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

I. 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 travellers 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).

II. 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 two major pieces of information:

1. Is being a superhost helpful in generating more revenue?
    a. response: estimated annual revenue
    b. 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
    a. response: estimated annual revenue
    b. predictors: superhost and all other variables in the dataset

III. 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 Scraping

In this project, we consider datasets from two places, *Los Angeles, California* and *Broward County, Florida*. 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 data 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, id of the data scrape). Then, we decided to drop all columns that had duplicate information (e.g., one of '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:

$$\boxed{\textit{Estimated Revenue} = \textit{Price of Listing} \ * \ \textit{Minimum Average Night Value} \ * \left(\textit{Reviews in the last 12 months} \ * \ \frac{100}{67} \right)}$$

To estimate annual revenue, we multiply price of the listing with:

3. minimum night value because it is the average minimum number of nights a renter stays on a listing.
4. reviews in the last 12 months because we believed around 67% of the times consumers leave the review after their stay (1).

Matching

To be able to analyse 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 regular hosts and super hosts as closely as possible. To do this, we decided to use DAME-FLAME.

DAME-FLAME ran a total number of iterations of 14 and stopped before iteration number 15, which would have yielded the same number of matches as iteration number 14. The total number of matched groups formed was 4,674. We were left with 3,098 unmatched treated units out of a total of 5,599 treated units and 4,812 unmatched control units out of a total of 6,985 control units.

Linear Regression

Once the matching process was complete, we were able to run the linear regressions. Our regressions were based on the DAME-FLAME output.

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:

$$\boxed{annual\ hotel\ revenue\ \sim C(host\_is\_superhost)}$$

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):

$$\boxed{annual\ hotel\ revenue\ \sim C(host\_is\_superhost) + C(state)}$$

Third, we ran a regression that included all confounding variables (except the ones used in revenue calculation) that were part of our analysis after data cleaning, to get an overview of which variables might be relevant for further analysis.

$$\boxed{\begin{array}{l} annual\ hotel\ revenue\ \sim C(host\_is\_superhost) + C(state) + host\_response\_time + host\_response\_rate + host\_acceptance\_rate + \\ C(host\_has\_profile\_pic) + C(host\_identity\_verified) + C(room\_type) + accommodates + bathrooms\_text + bedrooms + beds + \\ C(has\_availability) + number\_of\_reviews\_ltm + review\_scores\_rating + C(instant\_bookable) + caculated\_host\_listings\_count + essentials + \\ C(other\_amenities) + host\_experience \end{array}}$$

IV. SUMMARY STATISTICS

EDA

There were a total of 5,599 super hosts and 6,985 regular hosts in our dataset. The locational distribution of regular hosts and super hosts was similar for each location.

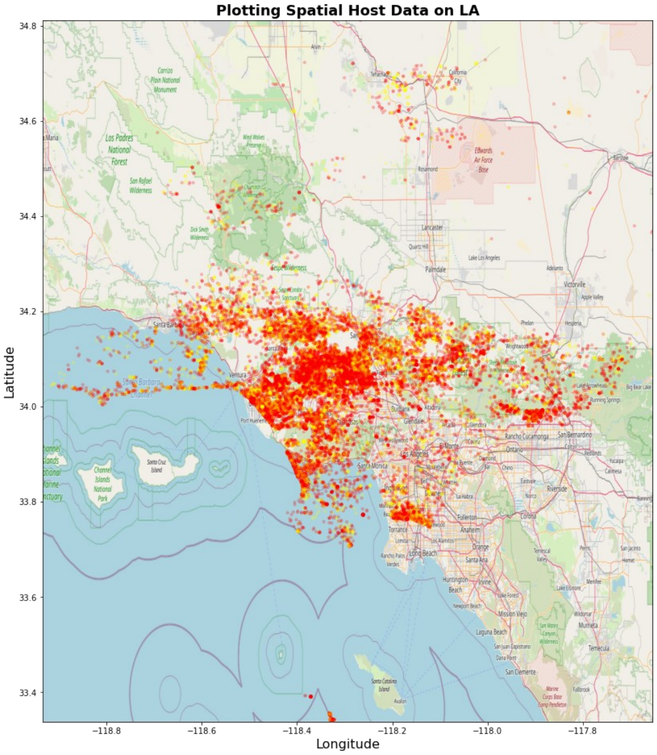


Figure 1: *Spatial distribution of regular hosts (yellow) and super hosts (red) across LA County*

