Host Listing Analysis for Airbnb

Team Bruins

airbnb

Project Overview

Data

- 8500 Airbnb house listing data in Seattle WA
- ~300 columns

Goals

- Help hosts to understand listing price, review scores, and requirements of being a superhost

Next step

- Test accuracy with prediction model

Data Overview

Airbnb data in Seattle WA

Listing	Host	Review
• Listing ID • **Price** • Location • **Property/Room Type** • **Amenities** • Cancellation policy • **Cleaning fee** • ...	• **Host ID** • Host URL • Host Name • **Superhost or Not** • Host photo • Location • Identity Verified • ...	• Total Rating • **Accuracy** • **Cleanliness** • **Checkin** • **Communication** • **Location** • **Value** • ...

Statement of Questions

Question 1

- As a host, what would be the key factors of getting a high price? What would be the key factors of getting a better review score?

Question 2

- As a host, what would be the key factors that would get me higher rating to qualify for a superhost?

Question 1

As a host, what would be the key factors of getting a higher listing **price**? What would be the key factors of getting a better **review score**?

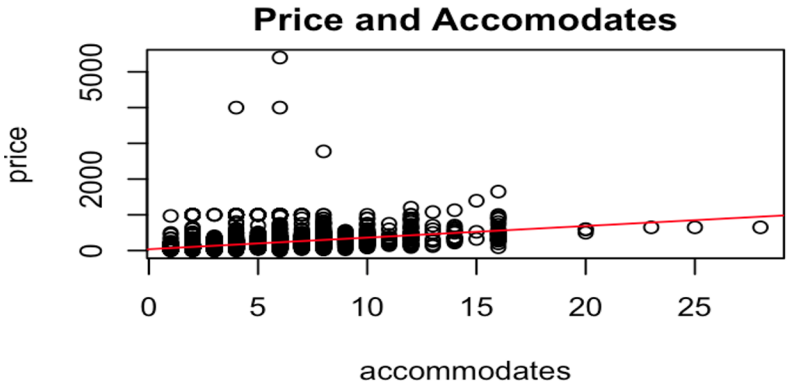
Model Selection - Why Linear Regression

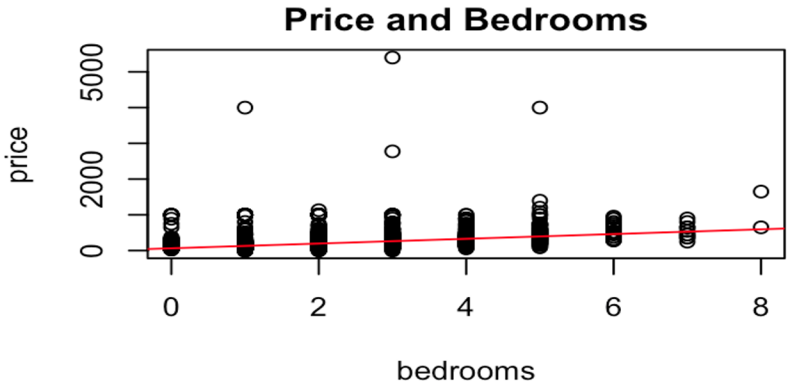
- Given the non-binary nature of our dependent variables (price, review scores) we ruled out Logistic regression and only looked at ANOVA and Linear Regression.
- We suspect a linear dose-response relationship between predictor and response for example between host response time and review scores and we expect a higher power to detect this relationship.
- The choice also supports assessment with dependent variables of non-factor nature.

