**Factors that influence Airbnb Superhost Status and Revenue Generation**

***Prepared by: Fides Schwartz, Jaya Khan, Satvik Kishore, Tego Chang***

***Unifying Data Science Spring 2022 – Team 10***

April 08, 2022

1. **INTRODUCTION**

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 (1). In addition, a lot of research has been conducted on its impact on customer segmentation (2), affordable housing (3), and consumer trust (4), among others.

Since its founding in 2008, approximately 500 million people have booked stays with Airbnb and in the summer of 2015 alone, 17 million people stayed in Airbnb accommodation (5).

While Airbnb has changed the landscape for travelers looking for cheap or unique accommodation, it has also provided hosts with an increase in monthly income (average of $7,350 in 2015, $9600 in 2021, with experienced hosts earning an average of $10,000; (6, 7), and there are ~4 million global hosts with Airbnb listings (7, 8).

The experienced hosts who make an average of $10,000 per year, are most likely in the superhost and Airbnb plus categories (9). Airbnb claims that one becomes a superhost by fulfilling four criteria (10):

1. 4.8+ star rating
2. 10+ completed stays in the last year
3. <1% cancellation rate
4. 90% response rate

Airbnb makes some of its data available publicly, which makes it accessible to thorough data science analysis (11).

1. **MOTIVATION FOR ANALYSIS**

With this project, we hope to find solutions for the travel and hospitality industry. The same ideas can be applied to other home and hotel rental companies such as Expedia and Booking.com.

We are interested in three major pieces of information:

1. Is being a superhost really helpful in generating more revenue?
   1. response: estimated annual revenue
   2. predictors: superhost (as defined by the data the dictionary provided by Airbnb) and location
2. Identify the factors that influence annual revenue and explore whether these correlate with the factors that influence superhost status
   1. response: estimated annual revenue
   2. predictors: all other variables in the dataset
3. **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://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit?usp=sharing>) that includes 73 variables.

*Data scraping*

In this project, we consider datasets from two places, Los Angeles, CA and Broward County, CA.

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. We believe that the two locations have enough similarities (both are liberal/progressive counties with a similar percentage of retired people (21% in LA, 23% in Broward County (retired people are the fastest growing marked for Airbnb), though LA is about double the size (~4 million inhabitants) as Broward County (~2 million inhabitants)) for this comparison to work.

We implemented an automated scraping program to ensure we can further stretch our research to multiple places in the US or even worldwide Airbnb hosts.

*Data cleaning*

Once our data collection was complete, we went on to processing and data wrangling. First, we excluded the columns that we assume to have no impacts on our response of interest - estimated annual revenue – logically (e.g., url-addresses of listings, scrape\_id). Then, we decided to drop all columns that had duplicate information (e.g., ‘bathrooms’ and ‘bathroom\_text’) and turn as many of the free-text columns into numerical or categorical values, as possible. Third, we found there were plenty of missing data in the dataset and decided to drop the columns with the most missing data.

As the following matching and the regression analysis will be conducted based on host, we aggregated the data for multiple listings by host (i.e., if a host has multiple listings, they are aggregated into one row) and locations (in this case, they are Los Angeles, CA and Broward County, FL), though we kept a column with the total number of listings for our matching process (assuming that hosts with e.g. 5 listings are more similar to each other than to a host with just one listing, even if not all of their listings are in the two towns we were investigating). During aggregation we always assumed that the highest category should be used (e.g., if a host has both an entire flat and a private room in a flat as listings, we aggregated to the entire flat category). The listing price was averaged among listings for one host.

Based on the data pre-processing above, we calculated our response variable, the estimated annual revenue of a host, according to the formula:

Estimated annual revenue per host = Sum (units owned by the host) Price of the unit i \* (number of reviews per month for the unit i) \* average number of reviews per host (= 72%)) \* average minimum nights required for the unit i

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

DAME-FLAME ran a total number of iterations of 39 and stopped before iteration number 40, which would have yielded the same number of matches as iteration number 39. The total number of matched groups formed was 433. We were left with 1625 unmatched treated units out of a total of 2590 treated units and 1,198 unmatched control units out of a total of 1,807 control units. The predictive error of covariates chosen was 28652034941291.805 and the number of matches made was 62 in iteration 39 and the total number of matches was 1,574.

Once the matching process was complete, we were able to run the linear regressions.

Our first regression after matching was a linear regression with annual revenue as our dependent variable and the categorical variable of whether or not a host is considered a superhost as our predictor:

*annual\_revenue ~ C(host\_is\_superhost)*

The second regression was a linear regression that used annual revenue as the dependent variable and both the superhost categorical variable and the regional variable of state (California vs Florida):

*annual\_revenue ~ + C(host\_is\_superhost) + state*

Third, we ran a regression that included all 21 variables that were part of our analysis after data cleaning, to get an overview of which variables might be relevant for further analysis.

