MQM Stats Final Project: Finding the Impact of Reviews and Predicting Prices on Airbnb

Team A15
Yehudi Baptiste, Rata Kiewkarnkha, Xinyue Li, Zhipeng Wang, Hirak Bhayani

Abstract

Reviews are said to be important for purchasing decisions on marketplace platforms. We test this claim by analyzing data from Airbnb, a renowned marketplace platform for short-term rentals. Specifically, we look at the relationship between price and reviews to determine if more reviews are associated with higher prices. We also construct a model to predict prices based on a listing's characteristics. We find that the number of reviews that a listing receives has no impact on its price but the overall ratings score it has positively impacts its price.

1. Background

Airbnb, Inc. is a renowned marketplace platform for people looking for accommodations on which hosts can list out vacant rooms or houses to guests for short-term rentals. There are over 6 million listings on the platform, across 100K cities and 191+ countries.[1]

The aim of this study is twofold. First, we want to see if the number and quality of reviews (i.e. ratings) that a listing receives impacts its price. Secondly, we want to see if we can build a model that predicts the price of a listing based on its characteristics.

This analysis was motivated by the assumption that having more reviews confers benefit to listings--by leading to higher prices or more bookings, for example. Indeed, we see a lot of effort put in by owners and Airbnb itself to encourage guests to leave reviews. We would like to see if these efforts are warranted by determining if more reviews are associated with higher prices, while also identifying other characteristics of a listing that impact price.

2. Business Understanding

This project seeks to answer two business questions. The main question is simple: do more reviews lead to higher prices and, if so, what characteristics of a listing are associated with more reviews. The second question is also simple: which characteristics of a listing can help predict its price.

This analysis will attempt to contribute to the larger debate on the value of user generated reviews in the purchasing decisions of customers on online marketplaces such as Airbnb by studying the relationship between reviews and prices for properties in San Francisco. The market for lodging in San Francisco is one of the largest and strongest in the United States.[2] Our model can help hosts predict their potential

[1] https://press.airbnb.com/fast-facts/
[2] https://www.hvs.com/article/7904-hvs-market-pulse-san-francisco-ca

share of this market. The findings from our analysis can also inform Airbnb's views, policies and efforts around reviews. The analysis of this report is not only relevant to the hospitality industry but also to any industry that operates on an online marketplace in which reviews are said to inform buying decisions.