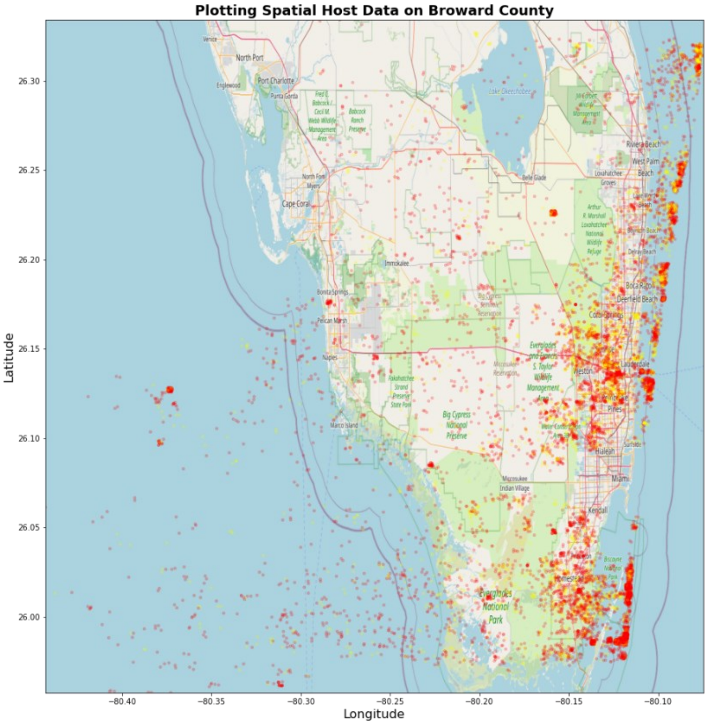
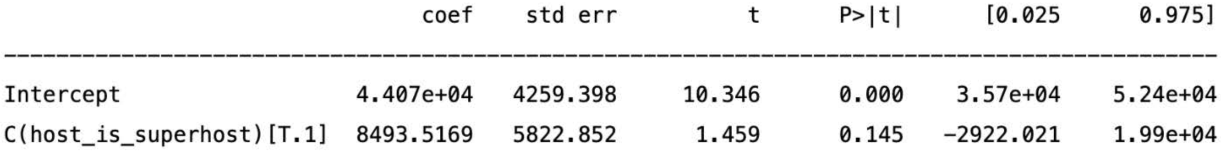


***Figure 2**: Spatial distribution of regular hosts (yellow) and super hosts (red) across Broward County, Florida*

The initial t-test to compare average annual revenue of super hosts and regular hosts yielded a difference of $ 19,293.45 with a t-statistic of 5.46 and a p-value of 0.000, so it looked like there was indeed a difference in average annual revenue with a statistically significant result. However, we wanted to ensure that there is no baseline difference between the two groups for which we performed matching.

Post DAME analysis, while we managed to remove the baseline difference between the two groups – super hosts and regular hosts, we couldn't find the statistically significant results from the first linear regression model with a statistically significant p-value (0.145).

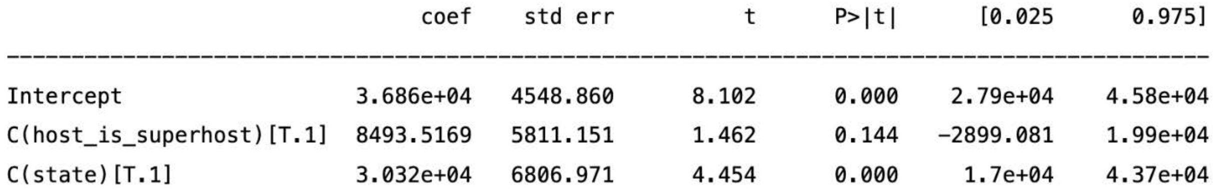
***Table 1**: Results of the regression analysis of the superhost status' influence on estimated annual revenue*



```
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                 4.407e+04   4259.398     10.346      0.000    3.57e+04    5.24e+04
C(host_is_superhost)[T.1] 8493.5169   5822.852      1.459      0.145   -2922.021    1.99e+04
```

Controlling for the state variable, our second regression model provided us the similar estimate on the causal effect as our first regression model, except that we noticed a statistically significant p-value (0.000) for the state variable – suggesting there is indeed a difference between the counties in the two states we examined.