Our last regressions were based on the DAME-FLAME output. Using the factors that were kept as predictors we ran a weighted regression on annual revenue:

*annual\_revenue ~ C(host\_response\_time) + C(host\_is\_superhost) + host\_total\_listings\_count + C(instant\_bookable) + beds + minimum\_nights\_avg\_ntm + reviews\_per\_month + C(city)*

In addition, we ran a regression without weighting and leaving out the variables that showed no statistical significance in the prior regression:

*annual\_revenue ~ C(host\_is\_superhost) + host\_total\_listings\_count + C(instant\_bookable) + beds + minimum\_nights\_avg\_ntm + reviews\_per\_month + C(city)*

1. **SUMMARY STATISTICS**

**EDA**

There were a total of 2,590 superhosts and 1,807 non-superhosts in our dataset and based on our chi-squared test, there was no statistically significant difference between the proportion of host-types between our two locations in California and Florida (1.0833446567931416e-69).

The initial t-test to compare annual revenue of superhosts and non-superhosts yielded a t-statistic of -1.356 with a p-value of 0.175, so it looked like there was no difference in annual revenue based on just the superhost label, when all other factors were matched.

Chart, box and whisker chart

Description automatically generated

Figure 1: Shows the boxplots of the hosts non-superhost (regular host) or superhost status in relation to the annual revenue. Visual inspection does not show a difference in annual revenue when accounting for all other influencing factors (e.g., number of listings per host, number of beds in residence).

This was reflected in the linear regression that we ran with annual revenue as dependent variable and the superhost label from Airbnb as predictor (**Table 1**), where the superhost label itself was not statistically significant as a predictor of annual revenue, though it was close to statistical significance at 0.052.

**Table 1**: Results of the regression analysis of the superhost status' influence on estimated annual revenue

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std. Error** | **t-value** | **p-value** | **95% CI** |
| Is\_superhost | 3394.4 | 1744.97 | 1.95 | 0.052 | -27.22 - 6816.02 |

We then added the predictor of state, in our case California vs Florida, into our regression, which was statistically significant with a p-value of almost zero (**Table 2**), while the superhost status stayed on the edge of statistical significance at 0.051. In addition, adding the state variable resulted in a slight decrease in standard error and a narrower 95% confidence interval for the superhost variable.

Both regressions showed the same amount of increase in annual revenue ($3,394.4) if a host was a superhost.

**Table 2**: Results of the regression analysis of the superhost status' and the regional factor’s (California vs Florida) influence on estimated annual revenue

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std. Error** | **t-value** | **p-value** | **95% CI** |
| Is\_superhost | 3394.4 | 1738.42 | 1.95 | 0.051 | -14.37 - 6803.16 |
| State | 9416.23 | 2036.06 | 4.63 | 0.000 | 5423.83 - 1.34e+04 |

To explore which variables in our dataset might have a significant influence other than being a superhost or the state an Airbnb is located in, we performed a regression with all of our 21 variables included (**Table 3**, shows only variables with statistically significant p-values). These seven variables reflected whether an Airbnb was instantly bookable, how many people it accommodates, how many bathrooms and bedrooms there are, the number of listings per host, and the review scores and host experience.

**Table 3**: Results from the regression analysis of annual revenue using all our variables as predictors. Only results with significant p-values are shown.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std. Error** | **t-value** | **p-value** | **95% CI** |
| Instantly bookable | 1590.57 | 546.63 | 2.910 | 0.004 | 518.77 – 2662.37 |
| Number of guests accommodated | 708.79 | 231.04 | 3.07 | 0.002 | 255.77 – 1161.8 |
| Number of bathrooms | 4425.42 | 519.2 | 8.5 | 0.000 | 3407.4 – 5443.5 |
| Number of bedrooms | -193.73 | 458.17 | -2.39 | 0.017 | -1992.1 - -195.36 |
| Review scores | 2466.53 | 1172.93 | 2.10 | 0.036 | 166.7 – 4766.36 |
| Number of listings per host | 2205.5 | 56.97 | 38.71 | 0.000 | 2903.8 – 2317.21 |
| Host experience | 0.65 | 0.267 | 2.42 | 0.016 | 0.12 – 1.18 |

Our regressions based on the variables that DAME-FLAME had yielded also showed no statistical significance for the superhost variable whether they were weighted (**Table 4**) or non-weighted (**Table 5**). The other variables that showed statistical significance were mostly included in both our full regression and the DAME-FLAME based regression, with the exception of average minimum nights (the minimum of consecutive nights one has to book the Airbnb), which was statistically significant with a p-value near 0.

