# **Different colored houses**

**Does an Airbnb Listing’s Annual Revenue Vary with Host Status?**

**Prepared by**

*Fides Schwartz*

*Jaya Khan*

*Satvik Kishore*

*Tego Chang*

*Unifying Data Science Spring 2022*



**Executive Summary**

In recent years, Airbnb hosts have been generating revenue to supplement or replace their regular salaries.

As per airbnb.com (Airbnb 2022), Airbnb awards the label of superhost to a regular host who fulfils a certain set of criteria for specialized service throughout the booking process and a guest's stay.

The question that immediately occurs is: "Does superhost status help Airbnb hosts generate more annual revenue?".

We find evidence that superhosts can generate more annual revenue per listing than regular hosts based on the Airbnb superhost label alone. On average, a superhost has an annual revenue of $3,807 per listing which is 2.53 times higher than the average revenue of a regular host with otherwise similar features, thus there seems to be a causal effect to the superhost label.

We validate these findings based on data from two select states - California and Florida - the states that generate the highest revenue on Airbnb US annually (Dogru, Mody et al. 2020).

Airbnb superhosts make an average of $2,041 more per listing annually than regular hosts



**Superhosts have higher annual revenue**

1. **INTRODUCTION**

* 4.8+

Star Rating

* 10+

Completed stays in the last year

* <1%

Cancellation rate

* 90%

Response rate

**AIrbnb Superhost Criteria**

You are planning your first post-COVID family vacation and are trying to decide how to book accommodation for five people and two dogs quickly, because you left your decision making to the last minute, since Centres for Disease Control guidance is changing by the minute. While browsing the Airbnb listings for Wilmington, you find several beautiful looking houses, just a ten-minute drive from the beach and pet friendly with a separate room for the in-laws and a large kitchen to make meals and memories together. All the hosts offer instant booking and have been verified and active for years.

But how do you make your decision now? Looking at the listings, only one of them has a superhost label, so you end up booking their residence, because you expect better service than the other hosts could provide.

This is, of course, a hypothetical scenario but one worth exploring through data analysis. Do people tend to book with Airbnb superhosts more frequently than with regular hosts, thus leading to higher annual revenue for superhosts over regular hosts with similar characteristics?

Airbnb is one of the most prominent companies of the so-called “sharing economy” or “peer-to-peer markets” together with household names such as Uber and TaskRabbit, and it has had an impact on how people book holidays and the hotel industry in the markets it has established itself in (Zervas, Proserpio et al. 2017). Since its founding in 2008, approximately 500 million people have booked stays with Airbnb (Airbnb 2015), so the question of how to attract the largest number of these potential customers is relevant to a host’s economic perspectives.

While Airbnb has changed the landscape for travellers looking for cheap or unique accommodation, it has also provided hosts with an increase in monthly income (increasing over the years (Poppick 2015, Management 2022)), and there are ~4 million global hosts with Airbnb listings (Lewis 2020, Management 2022), who are vying to take a slice of the huge holiday accommodation market.

To understand what might drive decision making in customers, there has been some research on the influence of multi-property listings, as well as location and

professionalism on Airbnb revenue in general and host revenue in particular (Lane and Woodworth 2016, Deboosere, Kerrigan et al. 2019, Kwok and Xie 2019, Xie and Mao 2019, Chattopadhyay and Mitra 2020, Xie, Heo et al. 2021).

**There is a causal effect from the superhost label on annual revenue per** listing

**Hypothesis**

Our hypothesis is, that hosts who have the superhost label awarded by Airbnb, can generate more annual revenue than other hosts with similar characteristics but without the superhost label, because customers elect to stay in their properties more frequently.

Airbnb makes some of its data available publicly, which makes it accessible to thorough data science analysis (Airbnb 2022).

1. **MOTIVATION FOR ANALYSIS**

With this project, we hope to find solutions for the travel and hospitality industry. The same idea of assigning a special category to some accommodations can be applied to other home booking sites and hotels that advertise their rooms on rental websites such as Expedia and Booking.com.