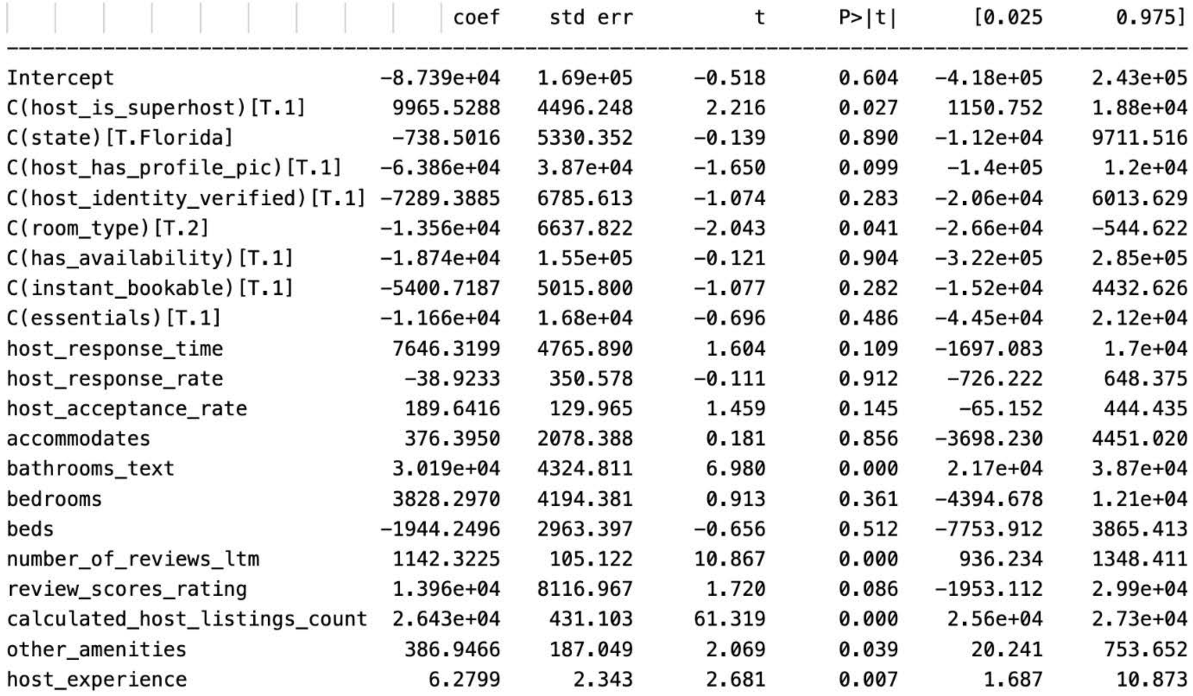
Table 2: *Results of the regression analysis of the superhost status' and the regional factor's (California vs Florida) influence on estimated annual revenue*



```
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                3.686e+04   4548.860      8.102      0.000    2.79e+04    4.58e+04
C(host_is_superhost)[T.1] 8493.5169  5811.151      1.462      0.144   -2899.081    1.99e+04
C(state)[T.1]            3.032e+04   6806.971      4.454      0.000     1.7e+04    4.37e+04
```

Later, when we controlled for all other confounding variables, we still had a statistically significant result (p-value = 0.027) – ***super hosts generate $ 9,965.53 more revenue than regular hosts.*** Other variables that are statistically significant are room type (entire, private, hotel, shared), number of bathrooms, number of reviews, host listings count, 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).