3. The Data

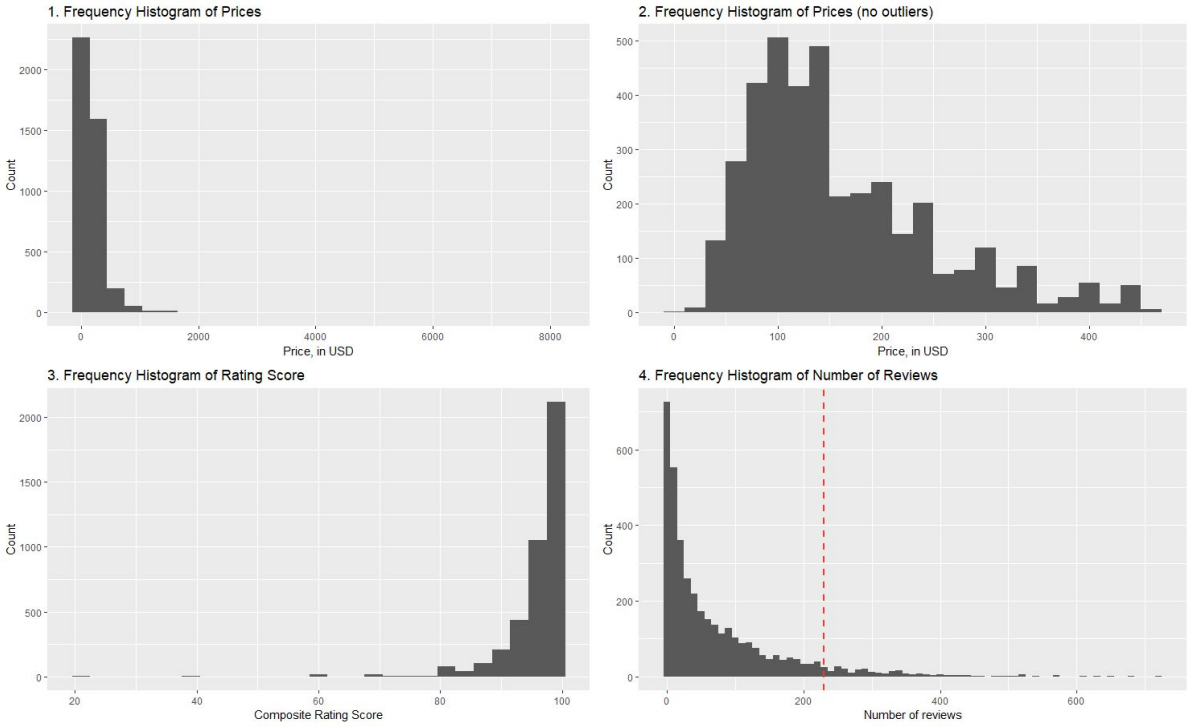
The data comes from insideairbnb.com, an independent platform that scrapes Airbnb listings for data and makes the data available to the public for analysis. Inside Airbnb includes data for many major states, cities, and regions around the world. We downloaded the data for San Francisco, a popular location with a contentious history with Airbnb.[3] One record reflects a listing that was on Airbnb the 8th of July 2019.

The original dataset includes 7,738 observations and 81 variables. We removed irrelevant and redundant columns such as country, which is the same for all records. We also filtered out inactive listings, which we defined as listings that haven't been reviewed in the last 8 months and do not have any availability in the next year.

4. Exploratory Data Analysis

4.1. Number of Reviews, Ratings & Price

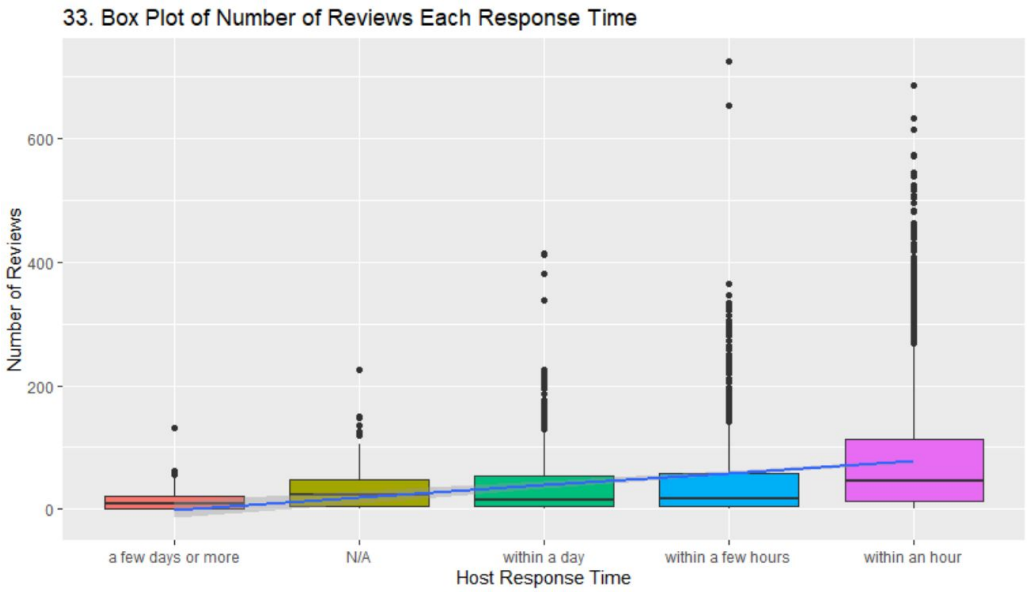
The first step in our exploratory data analysis is to look at the distribution of our three variables of interest: price, number of reviews, and ratings.