We are interested in one major piece of information for our causal inference:

**Is the label of superhost helpful in generating more revenue?**

Our response variable is thus the estimated annual revenue for listings clustered by hosts, and our predictors are (1) the Airbnb assigned superhost status and (2) all other host- and listing-based variables with statistical significance.

**LA**

Inhabitants:

**4** million

Superhosts:

**9,442**

Regular Hosts:

**6,992**

**Broward**

Inhabitants:

**2** million

Superhosts:

**3,187**

Regular Hosts:

**6,149**

**Counties**

1. **DATA**

**Overview**

We are using the data provided by Airbnb on the following website: <http://insideairbnb.com/get-the-data.html>. This data includes timepoints in March, June, September, and December of 2021 for 104 cities/regions that have Airbnb listings all over the world. Airbnb provides a data dictionary (<https://tinyurl.com/y7h9m4nu>) that includes 73 variables.

**Data Selection**

Since there are vast amounts of data that would all need different pre-treatment, we consider datasets from two places for this project. Since California and Florida are the two highest ranking states for Airbnb revenue generation in the USA, *Los Angeles, California* and *Broward County, Florida* were chosen (these were the two counties available for these states from Airbnb; (Dogru, Mody et al. 2020)).

We decided to compare the difference between these two locations at the beginning of our research on the relationships between superhost and the estimated revenue, because we wanted to include a locational component in the evaluation. Florida and California might attract very different types of travellers and those travellers may have disparate criteria for choosing to stay at an Airbnb listing.

**Data Cleaning**

Once our data collection was complete, we went on to data processing and data wrangling. First, we excluded the columns that we assume to have no impact on our response of interest – *estimated annual revenue per listing* – logically (e.g., URL-addresses of listings). Second, we decided to drop all columns that had duplicate information (e.g., ‘*bathrooms*’ which was numerical and ‘*bathroom\_text*’, which was a string). Third, we decided to drop the columns with the most missing data.

To estimate annual revenue, we multiplied the price of the listing with reviews in the last 12 months and accounted for the fact that only 67% of travellers leave a review after their stay (Zervas, Proserpio et al. 2017).

*Estimated Revenue = Price of Listing \* (Reviews in last 12 months \* 100/67)*

**Matching**

To be able to analyse what influence the superhost status has on the estimated annual revenue per listing, we needed to match all other factors that might influence revenue (e.g., location, size of property, star-rating) between the regular hosts and super hosts as closely as possible. To do this, we decided to use DAME-FLAME.

To prepare data for the matching process, several continuous variables had to be discretized and coarsened to limit the amount of detailed matching that would otherwise be attempted.

An initial run with 50 iterations yielded a stark increase in prediction error before the 10th iteration (**Appendix 1**), which led us to repeat the task with only 10 iterations, where we saw a notable increase in prediction error between the 3rd and 4th iteration (**Figure 1**) but no difference in model performance.

Based on these diagnostics, we decided to use the third iteration of our matching process. This included 10,223 matches after the variables pertaining to number of beds and bathrooms associated with a listing were dropped from the analysis.

This iteration yielded a total number of 1,617 matched groups.

Chart, line chart

Description automatically generated

**Figure 1:** Shows the plot of DAME diagnostics. The prediction error increases markedly between the 3rd and 4th iteration.

**Multiple Linear Regression**

Once the matching process was complete, we were able to run the linear regression analysis using fixed effects and clustering by host. Our regressions were based on the DAME-FLAME output.

We first ran a regression that included all variables (except the ones used in revenue calculation) that were part of our analysis after data cleaning.

Assessment of this model violated the assumption of normality with a Q-Q-plot that suggested an exponential distribution (**Figure 2**), which led us to run a second regression on log-transformed data.

This second regression satisfied the assumptions for linear regressions (**Appendix 2**), and we proceeded to use it for our statistical analysis.

*annual\_listing\_revenue ~ host\_is\_superhost + C(county) + C(host\_response\_time) + C(host\_acceptance\_rate) + host\_identity\_verified + C(room\_type) + C(accommodates) + C(bathrooms\_text) + C(bedrooms) + C(beds) + C(minimum\_nights) + C(maximum\_nights) + C(review\_scores\_rating) + C(review\_scores\_accuracy) + C(review\_scores\_cleanliness) + C(review\_scores\_checkin) + C(review\_scores\_communication) + C(review\_scores\_location) + C(review\_scores\_value) + instant\_bookable + C(calculated\_host\_listings\_count) + essentials + C(other\_amenities)*

1. **SUMMARY STATISTICS**

There was a total of 10,179 listings owned by superhosts and 15,591 listings owned by regular hosts in our dataset.