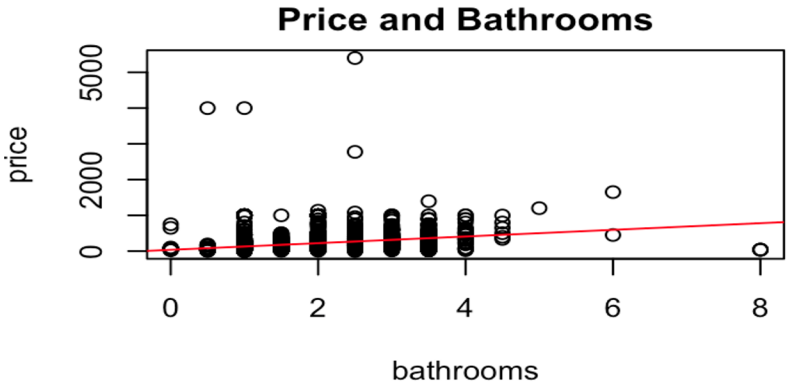
Q1a - Factors for Higher Price

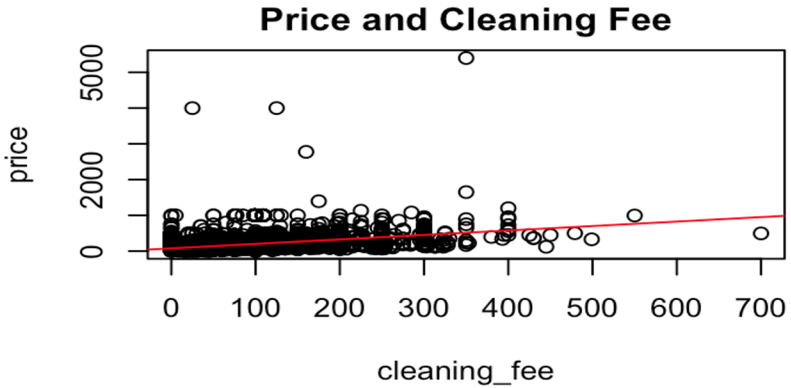
- Response: Price
- Predictors:
  - Accommodates
  - Bedrooms
  - Bathrooms
  - Cleaning fee