[3] San Francisco is Airbnb's home market. It is also the company's most mature market. Despite this fact, the two parties have had disputes. In 2017, the city passed a law that requires hosts of rentals of less than 30 day to be SF residents as well as register with the city.
https://www.airbnb.com/help/article/1849/san-franciscos-registration-process-frequently-asked-questions

4.1.1 Reviews

Reviews is the first dependent variable we look at. Reviews captures the number of written feedback left for a listing after checkout. A review can only be left only after a guest checkouts, meaning that reviews is (likely) highly correlated with bookings/reservations. A guest has up to 14 days to leave a review and only one review can be left per booking party[4]. Looking at figure 4, we see that the distribution of reviews is skewed to the right, with a mean of 70 and an std of 90.



Reviews and response time

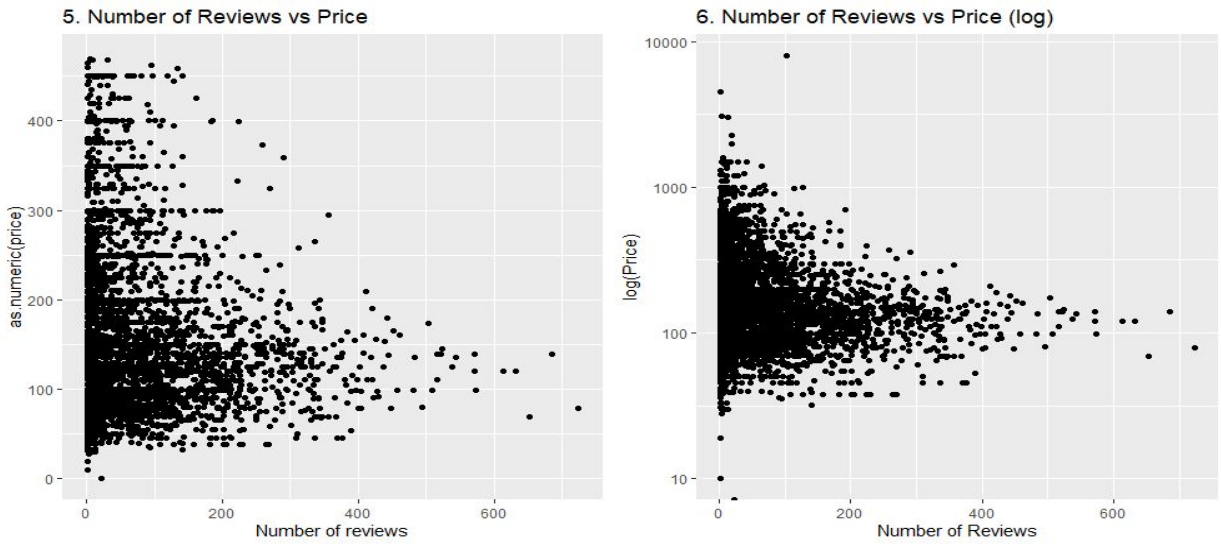
From figure 33 above, the more quickly the hosts respond to the guests, the more reviews the hosts will receive. However, it is also worth noticing that the variation in the number of reviews the hosts received increases when the response time decreases.

4.1.2. Ratings

In addition to written reviews, guests can leave numerical scores evaluating a listing across a number of categories (e.g. value, communication, and cleanliness). The individual scores are combined together to form a composite rating score. The distribution of ratings is skewed to the left, with a mean of 95 and an IQR of 5. Our data is consistent with the notion that ratings left of Airbnb are overwhelming positive, which Airbnb is notorious for.