The locational distribution of regular hosts and super hosts was similar for each location (**Appendix 3 and 4**).

Chart, line chart

Description automatically generated**Figure 2:** Shows the initial Q-Q-plot generated to test the assumption of normality in our data. The shape of the plot suggests that a log transformation of the data could improve model performance down the line.

**Causal Effect**

After removing the baseline differences between the two groups of hosts through our matching process, we found that our causal inference yields the following result:

The average difference in annual revenue between listings by hosts who we actually observe as superhosts and listings by hosts who we actually observe as regular hosts in a world where neither is a superhost is $2,041 (**Figure** **3**).

E(YT = 1|D = 1) – E(YT = 0|D = 0) = $2,041

Chart, bar chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

**Figure 3**: Log scale ATE on left and zoomed in ATE transformed back on right.

To explore which other variables might be relevant for revenue generation of an Airbnb listing, we generated a plot that visualizes both error bars and confidence intervals (**Figure 4**). We summarized the variables that seem to influence the revenue a listing generates, apart from the superhost label awarded by Airbnb in **Table 1**.

A picture containing antenna

Description automatically generated

**Figure 4**: Error Bars with 95% confidence interval as whiskers. If the whiskers cross the 0-line, the variable is not likely to be of statistical significance.

|  |  |  |
| --- | --- | --- |
|  | **Coefficient** | **SE** |
| **Intercept** | 11.137 | 0.204 |
| **Host is Superhost** | 0.930 | 0.037 |
| **Florida** | 0.394 | 0.054 |
| **Private room** | -1.193 | 0.055 |
| **Accommodates** | 0.662 | 0.055 |
| **No. Bathrooms** | 0.162 | 0.036 |
| **No. Bedrooms** | 0.447 | 0.080 |
| **No. Beds** | 0.162 | 0.036 |
| **Other Amenities** | 0.344 | 0.039 |
| **Min nights** | -1.782 | 0.041 |
| **Check-in score** | -0.643 | 0.118 |
| **Communication score** | -0.518 | 0.121 |
| **Listings/host [T1]** | -0.315 | 0.062 |
| **Listings/host [T2]** | -0.457 | 0.085 |
| **Listings/host [T3]** | -0.394 | 0.044 |
| **Response time [T1]** | -0.369 | 0.062 |
| **Response time [T2]** | -0.728 | 0.100 |

***Table 1****: Shows the variables that showed a statistically significant influence on annual revenue with p-values of lower than 0.001 on log transformed data*

**V. DISCUSSION**

We found that being a superhost does correlate with generating more estimated annual revenue per listing (*$2,041)*than being a non-superhost to the level of statistical significance.

The higher annual revenue generated by superhosts is in accordance with Airbnb’s communication of the overall average amount of money that is made over a year by all hosts ($9,600), and the amount they say an “experienced host” can expect to make in one year ($10,000), though our estimate is lower than this $4,000 difference.

*Limitations*

We did not use the whole dataset that Airbnb makes available on listings worldwide but focused instead on the data from the two states generating the most revenue. The analysis could be extended to all US states and worldwide locations that data is available for.

In addition, our analysis is limited by the quality of the data that Airbnb provides and by the fact that our data is cross-sectional, thus relying on the quality of our matching process. We could also consider a longitudinal analysis of hosts before and after they achieve superhost status.

In conclusion, we have found that there is a causal effect to the superhost label; the listings under the superhost label generate 2.53 times more revenue than ones under regular hosts and superhost status can thus be considered a relevant factor for revenue generation that hosts should aspire to.

**VI. REFERENCES**

Airbnb. (2015). "Airbnb Summer Travel Report 2015." Retrieved 04/03/2022, from <https://blog.atairbnb.com/wp-content/uploads/2015/09/Airbnb-Summer-Travel-Report-1.pdf>.