Regression Plots

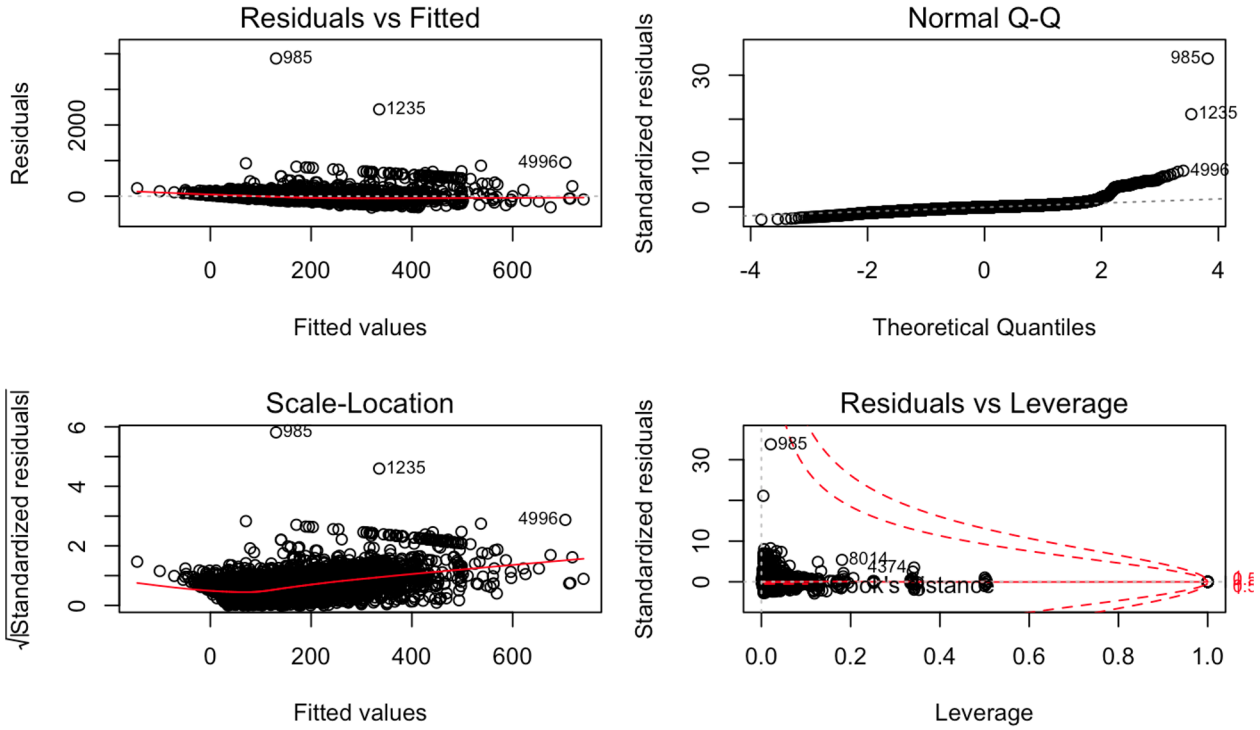








Assumption for Price



- Linearity assumption is safe
- Satisfactory homoscedasticity
- Satisfactory normality
- Some extreme outliers

Results/Analysis - Q1a for Price

- The number of bedrooms for a listing failed to attain significance
- Adjusted R-squared is low, much variability unaccounted for

  Possible reasons for this
  - Existence of a few extreme outliers
  - Lack of relevant predictors
  - Some relationships (e.g. interactions) unaccounted for
- Takeaways under assumptions of this model
  - More bathrooms (Coef = **22.48**) associate with higher prices than bedrooms (**4.06**)
  - The capacity to accommodate more guests associates with higher prices (Coef = **16.57**)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.90923    3.81203   4.173 3.03e-05 ***
accommodates 16.57441    1.20906  13.708  < 2e-16 ***
bedrooms      4.06097    2.82981   1.435    0.151
bathrooms    22.48364    3.39421   6.624 3.71e-11 ***
cleaning_fee  0.70464    0.03537  19.924  < 2e-16 ***
Multiple R-squared:  0.2435, Adjusted R-squared:  0.2432
```

Q1b - Factors for Review Scores

Response Variables: Review Scores [‘review_scores_value’, 0-10]

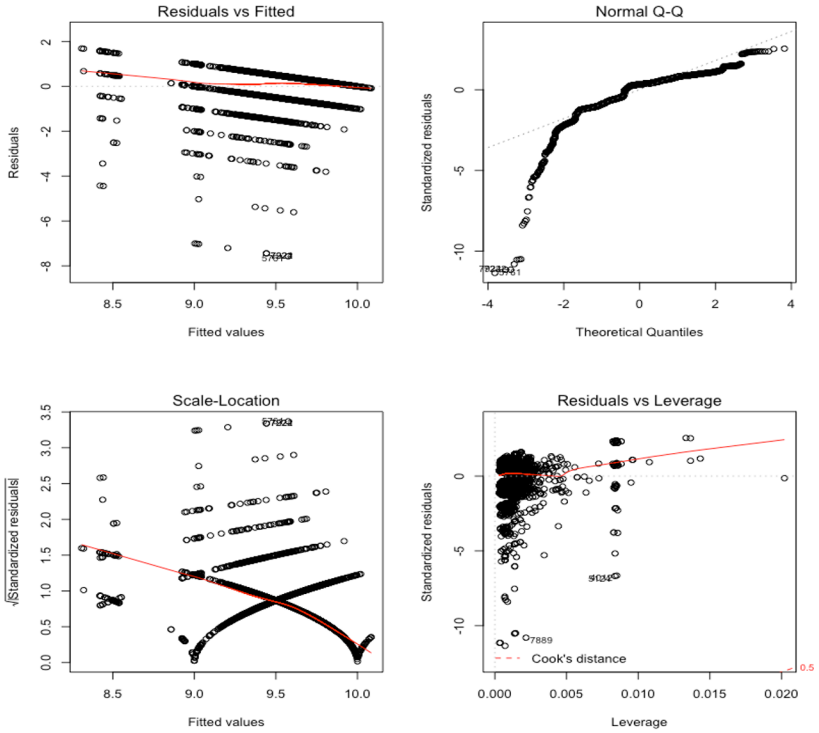
Predictor Variable Selection Steps:

- Intuitive Variable Selection from existing/engineered variables
- Assumptions Check
- Factor Selection through Stepwise [Forward/Backward elimination] Regression
- Final Results

Intuitive Variable Selection from Existing/Engineered Variables

- Price
- Accommodates
- Bedrooms
- Guests included
- Minimum nights
- Host listings count
- Host response rate
- Host is superhost
- Total Amenities

Assumption Check - Full Model



- Independence between the variables is assumed
- Heavy-tailed distribution is a deviation from Normality, but this is ignored due to the large size of the dataset
- Linearity and Variance assumptions hard to read

Factor Selection - Stepwise [Forward/Backward] Regression