[4] https://www.airbnb.com/help/article/13/how-do-reviews-work

4.1.3 Price



(footnote: fig.5 shows price with outliers removed)

The price of a listing reflects the cost of a night's stay. In our exploratory analysis, we plot number of reviews on price and see a lot of clustering on the right. This lead us to log transform the two variables and observe no clear relationship between price and reviews.

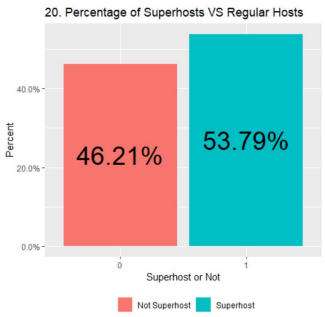
4.2. Highly Reviewed Listings & Superhost Analysis

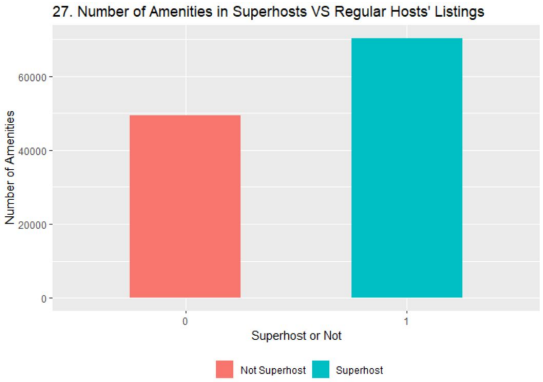
In the second part of our exploratory analysis, we look at how characteristics differ between highly reviewed listings and average reviewed listings as well as identify differences between listings of regular hosts and those of superhosts, who are hosts who receive a badge after meeting certain criteria.

4.2.1. Looking at popular properties : Highly reviewed listings

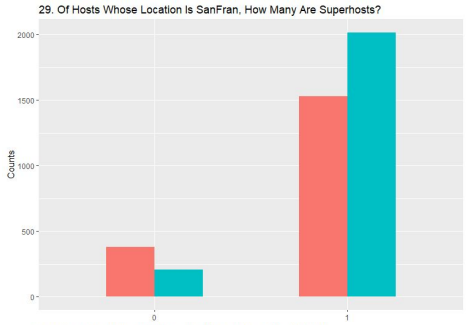
To deepen our exploration, we split our data into two groups: highly reviewed listings and average reviewed listings. We explore how these two groups differ across characteristics. Highly reviewed listings are those with reviews above the outlier cutoff point of 230 reviews. Since the number of reviews is a good proxy for bookings, we view these listings as the most popular places to stay in SF. On average, prices for highly reviewed listings are lower, with a smaller range (figure 22 below). Although this might seemingly challenge the assumption that more reviews leads to higher prices, we reason that this observation is likely explained by the fact that reviews and bookings are highly correlated.