Airbnb. (2022). "Get the Data." Retrieved 04/03/2022, from <http://insideairbnb.com/get-the-data.html>.

Airbnb. (2022). "Superhost: Recognizing the best in hospitality." Retrieved 04/03/2022, from <https://www.airbnb.com/d/superhost>.

Chattopadhyay, M. and S. K. Mitra (2020). "What Airbnb Host Listings Influence Peer-to-Peer Tourist Accommodation Price?" Journal of Hospitality & Tourism Research **44**(4): 597-623.

Deboosere, R., D. J. Kerrigan, D. Wachsmuth and A. El-Geneidy (2019). "Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue." Regional Studies, Regional Science **6**(1): 143-156.

Dogru, T., M. Mody, C. Suess, N. Line and M. Bonn (2020). "Airbnb 2.0: Is it a sharing economy platform or a lodging corporation?" Tourism Management **78**: 104049.

Kwok, L. and K. L. Xie (2019). "Pricing strategies on Airbnb: Are multi-unit hosts revenue pros?" International Journal of Hospitality Management **82**: 252-259.

Lane, J. and R. M. Woodworth (2016). "The sharing economy checks in: An analysis of Airbnb in the United States." CBRE Hotel’s Americas Research.

Lewis, T. (2020). "Airbnb Statistics (Growth, Revenue, Hosts + More!)." Retrieved 04/03/2022, from 8. <https://hostsorter.com/airbnb-statistics/>.

Management, i. (2022). "Airbnb Statistics." from <https://ipropertymanagement.com/research/airbnb-statistics>.

Poppick, S. (2015). "Airbnb Says Renting Your Place Is Like Getting a Big Raise." Retrieved 04/03/2022, from <https://money.com/airbnb-raise-income-report/>.

Xie, K., C. Y. Heo and Z. E. Mao (2021). "Do professional hosts matter? Evidence from multi-listing and full-time hosts in Airbnb." Journal of Hospitality and Tourism Management **47**: 413-421.

Xie, K. and Z. Mao (2019). "Locational strategy of professional hosts: Effect on perceived quality and revenue performance of Airbnb listings." Journal of Hospitality & Tourism Research **43**(6): 919-929.

Zervas, G., D. Proserpio and J. W. Byers (2017). "The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry." Journal of Marketing Research **54**(5): 687-705.

APPENDIX

**Methodology, exploratory data analysis and regression results**

To explore the causal effect that the superhost label has on annual revenue generation, we had to try and match all other variables that could influence a customer’s decision to book an accommodation as closely as possible, while keeping the criteria coarse enough for enough matches to be made between superhost and regular hosts.

Chart, line chart

Description automatically generated

**Appendix 1:** Initial DAME diagnostics showing the first increase in prediction error between 0 and 10 iterations.

We used several host and listing level variables for our regressions:

1. Host: average response time, acceptance rate, has profile picture, identity verified, number of listings under their administration.
2. Listing: room type, number of bathrooms, bedrooms and beds, amenities, review scores, minimum nights a guest has to book, whether a listing is instantly bookable, annual revenue (calculated earlier).

The results of our initial regression using data that was not log transformed yielded the following output:

![Table

Description automatically generated]()

We ran diagnostics on this data and realized that the assumption of normality for linear modelling were violated, thus we performed a log transformation and re-ran our diagnostics, which now showed much better results.

Chart, line chart

Description automatically generated

**Appendix** **2**: Q-Q-plot to evaluate the assumption of normality for log-transformed data.

There were no relevant leverage points in our data:

Chart, line chart

Description automatically generated

After the log transformation was performed, we received the following results on our regression:

![Table

Description automatically generated]()

The final regression we ran was added to cluster our results by host, so we could interpret our data more comprehensively:

![Table

Description automatically generated]()

Map

Description automatically generated

Map

Description automatically generated**Appendix 3**: Shows the distribution of listings belonging to superhosts (red circles) and listings belonging to regular hosts (blue triangles) in Broward County, Florida.

**Appendix 4**: Shows the distribution of listings belonging to superhosts (red circles) and listings belonging to regular hosts (blue triangles) in Broward