```
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
```

Stepwise regression [works through AIC minimization] narrows the list down to 6 variables.

Possible overfitting, needs to be checked for accuracy in further tests

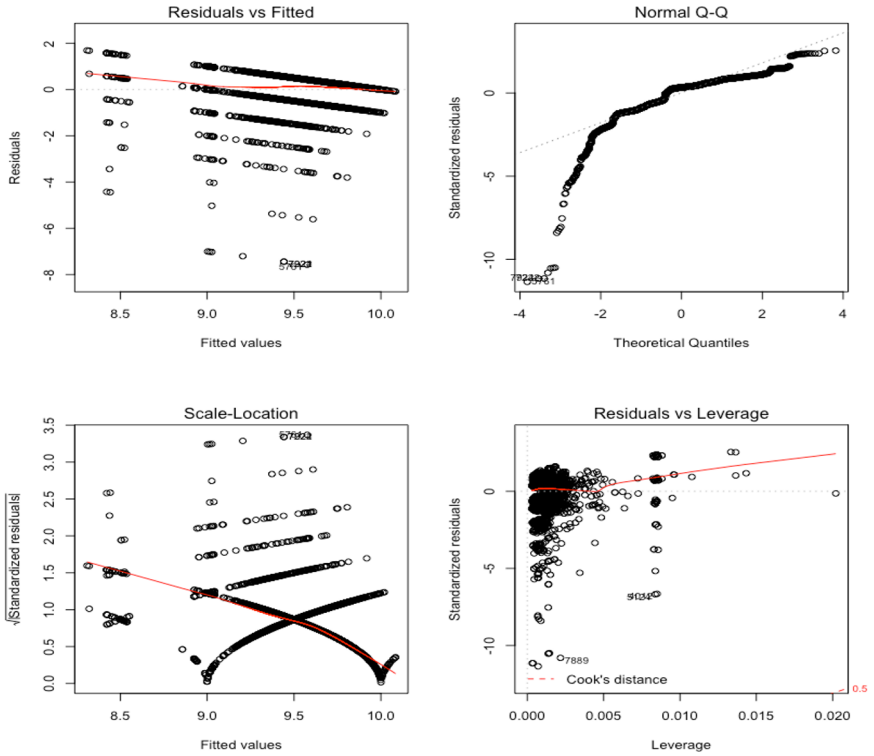
*AIC = - 2*log L + k * edf;*
L = likelihood
edf = equivalent degrees of freedom
k = number of parameters

Factor Selection - Stepwise [Forward/Backward] Regression

	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

Assumption Check - Fitted Model



- The observations are the same as the full model

Q1b Results Analysis for Review Scores

- Average number of reviews per listing: 45
- All predictors attained significance
- Adjusted R-squared is low, much variability unaccounted for

  Possible reasons for this
  - Lack of relevant predictors
  - Confounding relationships
  - Some relationships (e.g. interactions) unaccounted for
  - Weak adherence to assumptions, groupings of outliers
- Takeaways under assumptions of this model
  - Seattle Airbnb hosts are highly competent, judging by review scores (Intercept = **9.5**)
  - Being a Superhost (**.345**) and having more bedrooms (**.074**) associate most with review scores
  - Host response rate (**-.002**) and number of listings (**-.0007**) associate least with review scores
  - There are some complex challenges to modeling review scores, including selection bias

Future Steps for Prediction

- Isolate a main effect with more rigorous testing
- Account for confounding relationships when adding secondary effects
- Test for interactions around predictors
- Train and test the model

Question 2

As a host, what are the important factors that would get me higher ratings to qualify for being a superhost?

Criterion to Be a Superhost

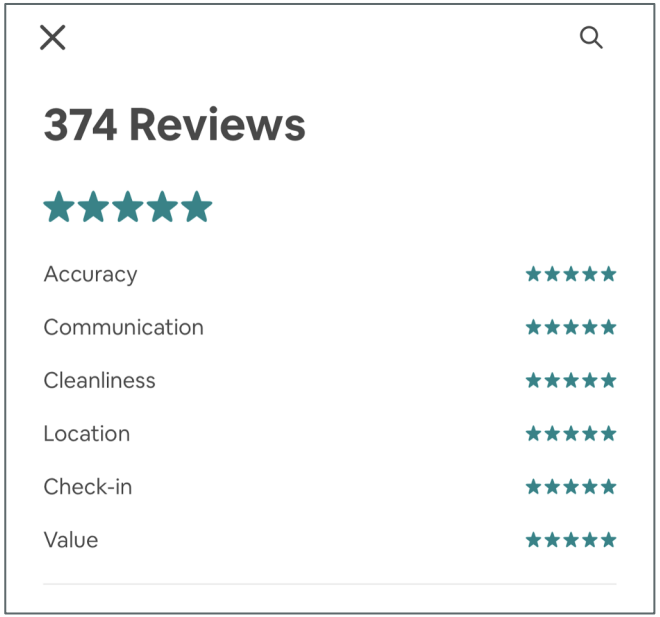
"Superhosts are experienced highly rated hosts who are committed to providing great stays for guests."

-- Airbnb

- Hosted at least 10 trips
- Maintained a 90% response rate or higher
- 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