Table 3: *Results of the weighted regression analysis of the variables kept by DAME-FLAME's influence on estimated annual revenue*



```
|    |    |    |    |    |    |    |    |    |coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
Intercept                         -8.739e+04   1.69e+05     -0.518      0.604   -4.18e+05    2.43e+05
C(host_is_superhost)[T.1]          9965.5288   4496.248      2.216      0.027    1150.752    1.88e+04
C(state)[T.Florida]                -738.5016   5330.352     -0.139      0.890   -1.12e+04    9711.516
C(host_has_profile_pic)[T.1]      -6.386e+04   3.87e+04     -1.650      0.099    -1.4e+05     1.2e+04
C(host_identity_verified)[T.1]    -7289.3885   6785.613     -1.074      0.283   -2.06e+04    6013.629
C(room_type)[T.2]                 -1.356e+04   6637.822     -2.043      0.041   -2.66e+04    -544.622
C(has_availability)[T.1]          -1.874e+04   1.55e+05     -0.121      0.904   -3.22e+05    2.85e+05
C(instant_bookable)[T.1]          -5400.7187   5015.800     -1.077      0.282   -1.52e+04    4432.626
C(essentials)[T.1]                -1.166e+04   1.68e+04     -0.696      0.486   -4.45e+04    2.12e+04
host_response_time                 7646.3199   4765.890      1.604      0.109   -1697.083     1.7e+04
host_response_rate                  -38.9233    350.578     -0.111      0.912    -726.222     648.375
host_acceptance_rate                189.6416    129.965      1.459      0.145     -65.152     444.435
accommodates                        376.3950   2078.388      0.181      0.856   -3698.230    4451.020
bathrooms_text                     3.019e+04   4324.811      6.980      0.000    2.17e+04    3.87e+04
bedrooms                           3828.2970   4194.381      0.913      0.361   -4394.678    1.21e+04
beds                              -1944.2496   2963.397     -0.656      0.512   -7753.912    3865.413
number_of_reviews_ltm              1142.3225    105.122     10.867      0.000     936.234    1348.411
review_scores_rating               1.396e+04   8116.967      1.720      0.086   -1953.112    2.99e+04
calculated_host_listings_count     2.643e+04    431.103     61.319      0.000    2.56e+04    2.73e+04
other_amenities                     386.9466    187.049      2.069      0.039      20.241     753.652
host_experience                       6.2799      2.343      2.681      0.007       1.687      10.873
```

Causal Effect

The average difference in annual revenue between hosts who we actually observe as super hosts and hosts who we actually observe as regular hosts in a world whether neither is superhost is $ 9965.53.