4.2.2. Superhost Analysis

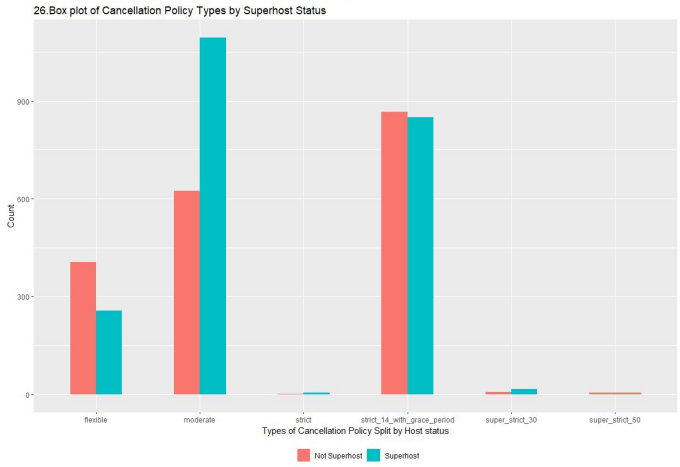




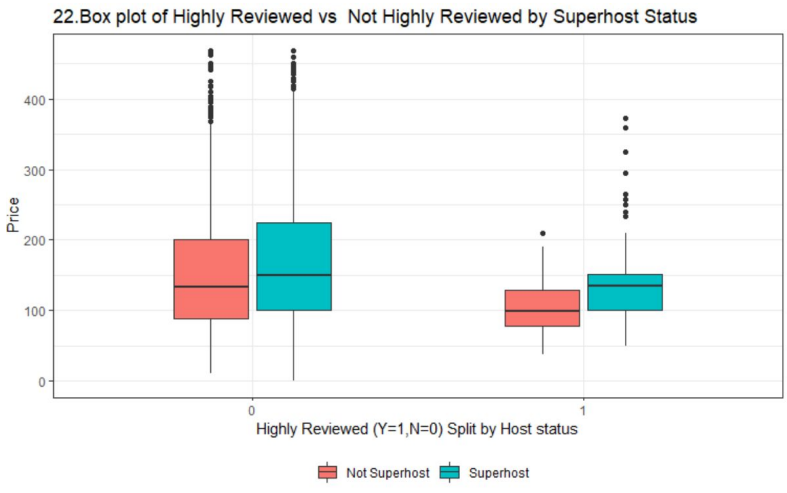
Superhost is a badge awarded to host who meet certain criteria. From figure 20, we see that slightly more than half of listings in SF are owned by superhosts. We can also see that, from figure 27, listings of superhosts offer more amenities.



Of all listings, 85.9% of them belong to a host that resides in San Francisco. And of those 85.9%, there are 489 (2,017-1,528) more listings owned by superhosts as compared to regular hosts. In contrast, of hosts who do not reside in San Francisco, (380-204) most are regular hosts.

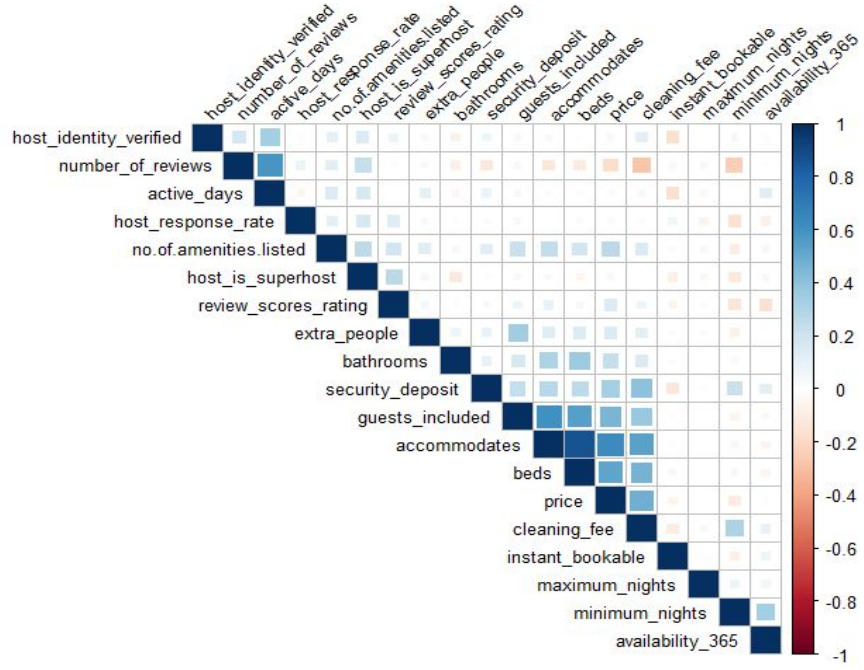


From figure 26, we see that most of the listings have flexible, moderate, and strict-with-grace-period cancellation policies. For moderate, the number of superhosts' listings are sharply higher that regulars' listings. However, for flexible and strict-with-grace-period, there are more listings of regular hosts.