- Completed each of the confirmed reservations without canceling

Six Star Rating Categories*

- **Accuracy.** How accurately did your listing page represent your space?
- **Communication.** How well did you communicate with your guest before and during their stay?
- **Cleanliness.** Did your guests feel that your space was clean and tidy?
- **Location.** How did guests feel about your neighborhood?
- **Check-in.** How smoothly did their check-in go?
- **Value.** Did your guest feel your listing provided good value for the price?



* https://www.airbnb.com/help/article/1257/how-do-star-ratings-work

Statistical Model: Logistic Regression

Hypothesis

The association between being a superhost and each of the six review categories.

Response host_is_superhost 0/1

Predictors Review scores in 6 categories

Model

glm(superhost~accuracy+communication+cleanliness+location+checkin+value, data=hostData,family=binomial)

Assumptions - Logistic Regression

1. Binary dependent variables

   Superhost response variable being 0/1.

1. Independent observation variables

   The study is an observation study, we could assume that each booking is independent from each other.

1. No strong intercorrelations among predictor variables

   Pearson correlation table and correlation matrix plot

Assumptions - Logistic Regression

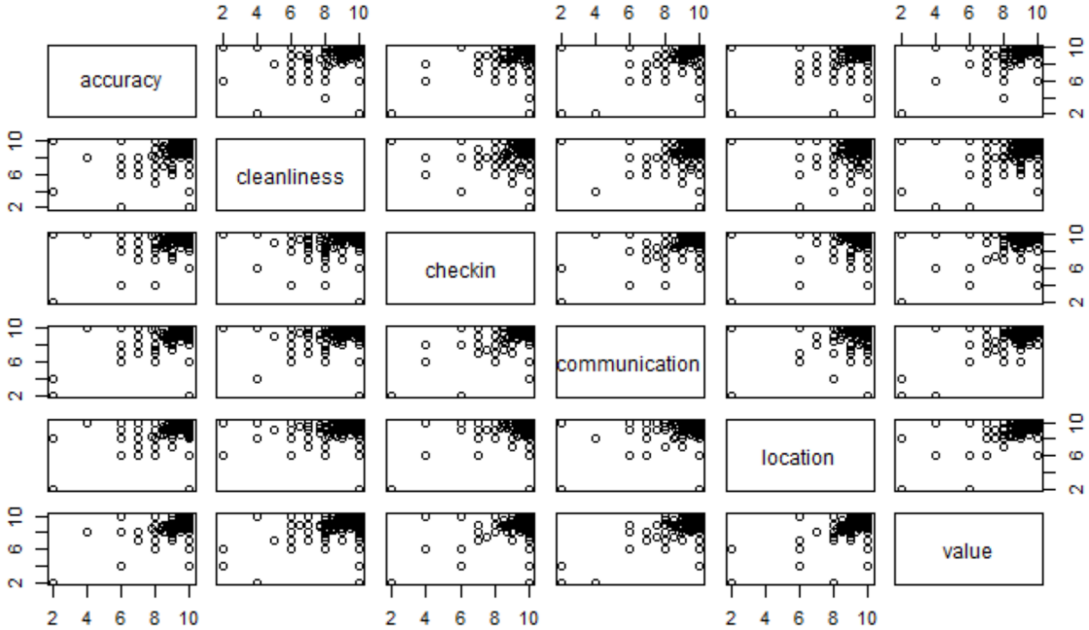
3. No strong intercorrelations among predictor variables.

	accuracy	cleanliness	checkin	communication	location	value
accuracy	1	0.5	0.439	0.53	0.313	0.532
cleanliness	0.5	1	0.337	0.326	0.225	0.487
checkin	0.439	0.337	1	0.601	0.228	0.356
communication	0.53	0.326	0.601	1	0.267	0.473
location	0.313	0.225	0.228	0.267	1	0.37
value	0.532	0.487	0.356	0.473	0.37	1

Pearson Correlation Table

Assumptions - Logistic Regression

3. No strong intercorrelations among predictor variables.



Correlation Matrix Plot

Assumptions - Logistic Regression

4. Linearity of independent variables and log odds

There is a linear relationship between the logit of the outcome and each predictor variables.

4. Large sample size