$$E(Y_{T=1}|D=1) - E(Y_{T=0}|D=0) = \$ \ 9965.53$$

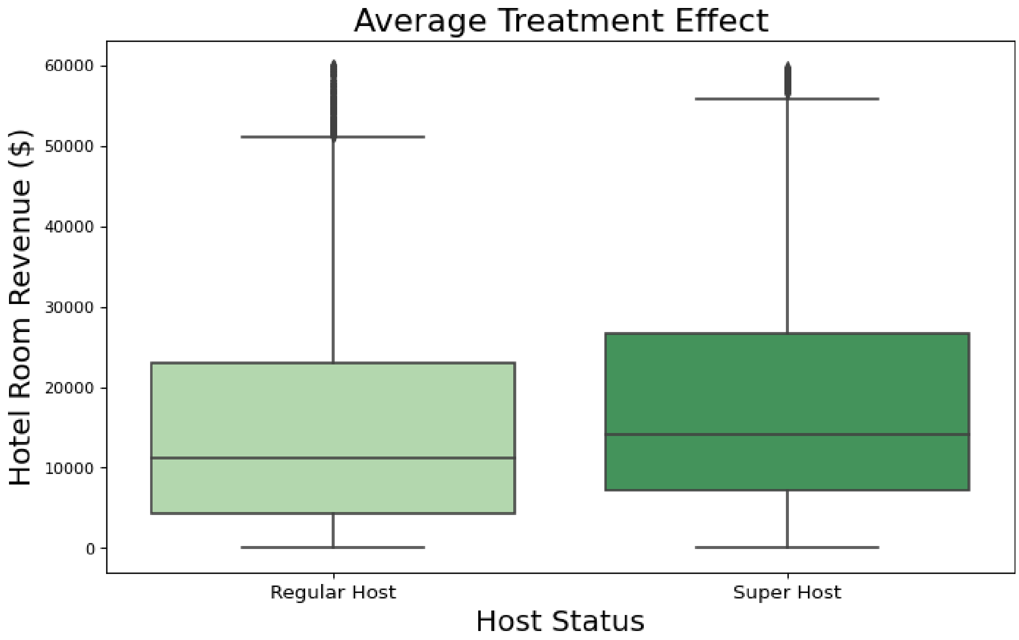


Figure 3: *Shows the boxplots of the hosts non-superhost (regular host) or superhost status in relation to the annual revenue. Visual inspection indeed shows a difference in annual revenue when accounting for all other influencing factors (e.g., number of listings per host, number of beds in residence).*

Conditional Treatment Effect

We also calculated treatment effect across every groups created by DAME FLAME. From Figure 4, however, it seems that there is not much can we infer about the conditional effect of "superhosts" on annual revenue from the dataset. Hence, we cannot conclude if there exists any treatment groups for which the impact of "superhosts" on annual revenue is different than the outcomes for other groups.

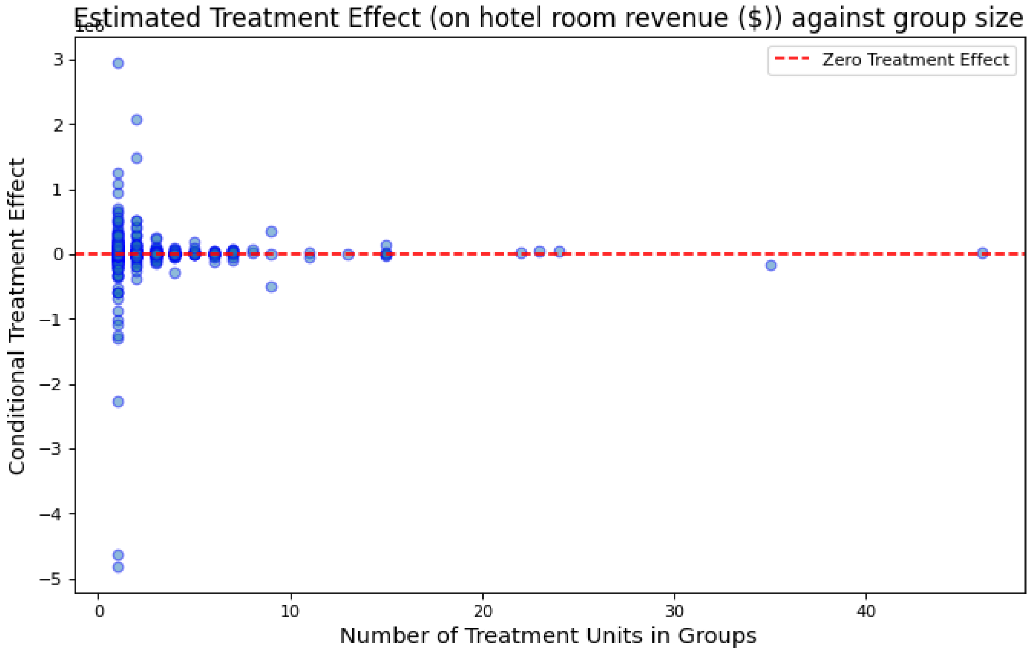


Figure 4: Output from DAME suggests that impact of "superhosts" on annual revenue is different than the outcomes for other groups.

V. CONCLUSION

In conclusion, we found that being a superhost does correlate with generating more estimated annual revenue (*$9,965.53)* than being a non-superhost to the level of statistical significance. Other variables that we found to influence annual revenue are room type (entire residences create more revenue), number of bathrooms (more bathrooms create more revenue), number of reviews in the last month (more reviews create more revenue), host listings count (more listings create more revenue), and host experience (more experience creates more revenue).

None of these correlate with the factors that influence whether a host is considered a superhost based on what Airbnb communicates.

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 higher than this $4,000 difference.

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