefit from preferred locations for their listings will have some advantage in pricing and superhost viability, partially due to increased demand. Exactly how much more viable this factors in, quantifiably, is a question for future research to answer. Seasonality effects are another potential topic for future work, and we would expect a highly informative answer to that question to be beyond the scope of the data set. However, all of this only becomes possible once better predictive model is successfully created.