There are 5302 observations (distinct hosts in Seattle) after averaging the review scores for hosts with more than one listing.

Coefficient Estimate Table

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-46.979	3.372	-13.934	0
accuracy	1.379	0.173	7.965	0
communication	0.893	0.275	3.249	0.001
cleanliness	0.961	0.092	10.481	0
location	0.253	0.087	2.895	0.004
checkin	0.946	0.237	3.987	0
value	0.31	0.084	3.707	0

Interpretation of the Association

	Estimate
(Intercept)	0
accuracy	3.971
communication	2.442
cleanliness	2.614
location	1.287
checkin	2.577
value	1.364

Exponentiated Coefficient Estimate

Accuracy: 3.971 - Holding other predictors fixed there is a 297.1% increase in odds of becoming a superhost for one-unit increase in accuracy rating.

Communication: 144.2% increase in odds

Cleanliness: 161.4% increase in odds

Location: 28.7% increase in odds

Check-in: 157.7% increase in odds

Value: 36.4 % increase in odds

Rank Rating Categories

	Estimate
(Intercept)	0
accuracy	3.971
communication	2.442
cleanliness	2.614
location	1.287
checkin	2.577
value	1.364

1. **Accuracy.** How accurately did listing page represent your space?
2. **Cleanliness.** Clean and tidy?
3. **Check-in.** How smoothly did check-in go?
4. **Communication.** How well did you communicate with your guest?
5. **Value.** Good value for the price?
6. **Location.** About your neighborhood?

Conclusion

1. Associated with higher listing price:
    a. Higher bathroom counts
    b. Ability to accommodate more guests
2. Associated with higher review scores
    a. Superhost status
    b. Utilize listings with more bedrooms
3. Priorities for becoming a superhost
    a. Maintain accurate listing information
    b. Ensure cleanliness of the space
    c. An easy check-in procedure
    d. Positive interactions with customers

* Reasonable price and great location also help in raising the overall review scores.

<u>Hosts want:</u>
Profit and ratings

<u>Guests want:</u>
Bedrooms, bathrooms, clean space and honesty

<u>Airbnb wants:</u>
Superhosts


SUPERHOST
like a boss

Thank you!