Taking Highly Reviewed part above into consideration, from figure 22 (Cut off price outliers), prices for highly reviewed listings are lower, with a smaller range for both superhosts and regular hosts on average.

5. Regression Analysis



To make sure there is no autocorrelation in our regression model, we first construct a correlation plot to identify independent variables that are highly correlated with one another as well as identify variables that could potentially have predictive power in our models--we confirm the relationships in the correlation plot with cortests. For example, we include the variable accommodates but exclude the variable beds because of its high correlation with accommodates.

Next, we do some grouping. Specifically, we reduce the number of levels for the variables neighborhood and property type by grouping the infrequent levels of each variable into an "others" category (Appendix 3). Specifically, for neighborhood we leave the top 6 and group the rest into the “others” category, and for property type we leave the top 2 and group the rest into the “others”category. This helps produce cleaner and easier to interpret regression outputs.

5.1 Model 1: Predicting Number of Reviews

number_of_reviews ~ bathrooms + active_days + minimum_nights + host_is_superhost + price + accommodates

Table 1: Regression of number of reviews

Variable	Coefficient	P value	Sig
(Intercept)	124.735	0.010	*
bathrooms	-66.645	0.007	**
active_days	0.038	0.018	*
minimum_nights	-2.148	0.019	*
host_is_superhost	40.335	0.068	.
price	-0.155	0.054	.
accommodates	8.989	0.177	

To select variables for a regression to predict reviews, we start with the stepwise method. We use all three approaches (i.e. forward selection, backward elimination, and bidirectional elimination) to pick the model with the lowest AIC.

The model with the lowest AIC includes the variables host is superhost, accommodates, bathrooms, price, minimum nights, and active days. Bathrooms, minimum nights, and active days are statistically significant. Residuals plot shows that homoscedasticity assumption is satisfied.

Statistical significance for active days makes sense. The older a listing is, the more bookings it is likely to have and thus the more reviews it is likely to have. Minimum nights' relationship to reviews is also clear. Listings that require higher minimum-night stays collect less bookings in a year and therefore less reviews. Explaining the relationship between number of bathrooms and reviews is more challenging, however.

5.2 Model 2: Predicting Price

Log(price) ~ host_response_rate + room_type + accommodates + bathrooms + security_deposit + minimum_nights + number_of_reviews + review_scores_rating + cancellation_policy + active_days + neighbourhood_cleansed

Table 2: Regression of price (log transformation)

Variable	Coefficient	P value	Sig
(Intercept)	1.861	0.152	
host_response_rate	0.541	0.090	.
room_typePrivate room	-0.367	0.004	**
accommodates	0.132	0.000	***
bathrooms	0.320	0.004	**
security_deposit	0.000	0.242	
minimum_nights	-0.016	0.000	***
number_of_reviews	-0.001	0.058	.
review_scores_rating	0.035	0.008	**
cancellation_policymoderate	-0.770	0.035	*
cancellation_policystrict_14_with_grace_period	-0.727	0.049	*
active_days	0.000	0.042	*
neighbourhood_cleansedBernal Heights	-0.389	0.029	*
neighbourhood_cleansedCastro/Upper Market	-0.014	0.916	
neighbourhood_cleansedDowntown/Civic Center	-0.044	0.905	
neighbourhood_cleansedMission	0.077	0.568	
neighbourhood_cleansedSouth of Market	0.198	0.303	
neighbourhood_cleansedWestern_Addition	-0.067	0.752	

We use the step function again to construct a model for price . Before interpreting our results, we test the assumptions of our model. By looking at the residuals plot for price, we find that the homoscedasticity assumption is not satisfied. (See Appendix 4). This leads us to perform a log transformation on price, and run the step function again to construct the best model for log(price). This regression seems to satisfy the homoscedasticity assumption better (See Appendix 4). The new regression includes host response rate, security deposit, room type, accommodates, bathrooms, minimum nights, number of reviews, review scores rating, cancellation policy and active days. Also, we add the variable neighborhood to show that we fix for price differences across neighborhoods.