Chart, line chart

Description automatically generated

Figure 2: Confirms our analysis, showing that for most treatment units, there is no estimated treatment effect (i.e., being a superhost) on annual revenue.

**Table 4**: Results of the weighted regression analysis of the variables kept by DAME-FLAME’s influence on estimated annual revenue

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std. Error** | **t-value** | **p-value** | **95% CI** |
| Host response time 1 | -6264.94 | 3632.17 | -1.73 | 0.085 | -1.34e+04 – 862.152 |
| Host response time 2 | -9317.66 | 4716.519 | -1.98 | 0.048 | -.186e+04 - -62.84 |
| Superhost | 291.18 | 1952.225 | 0.15 | 0.881 | -3539.50 – 4121.86 |
| Instantly bookable | 221.61 | 2178.529 | 0.15 | 0.0881 | -3539.50 – 4121.86 |
| City | 2364.7 | 2618.364 | 0.90 | 0.367 | -2773.09 - 7502.49 |
| Total listings by host | 1883.66 | 1110.630 | 1.7 | 0.090 | -295.64 - 4062.95 |
| Number of beds | 3006.39 | 960.053 | 3.13 | 0.002 | 1122.56 - 4890.21 |
| Average minimum nights | 1475.54 | 154.191 | 9.57 | 0.000 | 1172.99 - 1778.10 |
| Reviews per month | 2637.0923 | 760.099 | 3.469 | 0.001 | 1145.616 - 4128.568 |

**Table 5**: Results of the weighted regression analysis of the variables kept by DAME-FLAME’s influence on estimated annual revenue

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std. Error** | **t-value** | **p-value** | **95% CI** |
| Superhost | 9853.18 | 1.12e+04 | 0.883 | 0.377 | -1.2e+04 - 3.17e+04 |
| Instantly bookable | 2.79e+04 | 1.08e+04 | 2.592 | 0.010 | 6794.41 - 4.9e+04 |
| City | 2.251e+04 | 1.22e+04 | 1.845 | 0.065 | -1406.60 - 4.64e+04 |
| Total listings by host | 4326.96 | 101.74 | 42.529 | 0.000 | 4127.5 - 4526.425 |
| Number of beds | 9238.94 | 2818.08 | 3.278 | 0.001 | 3714.08 - 1.48e+04 |
| Average minimum nights | 1.275e+04 | 334.58 | 38.100 | 0.000 | 1.21e+04 - 1.34e+04 |
| Reviews per month | 2.294e+04 | 2758.01 | 8.317 | 0.000 | 1.75e+04 - 2.83e+04 |

Chart, line chart

Description automatically generated

Figure 3: Though there is no clear linear relationship between the reviews a host receives per month and their annual revenue, this visualization shows an overall increased revenue with more reviews per month (steepest increase in revenue is between ~5 and ~25 reviews per month).

**V. CONCLUSION**

In conclusion, we found that being a superhost does not correlate with generating more estimated annual revenue than being a non-superhost to the level of statistical significance, though there was a trend for higher annual revenue in superhosts of $3,394.4.

This 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; approximately $4,000 difference, similar to our regression output).

The seven variables we found that influenced annual revenue were whether an Airbnb was instantly bookable, how many people it accommodates, how many bathrooms and bedrooms there are, the number of listings per host, and the review scores and host experience. Only one of these correlates with the factors that influence whether a host is considered a superhost based on what Airbnb communicates (i. e. review scores).

**References**

1. Zervas G, Proserpio D, Byers JW. The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. Journal of Marketing Research. 2017;54(5):687-705. doi:10.1509/jmr.15.0204
2. Lutz C, Newlands G. Consumer segmentation within the sharing economy: The case of Airbnb. Journal of Business Research. Volume 88, 2018, Pages 187-196, ISSN 0148-2963, <https://doi.org/10.1016/j.jbusres.2018.03.019>
3. Barron K, Kung E, Proserpio D. The Sharing Economy and Housing Affordability: Evidence from Airbnb, 2018. Association for Computed Machinery, ISBN: 9781450358293. doi: 10.1145/3219166.3219180
4. Ert E, Fleischer A, Magen N. Trust and reputation in the sharing economy: The role of personal photos in Airbnb. Tourism Management, Volume 55, 2016, Pages 62-73, ISSN 0261-5177. doi:10.1016/j.tourman.2016.01.013.
5. Airbnb Summer Travel Report 2015
6. <https://money.com/airbnb-raise-income-report/>
7. <https://ipropertymanagement.com/research/airbnb-statistics>
8. <https://hostsorter.com/airbnb-statistics/>
9. <https://www.airbnb.com/help/article/2521/the-difference-between-airbnb-plus-and-superhost>
10. <https://www.airbnb.com/d/superhost>
11. <http://insideairbnb.com/get-the-data.html>

All websites were accessed on Sunday, April 03rd 2022