The model for price shows that, fixing for neighborhood and a few other factors, rating has a positive effect on price while number of reviews does not have a statistically significant impact. We also see that a stricter cancellation policy is associated with higher prices.

5.3 Model 3: Predicting Rating

Review_scores_rating ~ host_response_time + host_response_rate + host_is_superhost + accommodates + bathrooms + price + extra_people + minimum_nights

Table 3: Regression of ratings

Variable	Coefficient	P value	Sig
(Intercept)	81.262	<2e-16	***
host_response_rate	12.152	0.030	*
host_is_superhost	1.836	0.028	*
accommodates	-0.543	0.024	*
no.of.amenities.listed	0.045	0.165	
price	0.006	0.028	*
extra_people	0.023	0.101	
minimum_nights	0.051	0.108	

We used the same stepwise process to fit a model for ratings. The residuals plot shows that the homoscedasticity assumption is satisfied. The model produces a few interesting findings. We see that accommodates has a negative effect on ratings. We also see, not surprisingly, that higher response rates are associated with higher ratings. Specifically, holding a few factors constant, we see that a one percent increase in response rate leads to a 12 point increase in ratings.

5.4 Model 4: Predicting Superhost

host_is_superhost ~ review_scores_rating + extra_people + maximum_nights + number_of_reviews + require_guest_phone_verification + minimum_nights + room_type + security_deposit + active_days

Table 4: Regression of superhost

Variable	Coefficient	P value	Sig
(Intercept)	-42.560	0.013	*
review_scores_rating	0.463	0.009	**
extra_people	-0.038	0.009	**
maximum_nights	-0.002	0.045	*
number_of_reviews	0.013	0.042	*
require_guest_phone_verification	2.130	0.022	*
minimum_nights	-0.060	0.071	.
room_typePrivateroom	2.080	0.059	.
security_deposit	0.001	0.075	.
active_days	-0.001	0.089	.

Once our main models are constructed, we try to identify factors that can predict whether a host is a superhost or not. Because being a superhost is binary, we construct a logistic regression. The residuals plot shows that homoscedasticity assumption is satisfied.

As the model implies, holding the active days of listing, security deposit, room type and minimum nights constant, an increase of maximum nights or extra people of a listing would decrease the probability of being a superhost. On the other hand, rating and number of reviews is associated with an increase in the probability of being a superhost. Also, if the listing required guest phone verification, it is more likely to be a superhost listing.

6. Conclusion

We started our analysis being most interested in identifying the relationship between a listing's characteristics and the number of reviews it garners but out of our three models, we find that the model for reviews is the least interesting/informative. The model for reviews has the lowest coefficient of determination of the three main models we construct . It also produces the least insightful findings. On the other hand, the model for price produces the most interesting / interpretable results. We find that number of reviews do not have a statistically significant impact on price but ratings do. In other words, what matters, as the model suggests, is not how many written reviews a listing has but the feedback score that it gets via ratings.

7. Deployment

The results of this model suggest that hosts should focus their attention on garnering better ratings and not more reviews. Although ratings on Airbnb are already high on average, we see that ratings has a more statistically significant impact of price than does reviews. Firms / users that wish to use the findings in this study to inform their behavior / decisions on other marketplace platforms should be aware that the variable ratings in the dataset does not have much variability across the range of possible values that a score can take.

References

1. https://www.airbnb.com/help/article/1257/how-do-star-ratings-work
2. https://press.airbnb.com/fast-facts/
3. https://www.sfchronicle.com/business/article/SF-short-term-rentals-transformed-as-Airbnb-12617798.php
4. https://www.airbnb.com/help/article/1849/san-franciscos-registration-process-frequently-asked-questions
5. https://www.airbnb.com/help/article/13/how-do-reviews-work